Optimal Transport for Image Processing

Nicolas Papadakis

université de BORDEAUX

HABILITATION À DIRIGER DES RECHERCHES

ÉCOLE DOCTORALE MATHÉMATIQUES
ET INFORMATIQUE

Spécialité: Mathématiques Appliquées

Transport Optimal pour le Traitement d'Images

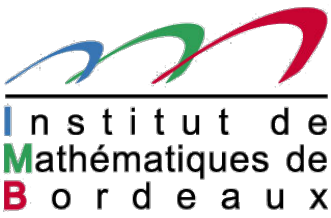
présentée par

Nicolas Papadakis

Soutenue le 17 décembre 2015

devant le jury composé de:

M.	Jean-François Aujol	Professeur de l'Université de Bordeaux	(Examinateur)
M.	Jérémie Bigot	Professeur de l'Université de Bordeaux	(Examinateur)
M.	Antonin Chambolle	Directeur de Recherches CNRS	(Rapporteur)
Me.	Julie Delon	Professeur de l'Université Paris Descartes	(Examinateur)
M.	Angelo Iollo	Professeur de l'Université de Bordeaux	(Examinateur)
Me.	Étienne Mémin	Directeur de Recherche Inria	(Examinateur)
M.	Patrick Pérez	Directeur de Recherche Inria, Technicolor	(Rapporteur)
M.	Christoph Schnörr	Professeur de l'Université d'Heidelberg	(Rapporteur)



Document préparé à
l'Institut de Mathématiques de Bordeaux (UMR 5251)
351 Cours de la Libération
33 405 Talence CEDEX

À Aurélie, Esteban et Anaé

Optimal Transport for Image Processing

Abstract

Optimal Transport is a well developed mathematical theory that defines robust metrics between probability distributions. The computation of optimal displacements between densities through the associated transport map makes this theory mainstream in several applicative fields.
For image processing applications, the transport map can for instance be used to compute geodesics between images or to transfer characteristics of one image to another. In this context, it is of major interest to preserve the nature of the observed objects, in order to synthesize images that are physically and visually plausible. In this document, generalized Optimal Transport distances including relaxation and regularization are considered to improve the modeling of image processing problems. New models and algorithms within the continuous and discrete formulations of Optimal Transport are presented.
With the continuous setting, the integration of physical regularization of the transport plan makes possible the interpolation of ocean images containing complex structures.
In the discrete setting, the regularization of the transport plan is considered for color transfer between images. Convex and non-convex models are proposed to define automatic methods that adapt the proportion of colors required to synthesize visually plausible images. These methods are extended to the computation of barycenters to deal with the color normalization of multiple images.
Finally, the entropy regularization of discrete Optimal Transport is used for image segmentation. A fast and convex model is designed to segment images, while respecting global color distribution constraints.

Keywords : Generalized Wasserstein distance, image processing, image interpolation, color transfer, segmentation, (non)convex optimization.

Résumé

Le Transport Optimal est une théorie mathématique très développée permettant de définir des métriques robustes entre distributions de probabilité. Le calcul du déplacement optimal entre densités par le plan de transport associé rend cette théorie attractive du point de vue applicatif.
En traitement d'images, le plan de transport peut par exemple être utilisé pour calculer des géodésiques entre images ou encore transférer des caractéristiques d'une image vers une autre. Dans ce contexte, il est important de préserver le nature des objets observés afin de synthétiser des images physiquement et visuellement plausibles.
Dans ce document, des distances de Transport Optimal généralisées sont considérées pour améliorer la modélisation de problèmes de traitement d'images. De nouveaux modèles et algorithmes sont présentés pour les formulations continues et discrètes du Transport Optimal.
Dans le cas continu, l'intégration de régularisation physique du plan de transport permet alors d'interpoler des images d'océan contenant des structures complexes.
Dans la formulation discrète, la régularisation du plan de transport est considérée pour le transfert de couleurs entre images. Des modèles convexes et non-convexes sont proposés pour définir des méthodes automatiques adaptant la proportion de couleur nécéssaire pour un transfert visuellement plausible. Ces méthodes sont également étendues aux barycentres afin d'aborder le problème de normalisation de couleurs entre plus de deux images.
La régularisation entropique du Transport Optimal discret est finalement utilisée pour la segmentation d'image. Un modèle convexe et rapide est proposé pour segmenter des images tout en respectant des contraintes globales sur les distributions de couleurs.

Mots-clefs : Distance de Wasserstein généralisée, interpolation d'images, transfert de couleur, segmentation, optimisation (non)convexe.

Contents

1 Introduction **11**

1.1 Motivations 11

1.2 Limitations of Optimal Transport in Imaging 15

1.3 Contributions and Organization of the document 16

I Continuous Optimal Transport **19**

2 Dynamic Optimal Transport **21**

2.1 Estimation of the Transport Map 21

2.2 Fluid Mechanics Formulation 23

2.3 Convexification 23

2.4 Discretization and Optimization 24

2.5 Limitations and Motivations 25

3 Generalizations for Image Interpolation **27**

3.1 Generalized Cost Functions 27

3.1.1 Interpolation between L^2-Wasserstein and H^{-1} 28

3.1.2 Modelling of obstacle 29

3.1.3 Anisotropic transport 30

3.2 Optimal Transport with physical priors 32

3.2.1 Non-convex coupling 32

3.2.2 Synthetic tests 33

3.2.3 Image interpolation in oceanography 35

Conclusion of Part I **37**

II Discrete Optimal Transport 39

4 Relaxation and Regularization of Optimal Transport 41

4.1 Discrete Formulation . . . 41

4.1.1 Monge-Kantorovitch Formulation . . . 42

4.1.2 Dual formulation . . . 43

4.2 Relaxed and Regularized Transport . . . 43

4.2.1 Relaxed Transport . . . 44

4.2.2 Case of cloud of points . . . 45

4.2.3 Mean Optimal Transport Map . . . 45

4.2.4 Discrete Regularized Transport . . . 46

4.2.5 Symmetric Regular Optimal Transport Formulation . . . 48

4.3 Approximate **MK** cost using Sinkhorn distances . . . 49

5 Application to Color Transfer 53

5.1 Convex and regularized Optimal Transport model . . . 54

5.1.1 Adaptive OT . . . 55

5.1.2 Color transfer experiments . . . 57

5.1.3 Extension to Barycenters . . . 60

5.1.4 Limitations . . . 62

5.2 Non-convex relaxation of regularized Optimal Transport . . . 64

5.2.1 Dispersion penalization . . . 64

5.2.2 Experiments . . . 66

5.3 Discussion . . . 67

6 Application to Image Segmentation 69

6.1 Convex histogram-based image segmentation . . . 70

6.2 Optimal Transport between non normalized histograms . . . 71

6.2.1 Wasserstein-ℓ_1 for $1-D$ histograms . . . 71

6.2.2 Approximate **MK** distance for general histograms . . . 72

6.2.3 Experiments . . . 74

6.3 Co-segmentation . . . 78

Conclusion of Part II **81**

General Conclusions and Perspectives **83**

A Non-smooth Optimisation **87**

A.1 Definitions 87

A.1.1 Subdifferential 88

A.1.2 Proximal operator 89

A.1.3 Conjugate function 89

A.2 Convex functionals 90

A.2.1 Primal Proximal Algorithms 90

A.2.2 Primal-Dual Proximal Algorithms 91

A.3 Non-convex functionals 93

A.3.1 Back to the Forward-Backward 93

A.3.2 Non-convex coupling 94

Bibliography **95**

Chapter 1

Introduction

This habilitation manuscript presents a selected subset of the research activities realized in the period 2011-2015. It focuses on the use of Optimal Transport for image processing purposes. These works have been initiated through the supervision of two PhD thesis and one post-doctorate.

In this introduction, the applicative and scientific motivations of the realized works are presented. The main limitations of the use of Optimal Transport for Image Processing are then exposed. The related contributions are finally briefly presented, before detailing the organization of the document.

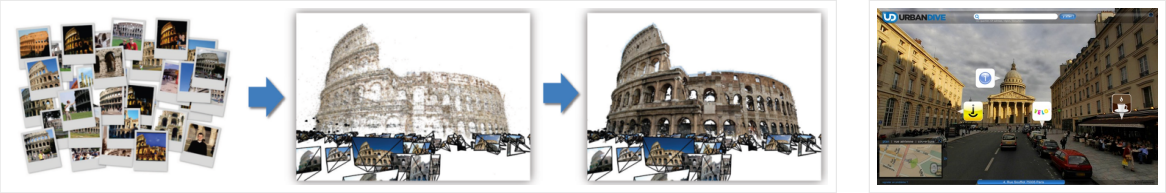
1.1 Motivations

Computer Vision Image and video contents are omnipresent in the world today, both in our personal and professional lives. Technology is in constant evolution and the amount of data is everyday dramatically increasing. Market data indicates that the number of acquisition mobile devices is going to considerably increase. In particular, a spectacular growth in video users for the years to come is expected. In 2016, more than 1.6 billion electronic devices capable of recording and sharing pictures and videos will be used all around the world. Billions of raw images and videos are consequently diffused on the Internet. For instance, approximately 2 millions of pictures are uploaded on Flickr on a daily basis and the volume of videos uploaded to Youtube is gigantic, and constantly increasing (300 hours per minute). However, most of this video content is not exploitable because it has not been properly edited.

Due to the diversity of applications in which images are involved, the type of data is completely heterogeneous: color images, 3D images or even multidimensional images (including transparency, depth, *etc*), animated sequence and videos, multispectral images, depth images, light-field, *etc*. Additionally, pictures are often captured at different resolutions and under different lighting conditions.

The need for generic methods to deal with this huge amount of data is increasing. The development of methods for a given application implies the control of the statistics of the images targeted by this application. An example is presented in Figure 1.1. Raw images diffused on the Internet are used for 3D reconstruction of buildings (using *e.g.*

geolocalized Flickr images) or virtual synthesis of 2D views of cities (Google street-view, Institut National de l'Information Géographique et Forestière, IGN). For these applications, image harmonization is a key step to merge heterogeneous data with different color statistics. In order to circumvent this challenging task, the current methods only select a small subset of available images that are sufficiently close in terms of color and illuminations. As a consequence, most of existing data is not even used, whereas it could lead to significant improvements.



(a) (b)

Figure 1.1 – Example of massive use of images. (a) Reconstruction of the Colosseum from Flickr images [2]. (b) UrbanDive navigation system of IGN.

In order to use more images in this context or simply when one looks for pictures on the Internet to illustrate professional or personal presentations, the huge quantity of images needs to be efficiently indexed and retrieved. This requires very fast and efficient comparison tools between descriptors of image statistics, which are also of interest for object segmentation, editing or harmonization.

Histograms are popularly used in image processing, computer vision and machine learning to represent complex visual objects. To that end, relevant features such as color, contours, orientations or textures are first extracted from images. Histograms are built from the quantization of the space of features into discrete bins. One can further define normalized histograms that describe the frequencies of each of the observed features. The problems of retrieving a similar image in a database or segmenting a particular object can thus be recasted as a problem of histograms comparison. In this context, even if adapted metrics can be learned by neural networks for specific application, generic robust comparison tools are still needed for a broad use.

While a lot of research effort is dedicated to the enhancement of image descriptors (by adding additional information such as text or context) in a "big data" perspective, the common metrics used for statistics comparison are still not developed enough to deal with various perturbation effects (such as histogram quantization, shifts, deformation, *etc*) or large amount of data outliers. Namely, spatial information is not efficiently considered as it is mainly used to enhance the feature description [196]. Including spatial information into the distance defined on the feature space is an interesting alternative to improve the quality of the results.

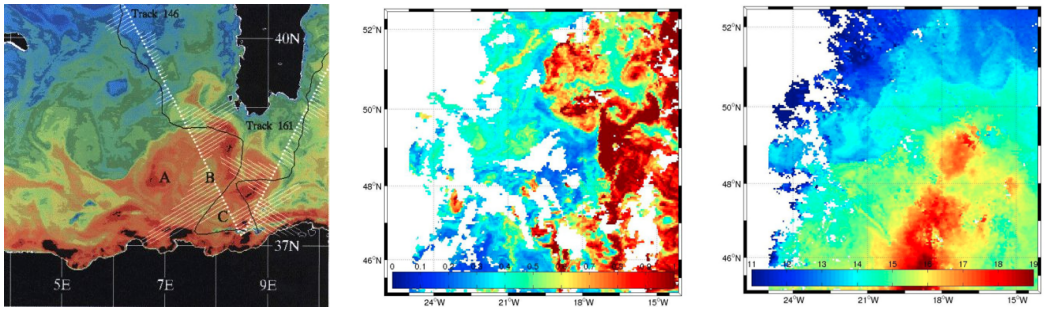
The Optimal Transport framework is the proper way to compare statistic distributions. In contrast to most distances from information theory, it takes into account the spatial location of the density modes [216]. By incorporating in its definition a ground metric between the features themselves, the Optimal Transport distance (that corresponds to the Wasserstein distance when the ground metric is the quadratic Euclidean distance) can compare sparse histograms even if their supports do not overlap signif-

icantly. Additionally, Optimal Transport provides a warping (the so-called transport plan) between histograms which can be used to perform image editing such as color transfer.

Optimal Transport framework has been shown to produce robust state of the art results for the comparison of statistical descriptors through the so-called Earth Mover's Distance [187]. Various works on color image retrieval [162] or artistic image indexation [121] have shown on small datasets (*i.e.* thousands of images) that Optimal Transport framework is intrinsically designed to carry this spatial information. However, unless considering some crude approximations and/or simple ground metrics, it involves prohibitive computational costs. Though Optimal Transport would lead to a significant gap in performance improvement, this computational limitation is clearly an obstacle to its study and large scale deployment, when considering the dimension of image databases or the data itself (multi-spectral images, patch representation, tensor-field imaging, *etc*).

Geosciences Since the late seventies, many satellites have been launched to improve our knowledge of the atmosphere and the oceans. Geostationary satellites provide, among other data, photographic images of the earth system. Sequences of such images show the dynamical evolution of identified meteorological or oceanic "objects": fronts, clouds, eddies, vortices, etc. The human vision can easily detect the dynamics in this kind of image sequences and it clearly has a strong predictive potential. This aspect is favoured by the fact that image data, contrary to many other measurements, is dense in space and time. Indeed the spatial resolution of current METEOSAT satellites is close to one kilometer and they produce an image every 15 minutes. This frequency will be improved up to one every 10 minutes (and even every 2.5 minutes for Europe only) with the upcoming third satellite generation. It implies a huge quantity of information which can be seen as an asset but also induces difficulties for the assimilation system that has to cope with such amount of data.

Satellite data is currently used in operational systems for calibrating Numerical Weather Prediction models, mainly through the assimilation of the radiance measured by the satellite at each pixel of the image. As illustrated in Figure 1.2, satellite images are related to physical quantities such as surface temperature, sea surface height, cloud pressure, chlorophyll concentration, etc.



(a) (b) (c)

Figure 1.2 – Different examples of satellite images used for data assimilation. (a) Altimetric reconstruction from JASON satellite data. (b) Ocean color/ Chlorophyll from the MODIS captor of ENVISAT satellite. (c) Sea surface Temperature from the MODIS captor of ENVISAT satellite.

In practice only a tiny percentage (about $3-5\%$) of total satellite (from polar orbiting and geostationary) data is used in operational systems, while being given low confidence with respect to synoptic data, collected by stations, aircrafts, radiosounding, ballons or drifters. Considering the cost of satellite observing systems (the cost of the launch of the Meteosat Third Generation is estimated to about 2.5 billion Euros) and of the infrastructures required for the collection of the data itself, improving their impact on forecasting systems is an important topic of research.

As a consequence, there is a need to provide pertinent distances to interpolate between the observed satellite images and the images of physical variables provided by the numerical models. The spatial information contained in satellite images is currently not taken into account, as the classical "pixel-to pixel" L^2 distance used in data assimilation is not adapted to deal with bad localization of structures such as clouds, vortex or fronts [209], [52]. If the same structure is contained in both images at different locations, an interpolation should deal with the spatial shift. While a L^2 interpolation would create two structures, Optimal Transport is intrinsically designed to correctly interpolate the position of the structure along the geodesic in the Wasserstein space.

Considering images as densities, interpolation between images can be realized with the Optimal Transport map [112]. It is thus an interesting framework in order to propose new metrics for data assimilation [159]. In this context, an important challenge is to model the topology of the considered domain (coasts or islands in oceanography) for estimating image interpolations that have a real physical meaning. In this context, even if Optical Flow methods can be considered for data assimilation to compute discrepancies between atmospheric images [68], they can not deal with complex domains in oceanography [27].

Optimal Transport Optimal Transport is a well developed mathematical theory that defines a family of metrics between probability distributions [105, 216]. These metrics measure the amplitude of an optimal displacement according to a so-called ground cost defined on the space supporting the distributions. The resulting distance is referred to as the Wasserstein distance in the case of L^p ground costs $C(x,y) = ||x-y||^p$. As shown in Figure 1.3, it measures the minimal effort required for filling a "remblai" $(-\mu_1)$ with a "déblai" (μ_0), *i.e.* transporting one distribution to another. The geometric nature of Optimal Transport, as well as the ability to compute optimal displacements between densities through the corresponding transport map T, make this theory progressively mainstream in several applicative fields.

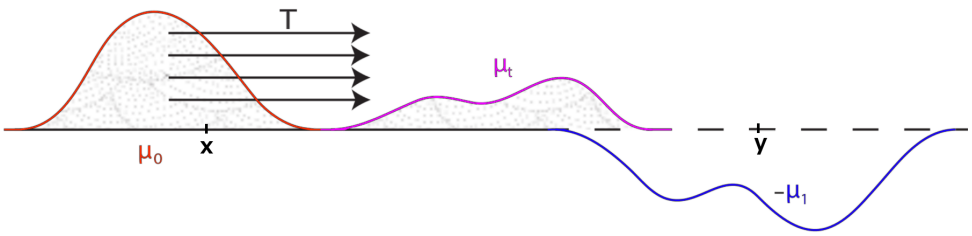


Figure 1.3 – Illustration of Optimal Transport between distributions μ_0 and μ_1 as introduced in [145] with the "déblais" and "remblais" problem . Interpolations $(\mu_t)_{t\in[0;1]}$ between μ_0 and μ_1 can be compted with the transport map T.

Early successes of applications of Optimal Transport were mostly theoretical, such as the study of shape recognition [106], the derivation of functional inequalities [67], the construction of solutions of non-linear partial differential equations [124] or the study of gradient flows in Wasserstein spaces [5].

In computer vision, the Wasserstein distance has been shown to outperform other metrics between distributions for machine learning tasks [187, 163, 75] or image segmentation [151, 165, 206].

In image processing, the warping provided by the Optimal Transport has been used for video restoration [82], color transfer [168], texture synthesis [95], optical nanoscopy [45] and medical imaging registration [112]. It has also been applied to interpolation in computer graphics [35, 199] and surface reconstruction in computational geometry [85]. Optimal Transport is either used to model various physical phenomena, such as for instance in astrophysics [101] and oceanography [17, 97].

1.2 Limitations of Optimal Transport in Imaging

The numerical resolution of the Optimal Transport problem raises several challenges. This is the main reason why it has been poorly used in image processing until recently. Computing Optimal Transport distance is only an easy task when dealing with one-dimensional densities, such as histograms of grayscale images, or multi-dimensional densities roughly discretized. However, when considering transport between images themselves or when dealing with statistics of color images, the densities are defined on higher dimensional spaces. Hence, the versatility and high quality of Optimal Transport distances come at a price for general image applications: they require the resolution of high dimensional problems that scale with the product of the dimensions of the discretized densities.

In this context, the fluid dynamic formulation of the Optimal Transport problem introduced in [19] is an interesting approach for dealing with higher dimensions. The entropic regularization proposed in [73] has also offered new perspectives as faster algorithms based on Sinkhorn distances can be designed for computing approximate Optimal Transport in larger dimensions. This last approach gives interesting approximations of the transport distance but leads to transport maps that may be too far from the optimal ones for transfer and interpolation purposes.

Hence, generalizations of the Optimal Transport distances are still necessary to adapt to applications involving image processing. For instance, when the transport plan is used to interpolate between densities that have very different "shapes" or contain data outliers, partial models able to relax the constraint of transporting the whole mass are required. In this context, another flaw for the use of Optimal Transport plan is that it is usually highly irregular. For the interpolation between images or histograms of features, dedicated regularizations of the transport plan are needed to preserve either the nature of the objects contained in the images or the modes of the statistics.

Such property is illustrated in Figure 1.4 with an example of color transfer between images. If the exact prescription of color is realized with the Optimal Transport map, visual artifacts may appear.

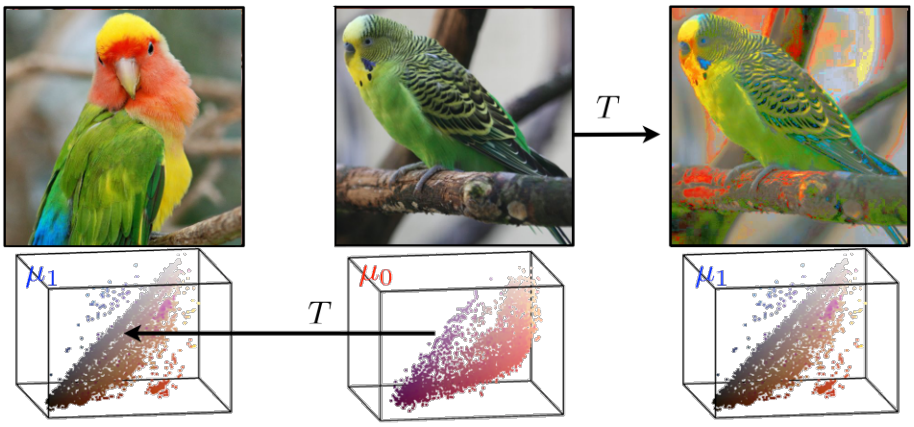


Figure 1.4 – Illustration of the limitation of Optimal Transport for color transfer. The Optimal Transport map T between the color distributions of the middle parrot μ_0 and the left parrot μ_1 is computed. The map is then used to modify the colors of the middle parrot in order to obtain the right one. As the transport map is highly irregular, very different colors are assigned to pixels of the background that were originally close both in spatial and color spaces.

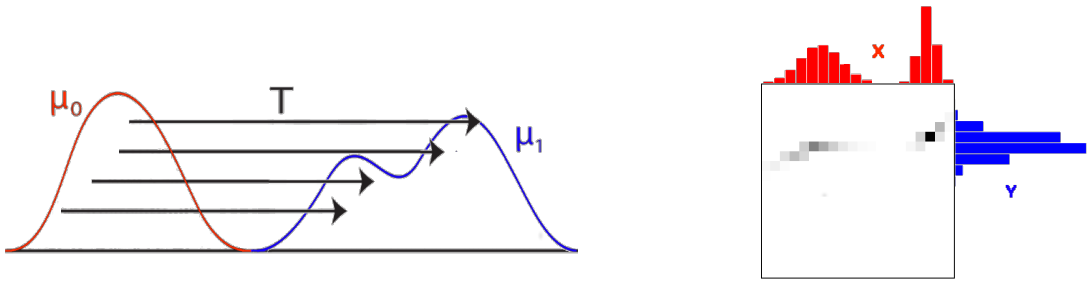
Finally, introducing Optimal Transport distances in more general image problems such as segmentation, involving histogram data fitting and spatial regularization in the image domain, is even more challenging. Spatial regularization can be incorporated through graph or variational modeling. However, graph based approaches are limited to simple bin-to-bin metrics between histograms [185] and cannot deal with Optimal Transport distances without leading to problems of gigantic size [200]. Slow hybrid methods that alternate optimization on graph and Optimal Transport computation should be designed. Variational methods may thus be more adapted to such problems, since the computation of Optimal Transport distances can be integrated into a general model that can directly be optimized.

1.3 Contributions and Organization of the document

In this document, some new variational image processing models using Optimal Transport distances are exposed. The main common contribution of the presented works is to propose original modeling of classical problems such as image interpolation, color transfer or image segmentation, while considering recent optimization methods able to solve these problems with "reasonable" computational cost (several seconds to several minutes).

For that purpose, extended Optimal Transport models including relaxation of the mass constraint and regularization of the transport map are introduced. These generalized Optimal Transport models have been designed in both continuous and discrete settings. The continuous formulation is applied to image interpolation, while the discrete one is used for color transfer and image segmentation.

The problems associated to both formulations of Optimal Transport are illustrated in Figure 1.5. In the continuous framework, the images are considered as continuous densities μ_0 and μ_1 defined on a domain Ω. The transport map is in this case a vector field, defined from Ω to Ω, that transfers μ_0 onto μ_1. The problem can be expressed in a dynamic way through a fluid mechanics formulation. In the discrete setting, normalized histograms of features extracted from images are considered. When computing the static transport between histograms X and Y discretized with M and N bins, an acceptable transport map is a $M \times N$ matrix describing a joint distribution P which marginals are the given histograms (*i.e.* $\sum_j P_{i,j} = X_i$ and $\sum_i P_{i,j} = Y_j$). Hence $P_{i,j}$ is the mass transferred from X_i to Y_j.



(a) Continuous transport velocity field $T : \Omega \to \Omega$

(b) Discrete transport matrix P coupling histograms X and Y

Figure 1.5 – Illustration of the transport map in the (a) continuous and (b) discrete formulations of the Optimal Transport problem. In (b), white color represents $P_{i,j} = 0$.

Overview of the manuscript The first part of this document is dedicated to some contributions in the continuous setting. Chapter 2 is based on [160] and presents an approach for the numerical estimation of the geodesic path between two densities according to the L^2 Wasserstein metric. Generalizations including additional physical constraints proposed during the PhD of R. Hug in [120] for application to the interpolation of ocean images are next presented in Chapter 3.

The second part focuses on contributions in the discrete setting. Chapter 4 presents the relaxed and regularized formalism proposed in [93] as well as entropic regularization recently introduced in the literature [73]. Chapters 5 and 6 are respectively dedicated to the application of these tools to color transfer between images and image segmentation. Chapter 5 gathers methods initiated during the Postdoctorate of S. Ferrandans [94, 178], while Chapter 6 reviews some models introduced during the PhD of R. Yıldızoğlu [221, 222, 176].

In order to make the manuscript self-content, standard first-order proximal algorithms dedicated to the optimization of non-smooth functionals used all along the document are finally presented in Appendix A.

Part I

Continuous Optimal Transport

Chapter 2

Dynamic Optimal Transport

This chapter is focused on the computation of geodesics for the Optimal Transport metric associated to the L^2 ground cost. It reviews various methods and extends the approach pioneered in [19] from the perspective of proximal operator splitting in convex optimization [160]. It shows the simplicity and efficiency of this method, which can easily be extended beyond the setting of Optimal Transport by considering various convex cost functions that will be presented in Chapter 3.

2.1 Estimation of the Transport Map

Let Ω be a convex and bounded domain of $\mathbb{R}^d$ and ρ_0, ρ_1 be two non-negative L^1 functions on Ω, of equal mass. We will assume without loss of generality that

$$\int_\Omega \rho_0(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_\Omega \rho_1(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1. \tag{2.1}$$

The original formulation of the Optimal Transport problem in [145] corresponds to minimizing the cost for transporting a ρ_0 onto ρ_1 using a map T. In the following, we restrict our exposition to maps $T : \Omega \subset \mathbb{R}^d \mapsto \Omega$. A valid transport map $T : \Omega \to \Omega$ then pushes forward the measure $\rho_0(\mathbf{x})\mathrm{d}\mathbf{x}$ onto $\rho_1(\mathbf{x})\,\mathrm{d}\mathbf{x}$, that is $T\#\rho_0 = \rho_1$ or for all bounded set $A \subset \Omega$:

$$\int_A \rho_1(\mathbf{x})d\mathbf{x} = \int_{T(\mathbf{x})\in A} \rho_0(\mathbf{x})d\mathbf{x}. \tag{2.2}$$

In term of densities, when T is smooth and one-to-one, the constraint $T\#\rho_0 = \rho_1$ corresponds to

$$\rho_0(\mathbf{x}) = \rho_1(T(\mathbf{x}))\,|\det(\nabla T(\mathbf{x}))| \tag{2.3}$$

where $\nabla T(\mathbf{x}) \in \mathbb{R}^{d\times d}$ is the differential of T at $\mathbf{x}$. This is known as the Jacobian equation. We call $\mathcal{T}(\rho_0, \rho_1)$ the set of transport maps that satisfy the constraint (2.3). The L^p Kantorovitch-Wassertein distance between ρ_0 and ρ_1 is then defined by

$$W_p(\rho_0, \rho_1)^p = \inf_{T\in\mathcal{T}(\rho_0,\rho_1)} \int |T(\mathbf{x}) - \mathbf{x}|^p \rho_0(\mathbf{x})d\mathbf{x}.$$

The L^p Monge-Kantorovitch problem corresponds to finding a mapping T such that this infimum is achieved. It can be recasted as

$$W_p(\rho_0, \rho_1)^p = \min_{T \in \mathcal{T}(\rho_0, \rho_1)} \int C(\mathbf{x}, T(\mathbf{x}))\rho_0(\mathbf{x})\, \mathrm{d}\mathbf{x} \tag{2.4}$$

where $C(\mathbf{x}_0, \mathbf{x}_1) \geq 0$ is the ground cost of transporting $\mathbf{x}_0 \in \Omega$ onto $\mathbf{x}_1 \in \Omega$.

Let us now detail how estimating the transport map T with Partial Differential Equations.

Optimal Transport and PDEs The Optimal Transport for the L^2 ground cost has a special structure. It can be shown to be uniquely defined (see *e.g.* [216] page 66) and the transport map is the gradient of a convex potential Ψ from Ω to $\mathbb{R}$:

$$T(\mathbf{x}) = \nabla\Psi(\mathbf{x}). \tag{2.5}$$

Hence, one can show that the mass transfer associated to a L^2 ground cost follows straight lines [41], which can be used for developing specific Lagrangian solvers [122].

Other class of methods relies on the fact that from relations (2.3) and (2.5), the convex function Ψ is solution of the Monge-Ampère equation:

$$\det(D^2\Psi)\rho_1(\nabla\Psi(\mathbf{x})) = \rho_0(\mathbf{x}).$$

This equation being highly nonlinear, numerical methods to solve the Monge-Kantorovitch problem based on discretization of the Monge-Ampère equation have been investigated [157, 135, 79, 154, 23, 24]. A major difficulty in these approaches is to deal with compactly supported densities, which requires a careful handing of the boundary conditions [102].

Another line of methods iteratively constructs mass preserving mappings converging to the Optimal Transport [6, 111, 38]. This explicitly constructs the so-called polar factorization of the initial map, see also [18] for a different approach. These PDE's based approaches to the resolution of the Optimal Transport have found several applications, such as image registration [112], density regularization [46], optical flow [63] and grid generation [204].

A last axis of research consists in using gradient flows where the gradient direction is computed according to the Wasserstein distance. This was initially proposed in [124] to build solutions to certain non-linear PDE's. This technique is now being used to design numerical approximation schemes for the solution of these equations, see for instance [51, 89, 96, 22].

