Regularized Optimal Transport and the ROT Mover's Distance

Arnaud Dessein, Nicolas Papadakis, Jean-Luc Rouas

Regularized Optimal Transport and the ROT Mover's Distance

Arnaud Dessein
Qucit, Bègles, France
arnaud.dessein@qucit.com

Nicolas Papadakis
IMB, CNRS, Bordeaux, France
nicolas.papadakis@math.u-bordeaux.fr

Jean-Luc Rouas
LaBRI, CNRS, Bordeaux, France
jean-luc.rouas@labri.fr

July 14, 2018

Abstract

This paper presents a unified framework for smooth convex regularization of discrete optimal transport problems. In this context, the regularized optimal transport turns out to be equivalent to a matrix nearness problem with respect to Bregman divergences. Our framework thus naturally generalizes a previously proposed regularization based on the Boltzmann-Shannon entropy related to the Kullback-Leibler divergence, and solved with the Sinkhorn-Knopp algorithm. We call the regularized optimal transport distance the rot mover's distance in reference to the classical earth mover's distance. By exploiting alternate Bregman projections, we develop the alternate scaling algorithm and non-negative alternate scaling algorithm, to compute efficiently the regularized optimal plans depending on whether the domain of the regularizer lies within the non-negative orthant or not. We further enhance the separable case with a sparse extension to deal with high data dimensions. We also instantiate our framework and discuss the inherent specificities for well-known regularizers and statistical divergences in the machine learning and information geometry communities. Finally, we demonstrate the merits of our methods with experiments using synthetic data to illustrate the effect of different regularizers, penalties and dimensions, as well as real-world data for a pattern recognition application to audio scene classification.

1 Introduction

A recurrent problem in statistical machine learning is the choice of a relevant distance measure to compare probability distributions. Various information divergences are famous, among which Euclidean, Mahalanobis, Kullback-Leibler, Itakura-Saito, Hellinger, χ^2, ℓ_p (quasi-)norm, total variation, logistic loss function, or more general Csiszár and Bregman divergences and parametric families of such divergences such as α- and β-divergences.

An alternative family of distances between probability distributions can be introduced in the framework of optimal transport (OT). Rather than performing a pointwise comparison of the

distributions, the idea is to quantify the minimal effort for moving the probability mass of one distribution to the other, where the transport plan to move the mass is optimized according to a given ground cost. This makes OT distances suitable and robust in certain applications, notably in the field of computer vision where the discrete OT distance, also known as earth mover's distance (EMD), has been popularized to compare histograms of features for pattern recognition tasks [41].

Despite its appealing theoretical properties, intuitive formulation, and excellent performance in various problems of information retrieval, the computation of the EMD involves solving a linear program whose cost quickly becomes prohibitive with the data dimension. In practice, the best algorithms currently proposed, such as the network simplex [1], scale at least with a super-cubic complexity. Embeddings of the distributions can be used to approximate the EMD with linear complexity [24, 28, 47], and the network simplex can be modified to run in quadratic time [25,31,38]. Nevertheless, the distortions inherent to such embeddings [35], and the exponential increase of costs incurred by such modifications, make these approaches inapplicable for dimensions higher than four. Instead, multi-scale strategies [36] and shortcut paths [43] can speed up the estimation of the exact optimal plan. These approaches are yet limited to particular convex costs such as ℓ_2, while other costs such as ℓ_1 and truncated or compressed versions are often preferred in practice for an increased robustness to data outliers [37–39]. For general applications, a gain in performance can also be obtained with a cost directly learned from labeled data [16]. The aforementioned accelerated methods that are dedicated to ℓ_2 or convex costs are thus not adapted in this context.

On another line of research, the regularization of the transport plan, for example via graph modeling [20], has been considered to deal with noisy data, though this latter approach does not address the computational issue of efficiency for high dimensions. In this continuity, an entropic regularization was shown to admit an efficient algorithm with quadratic complexity that speeds up the computation of solutions by several orders of magnitude, and to improve performance on applications such as handwritten digit recognition [15]. In addition, a tailored computation can be obtained via convolution for specific ground costs [49]. Since the introduction of the entropic regularization, OT has benefited from extensive developments in the machine learning community, with applications such as label propagation [50], domain adaptation [14], matrix factorization [55], dictionary learning [40, 42], barycenter computation [17], geodesic principal component analysis [8, 12, 45], data fitting [21], statistical inference [7], training of Boltzmann machines [33] and generative adversarial networks [4, 11, 23].

With the entropic regularization, the gain in computational time is only important for high dimensions or large levels of regularization. For low regularization, advanced optimization strategies can still be used to obtain a significant speed-up [42, 51]. It is also a well-known effect that the entropic regularization overspreads the transported mass, which may be undesirable for certain applications as in the case of interpolation purposes. An interesting perspective of these works, however, is that many more regularizers are worth investigating to solve OT problems both efficiently and robustly [9, 22, 34]. This is the idea we address in the present work, focusing on smooth convex regularization.

1.1 Notations

For the sake of simplicity, we consider distributions with same dimension d, and thus work with the Euclidean space $\mathbb{R}^{d\times d}$ of square matrices. It is straightforward, however, to extend all results for a different number of bins m, n by using rectangular matrices in $\mathbb{R}^{m\times n}$ instead. We denote the null matrix of $\mathbb{R}^{d\times d}$ by $\mathbf{0}$, and the matrix full of ones by $\mathbf{1}$. The Frobenius inner product

between two matrices $\boldsymbol{\pi}, \boldsymbol{\xi} \in \mathbb{R}^{d \times d}$ is defined by:

$$\langle \boldsymbol{\pi}, \boldsymbol{\xi} \rangle = \sum_{i=1}^{d} \sum_{j=1}^{d} \pi_{ij} \xi_{ij} \ . \tag{1}$$

When the intended meaning is clear from the context, we also write $\mathbf{0}$ for the null vector of $\mathbb{R}^d$, and $\mathbf{1}$ for the vector full of ones. The notation $\cdot^\top$ represents the transposition operator for matrices or vectors. The probability simplex of $\mathbb{R}^d$ is defined as follows:

$$\Sigma_d = \{\mathbf{p} \in \mathbb{R}_+^d \colon \mathbf{p}^\top \mathbf{1} = 1\} \ . \tag{2}$$

The operator $\mathrm{diag}(\mathbf{v})$ transforms a vector $\mathbf{v} \in \mathbb{R}^d$ into a diagonal matrix $\boldsymbol{\pi} \in \mathbb{R}^{d \times d}$ such that $\pi_{ii} = v_i$, for all $1 \leq i \leq d$. The operator $\mathrm{vec}(\boldsymbol{\pi})$ transforms a matrix $\boldsymbol{\pi} \in \mathbb{R}^{d \times d}$ into a vector $\mathbf{x} \in \mathbb{R}^{d^2}$ such that $x_{i+(j-1)d} = \pi_{ij}$, for all $1 \leq i, j \leq d$. The operator $\mathrm{sgn}(x)$ for $x \in \mathbb{R}$ returns $-1, 0, +1$, if x is negative, null, positive, respectively. Functions of a real variable, such as the absolute value, sign, exponential or power functions, are considered element-wise when applied to matrices. The max operator and inequalities between matrices should also be interpreted element-wise. Matrix divisions are similarly considered element-wise, whereas element-wise matrix multiplications, also known as Hadamard or Schur products, are denoted by $\odot$ to remove any ambiguity with standard matrix multiplications. Lastly, addition or subtraction of a scalar and a matrix should be understood element-wise by replicating the scalar.

1.2 Background and Related Work

Given two probability vectors $\mathbf{p}, \mathbf{q} \in \Sigma_d$, and a cost matrix $\boldsymbol{\gamma} \in \mathbb{R}_+^{d \times d}$ whose coefficients γ_{ij} represent the cost of moving the mass from bin p_i to q_j, the total cost of a given transport plan, or coupling, $\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})$ can be quantified as $\langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle$. An optimal cost is then obtained by solving a linear program:

$$d_{\boldsymbol{\gamma}}(\mathbf{p}, \mathbf{q}) = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})} \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle \ , \tag{3}$$

with the transport polytope of $\mathbf{p}$ and $\mathbf{q}$, also known as the polytope of couplings between $\mathbf{p}$ and $\mathbf{q}$, defined as the following polyhedron:

$$\Pi(\mathbf{p}, \mathbf{q}) = \{\boldsymbol{\pi} \in \mathbb{R}_+^{d \times d} \colon \boldsymbol{\pi} \mathbf{1} = \mathbf{p}, \boldsymbol{\pi}^\top \mathbf{1} = \mathbf{q}\} \ . \tag{4}$$

The EMD associated to the cost matrix $\boldsymbol{\gamma}$ is given by $d_{\boldsymbol{\gamma}}$ and is a true distance metric on the probability simplex Σ_d whenever $\boldsymbol{\gamma}$ is itself a distance matrix. In general, the optimal plans, or earth mover's plans, have at most $2d - 1$ nonzero entries, and consist either of a single vertex or of a whole facet of the transport polytope. One of the earth mover's plans can be obtained with the network simplex [1] among other approaches. For a general cost matrix $\boldsymbol{\gamma}$, the complexity of solving an OT problem scales at least in $O(d^3 \log d)$ for the best algorithms currently proposed, including the network simplex, and turns out to be super-cubic in practice as well.

[15] proposed a new family of OT distances, called Sinkhorn distances, from the perspective of maximum entropy. The idea is to smooth the original problem with a strictly convex regularization via the Boltzmann-Shannon entropy. The primal problem involves the entropic regularization as an additional constraint:

$$d'_{\boldsymbol{\gamma}, \alpha}(\mathbf{p}, \mathbf{q}) = \min_{\boldsymbol{\pi} \in \Pi_\alpha(\mathbf{p}, \mathbf{q})} \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle \ , \tag{5}$$

with the regularized transport polytope defined as follows:

$$\Pi_\alpha(\mathbf{p}, \mathbf{q}) = \{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q}) \colon E(\boldsymbol{\pi}) \leq E(\mathbf{p}\mathbf{q}^\top) + \alpha\} \ , \tag{6}$$

where $\alpha \geq 0$ is a regularization term and E is minus the Boltzmann-Shannon entropy as defined in (28). It is also straightforward to prove that we have:

$$\Pi_\alpha(\mathbf{p}, \mathbf{q}) = \{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q}) \colon K(\boldsymbol{\pi} \| \mathbf{1}) \leq K(\mathbf{p}\mathbf{q}^\top \| \mathbf{1}) + \alpha\} \ , \tag{7}$$

$$\Pi_\alpha(\mathbf{p}, \mathbf{q}) = \{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q}) \colon K(\boldsymbol{\pi} \| \mathbf{p}\mathbf{q}^\top) \leq \alpha\} \ . \tag{8}$$

where K is the Kullback-Leibler divergence as defined in (27). This enforces the solution to have sufficient entropy, or equivalently small enough mutual information, by constraining it to the Kullback-Leibler ball of radius $K(\mathbf{p}\mathbf{q}^\top \| \mathbf{1}) + \alpha$, respectively α, and center the matrix $\mathbf{1} \in \mathbb{R}_{++}^{d \times d}$, respectively the transport plan $\mathbf{p}\mathbf{q}^\top \in \mathbb{R}_{++}^{d \times d}$, which have maximum entropy. The dual problem exploits a Lagrange multiplier to relax the entropic regularization as a penalty:

$$d_{\boldsymbol{\gamma}, \lambda}(\mathbf{p}, \mathbf{q}) = \langle \boldsymbol{\pi}_\lambda^\star, \boldsymbol{\gamma} \rangle \ , \tag{9}$$

with the regularized optimal plan $\boldsymbol{\pi}_\lambda^\star$ defined as follows:

$$\boldsymbol{\pi}_\lambda^\star = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \ \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda E(\boldsymbol{\pi}) \ , \tag{10}$$

where $\lambda > 0$ is a regularization term. The problem can then be solved empirically in quadratic complexity with linear convergence using the Sinkhorn-Knopp algorithm [48] based on iterative matrix scaling, where rows and columns are rescaled in turn so that they respectively sum up to $\mathbf{p}$ and $\mathbf{q}$ until convergence. Finally, it is easy to prove that we have:

$$\boldsymbol{\pi}_\lambda^\star = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \ \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda K(\boldsymbol{\pi} \| \mathbf{1}) \ , \tag{11}$$

$$\boldsymbol{\pi}_\lambda^\star = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \ \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda K(\boldsymbol{\pi} \| \mathbf{p}\mathbf{q}^\top) \ . \tag{12}$$

This again shows that the regularization enforces the solution to have sufficient entropy, or equivalently small enough mutual information, by shrinking it toward the matrix $\mathbf{1}$ and the joint distribution $\mathbf{p}\mathbf{q}^\top$ which have maximum entropy.

[6] revisited the entropic regularization in a geometrical framework with iterative information projections. They showed that computing a Sinkhorn distance in dual form actually amounts to the minimization of a Kullback-Leibler divergence:

$$\boldsymbol{\pi}_\lambda^\star = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \ K(\boldsymbol{\pi} \| \exp(-\boldsymbol{\gamma}/\lambda)) \ . \tag{13}$$

Precisely, this amounts to computing the Kullback-Leibler projection of $\exp(-\boldsymbol{\gamma}/\lambda) \in \mathbb{R}_{++}^{d \times d}$ onto the transport polytope $\Pi(\mathbf{p}, \mathbf{q})$. In this context, the Sinkhorn-Knopp algorithm turns out to be a special instance of Bregman projection onto the intersection of convex sets via alternate projections. Specifically, we see $\Pi(\mathbf{p}, \mathbf{q})$ as the intersection of the non-negative orthant with two affine subspaces containing all matrices with rows and columns summing to $\mathbf{p}$ and $\mathbf{q}$ respectively, and we alternate projection on these two subspaces according to the Kullback-Leibler divergence until convergence.

[29] further studied this equivalence in the wider context of iterative proportional fitting. He notably showed that the Sinkhorn-Knopp and alternate Bregman projections can be extended to account for infinite entries in the cost matrix $\boldsymbol{\gamma}$, and thus null entries in the regularized optimal plan. Hence, it is possible to develop a sparse version of the entropic regularization to OT problems. This becomes interesting to store the $d \times d$ matrix variables and perform the required computations when the data dimension gets large.

[19] had already enlightened such an equivalence in the field of matrix analysis. They actually considered the estimation of contingency tables with fixed marginals as a matrix nearness problem based on the Kullback-Leibler divergence. In more detail, they use a rough estimate $\boldsymbol{\xi} \in \mathbb{R}_{++}^{d \times d}$ to produce a contingency table $\boldsymbol{\pi}^\star$ that has fixed marginals $\mathbf{p}, \mathbf{q}$ by Kullback-Leibler projection of $\boldsymbol{\xi}$ onto $\Pi(\mathbf{p}, \mathbf{q})$:

$$\boldsymbol{\pi}^\star = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \; K(\boldsymbol{\pi} \| \boldsymbol{\xi}) \;\; . \tag{14}$$

They showed that alternate Bregman projections specialize to the Sinkhorn-Knopp algorithm in this context. However, no relationship to OT problems was highlighted.

1.3 Contributions and Organization

Our main contribution is to formulate a unified framework for discrete regularized optimal transport (ROT) by considering a large class of smooth convex regularizers. We call the underlying distance the rot mover's distance (RMD) and show that a given ROT problem actually amounts to the minimization of an associated Bregman divergence. This allows the derivation of two schemes that we call the alternate scaling algorithm (ASA) and the non-negative alternate scaling algorithm (NASA), to compute efficiently the regularized optimal plans depending on whether the domain of the regularizer lies within the non-negative orthant or not. These schemes are based on the general form of alternate projections for Bregman divergences. They also exploit the Newton-Raphson method to approximate the projections for separable divergences. The separable case is further enhanced with a sparse extension to deal with high data dimensions. We also instantiate our two generic schemes with widely-used regularizers and statistical divergences.

The proposed framework naturally extends the Sinkhorn-Knopp algorithm for the regularization based on the Boltzmann-Shannon entropy [15], or equivalently the minimization of a Kullback-Leibler divergence [6], and their sparse version [29], which turn out to be special instances of ROT problems. It also relates to matrix nearness problems via minimization of Bregman divergences, and it is straightforward to construct more general estimators for contingency tables with fixed marginals than the classical estimator based on the Kullback-Leibler divergence [19]. Lastly, it brings some new insights between transportation theory [54] and information geometry [3], where Bregman divergences are known to possess a dually flat structure with a generalized Pythagorean theorem in relation to information projections.

The remainder of this paper is organized as follows. In Section 2, we introduce some necessary preliminaries. In Section 3, we present our theoretical results for a unified framework of ROT problems. We then derive the algorithmic methods for solving ROT problems in Section 4. We also discuss the inherent specificities of ROT problems for classical regularizers and associated divergences in Section 5. In Section 6, we provide experiments to illustrate our methods on synthetic data and real-world audio data in a classification problem. Finally, in Section 7, we draw some conclusions and perspectives for future work.

2 Theoretical Preliminaries

In this section, we introduce the required preliminaries to our framework. We begin with elements of convex analysis (Section 2.1) and of Bregman geometry (Section 2.2). We proceed with theoretical results for convergence of alternate Bregman projections (Section 2.3) and of the Newton-Raphson method (Section 2.4).

2.1 Convex Analysis

Let $\mathcal{E}$ be a Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. The boundary, interior and relative interior of a subset $\mathcal{X} \subseteq \mathcal{E}$ are respectively denoted by $\mathrm{bd}(\mathcal{X})$, $\mathrm{int}(\mathcal{X})$, and $\mathrm{ri}(\mathcal{X})$, where we recall that for a convex set $\mathcal{C}$, we have:

$$\mathrm{ri}(\mathcal{C}) = \{\mathbf{x} \in \mathcal{C} \colon \forall \mathbf{y} \in \mathcal{C}, \exists \lambda > 1, \lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in \mathcal{C}\} \ . \tag{15}$$

In convex analysis, scalar functions are defined over the whole space $\mathcal{E}$ and take values in the extended real number line $\mathbb{R} \cup \{-\infty, +\infty\}$. The effective domain, or simply domain, of a function f is then defined as the set:

$$\mathrm{dom}\, f = \{\mathbf{x} \in \mathcal{E} \colon f(\mathbf{x}) < +\infty\} \ . \tag{16}$$

A convex function f is proper if $f(\mathbf{x}) < +\infty$ for at least one $\mathbf{x} \in \mathcal{E}$ and $f(\mathbf{x}) > -\infty$ for all $\mathbf{x} \in \mathcal{E}$, and it is closed if its lower level sets $\{\mathbf{x} \in \mathcal{E} \colon f(\mathbf{x}) \leq \alpha\}$ are closed for all $\alpha \in \mathbb{R}$. If $\mathrm{dom}\, f$ is closed, then f is closed, and a proper convex function is closed if and only if it is lower semi-continuous. Moreover, a closed function f is continuous relative to any simplex, polytope or polyhedral subset in $\mathrm{dom}\, f$. It is also well-known that a convex function f is always continuous in the relative interior $\mathrm{ri}(\mathrm{dom}\, f)$ of its domain.

A function f is essentially smooth if it is differentiable on $\mathrm{int}(\mathrm{dom}\, f) \neq \emptyset$ and verifies $\lim_{k \to +\infty} \|\nabla f(\mathbf{x}_k)\| = +\infty$ for any sequence $(\mathbf{x}_k)_{k \in \mathbb{N}}$ from $\mathrm{int}(\mathrm{dom}\, f)$ that converges to a point $\mathbf{x} \in \mathrm{bd}(\mathrm{dom}\, f)$. A function f is of Legendre type if it is a closed proper convex function that is also essentially smooth and strictly convex on $\mathrm{int}(\mathrm{dom}\, f)$.

The Fenchel conjugate $f^\star$ of a function f is defined for all $\mathbf{y} \in \mathcal{E}$ as follows:

$$f^\star(\mathbf{y}) = \sup_{\mathbf{x} \in \mathrm{int}(\mathrm{dom}\, f)} \langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x}) \ . \tag{17}$$

The Fenchel conjugate $f^\star$ is always a closed convex function. Moreover, if f is a closed convex function, then $(f^\star)^\star = f$, and f is of Legendre type if and only if $f^\star$ is of Legendre type. In this latter case, the gradient mapping ∇f is a homeomorphism between $\mathrm{int}(\mathrm{dom}\, f)$ and $\mathrm{int}(\mathrm{dom}\, f^\star)$, with inverse mapping $(\nabla f)^{-1} = \nabla f^\star$, which guarantees the existence of dual coordinate systems $\mathbf{x}(\mathbf{y}) = \nabla f^\star(\mathbf{y})$ and $\mathbf{y}(\mathbf{x}) = \nabla f(\mathbf{x})$ on $\mathrm{int}(\mathrm{dom}\, f)$ and $\mathrm{int}(\mathrm{dom}\, f^\star)$.

Finally, we say that a function f is cofinite if it verifies:

$$\lim_{\lambda \to +\infty} f(\lambda \mathbf{x})/\lambda = +\infty \ , \tag{18}$$

for all nonzero $\mathbf{x} \in \mathcal{E}$. Intuitively, it means that f grows super-linearly in every direction. In particular, a closed proper convex function is cofinite if and only if $\mathrm{dom}\, f^\star = \mathcal{E}$.

2.2 Bregman Geometry

Let ϕ be a convex function on $\mathcal{E}$ that is differentiable on $\mathrm{int}(\mathrm{dom}\, \phi) \neq \emptyset$. The Bregman divergence generated by ϕ is defined as follows:

$$B_\phi(\mathbf{x} \| \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla \phi(\mathbf{y}) \rangle \ , \tag{19}$$

for all $\mathbf{x} \in \mathrm{dom}\, \phi$ and $\mathbf{y} \in \mathrm{int}(\mathrm{dom}\, \phi)$. We have $B_\phi(\mathbf{x} \| \mathbf{y}) \geq 0$ for any $\mathbf{x} \in \mathrm{dom}\, \phi$ and $\mathbf{y} \in \mathrm{int}(\mathrm{dom}\, \phi)$. If in addition ϕ is strictly convex on $\mathrm{int}(\mathrm{dom}\, \phi)$, then $B_\phi(\mathbf{x} \| \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$. Bregman divergences are also always convex in the first argument, and are invariant under adding an arbitrary affine term to their generator.

Bregman divergences are not symmetric and do not verify the triangle inequality in general, and thus are not necessarily distances in the strict sense. However, they still enjoy some nice geometrical properties that somehow generalize the Euclidean geometry. In particular, they verify a four-point identity similar to a parallelogram law:

$$B_\phi(\mathbf{x}\|\mathbf{y}) + B_\phi(\mathbf{x}'\|\mathbf{y}') = B_\phi(\mathbf{x}'\|\mathbf{y}) + B_\phi(\mathbf{x}\|\mathbf{y}') - \langle \mathbf{x} - \mathbf{x}', \nabla\phi(\mathbf{y}) - \nabla\phi(\mathbf{y}')\rangle \ , \tag{20}$$

for all $\mathbf{x}, \mathbf{x}' \in \operatorname{dom}\phi$ and $\mathbf{y}, \mathbf{y}' \in \operatorname{int}(\operatorname{dom}\phi)$. A special instance of this relation gives rise to a three-point property similar to a triangle law of cosines:

$$B_\phi(\mathbf{x}\|\mathbf{y}) = B_\phi(\mathbf{x}\|\mathbf{y}') + B_\phi(\mathbf{y}'\|\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}', \nabla\phi(\mathbf{y}) - \nabla\phi(\mathbf{y}')\rangle \ , \tag{21}$$

for all $\mathbf{x} \in \operatorname{dom}\phi$ and $\mathbf{y}, \mathbf{y}' \in \operatorname{int}(\operatorname{dom}\phi)$.

