Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the

Ruperto-Carola University of Heidelberg, Germany

for the degree of
Doctor of Natural Sciences

put forward by
M.Sc. Bernhard Schmitzer
born in Feuchtwangen

Date of oral examination: .....

Isometry Invariant Shape Priors for Variational Image Segmentation

Advisor: Prof. Dr. Christoph Schnörr

Zusammenfassung

Variationsmethoden spielen eine grundlegende Rolle in der mathematischen Bildverarbeitung als Bindeglied zwischen Modellen und Algorithmen. Es ist oft schwierig, ein gegebenes Modell auf ein handhabbares Optimierungsproblem abzubilden. Zuletzt wurde große Fortschritte bei der (näherungsweisen) Lösung lokaler Modelle durch konvexe Relaxationen gemacht. Doch nicht-lokale Modellierungsaspekte, wie z.B. die Form des gesuchten Objekts, stellen weiter eine große Herausforderung dar. In dieser Dissertation werden verschiedene mathematische Abstraktionen des Konzepts 'Form' studiert und, darauf basierend, Funktionale entwickelt, zur Einbringung von Vorwissen über die Objektform in variationelle Segmentierungsverfahren. Besondere Aufmerksamkeit wird der Invarianz der Funktionale unter Isometrien und der Kompatibilität mit den bereits existierenden konvexen Funktionalen für das Segmentierungsproblem gewidmet. Das Transportproblem wird als zentrales Werkzeug zur Modellierung und Berechnung von Korrespondenzen zwischen verschiedenen Formen eingesetzt. Auf Basis der berechneten Korrespondenzen kann ein aussagekräftiges Ähnlichkeitsmaß für Formen definiert werden. Unter diesem Blickwinkel wird zwischen den, in gewisser Weise komplementären, Darstellungen einer Form durch die Fläche die sie einnimmt oder durch ihre Kontur, eine Verbindung erkennbar. Dies ermöglicht die Kombination der jeweiligen Stärken. Ohne weiteres ist der implizierte Rechenaufwand der vorgestellten Funktionale sehr hoch. Daher werden geeignete hierarchische Optimierungsverfahren entwickelt.

Abstract

Variational methods play a fundamental role in mathematical image analysis as a bridge between models and algorithms. A major challenge is to formulate a given model as a feasible optimization problem. There has been a huge leap in that respect concerning local data models in the framework of convex relaxation. But non-local concepts such as the shape of a sought-after object are still difficult to implement. In this thesis we study mathematical representations for shapes and develop shape prior functionals for object segmentation based thereon. A particular focus is set on the isometry invariance of the functionals and the compatibility with existing convex functionals for image labelling. Optimal transport is used as a central modelling and computational tool to compute registrations between different shapes as a basis for a shape similarity measure. This point of view leads to a link between the two somewhat dual representations of a shape by the region it occupies and its outline, allowing to combine their respective strengths. Naïvely the computational complexity implied by the derived functionals is unfeasible. Therefore suitable hierarchical optimization methods are developed.

Acknowledgements

A lot of people have contributed in very different ways to this thesis. First off, I want to thank my advisor, Christoph Schnörr, who warmly welcomed me in the new field and showed me the ropes. His calmness, fairness and his ability to adapt to every student's own style made a great impression on me.

Then there is a long list of colleagues at the Image & Pattern Analysis Group, the Heidelberg Collaboratory for Image Processing and the Research Training Group for Graphical Models whom I wish to thank: Borislav Antic, Fabian Bachl, Florian Becker, Johannes Berger, Andreea Denitiu, Eva-Maria Didden, Johannes Dück & Dominic Edelmann, Jan Lellmann, Frank Lenzen, Jörg Kappes, Bernhard Kausler, Andreas Neufeld, Stefania Petra, Fabian Rathke, Bogdan Savchynskyy, Stefan Schmidt, Markus Speth, Karsten Staack, Paul Swoboda, Barbara Werner and Evelyn Wilhelm. Thank you, for discussions on mathematics and beyond, vital technical & administrative support, directions through the rat maze, for being great travel companions and sport buddies and in general for a lot of good times in Heidelberg or out and about!

Finally, I want to thank Ute for her rock-solid support.

Funding by the Deutsche Forschungsgemeinschaft via the research training group 1653 *Spatio / Temporal Graphical Models and Applications in Image Analysis* is also gratefully acknowledged.

Contents

List of Publications

1. B. Schmitzer and C. Schnörr. Weakly convex coupling continuous cuts and shape priors. In *Scale Space and Variational Methods (SSVM 2011)*, pages 423–434, 2012.

2. 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.

3. B. Schmitzer and C. Schnörr. A hierarchical approach to optimal transport. In *Scale Space and Variational Methods (SSVM 2013)*, pages 452–464, 2013.

4. B. Schmitzer and C. Schnörr. Object segmentation by shape matching with Wasserstein modes. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR 2013)*, pages 123–136, 2013.

5. B. Schmitzer and C. Schnörr. Contour manifolds and optimal transport. http://arxiv.org/abs/1309.2240, 2013. preprint.

Chapter 1

Introduction

1.1 Shape and Representation

1.1.1 Shape and Tasks

Shape is an omnipresent concept in mathematical image analysis and computer vision. It is difficult to find a precise definition for it, because of the vast set of mathematical and computational structures or *representations* that has been associated with the word 'shape': contours, indicator functions, pictorial structures, patches, histograms, point clouds, triangulated meshes and many more. However, there is a set of problems recurring in the context of shape. Typical examples are

- measuring similarity between shapes, i.e. finding a metric on shapes, applicable for classification and recognition,
- statistical analysis and modelling of distributions of shapes,
- finding meaningful registrations or interpolations between similar shapes,
- optimizing w.r.t. exterior criteria, be it technical specifications in product design or local appearance features for object localization and pose estimation in image data,
- the abstraction from geometric transformations, such as Euclidean isometries in the ambient space or non-isometric changes in the pose of articulated objects, that a shape can undergo while still perceptually remaining the same shape.

Each of the representations named above is more or less compatible with certain tasks, mathematically or computationally. So for each application the most suitable representation has to be picked with care.

1.1.2 Object Segmentation

Object segmentation is a prototypical problem related to shape. One is to locate and segment a particular sought-after object within a given image. It is a special case of the image labelling problem with two classes (fore- and background). In a variational framework every pixel of the image is assigned a label, based on local appearance information and a regularizer. The job of the regularizer is to exploit prior knowledge on the arrangement of labels to compensate for noisy local observations.

A popular a priori modelling assumption is that boundaries between regions of different labels are smooth and occur along lines in the image with a strong gradient. Such regularizers are local in the sense that the corresponding functionals are local functions of the labelling. From the perspective of probabilistic graphical models this means that labels in two disjoint regions that are separated by a third region are independent when conditioned on the separating region. Local regularizers have been studied in great depth and a lot of attention has been given to corresponding convex relaxations to obtain (approximately) optimal solutions.

But local regularizers can only compensate for local noise, i.e. noise that is statistically independent at different points of the image. This assumption is often not satisfied: faulty observations caused by illumination changes or occlusion clearly have long range correlations. At the same time, for the problem of object segmentation more detailed prior knowledge might be available that is not exploited by local regularizers: the shape of the sought after object. A regularizer that encourages the foreground region to have a particular shape is called *shape prior*.

The design of a shape prior entails most, if not all, the tasks we have listed above. We are looking for an image region that is optimal in terms of both appearance and shape regularization. Determining whether a given region looks like the sought-after shape requires a measure of shape similarity. Since not all sought-after instances will be exact copies of each other, one will need to be able to model the expected statistical intra-class variations. In many problems the location, orientation or scale of the object are not known beforehand, in other problems the same shape may be represented by equivalent but seemingly different mathematical objects (e.g. different parametrizations of the same curve), so the prior must be invariant w.r.t. these transformations. Last but not least, it has been demonstrated that the computation of registrations between shapes is a helpful basis for the construction of a similarity measure.

This makes choice of shape representation particularly difficult and in fact many variants have been tried, resulting in a variety of approaches with weaknesses and strengths in different aspects. The goal of the work, presented in this thesis, is to design and study shape representations and mathematically sound shape prior functionals based thereon that combine as much of the different advantages as possible.

1.2 Related Literature

1.2.1 Image Labelling

The image labelling problem with local regularizers is a fundamental field of research and many approaches for convex relaxations have been presented to obtain good approximate or even exactly optimal solutions to the inherently combinatorial problems. The prototypical convex local regularizer is *total variation*, of which, since its introduction to image analysis [ROF92], countless variations and extensions have been developed: application to vectorial data [BC98, LS11, PCCB09, GC10], regularization of higher order derivatives [BKP10], implementation of local ordering constraints [LVS10, SC11], anisotropic adaptive variants [GL10] and many more. The two class case has been shown to be solvable exactly through means of convex relaxation in the discrete [KZ04] and continuous case [CEN06].

A common feature of these approaches is that segmentation regions are represented by (relaxed) indicator functions or suitable generalizations thereof.

1.2.2 Shape Representations and Priors

The situation concerning non-local regularizers such as shape priors is not yet as satisfactory.

Indicator Functions. Shape priors formulated directly in terms of indicator functions have the obvious advantage of being computationally compatible with the well established convex variational framework of image labelling. A rich variety of data terms and local regularizers is readily available. Such approaches are however often rather simplistic and lack important features such as geometric invariance [SS12, KC11, AMH11, EHW12]. This may be owed to the fact that the linear structure provided by the vector space of real functions is not quite suitable for describing shapes: for example the linear interpolation between two shapes is nowhere a shape itself (see Fig. 1.1). Also, L^p-type metrics are no good measures of shape similarity: they only measure the area of difference, regardless where the differing regions are. A notable exception to this limitation is [LBR08], which circumvents the lack of a suitable linear structure and flexibility through vast numbers and smart optimization.

Parametrized Contours. Somewhat complimentary is the representation of regions by their outline contours. The set of parametrized contours can be treated as an infinite dimensional manifold [KM97]. This manifold, equipped with various metrics, has been studied theoretically in great depth in the context of shape analysis [MM06, YMSM08, SMSY11]. Local approximation by its tangent space yields a natural linear structure for applying machine learning

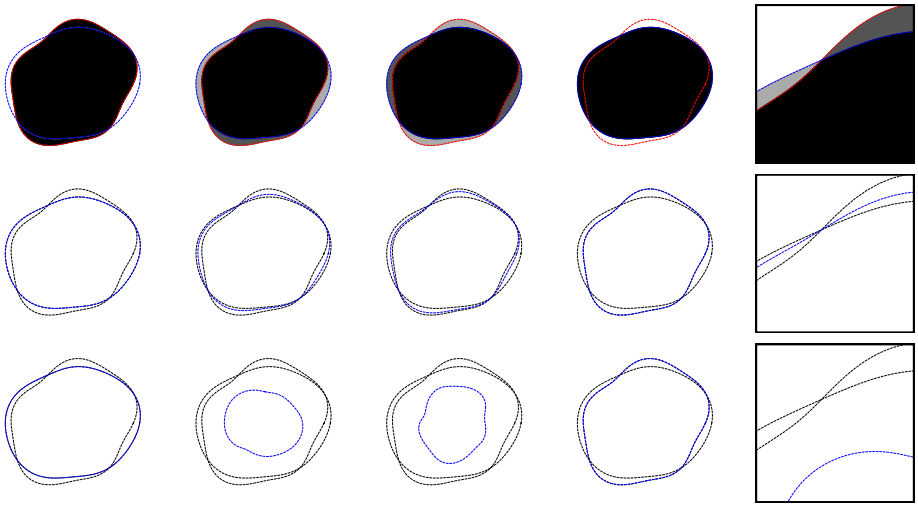


Figure 1.1: Linear structure on shape representations. From left to right: first four columns show a linear interpolation between two shapes in different representations. Fifth column shows a close-up of the second step for better visibility. **First row:** indicator functions (contours given for visual orientation). The intermediate objects are no indicator functions themselves. **Second row:** parametrized contours with compatible parametrizations. Blue contour describes a smooth interpolation between the two shapes given in black. **Third row:** similar contours with strongly differing parametrizations. The intermediate blue contours describe no meaningful interpolation.

techniques for elaborate statistical modelling [CKS03, CFK07]. But this representation, too, has its disadvantages: the parametrization ambiguity, while eliminated in theory by resorting to a suitable quotient manifold, remains a practical problem in implementations (see Fig. 1.1). In particular, it is more complicated as with indicator functions to evaluate and optimize the contour w.r.t. local image appearance features. Usually, internally the contour has to be converted into the corresponding region representation to evaluate the functional and optimization is only performed w.r.t. small local updates, resulting in non-convex models that are prone to get stuck in suboptimal local minima (e.g. [CKS03]).

Contour manifolds are closely related [MM07] to diffeomorphism groups [Tro98, TY05, You10], which are a popular tool in medical imaging [BMR13]. The relation between diffeomorphism groups and indicator functions is studied in [DZ11]. The difference in the natural linear structures on parametrized contours and indicator functions is illustrated in Figure 1.1.

Pictorial Structures. Another shape representation, which is popular especially in computer vision applications, are pictorial structures [FH05, KTZ05,

YV13]. These usually consist of a set of object parts and a list of parameters specifying their relative positions, thus being a rather compact and low-dimensional description. Although the functionals based thereon are usually non-convex, the low dimensionality helps in devising local [KTZ05] or clever global [YV13, FH05] optimization schemes, albeit possibly under certain modelling restrictions (e.g. tree graphs on the object parts).

Signatures. In shape classification and recognition tasks *signatures* haven proven to be a useful concept. By this term one usually refers to more compact representations that one computes from the 'actual' full shape representations, which are then easier to compare. Often such signatures are designed to be 'blind' w.r.t. certain classes of transformations and thus enable simple isometry invariant methods. Examples for signatures are Laplacian eigenvalues [RWP06], distributions of distances [OFCD02] and distributions of distributions of distances [RPC10]. Although some signatures uniquely determine a given object (up to the split off transformations), see e.g. [Kac66], it is generally difficult to reconstruct an object, given its signature. Thus, for shape segmentation the signature approach seems unsuitable.

1.2.3 Shape Matching

The problem of shape matching asks to find an optimal registration between two given shapes. For two objects of the same class this usually implies that the corresponding parts should be assigned to each other. This is the basis for applications such as shape interpolation, statistical analysis of sets of shapes [BFH+13] and for constructing similarity measures between shapes [Mém11], which can then again be used for classification and recognition.

Functionals for shape matching have been proposed based on physical deformation energies [HRWW12] or the notion of metric fidelity [EK03, MS05, BBK06]. An important property of these functionals is that the computed matchings are invariant under applying (non-rigid) isometries to either of the two objects. Consequently the implied optimization problems are computationally very demanding. They resemble the (bottleneck) quadratic assignment problem [Mém11], involve explicit optimization over the set of interpolating paths [RW] or very large linear programs [WSSC11], based on discrete exterior calculus [DHLM05].

Within this group we want to call particular attention to the framework of [Mém11], based on representing shapes by metric measure spaces (mm-spaces) and measuring their similarity with the Gromov-Hausdorff / Wasserstein distances [Gro07]. This representation is very general, encompassing indicator functions, super-pixel segmentations, point clouds, meshes and more abstract object such as graphs. The comparison is based purely on intrinsic metric information. Functionals based on physical deformation require existence of a

common ambient metric space into which the two compared objects can be embedded. For application on real-world objects this is certainly a natural assumption. The intrinsic methods in [Mém11] however require no such space and can thus also be applied to more abstract metric spaces with no natural common ambient space, such as graphs.

This literature impressively illustrates the power of shape matching for measuring similarity between shapes, and is thus very inspiring for the design of shape priors. But the computational complexity is already very high and clever optimization schemes are already required for practical application. Thus, when used naïvely, building a search for the optimal segmentation on top of the matching procedure will definitely render the methods unfeasible.

1.2.4 Optimal Transport

Optimal transport (OT) and the particular case of the linear assignment problem (LAP) are closely related optimization problems, dating back as far as the 18^{th} century. Their mathematical structure has been studied in great depth, see for example [RR98, Vil03, Vil09] for monographs on OT and [KV12, Sch03] for the LAP. A brief historical overview on optimal transport can be found in [Vil09, Chap. 3]. Today they are widespread tools in machine learning and image analysis, but they also allow modelling of problems with an economic context [Kuh55].

By explicitly taking into account a cost function on the base space, OT yields a distance on measures which is robust to quantization noise. This has proven very useful in computer vision applications [RTG00, PW09]. It has been successfully used to compute smooth interpolations between gray level images [HZTA04], to regularize densities [BFS12], for shape recognition [GM00, RPC10], and for texture analysis [FXPA12].

Computing meaningful registrations with standard OT requires alignment of the measures. To overcome this constraint, extension to an application dependent low-dimensional set of transformations was proposed in [CG99]. In [PT13] infinitesimal transformations have been incorporated in a linear programming formulation, allowing discrimination between systematic quantization errors and noise. Also an extension to measures of unequal mass was given [PW08].

The particular case of OT w.r.t. powers of the Euclidean distance gives rise to the rich structure of Wasserstein spaces [Vil09]. A fascinating facet is the differential geometric perspective of the 2-Wasserstein space on $\mathbb{R}^n$. Several partial differential equations can be interpreted as gradient flows therein [Ott01]. This led to viewing these spaces as Riemannian manifolds [Vil03, Lot08, AG13], which reveals connections between OT and fluid dynamics [BB97].

An interesting application of the Riemannian structure is given in [WSB+12], where a set of sample densities is projected into the appropriate tangent space to

exploit its linear structure for statistical analysis by standard machine learning techniques.

1.2.5 Solvers for Optimal Transport

Despite all its merits as a metric on measures and more generally as a modelling tool, optimal transport has the disadvantage of being computationally more expensive than simpler comparisons like L^p-distances, χ^2-like statistics or Mahalanobis distances. Thus, efficient solvers are critical for application. Optimal transport and the linear assignment problem can both be formulated as linear programs (LP) and thus they can in principle be solved by any LP solver. But of course there are also specialized algorithms available that exploit the particular structure of the problems.

Linear Assignment Problem. Famous 'classical' examples for algorithms to solve the linear assignment problem are the Hungarian method [Kuh55] and the 'Blossom algorithm' [Edm65], which can solve the linear assignment problem in at most $\mathcal{O}(N^3)$ steps, with N being the number of elements per set [Sch03]. See [Kol09] for a recent efficient implementation.

Another famous solver for the LAP is the auction algorithm by Bertsekas [Ber79], which is apt for parallelized implementation [Ber88], can be generalized to optimal transport [BC89] and which spawned development of more general min-cost flow algorithms [BE88]. A common feature of those algorithms is, that elements of one set submit 'bids' for elements in the other to achieve an assignment, somewhat analogous to an auction, hence the name.

Optimal Transport on 2-Wasserstein spaces in $\mathbb{R}^n$. For optimal transport in 2-Wasserstein spaces on $\mathbb{R}^n$ (see Sect. 2.4.1.3) there are dedicated solving algorithms based on the non-linear Monge formulation with the advantage that one need not handle a dense mass assignment matrix [HZTA04, CGS10]. These exploit the polar factorization theorem [Bre91], see also [Vil03, Chap. 3]. In [Mér11] an efficient multi-scale algorithm is given.

Approximate Methods. To save computation time, a lot of work has also been invested in approximation schemes. Proposed methods include equivalent metrics based on wavelets [SJ08], cost function thresholding [PW09], linearized approximation on the tangent space [WSB+12], and entropy regularization [Cut13].

1.3 Outline and Contribution

In preliminary work [SS12], state of the art techniques were used to directly incorporate a shape prior learned from examples into a convex variational seg-

mentation approach based on Markov random fields. The price to be paid for convexity was lack of invariance of the shape prior. This motivated the central topic of this thesis, the development of isometry invariant shape priors for object segmentation that are compatible with the convex variational framework for image labelling. This means the prior functionals should either be convex or global optimization of functionals including non-convex priors should still be feasible. Inspired by the success of shape matching as a foundation for similarity measures, the constructed priors will be based on matching for which we use optimal transport as computational tool. Hierarchical coarse to fine schemes will play an important role in this work, to find global optima of non-convex functionals and to make optimal transport computationally more efficient. The broad range of topics from the analysis of shape manifolds to efficient numerical implementation of combinatorial optimization algorithms is reflected in the breadth of mathematical areas that this work touches upon.