All these approaches are nevertheless limited to positive densities, or to non-negative densities with a convex support of the target density ρ_1, which is a real limitation for image processing applications. The **only numerical method that can deal with non-negative data** has been proposed by Y. Brenier et J.D. Benamou through a Fluid Mechanics formulation [19].

2.2 Fluid Mechanics Formulation

Instead of computing directly the transport, it is possible to consider the geodesic path between the two densities according to the Wasserstein metric (the so-called displacement interpolation [142]). For the L^2 ground cost, the geodesic path between the measures with densities ρ_0 and ρ_1 can be shown to have density $t \mapsto \rho(t, \mathbf{x})$ where the additional dimension $t \in [0, 1]$ parameterizes the path interpolating linearly between the identity map Id_d and the Optimal Transport map $\mathcal{T}$:

$$\rho_0(\mathbf{x}) = \rho(t, T_t(\mathbf{x}))\,|\det(\nabla T_t(\mathbf{x}))| \text{ where } T_t = (1-t)\,\mathrm{Id}_d + tT.$$

The geodesic can thus be computed by first obtaining the transport and then making the densities evolve. In [19], it has been demonstrated that this geodesic solves the following non-convex problem over the densities $\rho(t, \mathbf{x}) \in \mathbb{R}_+$ and a velocity field $v(t, \mathbf{x}) \in \mathbb{R}^d$ checking the continuity equation.

Theorem 1 (Benamou-Brenier)**.** *In case $p = 2$ the Wasserstein distance between ρ_0 and ρ_1 is such that:*

$$W(\rho_0, \rho_1) = \inf_{\rho, v} \int_\Omega \int_0^1 \rho(t, \mathbf{x}) |v(t, \mathbf{x})|^2 d\mathbf{x}dt \tag{2.6}$$

the infimum being taken on ρ, v verifying the non-linear constraint

$$C^0 = \{(\rho, v),\ \partial_t \rho + \mathrm{div}_{\mathbf{x}}(\rho v) = 0,\ \langle v, \vec{n}\rangle = 0,\ \rho(0, \cdot) = \rho_0,\ \rho(1, \cdot) = \rho_1\}, \tag{2.7}$$

with homogeneous Neumann conditions on the velocity field v, through $\vec{n}$ which is the normal of the domain Ω.

While the estimation of the L^2 Wasserstein distance (2.4) is a static problem over Ω, the introduction of the additional dimension t and the velocity field v leads to a dynamic problem involving kinetic energy (2.6) and continuity equation (2.7) on $\Omega \times [0; 1]$.

2.3 Convexification

Following [19] and introducing the change of variable $(\rho, v) \mapsto (\rho, m)$, where m is the momentum $m = \rho v$, we obtain a convex optimization problem over the couple (ρ, m):

$$\min_{(\rho,m)\in C} \mathcal{J}(\rho, m) := \int_0^1 \int_\Omega J(\rho, m) d\mathbf{x}dt \tag{2.8}$$

which deals the positivity constraint [49]:

$$J(\rho, m) = \begin{cases} \frac{||m||^2}{2\rho} & if\ \rho > 0 \\ 0 & if\ (\rho, m) = (0, 0) \\ +\infty & \text{otherwise.} \end{cases} \tag{2.9}$$

The set of constraint C becomes linear and reads:

$$C(\rho, m) = \{(\rho, m),\ \partial_t \rho + \mathrm{div}_{\mathbf{x}}\, m = 0,\ \langle m, \vec{n}\rangle = 0,\ \rho(0, \cdot) = \rho_0,\ \rho(1, \cdot) = \rho_1\}. \tag{2.10}$$

Denoting as $\iota_C(x) = 0$ if $x \in C$, $+\infty$ otherwise, the convex problem introduced in [19] reads:

$$(\rho^*, m^*) = \underset{\rho,m}{\operatorname{argmin}}\ \mathcal{J}(\rho, m) + \iota_C(\rho, m). \tag{2.11}$$

Notice that with this formulation, we are not able to compute the transport map and we only estimate the geodesic $\rho(x, t)$. The existence of minimizers of such a problem in the space of measures has been studied in [217]. The existence of L^2 minimizers for positive data (ρ_0, ρ_1) has been studied in [110]. The existence of L^2 minimizers for nonnegative data (ρ_0, ρ_1) has been shown during the PhD of Romain Hug [119], in which the uniqueness of the variable ρ has also been established.

As the problem (2.11) is convex and non-smooth, the Alternating Direction Method of Multipliers (ADMM, see (A.9)) algorithm was considered in [19] to compute a global minimizer. Nevertheless, other proximal method that are introduced in Appendix A can be used to compute a global minimizer.

2.4 Discretization and Optimization

In order to solve problem (2.11) over the sptaial domain of the image denoted as Ω, the dimension has to be increased with T time steps of artificial time t. The dimension of the variable ρ to estimate is therefore of size $|\Omega|T$.

In [160], it has been proposed to rely on a staggered grid for the discretization of the continuity equation (2.10), similarly to the discretization of PDE's in incompressible fluid dynamics (see for instance [113]). An interpolation operator $\mathcal{I}$ has then been considered for computing the cost function (2.9) on the regular grid. The use of a staggered grid is very natural in the context of the discretization of a divergence operator associated to a vector field on $\mathbb{R}^d$.

The basic idea is to perform an accurate evaluation of every partial derivative at prescribed nodes of a cartesian grid using standard centered finite differences. Hence, the discrete problem (2.11) can be reformulated as:

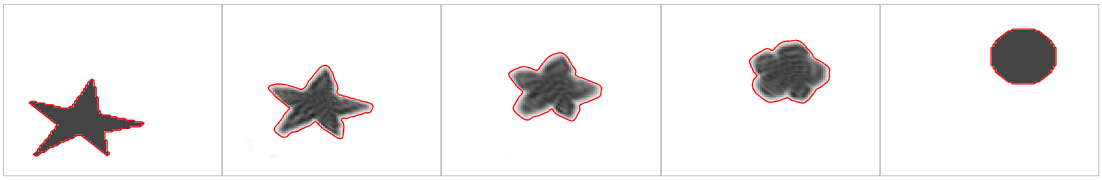
$$(\rho^*, m^*) = \underset{(\rho,m)\in\mathcal{E}_s}{\operatorname{argmin}}\ \mathcal{J}(\mathcal{I}(\rho, m)) + \iota_C(\rho, m), \tag{2.12}$$

The problem (2.12) consists in minimizing the sum of two non-smooth convex functionals, the first one being composed with a linear operator $\mathcal{I}$. We can thus make use of proximal splitting methods such as Douglas Rachford (A.6) or Primal Dual (A.15) algorithms to solve this problem. As illustrated in the Figure 2.2, faster convergence was observed considering these algorithms instead of the ADMM one. To apply such algorithms that are described in the Appendix, the fundamental point is to be able to compute the proximal operators (defined in relation A.2) of functions $\mathcal{J}$ and ι_C.

Since the functional $\mathcal{J}$ is separable, the proximal operator of $\mathcal{J}$ can be computed independently as the proximal operator of J for each point $(\tilde{\rho}, \tilde{m})(t, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^d$ with $(t, \mathbf{x}) \in [0; 1] \times \Omega$. As shown in [160], it simply requires to solve a third order polynomial in each grid point.

Finally, the proximal mapping of ι_C is Proj_C the orthogonal projector on the convex set C. It requires solving a Poisson equation on the centered grid with prescribed boundary conditions. It can be achieved with Fast Fourier Transform in $O(NP\log(NP))$ operations where N and P are number of spatial and temporal points, see [201].

Using this method, one can estimate the geodesic path between densities that vanishes on the spatial domain, as illustrated in Figure 2.1.

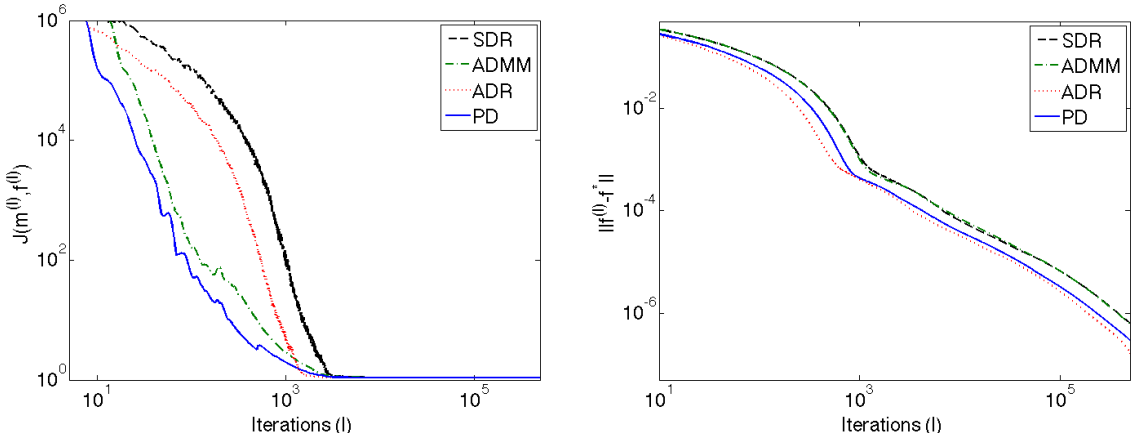


$t = 0$ $t = 1/4$ $t = 1/2$ $t = 3/4$ $t = 1$

Figure 2.1 – Transport between characteristic functions. Evolution of $\rho^\star(\cdot, t)$ for several values of t. The red curve denotes the boundary of the area with positive density.

2.5 Limitations and Motivations

A first limitation of this approach concerns the computational time. As exhibited in Figure 2.2, as the problem is not strictly convex, the convergence of the cost function $\mathcal{J}$ is fast but the convergence of the iterates (ρ^k, m^k) is much slower.



(a) $\mathcal{J}(\rho^k, m^k)$ (b) $||\rho^k - \rho^*||^2$

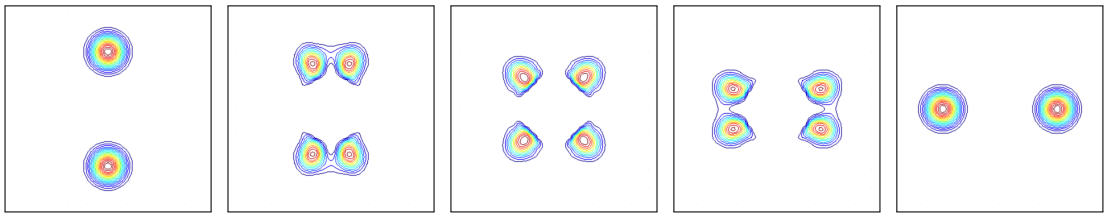
Figure 2.2 – Illustration of the convergence of the minimization process. (a) The cost function J reaches its minimum with a thousand iterations. (b) The convergence of the iterates ρ^k to ρ^* is slower. The reference solution ρ^* has been previously obtained with 10^8 iterations. Different proximal algorithms (ADMM, symmetric and asymmetric Douglas Rachford, Primal Dual) tunned with their best parameters are compared.

Numerical instabilities may also happen when input data are not smooth enough, as illustrated by Figure 2.3. This is nevertheless not a issue for the targeted application where data are images of the ocean representing continuous fluids.

ρ_0 $\rho(1/6)$ $\rho(1/3)$ $\rho(1/2)$ $\rho(2/3)$ $\rho(5/6)$ ρ_1

Figure 2.3 – Image interpolation between Gaspard Monge (ρ_0) and Leonid Kantorovitch (ρ_1). Numerical instabilities appear (mainly on the boundaries of the head) since the data are not smooth enough.

Next, when solving the Optimal Transport problem between two densities, we know that the mass transfer will follow straight lines. The structures contained in the original data can thus disappear along the optimal path. Such behavior is illustrated in Figure 2.4 that presents the optimal path between images containing two Gaussians.



$\rho(0) = \rho_0$ $\rho(1/4)$ $\rho(1/2)$ $\rho(3/4)$ $\rho(1) = \rho_1$

Figure 2.4 – Two Gaussians examples. The Gaussians, represnted by their level lines, split along the optimal path between ρ_0 and ρ_1.

One main objective for image interpolation is to incorporate some physical priors in order to preserve the structures contained in the data along the optimal path. This will be of main interest for geoscience imaging applications. Ocean and atmosphere, observed with satellites, are indeed driven by complex physical laws. The temporal interpolation of such satellite images is an important problem in this community and it should correspond to the underlying dynamics. As interpolation with Optimal Transport involves constant mass transport along straight lines, it is not a satisfactory solution.

In order to be able to process oceanographic images that contains "obstacles" such that continent or islands, the topology of the considered spatial model should be carefully taken into account.

Some generalizations are proposed in the next chapter to tackle such issues.

Chapter 3

Generalizations for Image Interpolation

In this Chapter, we consider generalized distances based on convex ground costs within the dynamical Optimal Transport formulation. In the aim of interpolating satellite images of the ocean in the presence of coasts, the computation of Optimal Transport on Riemannian manifolds modeling the spatial domain is tackled. Physical regularizations of the transport map are also considered. The numerical scheme proposed in the Chapter 2 can be used with only slight modifications with respect to the L^2-Wasserstein case.

3.1 Generalized Cost Functions

The formulation of the geodesic computation as a convex optimization problem presented in Chapter 2 enables the definition of various metrics obtained by changing the objective function. Transport with congested dynamics [48] or unbalanced transport between densities that have different mass [136, 60] are namely possible. One can also interpolate between different distances. An interpolation between the L^2-Wasserstein and L^2 distances is proposed in [20]. Lastly, an interpolation between L^2-Wasserstein and H^{-1} distances is described in [86]. This extension relies in a crucial manner on the convexity of the extended objective function, which enables a theoretical analysis to characterize minimizing geodesics [49].

Optimal Transport on Riemannian manifolds Many properties of the L^2-Wasserstein distance can be extended to the setting where the ground cost is the square of the geodesic distance on a Riemannian manifold. This includes in particular the existence and uniqueness of the transport map [141]. Displacement interpolation for transport on manifolds has the same variational characterization as the one introduced in [20] for Euclidean transport, see [217] for a detailed review of Optimal Transport on manifolds.

Displacement interpolation between two measures, each one composed of a single Dirac, amounts to computing a single geodesic curve on the manifold. Discretization and numerical solutions to this problem are numerous. A popular method is the Fast Marching algorithm introduced jointly by [195, 211] for isotropic Riemannian metrics

(*i.e.* when the metric at each point is a scalar multiple of the identity) discretized on a rectangular grid. This algorithm has been extended to compute geodesics on 2-D triangular meshes with only acute angles [128]. More general discretizations require the use of slower iterative schemes, see for instance [36].

Interpolation between pairs of measures in Riemannian manifolds generalizes to barycenters of a family of measures, see [127]. The numerical computation of Optimal Transport on manifolds has been less studied. For weighted sums of Diracs, displacement interpolation is achieved by solving a linear program to compute the coupling between the Diracs and then advancing the Diracs with the corresponding weights and constant velocity along the geodesics.

Hence, we have proposed in [160] to extend the method of [20] to solve for the displacement interpolation on a Riemannian manifold. Following [86, 49], we use a generalized cost function that allows one to compute geodesics that interpolate between the L^2-Wasserstein and the H^{-1} geodesics . To introduce further flexibility, we have also considered a spatially varying tensor [120], which corresponds to approximating a transportation problem on a Riemannian manifold.

To that end, the convex cost functional (2.9) is generalized, for $\beta \in [0;1]$ and $A(t,\mathbf{x})$ a symmetric positive definite tensor, as

$$J_\beta^A(\rho,m) = \begin{cases} \frac{\|Am\|^2}{2\rho^\beta} & \text{if } f>0, \\ 0 & \text{if } (m,\rho)=(0,0), \\ +\infty & \text{otherwise.} \end{cases} \tag{3.1}$$

and we now consider the problem:

$$(\rho^*,m^*) = \operatorname*{argmin}_{\rho,m} \int_0^1 \int_\Omega J_\beta^A(\rho,m)\,\mathrm{d}\mathbf{x}\,\mathrm{d}\,t + \iota_C(\rho,m). \tag{3.2}$$

The matrix $A(t,\mathbf{x})$ is of size $d \times d$ and represents the anisotropic penalization of the displacement energy. The isotropic case of constant weights $A = \mathrm{Id}_d$ is studied in [86, 49]. The case $\beta = 1$ corresponds to the Wasserstein L^2 distance. In a continuous (not discretized) domain, the value of the problem (3.2) for $\beta = 0$ is equal to the H^{-1} Sobolev norm over densities $\|\rho_0 - \rho_1\|_{H^{-1}}$, as detailed in [86]. In this case, the induced distance is an Hilbertian norm, and the corresponding geodesic is a linear interpolation of the input measures. Thus, for measures having densities, one obtains $\rho_t = (1-t)\rho_0 + t\rho_1$.

3.1.1 Interpolation between L^2-Wasserstein and H^{-1}

With this previous formulation (3.1), it is possible to interpolate between L^2-Wasserstein and H^{-1} geodesics. This is illustrated in Figure 3.1, which presents the level-lines of the estimated densities $\rho(t,\cdot)$ for $A = \mathrm{Id}_d$ and different values of β. It shows the evolution of the solution between a linear interpolation of the densities ($\beta = 0$) and a displacement interpolation with transport ($\beta = 1$).

With this β parameter, smoother interpolation are then obtained. As illustrated in Figure 3.2, it is of interest for interpolating between non-smooth images and avoid numerical instabilities.

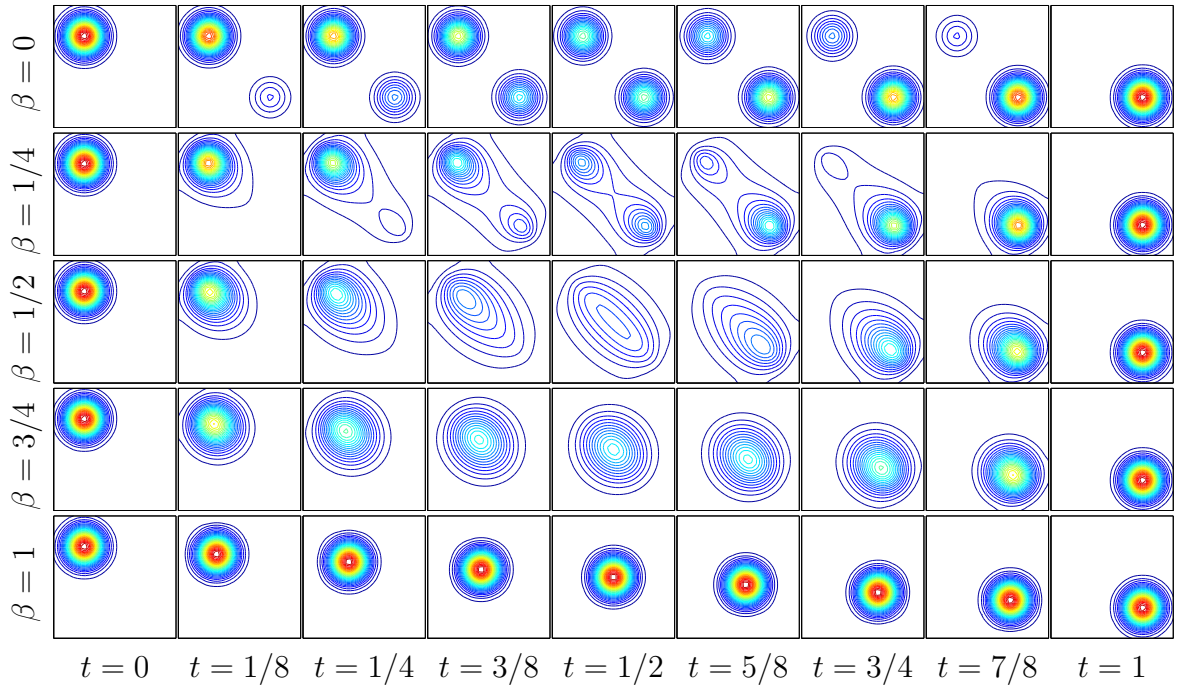


Figure 3.1 – Display of the level sets of $\rho(t,\cdot)$ for several value of t and β. For $t = 0$ and $t = 1$, $\rho(t,\cdot)$ exactly corresponds to ρ_0 and ρ_1.

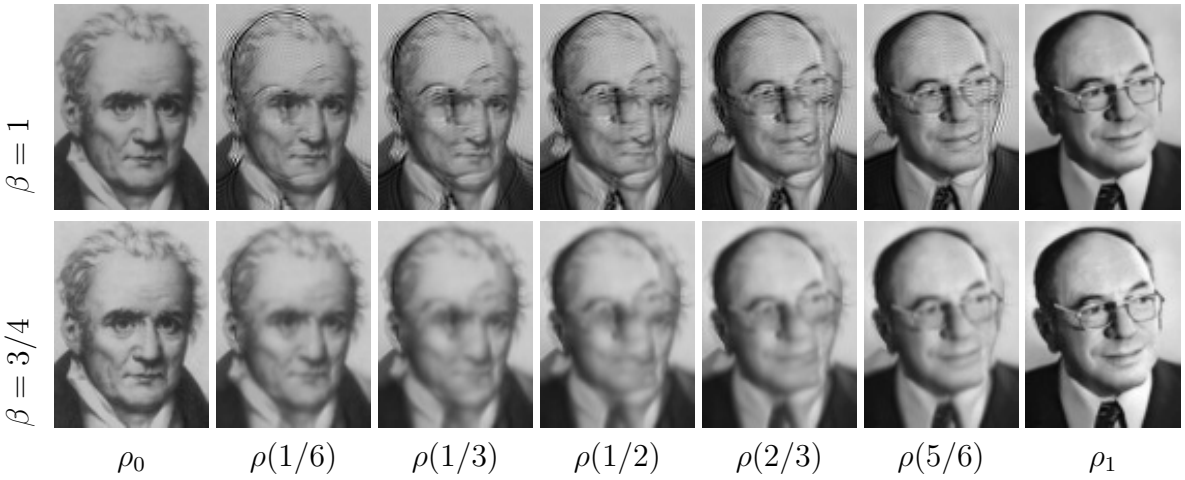


Figure 3.2 – Interpolation between Gaspard Monge (ρ_0) and Leonid Kantorovitch (ρ_1). The first line is the same as in Figure 2.3. On the second line, by taking $\beta = 0.75$, a smoother interpolation between images is estimated.

3.1.2 Modelling of obstacle

When $\beta = 1$ and the matrices $A(t,\mathbf{x}) = \mathrm{Id}_d\, w(\mathbf{x})$ are diagonal and constant in time, the solution of (2.12) discretizes the displacement interpolation between the densities (ρ_0, ρ_1) for a ground cost being the squared geodesic distance on a Riemannian manifold.

We exemplify this setting by considering Optimal Transport with obstacles, which corresponds to choosing weights w that are infinity on the obstacle $\mathcal{O} \subset \mathbb{R} \times \mathbb{R}^d$, *i.e.*

$$w(t,\mathbf{x}) = 1 + \iota_{\mathcal{O}}(t,\mathbf{x}) \in \{1, +\infty\}.$$

Note that with such definition, the obstacles can be dynamic, *i.e.* the weight w does not need to be constant in time. Figure 3.3 shows a first example where $\mathcal{O}$ is a 2-D ($d = 2$) static labyrinth map (the walls of the labyrinth being the obstacles are displayed in black). We use a $50 \times 50 \times 100$ discretization grid of the space-time domain $[0, 1]^3$ and the input measures (ρ_0, ρ_1) are Gaussians with standard deviations equal to 0.04. For Gaussians with such a small variance, this example shows that the displacement interpolation is located closely to the geodesic path between the centers of the gaussians.

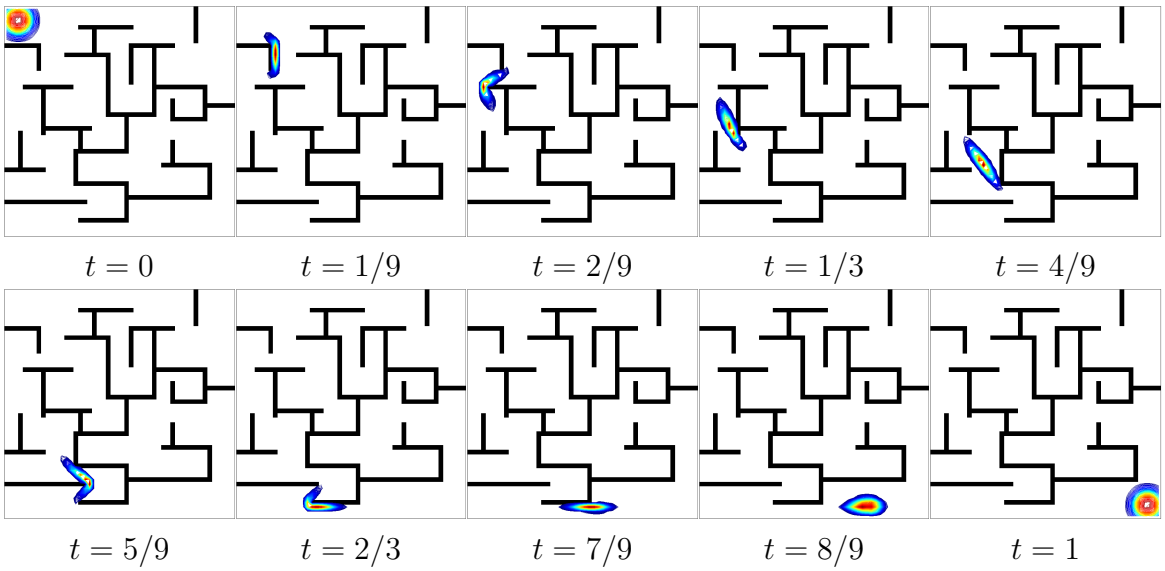


Figure 3.3 – Evolution of $\rho(t, \cdot)$ for several values of t, using a Riemannian manifold with weights $w(\mathbf{x})$ (constant in time) restricting the densities to lie within a 2-D static labyrinth map.

Isotropic Riemannian metrics corresponds to matrices $A(t, \mathbf{x}) = w(\mathbf{x}) \operatorname{Id}_d$ proportional to the identity at each point, but this extends to arbitrary Riemannian metrics $A(t, \mathbf{x}) = A(\mathbf{x})$. The existence of a solution of the Monge-Kantorovitch problem on Riemannian manifolds has been shown in [141]. The equivalence with a fluid mechanism formulation has been studied in [120]. In the case of a time dependent tensor $A(t, \mathbf{x})$, even if the numerical algorithms are still providing a seemingly sound solution (see Figure 3.4), there are no equivalence with any static Optimal Transport problem and the proof of existence of solutions remains an open subject.

Figure 3.4 shows a more complicated setting that includes a labyrinth with moving walls: a green wall appears at time $t = 1/4$ and a red one disappears at time $1/2$. The difference with respect to the previous example is the fact that w is now time dependent. This simple modification has a strong impact on the displacement interpolation. Indeed, the speed of propagation of the mean of the density is not constant anymore since the density measure is confined in a small area surrounded by the walls for $t \in [1/4, 1/2]$.

3.1.3 Anisotropic transport

The introduction of the anisotropy modeling through non diagonal matrices A is now illustrated. With adequate matrices A, it is now possible to define a polarization of

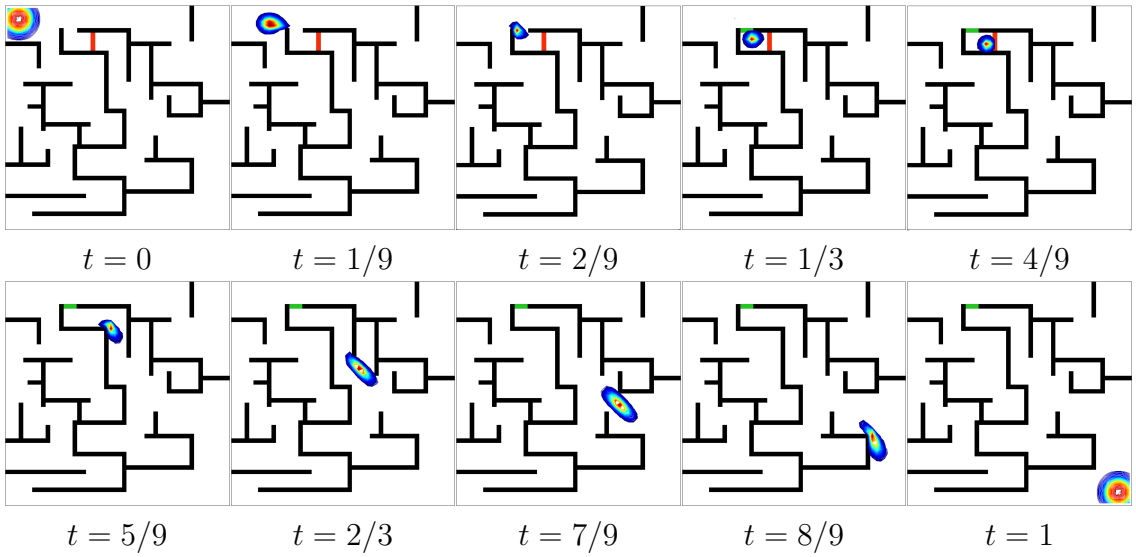


Figure 3.4 – Evolution of $\rho(t, \cdot)$ for several values of t, using a Riemannian manifold with weights $w(t, \mathbf{x})$ (evolving in time) restricting the densities to lie within a 2-D dynamic labyrinth map (*i.e.* with moving walls in green and red).

the space for modeling anisotropic domain priors. This problem is illustrated in Figure 3.5, which presents the optimal path between a horizontal and a vertical line on a square image. For these experiments, we defined different complex anisotropies, that are represented with the arrows of the first image of each line. If prior knowledge is available on the data, one can build a specific Riemannian manifold and the anisotropic model will therefore be able to simulate rigid or divergence free transports. This modelling is nevertheless not sufficient for more general purposes and we now propose to directly include physical prior into the model and not into the domain.

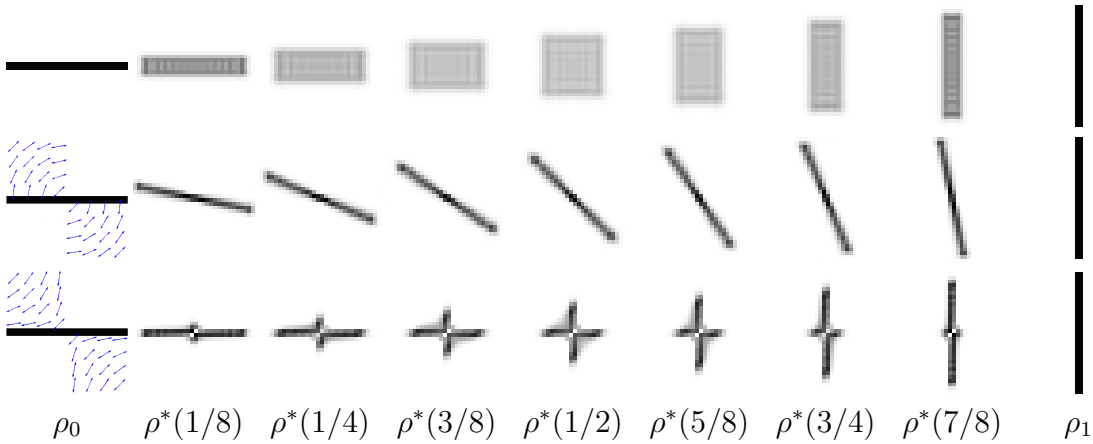


Figure 3.5 – Illustration of the mass $\rho^*(t)$ estimated between ρ_0 and ρ_1, through the computation of the transport costs defined by two different anisotropic domains A_k, illustrated by the blue directions in the first column of the two last rows. The first row is the isotropic transport.

