Noname manuscript No.
(will be inserted by the editor)

Fast Optimal Transport Regularized Projection and Application to Coefficient Shrinkage and Filtering

Antoine Rolet · Vivien Seguy

Accepted: 11/2020

Abstract This paper explores solutions to the problem of regularized projections with respect to the optimal transport metric. Expanding recent works on optimal transport dictionary learning and non-negative matrix factorization, we derive general purpose algorithms for projecting on any set of vectors with any regularization, and we further propose fast algorithms for the special cases of projecting onto invertible or orthonormal bases. Noting that pass filters and coefficient shrinkage can be seen as regularized projections under the Euclidean metric, we show how to use our algorithms to perform optimal transport pass filters and coefficient shrinkage. We give experimental evidence that using the optimal transport distance instead of the Euclidean distance for filtering and coefficient shrinkage leads to reduced artifacts and improved denoising results.

Keywords Optimal Transport, Coefficient Shrinkage, Sparse Decomposition, Wavelet Thresholding, Denoising

1 Introduction

Coefficient shrinkage has long been a staple method for signal denoising (Donoho, 1995; Kaur et al., 2002). In its simplest form, it consists in soft-thresholding the coefficients of a signal in the spectral domain (*e.g.* wavelet or Fourier), before going back to the signal domain. Let $\boldsymbol{x}$ be a vector representing a signal, D be the matrix representing a wavelet or Fourier basis, and $\boldsymbol{\lambda}$ be the coefficients of $\boldsymbol{x}$ in the spectral domain ($\boldsymbol{x} = D\boldsymbol{\lambda}$). Coefficient shrinkage of $\boldsymbol{x}$ is $D\theta_\alpha(\boldsymbol{\lambda}) = D\theta_\alpha(D^{-1}\boldsymbol{x})$, for some $\alpha \geq 0$ and with $\theta_\alpha \coloneqq \boldsymbol{\lambda} \mapsto \operatorname{sign}(\boldsymbol{\lambda})(\boldsymbol{\lambda} - \alpha)_+$. If D is orthonormal, which is the case for orthogonal wavelets and the discrete cosine transform, coefficient shrinkage amounts to the lasso problem:

$$\theta_\alpha(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{\lambda}} \|\boldsymbol{x} - D\boldsymbol{\lambda}\|_2^2 + \alpha\|\boldsymbol{\lambda}\|_1.$$

More generally, this problem falls in the scope of regularized least square problems:

$$\min_{\boldsymbol{\lambda}} \|\boldsymbol{x} - D\boldsymbol{\lambda}\|_2^2 + R(\boldsymbol{\lambda}),$$

which can also be thought of as a regularized Euclidean projection, where $\|\boldsymbol{x} - D\boldsymbol{\lambda}\|_2^2$ is a closeness term and R is used to enforce desired properties on $\boldsymbol{\lambda}$. Using an ℓ_1 norm as R leads to coefficient shrinkage, while an indicator function leads to pass-type filtering for example.

Using the Euclidean distance as the signal closeness term leads to artifacts on the reconstructed image $D\boldsymbol{\lambda}$. For example, filtering out high frequency components in the Fourier domain tends to create a "wave" pattern around sharp edges (Figure 1). In order to reduce these artifacts we propose to use instead the optimal transport distance, which instead of comparing images pixel-by-pixel, compute the best way to "transport" the intensity of the pixels of an image to fit the other image. This means that images are compared overall, instead of separately for each pixel, yielding less artifact on the reconstructed image, as shown in Figure 1c compared to Figure 1b.

This pdf is the accepted version. Please use the (improved) published version (https://link.springer.com/article/10.1007/s00371-020-02029-7) if possible.

Antoine Rolet
Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Kyoto, Japan E-mail: antoine.rolet@iip.ist.i.kyoto-u.ac.jp

V. Seguy
Nomad AI OU

(a) Original image *(b) Euclidean filter* *(c) Optimal transport filter*

Fig. 1: Effect of using the Euclidean or optimal transport distance as the closeness term for low pass filtering.

The optimal transport distance (*a.k.a.* the earth mover's distance or the Wasserstein distance) in its general form is a distance between probability measures. It has recently gained a lot of attention as a loss in optimization problems, with applications such as classification (Kusner et al., 2015; Frogner et al., 2015) and image generation (Arjovsky et al., 2017; Seguy et al., 2018). Optimal transport has also been used to tackle image processing problems, including Tartavel et al. (2016) for texture synthesis and image reconstruction, Rabin and Papadakis (2015) for foreground extraction or Solomon et al. (2015) for image interpolation.

In this paper, we study regularized projection of an image onto a fixed basis, or dictionary, where the reconstruction error is evaluated using the optimal transport distance. Projection onto a dictionary with respect to the optimal transport distance has been studied for musical note transcription (Flamary et al., 2016), and in the context of dictionary learning and non-negative matrix factorization (Sandler and Lindenbaum, 2009; Rolet et al., 2016, 2018). However these works did not consider the effect of different regularizers, sparsity-inducing or otherwise, nor did they analyze the qualitative effect of using the optimal transport as the reconstruction error for image processing specifically.

Our contributions. We give simple conditions on D and R for existence and unicity of the optimal transport regularized projection. We derive a method to compute this projection that can be used for any convex regularizer R and dictionary D. We further give fast algorithms for special cases depending on the properties of R and D. This allows us to perform pass-type filtering and sparse decomposition of images onto wavelet or Fourier bases, which was not possible using the previously existing methods of Rolet et al. (2016). Finally, we show how using the optimal transport distance as the reconstruction error leads to reduced artifacts for same level of sparsity when compared to the Euclidean distance.

This paper is organized as follows: in Section 2 we define the optimal transport distance between images, as well as the approximation that we use in order to make our problems tractable, and we introduce previous work in optimal transport dictionary learning. In Section 3 we proceed to give computational methods for solving optimal transport regularized projection. Building on these methods, we show in Section 4 how to perform optimal transport hard and soft thresholding and pass-type filtering, and compare optimal transport to the Euclidean distance in each case.

Notations

We denote matrices in upper-case, vectors in bold lower-case and scalars in lower-case. If M is a matrix, $M^\top$ is its transpose and $\mathrm{Im}(M)$ is the image of the linear map defined by M. $\mathbf{1}_n$ denotes the all-ones vector in $\mathbb{R}^n$; when the dimension can be deduced from context we simply write $\mathbf{1}$. For two matrices A and B of the same size, we denote their inner product $\langle A, B\rangle \coloneqq \mathrm{tr}\left(A^\top B\right)$, and their element-wise product as $A \odot B$. For a convex function f, f^* denotes its convex conjugate, defined as $f^*(\boldsymbol{x}) = \max_{\boldsymbol{y}} \langle \boldsymbol{x}, \boldsymbol{y}\rangle - f(\boldsymbol{y})$.

2 Background

In this section we first formalize the definition of optimal transport that we use throughout this paper and its regularized version. We then introduce relevant previous works in optimal transport non-negative matrix

factorization, which we build upon in further sections to perform optimal transport coefficient shrinkage.

2.1 Exact Optimal Transport

Definition. Let $\boldsymbol{x} \in \mathbb{R}_+^m$, $\boldsymbol{y} \in \mathbb{R}_+^n$. We define the polytope of transportation matrices between $\boldsymbol{x}$ and $\boldsymbol{y}$ as

$$U(\boldsymbol{x}, \boldsymbol{y}) = \left\{ T \in \mathbb{R}_+^{m \times n} \middle| \begin{matrix} T\mathbf{1} = \boldsymbol{x} \\ T^\top \mathbf{1} = \boldsymbol{y} \end{matrix} \right\}.$$

Note that $U(\boldsymbol{x}, \boldsymbol{y})$ is non-empty only if $\boldsymbol{x}^\top \mathbf{1} = \boldsymbol{y}^\top \mathbf{1}$ and $\boldsymbol{x}, \boldsymbol{y} \geq 0$. Let $C \in \mathbb{R}_+^{m \times n}$ be a matrix, where c_{ij} represents the cost of of moving weight from x_i to y_j, then the optimal transport cost between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as:

$$\mathrm{OT}(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} +\infty & \text{if } U(\boldsymbol{x}, \boldsymbol{y}) = \emptyset \\ \min\limits_{T \in U(\boldsymbol{x}, \boldsymbol{y})} \langle T, C \rangle & \text{otherwise.} \end{cases} \quad (1)$$

For a more complete introduction to optimal transport and an in-depth view of its application to different computational methods, we refer the reader to Peyré et al. (2019).