Suppose now that ϕ is of Legendre type, and let $\mathcal{C} \subseteq \mathcal{E}$ be a closed convex set such that $\mathcal{C} \cap \operatorname{int}(\operatorname{dom}\phi) \neq \emptyset$. Then, for any point $\mathbf{y} \in \operatorname{int}(\operatorname{dom}\phi)$, the following problem:

$$P_\mathcal{C}(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{C}} B_\phi(\mathbf{x}\|\mathbf{y}) \ , \tag{22}$$

has a unique solution, then called the Bregman projection of $\mathbf{y}$ onto $\mathcal{C}$. This solution actually belongs to $\mathcal{C} \cap \operatorname{int}(\operatorname{dom}\phi)$, and is also characterized as the unique point $\mathbf{y}' \in \mathcal{C} \cap \operatorname{int}(\operatorname{dom}\phi)$ that verifies the variational relation:

$$\langle \mathbf{x} - \mathbf{y}', \nabla\phi(\mathbf{y}) - \nabla\phi(\mathbf{y}')\rangle \leq 0 \ , \tag{23}$$

for all $\mathbf{x} \in \mathcal{C} \cap \operatorname{dom}\phi$. This characterization is equivalent to a well-known generalized Pythagorean theorem for Bregman divergences, which states that the Bregman projection of $\mathbf{y}$ onto $\mathcal{C}$ is the unique point $\mathbf{y}' \in \mathcal{C} \cap \operatorname{int}(\operatorname{dom}\phi)$ that verifies the following inequality:

$$B_\phi(\mathbf{x}\|\mathbf{y}) \geq B_\phi(\mathbf{x}\|\mathbf{y}') + B_\phi(\mathbf{y}'\|\mathbf{y}) \ , \tag{24}$$

for all $\mathbf{x} \in \mathcal{C} \cap \operatorname{dom}\phi$. When $\mathcal{C}$ is further an affine subspace, or more generally when the Bregman projection further belongs to $\operatorname{ri}(\mathcal{C})$, the scalar product actually vanishes:

$$\langle \mathbf{x} - \mathbf{y}', \nabla\phi(\mathbf{y}) - \nabla\phi(\mathbf{y}')\rangle = 0 \ , \tag{25}$$

leading to an equality in the generalized Pythagorean theorem:

$$B_\phi(\mathbf{x}\|\mathbf{y}) = B_\phi(\mathbf{x}\|\mathbf{y}') + B_\phi(\mathbf{y}'\|\mathbf{y}) \ . \tag{26}$$

A famous example of Bregman divergence is the Kullback-Leibler divergence, defined for matrices $\boldsymbol{\pi} \in \mathbb{R}_+^{d\times d}$ and $\boldsymbol{\xi} \in \mathbb{R}_{++}^{d\times d}$ as follows:

$$K(\boldsymbol{\pi}\|\boldsymbol{\xi}) = \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\pi_{ij}\log\left(\frac{\pi_{ij}}{\xi_{ij}}\right) - \pi_{ij} + \xi_{ij}\right) \ . \tag{27}$$

This divergence is generated by a function of Legendre type for $\boldsymbol{\pi} \in \mathbb{R}_+^{d\times d}$ given by minus the Boltzmann-Shannon entropy:

$$E(\boldsymbol{\pi}) = K(\boldsymbol{\pi}\|\mathbf{1}) = \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\pi_{ij}\log(\pi_{ij}) - \pi_{ij} + 1\right) \ , \tag{28}$$

with the convention $0 \log(0) = 0$. Another well-known example is the Itakura-Saito divergence, defined for matrices $\boldsymbol{\pi}, \boldsymbol{\xi} \in \mathbb{R}^{d \times d}_{++}$ as follows:

$$I(\boldsymbol{\pi} \| \boldsymbol{\xi}) = \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\frac{\pi_{ij}}{\xi_{ij}} - \log \left(\frac{\pi_{ij}}{\xi_{ij}} \right) - 1 \right) \ . \tag{29}$$

This divergence is generated by a function of Legendre type for $\boldsymbol{\pi} \in \mathbb{R}^{d \times d}_{++}$ given by minus the Burg entropy:

$$F(\boldsymbol{\pi}) = \sum_{i=1}^{d} \sum_{j=1}^{d} (\pi_{ij} - \log \pi_{ij} - 1) \ . \tag{30}$$

On the one hand, these examples belong to a particular type of so-called separable Bregman divergences between matrices on $\mathbb{R}^{d \times d}$, that can be seen as the aggregation of element-wise Bregman divergences between scalars on $\mathbb{R}$:

$$B_{\phi}(\boldsymbol{\pi} \| \boldsymbol{\xi}) = \sum_{i=1}^{d} \sum_{j=1}^{d} B_{\phi_{ij}}(\pi_{ij} \| \xi_{ij}) \ , \tag{31}$$

$$\phi(\boldsymbol{\pi}) = \sum_{i=1}^{d} \sum_{j=1}^{d} \phi_{ij}(\pi_{ij}) \ . \tag{32}$$

Often, all element-wise generators ϕ_{ij} are chosen equal, and are thus simply written as ϕ with a slight abuse of notation. Other examples of such divergences are discussed in Section 5, and include the logistic loss function generated by minus the Fermi-Dirac entropy, or the squared Euclidean distance generated by the Euclidean norm.

On the other hand, a classical example of non-separable Bregman divergence is half the squared Mahalanobis distance, defined for matrices $\boldsymbol{\pi}, \boldsymbol{\xi} \in \mathbb{R}^{d \times d}$ as follows:

$$M(\boldsymbol{\pi} \| \boldsymbol{\xi}) = \frac{1}{2} \mathrm{vec}(\boldsymbol{\pi} - \boldsymbol{\xi})^{\top} \mathbf{P} \, \mathrm{vec}(\boldsymbol{\pi} - \boldsymbol{\xi}) \ , \tag{33}$$

for a positive-definite matrix $\mathbf{P} \in \mathbb{R}^{d^2 \times d^2}$. This divergence is generated by a function of Legendre type for $\boldsymbol{\pi} \in \mathbb{R}^{d \times d}$ given by a quadratic form:

$$Q(\boldsymbol{\pi}) = \frac{1}{2} \mathrm{vec}(\boldsymbol{\pi})^{\top} \mathbf{P} \, \mathrm{vec}(\boldsymbol{\pi}) \ . \tag{34}$$

This example is also discussed in Section 5.

2.3 Alternate Bregman Projections

Let ϕ be a function of Legendre type with Fenchel conjugate $\phi^{\star} = \psi$. In general, computing Bregman projections onto an arbitrary closed convex set $\mathcal{C} \subseteq \mathcal{E}$ such that $\mathcal{C} \cap \mathrm{int}(\mathrm{dom}\, \phi) \neq \emptyset$ is nontrivial. Sometimes, it is possible to decompose $\mathcal{C}$ into the intersection of finitely many closed convex sets:

$$\mathcal{C} = \bigcap_{l=1}^{s} \mathcal{C}_l \ , \tag{35}$$

where the individual Bregman projections onto the respective sets $\mathcal{C}_1, \ldots, \mathcal{C}_s$ are easier to compute. It is then possible to obtain the Bregman projection onto $\mathcal{C}$ by alternate projections onto $\mathcal{C}_1, \ldots, \mathcal{C}_s$ according to Dykstra's algorithm.

In more detail, let $\sigma\colon \mathbb{N} \to \{1, \dots, s\}$ be a control mapping that determines the sequence of subsets onto which we project. For a given point $\mathbf{x}_0 \in \mathcal{C} \cap \operatorname{int}(\operatorname{dom}\phi)$, the Bregman projection $P_{\mathcal{C}}(\mathbf{x}_0)$ of $\mathbf{x}_0$ onto $\mathcal{C}$ can be approximated with Dykstra's algorithm by iterating the following updates:

$$\mathbf{x}_{k+1} \leftarrow P_{\mathcal{C}_{\sigma(k)}}(\nabla\psi(\nabla\phi(\mathbf{x}_k) + \mathbf{y}^{\sigma(k)})) \ , \tag{36}$$

where the correction terms $\mathbf{y}^1, \dots, \mathbf{y}^s$ for the respective subsets are initialized with the null element of $\mathcal{E}$, and are updated after projection as follows:

$$\mathbf{y}^{\sigma(k)} \leftarrow \mathbf{y}^{\sigma(k)} + \nabla\phi(\mathbf{x}_k) - \nabla\phi(\mathbf{x}_{k+1}) \ . \tag{37}$$

Under some technical conditions, the sequence of updates $(\mathbf{x}_k)_{k\in\mathbb{N}}$ then converges in norm to $P_{\mathcal{C}}(\mathbf{x}_0)$ with a linear rate. Several sets of such conditions have been studied, notably by [53], [5], [19].

We here use the conditions proposed by [19], which reveal to be the less restrictive ones in our framework. Specifically, the convergence of Dykstra's algorithm is guaranteed as soon as the function ϕ is cofinite, the constraint qualification $\operatorname{ri}(\mathcal{C}_1) \cap \dots \cap \operatorname{ri}(\mathcal{C}_s) \cap \operatorname{int}(\operatorname{dom}\phi) \neq \emptyset$ holds, and the control mapping σ is essentially cyclic, that is, there exists a number $t \in \mathbb{N}$ such that σ takes each output value at least once during any t consecutive input values. If a given $\mathcal{C}_l$ is a polyhedral set, then the relative interior can be dropped from the constraint qualification. Hence, when all subsets $\mathcal{C}_l$ are polyhedral, the constraint qualification simply reduces to $\mathcal{C} \cap \operatorname{int}(\operatorname{dom}\phi) \neq \emptyset$, which is already enforced for the definition of Bregman projections.

Finally, if all subsets $\mathcal{C}_l$ are further affine, then we can relax other assumptions. Notably, we do not require ϕ to be cofinite (18), or equivalently $\operatorname{dom}\psi = \mathcal{E}$, but only $\operatorname{dom}\psi$ to be open. The control mapping need not be essentially cyclic anymore, as long as it takes each output value an infinite number of times. More importantly, we can completely drop the correction terms from the updates, leading to a simpler technique known as projections onto convex sets (POCS):

$$\mathbf{x}_{k+1} \leftarrow P_{\mathcal{C}_{\sigma(k)}}(\mathbf{x}_k) \ . \tag{38}$$

2.4 Newton-Raphson Method

Let f be a continuously differentiable scalar function on an open interval $I \subseteq \mathbb{R}$. Assume f is increasing on a non-empty closed interval $[x^-, x^+] \subset I$, and write $y^- = f(x^-)$ and $y^+ = f(x^+)$. Then, for any $y \in [y^-, y^+]$, the equation $f(x) = y$ has at least one solution $x^\star \in [x^-, x^+]$. Such a solution can be approximated by iterative updates according to the Newton-Raphson method:

$$x \leftarrow \max\left\{x^-, \min\left\{x^+, x - \frac{f(x) - y}{f'(x)}\right\}\right\} \ , \tag{39}$$

where the fraction takes infinite values when $f'(x) = 0$ and $f(x) \neq y$, and a null value by convention when $f'(x) = 0$ and $f(x) = y$. It is well-known that the Newton-Raphson method converges to a solution $x^\star$ as soon as x is initialized sufficiently close to $x^\star$. Convergence is then quadratic provided that $f'(x^\star) \neq 0$. However, this local convergence has little importance in practice because it is hard to quantify the required proximity to the solution.

[52] elucidated results on global convergence of the Newton-Raphson method. He proved a necessary and sufficient condition of convergence for an arbitrary value $y \in [y^-, y^+]$ and from any starting point $x \in [x^-, x^+]$. This condition is that for any $a, b \in [x^-, x^+]$, $f(b) > f(a)$ implies:

$$f'(a) + f'(b) > \frac{f(b) - f(a)}{b - a} \ . \tag{40}$$

(A) Affine constraints	(B) Polyhedral constraints
(A1) ϕ is of Legendre type.	(B1) ϕ is of Legendre type.
(A2) $(0,1)^{d\times d} \subseteq \operatorname{dom}\phi$.	(B2) $(0,1)^{d\times d} \subseteq \operatorname{dom}\phi$.
(A3) $\operatorname{dom}\phi \subseteq \mathbb{R}_+^{d\times d}$.	(B3) $\operatorname{dom}\phi \nsubseteq \mathbb{R}_+^{d\times d}$.
(A4) $\operatorname{dom}\psi$ is open.	(B4) $\operatorname{dom}\psi = \mathbb{R}^{d\times d}$.
(A5) $\mathbb{R}_-^{d\times d} \subset \operatorname{dom}\psi$.	

Table 1: Set of assumptions for the considered regularizers ϕ.

In particular, a sufficient condition is that the underlying function f is an increasing convex or increasing concave function on $[x^-, x^+]$, or can be decomposed as the sum of such functions. In addition, if f satisfies the necessary and sufficient condition and is strictly increasing with $f'(x) > 0$ for all $x \in [x^-, x^+]$, then initializing with a boundary point $x^- \neq x^\star$ or $x^+ \neq x^\star$ ensures that the entire sequence of updates is interior to (x^-, x^+), so that we can actually drop the min and max truncation operators in the updates:

$$x \leftarrow x - \frac{f(x) - y}{f'(x)} \ . \tag{41}$$

3 Mathematical Formulation

In this section, we develop a unified framework to define ROT problems. We start by drawing some technical assumptions for our generalized framework to hold (Section 3.1). We then formulate primal ROT problems and study their properties (Section 3.2). We also formulate dual ROT problems and discuss their properties in relation to primal ones (Section 3.3). Finally, we provide some geometrical insights to summarize our developments in the light of information geometry (Section 3.4).

3.1 Technical Assumptions

Some mild technical assumptions are required on the convex regularizer ϕ and its Fenchel conjugate $\psi = \phi^*$ for the proposed framework to hold. Some assumptions relate to required conditions for the definition of Bregman projections and convergence of the algorithms, while others are more specific to ROT problems. In our framework, we also need to distinguish between two situations where the underlying closed convex set can be described as the intersection of either affine subspaces or polyhedral subsets. The two sets of assumptions (A) and (B) are summarized in Table 1.

For the first assumptions (A1) and (B1), we recall that a closed proper convex function is of Legendre type if and only if it is essentially smooth and strictly convex on the interior of its domain (Section 2.1). This is required for the definition of Bregman projections (Section 2.2). In addition, it guarantees the existence of dual coordinate systems on $\operatorname{int}(\operatorname{dom}\phi)$ and $\operatorname{int}(\operatorname{dom}\psi)$ via the homeomorphism $\nabla\phi = \nabla\psi^{-1}$:

$$\boldsymbol{\pi}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta}) \ , \tag{42}$$

$$\boldsymbol{\theta}(\boldsymbol{\pi}) = \nabla\phi(\boldsymbol{\pi}) \ . \tag{43}$$

With a slight abuse of notation, we omit the reparameterization to simply denote corresponding primal and dual parameters by $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.

The second assumptions (A2) and (B2) imply that $\mathrm{ri}(\Pi(\mathbf{p}, \mathbf{q})) \subset \mathrm{dom}\, \phi$ and ensure the constraint qualification $\Pi(\mathbf{p}, \mathbf{q}) \cap \mathrm{int}(\mathrm{dom}\, \phi) \neq \emptyset$ for Bregman projection onto the transport polytope, independently of the input distributions $\mathbf{p}, \mathbf{q}$ as long as they do not have null or unit entries. We assume hereafter that this implicitly holds, and discuss in the practical considerations (Section 4.6) how our methods actually generalize to deal explicitly with null or unit entries in the input distributions.

The third assumptions (A3) and (B3) separate between two cases depending on whether $\mathrm{dom}\, \phi$ lies within the non-negative orthant or not for the alternate Bregman projections (Section 2.3). In the former case, non-negativity is already ensured by the domain of the regularizer, so that the underlying closed convex set is made of two affine subspaces for the row and column sum constraints, and the POCS method can be considered. The fourth assumption (A4) thus requires that $\mathrm{dom}\, \psi$ be open for convergence of this algorithm. In the latter case, there is one additional polyhedral subset for the non-negative constraints and Dykstra's algorithm should be used. The fourth assumption (B4) hence further requires that $\mathrm{dom}\, \psi = \mathbb{R}^{d \times d}$, or equivalently that ϕ be cofinite (18), for convergence. In both cases, we remark that we necessarily have $\mathrm{dom}\, \psi = \mathrm{dom}\, \nabla \psi$.

The fifth assumption (A5) in the affine constraints ensures that $-\boldsymbol{\gamma}/\lambda$ belongs to $\mathrm{dom}\, \nabla \psi$ for definition of ROT problems, independently of the non-negative cost matrix $\boldsymbol{\gamma}$ and positive regularization term λ. Notice that this is already guaranteed by the fourth assumption in the polyhedral constraints. We also show in the sparse extension (Section 4.5) how to deal with infinite entries in the cost matrix $\boldsymbol{\gamma}$ for separable regularizers, so as to enforce null entries in the regularized optimal plan.

On the one hand, some common regularizers under assumptions (A) are the Boltzmann-Shannon entropy associated to the Kullback-Leibler divergence, the Burg entropy associated to the Itakura-Saito divergence, and the Fermi-Dirac entropy associated to the logistic loss function. To solve the underlying ROT problems, we employ our method called ASA based on the POCS technique, where alternate Bregman projections onto the two affine subspaces for the row and column sum constraints are considered (Section 4.3). On the other hand, examples under assumptions (B) include the Euclidean norm associated to the Euclidean distance, and the quadratic form associated to the Mahalanobis distance. For these ROT problems, we use our second method called NASA based on Dykstra's algorithm, where correction terms and a further Bregman projection onto the polyhedral non-negative orthant are needed (Section 4.4).

3.2 Primal Problem

We start our primal formulation with the following lemmas and definition for the RMD.

Lemma 1. *The regularizer ϕ attains its global minimum uniquely at $\boldsymbol{\xi}' = \nabla \psi(\mathbf{0})$.*

Proof. Using the assumptions (A4) and (A5), respectively (B4), we have that $\mathbf{0} \in \mathrm{dom}\, \psi = \mathrm{int}(\mathrm{dom}\, \psi)$. Thus, there exists a unique $\boldsymbol{\xi}' \in \mathrm{int}(\mathrm{dom}\, \phi)$ such that $\nabla \phi(\boldsymbol{\xi}') = \mathbf{0}$, or equivalently $\boldsymbol{\xi}' = \nabla \psi(\mathbf{0})$, via the homeomorphism $\nabla \phi = \nabla \psi^{-1}$ ensured by assumption (A1), respectively (B1). Hence, ϕ attains its global minimum uniquely at $\boldsymbol{\xi}'$ by strict convexity on $\mathrm{int}(\mathrm{dom}\, \phi)$. □

Lemma 2. *The restriction of the regularizer ϕ to the transport polytope $\Pi(\mathbf{p}, \mathbf{q})$ attains its global minimum uniquely at the Bregman projection $\boldsymbol{\pi}'$ of $\boldsymbol{\xi}'$ onto $\Pi(\mathbf{p}, \mathbf{q})$.*

Proof. Using the assumption (A2), respectively (B2), we have that $\Pi(\mathbf{p}, \mathbf{q}) \cap \mathrm{int}(\mathrm{dom}\, \phi) \neq \emptyset$. Since $\boldsymbol{\xi}' \in \mathrm{int}(\mathrm{dom}\, \phi)$ and $\Pi(\mathbf{p}, \mathbf{q})$ is a closed convex set, the Bregman projection $\boldsymbol{\pi}'$ of $\boldsymbol{\xi}'$ onto $\Pi(\mathbf{p}, \mathbf{q})$ according to the function ϕ of Legendre type is well-defined. Moreover, it is characterized

by the variational relation (23) as follows:

$$\langle \boldsymbol{\pi} - \boldsymbol{\pi}', \nabla\phi(\boldsymbol{\pi}')\rangle \geq 0 \ , \tag{44}$$

for all $\boldsymbol{\pi} \in \Pi(\mathbf{p},\mathbf{q}) \cap \operatorname{dom}\phi$. We also have $B_\phi(\boldsymbol{\pi}\|\boldsymbol{\pi}') > 0$ when $\boldsymbol{\pi} \neq \boldsymbol{\pi}'$ by strict convexity of ϕ on $\operatorname{int}(\operatorname{dom}\phi)$. As a result, we have:

$$\phi(\boldsymbol{\pi}) - \phi(\boldsymbol{\pi}') > \langle \boldsymbol{\pi} - \boldsymbol{\pi}', \nabla\phi(\boldsymbol{\pi}')\rangle \ . \tag{45}$$

Combining the two inequalities, we obtain $\phi(\boldsymbol{\pi}) > \phi(\boldsymbol{\pi}')$ and the restriction of ϕ to $\Pi(\mathbf{p},\mathbf{q})$ attains its global minimum uniquely at $\boldsymbol{\pi}'$. □

Lemma 3. *The restriction of the cost $\langle\cdot,\boldsymbol{\gamma}\rangle$ to the regularized transport polytope:*

$$\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q}) = \{\boldsymbol{\pi} \in \Pi(\mathbf{p},\mathbf{q})\colon \phi(\boldsymbol{\pi}) \leq \phi(\boldsymbol{\pi}') + \alpha\} \ , \tag{46}$$

where $\alpha \geq 0$, attains its global minimum.

Proof. The regularized transport polytope is the intersection of the compact set $\Pi(\mathbf{p},\mathbf{q})$ with a lower level set of ϕ which is also closed since ϕ is closed. Hence, $\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q})$ is compact and the restriction of $\langle\cdot,\boldsymbol{\gamma}\rangle$ to $\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q})$ attains its global minimum by continuity on a compact set. □

Definition 1. *The primal rot mover's distance is the quantity defined as:*

$$d'_{\boldsymbol{\gamma},\alpha,\phi}(\mathbf{p},\mathbf{q}) = \min_{\boldsymbol{\pi}\in\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q})} \langle \boldsymbol{\pi},\boldsymbol{\gamma}\rangle \ . \tag{47}$$

A minimizer ${\boldsymbol{\pi}'}^\star_\alpha$ is then called a primal rot mover's plan.

Remark 1. For the sake of notation, we omit the dependence on $\mathbf{p},\mathbf{q},\boldsymbol{\gamma},\phi$ in the index of primal rot mover's plans ${\boldsymbol{\pi}'}^\star_\alpha$.

The regularization enforces the associated minimizers to have small enough Bregman information $\phi({\boldsymbol{\pi}'}^\star_\alpha) \leq \phi(\boldsymbol{\pi}') + \alpha$ compared to the minimal one $\phi(\boldsymbol{\pi}')$ for transport plans. We also have a geometrical interpretation where the solutions are constrained to a Bregman ball whose center $\boldsymbol{\xi}'$ is the matrix with minimal Bregman information.

Proposition 4. *The regularized transport polytope is the intersection of the transport polytope with the Bregman ball of radius $B_\phi(\boldsymbol{\pi}'\|\boldsymbol{\xi}') + \alpha$ and center $\boldsymbol{\xi}'$:*

$$\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q}) = \{\boldsymbol{\pi} \in \Pi(\mathbf{p},\mathbf{q})\colon B_\phi(\boldsymbol{\pi}\|\boldsymbol{\xi}') \leq B_\phi(\boldsymbol{\pi}'\|\boldsymbol{\xi}') + \alpha\} \ . \tag{48}$$

Proof. Expanding the Bregman divergences from their definition (19), we obtain:

$$B_\phi(\boldsymbol{\pi}\|\boldsymbol{\xi}') = \phi(\boldsymbol{\pi}) - \phi(\boldsymbol{\xi}') - \langle \boldsymbol{\pi} - \boldsymbol{\xi}', \nabla\phi(\boldsymbol{\xi}')\rangle \ , \tag{49}$$

$$B_\phi(\boldsymbol{\pi}'\|\boldsymbol{\xi}') = \phi(\boldsymbol{\pi}') - \phi(\boldsymbol{\xi}') - \langle \boldsymbol{\pi}' - \boldsymbol{\xi}', \nabla\phi(\boldsymbol{\xi}')\rangle \ . \tag{50}$$

Since $\nabla\phi(\boldsymbol{\xi}') = \mathbf{0}$, the last terms with scalar products vanish, leading to:

$$\phi(\boldsymbol{\pi}) - \phi(\boldsymbol{\pi}') = B_\phi(\boldsymbol{\pi}\|\boldsymbol{\xi}') - B_\phi(\boldsymbol{\pi}'\|\boldsymbol{\xi}') \ . \tag{51}$$

Therefore, in the definition (46) of $\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q})$, we have $\phi(\boldsymbol{\pi}) \leq \phi(\boldsymbol{\pi}') + \alpha$ if and only if $\boldsymbol{\pi}$ is in the Bregman ball of radius $B_\phi(\boldsymbol{\pi}'\|\boldsymbol{\xi}') + \alpha$ and center $\boldsymbol{\xi}'$. □

Under some additional conditions, this geometrical interpretation still holds with a Bregman ball whose center $\boldsymbol{\pi}'$ has minimal Bregman information for transport plans.