3.2 Optimal Transport with physical priors

With the previous formulation, it is now possible to deal with islands for processing ocean images. Our underlying objective is to incorporate physic priors into the Optimal Transport model. More precisely, for general images, we would like to consider incompressible or rigid transports. Such transports can indeed prevent the object contained within the data from splitting along the computed paths, as for classical transport of Figure 2.4. The physical constraints can be characterized using the velocity, that is not a variable of the problem (2.11) anymore and should be reintroduced. We thus come back to the classic Optimal Transport framework with $\beta = 1$ and $A = \mathrm{Id}_d$.

3.2.1 Non-convex coupling

In order to have a coupling between the variables (ρ, m) and v, a natural idea is to consider the term $\iota_D(\rho, m, v) = 0$, with the set $D = \{m = \rho v\}$. As D is not convex and ι_D non-smooth, we rather consider a differentiable coupling[1]:

$$K(\rho, m, v) = \frac{1}{2} \int_\Omega \int_0^1 ||m - \rho v||^2 \, \mathrm{d}\mathbf{x} \, \mathrm{d}\, t.$$

In [120], specific kind of transports were promoted regarding velocity priors that depend on the targeted application, as can be done with physical regularization of Optical Flow for fluid image sequences [69, 223]. A new functional term $R(v)$ is thus introduced. This term can be defined as:

- Divergence-free: $R(v) = \iota_{C^v}(v), C^v = \{\mathrm{div}_\mathbf{x}\, v(t,\cdot) = 0,\ \langle v(t,\cdot), \vec{n}\rangle = 0,\ \forall\, t \in [0;1]\}$,
- Incompressible penalization: $R(v) = \int_0^1 ||\,\mathrm{div}_\mathbf{x}(v(t,\cdot))||^2_{L^2(\Omega)} \,\mathrm{d}\, t$,
- Rigid penalization: $R(v) = \int_0^1 ||(\nabla_\mathbf{x} v(t,\cdot) + (\nabla_\mathbf{x} v(t,\cdot))^T)/2||^2_{L^2(\Omega)} \,\mathrm{d}\, t$,
- Translation penalization: $R(v) = \int_0^1 ||(\nabla_\mathbf{x} v(t,\cdot)||^2_{L^2(\Omega)} \,\mathrm{d}\, t$.

By *translation penalization*, we mean a penalization of the deviation of a velocity field from translations. These terms are convex and the three last ones are differentiable. The existence of minimizers for the translation case has been shown in [120]. The generalized Optimal Transport model we are interested in now reads:

$$\min_{\rho,m,v} F(\rho, m, v) := \mathcal{J}(\mathcal{I}(\rho, m)) + \iota_C(\rho, m) + \lambda K(\mathcal{I}(\rho, m), \mathcal{I}(v)) + \alpha R(v) \tag{3.3}$$

where $\lambda, \alpha \geq 0$ respectively weight the coupling between variables and the velocity regularization term. The problem (3.3) is not convex in (ρ, m, v) but it is separately convex in (ρ, m) and in v. As the coupling is differentiable and the non-smooth terms are separable, following [210], we can perform a block coordinate descent with Algorithm (A.19) and minimize alternatively each convex problem in (ρ, m) and in v to obtain a critical point of the joint problem (3.3). Each subproblem can be solved with Algorithms presented in Appendix A.

[1] In [139], a different choice has been made and the problem is solved on (ρ, v) while penalizing the non convex constraint $||\partial\rho + \mathrm{div}(\rho v)||^2$.

3.2.2 Synthetic tests

First of all, we compare in Figure 3.6 the results obtained with the incompressible, translation and rigid penalizations on the two Gaussians example. Contrary to standard transport shown in Figure 2.4, such physical priors prevent the mass from splitting along the computed path. The Gaussians are deformed with the divergence-free prior, but it can be seen that the length of the level lines of the densities are preserved along the path. It is also important to underline that both translation and rigid penalizations keep the exact shapes of the two Gaussians along time. One can see in Figure 3.7 representing both computed paths, that the rigid penalization really performs a rotation and not a translation, so that the optimal path is no more composed of straight lines.

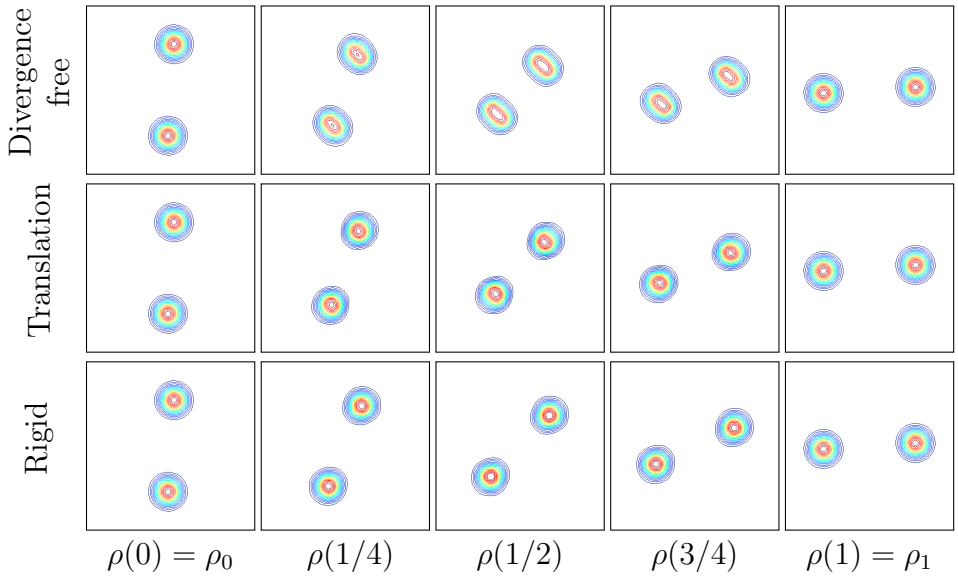
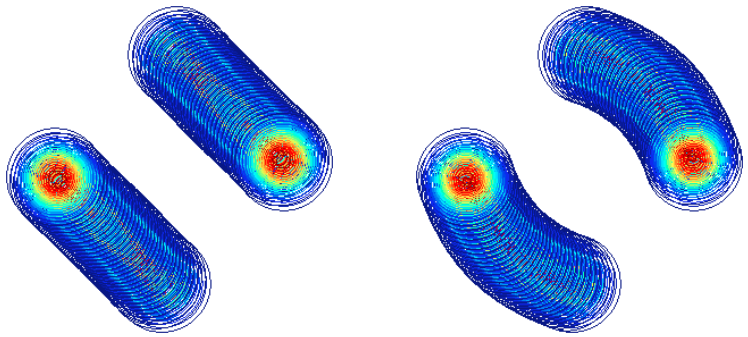


Figure 3.6 – Two Gaussians experiments with penalization and a null initialization. Plot of the isolevels of the density $\rho(t)$ along the different computed optimal path. The top line is realized with incompressibility (*i.e.* divergence free) penalization, the second with a translation penalization and the last one with a rigid penalization.



Translation penalization Rigid penalization

Figure 3.7 – Two Gaussians experiment. Plot of the whole trajectory computed with the translation and the rigid penalization models. The rigidity here involves a rotation.

In the example of Figure 3.8 that presents a rotating bar, the rigid penalization (last line) recovers a quasi-rotation, which better preserves the prior physics with respect to pure Optimal Transport (first line). As expected, it can also be observed in Figure 3.9 that the length of the level lines of the estimated density are preserved with the incompressible and rigid penalization approaches. We refer the reader to [39, 120] for more synthetic examples involving such penalizations.

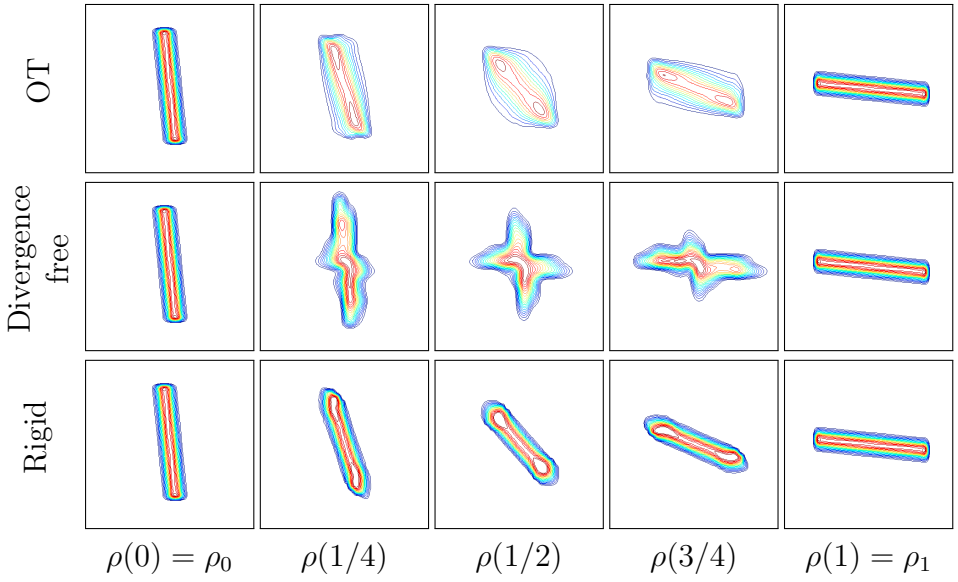


Figure 3.8 – Bar experiment. Plot of the isolevels of the density $\rho(t)$ along the optimal path computed with different approaches. The first line is the Optimal Transport, the second is the proposed approach with an incompressible penalization and the last one with a rigid penalization that conserves the nature of the object to transport.

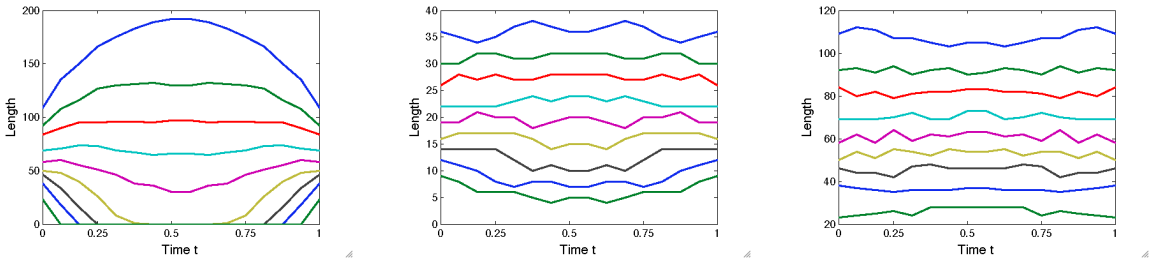


Figure 3.9 – Bar experiment. Evolution of the length of the upper level lines of the estimated density along time t: $|\rho(t, \mathbf{x}) > i/10|$, for $i = 1 \dots 9$. The left plot is the classic Optimal Transport, the middle one corresponds to the incompressible penalization and the right one corresponds to a rigid penalization. Penalizing the norm of the velocity makes the level lines almost preserved along the computed path.

3.2.3 Image interpolation in oceanography

In order to study the state of the oceans, snapshot images are produced by operational numerical codes such as the Ocean Circulation Model NEMO (http://www.nemo-ocean.eu/) . The synthesis process is nevertheless time-consuming and the oceanographers would like to create a few images and then realize a temporal interpolation between these images. This would be of particular interest to visualize and diffuse movies of dynamic structures of the ocean. The main issue comes from the coast that appears in a lot of interesting places, as illustrated in Figure 3.10. It makes useless classical image registration techniques such as optical flow or diffeomorphism estimations that can not deal with such complex domain.

With our formulation, the image domain can be represented as a Riemannian manifold, by taking $A(t, \mathbf{x}) = w(\mathbf{x}) \, \mathrm{Id}_2$. The variable $w(\mathbf{x})$ describes the manifold. It is set to 1 in the ocean and to a huge value in the land in order to restrict the transport into the ocean.

The successive optimal paths between 10 pairs of Sea Surface Height images of size 843×516 produced by the NEMO model have been computed with a temporal discretization of 9 steps. The brightest colors correspond to the highest sea height. The image sequence illustrates the Agulhas Current and the creation of vortexes in Cap Point. As shown in Figure 3.10, that presents the computed path between two consecutive images, by adding the proposed divergence-free penalization, we can induce some rotational prior within the transport estimation and recover the creation of vortexes.

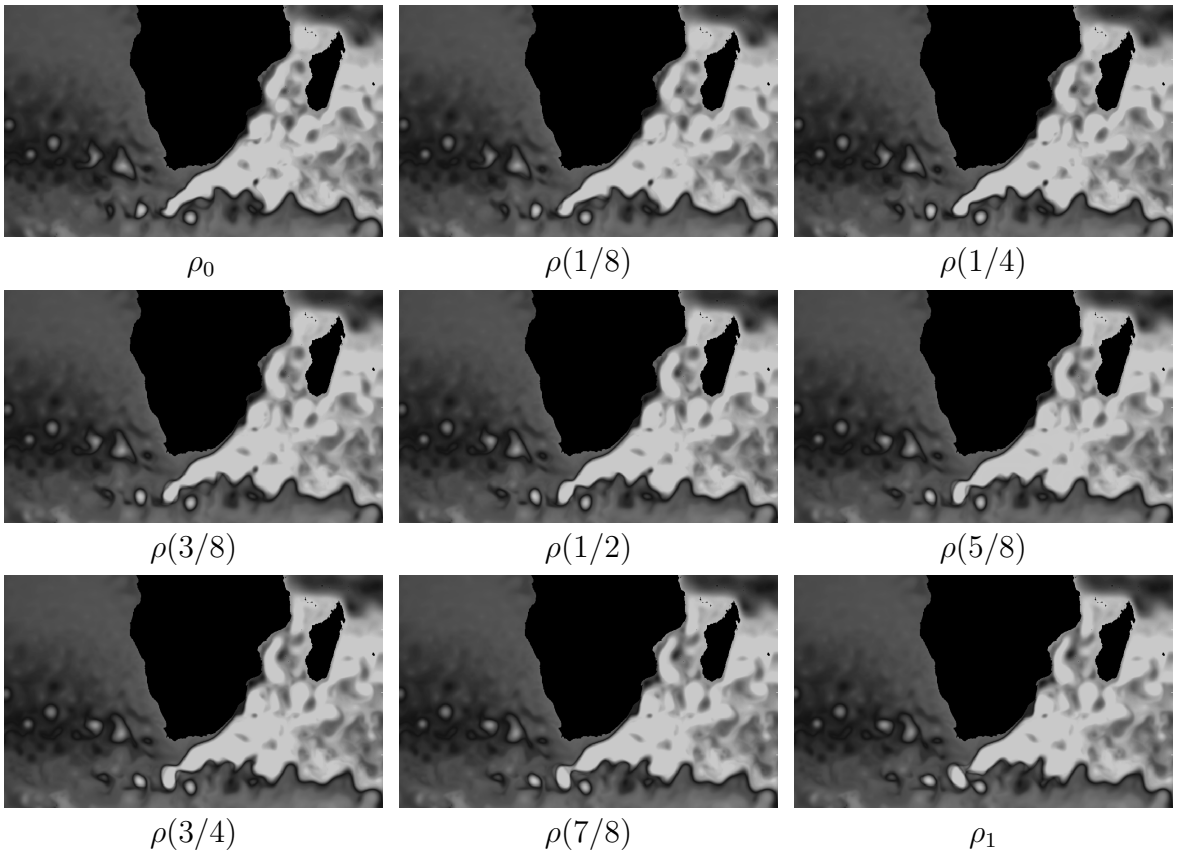


Figure 3.10 – Interpolation of Sea Surface Height images (ρ_0 and ρ_1) in oceanography: the creation of vortexes in Cap Point is finely estimated.

Conclusion of Part I

In this first part, we have shown how proximal splitting schemes offer an elegant and unifying framework to describe computational methods to solve the dynamical Optimal Transport with an Eulerian discretization. It allowed us to extend the original method of [19] in several directions, most notably the use of staggered grid discretization and the introduction of generalized, spatially variant, cost functions.

We have studied generalized Optimal Transport models which attach a multiphysics model to the images to be registrated. This is of particular interest for image interpolation purposes, where the results obtained using a simple minimization of a kinetic energy under some constraints do not preserve image characteristics along the optimal interpolation path, which is not physical. These generalizations are not limited by the expressions we consider here and others physical terms can also be considered taking into account more complex physics.

Promising results have been obtained on high-resolution oceanographic images. Future works will therefore be dedicated to the modeling of more accurate physical priors.

From the optimization point of view, we also focused on the minimization of non-convex and non-smooth functionals. Contrary to the block coordinate descent method we considered in (3.3), defining an algorithm without inner loops (on (ρ, m) and v) would be of main interest to obtain a faster computation of minima of our non-convex problems. Recent advances of proximal splitting methods have been made for related problems [8, 31], but they are still limited to simple functionals and would still require inner loops to deal with our generalized Optimal Transport formulation.

Finally, for the anisotropic Optimal Transport, we plan to study the existence of minimizers when the anisotropy varies with respect to time. This is a challenging problem as, in this case, there are no correspondences with any static Monge problem.

Notice that this method has found interest in the Computer Graphics community. Our original Matlab code has indeed been optimally reimplemented in C++ [32], where $56\times$ speed up has been observed. It is therefore possible to estimate a good geodesic path discretized with 256 timesteps between 256×256 grayscale images in 3 minutes. An extension of the proposed model to the interpolation of color images have also been proposed in [98].

Part II

Discrete Optimal Transport

Chapter 4

Relaxation and Regularization of Optimal Transport

An easiest way to discretize and compute numerically Optimal Transport is to consider finite sums of weighted Diracs. In this specific case, the Optimal Transport is a multi-valued map between the Diracs locations. The dimension of the transport map then scales with the product of the dimensions of the discretized densities. Despite being numerically intensive for finely discretized distributions, this discrete transport framework has found many applications. It first includes image retrieval with the well-kwown Earth Mover's Distance [187], and also color transfer between images [168], shape retrieval [181] or surface reconstruction [78] and interpolation for computer graphics [35].

As already seen in the previous part, the Optimal Transport map between complicated densities is usually irregular, which is a real drawback for interpolating between densities or for transfer purposes. In this chapter, we first recall the discrete setting of Optimal Transport and next present some regularization approaches that will be used for image processing purposes in Chapters 5 and 6.

4.1 Discrete Formulation

Monge's original formulation of the Optimal Transport problem aims at minimizing the cost for transporting a distribution μ onto another distribution ν using a map T

$$\min_T \int_X c(x, T(x)) \mathrm{d}\mu(x), \quad \text{where} \quad T\#\mu = \nu. \tag{4.1}$$

Here, μ, ν are measures in $\mathbb{R}^d$, $T : \mathbb{R}^d \to \mathbb{R}^d$ is a μ-measurable function, $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is a $\mu \otimes \nu_Y$-measurable function, and $\#$ is the push forward operator. We now consider the discrete formulation of the optimal mass transportation problem between a pair of histograms μ and ν. In the most general setting, histograms can be viewed as weighted sum of dirac distributions

$$\mu = \sum_{i=1}^{N_X} \mu_i \delta_{X_i} \quad \text{and} \quad \nu = \sum_{j=1}^{N_Y} \nu_j \delta_{Y_j},$$

where δ_X is the Dirac measure at location $X \in \mathbb{R}^d$ and $X = \{X_i \in \mathbb{R}^n\}_{i=1}^{N_X}$ and $Y = \{Y_j \in \mathbb{R}^n\}_{j=1}^{N_Y}$ are point-clouds in a feature space $\mathbb{R}^d$ (for instance position, color, gabor filter coefficients, etc). These points are often referred to as "bins" when considering a regular grid. Without loss of generality, we assume from now that the two histograms have positive weights $\mu_i, \nu_j > 0$, the same number of bins ($N_X = N_Y = N$) and the same total mass $\sum_i \mu_i = \sum_j \nu_j$.

4.1.1 Monge-Kantorovitch Formulation

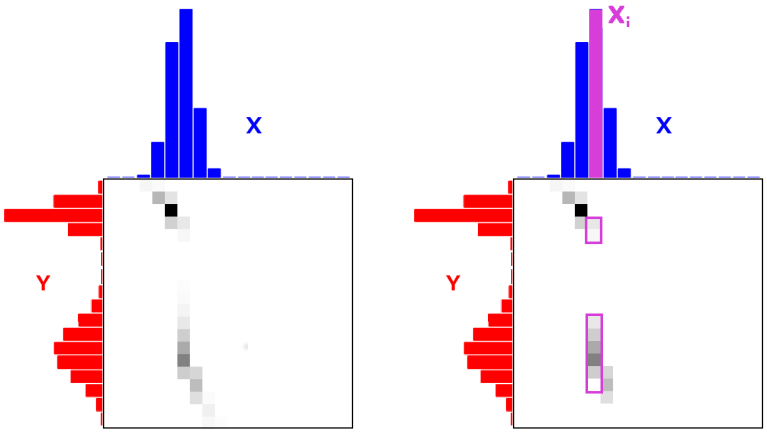
Given a fixed cost matrix $C = (C_{i,j})_{i,j}$ measuring a distance between locations X_i and Y_j, we can consider the Kantorovitch formulation of the Optimal Transport problem, which corresponds to estimating the Optimal Transport plan P^* between histograms. The Monge-Kantorovitch distance **MK** reads [125]

$$\mathbf{MK}(\mu, \nu) = \langle P^\star, C \rangle = \min_{P \in \mathcal{S}(\mu,\nu)} \left\{ \langle P, C \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} P_{i,j} C_{i,j} \right\}, \tag{4.2}$$

where the set of admissible transport matrices $\mathcal{S}(\mu, \nu)$ is the set of measures defined on the product space of considered histograms which marginals are given by μ and ν:

$$\mathcal{S}(\mu, \nu) := \{P \in \mathbb{R}_+^{N \times N}, P\mathbf{1}_N = \mu \text{ and } P^T \mathbf{1}_N = \nu\}. \tag{4.3}$$

As illustrated in Figure 4.1, the value $P_{i,j}$ of matrix P then corresponds to the proportion of the mass μ_i of cluster X_i that is transferred to the cluster Y_j. If $C_{i,j} = \|X_i - Y_j\|^p$, the value of the optimization problem (4.2) is called the L^p-Wasserstein distance (to the power p) and is denoted $W_p(\mu, \nu)^p = \mathbf{MK}(\mu, \nu)$. It can be shown that W_p defines a distance on the set of distributions that have moments of order p.



(a) (b)

Figure 4.1 – (a) Optimal Transport matrix P between histograms X (in blue) and Y (in red). The white color correspond to $P_{i,j} = 0$. (b) The transport map does not impose one-to-one mappings betweens bins and the bin X_i is transported to several bins Y_j.

Problem (4.2) can be reformulated as a linear program that scales with the product of the dimensions of the consider histogram. The computation of the transport cost is therefore limited in practice to histograms discretized with a low number of bins.

Linear programming methods (interior-point method [150], simplex algorithm [77]) or first-order methods (Proximal Algorithms presented in Appendix A, Frank–Wolfe algorithm, as proposed in [225]) can be applied to estimate the transport matrix P. Faster schemes exist for specific cost functions $C_{X,Y}$, such as for instance convex cost of the distance on the line (where it boils down to sorting the positions as $d = 1$) and the circle [83], concave costs on the line [84], the ℓ^1 distance [133].

In case $d > 1$, various approximations of the transportation distance have been proposed using Kantorovitch-Rubinstein discrepency [197] or iterative explicit $1D$ computations [33]. The computation can also be accelerated with multi-scale clustering [144, 190, 155], using the structure of the cost function [189] or considering Voronoi diagrams [12, 132].

Remark 1 (Transport between clouds of points)**.** *In the case where the measures are discrete, have the same number of points, and all points have the same mass, μ and ν are cloud of points and the transport T between them is a one-to-one assignment, which corresponds to optimizing over the set of permutation matrix P in the Kantorovitch formulation* (4.2)*. Standard combinatorial methods such as the Hungarian Algorithm [130] or the Auction Algorithm [26] can thus be used to compute the transport matrix. Notice that the set of permutation matrices is not convex, so one can rely on its convex hull, the set of bi-stochastic matrices. It is shown that the relaxation over this set is tight, meaning that there exists a solution of* (4.2) *which is a binary matrix, hence being also a solution of the original non-convex problem over permutation matrices, see [216].*

4.1.2 Dual formulation

An alternative for solving the Optimal Transport problem (4.2) consists in looking at its dual formulation:

$$\mathbf{MK}(\mu,\nu) = \max_{\substack{u\in\mathbb{R}^N,\, v\in\mathbb{R}^N \\ \text{s.t. } u_i+v_j\leq C_{i,j}\,,\,\forall\, i,j}} \mu^T u + \nu^T v, \tag{4.4}$$

which comes from the Legendre-Frenchel transform of the primal problem:

$$\mathbf{MK}^*(u,v) = \iota_{u_i+v_j\leq C_{i,j}}(u,v) \tag{4.5}$$

where ι_K stands for the characteristic function of the convex set K. In this case the number of unknowns boils down to $2N$ but the number of linear constraints is N^2. Hence, solving the dual problem is as hard as solving the primal one. This dual formulation presents some advantages for particular image processing applications. It can for instance be considered to convexify vecto-valued labelling problems [202].

4.2 Relaxed and Regularized Transport

For applications such as color transfer, the regularization of the transport map is a necessary ingredient to provide visually plausible images. To that end, a simple idea consists

in adding a regularization penalty to the Optimal Transport energy (4.2). However, it leads to difficult non-convex problems, that have not yet been solved in a satisfying manner either theoretically or numerically.

Graph regularization and matching Based on concepts developed in manifold learning [208], regularizations built on top of a graph structure connecting neighboring points in the input density are classically used in imaging applications [92]. With graphs, one can design regularizations that are adapted to the geometry of the input density, that often has a manifold-like structure.

This idea of graph-based regularization of Optimal Transport can be interpreted as a soft version of the graph matching problem, which is at the heart of many computer vision tasks, see [16, 227]. Graph matching is a quadratic assignment problem, known to be NP-hard. Several convex approximations have been proposed, including for instance linear programming [4] and SDP programming [188].

Transport relaxation In the recent work of [138], it has been shown that in 1-D, no regularization is possible if one maintains a one-to-one assignment between the two densities. This is a first motivation for introducing a relaxed transport which is not a bijection between the densities. Another (more practical) motivation is that relaxation is crucial to solve imaging problems such as color transfer. Indeed, the color distributions of natural images are multi-modals. An ideal color transfer should match the modes together. This cannot be achieved by classical Optimal Transport because these modes often do not have the same mass.

A typical example is for two images with strong foreground and background dominant colors (thus having bi-modal densities) but where the proportion of pixels in foreground and background are not the same. Such simple examples cannot be handled properly with Optimal Transport. Monitoring the variation of mass between the matched densities requires an appropriate relaxation of the mass conservation constraint. Mass conservation relaxation is related to the relaxation of the bijectivity constraint in graph matching, for which a convex formulation is proposed in [226].

In the previous section, we introduced the Monge-Kantorovitch formulation for the computation of the Optimal Transport between two distributions as the minimization of the energy (4.2). In this section, we modify this energy in order to obtain a regular Optimal Transport mapping. We will also consider the recent entropic regularization of OT proposed in [73] that presents different and interesting properties, namely in terms of computational cost.

4.2.1 Relaxed Transport

In many applications in imaging, strict mass conservation should be avoided. As a consequence, it is not desirable to impose a one-to-one mapping between the points in μ and ν. The relaxation proposed in [94] allows each point of μ to be transported to multiple points of ν and vice versa. For this purpose, the Optimal Transport problem (4.2) is modified as:

$$\min_{P \in \mathcal{S}_\kappa(\mu,\nu)} \langle C_{X,Y},\, P \rangle \tag{4.6}$$

where

$$\mathcal{S}_\kappa(\mu,\nu) = \left\{ P \in [0,1]^{N\times N} \setminus \begin{array}{l} k_X\mu \leq P\mathbf{1} \leq K_X\mu, \\ k_Y\nu \leq P^T\mathbf{1} \leq K_Y\nu, \end{array} \mathbf{1}^T P\mathbf{1} = M \right\}. \tag{4.7}$$

In this new problem, $\kappa = (k_X, K_X, k_Y, K_Y) \in (\mathbb{R}^+)^4$ are the parameters of the method. They describe the minimum (k_X, k_Y) and maximum (K_X, K_Y) proportion of each bin that can be transported. To impose the total amount of mass M transported between the densities, it is necessary to further set the constraint $\mathbf{1}^T P\mathbf{1} = M$, where $M > 0$ is a parameter, basically taken as $M = 1$ for transporting all the mass from μ to ν. To ensure that $\mathcal{S}_\kappa$ is non empty, the condition $\max(k_X, k_Y) \leq M \leq \min(K_X, K_Y)$ has to be fulfilled. Notice that problem (4.6) can still be solved using standard linear programming algorithms.

4.2.2 Case of cloud of points

In the special case where μ and ν are cloud of points, one can consider $\mu = \nu = \mathbf{1}$. The following proposition shows that an optimal P is binary when the parameters κ are integers. Such a binary P can be interpreted as a set of pairwise assignments between the points in X and Y. Note that this is not true in general when the parameters κ are not integers.

Proposition 1. *For $(k_X, K_X, k_Y, K_Y, M) \in (\mathbb{N}^*)^5$, there exists a solution $\tilde{P}$ of (4.6) which is binary,* i.e. *$\tilde{P} \in \{0,1\}^{N\times N}$.*

This property is due to the fact that binary matrices $P \in \mathcal{S}_\kappa(\mu,\nu)$ are totally unimodular and the polytope $\mathcal{S}_\kappa(\mu,\nu)$ has integer vertices for integer values κ. Since there is always a solution of the linear program (4.6) which is a vertex of $\mathcal{S}_\kappa(\mu,\nu)$, it has coefficients in $\{0,1\}$ [193].

It is illustrated in Figure 4.2. Given a set of points X (in blue) and Y (in red), the optimal P is computed solving (4.6) for different values of κ. For each values of κ, a line between X_i and Y_j is drawn if the value of the associated optimal $P_{i,j} > 0.1$, solid if $P_{i,j} = 1$, and dashed otherwise. For non integer values of K_X, K_Y, the mappings $P_{i,j}$ are in $[0,1]$ while for integer values, $P_{i,j} \in \{0,1\}$. As the values of K_X, K_Y are increased (Figure 4.2, right), the points in X tend to be mapped to the closer points in Y.

