A Swiss Army Knife for Minimax Optimal Transport

Sofien Dhouib [1] **Ievgen Redko** [2] **Tanguy Kerdoncuff** [2] **Rémi Emonet** [2] **Marc Sebban** [2]

Abstract

The Optimal transport (OT) problem and its associated Wasserstein distance have recently become a topic of great interest in the machine learning community. However, its underlying optimization problem is known to have two major restrictions: (i) it strongly depends on the choice of the cost function and (ii) its sample complexity scales exponentially with the dimension. In this paper, we propose a general formulation of a minimax OT problem that can tackle these limitations by jointly optimizing the cost matrix and the transport plan, allowing us to define a robust distance between distributions. We propose to use a cutting-set method to solve this general problem and show its links and advantages compared to other existing minimax OT approaches. Additionally, we use this method to define a notion of stability allowing us to select the ground metric robust to bounded perturbations. Finally, we provide an experimental study highlighting the efficiency of our approach.

1. Introduction

In many scientific areas, we are often confronted with a necessity of comparing different objects to assess their relatedness. In machine learning, for instance, these objects may be individual data points in similarity based classification (e.g., k-nearest neighbors (Cover & Hart, 2006), non-linear support vector machines (Boser et al., 1992)) or probability distributions in generative modelling (Goodfellow et al., 2014) and hypothesis testing. For this latter case, the optimal transportation (OT) metric (also called the Wasserstein distance) has recently emerged as a powerful tool used to compare complex objects based on the OT problem (Monge, 1781) that roughly quantifies the minimal amount of effort required to transform one distribution into another. Several key features of this metric lead to its widespread use in many different applications and setups (Gramfort et al., 2015; Kusner et al., 2015; Bonneel et al., 2016; Courty et al., 2017; Laclau et al., 2017). First, it takes into account the geometry of the underlying data distributions by the means of pairwise costs calculated for the points that they are supported on. Second, it allows to compare distributions with disjoint supports thus avoiding the vanishing gradient problem (Arjovsky et al., 2017) when used as a loss function.

In this paper, we study a general formulation of the OT problem with a minimax objective function where one seeks an OT plan with respect to (w.r.t.) the worst possible ground metric (also called cost function) belonging to an arbitrary and possibly infinite convex set. Such a minimax formulation is of a particular interest as it has been shown previously (i) to reduce the sample complexity and increase the robustness to noise of the original OT problem for high-dimensional data (Paty & Cuturi, 2019), (ii) to allow to consider submodular cost functions (Alvarez-Melis et al., 2018) and (iii) to use it as a loss in generative models (Genevay et al., 2018). We advance the study of the minimax OT further by providing the following contributions. First, for an infinite set of cost functions defined by a Mahalanobis distance, we reformulate the minimax OT problem as a minimization of the arbitrary dual norm of the matrix of second-order displacements and show how one can use it to smoothly interpolate between the original OT problem and the minimax formulation of (Paty & Cuturi, 2019) included as a special case. Second, we provide a generic solver for minimax OT for both regularized and unregularized minimax OT problems and for both finite and infinite families of cost functions contrary to previous work (Paty & Cuturi, 2019; Alvarez-Melis et al., 2018) that considered the differentiable and strictly convex regularized OT problem only. Finally, we introduce the notion of cost matrix stability and solve its underlying optimization problem. It consists in finding a cost function from a list of possible candidates that leads to a stable transportation cost in its unit ball neighborhood.

The rest of this paper is organized as follows. In Section 2, we provide the necessary introductory definitions related

[1] Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69100, LYON, France [2] Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France. Correspondence to: Sofien Dhouib <sofiane.dhouib@creatis.insa-lyon.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

to the OT problem and the notations used throughout the paper. We then present in Section 3 the main contributions of this paper including a general minimax formulation for the OT problem with an arbitrary convex compact set of cost matrices and discuss an optimization procedure that can be used to solve it as well as its theoretical guarantees. We further proceed by considering the important special cases of the previously introduced problem and showing their relationship to other works on subject. Finally, in Section 4, we present an experimental evaluation of our approach for several considered use-cases.

2. Preliminary Knowledge

Optimal transport Optimal transport (OT) can be seen as the search for a transportation plan that moves (transports) a probability measure μ_1 onto another measure μ_2 with a minimum cost measured by some function $c : (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} \mapsto c(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+$, where $\mathcal{X}$ and $\mathcal{Y}$ are some complete metric spaces that, in most applications, are taken to be Euclidean spaces. More formally, the Kantorovich (Kantorovich, 1942) formulation of OT seeks for an optimal coupling γ having marginals μ_1 and μ_2, which minimizes the following quantity:

$$W_c(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1,\mu_2)} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\gamma} [c(\mathbf{x}, \mathbf{y})] \,, \qquad (1)$$

where $c(\mathbf{x}, \mathbf{y})$ is the cost of moving $\mathbf{x}$ to $\mathbf{y}$ (drawn from distributions μ_1 and μ_2, respectively). When c is the squared Euclidean distance, we write W_2. In the discrete version of the problem, *i.e.* when μ_1 and μ_2 are defined as empirical measures supported on vectors $\{\mathbf{x}_i\}_{i=1}^m$, $\{\mathbf{y}_j\}_{j=1}^n$ in $\mathbb{R}^d$ with probability vectors $\mathbf{r} \in \Delta_m$ and $\mathbf{c} \in \Delta_n$, the previous problem can be expressed as follows:

$$\mathbf{P}^* \in \operatorname*{argmin}_{\mathbf{P}\in\Pi(\mathbf{r},\mathbf{c})} \langle \mathbf{P}, \mathbf{C} \rangle_F, \qquad (2)$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product, $\mathbf{C} \in \mathbb{R}_+^{m\times n}$ is a cost matrix representing the pairwise costs of transporting $\mathbf{x}_i$ to $\mathbf{y}_j$ and $\mathbf{P}$ is a matrix of size $m \times n$ belonging to the transportation polytope $\Pi(\mathbf{r}, \mathbf{c})$ (also called Birkhoff polytope for $m = n$) defined as $\Pi(\mathbf{r}, \mathbf{c}) = \{\mathbf{P} \in \mathbb{R}_+^{m\times n}; \mathbf{P}\mathbf{1}_n = \mathbf{c}; \mathbf{P}^T\mathbf{1}_m = \mathbf{r}\}$. Note that (2) is a linear programming (LP) problem with equality constraints, but its dimensions scale quadratically with the size of the sample. Alternatively, one can consider a regularized version of the problem (Cuturi, 2013), which has the extra benefit of being faster to solve.