The main contributions are:

- Two approaches to construct isometry invariant shape priors are presented, based on matching a template onto the foreground of the image. Shapes are represented by metric measure spaces, thus ensuring compatibility with segmentation functionals based on indicator functions.

- The shape measure representation is introduced. This special class of mm-spaces forms a Riemannian manifold which is shown to be diffeomorphic to the manifold of closed simple contours modulo parametrization. This allows the combination of the strengths of the shape representations as contours and indicator functions and consequently the design of advantageous shape priors.

- A hierarchical sparse / dense hybrid variant of the auction algorithm for optimal transport is developed, that can solve optimal transport problems to global optimality by considering only a small fraction of all potential assignment pairs and thus achieving a significant decrease in run-time on typical practical problems. This is not only important for efficient application of the two developed shape prior functionals but applies to optimal transport problems in general.

The following thesis is organized as follows: In **Chapter 2** the necessary mathematical background is established. We touch upon supervised image segmentation, the linear and quadratic assignment problem, optimal transport and in particular the Riemannian structure of the 2-Wasserstein space, isometry invariant shape matching with the Gromov-Hausdorff / -Wasserstein distances and some basic facts on contour manifolds, flow-fields, their induced diffeomorphisms and Poisson's equation.

Chapter 3 starts with the first approach to design an isometry invariant shape prior, based on a convex relaxation of the Gromov-Wasserstein distance.

According to the nomenclature established in Sect. 2.5.2.3 this is an *intrinsic* method. Combination of the shape prior with other convex functionals is discussed. In particular it is shown for a discrete setting, that combination of the shape prior with a linear appearance term implies existence of a binary optimal segmentation and a bijective assignment between image foreground and shape template. Some numerical experiments are given and the breakdown point of the relaxation is illustrated, when optimal segmentation and assignment become spatially very irregular. In the second part of the chapter a method for tightening the relaxation is discussed. This method applies to general instances of the quadratic assignment problem. The benefit is demonstrated numerically.

The second approach is presented in **Chapter 4**. In the first part the set of shape measures is introduced and their Riemannian structure, as a 'submanifold' of the 2-Wasserstein space on $\mathbb{R}^2$ is pointed out. It is then established that this structure is in fact a manifold in the precise sense of [KM97] and that it is diffeomorphic to the manifold of closed simple contours modulo parametrization. In the second part of the chapter a shape prior based on the tangent space approximation of the shape measure manifold is constructed. This allows for explicit modelling of isometries (thus making it an *extrinsic* approach, according to Sect. 2.5.2.3) as well as statistical variations. The resulting overall functional is non-convex, but since non-convexity is confined to a low-dimensional variable, this can still be optimized to global optimality through hierarchical relaxations and branch & bound.

In **Chapter 5** a sparse / dense hybrid variant of the auction algorithm is developed that solves large dense optimal transport problems to global optimality while only explicitly considering a sparse subset of assignment pairs. Equivalence with the dense problem is ensured via hierarchical consistency checks. Proper initializations for the initial sparse subsets are obtained via a coarse-to-fine multi-scale scheme. We show that the worst-case computational complexity of the algorithm is only slightly increased by the proposed alternations but illustrate on realistic test problems that the average runtime is significantly decreased.

1.4 Notation

An overview on mathematical notation will be given, sorted by area.

Calculus. The space of test functions $\mathcal{D} = C_0^\infty(\mathbb{R}^2)$ is the space of smooth real functions on $\mathbb{R}^2$ with compact support. For a multi-index $I = (i_1, i_2, \dots, i_n)$ denote by $\partial_I \varphi = \partial_{i_1} \partial_{i_2} \dots \partial_{i_n} \varphi$ the corresponding partial derivative and by $|I| = n$ its order. Given a differentiable map $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$, we write J_φ for the Jacobian matrix.

Measures. For a measurable set $Y \subset \mathbb{R}^n$ we denote by $\mathcal{L}_Y$ the restriction of the Lebesgue measure to Y. In some cases, when a discrete Y approximates a continuous domain, $\mathcal{L}_Y$ gives the discrete approximation of the Lebesgue measure, i.e. an area weighted counting measure. The volume of a measurable set $\Omega \subset \mathbb{R}^2$ w.r.t. the Lebesgue measure is denoted by $|\Omega|$. For a discrete set A by $|A|$ we denote its cardinality. For a measurable space A denote by $\mathrm{Meas}(A)$ the set of non-negative measures on A and by $\mathrm{Prob}(A)$ the set of probability measures. For a measurable map $f : A \to B$ by $f_\sharp\mu$ we denote the push-forward of measure μ from A to B, defined by $(f_\sharp\mu)(\sigma) = \mu(f^{-1}(\sigma))$ for all measurable $\sigma \subset B$. We write $\operatorname{spt}\mu$ for the support of the measure μ, which is the smallest closed set such that $\mu(A \setminus \operatorname{spt}\mu) = 0$. Throughout this dissertation we will assume that all measures on $\mathbb{R}^2$ are absolutely continuous, i.e. they yield zero on any Lebesgue negligible set. This is equivalent to the existence of a locally integrable map ρ such that

$$\mu(A) = \int_A \rho \, dx \,.$$

ρ is called density of μ and is unique Lebesgue almost everywhere.

Differential Geometry, Contour Manifolds and Shape Measures. For a smooth manifold M we denote by TM its tangent bundle and for $x \in M$ by T_xM the tangent space at x. Generally for any fiber bundle $\mathcal{F}M$ over M we denote the fibre at x by $\mathcal{F}_xM$. For a differentiable map $f : M \to N$ between two smooth manifolds, denote by $\mathcal{D}f : TM \to TN$ the differential.

Sobolev Spaces. Denote by $H(\mathrm{div}, \Omega)$ the space of square-integrable vector fields $u : \Omega \to \mathbb{R}^2$, with square integrable divergence

$$\operatorname{div} u = \nabla \cdot u = \sum_{i=1}^{2} \frac{\partial}{\partial x_i} u_i \,.$$

This is a Hilbert space with scalar product and norm

$$\langle u, v \rangle_{\mathrm{div};\Omega} = \int_\Omega u \cdot v \, dx + \int_\Omega (\operatorname{div} u)(\operatorname{div} v) \, dx, \qquad \|u\|^2_{\mathrm{div};\Omega} = \langle u, u \rangle_{\mathrm{div};\Omega} \,.$$

For some Sobolev space W, by $[u]$ we denote the equivalence class of functions $u \in W$ that only differ by a constant. They form a unique element of the quotient space $W/\mathbb{R}$. The corresponding norm is given by

$$\|[u]\|_{W/\mathbb{R}} = \inf\{\|u\|_W : u \in [u]\} \,.$$

Miscellaneous and Collisions. By $\mathcal{O}(\dots)$ we denote the Landau Big O symbol for asymptotic scaling. pr_A denotes the projection map onto the set A.

Due to the variety of different fields, it was not possible to completely avoid collision in nomenclature throughout this dissertation. However, these collisions only occur for vastly different objects, thus the meaning of each symbol should still be unambiguous when considering the context. The most prominent cases are:

- The symbol α refers to both a dual variable for linear assignment and optimal transport, as well as to tangent vectors on the manifold of shape measures.
- a refers to normal deformation fields that represent tangent vectors on the manifold of contours, as well as to assignment fields, as introduced in the context of QAP diffusion (Sect. 3.2).
- The symbol c refers to both cost functions in linear assignment and optimal transport problems, as well as to parametrized smooth contours.

Chapter 2

Mathematical Background

2.1 Supervised Image Segmentation

The initial motivation for the work presented in this thesis comes from the problem of supervised image segmentation or labelling. We provide a brief review to obtain the necessary background for our work. For more details see for example [LS11].

2.1.1 The Generic Functional

The problem consists of assigning to each pixel of an image a label from a pre-specified set (for example {'gras', 'house', 'sky', ...}). A typical functional for such a task looks like this [LS11, Sect. 1]: Let Ω be the image domain (either described by a discrete set of pixels or by a continuous subset of $\mathbb{R}^2$) and $\{1, \dots, l\}$ the set of labels. We are looking for a function $\ell : \Omega \to \{1, \dots, l\}$ that minimizes an energy of the following type:

$$E(\ell) = \int_\Omega s\big(x, \ell(x)\big)\, dx + R(\ell) \tag{2.1.1}$$

The first term is usually referred to as *data term* and the second as *regularizer*. The value $s\big(x, \ell(x)\big)$ describes how well the label $\ell(x)$ fits to the pixel x according to local appearance information (for a review on local image features see for example [CRD07]). Without a regularizer each pixel would choose its best label purely on the data term. This is obviously highly sensitive to noise. The regularizer introduces additional prior knowledge into the problem: in many applications it is designed to encourage smooth boundaries between regions of different labels or to penalize unwanted transitions between certain labels (e.g. 'house' cannot be above 'sky') to at least partially compensate for noise in the data term (cf. Sect. 1.2.1).

2.1.2 Convex Relaxation

The problem of minimizing (2.1.1) is combinatorial and hence in general not feasible in this formulation. A lot of attention has been focussed on finding good convex relaxations via which one can find optimal or at least approximately optimal solutions.

One important step is to convexify the feasible set. This can be achieved by the simplex relaxation [LS11, Sect. 1]. First one identifies the i-th label with the i-th unit vector $e_i \in \mathbb{R}^l$. One can then rewrite (2.1.1) as

$$E(u) = \int_\Omega \langle s(x), u(x) \rangle \, dx + R(u) \tag{2.1.2}$$

where $u : \Omega \to \{e_1, \dots, e_l\}$ and $s(x) = \big(s(x,1), \dots, s(x,l)\big)^\top$.

The feasible set can then be convexified by allowing u to take values in the convex hull of the unit vectors which is the l-dimensional unit simplex

$$\Delta_l = \operatorname{conv}\big(\{e_1, \dots, e_l\}\big) = \left\{ v \in \mathbb{R}^l \,:\, v \geq 0 \wedge \sum_{i=1}^{l} v_i = 1 \right\}. \tag{2.1.3}$$

Remark 2.1.1. (Linear data term) Note that by replacing the labelling function ℓ in (2.1.1) with the vector valued function u in (2.1.2) the data term became linear in u, regardless of the original function $s\big(x, \ell(x)\big)$. Hence, even complicated appearance models can be described by a linear data term.

A regularizer that has been studied very intensely is *total variation* [ROF92] (cf. Sect. 1.2.1). For a differentiable scalar function u on a domain $\Omega \in \mathbb{R}^n$ it is given by

$$\mathrm{TV}(u) = \int_\Omega \|\nabla u(x)\| \, dx \,. \tag{2.1.4}$$

An extension to non-smooth u is given via the dual of the norm [LS11]:

$$\mathrm{TV}(u) = \sup \left\{ \int_\Omega u \cdot \operatorname{div} v \, dx \,:\, v \in C_c^\infty(\Omega, \mathbb{R}^n) \text{ s.t. } \|v(x)\| \leq 1 \; \forall \, x \in \Omega \right\} \tag{2.1.5}$$

This reduces to (2.1.4) for differentiable u.

For more than two labels problem (2.1.1) is usually NP-hard [BVZ01], so the convex relaxations cannot be expected to be tight. But the particular case of two labels has been shown to be solvable exactly by means of convex relaxation for R being the total variation, both in a discrete [KZ04] and a continuous setting [CEN06]. For the application of object segmentation two classes are sufficient: fore- and background. Then the simplex relaxation is equivalent to

using a continuous function $u : \Omega \to [0, 1]$ with 0 corresponding to one label and 1 to the other. A corresponding energy reads

$$E(u) = \int_\Omega s(x)\, u(x)\, dx + \mathrm{TV}(u)\,, \tag{2.1.6}$$

to be optimized subject to the constraint

$$0 \leq u(x) \leq 1 \text{ for all } x \in \Omega\,. \tag{2.1.7}$$

We will, in the course of this dissertation, construct regularizers based on the notion of optimal transport (Sect. 2.4). Thus, we will interpret u as the density of a measure w.r.t. the Lebesgue measure on Ω and consequently reformulate the box constraint (2.1.7) for u directly in terms of measures. For compatibility we will sometimes need to constrain the mass of this measure to be fixed. We define:

Definition 2.1.2 (Segmentation measures)**.** For a space Z, a fixed mass $M > 0$ and a non-negative reference measure $\mathcal{L}_Z$ on Z we define the set of *segmentation measures*

$$\begin{aligned}\mathrm{SegMeas}(Z, M) = \{\nu \in \mathrm{Meas}(Z) :\ & \nu(Z) = M \wedge \nu(\sigma_Z) \leq \mathcal{L}_Z(\sigma_Z) \\ & \text{for all measurable } \sigma_Z \subset Z\}\,.\end{aligned} \tag{2.1.8}$$

For $Z \subset \mathbb{R}^n$ the reference measure $\mathcal{L}_Z$ will usually be the Lebesgue measure, for discrete sets it will be a weighted counting measure. Segmentation measures will be the feasible sets in several of the variational approaches we introduce later on.

2.2 The Linear Assignment Problem

2.2.1 Definition and Basic Properties

The *(linear) assignment problem* (LAP) [KV12, Sch03] is a fundamental class of optimization problems, both interesting as a theoretical object of study and as a practical tool for modelling in various fields. It is an important tool for shape matching in the work that this thesis is based upon. Let $X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_N\}$ be two sets of equal finite cardinality $|X| = |Y| = N$ and let $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ be a cost function where $c(x, y)$ gives the cost of assigning element x to y, a value $c(x, y) = \infty$ indicating that this particular assignment is impossible altogether.

Let

$$\mathcal{N} = \{(x, y) \in X \times Y :\ c(x, y) < \infty\}\,. \tag{2.2.1}$$

We call $\mathcal{N}$ the set of neighbours and write

$$\mathcal{N}(x) = \{y \in Y \,:\, (x, y) \in \mathcal{N}\} \tag{2.2.2}$$

and similarly $\mathcal{N}(y)$.

The LAP asks to find the one-to-one assignment between X and Y that inflicts the minimal total assignment cost. One-to-one assignments between two sets of equal cardinality correspond to permutations. Hence, the LAP can be formulated as an optimization problem over the set Perm_N of permutation matrices of dimension $N \times N$:

$$\min \left\{ \sum_{i,j=1}^{N} c(x_i, y_j)\, P_{ij} : P \in \mathrm{Perm}_N \right\} \tag{2.2.3}$$

We call a problem *feasible* if there is at least one permutation P with a finite cost.

2.2.1.1 Linear Programming Formulation

The LAP, as stated in (2.2.3), is a combinatorial optimization problem over a discrete set of permutations. We will now consider a reformulation as a linear program based on doubly stochastic matrices and the Birkhoff-von-Neumann theorem. Recall that the set of doubly stochastic matrices of dimension $N \times N$, denoted by Stoch_N, is defined by

$$\begin{aligned} \mathrm{Stoch}_N = \Bigg\{ P \in \mathbb{R}^{N \times N} \,:\, & P_{ij} \geq 0 \;\forall\, 1 \leq i, j \leq N \\ & \wedge \sum_{i=1}^{N} P_{ij} = 1 \;\forall\, 1 \leq j \leq N \\ & \wedge \sum_{j=1}^{N} P_{ij} = 1 \;\forall\, 1 \leq i \leq N \Bigg\}. \end{aligned} \tag{2.2.4}$$

Theorem 2.2.1 (Birkhoff-von-Neumann Theorem [KV12, Cor. 11.5])**.** *Any doubly stochastic matrix is a convex combination of permutation matrices.*

Consider then optimizing (2.2.3) over Stoch_N instead of Perm_N:

$$\inf \left\{ \sum_{i,j=1}^{N} c(x_i, y_j)\, P_{ij} : P \in \mathrm{Stoch}_N \right\} \tag{2.2.5}$$

From Theorem 2.2.1 immediately follows:

Corollary 2.2.2 (Relation of (2.2.3) and (2.2.5))**.** *The problems* (2.2.3) *and* (2.2.5) *have the same optimal value. The linear program* (2.2.5) *always has an optimizer which is a permutation matrix.*

Remark 2.2.3 (Existence of non-binary optimizers)**.** However, in the case that there are multiple optimal permutations for (2.2.3), any convex combination of the optimal permutation matrices, which need not be a permutation matrix itself, will also be an optimizer to (2.2.5).

2.2.1.2 Dual Problem

Rewriting the LAP as a linear program allows to derive a corresponding dual problem which is given by

$$\sup \left\{ \sum_{i=1}^{N} \alpha(x_i) + \sum_{j=1}^{N} \beta(y_j) \; : \; \alpha : X \to \mathbb{R},\, \beta : Y \to \mathbb{R} \right.$$
$$\left. \text{s.t. } \alpha(x) + \beta(y) \leq c(x,y) \; \forall \; (x,y) \in X \times Y \right\} . \tag{2.2.6}$$

Note that for fixed β the best choice for a corresponding α is given by

$$\alpha(x) = \min\{c(x,y) - \beta(y) \; : \; y \in \mathcal{N}(x)\} . \tag{2.2.7}$$

Duality implies that for any feasible matrix $P \in \text{Stoch}_n$ of (2.2.5) and any feasible pair of functions (α, β) of (2.2.6) one has

$$\sum_{i,j=1}^{N} c(x_i, y_j)\, P_{ij} \geq \sum_{i=1}^{N} \alpha(x_i) + \sum_{j=1}^{N} \beta(y_j) \tag{2.2.8}$$

with equality when P and (α, β) are primal, resp. dual optimizers. Further, for any primal optimizer P and dual optimizers (α, β) one has

$$P_{ij} > 0 \Rightarrow \alpha(x_i) + \beta(y_j) = c(x_i, y_j) . \tag{2.2.9}$$

This relation is called *complimentary slackness.*

2.2.1.3 Solvers for the LAP

The formulation of the LAP as a linear program allows in principle to apply any linear programming algorithm for its solution. This can be very convenient since such algorithms are available on many platforms. However, there are certain limitations: first, a general linear programming algorithm may not be able to exploit the particular structure of the constraints. This can be critical for the performance on large problems. Second, when there are multiple optimal

permutations to (2.2.3), an interior point method may not yield a binary solution in Perm_N but some non-binary convex-combination (see Remark 2.2.3) which does not immediately imply a one-to-one assignment between the two sets.

As mentioned in Sect. 1.2.5 there are also dedicated combinatorial solvers to the LAP that heavily exploit the particular problem structure, such as the Hungarian method and the auction algorithm. In Sect. 2.4.3 we will discuss an extension of the auction algorithm for more general optimal transport problems and in Chapter 5 we will introduce a modified version of the auction algorithm to increase performance on typical optimal transport and linear assignment problems. Therefore, we will in the next Section present the basic version of the auction algorithm.

2.2.2 Auction Algorithm

One possible method for solving linear assignment problems is the auction algorithm [Ber79]. Details on the intuitive motivation, descriptions and analysis of the properties of the various forms of the algorithm are given in the references listed in Sect. 1.2.5. Here we give the description from [Ber88, Sect. 2].

2.2.2.1 Sign Conventions

In the literature the LAP is sometimes seen as minimizing the assignment cost between two sets and sometimes as maximizing an assignment score. Naturally both conventions are equivalent and connected by sign inversion. We choose the viewpoint of minimizing an assignment cost to keep signs compatible with the notion of optimal transport. Hence, note that we flipped the signs relative to the original presentation of the algorithm. In the following the comparison to an auction is no longer very intuitive because the lowest bid gets accepted.