Optimal transport distance between images. In this paper, we use optimal transport to compute distances between images. We either have grayscale images, or color images for which we treat each color component independently. In any case, $\boldsymbol{x}$ is a vector representing intensity levels for each pixel of an $n \times m$ image, *i.e.* x_i is the intensity of the pixel located at coordinates $\mathcal{C}(i) \coloneqq (\lfloor i/m \rfloor, i \,\%\, m)$ in the image, where $\lfloor \cdot \rfloor$ is the integer part and $\%$ is the remainder operator. We use as the cost matrix C the matrix of pairwise squared Euclidean distances between the locations of the pixels, that is $c_{ij} = \|\mathcal{C}(i) - \mathcal{C}(j)\|_2^2$.

Our goal in this paper is to minimize a function of the form

$$f_{\boldsymbol{x}}(\boldsymbol{\lambda}) = \mathrm{OT}(\boldsymbol{x}, D\boldsymbol{\lambda}) + R(\boldsymbol{\lambda}),$$

where $\boldsymbol{x}$ is a vector representation of an image in $\mathbb{R}_+^n$, D a basis of $\mathbb{R}^n$, $g(\boldsymbol{\lambda}) = \mathrm{OT}(\boldsymbol{x}, D\boldsymbol{\lambda})$ a data-fitting term for the projection on D and R a regularizer. Direct optimization of such a function is tedious, because $\mathrm{OT}(\boldsymbol{x}, \cdot)$ is not differentiable, and the computation of a subgradient requires solving the linear program defined in Equation (1), for which best known algorithms run in $\mathcal{O}(n^3 \log n)$ (Orlin, 1997).

In order to alleviate this problem, we use a smooth approximation which is obtained by adding an entropy term to the optimal transport problem.

2.2 Entropy Regularized Optimal Transport

We propose to use an entropy regularized version of the optimal transport to solve optimization problems involving $\mathrm{OT}(\boldsymbol{x}, \cdot)$. The advantage of using this regularized optimal transport is twofold. First, similarly to Cuturi and Peyré (2016); Rolet et al. (2016), we take advantage of its smooth convex conjugate to derive dual problems that can be solved efficiently. Additionally, it lets us use further accelerations due to the special form of the cost matrix C in the case of optimal transport between images.

Definition. The entropy regularized optimal transport was proposed by Cuturi (2013) as a fast approximation of the optimal transport. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^n$, $\gamma > 0$, we define the entropy regularized optimal transport between $\boldsymbol{x}$ and $\boldsymbol{y}$ as:

$$\mathrm{OT}_\gamma(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} +\infty \text{ if } U(\boldsymbol{x}, \boldsymbol{y}) = \emptyset \\ \min\limits_{T \in U(\boldsymbol{x}, \boldsymbol{y})} \langle T, C \rangle + \gamma E(T) \text{ otherwise,} \end{cases} \quad (2)$$

where $E(T) \coloneqq \langle T, \log(T) \rangle$ is the entropy of T.

In recent years, entropy-regularized optimal transport has gained popularity as a proxy for the optimal transport as a loss in optimization problems (Gramfort et al., 2015; Frogner et al., 2015; Seguy et al., 2018) due to both its simplicity and good properties with respect to convex optimization. Indeed, contrary to exact optimal transport, the entropy regularized version is differentiable everywhere, and the simple form of its convex conjugate allows to derive tractable duals for many optimization problems involving OT_γ.

Convex conjugate. Let $\mathrm{OT}_\gamma^\star(\boldsymbol{x}, \cdot)$ be the convex conjugate of $\mathrm{OT}_\gamma(\boldsymbol{x}, \cdot)$:

$$\mathrm{OT}_\gamma(\boldsymbol{x}, \boldsymbol{y}) = \max_{\boldsymbol{h}} \langle \boldsymbol{y}, \boldsymbol{h} \rangle - \mathrm{OT}_\gamma(\boldsymbol{x}, \boldsymbol{h}).$$

Cuturi and Peyré (2016) showed that $\mathrm{OT}_\gamma^\star(\boldsymbol{x}, \cdot)$ can be expressed in closed form. Furthermore it is differentiable, its gradient is γ-Lipschitz and can also be expressed in closed form:

$$\mathrm{OT}_\gamma^\star(\boldsymbol{x}, \boldsymbol{y}) = \gamma \left(E(\boldsymbol{x}) + \langle \boldsymbol{x}, \log K\boldsymbol{\alpha} \rangle \right),$$
$$\nabla_{\boldsymbol{y}} \mathrm{OT}_\gamma^\star(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\alpha} \odot \left(K^\top \frac{\boldsymbol{x}}{K\boldsymbol{\alpha}} \right),$$

where $K \coloneqq e^{-C/\gamma}$ and $\boldsymbol{\alpha} \coloneqq e^{\boldsymbol{y}/\gamma}$.

Rolet et al. (2016) make use of the simple form of this convex conjugate and its gradient to derive fast a algorithm for the optimal transport dictionary learning problem. Section 3.2 showcases the computational gain

of using dual methods over primal ones on a simple regression problem.

The bottleneck in computing these formulas is the multiplication with matrix K. Supposing we are working with square images of size m, then $\boldsymbol{x}$ is of size $n = m^2$ and the complexity of multiplying with matrix K is $\mathcal{O}(n^2) = \mathcal{O}(m^4)$. Moreover storing matrix K also has a space complexity of $\mathcal{O}(n^2)$.

Accelerations. Since we use as a the cost matrix C the matrix of pairwise Euclidean distances on a grid representing the pixel locations of images, multiplications with matrix K and $K^\top$ are simply Gaussian convolutions of standard deviation $\sigma^2 = \gamma$ (Solomon et al., 2015, ¶5.). This allows us to compute $\mathrm{OT}_\gamma^\star$ in $\mathcal{O}(n \log n)$ instead of $\mathcal{O}(n^2)$, and to not store the matrix K in memory. Figure 2 shows experimental times for multiplicating K with a vector, implementing this operation as either a convolution or an actual matrix multiplication. For images of size lower or equal to 16, the matrix multiplication may be faster. This can be useful for example in dictionary learning or any other task in which images are divided into small patches. In this paper however we consider full images, accordingly we use the acceleration of Solomon et al. (2015) in all the results we report.

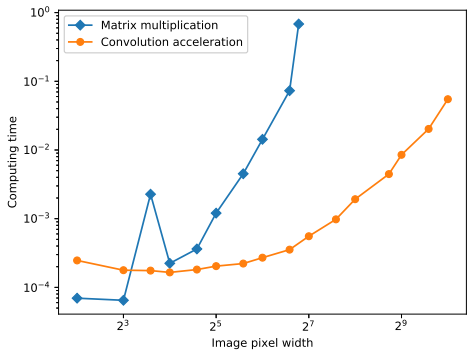


Fig. 2: Computational time for multiplication with K for a square image with respect to its width (log-scale)

Effect of using the entropy-regularized optimal transport. Since the entropy regularized optimal transport is not a distance, given an input $\boldsymbol{x}$, the point $\boldsymbol{y}$ which minimizes $\mathrm{OT}_\gamma(\boldsymbol{x}, \boldsymbol{y})$ is not $\boldsymbol{x}$:

Lemma 1 (Closest point) *Let $\boldsymbol{x} \in \mathbb{R}_+^n$,*

$$\operatorname*{argmin}_{\boldsymbol{y} \in \mathbb{R}_+^n} \mathrm{OT}_\gamma(\boldsymbol{x}, \boldsymbol{y}) = \frac{K^\top \boldsymbol{x}}{K\mathbf{1}}$$

Proof Let $g := \boldsymbol{x} \mapsto 0$, Fenchel duality tells us that

$$\begin{aligned}\min_{\boldsymbol{y}} \mathrm{OT}_\gamma(\boldsymbol{x}, \boldsymbol{y}) + g(\boldsymbol{y}) &= \max_{\boldsymbol{h}=\mathbf{0}} \mathrm{OT}_\gamma^*(\boldsymbol{x}, \boldsymbol{h}) \\ &= \mathrm{OT}_\gamma^*(\boldsymbol{x}, \mathbf{0})\end{aligned}$$

The primal-dual relationship gives us

$$\boldsymbol{y}^\star = \nabla \mathrm{OT}_\gamma^*(\boldsymbol{x}, \mathbf{0}) = \mathbf{1} \odot \left(K^\top \frac{\boldsymbol{x}}{K\mathbf{1}}\right).$$

In the case where $\boldsymbol{x}$ is an image and C is the matrix of squared Euclidean distances between pixel locations, this means that the closest point to any point with respect to the regularized optimal transport is simply a Gaussian blur of standard deviation $\sigma^2 = \gamma$, rescaled to have the same total intensity as the original image. Based on this observation, we set the regularization parameter γ of the entropy-regularized optimal transport to 0.1 in all of our results of Section 4, so that the closest point would be a blur of standard deviation 0.1 pixel, which is invisible to the naked eye.