Proposition 5. *If $\boldsymbol{\pi}' \in \mathrm{ri}(\Pi(\mathbf{p},\mathbf{q}))$, then the regularized transport polytope is the intersection of the transport polytope with the Bregman ball of radius α and center $\boldsymbol{\pi}'$:*

$$\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q}) = \{\boldsymbol{\pi} \in \Pi(\mathbf{p},\mathbf{q}) \colon B_\phi(\boldsymbol{\pi}\|\boldsymbol{\pi}') \leq \alpha\} \ . \tag{52}$$

Proof. Since $\boldsymbol{\pi}' \in \mathrm{ri}(\Pi(\mathbf{p},\mathbf{q}))$, there is equality in the generalized Pythagorean theorem (26):

$$B_\phi(\boldsymbol{\pi}\|\boldsymbol{\xi}') = B_\phi(\boldsymbol{\pi}\|\boldsymbol{\pi}') + B_\phi(\boldsymbol{\pi}'\|\boldsymbol{\xi}') \ . \tag{53}$$

The regularized transport polytope as seen from (48) is then the intersection of the transport polytope $\Pi(\mathbf{p},\mathbf{q})$ with the Bregman ball of radius α and center $\boldsymbol{\pi}'$. □

Remark 2. The proposition also holds trivially when the global minimum is attained on the transport polytope, that is, when $\boldsymbol{\xi}' = \boldsymbol{\pi}'$.

Corollary 6. *Under assumptions (A), the regularized transport polytope is the intersection of the transport polytope with the Bregman ball of radius α and center $\boldsymbol{\pi}'$:*

$$\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q}) = \{\boldsymbol{\pi} \in \Pi(\mathbf{p},\mathbf{q}) \colon B_\phi(\boldsymbol{\pi}\|\boldsymbol{\pi}') \leq \alpha\} \ . \tag{54}$$

Proof. This is a result of $\boldsymbol{\pi}' \in \Pi(\mathbf{p},\mathbf{q}) \cap \mathrm{int}(\mathrm{dom}\,\phi) = \mathrm{ri}(\Pi(\mathbf{p},\mathbf{q}))$ when $\mathrm{dom}\,\phi \subseteq \mathbb{R}_+^{d\times d}$. Indeed, we then have $\mathrm{ri}(\Pi(\mathbf{p},\mathbf{q})) \subset \Pi(\mathbf{p},\mathbf{q})$ and $\mathrm{ri}(\Pi(\mathbf{p},\mathbf{q})) \subset \mathrm{int}(\mathrm{dom}\,\phi)$, so that $\mathrm{ri}(\Pi(\mathbf{p},\mathbf{q})) \subseteq \Pi(\mathbf{p},\mathbf{q}) \cap \mathrm{int}(\mathrm{dom}\,\phi)$. Conversely, let $\boldsymbol{\pi} \in \Pi(\mathbf{p},\mathbf{q}) \cap \mathrm{int}(\mathrm{dom}\,\phi)$ so that $\boldsymbol{\pi} \in \mathbb{R}_{++}^{d\times d}$. Then, for a given $\overline{\boldsymbol{\pi}} \in \Pi(\mathbf{p},\mathbf{q})$, let us pose $\boldsymbol{\pi}_\lambda = \lambda\boldsymbol{\pi} + (1-\lambda)\overline{\boldsymbol{\pi}}$ for $\lambda > 1$. We easily have $\boldsymbol{\pi}_\lambda \mathbf{1} = \mathbf{p}$ and $\boldsymbol{\pi}_\lambda^\top \mathbf{1} = \mathbf{q}$. Moreover, since all entries of $\boldsymbol{\pi}$ are positive and that of $\overline{\boldsymbol{\pi}}$ are non-negative, we can always choose a given λ sufficiently close to 1 such that $\boldsymbol{\pi}_\lambda \in \mathbb{R}_+^{d\times d}$. We then have $\boldsymbol{\pi}_\lambda \in \Pi(\mathbf{p},\mathbf{q})$ so that $\boldsymbol{\pi} \in \mathrm{ri}(\Pi(\mathbf{p},\mathbf{q}))$ as characterized by (15), and thus $\Pi(\mathbf{p},\mathbf{q}) \cap \mathrm{int}(\mathrm{dom}\,\phi) \subseteq \mathrm{ri}(\Pi(\mathbf{p},\mathbf{q}))$. □

Remark 3. Under assumptions (B), the Bregman projection $\boldsymbol{\pi}'$ does not necessarily lie within $\mathrm{ri}(\Pi(\mathbf{p},\mathbf{q}))$. Hence, the geometrical interpretation in terms of a Bregman ball might break down, although the solutions are still constrained to have a small enough Bregman information above that of $\boldsymbol{\pi}'$.

Although Sinkhorn distances verify the triangular inequality when $\boldsymbol{\gamma}$ is a distance matrix, thanks to specific chain rules and information inequalities for the Bolzmann-Shannon entropy and Kullback-Leibler divergence, it is not necessarily the case for the RMD with other regularizations, even for separable regularizers. Hence, the RMD does not provide a true distance metric on Σ_d in general even if $\boldsymbol{\gamma}$ is a distance matrix. Nonetheless, the RMD is symmetric as soon as ϕ is invariant by transposition, which holds for separable regularizers $\phi_{ij} = \phi$, and $\boldsymbol{\gamma}$ is symmetric. We now study some properties of the RMD that hold for general regularizers.

Property 1. *The primal rot mover's distance $d'_{\boldsymbol{\gamma},\alpha,\phi}(\mathbf{p},\mathbf{q})$ is a decreasing convex and continuous function of α.*

Proof. The fact that it is decreasing is a direct consequence of the regularized transport polytope $\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q})$ growing with α. The convexity can be proved as follows. Let $\alpha_0, \alpha_1 \geq 0$, and $0 < \lambda < 1$. We pose $\alpha_\lambda = (1-\lambda)\alpha_0 + \lambda\alpha_1 \geq 0$. We also choose arbitrary rot mover's plans $\boldsymbol{\pi}'^\star_{\alpha_0}, \boldsymbol{\pi}'^\star_{\alpha_1}, \boldsymbol{\pi}'^\star_{\alpha_\lambda}$. We finally pose $\boldsymbol{\pi}_\lambda = (1-\lambda)\boldsymbol{\pi}'^\star_{\alpha_0} + \lambda\boldsymbol{\pi}'^\star_{\alpha_1}$. By convexity of ϕ, we have:

$$\begin{aligned}
\phi(\boldsymbol{\pi}_\lambda) &\leq (1-\lambda)\phi(\boldsymbol{\pi}'^\star_{\alpha_0}) + \lambda\phi(\boldsymbol{\pi}'^\star_{\alpha_1}) && (55)\\
&\leq (1-\lambda)(\alpha_0 + \phi(\boldsymbol{\pi}')) + \lambda(\alpha_1 + \phi(\boldsymbol{\pi}')) && (56)\\
&= \alpha_\lambda + \phi(\boldsymbol{\pi}') \ . && (57)
\end{aligned}$$

Hence, $\boldsymbol{\pi}_\lambda \in \Pi_{\alpha_\lambda,\phi}(\mathbf{p},\mathbf{q})$, and by construction we have $\langle {\boldsymbol{\pi}'}^\star_{\alpha_\lambda}, \boldsymbol{\gamma}\rangle \leq \langle \boldsymbol{\pi}_\lambda, \boldsymbol{\gamma}\rangle$, or equivalently:

$$\langle {\boldsymbol{\pi}'}^\star_{\alpha_\lambda}, \boldsymbol{\gamma}\rangle \leq (1-\lambda)\langle {\boldsymbol{\pi}'}^\star_{\alpha_0}, \boldsymbol{\gamma}\rangle + \lambda\langle {\boldsymbol{\pi}'}^\star_{\alpha_1}, \boldsymbol{\gamma}\rangle \ . \tag{58}$$

The continuity for $\alpha > 0$ is a direct consequence of convexity for $\alpha > 0$, since a convex function is always continuous on the relative interior of its domain. Lastly, the continuity at $\alpha = 0$ can be seen as follows. Let $(\alpha_k)_{k\in\mathbb{N}}$ be a sequence of positive numbers that converges to 0. We choose arbitrary rot mover's plans $({\boldsymbol{\pi}'}^\star_{\alpha_k})_{k\in\mathbb{N}}$. By compactness of $\Pi(\mathbf{p},\mathbf{q})$, we can extract a subsequence of rot mover's plans that converges in norm to a point $\boldsymbol{\pi}'^\star \in \Pi(\mathbf{p},\mathbf{q})$. For the sake of simplicity, we do not relabel this subsequence. By construction, we have $\phi(\boldsymbol{\pi}') \leq \phi({\boldsymbol{\pi}'}^\star_{\alpha_k}) \leq \phi(\boldsymbol{\pi}') + \alpha_k$, and $\phi({\boldsymbol{\pi}'}^\star_{\alpha_k})$ converges to $\phi(\boldsymbol{\pi}')$. By lower semi-continuity of ϕ, we thus have $\phi(\boldsymbol{\pi}'^\star) \leq \phi(\boldsymbol{\pi}')$. Since the global minimum of ϕ on $\Pi(\mathbf{p},\mathbf{q})$ is attained uniquely at $\boldsymbol{\pi}'$, we must have $\boldsymbol{\pi}'^\star = \boldsymbol{\pi}'$, and the original sequence also converges in norm to $\boldsymbol{\pi}'$. By continuity of the total cost $\langle\cdot,\boldsymbol{\gamma}\rangle$ on $\mathbb{R}^{d\times d}$, $\langle {\boldsymbol{\pi}'}^\star_{\alpha_k}, \boldsymbol{\gamma}\rangle$ converges to $\langle \boldsymbol{\pi}', \boldsymbol{\gamma}\rangle$. Hence, the limit of the RMD when α tends to 0 from above is $\langle \boldsymbol{\pi}', \boldsymbol{\gamma}\rangle$, which equals the RMD for $\alpha = 0$ as shown in the next property. □

Property 2. *When $\alpha = 0$, the primal rot mover's distance reduces to:*

$$d'_{\boldsymbol{\gamma},0,\phi}(\mathbf{p},\mathbf{q}) = \langle \boldsymbol{\pi}', \boldsymbol{\gamma}\rangle \ , \tag{59}$$

and the unique primal rot mover's plan is the transport plan with minimal Bregman information:

$${\boldsymbol{\pi}'}^\star_0 = \boldsymbol{\pi}' \ . \tag{60}$$

Proof. Since $\boldsymbol{\pi}'$ is the unique global minimizer of ϕ on $\Pi(\mathbf{p},\mathbf{q})$, the regularized transport polytope reduces to the singleton $\Pi_{0,\phi}(\mathbf{p},\mathbf{q}) = \{\boldsymbol{\pi} \in \Pi(\mathbf{p},\mathbf{q})\colon \phi(\boldsymbol{\pi}) \leq \phi(\boldsymbol{\pi}')\} = \{\boldsymbol{\pi}'\}$. The property follows immediately. □

Property 3. *When α tends to $+\infty$, the primal rot mover's distance converges to the earth mover's distance:*

$$\lim_{\alpha\to+\infty} d'_{\boldsymbol{\gamma},\alpha,\phi}(\mathbf{p},\mathbf{q}) = d_{\boldsymbol{\gamma}}(\mathbf{p},\mathbf{q}) \ . \tag{61}$$

Proof. Let $\boldsymbol{\pi}^\star \in \Pi(\mathbf{p},\mathbf{q})$ be an earth mover's plan so that $d_{\boldsymbol{\gamma}}(\mathbf{p},\mathbf{q}) = \langle \boldsymbol{\pi}^\star, \boldsymbol{\gamma}\rangle$. By continuity of the total cost $\langle\cdot,\boldsymbol{\gamma}\rangle$ on $\mathbb{R}^{d\times d}$, we have that for all $\epsilon > 0$, there exists an open neighborhood of $\boldsymbol{\pi}^\star$ such that $\langle \boldsymbol{\pi}, \boldsymbol{\gamma}\rangle \leq \langle \boldsymbol{\pi}^\star, \boldsymbol{\gamma}\rangle + \epsilon$ for any transport plan $\boldsymbol{\pi}$ within this neighborhood. We can always choose a transport plan such that $\boldsymbol{\pi} \in \mathrm{ri}(\Pi(\mathbf{p},\mathbf{q}))$. Since $\mathrm{ri}(\Pi(\mathbf{p},\mathbf{q})) \subset \mathrm{dom}\,\phi$, $\phi(\boldsymbol{\pi})$ is finite and we can fix $\alpha_\epsilon = \phi(\boldsymbol{\pi}) - \phi(\boldsymbol{\pi}') \geq 0$. Hence, $\boldsymbol{\pi} \in \Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q})$ for any $\alpha \geq \alpha_\epsilon$, and we have $d_{\boldsymbol{\gamma}}(\mathbf{p},\mathbf{q}) \leq d'_{\boldsymbol{\gamma},\alpha,\phi}(\mathbf{p},\mathbf{q}) \leq \langle \boldsymbol{\pi}, \boldsymbol{\gamma}\rangle \leq d_{\boldsymbol{\gamma}}(\mathbf{p},\mathbf{q}) + \epsilon$. □

Property 4. *If $[0,1)^{d\times d} \subseteq \mathrm{dom}\,\phi$, then there exists a minimal $\alpha' \geq 0$ such that for all $\alpha \geq \alpha'$, the primal rot mover's distance reduces to the earth mover's distance:*

$$d'_{\boldsymbol{\gamma},\alpha,\phi}(\mathbf{p},\mathbf{q}) = d_{\boldsymbol{\gamma}}(\mathbf{p},\mathbf{q}) \ . \tag{62}$$

Proof. The extra condition guarantees that $\Pi(\mathbf{p},\mathbf{q}) \subset \mathrm{dom}\,\phi$, and thus that ϕ is bounded on the closed set $\Pi(\mathbf{p},\mathbf{q})$. The property is then a direct consequence of $\Pi_{\alpha,\phi}(\mathbf{p},\mathbf{q}) = \Pi(\mathbf{p},\mathbf{q})$ for α large enough. □

Property 5. *If $[0,1)^{d\times d} \subseteq \mathrm{dom}\,\phi$ and ϕ is strictly convex on $[0,1)^{d\times d}$, then the unique primal rot mover's plan for $\alpha = \alpha'$ is the earth mover's plan $\boldsymbol{\pi}^\star_0$ with minimal Bregman information:*

$${\boldsymbol{\pi}'}^\star_{\alpha'} = \boldsymbol{\pi}^\star_0 \ . \tag{63}$$

Proof. First, we recall that the set of earth mover's plans $\boldsymbol{\pi}^\star$ is either a single vertex or a whole facet of $\Pi(\mathbf{p}, \mathbf{q})$. Hence, it forms a closed convex subset in $\Pi(\mathbf{p}, \mathbf{q})$, and there is a unique earth mover's plan $\boldsymbol{\pi}_0^\star$ with minimal Bregman information by strict convexity of ϕ on this subset. Second, it is trivial that all primal rot mover's plan $\boldsymbol{\pi}'^\star_{\alpha'}$ must be earth mover's plans. If there is a single vertex as earth mover's plan, then the property follows immediately. Otherwise, we can see the property geometrically as follows. The whole facet of earth mover's plans is orthogonal to $\boldsymbol{\gamma}$. Nevertheless, by strict convexity of ϕ on $[0,1)^{d\times d}$, the facet must be tangent to $\Pi_{\alpha',\phi}(\mathbf{p}, \mathbf{q})$ at the unique earth mover's plan $\boldsymbol{\pi}_0^\star$ with minimal Bregman information $\phi(\boldsymbol{\pi}_0^\star) = \phi(\boldsymbol{\pi}') + \alpha'$, and $\boldsymbol{\pi}_0^\star$ is also the rot mover's plan $\boldsymbol{\pi}'^\star_{\alpha'}$. Another way to prove the property more formally is as follows. Suppose that a primal rot mover's plan $\boldsymbol{\pi}^\star$ is not the earth mover's plan with minimal Bregman information. We thus have $\phi(\boldsymbol{\pi}_0^\star) < \phi(\boldsymbol{\pi}^\star) \leq \phi(\boldsymbol{\pi}') + \alpha'$. We can then choose a smaller α' such that $\phi(\boldsymbol{\pi}_0^\star) \leq \phi(\boldsymbol{\pi}') + \alpha'$ and the RMD still equals the EMD for this smaller value, and actually all values in between by monotonicity. This leads to a contradiction and $\boldsymbol{\pi}_0^\star$ must be the earth mover's plan with minimal Bregman information. □

Remark 4. When $\alpha > \alpha'$, the regularized transport polytope might grow to include several earth mover's plans with different Bregman information, which are then all minimizers for the RMD. When we do not have strict convexity outside $(0,1)^{d\times d}$, there might also be multiple earth mover's plans with minimal Bregman information.

If $[0,1)^{d\times d} \subseteq \operatorname{dom}\phi$, then it is easy to check that the strict convexity of ϕ on $[0,1)^{d\times d}$ is always verified when ϕ is separable under assumptions (A) or (B), or when $[0,1)^{d\times d} \subset \operatorname{int}(\operatorname{dom}\phi)$ under assumptions (B). This holds for almost all typical regularizers, notably for all regularizers considered in this paper except from minus the Burg entropy as defined in (30) and associated to the Itakura-Saito divergence in (29). For this latter regularizer, the solutions for an increasing α all lie within $\operatorname{ri}(\Pi(\mathbf{p}, \mathbf{q}))$, and the RMD never reaches the EMD. In such cases where the minimal α' does not exist, we can use the convention $\alpha' = +\infty$ since the RMD always converges to the EMD in the limit when α tends to $+\infty$. We can then prove that there is a unique rot mover's plan $\boldsymbol{\pi}'^\star_\alpha$ as long as $0 < \alpha < \alpha'$, which can be seen informally as follows. The solutions geometrically lie at the intersection of $\Pi_{\alpha,\phi}(\mathbf{p}, \mathbf{q})$ and of a supporting hyperplane with normal $\boldsymbol{\gamma}$. By strict convexity of ϕ on $\operatorname{ri}(\Pi(\mathbf{p}, \mathbf{q}))$, this intersection is a singleton inside the polytope. When the intersection reaches a facet, the only facet that can coincide locally with the hyperplane is the one that contains the earth mover's plans. Hence, we also have a singleton on the boundary of the polytope before reaching an earth mover's plan. We formally prove this uniqueness result next by exploiting duality.

3.3 Dual Problem

We now present the following two lemmas before defining our dual formulation for the RMD.

Lemma 7. *The regularized cost $\langle \cdot, \boldsymbol{\gamma}\rangle + \lambda\phi(\cdot)$, where $\lambda > 0$, attains its global minimum uniquely at $\boldsymbol{\xi} = \nabla\psi(-\boldsymbol{\gamma}/\lambda)$.*

Proof. The regularized cost is convex with same domain as ϕ, and is strictly convex on $\operatorname{int}(\operatorname{dom}\phi)$. Thus, it attains its global minimum at a unique point $\boldsymbol{\xi} \in \operatorname{int}(\operatorname{dom}\phi)$ if and only if $\boldsymbol{\gamma} + \lambda\nabla\phi(\boldsymbol{\xi}) = \mathbf{0}$, or equivalently $\nabla\phi(\boldsymbol{\xi}) = -\boldsymbol{\gamma}/\lambda$. By assumptions (A4) and (A5), respectively (B4), $-\boldsymbol{\gamma}/\lambda \in \operatorname{dom}\nabla\psi$, so that the global minimum is attained uniquely at $\boldsymbol{\xi} = \nabla\psi(-\boldsymbol{\gamma}/\lambda)$ in virtue of the homeomorphism in (42) and (43). □

Lemma 8. *The restriction of the regularized cost $\langle \cdot, \boldsymbol{\gamma}\rangle + \lambda\phi(\cdot)$ to the transport polytope $\Pi(\mathbf{p}, \mathbf{q})$ attains its global minimum uniquely.*

Proof. We notice that the regularized cost is equal to a Bregman divergence up to a positive factor and additive constant:

$$\langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda \phi(\boldsymbol{\pi}) - \lambda \phi(\boldsymbol{\xi}) = \lambda B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}) \ . \tag{64}$$

Hence, its minimization over the closed convex set $\Pi(\mathbf{p}, \mathbf{q})$ is equivalent to the Bregman projection of $\boldsymbol{\xi} \in \operatorname{int}(\operatorname{dom} \phi)$ onto $\Pi(\mathbf{p}, \mathbf{q})$ according to the function ϕ of Legendre type. Since $\Pi(\mathbf{p}, \mathbf{q}) \cap \operatorname{int}(\operatorname{dom} \phi) \neq \emptyset$, this projection exists and is unique. □

Definition 2. *The dual rot mover's distance is the quantity defined as:*

$$d_{\boldsymbol{\gamma}, \lambda, \phi}(\mathbf{p}, \mathbf{q}) = \langle \boldsymbol{\pi}^\star_\lambda, \boldsymbol{\gamma} \rangle \ , \tag{65}$$

where the dual rot mover's plan $\boldsymbol{\pi}^\star_\lambda$ is given by:

$$\boldsymbol{\pi}^\star_\lambda = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \ \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda \phi(\boldsymbol{\pi}) \ . \tag{66}$$

Remark 5. For the sake of notation, we omit the dependence on $\mathbf{p}, \mathbf{q}, \boldsymbol{\gamma}, \phi$ in the index of dual rot mover's plans $\boldsymbol{\pi}^\star_\lambda$.

We proceed with the following proposition that enlightens the relation between the RMD and associated Bregman divergence.

Proposition 9. *The dual rot mover's plan is the Bregman projection of $\boldsymbol{\xi}$ onto the transport polytope:*

$$\boldsymbol{\pi}^\star_\lambda = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \ B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}) \ . \tag{67}$$

Proof. This is a consequence of the proof for Lemma 8. Indeed, from the definition in (66), we see that the rot mover's plan also minimizes (64). Therefore, it is the unique Bregman projection of $\boldsymbol{\xi}$ onto the transport polytope. □

We have a geometrical interpretation where the regularization shrinks the solution toward the matrix $\boldsymbol{\xi}'$ that has minimal Bregman information.

Proposition 10. *The dual rot mover's plan $\boldsymbol{\pi}^\star_\lambda$ can be obtained as:*

$$\boldsymbol{\pi}^\star_\lambda = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \ \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}') \ . \tag{68}$$

Proof. Developing the Bregman divergence based on its definition (19), we have:

$$B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}') = \phi(\boldsymbol{\pi}) - \phi(\boldsymbol{\xi}') - \langle \boldsymbol{\pi} - \boldsymbol{\xi}', \nabla \phi(\boldsymbol{\xi}') \rangle \ . \tag{69}$$

Since $\nabla \phi(\boldsymbol{\xi}') = \mathbf{0}$, the last term with scalar product vanishes and we are left out with $\phi(\boldsymbol{\pi})$ plus a constant term with respect to $\boldsymbol{\pi}$. Hence, we can replace $\phi(\boldsymbol{\pi})$ by $B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}')$ in the minimization (66) that defines $\boldsymbol{\pi}^\star_\lambda$. □

Under some additional conditions, this interpretation can also be seen as shrinking toward the transport plan $\boldsymbol{\pi}'$ with minimal Bregman information.

Proposition 11. *If $\boldsymbol{\pi}' \in \operatorname{ri}(\Pi(\mathbf{p}, \mathbf{q}))$, then the dual rot mover's plan $\boldsymbol{\pi}^\star_\lambda$ can be obtained as:*

$$\boldsymbol{\pi}^\star_\lambda = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})}{\operatorname{argmin}} \ \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda B_\phi(\boldsymbol{\pi} \| \boldsymbol{\pi}') \ . \tag{70}$$

Proof. If $\boldsymbol{\pi}' \in \mathrm{ri}(\Pi(\mathbf{p}, \mathbf{q}))$, then we have equality in the generalized Pythagorean theorem (26), leading to:

$$B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}') = B_\phi(\boldsymbol{\pi} \| \boldsymbol{\pi}') + B_\phi(\boldsymbol{\pi}' \| \boldsymbol{\xi}') \ . \tag{71}$$

Since the last term is constant with respect to $\boldsymbol{\pi}$, we can replace $B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}')$ by $B_\phi(\boldsymbol{\pi} \| \boldsymbol{\pi}')$ in the minimization (68) that characterizes $\boldsymbol{\pi}^\star_\lambda$. □

Remark 6. The proposition also holds trivially when the global minimum is attained on the transport polytope, that is, when $\boldsymbol{\xi}' = \boldsymbol{\pi}'$.

Corollary 12. *Under assumptions (A), the dual rot mover's plan $\boldsymbol{\pi}^\star_\lambda$ can be obtained as:*

$$\boldsymbol{\pi}^\star_\lambda = \operatorname*{argmin}_{\boldsymbol{\pi} \in \Pi(\mathbf{p}, \mathbf{q})} \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda B_\phi(\boldsymbol{\pi} \| \boldsymbol{\pi}') \ . \tag{72}$$

Proof. This is a result of $\boldsymbol{\pi}' \in \Pi(\mathbf{p}, \mathbf{q}) \cap \mathrm{int}(\mathrm{dom}\, \phi) = \mathrm{ri}(\Pi(\mathbf{p}, \mathbf{q}))$ when $\mathrm{dom}\, \phi \subseteq \mathbb{R}_+^{d \times d}$, as shown in the proof of Corollary 6. □

In the sequel, we also extend naturally the definition of the dual RMD for $\lambda = 0$ as the EMD. We then do not necessarily have uniqueness of dual rot mover's plans for $\lambda = 0$, and the geometrical interpretation in terms of a Bregman projection does not hold anymore for $\lambda = 0$. However, we have the following theorem based on duality theory that shows the equivalence between primal and dual ROT problems.