Minimax OT Two other studies considered the minimax formulation of the OT problem in a setting similar to ours. In the first one, (Paty & Cuturi, 2019) showed that one can see the OT problem, with c taken to be the squared Euclidean distance, as a trace minimization problem of the second-order displacement matrix defined for any $\gamma \in \Pi(\mu_1, \mu_2)$ defined as:

$$W_2^2(\mu_1, \mu_2) = \min_{\gamma\in\Pi(\mu_1,\mu_2)} \operatorname{Tr}(V_\gamma),$$

$$\mathbf{V}_\gamma := \int_{\mathcal{X}\times\mathcal{Y}} (\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T d\gamma(\mathbf{x}, \mathbf{y}).$$

The authors of (Paty & Cuturi, 2019) further used this expression of the 2-Wasserstein distance to introduce the Subspace Robust Wasserstein (SRW) distance as follows:

$$\mathcal{S}_k^2(\mu_1, \mu_2) := \min_{\gamma\in\Pi} \operatorname{Tr}^k(\mathbf{V}_\gamma) = \min_{\gamma\in\Pi} \sum_{i=1}^k \lambda_i(\mathbf{V}_\gamma), \qquad (3)$$

where λ_i are the $k \le d$ largest eigenvalues of $\mathbf{V}_\gamma$. Note that considering only the maximization over the $k \le d$ largest eigenvalues allows to learn a cost matrix of a reduced rank thus tackling the curse of dimensionality issue of calculating the Wasserstein distance for high-dimensional data.

A somehow different way of using the minimax formulation of OT was proposed in (Alvarez-Melis et al., 2018) for c taken to be a submodular function $F : 2^V \to \mathbb{R}$ with V denoting a certain set of available items. In this case, taking the Lovász extension f of F leads to the following optimization problem:

$$\text{StrOT}(\mu_1, \mu_2) := \min_{\mathbf{P}\in\Pi} \max_{\mathbf{C}\in\mathcal{B}_F} \langle \mathbf{P}, \mathbf{C} \rangle \,,$$

where $\mathcal{B}_F$ is the base polytope of F defined as $\mathcal{B}_F = \{y \in \mathbb{R}^{|V|} | y(V) = F(V); y(S) \le F(S), \forall S \subseteq V\}$. A game-theoretic interpretation of this formulation is to consider two players, where Player 1 aims at aligning the two distributions by picking a coupling matrix $\mathbf{P}$, while Player 2 resists to it by choosing the cost matrix $\mathbf{C}$ from the set of admissible costs $\mathcal{B}_F$. When F is a modular function, the size of $\mathcal{B}_F$ is 1 thus recovering the original OT problem.

Other related work Three other papers presented an OT-based minimax formulation distantly related to ours. In (Genevay et al., 2018), the authors studied a generative model that uses Sinkhorn divergence as a fitting criterion and proposed to learn a cost function in this framework. Their problem is intrinsically different from ours as we do not consider the density fitting problem where one optimises the parameters of the fitted distribution. On the other hand, in (Li et al., 2019), the authors reduced the regularized OT formulation with relaxed marginal constraints into a minimax problem. Their formulation, however, is also different from ours as it does not seek to learn a cost matrix. Finally, the line of work on the Wasserstein distributionally robust optimization (Kuhn et al., 2019) is also very dissimilar to our paper as this latter considers finding the best estimator of a density from a Wasserstein ball of a certain radius.

We now proceed to the presentation of our contributions.

3. Robust Optimal Transport with a Convex Set of Cost Matrices

Below, we formulate the general robust OT problem and highlight its properties in several cases of interest. We further propose and theoretically analyze a general algorithm that can be used to solve it.

3.1. Problem formulation

Let $\mathcal{C}$ be an arbitrary set of cost functions defined over $\mathcal{X}\times\mathcal{Y}$. This set may represent, for instance, a convex combination of cost function candidates provided by several experts, or it can be described by an infinite set of parameters. We impose no particular constraints on the cost functions belonging to $\mathcal{C}$ as long as the corresponding Kantorovich problems admit a solution. We now consider the following minimax problem:

$$\text{RKP}(\Pi, \mathcal{C}) = \min_{\gamma\in\Pi} \max_{c\in\mathcal{C}} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y}\sim\gamma} \left[c(\mathbf{x},\mathbf{y})\right], \tag{4}$$

where we look for a coupling γ^* that is robust to the choice of a cost function $c \in \mathcal{C}$, by considering the worst achievable transportation cost. We denote the value at the solution of this problem by $\text{RKP}(\Pi, \mathcal{C})$ where RKP stands for robust Kantorovich problem. We abuse the notation and use $\text{RKP}(\mathcal{P}, \mathcal{C})$ for any set $\mathcal{P} \subset \Pi$ (even non convex) to denote $\text{RKP}(\text{conv}(\mathcal{P}), \mathcal{C})$, i.e., solving for $\gamma \in \text{Conv}\,(\mathcal{P})$. We also extend the notation W_c, presented before by defining $W_{\mathcal{C}}(\mu_1, \mu_2) := \text{RKP}(\Pi, \mathcal{C})$.

3.2. Choice of $\mathcal{C}$

Below, we consider two possible choices for the convex set $\mathcal{C}$. First, we study the infinite family of Mahalanobis distance cost matrices widely used in the metric learning literature (Bellet et al., 2015). Second, we consider a convex hull of a finite family of cost functions as in the example given above.

3.2.1. Infinite family of Mahalanobis distances

For any $\mathbf{u} = (u_1, ..., u_d) \in \mathbb{R}^d$ and any $\mathbf{M} \in \mathbb{R}^{d\times d}$, we define their respective p-norm and Schatten p-norm as

$$\|\mathbf{u}\|_p^p = \sum_{1\leq i\leq d} |u_i|^p, \ \|\mathbf{M}\|_p^p = \sum_{1\leq i\leq d} \sigma_i^p(\mathbf{M}),$$

where $p \in [1, +\infty]$ and $\{\sigma_i(\mathbf{M})\}$ are $\mathbf{M}$'s singular values. In particular, if $\mathbf{M} \in \mathcal{S}_+^d(\mathbb{R})$, where $\mathcal{S}_+^d(\mathbb{R})$ denotes the set of symmetric positive semi-definite matrices (PSD), then $\|\mathbf{M}\|_p = \text{Tr}(\mathbf{M}^p)^{\frac{1}{p}}$. We also recall that the dual of a $p-$norm (resp. Schatten $p-$norm) is the $q-$norm (resp. the Schatten $q-$norm) with q equal to $\frac{p}{p-1}$ if $p > 1$, to ∞ if $p = 1$ and to 1 if $p = \infty$.