In the case where $\boldsymbol{x}$ is not an image, Blondel et al. (2018) gives lower and upper bounds for the approximation given by a regularized transport, where the regularization can be the entropy or the squared Euclidean norm.

2.3 Optimal Transport Dictionary Learning

Regularized projection on a linear subspace can be seen as a part of the wider problem of regularized dictionary learning. Let $X \in \mathbb{R}_+^{n \times t}$ and $k \in \mathbb{N}$, the dictionary learning problem is

$$\min_{\lambda \in \mathbb{R}^{k \times t}} \min_{D \in \mathbb{R}^{n \times k}} \sum_i \ell(\boldsymbol{x}_i, D\boldsymbol{\lambda}_i) + R_1(\lambda) + R_2(D). \quad (3)$$

Particular cases where ℓ is either the Euclidean distance or the Kullback-Leibler divergence have been studied extensively. In particular, restricting λ and D to non-negative values lead to non-negative matrix factorization (NMF, Lee and Seung, 2001). Sparsity-inducing regularizations has been shown to yield good results for classification (Ataee and Mohseni, 2020), and image denoising and inpainting (Mairal et al., 2009), .

The optimal transport projection problem that interests us in this work is the sub-problem of Problem 3 where the dictionary D fixed and $\ell = \mathrm{OT}_\gamma$. Sandler and Lindenbaum (2009) showed that the optimal transport projection problem when $\gamma = 0$ and $R = 0$ is a linear program. However this linear program is in very high dimension with many constraints and is considered non-tractable even when n is relatively small.

Rolet et al. (2016) used the conjugate of OT_γ to get dual problems that can be solved efficiently. Their method was also used in Rolet et al. (2018) to perform NMF on sound data in the STFT domain, leading to good results in source separation and denoising. This approach is especially suited to image processing, since it lets us use the accelerations discussed in Section 2.2.

3 Optimal Transport Regularized Projection

We now show how to solve regularized optimal transport projection problems.

Let us fix $D \in \mathbb{R}_+^{n\times k}$, $\boldsymbol{x} \in \mathbb{R}_+^n$, and let R be a convex function. The regularized optimal transport projection of $\boldsymbol{x}$ onto D is the solution of

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^k} \mathrm{OT}_\gamma(\boldsymbol{x}, D\boldsymbol{\lambda}) + R(\boldsymbol{\lambda}). \tag{4}$$

Rolet et al. (2016) proposed fast dual methods for this problem either without a regularizer, or where R is the entropy in order to enforce non-negativity, in the context of NMF. We extend their methods for convex regularizers R with a smooth convex conjugate R^*. Further more we propose new methods for solving this problem when R^* is not smooth but D is orthonormal or simply invertible. These methods work as long as we have access to the proximal operator or R^*, either through a formula or a tractable algorithm. Finally we propose a general method which only requires a computable proximal operator for R.

3.1 Existence and unicity

We start by giving simple existence and unicity conditions for the solutions of Problem 4. Let

$$f : \boldsymbol{\lambda} \mapsto \mathrm{OT}_\gamma(\boldsymbol{x}, D\boldsymbol{\lambda}).$$

We can get simple existence conditions for the solutions of Problem 4 based on the domains of f and R, which we call dom_f and dom_R respectively.

Proposition 1 *If D is full rank and $\mathrm{Im}(D)\cap\mathbb{R}_+^k \neq \{0\}$, then dom_f is compact and non-empty.*

Proof Suppose that D is full-rank and $\mathrm{Im}(D) \cap \mathbb{R}_+^k \neq \{0\}$. Let $\boldsymbol{a} \in \mathrm{Im}(D) \cap \mathbb{R}_+^k$ such that $\boldsymbol{\lambda} \neq 0$. Let $\boldsymbol{b} = \frac{\|\boldsymbol{x}\|_1}{\|\boldsymbol{a}\|_1}\boldsymbol{a}$, we have $D\boldsymbol{b} \geq 0$ and $\|D\boldsymbol{b}\|_1 = \|\boldsymbol{x}\|_1$ so $\boldsymbol{b} \in \mathrm{dom}_f$ and dom_f is not empty.

Let us now prove that dom_f is compact. $\mathrm{dom}_f = \left\{\boldsymbol{\lambda} | D\boldsymbol{\lambda} \geq 0,\, \mathbf{1}^\top D\boldsymbol{\lambda} = \mathbf{1}^\top \boldsymbol{x}\right\}$ is a polyhedron defined as an intersection of an hyperplane and closed half-spaces. It is thus closed and as a subset of $\mathbb{R}^k$, it is compact *iif* it is unbounded, which for a polyhedron is equivalent to not containing any half line.

Let δ be a half-line, we will show that δ is not included in dom_f. There exist some vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$ with $\boldsymbol{b} \neq 0$, such that $\delta = \{\boldsymbol{a} + \beta\boldsymbol{b} | \beta \geq 0\}$.

Since D is full rank, $D\boldsymbol{b} \neq 0$. Let $0 < i \leq k$ such that $(D\boldsymbol{b})_i \neq 0$. There are three possible cases:

- $\boldsymbol{a}$ is not in dom_f, then δ is not included in dom_f.
- $\boldsymbol{a} \in \mathrm{dom}_f$ and $(D\boldsymbol{b})_i > 0$:
  Since $\boldsymbol{a} \in \mathrm{dom}_f$, we now that $(D\boldsymbol{a})_i \leq \|\boldsymbol{x}\|_1$. Let $\beta = \frac{\|\boldsymbol{x}\|_1 - (D\boldsymbol{a})_i + 1}{((D\boldsymbol{b})_i)}$, $(D(\boldsymbol{a}+\beta\boldsymbol{b}))_i = \|\boldsymbol{x}\|_1 + 1 > \|\boldsymbol{x}\|_1$, so $\boldsymbol{a} + \beta\boldsymbol{b}$ is not in dom_f and δ is not included in dom_f.
- $\boldsymbol{a} \in \mathrm{dom}_f$ and $(D\boldsymbol{b})_i < 0$:
  Since $\boldsymbol{a} \in \mathrm{dom}_f$, we now that $(D\boldsymbol{a})_i \geq 0$. Let $\beta = \frac{-(D\boldsymbol{a})_i - 1}{((D\boldsymbol{b})_i)}$, $(D(\boldsymbol{a}+\beta\boldsymbol{b}))_i = -1 < 0$, so $\boldsymbol{a}+\beta\boldsymbol{b}$ is not in dom_f and δ is not included in dom_f.

This shows that δ is not included in dom_f, As a closed polyhedron which contains no half-line, dom_f is bounded and thus compact. □

Proposition 2 (Existence) *Let R be a convex function. If $\mathrm{dom}_R \cap \mathrm{dom}_f$ is not empty and compact, then Problem 4 has a solution.*

Proof The conditions directly imply that Problem 4 is a convex problem over a non-empty compact set, so it has a solution. □

Unicity of a solution is follows from strict convexity of either f or R.

Proposition 3 (Unicity) *Let R be a convex function, $\gamma > 0$. If D is full rank, Problem 4 has at most one solution.*

Proof Suppose that D is full-rank, then it defines an injective linear map. Since $\gamma > 0$, $\mathrm{OT}_\gamma(\boldsymbol{x}, \cdot)$ is strictly convex. f is then strictly convex, and since R is convex the objective of Problem 4 is strictly convex. As a result it can have at most one solution. □

The previous result is only valid for the entropy-regularized optimal transport. We can get unicity of a solution with exact transport by restricting R to strictly convex functions:

Proposition 4 (Unicity II) *Let R be stricly convex function, $\gamma \geq 0$. Problem 4 has at most one solution.*

Proof $\mathrm{OT}_\gamma(\boldsymbol{x}, \cdot)$ is convex, so f is convex too and the objective of Problem 4 is strictly convex. As a result it can have at most one solution. □

In many of our applications, D is invertible and $\gamma > 0$. Proposition 3 then implies that Problem 4 has at most a solution. According to Proposition 1, dom_f is compact and non-empty. If we further have $\mathrm{dom}_R = \mathbb{R}^k$ for example, $\mathrm{dom}_R \cap \mathrm{dom}_f$ is not empty and compact and Problem 4 has a solution according to Proposition 2.

In the remainder of this paper, we assume existence and unicity for the wording of our results. However these results hold whether existence and unicity conditions are actually satisfied or not.

3.2 Dual Problem

We now derive a dual problem for Problem 4 which is the basis of most of the methods used in this paper.