2.2.2.2 Primal-Dual Method

The auction algorithm is a primal-dual method since it works with both variables for the primal as well as the dual problem. The primal variable P will be taken from the set of 'incomplete permutation matrices' Perm'_N:

$$\mathrm{Perm}'_N = \left\{ P \in \{0,1\}^{N\times N} \;:\; \sum_{i=1}^{N} P_{ij} \leq 1 \;\forall\, 1 \leq j \leq N \right. \\ \left. \wedge \sum_{j=1}^{N} P_{ij} \leq 1 \;\forall\, 1 \leq i \leq N \right\}. \tag{2.2.10}$$

This is equivalent to the notion of *assignment* defined in [Ber88, Sect. 2]. Note that Perm'_N includes matrices that are primal feasible in neither (2.2.3) nor

(2.2.5). P will be initialized with all entries being 0. During the iterations of the algorithm assignments between elements of X and Y are added and removed (i.e. the corresponding entries of P will be set to 1 or 0 respectively). The algorithm will terminate when P is primal feasible and therefore corresponds to a complete one-to-one assignment between X and Y.

Additionaly, the state of the algorithm will be represented by a dual variable $\beta : Y \to \mathbb{R}$ (Sect. 2.2.1.2). The second dual variable α will be held implicitly via (2.2.7), i.e. the dual variables will by construction always be dual feasible. The initialization of β is arbitrary, reasonable choices and dependency of the algorithm runtime on the initialization are briefly discussed in Sect. 2.2.2.6.

2.2.2.3 ε-Complimentary Slackness

A key property of the auction algorithm is that the complimentary slackness condition (2.2.9) does not hold strictly throughout the iterations and on termination. Instead, the weaker condition

$$P_{ij} > 0 \Rightarrow \alpha(x_i) + \beta(y_j) \geq c(x_i, y_j) - \varepsilon \tag{2.2.11}$$

called *ε-complimentary slackness* for a parameter $\varepsilon > 0$ is satisfied throughout execution [Ber88, Sect. 2]. Positivity of ε is essential for convergence of the algorithm. As long as $\varepsilon < \delta c/N$ the primal variable $P \in \mathrm{Perm}_N$, that the algorithm returns upon termination, will be an optimizer of (2.2.3). Here, δc is the smallest absolute difference between to distinct values of c. A proof for this is given for example in [Ber88, Prop. 1]. There $\varepsilon < 1/N$ is required for optimality and it was assumed that c only takes integer values to ensure that the cost difference between an optimal permutation and a suboptimal one is at least 1. This generalizes directly to the case with $\varepsilon < \delta c/N$ discussed here.

2.2.2.4 Main Loop

The main loop of the algorithm consists of two stages: *bidding* and *assignment*. During the bidding phase elements of X locally determine their most suitable assignment partner in Y and propose a corresponding dual variable change. After that, during the assignment phase, for each $y \in Y$ the best proposed dual variable change and the corresponding assignment in P is implemented. The two stages will be iterated until P is a complete permutation. If the problem is feasible the algorithm will terminate after a finite number of iterations [Ber88, Prop. 2]. Different x do not interact during the bidding phase and neither do different y during the assignment phase. Thus both stages can be easily parallelized. We will now discuss both stages in detail.

Bidding Phase. For every $x \in X$ that is unassigned under P (i.e. the corresponding row of P contains only 0s):

Compute the corresponding value of $\alpha(x)$ as given by (2.2.7):

$$\alpha(x) = \min\{c(x,y) - \beta(y) : \ y \in \mathcal{N}(x)\} \tag{2.2.12}$$

and find a minimizer y^*. Determine also the slack of the 'second nearest' constraint:

$$\alpha'(x) = \min\{c(x,y) - \beta(y) : \ y \in \mathcal{N}(x) \setminus \{y^*\}\} \tag{2.2.13}$$

Then element $x \in X$ submits a bid for element $y^* \in Y$ with value

$$b_{xy^*} = c(x,y^*) - \alpha'(x) - \varepsilon\,. \tag{2.2.14}$$

Assignment Phase. For each $y \in Y$ let $\mathcal{B}(y)$ be the set of $x \in X$ from which y received a bid in the bidding phase of the iteration. If $\mathcal{B}(y)$ is nonempty, decrease $\beta(y)$ to the lowest bid

$$\beta(y) := \min_{x \in \mathcal{B}(y)} b_{xy}\,, \tag{2.2.15}$$

set the column in P corresponding to y to 0 and the entry in P corresponding to the assignment (x^*, y) to 1, where x^* is some element in $\mathcal{B}(y)$ attaining the minimum in (2.2.15). If $\mathcal{B}(y)$ is empty, $\beta(y)$ is left unchanged.

2.2.2.5 Complexity Analysis

For the application of the algorithm one is interested in its average computational complexity. This is very difficult to estimate theoretically. Instead usually a worst case bound is given and evaluation of the average runtime is done through numerical experiments.

The details of a theoretical complexity analysis depend on the considered variant of the auction algorithm. It usually consists of two parts: bounding the number of bids and estimating the cost per bid.

The maximum number of bids is typically estimated via the implied minimum decrease of the dual variable at the receiving node. For the more general minimum cost flow algorithm such an estimate is given by [BE88, Lemma 5], of which the sequential auction algorithm is a special case (Sect. 6, ibid.). This estimate bounds the number of accepted bids by

$$\mathcal{O}(N^2 \cdot C/\varepsilon + N^2) \tag{2.2.16}$$

with

$$C = \big(\max_{x,y} c(x,y)\big) - \big(\min_{x,y} c(x,y)\big)\,. \tag{2.2.17}$$

For the sequential variant the number of accepted bids is equal to the number of submitted bids. For the parallel variant discussed here, some submitted bids

may not be accounted for when multiple bids are submitted to the same $y \in Y$ but only one is accepted. Apart from the obvious benefit of potential parallel implementation, the parallel variant is also more amenable for the extensions discussed in Chapter 5.

For $\varepsilon < \delta c/N$ as required for global optimality (2.2.16) yields

$$\mathcal{O}(N^3 \cdot C/\delta c) \tag{2.2.18}$$

where we have only kept the dominant first term.

For dense problems the cost per bid is

$$\mathcal{O}(N) \tag{2.2.19}$$

since one must scan all $y \in Y$ to determine (2.2.12) and (2.2.13).

The estimated cost per bid (2.2.19) is accurate. The bound (2.2.16) on the number of bids is typically quite conservative. But one often observes a critical dependency of the number of bids on the ratio C/ε, causing long runtimes. A fix for this is discussed in the next section.

In Chapter 5 we will present an extension of the auction algorithm that attempts to reduce the average cost per bid.

2.2.2.6 Epsilon-Scaling

Estimate (2.2.16) is computed by bounding the maximal change in the dual variable $\beta(y)$ at a given $y \in Y$ and dividing this by the minimum amount of change of $\beta(y)$ per accepted bid. One quickly finds that upon accepting a bid the dual variable $\beta(y)$ of the receiving node decreases by at least ε.

It may happen that the algorithm gets stuck in a phase where β-changes only occur in the smallest possible size, a phenomenon dubbed *price haggling* [BE88]. For $\varepsilon < \delta c/N$, as required for global optimality, this can result in very long runtimes. *ε-scaling* has been proposed to alleviate this problem [BE88, Sect. 5] which we will briefly describe here.

When the auction algorithm is run with an initial β_{init} for which we know that a feasible primal variable exists for which ε-complimentary slackness holds for a parameter $\gamma > 0$, then the bound (2.2.16) can be refined to [BE88, Lemmata 4(a) and 5]

$$\mathcal{O}(N^2 \cdot \gamma/\varepsilon + N^2)\,. \tag{2.2.20}$$

Based on this refined bound one can make the algorithm more efficient by first running it with a large value for ε and successively re-running it with smaller values. Let θ be a given constant, in practice somewhere between 2 and 10. First the algorithm is run with $\varepsilon_0 = C/\theta$ and with some constant initial $\beta_{0,\text{init}}$. We know from (2.2.16) that the number of bids is bounded by $\mathcal{O}(N^2 \cdot \theta + N^2)$. The returned dual variable $\beta_{0,\text{fin}}$ and the assignment matrix

$P_{0,\text{fin}}$ satisfy the ε-complimentary slackness condition for $\varepsilon_0 = C/\theta$ which is probably not sufficient for guaranteeing global optimality. Instead, run the auction algorithm again with $\varepsilon_1 = \varepsilon_0/\theta$ and $\beta_{1,\text{init}} = \beta_{0,\text{fin}}$. Through (2.2.20) the number of bids is again bounded by $\mathcal{O}(N^2 \cdot \theta + N^2)$. By repeating this loop until $\varepsilon_i = \varepsilon_0/\theta^i < \delta c/N$ one can obtain a globally optimal solution. It is readily computed that one needs of the order

$$\mathcal{O}\big(\log(N \cdot C/\delta c)/\log\theta\big) \tag{2.2.21}$$

runs where $\mathcal{O}(\log\theta) = \mathcal{O}(\theta) = 1$. Summing up the number of bids over the iterations yields

$$\mathcal{O}\big(\log(N \cdot C/\delta c) \cdot N^2\big) \tag{2.2.22}$$

which is clearly more efficient than the unscaled approach (2.2.18). The effect of ε-scaling is not only relevant for decreasing the theoretical worst case complexity but is also critical for practical application.

Remark 2.2.4 (Primal variable initialization)**.** It seems like a waste of information always to initialize the auction algorithm with the empty assignment after reducing ε. With some effort one could certainly establish an at least partial assignment matrix P which satisfies ε-complimentary slackness with the new dual variable $\beta_{i,\text{init}}$ and ε_i. In [BE88, Sect. 5] a method to initialize the primal variable upon decreasing ε is discussed. However, we find in practice that the first iterations of the auction algorithm quickly restore an almost complete assignment from the suitably initialized $\beta_{i,\text{init}}$ and thus find empty initialization sufficient.

2.3 The Quadratic Assignment Problem

Although the LAP is a widespread tool that lends itself to modelling optimization problems in many fields, it is inherently limited in its applicability by the linearity of the cost function. Interaction between different assignments takes only place through the one-to-one constraint. We will now introduce the more general *quadratic assignment problem* (QAP). The QAP will be of interest to us for its relation to shape matching (Sect. 2.5). For a good introduction to the topic see for example [BÇPP98]. We summarize here what is relevant for this dissertation.

2.3.1 Definition and Basic Properties

2.3.1.1 The Koopmans-Beckmann QAP

The QAP was used by Koopmans and Beckmann in 1957 [KB57] as a model for describing the cost of assigning a set of plants or factories to a set of prespecified locations. There is a linear term, just as with the LAP, that models

the cost of building a given factory at a given location. But more importantly between the factories there are flows of goods and naturally between locations there are transport costs per unit of goods. The costs caused by transportation of goods between the sites is modelled by a quadratic term.

Similarly as with the LAP, let $X = \{x_1, \dots, x_N\}$ be the set of factories and $Y = \{y_1, \dots, y_N\}$ the set of sites. Let F_{ij} be the matrix of flows between factories x_i and x_j, let D_{ij} be the matrix of distances between the sites y_i and y_j, i.e. the transportation costs, and let B_{ij} be the cost of constructing factory x_i at site y_j. Again, we will describe an assignment between factories and sites by a permutation matrix. $P_{ij} = 1$ means factory x_i is built at site y_j. Finding the best assignment then corresponds to solving

$$\inf \left\{ \sum_{i,j,k,l=1}^{N} F_{ij}\, D_{kl}\, P_{ik}\, P_{jl} + \sum_{i,j=1}^{N} B_{ij} P_{ij} \;:\; P \in \mathrm{Perm}_N \right\}. \tag{2.3.1}$$

The first term is quadratic in P, the array of coefficients being constructed from the flow matrix F and the distance matrix D.

2.3.1.2 The Lawler QAP

More generally, one can just pick any array of coefficients as costs for pairs of assignments. This was proposed by Lawler [Law63]. For a cost array $d \in \mathbb{R}^{(N\times N)^2}$ the problem is given by

$$\inf \left\{ \sum_{i,j,k,l=1}^{N} d_{ijkl}\, P_{ij}\, P_{kl} \;:\; P \in \mathrm{Perm}_N \right\}. \tag{2.3.2}$$

Obviously by choosing $d_{ijkl} = F_{ik}\, D_{jl}$ one can describe the flow / transportation costs and by setting $d_{ijij} = B_{ij}$ one can incorporate linear costs. So any Koopmans-Beckmann QAP can be formulated as a Lawler QAP.

A variant of the QAP is the bottleneck QAP [BÇPP98, Sect. 13.1] where we do not ask for the total cost but for the maximal cost that is inflicted by one pair of assignments:

$$\inf \left\{ \max \left\{ d_{ijkl}\, P_{ij}\, P_{kl} \;:\; 1 \le i,j,k,l \le N \right\} \;:\; P \in \mathrm{Perm}_N \right\} \tag{2.3.3}$$

The Gromov-Wasserstein distances that we will introduce in Sect. 2.5 will be formally similar to the QAP and the bottleneck QAP.

2.3.1.3 Computational Complexity

Due to the non-linearity of the objective function, relaxing the QAP from permutation matrices to doubly stochastic matrices as with the LAP (Corollary 2.2.2) will not allow reformulation as a linear program analogous to (2.2.5).

In fact, the QAP is considered a 'very hard' problem. More precisely, it is strongly NP-hard and existence of an algorithm for finding an approximate solution which is sub-optimal by at most a pre-specified factor would imply $P = NP$ [BÇPP98, Thm. 3.2]. Several other difficult problems, such as the *travelling salesman problem* can be formulated as special cases of the QAP (Sect. 13.4, ibid.).

Hence, in particular for the problem sizes relevant in shape matching, approximate methods are indispensable. We will give a short overview on lower bounds in the next section.

2.3.2 Lower Bounds

Lower bounds for the QAP are used to construct exact branch and bound algorithms [BÇPP98, Sect. 7.1]. They can however be used in their own right to obtain approximate solutions. In [KÇCE99] a framework of constructing lower bounds from a linearized variant of the QAP is presented.

2.3.2.1 Linearization

First the QAP is rewritten as an *integer linear program* (ILP):

Definition 2.3.1 (Integer linear program formulation of the QAP [KÇCE99, Sect. 2])**.** An equivalent formulation of the QAP (2.3.2) is given by the following linear integer program:
Minimize:

$$\sum_{i,j=1}^{N} \sum_{k,l=1}^{N} d_{ijkl}\, P'_{ijkl} \tag{2.3.4a}$$

Subject to:

$$P \in \mathrm{Perm}_N \tag{2.3.4b}$$

$$\sum_{l=1}^{N} P'_{ijkl} = P_{ij} \qquad \forall\, 1 \le i,j,l \le N \tag{2.3.4c}$$

$$\sum_{k=1}^{N} P'_{ijkl} = P_{ij} \qquad \forall\, 1 \le i,j,k \le N \tag{2.3.4d}$$

$$P'_{ijkl} = P'_{klij} \qquad \forall\, 1 \le i,j,k,l \le N,\ i \cdot N + j < k \cdot N + l \tag{2.3.4e}$$

$$P'_{ijkl} \ge 0 \qquad \forall\, 1 \le i,j,k,l \le N \tag{2.3.4f}$$

$$P'_{ijij} = P_{ij} \qquad \forall\, 1 \le i,j \le N \tag{2.3.4g}$$

$$P'_{ijil} = 0 \qquad \forall\, 1 \le i,j,l \le N,\ j \ne l \tag{2.3.4h}$$

$$P'_{ijkj} = 0 \qquad \forall\, 1 \le i,j,k \le N,\ i \ne k \tag{2.3.4i}$$

Remark 2.3.2. Obviously the variables P'_{ijil} and P'_{ijkj} for $j \neq l$ and $i \neq k$, set to zero by constraints (2.3.4h, 2.3.4i) and the variables P'_{ijij} which are just copies of P_{ij} could be eliminated from the linear program, consequently the constraints (2.3.4c, 2.3.4d) could be dropped for $i = k$ and $j = l$. This will yield the formulation given in [KÇCE99], where we have absorbed the linear costs into the 'diagonal' of d.

2.3.2.2 The Gilmore-Lawler Bound

The well known Gilmore-Lawler bound can be obtained from (2.3.4) by dropping the constraints (2.3.4e), known as *complimentary constraints*.

Definition 2.3.3 (Gilmore-Lawler bound [KÇCE99, Sect. 2.2])**.** The Gilmore-Lawler bound for the QAP (2.3.2) is given by

$$\inf \left\{ \sum_{i,j=1}^{N} c_{ij} P_{ij} \ : \ P \in \mathrm{Perm}_N \right\} \tag{2.3.5a}$$

where

$$c_{ij} = \inf \left\{ \sum_{k,l=1}^{N} d_{ijkl} P'_{ijkl} \ : \ P'_{ij} \in \mathrm{Perm}_N,\ P'_{ijij} = 1 \right\}. \tag{2.3.5b}$$

Remark 2.3.4 (Relation to the linearized QAP and interpretation)**.** The equivalence with (2.3.4) minus the constraints (2.3.4e) can be seen rather directly: when computing c_{ij} we first fix the hypothetical assumption $P_{ij} = 1$ which turns constraints (2.3.4c, 2.3.4d, 2.3.4f) into the constraint $P'_{ij} \in \mathrm{Stoch}_N$, which is by virtue of Corollary 2.2.2 equivalent to $P'_{ij} \in \mathrm{Perm}_N$ and constraints (2.3.4g, 2.3.4h, 2.3.4i) into the constraint $P'_{ijij} = 1$. The final optimization in (2.3.5a) is then an optimization over the best set of consistent hypotheses of which $P_{ij} = 1$ are true.

Remark 2.3.5 (Complexity of the Gilmore-Lawler bound)**.** Given that the LAP can be solved in $\mathcal{O}(N^3)$ (Sect. 1.2.5) and the Gilmore-Lawler bound requires the computation of $\mathcal{O}(N^2)$ entries of c, its computational complexity for general d is $\mathcal{O}(N^5)$. However, for the particular case of the Koopmans-Beckmann QAP (2.3.1), the effective cost function entries c_{ij} can be computed by a simple sorting of lists, thus reducing the dominant contribution to the complexity to $\mathcal{O}(N^3)$. See [BÇPP98, Sect. 6.1 and Prop. 6.1] or [Gil62].

Remark 2.3.6 (A Dual Framework for QAP Bounds)**.** Usually one observes that the tightness of the Gilmore-Lawler bound quickly decreases with growing problem size. Hence, tighter bounds have been developed by many researchers. In [KÇCE99] it is shown that many previously presented bounds, including the GLB itself, can be viewed within a unified framework as approaches to approximately solving a dual of the continuous relaxation of (2.3.4).

2.4 Optimal Transport

2.4.1 Definition and Basic Properties

The problem of optimal transport (OT) is a natural generalization of the linear assignment problem. Let X and Y be two measurable spaces and let $\mu \in \mathrm{Prob}(X), \nu \in \mathrm{Prob}(Y)$ be two probability measures on X and Y respectively. Let further $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ be a function where $c(x, y)$ gives the cost of transporting one unit of mass from x to y. Just as with the LAP a value $c(x, y) = \infty$ indicates that no mass may be transported from x to y. OT asks for finding a *transport plan* for the mass from (X, μ) onto (Y, ν) that causes the least possible total transportation cost. Such plans are described by measures on the product space $X \times Y$:

$$\Pi(\mu, \nu) = \Big\{ \pi \in \mathrm{Prob}(X \times Y) \ : \ \pi(\sigma_X \times Y) = \mu(\sigma_X) \\ \wedge\, \pi(X \times \sigma_Y) = \nu(\sigma_Y) \\ \text{for all measurable } \sigma_X \subset X,\, \sigma_Y \subset Y \Big\} \tag{2.4.1}$$

The set $\Pi(\mu, \nu)$ is referred to as *transport plans* or *couplings* between μ and ν. It always contains at least the product measure of μ and ν. For any $\pi \in \Pi(\mu, \nu)$ the weight $\pi(\sigma_X \times \sigma_Y)$ tells how much mass is transported from $\sigma_X \subset X$ to $\sigma_Y \subset Y$. The total transportation cost associated with a coupling π is given by integration of c w.r.t. π:

$$\int_{X \times Y} c(x, y)\, d\pi(x, y) \tag{2.4.2}$$

Hence, the optimal transport problem can be written as

$$\inf \left\{ \int_{X \times Y} c(x, y)\, d\pi(x, y) \ : \ \pi \in \Pi(\mu, \nu) \right\} . \tag{2.4.3}$$

Just as with the LAP we call the problem *feasible* if there is at least one coupling with a finite cost.