We now define $\mathcal{C}$ as a family of Mahalanobis cost functions, indexed by bounded matrices $\mathbf{M}$:

$$\mathcal{C} = \{c^{\mathbf{M}} : (\mathbf{x},\mathbf{y}) \mapsto (\mathbf{x}-\mathbf{y})^T\mathbf{M}(\mathbf{x}-\mathbf{y}); \|\mathbf{M}\|_p \leq 1\}. \tag{5}$$

We can now state the following proposition.[1]

Proposition 1. *Let $\mathcal{C}$ be defined as in (5) for $\mathbf{M} \in \mathcal{S}_+^d(\mathbb{R})$. Then, $\mathcal{C}$ is a convex compact set of cost functions and for any $p \in [1, +\infty]$, $q = \frac{p}{p-1}$ the following holds:*

1. $\text{RKP}(\Pi, \mathcal{C}) = \min_{\gamma\in\Pi} \|\mathbf{V}_\gamma\|_q$. *In particular, we have:*

$$\text{RKP}(\Pi, \mathcal{C}) = \begin{cases} W_2^2(\mu_1, \mu_2), & \text{if } q = 1, \\ \mathcal{S}_1^2(\mu_1, \mu_2), & \text{if } q = \infty. \end{cases}$$

2. *For any $\gamma \in \Pi$, $\|\mathbf{M}^*\|_p = 1$ and*

$$\mathbf{M}^* = \mathop{\text{argmax}}_{\mathbf{M}\in\mathcal{S}_+^d,\ \|\mathbf{M}\|_p\leq 1} \langle \mathbf{V}_\gamma, \mathbf{M}\rangle = \left(\frac{\mathbf{V}_\gamma}{\|\mathbf{V}_\gamma\|_q}\right)^{\frac{q}{p}}.$$

Proof idea. We use the fact that $\mathcal{C}$ is the image of a convex compact set of $\mathbb{R}^{d\times d}$ by a linear mapping to prove its convexity and compactness. Point 1 is a consequence of the equality case of Hölder's inequality, the positive semi-definiteness of matrix $\mathbf{V}_\gamma$ and the fact that the Schatten p-norm is the classic p-norm for the vector of a matrix's singular values, which tends to the $\infty-$norm as $q \to \infty$. The second point is a direct consequence of the equality case of Hölder's inequality for Schatten p-norms (Magnus, 1987) using the fact that $\mathbf{V}_\gamma$ is PSD. □

This theorem highlights several novel insights. First, it provides a different point of view for a general minimax OT problem with the infinite family of Mahalanobis distances. In particular, it shows that the original OT problem can be seen as a minimax problem when one takes the least restrictive infinity norm for the bound on the matrix parameterizing the Mahalanobis distance, while SRW with $k = 1$ corresponds to the case of the $||\cdot||_1$ norm[2]. This observation is illustrated in Figure 1 where we smoothly interpolate between the two boundary cases by solving (4) with intermediate values of q. We note that such an interpolation may have interesting implications in practice when one seeks for an explicit control between the original and the minimax OT problems. Second, the optimal expression for $\mathbf{M}^*$ shows that it is proportional to $\mathbf{V}_\gamma$ and if this latter captures the displacement in lower dimensions, then $\mathbf{M}^*$ is expected to do so too. This follows from $\mathbf{M}^*$ being a linear combination of $(\mathbf{x}_i - \mathbf{y}_j)(\mathbf{x}_i - \mathbf{y}_j)^T$, where $\{i, j\}$ are indices for which $\gamma_{ij} > 0$, making its image included in the span of

[1] All detailed proofs are provided in the supplementary material.

[2] Other values of k for SRW are also covered when using a truncated Schatten p-norm.

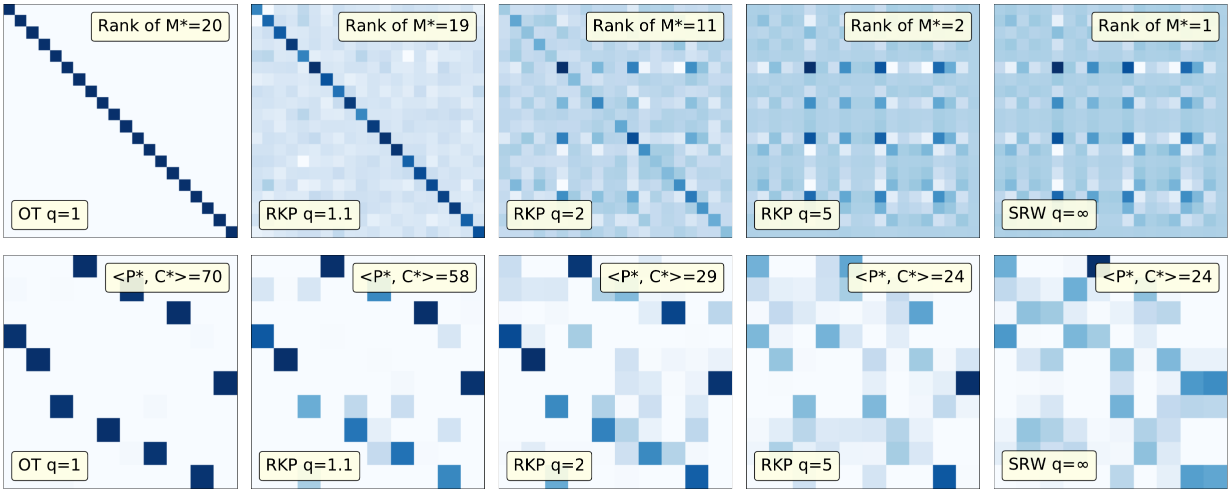

Figure 1. Interpolation between OT and SRW$_{k=1}$ on a binary toy classification problem with each class consisting of 5 points sampled from Gaussians centered on the edge of a 10-dimensional hypercube with $\sigma = 1$ with 10 additional random noise features. The transport is computed between the 2 classes using the setting of Proposition 1 with $q \in \{1, 1.1, 2, 5, \infty\}$. **(top row)** Mahalanobis matrices $\mathbf{M}^*$ and their rank; (bottom row) Couplings $\mathbf{P}^*$ and the associated value of the Wasserstein distance. Note that for this visualization we use a Frank Wolfe algorithm presented in the Supplementary material.

$\{(\mathbf{x}_i - \mathbf{y}_j); \gamma_{ij} > 0\}$, i.e., the span of displacement directions. This intuition is confirmed in our experiments where we show that even without the rank constraint, solving (4) results in a matrix of a reduced rank.

Finally, below we use this result to show that in the case of the Frobenius norm, the PSD property of the learned matrix $\mathbf{M}$ is obtained for free without imposing any additional constraint on the set $\mathcal{C}$.

Corollary 1 (Euclidean norm case)**.** *Let $\mathcal{C}$ be defined with $p = 2$ in (5) and let $\mathbf{M}^* = \operatorname{argmax}_{\|\mathbf{M}\|_2 \leq 1} \langle \mathbf{V}_\gamma, \mathbf{M}\rangle$. Then $\mathbf{M}^* = \dfrac{\mathbf{V}_\gamma}{\|\mathbf{V}_\gamma\|_2}$, thus $\mathbf{M}^*$ is PSD and $\|\mathbf{M}^*\|_2 = 1$.*

This corollary shows that the case $p = 2$ (Frobenius norm) can be very convenient in practice as PSD constraints increase considerably the computational burden of any optimization problem, yet they are necessary for the obtained cost function to be a true metric.

To conclude the theoretical analysis of the considered case for the minimax problem, we establish a general bound on $\mathrm{RKP}(\Pi, \mathcal{C})$ in terms of the original 2-Wasserstein distance.

Corollary 2. *With the assumptions from Proposition 1, the following inequality holds for any $p \in [1, +\infty]$:*

$$\frac{1}{d^{\frac{1}{p}}} W_2^2(\mu_1, \mu_2) \leq W_{\mathcal{C}}(\mu_1, \mu_2) \leq W_2^2(\mu_1, \mu_2).$$

Note that compared to a similar bound given in (Paty & Cuturi, 2019, Proposition 2) for the SRW distance, our result does not involve the k term in the left-hand side as we do not impose any explicit constraint on the rank of $\mathbf{M}$.