4.2.3 Mean Optimal Transport Map

As stated before, when general distributions are concerned, optimal matrices P minimizing (4.6) are in general non binary and their non zero entries do not define one-to-one maps between the points of X and Y. As illustrated in Figure 4.3, it is however possible to define a map T from X to Y by mapping each point X_i to a weighted barycenter of its neighbors in Y as defined by P. This corresponds to defining

$$T(X_i) = \frac{\sum_{j=1}^N P_{i,j} Y_j}{\sum_{j=1}^N P_{i,j}} \tag{4.8}$$

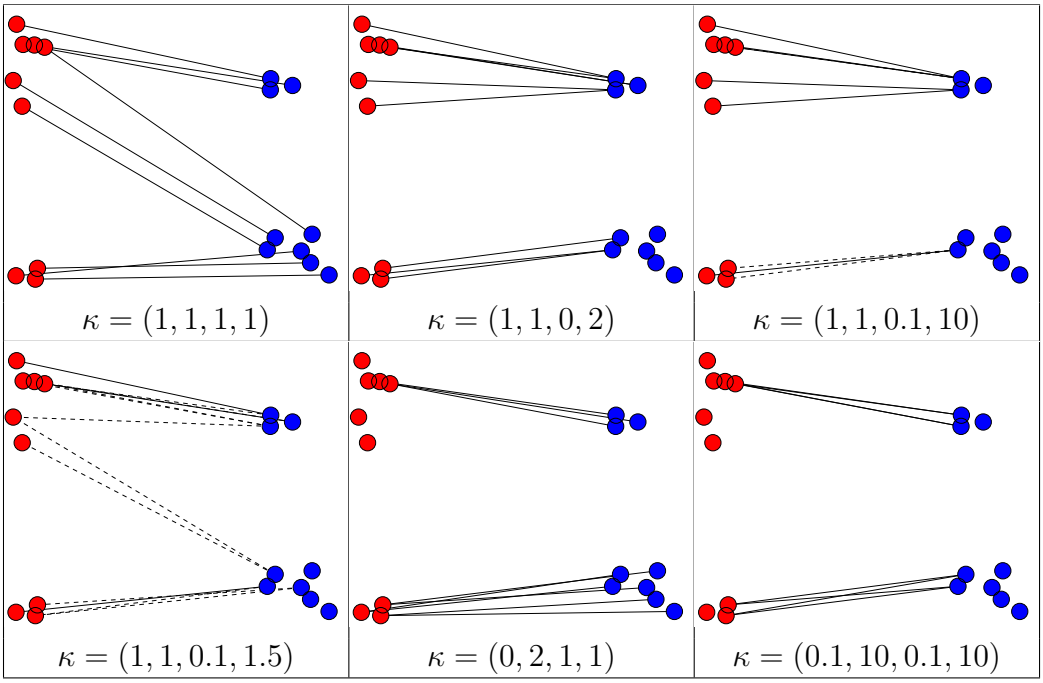


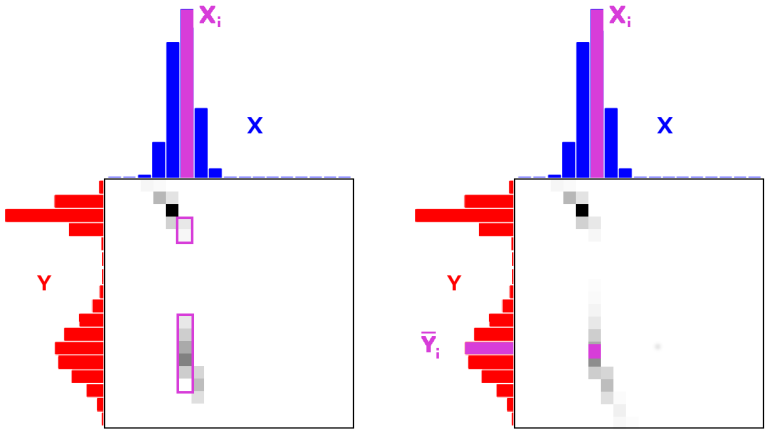
Figure 4.2 – Relaxed transport computed between X (blue dots) and Y (red dots) for different values of κ. Note that $\kappa = (1, 1, 1, 1)$ corresponds to classical Optimal Transport. The mappings $P_{i,j}$ that relate X_i and Y_j are plotted as line segments. The segments are dashed if $P_{i,j} \in]0.1, 1[$ and solid if $P_{i,j} = 1$. The parameter $\kappa = (1, 1, 0, 2)$ mean that the sum of links connecting each red point to blue points equals 1, while the sum of links connecting each blue point to red points is in $[0; 2]$. In this case, as all the parameters are integers, assignments are obtained (solid lines) and each red point is assigned to exactly one blue point, whereas each blue point can be connected to 0, 1 or 2 red points.

which in vectorial form can be expressed as $T(X_i) = ((\text{diag}(P\mathbf{1}))^{-1}PY)_i$, where the operator $\text{diag}(v)$ creates a diagonal matrix in $\mathbb{R}^{N\times N}$ with the vector $v \in \mathbb{R}^N$ on the diagonal. To insure that the map is well defined, we impose that $k_X > 0$. Note that it is possible to define a map from Y to X by replacing P by P^T in the previous formula and exchanging the roles of X and Y.

4.2.4 Discrete Regularized Transport

So far, a transport problem where the mass conservation constraint is relaxed has been presented. The second step is to define its regularization. A classic way of imposing regularity on a mapping $V : \mathbb{R}^d \to \mathbb{R}^d$ is to measure the amplitude of its derivatives. Two examples for continuous functions are the quadratic Tikhonov regularizations such as the Sobolev semi-norm $\|\nabla V\|^2$, and the anisotropic total variation semi-norm $\|\nabla V\|_1$ regularization. However, the differential operator ∇ cannot be applied directly to a point clouds X_i due to the lack of neighborhood definition. To extend the definition of the gradient operator, a graph structure on each point cloud has to be designed.

In the current setting, the goal is to regularize the discrete map T introduced in (4.8),



(a) (b)

Figure 4.3 – Computation of the mean transport map T from the Optimal Transport matrix P between histograms X (in blue) and Y (in red). (a) The bin X_i is transported to several bins Y_j. (b) The mean transport from bin X_i is $T(X_i) = \bar{Y}_i$, obtained from relation (4.8). The white color correspond to $P_{i,j} = 0$.

which is only defined at the location of the points as $X_i \mapsto \tilde{V}_i = X_i - T(X_i) = X_i - \mathrm{diag}(P\mathbf{1})^{-1}(PY)_i$. To avoid the normalization $\mathrm{diag}(P\mathbf{1})$ (which typically leads to non-convex optimization problems), the regularity can be imposed on the map $X_i \mapsto V_i = \mathrm{diag}(P\mathbf{1})X_i - (PY)_i$. This switch also has the advantage of imposing a stronger regularization in regions with large weights and to further regularize the variations of the weights $P\mathbf{1} \in \mathbb{R}^N$. These are actually interesting features to reduce artifacts for imaging applications.

Gradient on Graphs A natural way to define a gradient on a point cloud X is by using the gradient on a weighted graph $\mathcal{G}_X = (X, E_X, W_X)$ where $E_X \subset \{1, \dots, N\}^2$ is the set of edges and W_X is the set of weights, $W_X = (w_{i,j})_{i,j=1}^N : \{1, \dots, N\}^2 \mapsto \mathbb{R}^+$, satisfying $w_{i,j} = 0$ if $(i,j) \notin E_X$. The edges of this graph are defined depending on the application. A typical example is the n-nearest neighbor graph, where every vertex X_i is connected to X_j if X_j is one of the n-closest points to X_i in X, creating the edge $(i,j) \in E_X$, with a weight $w_{i,j}$. Because the edges are directed, the adjacency matrix is not symmetric.

The gradient operator on $\mathcal{G}_X$ is defined as $G_X : \mathbb{R}^{N \times d} \to \mathbb{R}^{P \times d}$, where $P = \|E_X\|$ is the number of edges and where, for each $V = (V_i)_{i=1}^N \in \mathbb{R}^d$,

$$G_X V = (w_{i,j}(V_i - V_j))_{(i,j) \in E_X} \in \mathbb{R}^{P \times d}.$$

A classic choice for the weights to ensure consistency with the directional derivative is $w_{i,j} = \|X_i - X_j\|^{-1}$, see for instance [107].

Regularity Term The regularity of a transport map $V \in \mathbb{R}^{N\times d}$ is measured according to some norm of $G_X V$, that we choose here for simplicity to be the following

$$J_{p,q}(G_X V) = \sum_{(i,j)\in E_x} \Big(\|w_{i,j}(V_i - V_j)\|_q \Big)^p,$$

where $\|.\|_q$ is the ℓ^q norm in R^d.

The case $(p,q) = (1,1)$ is the graph anisotropic total variation, $(p,q) = (2,2)$ is the graph Sobolev semi-norm, and $(p,q) = (1,2)$ is the graph isotropic total variation, see for instance [92] for applications of these functionals to imaging problems such as image segmentation and regularization. In practice, we observe that the optimal matrix P is somehow sparser with the total variation prior, *i.e.* the resulting weights are more concentrated on a few entries.

4.2.5 Symmetric Regular Optimal Transport Formulation

Given two histograms μ and ν, a relaxed and regularized Optimal Transport mapping between them can be computed. In order to make the mapping regular with respect to both histograms, the displacement fields to be regularized are:

$$T_{X,Y}(P) = \mathrm{diag}(P\mathbf{1})X - PY \quad \text{and} \quad T_{Y,X}(P^*) = \mathrm{diag}(P\mathbf{1})Y - P^T X.$$

In order to obtain a partial matching that is regular according to X and Y, two graphs $\mathcal{G}_X$ and $\mathcal{G}_Y$ are created (as described Page 47) and the corresponding gradient operators are denoted $G_X \in \mathbb{R}^{n_X\times N}$ and $G_Y \in \mathbb{R}^{n_Y\times N}$, where n_X and n_Y are the number of edges in the respective graphs. The symmetric regularized discrete Optimal Transport energy proposed in [93] is defined as:

$$\min_{P\in\mathcal{S}_\kappa} \langle P,\, C_{X,Y}\rangle + \lambda_X J_{p,q}(G_X T_{X,Y}(P)) + \lambda_Y J_{p,q}(G_Y T_{Y,X}(P)), \tag{4.9}$$

where $(\lambda_X, \lambda_Y) \in (\mathbb{R}^+)^2$ controls the desired amount of regularity. The case $\kappa = (1,1,1,1)$ and $(\lambda_X, \lambda_Y) = (0,0)$ corresponds to the usual Optimal Transport defined in (4.2), and $(\lambda_X, \lambda_Y) = (0,0)$ corresponds to the un-regularized formulation (4.6).

By setting the parameters $q = p = 1$ to obtain the anisotropic Total Variation (TV) norm, problem (4.9) can again be re-written as a linear program by introducing the auxiliary variables $U_X \in \mathbb{R}^{n_X\times d}$ and $n_Y \in \mathbb{R}^{P_Y\times d}$

$$\begin{aligned} \min_{P,U_X,U_Y} \quad & \langle C_{X,Y},\, P\rangle + \lambda_X\langle U_X,\, \mathbf{1}\rangle + \lambda_Y\langle U_Y,\, \mathbf{1}\rangle \\ \text{subject to} \quad & \begin{cases} -U_X \le G_X(PY - \mathrm{diag}(P\mathbf{1})X) \le U_X, \\ -U_Y \le G_Y(P^T X - \mathrm{diag}(P^T\mathbf{1})Y) \le U_Y, \\ \quad P \in \mathcal{S}_\kappa. \end{cases} \end{aligned} \tag{4.10}$$

In Figure 4.4, the influence of the parameters κ and (λ_X, λ_Y), from equation (4.9) are illustrated on a synthetic example. For $\lambda_X = \lambda_Y = 0$ one obtains the relaxed symmetric

Optimal Transport solution, where the transport maps the points in X to the closest point on Y, and vice versa. As we increase the values of λ_X and λ_Y to 0.001, we can see how the regularization affects the mapping. Let us analyze $J_{p,q}(G_X T_{X,Y}(P)) = \|G_X \operatorname{diag}(P\mathbf{1})X - G_X PY\|^2$, for instance. The term $G_X \operatorname{diag}(P\mathbf{1})X$ is measuring the regularity of the weights $\operatorname{diag}(P\mathbf{1})$ on X and the consequence is that for $\lambda_X = \lambda_Y = 0.001$ there are plenty of connections with low weight (there are few solid lines), while for $\lambda_X = \lambda_Y = 0$ there are several mappings with $P_{i,j} = 1$ (solid lines). So, the regularization promotes a spreading of the matchings.

The minimum of $J_{p,q}(G_X T_{X,Y}(P))$ is reached when $G_X \operatorname{diag}(P\mathbf{1})X = G_X PY$, that is, when the graph structure of X has the same shape as the graph structure of PY, which both can be observed in the last column and row. For high values of $\lambda_X = \lambda_Y$ the matchings tend to link the clusters by their shape, that is, the big cluster on X with the big cluster of Y, and similarly for the small clusters (note that the links with higher value are between the small clusters).

We will show in the next chapter how such framework can be efficiently applied to color transfer. In the meanwhile, we finally present the entropic regularization of Optimal Transport that will be mainly used for segmentation purposes.

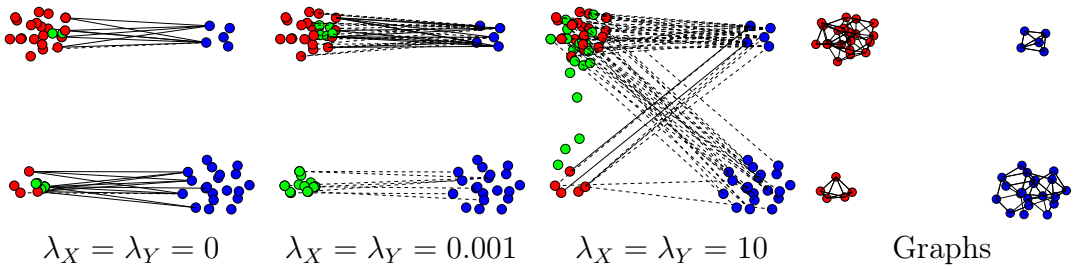


Figure 4.4 – Given two sets of points X (in blue) and Y (in red), we show the points $Z = \operatorname{diag}(P\mathbf{1})^{-1}PY$ (in green), and the mappings $P_{i,j}$ as line segments connecting X_i and Y_j, which are dashed if $P_{i,j} \in]0.1, 1[$ and solid if $P_{i,j} = 1$. The results were obtained with the relaxed and regularized Optimal Transport formulation, setting the parameters to $\kappa = (0.1, 8, 0.1, 8)$. Note the influence of a change in λ_X and λ_Y on the final result: with no regularization ($\lambda_X = \lambda_Y = 0$) only few points in the data set are matched. The introduction of regularization ($\lambda_X = \lambda_Y = 0.001$) spreads the connections among the clusters, while maintaining the cluster-to-cluster matching. For a high value of $\lambda_X = \lambda_Y = 10$, the regularization tends to match the clusters with similar shape with each other, where the shape is defined by the graph structure. The graphs $\mathcal{G}_X$ and $\mathcal{G}_Y$ are represented with the nodes on blue and red respectively, and the edges as solid lines.

4.3 Approximate MK cost using Sinkhorn distances

The entropy smoothing of Optimal Transport proposed in [218, 73] reads

$$\mathbf{MK}_\lambda(\mu, \nu) = \langle P^\lambda, C \rangle = \min_{P \in \mathcal{S}(\mu,\nu)} \left\{ \langle P, C \rangle - \frac{1}{\lambda} h(P) \right\}, \tag{4.11}$$

where the entropy of the matrix P is defined as $h(P) := -\sum_{i,j} P_{i,j} \log P_{i,j}$. As shown in [183], it acts as a barrier function that ensures the positivity of $P_{i,j}$ while making the problem strongly convex. The transport cost $\mathbf{MK}_\lambda(\mu, \nu)$ is called the Sinkhorn distance. Denoting as $K = e^{-\lambda C}$, it can be shown [198, 73] that the optimal transport map can be written as $P^\lambda = diag(u)Kdiag(v)$ and can be estimated with the following fixed point iteration algorithm on u and v:

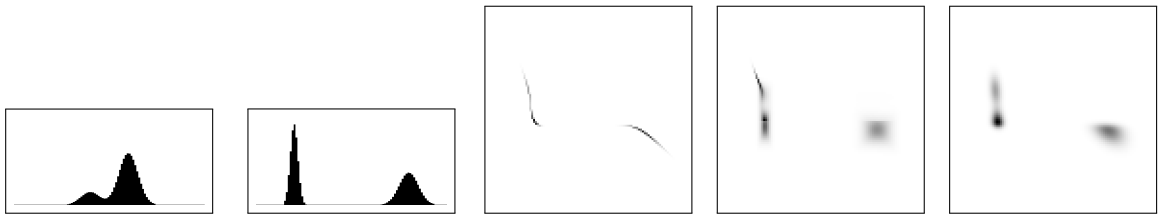
$$u_i = \frac{\mu_i}{\sum_j K_{i,j} v_j}, \quad v_j = \frac{\nu_j}{\sum_i K_{i,j} u_i} \tag{4.12}$$

One should note that such type of normalization, *i.e.* $x \mapsto \frac{x}{\|x\|_1}$ is the projection on the ℓ_1 ball (or equivalently probability simplex) using the Kullback-Leibler divergence [21]. Indeed, another way to express the regularization term is:

$$\mathbf{MK}_\lambda(\mu, \nu) = \min_{P \in \mathcal{S}(\mu,\nu)} \langle P, C \rangle + \frac{1}{\lambda} \mathbf{KL}(P \| \mathbf{1}).$$

When regular grids X_i and Y_j are involved and the cost $C_{i,j} \propto X_i - Y_j$, the previous algorithm (4.12) corresponds to convolutions with the kernel $K_{i,j} = e^{-\lambda C_{i,j}}$. For $C_{i,j} = ||X_i - Y_j||^2$, one recovers convolutions with Gaussian Kernels that can be implemented separately in each dimension thus leading to very fast algorithms [199].

In Figure 4.5, the matrices P of transport flows computed with the different regularization methods are illustrated. The null values $P_{i,j}$ are displayed in white color. These matrices describe how the mass from the bins of the input histogram (in rows) are sent to the bins of the target histogram (in columns). As a bin μ_i corresponds to the line i of the matrix, the variance of the transport map is larger when a lot of columns j are active (*i.e.* $P_{i,j} > 0$) and correspond to very different bins Y_j.



(a) Histogram a (b) Histogram b (c) Optimal Transport (d) Relaxed/ regularized (e) Entropy regularization

Figure 4.5 – Illustration of the effect of different regularizations of Optimal Transport computed between histograms (a) and (b). The standard Optimal Transport map is shown in (c). Relaxaed and Regularized Optimal Transport flow of [94] is presented in (d) and the entropic reguarization of [73] in (e).

Even if both regularizations have very different origins, we can see on this illustration that the influence over the transport matrix P is quite similar, as the mass of each bin of one histogram is spread to several bins on the other histogram.

The parameter λ of the entropy regularization [73] directly measures this dispersion. In the asymptotic case $\lambda \to 0$, the transport matrix is the one with the minimum entropy: $P = \mu\nu^T$. In the case $\lambda \to \infty$, one recovers the standard transport matrix (Figure 4.5 (c)), which is the one of maximum entropy. In practice the convergence of

the algorithm (4.12) that estimates the $2N$ variables is faster for small values of λ. A trade-off between computational cost and quality of the approximate transport has to be chosen.

On the other hand, even if it leads to slower algorithms (since a linear program with N^2 variables has to be solved), the explicit regularization of the transport plan [94] has a physical meaning which presents some advantages as will be seen later.

Dual formulation As shown in [75], the Fenchel transform of (4.11) reads

$$\mathbf{MK}^*_\lambda(u,v) = \frac{1}{\lambda}\sum_{i,j}\left(e^{-\lambda(C_{i,j}-u_i-v_j)-1}\right). \tag{4.13}$$

Note that the indicator function for condition $C_{i,j}-\alpha\mu_i-\beta\nu_j$ of the standard transform (4.5) has been smoothed by an exponential function. As noticed in [76], such function is smooth and its gradient can be computed explicitly, which is a main advantage for incorporating **MK** terms in more general functionals to be minimized.

More generally, generalized constraints can be naturally integrated with the dual formulation of the entropic regularization. A simple example concerns unnormalized histograms μ and ν. In order to impose them to have a total mass of M, the regularized Optimal Transport function has been extended in [176] as follows:

$$\mathbf{MK}_{\lambda,M}(\mu,\nu) = \min_{P\in\mathcal{S}(\mu,\nu)}\left\{\langle P,\, C\rangle - \frac{M}{\lambda}h\left(\frac{P}{M}\right)\right\} + \iota_{H_M}(\mu,\nu), \tag{4.14}$$

where

$$H_M := \{\mu,\nu\in\mathbb{R}^N_+, \langle\mu,\mathbf{1}_N\rangle = \langle\nu,\mathbf{1}_N\rangle = M\}. \tag{4.15}$$

The Fenchel transform of the regularized transport cost (4.14) reads:

$$\mathbf{MK}^*_{\lambda,M}(u,v) = \frac{M}{\lambda}\log\sum_{i,j}\left(e^{-\lambda(C_{i,j}-u_i-v_j)}\right), \tag{4.16}$$

which directly encodes the mass constraint.

In the next chapters, we will see how these different regularizations can be used to deal with image processing problems.

Chapter 5

Application to Color Transfer

Many image processing applications require the modification or the prescription of some characteristics (colors, frequencies or wavelet coefficients) of a given image, while preserving other features. Statistics to be prescribed may come from prior knowledge, or more generally, are learned from an example. In such a case, another image is selected from a database to define a template. Such a framework arises for image enhancement, inpainting, colorization of grayscale or infrared images, tone mapping, color grading or color transfer. In this paper, we will focus on this last application through histogram transfer between images.

Color transfer consists in modifying an image to match the color palette of another one, while preserving its geometry. In the literature, the different interpretations and definitions of color palettes have led to various algorithms. In the following, we only consider unsupervised approaches. Since the seminal work of [184], methods have been designed to transfer some simple color statistics (*i.e.* the mean and standard deviation) in any color space [220]. These works share some common features with texture synthesis by example (see in particular [171]). More general approaches match the complete histogram of features from two images. When considering grayscale images, the problem is known as 1-D *histogram specification*. This framework has been extended for color histograms, using for instance 1-D ΔE-color index [146].

Histogram matching *via* optimal transportation There exist strong links between histogram specification and the Monge-Kantorovich Optimal Transport problem. Histogram transfer between grayscale images indeed corresponds to the application of the 1-D Optimal Transport plan between the gray levels of the pixels of the images, as noted by [81] for the case of histogram equalization. Some approaches to find fast approximate solution of OT for color images were investigated in [168, 180].

Spatial information As illustrated in Figure 5.1 (c), the exact transfer of color palette is generally not satisfying for practical applications in image processing [172]. The color distributions to be matched may have very different shapes, so that outliers generally appear in the processed image. Moreover, as the process is performed in the color space, it may not transfer coherent colors to neighboring pixels. It results in undesirable artifacts, such as JPEG compression blocks, enhanced noise, saturation or contrast inversion that should be corrected with specific post-processings [168, 174, 203]. As

a consequence, various models have been designed to incorporate some regularity priors on the image domain, such as Total Variation [153]. Color transfer may be formalized [161] as a variational problem in the image domain, in order to directly incorporate a spatial regularization of colors.

(a) I_X (b) I_Y (c) OT transfer (d) Post-processing

Figure 5.1 – Artifacts of color transfer with Optimal Transport. (a) Input image I_x. (b) Target color palette I_Y. (c) Exact color specification with optimal transport. (d) Post-processing with spatial regularization.

While spatial regularization suppresses small artifacts due to exact histogram specification (see Figure 5.1 (d)), it cannot handle strong artifacts due to an irregular OT map [179] and deal with potential very different modes (*i.e.* the dominant colors) between each distribution. For instance, one would like to only transfer a limited quantity of purple color into the flower example of Figure 5.1.

Hence, we investigate in this chapter the use of relaxed and regularized Optimal Transport for proposing models robust to color outliers that directly include spatial information in the image domain.

5.1 Convex and regularized Optimal Transport model

This section shows how the relaxed and regularized Optimal Transport formulation introduced in Chapter 4 can be applied to color transfer, and how the regularization and the relaxation improve the results obtained by previous methods.

The color transfer problem consists in modifying a source image I_X so that its colors match the colors of a target image I_Y. We denote as $I_X : \Omega \subset \mathbb{Z}^2 \mapsto \Sigma \subset \mathbb{R}^3$, where Ω is the regular pixel grid and Σ is the quantized $3D$ RGB color space. The problem is to find a new image I_Z whose geometry is as close as possible to the source image I_X and whose color distribution is close to the one of the exemplar image I_Y. The color histogram of such an image I_X can be estimated using the empirical distribution μ_X. The goal of color transfer algorithms is to compute a transformation T such that for

all pixel $x \in \Omega$, $I_Z(x) = T(I_X(x))$, where the new empirical distribution μ_Z is close (or equal) to μ_Y. Figure 5.2 shows an example where I_X, I_Y are the original input images, the second row displays the 2-D projection of the 3-D distribution of pixels μ_X and μ_Y, and in the third column, we show the distribution μ_Z which is the result of applying T to I_X, where T is computed using the Optimal Transport framework. The associated image I_Z has the geometry of I_X and the color palette (3-D histogram) of I_Y.

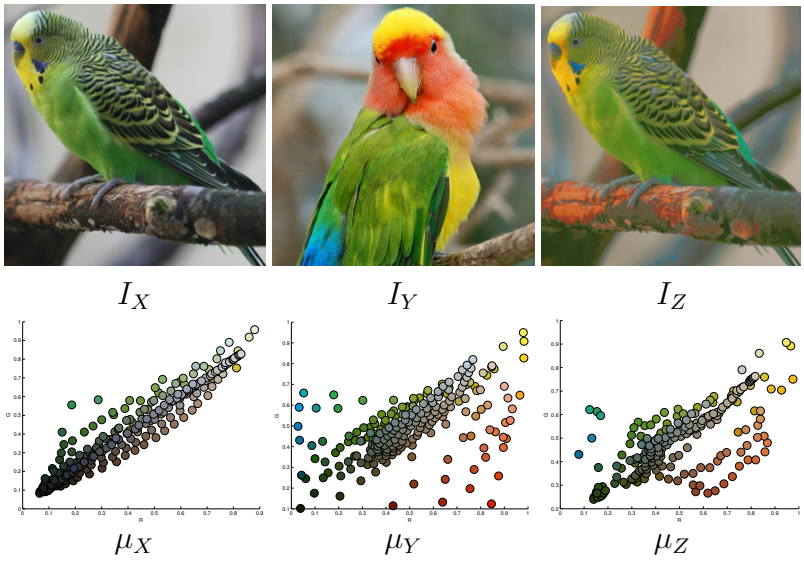


Figure 5.2 – Example of color transfer. Given images I_X and I_Y with their corresponding 3-D RGB color distributions μ_X and μ_Y (represented here using the 2-D projection of every pixel in the RG plane), the goal is to define an image I_Z that has the geometry of I_X and a histogram μ_Z that is similar to μ_Y.

Problem Formulation We consider the bipartite weighted matching problem between two histograms. Let μ and ν be two discrete and normalized distributions of features $X \in \mathbb{R}^{N_X \times d}$ and $Y \in \mathbb{R}^{N_Y \times d}$, respectively. Taking $N_X = N_Y = N$, their corresponding weights are $\mu \in \mathbb{R}^N$ and $\nu \in \mathbb{R}^N$, so that: $\mu = \sum_i \mu_i \, \delta_{X_i}$ and $\nu = \sum_j \nu_j \, \delta_{Y_j}$. We look for a mapping $T : \mathbb{R}^d \mapsto \mathbb{R}^d$ between X and Y, referred to as *transfer* map in the following. In order to transport μ towards ν, we define $T\#\mu = \sum_i \mu_i \delta_{T(X_i)}$. For color transfer, the mapping should have the following properties: *(i)* $T\#\mu$ should be close to or match the target distribution ν, *(ii)* T should be regular in some sense.

5.1.1 Adaptive OT

Relaxing the mass conservation constraint is crucial in order to better match the modes (*i.e.* the dominant colors) of each distribution. Regularizing the transport is also important to reduce colorization artifacts. Inspired from functional (4.9), the regularized model proposed in [175] aims at tackling the aforementioned issues by relaxing the matching constraint towards the target distribution ν on the matrix P.

This is done by introducing the set of acceptable transfer maps $\mathcal{P}_\kappa(\mu,\nu)$:

$$\mathcal{P}_\kappa(\mu,\nu) = \{P \in [0,1]^{N\times N} \setminus P\,\mathbf{1}_N = \mu,\ P^T\mathbf{1}_N = \kappa\otimes\nu\}$$
$$\text{where } \kappa\in\mathbb{R}^N,\ \kappa\geq 0, \kappa\otimes\nu = (\kappa_j\,\nu_j)_j \text{ and s.t. } \langle\kappa,\,\nu\rangle \geq 1.$$

Observe here that the second constraint involves a set of local relaxation parameters $\kappa = (\kappa_j)$. With respect to the previous admissible set (4.7), the relaxation with these parameters allow the model to increase or decrease the capacity value of the corresponding feature Y_j of the target histogram μ_Y; the maximum amount of mass from μ that Y_j can receive is now bounded by $\kappa_j\,\nu_j$. If $\kappa = \mathbf{1}_{N_Y}$, we get the previous Optimal Transport problem. The last constraint $\langle\kappa,\,\nu\rangle \geq 1$ prevents the set of acceptable transport maps $\mathcal{P}_\kappa(\mu,\nu)$ from being empty. With this relaxed parameterization, a model that locally adapt itself to the data can be defined.

Convex formulation When considering color transfer, we would like the transport to be piecewise constant, since we do not want to penalize color shift. To that end, as in (4.8), we consider the regularization of the average transport displacement $T(X_i) = (D_\mu PY - X)_i$, where $D_\mu = diag(\mu^{-1})$. As in the previous chapter, this regularization is done on the gradient G_X of the average transport T, that is computed on the graph created from the nodes X_i. Moreover, in order to control that the relaxed and regularized transported color palette is close to the target, an additional term $\|\kappa - \mathbf{1}_N\|$ is added to check that the relaxation is tight, *i.e.* that the final transported density will be close enough to the proportion of colors in ν.