Theorem 13. *For all $\alpha > 0$, there exists $\lambda \geq 0$ such that the primal and dual rot mover's distances are equal:*

$$d'_{\boldsymbol{\gamma}, \alpha, \phi}(\mathbf{p}, \mathbf{q}) = d_{\boldsymbol{\gamma}, \lambda, \phi}(\mathbf{p}, \mathbf{q}) \ . \tag{73}$$

Moreover, if $\alpha < \alpha'$, then a corresponding value is such that $\lambda > 0$, and the primal and dual rot mover's plans are unique and equal:

$$\boldsymbol{\pi}'^\star_\alpha = \boldsymbol{\pi}^\star_\lambda \ . \tag{74}$$

Proof. The primal problem can be seen as the minimization $p^\star$ of the cost $\langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle$ on $\Pi(\mathbf{p}, \mathbf{q})$ subject to $\phi(\boldsymbol{\pi}) - \phi(\boldsymbol{\pi}') - \alpha \leq 0$. The domain of this constrained convex problem is $\mathcal{D} = \Pi(\mathbf{p}, \mathbf{q}) \cap \mathrm{dom}\, \phi \neq \emptyset$. The Lagrangian on $\mathcal{D} \times \mathbb{R}$ is given by $\mathcal{L}(\boldsymbol{\pi}, \lambda) = \langle \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle + \lambda(\phi(\boldsymbol{\pi}) - \phi(\boldsymbol{\pi}') - \alpha)$, and its minimization over $\mathcal{D}$ for a fixed $\lambda \geq 0$ has the same solutions $\boldsymbol{\pi}^\star$ as the dual problem. In addition, Slater's condition for convex problems, stating that there is a strictly feasible point in the relative interior of the domain, is verified as long as $\alpha > 0$. Indeed, we have $\mathrm{ri}(\mathcal{D}) = \mathrm{ri}(\Pi(\mathbf{p}, \mathbf{q}))$. The existence of a strictly feasible point $\phi(\boldsymbol{\pi}) < \phi(\boldsymbol{\pi}') + \alpha$ then holds by continuity of ϕ at $\boldsymbol{\pi}' \in \mathrm{int}(\mathrm{dom}\, \phi)$. As a result, we have strong duality with a zero duality gap $p^\star = d^\star$, where $d^\star$ is the maximization of $g(\lambda)$ subject to $\lambda \geq 0$. Moreover, if $d^\star$ is finite, then it is attained at least once at a point $\lambda^\star$. This is the case since we already know that $p^\star$ is finite. Since $p^\star$ is also attained at least once at a point $\boldsymbol{\pi}^\star$ solution of the primal problem, we have the following chain:

$$\begin{aligned}
p^\star &= d^\star && (75)\\
&= \min_{\boldsymbol{\pi} \in \mathcal{D}} \mathcal{L}(\boldsymbol{\pi}, \lambda^\star) && (76)\\
&\leq \mathcal{L}(\boldsymbol{\pi}^\star, \lambda^\star) && (77)\\
&= \langle \boldsymbol{\pi}^\star, \boldsymbol{\gamma} \rangle + \lambda^\star(\phi(\boldsymbol{\pi}^\star) - \phi(\boldsymbol{\pi}') - \alpha) && (78)\\
&\leq \langle \boldsymbol{\pi}^\star, \boldsymbol{\gamma} \rangle && (79)\\
&= p^\star \ . && (80)
\end{aligned}$$

Therefore, all inequalities are in fact equalities, $\boldsymbol{\pi}^\star$ also minimizes the Lagrangian over $\mathcal{D}$ and thus is a solution of the dual problem. In other words, the primal and dual RMD for α and $\lambda^\star$ are equal, and the primal solutions must be dual solutions too. For $\alpha < \alpha'$, the RMD has not reached the EMD yet, and thus we must have $\lambda^\star > 0$. Hence, the dual solution is unique, so that the primal solution is unique too and equal to the dual one. □

Remark 7. Corresponding values of α and λ depend on $\mathbf{p}, \mathbf{q}, \boldsymbol{\gamma}, \phi$. In addition, there might be multiple values of λ that correspond to a given α.

Again, the RMD does not verify the triangular inequality in general, and hence does not provide a true distance metric on Σ_d even if $\boldsymbol{\gamma}$ is a distance matrix. Nevertheless, we still have the result that the RMD is symmetric as soon as ϕ is invariant by transposition, which holds for separable regularizers $\phi_{ij} = \phi$, and $\boldsymbol{\gamma}$ is symmetric. We also obtain properties for the dual RMD that are similar to the ones for the primal RMD.

Property 6. *The dual rot mover's distance $d_{\boldsymbol{\gamma},\lambda,\phi}(\mathbf{p}, \mathbf{q})$ is an increasing and continuous function of λ.*

Proof. The fact that it is increasing can be seen as follows. Let $0 \leq \lambda_1 < \lambda_2$. By construction, we have the following inequalities:

$$\langle \boldsymbol{\pi}^\star_{\lambda_1}, \boldsymbol{\gamma} \rangle + \lambda_1 \phi(\boldsymbol{\pi}^\star_{\lambda_1}) \leq \langle \boldsymbol{\pi}^\star_{\lambda_2}, \boldsymbol{\gamma} \rangle + \lambda_1 \phi(\boldsymbol{\pi}^\star_{\lambda_2}) \ , \tag{81}$$

$$\langle \boldsymbol{\pi}^\star_{\lambda_2}, \boldsymbol{\gamma} \rangle + \lambda_2 \phi(\boldsymbol{\pi}^\star_{\lambda_2}) \leq \langle \boldsymbol{\pi}^\star_{\lambda_1}, \boldsymbol{\gamma} \rangle + \lambda_2 \phi(\boldsymbol{\pi}^\star_{\lambda_1}) \ . \tag{82}$$

Subtracting these two inequalities, we obtain that $\phi(\boldsymbol{\pi}^\star_{\lambda_1}) \geq \phi(\boldsymbol{\pi}^\star_{\lambda_2})$. Reinserting this result in the first inequality, we finally get $\langle \boldsymbol{\pi}^\star_{\lambda_1}, \boldsymbol{\gamma} \rangle \leq \langle \boldsymbol{\pi}^\star_{\lambda_2}, \boldsymbol{\gamma} \rangle$. The continuity of the dual RMD results from that of the primal RMD. Let $\lambda \geq 0$, and choose an arbitrary dual rot mover's plan $\boldsymbol{\pi}^\star_\lambda$ and earth mover's plan $\boldsymbol{\pi}^\star$. On the one hand, we have $\langle \boldsymbol{\pi}^\star, \boldsymbol{\gamma} \rangle \leq \langle \boldsymbol{\pi}^\star_\lambda, \boldsymbol{\gamma} \rangle$. On the other hand, we have $\langle \boldsymbol{\pi}^\star_\lambda, \boldsymbol{\gamma} \rangle + \lambda \phi(\boldsymbol{\pi}^\star_\lambda) \leq \langle \boldsymbol{\pi}', \boldsymbol{\gamma} \rangle + \lambda \phi(\boldsymbol{\pi}')$, and thus $\langle \boldsymbol{\pi}^\star_\lambda, \boldsymbol{\gamma} \rangle \leq \langle \boldsymbol{\pi}', \boldsymbol{\gamma} \rangle + \lambda(\phi(\boldsymbol{\pi}') - \phi(\boldsymbol{\pi}^\star_\lambda)) \leq \langle \boldsymbol{\pi}', \boldsymbol{\gamma} \rangle$. Suppose we have a discontinuity of the dual RMD at λ. Then by monotonicity, there is a value $\langle \boldsymbol{\pi}^\star, \boldsymbol{\gamma} \rangle < d < \langle \boldsymbol{\pi}', \boldsymbol{\gamma} \rangle$ that is not in the image of the dual RMD. But d is in the image of the primal RMD for a given $\alpha > 0$ by continuity. It means that $\nexists \lambda > 0$ such that $\langle \pi^\star_\lambda, C \rangle = d$, whereas, by continuity of the primal problem, we know that there exist $\alpha > 0$ such that $\langle \pi^\star_\alpha, C \rangle \geq d$. This is in contradiction with the duality result in Theorem 13, which implies that the image of the primal RMD for $\alpha > 0$ must be included in that of the dual RMD for $\lambda \geq 0$. □

Property 7. *When λ tends to $+\infty$, the dual rot mover's distance converges to:*

$$\lim_{\lambda \to +\infty} d_{\boldsymbol{\gamma},\lambda,\phi}(\mathbf{p}, \mathbf{q}) = \langle \boldsymbol{\pi}', \boldsymbol{\gamma} \rangle \ , \tag{83}$$

and the dual rot mover's plan converges in norm to the transport plan with minimal Bregman information:

$$\lim_{\lambda \to +\infty} \boldsymbol{\pi}^\star_\lambda = \boldsymbol{\pi}' \ . \tag{84}$$

Proof. Let $(\lambda_k)_{k \in \mathbb{N}}$ be a sequence of positive numbers that tends to $+\infty$, and $(\boldsymbol{\pi}^\star_{\lambda_k})_{k \in \mathbb{N}}$ the associated rot mover's plans. By compactness of $\Pi(\mathbf{p}, \mathbf{q})$, we can extract a subsequence of rot mover's plans that converges in norm to a point $\boldsymbol{\pi}^\star \in \Pi(\mathbf{p}, \mathbf{q})$. For the sake of simplicity, we do not relabel this subsequence. By construction, we have $\langle \boldsymbol{\pi}^\star_{\lambda_k}, \boldsymbol{\gamma} \rangle + \lambda_k \phi(\boldsymbol{\pi}') \leq \langle \boldsymbol{\pi}^\star_{\lambda_k}, \boldsymbol{\gamma} \rangle + \lambda_k \phi(\boldsymbol{\pi}^\star_{\lambda_k}) \leq \langle \boldsymbol{\pi}', \boldsymbol{\gamma} \rangle + \lambda_k \phi(\boldsymbol{\pi}')$. The scalar products are bounded, so dividing the inequalities by λ_k and taking the limit, we obtain that $\phi(\boldsymbol{\pi}^\star_{\lambda_k})$ converges to $\phi(\boldsymbol{\pi}')$. By lower semi-continuity of ϕ, we thus have $\phi(\boldsymbol{\pi}^\star) \leq \phi(\boldsymbol{\pi}')$. Since the global minimum of ϕ on $\Pi(\mathbf{p}, \mathbf{q})$ is attained uniquely at $\boldsymbol{\pi}'$, we must

have $\boldsymbol{\pi}^\star = \boldsymbol{\pi}'$, and the original sequence also converges in norm to $\boldsymbol{\pi}'$. Hence, the dual rot mover's plan $\boldsymbol{\pi}_\lambda$ converges in norm to $\boldsymbol{\pi}'$ when λ tends to $+\infty$. By continuity of the total cost $\langle \cdot, \boldsymbol{\gamma} \rangle$ on $\mathbb{R}^{d \times d}$, $\langle \boldsymbol{\pi}^\star_{\lambda_k}, \boldsymbol{\gamma} \rangle$ converges to $\langle \boldsymbol{\pi}', \boldsymbol{\gamma} \rangle$. Hence, the limit of the RMD when λ tends to $+\infty$ is $\langle \boldsymbol{\pi}', \boldsymbol{\gamma} \rangle$. □

Property 8. *When λ tends to 0, the dual rot mover's distance converges to the earth mover's distance:*

$$\lim_{\lambda \to 0} d_{\boldsymbol{\gamma},\lambda,\phi}(\mathbf{p}, \mathbf{q}) = d_{\boldsymbol{\gamma}}(\mathbf{p}, \mathbf{q}) \ . \tag{85}$$

Proof. This is a direct consequence of the dual RMD being continuous at $\lambda = 0$. □

Property 9. *If $[0,1)^{d \times d} \subseteq \operatorname{dom} \phi$ and ϕ is strictly convex on $[0,1)^{d \times d}$, then the dual rot mover's plan converges in norm when λ tends to 0 to the earth mover's plan $\boldsymbol{\pi}^\star_0$ with minimal Bregman information:*

$$\lim_{\lambda \to 0} \boldsymbol{\pi}^\star_\lambda = \boldsymbol{\pi}^\star_0 \ . \tag{86}$$

Proof. Let $(\lambda_k)_{k \in \mathbb{N}}$ be a sequence of positive numbers that converges to 0, and $(\boldsymbol{\pi}^\star_{\lambda_k})_{k \in \mathbb{N}}$ the associated rot mover's plans. By compactness of $\Pi(\mathbf{p}, \mathbf{q})$, we can extract a subsequence of rot mover's plans that converges in norm to a point $\boldsymbol{\pi}^\star \in \Pi(\mathbf{p}, \mathbf{q})$. For the sake of simplicity, we do not relabel this subsequence. By construction, we have $\langle \boldsymbol{\pi}^\star_0, \boldsymbol{\gamma} \rangle + \lambda_k \phi(\boldsymbol{\pi}^\star_{\lambda_k}) \leq \langle \boldsymbol{\pi}^\star_{\lambda_k}, \boldsymbol{\gamma} \rangle + \lambda_k \phi(\boldsymbol{\pi}^\star_{\lambda_k}) \leq \langle \boldsymbol{\pi}^\star_0, \boldsymbol{\gamma} \rangle + \lambda_k \phi(\boldsymbol{\pi}^\star_0)$. The regularizer ϕ is continuous on the polytope $\Pi(\mathbf{p}, \mathbf{q}) \subseteq \operatorname{dom} \phi$, so taking the limit, we obtain that $\langle \boldsymbol{\pi}^\star_{\lambda_k}, \boldsymbol{\gamma} \rangle$ converges to $\langle \boldsymbol{\pi}^\star_0, \boldsymbol{\gamma} \rangle$. Therefore, $\boldsymbol{\pi}^\star$ must be an earth mover's plan. Now dividing by λ_k and taking the limit, we obtain that $\phi(\boldsymbol{\pi}^\star) \leq \phi(\boldsymbol{\pi}^\star_0)$. Since $\boldsymbol{\pi}^\star_0$ is the unique earth mover's plan with minimal Bregman information, we must have $\boldsymbol{\pi}^\star = \boldsymbol{\pi}^\star_0$. □

3.4 Geometrical Insights

Our primal and dual formulations enlighten some intricate relations between optimal transportation theory [54] and information geometry [3], where Bregman divergences are known to possess a dually flat structure with a generalized Pythagorean theorem for information projections. A schematic view of the underlying geometry for ROT problems is represented in Figure 1, and can be discussed as follows.

Our constructions start from the global minimizer $\boldsymbol{\xi}'$ of the regularizer ϕ (Lemma 1). The Bregman projection $\boldsymbol{\pi}'$ of $\boldsymbol{\xi}'$ onto the transport polytope $\Pi(\mathbf{p}, \mathbf{q})$ has minimal Bregman information on $\Pi(\mathbf{p}, \mathbf{q})$ (Lemma 2). The linear cost restricted to the regularized transport polytope $\Pi_{\alpha,\phi}(\mathbf{p}, \mathbf{q})$ also attains its global minimum (Lemma 3). Such a minimizer $\boldsymbol{\pi}'^\star_\alpha$ is a primal rot mover's plan (Definition 1). We can interpret $\Pi_{\alpha,\phi}(\mathbf{p}, \mathbf{q})$ as the intersection of $\Pi(\mathbf{p}, \mathbf{q})$ with the Bregman ball of radius $B_\phi(\boldsymbol{\pi}' \| \boldsymbol{\xi}') + \alpha$ and center $\boldsymbol{\xi}'$ (Proposition 4). In certain cases, $\Pi_{\alpha,\phi}(\mathbf{p}, \mathbf{q})$ is also the intersection of $\Pi(\mathbf{p}, \mathbf{q})$ with the Bregman ball of radius α and center $\boldsymbol{\pi}'$, as a result of the generalized Pythagorean theorem $B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}') = B_\phi(\boldsymbol{\pi} \| \boldsymbol{\pi}') + B_\phi(\boldsymbol{\pi}' \| \boldsymbol{\xi}')$ (Proposition 5, Corollary 6). All in all, this enforces the solutions to have small enough Bregman information, by constraining them to lie close to the matrix $\boldsymbol{\xi}'$ or transport plan $\boldsymbol{\pi}'$ with minimal Bregman information.

In our developments, we next introduce the global minimizer $\boldsymbol{\xi}$ of the regularized cost (Lemma 7). The regularized cost restricted to $\Pi(\mathbf{p}, \mathbf{q})$ also attains its global minimum uniquely (Lemma 8). This minimizer defines the dual rot mover's plan $\boldsymbol{\pi}^\star_\lambda$ (Definition 2). Actually, $\boldsymbol{\pi}^\star_\lambda$ can be seen as the Bregman projection of $\boldsymbol{\xi}$ onto $\Pi(\mathbf{p}, \mathbf{q})$ (Proposition 9). The regularization by the Bregman information is also equivalent to regularizing the solution toward $\boldsymbol{\xi}'$ (Proposition 10). In some cases, this can also be seen as regularizing toward $\boldsymbol{\pi}'$, as a result of the generalized Pythagorean theorem $B_\phi(\boldsymbol{\pi} \| \boldsymbol{\xi}') = B_\phi(\boldsymbol{\pi} \| \boldsymbol{\pi}') + B_\phi(\boldsymbol{\pi}' \| \boldsymbol{\xi}')$ (Proposition 11, Corollary 12). Again,

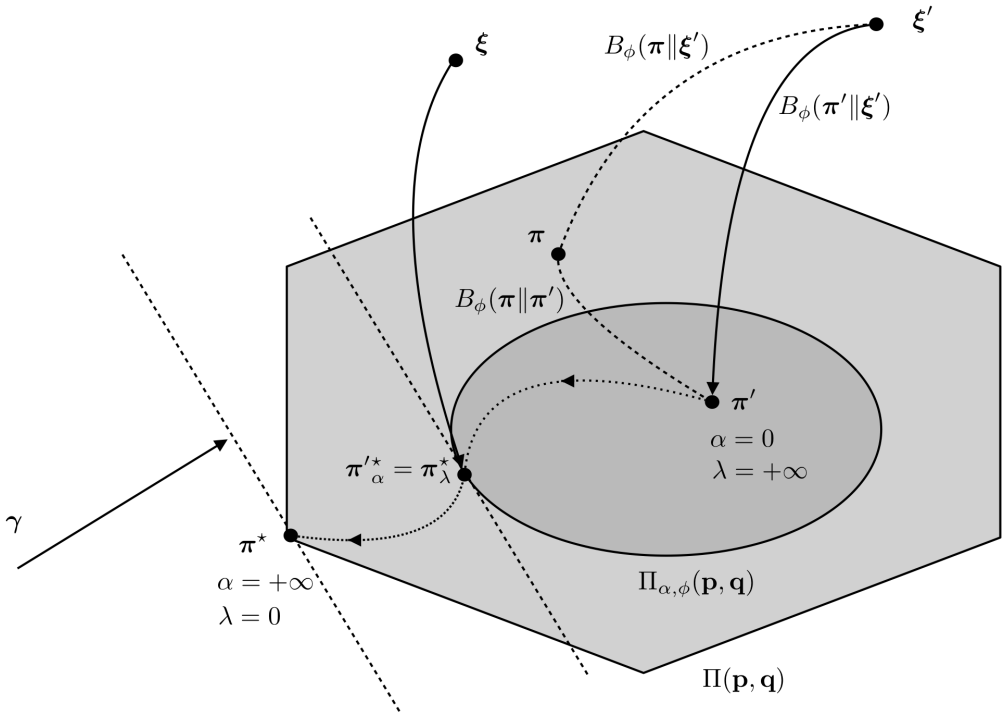


Figure 1: Geometry of regularized optimal transport.

this enforces the solutions to have small enough Bregman information, by shrinking them toward the matrix $\boldsymbol{\xi}'$ or transport plan $\boldsymbol{\pi}'$ with minimal Bregman information.

We have duality between the primal and dual formulations, so that primal and dual rot mover's plans follow the same path on $\Pi(\mathbf{p},\mathbf{q})$ from no regularization ($\alpha = +\infty$, $\lambda = 0$) to full regularization ($\alpha = 0$, $\lambda = +\infty$) (Theorem 13). In the limit of no regularization, we obviously retrieve an earth mover's plan $\boldsymbol{\pi}^\star$ for the cost matrix $\boldsymbol{\gamma}$. By duality, it is also intuitive that the additional constraint for the primal formulation, seen in the equivalent forms of $\phi(\boldsymbol{\pi})$, $B_\phi(\boldsymbol{\pi}\|\boldsymbol{\xi}')$ or $B_\phi(\boldsymbol{\pi}\|\boldsymbol{\pi}')$, leads to an analog penalty for the dual formulation in the same respective form.

Since $\boldsymbol{\xi}' = \mathbf{1}$, $\boldsymbol{\pi}' = \mathbf{p}\mathbf{q}^\top$, $\boldsymbol{\xi} = \exp(-\boldsymbol{\gamma}/\lambda)$, for minus the Boltzmann-Shannon entropy and Kullback-Leibler divergence, we retrieve the existing results discussed in Section 1.2 as a specific case [6,15]. In addition, we can readily generalize the estimation of contingency tables with fixed marginals to a matrix nearness problem based on other divergences than the Kullback-Leibler divergence [19]. Given a rough estimate $\boldsymbol{\xi} \in \operatorname{int}(\operatorname{dom}\phi)$, a contingency table with fixed marginals $\mathbf{p},\mathbf{q}$ can be estimated by Bregman projection of $\boldsymbol{\xi}$ onto $\Pi(\mathbf{p},\mathbf{q})$:

$$\boldsymbol{\pi}^\star = \operatorname*{argmin}_{\boldsymbol{\pi}\in\Pi(\mathbf{p},\mathbf{q})} B_\phi(\boldsymbol{\pi}\|\boldsymbol{\xi}) \ . \tag{87}$$

This simply amounts to solving a dual ROT problem with an arbitrary penalty $\lambda > 0$ and a cost matrix $\boldsymbol{\gamma} = -\lambda\nabla\phi(\boldsymbol{\xi})$.

Finally, since Bregman divergences are invariant under adding an affine term to their generator, it is straightforward to generalize ROT problems by shrinking toward an arbitrary prior matrix $\boldsymbol{\xi},\boldsymbol{\xi}' \in \operatorname{int}(\operatorname{dom}\phi)$, or transport plan $\boldsymbol{\pi}' \in \Pi(\mathbf{p},\mathbf{q})$. This is indeed equivalent to translating the regularizer by the appropriate amount $\phi(\boldsymbol{\pi}) + \langle\boldsymbol{\pi},\boldsymbol{\delta}\rangle$, so that the global minimizer is

now attained at the desired point. Equivalently, this amounts to translating the cost matrix as $\boldsymbol{\gamma} + \lambda\boldsymbol{\delta}$ instead.

4 Algorithmic Derivations

In this section, we introduce algorithmic methods to solve ROT problems. We focus without lack of generality on the dual problem, which can be solved efficiently via alternate Bregman projections. The primal problem can then easily be solved for $0 < \alpha < \alpha'$ by a bisection search on $\lambda > 0$. For $\alpha = 0$, we could simply use alternate Bregman projections to project $\nabla\psi(\mathbf{0})$ instead of $\nabla\psi(-\boldsymbol{\gamma}/\lambda)$ in virtue of Lemmas 1 and 2, which actually corresponds to the special case $\boldsymbol{\gamma} = \mathbf{0}$ in our algorithms, though this is not really relevant in practice since this completely removes the linear influence of the total cost from the ROT problem. In the limit $\alpha \geq \alpha'$, a classical OT solver such as the network simplex can directly be used. We first study the underlying Bregman projections in their generic form (Section 4.1) and specifically develop the case of separable divergences (Section 4.2). We then derive the two generic schemes of ASA (Section 4.3) and NASA (Section 4.4) to solve dual ROT problems, depending on whether the domain of the smooth convex regularizer lies within the non-negative orthant or not. We also enhance both algorithms in the separable case with a sparse extension (Section 4.5), and finally discuss some practical considerations of our methods (Section 4.6). To simplify notations, we omit the penalty value λ in the index and simply write $\boldsymbol{\pi}^\star$ for the rot mover's plan.