3.2.2. Finite set of cost functions

Let $\{c_1, ..., c_K\}$ denote a family of candidate cost functions, and let $\mathcal{C} = \mathrm{Conv}\left(\{c_1, ..., c_K\}\right)$ meaning that $\mathcal{C}$ is a convex compact space as it is the convex combination of a finite set. As mentioned in Section 2, the optimization of the OT problem with a submodular function F taken as a cost function can be equivalently seen as a minimax OT problem of the following form:

$$\min_{\mathbf{P}\in\Pi} \max_{\mathbf{C}\in\mathcal{B}_F} \langle \mathbf{P}, \mathbf{C}\rangle,$$

where $\mathcal{B}_F$ is the base polytope of F. We note that the number of vertices of $\mathcal{B}_F$ is finite and thus one can show that the StrOT distance is a particular case of our problem (4) when $\mathcal{C}$ is a finite set of cost functions, i.e.

$$\mathrm{RKP}(\Pi, \mathrm{Conv}\left(\mathcal{B}_F\right)) = \mathrm{StrOT}(\mu_1, \mu_2).$$

This result establishes the link between our general formulation and that considered in (Alvarez-Melis et al., 2018).

3.3. Proposed optimization strategy

We now propose a general solution for optimising (4) in the discrete case where $\mathcal{X}$ and $\mathcal{Y}$ are identified respectively with finite sets $\{\mathbf{x}_i\}_{i=1}^m$ and $\{\mathbf{y}_j\}_{j=1}^n$, while $\mathcal{C}$ is identified with an arbitrary convex set of cost matrices with entries $\mathbf{C}_{ij} = c(\mathbf{x}_i, \mathbf{y}_j)$. Since $\mathcal{X}$ and $\mathcal{Y}$ are finite, hence bounded, all results from Section 3.2 hold in the discrete case.

To proceed, we first note that in our case we cannot apply the optimization techniques used in (Paty & Cuturi, 2019; Alvarez-Melis et al., 2018) as they both consider the differentiable regularized OT problem in their minimax formulations contrary to our non-differentiable unregularized one. To deal with the latter, we propose to adapt the cutting set method presented in (Mutapcic & Boyd, 2009) for robust optimization to Problem (4) that allows us to cover both unregularized and regularized minimax OT problems. In a nutshell, this method consists in alternating between solving a worst-case problem and the corresponding sampled robust minimization problem w.r.t. a set of constraints that grows linearly with iterations and requires for optimized functions to be convex only. In application to (4), the high level idea of the proposed algorithm thus would be to solve the maximization problem over $\mathcal{C}$ w.r.t. a set $\mathcal{P} \subset \Pi$ and add one transportation matrix to $\mathcal{P}$ at each iteration. The implementation of this idea, however, is not straightforward and requires two obstacles to be addressed. First, the original algorithm presented by the authors allows to solve a minimax problem of the form $\min_{\mathbf{C}\in\mathcal{C}} \max_{\mathbf{P}\in\Pi} = -\max_{\mathbf{C}\in\mathcal{C}} \min_{\mathbf{P}\in\Pi}$ and thus requires from us to prove $\min_{\mathbf{P}\in\Pi} \max_{\mathbf{C}\in\mathcal{C}} = \max_{\mathbf{C}\in\mathcal{C}} \min_{\mathbf{P}\in\Pi}$ in order to apply it. Second, and similar to the projected supergradient algorithm proposed for SRW, the authors of (Mutapcic & Boyd, 2009) disregard the optimal solution for the variable over which the minimization is performed, i.e., $\mathbf{P}^*$ in our case, and provide a solution for $\mathbf{C}^*$ only. To address these issues, we now present the following result.

Proposition 2. *Let $\mathcal{P}$ be a finite subset of Π. Then, the following holds:*

1. $\mathrm{RKP}(\mathcal{P},\mathcal{C}) := \mathrm{RKP}(\mathrm{Conv}(\mathcal{P}),\mathcal{C})$ *has a saddle point* $(\mathbf{P}^*,\mathbf{C}^*)$ *verifying:*

$$\langle \mathbf{P}^*, \mathbf{C}^* \rangle_F = \min_{\mathbf{P}\in\mathrm{Conv}(\mathcal{P})} \max_{\mathbf{C}\in\mathcal{C}} \langle \mathbf{P}, \mathbf{C} \rangle = \max_{\mathbf{C}\in\mathcal{C}} \min_{\mathbf{P}\in\mathcal{P}} \langle \mathbf{P}, \mathbf{C} \rangle . \quad (6)$$

2. $\mathrm{RKP}(\mathcal{P},\mathcal{C})$ *is equivalent to*

$$\begin{aligned} &\mathbf{C}^* \in \mathrm{argmax}_{\mathbf{C}\in\mathcal{C},\mu\geq 0}\ \mu, \\ &\textit{s.t.}\ \langle \mathbf{P}, \mathbf{C} \rangle \geq\ \mu,\ \forall \mathbf{P} \in \mathcal{P}. \end{aligned} \quad (7)$$

3. $\mathbf{P}^* = \sum_{l=1}^{|\mathcal{P}|} q_l \mathbf{P}_l$, *where* $\mathbf{Q} = \{q_l\}_{l=1}^{|\mathcal{P}|}, \sum_i q_i = 1$, *are dual variables of* (7).

Proof idea. Point 1 is an application of Sion's minimax theorem (Sion, 1958). Point 2 is a reformulation of the right hand side of Equation (6). The last point, $\mathbf{P}^*$'s expression, is a consequence of the Lagrange duality. □

Several remarks are in order here. First, we note that solving Problem (7) directly is intractable in practice for a sufficiently large n as its number of constraints (size of $\mathcal{P}$) grows extremely fast with the number of points (e.g., equal to $n!$ for $m = n$). This motivates the use of the cutting plane algorithm that gradually increases the size of the set $\mathcal{P}$ with iterations and allows to solve intermediate problems with a reduced number of constraints efficiently. Second, the theorem is valid for any finite subset $\mathcal{P}$ of Π so that 1) solving $\mathrm{RKP}(\Pi,\mathcal{C})$ can be done by setting $\mathcal{P}$ to the set of vertices of Π and 2) solving the regularized minimax formulation with added convex regularizer on $\mathbf{P}$ is covered by considering $\mathrm{RKP}(\tilde{\Pi},\mathcal{C})$, where $\tilde{\Pi}$ is a convex compact subset of Π.