Proposition 5 *The solution $\boldsymbol{\lambda}^\star$ of Problem 4 satisfies the primal-dual relationship*

$$D\boldsymbol{\lambda}^\star = \nabla \mathrm{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}^\star) \tag{5}$$

where $\boldsymbol{h}^\star$ is the solution of the dual problem

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \mathrm{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}) + R^*(-D^\top \boldsymbol{h}). \tag{6}$$

Proof The proof follows the same path as in Rolet et al. (2016), where R was the non-negative entropy. We rewrite Problem (4) as:

$$\min_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^k \\ \boldsymbol{p}\in\mathbb{R}^n_+ \\ D\lambda = p}} \mathrm{OT}_\gamma(\boldsymbol{x}, \boldsymbol{p}) + R(\boldsymbol{\lambda}).$$

It is a convex problem with linear constraints so strong duality holds, the problem is then:

$$\max_{\boldsymbol{h}\in\mathbb{R}^n} \min_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^k \\ \boldsymbol{p}\in\mathbb{R}^n_+}} \mathrm{OT}_\gamma(\boldsymbol{x}, \boldsymbol{p}) - \langle \boldsymbol{h}, \boldsymbol{p} - D\boldsymbol{\lambda}\rangle + R(\boldsymbol{\lambda}).$$

By definition of OT^*_γ, we get

$$\max_{\boldsymbol{h}\in\mathbb{R}^n} \min_{\boldsymbol{\lambda}\in\mathbb{R}^k} -\mathrm{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}) + \langle \boldsymbol{h}, D\boldsymbol{\lambda}\rangle + R(\boldsymbol{\lambda}) \tag{7}$$

$$-\min_{\boldsymbol{h}\in\mathbb{R}^n} \max_{\boldsymbol{\lambda}\in\mathbb{R}^k} \mathrm{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}) + \left\langle -D^\top \boldsymbol{h}, \boldsymbol{\lambda}\right\rangle - R(\boldsymbol{\lambda}) \tag{8}$$

Noting that the right side is the convex conjugate of R, we get the dual problem:

$$-\min_{\boldsymbol{h}\in\mathbb{R}^n} \mathrm{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}) + R^*(-D^\top \boldsymbol{h}). \tag{9}$$

Problem (6) is simply the Fenchel dual of the original problem, the primal-dual relationship in Equation (5) can be recovered from the first order conditions of Problem 7 with respect to variable $\boldsymbol{h}$.

If R^* is smooth and its gradient can be computed efficiently, we can solve Problem (6) with an accelerated gradient method (Nesterov, 1983).

3.3 Saddle Point Problem

If R^* is not smooth, or if its gradient is expensive to compute, we can still compute the projection by finding the saddle point in Problem (8). We propose to do this with a primal-dual approach such as Condat (2013) or Lorenz and Pock (2015). We use the algorithm defined in Theorem 9 of Lorenz and Pock (2015) to make use of preconditioning. Following their notations, we set:

$$\begin{cases} Q = \mathrm{OT}^*_\gamma(\boldsymbol{x}, \cdot), & G = 0, \quad K = -D^\top, \\ F^* = R, & P^* = 0, \quad \alpha_k = 0, \forall k. \end{cases}$$

This leads to updates:

$$\begin{cases} \boldsymbol{h}^{k+1} = \boldsymbol{h}^k - \tau(\nabla \mathrm{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}^k) - D\boldsymbol{\lambda}^k) \\ \boldsymbol{\xi}^{k+1} = 2\boldsymbol{h}^{k+1} - \boldsymbol{h}^k \\ \boldsymbol{\lambda}^{k+1} = \mathrm{prox}_{\sigma R}(\boldsymbol{\lambda}^k - \sigma D^\top \boldsymbol{\xi}^{k+1}), \end{cases}$$

where prox_f denotes the proximal operator of a function f. Solving the saddle-point problem in that way tends to be slow compared to full dual approaches, as we show in Section 3.5. We now focus on special conditions which allow to expand on Proposition 5.

3.4 Special Case: Invertible Dictionary

In the case where R^* is not smooth, we cannot solve Problem (6) directly with first order methods. However if D is invertible we can rewrite the problem and solve it with proximal methods.

Proposition 6 *Let $D \in \mathbb{R}^{n\times n}$ be an invertible matrix. The solution $\boldsymbol{\lambda}^\star$ of Problem 4 satisfies*

$$\boldsymbol{\lambda}^\star = D^{-1}\nabla \mathrm{OT}^*_\gamma(\boldsymbol{x}, -D^{\top -1}\boldsymbol{g}^\star) \tag{10}$$

where $\boldsymbol{g}^\star$ is the solution of

$$\min_{\substack{\boldsymbol{g}\in\mathbb{R}^n \\ -D^\top \boldsymbol{h} = \boldsymbol{g}}} \mathrm{OT}^*_\gamma(\boldsymbol{x}, -D^{\top -1}\boldsymbol{g}) + R^*(\boldsymbol{g}). \tag{11}$$

Proof Problem 11 is obtained by the change of variable $-D^\top \boldsymbol{h} = \boldsymbol{g}$ in Problem 6. This same change of variable gives us $D\boldsymbol{\lambda}^\star = \nabla \mathrm{OT}^*_\gamma(\boldsymbol{x}, -D^{\top -1}\boldsymbol{g}^\star)$.

Assuming that we have access to the proximal operator of R^*, we can solve Problem 11 efficiently with a proximal method such as FISTA(Beck and Teboulle, 2009).

Orthonormal dictionary. In the case where D is orthonormal, the problem of learning the coefficients can be solved with the invertible special case.

Table 1: Algorithms available based on the properties of R and D

Conditions	Method	Gradient	Proximal operator	Primal-dual relationship
R^* differentiable	accelerated gradient[1]	$\nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}) - D\nabla R^*(-D^\top \boldsymbol{h})$	Not used	$D\boldsymbol{\lambda}^\star = \nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}^\star)$
D invertible	FISTA[2]	$-D^{-1}\nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, -D^{-1\top}\boldsymbol{g})$	$\operatorname{prox}_{R^*}(\boldsymbol{g})$	$\boldsymbol{\lambda}^\star = D^{-1}\nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, -D^{\top -1}\boldsymbol{g}^\star)$
D orthonormal	FISTA[2]	$\nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h})$	$-D \operatorname{prox}_{R^*}(-D^\top \boldsymbol{h})$	$\boldsymbol{\lambda}^\star = D^\top \nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}^\star)$
None	forward-backward splitting[3]	$\nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h})$	$\operatorname{prox}_R(\boldsymbol{h})$	$\boldsymbol{\lambda}^\star$ is already available

Another solution arises if we rewrite Problem (6) as

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}) + \Pi(\boldsymbol{h}), \tag{12}$$

where $\Pi(\boldsymbol{h}) = R^*(-D^\top \boldsymbol{h})$. We can solve this new problem with FISTA. Indeed, since D is orthonormal, the proximal operator prox_Π of Π can be computed easily. By definition we have

$$\operatorname{prox}_\Pi(\boldsymbol{h}) = \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^n} \|\boldsymbol{h} - \boldsymbol{y}\|^2 - R^*(-D^\top \boldsymbol{y}).$$

Using the change of variable $\boldsymbol{z} = -D^\top \boldsymbol{y}$, we have

$$\operatorname{prox}_\Pi(\boldsymbol{h}) = -D \operatorname*{argmin}_{\boldsymbol{z}\in\mathbb{R}^n} \|\boldsymbol{h} + D\boldsymbol{z}\|^2 - R^*(\boldsymbol{z}).$$

Since D is orthonormal, it follows that

$$\begin{aligned}\|\boldsymbol{h} + D\boldsymbol{z}\|^2 &= \| - D^\top \boldsymbol{h} - D^\top D\boldsymbol{z}\|^2 \\ &= \| - D^\top \boldsymbol{h} - \boldsymbol{z}\|^2.\end{aligned}$$

We can thus compute the proximal operator of Π from that of R^*:

$$\begin{aligned}\operatorname{prox}_\Pi(\boldsymbol{h}) &= -D \operatorname*{argmin}_{\boldsymbol{z}\in\mathbb{R}^n} \| - D^\top \boldsymbol{h} - \boldsymbol{z}\|^2 - R^*(\boldsymbol{z}) \\ &= -D \operatorname{prox}_{R^*}(-D^\top \boldsymbol{h}).\end{aligned}$$

The primal-dual relationship becomes

$$\boldsymbol{\lambda}^\star = D^\top \nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}^\star). \tag{13}$$

We sum up our proposed methods in Table 1.