Introducing the regularization parameter $\lambda \geq 0$ and the fidelity one $\rho \geq 0$, the model derived from (4.9) is the following convex optimization problem with linear constraints:

$$\{P^\star,\kappa^\star\} \in \underset{\substack{P\in\mathcal{P}_\kappa(\mu,\nu)\\ \kappa\in\mathbb{R}^N,\kappa\geq 0,\langle\kappa,\,\nu\rangle\geq 1}}{\operatorname{argmin}} \langle C_{XY},\,P\rangle + \lambda\,\|G_X(D_\mu PY - X))\|_1 + \rho\|\kappa - \mathbf{1}\|_1. \qquad (5.1)$$

This energy can be minimized with a linear solver by introducing auxiliary variables to deal with the L^1 norms, as detailed in (4.10). This formulation generalizes the Optimal Transport model proposed in [94] on cloud of points, to handle distributions. Observe that the transport is here automatically relaxed, since we do not require the user to tune the capacity vector κ.

Spatio-color clustering To make the optimization problem (5.1) tractable for histograms obtained from large scale images, a joint clustering is performed on the pixel grid and on the color space ($d = 5$). The images I_X and I_Y are quantized to generate X and Y with N clusters. We denote $X_i = (x_i, c_i) \in \mathbb{R}^5$, to specify the spatial component ($x_i \in \Omega$) and the color component ($c_i \in \Sigma$). These values (x_i, c_i) represent the mean spatial and color values of the set of pixels in I_X (and equivalently for I_Y) assigned to cluster X_i. As illustrated in Figure 5.3 (a-b), this assignment is computed using the fast super-pixels method [1]. Note that each cluster X_i has an assigned group of pixels $\{\mathcal{N}_i \subset \Omega\times\Sigma\}_i$ that can have a different amount $|\mathcal{N}_i|$ of pixels. We define the cluster weight as $\mu_i = |\mathcal{N}_i|/|\Omega|$.

Thus, we consider from now the measure $\mu : X = (x,c) \mapsto \sum_i \mu_i\delta_{X_i}(X) = \sum_i \mu_i\delta_{x_i}(x)\delta_{c_i}(c)$, and similarly for the measure ν obtained from image I_Y.

Graph With this clustering involving spatial information of color location, color consistency in the image domain can be ensured through graph regularization. A graph is created from the nodes X_i that are assumed to be linked to P nodes X_j. The gradient $G_X V \in \mathbb{R}^{N\times P\times d}$ of a field $V = \{V_i^l\} \in \mathbb{R}^{N\times d}$ on $\mathcal{G}_X$ is computed at point X_i as

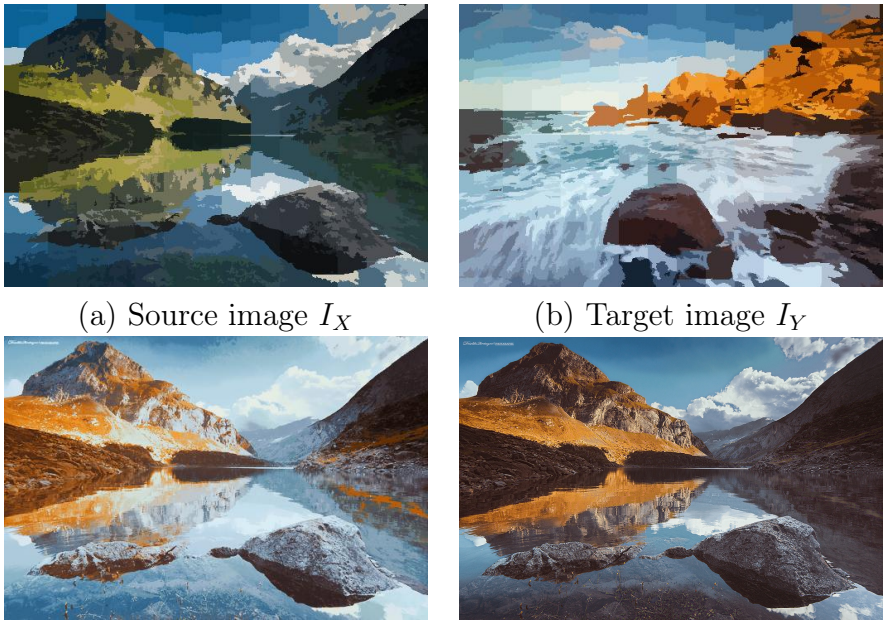
$$(G_X V)_i = (w_{ij}(V_i - V_j))_{j)} \in \mathbb{R}^{P\times d},$$

where the weight w_{ij} between cluster X_i and X_j is non null if the nodes are linked in the graph. In this case, the weight relies on the clusters' distances: $w_{ij} \propto \exp(-||X_i - Y_j||)$.

Image Synthesis As we work at a super-pixel scale to speed-up the Optimal Transport computation, the last step is to synthesize a new full resolution image I_Z from the source image I_X using the new color palette $T\#\mu$. Following [207], a maximum likelihood estimation can be used to incorporate geometrical information from the source image I_X into the synthesis process. For each pixel $\tilde{X} = (x, I_X(x))$ of the source image, the idea is to use the likelihood to all clusters $\{X_i\}_i$ through the weights $\{\omega_i\}_i$ in order to compute a linear combination of the estimated transferred colors $T(X_i)$. Hence, we finally get for each pixel $x \in \Omega$: $(x, I_Z(x)) = Z_i(x) = \frac{1}{W(\tilde{X})}\sum_i w_i(\tilde{x})T(X_i)$, with $w_i(\cdot) = \exp\left(-\frac{1}{2}\|\cdot - X_i\|^2\right)$, and the normalizing factor $W(\tilde{X}) = \sum_i w_i(\tilde{X})$.

5.1.2 Color transfer experiments

A first example is presented in Figure 5.3, where the super-pixel segmentation of the input (a) and target images (b) are first displayed. The color transfers obtained without (c) and with (d) the automatic tuning of the relaxation parameters are then presented.



(a) Source image I_X (b) Target image I_Y

(c) Optimal Transport (d) Adaptive model

Figure 5.3 – Color transfer process. Super-pixel segmentation of (a) source image I_X and (b) target image I_y. Color transfer without (c) and with (d) relaxation (*i.e.* $\rho = \infty$ and $\rho = 10^3$). *Photo credit: Nicolas Le Dilhuit*

A comparison with state-of-the art methods [168], [161, 94] is given in Figure 5.4. Notice that the simplified optimal transport of colors with Principal Component Analysis of [168] has been improved with a post-processing [174]. Thanks to the super-pixel segmentation, the process only requires a few seconds whereas the methods of the literature last a few minutes. Contrary to [168, 161], the target color palette is not perfectly matched, since the model adapts itself to the source image data. It leads to a transfer visually more plausible, as illustrated in the flower example of Figure 5.5 where the correct proportion of purple has been selected. With respect to [94], the improvement of the spatial regularization involved by the super-pixel clustering removes artifacts that appear for instance in the background. Figure 5.6 finally presents examples of color transfer on HDR images.

I_X I_Y [94]

[168] [161] [175]

Figure 5.4 – Comparison of the adaptive approach with [168] (with the post-processing of [174]), [161] and [94]. *Photo credit: Elektrofisch*

(a) I_X (b) I_Y (c) Adaptive model Relaxed palette

Figure 5.5 – Adaptive color transfer with relaxed Optimal Transport. (a) Input image I_x. (b) Target color palette I_Y. (c) Transfer with adaptive relaxation. (d) Relaxed color palette $\kappa \otimes \nu$ (the darker, the less a color is used). The model adapts itself to the color distributions and automatically selects the adequate proportion of color to transfer.

(a) (b)

Figure 5.6 – Examples of color transfer. (a) source image u and target color image v. (b) Final image w with post-processing. *Photo credit: 'Waves' photo by Daniele Zedda, 'Grafiti' and 'fire exits' by Thomas Leuthard, 'Canyon' by Chen Su, 'Desert' by Steve Yabek, and 'Creek' by Nicolas Le Dilhuit.*

5.1.3 Extension to Barycenters

Until now, we focused on the computation of transport between pairs of distributions. Dealing with more that 2 distributions can nevertheless be useful to compute barycenters of color palettes for color normalization of multiple images.

Geodesics between pairs of distributions can indeed be extended to barycenters between an arbitrary finite collection of distributions. Existence and uniqueness of this barycenter is studied in [3]. Computing the barycenter between discrete distributions requires the resolution of a linear program that corresponds to a multi-marginal optimal transportation, as proved in [3]. However, in contrast with the case of two distributions, the case of un-weighted sums of Diracs is not anymore equivalent to an assignment problem, which is known to be NP-hard [47]. Computing numerically this barycenter for large scale problems can however be tackled using a non-convex formulation solved with a Lagrangian discretization, which finds applications in image processing [180].

Let us now detail how the previous model can be extended to find a new distribution in-between two or more distributions. Given a set of $r = 1 \cdots R$ input histograms $\mu_r \in \mathbb{R}^N$ associated to clusters $Y_r \in \mathbb{R}^{N \times d}$, and weights $(\rho_r)_{r \in R} \in (\mathbb{R}^+)^R$, the regularized histogram barycenter μ on the fixed cluster X is a local minimizer of

$$\min_{\mu} \sum_{r \in R} \rho_r E(\mu, \mu_r), \tag{5.2}$$

where

$$E(\mu, \mu_r) = \min_{\substack{P_r \in \mathcal{P}_{\kappa_r}(\mu, \mu_r) \\ \kappa_r \in \mathbb{R}^N, \kappa_r \geq 0, \langle \kappa_r, \nu \rangle \geq 1}} \langle C_{XY},\, P_r \rangle + \lambda \left\| G_X (D_\mu P_r Y_r - X) \right\|_1 + \rho \| \kappa_r - \mathbf{1} \|_1. \tag{5.3}$$

The minimization of (5.2) can be performed by doing a joint minimization on both the barycenter histogram μ and the set of matrices P_r and relaxation parameters κ_r. This is a non-convex optimization problem. Fortunately, it is separately convex with respect to each of separate group of variables μ and $(P_r, \kappa_r)_{r \in R}$, so one can use the block coordinate descent algorithm (A.19) that iteratively minimizes with respect to each of these variables.

The updates of $(P_r, \kappa_r)_{r \in R}$ can be done in parallel, still using linear solvers. The update of μ is also a linear program which can be solved using for instance interior point solvers [150]. An alternative option is to use first order proximal splitting schemes [56] as algorithm (A.15), that is well tailored for such highly structured problems. The convergence of such process has been shown in [93].

Remark 2. *In the cloud of points formulation where all the mass is uniformly distributed over the clusters X and Y_r, the previous problem can be reformulated in terms of clusters positions X instead of histogram weights μ. In the case where $\lambda = 0$, $\rho \to +\infty$, one recovers the barycenter over the Wasserstein space.*

Experiments Figure 5.7 shows the results of histogram barycenter computed between images I_X and I_Y and then transferred to image I_X. The geometry of I_X is maintained

(row #1) and its histogram matches the computed barycenter distribution. Note how the colors change smoothly from $(1, 0)$ to $(0, 1)$ without generating artifacts and match the color and contrast of image I_Y for $\rho = (0, 1)$. The change in contrast is specially visible for the (b) wheat image.

Figure 5.7 – Results for the barycenter algorithm on different images. Note how as ρ approaches $(0, 1)$, the histogram of the barycenter image becomes similar to the histogram of I_Y.

As a final example, we show in Figure 5.8 how this method can be applied as a color

normalization preprocessing before comparing/registering images of the same object obtained under different illumination conditions.

Color normalization is the process of imposing the same color palette on a group of images. This color palette is always somehow related to the color palettes of the original images. For instance, if the goal is to cancel the illumination of a scene (avoid color cast), then the imposed histogram should be the histogram of the same scene illuminated with white light. Of course, in many occasions this information is not available. Following [161], an in-between histogram is chosen as the regularized Optimal Transport barycenter and then transferred to each image with the corresponding transport matrix P_r.

Finally notice that the entropic regularization of optimal transport can also be considered to compute regularized histogram barycenters in a very efficient way [199].

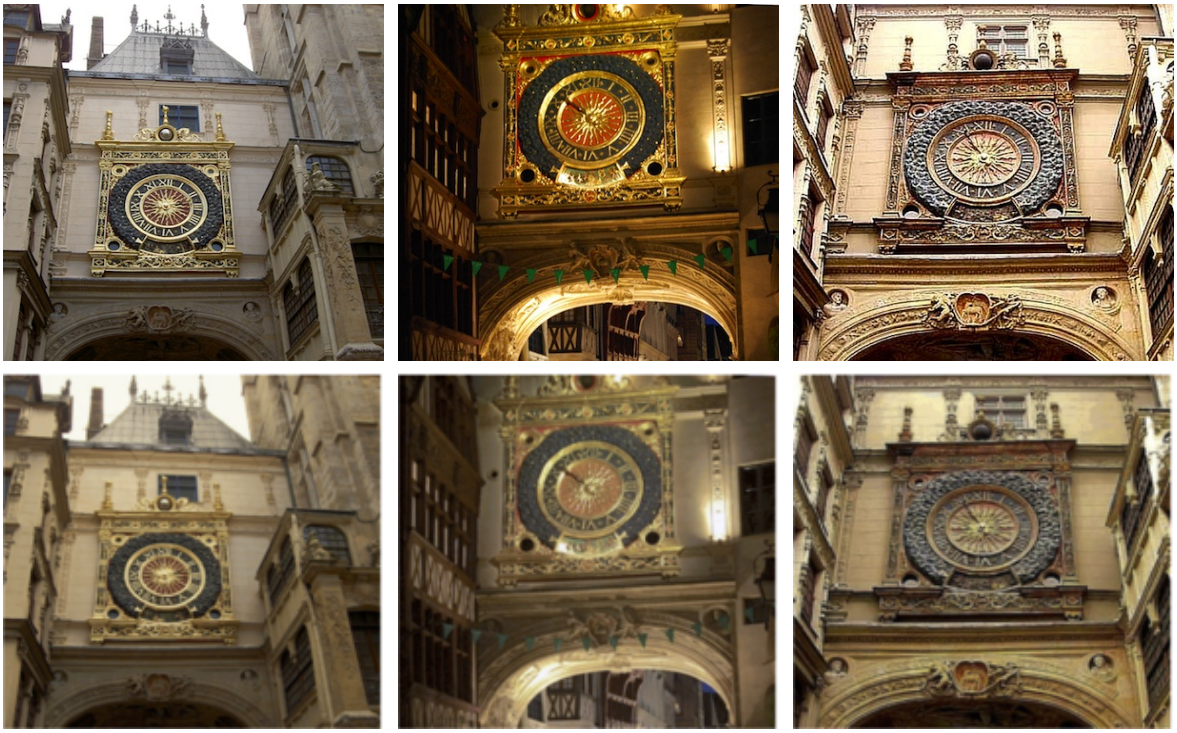


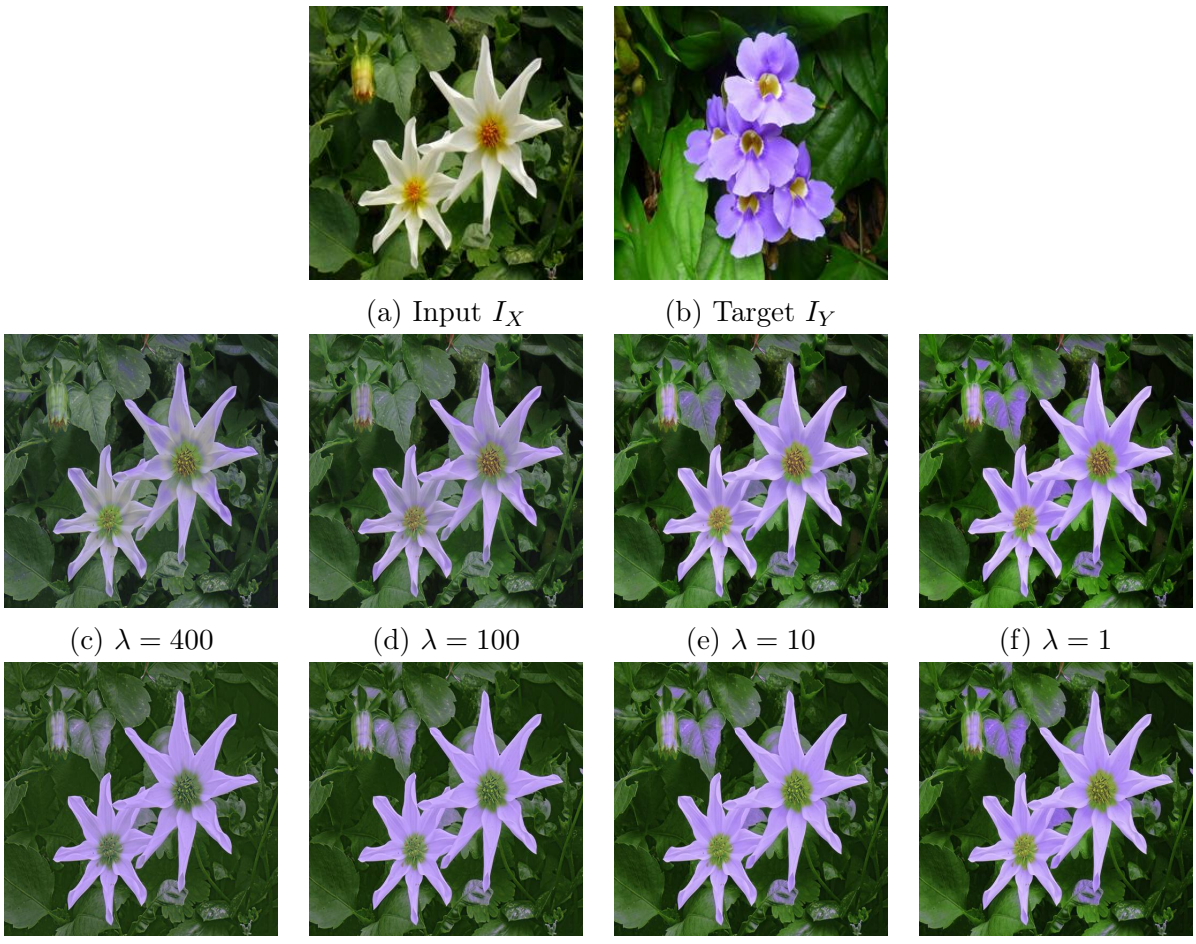
Figure 5.8 – The proposed method can be applied as a preprocessing step in a pipeline for objects detection or image registration, where canceling illumination is important. On the first row, a set of pictures of the same object taken at different hours of the day or night is displayed. On the second row, the histogram barycenter has been transfered to the 3 images. Note how the algorithm is able to normalize the illumination conditions of all the images.

5.1.4 Limitations

Due to the one-to-many relaxation and the fact that only the gradient of the *average transport* flow is penalized, the regularization does not prevent the transport map to associate very different colors to a single pixel or cluster. This leads to undesirable

results such as color mixing or color inconsistencies in the modified image. Indeed, new drab colors that do not exist in the target palette ν can be created with regularization. This is illustrated in Figure 5.9, by increasing the regularization parameter λ of model (5.1) or when using instead the entropic regularization of model (4.11).

These experiments show that a large regularization involves an important mixing between colors. Hence, the incorporation of a non-convex constraint that minimizes the variance of colors assigned to each cluster has been proposed in [178].



(a) Input I_X (b) Target I_Y

(c) $\lambda = 400$ (d) $\lambda = 100$ (e) $\lambda = 10$ (f) $\lambda = 1$

(g) $\gamma = 0.1$ (h) $\gamma = 0.05$ (i) $\gamma = 0.025$ (j) $\gamma = 0.01$

Figure 5.9 – Creation of drab colors with the regularization. The influence of the regularization of the average transport (5.1) and the entropic regularization of [73] are exhibited for color transfer. The second row illustrates the tradeoff between spatial regularity and color fidelity with decreasing regularization parameter λ in (5.1). In the last row, a similar tradeoff is obtained with the method of [73], when tuning the regularization parameter $\gamma = 1/\lambda$ in (4.11).

5.2 Non-convex relaxation of regularized Optimal Transport

By using linear programming to optimize the regularized problem (5.1), the dimension of the variables to estimate is greatly increased as one additional variable is needed for each regularization constraint. Simplifications of the regularization term (through the mean transport) is thus needed to reduce the complexity, as illustrated in Figure 4.3. As previously noticed, such regularization limits the inter-cluster color dispersion but it induces the creation of new drab colors since an important interpolation of the target color palette may occur (*i.e.* the intra-cluster color variance may be large with the one-to-many assignment). To cope with this issue, the penalization of the dispersion of assigned colors with a non-convex constraint has been introduced in [178]. With this new formulation, a different optimization tool which decreases the dimension of the problem with respect to linear programming can be considered in order to only deal with the estimation of the transport matrix without considering additional variables for the constraints as in (4.10).

However, to be able to define an algorithm that efficently deals with such non-convex penalization of the color dispersion, differentiable funtional terms have to be considered. Hence, the relaxation here presented is different from the capacity relaxation of the target histogramt proposed in previous section. The closeness to the target histogram is now imposed through a data fidelity term which makes easier the control of the results while simplifying the projection onto the set of acceptable transport matrices.

5.2.1 Dispersion penalization

In order to deal with the aforementioned limitations which are intrinsic to the optimal transport framework, the following relaxed and regularized optimal transport problem can be considered:

$$P^\star \in \underset{P \in \mathcal{P}_\mu}{\operatorname{argmin}} \langle C_{X,Y}, P \rangle + \rho F_\nu(P) + \lambda R(P) + \alpha D(P) \tag{5.4}$$

where:

- $\mathcal{P}_\mu = \{\mathbb{R}^{N \times N}, P \geq 0, P\mathbf{1}_N = \mu\}$ is the convex set of right stochastic matrices (where each row sums to the corresponding bin value in μ),
- $F_\nu(P)$ is the fidelity term w.r.t the target histogram ν,
- $R(P)$ is the regularization prior on the average transport mapping,
- $D(P)$ measures the dispersion of the average transport map.

To solve efficiently this problem with projected gradient descent, differentiable functions F, R and P are considered.

Linear Constraint The linear constraint $P \in \mathcal{P}_\mu$ may be incorporated using an indicator function $\iota_{\mathcal{P}_\mu}(P) = 0$ if $P \in \mathcal{P}_\mu$ and $+\infty$ otherwise. As the constraint on acceptable transport matrices $\mathcal{P}_\mu$ only concerns the rows, the projection onto this set

is simple. The Euclidean orthogonal projector $\text{Proj}_{\mathcal{P}_\mu}$ onto the simplex $\mathcal{P}_\mu$ can be computed in $O(nm\log(m))$ by using, for each row of the matrix, a variant of the Euclidean projector onto the probability simplex (see for instance [59]).

Fidelity term As the set of acceptable transport matrices $\mathcal{P}_\mu$ does not take into account the target distribution, an additional constraint is required to make sure that the transported histogram $P^T\mathbf{1}_N$ is close enough to the target histogram ν. To do so, we rely on the Pearson's χ^2 statistics, which writes for any bistochastic matrix $P \in \mathcal{P}_\mu$

$$F_\nu(P) \quad = \tfrac{1}{2}\chi^2_\nu(P^T\mathbf{1}_n) = \tfrac{1}{2}\left\|\mathbf{1}_n^T P D_\nu^{1/2}\right\|^2 - \tfrac{1}{2}, \tag{5.5}$$

where $D_\mu = diag(\nu^{-1})$ We assume from here, without loss of generality, that ν has non empty bins. Observe that the corresponding fidelity term can be interpreted as a weighted L^2 metric, which further penalizes bins of the target histogram that has small values. This will prevent the model from using very rare features in the exemplar image.

Regularity term A Tikhonov regularization of the gradient of a flow V can be used for incorporating spatial information from the input feature distribution in the gradient operator G_X defined on the graph of clusters $\mathcal{G}_X$. The norm of the gradient of the mean flow $V = D_{h_u}PY - X$ can therefore be measured as:

$$R(P) = \frac{1}{2}\|D_\mu G_X(D_\mu PY - X)\|_2^2, \tag{5.6}$$

where the value of the gradient norm is weighted by the corresponding histogram bin value: $({D_\mu}^{-1})_{ii} = \mu_i$.

Dispersion term As pointed out before, the regularized transfer induces a high variability of color assigned to each input color. The simple idea is thus to consider the minimization of the variance of the flow. Denoting $\overline{Y} = D_\mu PY$, where $\overline{Y}_i$ is the mean color assigned to cluster X_i, the intra-cluster variance of the color assigned to X_i is defined as $\text{Var}(Y)_i = (\overline{Y^2})_i - \overline{Y}_i^2 = \frac{1}{\mu_i}\sum_j P_{i,j}\|Y_j\|^2 - \left\|\frac{1}{\mu_i}\sum_j P_{i,j}Y_j\right\|^2$.

This variance is finally penalized with respect to each cluster weight to obtain the last functional term:

$$D(P) = \sum_i \mu_i \,\text{Var}(Y)_i.$$

Minimizing this term will promote sparser transport matrices and will prevent strong color mixing in the tackled applications.

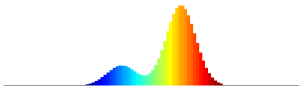
Optimization Because of this last term penalizing the variance of the flow, the objective function (5.4) is not convex in P. The inertial Forward-Backward algorithm (A.17) can be used to find a critical point of this non-convex problem [8, 156], since it contains a convex non-smooth term (the linear constraint $\iota_{\mathcal{P}_\mu}(P)$) and the sum of differentiable terms $G(P) = \langle C_{XY},\, P\rangle + \lambda R(P) + \rho F(P) + \alpha D(P)$, which all have a Lipschitz gradient on the set $\mathcal{P}_\mu$.

5.2.2 Experiments

Figure 5.10 illustrates the interest of the proposed approach to cancel the effect of transport dispersion. The first row shows input and target histograms, respectively in (a) and (b). The colormap in (a) is arbitrary, and used to illustrate the location reach with the optimal transport map in (b). Observe that the mapping is highly irregular, sending neighboring colors (orange) in two distant locations, as illustrated in (f). Figures (c) and (d), respectively proposed in [94] and [73], provides more regular mapping of *the average flow*, as illustrated in (g) and(h), but the colors are spread out to many different locations.

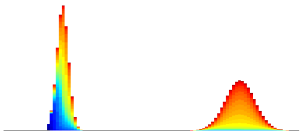
The non-convex model, illustrated in (e), relaxes the matching constraint while controlling the variance of the mapping and the χ^2 statistics with the desired output histogram. This leads to sparser transport flow (i) and almost no color mixing.

The matrices P of transport flows computed with the different methods are illustrated in Figures (f) to (i). The null values $P_{i,j}$ are displayed in white color. These matrices describe how the mass from the bins of the input histogram (in rows) are sent to the bins of the target histogram (in columns). As a bin μ_i corresponds to the line i of the matrix, the variance of this bin is larger when a lot of columns j are active (*i.e.* $P_{i,j} > 0$) and corresponds to very different features Y_j.

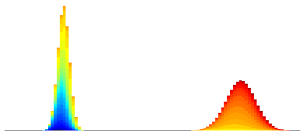


(a) Input histogram

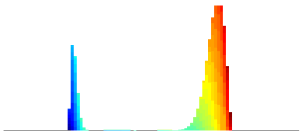
(b) Output histogram using OT



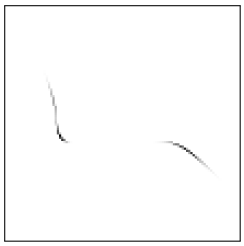
(c) Gradient-based Regularization (5.1)



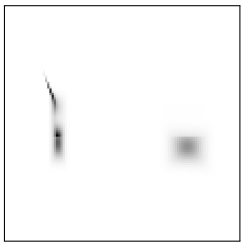
(d) Entropy-based Regularization (4.11)



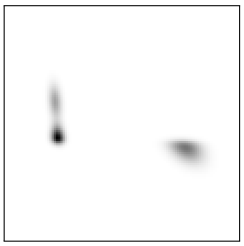
(e) Non-convex model



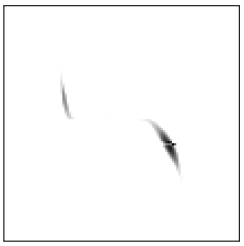
(f) OT flow



(g) Gradient regularized OT flow (5.1)



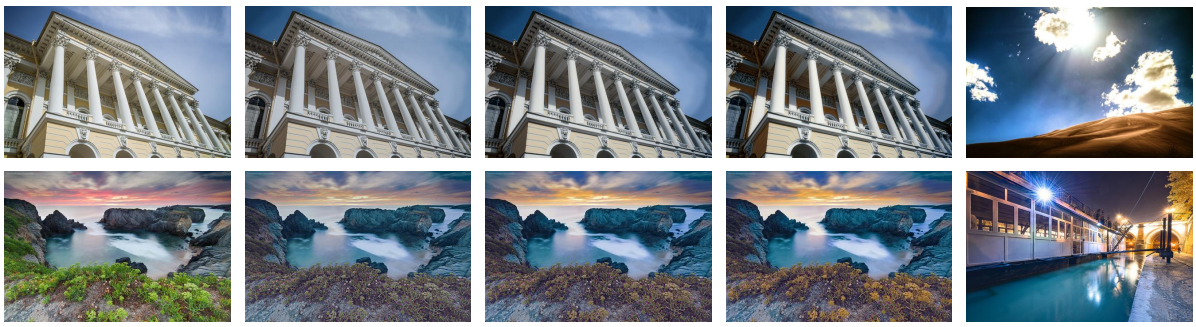
(h) Entropy regularized OT flow (4.11)



(i) Non-convex regularization (5.4)

Figure 5.10 – Illustration with 1D histograms of various models for regularized optimal transport.

As illustrated in Figure 5.11, the colors of the target palette are better recovered with increasing penalization of the color variance. When no penalization is applied to the color variance (*i.e.* $\alpha = 0$), it corresponds to the model of (5.1). By monitoring the capacity of the target histogram and regularizing the average flow [175], the synthesized images look plausible (5.11b) but they contain new drab colors (that do not exist in the target image) and they are over-smoothed. On the other hand, the transfer is far better when the color variance is penalized (with high values of α). In this case, the final images only contain the colors of the target images.



(a) Input (b) $\alpha = 0$ (c) $\alpha = 10$ (d) $\alpha = 100$ (e) Exemplar

Figure 5.11 – Color transfer. The colors of the exemplar images (e) are transfered to the input images (a).

5.3 Discussion

In the future, one interesting aspect to study is the behavior of the non-convex dispersion penalization with respect to the convex entropic regularization. The combination of non-convex terms with the entropy has for instance been proposed for the domain adaptation problem [70]. At first sight, it may seem contradictory to minimize the variance of the flow while maximizing its entropy. Nevertheless, the variance penalization corresponds to the minimization of a local entropy, on the rows of the transport matrix, whereas the method of [73] corresponds to the maximization of the global entropy of the matrix. Optimization methods dedicated to problems formulated as the difference of convex functions, known as DC programming methods [118], may thus be an interesting way to compute efficiently the related transport maps.