Algorithm 1 Cutting set method for $\mathrm{RKP}(\Pi,\mathcal{C})$ with constraint elimination

```
Input: maxIt, $\mathcal{C}$, $\mathcal{P}_0 \subset \Pi$, thd1, thd2
$t, l \leftarrow 0$
$err, \mu_{-1} \leftarrow \infty$
while $t <$ maxIt and $err >$ thd1 and $\frac{\mu_{t-1}-\mu_t}{\mu_{t-1}} >$ thd1$^2$ do
    Solve (7) to obtain $(\mu_t, \mathbf{C}_t), \mathbf{Q}$
    for $l$ in $\{0, ..., |\mathcal{P}_t| - 1\}$ do
        if $q_l \leq$ thd2 then
            $\mathcal{P}_t \leftarrow \mathcal{P}_t \setminus \{\mathbf{P}_l\}$
            $\mathbf{Q} \leftarrow \mathbf{Q} \setminus \{\mathbf{q}_l\}$
    Find $\mathbf{P}_t \in \mathrm{argmin}_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \mathbf{C}_t \rangle$
    $l \leftarrow \max(l, \langle \mathbf{P}_t, \mathbf{C}_t \rangle)$
    $err \leftarrow (\mu_t - l)/l$
    $\mathcal{P}_{t+1} = \mathcal{P}_t \cup \{\mathbf{P}_t\}$
    $t \leftarrow t + 1$
return $\sum_{l=0}^{|\mathcal{P}_t|-1} q_l \mathbf{P}_l$, $\mathbf{C}_t$
```

Our final algorithm inspired by (Mutapcic & Boyd, 2009, Section 5.1) then boils down to alternately performing the following two steps for $t \in \{0, \ldots, \mathrm{maxIt}\}$:

Step 1. Find $\mathbf{C}_t$ solving (7) over $(\mathcal{P}_t, \mathcal{C})$, where $\mathcal{P}_t$ is a finite subset of Π; let μ_t be the value at the solution.

Step 2. For a fixed matrix $\mathbf{C}_t$ obtained at **Step 1**, find $\mathbf{P}_t \in \mathrm{argmin}_{\mathbf{P}\in\Pi} \langle \mathbf{P}, \mathbf{C}_t \rangle$.

Step 2 of each iteration can make use of any efficient algorithm for solving the classic unregularized optimal transport. Empirically, we observed that even the approximate solutions obtained by solving the entropy regularized formulation of the optimal transport problem ensure the convergence. We further use the constraint dropping strategy (Mutapcic & Boyd, 2009, Sec. 5.3.2) and provide a complete pseudo-code for our algorithm in Algorithm 1, where thd1 and thd2 respectively control the stopping criterion and the constraint elimination. The proposed algorithm is generic and can also be used to solve the problems underlying the SRW and StrOT distances seen previously. Moreover, it acts as a meta-algorithm by implicitly choosing (or learning depending on the construction of the set $\mathcal{C}$) the "right" cost function. This differs from other existing methods on learning the cost matrix in the OT framework (Cuturi & Avis, 2014; Zhao & Zhou, 2018) that usually learn this latter using the *a priori* similarity between the histograms.

Finally, Algorithm 1 is guaranteed to converge in a finite number of iterations with the latter being upper-bounded thanks to the following proposition.

Proposition 3. *Let T be the number of iterations required by Algorithm 1 to reach error $err(T) \leq \mathrm{thd1}$. Then,*

$$T \leq \left(\frac{\mathrm{diam}_\infty(\mathcal{C}) + \mathrm{RKP}(\mathcal{P}_0, \mathcal{C})}{2.\mathrm{thd1}} + 1 \right)^{\dim(\mathcal{C})+1}$$

where $\mathrm{diam}_\infty(\mathcal{C}) := \sup_{\mathbf{C}^1, \mathbf{C}^2 \in \mathcal{C}, i, j} \left| \mathbf{C}^1_{ij} - \mathbf{C}^2_{ij} \right|$ *and* $\dim(\mathcal{C})$ *is the dimension of the affine hull of $\mathcal{C}$. Also, $\forall t \geq 0$, we have that $0 \leq \mathrm{RKP}(\mathcal{P}_t, \mathcal{C}) - \mathrm{RKP}(\Pi, \mathcal{C}) \leq err(t)$.*

Proof idea. We adapt the proof technique presented in (Mutapcic & Boyd, 2009, Section 5.2) to our case, after rewriting the right hand side of Equation (6) as as

$$\min_{\mathbf{C} \in \mathcal{C}} \max_{\mathbf{P} \in \mathcal{P}_t} \left(- \langle \mathbf{P}, \mathbf{C} \rangle_F \right)$$

to make our problem coincide with the authors' formulation. □

This theorem offers interesting insights regarding the convergence speed of the proposed algorithm. First, it introduces the dependence of the latter on $\mathrm{diam}_\infty(\mathcal{C})$, which can be interpreted as a degree of disagreement between the cost matrices in $\mathcal{C}$ so that one may need more iterations to reach precision err when they disagree. Second, the presence of the value of the initial nominal problem $\mathrm{RKP}(\mathcal{P}_0, \mathcal{C})$ reflects the influence of the initialization $\mathcal{P}_0$. Finally, when $\mathcal{C}$ lies in a subspace of a much smaller dimension than $m \times n$ (i.e., in case of the Mahalanobis distance, $\mathcal{C}$ is the image of $d \times d$ matrices by a linear mapping, while for the finite number of matrices, $\dim(\mathcal{C})$ is $\dim(\mathrm{span}(\mathbf{C}_1, ..., \mathbf{C}_n)) - 1)$, the algorithm needs much less iterations as highlighted by the presence of $\dim(\mathcal{C})$ in the exponent.

3.4. Variations for different choices of $\mathcal{C}$

Below, we express the maximization problem (7) over $\mathcal{P}_t \times \mathcal{C}$ at step $t \geq 0$ of Algorithm 1, for both choices of $\mathcal{C}$ considered in Section 3.2, in a more convenient way.

Proposition 4 (Finite set $\mathcal{C}$)**.** *Let $\mathcal{C} = \mathrm{Conv}\left(\{\mathbf{C}_1, ..., \mathbf{C}_d\}\right)$. Then, for $t \geq 0$, solving the problem given in (7) over $\mathcal{P}_t \times \mathcal{C}$ is equivalent to the following linear program*

$$\begin{aligned} &\min_{\mathbf{p} \in \mathbb{R}^d_+} \mathbf{1}_d^T \mathbf{p} \\ &s.t. \mathbf{G}\mathbf{p} \geq \mathbf{1}_{|\mathcal{P}_t|}, \end{aligned} \tag{8}$$

where $\mathbf{G} \in \mathbb{R}^{|\mathcal{P}_t| \times d}$ with $\mathbf{G}_{kl} = \langle \mathbf{P}_k, \mathbf{C}_l \rangle$. Moreover,

$$\mathbf{C}^* = \frac{\sum_{k=1}^d p_k^* \mathbf{C}_k}{\sum_{k=1}^d p_k^*}, \qquad \mathbf{P}^* = \frac{\sum_l^{|\mathcal{P}_t|} q_l^* \mathbf{P}_l}{\sum_l^{|\mathcal{P}_t|} q_l^*},$$

where $\mathbf{p}^*$ and $\mathbf{q}^*$ are optimal solutions of (8) and its dual.