[1] Nesterov (1983)
[2] Beck and Teboulle (2009)
[3] Lorenz and Pock (2015)

3.5 Time Comparisons

Primal VS dual algorithms. In order to compare computation times between a direct primal method and our dual algorithms, we have to select a problem for which similar algorithms can be used, FISTA in this case. In particular we need a problem which can be divided into a smooth part and a part for that has a tractable proximal operator. Let us consider the simple following problem:

$$\min_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^k \\ \mathbf{1}^\top D\boldsymbol{\lambda} = \mathbf{1}^\top \boldsymbol{x} \\ D\boldsymbol{\lambda}\geq 0}} \operatorname{OT}_\gamma(\boldsymbol{x}, D\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2,$$

where D is the an orthonormal matrix[4]. We can project any $\boldsymbol{\lambda}$ on the constraint $\mathbf{1}^\top D\boldsymbol{\lambda} = \mathbf{1}^\top \boldsymbol{x}$ by projecting $D\boldsymbol{\lambda}$ on the non-negative part of the ℓ_1 sphere of radius $\mathbf{1}^\top \boldsymbol{x}$, and then applying $D^\top$ to the result. The objective is differentiable, we compute the optimal transport part and its gradient with the Sinkhorn algorithm (Cuturi, 2013). This algorithm's computational bottleneck is also the multiplication with $K = e^{\frac{C}{\gamma}}$, so it benefits from the convolution acceleration defined in Section 2.2 as much as our dual methods do.

We also solve the problem with the invertible case of Section 3.4, with $R(\boldsymbol{\lambda}) = \alpha\|\boldsymbol{\lambda}\|_2^2$. We then have $R^*(\boldsymbol{h}) = \frac{\alpha}{4}\|\boldsymbol{h}\|_2^2$ and $\operatorname{prox}_{R^*}(\boldsymbol{h}) = \frac{\boldsymbol{h}}{1+\alpha/2}$.

Figure 3 shows a time comparison of the FISTA algorithm used to solve the primal or dual problem, with either a fixed step-size or a step-size chosen by backtracking line-search. As the figure shows, our dual algorithm is orders of magnitude faster in any of the settings. For both methods, the backtracking line-search heuristic for choosing the step-size leads to faster convergence. However for the primal method, the precision σ to which we solve the regularized transport problem has a direct influence on the quality of the gradient. As a result backtracking line-search is not able to select

[4] Since D is orthonormal, the problem is actually equivalent to simply $\min_{\boldsymbol{\lambda}} \operatorname{OT}_\gamma(\boldsymbol{x}, \boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2$.

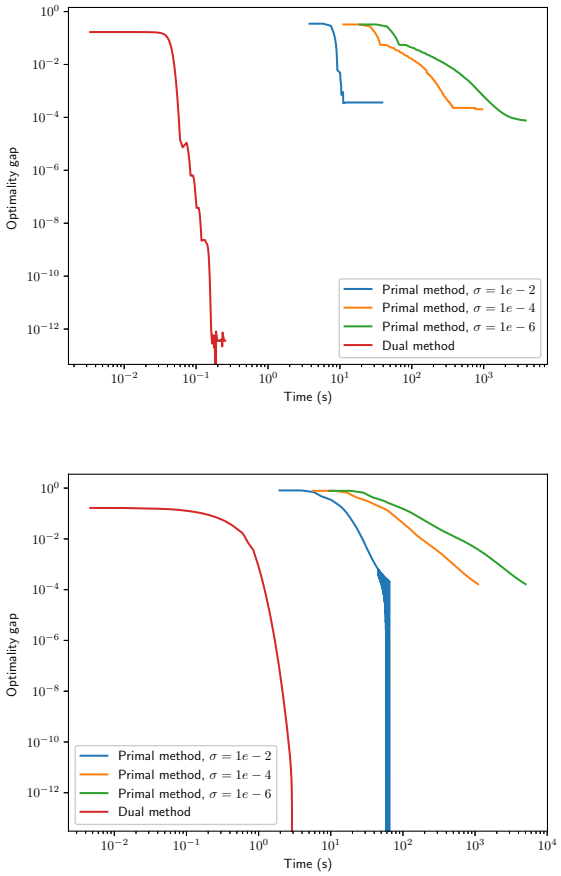


Fig. 3: Optimality gap with respect to time for a simple l2-regularized projection with a primal approach or our dual approach. Top: FISTA with backtracking line-search. Bottom: FISTA with a fixed step-size.

positive step-size after getting close to the optimal solution when the precision of the Sinkhorn algorithm is too low.

Saddle point VS dual algorithms. We now compare computation time for an optimal transport regularized projection problem using a primal-dual approach and a fully dual approach. We perform optimal transport coefficient shrinkage on the DCT coefficients of a 256×256 image using our dual approaches of Section 3.4 and the saddle point approach of Section 3.3. Although the saddle-point approach has the advantage of being valid for any dictionary D, Figure 4 shows that it is orders of magnitude slower to converge than dual approaches.

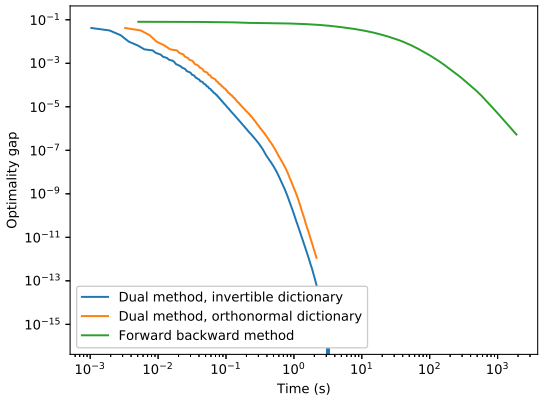


Fig. 4: Computation time for a same sparse projection problem with a primal-dual method or dual methods.

4 Applications

In this section, we show how to use the fast regularized projection methods we derived to perform optimal transport filtering, coefficient shrinkage and hard thresholding. We examine qualitative and quantitative differences of using optimal transport instead of the Euclidean distance on different image processing tasks, namely low-pass filtering, compressing and denoising.

4.1 Optimal Transport Filtering

Filtering is the process of setting a subset of the components of $\boldsymbol{\lambda}$ to 0. Let $\mathcal{N}$ be the set of indices of the components that we want to filter out, optimal transport filtering is performed by solving

$$\min_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^k \\ \lambda_i=0\ \forall i\in\mathcal{N}}} \mathrm{OT}_\gamma(\boldsymbol{x}, D\boldsymbol{\lambda}).$$

Although this problem can be solved by removing all non-relevant columns in D and using the non-regularized algorithm in Rolet et al. (2016), in the case of filtering components of DCT or wavelet transforms we can make use of our orthonormal dictionary case of Section 3.2 to get a simpler algorithm. Indeed, the filtering can be rewritten as a regularized projection on an orthonormal basis:

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^k} \mathrm{OT}_\gamma(\boldsymbol{x}, D\boldsymbol{\lambda}) + F_{\mathcal{N}}(\lambda)$$

where

$$F_{\mathcal{N}}(\boldsymbol{\lambda}) = \begin{cases} 0 & \text{if } \forall i \in \mathcal{N} \text{ s.t. } \lambda_i = 0 \\ \infty & \text{otherwise.} \end{cases}$$

The convex conjugate of $F_{\mathcal{N}}$ is

$$F^*_{\mathcal{N}}(\boldsymbol{h}) = \begin{cases} 0 & \text{if } \forall i \in \overline{\mathcal{N}} \text{ s.t. } \lambda_i = 0 \\ \infty & \text{otherwise.} \end{cases}$$

Fig. 5: Low-pass filtering of the DTC coefficients. Left: original image; Center: Euclidean filtering; Right: Optimal Transport filtering. Top: keeping the $1/16^{th}$ lowest frequencies. Bottom: keeping the $1/4^{th}$ lowest frequencies

$F^*_{\mathcal{N}}(\boldsymbol{h})$ is not differentiable, however its proximal operator is easy to compute:

$$\operatorname{prox}_{F^*_{\mathcal{N}}}(\boldsymbol{h})_i = \begin{cases} 0 & \text{if } i \in \overline{\mathcal{N}} \\ h_i & \text{otherwise.} \end{cases}$$

which is simply a regular filter on the complementary components to those described by $\mathcal{N}$. Thus we can solve the dual problem:

$$\min_{\boldsymbol{h} \in \mathbb{R}^n} \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}) + F^*_{\mathcal{N}}(-D^\top \boldsymbol{h}),$$

and recover $\lambda^\star$ through the primal-dual relationship: $\boldsymbol{\lambda}^\star = D^\top \nabla \operatorname{OT}^*_\gamma(\boldsymbol{x}, \boldsymbol{h}^\star)$.

We use this method to perform low-pass filters on images, and compare the results with regular low-pass filtering, which can be viewed as a regularized projection *w.r.t.* the Euclidean distance with the same regularizer.