Remark 2.4.1 (Extension to non-probability measures)**.** Optimal transport can be extended to non-probability measures by rescaling. Let $0 < M < \infty$ such that $\mu/M \in \mathrm{Prob}(X)$, $\nu/M \in \mathrm{Prob}(Y)$. Then (2.4.3) and all successive concepts in this section naturally extend to μ and ν.

2.4.1.1 Dual Problem

Intuitively, since OT is a linear program (although potentially infinite dimensional) there exists a corresponding dual problem. It is given by [Vil09, Chap. 5]

$$\sup \left\{ \int_X \alpha(x)\, d\mu(x) + \int_Y \beta(y)\, d\nu(y) \;:\; \alpha \in L^1(\mu),\, \beta \in L^1(\nu) \right. \\ \left. \text{s.t. } \alpha(x) + \beta(y) \leq c(x,y) \right\}. \tag{2.4.4}$$

For any primal feasible coupling $\pi \in \Pi(\mu,\nu)$ and a dual feasible pair (α,β) we find immediately:

$$\begin{aligned} \int_X \alpha(x)\, d\mu(x) + \int_Y \beta(y)\, d\nu(y) &= \int_{X\times Y} \big(\alpha(x) + \beta(y)\big) d\pi(x,y) \\ &\leq \int_{X\times Y} c(x,y) d\pi(x,y) \end{aligned} \tag{2.4.5}$$

Hence, the optimum of the dual problem is not greater than the optimum of the primal problem.

2.4.1.2 Regularity and Existence of Optimal Couplings and Dual Optimizers

The setting of optimal transport is much more general than that of the linear assignment problem. Hence, for a more detailed study further assumptions on the regularity of the measurable spaces X, Y and the cost function c are required. Let us fix such a setting.

Definition 2.4.2 (Regularity setting for optimal transport)**.** We define the following regularity setting:

(i) (X,μ) and (Y,ν) are Polish spaces, that is complete, separable metric spaces equipped with their Borel σ-algebra.

(ii) c is lower semi-continuous and bounded from below.

(iii) There exists at least one coupling with a finite cost.

One then has the following results:

Theorem 2.4.3 (Existence of Optimal Couplings [Vil09, Thm. 4.1])**.** *Under the assumptions of Definition 2.4.2 the optimal transport problem (2.4.3) has a minimizer in* $\Pi(\mu,\nu)$*.*

Theorem 2.4.4 (Duality and Existence of Dual Optimizers [Vil09, Thm. 5.10])**.** *Under the assumptions of Definition 2.4.2 the dual optimal transport problem (2.4.4) has an optimizer. The optimal values of the primal and dual problem are*

identical. Moreover, if π is an optimal coupling and (α, β) are dual optimizers then

$$\alpha(x) + \beta(y) = c(x, y) \quad \pi\text{-almost surely.} \tag{2.4.6}$$

2.4.1.3 Wasserstein Spaces

If $X = Y$ is a metric space with a metric d then optimal transport induces a family of metrics on $\mathrm{Prob}(X)$ for $p \in [1, \infty[$ through

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} d(x, y)^p \, d\pi(x, y) \right)^{1/p}. \tag{2.4.7}$$

A proof that W_p satisfies the axioms of a metric can be found in [Vil09, Chap. 6]. The space of probability measures on X with finite moments of p-th order is called the *Wasserstein space* $\mathcal{W}_p(X)$ of order p.

There are several notions of convergence on $\mathcal{W}_p(X)$, defined through convergence of integrals w.r.t. functions. What we will introduce as *narrow convergence* on $\mathrm{Prob}(X)$ is sometimes referred to also as *weak convergence*. This must be carefully distinguished however from the slightly stronger notion of weak convergence on $\mathcal{W}_p(X)$.

Definition 2.4.5 (Narrow convergence on Prob [AGS08, Sect. 5.1])**.** Let X be a Polish space. A sequence $(\mu_k)_{k \in \mathbb{N}}$ in $\mathrm{Prob}(X)$ is *narrowly* convergent to $\mu \in \mathrm{Prob}(X)$ if

$$\int_X \phi \, d\mu_k \to \int_X \phi \, d\mu \tag{2.4.8}$$

for every function $\phi \in C_b^0(X)$, the space of continuous and bounded real functions on X.

Using this, we can establish weak convergence on $\mathcal{W}_p$.

Definition 2.4.6 (Weak convergence in $\mathcal{W}_p$ [Vil09, Def. 6.8])**.** Let (X, d) be a Polish space and $p \in [0, \infty[$. A sequence $(\mu_k)_{k \in \mathbb{N}}$ in $\mathcal{W}_p(X)$ is said to converge weakly to some $\mu \in WS_p(X)$ if

$$\mu_k \text{ converges narrowly to } \mu \tag{2.4.9}$$

and

$$\int_X d(x_0, x)^p \, d\mu_k(x) \to \int_X d(x_0, x)^p \, d\mu(x) \tag{2.4.10}$$

for any $x_0 \in X$.

Theorem 2.4.7 (W_p metrizes $\mathcal{W}_p$ [Vil09, Thm. 6.9])**.** *W_p metrizes the topology of weak convergence on $\mathcal{W}_p(X)$. That is*

$$W_p(\mu_k, \mu) \to 0 \quad \Leftrightarrow \quad \mu_k \text{ converges weakly to } \mu\,. \tag{2.4.11}$$

2.4.1.4 Relation to the LAP

It is easy to see that optimal transport reduces to the linear assignment problem when X and Y are finite sets of equal cardinality, equipped with the discrete topology, and μ and ν are just the normalized counting measures thereon. The set $\Pi(\mu, \nu)$ will then correspond to the set of permutation matrices $\mathrm{Perm}_{|X|}$, primal and dual problems of OT will coincide with their LAP counterparts (up to normalization by a factor of N) and (2.4.6) will turn into (2.2.9).

Conversely, the LAP can be used to approximate optimal transport problems. We will now briefly sketch a discretization scheme. Let $\hat{X}_N$ and $\hat{Y}_N$ be sets of N points sampled from X and Y according to the probabilities μ and ν. Then by the law of large numbers, with probability 1 the measures

$$\hat{\mu}_N = \frac{1}{|\hat{X}|} \sum_{x \in \hat{X}} \delta_x \,, \qquad \hat{\nu}_N = \frac{1}{|\hat{Y}|} \sum_{y \in \hat{Y}} \delta_y \tag{2.4.12}$$

converge to μ and ν respectively with increasing sampling number $N \to \infty$ in the sense of narrow convergence (cf. Sect. 2.4.1.3). Prokhorov's theorem then implies that a sequence $\hat{\pi}_N \in \Pi(\hat{\mu}_N, \hat{\nu}_N)$ of couplings has a subsequence which converges weakly to some $\pi \in \Pi(\mu, \nu)$. For sufficiently regular (e.g. continuous, bounded) cost functions this implies convergence of the optimal transport cost. For more details on this kind of approximation see for example step 2 of the proof of Thm. 5.10 in [Vil09].

An additional approximation step might be to divide the spaces X and Y into bins $\{X_i\}_i$, $\{Y_i\}_i$, that is

$$X = \bigcup_i X_i, \qquad i \neq j \quad \Rightarrow \quad X_i \cap X_j = \emptyset \,, \tag{2.4.13}$$

$$Y = \bigcup_i Y_i, \qquad i \neq j \quad \Rightarrow \quad Y_i \cap Y_j = \emptyset \tag{2.4.14}$$

and 'round' the cost function according to the bins such that

$$x, x' \in X_i,\, y, y' \in Y_j \quad \Rightarrow \quad c(x, y) = c(x', y') \,. \tag{2.4.15}$$

Then there will be groups of samples in $\hat{X}_n$ and $\hat{Y}_n$ that will have the same rounded cost among each other. We will see in Sect. 2.4.3 that this helps to simplify numerical solution of the approximating assignment problem. By suitable refinement of the bins, such that the rounded cost function converges to the original one in a suitable sense, convergence of the optimal transport cost can still be ensured.

For the particular case $X = Y$ and $c(x, y) = d(x, y)^p$ as on a Wasserstein space, one can use the triangle equation to show convergence of the approximated cost (cf. [Mém11, Remark 2.4]).

2.4.1.5 Monge Formulation of Optimal Transport

The first mathematical treatment of the optimal transport problem dates back to the works of the French mathematician Gaspard Monge from the 18th century [Vil09, Chap. 3] . His formulation was not based on couplings but in terms of a transport map $T : X \to Y$. $T(x)$ indicates where the mass in $x \in X$ is transported to. This is a *deterministic assignment* in the sense that mass from x can, unlike as in the more general coupling formulation, only be transported to a single location. Transporting the measure μ under the map T will yield the measure $T_\sharp\mu$ (assuming T is measurable) which must equal the desired target measure ν. It is easy to check that a feasible transport map (in the sense that the push-forward of μ via T yields ν) induces a feasible coupling by

$$(\mathrm{id}, T)_\sharp\mu \in \Pi(\mu, \nu) \tag{2.4.16}$$

where $(\mathrm{id}, T) : X \to X \times Y$ denotes the product map $x \mapsto \big(x, T(x)\big)$. The cost induced by T is

$$\int_X c\big(x, T(x)\big)\, d\mu(x) = \int_{X\times Y} c(x, y)\, d\big((\mathrm{id}, T)_\sharp\mu\big)(x, y)\,, \tag{2.4.17}$$

which equals the cost of the induced coupling. The formulation of Monge would require to optimize the set of feasible transport maps T, subject to the push-forward constraint and suitable regularity conditions. Neither the feasible set nor the objective function are particularly easy to handle (compared to polyhedrons in some real vector space and linear functionals) nor is it in general clear whether a feasible map exists at all. By virtue of (2.4.16) and (2.4.17) we know however that the deterministic cases are a subset of all allowed assignments in the coupling formulation. A popular question is thus under what conditions the optimizer to (2.4.3) will be deterministic and induced by a transport map.

For this dissertation the following basic case is sufficient:

Theorem 2.4.8 (Existence of Monge Optimizers [Vil03, Thm. 2.12])**.** *Let $X = Y = \mathbb{R}^n$ and let μ, ν be absolutely continuous measures (w.r.t. Lebesgue measure) with finite second order moments. Then for the cost function $c(x, y) = \|x - y\|^2$ there is a unique optimal coupling π which is induced by a transport map $T : X \to Y$.*

For this particular setup in fact much more is known, e.g. that T is the gradient of a convex function on $\mathbb{R}^n$. More details on this special case are for example found in [Vil03, Chap. 3].

2.4.2 The Riemannian Structure of Wasserstein Spaces

The 2-Wasserstein space in $\mathbb{R}^n$ has a structure akin to a Riemannian manifold, allowing for smooth deformation of measures along flow-fields. We will later use

this structure to model deformations of shape templates analogous to contour manifolds. Hence, we now give some required background.

2.4.2.1 Displacement Interpolation

We now fix the case $X = Y = \mathbb{R}^n$, $c(x, y) = \|x - y\|^p$, $p \in [1, \infty[$. What has been discussed so far on optimal transport can be considered to be from a rather *static perspective*. We were only concerned with the initial and final locations of each mass 'atom', as only these coordinates appear in the cost function c: when mass is transported from x to y the cost per unit is given by $c(x, y)$, a function of x and y.

We will now discuss a more *dynamical* point of view. Let us start by an informal discussion of the underlying idea. A more rigorous treatment will then follow.

In many applications the cost function can be written as minimal value of an *action functional* $\mathcal{A}$ evaluated on curves $\gamma : [0, 1] \to X$ joining x and y (i.e. $\gamma(0) = x$, $\gamma(1) = y$), minimized over a suitable regularity class of curves.

For example, the cost function $c(x, y) = \|x-y\|^p$, $p \in [1, \infty[$ can be expressed as a minimization problem over differentiable curves from x to y:

$$c(x, y) = \inf \left\{ \mathcal{A}(\gamma) \,:\, \gamma \in C^1([0, 1], X),\, \gamma(0) = x,\, \gamma(1) = y \right\} \tag{2.4.18}$$

with

$$\mathcal{A}(\gamma) = \int_0^1 \|\partial_t \gamma(t)\|^p \, dt \,. \tag{2.4.19}$$

For $p > 1$ the minimizing curves γ will be straight lines from x to y, parametrized with constant speed. For $p = 1$ any parametrization of the straight line will be minimizing.

Now, when mass is transported from x to y in an optimal coupling between two measures μ, ν, one can think of that mass as not just 'teleporting' from x to y, but as 'travelling smoothly' along the joining action-minimizing curve. In the example above that would be the straight line geodesics.

Informally, 'keeping track' of how each 'mass atom' moves along its action minimizing curve from the initial to its final location will provide a path of measures

$$[0, 1] \ni t \mapsto \rho_t \in \mathrm{Prob}(X), \quad \rho_0 = \mu, \quad \rho_1 = \nu \tag{2.4.20}$$

which is called *displacement interpolation* between μ and ν. This is clearly distinct from the *linear interpolation* based on the vector space structure of measures: $\rho_{\mathrm{lin},t} = (1 - t) \cdot \mu + t \cdot \nu$. In the context of optimal transport, displacement interpolation seems much more natural (Fig. 2.1).

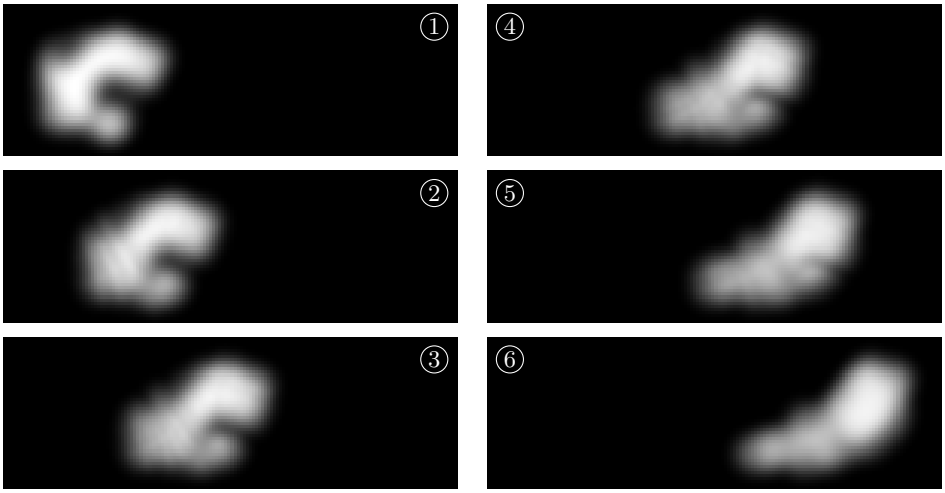


Figure 2.1: Displacement interpolation between two measures. Gray level indicates density, frame numbers indicate temporal ordering. Instead of a simple fade out/in (cf. Fig. 1.1), with displacement interpolation one obtains more natural intermediate measures: mass is actually 'moving' through the image plane from its initial to its final location.

For a thorough and rather general introduction to displacement interpolation we refer the reader to [Vil09, Chap. 7]. Here will will just briefly indicate the main results: under suitable regularity assumptions on $X, c, \mathcal{A}$ and $\mu, \nu \in \mathrm{Prob}(X)$ a displacement interpolation exists. It is related to transporting mass from μ to ν along action minimizing curves and it is itself an action minimizing path in $\mathrm{Prob}(X)$ in a suitable sense (Thm. 7.21, ibid.). Moreover, on a Wasserstein space (Sect. 2.4.1.3), mass is transported along geodesics in the underlying metric space X and the displacement interpolation is itself a geodesic in $\mathcal{W}_p(X)$ (Cor. 7.22, ibid.).

In the next section we will adopt a more regular setting and discuss displacement interpolation in more mathematical detail.

2.4.2.2 Absolute Continuity and Continuity Equation

In the last section, due to the lack of regularity assumptions, we had to be very vague about the relationship between geodesics in X and displacement interpolation on $\mathrm{Prob}(X)$. We will now assume additional regularity and make the relation more explicit.

As before, let $X = Y = \mathbb{R}^n$. We will now fix $c(x, y) = \|x - y\|^2$ and further assume that μ and ν are absolutely continuous measures (i.e. they have a density).

Additionally, we introduce the notion of absolute continuous curves in a metric space.

Definition 2.4.9 (Absolutely continuous paths [AGS08, Def. 1.1.1])**.** Let (Z, d) be a complete metric space and let $[0,1] \ni t \mapsto z_t \in Z$ be a path in Z. $(z_t)_t$ is said to be absolutely continuous if there exists a function $m \in L^1([0,1])$ such that

$$d(z_s, z_t) \leq \int_s^t m(r)\, dr \quad \forall \quad 0 < s \leq t < 1\,. \tag{2.4.21}$$

An important tool for analyzing displacement interpolation in $\mathbb{R}^n$ is the continuity equation which describes the evolution of the interpolation ρ_t through a fluid-dynamics like PDE.

Definition 2.4.10 (Continuity equation [AGS08, Sect. 8.1])**.** Let $[0,1] \ni t \mapsto \mu_t \in \mathrm{Prob}(\mathbb{R}^n)$ be a measure path and $[0,1] \ni t \mapsto \alpha_t \in L^2(\mu_t, \mathbb{R}^n)$ a path of flow-fields such that

$$\int_0^1 \|\alpha_t\|_{L^2(\mu_t, \mathbb{R}^n)}\, dt < \infty\,. \tag{2.4.22}$$

Then we say (μ_t, α_t) satisfy the continuity equation

$$\frac{d}{dt}\mu_t + \nabla\,(\alpha_t\, \mu_t) = 0 \tag{2.4.23a}$$

on $\mathbb{R}^n \times [0,1]$ in the sense of distributions if

$$\int_0^1 \int \big(\partial_t \phi + \langle \nabla\phi, \alpha_t\rangle\big) d\mu_t\, dt = 0 \tag{2.4.23b}$$

for all test functions $\phi \in C_0^\infty(\mathbb{R}^n \times]0,1[)$.

Remark 2.4.11. From (2.4.23b) follows immediately that

$$\int \langle \nabla\phi, \alpha_t\rangle d\mu_t \tag{2.4.24}$$

is the weak time derivative of

$$\int \phi\, d\mu_t \tag{2.4.25}$$

for any $\phi \in \mathcal{D}$. That is, we can express derivatives along measure paths of the function $\mu \mapsto \int \phi\, d\mu$ in terms of the flow-field α_t. So α_t seems to encode first-order deformation information of μ_t.

The continuity equation can be used to characterize absolutely continuous paths of measures.

Theorem 2.4.12 (Characterization of absolutely continuous curves in $\mathcal{W}_2(\mathbb{R}^n)$ [AGS08, Thm. 8.3.1])**.** *When $t \mapsto \mu_t$ is an absolutely continuous measure path then there exists a flow-field $t \mapsto \alpha_t$ such that t-a.e. $\alpha_t \in L^2(\mu_t, \mathbb{R}^n)$ and (μ_t, α_t) satisfy the continuity equation (2.4.23) in the sense of distributions.*

Conversely, if a narrowly continuous curve $t \mapsto \mu_t$ satisfies the continuity equation with some flow-field $t \mapsto \alpha_t$ with $t \mapsto \|\alpha_t\|_{L^2(\mu_t,\mathbb{R}^n)} \in L^1([0,1])$ then μ_t is absolutely continuous and $\|\alpha_t\|_{L^2(\mu_t,\mathbb{R}^n)}$ is an admissible bounding function in Definition 2.4.9.

