23rd International Symposium on Mathematical Theory of Networks and Systems
Hong Kong University of Science and Technology, Hong Kong, July 16-20, 2018

Optimal mass transport for regularizing inverse problems using generalized Sinkhorn iterations

Johan Karlsson[1] and Axel Ringh[1]

***Abstract*— The optimal mass transport problem has recently gained significant interest in application areas such as signal processing, image processing, and computer vision. Although the problem can be phrased as a linear program, in many cases the resulting optimization problem is intractable due to the vast number of variables. This issue was recently addressed by introducing a perturbation in terms of an entropic barrier term and solving the resulting optimization problem using Sinkhorn iterations. In this extended abstract, based on [23], we extend this to incorporate a class of optimization problems involving an optimal transport cost. In particular we show that the proximal operator of the optimal transport cost can be computed, also for large problems, using Sinkhorn-type iterations. By using a splitting framework, this is then used to solve inverse problems where the optimal mass transport cost is used for incorporating a priori information. We illustrate the method on a problem in limited angle computerized tomography, where a priori information is used to compensate for missing measurements.**

I. Introduction

The optimal mass transport problem is sometimes refereed to as the Monge-Kantorovich problem after the founding fathers Gaspard Monge and Leonid Kantorovich [30]. In the optimal transport problem the aim is to transform one function (distribution) into another by moving the mass of the function in a way that minimizes the cost of the movement. This minimal cost of movement can be used as a distance for comparing functions, which provides a geometric framework that can be used in many contexts. The latter is also reflected in that the approach lately has gained much interest in several application fields such as signal processing [15], [16], [19], image processing [18], [17], [21], computer vision and machine learning [2], [24], [29]. For an overview of the optimal mass transport problem, see, e.g., [30].

A drawback with the optimal transport problem is that it is often hard to solve. Monge's original formulation is a nonconvex optimization problem, while the formulation due to Kantorovich often results in large-scale optimization problems that are intractable (impossible) to solve with standard methods. Recently, a technique for approximating a solution to the Kantorovich formulation was suggested in [9]. In this approximation an entropic barrier term is added to the cost function, whereafter the resulting optimization problem is solved using the so called Sinkhorn iterations.

*This work was supported by the Swedish Research Council (VR) grant 2014-5870, the Swedish Foundation of Strategic Research (SSF) grant AM13-0049, National Science Foundation (NSF) grant CCF-1218388, and the Center for Industrial and Applied Mathematics (CIAM).

[1]Division of Optimization and Systems Theory, Department of Mathematics, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden. johan.karlsson@math.kth.se, aringh@kth.se

In this extended abstract, which is based on [23], we use and extend this idea. In particular, we consider variational problems that contain an optimal mass transport cost, i.e., problems of the form

$$\min_{\mu_{\text{est}}} \quad T_\epsilon(\mu_0, \mu_{\text{est}}) + g(\mu_{\text{est}}), \tag{1}$$

where T_ϵ is the entropy regularized optimal mass transport cost, and g both quantify data miss-match and can contain other regularization terms. These kind of problems occur frequently in regularization of inverse problems, in which μ_0 is typically a given prior and the minimizing argument μ_{est} is the sought reconstruction. However, they are also common in other settings, e.g., when solving gradient flow problems [5], [28] in the Jordan-Kinderlehrer-Otto framework [20], and thus they have been studied before. In [5] the authors consider a fluid dynamics formulation [4], and in [28] an entropic proximal operator is used for solving similar problems. In this work we will use a Douglas-Rachford type method [12], [7], which is a so called variable splitting technique in convex optimization [3], [11]. In order to do so we propose a new fast iterative computational method for computing the proximal operator of $T_\epsilon(\mu_0, \cdot)$ based on Sinkhorn-type iterations. This allows us to, e.g., solve large scale inverse problems of interest in medical imaging.