4.1 Generic Projections

The closed convex transport polytope $\Pi(\mathbf{p}, \mathbf{q})$ is the intersection of the non-negative orthant:

$$\mathcal{C}_0 = \mathbb{R}_+^{d \times d} \ , \tag{88}$$

which is a polyhedral subset, with two affine subspaces:

$$\mathcal{C}_1 = \{\boldsymbol{\pi} \in \mathbb{R}^{d \times d} \colon \boldsymbol{\pi}\mathbf{1} = \mathbf{p}\} \ , \tag{89}$$

$$\mathcal{C}_2 = \{\boldsymbol{\pi} \in \mathbb{R}^{d \times d} \colon \boldsymbol{\pi}^\top\mathbf{1} = \mathbf{q}\} \ . \tag{90}$$

The Bregman projection $\boldsymbol{\pi}^\star$ onto $\Pi(\mathbf{p}, \mathbf{q})$ can then be obtained by alternate Bregman projections onto $\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2$, where we expect that these latter projections are easier to compute.

On the one hand, the Karush-Kuhn-Tucker conditions for Bregman projection $\boldsymbol{\pi}_0^\star$ of a given matrix $\overline{\boldsymbol{\pi}} \in \operatorname{int}(\operatorname{dom}\phi)$ onto $\mathcal{C}_0$ are necessary and sufficient, and write as follows:

$$\boldsymbol{\pi}_0^\star \geq \mathbf{0} \ , \tag{91}$$

$$\nabla\phi(\boldsymbol{\pi}_0^\star) - \nabla\phi(\overline{\boldsymbol{\pi}}) \geq \mathbf{0} \ , \tag{92}$$

$$(\nabla\phi(\boldsymbol{\pi}_0^\star) - \nabla\phi(\overline{\boldsymbol{\pi}})) \odot \boldsymbol{\pi}_0^\star = \mathbf{0} \ . \tag{93}$$

While these conditions are nontrivial to solve in general, we shall see that they admit an elegant solver specific to the non-separable squared Mahalanobis distances defined in (33) and generated by the quadratic form in (34). In addition, they also greatly simplify for separable divergences, which encompass all other divergences used in this paper.

On the other hand, the Lagrangians with Lagrange multipliers $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^d$ for the Bregman projections $\boldsymbol{\pi}_1^\star$ and $\boldsymbol{\pi}_2^\star$ of a given matrix $\overline{\boldsymbol{\pi}} \in \operatorname{int}(\operatorname{dom}\phi)$ onto $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively write as follows:

$$\mathcal{L}_1(\boldsymbol{\pi}, \boldsymbol{\mu}) = \phi(\boldsymbol{\pi}) - \langle \boldsymbol{\pi}, \nabla\phi(\overline{\boldsymbol{\pi}}) \rangle + \boldsymbol{\mu}^\top(\boldsymbol{\pi}\mathbf{1} - \mathbf{p}) \ , \tag{94}$$

$$\mathcal{L}_2(\boldsymbol{\pi}, \boldsymbol{\nu}) = \phi(\boldsymbol{\pi}) - \langle \boldsymbol{\pi}, \nabla\phi(\overline{\boldsymbol{\pi}}) \rangle + \boldsymbol{\nu}^\top (\boldsymbol{\pi}^\top \mathbf{1} - \mathbf{q}) \ . \tag{95}$$

Their gradients are given on $\operatorname{int}(\operatorname{dom}\phi)$ by:

$$\nabla\mathcal{L}_1(\boldsymbol{\pi}, \boldsymbol{\mu}) = \nabla\phi(\boldsymbol{\pi}) - \nabla\phi(\overline{\boldsymbol{\pi}}) + \boldsymbol{\mu}\mathbf{1}^\top \ , \tag{96}$$

$$\nabla\mathcal{L}_2(\boldsymbol{\pi}, \boldsymbol{\nu}) = \nabla\phi(\boldsymbol{\pi}) - \nabla\phi(\overline{\boldsymbol{\pi}}) + \mathbf{1}\boldsymbol{\nu}^\top \ , \tag{97}$$

and vanish at $\boldsymbol{\pi}_1^\star, \boldsymbol{\pi}_2^\star \in \operatorname{int}(\operatorname{dom}\phi)$ if and only if:

$$\boldsymbol{\pi}_1^\star = \nabla\psi(\nabla\phi(\overline{\boldsymbol{\pi}}) - \boldsymbol{\mu}\mathbf{1}^\top) \ , \tag{98}$$

$$\boldsymbol{\pi}_2^\star = \nabla\psi(\nabla\phi(\overline{\boldsymbol{\pi}}) - \mathbf{1}\boldsymbol{\nu}^\top) \ . \tag{99}$$

By duality, the Bregman projections onto $\mathcal{C}_1, \mathcal{C}_2$ are thus equivalent to finding the unique vectors $\boldsymbol{\mu}, \boldsymbol{\nu}$, such that the rows of $\boldsymbol{\pi}_1^\star$ sum up to $\mathbf{p}$, respectively the columns of $\boldsymbol{\pi}_2^\star$ sum up to $\mathbf{q}$:

$$\nabla\psi(\nabla\phi(\overline{\boldsymbol{\pi}}) - \boldsymbol{\mu}\mathbf{1}^\top)\mathbf{1} = \mathbf{p} \ , \tag{100}$$

$$\nabla\psi(\nabla\phi(\overline{\boldsymbol{\pi}}) - \mathbf{1}\boldsymbol{\nu}^\top)^\top\mathbf{1} = \mathbf{q} \ . \tag{101}$$

Similarly, solving for the Lagrange multipliers is an expensive problem in general, since the search space is of dimension d and we evaluate matrix functions of size $d \times d$. This is because a given entry μ_i, ν_j can actually modify any entry of the $d \times d$ matrix functions being evaluated. Again, we shall see that they can nevertheless be computed efficiently for separable divergences as well as the non-separable Mahalanobis distances.

4.2 Separable Case

Assuming that the regularizer ϕ is separable, the underlying Bregman projections can be computed more efficiently. To keep notations simple, we focus on separable divergences with same element-wise regularizer, and thus chiefly omit the indices $\phi_{ij} = \phi$. We emphasize, however, that it is straightforward to apply all our methods for separable divergences with different element-wise regularizers, which notably enables weighting a given element-wise regularizer.

In case of separability, the Karush-Kuhn-Tucker conditions for projection onto $\mathcal{C}_0$ simplify to provide a closed-form solution on primal parameters:

$$\pi_{0,ij}^\star = \max\{0, \overline{\pi}_{ij}\} \ . \tag{102}$$

Since ϕ' is increasing, this is equivalent on dual parameters to:

$$\theta_{0,ij}^\star = \max\{\phi'(0), \overline{\theta}_{ij}\} \ . \tag{103}$$

Now turning to projections onto $\mathcal{C}_1, \mathcal{C}_2$ for primal parameters $\pi_{1,ij}^\star, \pi_{2,ij}^\star$, we can divide the initial problems into d parallel subproblems in search space of dimension 1 each. This is much more efficient to solve than in the non-separable case. This can be summarized as looking for d separate Lagrange multipliers μ_i, respectively ν_j, such that:

$$\sum_{j=1}^{d} \psi'(\overline{\theta}_{ij} - \mu_i) = p_i \ , \tag{104}$$

$$\sum_{i=1}^{d} \psi'(\overline{\theta}_{ij} - \nu_j) = q_j \ . \tag{105}$$

Finding the optimal values $\mu_i, \nu_j \in \mathbb{R}$ through ψ' and the sums over rows or columns, however, is still nontrivial in general.

An analytical solution can be obtained in specific cases. Intuitively, we need to factor μ_i, ν_j out of ψ' as additive or multiplicative terms. This is related to Pexider's functional equations, which hold only for functions with a linear form $\psi'(\theta) = a\theta + b$, or exponential form $\psi'(\theta) = a\exp(b\theta)$, with $a, b \in \mathbb{R}$. This leads to regularizers with a quadratic form $\phi(\pi) = a\pi^2 + b\pi + c$, or entropic form $\phi(\pi) = a\pi\log\pi + b\pi + c$, with $a, b, c \in \mathbb{R}$. The constants a, b actually only scale and translate the cost matrix, whereas the constant c has no effect. Referring to Table 1, the quadratic case holds under assumptions (B), and thus requires Dykstra's algorithm for alternate Bregman projections with correction terms to ensure non-negativity by projection onto the polyhedral non-negative orthant. The entropic case holds under assumptions (A), using the POCS technique for alternate Bregman projection with no correction terms since the non-negativity is already ensured by the domain of the regularizer. The latter case reduces to the regularization of [15] and [6], so that we actually end up with the Sinkhorn-Knopp algorithm. Hence, the Euclidean norm associated to the squared Euclidean distance, and the entropic case associated to the Kullback-Leibler divergence, are reasonably the only two existing analytical schemes to find the sum constraint projections. For other ROT problems, available solvers for line search can be employed instead.

For simplicity, we assume hereafter that ψ is twice continuously differentiable with ψ'' positive and ψ' verifying the necessary and sufficient condition (40) on its whole domain. Therefore, we can use the Newton-Raphson method with guarantees of global convergence. This encompasses most of the common regularizers, and notably all regularizers used in this paper except from the Fermi-Dirac entropy, ℓ_p norms and Hellinger distance. When the condition (40) for global convergence is not met on the whole domain, it is still possible to apply the Newton-Raphson method after careful initialization, so as to restrict to a smaller interval where the condition holds. This is discussed in more detail with practical examples for the Fermi-Dirac entropy, ℓ_p norms and Hellinger distance in Section 5, where the first-order derivatives are increasing convex on half of the domain and increasing concave on the other half. When the second-order derivatives do not exist, are not continuous or vanish at some points, a similar strategy can be applied. This is again discussed for the ℓ_p norms in Section 5, where the second-order derivative is undefined or vanishes at 0 depending on the value of the parameter. If such an initialization is not possible, then a bisection search can always be applied instead of the Newton-Raphson method.

To apply the Newton-Raphson method, we exploit the following functions:

$$f(\mu_i) = -\sum_{j=1}^{d} \psi'(\overline{\theta}_{ij} - \mu_i) \ , \tag{106}$$

$$g(\nu_j) = -\sum_{i=1}^{d} \psi'(\overline{\theta}_{ij} - \nu_j) \ , \tag{107}$$

defined respectively on the open intervals $(\hat{\theta}_i - \overline{\theta}, +\infty)$ and $(\check{\theta}_j - \overline{\theta}, +\infty)$, where $0 < \overline{\theta} \leq +\infty$ is such that $\operatorname{dom}\psi = (-\infty, \overline{\theta})$, and $\hat{\theta}_i = \max\{\overline{\theta}_{ij}\}_{1\leq j\leq d}$, $\check{\theta}_j = \max\{\overline{\theta}_{ij}\}_{1\leq i\leq d}$. Their continuous derivatives are given by:

$$f'(\mu_i) = \sum_{j=1}^{d} \psi''(\overline{\theta}_{ij} - \mu_i) \ , \tag{108}$$

$$g'(\nu_j) = \sum_{i=1}^{d} \psi''(\overline{\theta}_{ij} - \nu_j) \ , \tag{109}$$

and are positive, so that f, g are strictly increasing on their whole domain, and thus on any closed interval with endpoints consisting of a feasible point and a solution. By construction, f, g also verify the necessary and sufficient condition (40) for global convergence, and we know that there are unique solutions to $f(\mu_i) = -p_i$ and $g(\nu_j) = -q_j$. Hence, the Newton-Raphson updates:

$$\mu_i \leftarrow \mu_i + \frac{\sum_{j=1}^{d} \psi'(\overline{\theta}_{ij} - \mu_i) - p_i}{\sum_{j=1}^{d} \psi''(\overline{\theta}_{ij} - \mu_i)} \ , \tag{110}$$

$$\nu_j \leftarrow \nu_j + \frac{\sum_{i=1}^{d} \psi'(\overline{\theta}_{ij} - \nu_j) - q_j}{\sum_{i=1}^{d} \psi''(\overline{\theta}_{ij} - \nu_j)} \ , \tag{111}$$

converge to the optimal solutions with a quadratic rate for any feasible starting points. By construction, we also know that initialization can be done with $\mu_i \leftarrow 0$, $\nu_j \leftarrow 0$. To avoid storing the intermediate Lagrange multipliers, the updates can then directly be written on dual parameters:

$$\theta^\star_{1,ij} \leftarrow \theta^\star_{1,ij} - \frac{\sum_{j=1}^{d} \psi'(\theta^\star_{1,ij}) - p_i}{\sum_{j=1}^{d} \psi''(\theta^\star_{1,ij})} \ , \tag{112}$$

$$\theta^\star_{2,ij} \leftarrow \theta^\star_{2,ij} - \frac{\sum_{i=1}^{d} \psi'(\theta^\star_{2,ij}) - q_j}{\sum_{i=1}^{d} \psi''(\theta^\star_{2,ij})} \ , \tag{113}$$

after initialization by $\theta^\star_{1,ij} \leftarrow \overline{\theta}_{ij}$, $\theta^\star_{2,ij} \leftarrow \overline{\theta}_{ij}$.

4.3 Alternate Scaling Algorithm

Under assumptions (A), we can drop the non-negative constraint since it is already ensured by $\operatorname{dom} \phi \subseteq \mathbb{R}_+^{d \times d}$ (Table 1). The POCS technique in its basic form (38) then states that the projection of $\boldsymbol{\xi}$ onto $\Pi(\mathbf{p}, \mathbf{q})$ can be obtained by alternate Bregman projections onto the affine subspaces $\mathcal{C}_1$ and $\mathcal{C}_2$ with linear convergence. Clearly, the underlying control mapping takes each output value an infinite number of times. Since we have just two sets, the only possible alternative in the control mapping is to swap the order of projections starting from $\mathcal{C}_2$ instead of $\mathcal{C}_1$, which actually amounts to swapping the input distributions $\mathbf{p}, \mathbf{q}$ and transposing the cost matrix $\boldsymbol{\gamma}$, to obtain the transposed of the rot mover's plan. We thus focus on the first choice without lack of generality.

Starting from $\boldsymbol{\xi}$ and writing the successive vectors $\boldsymbol{\mu}^{(k)}, \boldsymbol{\nu}^{(k)}$ along iterations, we have the following sequence:

$$\nabla\psi(-\boldsymbol{\gamma}/\lambda) \to \nabla\psi\left(-\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top\right) \tag{114}$$

$$\to \nabla\psi\left(-\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\right) \tag{115}$$

$$\to \dots \tag{116}$$

$$\to \nabla\psi\left(-\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top} - \dots - \boldsymbol{\mu}^{(k)}\mathbf{1}^\top\right) \tag{117}$$

$$\to \nabla\psi\left(-\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top} - \dots - \boldsymbol{\mu}^{(k)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\right) \tag{118}$$

Algorithm 1 Alternate scaling algorithm.

$\boldsymbol{\theta}^\star \leftarrow -\boldsymbol{\gamma}/\lambda$
repeat
  $\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \boldsymbol{\mu}\mathbf{1}^\top$, where $\boldsymbol{\mu}$ uniquely solves $\nabla\psi(\boldsymbol{\theta}^\star - \boldsymbol{\mu}\mathbf{1}^\top)\mathbf{1} = \mathbf{p}$
  $\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\nu}^\top$, where $\boldsymbol{\nu}$ uniquely solves $\nabla\psi(\boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\nu}^\top)^\top\mathbf{1} = \mathbf{q}$
until convergence
$\boldsymbol{\pi}^\star \leftarrow \nabla\psi(\boldsymbol{\theta}^\star)$

Algorithm 2 Alternate scaling algorithm in the separable case.

$\boldsymbol{\theta}^\star \leftarrow -\boldsymbol{\gamma}/\lambda$
repeat
  repeat
    $\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \frac{\psi'(\boldsymbol{\theta}^\star)\mathbf{1} - \mathbf{p}}{\psi''(\boldsymbol{\theta}^\star)\mathbf{1}}\,\mathbf{1}^\top$
  until convergence
  repeat
    $\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \mathbf{1}\,\frac{\mathbf{1}^\top\psi'(\boldsymbol{\theta}^\star) - \mathbf{q}^\top}{\mathbf{1}^\top\psi''(\boldsymbol{\theta}^\star)}$
  until convergence
until convergence
$\boldsymbol{\pi}^\star \leftarrow \psi'(\boldsymbol{\theta}^\star)$

$$\rightarrow \dots \tag{119}$$
$$\rightarrow \boldsymbol{\pi}^\star \ . \tag{120}$$

In other terms, we obtain the rot mover's plan $\boldsymbol{\pi}^\star$ by scaling iteratively the rows and columns of the successive estimates through $\nabla\psi$. An efficient algorithm, called ASA, is to store a unique $d \times d$ matrix in dual parameter space and update it by alternating the projections in primal parameter space (Algorithm 1). The updates have a complexity in $O(d^2)$ once the vectors $\boldsymbol{\mu}, \boldsymbol{\nu}$ are obtained.

In the separable case, the projections can be obtained by iterating the respective Newton-Raphson update steps, which can be written compactly with matrix and vector operations (Algorithm 2). The complexity for the updates are now clearly in $O(d^2)$. In more detail, each update step features one vector row or column replication, one vector element-wise division, one vector subtraction, one matrix subtraction, two matrix row or column sums, and two element-wise matrix function evaluations. Because of separability, we can expect the required number of iterations for convergence in the different loops to be independent of the data dimension, and thus expect a quadratic empirical complexity as well.

4.4 Non-negative Alternate Scaling Algorithm

Under assumptions (B), we must now include the non-negative constraint since $\operatorname{dom}\phi \not\subseteq \mathbb{R}_+^{d\times d}$ (Table 1). We suggest to ensure non-negativity of each update, and thus follow a cycle of projections onto $\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_0, \mathcal{C}_2$. The underlying control mapping is a fortiori essentially cyclic. For practical reasons, we also ensure non-negativity of the output solution with a final projection onto $\mathcal{C}_0$. Again, swapping the order of projections onto $\mathcal{C}_1, \mathcal{C}_2$ is equivalent to swapping the input distributions $\mathbf{p}, \mathbf{q}$ and transposing the cost matrix $\boldsymbol{\gamma}$ to obtain the transposed of the rot mover's plan. Other control mappings could also be exploited, for example by ensuring non-negativity every two or more sum constraint projections. We do not discuss such variants here and focus

Algorithm 3 Non-negative alternate scaling algorithm.

$\boldsymbol{\theta}^\star \leftarrow -\boldsymbol{\gamma}/\lambda$
$\boldsymbol{\vartheta} \leftarrow \mathbf{0}$
$\boldsymbol{\varrho} \leftarrow \mathbf{0}$
$\boldsymbol{\varsigma} \leftarrow \mathbf{0}$
$\overline{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^\star + \boldsymbol{\vartheta}$
$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ uniquely solves $\nabla\psi(\boldsymbol{\theta}) \geq \mathbf{0}$, $\boldsymbol{\theta} \geq \overline{\boldsymbol{\theta}}$, $(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) \odot \nabla\psi(\boldsymbol{\theta}) = \mathbf{0}$
$\boldsymbol{\vartheta} \leftarrow \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star$
repeat
  $\overline{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^\star + \boldsymbol{\varrho}$
  $\boldsymbol{\theta}^\star \leftarrow \overline{\boldsymbol{\theta}} - \boldsymbol{\mu}\mathbf{1}^\top$, where $\boldsymbol{\mu}$ uniquely solves $\nabla\psi(\overline{\boldsymbol{\theta}} - \boldsymbol{\mu}\mathbf{1}^\top)\mathbf{1} = \mathbf{p}$
  $\boldsymbol{\varrho} \leftarrow \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star$
  $\overline{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^\star + \boldsymbol{\vartheta}$
  $\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ uniquely solves $\nabla\psi(\boldsymbol{\theta}) \geq \mathbf{0}$, $\boldsymbol{\theta} \geq \overline{\boldsymbol{\theta}}$, $(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) \odot \nabla\psi(\boldsymbol{\theta}) = \mathbf{0}$
  $\boldsymbol{\vartheta} \leftarrow \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star$
  $\overline{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^\star + \boldsymbol{\varsigma}$
  $\boldsymbol{\theta}^\star \leftarrow \overline{\boldsymbol{\theta}} - \mathbf{1}\boldsymbol{\nu}^\top$, where $\boldsymbol{\nu}$ uniquely solves $\nabla\psi(\overline{\boldsymbol{\theta}} - \mathbf{1}\boldsymbol{\nu}^\top)^\top\mathbf{1} = \mathbf{q}$
  $\boldsymbol{\varsigma} \leftarrow \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star$
  $\overline{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^\star + \boldsymbol{\vartheta}$
  $\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ uniquely solves $\nabla\psi(\boldsymbol{\theta}) \geq \mathbf{0}$, $\boldsymbol{\theta} \geq \overline{\boldsymbol{\theta}}$, $(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) \odot \nabla\psi(\boldsymbol{\theta}) = \mathbf{0}$
  $\boldsymbol{\vartheta} \leftarrow \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star$
until convergence
$\boldsymbol{\pi}^\star \leftarrow \nabla\psi(\boldsymbol{\theta}^\star)$

on the above-mentioned sequence. The non-negative orthant being polyhedral but not affine, we also need to incorporate correction terms $\boldsymbol{\vartheta}, \boldsymbol{\varrho}, \boldsymbol{\varsigma}$ for all three projections. In more detail, the projections are computed after correction so that we do not directly project the obtained updates $\boldsymbol{\theta}^\star$ but the corrected updates $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star + \boldsymbol{\vartheta}$, $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star + \boldsymbol{\varrho}$, and $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star + \boldsymbol{\varsigma}$ for the respective subsets. The correction terms are also updated as the difference $\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star$ between the projected point and its projection. Dykstra's algorithm (36) for Bregman divergences with corrections (37) then guarantees that the projection of $\boldsymbol{\xi}$ onto $\Pi(\mathbf{p}, \mathbf{q})$ is obtained with linear convergence.

A general algorithm, called NASA, is to store $d \times d$ matrices for projected points, projections and correction terms in dual parameter space, update them accordingly and finally go back to primal parameter space (Algorithm 3). The updates have a complexity in $O(d^2)$ once the Karush-Kuhn-Tucker conditions are solved or Lagrange multipliers $\boldsymbol{\mu}, \boldsymbol{\nu}$ are obtained.

In the separable case, the non-negativity constraint can be obtained analytically and the sequence of updates greatly simplifies. Starting from $\boldsymbol{\xi}$ and writing the successive vectors $\boldsymbol{\mu}^{(k)}, \boldsymbol{\nu}^{(k)}$ along iterations, we have:

$$\begin{aligned}
\psi'(-\boldsymbol{\gamma}/\lambda) &\to \psi'\big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda\}\big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda\} - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top\}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top\} - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\}\Big)
\end{aligned}$$

Algorithm 4 Non-negative alternate scaling algorithm in the separable case.

$\widetilde{\boldsymbol{\theta}} \leftarrow -\boldsymbol{\gamma}/\lambda$
$\boldsymbol{\theta}^\star \leftarrow \max\{\phi'(\mathbf{0}), \widetilde{\boldsymbol{\theta}}\}$
repeat
  $\boldsymbol{\tau} \leftarrow \mathbf{0}$
  repeat
    $\boldsymbol{\tau} \leftarrow \boldsymbol{\tau} + \frac{\psi'(\boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top)\mathbf{1} - \mathbf{p}}{\psi''(\boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top)\mathbf{1}}$
  until convergence
  $\widetilde{\boldsymbol{\theta}} \leftarrow \widetilde{\boldsymbol{\theta}} - \boldsymbol{\tau}\mathbf{1}^\top$
  $\boldsymbol{\theta}^\star \leftarrow \max\{\phi'(\mathbf{0}), \widetilde{\boldsymbol{\theta}}\}$
  $\boldsymbol{\sigma} \leftarrow \mathbf{0}$
  repeat
    $\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} + \frac{\mathbf{1}^\top \psi'(\boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\sigma}^\top) - \mathbf{q}^\top}{\mathbf{1}^\top \psi''(\boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\sigma}^\top)}$
  until convergence
  $\widetilde{\boldsymbol{\theta}} \leftarrow \widetilde{\boldsymbol{\theta}} - \mathbf{1}\boldsymbol{\sigma}^\top$
  $\boldsymbol{\theta}^\star \leftarrow \max\{\phi'(\mathbf{0}), \widetilde{\boldsymbol{\theta}}\}$
until convergence
$\boldsymbol{\pi}^\star \leftarrow \psi'(\boldsymbol{\theta}^\star)$

$$
\begin{aligned}
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\} + \boldsymbol{\mu}^{(1)}\mathbf{1}^\top - \boldsymbol{\mu}^{(2)}\mathbf{1}^\top\right)\\
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(2)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\}\right)\\
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(2)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\} + \mathbf{1}\boldsymbol{\nu}^{(1)\top} - \mathbf{1}\boldsymbol{\nu}^{(2)\top}\right)\\
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(2)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(2)\top}\}\right)\\
&\to \dots\\
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\}\right)\\
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\} + \boldsymbol{\mu}^{(k)}\mathbf{1}^\top - \boldsymbol{\mu}^{(k+1)}\mathbf{1}^\top\right)\\
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k+1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\}\right)\\
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k+1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\} + \mathbf{1}\boldsymbol{\nu}^{(k)\top} - \mathbf{1}\boldsymbol{\nu}^{(k+1)\top}\right)\\
&\to \psi'\left(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k+1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k+1)\top}\}\right)\\
&\to \dots\\
&\to \boldsymbol{\pi}^\star \ .
\end{aligned}
$$

An efficient algorithm then exploits the differences $\boldsymbol{\tau}^{(k)} = \boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(k-1)}$ and $\boldsymbol{\sigma}^{(k)} = \boldsymbol{\nu}^{(k)} - \boldsymbol{\nu}^{(k-1)}$ to scale the rows and columns (Algorithm 4). We store $d \times d$ matrices as well as difference vectors instead of correction matrices. The algorithm can then be interpreted as producing interleaved updates between the projections according to the max operator and according to the respective scalings. The updates in NASA now clearly have a complexity in $O(d^2)$ when using the Newton-Raphson method for scaling, with similar matrix and vector operations to ASA in the separable case, and an expected empirical complexity that is quadratic.