The continuity equation allows us to reformulate the optimal transport problem explicitly in terms of a displacement interpolation, as discussed in the last section:

Proposition 2.4.13 (Benamou-Brenier formula [AG13, Prop. 2.30])**.** *Let $\mu, \nu \in \mathcal{W}_2(\mathbb{R}^n)$ be absolutely continuous measures with compact support. Then*

$$W_2(\mu,\nu) = \inf\left\{ \int_0^1 \|\alpha_t\|_{L^2(\rho_t)} dt \right\}, \tag{2.4.26}$$

where the infimum is taken among all (ρ_t, α_t) such that ρ_t is a weakly continuous measure path with $\rho_0 = \mu$ and $\rho_1 = \nu$ and (ρ_t, α_t) are a distributional solution of the continuity equation (2.4.23).

For absolutely continuous measures on $\mathbb{R}^n$ it is rather simple to verify that minimizers to (2.4.26) are in fact displacement interpolations. Let $\mu, \nu \in \mathcal{W}_2(\mathbb{R}^n)$ be two absolutely continuous measures and let T be the optimal transport map of the Monge problem with $c(x,y) = \|x-y\|^2$, that is

$$\nu = T_\sharp \mu \qquad \text{and} \qquad W_2^2(\mu,\nu) = \int \|T(x) - x\|^2 \, d\mu(x)\,. \tag{2.4.27}$$

Then consider the measure path

$$\rho_t = \big((1-t)\cdot \mathrm{id} + t \cdot T\big)_\sharp \mu \tag{2.4.28}$$

and the flow-field path

$$\alpha_t = (T - \mathrm{id}) \circ \big((1-t)\cdot \mathrm{id} + t\cdot T\big)^{-1}\,. \tag{2.4.29}$$

It is easy to see that the pair (ρ_t, α_t) is feasible for the optimization problem (2.4.26) and that it attains the minimum. By construction we see that mass in ρ_t travels along constant speed straight lines $\gamma_t = \big((1-t)\cdot \mathrm{id} + t\cdot T\big)$, which are the action minimizing curves in this setup.

2.4.2.3 Tangent Space and Riemannian Inner Product

Let us now have a closer look at the continuity equation (2.4.23): assume (ρ_t, α_t) is a pair of measure and flow field that satisfy (2.4.23). Then, as discussed in Remark 2.4.11 we know that α_t encodes first order deformation information on μ_t.

Now let us turn to the Benamou-Brenier formula (2.4.26): it states that the distance between two elements μ, ν is given by finding the shortest path between the two, where path length is measured by integrating along the path a norm of something that encodes first-order deformation information.

These considerations have motivated Otto to view $\mathcal{W}_2(\mathbb{R}^n)$ informally as an infinite dimensional Riemannian manifold [Ott01], see also [Vil03, Sect. 8.1.2] for a discussion and [AG13, Sect. 2.3.2] for further analogies.

A problem that remains with this perspective is ambiguity of the deformation encoding: if $\nabla \cdot (\beta_t \, \mu_t) = 0$ in a distributional sense t-a.e., then $(\rho_t, \alpha_t + \beta_t)$ will satisfy the continuity equation when (ρ_t, α_t) does. So there are many ways to encode the same measure deformation through different flow-fields. However, there is a natural selection principle: it turns out that minimizers to (2.4.26) lie in a particular subspace which we can interpret as representing the tangent space of the manifold.

Definition 2.4.14 (The tangent space [AG13, Def. 2.31])**.** Let $\mu \in \mathcal{W}_2(\mathbb{R}^n)$. Then the tangent space $\mathrm{Tan}(\mu)$ at μ is defined as

$$\mathrm{Tan}(\mu) = \overline{\{\nabla u : \; u \in \mathcal{D}\}}^{L^2(\mu)} . \tag{2.4.30}$$

Remark 2.4.15. The flow-fields $\alpha \in \mathrm{Tan}(\mu)$ should not so much be thought of as tangent vectors themselves but more as representatives. We will refer to functions u whose gradients represent tangent vectors as *potential functions*.

Comparing with (2.4.26) will tells us how the metric inner product on $\mathrm{Tan}(\mu)$ should look like. The expression $\|\alpha\|_{L^2(\mu)}$ can be interpreted as the (pseudo-)norm of α, induced by the following inner product on $\mathrm{Tan}(\mu)$:

$$\langle \alpha_1, \alpha_2 \rangle_{L^2(\mu)} = \int \langle \alpha_1, \alpha_2 \rangle_{\mathbb{R}^2} \, d\mu \qquad \text{for } \alpha_1, \alpha_2 \in \mathrm{Tan}(\mu) . \tag{2.4.31}$$

2.4.2.4 A Geodesic Equation for Optimal Transport

So far our considerations on $\mathcal{W}_2(\mathbb{R}^n)$ as a Riemannian manifold have been merely in terms of analogies. In [Lot08] the set of probability measures with smooth density functions is treated as an infinite dimensional manifold in the precise sense of [KM97]. Expressions for typical notions in differential geometry, such as the Levi-Civita connection, parallel transport or the geodesic equation are derived. It is however a very tedious task to extend these results in formally rigorous way to less smooth settings.

We will now recall some of the results from [Lot08] for the particular case of optimal transport on a compact subset of $\mathbb{R}^n$. Here, we will not explicitly denote limitation but simply assume that we are on some compact subset, but the set is large enough such that for all our purposes it looks like the whole $\mathbb{R}^n$. We consider the set $\mathrm{Prob}^\infty(\mathbb{R}^n)$ of measures that are absolutely continuous and have a smooth density function. Hence, concepts like the continuity equation can be expressed directly in terms of flow-fields and density functions, and one need not fall back on a distributional formulation.

The tangent space at a point $\mu \in \mathrm{Prob}^\infty$ is isomorphic to

$$\mathrm{Tan}^\infty(\mu) = \{\nabla u : u \in C^\infty(\mathbb{R}^n)\}\,, \tag{2.4.32}$$

where due to the additional smoothness, in contrast to Definition 2.4.14 we do not consider the L^2-completion. Any element in the tangent space describes a local deformation of the footpoint measure as discussed in Remark 2.4.11. In fact, due to the additional regularity, for $\alpha \in \mathrm{Tan}^\infty(\mu)$ we can explicitly model this first order deformation through

$$t \mapsto (\mathrm{id} + t \cdot \alpha)_\sharp \mu\,. \tag{2.4.33}$$

Note that this path and α satisfy the continuity equation at $t = 0$.

More generally, a function $u \in \mathcal{D}$ and its gradient not only represent a tangent vector at one footpoint μ, one can think of them as representing a tangent vector at any footpoint on $\mathrm{Prob}^\infty(\mathbb{R}^n)$, i.e. a vector field [Lot08, Sect. 2]. One then finds:

Proposition 2.4.16 (Covariant Derivative on $\mathrm{Prob}^\infty(\mathbb{R}^n)$ [Lot08, Sect. 2])**.** *Let $\nabla u_1, \nabla u_2$ represent two vector fields on $\mathrm{Prob}^\infty(\mathbb{R}^n)$. Then the Levi-Civita covariant derivative $\overline{\nabla}_{u_1} u_2$ of u_2 w.r.t. u_1 is given by*

$$(\overline{\nabla}_{u_1} u_2)_i = \sum_{j=1}^{n} (\partial_j u_1)(\partial_i \partial_j u_2)\,. \tag{2.4.34}$$

This can be interpreted as the change that the flow field ∇u_2 exhibits when being pushed along the flow induced by flow-field ∇u_1.

This covariant derivative only applies to vector fields on $\mathrm{Prob}^\infty(\mathbb{R}^n)$ for which the vector at any footpoint is represented by the gradient of the same static function u_2. In general, this function can also change throughout the manifold $\mathrm{Prob}^\infty(\mathbb{R}^n)$. Let μ_t be a path in $\mathrm{Prob}^\infty(\mathbb{R}^n)$ with tangent vector ∇u_t at time t and let ∇w_t be a vector field on the path μ_t with potential function w_t at footpoint μ_t. Then the covariant derivative of ∇w_t at μ_t w.r.t. ∇u_t is given by

$$(\overline{\nabla}_{u_t} w_t)_i = \sum_{j=1}^{2} (\partial_j u_t)(\partial_i \partial_j w_t) + \partial_t \partial_i w_t\,. \tag{2.4.35}$$

This is the change of the vector field as given by Proposition 2.4.16, plus the change induced by the change of the potential function w_t along the path.

Since a vector ∇u_t is specified by the unique potential u_t, up to a constant, one can for $w_t = u_t$, by suitably fixing this constant, express the covariant derivative directly in terms of the potential function:

$$(\overline{\nabla}_{u_t} u_t) = \nabla \left(\frac{1}{2} \|\nabla u_t\|^2 + \partial_t u_t \right) . \tag{2.4.36}$$

Setting this, the covariant derivative of u_t along itself, to zero, one finds the geodesic equation on $\mathrm{Prob}^\infty(\mathbb{R}^n)$ in terms of the potential function u_t:

Proposition 2.4.17 ([Lot08, Prop. 4])**.** *The geodesic equation on* $\mathrm{Prob}^\infty(\mathbb{R}^n)$ *in terms of the potential function* u_t *is given by*

$$\partial_t u_t + \frac{1}{2} \|\nabla u_t\|^2 = 0 . \tag{2.4.37}$$

Remark 2.4.18. This geodesic equation is in fact known to be satisfied in a more general setting as $\mathrm{Prob}^\infty(\mathbb{R}^n)$. For absolutely continuous measures it is also satisfied by minimizers to (2.4.26).

2.4.3 Auction Algorithm for Optimal Transport

We have now gone through various theoretical properties and aspects of optimal transport in a continuous setting. At the end of the day however, we will have to solve a discretized approximation on a computer.

2.4.3.1 Mass Splitting

Let us now consider the following setup: X and Y are finite, $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ is a cost function with smallest difference between non-equal values $\delta c > 0$ (cf. Sect. 2.2.2.3), the measures $\mu \in \mathrm{Prob}(X)$, $\nu \in \mathrm{Prob}(Y)$ shall be decomposable into some 'mass atom', $\Delta m > 0$, such that the mass assigned to any $x \in X, y \in Y$ will be an integer multiple of Δm. This corresponds to an approximation described in 2.4.1.4, obtained through sampling and binning, where we have just joined all samples that were within the same bin.

In principle we can translate this approximation into an LAP by splitting every $x \in X, y \in Y$ into an appropriate number of nodes, each carrying one atom Δm of mass. Then any LAP solver can be used for solving (see Sect. 2.2.1.3), in particular the auction algorithm (Sect. 2.2.2). By applying suitable data structures this splitting can be made implicit and the auction algorithm does not actually need to handle each mass atom separately. For example permutation matrices will be replaced by couplings. Also, some modifications in the bidding process are advisable to prevent inefficient competition between atoms originating from the same elements of X. Such a reformulation is given in [BC89, Sect. 4], which we will now discuss. As with the standard auction algorithm we flipped the signs relative to the original presentation (cf. Sect. 2.2.2.1).

2.4.3.2 Implicit Mass Splitting

We will treat any measure in $\mathrm{Prob}(X), \mathrm{Prob}(Y), \mathrm{Prob}(X \times Y)$ as vector of dimension $|X|, |Y|, |X| \cdot |Y|$ respectively and address its elements by $x \in X, y \in Y, (x, y) \in X \times Y$ respectively. For the description of the algorithm it is convenient, to extend the set X by some placeholder element $\Diamond$, standing for 'unassigned'. We then initialize the algorithm with $\pi(x, y) = 0$ for all $(x, y) \in (X \setminus \{\Diamond\}) \times Y$ and $\pi(\Diamond, y) = \nu(y)$ for all y, being interpreted as: all mass from y is currently considered 'unassigned'. All pairs $(\Diamond, y)$ for $y \in Y$ will be added to $\mathcal{N}$. After convergence the row $\pi(\Diamond, \cdot)$ can be dropped from π. So the variable π will not be a feasible coupling between μ and ν when running the algorithm. In analogy to the standard auction algorithm it can be viewed as an 'incomplete coupling' (compare with (2.2.10)).

The auction algorithm implicitly splits any $x \in X$ and $y \in Y$ into small mass atoms and solves the LAP on these atoms. As a result the dual variable β may no longer be constant 'within' every y. Instead there will be one variable β for every pair (x, y). Set all $\beta(x, y) = \beta_{\mathrm{init}}(y)$ for $(x, y) \in (X \cup \{\Diamond\}) \times Y$ for some initially given $\beta_{\mathrm{init}}(\cdot)$.

During iterations a 'homogeneous' (in the sense that it has a unique value for every y) dual variable β can be obtained by

$$\beta(y) = \max\{\beta(x', y) \,:\, x' \in X \cup \{\Diamond\}, \pi(x', y) > 0\}\,. \tag{2.4.38}$$

One finds that β (2.4.38) and the corresponding α via (2.2.7) satisfy ε-complimentary slackness (2.2.11) throughout running of the algorithm. The returned coupling π is optimal if

$$\varepsilon < \delta c / \min\{|X|, |Y|\}\,. \tag{2.4.39}$$

For a detailed discussion and for a stepwise development from the standard auction algorithm via an explicit splitting into mass atoms to the implicit splitting we refer to [BC89].

2.4.3.3 Main Loop

We can now describe the main loop of the algorithm.

Bidding Phase. For each $x \in X \setminus \{\Diamond\}$ where $\sum_y \pi(x, y) < \mu(x)$, consider the collection

$$\begin{aligned}\mathcal{R}(x) = \Big\{ & c(x, y) - \beta(x', y) \,:\, y \in \mathcal{N}(x), \\ & x' \in \big(X \cup \{\Diamond\} \setminus \{x\}\big) \cap \mathcal{N}(y),\, \pi(x', y) > 0 \Big\}\end{aligned} \tag{2.4.40}$$

and assume that the entries are arranged in ascending order, i.e. we have

$$\mathcal{R}(x) = \Big\{ c(x, y_1) - \beta(x'_1, y_1), \ldots, c(x, y_{|\mathcal{R}(x)|}) - \beta(x'_{|\mathcal{R}(x)|}, y_{|\mathcal{R}(x)|}) \Big\} \tag{2.4.41}$$

with $c(x, y_i) - \beta(x'_i, y_i) \leq c(x, y_{i+1}) - \beta(x'_{i+1}, y_{i+1})$ for all $1 \leq i \leq |\mathcal{R}(x)| - 1$, and possibly $x'_i = \Diamond$ for some i.

Let m be the smallest integer that satisfies

$$\pi(x'_1, y_1) + \ldots + \pi(x'_m, y_m) > \mu(x) - \sum_{y \in \mathcal{N}(x)} \pi(x, y)\,. \tag{2.4.42}$$

Define new coupling entries $\hat{\pi}(x, y)$ as follows:

$$\hat{\pi}(x,y) = \begin{cases} \pi(x,y) & \text{if } y \neq y_1, \ldots, y_m \\ \pi(x,y) + \sum_{i=1,\ldots,m-1\,:\,y_i = y} \pi(x'_i, y) & \text{if } y \in \{y_1, y_2, \ldots, y_{m-1}\} \\ \mu(x) - \sum_{y' \in \mathcal{N}(x) \setminus \{y_m\}} \hat{\pi}(x, y') & \text{if } y = y_m \end{cases} \tag{2.4.43}$$

The value $\alpha'(x)$, corresponding to (7) in the LAP variant, is given by

$$\alpha'(x) = c(x, y_m) - \beta(x'_m, y_m) \tag{2.4.44}$$

if $\hat{\pi}(x, y) > 0$ for more than one $y \in Y$. Otherwise set

$$\alpha'(x) = c(x, y_{m'}) - \beta(x'_{m'}, y_{m'}) \tag{2.4.45}$$

where m' is the smallest integer which satisfies $y_{m'} \neq y_1$.

x then submits a bid for any y where $\hat{\pi}(x, y) > 0$ with value $b_{xy} = c(x, y) - \alpha'(x) - \varepsilon$.

Assignment Phase. For each $y \in Y$, let $\mathcal{B}(y)$ be the set of $x \in X$ for which y received a bid, corresponding to a positive hypothetical coupling entry $\hat{\pi}$. Assume that $\mathcal{B}(y)$ is ordered in ascending bid value, i.e. for $\bar{n} = |\mathcal{B}(y)|$ one has

$$\mathcal{B}(y) = \{x_1, x_2, \ldots, x_n\} \text{ where } b_{x_i y} \leq b_{x_{i+1} y} \text{ for all } i = 1, \ldots, \bar{n} - 1\,. \tag{2.4.46}$$

Let $\bar{m} = \bar{n}$ if

$$\sum_{i=1}^{\bar{n}} \hat{\pi}(x_i, y) \leq \nu(y) \tag{2.4.47}$$

and, otherwise, let $\bar{m}$ be the smallest integer that satisfies

$$\sum_{i=1}^{\bar{m}} \hat{\pi}(x_i, y) > \nu(y)\,. \tag{2.4.48}$$

Then update the coupling π for $x \in \mathcal{N}(y)$ as follows:

$$\pi(x,y) := \begin{cases} \hat{\pi}(x,y) & \text{if } x \in \{x_1, \dots, x_{\bar{m}-1}\} \\ \hat{\pi}(x,y) - \max\left\{0, \sum_{i=1}^{\bar{m}} \hat{\pi}(x_i,y) - \nu(y)\right\} & \text{if } x = x_{\bar{m}-1} \\ 0 & \text{else} \end{cases} \tag{2.4.49}$$

and update β by setting $\beta(x,y) = b_{xy}$ for $x \in \{x_1, \dots, x_{\bar{m}}\}$.

The two stages are repeated, until $\pi(\lozenge, y) = 0$ for all y.

Remark 2.4.19 (Complexity of auction algorithm for optimal transport)**.** The cost per bid in the auction algorithm for the linear assignment problem was $\mathcal{O}(n)$. The worst case cost of determining (2.4.40) is $\mathcal{O}(n^2)$ for finding the candidates and some more for sorting, leading to a higher complexity bound. Via [BE88] an algorithm for OT could be constructed with the same complexity bound as for the linear assignment problem. It turns out however that in practice the modifications made here make the algorithm robust to price haggling (see Sect. 2.2.2.6) and thus lead to a better average complexity on practical problems.

2.4.4 Assignment Regularization

Under sufficient regularity conditions the Monge formulation of optimal transport guarantees that the optimal coupling is deterministic and induced by a transport map T (Sect. 2.4.1.5). In the case of the LAP we know from Corollary 2.2.2 that an optimal permutation matrix exists, which can also be interpreted as deterministic assignment T. Optimal transport and linear assignment are often used on geometric problems. It is then usually a reasonable modelling assumption that the map T is spatially regular. However such spatial regularity is not explicitly implemented into the linear matching approach. Corresponding extensions have therefore been proposed.

In this section we will briefly go through the approach presented in [FPPA13]. For simplicity we consider a partial linear assignment problem between two sets X and Y with $|X| \leq |Y|$ and cost function $c : X \times Y \to \mathbb{R}$. By extension of X with dummy nodes this can be solved through the following linear program (cf. Proposition 3.1.6 later on):

$$\inf \left\{ \sum_{(x,y)\in X\times Y} c(x,y)\,\pi(x,y) \;:\; \pi \in \mathbb{R}_+^{|X|\times|Y|} \text{ s.t. } \sum_{y\in Y} \pi(x,y) = 1 \;\forall\, y \in Y \right.$$
$$\left. \sum_{x\in X} \pi(x,y) \leq 1 \;\forall\, x \in X \right\} \tag{2.4.50}$$

We now want to extend the functional by a geometric regularizer on the assignment T. Since we are not optimizing over T, but over the coupling matrix

π, we must try to extract T from π. We assume that X and Y are embedded into some vector space (for example the image plane). Then from π we can extract for every $x \in X$ the mean assignment

$$\hat{T}(x, \pi) = \sum_{y \in Y} \pi(x, y)\, y\,. \tag{2.4.51}$$

The definition in [FPPA13, Sect. 3.2] is slightly more general since the X-marginal of the coupling is not fixed during optimization and one needs to normalize $\hat{T}(x)$ by the assigned mass.

$$\hat{V}(x, \pi) = \hat{T}(x, \pi) - x \tag{2.4.52}$$

is then the vector field of mean relative mass displacement. For a vector field on X and neighbourhood weights $w : X \times X \to \mathbb{R}^+$ one can define a gradient on vector fields by

$$\operatorname{grad}(V, x, x') = w(x, x') \left(V(x') - V(x)\right). \tag{2.4.53}$$

If X is a regular pixel grid, $w(x, x')$ can be chosen 1 for adjacent pixels and 0 else. On point clouds and super-pixels a suitable neighbourhood graph will have to be introduced on X and the weights set accordingly, for example $w(x, x') = \|x - x'\|^{-1}$ for adjacent nodes, to obtain consistency with the directional derivative. In [FPPA13] a regularizer of the form

$$J_p\big(\hat{V}(\cdot, \pi)\big) = \sum_{x, x' \in X} \Big\| \operatorname{grad}\big(\hat{V}(\cdot, \pi), x, x'\big) \Big\|^p \tag{2.4.54}$$

for some exponent $p \geq 1$ is then proposed and introduced into the linear optimal transport functional, yielding

$$E(\pi) = \sum_{(x,y) \in X \times Y} c(x, y)\, \pi(x, y) + J_p\big(\hat{V}(\cdot, \pi)\big)\,. \tag{2.4.55}$$

This functional is no longer linear in π and can no longer be optimized with an optimal transport solver. Nevertheless it is still convex and can be tackled for reasonable instance sizes.