A. Notation

Most operations in this paper are defined elementwise. In particular we use $\odot$, $./$, exp, and log, to denote elementwise multiplication, division, exponential function, and logarithm function, respectively. We also use $\leq$ ($<$) to denote elementwise inequality (strict). Finally, let $\mathbf{1}_n$ denote the $n \times 1$ (column) vector of ones.

II. The optimal transport problem

The Kantorovich formulation of the optimal transport problem can be stated as follows: let $\Omega \subset \mathbb{R}^d$ be a compact set, and let μ_0 and μ_1 be two measures defined on Ω, with the same total mass. The optimal transport cost between μ_0 and μ_1, denoted $T(\mu_0, \mu_1)$, is defined as

$$\begin{aligned} T(\mu_0, \mu_1) = \quad & \min_{dM \geq 0} \quad \int_{(x_0,x_1)\in\Omega\times\Omega} c(x_0, x_1) dM(x_0, x_1) \\ & \text{subject to} \quad \mu_0(x_0)dx_0 = \int_{x_1\in\Omega} dM(x_0, x_1), \\ & \qquad\qquad\quad \mu_1(x_1)dx_1 = \int_{x_0\in\Omega} dM(x_0, x_1). \end{aligned} \tag{2}$$

Here, $c : \Omega \times \Omega \to \mathbb{R}_+$ is a given function that describes the cost for transporting a unit mass from one point to another,

and M is a nonnegative measure defined on $\Omega \times \Omega$, called the transference plan, describing how to transport the mass between μ_0 and μ_1.

As mentioned in the introduction, the optimal transport cost $T(\cdot,\cdot)$ can be used to to define a distance between two measures. To this end, let $c(x_0, x_1) = d(x_0, x_1)^p$ for $p \geq 1$ and where d is a metric on Ω. Then $W_p(\mu_0, \mu_1) := T(\mu_0, \mu_1)^{1/p}$ defines a metric (the Wasserstein p-metric) on the set of nonnegative measures on Ω with fixed mass [30, Theorem 6.9]. Furthermore, T (and W_p) is weak* continuous on this set, and in contrast to standard L^p metrics, optimal transport distance does not only compare the functions pointwise. Instead it quantifies the length that the mass is moved, which makes the distance natural for quantifying uncertainty and modelling deformations [13], [19], [22].

One way to solve the optimal transport problem in applications is to discretize Ω into a finite number of n points and solve the corresponding finite-dimensional linear programming problem. This problem takes the form

$$\begin{aligned} T(\mu_0, \mu_1) = \quad & \min_{M \geq 0} \quad \operatorname{trace}(C^T M) \\ & \text{subject to} \quad M\mathbf{1}_n = \mu_0, \quad M^T\mathbf{1}_n = \mu_1 \end{aligned} \tag{3}$$

where the matrix $M \in \mathbb{R}_+^{n\times n}$ corresponds to the transference plane and is defined by $M := [m_{ij}]_{ij}$, where m_{ij} denotes the amount of mass transported from point $x_{(i)}$ to $x_{(j)}$. Moreover, $C = [c_{ij}]_{ij}$, where $c_{ij} = c(x_{(i)}, x_{(j)})$ is the transportation cost from $x_{(i)}$ to $x_{(j)}$. As mentioned before, the issue with this formulation is the vast number of variables, a total of n^2. Thus if one seek to solve the optimal transport problem between two 256×256 images this results in $(256^2)^2$ variables, which is more than $4 \cdot 10^9$.