4.5 Sparse Extension

In the separable case, it is possible to develop a sparse extension of both our methods ASA and NASA. Storing and updating full $d \times d$ matrices becomes expensive with the data dimension. Instead, we allow for infinite entries in the cost matrix $\boldsymbol{\gamma}$, meaning that the transport of mass between certain bins is proscribed. As a result, the corresponding entries of $\boldsymbol{\pi}^\star$ must be null. Eventually, we can drop all these entries so that we just need to store and update the remaining ones. The RMD via the Frobenius inner product $\langle \boldsymbol{\pi}^\star, \boldsymbol{\gamma} \rangle$ is then computed without accounting for discarded entries, or equivalently by setting indefinite element-wise products $0 \times \infty = 0$ by convention, so it naturally costs nothing to move no mass on a path that is forbidden. This leads to an expected complexity in $O(r)$, where r is the number of finite entries in $\boldsymbol{\gamma}$. Typically, r can be chosen in the order of magnitude of d, so as to obtain a linear instead of quadratic empirical complexity.

In practice, both ASA and NASA are compatible with this strategy. We always have $\lim_{\theta \to -\infty} \psi'(\theta) = 0$ under assumptions (A) for ASA. Under assumptions (B) for NASA, this limit might be finite or infinite but is necessarily negative, so also leads to 0 after enforcing non-negativity by projection onto the non-negative orthant. As a result, the obtained sequence of projections preserves the desired zeros in both algorithms, and an infinite element-wise cost does lead to no mass transport at all between the corresponding bins. In theory, we can understand this extension in light of the dual formulation seen as a Bregman projection in (67). Under assumptions (B), we always have $0 \in \operatorname{int}(\operatorname{dom} \phi)$ and thus $B_\phi(0 \| 0) = 0$. Hence, Dykstra's algorithm is readily applicable in the sparse version. Under assumptions (A), however, we have $0 \notin \operatorname{int}(\operatorname{dom} \phi)$, and even sometimes $0 \notin \operatorname{dom} \phi$ as for the Itakura-Saito divergence. We can nonetheless extend the domain of the element-wise divergence at the origin by continuity on the diagonal, that is, by setting it null as $B_\phi(0 \| 0) = 0$. This is akin to considering absolutely continuous measures, also known as dominated measures, and Radon-Nikodym derivatives to generalize the definition of Bregman divergences. [29] then showed that the POCS method still holds with this convention by introducing a notion of locally affine spaces.

With such a sparse extension, however, we must take care that a sparse solution does exist, meaning that there is a transport plan in the transport polytope that has the desired zeros. For example, if all entries of $\boldsymbol{\gamma}$ are infinite, then there are obviously no possible sparse solutions since we enforce all entries of the plan to be null. A necessary condition for the existence of a sparse solution is that for any entry q_j, all entries p_k from which we are allowed to transport mass must provide enough total mass to fill q_j completely. Similarly, for any entry p_i, all entries q_k to which we are allowed to transport mass must require enough total mass to empty p_i completely. Unfortunately, sufficient conditions are not so intuitive. [27, Theorem 4.1] studied such problems thoroughly and elucidated several necessary and sufficient conditions for sparse solutions to exist, but these conditions are nontrivial to use from in practice. [29] advocates trying first to compute a solution with the desired sparsity, and if no solution can be found, then gradually reduce sparsity until a solution is found. This might still speed up computation drastically because of the linear instead of quadratic complexity. Lastly, we remark that it is not evident to propose a sparse extension for the non-separable case in general, since a given entry of $\boldsymbol{\gamma}$ might influence all entries of $\boldsymbol{\pi}^\star$.

4.6 Practical Considerations

As noticed by [15] and [6], the Sinkhorn-Knopp algorithm might fail to converge because of numerical instability when the penalty λ gets small. In particular, unless taking special care of numerical stabilization [44], a direct limitation is the machine precision under which some entries of $\exp(-\boldsymbol{\gamma}/\lambda)$ are represented as zeros in memory. Such issues occur similarly for other regu-

larizations, notably via the representation $\nabla\psi(-\boldsymbol{\gamma}/\lambda)$ of the unconstrained solution to project. Therefore, the proposed methods are actually competitive in a range where the penalty λ is not too small, and for which the rot mover's plan $\boldsymbol{\pi}^\star$ exhibits a significant amount of smoothing. Hence, we do not target the same problems as traditional schemes such as interior point methods or the network simplex.

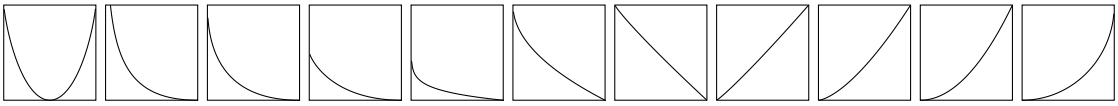
In addition, the different Bregman projections in our algorithms are most of the time approximate up to a given tolerance depending on the termination criterion used for convergence. Exceptions occur for the sum constraints with the Euclidean distance or Kullback-Leibler divergence, as well as the non-negativity constraints in the separable case, which are obtained analytically. A natural question to raise is then whether our algorithms still converge when the projections are approximate only. However, this is relatively hard to answer in theory. We did not observe in practice any problem of convergence when using sufficiently good approximations. Furthermore, first approximations can be quite rough without affecting convergence as long as final approximations are good enough. Sometimes, even alternating a single or two steps of the Newton-Raphson method throughout the main iterations the algorithm still works, though this is not systematic. Thus, we advocate for safety to use a tight tolerance for the auxiliary projections.

We also observed numerical instability of the Newton-Raphson updates for separable divergences under assumptions (A). This is due to the denominator being based on ψ'' with limit $\lim_{\theta\to-\infty}\psi''(\theta) = 0$, that is, for entries π close to zero. It is possible, however, to make the updates of μ_i, ν_j much more stable in practice by using the max truncation operator, despite theoretical guarantees of convergence without it. Specifically, we know that the entries $\pi^\star_{1,ij}$ must lie between 0 and p_i, and $\pi^\star_{2,ij}$ between 0 and q_j. Hence, we can lower bound μ_i and ν_j by $\hat{\theta}_i - \phi'(p_i)$ and $\hat{\theta}_j - \phi'(q_j)$, respectively. Interestingly, this also speeds up the convergence of the updates significantly when the initialization by 0 is far from the actual solution.

A possible termination criterion for the main and auxiliary iterations is to compute the marginal difference between the updated matrix and $\mathbf{p}, \mathbf{q}$. In the auxiliary iterations for the two scaling projections, we compare the sums of rows or columns to $\mathbf{p}$ or $\mathbf{q}$ respectively, and in the main iterations of the algorithm, we compare both marginals simultaneously. Typically, we can use the ℓ_p (quasi-)norm with $0 < p \leq +\infty$ to assess the marginal difference, and the auxiliary tolerance should be at least the main one for sufficient precision of the approximations. Two alternative quantities in absolute or relative scales can also be used for termination, either the variation with ℓ_p (quasi-)norm in the updated matrix or in the updated distance. Here the auxiliary tolerance should be at least the square of the main one. This seems reasonable for $\boldsymbol{\pi}^\star$ given the quadratic rate of convergence for the Newton-Raphson method versus the linear one for alternate Bregman projections, as well as for $\langle\boldsymbol{\pi}^\star, \boldsymbol{\gamma}\rangle$ under the Cauchy-Schwarz inequality. In all cases, convergence can be checked either after each iteration or after a given number of iterations to reduce the underlying cost of computing the termination criterion. We can also fix a maximum number of main and auxiliary iterations to limit the overall running time.

Regarding implementation, the matrix and vector operations used for ASA and NASA in the separable case are well-suited for fast calculation on a GPU and for processing of multiple input distributions in parallel. By working directly in the primal parameter space, the Sinkhorn-Knopp algorithm is also readily suited for dealing with sparse plans, based on existing libraries. In more general ROT problems, however, a specific library should be written for the sparse extension because null entries in the transport plan are not represented by null entries in the dual parameter space, so that tailored data structures and operations for such matrices need to be coded. Therefore, we only implemented and will focus in our experiments on the non-sparse version of our methods.

Finally, although we implicitly assumed throughout that the entries of $\mathbf{p}$ and $\mathbf{q}$ are strictly



—	$\beta = 0$	$\beta = 0.5$	$\beta = 1$	$p = 0.1$	$p = 0.5$	$p = 0.9$	$p = 1.1$	$p = 1.5$	$p = 2$	—
FDLOG	BETA	BETA	BETA	LPQN	LPQN	LPQN	LPN	LPN	LPN	HELL
—	BIS	—	BSKL	—	—	—	—	—	EUC	—

Figure 2: Separable regularizers on $(0, 1)$.

comprised between 0 and 1 for theoretical issues, it is often possible in practice to deal explicitly with null or unit entries in the input distributions. Intuitively, no mass can be moved from or to a null entry, so the transport plans have null rows and columns for the corresponding null entries of $\mathbf{p}$ and $\mathbf{q}$, respectively. In the separable case, we can thus simply remove these entries, solve the reduced ROT problem, and reinsert the corresponding null entries in the rot mover's plan $\boldsymbol{\pi}^\star$. The same reasoning as for the sparse extension can be made to show that our two algorithms still hold with this strategy from a theoretical standpoint. In the non-separable case, however, this is not as straightforward again because the influences of the different entries of $\boldsymbol{\pi}^\star$ are interleaved through the regularizer ϕ. Nonetheless, as long as we have $[0, 1)^{d \times d} \subset \operatorname{int}(\operatorname{dom} \phi)$ under assumptions (B), then we have $\Pi(\mathbf{p}, \mathbf{q}) \subset \operatorname{int}(\operatorname{dom} \phi)$ and we can apply NASA without modification. This is notably the case for the Mahalanobis distances whose domain is $\mathbb{R}^{d \times d}$. For a non-separable regularizer under assumptions (A), it is not easy to account for null entries because the constraint qualification $\Pi(\mathbf{p}, \mathbf{q}) \cap \operatorname{int}(\operatorname{dom} \phi) \neq \emptyset$ never holds due to mandatory null entries in the transport plans. Nevertheless, common regularizers under assumptions (A), including the ones used in this paper, are separable in general. Lastly, it is direct to cope with unit entries in $\mathbf{p}$ or $\mathbf{q}$ in all cases, since the transport polytope then reduces to a singleton, so that there is a unique transport plan $\mathbf{p}\mathbf{q}^\top$ which is the rot mover's plan.

5 Classical Regularizers and Divergences

In this section, we discuss the specificities of the ASA (Algorithm 1) and NASA (Algorithm 3) methods to solve ROT problems for classical regularizers and associated divergences. We start with several separable regularizers under assumptions (A), based on the Boltzmann-Shannon entropy related to the Kullback-Leibler divergence (`BSKL`, Section 5.1), the Burg entropy related to the Itakura-Saito divergence (`BIS`, Section 5.2), and the Fermi-Dirac entropy related to a logistic loss function (`FDLOG`, Section 5.3), as well as the parametric families of β-potentials related to the β-divergences (`BETA`, Section 5.4). We then discuss the separable ℓ_p quasi-norms (`LPQN`, Section 5.5), which require a slight adaptation of assumptions (A). We also consider separable regularizers under assumptions (B) related to ℓ_p norms (`LPN`, Section 5.6), as well as the Euclidean norm related to the Euclidean distance (`EUC`, Section 5.7) and the Hellinger distance (`HELL`, Section 5.8). Finally, we study a non-separable regularizer under assumptions (B) via quadratic forms in relation to Mahalanobis distances (Section 5.9). We plot all separable regularizers in Figure 2. All regularizers and their corresponding divergences are also summed up in Table 2. Lastly, we provide in Table 3 the related terms based on derivatives that are needed to instantiate the separable versions of ASA (Algorithm 2) or NASA (Algorithm 4) accordingly.

$\phi(\pi)$ / $\phi(\boldsymbol{\pi})$	$B_\phi(\pi\|\xi)$ / $B_\phi(\boldsymbol{\pi}\|\boldsymbol{\xi})$	$\operatorname{dom}\phi$	$\operatorname{dom}\psi$
Boltzmann-Shannon entropy $\pi\log\pi-\pi+1$	*Kullback-Leibler divergence* $\pi\log\frac{\pi}{\xi}-\pi+\xi$	$\mathbb{R}_+$	$\mathbb{R}$
Burg entropy $\pi-\log\pi-1$	*Itakura-Saito divergence* $\frac{\pi}{\xi}-\log\frac{\pi}{\xi}-1$	$\mathbb{R}_{++}$	$(-\infty,1)$
Fermi-Dirac entropy $\pi\log\pi+(1-\pi)\log(1-\pi)$	*Logistic loss function* $\pi\log\frac{\pi}{\xi}+(1-\pi)\log\frac{1-\pi}{1-\xi}$	$[0,1]$	$\mathbb{R}$
β-potentials $(0<\beta<1)$ $\frac{1}{\beta(\beta-1)}(\pi^\beta-\beta\pi+\beta-1)$	*β-divergences* $\frac{1}{\beta(\beta-1)}(\pi^\beta+(\beta-1)\xi^\beta-\beta\pi\xi^{\beta-1})$	$\mathbb{R}_+$	$(-\infty,\frac{1}{1-\beta})$
ℓ_p quasi-norms $(0<p<1)$ $-\pi^p$	$-\pi^p+p\pi\xi^{p-1}-(p-1)\xi^p$	$\mathbb{R}_+$	$\mathbb{R}_{--}$
ℓ_p norms $(1<p<+\infty)$ $\lvert\pi\rvert^p$	$\lvert\pi\rvert^p-p\pi\ \operatorname{sgn}(\xi)\lvert\xi\rvert^{p-1}+(p-1)\lvert\xi\rvert^p$	$\mathbb{R}$	$\mathbb{R}$
Euclidean norm $\frac{1}{2}\pi^2$	*Euclidean distance* $\frac{1}{2}(\pi-\xi)^2$	$\mathbb{R}$	$\mathbb{R}$
Hellinger distance $-(1-\pi^2)^{\frac{1}{2}}$	$(1-\pi\xi)(1-\xi^2)^{-\frac{1}{2}}-(1-\pi^2)^{\frac{1}{2}}$	$[-1,1]$	$\mathbb{R}$
Quadratic forms $(\mathbf{P}\succ\mathbf{0})$ $\frac{1}{2}\operatorname{vec}(\boldsymbol{\pi})^\top\mathbf{P}\operatorname{vec}(\boldsymbol{\pi})$	*Mahalanobis distances* $\frac{1}{2}\operatorname{vec}(\boldsymbol{\pi}-\boldsymbol{\xi})^\top\mathbf{P}\operatorname{vec}(\boldsymbol{\pi}-\boldsymbol{\xi})$	$\mathbb{R}^{d\times d}$	$\mathbb{R}^{d\times d}$

Table 2: Convex regularizers and associated Bregman divergences.

5.1 Boltzmann-Shannon Entropy and Kullback-Leibler Divergence

Assumptions (A) hold for minus the Boltzmann-Shannon entropy $\pi\log\pi-\pi+1$ associated to the Kullback-Leibler divergence. Hence, the ROT problem can be solved with the ASA scheme. In addition, the updates in the POCS technique can be written analytically, leading to the Sinkhorn-Knopp algorithm. Specifically, the two projections amount to normalizing in turn the rows and columns of $\boldsymbol{\pi}^\star$ so that they sum up to $\mathbf{p}$ and $\mathbf{q}$ respectively:

$$\boldsymbol{\pi}^\star \leftarrow \operatorname{diag}\left(\frac{\mathbf{p}}{\boldsymbol{\pi}^\star\mathbf{1}}\right)\boldsymbol{\pi}^\star \ , \tag{121}$$

$$\boldsymbol{\pi}^\star \leftarrow \boldsymbol{\pi}^\star\operatorname{diag}\left(\frac{\mathbf{q}}{\boldsymbol{\pi}^{\star\top}\mathbf{1}}\right) \ . \tag{122}$$

This can be optimized by remarking that the iterates $\boldsymbol{\pi}^{\star(k)}$ after each couple of projections verify:

$$\boldsymbol{\pi}^{\star(k)} = \operatorname{diag}(\mathbf{u}^{(k)})\boldsymbol{\xi}\operatorname{diag}(\mathbf{v}^{(k)}) \ , \tag{123}$$

where $\boldsymbol{\xi}=\exp(-\boldsymbol{\gamma}/\lambda)$, and vectors $\mathbf{u}^{(k)},\mathbf{v}^{(k)}$ satisfy the following recursion:

$$\mathbf{u}^{(k)} = \frac{\mathbf{p}}{\boldsymbol{\xi}\mathbf{v}^{(k-1)}} \ , \tag{124}$$

$$\mathbf{v}^{(k)} = \frac{\mathbf{q}}{\boldsymbol{\xi}^\top\mathbf{u}^{(k)}} \ , \tag{125}$$

$\phi(\pi)$	$\phi'(\pi)$	$\psi'(\theta)$	$\psi''(\theta)$
Boltzmann-Shannon entropy			
$\pi \log \pi - \pi + 1$	$\log \pi$	$\exp \theta$	$\exp \theta$
Burg entropy			
$\pi - \log \pi - 1$	$1 - \pi^{-1}$	$(1-\theta)^{-1}$	$(1-\theta)^{-2}$
Fermi-Dirac entropy			
$\pi \log \pi + (1-\pi)\log(1-\pi)$	$\log \frac{\pi}{1-\pi}$	$\frac{\exp \theta}{(1+\exp \theta)}$	$\frac{\exp \theta}{(1+\exp \theta)^2}$
β-potentials $(0 < \beta < 1)$			
$\frac{1}{\beta(\beta-1)}(\pi^\beta - \beta\pi + \beta - 1)$	$\frac{1}{\beta-1}(\pi^{\beta-1} - 1)$	$((\beta-1)\theta + 1)^{\frac{1}{\beta-1}}$	$((\beta-1)\theta + 1)^{\frac{1}{\beta-1}-1}$
ℓ_p quasi-norms $(0 < p < 1)$			
$-\pi^p$	$-p\pi^{p-1}$	$p^{-\frac{1}{p-1}}(-\theta)^{\frac{1}{p-1}}$	$-\frac{p^{-\frac{1}{p-1}}}{p-1}(-\theta)^{\frac{1}{p-1}-1}$
ℓ_p norms $(1 < p < +\infty)$			
$\lvert\pi\rvert^p$	$p\ \mathrm{sgn}(\pi)\lvert\pi\rvert^{p-1}$	$p^{-\frac{1}{p-1}}\ \mathrm{sgn}(\theta)\lvert\theta\rvert^{\frac{1}{p-1}}$	$\frac{p^{-\frac{1}{p-1}}}{p-1}\lvert\theta\rvert^{\frac{1}{p-1}-1}$
Euclidean norm			
$\frac{1}{2}\pi^2$	π	θ	1
Hellinger distance			
$-(1-\pi^2)^{\frac{1}{2}}$	$\pi(1-\pi^2)^{-\frac{1}{2}}$	$\theta(1+\theta^2)^{-\frac{1}{2}}$	$(1+\theta^2)^{-\frac{3}{2}}$

Table 3: Separable regularizers and related terms based on derivatives.

with convention $\mathbf{v}^{(0)} = \mathbf{1}$. This allows a fast implementation by performing only matrix-vector multiplications using a fixed matrix $\boldsymbol{\xi} = \exp(-\boldsymbol{\gamma}/\lambda)$. We can further save one element-wise vector multiplication per update:

$$\mathbf{u} \leftarrow \frac{\mathbf{1}}{\operatorname{diag}\left(\frac{\mathbf{1}}{\mathbf{p}}\right)\boldsymbol{\xi}\,\mathbf{v}}\ , \tag{126}$$

$$\mathbf{v} \leftarrow \frac{\mathbf{1}}{\operatorname{diag}\left(\frac{\mathbf{1}}{\mathbf{q}}\right)\boldsymbol{\xi}^\top\,\mathbf{u}}\ , \tag{127}$$

where the matrices $\operatorname{diag}\left(\frac{\mathbf{1}}{\mathbf{p}}\right)\boldsymbol{\xi}$ and $\operatorname{diag}\left(\frac{\mathbf{1}}{\mathbf{q}}\right)\boldsymbol{\xi}^\top$ are precomputed and stored.

5.2 Burg Entropy and Itakura-Saito Divergence

Assumptions (A) also hold for minus the Burg entropy $\pi - \log \pi - 1$ associated to the Itakura-Saito divergence, so the ROT problem can be solved with the ASA scheme. Eventually, the Newton-Raphson steps to update the alternate projections in POCS technique can be written as follows:

$$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \frac{(1-\boldsymbol{\theta}^\star)^{-1}\,\mathbf{1} - \mathbf{p}}{(1-\boldsymbol{\theta}^\star)^{-2}\,\mathbf{1}}\,\mathbf{1}^\top\ , \tag{128}$$

$$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \mathbf{1}\,\frac{\mathbf{1}^\top \left(1-\boldsymbol{\theta}^\star\right)^{-1} - \mathbf{q}^\top}{\mathbf{1}^\top \left(1-\boldsymbol{\theta}^\star\right)^{-2}} \quad . \tag{129}$$

Each step can be optimized by computing first an element-wise matrix inverse $\left(1-\boldsymbol{\theta}^\star\right)^{-1}$ for the numerator, and then performing an element-wise matrix multiplication of this matrix by itself to obtain a matrix for the denominator instead of applying an additional element-wise matrix power. Since ψ' is convex and strictly increasing with ψ'' positive everywhere, the convergence of the updates is guaranteed.

5.3 Fermi-Dirac Entropy and Logistic Loss Function

Assumptions (A) again hold for minus the Fermi-Dirac entropy $\pi \log \pi + (1-\pi)\log(1-\pi)$, also known as bit entropy, associated to a logistic loss function. The ROT problem can thus be solved with the ASA scheme, and the Newton-Raphson steps to update the alternate projections in the POCS technique can be written as follows:

$$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \frac{\frac{\exp \boldsymbol{\theta}^\star}{1+\exp \boldsymbol{\theta}^\star}\,\mathbf{1} - \mathbf{p}}{\frac{\exp \boldsymbol{\theta}^\star}{(1+\exp \boldsymbol{\theta}^\star)^2}\,\mathbf{1}}\,\mathbf{1}^\top \quad , \tag{130}$$

$$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \mathbf{1}\,\frac{\frac{\exp \boldsymbol{\theta}^\star}{1+\exp \boldsymbol{\theta}^\star} - \mathbf{q}^\top}{\mathbf{1}^\top \frac{\exp \boldsymbol{\theta}^\star}{(1+\exp \boldsymbol{\theta}^\star)^2}} \quad . \tag{131}$$

Each step can be optimized by storing first the element-wise matrix exponential $\exp \boldsymbol{\theta}^\star$, then applying an element-wise matrix division by the temporary matrix $1+\exp \boldsymbol{\theta}^\star$ to obtain a matrix for the numerator, and lastly performing an element-wise matrix division of these two matrices to obtain a matrix for the denominator and thus save an additional element-wise matrix power as well as several element-wise matrix exponentials. However, even if ψ' is strictly increasing with ψ'' positive everywhere, ψ' is neither convex nor concave and does not verify the necessary and sufficient condition (40) for global convergence of the Newton-Raphson method.