The benefit of this type of geometric regularization has been demonstrated on the problem of color transfer [FPPA13] and it is likely to improve results on other geometric assignment problems.

2.5 The Gromov-Wasserstein Distance

We have discussed earlier (Sect. 1.2.3) the significance of matchings between shapes as a means of data interpretation and as a basis for a similarity measure.

As was pointed out, 'naïve' linear approaches are limited in their applicability, in particular when it comes to invariance w.r.t. geometric transformations. The *metric approach to shape matching* is a framework particularly devised to obtain this kind of invariance. Central tools are the *Gromov-Hausdorff* and the *Gromov-Wasserstein* distances. A very good motivation, overview and mathematical study of the Gromov-Hausdorff/Wasserstein distances is given in [Mém11]. We will now briefly go through the results that are most relevant to this work.

2.5.1 The (Gromov-)Hausdorff Distance

2.5.1.1 The Hausdorff Distance

We have already touched upon the importance of choosing a good representation for shape. In the following we will describe shapes by metric spaces. This comes quite natural, as we can for example model 'the world' or a given image as Euclidean spaces. The metric space describing an object is then simply the restriction of the ambient space to the object. One could then compare objects by the *Hausdorff distance*. Let Z be the ambient space with metric d and let $X, Y \subset Z$ be subspaces, describing two objects. The Hausdorff distance is defined by

$$D_{\mathrm{H}}(X,Y) = \max\left\{ \sup_{x\in X} \inf_{y\in Y} d(x,y), \sup_{y\in Y} \inf_{x\in X} d(x,y) \right\}. \tag{2.5.1}$$

By introducing the concept of *correspondences* we can reformulate (2.5.1) to make the importance of shape matching for shape similarity more explicit.

Definition 2.5.1 (Correspondence [Mém11, Def. 2.1])**.** For non-empty sets X and Y, a subset $R \subset X \times Y$ is a *correspondence* (between X and Y) if and only if

- $\forall\, x \in X$ there exists $y \in Y$ s.t. $(x,y) \in R$,
- $\forall\, y \in Y$ there exists $x \in X$ s.t. $(x,y) \in R$.

Let $\mathrm{Corr}(X,Y)$ denote the set of all possible correspondences between X and Y.

A correspondence in $\mathrm{Corr}(X,Y)$ can be interpreted as a multivalued assignment between X and Y. We then find:

Proposition 2.5.2 ([Mém11, Prop. 2.1])**.** *For a compact metric space (Z,d) the Hausdorff distance between two sets $X, Y \subset Z$ can be expressed as*

$$D_{\mathrm{H}}(X,Y) = \inf_{R\in \mathrm{Corr}(X,Y)} \sup_{(x,y)\in R} d(x,y)\,. \tag{2.5.2}$$

2.5.1.2 Geometric Invariance

It is obvious that the standard Hausdorff distance is not a good measure of shape similarity. For a given ambient space (Z, d), when Y is the image of X under translation we intuitively say X and Y are of the same shape. Nonetheless we will find $D_{\mathrm{H}}(X, Y) > 0$. A solution could be to explicitly take into account *isometries* on the ambient space Z:

$$\mathcal{T} = \left\{\varphi : Z \to Z : \; d\big(\varphi(z), \varphi(z')\big) = d(z, z') \; \forall \, z, z' \in Z\right\} \tag{2.5.3}$$

We can then define an isometry invariant extension of the Hausdorff distance by

$$D_{\mathrm{H}}^{\mathcal{T}}(X, Y) = \inf_{\varphi \in \mathcal{T}} D_{\mathrm{H}}\big(X, \varphi(Y)\big) \, . \tag{2.5.4}$$

Naturally this is not limited to the Hausdorff distance, but can be done for any non-invariant similarity measure. Also, in principle the set of transformations $\mathcal{T}$ need not be restricted to rigid isometries: for example, one could also take into account rescaling and transformations that correspond to changes in pose. Due to the explicit role of the ambient space and its invariance transformations we will refer to this type of invariance as *extrinsic*. We will next develop a corresponding *intrinsic* notion.

2.5.1.3 Intrinsic Invariance

For the Hausdorff distance and its isometry invariant extension (2.5.4) we have represented a shapes X, Y as a metric spaces that were subspaces of an ambient space Z, the metric being induced by restrictions. Let us now consider X as a metric space in its own right, with no reference to a distinct ambient space and try to find a direct means of comparison. This has some important advantages:

First, one can choose a metric on X that is more meaningful than the restricted ambience metric of Z. For example let $Z = \mathbb{R}^n$ and let X be some articulated object. Then, when we change the pose of X, the Euclidean distance between points of X, as induced by the embedding $X \to Z$ may change significantly, while X basically remains the same object. If instead we equip X with a geodesic metric, based on the length of paths running only within X, then distances between points in X will be approximately invariant under pose changes of X (Fig. 2.2).

Second, the notion of an ambient space might not always be available, for example if X were a graph. Thus, by working without an ambient space the approach will become applicable to a wider set of problems.

This proposal implies that information on the shape of X will no longer be encoded in both the metric on X *and* a set of allowed transformations on Z, but *exclusively* in the metric of X, which one might find more aesthetically pleasing.

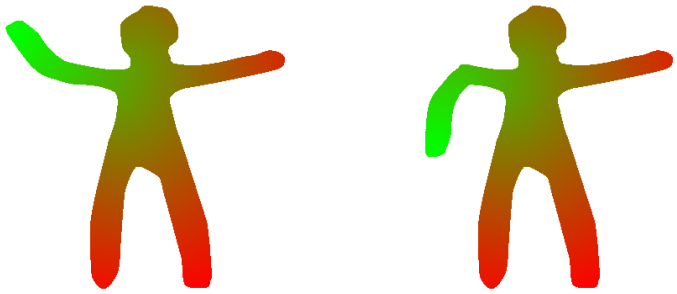


Figure 2.2: Pose invariance of geodesic metric: a humanoid shape is given in two different poses. Color coding gives the geodesic distance to a point in the left hand. The distance is approximately equal in both versions, despite the pose change.

A way to directly compare the structure of two metric spaces is the Gromov-Hausdorff distance, which we will discuss next.

2.5.1.4 The Gromov-Hausdorf Distance

Definition 2.5.3 (Gromov-Hausdorff distance [Gro07, Def. 3.4])**.** Let X, Y be metric spaces. Then the Gromov-Hausdorff distance between them is given by

$$D_{\mathrm{GH}}(X,Y) = \inf_{Z,f,g} D_{\mathrm{H}}\big(f(X), g(Y)\big) \tag{2.5.5}$$

where the infimum runs over all metric spaces Z and isometric embeddings $f : X \to Z$, $g : Y \to Z$.

The Gromov-Hausdorff distance is a metric on the space of equivalence classes of compact metric spaces ([Gro07, Rem. $3.11\frac{1}{2}$], also [Mém11, Prop. 4.1]). Of course formula (2.5.5) is impractical from a numerical point of view: minimizing over the set of metric spaces is certainly not explicitly feasible on a computer. [Mém11, Table 1] gives a list of equivalent formulations. We will here focus on the following form in terms of correspondences:

$$D_{\mathrm{GH}}(X,Y) = \inf_{R \in \mathrm{Corr}(X,Y)} \sup_{(x,y),(x',y') \in R} \Gamma(x,y,x',y') \tag{2.5.6}$$

with

$$\Gamma(x,y,x',y') = |d_X(x,x') - d_Y(y,y')| \tag{2.5.7}$$

where d_X, d_Y are the metrics on X, Y. This is considerably more concrete than (2.5.5). However, due to the combinatorial nature (2.5.6) is still very difficult to handle from a numerical point of view. We will next discuss a smooth relaxation based on concepts from mass transportation.

2.5.2 The Gromov-Wasserstein Distance

2.5.2.1 Mass Transport Relaxation

Let Z be a compact Polish space, let $\mu, \nu \in \mathrm{Prob}(Z)$ and let $X = \operatorname{spt}\mu$, $Y = \operatorname{spt}\nu$. In Sect. 2.4.1.3 we have introduced the Wasserstein distances W_p, $p \in [1, \infty[$ on sufficiently regular probability measures on Z by minimization over the set of couplings (2.4.1). In Sect. 2.5.1.1 the Hausdorff distance was defined for subsets of Z by minimizing over the set of correspondences (Def. 2.5.1). We see that both distances are constructed in terms of minimization over a suitable notion of matching between the two objects in question. Let us have a closer look at the relation between W_p and D_{H}.

From [Mém11, Lem. 2.2] we know that

$$\pi \in \Pi(\mu, \nu) \qquad \Rightarrow \qquad \operatorname{spt}\pi \in \mathrm{Corr}(\operatorname{spt}\mu, \operatorname{spt}\nu)\,.$$

A reasonable extension of the definition of W_p to $p = \infty$ is given by (cf. [Mém11, Def. 2.4])

$$W_\infty(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \sup_{(x,y)\in \operatorname{spt}(\pi)} d(x, y)\,. \tag{2.5.8}$$

So we can conclude [Mém11, Cor. 2.1] that

$$D_{\mathrm{H}}(X, Y) \leq W_\infty(\mu, \nu)\,. \tag{2.5.9}$$

We can thus interpret the Wasserstein distances as a relaxation of the Hausdorff distance. There are two important differences: First, the optimization in D_{H} is combinatorial, the problem underlying the computation of W_p for finite p is smooth and convex. Second, the Hausdorff distance is defined between sets, the Wasserstein distances are defined between measures. The additional information contained in the measure can be interpreted as a description of importance of elements in the support.

We will now try to find a relaxation of the Gromov-Hausdorff distance based on mass transportation that is related to the Gromov-Hausdorff distance in the way that the Wasserstein distances are related to the Hausdorff distance. To that end we will from now on represent shapes by both a metric and a measure.

Definition 2.5.4 ([Mém11, Def. 5.1])**.** A *metric measure space* (mm-space) is a triple (X, d, μ) where

- (X, d) is a compact metric space,
- and μ is a Borel probability measure on X, i.e. $\mu(X) = 1$, and μ has full support: $\operatorname{spt}\mu = X$.

When it is clear from the context, sometimes the triple (X, d, μ) will be denoted only by X or μ. Two mm-spaces (X, d_X, μ), (Y, d_Y, ν) are *isomorphic* if there exists an isometry $\psi : X \to Y$ such that $\psi_\sharp \mu = \nu$.

We are then ready to define:

Definition 2.5.5 (Gromov-Wasserstein distance [Mém11, Def. 5.7])**.** For two mm-spaces $(X, d_X, \mu), (Y, d_Y, \nu)$ and $p \in [1, \infty]$ the Gromov-Wasserstein distance $D_{\mathrm{GW},p}(\mu, \nu)$ is given by

$$D_{\mathrm{GW},p}(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} J_p(\pi) \tag{2.5.10a}$$

with

$$J_p(\pi) = \frac{1}{2} \left(\int_{X\times Y} \int_{X\times Y} \Gamma(x,y,x',y')^p \, d\pi(x,y) \, d\pi(x',y') \right)^{1/p} \tag{2.5.10b}$$

for $p < \infty$ and

$$J_\infty(\pi) = \frac{1}{2} \sup_{(x,y),(x',y') \in \operatorname{spt} \pi} \Gamma(x,y,x',y') \,. \tag{2.5.10c}$$

Some central properties of $D_{\mathrm{GW},p}$ are proved in [Mém11, Thm. 5.1], in particular that it metrizes the collection of isomorphism classes of mm-spaces and that

$$D_{\mathrm{GH}}(X,Y) \leq D_{\mathrm{GW},\infty}(\mu,\nu) \,, \tag{2.5.11}$$

which is the equivalent of (2.5.9).

Remark 2.5.6 (Relation to the QAP)**.** Let X and Y be discrete, $|X| = |Y| < \infty$, and μ, ν be the (normalized) counting measures on X and Y. Then the problem (2.5.10) for $p < \infty$ is related to the quadratic assignment problem (2.3.2) (the quadratic bottleneck problem (2.3.3) for $p = \infty$) as (2.2.5) is related to (2.2.3). That is, the set of permutation matrices to optimize over is relaxed to the set of doubly stochastic matrices. Since the objective functional is not linear (and the quadratic form is in general non-definite), its optimizers need not lie at the vertices of the feasible polytope. So there is no equivalent to Corollary 2.2.2. Also this means that computing $D_{\mathrm{GW},p}$ is not necessarily NP-hard.

We should still regard the formal similarity as a warning that computing $D_{\mathrm{GW},p}$ is not an easy problem.

2.5.2.2 Lower Bounds

The computational complexity of solving the problem underlying $D_{\mathrm{GW},p}$ (or to find sufficiently good suboptimal solutions) suggest the application of less expensive lower bounds, in particular when many problem instances need to be solved or when $D_{\mathrm{GW},p}$ is only a part of the total problem, as it will be in Sect. 3.1.

In [Mém11] three lower bounds are discussed. We will now discuss the one we will later apply in Sect. 3.1.

Definition 2.5.7 (Relaxed linearization of the Gromov-Wasserstein distance [Mém11, Def. 6.3])**.** Let (X, d_X, μ), (Y, d_Y, ν) be two metric measure spaces and $p \in [1, \infty[$. Then define

$$D_{\mathrm{LinGW},p}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \frac{1}{2}\left(\int_{X\times Y} c_{\mathrm{LinGW},p}(\mu,\nu;x,y)\,d\pi(x,y)\right)^{1/p} \tag{2.5.12a}$$

with

$$c_{\mathrm{LinGW},p}(\mu,\nu;x,y) = \inf_{\pi'\in\Pi(\mu,\nu)} \int_{X\times Y} \Gamma(x,y,x',y')^p\,d\pi'(x',y')\,. \tag{2.5.12b}$$

We immediately find:

Proposition 2.5.8 ([SS13c, Prop. 4.1])**.**

$$D_{\mathrm{LinGW},p}(\mu,\nu) \leq D_{\mathrm{GW},p}(\mu,\nu) \tag{2.5.13}$$

Proof. Problem (2.5.12) can be transformed into problem (2.5.10) by adding the additional constraint $\pi' = \pi$ for all $(x, y) \in X \times Y$ in (2.5.12b). □

Remark 2.5.9 (Relation to Gilmore-Lawler QAP bound)**.** Adopting the scenario of Remark 2.5.6 the bound $D_{\mathrm{LinGW},p}$ corresponds to the Gilmore-Lawler bound (Def. 2.3.3). There is a slight difference in the constraints for computing $c_{\mathrm{LinGW},p}(x, y)$: in the Gilmore-Lawler bound the assignment $\pi'(x, y) = \mu(x) = \nu(y)$ is fixed. Beyond that, both objective functions (2.5.12a) and (2.5.12b) are linear, the binary-ness constraint of the QAP does not matter and both relaxations (Gilmore-Lawler to QAP, $D_{\mathrm{LinGW},p}$ to $D_{\mathrm{GW},p}$) are identical (cf. Corollary 2.2.2).

Remark 2.5.10 (Interpretation of $D_{\mathrm{LinGW},p}$)**.** As with the Gilmore-Lawler bound the interpretation of this relaxation is very simple (cf. Remark 2.3.4): for the computation of $c_{\mathrm{LinGW},p}(\mu, \nu; x, y)$ we fix the hypothetical assignment $x \leftrightarrow y$ and see how well the rest of the mm-spaces can be matched w.r.t. the distances to x and y respectively. The final matching with $c_{\mathrm{LinGW},p}(\mu, \nu; \cdot, \cdot)$ as cost function corresponds then to finding the optimal consistent set of assignment hypotheses.

All involved sub-problems for computing $D_{\mathrm{LinGW},p}$ are standard linear programs. However, in a numerical implementation naïvely $|X| \cdot |Y|$ sub-problems for computing $c_{\mathrm{LinGW},p}(\mu, \nu; x, y)$ need to be solved, each containing $|X| \cdot |Y|$ variables. The cost for this quickly becomes prohibitive when the cardinality of X and Y increases. Throughout this thesis we will repeatedly address the question of how $D_{\mathrm{LinGW},p}$ and various extensions can be computed more efficiently. We will start by the following observation, which helps greatly reducing the dimensionality of the sub-problems to compute $c_{\mathrm{LinGW},p}(\mu, \nu; x, y)$ in numerical implementations.

Definition 2.5.11 (Radial mass distribution)**.** For an mm-space (X, d_X, μ) define the measure valued map

$$\mathrm{RadDist}_X : X \to \mathrm{Prob}(\mathbb{R}_+), \qquad x \mapsto d_X(x, \cdot)_\sharp \mu\,, \tag{2.5.14}$$

which assigns every point x the radial distribution of mass around that point.

In Sect. 3.1.4 it will be shown that $c_{\mathrm{LinGW},p}(\mu, \nu; x, y)$ can be computed by solving a suitable optimal transport problem between $\mathrm{RadDist}_X(x)$ and $\mathrm{RadDist}_Y(y)$. This means $c_{\mathrm{LinGW},p}(\mu, \nu; x, y)$ can be computed by solving an optimal transport problem on the real line, which can for larger problems be discretized with reasonable accuracy by much fewer bins than both the cardinality of X and Y, reducing the involved computational cost.

A weaker relation between $c_{\mathrm{LinGW},p}$ and $\mathrm{RadDist}_{X,Y}$ that allows for an even cheaper lower bound on $D_{\mathrm{LinGW},p}$ is proved in [Mém11, Cor. 6.3]. In Example 5.6, ibid., it is shown that the radial mass distributions do not uniquely characterize a mm-space. So this relaxation is no longer a distance, because different mm-spaces can have the same radial mass distributions and thus zero $D_{\mathrm{LinGW},p}$-'distance' between them.

Remark 2.5.12 (Connection to the Koopmans-Beckmann QAP (2.3.1))**.** For the particular case $p = 2$ and the setup of Remark 2.5.6 the Gromov-Wasserstein distance resembles the Koopmans-Beckmann QAP (2.3.1):

Up to renormalization of the permutation matrices (to turn them into probability measures) computing the Gromov-Wasserstein distance boils down to

$$\inf \left\{ \frac{1}{2} \sum_{i,j,k,l=1}^{n} \big(d_X(x_i, x_j) - d_Y(y_k, y_l)\big)^2 P_{ik}\, P_{jl} \;:\; P \in \mathrm{Stoch}_n \right\}. \tag{2.5.15}$$

We can rewrite the objective function as follows:

$$\frac{1}{2} \sum_{i,j,k,l=1}^{n} \big(d_X(x_i, x_j) - d_Y(y_k, y_l)\big)^2 P_{ik}\, P_{jl} \tag{2.5.16}$$

$$= \sum_{i,j,k,l=1}^{n} \left(\frac{1}{2} \big(d_X(x_i, x_j)\big)^2 - d_X(x_i, x_j)\, d_Y(y_k, y_l) + \frac{1}{2} \big(d_Y(y_k, y_l)\big)^2 \right) P_{ik}\, P_{jl} \tag{2.5.17}$$

For any feasible $P \in \mathrm{Perm}_n$ the first and the third term will be constant and can be dropped during optimization. We are left with the middle term which takes the required form. Then, according to Remark 2.3.5 one can greatly reduce the computational effort for computing the relaxed Gromov-Wasserstein distance.