The approach suggested in [9] to circumvent this issue is to introduce the entropic regularizing term $D(M) = \sum_{i,j=1}^n (m_{ij}\log(m_{ij}) - m_{ij} + 1)$. The resulting optimization problem looks like

$$\begin{aligned} T_\epsilon(\mu_0, \mu_1) = \quad & \min_{M \geq 0} \quad \operatorname{trace}(C^T M) + \epsilon D(M) \\ & \text{subject to} \quad M\mathbf{1}_n = \mu_0, \quad M^T\mathbf{1}_n = \mu_1, \end{aligned} \tag{4}$$

and for ϵ small this is a good approximation to (3). Moreover, one can show that the solution to (4) takes the form

$$M = \operatorname{diag}(u) K \operatorname{diag}(v), \tag{5}$$

where $K = \exp(-C/\epsilon)$ is known, and $u, v \in \mathbb{R}_+^n$ are unknown. This reduces the number of variables from n^2 to $2n$, and the two vectors u and v can be computed iteratively by so called Sinkhorn iterations:

$$u = \mu_0./(Kv), \quad v = \mu_1./(K^T u).$$

III. The dual problem and generalized Sinkhorn iterations

One can show that the Sinkhorn iterations are in fact identical to block-coordinate ascent of the corresponding Lagrangian dual problem see [23], [10], [28]. We use this to derive algorithms similar to Sinkhorn iterations for solving problems on the form (1), and especially an algorithm for computing the proximal operator of $T_\epsilon(\mu_0, \cdot)$. However, in order for (1) to be well-defined and convex, we make the following assumption.

Assumption 1: Let g be a proper, convex and lower semi-continuous function that is finite in at least one point with mass equal to μ_0, i.e., $g(\mu_{\text{est}}) < \infty$ for some μ_{est} with $\sum_{i=1}^{n_0} \mu_0(i) = \sum_{j=1}^{n_1} \mu_{\text{est}}(j)$.

By putting the definition (4) into (1) and relaxing the constraints we can show the following proposition.

Proposition 1: Let $\mu_0 > 0$ be given and let g satisfy Assumption 1. Then the Lagrange dual of (1) is given by

$$\begin{aligned} \max_{\lambda_0, \lambda_1} \quad & \epsilon n^2 + \lambda_0^T \mu_0 - g^*(-\lambda_1) \\ & - \epsilon \exp(\lambda_0^T/\epsilon) \exp(-C/\epsilon) \exp(\lambda_1/\epsilon) \end{aligned} \tag{6}$$

and strong duality holds.

To find the solution to (6) we use block-coordinate ascent. This is done by considering the optimality conditions for (6), which is that zero must be a (sub)gradient of the cost function [6, pp. 711-712].

Lemma 1: For a fixed λ_1, then λ_0 is the maximizing vector of (6) if

$$\mu_0 = \exp(\lambda_0/\epsilon) \odot (\exp(-C/\epsilon)\exp(\lambda_1/\epsilon)). \tag{7a}$$

Similarly, for a fixed λ_0, then λ_1 is the maximizing vector of (6) if

$$0 \in \partial g^*(-\lambda_1) - \exp(\lambda_1/\epsilon) \odot \left(\exp(-C^T/\epsilon)\exp(\lambda_0/\epsilon)\right). \tag{7b}$$

The proximal operator of the transportation cost $T_\epsilon(\mu_0, \cdot)$ is defined as

$$\operatorname{Prox}^\sigma_{T_\epsilon(\mu_0,\cdot)}(\mu_1) := \arg\min_{\mu_{\text{est}}} T_\epsilon(\mu_0, \mu_{\text{est}}) + \frac{1}{2\sigma}\|\mu_{\text{est}} - \mu_1\|_2^2.$$

This is a special case of (1), where the data fitting term and the corresponding conjugate functional are

$$g(\mu) = \frac{1}{2\sigma}\|\mu - \mu_1\|_2^2, \qquad g^*(\lambda) = \lambda^T\left(\mu_1 + \frac{\sigma}{2}\lambda\right).$$

Therefore, Lemma 1 can be used in order to find the optimal solution. By alternatingly solving (7a) and (7b), for fixed λ_1 and λ_0 respectively, we obtain a dual block-coordinate ascent algorithm for solving the dual problem (6). The algorithm is shown in Algorithm 1, where ω denotes the elementwise Wright ω function, i.e., the function mapping $x \in \mathbb{R}$ to $\omega(x) \in \mathbb{R}_+$ for which $x = \log(\omega(x)) + \omega(x)$ [8]. The properties of the algorithm are summarized in the following theorem.