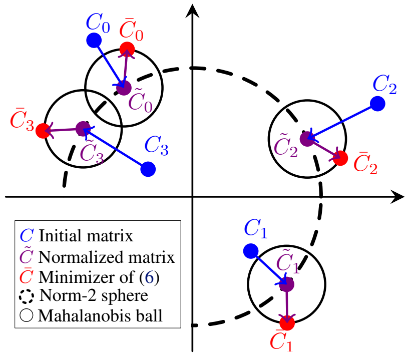

Figure 2. Illustration of the notion of matrix cost stability. Every matrix $\mathbf{C}_i$ is normalized so as to get a matrix $\tilde{\mathbf{C}}_i$ which lies on the norm-2 sphere. $\bar{\mathbf{C}}_i$ is the minimizer of Problem 6. The stability (Definition 1) comes from the difference of the cost transports induced by $\bar{\mathbf{C}}_i$ and $\tilde{\mathbf{C}}_i$.

For the case of the infinite family of Mahalanobis distances, we propose a more general result that considers the following set of non-centered Mahalanobis distances:

$$\begin{aligned} \mathcal{C}_\mathbf{C} = \{ \mathbf{C} + \mathbf{E}^\mathbf{M} \in & \mathbb{R}^{m \times n} \mid \mathbf{E}^\mathbf{M}_{ij} = (\mathbf{x}_i - \mathbf{y}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{y}_j); \\ & \mathbf{M} \in \mathcal{S}^d_+(\mathbb{R}); \|\mathbf{M}\|_p \leq r \}. \end{aligned} \tag{9}$$

for an arbitrary radius $r > 0$.

Proposition 5 (Non centered family of Mahalanobis distances)**.** *For a fixed $\mathbf{C}$, let $\mathcal{C}_\mathbf{C}$ be defined as in (9). Then, for $t \geq 0$, solving (7) over $\mathcal{P}_t \times \mathcal{C}_\mathbf{C}$, is equivalent to solving the following convex program*

$$\min_{\mathbf{P} \in \mathrm{Conv}(\mathcal{P}_t)} r \|\mathbf{V}_\mathbf{P}\|_q + \sum_{ij} (\mathbf{P})_{ij} (\mathbf{C})_{ij}. \tag{10}$$

Moreover, if $\mathbf{P}^*$ is an optimal solution of (10), then $\mathbf{M}^*$ is as in Proposition 1 with γ replaced by $\mathbf{P}^*$.

In the following, we consider the case of $p = 2$. By Corollary 1, $\mathbf{M}^*$ is PSD even without imposing such a constraint.

3.5. Towards the notion of stability of cost matrices

In this section, we define a new notion of OT stability for a cost matrix $\mathbf{C}$ based on a non-centered convex set $\mathcal{C}_\mathbf{C}$.

Definition 1. *For a cost matrix $\mathbf{C}$ and its associated convex set $\mathcal{C}_\mathbf{C}$ introduced in (9), for some $r > 0$, we define the stability $\mathcal{WS}_{\mathbf{C},r}$ as follows:*

$$\mathcal{WS}_{\mathbf{C},r} = W_{\mathcal{C}_\mathbf{C}}(\mu_1, \mu_2) - W_\mathbf{C}(\mu_1, \mu_2)$$

$$=\min_{\mathbf{P}\in\Pi}\max_{\|\mathbf{M}\|\leq r}\left\langle \mathbf{P},\mathbf{C}+\mathbf{E}^{\mathbf{M}}\right\rangle-\min_{\mathbf{P}\in\Pi}\langle\mathbf{P},\mathbf{C}\rangle\,.$$

Roughly speaking, Definition 1 tells us that the Wasserstein distance between μ_1 and μ_2 associated with a stable cost matrix $\mathbf{C}$ should not differ much from the Wasserstein distance calculated based on the worst cost matrix in the neighborhood of $\mathbf{C}$. Note that the latter is defined as a Mahalanobis ball allowing us to define the stability of $\mathbf{C}$ w.r.t. the finite sets $\mathcal{X}$ and $\mathcal{Y}$. To be able to compare different stabilities for a family of cost matrices $\{\mathbf{C}_i\}_{i=1}^K$, we normalize each $\mathbf{C}_i$ either by diving its elements by its Frobenius norm or by the associated transport cost $W_{\mathbf{C}_i}(\mu_1,\mu_2)$. Figure 2 illustrates the intuition behind the notion of cost matrix stability where the Frobenius norm is used for the normalization.

4. Experiments

In this section, we first illustrate our algorithm's speed of convergence and compare it to solving the original LP problem from (7). Then, we reproduce a simulated problem from (Paty & Cuturi, 2019) to assess the algorithm's ability to correctly identify the subspace of a lower dimensionality in which the transformation between the two samples lies. In what follows, we concentrate on comparing our approach with the authors' implementation of SRW while leaving aside the comparison with StrOT for which the implementation is not publicly available. The second part of our experiments is related to the notion of stability defined in Section 3.5. We first bring to light a correlation between the stability and the noise resistance of a cost matrix. Then, we show that selecting the most stable matrix allows to efficiently transport colors between two images in a color transfer task. The code for the different experiments is available on this link[3].

4.1. Convergence and execution time

We consider the case where $\mathcal{C}$ is the set of convex combinations of a given number of cost matrices denoted as $|\mathcal{C}|$. The convergence of Algorithm 1 is illustrated in Figure 3 (left) by plotting the evolution of the quantity $err(t) := |\mu_t - \langle\mathbf{P}_t,\mathbf{C}_t\rangle|$ along the iterations for $|\mathcal{C}|\in\{10,40,90\}$. From this plot, we see that the convergence becomes slower as $|\mathcal{C}|$ grows, which is expected because μ_t is the value at the solution of Problem (7) over $\operatorname{Conv}(\mathcal{P}_t)\times\mathcal{C}$. Second, for $|\mathcal{C}|=10$, Algorithm 1 already achieves an error $err(t)\leq 10^{-10}$ after $t=100$ iterations. This confirms that $\mathcal{P}_t$ does not have to grow until it becomes the set of vertices of Π, as $|\mathcal{P}_{100}|\leq|\mathcal{P}_0|+100\ll n! = 100!$. We also test our algorithm with the entropic regularization of the transport matrix with $\lambda\in\{1,0.1,0.01\}$ as regularization parameter, using Sinkhorn algorithm (Cuturi, 2013) for $|\mathcal{C}|=40$. For this setting, we initialize it with $\mathcal{P}_0=\{\frac{1}{mn}\mathbf{1}_m\mathbf{1}_n^T\}$ for any $\lambda>0$, as this set $\mathcal{P}_0$ is included in the feasible set of entropy-regularized transport (as suggested in the discussion of Proposition 2). Interestingly, we have noticed that the algorithm does not converge if $\mathcal{P}_0$ is a subset of the vertices of transportation polytope Π in the regularized case. The results of this experiment are reported in Figure 3 (middle), where we observe the convergence even with the entropy regularization. Additionally, we note that due to the linearity of the mapping $\langle\cdot,\mathbf{P}\rangle$ for all $\mathbf{P}\in\Pi$, Problem (4) can be reformulated as the following LP:

$$\begin{aligned}&\min_{\mathbf{P}\in\Pi,\eta\geq 0}\ \eta,\\ &\text{s.t. } \langle\mathbf{P},\mathbf{C}_l\rangle\leq\eta \qquad \forall 1\leq l\leq d.\end{aligned}$$

It turns out that this is nothing more than the dual of Problem (7). Under this formulation, solving $\mathrm{RKP}(\Pi,\mathcal{C})$ becomes tractable for $m=n=100$ and $|\mathcal{C}|\in\{10,20,...,90\}$ and allows us to compare the execution time of solving the LP problem to that of our algorithm in Figure 3 (right). As the number of candidate matrices grows, our algorithm becomes much more efficient than solving the full LP problem. This is rather expected since at each iteration, it solves a linear program with much less constraints (the problem is restricted to $\operatorname{Conv}(\mathcal{P}_t)\times\mathcal{C}$ instead of $\Pi\times\mathcal{C}$) and it leverages efficient algorithms for solving problem (2).