2.5.2.3 A Discussion

In the previous sections we have discussed two alternative approaches for constructing transformation invariant shape similarity measures based on matching: *extrinsic* and *intrinsic*. In Sect. 2.5.1.2 explicit modelling of invariance via optimization of a non-invariant similarity measure over a corresponding set of transformations has been proposed. In Sections 2.5.1.3 and 2.5.1.4 invariance was obtained by completely stripping of any reference to the ambient space of the object, encoding all information about shape in the intrinsic metric.

In this thesis we will use both methods to construct invariant geometric shape priors. In Sect. 3.1 we will present a functional based on the linear relaxation of the Gromov-Wasserstein distance (Def. 2.5.7), an intrinsic approach. An extrinsic approach with explicit modelling of invariance transformations through the Riemannian structure of optimal transport (Sect. 2.4.2) is developed in Chapter 4. Here, let us briefly compare the advantages and difficulties in both variants.

The intrinsic variant works without optimization over a set of transformations. This allows for a completely convex approach through suitable relaxations, however, at the cost of losing some of the descriptive power. For one, the metric will have to be static and cannot depend on the object's current configuration. This limits the implementation of pose invariance. Second, due to the absence of an explicit low-dimensional variable describing the transformation underlying the shape matching, geometric regularity of the shape matching is a difficult challenge which we address in Sect. 3.2.

The explicit transformation modelling of the extrinsic approach leaves us with a non-convex functional which cannot be optimized in a straight-forward fashion. In return, we gain more flexibility in shape modelling: shape similarity is no longer solely based on the near-isometry assumption. Instead we can also model non-isometric variations. Also, regularity of the underlying assignments is automatically encouraged. It is shown in Sect. 4.1 that from a modelling perspective this approach is equivalent to well-established contour manifolds.

Both approaches address the difficult problem of simultaneous shape optimization and matching. Hence, we cannot expect to find functionals that are computationally cheap to optimize. By substituting the combinatorial similarity measures with their mass transport relaxations (cf. Sect. 2.5.2.1) both problems can be turned into computationally more approachable forms. But the naïve complexities even of the approximations that we use are still huge.

It is not uncommon however for problems inspired by computer vision, that while the general problem class is generally unfeasible, practical problem instances are de facto much simpler and one can hope for efficient solvers for at least those cases. The key question is then how to 'extract the simpleness'. In Chapter 5 hierarchical coarse-to-fine schemes are proposed as a means of doing just that.

2.6 Contour Manifolds, Flows, Diffeomorphisms and Poisson's Equation

In Chapter 4 we will introduce a connection between the shape representations in terms of parametrized contours and indicator functions, to combine the advantages of the former in terms of shape modelling with those of the latter in terms of local appearance matching. More precisely we will show that the manifold of closed simple contours is diffeomorphic to a submanifold of the 'manifold' of measures as introduced in Sect. 2.4.2.

In this section we will first summarize some facts on this contour manifold. Then some background on flows and their induced diffeomorphisms as well as on Poison's equation are given. These tools are required to establish the diffeomorphism in Sect. 4.1. A shape prior based thereon is then discussed in Sect. 4.2.

2.6.1 Contour Manifolds

The set of embeddings of the unit circle S^1 into $\mathbb{R}^2$ can formally be treated as infinite dimensional manifold. A comprehensive framework is laid out in [KM97] and an overview is given for example in [MM06]. We now summarize the facts that are relevant for this thesis.

Definition 2.6.1 (Space of smooth mappings, manifold of embeddings, manifold of submanifolds [SS13a, Def. 2.1])**.** Denote by $C^\infty(S^1, \mathbb{R}^2)$ the vector space of smooth mappings from S^1 into $\mathbb{R}^2$, equipped with the topology of uniform convergence in all derivatives and spatial components. By Emb we denote the set of C^∞-embeddings $S^1 \to \mathbb{R}^2$. It is an open submanifold of $C^\infty(S^1, \mathbb{R}^2)$. Its tangent bundle $T\text{Emb}$ is given by $\text{Emb} \times C^\infty(S^1, \mathbb{R}^2)$. Let Diff be the Lie group of C^∞-diffeomorphisms on S^1. Then, by B we denote the quotient set Emb/Diff of equivalence classes in Emb, two contours in Emb being equivalent if there is a reparametrization in Diff that transforms one into the other by right composition. That is for $c_1, c_2 \in \text{Emb}$ have $c_1 \sim c_2$ if $c_2 = c_1 \circ \varphi$ for some $\varphi \in \text{Diff}$.

The set B of equivalence classes $[c]$ on Emb is then itself a smooth manifold and the continuous map

$$\pi : \text{Emb} \to \text{B}, \qquad c \mapsto [c] \tag{2.6.1}$$

that takes contours to their equivalence class is a principal bundle with total space Emb, base space B and structure group Diff.

The relevant parts of [KM97] for this definition are: Sect. 6.1, Thm. 42.1 for the structure of $C^\infty(S^1, \mathbb{R}^2)$, Thm. 44.1 for the principal bundle $(\text{Emb}, \pi, \text{B}, \text{Diff})$.

For a given contour $c \in \text{Emb}$ we denote by $n_c \in C^\infty(S^1, \mathbb{R}^2)$ its *outward pointing unit-normal field*.

Definition 2.6.2 (Vertical and horizontal bundle, horizontal lifting [KM97, Sect. 37] [SS13a, Def. 2.2])**.** The vertical bundle on Emb with respect to π, $V\mathrm{Emb} = \ker \mathcal{D}\pi$, is at each point $c \in \mathrm{Emb}$ the set of tangent vectors

$$V_c\mathrm{Emb} = \left\{ a \in T_c\mathrm{Emb} = C^\infty(S^1, \mathbb{R}^2) : \ \langle a(\theta), n_c(\theta) \rangle_{\mathbb{R}^2} = 0 \, \forall \, \theta \in S^1 \right\} \quad (2.6.2)$$

which are locally orthogonal to the normal field n_c on c. A corresponding choice for the horizontal bundle is then given by

$$H_c\mathrm{Emb} = \left\{ a \cdot n_c : \ a \in C^\infty(S^1, \mathbb{R}) \right\} . \quad (2.6.3)$$

This is the orthogonal complement of $V_c\mathrm{Emb}$ w.r.t. the L^2-inner product on $T_c\mathrm{Emb}$.

The projection $\pi : \mathrm{Emb} \to \mathrm{B}$ induces an isomorphism

$$\pi_{c,*} : H_c\mathrm{Emb} \to T_{\pi(c)}\mathrm{B} \quad (2.6.4)$$

whose inverse is referred to as *horizontal lift*. For every tangent vector $v \in T_{\pi(c)}\mathrm{B}$ there is a unique horizontal vector field $a \in H_c\mathrm{Emb}$ such that $\pi_{c,*}(a) = v$.

Lemma 2.6.3 (Horizontal Parametrization [SS13a, Lemma 2.3])**.** *Any C^1 contour-family $[0,1] \ni t \mapsto c_t \in \mathrm{Emb}$ can be reparametrized such that $\dot{c}_t = a_t \cdot n_{c_t}$, $a_t \in C^\infty(S^1, \mathbb{R})$, i.e. such that the temporal deformation is normal to the contour and the tangent vectors lie in the horizontal bundle.*

The proof is analogous to that of the proposition in [MM06, Sect. 2.5], see also discussion ibid., Sect. 2.3.

Remark 2.6.4. Based on Lemma 2.6.3, in the course of this paper, we will always describe contour deformations by scalar fields, that give the local deformation along the normal field, i.e. within the horizontal bundle. By aid of the unit normal-field on contours we will canonically identify

$$H\mathrm{Emb} \cong \mathrm{Emb} \times C^\infty(S^1, \mathbb{R}) . \quad (2.6.5)$$

We show next that diffeomorphisms $\varphi \in \mathrm{Diff}$ preserve horizontal lifting.

Proposition 2.6.5 ([SS13a, Prop. 2.5])**.** *For any $v \in T_{[c]}\mathrm{B}$ and $\varphi \in \mathrm{Diff}$ have*

$$\left(\pi_{c,*}^{-1}(v)\right) \circ \varphi = \pi_{c\circ\varphi,*}^{-1}(v) . \quad (2.6.6)$$

This implies that any element in the tangent bundle $T\mathrm{B}$ can be represented by an equivalence class $[(c,a)]$ in $H\mathrm{Emb}$, equivalence $(c_1, a_1) \sim (c_2, a_2)$ holding when there is some $\varphi \in \mathrm{Diff}$ such that $c_2 = c_1 \circ \varphi$ and $a_2 = a_1 \circ \varphi$.

Proof. Let $c_1 \in \mathrm{Emb}$ and $c_2 = c_1 \circ \varphi$ for some $\varphi \in \mathrm{Diff}$. Let further $v \in T_{[c_1]}\mathrm{B}$ and let $a_1 = \pi^{-1}_{c_1,*}(v)$. There is then a horizontal C^1-path $c_{1,t}$ in Emb with $\partial_t c_{1,t}|_{t=0} = a_1$. The path $c_{1,t} \circ \varphi$ is a horizontal path through c_2 at $t = 0$. By differentiation we find that it is tangent to $a_2 = a_1 \circ \varphi$ in $t = 0$ with $a_2 \in H_{c_2}\mathrm{Emb}$. Since $\pi(c_{1,t}) = \pi(c_{2,t})$ for all times, we must have that $\mathcal{D}\pi(c_1, a_1) = \mathcal{D}\pi(c_2, a_2) = ([c_1], v)$. Hence $\pi_{c_2,*}(a_2) = v$ and therefore $a_2 = \pi^{-1}_{c_2,*}(v)$. This establishes (2.6.6).

Hence, analogously to Emb, we introduce an equivalence relation on $H\mathrm{Emb}$ by stating $(c_1, a_1) \sim (c_2, a_2)$ if $c_2 = c_1 \circ \varphi$ and $a_2 = a_1 \circ \varphi$ for some $\varphi \in \mathrm{Diff}$. We can then represent the point $([c_1], v) \in T\mathrm{B}$ by the equivalence class $[(c_1, a_1)]$ in $H\mathrm{Emb}$. By virtue of (2.6.6) all elements in $[(c_1, a_1)]$ consistently represent the same element $([c_1], v)$ and by virtue of horizontal lifting we know that every point in $([c], v) \in T\mathrm{B}$ has a representing equivalence class with one element in $H_c\mathrm{Emb}$ for every $c \in [c]$. □

Finally we need to establish how to verify convergence on B.

Proposition 2.6.6 (Convergence on B [SS13a, Prop. 2.6])**.** *A sequence* $[c_n]$ *in* B *converges to some* $[c] \in \mathrm{B}$ *if and only if there is a sequence* c'_n *and a point* c' *in* Emb *with* $c'_n \in [c_n]$ *for all* n *and* $c' \in [c]$ *such that* $c'_n \to c'$ *in* Emb*.*

Proof. The 'if' part follows immediately from the continuity of π. The 'only if' part works as follows: let U be an open neighbourhood of $[c]$ in B such that $\pi^{-1}(U) \simeq U \times \mathrm{Diff}$. Then, since $[c_n] \to [c]$, all $[c_n]$ will eventually lie in U. We can then pick any element φ from Diff and employ the local isomorphism of the fiber bundle to turn the sequence $([c_n], \varphi)$ into some sequence in Emb converging to the c corresponding to $([c], \varphi)$. □

2.6.2 Flows and Diffeomorphisms

Flow-fields and the diffeomorphisms they induce are important tools in this thesis because of the way they act on subsets of $\mathbb{R}^2$. We collect some corresponding facts.

Let B denote the open unit ball in $\mathbb{R}^2$ centered at the origin and

$$B_0 = \{x \in B : x_1 = 0\}, \quad B_+ = \{x \in B : x_1 > 0\}, \quad B_-\{x \in B : x_1 < 0\}\,. \tag{2.6.7}$$

Definition 2.6.7 ([DZ11, Def. 3.1] [SS13a, Def. 2.7])**.** A subset $\Omega \subset \mathbb{R}^2$ is *locally of class* C^k if for any $x \in \partial\Omega$ there exists a neighbourhood $U(x)$ of x and a map $g_x \in C^k\big(U(x), B\big)$ with inverse $g^{-1} \in C^k\big(B, U(x)\big)$ such that

$$g_x\big(\operatorname{int}\Omega \cap U(x)\big) = B_+, \quad g_x\big(\partial\Omega \cap U(x)\big) = B_0\,. \tag{2.6.8}$$

If g_x and g_x^{-1} are also bi-Lipschitzian for all $x \in \partial\Omega$ then Ω is said to be *locally k-Lipschitzian.*

If an open set Ω of class C^∞ is simply connected, its boundary $\partial\Omega$ is diffeomorphic to S^1 and can be parametrized by a map $c \in \mathrm{Emb}$.

In the context of image segmentation an important type of shape functionals is the integration of a given function over the interior of the shape. The following Lemma gives the derivative of such an integration in the contour representation w.r.t. a contour deformation.

Lemma 2.6.8 (Shape Derivative)**.** *For a family of contours* $[0,1] \ni t \mapsto c_t \in \mathrm{Emb}$ *which is* C^1 *in time and for some* $\phi \in C^\infty_{\mathrm{loc}}(\mathbb{R}^2)$

$$\frac{d}{dt}\int_{\Omega(c_t)} \phi \, dt = \int_{\partial\Omega(c_t)} \phi \, \langle \dot{c}_t \circ c_t^{-1}, n_{c_t} \circ c_t^{-1} \rangle \, ds \,. \tag{2.6.9}$$

Proof. By virtue of [DZ11, Chap. 4, Sect. 3.3.2], for any given time $\tilde{t} \in [0,1]$, we can use the normal component of the time derivative of c_t and extend it to some C^∞-field around the boundary. This field will only describe the deformation correctly up to first order in t, being exact only in $\tilde{t}$ itself. This is however sufficient to apply [DZ11, Chap. 9, Thm. 4.2] for this instant. As we can do this for any t, the proof is complete. □

Now we make some definitions similar to [You10, Sect. 8.2.1]. The goal is to establish existence and regularity of diffeomorphisms associated to flow-fields by integration.

For some bounded open $\Omega \subset \mathbb{R}^2$ and a positive integer p we denote by $C^p_0(\Omega, \mathbb{R}^2)$ the Banach space of continuously differentiable vector fields α on Ω, such that the support of α and its derivatives up to p-th order lies within Ω. Denote the corresponding norm by

$$\|\alpha\|_{\Omega,p,\infty} = \sum_{I\colon |I|\le p} \|\partial_I \alpha\|_{\Omega,\infty} \tag{2.6.10}$$

with $\|\cdot\|_{\Omega,\infty}$ denoting the supremum-norm on Ω.

We then define the set of absolutely integrable functions from $[0,1]$ to $C^p_0(\Omega, \mathbb{R}^2)$ by

$$\mathcal{X}_p(\Omega) = \{\alpha : [0,1] \to C^p_0(\Omega, \mathbb{R}^2) : \ \|\alpha\|_{\mathcal{X}_p(\Omega)} < \infty\}$$

with

$$\|\alpha\|_{\mathcal{X}_p(\Omega)} = \int_0^1 \|\alpha_t\|_{\Omega,p,\infty} \, dt \,. \tag{2.6.11}$$

Given these regularity conditions, we find:

Theorem 2.6.9 ([You10, Thms. 8.7,8.9])**.** *A flow-field path* $\alpha \in \mathcal{X}_p(\Omega)$ *induces a family of diffeomorphisms* φ_t, $t \in [0,1]$, *on* Ω *via the differential equation*

$$\partial_t \varphi_t = \alpha_t \circ \varphi_t, \qquad \varphi_0 = \mathrm{id} \,. \tag{2.6.12}$$

φ_t is p-times differentiable and for all I with $|I| \leq p$ have

$$\partial_t \partial_I \varphi_t = \partial_I (\alpha_t \circ \varphi_t) \tag{2.6.13}$$

with corresponding initial conditions.

We will later need the following small Lemma, based on the theorem above.

Lemma 2.6.10 (Uniform convergence of φ_t)**.** *For $\alpha \in \mathcal{X}_p(\Omega)$ the corresponding family of diffeomorphisms φ_t according to Theorem 2.6.9 is continuous in time w.r.t. uniform convergence in its derivatives up to p-th order.*

Proof. In 0-th order we have

$$\|\varphi_{t_1} - \varphi_{t_2}\|_{\Omega,\infty} \leq \int_{t_1}^{t_2} \|\partial_t \varphi_t\|_{\Omega,\infty}\, dt = \int_{t_1}^{t_2} \|\alpha_t\|_{\Omega,\infty}\, dt \to 0 \quad \text{as} \quad t_1 \to t_2\,. \tag{2.6.14}$$

For all orders from 1 up to p the proof can be established by induction: according to Theorem 2.6.9 for any multi-index I, $1 \leq |I| \leq p$ have

$$\partial_t \partial_I \varphi_t = \partial_I (\alpha_t \circ \varphi_t) \tag{2.6.15}$$

Using [You10, Lemma 8.3] to disentangle the expression one finds

$$= \sum_j \big((\partial_j \alpha_t) \circ \varphi_t\big)\big(\partial_I (\varphi_t)_j\big) + C_t \tag{2.6.16}$$

where C_t is a combination of derivatives up to order $|I|-1$ of φ_t and of derivatives up to order $|I|$ of α_t. Assuming the Lemma holds for orders up to $|I|-1$ and using the assumption $\alpha \in \mathcal{X}_p(\Omega)$, we can find some bound $\hat{C}_t$ with $\hat{C}_t \geq \|C_t\|_{\Omega,\infty}$ and $\int_0^1 \hat{C}_t dt < \infty$. Consider then the following ODE:

$$\partial_t \hat{\varphi}_t = \hat{\alpha}_t\, \hat{\varphi}_t + \hat{C}_t \qquad \text{with} \qquad \hat{\alpha}_t = \sum_j \|\partial_j \alpha_t\|_{\Omega,\infty} \tag{2.6.17}$$

For some initial condition $\hat{\varphi}_0 \geq \max_j \|\partial_I (\varphi_0)_j\|_{\Omega,\infty}$ the continuous solution $\hat{\varphi}_t$ to (2.6.17) will satisfy $\hat{\varphi}_t \geq \max_j \|\partial_I (\varphi_t)_j\|_{\Omega,\infty}$. One then has

$$\|\partial_I \varphi_{t_1} - \partial_I \varphi_{t_2}\|_{\Omega,\infty} \leq \int_{t_1}^{t_2} \|\partial_t \partial_I \varphi_t\|_{\Omega,\infty}\, dt \leq \int_{t_1}^{t_2} \hat{\alpha}_t\, \hat{\varphi}_t + \hat{C}_t\, dt \to 0 \text{ as } t_1 \to t_2\,. \tag{2.6.18}$$

For $|I| = 1$ one finds $C_t = 0$, i.e. the first step holds. The higher orders then follow from induction. □

2.6.3 Poisson's Equation

In Section 4.1 the differential of the diffeomorphism between the manifold of contours and the counterpart on the measure manifold will be defined through solutions to Poisson's equation with appropriate data terms. For the analysis of this map general facts on existence, uniqueness and smoothness properties of solutions to Poisson's equation will be used.

We assume for now that $\Omega \subset \mathbb{R}^2$ is a simply connected, bounded, open set with a Lipschitz-continuous boundary.

Lemma 2.6.11 ([GR86, Thm. 2.5])**.** *The operator* $\gamma_n \colon C^\infty(\overline{\Omega}, \mathbb{R}^2) \to C^\infty(\partial\Omega)$ *mapping a vector to its normal component on the boundary, can be continuously extended to an operator* $\tilde{\gamma}_n \colon H(\mathrm{div}; \Omega) \to H^{-1/2}(\partial\Omega)$.