Theorem 1: The variables (λ_0, λ_1) in Algorithm 1 converges to the optimal solution of the dual problem (6), where $g(\mu) = \frac{1}{2\sigma}\|\mu - \mu_1\|_2^2$. Furthermore, the convergence rate is locally q-linear.

Remark 1: The bottlenecks in Algorithm 1 are the multiplications with the matrices K and K^T, since all other operations are elementwise. However, in many cases of interest the structures of K can be exploited for fast computations. This is true in particular when the discretization points $x_{(i)}$ are on a regular grid and the cost function is translation invariant. In this case the matrix C, and thus also K, is a multilevel Toeplitz-block-Toeplitz matrix, and the multiplication can be performed in $\mathcal{O}(n\log(n))$ using the fast Fourier transform.

Algorithm 1 Generalized Sinkhorn algorithm for evaluating the proximal operator of $T_\epsilon(\mu_0, \cdot)$.

Input: ϵ, C, λ_0, μ_0, μ_1

$K = \exp(-C/\epsilon)$

while Not converged **do**

  $\lambda_0 \leftarrow \epsilon \log\left(\mu_0 ./ (K \exp(\lambda_1/\epsilon))\right)$

  $\lambda_1 \leftarrow \frac{\mu_1}{\sigma} - \epsilon\omega\left(\frac{\mu_1}{\sigma\epsilon} + \log\left(K^T \exp(\lambda_0/\epsilon)\right) - \log(\sigma\epsilon)\right)$

end while

Output: $\mu_{\text{est}} \leftarrow \exp(\lambda_1/\epsilon) \odot (K^T \exp(\lambda_0/\epsilon))$

IV. Douglas-Rachford splitting and example in computerized tomography

In computerized tomography (CT), which is an imaging modality frequently used in medical imaging [25], [26], the object under investigation is probed with X-rays. Since different materials attenuate X-rays to different degrees, the intensities of the incoming and outgoing X-rays contain information of the material content and distribution. Mathematically, if $\mu_{\text{true}}(x)$ is the attenuation in the point x in the object then a set of measurements in CT corresponds to the line integral of μ_{true} along a limited set of lines. The operator that maps μ_{true} to the line integrals is called a partial Radon transform operator. This is a linear operator, and if A is a partial Radon transform operator the problem in CT is to reconstruct μ_{true} from measurements

$$b = A(\mu_{\text{true}}) + \text{noise}.$$

However, this is an ill-posed inverse problem [14, p. 40], in particular if the set of measurements is small or limited to certain angles. Hence regularization is needed to obtain an estimate μ_{est} of μ_{true}. This is often achieved by so called variational regularization.

A. Douglas-Rachford splitting for problems of type (1)

The variational regularization we consider here is a total variation (TV) reconstruction, but where we also use optimal transport in order to incorporate prior information. In particular, we consider the problem

$$\begin{aligned} \min_{\mu_{\text{est}}} \quad & \gamma T_\epsilon(\mu_0, \mu_{\text{est}}) + \|\nabla \mu_{\text{est}}\|_{2,1} \\ \text{subject to} \quad & \|A\mu_{\text{est}} - b\|_2 \leq \kappa, \end{aligned} \tag{8}$$

where μ_0 is a prior, κ quantifies the allowed measurement error, and γ determines the trade off between the optimal transport prior and the TV-regularization.

Douglas-Rachford splitting is an operator splitting technique for solving a rather general class of convex optimization problems. In [7] the authors consider Douglas-Rachford splitting for a family of problems that include problems of the form (1). To apply this algorithm one needs to be able to compute the proximal operators of the involved functionals, and since we can evaluate the proximal operator of the optimal transport term we can apply the Douglas-Rachford splitting for solving (8) (see [23] for details).