Finally, it is important to underline that the regularization of the average optimal transport map allows removing artifacts for color transfer purpose, but its related computational cost is prohibitive and limited in practice to histograms discretized with a few hundreds of bins. On the other hand, the entropic regularization is not designed to deal with transfer artifacts, but it leads to extremely fast computations of approximate transport distances between large scale histograms. Hence, we will see in the next chapter that such regularization is very efficient for involving transport distances in more general problems.

Chapter 6

Application to Image Segmentation

Image segmentation has been the subject of active research for more than 20 years (see *e.g.* [10, 71] and references therein). For instance, we can refer to the seminal work of [147], or to its very popular approximation with level sets developed in [58]. This last work provides a very flexible algorithm to segment an image into two homogeneous regions, each one being characterized by its mean gray level value.

In the case of textured images, a lot of extensions of [58] have been proposed to enhance the mean value image segmentation model. For instance, local histograms are used in [228, 151], Gabor filters in [213], wavelet packets in [11] and textures are characterized thanks to the structure tensor in [43, 186].

When considering the global histograms of the regions to segment, there also exist a large range of methods in the literature [9, 126, 44, 116] also based on [58]. Recent works make use of the the Wasserstein distance [165] to compare globally the histograms. It is important to notice that this class of approaches involves complex shape gradient computations [80] for the level set evolution equation. Moreover, as these methods all rely on the evolution of a level set function [158], it leads to non-convex methods that are sensitive to the initialization choice and only a local minimizer of the associated energy is computed.

Recently, convexification methods have been proposed to tackle this problem, as in [152, 53, 54, 25, 170, 56, 42, 221]. The original model [58] can indeed be convexified, and a global solution can be efficiently computed, for instance with a primal-dual algorithm. Thanks to the coarea formula, a simple thresholding of this global solution provides a global minimizer of the original non-convex problem. Such approaches have not been developed yet for global histogram segmentation with length boundary regularization.

Other models as in [185, 214, 13, 109, 224] that use graph-based methods and max-flow formulations [173], obtain good minima without level-sets. Nevertheless, these algorithms just consider simple bin-to-bin distances between histograms (for instance ℓ_1 or Bhattacharyya) which are not robust enough to deal with coarse quantification of histograms or data outliers.

The use of Optimal Transport for image segmentation has been first investigated in [151] for comparing local $1D$ histograms. In [165, 143], active contours approaches

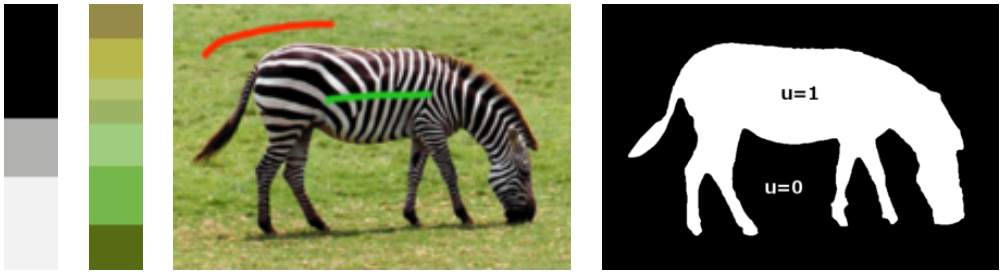
using the Wasserstein distance for comparing global multi-dimensional histograms of the region of interest have then been proposed. Again, such non-convex methods are sensitive to the initial contour. Moreover, their computational cost is too important (several minutes) even if they include some approximations of the Wasserstein distance.

In order to include more robust metrics between histograms while being independent of any initialization choice, we now describe some convex [206] and fast [222, 176, 177] segmentation models including Optimal Transport between global color histograms.

6.1 Convex histogram-based image segmentation

Let $I : x \in \Omega \mapsto I(x) \in \mathbb{R}^n$ be a color image, defined over the N-pixels domain Ω ($N = |\Omega|$), and $\mathcal{F}$ a feature-transform of n-dimensional descriptors (*i.e.* colors, gradient orientations...) such that $\mathcal{F}I(x) \in \mathbb{R}^d$. The feature space F is assumed to be clustered in M regular bins $F_j \in \mathbb{R}^d$. Hence, one can design a linear assignment operator H of size $M \times N$ such that for each pixel x, $H(x, j) = 1$ if $\mathcal{F}I(x) \in F_j$ and 0 otherwise.

The objective is to segment the image I into 2 regions, by defining a binary segmentation map $u : \Omega \mapsto \{0, 1\}$ of the whole image domain with respect to two fixed probability distributions of features μ_1 for the foreground and μ_0 for the background defined in F. This prior information is commonly obtained by scribbles given by a user in an interactive segmentation process [40]. In a local model, the distance to the scribble information is considered to perform the segmentation, using for instance local histogram comparison [30, 151] or geodesic distances [72]. As illustrated in Figure 6.1, the goal is here different as global statistics of the regions to segment are compared.



μ_1 μ_0 I u

Figure 6.1 – Segmentation problem. The objective is to find a segmentation of the image I into 2 regions with a binary map u such that the color distributions of the region defined by $u = 1$ (resp. $u = 0$) matches the prior color distribution μ_1 (resp. μ_0) obtained from the given scribbles visible in image I.

As proposed in [222], the following model is considered:

$$J(u) = D(\mu_1, \pi(u)) + D(\mu_0, \pi(\mathbf{1} - u)) + \rho\, TV(u) \tag{6.1}$$

where $TV(u) = \int |Du|$ is the total variation of u measuring the length of its level sets and $\rho \geq 0$ is the regularization parameter. The function D is a convex distance between feature distributions. Finally, $\pi(u)$ is the empirical discrete probability distribution of

features $\mathcal{F}I$ using the binary map u which writes:

$$\pi(u) : u \in \mathbb{R}^N \mapsto \frac{Hu}{|u|_1} \in \mathbb{R}^M,$$

where $|u|_1 = \sum_x |u(x)| = \langle u, \mathbf{1} \rangle = \mathbf{1}^T u$ for positive values of u denotes the size of the region defined by $u = 1$, *i.e.* the mass of the histogram Hu. The functional (6.1) is highly non-convex since π is a non linear operator, and the objective is to find a minimum over the non-convex set $\{0,1\}^N$.

In order to avoid the use of combinatorial algorithms for solving this problem, a convex relaxed model has been proposed in [222] and [206] by considering the minimization over the convex set $u \in [0,1]^N$ and the comparison of the non normalized color histograms instead of normalized distributions:

$$J(u) = D(\mu_1 \mathbf{1}^T u, Hu) + D(\mu_0 \langle \mathbf{1} - u, \mathbf{1} \rangle, H\mathbf{1}^T(\mathbf{1} - u)) + \rho TV(u). \tag{6.2}$$

The distribution μ_1 (resp. μ_0) is here normalized with respect to the mass of the histogram Hu (resp. $H(\mathbf{1} - u)$).

Once a global minimizer u^* of problem (6.2) has been computed, a threshold $\epsilon \in [0;1[$ is required to get a binary segmentation of the image into two regions $u_\epsilon(x) =$ if $u^*(x) \geq \epsilon$ and 0 otherwise. Notice that contrary to local data terms [53, 152], the thresholded solution may lead to a greater value of the functional, *i.e.* $J(u_\epsilon(x)) \geq J(u^*)$, since the coarea formula does not hold for general nonlinear functions D involving linear operators H, $\mu_1 \mathbf{1}^T$ and $\mu_0 \mathbf{1}^T$. One other solution could be to rely on integer linear programming with methods such as branch-and cut [149], in order to directly estimate a global binary solution of problem 6.2.

In the following, the use of distances D related to Optimal Transport are described.

6.2 Optimal Transport between non normalized histograms

In order to have a robust distance between the non-normalized histogram, one would like to take D as the Wasserstein distance. Nevertheless, as the computation of Optimal Transport involves prohibitive computational costs, some approximations are now described.

6.2.1 Wasserstein-ℓ_1 for $1-D$ histograms

When considering $1-D$ histograms μ and ν defined over a $d = 1$ dimensional space (for instance distributions of gray levels), the ℓ_p Wasserstein distance reads [216]:

$$W_p^p(\mu, \nu) = ||C_\mu^{-1} - C_\nu^{-1}||_p,$$

where C_μ is the cumulative distribution function of μ and C_μ^{-1} its pesudo-inverse. In the case $p = 1$, the disance can be simplified as:

$$W_1^1(\mu, \nu) = ||C_\mu - C_\nu||_1.$$

As noticed in [222], by defining the $M \times M$ cumulative matrix $C(i, j) = 1$ if $i \leq j$ and 0 otherwise, the Wasserstein- ℓ_1 distance between histograms can be included in the model (6.2) by taking:

$$D(\mu, \nu) = ||C(\mu - \nu)||_1.$$

Proximal algorithm such as the Primal-Dual one (A.15) can be thus used to compute a global minimizer u^* of the associated non-smooth convex problem. An illustration of the application of the method is given in Figure 6.2. Notice that similar results can be obtained with graph approaches [185] when ℓ_1 norm between $1D$ histograms are involved.

(a) Initial image (b) Segmentation

Figure 6.2 – Segmentation of a natural image. The input histogram prior μ_1 (resp. μ_0) is given by the white (resp. green) area. The final contour is obtained by thresholding the solution with $\epsilon = 0.5$.

6.2.2 Approximate MK distance for general histograms

For general feature space F of dimension $d > 1$, the previous reformulation in term of cumulative histograms does not hold. Hence, we would like to consider D as the Monge-Kantorovitch distance between the histograms $\mu, \nu \in \mathbb{R}^M$:

$$\mathbf{MK}(\mu, \nu) := \min_{P \in \mathcal{S}(\mu,\nu)} \left\{ \langle P, C \rangle = \sum_{i=1}^{M} \sum_{j=1}^{M} P_{i,j} C_{i,j} \right\}. \tag{6.3}$$

where the sets of admissible transport matrices is

$$\mathcal{S}(\mu, \nu) := \{P \in \mathbb{R}_+^{M \times M}, P\mathbf{1}_M = \mu \text{ and } P^T \mathbf{1}_M = \nu\}. \tag{6.4}$$

and the $M \times M$ matrix C is defined as $C_{i,j} = ||F_i - F_j||^2$, the distance between bins F_i and F_j.

The **MK** distance can naturally be incorporated into the model (6.2), as proposed in [206]. However, as the **MK** distance is not smooth, when solving the associated

problem with a proximal algorithm, one requires, at each iteration of the algorithm, the computation of the proximal operators of $\mathbf{MK}(\mu_1 \mathbf{1}^T u, Hu)$ and $\mathbf{MK}(\mu_1 \mathbf{1}^T(\mathbf{1}-u), H(\mathbf{1}-u))$:

$$\mathrm{Prox}_{\mathbf{MK}(\mu_1 \mathbf{1}^T \cdot, H\cdot)}(v) = \underset{u}{\mathrm{argmin}} \frac{1}{2}||u-v||^2 + \mathbf{MK}(\mu_1 \mathbf{1}^T u, Hu),$$

which is a quadratic programming problem that involves inner loops in the process. It has been suggested in [176] to rely on the entropic regularization of Optimal Transport by extending the works of [76]. The entropy-regularized Optimal Transport problem is now reformulated as the Sinkhorn distance:

$$\mathbf{MK}_{\lambda,\leq N}(\mu,\nu) := \min_{P \in \mathcal{S}(\mu,\nu)} \left\{ \langle P,\, C\rangle - \tfrac{N}{\lambda} h(P/N) \right\} + \iota_{\mathcal{H}_{\leq \mathcal{N}}}(\mu,\nu), \tag{6.5}$$

where the entropy of the matrix P is defined as $h(P) := -\langle P, \log P\rangle$ and the set $\mathcal{H}_{\leq \mathcal{N}}$ of admissible histograms prescribes the masses of histograms, that are bounded above by N, the number of pixels of the image domain Ω:

$$\mathcal{H}_{\leq N} := \left\{ \mu \in \mathbb{R}^M, \nu \in \mathbb{R}^M \,\middle|\, \mu \geq 0, \nu \geq 0, \langle \mu,\, \mathbf{1}\rangle = \langle \nu,\, \mathbf{1}\rangle \leq N \right\}. \tag{6.6}$$

By introducing the matrix-valued function $Q_\lambda(\alpha,\beta)_{i,j} = e^{\lambda(\alpha_i + \beta_j - C_{i,j}) - 1}$ defined for $\alpha, \beta \in \mathbb{R}^M$, one can show that the Legendre transform of $\mathbf{MK}_{\lambda,\leq N}$ reads:

$$\mathbf{MK}^*_{\lambda,\leq N}(\alpha,\beta) = \begin{cases} \frac{N}{\lambda}\langle Q_\lambda(\alpha,\beta), \mathbf{1}\rangle & \text{if } \langle Q_\lambda(\alpha,\beta), \mathbf{1}\rangle \leq 1 \\ \frac{N}{\lambda}\log\langle Q_\lambda(\alpha,\beta), \mathbf{1}\rangle + \frac{N}{\lambda} & \text{if } \langle Q_\lambda(\alpha,\beta), \mathbf{1}\rangle \geq 1. \end{cases} \tag{6.7}$$

The constraint $\iota_{\mathcal{H}_{\leq \mathcal{N}}}$ is thus absorbed in this dual formulation. A last interesting point to notice is that this dual function is differentiable and its gradient can be explicitly written as:

$$\nabla \mathbf{MK}^*_{\lambda,\leq N}(\alpha,\beta) = \begin{cases} N & (Q_\lambda(\alpha,\beta)\mathbf{1}, Q_\lambda(\alpha,\beta)\mathbf{1}) & \text{if } \langle Q_\lambda(\alpha,\beta), \mathbf{1}\rangle \leq 1 \\ \frac{N}{\langle Q_\lambda(\alpha,\beta), \mathbf{1}\rangle} & \left(Q_\lambda(\alpha,\beta)\mathbf{1}, Q_\lambda(\alpha,\beta)^T\mathbf{1}\right) & \text{if } \langle Q_\lambda(\alpha,\beta), \mathbf{1}\rangle \geq 1 \end{cases}. \tag{6.8}$$

Moreover, it can be shown that the gradient $\nabla \mathbf{MK}^*_{\lambda,\leq N}$ is a Lipschitz continuous function of constant bounded by $2\lambda N$.

We emphasis here that we retrieve a similar expression to the one originatively demonstrated in [76], where the authors consider the Sinkhorn distance on the probability simplex (*i.e.* the special case where $N = 1$ and $\langle Q, \mathbf{1}\rangle = 1$).

Optimization The general final problem we want to solve can be expressed as:

$$\min_u \mathbf{MK}_{\lambda,\leq N}\big(\mu_1 \mathbf{1}^T u, Hu\big) + \mathbf{MK}_{\lambda,\leq N}\big(\mu_0 \mathbf{1}^T(\mathbf{1}-u), H(\mathbf{1}-u)\big) + \rho\, TV(u) + \iota_{[0,1]^N}(u). \tag{6.9}$$

Using the Legendre–Fenchel transform, the problem (6.9) can be reformulated as:

$$\min_u \max_{\substack{\alpha_1,\beta_1 \\ \alpha_0,\beta_0,z}} \langle \mu_1 \mathbf{1}^T u, \alpha_1\rangle + \langle Hu, \beta_1\rangle + \langle \mu_0 \mathbf{1}^T(\mathbf{1}-u), \alpha_0\rangle + \langle H(\mathbf{1}-u), \beta_0\rangle + \langle \nabla u, z\rangle$$
$$+ \iota_{[0,1]^N}(u) - \mathbf{MK}^*_{\lambda,\leq N}(\alpha_1,\beta_1) - \mathbf{MK}^*_{\lambda,\leq N}(\alpha_0,\beta_0) - \iota_{\|\cdot\|\leq\rho}(z),$$

where the dual variables α_i, β_i for the **MK** term are in $\mathbb{R}^N$ and the dual variable z for the TV term is in $\mathbb{R}^{2N}$.

This problem can be efficiently optimized with the Preconditionned formulation [137] of the Primal-Dual algorithm (A.15) proposed in [57] that combines implicit schemes with the proximal operators of $\iota_{[0,1]^N}(u)$ and $\iota_{\|.\|\leq\rho}(z)$ and explicit schemes with the gradient of the other terms. Notice that the process does not require the computation of Optimal Transport and only involves variables of dimensions scaling with N or M.

However, the time steps of this algorithm are bounded by the inverse of the Lipschitz constant of the differentiable terms which is proportional to $\nabla\mathbf{MK}^*_{\lambda,\leq N} = 2\lambda N$. It thus involves a slow explicit gradient ascent in the dual update (A.15) for accurate approximations of the **MK** distance (*i.e.* large values of λ). In practice, good results are nevertheless obtained in a few hundred iterations, for relatively high values of λ.

Remark 3. *As noticed in [176], one can resort to the bidualization of* $\boldsymbol{MK}$ *or* $\boldsymbol{MK}^*_{\lambda,\leq N}$ *in order to remove the dependency of time steps to* λ *and have algorithms without inner loops. This nevertheless induces a new dual variable which dimension scales with* M^2*, the dimension of the transport matrix* P*.*

6.2.3 Experiments

In this experimental section, exemplar regions are defined by the user with scribbles. These regions are only used to built prior histograms, so erroneous labeling is tolerated. Histograms μ_1 and μ_0 are built using hard-assignment on $M = 8^d$ clusters. We use either RGB color features ($\mathcal{F} = \mathrm{Id}$ and $n = d = 3$) or the gradient norm of color features ($\mathcal{F} = \|\nabla.\|$ and again $n = d = 3$). The cost matrix is defined from the Euclidean metric $\|\cdot\|$ in $\mathbb{R}^d$ space. Final regions are obtained with threshold $\epsilon = 0.5$. A few seconds are required to run 100 iterations and segment a 1 Megapixel color image.

Comparison with a local histogram-based approach In order to illustrate the importance of global histogram comparisons over local ones, a local histogram segmentation model is introduced. It looks at the local probability of a pixel to belong to a class (foreground or background). This can be formulated with the functional:

$$J_{local}(u) = TV(u) + \sum_x u(x) \left(\| h_0 - h_{V(x)} \|_1 - \| h_1 - h_{V(x)} \|_1 \right), \tag{6.10}$$

where $h_{V(x)}$ is the histogram estimated on a neighborhood $V(x)$ of $x \in \Omega$. Notice that the data term can be computed pointwise (once the map of distances with local histograms has been calculated for the local model). This is a main difference with the previous data term which values depend on the whole state u. On the other hand, it is important to underline that such functionals can be globally minimized on $\{0;1\}^N$, using graph methods [40] or convexification approaches [53, 152, 25, 56].

In Figure 6.3, we can see that the reference color distributions are here not homogeneous in the sense that the histogram of the whole region is not similar to the ones computed in local neighborhoods. Here the orange colors are more probable in the region related to the butterfly. With the local model, the flowers are thus classified as the butterfly, and the darker regions are segmented as being in the background. This

example illustrates the importance of global histogram comparisons to get a global segmentation of an image. Indeed, the convex model (c) is able to recover the butterfly, whereas the local approach (b) completely fails.

(a) Original image (b) Local model (c) Global model

Figure 6.3 – Segmentation of an image where local histograms (9×9 neighborhoods) are different from the global ones. The user defines scribbles which indicate the object to be segmented (here in white) and the background to be discarded (in green). The segmentation fails for the local histogram model as it classifies the orange areas in the first class and the darker ones in the second class.

Comparison with a shape gradient approach The advantage of the convex model that does not depend on the initialization is finally illustrated through a comparison with the Wasserstein Active Contour method proposed in [165].This approach consists in deforming a level set function in order to minimize globally the Wasserstein distance between the reference histograms and the one induced by the segmentation. To make the level set evolve, this formulation requires complex shape gradients computations. As shown in Figure 6.4, even if this model can give good segmentations that are close to the one obtained in Figure 6.3 (c), its initialization may be a critical step as really different segmentations are obtained with very similar initializations.

(a) (b) (c) (d)

Figure 6.4 – The Wasserstein active contours method [165] has been initialized in two different ways (a-b), the corresponding segmentations being presented in (c-d). When carefully initialized, it leads to a segmentation close to the one obtained with the convex approach (see Figure 6.3 (c)).

General results Figure 6.5 shows the influence of the threshold ϵ used to get a binary segmentation. A comparison with the bin-to-bin ℓ_1 distance between color histograms is then given in Figure 6.6. This underlines the robustness of Optimal Transport distance with respect to bin-to-bin distance. Contrary to Optimal Transport, when a color is not present in the reference histograms, the ℓ^1 distance does not take into account the color distance between bins which can lead to incorrect segmentation.

This robustness is further illustrated in Figure 6.7. It is indeed possible to use a prior histogram from a different image, even with a different clustering of the feature space.

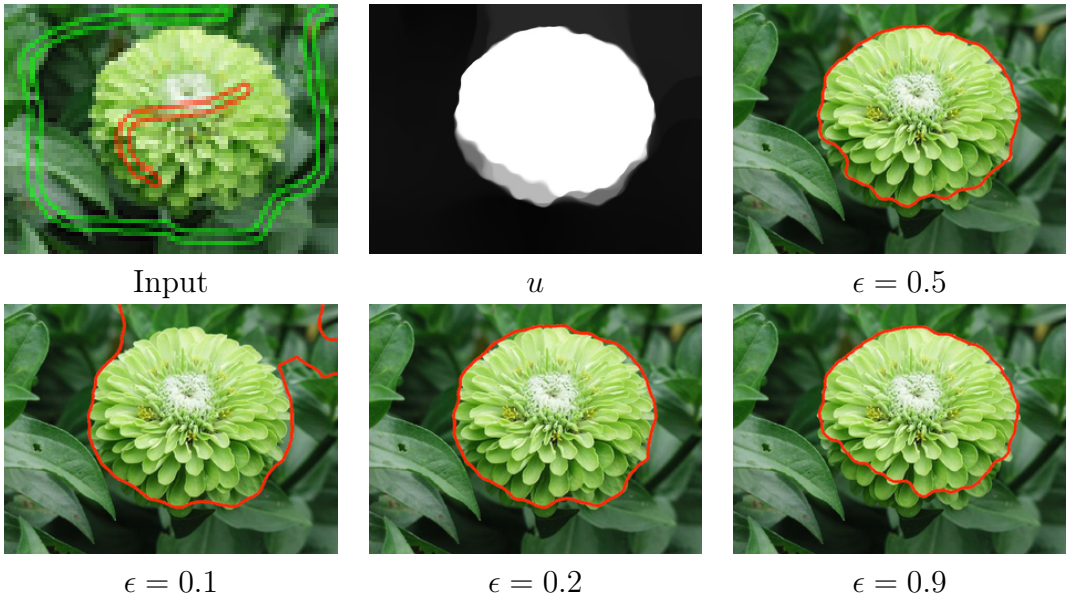


Figure 6.5 – Illustration of the influence of the threshold parameter ϵ on the segmentation result. The output image u is a regularized weight map that gives the probability of a pixel to belong to the object. This probability map u is finally thresholded with a parameter ϵ to segment the input image into a region, which contour is displayed in red. The value $\epsilon = 0.5$ is used by default, but other strategies may be defined, such as selecting the threshold value that minimizes the non-relaxed energy (6.1).

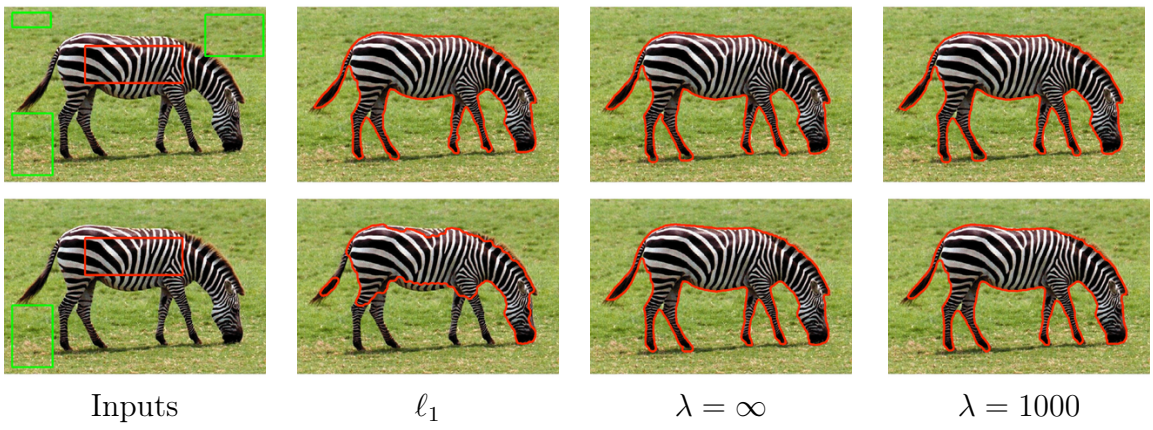
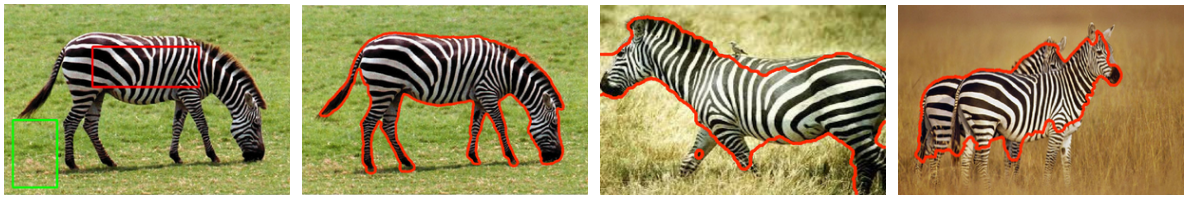


Figure 6.6 – Comparison of the segmentation results obtained from the segmentation models using $\mathbf{MK}_\lambda$ distances, together with the bin-to-bin ℓ_1 distance. The same regularization parameter ρ is used for every segmentations. Note that the Optimal Transport similarity measure is a more robust statistical metric between histograms than ℓ_1.

Note that it is not possible with a bin-to-bin metric, which requires the same clustering. Figure 6.8 shows comparisons between the non-regularized model, quite fast but high dimensional model, with the regularized model, using a low dimensional formulation. One can see that setting a large value of λ gives interesting results. On the other hand, using a very small value of λ always yields poor segmentation results.

Some last examples on texture segmentation are presented in Figure 6.9 where the



Input histograms Image 1 Image 2 Image 3

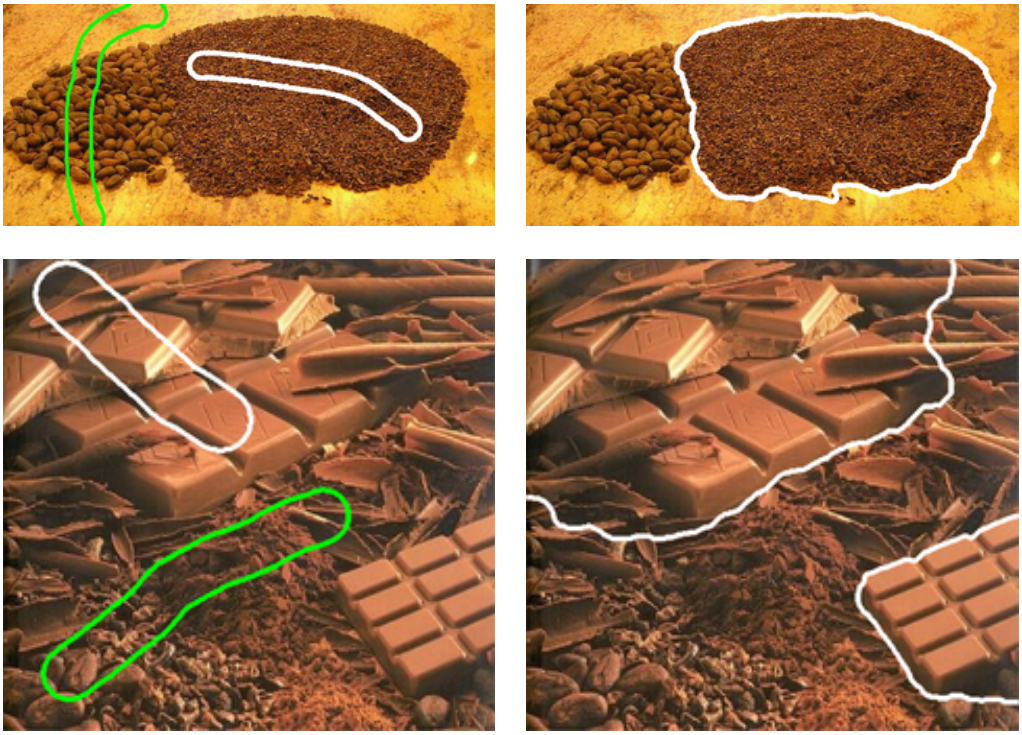
Figure 6.7 – Illustration of the interest of Optimal Transport for the comparison of histograms. Its robustness makes it possible to use prior histograms from different images (here histograms are taken from image 1 and used to segment images 2 and 3), even with a different clustering of feature space. Note that it is not possible with bin-to-bin metric, which requires the same clustering.

Input $\lambda = \infty$ $\lambda = 100$ $\lambda = 10$

Figure 6.8 – Comparison of segmentations obtained from the proposed models. The input areas are used to compute the reference color distributions μ_1 and μ_0. The non-regularized model corresponds to $\lambda = +\infty$. The influence of entropy regularization of Optimal Transport is shown with decreasing parameters λ.

proposed method is perfectly able to recover the textured areas. We consider in this example the joint histograms of gradient norms on the 3 color channels. Note that the complexity of the algorithm is the same as for color features, as long as we use the same number of clusters to quantize the feature space.

The natural extension of the presented 2 region-based model is the definition of a convex multi-phase segmentation model using Optimal Transport, as done for image restoration in [205]. In this case, the problem of thresholding the final solution to obtain a labeling map may have to be studied more in details. Indeed, even in the case of local



Input $\lambda = \infty$

Figure 6.9 – Texture segmentation using joint histograms of color gradient norms.

data terms, the final binarization of the solution of the multi-phase problem does not ensure to recover a global minima of the original non-convex energy [170].

6.3 Co-segmentation

As proposed in [206], the previous framework can be extended to deal with the unsupervised co-segmentation of multiple images. The problem of co-segmentation [215] consists in segmenting simultaneously multiple images that contain the same object of interest without any prior information. In [177], the regularized Optimal Transport distance has been considered to solve this problem. Let us consider K images I^k defined over domains Ω^k, and H^k the associated assignment operators that match each pixel of images I^k to a bin of the feature space.