Nevertheless, ψ' is convex on $\mathbb{R}_-$ and concave on $\mathbb{R}_+$. It thus divides for a given $1 \leq i \leq d$, respectively $1 \leq j \leq d$, the real line into at most $d+1$ intervals $-\infty < \hat{\theta}_i^{(1)} \leq \hat{\theta}_i^{(2)} \leq \cdots \leq \hat{\theta}_i^{(d-1)} \leq \hat{\theta}_i^{(d)} < +\infty$, respectively $-\infty < \check{\theta}_j^{(1)} \leq \check{\theta}_j^{(2)} \leq \cdots \leq \check{\theta}_j^{(d-1)} \leq \check{\theta}_j^{(d)} < +\infty$, with the values $\left(\hat{\theta}_i^{(k)}\right)_{1\leq k\leq d}$ from row i of $\overline{\boldsymbol{\theta}}$, respectively $\left(\check{\theta}_j^{(k)}\right)_{1\leq k\leq d}$ from column j of $\overline{\boldsymbol{\theta}}$, sorted in increasing order. On each of these intervals, the necessary and sufficient condition (40) is verified since we can decompose $f(\mu_i)$, respectively $g(\nu_j)$, as the sum of an increasing convex and an increasing concave function. Hence, we have global convergence on the interval that contains the solution. It is further possible to restrict the search to the two last intervals only. Indeed, we have $\sum_{j=1}^d \psi'(\overline{\theta}_{ij} - \hat{\theta}_i^{(d-1)}) \geq \psi'(\hat{\theta}_i^{(d-1)} - \hat{\theta}_i^{(d-1)}) + \psi'(\hat{\theta}_i^{(d)} - \hat{\theta}_i^{(d-1)}) \geq 2\psi'(0) = 1$, so that $\hat{\theta}_i^{(d-1)} < \mu_i < +\infty$. Similarly, we have $\sum_{i=1}^d \psi'(\overline{\theta}_{ij} - \check{\theta}_j^{(d-1)}) \geq \psi'(\check{\theta}_j^{(d-1)} - \check{\theta}_j^{(d-1)}) + \psi'(\check{\theta}_j^{(d)} - \check{\theta}_j^{(d-1)}) \geq 2\psi'(0) = 1$, so that $\check{\theta}_j^{(d-1)} < \nu_j < +\infty$. As a result, it suffices to initialize μ_i with $\hat{\theta}_i = \hat{\theta}_i^{(d)} = \max\left\{\overline{\theta}_{ij}\right\}_{1\leq j\leq d}$, respectively ν_j with $\check{\theta}_j = \check{\theta}_j^{(d)} = \max\left\{\overline{\theta}_{ij}\right\}_{1\leq i\leq d}$, to guarantee convergence of the updates.

5.4 β-potentials and β-divergences

Assumptions (A) hold for the β-potentials $(\pi^\beta - \beta\pi + \beta - 1)/(\beta(\beta-1))$ with $0 < \beta < 1$, associated to the so-called β-divergences. Hence, the ROT problem can be solved with the ASA scheme,

and the Newton-Raphson steps to update the alternate projections in the POCS technique can be written as follows:

$$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \frac{((\beta-1)\boldsymbol{\theta}^\star+1)^{\frac{1}{\beta-1}}\,\mathbf{1}-\mathbf{p}}{((\beta-1)\boldsymbol{\theta}^\star+1)^{\frac{1}{\beta-1}-1}\,\mathbf{1}}\,\mathbf{1}^\top \enspace , \tag{132}$$

$$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star - \mathbf{1}\,\frac{\mathbf{1}^\top\left((\beta-1)\boldsymbol{\theta}^\star+1\right)^{\frac{1}{\beta-1}}-\mathbf{q}^\top}{\mathbf{1}^\top\left((\beta-1)\boldsymbol{\theta}^\star+1\right)^{\frac{1}{\beta-1}-1}} \enspace . \tag{133}$$

Each step can be optimized by computing first the temporary matrix $(\beta-1)\boldsymbol{\theta}^\star+1$, then applying an element-wise matrix power of $1/(\beta-1)-1$ to this temporary matrix to obtain a matrix for the denominator, and lastly performing an element-wise matrix multiplication of these two matrices to obtain a matrix for the numerator and thus save one element-wise matrix power. Since ψ' is convex and strictly increasing with ψ'' positive, the convergence of the updates is guaranteed.

Interestingly, the regularizer tends to minus the Burg and Boltzmann-Shannon entropies in the limit $\beta=0$ and $\beta=1$, respectively. Therefore, the β-divergences interpolate between the Itakura-Saito and Kullback-Leibler divergences. We finally remark that the regularizer can also be defined for other values of the parameter β using the same formula, but do not verify assumptions (A) for these values.

5.5 ℓ_p quasi-norms

Considering regularizers $-\pi^p$ with $0<p<1$, all assumptions (A) are verified except from (A5) since $\mathbb{R}_-^{d\times d}\not\subset\operatorname{dom}\psi=\mathbb{R}_{--}^{d\times d}$. Hence, our primal formulation does not hold here because $\mathbf{0}\notin\operatorname{dom}\nabla\psi$. However, it is straightforward to check that our dual formulation for ROT problems with the ASA scheme can still be applied as long as the cost matrix $\boldsymbol{\gamma}$ does not have null entries so that $-\boldsymbol{\gamma}/\lambda\in\operatorname{dom}\nabla\psi$. Eventually, the Newton-Raphson steps to update the alternate projections in the POCS technique can be written as follows:

$$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star + \frac{(-\boldsymbol{\theta}^\star)^{\frac{1}{p-1}}\,\mathbf{1}-p^{\frac{1}{p-1}}\,\mathbf{p}}{\frac{1}{p-1}(-\boldsymbol{\theta}^\star)^{\frac{1}{p-1}-1}\,\mathbf{1}}\,\mathbf{1}^\top \enspace , \tag{134}$$

$$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}^\star + \mathbf{1}\,\frac{\mathbf{1}^\top(-\boldsymbol{\theta}^\star)^{\frac{1}{p-1}}-p^{\frac{1}{p-1}}\,\mathbf{q}^\top}{\frac{1}{p-1}\,\mathbf{1}^\top(-\boldsymbol{\theta}^\star)^{\frac{1}{p-1}-1}} \enspace . \tag{135}$$

Each step can be optimized by computing first the temporary matrix $-\boldsymbol{\theta}^\star$, then applying an element-wise matrix power of $1/(p-1)-1$ to obtain a matrix for the denominator, and lastly performing an element-wise matrix multiplication of these two matrices to obtain a matrix for the numerator and thus save one element-wise matrix power. Since ψ' is convex and strictly increasing with ψ'' positive everywhere, the convergence of the updates is guaranteed.

5.6 ℓ_p norms

Assumptions (B) hold for the ℓ_p norms $|\pi|^p$ with $1<p<+\infty$, so the ROT problem can be solved with the NASA scheme. For $p\neq 2$, the Newton-Raphson steps to update the alternate projections in Dykstra's algorithm can be written as follows:

$$\boldsymbol{\tau} \leftarrow \boldsymbol{\tau} + \frac{\left\{\operatorname{sgn}(\boldsymbol{\theta}^\star-\boldsymbol{\tau}\mathbf{1}^\top)\odot|\boldsymbol{\theta}^\star-\boldsymbol{\tau}\mathbf{1}^\top|^{\frac{1}{p-1}}\right\}\mathbf{1}-p^{\frac{1}{p-1}}\,\mathbf{p}}{\frac{1}{p-1}|\boldsymbol{\theta}^\star-\boldsymbol{\tau}\mathbf{1}^\top|^{\frac{1}{p-1}-1}\,\mathbf{1}} \enspace , \tag{136}$$

$$\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} + \frac{\mathbf{1}^\top \left\{ \operatorname{sgn}(\boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top) \odot |\boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top|^{\frac{1}{p-1}} \right\} - p^{\frac{1}{p-1}}\, \mathbf{q}^\top}{\frac{1}{p-1}\, \mathbf{1}^\top |\boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top|^{\frac{1}{p-1}-1}} \ . \tag{137}$$

Denoting $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top$ or $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\sigma}^\top$ in the respective updates, each step can be optimized by computing first the temporary matrix $|\overline{\boldsymbol{\theta}}|$, then applying an element-wise matrix power of $1/(p-1)-1$ to obtain a matrix for the denominator, and lastly performing an element-wise matrix multiplication of these two matrices and of $\operatorname{sgn}\overline{\boldsymbol{\theta}}$ to obtain a matrix for the numerator and thus save one element-wise matrix power as well as several vector replications and matrix subtractions. However, even if ψ' is strictly increasing with $\psi'' > 0$ on $\mathbb{R}^*$, ψ' is neither convex nor concave and does not verify the necessary and sufficient condition (40) for global convergence of the Newton-Raphson method. Moreover, ψ'' vanishes at 0 for $p < 2$, and ψ' is not differentiable at 0 for $p > 2$.

Nevertheless, ψ' is concave on $\mathbb{R}_-$ and convex on $\mathbb{R}_+$ for $p < 2$, as well as convex on $\mathbb{R}_-$ and concave on $\mathbb{R}_+$ for $p > 2$. It thus divides for a given $1 \leq i \leq d$, respectively $1 \leq j \leq d$, the real line into at most $d+1$ intervals $-\infty < \hat{\theta}_i^{(1)} \leq \hat{\theta}_i^{(2)} \leq \cdots \leq \hat{\theta}_i^{(d-1)} \leq \hat{\theta}_i^{(d)} < +\infty$, respectively $-\infty < \check{\theta}_j^{(1)} \leq \check{\theta}_j^{(2)} \leq \cdots \leq \check{\theta}_j^{(d-1)} \leq \check{\theta}_j^{(d)} < +\infty$, with the values $(\hat{\theta}_i^{(k)})_{1 \leq k \leq d}$ from row i of $\overline{\boldsymbol{\theta}}$, respectively $(\check{\theta}_j^{(k)})_{1 \leq k \leq d}$ from column j of $\overline{\boldsymbol{\theta}}$, sorted in increasing order. The necessary and sufficient condition (40) is verified on the interior of each of these intervals since we can decompose $f(\mu_i)$, respectively $g(\nu_j)$, as the sum of an increasing convex and an increasing concave function. Hence, we have global convergence on the interior of the interval that contains the solution. In both cases, we must remove the finite endpoints to ensure differentiability of ψ' and positivity of ψ''. It is also further possible to prune the last interval from the search. Indeed, we have $\sum_{j=1}^d \psi'(\overline{\theta}_{ij} - \hat{\theta}_i^{(d)}) \leq \sum_{j=1}^d \psi'(0) = 0$, so that $\mu_i < \hat{\theta}_i = \hat{\theta}_i^{(d)} = \max\{\overline{\theta}_{ij}\}_{1 \leq j \leq d}$. Similarly, we have $\sum_{i=1}^d \psi'(\overline{\theta}_{ij} - \check{\theta}_j^{(d)}) \leq \sum_{i=1}^d \psi'(0) = 0$, so that $\nu_j < \check{\theta}_j = \check{\theta}_j^{(d)} = \max\{\overline{\theta}_{ij}\}_{1 \leq i \leq d}$. Lastly, we can restrict the first interval with a finite lower bound instead. Indeed, we have $\sum_{j=1}^d \psi'(\overline{\theta}_{ij} - \hat{\theta}_i^{(1)} + \phi'(p_i/d)) \geq \sum_{j=1}^d \psi'(\phi'(p_i/d)) = p_i$, so that $\mu_i \geq \hat{\theta}_i^{(1)} - \phi'(p_i/d)$. Similarly, we have $\sum_{i=1}^d \psi'(\overline{\theta}_{ij} - \check{\theta}_j^{(1)} + \phi'(q_j/d)) \geq \sum_{i=1}^d \psi'(\phi'(q_j/d)) = q_j$, so that $\nu_j \geq \check{\theta}_j^{(1)} - \phi'(q_j/d)$. As a result, we can perform at most d binary searches in parallel to determine within which of the remaining bounded intervals the solutions μ_i, respectively ν_j, lie. Initialization is then done with the midpoint to guarantee convergence of the updates. A given search thus requires a worst-case logarithmic number of tests, each of which requires a linear number of operations, for a total complexity in $O(d^2 \log d)$ instead of $O(d^2)$ if no such binary search were needed.

Now for $p = 2$, the regularizer specializes to the Euclidean norm, leading to the squared Euclidean distance as the associated divergence. In addition, the formula for ψ'' still holds with the convention $0^0 = 1$, and ψ'' is actually constant equal to $1/2$. Eventually, the projections can be written in closed form, and we can resort to the analytical algorithm derived in the next example specifically for the Euclidean distance, after doubling the penalty λ to account for the regularizer being halved.

5.7 Euclidean Norm and Euclidean Distance

Assumptions (B) hold for half the Euclidean norm $\pi^2/2$ associated to half the squared Euclidean distance. Therefore, the ROT problem can be solved with the NASA scheme, where Dykstra's algorithm can actually be written in closed form. Specifically, the non-negative projection reduces to:

$$\boldsymbol{\pi}^\star \leftarrow \max\{\mathbf{0}, \widetilde{\boldsymbol{\pi}}\} \ , \tag{138}$$

and is interleaved with the scaling projections which amount to offsetting the rows and columns of $\widetilde{\boldsymbol{\pi}}$ by an amount such that the rows and columns of $\boldsymbol{\pi}^\star$ sum up to $\mathbf{p}$ and $\mathbf{q}$ respectively:

$$\widetilde{\boldsymbol{\pi}} \leftarrow \widetilde{\boldsymbol{\pi}} - \frac{1}{d}(\boldsymbol{\pi}^\star \mathbf{1} - \mathbf{p})\,\mathbf{1}^\top \ , \tag{139}$$

$$\widetilde{\boldsymbol{\pi}} \leftarrow \widetilde{\boldsymbol{\pi}} - \frac{1}{d}\,\mathbf{1}\,(\mathbf{1}^\top \boldsymbol{\pi}^\star - \mathbf{q}^\top) \ . \tag{140}$$

As a remark, we notice that half the squared Euclidean distance can be seen as a β-divergence using the provided formula for $\beta = 2$. However, the β-divergence generated is not of Legendre type because the domain is restricted to $\mathbb{R}_+$, whereas it could actually be extended to $\mathbb{R}$ so that the regularizer would then be of Legendre type. This is why we fall under assumptions (B) rather than assumptions (A) in this case.

5.8 Hellinger Distance

Assumptions (B) hold for the regularizer $-(1-\pi^2)^{\frac{1}{2}}$ akin to a Hellinger distance. Hence, the ROT problem can be solved with the NASA scheme, and the Newton-Raphson steps to update the alternate projections in Dykstra's algorithm can be written as follows:

$$\boldsymbol{\tau} \leftarrow \boldsymbol{\tau} + \frac{\left\{(\boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top) \odot \left(1 + (\boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top)^2\right)^{-\frac{1}{2}}\right\}\mathbf{1} - \mathbf{p}}{\left(1 + (\boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top)^2\right)^{-\frac{3}{2}}\mathbf{1}} \ , \tag{141}$$

$$\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} + \frac{\mathbf{1}^\top\left\{(\boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\sigma}^\top) \odot \left(1 + (\boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\sigma}^\top)^2\right)^{-\frac{1}{2}}\right\} - \mathbf{q}^\top}{\mathbf{1}^\top\left(1 + (\boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\sigma}^\top)^2\right)^{-\frac{3}{2}}} \ . \tag{142}$$

Denoting $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star - \boldsymbol{\tau}\mathbf{1}^\top$ or $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star - \mathbf{1}\boldsymbol{\sigma}^\top$ in the respective updates, each step can be optimized by computing first the temporary matrix $1/(1+\overline{\boldsymbol{\theta}}^2)$, then applying an element-wise matrix square root to this temporary matrix, performing an element-wise matrix multiplication of these two matrices to obtain a matrix for the denominator, and lastly an element-wise matrix multiplication of the temporary matrix with $\overline{\boldsymbol{\theta}}$ to obtain a matrix for the numerator and thus save one element-wise matrix power as well as several vector replications and matrix subtractions. However, even if ψ' is strictly increasing with ψ'' positive everywhere, ψ' is neither convex nor concave and does not verify the necessary and sufficient condition (40) for global convergence of the Newton-Raphson method.

Nevertheless, ψ' is convex on $\mathbb{R}_-$ and concave on $\mathbb{R}_+$. It thus divides for a given $1 \leq i \leq d$, respectively $1 \leq j \leq d$, the real line into at most $d+1$ intervals $-\infty < \hat{\theta}_i^{(1)} \leq \hat{\theta}_i^{(2)} \leq \cdots \leq \hat{\theta}_i^{(d-1)} \leq \hat{\theta}_i^{(d)} < +\infty$, respectively $-\infty < \check{\theta}_j^{(1)} \leq \check{\theta}_j^{(2)} \leq \cdots \leq \check{\theta}_j^{(d-1)} \leq \check{\theta}_j^{(d)} < +\infty$, with the values $(\hat{\theta}_i^{(k)})_{1\leq k\leq d}$ from row i of $\overline{\boldsymbol{\theta}}$, respectively $(\check{\theta}_j^{(k)})_{1\leq k\leq d}$ from column j of $\overline{\boldsymbol{\theta}}$, sorted in increasing order. On each of these intervals, the necessary and sufficient condition (40) is verified since we can decompose $f(\mu_i)$, respectively $g(\nu_j)$, as the sum of an increasing convex and an increasing concave function. Hence, we have global convergence on the interval that contains the solution. It is further possible to prune the last interval from the search. Indeed, we have $\sum_{j=1}^d \psi'(\overline{\theta}_{ij} - \hat{\theta}_i^{(d)}) \leq \sum_{j=1}^d \psi'(0) = 0$, so that $\mu_i < \hat{\theta}_i = \hat{\theta}_i^{(d)} = \max\{\overline{\theta}_{ij}\}_{1\leq j\leq d}$. Similarly, we have $\sum_{i=1}^d \psi'(\overline{\theta}_{ij} - \check{\theta}_j^{(d)}) \leq \sum_{i=1}^d \psi'(0) = 0$, so that $\nu_j < \check{\theta}_j = \check{\theta}_j^{(d)} = \max\{\overline{\theta}_{ij}\}_{1\leq i\leq d}$. Lastly, we can restrict the first interval with a finite lower bound instead. Indeed, we have

$\sum_{j=1}^{d} \psi'(\overline{\theta}_{ij} - \hat{\theta}_i^{(1)} + \phi'(p_i/d)) \geq \sum_{j=1}^{d} \psi'(\phi'(p_i/d)) = p_i$, so that $\mu_i \geq \hat{\theta}_i^{(1)} - \phi'(p_i/d)$. Similarly, we have $\sum_{i=1}^{d} \psi'(\overline{\theta}_{ij} - \check{\theta}_j^{(1)} + \phi'(q_j/d)) \geq \sum_{i=1}^{d} \psi'(\phi'(q_j/d)) = q_j$, so that $\nu_j \geq \check{\theta}_j^{(1)} - \phi'(q_j/d)$. As a result, we can perform d binary searches in parallel to determine within which of the remaining intervals the solutions μ_i, respectively ν_j, lie. Initialization is then done with the midpoint to guarantee convergence of the updates. A given search requires a worst-case logarithmic number of tests, each of which requires a linear number of operations, for a total complexity in $O(d^2 \log d)$ instead of $O(d^2)$ if no such binary search were needed.

5.9 Quadratic Forms and Mahalanobis Distances

Assumptions (B) hold for the quadratic forms $(1/2)\operatorname{vec}(\boldsymbol{\pi})^\top \mathbf{P} \operatorname{vec}(\boldsymbol{\pi})$ with positive-definite matrix $\mathbf{P} \in \mathbb{R}^{d^2 \times d^2}$, associated to the Mahalanobis distances, so the ROT problem can be solved with the NASA scheme. For a diagonal matrix $\mathbf{P}$, the regularizer is separable and the Newton-Raphson steps to update the alternate projections in Dykstra's algorithm are similar to that for the Euclidean distance with appropriate weights. For a non-diagonal matrix $\mathbf{P}$, however, the regularizer is not separable anymore and we must resort to the generic NASA scheme.

In this general case, the scaling projections amount to convex quadratic programs with linear equality constraints. They can be solved using classical techniques such as the range-space and null-space approaches, Krylov subspace methods or active set strategies. The non-negative projection reduces to a convex quadratic program with a linear inequality constraint. It can be solved elegantly with an iterative algorithm for non-negative quadratic programming proposed by [46] using multiplicative updates with a complexity in $O(d^4)$. All in all, we recommend using a sparse matrix $\mathbf{P}$ with a block-diagonal structure and an order of magnitude of d^2 non-null entries, so as to obtain a quadratic instead of quartic empirical complexity.

6 Experimental Results

In this section, we present the results of our methods on different experiments. We first design an synthetic test to showcase the behavior of different regularizers and penalties on the output solutions or computational times (Section 6.1). We then consider a pattern recognition application to audio scene classification on a real-world dataset (Section 6.2).

6.1 Synthetic Data

We start by visualizing the effects of different regularizers ϕ and varying penalties λ on synthetic data. For the input distributions, we discretize and normalize continuous densities on a uniform grid $(x_i)_{1 \leq i \leq d}$ of $[0, 1]$ with dimension $d = 256$. We use for $\mathbf{p}$ a univariate normal with mean 0.5 and variance 0.2, and for $\mathbf{q}$ a mixture of two normals with equal weights, respective means 0.25 and 0.75, and same variance 0.1. We set the cost matrix $\boldsymbol{\gamma}$ as the squared Euclidean distance $\gamma_{ij} = (x_i - x_j)^2$ on the grid. The input distributions (bottom left and top right), cost matrix (top left) and unique earth mover's plan (bottom right) computed for classical OT using the solver of [41] with standard settings, are shown in Figure 3.

We test all separable regularizers ϕ introduced in Section 5. Because these regularizers have different ranges in the sensible values of the rot mover's plans $\boldsymbol{\pi}^\star$, we manually tune the penalties λ so that they feature similar amounts of regularization. For ease of comparison, we set $\lambda = \overline{\lambda}\,\lambda'$, with $\overline{\lambda}$ constant for each ϕ, and λ' varying similarly for all ϕ. The limit case when λ tends to infinity is simply obtained by setting $\boldsymbol{\gamma}/\lambda = \mathbf{0}$ in the algorithms, except from ℓ_p quasi-norms for which we use $\lambda = 10^{10}$. The null values of $\boldsymbol{\gamma}$ are also fixed to 10^{-12} for ℓ_p quasi-norms. We do

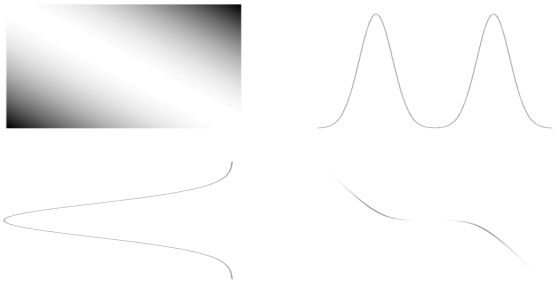


Figure 3: Earth mover's plan $\boldsymbol{\pi}^\star$ for the cost matrix $\boldsymbol{\gamma}$ and input distributions $\mathbf{p}, \mathbf{q}$.

not limit the number of iterations in the different algorithms, and use a small tolerance of 10^{-8} for convergence with the ℓ_∞ norm on the marginal difference checked after each iteration as a termination criterion.

The rot mover's plans obtained for ROT for $d = 256$ with the different regularizers and penalties are visualized in Figure 4. We first observe that all rot mover's plans converge to the earth mover's plan for low values of the penalty as shown theoretically in Property 9. Nevertheless, the rot mover's plans exhibit different shapes depending on the regularizers for intermediary and large values of the penalty. In the limit when the penalty grows to infinity, we obtain the transport plan with minimal Bregman information as shown theoretically in Property 7. In particular, this leads to $\mathbf{p}\mathbf{q}^\top$ with an ellipsoidal shape for `BSKL` (Boltzmann-Shannon entropy and Kullback-Leibler divergence), meaning that the mass is relatively spread among neighbor bins. The same pattern is observed for `FDLOG` (Fermi-Dirac entropy and logistic loss function), which can be explained in this synthetic example by the rot mover's plans having low values and the two regularizers being equivalent up to a constant in the neighborhood of zero. The profile gets more rectangular for `BIS` (Burg entropy and Itakura-Saito divergence), implying that the mass is even more spread across the different bins. Using an intermediary value $\beta = 0.5$ in `BETA` (β-potentials and β-divergences) allows the interpolation between these two limits of a rectangle for $\beta = 1$ and an ellipsoid for $\beta = 0$, so that the parameter β actually helps to control the spread of mass in the regularization. We observe similar results for `LPQN` (ℓ_p quasi-norms) with an ellipsoid for $p = 0.9$, a rectangle for $p = 0.1$, and a shape in between for $p = 0.5$. When the power parameter further increases in `LPN` (ℓ_p norms), we obtain new shapes that feature less spread of mass. These shapes for $p = 1.1$ and $p = 1.5$ now interpolate up to a lozenge for $p = 2$ in `EUC` (Euclidean norm and Euclidean distance), so that the parameter p also provides control on the spread of mass. A similar diamond profile is obtained for `HELL` (Hellinger distance), which is due again to the rot mover's plans having low values and the two regularizers being equivalent up to a constant in the neighborhood of zero. Lastly, we remark that varying the penalty between the two extremes allows a smooth interpolation of the earth mover's plan and optimal plan with minimal Bregman information, while keeping similar shapes and effects in terms of spreading of mass.