B. Numerical simulation

In this example we consider the Shepp-Logan phantom shown in figure 1a, of resolution 256×256 pixels, and assume that the deformed image in figure 1b, also 256×256 pixels, has been reconstructed previously from a detailed CT scan of the patient. We want to use this deformed image as prior information μ_0 in the problem (8) in order to improve the reconstruction. In the example, data from the phantom is obtained from 350 parallel lines, from 30 equidistant angles in the interval $[\pi/4,\ 3\pi/4]$. Moreover, on this data set 5% white Gaussian noise is added.[1]

The reconstruction is shown in figure 2c, together with a TV-reconstruction in figure 2a and a reconstruction where the prior information is incorporated using the ℓ_2 distance in figure 2b. From the figures we see that the TV-reconstruction suffers from artifacts and sever vertical blurring due to poor vertical resolution resulting from the limited angle measurements. For the ℓ_2 reconstruction some details are visible, however these are at the same locations as in the prior and does not adjust according to the measurements from the phantom. Considerable artifacts also appear in this reconstruction, typically as fade-in-fade-out effects where the prior and the data do not match. This effect can not be mitigated by the choice of regularization parameter, but is inherent in that ℓ_2 is a pointwise and strong metric (see [23]). For the reconstruction with optimal mass transport prior some blurring occurs, especially in the top and the bottom of the image. However, the overall shape is better preserved compared to the other reconstructions. Fine details are not visible, but the major features are better estimated compared to the TV- and ℓ_2-reconstructions. This example illustrates how one can improve the reconstruction by incorporating prior information, but without the fade-in-fade-out effects that typically occurs when using a strong metric such as ℓ_2 for regularization.

References

[1] J. Adler, H. Kohr, and O. Öktem. ODL - a Python framework for rapid prototyping in inverse problems. *In preparation, KTH, Royal Institute of Technology. Code and documentation available online:* `https://github.com/odlgroup/odl`.

[2] J. Adler, A. Ringh, O. Öktem, and J. Karlsson. Learning to solve inverse problems using Wasserstein loss. *arXiv preprint arXiv:1710.10898*, 2017.

[3] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011.

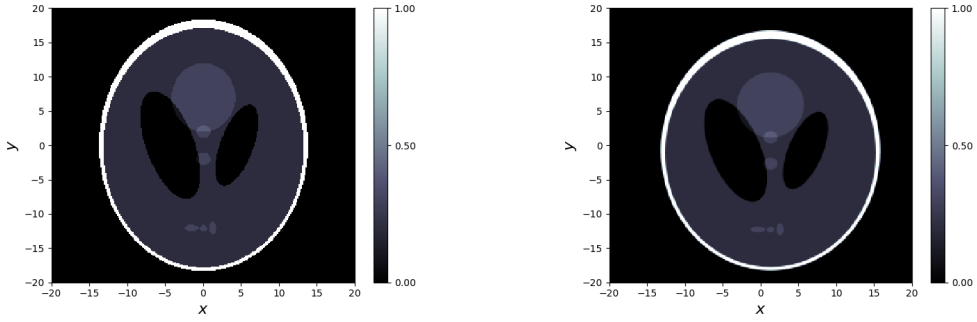
[4] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[5] J.-D. Benamou, G. Carlier, and M. Laborde. An augmented Lagrangian approach to Wasserstein gradient flows and applications. *ESAIM: Proceedings and Surveys*, 54:1–17, 2016.