4.2. Comparison to SRW

In this series of experiments, we consider the fragmented hypercube dataset studied in (Paty & Cuturi, 2019) and earlier in (Forrow et al., 2019) and compare RKP to both the SRW and the Wasserstein distances. To proceed, let $\{\mathbf{e}_l\}_{1\leq l\leq d}$ be the canonical basis of $\mathbb{R}^d$ and let $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^m$ and $\mathcal{Y}=\{\mathbf{y}_j\}_{i=j}^n$ be two finite sets drawn i.i.d. from the uniform distribution over the d-dimensional hypercube $\mathcal{U}([-1,1]^d)$ and its pushforward distribution under the mapping $T: x\mapsto x+2\operatorname{sgn}(x)\odot(\sum_{i=1}^k\mathbf{e}_i)$, where $\odot$ denotes elementwise multiplication and $k\in[\![1,d]\!]$, respectively. Therefore, by construction, there are k relevant features and $d-k$ features that contain no useful information. Depending on the choice of $\mathcal{C}$, two cases of our algorithm are tested: 1) squared Euclidean distance after projecting on all combinations of two vectors of the canonical basis $\mathcal{C}=\{\mathbf{C}_{s,l}\in\mathbb{R}^{m\times n}|(\mathbf{C}_{s,l})_{ij}=((\mathbf{x}_i-\mathbf{y}_j)^T(\mathbf{e}_s+\mathbf{e}_l))^2; 1\leq s<l\leq d\}$ and 2) the Mahalanobis ball centered at 0 as defined in Section 3.2.1. Note that in this latter case $\operatorname{rank}(\mathbf{M}^*)=k$.

Figure 4 (left) reproduces the experiments of (Paty & Cuturi, 2019) and shows that the original OT (bottom left) is sensitive to noise, while both SRW and RKP (for the 2 configurations considered) are able to recover the true pushforward transformation. However, while SRW requires a hyperparameter k to constrain the rank of the Mahalanobis matrix,

[3] https://github.com/sofiendhouib/minimax_OT.

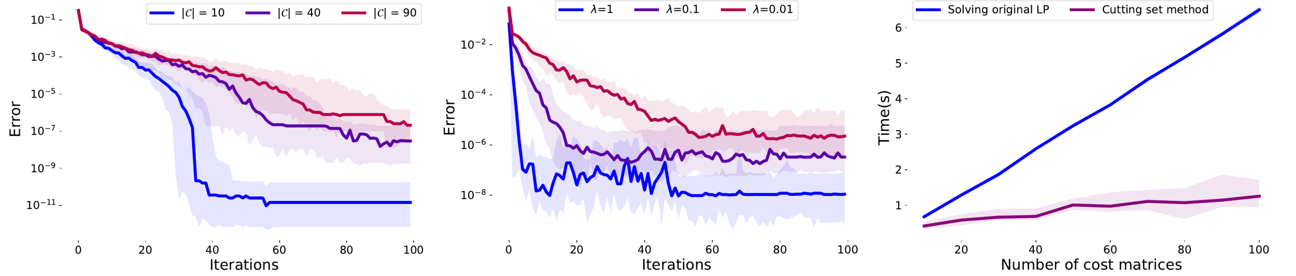

Figure 3. **(left)** Evolution of the error along the iterations for $|\mathcal{C}| \in \{10, 40, 90\}$; **(middle)** Evolution of the error with a regularization parameter $\lambda \in \{1, 0.1, 0.01\}$; **(right)** Execution time of our algorithm vs solving the original LP problem with $|\mathcal{C}| \in \{10, 20, ..., 90\}$ and $n = m = 100$. The experiments are repeated 30 times. The median and the interval between the first and third quartiles are reported.

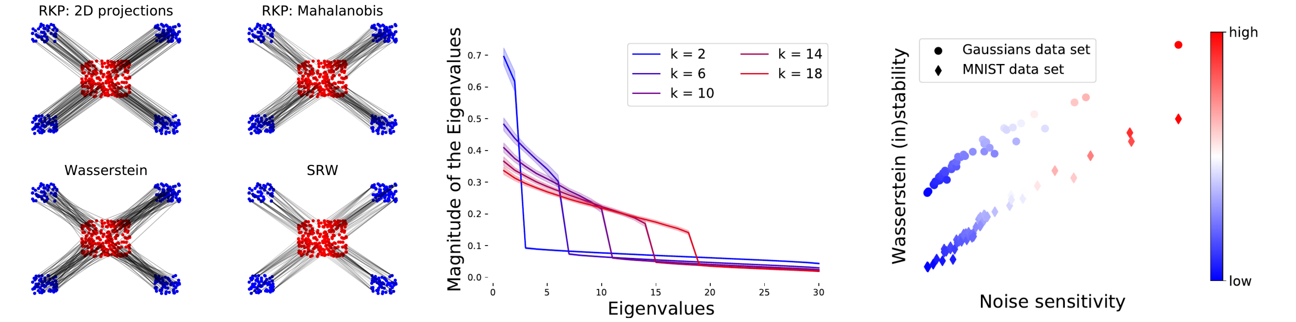

Figure 4. **(left)** Results obtained on the fragmented hypercube for $m = n = 250$, $d = 30$ and $k = 2$ with **(top row)** our approach with 2D projections and Mahalanobis distances; **(bottom row)** Original OT problem and SRW method of (Paty & Cuturi, 2019); **(middle)** Sorted eigenvalues of $\mathbf{M}^*$ obtained using RKP averaged over 100 runs for different values of k reveals a phase transition between k dominant and the $k + 1$ eigenvalues; **(right)** Correlation between the stability and the sensitivity to noise.