Experimental results. Figure 5 shows the result of applying a low-pass filter on a 256×256 image, keeping either the $1/16^{th}$ or $1/4^{th}$ coefficients of its discrete cosine transform (DCT) of lowest frequency. We set the regularization parameter γ of the entropy-regularized optimal transport to 0.1, meaning that an optimal transport pass filter of full bandwidth would correspond to a Gaussian blur of standard deviation 0.1 pixel (see Section 2.2), which is almost invisible to the naked eye.

Both filtering methods show the wave-like patterns around edges in the image typical of DCT filtering, however these are more pronounced in the case of the classical, "Euclidean" filtering.

4.2 Coefficients Shrinkage and Thresholding

Let $\boldsymbol{x} \in \mathbb{R}^n_+$ be a non-negative vector and $D \in \mathbb{R}^{n \times n}$ be an invertible matrix, typically representing a discrete wavelet or fourier basis. Coefficient shrinkage of $\boldsymbol{x}$ usually refers to soft-thresholding of the coefficients $\boldsymbol{\lambda} = D^{-1}\boldsymbol{x}$ defined as:

$$\mathcal{S}_\alpha(\boldsymbol{\lambda}) = \operatorname{sign}(\boldsymbol{\lambda}) \odot \max\left\{|\boldsymbol{\lambda}| - \alpha, 0\right\}.$$

In the case where D is orthonormal, $\mathcal{S}_\alpha(D^{-1}\boldsymbol{x})$ is also the solution of the l_1 regularized Euclidean projection on D:

$$\mathcal{S}_\alpha(D^{-1}\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{\lambda}} \|\boldsymbol{x} - D\boldsymbol{\lambda}\|_2^2 + \alpha \|\boldsymbol{\lambda}\|_1 .$$

Hard thresholding on the other hand, is defined as

$$\mathcal{H}_\alpha(\boldsymbol{\lambda})_i = \begin{cases} \lambda_i & \text{if } |\lambda_i| > \alpha \\ 0 & \text{otherwise.} \end{cases}$$

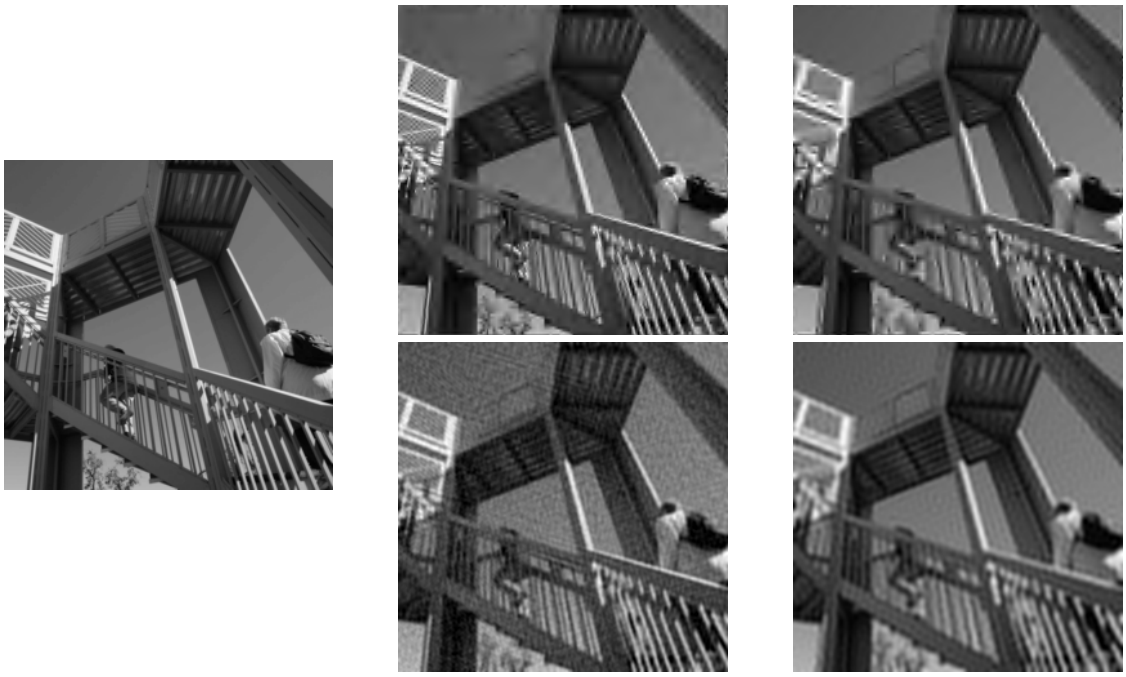


Fig. 6: Compression with Euclidean or Optimal Transport hard thresholding with biorthogonal spline wavelets of order 2 and dual order 4. Sparsity is set to 95%. Left: original image; Center: Euclidean hard thresholding; Right: Optimal Transport hard thresholding. Top: biorthogonal spline wavelets decomposition. Bottom: DCT decomposition.

Non-zero coefficients are the same for both hard and soft thresholding. If D is orthonormal, $\mathcal{H}_\alpha(D^{-1}\boldsymbol{x})$ is also the solution of the l_0 regularized euclidean projection on D:

$$\operatorname*{argmin}_{\boldsymbol{\lambda}} \|\boldsymbol{x} - D\boldsymbol{\lambda}\|_2^2 + \alpha \|\boldsymbol{\lambda}\|_0 .$$

Optimal transport shrinkage. We mirror this definition of shrinkage to define the optimal transport shrinkage of $\boldsymbol{x}$ as

$$\operatorname*{argmin}_{\boldsymbol{\lambda}} \mathrm{OT}_\gamma(\boldsymbol{x}, D\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_1$$

for some $\alpha > 0$. This problem can be solved efficiently through one of its dual, *i.e.* Problem (11) or Problem (12) with $R := \boldsymbol{\lambda} \mapsto \alpha\|\boldsymbol{\lambda}\|_1$. The convex conjugate of R^* of R is an indicator of the l_∞ ball of radius α, and its proximal is a projection on that same ball:

$$R^*(\boldsymbol{h}) = \begin{cases} 0 & \text{if } \|\boldsymbol{h}\|_\infty \leq \alpha \\ \infty & \text{otherwise,} \end{cases}$$
$$\mathrm{prox}_R^*(\boldsymbol{h}) = \mathrm{sign}(\boldsymbol{h}) \odot \min(|\boldsymbol{h}|, \alpha).$$

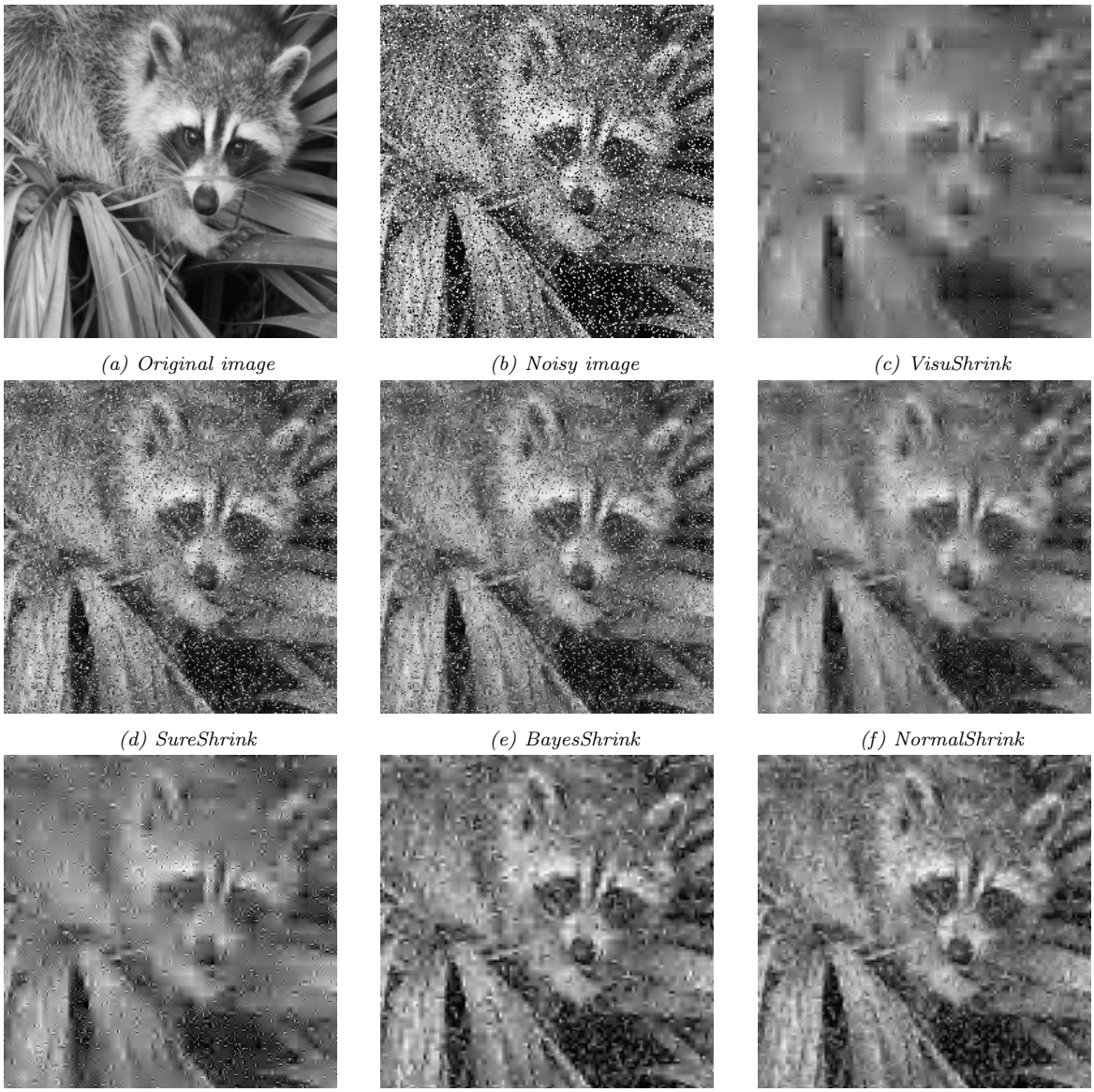
We recover $\boldsymbol{\lambda}^\star$ from the primal-dual relationships defined in Equation (10) or Equation (13). Because of machine precision, and of the fact that we can never solve the dual exactly, the coefficients we recover are not sparse, but a lot of them are very close to 0. We can however recover the sparsity pattern of $\boldsymbol{\lambda}^\star$ with the first order conditions for Problem (8) with respect to $\boldsymbol{\lambda}$. Indeed, these first order conditions are $-D^\top \boldsymbol{h}^\star \in \nabla R(\boldsymbol{\lambda}^\star)$, *i.e.*:

$$-D^\top \boldsymbol{h}^\star \in \left\{ \boldsymbol{a} \in \mathbb{R}^k \middle| \begin{matrix} -\alpha \leq a_i \leq \alpha & \text{if } \lambda_i^\star = 0 \\ a_i = \mathrm{sign}(\lambda_i^\star)\alpha & \text{otherwise} \end{matrix} \right\}.$$

Accordingly, we can set $\lambda_i^\star$ to 0 for all i such that $\left|(D^\top \boldsymbol{h}^\star)_i\right| < \alpha$.

Since $\|\cdot\|_1$ is convex, has full domain and D is full rank, the optimal transport coefficient shrinkage problem has a unique solution according to Proposition 3 and Proposition 2.

Optimal transport hard thresholding. Since the ℓ_0 norm is not a convex function, we do not have a method to solve the ℓ_0-regularized optimal transport projection. We define hard thresholding of the coefficients by analogy with the Euclidean case, based on the fact that the sparsity pattern for hard and soft thresholding is the same. In other words, hard thresholding corresponds to a pass filters on the non-zero coefficients



(a) Original image *(b) Noisy image* *(c) VisuShrink*

(d) SureShrink *(e) BayesShrink* *(f) NormalShrink*

(g) newThresh, $\alpha = 3.7$ *(h) OT hard thresholding* *(i) OT Shrinkage*

Fig. 7: Denoising of salt-and-pepper noise of level $\sigma = 10\%$ with Daubechies wavelets of order 2.

of the soft-thresholding operator. In terms of optimization problems, this means that if $\boldsymbol{\lambda}^\star$ is the solution of

$$\min_{\boldsymbol{\lambda}} \|\boldsymbol{x} - D\boldsymbol{\lambda}\|_2^2 + \alpha \|\boldsymbol{\lambda}\|_1 ,$$

and denoting $\mathcal{N} = \{i \text{ s.t. } \lambda_i^\star = 0\}$, then

$$\mathcal{H}_\alpha(D^{-1}\boldsymbol{x}) = \underset{\forall i \in \mathcal{N},\, \lambda_i = 0}{\operatorname{argmin}} \|\boldsymbol{x} - D\boldsymbol{\lambda}\|_2 .$$

Similarly, for $\alpha > 0$, we define the optimal transport hard thresholding as the optimal transport pass filter on the non-zero coefficients of

$$\operatorname*{argmin}_{\boldsymbol{\lambda}} \mathrm{OT}_\gamma(\boldsymbol{x}, D\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_1.$$

We compute the pass filter using the method defined in Section 4.1. Hard thresholding allows us to get better results when the level of noise on the signal is low. We analyze in the remainder of this section the effect of using optimal transport instead of the usual implicit Euclidean distance when performing hard and soft thresholding for either compression or denoising.

Noise	σ	normalShrink	sureShrink	bayesShrink	visuShrink	newThresh	OT hard	OT soft
Salt & pepper	5%	0.384	0.353	0.347	0.318	0.328	0.520	**0.545**
	10%	0.333	0.257	0.277	0.191	0.238	0.403	**0.441**
	15%	0.259	0.229	0.278	0.114	0.173	0.319	**0.350**
Gaussian	0.2	0.641	**0.706**	0.704	0.403	0.533	0.666	0.701
	0.3	0.532	0.604	0.600	0.312	0.424	0.581	**0.613**
	0.4	0.459	0.525	0.523	0.256	0.357	0.505	**0.537**

(a) SSIM

Noise	σ	normalShrink	sureShrink	bayesShrink	visuShrink	newThresh	OT hard	OT soft
Salt & pepper	5%	18.826	16.965	16.752	20.140	20.212	22.086	**22.911**
	10%	19.751	16.256	16.996	18.535	19.084	20.631	**21.077**
	15%	19.157	17.310	18.610	17.310	18.069	19.343	**19.781**
Gaussian	0.2	25.123	**25.903**	25.849	22.177	23.775	24.844	25.698
	0.3	23.594	24.163	24.172	20.889	22.288	23.308	**24.190**
	0.4	22.644	23.098	23.097	20.054	21.307	22.328	**23.171**

(b) pSNR

Table 2: Denoising scores for different wavelet thresholding methods

Compressing. Hard thresholding can be used to perform compressing, where the goal is to represent an image with as few coefficients as possible, while retaining good image quality.

Figure 6 shows the effect of optimal transport and Euclidean hard thresholding on the coefficients of either biorthogonal spline wavelets (Cohen et al., 1992) or DCT decomposition, where 5% of the coefficients are kept. Again we observe higher levels of artifacts with Euclidean thresholding.

For the biorthogonal spline wavelet decomposition, these artifact are especially visible in low contrast areas such as the background. As a result the fence-like structure at the top of the image has almost disappeared with Euclidean thresholding, but is still visible with optimal transport.

Denoising. We now examine how optimal transport thresholding compares to other wavelet coefficient shrinkage methods for image denoising. Many of the standard wavelet methods for image denoising perform either a soft or hard thresholding on the coefficients, which makes them inherently Euclidean sparse projection methods. Their main difference is on how to select the threshold. We compare our methods to visuShrink (Donoho and Johnstone, 1994), which selects one global threshold for the image, and *adaptive* methods which select a threshold for each wavelet decomposition level: sureShrink (Donoho and Johnstone, 1995), bayesShrink (Chang et al., 2000) and normalShrink (Kaur et al., 2002). We also compare our method with Dehda and Melkemi (2017), a thresholding method which uses a smooth thresholding function which can be seen as a trade-off between soft and hard thresholding. We call this method "newThresh" in the experiment.

With optimal transport, adaptive thresholding could be achieved by using either a weighted ℓ_1-norm or a block-sparse regularizer. However we found that this doesn't improve significantly upon simple ℓ_1-norm regularization and we only report the results of "global" thresholding here for simplicity.

For this experiment, we corrupt a 256×256 image with either a Gaussian or a salt-and-pepper noise with several noise levels σ. In the case of the Gaussian noise, σ is the variance and is taken to be 0.2, 0.3 or 0.4 times the mean intensity of the image. For the salt-and-pepper noise, $\sigma \in \{5\%, 10\%, 15\%\}$ is the proportion of pixels that are set to 0, and the same number are set to the maximum intensity (255). We perform coefficient shrinkage on the coefficients of Daubechies wavelets of order 2 (Daubechies, 1992) of the noisy image.