We next report in Table 4 the computational times required to reach convergence for the different regularizers and penalties. As a stopping criterion, we use the relative variation with tolerance 10^{-2} in ℓ_2 norm for the main loop of alternate Bregman projections, and the absolute variation with tolerance 10^{-5} in ℓ_2 norm for the auxiliary loops of the Newton-Raphson method. We use the same synthetic data as above but also vary the dimension d to assess its influence on speed. As already observed specifically for Sinkhorn distances [15], computing ROT distances is faster for important regularization with larger values of λ. The regularizers under assumptions

Figure 4: Rot mover's plans $\boldsymbol{\pi}^\star$ for different regularizers ϕ and penalties $\lambda = \overline{\lambda}\,\lambda'$.

Algorithm	ϕ	β/p	$\overline{\lambda}/\lambda'$	d = 128			256			512		
				10^{-2}	10^{-1}	10^{+0}	10^{-2}	10^{-1}	10^{+0}	10^{-2}	10^{-1}	10^{+0}
RMD	FDLOG	—	10^{-2}	0.366	0.079	0.044	1.091	0.311	0.116	1.865	0.571	0.273
RMD	BSKL BETA	1.00	10^{-2}	0.105	0.013	0.008	0.259	0.055	0.017	0.680	0.100	0.038
RMD	BETA	0.50	10^{-4}	0.971	0.102	0.044	1.922	0.251	0.147	3.526	1.339	0.281
RMD	BIS BETA	0.00	10^{-6}	0.916	0.106	0.019	1.466	0.108	0.053	2.598	0.398	0.096
RMD	LPQN	0.10	10^{-4}	0.968	0.068	0.055	0.732	0.173	0.152	0.416	0.309	0.305
RMD	LPQN	0.50	10^{-3}	0.404	0.057	0.042	0.778	0.163	0.160	0.780	0.305	0.304
RMD	LPQN	0.90	10^{-1}	0.226	0.047	0.040	0.751	0.178	0.131	1.110	2.492	0.214
RMD	LPN	1.10	10^{+0}	1.570	0.349	0.148	5.941	1.557	0.492	6.357	0.293	0.926
RMD	LPN	1.50	10^{+1}	0.399	0.099	0.053	1.170	0.474	0.166	6.688	2.163	0.532
RMD	EUC LPN	2.00	10^{+2}	0.074	0.043	0.043	0.253	0.240	0.237	7.308	3.190	0.966
RMD	HELL	—	10^{+2}	0.197	0.097	0.087	0.429	0.316	0.299	5.570	1.826	0.823
EMD OLD	—	—	—	—	0.231	—	—	1.912	—	—	10.95	—
EMD NEW	—	—	—	—	0.003	—	—	0.011	—	—	0.076	—

Algorithm	ϕ	β/p	$\overline{\lambda}/\lambda'$	d = 1024			2048			4096		
				10^{-2}	10^{-1}	10^{+0}	10^{-2}	10^{-1}	10^{+0}	10^{-2}	10^{-1}	10^{+0}
RMD	FDLOG	—	10^{-2}	4.156	3.517	1.410	15.85	9.663	5.109	54.19	33.87	17.24
RMD	BSKL BETA	1.00	10^{-2}	2.992	0.705	0.192	13.42	1.923	0.630	49.95	7.074	2.548
RMD	BETA	0.50	10^{-4}	8.015	2.769	0.888	42.95	7.538	3.557	101.3	21.13	10.86
RMD	BIS BETA	0.00	10^{-6}	4.439	0.777	0.550	6.590	3.262	2.218	41.80	12.96	6.742
RMD	LPQN	0.10	10^{-4}	4.068	2.174	1.291	6.962	4.890	4.334	51.15	15.86	14.02
RMD	LPQN	0.50	10^{-3}	7.819	4.198	1.314	26.34	6.129	4.301	53.74	15.65	11.98
RMD	LPQN	0.90	10^{-1}	3.584	2.264	1.054	13.80	4.571	3.285	43.83	14.22	11.51
RMD	LPN	1.10	10^{+0}	9.110	4.924	1.956	38.98	16.47	8.400	145.6	65.95	32.82
RMD	LPN	1.50	10^{+1}	18.97	9.509	2.539	61.92	20.41	9.314	236.6	77.87	45.94
RMD	EUC LPN	2.00	10^{+2}	11.90	5.805	2.161	31.43	14.22	4.906	117.1	50.88	27.67
RMD	HELL	—	10^{+2}	18.22	6.629	3.456	35.45	20.03	7.199	205.0	48.42	31.75
EMD OLD	—	—	—	—	85.56	—	—	482.2	—	—	$+\infty$	—
EMD NEW	—	—	—	—	0.482	—	—	2.760	—	—	13.23	—

Table 4: Computational times in seconds required to reach convergence for different regularizers ϕ and penalties $\lambda = \overline{\lambda}\,\lambda'$, with varying dimensions d.

(A) do not require the extra projections onto the non-negative orthant, and thus intuitively require less computational effort than the ones that verify assumptions (B). In addition, we notice that when the projections have closed-form expressions, the algorithms are also faster. The results further illustrate the influence of the data dimension d and the difference between ROT and classical OT performances. For a low dimension d, the RMD is competitive with EMD in his historical implementation `EMD OLD` [41]. The super-cubic complexity of the EMD with `EMD OLD` becomes prohibitive as the data dimension increases in contrast to the RMD which scales better. It should nevertheless be underlined that for reasonable dimensions, fast computation of the EMD can be obtained with a more recent, optimized implementation of the network simplex solver `EMD NEW` [10]. For higher dimensions, the super-cubic complexity makes `EMD NEW` less attractive, though it stays competitive with the RMD under a dimension $d = 4096$.

As a consequence, a numerical alternative to our algorithms for solving ROT problems with

reasonable dimensions is to rely on conditional gradient methods similar to [20]. Indeed, such methods imply the iterative resolution of linearized ROT problems, that can be reformulated as EMD problems and therefore be solved with the fast network simplex approach [10]. Lastly, for a fair interpretation of the above timing results, we must mention that the two EMD schemes tested were run under MATLAB from native C/C++ implementations[1][2] via compiled MEX files[3][4]. Hence, these EMD codes are quite optimized in comparison to our pure MATLAB prototype codes[5] for the RMD. It is thus plausible that optimized C/C++ implementations of our algorithms would be even more competitive in this context.

6.2 Audio Classification

We now assess our methods in the context of audio classification, and specifically address the task of acoustic scene classification where the goal is to assign a test recording to one of predefined classes that characterizes the environment in which it was captured. We consider the framework of the DCASE 2016 IEEE AASP challenge with the TUT Acoustic Scenes 2016 database [32]. The data set consists of audio recordings at 44.1 kHz sampling rate and 24-bit resolution. The metadata contains ground-truth annotations on the type of acoustic scene for all files, with a total of 15 classes: home, office, library, café/restaurant, grocery store, city center, residential area, park, forest path, beach, car, train, bus, tram, metro station. The audio material is cut into 30-second segments, and is split into two subsets of 75%–25% containing respectively 78–26 segments per class for development and evaluation, resulting in a total of 1170–390 files for training and testing. A 4-fold cross-validation setup is given with the training set. The classification accuracy, that is, the number of correctly classified segments among the total number of segments, is used as a score to evaluate systems.

A baseline system is also provided with the database for comparison. This system is based on Mel-frequency cepstral coefficient (MFCC) timbral features with Gaussian mixture model (GMM) classification. One GMM with diagonal covariance matrix is learned per class by expectation-maximization (EM), after concatenating and normalizing in mean and variance the extracted MFCCs from the training segments in that class. A test file is assigned to the class whose trained GMM leads to maximum likelihood for the extracted MFCCs for that file, where the MFCCs are considered as independent samples and normalized with the learned mean and variance for the respective classes. The baseline system is ran with its default parameters: 40 ms frame size, 20 ms hop size, 60-dimensional MFCCs comprising 20 static (including energy) plus 20 delta and 20 acceleration coefficients extracted with standard settings in RASTAMAT, 16 GMM components learned with standard settings in VOICEBOX.

Since MFCCs potentially take negative values, OT tools cannot be applied directly to this kind of features. Therefore, the common approach is to compute OT appropriately on GMMs estimated from MFCCs instead. Our proposed system follows this principle, and is implemented in the very same pipeline as the baseline for a fair comparison, with the following differences. One GMM is learned by EM for each training segment instead of class. Any normalization on the MFCCs per class is thus removed. Since less components are typically required to model one segment compared to one class, the spurious GMM components are further discarded as post-processing by keeping only those with weight and variances all greater than 10^{-2}. Instead of applying a GMM classifier, all individual models are exploited to train a support vector machine

[1] `http://robotics.stanford.edu/~rubner/emd/default.htm`
[2] `http://liris.cnrs.fr/~nbonneel/FastTransport/`
[3] `https://github.com/francopestilli/life/tree/master/external/emd`
[4] `https://arolet.github.io/code/`
[5] `https://www.math.u-bordeaux.fr/~npapadak/GOTMI/codes.php`

(SVM) classifier. An exponential kernel for the SVM is designed by introducing a distance between two mixtures P, Q based on the RMD as follows:

$$\kappa(P, Q) = \exp(-d_{\boldsymbol{\gamma},\lambda,\phi}(\boldsymbol{\omega}, \boldsymbol{\upsilon})/\tau) \ , \tag{143}$$

where the exponential decay rate $\tau > 0$ is a kernel parameter, and $\boldsymbol{\omega}, \boldsymbol{\upsilon} \in \Sigma_d$ are the respective weights of the $d = 16$ (or less) components for the two GMMs P, Q. The cost matrix $\boldsymbol{\gamma} \in \mathbb{R}_+^{d \times d}$ depends on P, Q and is the square root of a symmetrized Kullback-Leibler divergence, called the Jeffrey divergence, between the pairwise Gaussian components:

$$\gamma_{ij} = \sqrt{\frac{1}{4} \sum_{k=1}^{l} \frac{(\sigma_{ik}^2 - \varsigma_{jk}^2)^2 + (\sigma_{ik}^2 + \varsigma_{jk}^2)(\mu_{ik} - \nu_{jk})^2}{\sigma_{ik}^2 \varsigma_{jk}^2}} \ , \tag{144}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i^2$, respectively $\boldsymbol{\nu}_j$ and $\boldsymbol{\varsigma}_j^2$, are the means and variances of the $l = 60$ MFCC features for component i in the first mixture P, respectively component j in the second mixture Q. The SVM classifier is implemented with standard settings in LIBSVM, and requires an additional soft-margin parameter $C > 0$ to be tuned. Notice that, even if the kernel is not positive-definite, LIBSVM is still able to provide a relevant classification by guaranteeing convergence to a stationary point [2, 26, 30]. All separable regularizers ϕ from Section 5 with different penalties $\lambda > 0$ are tested for the RMD in comparison to the EMD. The two distances between $\mathbf{p}, \mathbf{q}$ and $\mathbf{q}, \mathbf{p}$ with cost matrix $\boldsymbol{\gamma}$ transposed are computed and averaged, so as to remove any asymmetry due to practical issues. The number of iterations is limited to 100 for the main loop of the algorithm and to 10 for the auxiliary loops of the Newton-Raphson method, and the tolerance is set to 10^{-6} in all loops for convergence with the ℓ_∞ norm on the marginal difference checked after each iteration as a termination criterion. The parameters τ, $C \in 10^{\{-1,+0,+1,+2\}}$ and penalty $\lambda \in \Lambda$, where Λ is a manually chosen set of four successive powers of ten depending on the range of the regularizer ϕ, are tuned automatically by cross-validation.

The obtained results on this experiment in terms of accuracy per system are reported in Table 5. The optimal penalties $\lambda \in \Lambda$ selected by cross-validation for each regularizer ϕ are also included, while the optimal parameters τ, C are not displayed since they actually all equal 10^{+1} independently of the kernel used. We first notice that the proposed system `SVM` (support vector machine classifier) consistently outperforms the baseline system `GMM` (Gaussian mixture model classifier). This proves the benefits of incorporating individual information per sound via an SVM rather than exploiting global information per class with a GMM. This further demonstrates the relevance of OT and more general ROT problems for the design of kernels between GMMs in the SVM pipeline. We also notice that `RMD` (rot mover's distance kernel) is at least competitive with `EMD` (earth mover's distance kernel) for all proposed regularizers, except from `EUC` which does not perform as well. This might be a consequence of the regularization profile for `EUC`, or equivalently `LPN` with $p = 2$, which does not spread enough mass across similar bins, implying a lack of robustness to slight variations in the means and variances of the GMM components. Reducing the power parameter in `LPN` brings back to a competitive system with `EMD` for $p = 1.1$, and even a better trade-off with improved accuracy for $p = 1.5$. We obtain similar results for `LPQN` with $p = 0.9$ and $p = 0.5$, with now the best compromise for the lowest power value $p = 0.1$ which clearly outperforms `EMD`. As a remark, the accuracy for `LPN` and `LPQN` is not unimodal with respect to p which controls the spread of mass in the regularization. We suspect this is because the performance is a function of both the spread of mass and the amount of regularization, whose coupling allows for similar compromises in terms of results within different regimes of use. Concerning `BETA` now, we observe that the existing Sinkhorn-Knopp algorithm `BSKL` for $\beta = 1$ does not improve the accuracy compared to `EMD`.

Classifier		— ϕ —		β/p	Λ	λ	Accuracy
GMM	—	—	—	—	—	—	77.2%
SVM	EMD	—	—	—	—	—	81.3%
	RMD	—	FDLOG	—	$10^{\{-2,-1,+0,+1\}}$	10^{-1}	81.0%
		BSKL	BETA	1.00	$10^{\{-2,-1,+0,+1\}}$	10^{+0}	81.3%
		—	BETA	0.50	$10^{\{-3,-2,-1,+0\}}$	10^{-2}	81.5%
		BIS	BETA	0.00	$10^{\{-4,-3,-2,-1\}}$	10^{-2}	81.3%
		—	LPQN	0.10	$10^{\{-2,-1,+0,+1\}}$	10^{-1}	82.1%
		—	LPQN	0.50	$10^{\{-2,-1,+0,+1\}}$	10^{-1}	81.3%
		—	LPQN	0.90	$10^{\{-2,-1,+0,+1\}}$	10^{+0}	81.0%
		—	LPN	1.10	$10^{\{-1,+0,+1,+2\}}$	10^{+0}	81.0%
		—	LPN	1.50	$10^{\{+0,+1,+2,+3\}}$	10^{+1}	81.8%
		EUC	LPN	2.00	$10^{\{+1,+2,+3,+4\}}$	10^{+3}	77.4%
		—	HELL	—	$10^{\{+1,+2,+3,+4\}}$	10^{+2}	82.8%

Table 5: Results of the experiment on audio classification.

Increasing the spread of mass with $\beta = 0$ in BIS is even worse. The best performance is obtained with a range in between for $\beta = 0.5$, which slightly improves results over EMD. Using LOG here slightly degrades the performance compared to EMD and BSKL. Interestingly, the overall best accuracy on this application is obtained for HELL which beats all other systems, including EUC, by a safe margin. In contrast to the experiment on synthetic data with dimension 256 presented in Section 6.1, where both BSKL and LOG, respectively EUC and HELL, behave similarly due to equivalence up to a constant for low values in the transport plans, the range of the transport plans here is much higher since the dimension of the input distributions is at most 16 (typically less than 10). This raises the importance of choosing a good regularizer depending on the actual task and its inherent design criteria such as the data dimension.

7 Conclusion

In this paper, we formulated a unified framework for smooth convex regularization of discrete OT problems. We also derived some algorithmic methods to solve such ROT problems, and detailed their specificities for classical regularizers and associated divergences from the literature. We finally designed a synthetic experiment to illustrate our proposed methods, and proved the relevance of ROT problems and the RMD on a real-world application to audio scene classification. The obtained results are encouraging for further development of the present work, and we now discuss some interesting perspectives for future investigation.

Firstly, we want to assess the effect of other regularizers on the solutions, notably when adding an affine term. From a geometrical viewpoint, such a transformation is equivalent to simply translating the cost matrix, with no effect on the Bregman divergence itself. For a given regularizer, we could therefore parametrize a whole family of interpolating regularizers, and tune the translation parameter according to the application. In particular, a recent work developed independently of ours makes use of Tsallis entropies to regularize OT problems with ad hoc solvers [34]. These regularizers could be integrated readily to our more general framework based on alternate Bregman projections, since Tsallis entropies are equivalent to β-potentials and ℓ_p (quasi)-norms up to an affine term.

In another direction, we would like to extend some theoretical results that hold for the

Boltzmann-Shannon entropy and associated Kullback-Leibler divergence. Specifically, it is known that the related rot mover's plan converges in norm to the earth mover's plan with an exponential rate as the penalty decreases [13]. It is not straightforward, however, to generalize this to other regularizers and divergences. In addition, it would be worth elucidating some technical restrictions under which metric properties such as the triangular inequality can be proved similarly to Sinkhorn distances.

We also plan to study other pattern recognition tasks in text, image and audio signal processing. Intuitive possibilities include retrieval and classification for various kinds of data modeled via histograms of features or GMMs. Among potential approaches, this can be addressed by exploiting the RMD either directly in a nearest-neighbor search, or in the design of kernels for an SVM as done here for acoustic scenes. For such tasks, it would be relevant to provide insight into the choice of a good regularizer for the actual problem, or develop methods for automatic tuning of regularization parameters, and for learning the cost matrix from the data as can be done for the EMD [16]. Even if we mostly focused on separable regularizers, it would be relevant to further use the quadratic forms associated to Mahalanobis distances in certain applications, and maybe propose a parametric learning scheme for the quadratic regularizer from the data.

Lastly, a more prospective idea is to use the RMD instead of Sinkhorn distances in the recent works built on the entropic regularization mentioned in Section 1. We also think that variational ROT problems could be formulated for statistical inference, notably parameter estimation in finite mixture models by minimizing loss functions based on the RMD [18]. This would leverage new applications of our ROT framework for more general machine learning problems. Such developments are yet involved and require some theoretical effort before reaching enough maturity to address practical setups.

Acknowledgments This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the "Investments for the future" Program IdEx Bordeaux (ANR-10-IDEX-03-02), Cluster of excellence CPU and the GOTMI project (ANR-16-CE33-0010-01). The authors would like to thank Annamaria Mesaros for her kind help with the evaluation on the DCASE 2016 IEEE AASP challenge, Charles-Alban Deledalle for his valuable advice on the use of the computing platform PlaFRIM, and Marco Cuturi for the insightful discussions about this work.

References

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[2] I. Alabdulmohsin, X. Gao, and X. Zhang. Support vector machines with indefinite kernels. In *Asian Conference on Machine Learning (ACML)*, pages 32–47, 2014.

[3] S.-i. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, USA, 2000.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. Technical report, arXiv:1701.07875, 2017.

[5] H. H. Bauschke and A. S. Lewis. Dykstra's algorithm with Bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[6] 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.

[7] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Inference in generative models using the Wasserstein distance. Technical report, arXiv:1701.05146, 2017.

[8] J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic PCA in the Wasserstein space. Technical report, arXiv:1307.7721, 2013.

[9] M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. Technical report, arXiv:1710.06276, 2017.

[10] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Transactions on Graphics*, 30(6):158:1–158:12, 2011.

[11] O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schölkopf. From optimal transport to generative modeling: the VEGAN cookbook. Technical report, arXiv:1705.07642, 2017.

[12] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis. Log-PCA versus geodesic PCA of histograms in the Wasserstein space. Technical report, arXiv:1708.08143, 2017.

[13] R. Cominetti and J. San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1–3):169–187, 1994.

[14] 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, 2015.

[15] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 2292–2300, 2013.

[16] M. Cuturi and D. Avis. Ground metric learning. *Journal of Machine Learning Research*, 15(1):533–564, 2014.

[17] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[18] A. Dessein, N. Papadakis, and C.-A. Deledalle. Parameter estimation in finite mixture models by regularized optimal transport: A unified framework for hard and soft clustering. Technical report, arXiv:1711.04366, 2017.

[19] I. S. Dhillon and J. A. Tropp. Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007.

[20] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[21] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein loss. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 2053–2061, 2015.

[22] A. Galichon and B. Salanié. Cupid's invisible hand: Social surplus and identification in matching models. Technical report, SSRN:1804623, 2015.

[23] A. Genevay, G. Peyré, and M. Cuturi. GAN and VAE from an optimal transport point of view. Technical report, arXiv:1706.01807, 2017.

[24] K. Grauman and T. Darrell. Fast contour matching using approximate earth mover's distance. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 220–227, 2004.

[25] J. Gudmundsson, O. Klein, C. Knauer, and M. Smid. Small Manhattan networks and algorithmic applications for the earth mover's distance. In *European Workshop on Computational Geometry (EuroCG)*, pages 174–177, 2007.

[26] B. Haasdonk. Feature space interpretation of SVMs with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):482–492, 2005.

[27] M. Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. Technical report, arXiv:1609.06349, 2016.

[28] P. Indyk and N. Thaper. Fast image retrieval via embeddings. In *International Workshop on Statistical and Computational Theories of Vision (SCTV)*, 2003.

[29] S. Kurras. Symmetric iterative proportional fitting. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 526–534, 2015.

[30] H.-T. Lin and C.-J. Lin. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical report, National Taiwan University, 2003.

[31] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853, 2007.

[32] A. Mesaros, T. Heittola, and T. Virtanen. TUT database for acoustic scene classification and sound event detection. In *European Signal Processing Conference (EUSIPCO)*, pages 1128–1132, 2016.

[33] G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein training of restricted Boltzmann machines. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 3718–3726, 2016.

[34] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen. Tsallis regularized optimal transport and ecological inference. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2387–2393, 2018.

[35] A. Naor and G. Schechtman. Planar earthmover is not in l_1. *SIAM Journal on Computing*, 37(3):804–826, 2007.

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

[37] O. Pele and M. Werman. A linear time histogram metric for improved SIFT matching. In *European Conference on Computer Vision (ECCV)*, pages 495–508, 2008.

[38] O. Pele and M. Werman. Fast and robust earth mover's distances. In *IEEE International Conference on Computer Vision (ICCV)*, pages 460–467, 2009.

[39] J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM Journal on Imaging Sciences*, 2(3):931–958, 2009.

[40] A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 630–638, 2016.

[41] 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.

[42] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngolè, D. Coeurjolly, M. Cuturi, P. Gabriel, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[43] B. Schmitzer. A sparse multi-scale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, 2016.

[44] B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. Technical report, arXiv:1610.06519, 2016.

[45] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 3312–3320, 2015.

[46] F. Sha, Y. Lin, L. K. Saul, and D. D. Lee. Multiplicative updates for nonnegative quadratic programming. *Neural Computation*, 19(8):2004–2031, 2007.

[47] S. Shirdhonkar and D. W. Jacobs. Approximate earth mover's distance in linear time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[48] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[49] 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 Transactions on Graphics*, 34(4):66:1–66:11, 2015.

[50] J. Solomon, R. M. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning (ICML)*, pages 306–314, 2014.

[51] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis. Overrelaxed sinkhorn-knopp algorithm for regularized optimal transport. Technical report, arXiv:1711.01851, 2017.

[52] L. Thorlund-Petersen. Global convergence of Newton's method on an interval. *Mathematical Methods of Operations Research*, 59(1):91–110, 2004.

[53] P. Tseng. Dual coordinate ascent methods for non-strictly convex minimization. *Mathematical Programming*, 59(1–3):231–247, 1993.

[54] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Comprehensive Studies in Mathematics*. Springer, Berlin Heidelberg, Germany, 2009.

[55] G. Zen, E. Ricci, and N. Sebe. Simultaneous ground metric learning and matrix factorization with earth mover's distance. In *International Conference on Pattern Recognition (ICPR)*, pages 3690–3695, 2014.