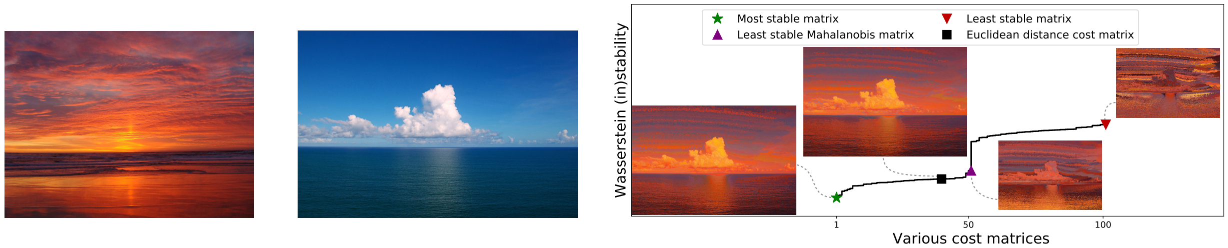

Figure 5. **(left)** Source (ocean) and target (sky) images considered as probability distributions; **(right)** Cost matrices sorted by Wasserstein stability. The first 50 are Mahalanobis cost matrices, while the last 50 are random cost matrices.

our method is parameter-free since k is found automatically as illustrated in Figure 4 (middle). In this figure, we plot the eigenvalues of $\mathbf{M}^*$ for different values of k and observe that the eigengap between the k largest eigenvalues and the $(k+1)^{th}$ eigenvalue clearly reveals that rank$(\mathbf{M}^*) = k$.

4.3. Stability and noise sensitivity

Below, we illustrate the correlation between the cost matrix stability and the sensitivity of the Wasserstein distance to the presence of noise using both toy and a real-world datasets. The latter one is composed of 100 zeros and 100 ones coming from the MNIST dataset, after reducing its dimensionality to 10 with UMAP (McInnes et al., 2018). The former consists of 100 points drawn from two 10-dimensional Gaussian distributions centered at $\mathbf{0}_{10}$ and $3 \times \mathbf{1}_{10}$ respectively with unit variance. For both datasets, we generate a family of cost matrices $\{\mathbf{C}_i\}_{i=1}^{50}$ based on random Mahalanobis distances with different norms, normalize them so that their Frobenius norm equals to 1 and compute $\mathcal{WS}_{\mathbf{C}_i, r=0.01}$ from Definition 1 for all i. To introduce noise to each $\mathbf{C}_i$, we add a random Mahalanobis cost matrix $\mathbf{E}^\mathbf{N}$ with $\|\mathbf{N}\|_2 = r$ to it and compute the noise sensitivity defined as:

$$\mathcal{NS}_{\mathbf{C}_i} = \left| \min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C}_i \rangle - \min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C}_i + \mathbf{E}^\mathbf{N} \rangle \right|.$$

Note that we apply a Mahalanobis noise which has the advantage of taking into account the point distributions and can be applied on any matrix $\mathbf{C}_i$. Figure 4 (right) presents the results of this experiment averaged over 200 runs and shows a clear correlation between the stability and noise sensitivity indicating that the most stable matrices are more noise tolerant. Other experiments on the MNIST dataset provided in the supplementary material show a similar behavior.

4.4. Color transfer

In this last experiment, we show how we can benefit from the notion of stability to address a color transfer task where the goal is to transfer the colors from a blueish sky image to the reddish ocean image shown in Figure 5 (left). Here, we use OT between the sets of pixels in the RGB space extracted from both images. For the sake of efficiency, we consider only 200 pixels from each image and generalize the obtained OT mapping to the remaining pixels following the method detailed in (Ferradans et al., 2014). As before, we use $\{\mathbf{C}_i\}_{i=1}^{50}$ as meaningful cost matrices and add 50 completely random matrices that are unrelated to the considered task. The results of this experiment given in Figure 5 (right) show a significant gap in terms of stability between the Mahalanobis matrices (the first 50 matrices on the x-axis) and the random ones (the last 50). This tends to highlight the fact that the stability can be used as a criterion to select a good cost matrix, and therefore to induce a relevant Mahalanobis distance. This also holds in terms of visual perception as illustrated in Figure 5 (right). Even if the most stable matrix is visually very similar to the Euclidean one, a finer evaluation reveals more discontinuities in the center of the picture, on the water.

5. Conclusion

In this paper, we study a general formulation of the minimax OT problem that consists in optimizing over the coupling matrix w.r.t. the worst cost function from a certain convex set of cost functions. When the latter is given by an infinite family of Mahalanobis distances, we highlight several novel properties of the considered problem and characterize the different features of its solutions. We further show how the underlying optimization problem can be solved in practice using a variation of a cutting set algorithm with theoretical guarantees regarding its convergence speed. Finally, we define a new notion of stability for cost matrices in OT based on the studied minimax problem and reveal a correlation between this stability and the noise resistance of the matrices. This leads to a criterion that can be used to select a relevant cost function which has been shown to be efficient on both toy and real-world data. A promising line of research might be to find the most stable cost matrix from a continuous set.

Acknowledgements

This work was supported by the following projects: AURA project TADALoT (Pack Ambition 2017, 17 011047 01) and IDEXLYON project ACADEMICS (ANR-16-IDEX-0005).

References

Alvarez-Melis, D., Jaakkola, T., and Jegelka, S. Structured optimal transport. In *AISTATS*, pp. 1771–1780, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *ICML*, volume 70, pp. 214–223, 2017.

Bellet, A., Habrard, A., and Sebban, M. *Metric Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2015.

Bonneel, N., Peyré, G., and Cuturi, M. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71:1–71:10, 2016.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *COLT*, pp. 144–152, 1992.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. In *NIPS*, pp. 3733–3742, 2017.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Transactions Information Theory*, 13(1): 21–27, 2006.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pp. 2292–2300, 2013.

Cuturi, M. and Avis, D. Ground metric learning. *J. Mach. Learn. Res.*, 15(1), 2014.

Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. Regularized Discrete Optimal Transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., and Weed, J. Statistical optimal transport via factored couplings. In *AISTATS*, pp. 2454–2465, 2019.

Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *AISTATS*, pp. 1608–1617, 2018.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.

Gramfort, A., Peyré, G., and Cuturi, M. Fast optimal transport averaging of neuroimaging data. In *IPMI*, pp. 261–272, 2015.

Kantorovich, L. On the translocation of masses. *Doklady of the Academy of Sciences of the USSR*, 37:199–201, 1942.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *CoRR*, abs/1908.08729, 2019.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *ICML*, volume 37, pp. 957–966, 2015.

Laclau, C., Redko, I., Matei, B., Bennani, Y., and Brault, V. Co-clustering through optimal transport. In *ICML*, pp. 1955–1964, 2017.

Li, R., Ye, X., Zhou, H., and Zha, H. Learning to match via inverse optimal transport. *J. Mach. Learn. Res.*, 20: 80:1–80:37, 2019.

Magnus, J. R. A representation theorem for (trAp)1/p. Other publications TiSEM, Tilburg University, School of Economics and Management, 1987.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Academie Royale des Sciences*, pp. 666–704, 1781.

Mutapcic, A. and Boyd, S. P. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods and Software*, 24:381–406, 2009.

Paty, F. and Cuturi, M. Subspace robust wasserstein distances. In *ICML*, pp. 5072–5081, 2019.

Sion, M. On general minimax theorems. *Pacific J. Math.*, 8 (1):171–176, 1958.

Zhao, P. and Zhou, Z. Label distribution learning by optimal transport. In *AAAI*, pp. 4506–4513, 2018.