[6] D. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

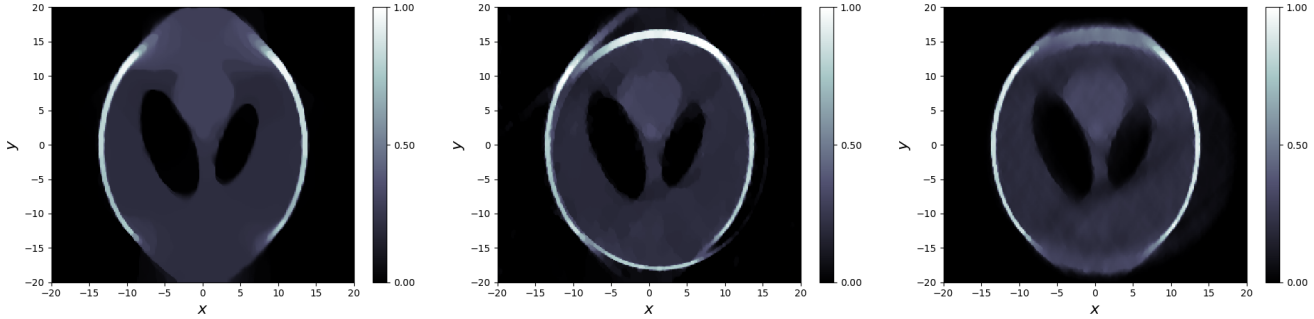
[7] R.I. Boţ and C. Hendrich. A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM Journal on Optimization*, 23(4):2541–2565, 2013.

[1] The example have been implemented and solved using ODL [1], and the ray transform computations are performed by the GPU-accelerated version of ASTRA. [27]. Moreover, all code used is available online, see [23].



(a) Phantom. (b) Prior.

Fig. 1. Figure showing (a) the Shepp-Logan phantom, (b) a deformed Shepp-Logan used as prior. Gray scale values are shown to the right of each image.



(a) TV reconstruction. (b) Reconstruction with ℓ_2 prior. (c) Reconstruction with optimal transport prior.

Fig. 2. Reconstructions using different methods. (a) reconstruction using TV-regularization, (b) reconstruction with ℓ_2^2-prior and TV-regularization, and (c) reconstruction with optimal transport prior and TV-regularization.

[8] R.M. Corless and D.J. Jeffrey. The Wright ω function. In *Artificial intelligence, automated reasoning, and symbolic computation*, pages 76–89. Springer, 2002.

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

[10] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning (ICML-14)*, pages 685–693, 2014.

[11] J. Eckstein. *Splitting methods for monotone operators with applications to parallel optimization*. PhD thesis, 1989. Department of Civil Engineering, Massachusetts Institute of Technology.

[12] J. Eckstein and D.P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[13] F. Elvander, A. Jakobsson, and J. Karlsson. Interpolation and extrapolation of toeplitz matrices via optimal mass transport. *arXiv preprint arXiv:1711.03890*, 2017.

[14] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Kluwer Academic Publisher, 2000.

[15] B. Engquist and B.D. Froese. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5), 2014.

[16] T.T. Georgiou, J. Karlsson, and M.S. Takyar. Metrics for power spectra: an axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3):859–867, 2009.

[17] F. De Goes, D. Cohen-Steiner, P. Alliez, and M. Desbrun. An optimal transport approach to robust reconstruction and simplification of 2d shapes. In *Computer Graphics Forum*, volume 30, pages 1593–1602. Wiley Online Library, 2011.

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

[19] X. Jiang, Z.-Q. Luo, and T.T. Georgiou. Geometric methods for spectral analysis. *IEEE Transactions on Signal Processing*, 60(3):1064–1074, 2012.

[20] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[21] T. Kaijser. Computing the Kantorovich distance for images. *Journal of Mathematical Imaging and Vision*, 9(2):173–191, 1998.

[22] J. Karlsson and T.T. Georgiou. Uncertainty bounds for spectral estimation. *IEEE Transactions on Automatic Control*, 58(7):1659–1673, 2013.

[23] J. Karlsson and A. Ringh. Generalized sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4):1935–1962, 2017.

[24] 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.

[25] F. Natterer. *The mathematics of computerized tomography*, volume 32 of *Classics in Applied Mathematics*. SIAM, 2001.

[26] F. Natterer and F. Wübbeling. *Mathematical methods in image reconstruction*. SIAM, 2001.

[27] W.J. Palenstijn, K.J. Batenburg, and J. Sijbers. Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs). *Journal of structural biology*, 176(2):250–253, 2011.

[28] G. Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[29] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[30] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.