The co-segmentation functional defined over the segmentation maps $u = (u_k)$ reads:

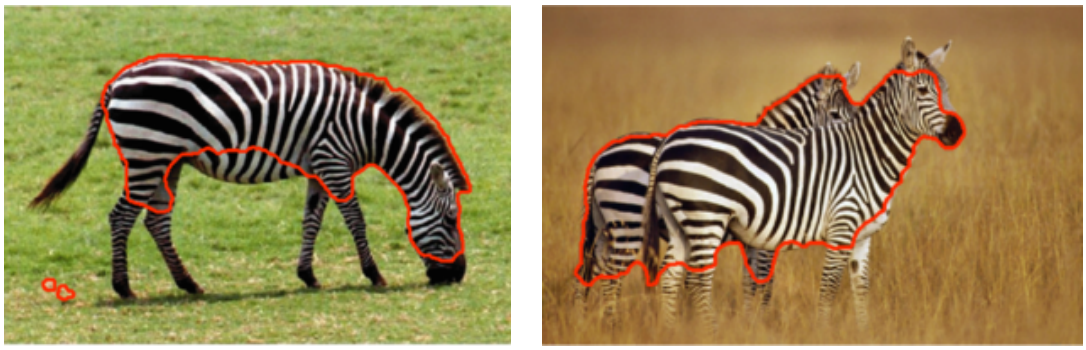
$$J(u) := \iota_{\mathcal{H}(u)} + \sum_{k=1}^{K} \left(\sum_{l>k} \mathbf{MK}_\lambda \left(H^k u^k, H^l u^l \right) + TV(u^k) - \alpha \left\| u^k \right\|_1 \right) \tag{6.11}$$

where $-\alpha \left\| u^k \right\|_1$ is a balloon force that prevents from estimating empty areas. In this model, the histograms of regions defined by u^k on each image I^k are all compared with the $\mathbf{MK}_\lambda$ distance. The common mass of all histograms is ensure through the convex

set of admissible maps:

$$\mathcal{H}(u) = \{u = (u_k), \text{ s.t. } u_k \geq 0, \sum_{\Omega^k} u^k = constant\}.$$

As illustrated in Figure 6.10, this model does not require any prior and automatically find regions in the different images with common color histograms. One main limitation is that the area of the two regions has to be the same through the constraint $\iota_{\mathcal{H}}$, which makes the method not invariant to different scales of the same object. The introduction of unbalanced transport distances [61] able to compare histograms of different masses is thus the next step to define a more flexible model.



u^1 u^2

Figure 6.10 – Example of co-segmentation of two images.

Finally notice that the term of functional (6.11) that compares all pairs of histogram with the $\mathbf{MK}_\lambda$ distance is linked to the problem of estimating the regularized barycenter between K histograms μ^k:

$$\min_{\mu} \sum_{k} \mathbf{MK}_\lambda(\mu^k, \mu),$$

which is another interesting aspect to study since regularized barycenters can be computed efficiently with a parallel implementation [199].

Conclusion of Part II

In this second part, regularized discrete Optimal Transport distances have been presented and applied to color transfer and image segmentation. The aim of the presented works is to enhance the modeling of these image procesing applications using Optimal Transport, while defining fast and unsupervised methods able to automatically process given images.

With the integration of regularization of the transport map, it is possible to design relaxed model able to adapt themselves to heterogeneous color histograms for color transfer purposes. Spatio-color clustering is an important element to decrease the dimension of the Optimal Transport problem, while facilitating the incorporation of spatial consistency in the image domain, and preventing from additional post-processing. The penalization of the color dispersion has been considered with a non-convex model that encourages one-to-one bin associations. Contrary to average transport or entropic regularization of Optimal Transport, it prevents from large mixing of colors contained in the palette of the target image. Notice that the presented methods can be pushed forward to consider other features such that color, gray or gradient patches around each pixel, in order to transfer texture, grain, tone or stylish attributes [91, 117, 164, 219, 99].

It is important to underline that there exist semi-automatic color transfer methods using high level segmentation and/or semantic analysis, while relying on standard Optimal Transport [34, 100]. They give fast and accurate color transfer results even for videos. However, enhancing the Optimal Transport framework is an important step to make these semi-automatic methods more robust to color outliers.

For the segmentation application, convex methods using user's information and comparing global color distributions have been proposed. The unsupervised automatic cosegmentation of multiple images is a challenge that has been partially addressed. Graph based segmentation approaches can not deal directly with Optimal Transport distances without involving algorithms with inner loops or graph of gigantic size [200]. In this context, the dual formulation of the Monge-Kantorivitsch problem is a very relevant solution to propose fast algorithms that directly include the minimization of the transport distances without computing explicitly the transport. Indeed, unavoidable constraints (as the unknown size of the region to segment) can be absorbed within the dual formulation. Moreover, the entropy regularization provides explicit expressions of the gradient of Monge-Kantorovitch distances that are useful for the minimization of general variational problems involving Optimal Transport terms.

General Conclusions and Perspectives

As shown all along the document, generalized Optimal Transport distances including relaxation and regularization can be considered to improve the accuracy performances of image processing applications. New models and algorithms for the continuous and discrete formulations of Optimal Transport have been presented.

Dynamic Optimal Transport With the fluid dynamic setting, one can compute geodesics between continuous and vanishing densities. The integration of physical regularization of the transport plan is explicit with this formulation and it has made possible the interpolation of ocean images containing complex structures. These works open new possibilities to integrate Optimal Transport distances in numerical weather prediction systems.

Discrete Optimal Transport In the discrete setting, the regularization of the mean transport plan has been considered for application to color transfer between images. Convex and non-convex models have been proposed in this context in order to define state-of-the-art automatic methods that adapt the proportion of colors contained in the given palette. These methods have also been extended to the computation of barycenters to deal with color normalization of multiple images.

Finally, the entropy regularization of discrete Optimal Transport has been used for image segmentation. A fast and convex model has been designed to segment images, while respecting global color constraints. Whereas it does not reach yet the computational time of global graph-based segmentation approaches, this model is more robust to color histogram quantization thanks to the use of the Optimal Transport distance.

Pros and cons of both formulations With the dynamical formulation, high dimensional continuous and non-negative images can be interpolated and the regularization of the transport map can be done explicitly. Nevertheless, the convergence of this process is slow. On the other hand, the discrete setting can deal with any kind of distributions, but it is limited to small dimensions unless considering the entropic regularization. However, even if such process gives good and fast estimation of the transport cost, it requires positive densities and the transport map is in practice too smooth to be directly used for interpolation purposes.

Future works Problems involving Optimal Transport distances have the potential to impact a very wide range of applications: data assimilation in geophysics [97], sound processing, unsupervised clustering through the computation of barycenters [75], statistical analysis of population estimators [29], computer graphics [35, 199].

Hence, there is a need to provide regularized Optimal Transport distances that can be efficiently minimized using standard optimization algorithms. The recent development of proximal algorithms for non-convex problems (see Appendix A) offers new perspectives for integrating generalized transport distances in a larger range of applications. A specific work is also required to adapt generic models to the targeted applications. Hence, the main perspectives of the presented works will concern the following points.

Geoscience For assimilating satellite images in geosciences, the first direction to investigate is the integration of mass relaxation constraint [166]. There is indeed no reason for an image to have the same mass than the corresponding physical variable in a numerical model. This problem, known as unbalanced Optimal Transport [136] is crucial for applying the dynamical Optimal Transport framework to real applications, as noticed in [60] for biological images.

The main limitation of this dynamical approach is the related computational cost. However, when considering adequate basis for decomposing the density and the velocity, faster algorithms can be designed [115]. The integration of the entropic regularization could be a solution to propose accelerated methods but it should include the modeling of advanced physical priors that is necessary to preserve the topology of the structures contained in the images [139].

All these issues are fundamental problems to solve for integrating Optimal Transport in operational data assimilation systems for numerical weather prediction.

Image Processing Other direct perspectives of the works presented in this manuscript are the development of discrete models for transferring general features, multi-phase segmentation models, unbalanced models [103, 61] for co-segmentation or the integration of additional priors such as the shape of the objects [191, 192].

The interest of Optimal Transport is nevertheless not limited to image segmentation or color transfer and its relation with image denoising or cartoon decomposition for instance has already been shown [131]. A yet not studied problem concerns its use in non-local models. The introduction of Optimal Transport for defining regular non-local weight distributions between adjacent pixels is for instance an option to enhance the construction of nonlocal graphs.

Computer Vision and Machine Learning Various works on color image retrieval [162] or artistic image indexation [121] have shown on small datasets that the Optimal Transport framework is able to increase retrieval performances. Spatial information is typically only considered to enhance the feature description [196] but not used in the metric itself. Partial transport models, that relax the constraint of transporting the whole histograms mass, are nevertheless still required to increase the robustness to data outliers.

The development of approximate Monge-Kantorovitch distances with the entropic regularization has made possible the real-time comparison of histograms quantized with thousands of bins. In order to transfer such tools for data retrieval, a research effort is still required. It involves the learning of the ground cost [74] inside the transport distance as well as efficient implementation, to reach the performances of local descriptors and more simple metrics in optimized frameworks [123].

Data analysis and Clustering The definition and computation of barycenters relative to Optimal Transport distance is a major issue for proposing clustering algorithms adapted to the analysis of structured data. The notion of mean and variance with respect to the Optimal transport distance is then necessary. Recent studies on Wasserstein barycenters [3] have led to practical algorithms both for standard [50] or regularized barycenters [93] [200].

An interesting direction of research is the definition of geodesic Principal Component Analysis in the Wasserstein space [28]. First algorithmic works [194] have been developed with applications to color transfer, but this subject deserves to be explored more in depth to propose new clustering algorithms.

Audio Processing The transfer of Optimal Transport tools into the audio community is an open subject of research. As for image processing, the transport distance could define a robust dissimilarity measure and the transport map could be used for interpolation purposes (voice, prosody, music, etc).

A fundamental difference with images is that the audio signal is described with phase and modulus, and can not be represented by non negative densities to apply the Optimal Transport framework. Computing transport distances between distributions of features extracted from audio signal is nevertheless possible. Dictionary learning models could also be considered [76] to separate different sources presents in a signal. For audio interpolation, the use of the Optimal Transport map is more challenging and dedicated tools should be defined. Works on phase retrieval from signal modulus [140] are thus an interesting direction to handle such issue.

In conclusion This document presents contributions in the field of Optimal Transport for image processing that have been realized thanks to the help of a lot of colleagues that I would like to thank: Jean-François Aujol, Afaf Bouharguane, Vincent Chabot, Sira Ferradans, Romain Hug, Emmanuel Maitre, Quentin Mérigot, Maëlle Nodet, Edouard Oudet, Gabriel Peyré, Julien Rabin, Arthur Vidard and Romain Yıldızoğlu.

I acknowledge Antonin Chambolle, Patrick Pérez and Christoph Schnörr for accepting to review this thesis as well as Jean-François Aujol, Jérémie Bigot, Julie Delon, Angelo Iollo and Étienne Mémin for agreeing to be committee members.

I also thank Pablo Arias, Benjamin Berkels, Eric Blayo, Adrien Boussicault, Pierre-Antoine Bouttier, Aurélie Bugeau, Christoph Brune, Elsa Cazelles, Thierry Colin, Pierrick Coupé, MArco Cuturi, Laurent Debreu, Charles Deledalle, Jérémie Demange, Arnaud Dessein, Charles Dossal, Baudouin Denis de Senneville, Vincent Duval, Jalal Fadili, Nelson Feyeux, Lucile Gaultier, Guy Gilboa, Rémi Giraud, Morgane Henry, François-Xavier Le Dimet, Florian Lemarié, Alexandros Makris, Simon Masnou, Marc Nicodème, Mila Nikolova, Fabien Pierre, Anne Pierson, Antoine Roler, Jean-Luc Rouas, Joseph Salmon, Olivier Saut, Vivien, Seguy, Camille Sutour, Samuel Vaiter, Jacques Verron, François-Xavier Vialard, Vinh-Tong Ta and Cornel Zachiu for exciting (and also scientific) help, advices and discussions in parallel of the works presented here.

My special thoughts go to Vicent Caselles who, among many other things, introduced me to this fascinating topic of research.

Appendix A

Non-smooth Optimisation

We consider functionals that can be written as the sum of different terms being non-smooth or non-convex, which is a common setup for problems in image processing. The objectif of this chapter is to review classical proximal splitting algorithms dedicated to the estimation of global or local minima of such functionals. Proximal splitting schemes are indeed first order optimization methods that present good performances for large scale problems, as the ones that will be considered in Parts II and III of this document. This chapter is inspired from [14, 87] in which rigorous proofs can be found.

A.1 Definitions

In the following, E is an Euclidean vectorial space of finite dimension equipped with an inner product $\langle \cdot, \cdot \rangle$ and a norm $||.|| = \langle \cdot, \cdot \rangle^{\frac{1}{2}}$.

Definition 1 (Domain). *Let f be a function defined from E to $\bar{\mathbb{R}} = \mathbb{R} \cup +\infty$, we denote the domain of f as $dom(f) = \{x \in E \text{ such that } f(x) \neq +\infty\}$.*

Definition 2 (Convexity). *Let f be a function defined from E to $\bar{\mathbb{R}}$, f is convex iff for all pairs $(x, y) \in E \times E$ and $\forall \lambda \in [0, 1]$, $f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda) f(y)$. If the inequality is strict, then f is strictly convex.*

Definition 3 (Coercivity). *A function f defined on E is coercive if $\lim_{\|x\| \to +\infty} f(x) = +\infty$.*

Definition 4 (Lower semicontinuity). *A function f defined from E to $\mathbb{R} \cup +\infty$ is lower semicontinuous (l.s.c) if, $\forall x \in E$, $\liminf_{y \to x} f(y) \geqslant f(x)$.*

Definition 5 (Proper). *A function f from E to $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ is proper if $\exists x \in E$ such that $f(x) < +\infty$ and $f(x) > -\infty$, $\forall x \in E$.*

Proposition 2. *Let J be a convex, proper, l.s.c and coercive function defined on E with values in $\bar{\mathbb{R}}$, then J is bounded from below and admits at least one minimizer. If J is strictly convex then the minimizer is unique.*

In the following, we will consider smooth and non-smooth functions. We will refer to a smooth function as soon as it can be differentiated once $\forall x \in E$, *i.e.* $f \in \mathcal{C}^k(E, \bar{\mathbb{R}})$, for $k \geq 1$. A smooth functional is thus differentiable and its gradient will be denoted as ∇f. Hence, a function f is called **non-smooth** if is at most in $\mathcal{C}^0(E, \bar{\mathbb{R}})$.

Definition 6 (Lipschitz)**.** *A function f defined from E to $\mathbb{R} \cup +\infty$, is L-Lipschitz is there exists $L > 0$ such that $\forall (x, y) \in E \times E$:*

$$||f(x) - f(y)|| \leq L||x - y||.$$

if $L < 1$ then f is said firmly non-expansive.

In order to minimize a convex and differentiable function that admits minimizers, gradient descent algorithms are standard optimisation tools. Second-order methods such as the Newton's one, can be considered to algorithms with faster convergence rates. Before detailing the non-smooth context, we finally recall the explicit first order gradient descent algorithm.

Theorem 2 (Gradient descent Algorithm)**.** *Let f be a convex, proper, l.s.c and differentiable function defined from E to $\mathbb{R} \cup +\infty$ with a L-Lipschitz gradient. Let τ be a real parameter checking $\tau < \frac{1}{L}$, so that the operator $Id - \tau \nabla f$ is firmly non-expansive. Let $x_0 \in E$, then the sequence $(x_n)_{n \in \mathbb{N}}$ defined $\forall n \in \mathbb{N}$ as*

$$x_{n+1} = (Id - \tau \nabla f)(x_n), \tag{A.1}$$

converges to a minimizer of f.

A.1.1 Subdifferential

In order to consider the minimization of non-smooth functionals, we now recall standard results.

Definition 7 (Subdifferential)**.** *Let f be a convex function defined from E to $\mathbb{R} \cup +\infty$, the subdifferential of f at point $x \in E$ is the multivalued operator that associates to x the set:*

$$\partial f(x) = \{u \in E \text{ such that } \forall y \in E,\ \langle y - x, u \rangle + f(x) \leqslant f(y)\}.$$

The subdifferential of a function at a point x can be empty or it can be a convex set. If f is convex et differentiable $\forall x \in E$, the subdifferential is reduced to the singleton $\{\nabla f(x)\}$, the gradient of f at point x.

The minimizers of non-smooth convex functions can be characterized with the subdifferential.

Theorem 3. *Let f be a convex, l.s.c and proper function defined from E to $\mathbb{R} \cup +\infty$, then*

$$\operatorname{argmin} f = zer\ \partial f = \{x \in E \text{ such that } 0 \in \partial f(x)\}.$$

The subdifferential also has the summability property:

Lemma 1. *Let f and g be two convex, l.s.c and proper functions. If dom $(f) \cap$ dom $(g) \neq \emptyset$ then, for $J = f + g$, we have $\partial J = \partial f + \partial g$.*

A.1.2 Proximal operator

The proximal operator associated to a convex function is defined as follows.

Definition 8 (Proximal operator)**.** *Let f be a convex function and $x \in E$, $Prox_f(x)$ is the unique minimizer of the strictly convex function:*

$$x \mapsto f(y) + \frac{1}{2} \|x - y\|^2 . \tag{A.2}$$

The proximal operator can be seen as the generalization of the concept of the projection onto a convex set. Indeed, if f is the characteristic function of a close convex set C then $\text{Prox}_f(x)$ is the orthogonal projection of x onto C.

In the following, we will say that f **is simple if its proximal operator can be computed explicitly**.

When f is not a convex function; then the proximal operator may not be defined or it can be multivalued. The convexity of f is a sufficient but not necessary condition for the existence and uniqueness of $\text{Prox}_f(x)$. In the following, a function will be said proximable if the proximal operator $\text{Prox}_f(x)$ is defined for every point $x \in E$. A proximable function can therefore be non-convex.

An important property of the proximal operator is now given.

Proposition 3. *Let f be a convex, l.s.c and proper function defined on E, then $Prox_f$ is a 1-Lipschitz operator.*

With the proximal operator, it is thus possible to minimize non-smooth convex functional through an implicit descent method.

Theorem 4 (Proximal Algorithm)**.** *Let f be a convex, l.s.c and proper function defined from E to $\mathbb{R} \cup +\infty$, then the fixed points of $Prox_f$ are minimizers of f. Hence, $\forall \tau > 0$ and $x_0 \in E$, the sequence $(x_n)_{\in \mathbb{N}}$ defined as*

$$x_{n+1} = Prox_{\tau f}(x_n)$$

converges to a minimizer of f.

A.1.3 Conjugate function

A large range of approaches use the duality of convex functionals to design minimization algorithms. For all function f from E to $\bar{\mathbb{R}}$, one can associate a conjugate function f^* through the following relation.

Definition 9 (Convex conjugate). *Let f be a function defined from E to $\bar{\mathbb{R}}$, the Fenchel conjugate of f is the convex function defined from E to $\bar{\mathbb{R}}$ by*

$$f^*(u) = \sup_{x \in E}(\langle x, u\rangle - f(x)). \tag{A.3}$$

Conjugate functions plays an essential role in the optimisation of convex functionals. There exists strong relations between conjugation subdifferential and proximal operators, from which the equivalence between primal and dual problems can be established. One main property of this transform is the following.

Theorem 5 (Biconjugate). *All function f defined on E with values in $[-\infty, +\infty]$, convex, l.s.c and proper is equal o its biconjugate: $f^{**} = f$.*

One useful corollary can be shown from this relation.

Corollary 1 (Moreau's identity). *If f is convex, l.s.c and proper, then , for all $\tau > 0$: $Prox_{\tau f}(x) + \tau Prox_{f^*/\tau}(x/\tau) = x$, that is to say $Prox_f + Prox_{f^*} = Id$.*

A.2 Convex functionals

Standard algorithms for the minimization of convex and non-smooth functionals are now introduced. We will assume that the considered functions admit at leat one minimizer.

A.2.1 Primal Proximal Algorithms

In this subsection, we first consider the problem:

$$\min_{x \in E} J(x) = f(x) + g(x), \tag{A.4}$$

where f and g are convex, l.s.c, proper, coercive and bounded from below functions.

We first assume that f is differentiable with a L-Lipschitz gradient and that the condition of Lemma 1 is satisfied so that the subdifferential of J is equal to the sum of the subdifferentials of f and g for all point of the domain of ∂J. From Theorem 3, a minimizer x^* of J then checks $-\nabla f(x^*) \in \partial g(x^*)$. Under these assumptions, a minimizer of J can be estimated with the Forward-Backward (FB) Algorithm.

Theorem 6 (Forward-Backward Algorithm). *Let $J = f + g$ be the sum of two convex, l.s.c, proper and coercive functions. Let f be differentiable with a L-Lipschitz gradient. Let $\tau < \dfrac{1}{L}$ and $x_0 \in E$ then the sequence $(x_n)_{n \in \mathbb{N}}$ defined $\forall n \in \mathbb{N}$ as*

$$x_{n+1} = Prox_{\tau g}(Id - \tau \nabla f)(x_n) \tag{A.5}$$

converges to a minimizer of J.

A complete description of this algorithm can be found in [65]. Notice that using an sadequate sequence of time steps τ_n, it is possible to accelerate its convergence [15, 55].

The FB algorithm can also be generalized to deal with a differentiable convex function and the sum of proximable convex ones [182].

When both functions f and g are non-smooth, the standard Douglas-Rachford (DR) algorithm [134, 64] relying on the proximal operators of the two functions can be considered.

Theorem 7 (Douglas-Rachford Algorithm)**.** *Let $J = f + g$ be the sum of two convex, l.s.c, proper and coercive functions. Let $\mu \in [0, 2]$, $\tau > 0$, $x_0 \in E$ and $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ and $(z_n)_{n\in\mathbb{N}}$ the sequences defined as*

$$\left\{\begin{array}{ll} x_n = & Prox_{\tau g}(y_n), \\ z_n = & Prox_{\tau f}(2x_n - y_n), \\ y_{n+1} = & y_n + \mu(z_n - x_n), \end{array}\right. \tag{A.6}$$

then there exists $x^* \in E$ minimizer of J such that the sequence (x_n) converges to x^*.

The Alternating Direction Method of Multipliers (ADMM) is another classical optimization method for that purpose. This approach consists in introducing an auxiliary variable $y \in E$ and dealing with the constraint $y = x$ with a Lagrange multiplier $z \in E$. An augmented Lagrangian is then defined for $\tau > 0$ and $\forall (x, y, z) \in E \times E \times E$ as

$$\mathcal{L}_\tau(x, y, z) = f(y) + g(x) + \langle z, x - y\rangle + \frac{\tau}{2}\|x - y\|^2 . \tag{A.7}$$

One can show that problem (A.4) is equivalent to

$$\min_{(x,y)} \max_{z} \mathcal{L}_\tau(x, y, z). \tag{A.8}$$

Theorem 8 (ADMM Algorithm)**.** *Let $J = f + g$ be the sum of two convex, l.s.c, proper and coercive functions. Let $\tau > 0$, $(y_0, z_0) \in E \times E$ and $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ et $(z_n)_{n\in\mathbb{N}}$ the sequences defined as*

$$\left\{\begin{array}{ll} x_{n+1} & = \operatorname{argmin}_x \mathcal{L}_\tau(x, y_n, z_n), \\ y_{n+1} & = \operatorname{argmin}_y \mathcal{L}_\tau(x_{n+1}, y, z_n), \\ z_{n+1} & = z_n + \tau(x_{n+1} - y_{n+1}), \end{array}\right. \tag{A.9}$$

then there exists $z^* \in E$ minimizer of J such that the sequence (z_n) converges to z^*.

More details can be found in the reference papers [104, 108]. The ADMM algorithm can be reformulated in term of proximal operators of f and g, which makes it equivalent to the DR algorithm [90]. More precisely, the sequence of iterates of DR with $\mu = 1$ for the primal probleml $f + g$ are identical to the sequence of iterates of ADMM for the problem $f^* + g^*$ [160].

A.2.2 Primal-Dual Proximal Algorithms

We now consider the minimization of two convex functions, one of them being composed with a linear operator K:

$$\min_{x\in E} J(x) = f(Kx) + g(x), \tag{A.10}$$

where K is defined on E with values in F, f defined from F to $\bar{\mathbb{R}}$ and g defined from E to $\bar{\mathbb{R}}$. We now assume that the proximal operator of $f \circ K$ can not be computed explicitly, whereas f and g are simple (*i.e.* their proximal operators can be computed explicitly). In order to define a simple algorithm for minimizing J in this case, one can consider DR or ADMM algorithms by introducing an auxiliary variable $y \in F$ and solve the following problem:

$$\min_{x \in E} f(Kx) + g(x) = \min_{x \in E, y \in F} f(y) + g(x) + \iota_K(x, y), \tag{A.11}$$

where h is the characteristic function of the linear constraint $Kx = y$:

$$\iota_K(x, y) = \begin{cases} 0 & \text{if } Kx = y \\ +\infty & \text{otherwise.} \end{cases}$$

By denoting $\tilde{f} = f + g$ and $\tilde{g} = \iota_K$, the Douglas-Rachford Algorithm can indeed be directly applied to the function $\tilde{f} + \tilde{g}$ with the variable $z = (x, y)$. The proximal operator of $\tilde{f}(z)$ can be computed separately for the two variables x and y through the proximal operators of f and g. The proximal operator of $\tilde{g}(z)$ can be computed explicitly as the orthogonal projection onto the linear constraint $Kx = y$.

The *Primal-Dual* formulation is another way to deal with such problem. The convex conjugate (A.3) can indeed be used to separate the function f from the operator K:

$$\forall x \in E, \quad f(Kx) = \sup_{y \in F}(\langle y, Kx \rangle - f^*(y)) = \sup_{y \in F}(\langle K^* y, x \rangle - f^*(y)) \tag{A.12}$$

where K^* denotes the adjoint operator of K defined from F to E. Injecting this expression in the problem (A.10), ones gets a new problem where the function f has been dualized. As soon as the subdifferential of f^* is no empty for all point of its domain, the above supremum is a maximum. This will be the case for all the functionals we will consider, so that we can turn into the general saddle-point problem $(x^*, y^*) \in E \times F$:

$$\min_{x \in E} \max_{y \in F} \langle K^* y, x \rangle - f^*(y) + g(x). \tag{A.13}$$

Notice that problems (A.10) and (A.13) are equivalent in the sense that if x is a solution of (A.10) then there exists y such that the pair (x, y) is solution of (A.13). Conversely, if (x, y) is a solution of (A.13) then x is a solution of (A.10).

In [56], a variant of the Arrow-Hurwicz Algorithm [7] has been proposed to deal with this problem. Such algorithm is equivalent to the DR one, if $K = Id$. We now recall a more general expression of this algorithm [114, 212, 66, 37, 57, 137, 88], able to deal with an additional differentiable convex function h:

$$\min_{x \in E} \max_{y \in F} \langle K^* y, x \rangle - f^*(y) + g(x) + h(x). \tag{A.14}$$

Theorem 9 (Primal-Dual Algorithm)**.** *Let f, g and h be three convex, l.s.c and proper functions, f defined from F to $[-\infty, +\infty]$, g defined from E tp $[-\infty, +\infty]$, h differentiable defined from E to $[-\infty, +\infty]$ with a L-Lipschitz gradient and K a linear operator defined from F to E. Let $L_K = \|K\|$, $\tau > 0$ and $\sigma > 0$ such that $(1/\tau + L)/\sigma > L_K^2$.*

Let $(x_0, y_0) \in E \times F$ *and* $\bar{x}_0 = x_0$ *and the sequences* $(x_n)_{n\in\mathbb{N}}$, $(\bar{x}_n)_{n\in\mathbb{N}}$, *et* $(y_n)_{n\in\mathbb{N}}$ *defined as*

$$\left\{\begin{array}{rcl} x_{n+1} & = & Prox_{\tau g}(x_n - \tau(K^* y_n + \nabla h(x_n)), \\ \bar{x}_{n+1} & = & 2x_{n+1} - x_n, \\ y_{n+1} & = & Prox_{\sigma f^*}(y_n + \sigma K \bar{x}_{n+1}), \end{array}\right. \tag{A.15}$$

then the sequence (x_n, y_n) *converges to a solution* (x^*, y^*) *of* (A.14).

The time steps of this algorithm can be defined automatically with respect to the operator K by considering diagonal preconditioning matrices T and Σ instead of scalar values τ and σ [169, 137]. The algorithm can also deal with several functions including different linear operators $\sum_i f_i(K_i x)$. Notice that better "worst case convergence rates" can be obtained by using adequate strategies involving extragradient steps [88], overrelaxation parameters [57] or inertial terms [137].

A final important point to underline is that the definition of the proximal operator is not limited to the L_2 norm (A.2) and any strongly convex function ψ can be considered:

$$\mathrm{Prox}_f^{\psi}(x) = \underset{y}{\mathrm{argmin}}\ f(y) + \psi(x, y).$$

Hence, depending of the function f, a good choice of the metric ψ can lead to an explicit expression of the proximal operator [57].

A.3 Non-convex functionals

We now return to the problem (A.4):

$$\min_{x\in E} J(x) = f(x) + g(x). \tag{A.16}$$

When f and or g are non-convex and one of the two functions is non-smooth, we would like to estimate a local minima of J.

A.3.1 Back to the Forward-Backward

In the non-convex setting, local minima can not be characterized by the subdifferential that it is only defined for convex functions. The set of critical points of J, that include its local minima, can nevertheless be characterized by some local properties derived from the Fréchet subdifferential. When J satisfies the so-called Kurdika-Lojasiewicz (KL) property, the authors of [8] have shown that the Forward-Backward Algorithm (A.5) converges to a critical point of J. We refer to [8, 87] for more details on this technical assumption. Since the KL property is checked by all semi algebraic function, it it will be verified by all the considered image processing functionals introduced in the present document.

An acceleration of the Forward-Backward Algorithm has been proposed for this non-convex setting [156]. This approach relies on an inertial force, derived from the Nesterov acceleration dor the gradient descent [150]:

Theorem 10 (Inertial Forward-Backward Algorithm)**.** *Let f and g be two l.s.c and proper functions. Let f be differentiable with a L-Lipschitz gradient, g be proximable and $J = f + g$ verifying the KL property. Let $\beta \in [0; 1[$, $\tau < 2(1-\beta)/L$ and $x_0 \in E$, then the sequence $(x_n)_{n\in\mathbb{N}}$ defined as*

$$x_{n+1} = Prox_{\tau g}(x_n - \tau \nabla f(x_n)) + \beta(x_n - x_{n-1}) \tag{A.17}$$

converges to a critical point $x^\star$ of J.

The Forward-Backward can also be extended to deal with several proximable terms $\sum_i g(x)$ [62]. Finally notice that if f is non-convex but c-semi convex (such that $f(x) + c||x||^2$ is convex), other proximal algorithms in the primal [129] or primal-dual [148] formulations also ensure the convergence to a critical point of J.

A.3.2 Non-convex coupling

For some particular applications, we finally consider the following problem:

$$\min_{x\in E, y\in E'} J(x, y) = f(x) + g(y) + h(x, y), \tag{A.18}$$

where f and g are two convex functions and h is non-convex but differentiable. We make the additional assumption that J is separately convex in x and y but not in (x, y). As the function h may not be semiconvex, the abovementioned approaches can not be applied. The estimation of a critical point of such functional is ensured by an alternate minimization by blocks [210]:

Theorem 11 (Block Coordinates Descent Algorithm)**.** *Let $J(x, y) = f(x) + g(y) + h(x, y)$ such that for all fixed $(\tilde{x}, \tilde{y}) \in E \times E'$, $J(., \tilde{y})$ et $J(\tilde{x}, .)$ are convex, l.s.c and proper functions. Let $(x_0, y_0) \in E \times E'$, then the sequence $(x_n, y_n)_{n\in\mathbb{N}}$ defined by*

$$\begin{cases} x_{n+1} &= \underset{x}{\operatorname{argmin}}\ f(x) + h(x, y_n), \\ y_{n+1} &= \underset{y}{\operatorname{argmin}}\ g(y) + h(x_{n+1}, y), \end{cases} \tag{A.19}$$

converge to a critical point (x^*, y^*) of J.