Figure 7 shows the images produced by the different thresholding methods for a salt-and-pepper noise. Similarly to the low-pass filtering and compression experiments, optimal transport based thresholding shows less wavelet artifacts. Hard thresholding appears to produce images that are sharper, but also more corrupted.

Table 2 reports the pSNR and SSIM (Wang et al., 2004) scores for each method and each noise. Our methods and newThresh each have a free parameter. For newThresh, we report the best score among 15 candidate shape parameters α in a log-scale interval from $1e-4$ to $1e4$. For optimal transport methods, we report the best score among 10 candidate regularization parameters α in the interval from 0.25 to 6. Optimal transport shrinkage improves upon all other methods for both denoising scores, except for a small intensity Gaussian noise, for which sureShrink and baryesShrink perform slightly better. In particular, optimal transport

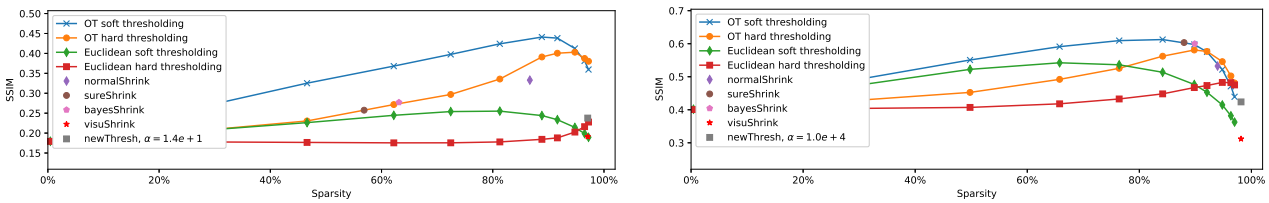


Fig. 8: Comparison of SSIM values as a fonction of sparsity. Competitors produce a single image, and are represented as a point. Top: salt-and-pepper noise with $\sigma = 10\%$. Bottom: Gaussian noise with of $\sigma = 0.3$.

brings significant improvement for a salt-and-pepper noise.

We now investigate how sparsity of the coefficients of the denoised image impacts the denoising score. Figure 8 plots the SSIM score with respect to sparsity for all methods. We first observe that if we only compare optimal transport thresholding to its Euclidean counterpart, optimal transport achieves a higher SSIM across all sparsities for both noises. Furthermore, we see that for the salt-and-pepper noise, optimal transport shrinkage achieves better results than other methods, even without selecting the sparsity carefully: all points of the blue curve above 40% sparsity have a better SSIM score than competitors, excluding optimal transport hard thresholding. With a Gaussian noise, which sureShrink and bayesShrink are optimized for, we can see that optimal transport shrinkage is still competitive, and achieves similar results as sureShrink and bayesShrink for the same sparsity.

With optimal transport shrinkage, the denoising output for an image is not defined uniquely, but rather is a function of sparsity (or of the regularization parameter). This is a good thing in a supervised setting, in which a user can modify the regularization parameter α until they are satisfied with the output. However it also means that in an unsupervised, or automated setting, we need a way to select the sparsity level based on the image. Based on Figure 8, a simple solution would be to pick any of the sparsities obtained by the outputs of sureShrink, bayesShrink and normalShrink, or their average sparsity.

5 Conclusion

In this paper we showed how to perform a regularized projection of a signal onto a fixed dictionary with respect to the optimal transport distance. We showed that while the general saddle point method is slow, we can get faster algorithms when either the regularizer's convex conjugate is differentiable or the dictionary is invertible. This last case allows us to perform sparse signal decomposition in various domains, including the discrete Fourier domain or wavelets. In practice, our results show that this optimal transport coefficients shrinkage yields less artifacts than coefficient shrinkage, where the signal is projected with respect to the Euclidean distance. For image denoising, it also outperforms other widely used wavelet based methods such as BayesShrink and SureShrink, especially for images corrupted with non-Gaussian noise.

We believe this showcases the need for further research in the area of optimal transport sparse projection. In particular fast algorithms for sparse projection onto a non-invertible dictionary would open the way to optimal transport sparse dictionary learning, allowing to expand on existing results with standard optimal transport dictionary learning in natural language processing and sound processing.

6 Declarations

6.1 Funding

This work was partly supported by JSPS KAKENHI Grant Number 17H01788.

6.2 Conflicts of interest

Not applicable

6.3 Availability of data and material

Not applicable

6.4 Code availability

A python library for our methods and all scripts necessary to reproduce our figures and results will be made available on the author's website upon publication of this paper.

6.5 Authors' contributions

AR and VS designed the research and wrote the paper. Experiments were performed by AR. All authors read and approved the final manuscript.

References

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. PMLR, Proceedings of Machine Learning Research, vol 70, pp 214–223

Ataee Z, Mohseni H (2020) Structured dictionary learning using mixed-norms and group-sparsity constraint. The Visual Computer 36(8):1679–1692

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2(1):183–202

Blondel M, Seguy V, Rolet A (2018) Smooth and sparse optimal transport. In: International Conference on Artificial Intelligence and Statistics, pp 880–889

Chang SG, Yu B, Vetterli M (2000) Adaptive wavelet thresholding for image denoising and compression. IEEE transactions on image processing 9(9):1532–1546

Cohen A, Daubechies I, Feauveau JC (1992) Biorthogonal bases of compactly supported wavelets. Communications on pure and applied mathematics 45(5):485–560

Condat L (2013) A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. Journal of Optimization Theory and Applications 158(2):460–479

Cuturi M (2013) Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems, pp 2292–2300

Cuturi M, Peyré G (2016) A smoothed dual approach for variational wasserstein problems. SIAM Journal on Imaging Sciences 9(1):320–343

Daubechies I (1992) Ten lectures on wavelets, vol 61. Siam

Dehda B, Melkemi K (2017) Image denoising using new wavelet thresholding function. Journal of Applied Mathematics and Computational Mechanics 16(2)

Donoho DL (1995) De-noising by soft-thresholding. IEEE transactions on information theory 41(3):613–627

Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. biometrika 81(3):425–455

Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association 90(432):1200–1224

Flamary R, Févotte C, Courty N, Emiya V (2016) Optimal spectral transportation with application to music transcription. In: Advances in Neural Information Processing Systems, pp 703–711

Frogner C, Zhang C, Mobahi H, Araya M, Poggio TA (2015) Learning with a wasserstein loss. In: Advances in Neural Information Processing Systems, pp 2053–2061

Gramfort A, Peyré G, Cuturi M (2015) Fast optimal transport averaging of neuroimaging data. In: International Conference on Information Processing in Medical Imaging, Springer, pp 261–272

Kaur L, Gupta S, Chauhan R (2002) Image denoising using wavelet thresholding. In: ICVGIP, vol 2, pp 16–18

Kusner M, Sun Y, Kolkin N, Weinberger K (2015) From word embeddings to document distances. In: International conference on machine learning, pp 957–966

Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562

Lorenz DA, Pock T (2015) An inertial forward-backward algorithm for monotone inclusions. Journal of Mathematical Imaging and Vision 51(2):311–325

Mairal J, Bach F, Ponce J, Sapiro G (2009) Online dictionary learning for sparse coding. In: Proceedings of the 26th annual international conference on machine learning, pp 689–696

Nesterov Y (1983) A method of solving a convex programming problem with convergence rate $o(1/k^2)$. In: Soviet Mathematics Doklady, vol 27, pp 372—376

Orlin J (1997) A polynomial time primal network simplex algorithm for minimum cost flows. Mathematical Programming 78(2):109–129

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

Rabin J, Papadakis N (2015) Convex color image segmentation with optimal transport distances. In: International Conference on Scale Space and Variational Methods in Computer Vision, Springer, pp 256–269

Rolet A, Cuturi M, Peyré G (2016) Fast dictionary learning with a smoothed wasserstein loss. In: Artificial Intelligence and Statistics, pp 630–638

Rolet A, Seguy V, Blondel M, Sawada H (2018) Blind source separation with optimal transport non-negative matrix factorization. EURASIP Journal on Advances in Signal Processing 2018(1):53

Sandler R, Lindenbaum M (2009) Nonnegative matrix factorization with earth mover's distance metric. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, pp 1873–1880

Seguy V, Damodaran BB, Flamary R, Courty N, Rolet A, Blondel M (2018) Large-scale optimal transport

and mapping estimation. In: International Conference on Learning Representations (ICLR)

Solomon J, de Goes F, Peyré G, Cuturi M, Butscher A, Nguyen A, Du T, Guibas L (2015) Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. ACM Transactions on Graphics 34(4)

Tartavel G, Peyré G, Gousseau Y (2016) Wasserstein loss for image synthesis and restoration. SIAM Journal on Imaging Sciences 9(4):1726–1755

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4):600–612