This approach nevertheless requires to solve sub-optimization problems at each iteration of the algorithm. For some very specific functionals (A.18), these intern loops can be avoided and primal-dual approaches can be considered [167].

In fact, as soon as f and g are proximable functions (even if they are not convex), h is a $\mathcal{C}^2$ function and the gradient $\nabla_x h(., y)$ for y fixed is L_y-Lipschitz (resp. $\nabla_y h(x, .)$ is L_x-Lipschitz), a proximal alternating algorithm can be applied [31]. This last approach is related to the application of the Forward-Backward algorithm on the problem $\tilde{f}(z) = f(x) + g(y)$ and $\tilde{g}(z) = h(z)$ with the variable $z = (x, y)$.

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing Rome. *IEEE Computer*, 43(6):40–47, 2010.

[3] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[4] H. A. Almohamad and S. O. Duffuaa. A linear programming approach for the weighted graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5):522–525, 1993.

[5] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2006.

[6] S. Angenent, S. Haker, and A. Tannenbaum. Minimizing flows for the Monge-Kantorovich problem. *SIAM Journal on Mathematical Analysis*, 35(1):61–97, 2003.

[7] K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in linear and non-linear programming*. Stanford University Press, 1958.

[8] H. Attouch, J. Bolte, and B. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[9] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson. Image segmentation using active contours: Calculus of variations or shape gradients? *SIAM Applied Mathematics*, 63(6):2128–2154, 2003.

[10] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing*, volume 147 of *Applied Mathematical Sciences*. Springer US, 2002.

[11] J.-F. Aujol, G. Aubert, and L. Blanc-Féraud. Wavelet-based level set evolution for classification of textured images. *IEEE Transactions on Image Processing*, 12(12):1634–1641, 2003.

[12] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares partitioning. In *Symposium on Computational Geometry*, pages 350–357, 1992.

[13] I. Ayed, H. Chen, K. Punithakumar, I. Ross, and S. Li. Graph cut segmentation with a global constraint: Recovering region distribution via a bound of the bhattacharyya measure. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 3288–3295, 2010.

[14] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

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

[16] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[17] J. Benamou and Y. Brenier. Weak existence for the semigeostrophic equations formulated as a coupled Monge-Ampere/transport problem. *SIAM Journal on Applied Mathematics*, 58(5):1450–1461, 1998.

[18] J.-D. Benamou. A domain decomposition method for the polar factorization of vector-valued mappings. *SIAM Journal on Numerical Analysis*, 32(6):1808–1838, 1995.

[19] 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.

[20] J.-D. Benamou and Y. Brenier. Wasserstein optimal mapping between prescribed densities functions. *Journal of Optimization Theory and Applications*, 111(2):255–271, 2001.

[21] 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.

[22] J.-D. Benamou, G. Carlier, Q. Mérigot, and E. Oudet. Discretization of functionals involving the Monge-Amp\‘ere operator. Preprint arXiv:1408.4536, 2014.

[23] J.-D. Benamou, B. D. Froese, and O. A. M. Numerical solution of the optimal transportation problem via viscosity solutions for the Monge-Ampère equation. Preprint arXiv:1208.4873, 2012.

[24] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Numerical solution of the optimal transportation problem using the Monge-Ampère equation. *Journal of Computational Physics*, 260:107–126, 2014.

[25] B. Berkels. An unconstrained multiphase thresholding approach for image segmentation. In *Scale Space and Variational Methods in Computer Vision (SSVM'09)*, volume 5567, pages 26–37, 2009.

[26] D. Bertsekas. The auction algorithm: A distributed relaxation method for the assignment problem. *Annals of Operations Research*, 14:105–123, 1988.

[27] S. Beyou, A. Cuzol, S. S. Gorthi, and E. Mémin. Weighted ensemble transform kalman filter for image assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, 65, 2013.

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

[29] J. Bigot and T. Klein. Consistent estimation of a population barycenter in the wasserstein space. Preprint arXiv:11212.2562, 2012.

[30] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *European Conference on Computer Vision (ECCV'04)*, 2004.

[31] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[32] N. Bonneel. Faster and specialized c++ implementation of "optimal transport with proximal splitting", 2013.

[33] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[34] N. Bonneel, K. Sunkavalli, S. Paris, and H. Pfister. Example-based video color grading. *ACM Transactions on Graphics*, 32(4):39:1–39:12, 2013.

[35] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich. Displacement interpolation using Lagrangian mass transport. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2011)*, 30(6), 2011.

[36] F. Bornemann and C. Rasch. Finite-element discretization of static hamilton-jacobi equations based on a local variational principle. *Comput and Visualization in Science*, 9(2):57–69, 2006.

[37] R. I. Bot, E. R. Csetnek, and A. Heinrich. On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. Preprint arXiv:1303.2875, 2013.

[38] A. Bouharguane, A. Iollo, and L. Weynans. Numerical solution of the Monge–Kantorovich problem by density lift-up continuation. *ESAIM Mathematical Modelling and Numerical Analysis*, 49(6):1577–1592, 2015.

[39] A. Bouharguane, E. Maitre, E. Oudet, and N. Papadakis. Multiphysics optimal transportation and image analysis. Preprint hal-00740671, 2012.

[40] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *IEEE International Conference on Computer Vision (ICCV'01)*, volume 1, pages 105–112, 2001.

[41] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[42] E. Brown, T. F. Chan, and X. Bresson. Completely convex formulation of the Chan-Vese image segmentation model. *International Journal of Computer Vision*, 98(1):103–121, 2012.

[43] T. Brox, M. Rousson, R. Deriche, and J. Weickert. Unsupervised segmentation incorporating colour, texture, and motion. In *Computer Analysis of Images and Patterns*, volume 2756, pages 353–360, 2003.

[44] T. Brox and J. Weickert. Level set segmentation with multiple regions. *IEEE Transactions on Image Processing*, 15(10):3213–3218, 2006.

[45] C. Brune. *4D imaging in tomography and optical nanoscopy*. PhD thesis, University of Münster, Germany, 2010.

[46] M. Burger, M. Franek, and C.-B. Schönlieb. Regularized regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress*, 2011.

[47] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.

[48] G. Buttazzo, C. Jimenez, and E. Oudet. An optimization problem for mass transportation with congested dynamics. *SIAM Journal on Control and Optimization*, 48(3):1961–1976, 2009.

[49] P. Cardaliaguet, G. Carlier, and B. Nazaret. Geodesics for a class of distances in the space of probability measures. *Calculus of Variations and Partial Differential Equations*, 526:1–26, 2012.

[50] G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. Preprint arXiv: 1411.3602, 2014.

[51] J. Carrillo and J. Moll. Numerical simulation of diffusive and aggregation phenomena in nonlinear continuity equations by evolving diffeomorphisms. *SIAM Journal on Scientific Computing*, 31(6):4305–4329, 2010.

[52] V. Chabot, M. Nodet, N. Papadakis, and A. Vidard. Accounting for observation errors in image data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, 67:19, 2015.

[53] A. Chambolle. An algorithm for mean curvature motion. *Interfaces and Free Boundaries. Mathematical Modelling, Analysis and Computation*, 6(2):195–218, 2004.

[54] A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3):288–307, 2009.

[55] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

[56] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[57] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, pages 1–35, 2015.

[58] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.

[59] Y. Chen and X. Ye. Projection onto a simplex. Preprint arXiv:1101.6081, 2011.

[60] L. Chizat, B. Schmitzer, G. Peyré, and F.-X. Vialard. An Interpolating Distance between Optimal Transport and Fischer-Rao. Preprint arXiv: 1506.06430, 2015.

[61] L. Chizat, B. Schmitzer, G. Peyré, and F.-X. Vialard. Unbalanced Optimal Transport: Geometry and Kantorovich Formulation. Preprint arXiv: 1508.05216, 2015.

[62] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. A block coordinate variable metric forward-backward algorithm. Preprint hal-00945918, 2014.

[63] P. Clarysse, B. Delhay, M. Picq, and J. Pousin. Optimal extended optical flow subject to a statistical constraint. *Journal of Computational and Applied Mathematics*, 234(4):1291–1302, 2010.

[64] P. L. Combettes and J.-C. Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564 –574, 2007.

[65] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

[66] L. Condat. 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, 2013.

[67] D. Cordero-Erausquin, B. Nazaret, and C. Villani. A mass-transportation approach to sharp Sobolev and Gagliardo-Nirenberg inequalities. *Advances in Mathematics*, 182(2):307–332, 2004.

[68] T. Corpetti, P. Héas, E. Mémin, and N. Papadakis. Pressure image assimilation for atmospheric motion estimation. *Tellus A: Dynamic Meteorology and Oceanography*, 61(1):160–178, 2009.

[69] T. Corpetti, É. Mémin, and P. Pérez. Dense estimation of fluid flows. *EEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):365–380, 2002.

[70] N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases*, volume 8724, pages 274–289, 2014.

[71] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *International Journal of Computer Vision*, 72(2):215, 2007.

[72] A. Criminisi, T. Sharp, C. Rother, and P. Pérez. Geodesic image and video editing. *ACM Transactions on Graphics*, 29(134):1–134, 2010.

[73] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Conference on Neural Information Processing Systems (NIPS'13)*, pages 2292–2300, 2013.

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

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

[76] M. Cuturi, G. Peyré, and A. Rolet. A smoothed dual approach for variational wasserstein problems. Preprint arXiv:1503.02533, 2015.

[77] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ, 1963.

[78] F. de Goes, D. Cohen-Steiner, P. Alliez, and M. Desbrun. An optimal transport approach to robust reconstruction and simplification of 2D shapes. *Computer Graphics Forum*, 30(5):1593–1602, 2011.

[79] E. J. Dean and R. Glowinski. An augmented Lagrangian approach to the numerical solution of the dirichlet problem for the elliptic Monge-Ampère equation in two dimensions. *Electronic Transactions on Numerical Analysis*, 22:71–96, 2006.

[80] M. C. Delfour and J.-P. Zolésio. *Shapes and geometries: analysis, differential calculus, and optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[81] J. Delon. Midway image equalization. *Journal of Mathematical Imaging and Vision*, 21:119–134, 2004.

[82] J. Delon. Movie and video scale-time equalization application to flicker reduction. *IEEE Transactions on Image Processing*, 15(1):241–248, 2006.

[83] J. Delon, J. Salomon, and A. Sobolevski. Fast transport optimization for Monge costs on the circle. *SIAM Journal on Applied Mathematics*, 70(7):2239–2258, 2010.

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

[85] J. Digne, D. Cohen-Steiner, P. Alliez, F. de Goes, and M. Desbrun. Feature-preserving surface reconstruction and simplification from defect-laden point sets. *Journal of Mathematical Imaging and Vision*, 48(2):369–382, 2014.

[86] J. Dolbeault, B. Nazaret, and G. Savaré. A new class of transport distances between measures. *Calculus of Variations and Partial Differential Equations*, 34(2):193–231, 2009.

[87] C. Dossal. Introduction à l'optimisation convexe non lisse pour le traitement d'images, 2015.

[88] Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209 – 214, 2015.

[89] B. During, D. Matthes, and J. Milisic. A gradient flow scheme for nonlinear fourth order equations. *Discrete and Continuous Dynamical Systems - Series B*, 14(3):935–959, 2010.

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

[91] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *IEEE International Conference on Computer Vision (ICCV'99)*, volume 2, pages 1033–1038, 1999.

[92] A. Elmoataz, O. Lezoray, and S. Bougleux. Nonlocal discrete regularization on weighted graphs: A framework for image and manifold processing. *IEEE Transactions on Image Processing*, 17(7):1047–060, 2008.

[93] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Science*, 7(1):212–238, 2014.

[94] S. Ferradans, N. Papadakis, J. Rabin, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. In *Scale Space and Variational Methods in Computer Vision (SSVM'13)*, pages 428–439, 2013.

[95] S. Ferradans, G.-S. Xia, G. Peyré, and J.-F. Aujol. Static and dynamic texture mixing using optimal transport. In *Proc. SSVM'13*, volume 7893 of *Lecture Notes in Computer Science*, pages 137–148. Springer Berlin Heidelberg, 2013.

[96] L. Ferragut and I. Asensio. Mixed finite element methods for a class of nonlinear reaction diffusion problems. *Neural, Parallel, and Scientific Computations*, 10(1):91–112, 2002.

[97] N. Feyeux, M. Nodet, and A. Vidard. Application of optimal transport to data assimilation. Special semester on new trends in calculus of variation, 2014.

[98] J. H. Fitschen, F. Laus, and G. Steidl. Dynamic optimal transport with mixed boundary condition for color image processing. In *International Conference on Sampling Theory and Applications (SampTA'15)*, 2015.

[99] O. Frigo, N. Sabater, J. Delon, and P. Hellier. Motion Driven Tonal Stabilization. Preprint hal-01150558, 2015.

[100] O. Frigo, N. Sabater, V. Demoulin, and P. Hellier. Optimal transportation for example-guided color transfer. In *Asian Conference on Computer Vision (ACCV'14)*, volume 9005, pages 655–670, 2014.

[101] U. Frisch, s. Matarrese, R. Mohayaee, and A. Sobolevski. A reconstruction of the initial conditions of the universe by optimal mass transportation. *Nature*, 417:260–262, 2002.

[102] B. D. Froese. A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions. *SIAM Journal on Scientific Computing*, 34(3):A1432–A1459, 2012.

[103] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a wasserstein loss. Preprint arXiv: 1506.05439, 2015.

[104] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[105] W. Gangbo and R. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[106] W. Gangbo and R. J. McCann. Shape recognition via wasserstein distance. *Quarterly of Applied Mathematics*, LVIII(4):705–737, 2000.

[107] G. Gilboa and S. Osher. Nonlocal linear image regularization and supervised segmentation. *SIAM Multiscale Modeling and Simulation*, 6(2):595–630, 2007.

[108] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis*, 9(R2):41–76, 1975.

[109] L. Gorelick, F. R. Schmidt, Y. Boykov, A. Delong, and A. Ward. Segmentation with non-linear regional constraints via line-search cuts. In *European Conference on Computer Vision (ECCV'12)*, pages 583–597, 2012.

[110] K. Guittet. On the time-continuous mass transport problem and its approximation by augmented lagrangian techniques. *SIAM Journal on Numerical Analysis*, 41(1):382–399, 2004.

[111] E. Haber, T. Rehman, and A. Tannenbaum. An efficient numerical method for the solution of the l_2 optimal mass transfer problem. *SIAM Journal on Scientific Computing*, 32(1):197–211, 2010.

[112] 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.

[113] F. H. Harlow and J. E. Welch. Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Physics of Fluids*, 8(12):2182–2189, 1965.

[114] B. He and X. Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

[115] M. Henry, E. Maitre, and V. Perrier. Optimal transport using Helmholtz-Hodge decomposition and first-order primal-dual algorithms. Preprint hal-01134194, 2015.

[116] A. Herbulot, S. Jehan-Besson, S. Duffner, M. Barlaud, and G. Aubert. Segmentation of vectorial image features using shape gradients and information measures. *Journal of Mathematical Imaging and Vision*, 25(3):365–386, 2006.

[117] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01)*, pages 327–340, 2001.

[118] R. Horst and N. Thoai. Dc programming: Overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

[119] R. Hug, E. Maitre, and N. Papadakis. Mathematical and numerical analysis of the dynamic optimal transport problem. In preparation, 2015.

[120] R. Hug, E. Maitre, and N. Papadakis. Multiphysics optimal transportation and image interpolation. *ESAIM Mathematical Modelling and Numerical Analysis*, 49(6):1671–1692, 2015.

[121] T. Hurtut, Y. Gousseau, and F. Schmitt. Adaptive image retrieval based on the spatial organization of colors. *Computer Vision and Image Understanding*, 112(2):101–113, 2008.

[122] A. Iollo and D. Lombardi. A lagrangian scheme for the solution of the optimal mass transfer problem. *Journal of Computational Physics*, 230(9):3430–3442, 2011.

[123] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.

[124] 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.

[125] L. Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.

[126] J. Kim, J. W. Fisher, A. Yezzi, M. Cetin, and A. S. Willsky. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Transactions on Image Processing*, 14(10):1486–1502, 2005.

[127] Y.-H. Kim and B. Pass. Multi-marginal optimal transport on Riemannian manifolds. Preprint arXiv:1303.6251, 2013.

[128] R. Kimmel and J. Sethian. Computing geodesic paths on manifolds. *Proc. of the National Academy of Sciences*, 95(15):8431–8435, 1998.

[129] M. Kowalski. Thresholding rules and iterative shrinkage/thresholding algorithm: A convergence study. In *IEEE International Conference on Image Processing (ICIP'14)*, 2014.

[130] H. W. Kuhn. The Hungarian method of solving the assignment problem. *Naval Res. Logistics Quart.*, 2:83–97, 1955.

[131] J. Lellmann, D. A. Lorenz, C. Schönlieb, and T. Valkonen. Imaging with kantorovich-rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[132] B. Levy. A numerical algorithm for l_2 semi-discrete optimal transport in 3d. Preprint arXiv: 1409.1279, 2014.

[133] 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.

[134] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *Siam J. Numer. Anal.*, 16:964–979, 1979.

[135] G. Loeper and F. Rapetti. Numerical solution of the Monge-Ampère equation by a Newton's algorithm. *Comptes Rendus Mathematique*, 340(4):319–324, 2005.

[136] D. Lombardi and E. Maitre. Eulerian models and algorithms for unbalanced optimal transport. *ESAIM Mathematical Modelling and Numerical Analysis*, 49(6):1717–1744, 2015.

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

[138] J. Louet and F. Santambrogio. A sharp inequality for transport maps in via approximation. *Applied Mathematics Letters*, 25(3):648–653, 2012.

[139] J. Maas, M. Rumpf, C. Schönlieb, and S. Simon. A generalized model for optimal transport of images including dissipation and density modulation. Preprint arXiv: 1504.01988, 2015.

[140] S. Mallat and I. Waldspurger. Phase retrieval for the Cauchy wavelet transform. Preprint arXiv: 1404.1183, 2014.

[141] R. McCann. Polar factorization of maps on Riemannian manifolds. *Geometric & Functional Analysis*, 11(3):589–608, 2001.

[142] R. J. McCann. A convexity principle for interacting gases. *advances in mathematics*, 128(1):153–179, 1997.

[143] C. Mendoza, J.-A. Perez-Carrasco, A. Saez, B. Acha, and C. Serrano. Linearized multidimensional earth-mover's-distance gradient flows. *IEEE Transactions on Image Processing*, 22(12):5322–5335, 2013.

[144] Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[145] G. Monge. Mémoire sur la théorie des déblais et de remblais. *Académie Royale des Sciences de Paris*, page 666–704, 1781.

[146] J. Morovic and P.-L. Sun. Accurate 3d image colour histogram transformation. *Pattern Recognition Letters*, 24(11):1725–1735, 2003.

[147] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42:577–685, 1989.

[148] T. Möllenhoff, E. Strekalovskiy, M. Moeller, and D. Cremers. The primal-dual hybrid gradient method for semiconvex splittings. *SIAM Journal on Imaging Sciences*, 8(2):827–857, 2015.

[149] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization*. Wiley-Interscience, 1988.

[150] Y. E. Nesterov and A. S. Nemirovsky. *Interior Point Polynomial Methods in Convex Programming: Theory and Algorithms*. SIAM Publishing, 1993.

[151] K. Ni, X. Bresson, T. F. Chan, and S. Esedoglu. Local histogram based segmentation using the wasserstein distance. *International Journal of Computer Vision*, 84(1):97–111, 2009.

[152] M. Nikolova, S. Esedoglu, and T. F. Chan. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal on Applied Mathematics*, 66(5):1632–1648, 2006.

[153] M. Nikolova, Y.-W. Wen, and R. H. Chan. Exact histogram specification for digital images using a variational approach. *Journal of Mathematical Imaging and Vision*, 46(3):309–325, 2013.

[154] A. M. Oberman. Wide stencil finite difference schemes for the elliptic Monge-Ampere equation and functions of the eigenvalues of the hessian. *Discrete and Continuous Dynamical Systems - Series B*, 1(10):221–238, 2008.

[155] A. M. Oberman and Y. Ruan. An efficient linear programming method for Optimal Transportation. Preprint arXiv:1509.03668, 2015.

[156] P. Ochs, Y. Chen, T. Brox, and T. Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[157] V. I. Oliker and L. D. Prussner. On the numerical solution of the equation $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 = f$ and its discretizations. *Numerische Mathematik*, 3(54):271–293, 1988.

[158] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.

[159] N. Papadakis, V. Chabot, A. Makris, M. Nodet, and A. Vidard. Assimilation d'images et de structures. In *Colloque du Groupe d'Etudes du Traitement du Signal et des Images (GRETSI'13)*, 2013.

[160] N. Papadakis, G. Peyré, and E. Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[161] N. Papadakis, E. Provenzi, and V. Caselles. A variational model for histogram transfer of color images. *IEEE Transactions on Image Processing*, 20(6):1682–1695, 2011.

[162] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *European Conference on Computer Vision (ECCV'08)*, pages 495–508, 2008.

[163] O. Pele and M. Werman. Fast and robust earth mover's distances. In *IEEE International Conference on Computer Vision (ICCV'09)*, pages 460–467, 2009.

[164] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *ACM Transactions on Computer Graphics and Interactive Techniques (SIGGRAPH'03)*, pages 313–18, 2003.

[165] G. Peyré, J. Fadili, and J. Rabin. Wasserstein active contours. In *IEEE International Conference on Image Processing (ICIP'12)*, 2012.

[166] B. Piccoli and F. Rossi. Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

[167] F. Pierre, J.-F. Aujol, A. Bugeau, N. Papadakis, and V.-T. Ta. Luminance-chrominance model for image colorization. *SIAM Journal on Imaging Sciences*, 8(1):536–563, 2015.

[168] F. Pitié and A. Kokaram. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In *European Conference on Visual Media Production (IETCVMP'07)*, pages 1–9, 2007.

[169] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *IEEE International Conference on Computer Vision (ICCV'11)*, pages 1762–1769, 2011.

[170] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM Journal of Imaging Sciences*, 3(4):1122–1145, 2010.

[171] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *Internationall Journal of Computer Vision*, 40(1):49–71, 2000.

[172] T. Pouli and E. Reinhard. Progressive color transfer for images of arbitrary dynamic range. *Computer Graphics*, 35(1):67–80, 2011.

[173] K. Punithakumar, J. Yuan, I. Ben Ayed, S. Li, and Y. Boykov. A convex max-flow approach to distribution-based figure-ground separation. *SIAM Journal on Imaging Sciences*, 5(4):1333–1354, 2012.

[174] J. Rabin, J. Delon, and Y. Gousseau. Removing artefacts from color and contrast modifications. *IEEE Transactions on Image Processing*, 20(11):3073–3085, 2011.

[175] J. Rabin, S. Ferradans, and N. Papadakis. Adaptive color transfer with relaxed optimal transport. In *IEEE International Conference on Image Processing (ICIP'14)*, 2014.

[176] J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *Scale Space and Variational Methods in Computer Vision (SSVM'15)*, pages 241–252, 2015.

[177] J. Rabin and N. Papadakis. Cosegmentation d'images non-supervisée utilisant les distances de sinkhorn. In *Colloque du Groupe d'Etudes du Traitement du Signal et des Images (GRETSI'15)*, 2015.

[178] J. Rabin and N. Papadakis. Non-convex relaxation of optimal transport for color transfer. In *Conference on Geometric Science of Information (GSI'15)*, 2015.

[179] J. Rabin and G. Peyré. Wasserstein regularization of imaging problem. In *IEEE International Conderence on Image Processing (ICIP'11)*, pages 1541–1544, 2011.

[180] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision (SSVM'11)*, pages 435–446, 2011.

[181] J. Rabin, G. Peyré, and L. Cohen. Geodesic shape retrieval via optimal mass transport. In *European Conference on Computer Vision (ECCV'10)*, pages 771–784, 2010.

[182] H. Raguet, J. Fadili, and G. Peyré. Generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

[183] A. Rangarajan, A. L. Yuille, S. Gold, and E. Mjolsness. A convergence proof for the softassign quadratic assignment algorithm. In *Neural Information Processing Systems (NIPS'96)*, pages 620–626, 1996.

[184] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Transactions on Computer Graphics and Applications*, 21(5):34–41, 2001.

[185] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 993–1000, 2006.

[186] M. Rousson, T. Brox, and R. Deriche. Active unsupervised texture segmentation on a diffusion based feature space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)*, 2003.

[187] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision (ICCV'98)*, pages 59–66, 1998.

[188] C. Schellewald, S. Roth, and C. Schnörr. Evaluation of a convex relaxation to a quadratic assignment matching approach for relational object views. *Image and Vision Computing*, 25(8):1301–1314, 2007.

[189] B. Schmitzer. A sparse algorithm for dense optimal transport. In *Scale Space and Variational Methods in Computer Vision (SSVM'15)*, pages 629–641, 2015.

[190] B. Schmitzer and C. Schnörr. A hierarchical approach to optimal transport. In *Scale Space and Variational Methods in Computer Vision (SSVM'13)*, pages 452–464, 2013.

[191] B. Schmitzer and C. Schnörr. Modelling convex shape priors and matching based on the gromov-wasserstein distance. *Journal of Mathematical Imaging and Vision*, 46(1):143–159, 2013.

[192] B. Schmitzer and C. Schnörr. Globally optimal joint image segmentation and shape matching based on wasserstein modes. *Journal of Mathematical Imaging and Vision*, 52(3):436–458, 2015.

[193] A. Schrijver. *Theory of linear and integer programming.* John Wiley & Sons, Inc., 1986.

[194] V. Seguy and M. Cuturi. An Algorithmic Approach to Compute Principal Geodesics in the Wasserstein Space. Preprint arXiv: 1506.07944, 2015.

[195] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proc. of the National Academy of Sciences*, 93(4):1591–1595, 1996.

[196] G. Sharma, F. Jurie, and C. Schmid. Expanded parts model for human attribute and action recognition in still images. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'14)*, pages 652–659, 2014.

[197] S. Shirdhonkar and D. Jacobs. Approximate earth mover's distance in linear time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, 2008.

[198] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[199] 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. In *ACM Transactions on Graphics (SIGGRAPH'15)*, 2015.

[200] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *International Conference on machine Learning (ICML'14)*, pages 306–314, 2014.

[201] G. Strang. *Linear Algebra and Its Applications*. Brooks Cole, 1988.

[202] E. Strekalovskiy, A. Chambolle, and D. Cremers. Convex relaxation of vectorial problems with coupled regularization. *SIAM Journal on Imaging Sciences*, 7(1):294–36, 2014.

[203] Z. Su, K. Zeng, L. Liu, B. Li, and X. Luo. Corruptive artifacts suppression for example-based color transfer. *IEEE Transactions on Multimedia*, 16(4):988–999, 2014.

[204] M. Sulman, J. F. Williams, and R. D. Russell. Optimal mass transport for higher dimensional adaptive grid generation. *Journal of Computational Physics*, 230(9):3302–3330, 2011.

[205] P. Swoboda and C. Schnörr. Convex variational image restoration with histogram priors. *SIAM Journal on Imaging Sciences*, 6(3):1719–1735, 2013.

[206] P. Swoboda and C. Schnörr. Variational image segmentation and cosegmentation with the wasserstein distance. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR'13)*, pages 321–334, 2013.

[207] Y.-W. Tai, J. Jia, and C.-K. Tang. Local color transfer via probabilistic segmentation by expectation-maximization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 747–754, 2005.

[208] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[209] O. Titaud, A. Vidard, I. Souopgui, and F.-X. Le Dimet. Assimilation of Image Sequences in Numerical Models. *Tellus A: Dynamic Meteorology and Oceanography*, 62(1):30–47, 2010.

[210] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[211] J. Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE Transactions on Automatic Control*, 40(9):1528–1538, Step. 1995.

[212] B. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[213] L. A. Vese and T. F. Chan. A multiphase level set framework for image segmentation using the mumford and shah model. *International Journal of Computer Vision*, 50(3):271–293, 2002.

[214] S. Vicente, V. Kolmogorov, and C. Rother. Joint optimization of segmentation and appearance models. In *IEEE International Conference on Computer Vision (ICCV'09)*, pages 755–762, 2009.

[215] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation revisited: Models and optimization. In *European Conference on Computer Visio (ECCV'10)*, pages 465–479, 2010.

[216] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

[217] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, 2008.

[218] A. G. Wilson. The Use of Entropy Maximising Models, in the Theory of Trip Distribution, Mode Split and Route Split. *Journal of Transport Economics and Policy*, 3(1):108–126, 1969.

[219] G.-S. Xia, S. Ferradans, G. Peyré, and J.-F. Aujol. Synthesizing and mixing stationary gaussian texture models. *SIAM Journal on Imaging Sciences*, 7(1):476–508, 2014.

[220] X. Xiao and L. Ma. Color transfer in correlated color space. In *ACM International Conference on Virtual Reality Continuum and Its Applications (VRCIA'06)*, pages 305–309, 2006.

[221] R. Yıldızoğlu, J.-F. Aujol, and N. Papadakis. Active contours without level sets. In *IEEE International Conference on Image Processing (ICIP'12)*, pages 2549–2552, 2012.

[222] R. Yıldızoğlu, J.-F. Aujol, and N. Papadakis. A convex formulation for global histogram based binary segmentation. In *International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR'13)*, pages 335–349, 2013.

[223] J. Yuan, C. Schnörr, and E. Mémin. Discrete orthogonal decomposition and variational fluid flow estimation. *Journal of Mathematical Imaging and Vision*, 28(1):67–80, 2007.

[224] Y. Yuan, E. Ukwatta, X. Tai, A. Fenster, and C. Schnörr. A fast global optimization-based approach to evolving contours with generic shape prior. *UCLA Tech. Report CAM*, 12-38, 2012.

[225] M. Zaslavskiy, F. Bach, and J.-P. Vert. A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2227–2242, 2009.

[226] M. Zaslavskiy, F. Bach, and J.-P. Vert. Many-to-many graph matching: a continuous relaxation approach. In *European Conference on Machine Learning and Practice of Knowledge Discovery in Databases (ECML PKDD'10)*, pages 515–530, 2010.

[227] Y. Zheng and D. Doermann. Robust point matching for nonrigid shapes by preserving local neighborhood structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):643–649, 2006.

[228] S. Zhu, T. Lee, and A. Yuille. Region competition: unifying snakes, region growing, energy/bayes/mdl for multi-band image segmentation. In *IEEE International Conference on Computer Vision (ICCV'95)*, pages 416 –423, 1995.