The following Green's formula holds:

$$\int_\Omega u \cdot (\nabla q)\, dx + \int_\Omega (\mathrm{div}\, u) q\, dx = \int_{\partial\Omega} (u \cdot n) q\, ds, \qquad u \in H(\mathrm{div}; \Omega), \quad q \in H^1(\Omega), \tag{2.6.19}$$

where the integral on the r.h.s. is understood as the duality pairing between $H^{-1/2}(\partial\Omega)$ and $H^{1/2}(\Omega)$.

The following Lemma can be deduced from the basic theory of elliptic PDEs [GR86, Prop. 1.2, Cor. 2.7].

Lemma 2.6.12. *Let* $g \in H^{-1/2}(\partial\Omega)$ *be given. Then the mapping* $\tilde{F} \colon H^{-1/2}(\partial\Omega) \to H(\mathrm{div}; \Omega)$ *given by* $\alpha = \tilde{F}(g) = \nabla u$*, where* u *is up to a constant the unique solution to the Neumann problem*

$$\Delta u = f \ \text{ in } \Omega, \qquad \frac{\partial u}{\partial n} = g \ \text{ on } \partial\Omega, \qquad f = \frac{1}{|\Omega|} \int_{\partial\Omega} g ds, \tag{2.6.20}$$

maps g *to the unique vector field* α *with constant divergence* $\mathrm{div}\, \alpha = f$ *and normal component* $\tilde{\gamma}_n(\alpha) = n \cdot \alpha = g$.

The solution u to (2.6.20) inherits additional regularity of the data as follows.

Theorem 2.6.13 ([GR86, Thm. 1.10])**.** *Let* $\Omega \subset \mathbb{R}^2$ *be a locally* $(m+1)$*-Lipschitzian domain with boundary* $\partial\Omega$ *and assume that the data* f *and* g *satisfy*

$$f \in W^{m,p}(\Omega), \qquad g \in W^{m+1-1/p,p}(\partial\Omega), \qquad 1 < p < \infty. \tag{2.6.21}$$

Then $[u] \in W^{m+2,p}(\Omega)/\mathbb{R}$ *and there exists a constant* $C = C(m, p, \Omega)$ *such that*

$$\|[u]\|_{W^{m+2,p}(\Omega)/\mathbb{R}} \leq C(\|f\|_{W^{m,p}(\Omega)} + \|g\|_{W^{m+1-1/p,p}(\partial\Omega)}). \tag{2.6.22}$$

Based on the assumptions of Theorem 2.6.13 and the Sobolev embedding theorem [GR86, Thm. 1.3], the continuous injection

$$W^{m,p}(\Omega) \hookrightarrow C^n(\overline{\Omega}), \qquad 1/p < (m-n)/2 \tag{2.6.23}$$

holds.

Remark 2.6.14. Theorem 2.6.13 in fact holds for more general differential operators than the Laplacian. The weak solution to (2.6.20) is given by the minimizer of

$$E(u,\Omega,f,g) = \frac{1}{2}\int_\Omega \|\nabla u\|^2\,dx + \int_\Omega f\,u\,dx - \int_{\partial\Omega} g\,u\,ds \tag{2.6.24}$$

over $u \in H^1(\Omega)/\mathbb{R}$. More generally, Theorem 2.6.13 holds for minimizers of functionals of the type

$$E(u,\Omega,A,f,g) = \frac{1}{2}\int_\Omega \langle A\nabla u,\nabla u\rangle dx + \int_\Omega f\,u\,dx - \int_{\partial\Omega} g\,u\,ds \tag{2.6.25}$$

with spatially varying matrix $A \in C^{m+1}(\overline{\Omega},\mathbb{R}^{2\times 2})$ such that there are constants $0<\lambda<\Lambda$ with

$$\lambda\|r\|^2_{\mathbb{R}^2} \le \langle A(x)r,r\rangle \le \Lambda\|r\|^2_{\mathbb{R}^2} \tag{2.6.26}$$

for all $x \in \overline{\Omega}$ and $r \in \mathbb{R}^2$.

In particular the dependency of the constant C on A is only through the parameters λ,Λ and on upper bounds to the supremum norms $\sup_{x\in\overline{\Omega}}|\partial_I A_{ij}(x)|$ of derivatives up to order $m+1$ of coefficients of A. That is, for a set of matrices for which common λ,Λ and common upper bounds can be found, the same constant C in Theorem 2.6.13 can be applied uniformly. For a detailed exposition of regularity results based on C^∞ assumptions, we refer to [Sch59, ADN59].

Now we adopt the settings for the main part of this paper, that is Ω is of class C^∞, bounded and simply connected, f is constant as in (2.6.20) and $g \in C^\infty(\partial\Omega)$. First we establish additional regularity of the images $\tilde{F}(g)$ in Lemma 2.6.12. If $m,p\to\infty$, as in this setting, then $n < m - 2/p \to \infty$. In view of the isomorphism established by equation (2.6.19) due to Lemma 2.6.12, we conclude

Proposition 2.6.15. *Let $\Omega \subset \mathbb{R}^2$ be a C^∞ domain, $g \in C^\infty(\partial\Omega)$ and $f = |\Omega|^{-1}\int_{\partial\Omega} g\,ds$. Then there is a unique vector field $\alpha \in C^\infty(\overline{\Omega},\mathbb{R}^2)$ with $\langle\alpha,n\rangle = g$ and $\operatorname{div}\alpha = f$, given by the unique solution to the corresponding Neumann problem (2.6.20).*

Chapter 3

Relaxed Gromov-Wasserstein Distance

3.1 Relaxed Gromov-Wasserstein Distance

3.1.1 Setup and Overview

3.1.1.1 A Proposal for a Shape Prior Functional

In Section 2.1 we have briefly discussed some standard functionals for supervised image segmentation based on local criteria such as appearance and local boundary regularity. We have discussed the need to enhance such functionals by a shape prior term to deal with noisy or otherwise corrupted local information. In Section 2.5 the Gromov-Wasserstein distance was introduced as an isometry invariant distance on metric measure spaces, based on matching. We will now use this for the construction of a shape prior which is intrinsically invariant under geometric transformations (recall the discussion in Sect. 2.5.2.3).

Let Y be the image domain in which we want to locate and match the sought-after object and let d_Y be a suitable metric thereon. As discussed in Sect. 2.1.2 we will describe the object location by an indicator function $u : Y \to \{0, 1\}$, which we relax to the interval $[0, 1]$ to simplify optimization. Since we want to use the mass transport based Gromov-Wasserstein distance for shape regularization, we interpret u as the density of a corresponding measure ν w.r.t. the Lebesgue measure $\mathcal{L}_Y$ on Y. The feasible set for ν will be given by $\mathrm{SegMeas}(Y, M)$, as defined in (2.1.8). The role of M will be explained shortly.

Note that this is conceptually different from matching approaches where a certain image feature (usually intensity or gray-level) is directly converted into a density. To handle image appearance information and local regularity of the segmentation, we introduce a function $G : \mathrm{Meas}(Y) \to \mathbb{R}$ that contains typical local components of a segmentation functional, as discussed in Sect. 2.1. We will assume G is convex. The total functional will be comprised of G and a shape prior.

If computational complexity was not an issue, a potential variational approach to image segmentation is

$$\inf_{\nu \in \mathrm{SegMeas}(Y,M)} E_{\mathrm{GW}}(\nu) \qquad \text{with} \qquad E_{\mathrm{GW}}(\nu) = G(\nu) + D_{\mathrm{GW},p}(\mu,\nu)^p \quad (3.1.1)$$

The the triple (X, d_X, μ) is a mm-space that plays the role of prior knowledge by representing a prototype of the sought-after shape, it will be referred to as *template*. X and the metric d_X will model the geometry of the object and μ will be the Lebesgue measure on X with constant density 1 to indicate that 'all of X is part of the template'. The constant

$$M = \mu(X) \quad (3.1.2)$$

is the total mass of μ. A segmentation proposal ν will have a low shape prior energy, if it can be matched onto the template with little geometric distortion via $D_{\mathrm{GW},p}$.

Concerning the discussion in Sect. 1.2.2 we note that both the shape prior as well as the remaining segmentation functional are functions of ν and no conversion between different shape representations is necessary.

Remark 3.1.1 (Generality of functional)**.** The setup of functional (3.1.1) can be applied to very different data structures. X and Y can be open sets in $\mathbb{R}^2$, describing continuous templates and images. Then μ would be, as indicated, the Lebesgue measure on X and $\mathcal{L}_Y$ in (2.1.2) would be the Lebesgue measure on Y. Alternatively X and Y could be discrete sets of pixels in $\mathbb{R}^2$ or point clouds in $\mathbb{R}^n$, then μ and $\mathcal{L}_Y$ should be chosen to be the respective uniform counting measures on X and Y. If X and Y represent an over-segmentation of some data (i.e. super-pixels or voxels), μ and $\mathcal{L}_Y$ would be weighted counting measures, the weights representing the area/volume of each cell.

Of course however, computational complexity is in fact a crucial issue. Computing the value of $D_{\mathrm{GW},p}$ requires solving a non-convex optimization problem and it is non-convex as a function of ν, thus the problem (3.1.1) is not practical.

3.1.1.2 Chapter Overview

In Section 3.1.2 we will discuss a convex relaxation, based on Definition 2.5.7. This will render the problem feasible while preserving the important isometry invariance. The particular case of a linear appearance term G will be treated in Section 3.1.3. Then the relaxed functional will be an optimal transport problem and we will prove for the discrete case that the segmentation measure ν is binary (i.e. its density is binary) and the optimal coupling $\pi \in \Pi(\mu,\nu)$ induces a bijection between the template and the segmented region (Sect. 3.1.3.1). In Sect. 3.1.3.2 we discuss inhomogeneous appearance models where the corresponding functionals do no longer just depend on the marginal ν. This allows

modelling additional prior information about the sought-after object. A reformulation of the computation of the linear cost function (cf. Def. 2.5.7) in terms of radial mass distributions is discussed in Sect. 3.1.4 (see also Proposition 3.1.9). This is important both from the aspect of interpreting the relaxation as well as for efficient numerical implementation. Numerical experiments for illustrating all discussed aspects of the proposed approach will be presented in Section 3.1.5.

It is to be expected that the presented relaxation will be too loose in difficult cases with very noisy appearance data. In practice this results in spatially very irregular assignments between template and image. Therefore, to strengthen the relaxation, in Sect. 3.2 some proposals to enforce regularity of the assignment are discussed.

3.1.2 Linearization

3.1.2.1 Linearizing the Gromov-Wasserstein Distance

The first thing which may come to mind to make (3.1.1) feasible is to replace $D_{\mathrm{GW},p}$ by $D_{\mathrm{LinGW},p}$ (Def. 2.5.7):

$$E_{\mathrm{LinGW}}(\nu) = G(\nu) + D_{\mathrm{LinGW},p}(\mu,\nu)^p\,. \tag{3.1.3}$$

By virtue of Proposition 2.5.8 we immediately find that (3.1.3) is a lower bound to (3.1.1), for an interpretation recall Remark 2.5.10. The optimization problem involved in evaluating this functional is convex. However it is itself not convex as a function of ν. So while we can for a segmentation proposal ν compute the corresponding score, we still cannot easily find the globally optimal proposal.

3.1.2.2 Static Cost Function

The non-convexity arises from the dependency of $c_{\mathrm{LinGW},p}$ on ν. We therefore try to estimate a *static* cost function (in the sense that it is no longer a function of ν) by exploiting an appearance model.

Consider the following definition:

Definition 3.1.2 (Relaxed Gromov-Wasserstein distance with static cost function [SS13c, Def. 4.2])**.**

$$D_{\mathrm{LinGW},p}(\mu,\nu,G) = \inf_{\pi\in\Pi(\mu,\nu)} \frac{1}{2}\left(\int_{X\times Y} c_{\mathrm{LinGW},p}(\mu,G;x,y)\,d\pi(x,y)\right)^{1/p} \tag{3.1.4}$$

with

$$c_{\mathrm{LinGW},p}(\mu,G;x,y) = \inf_{\nu'\in\mathrm{SegMeas}(Y,M)} \Big(c_{\mathrm{LinGW},p}(\mu,\nu';x,y) + G(\nu')\Big) \tag{3.1.5}$$

Recall that $\mathrm{SegMeas}(Y, M)$ is the set of segmentation measures on Y with total mass M (Definition 2.1.2).

Compared to $D_{\mathrm{LinGW},p}(\mu, \nu)$ this relaxation goes one step further: for each pair (x, y) to compute $c_{\mathrm{LinGW},p}(\mu, G; x, y)$ not only a locally optimal coupling π' is estimated as in $c_{\mathrm{LinGW},p}(\mu, \nu; x, y)$, but a locally optimal segmentation $\nu' \in \mathrm{SegMeas}(Y, M)$, based on the appearance model G and the metric matching in $c_{\mathrm{LinGW},p}(\mu, \nu'; x, y)$. This potentially introduces inconsistencies as the various optimal marginals ν' estimated for different pairs (x, y) need not agree. The relaxation is however necessary to obtain a functional that is convex in the segmentation proposal ν.

We can relate the relaxations $D_{\mathrm{LinGW},p}(\mu, \nu)$ and $D_{\mathrm{LinGW},p}(\mu, \nu, G)$ as follows:

Proposition 3.1.3 ([SS13c, Prop. 4.2])**.**

$$D_{\mathrm{LinGW},p}(\mu, \nu, G)^p \leq D_{\mathrm{LinGW},p}(\mu, \nu)^p + M \cdot G(\nu) \tag{3.1.6}$$

where $M = \mu(X)$.

Proof. By definition for each $\nu \in \mathrm{SegMeas}(Y, M)$ (recall: $M = \mu(X)$) we have

$$c_{\mathrm{LinGW},p}(\mu, G; x, y) \leq \Big(c_{\mathrm{LinGW},p}(\mu, \nu; x, y) + G(\nu) \Big) \,. \tag{3.1.7}$$

Thus we have for any $\pi \in \Pi(\mu, \nu)$

$$\int_{X \times Y} c_{\mathrm{LinGW},p}(\mu, G; x, y) \, d\pi(x, y) \tag{3.1.8}$$

$$\leq \int_{X \times Y} \Big(c_{\mathrm{LinGW},p}(\mu, \nu; x, y) + G(\nu) \Big) \, d\pi(x, y) \tag{3.1.9}$$

$$= \int_{X \times Y} c_{\mathrm{LinGW},p}(\mu, \nu; x, y) \, d\pi(x, y) + M \cdot G(\nu) \,. \tag{3.1.10}$$

The claim follows. □

Analogous to (3.1.1) this leaves us with the following overall optimization problem:

$$\inf_{\nu \in \mathrm{SegMeas}(Y, M)} E_{\mathrm{LinGW,static}}(\nu) \tag{3.1.11a}$$

with

$$E_{\mathrm{LinGW,static}}(\nu) = D_{\mathrm{LinGW},p}(\mu, \nu, \lambda_1 \cdot G)^p + \lambda_2 \cdot G(\nu) \tag{3.1.11b}$$

The appearance model G will first be used to estimate the static linear cost function $c_{\mathrm{LinGW},p}(\mu, G; \cdot, \cdot)$ and can be used again in the final optimization over ν. The constants $\lambda_1, \lambda_2 \geq 0$ control the weight of the two influences.

Summarizing the two relaxation steps (3.1.1) → (3.1.3) → (3.1.11) we conclude:

$$D_{\mathrm{LinGW},p}(\mu,\nu,\lambda_1\cdot G)^p+\lambda_2\cdot G(\nu) \tag{3.1.12a}$$
$$\leq D_{\mathrm{LinGW},p}(\mu,\nu)^p+(\lambda_1+\lambda_2\cdot M)\,G(\nu) \tag{3.1.12b}$$
$$\leq D_{\mathrm{GW},p}(\mu,\nu)^p+(\lambda_1+\lambda_2\cdot M)\,G(\nu) \tag{3.1.12c}$$

The relaxation (3.1.12c) → (3.1.12b) is necessary because evaluation of (3.1.12c) involves a non-convex problem. (3.1.12b) → (3.1.12a) is required because (3.1.12b) is still non-convex as a function of ν.

Remark 3.1.4. Note that these bounds hold for any convex appearance term G. Also, in principle different appearance models G_1 and G_2 could be used for the cost function estimation and the final matching.

3.1.2.3 Metric on Y

In Section 2.5 it was discussed how geometric invariance can be achieved by imposing a suitable metric on the mm-spaces to be compared. For rigid objects the Euclidean metric of the ambient space is the obvious option, enabling recognition of an object in any translated, rotated or flipped state. For articulated objects that can appear in different poses the geodesic metric is a more appropriate choice. It can be computed efficiently by fast marching algorithms [Set95]. To increase robustness to topological noise, more robust variants as for example the diffusion metric [CL06] can be applied (see for example [BBK+10]).

Having said this it must be pointed out that there is a fundamental issue about the geodesic (and diffusion) metric in the context of our application: the geodesic metric on Y is given by the shortest path between two points that runs within the object. The area of the object in the image is described by the segmentation measure ν. This implies that the metric fidelity coefficients $\Gamma(x,y,x',y')$ (2.5.7) depend on ν in a highly non-trivial fashion. This leads to severe computational obstacles: The feasible set for ν, $\mathrm{SegMeas}(Y,M)$, contains non-binary measures for which it is unclear how a meaningful notion of geodesic distance should be defined. And of course this dependency of the metric on the segmentation would result in a highly non-convex functional. To obtain feasible problems, the metric on Y must be fixed beforehand. In most cases this limits application to the Euclidean metric. We will present in Sect. 3.1.5 some experiments with particular cases where this issue can be partially circumvented and more sophisticated metrics can be applied.

3.1.3 Linear Appearance Term

We have just presented a way how a generic appearance model, based on local features, can be combined with relaxations to the Gromov-Wasserstein distance to incorporate prior knowledge on shape. Let us now consider the particular case of a linear appearance term and investigate its specific properties.

3.1.3.1 Bijective Shape Matching

In this section we will assume that X and Y are finite spaces with $|X| \leq |Y|$. As done earlier, we can then identify measures on X, Y and $X \times Y$ with their densities which we can treat as vectors of dimensions $|X|, |Y|$ and $|X| \cdot |Y|$ respectively and index their entries by elements in X, Y and $X \times Y$. Assume that μ is the counting measure on X, that is each point $x \in X$ carries unit mass: $\mu(x) = 1$ for all $x \in X$. As reference measure $\mathcal{L}_Y$ on Y for defining the set of segmentation measures (see Definition 2.1.2) we pick the counting measure on Y.

Remark 3.1.5. The assumption $|X| \leq |Y|$ is certainly reasonable for the application of shape segmentation: (Y, d_Y) represents the whole image of which the object we are looking for, its shape described by (X, d_X) and its location given by ν, only takes up a fraction. If the object was larger than the image, we could not make out its shape anyway.

We now consider the case when G is linear:

$$G(\nu) = \int_Y g(y)\, d\nu(y) = \sum_{y \in Y} g(y)\, \nu(y) \tag{3.1.13}$$

The coefficients $g(y) \in \mathbb{R}$ indicate for each pixel its affinity for foreground ($g(y) < 0$) or background ($g(y) > 0$). Recall from Remark 2.1.1 that also a simple linear term allows incorporation of sophisticated appearance models.

In this case we find:

Proposition 3.1.6 (Binary shape segmentation and deterministic matching [SS13c, Props. 4.3 and 4.4])**.** *For discrete finite X and Y with $|X| \leq |Y|$ with μ being the counting measure on X and the segmentation reference measure $\mathcal{L}_Y$ (see Definition 2.1.2) being the counting measure on Y for a real finite cost function $c : X \times Y \to \mathbb{R}$ and for G being linear as in (3.1.13), for the problem*

$$\inf_{\nu \in \mathrm{SegMeas}(Y,M)} E(\nu) \quad \text{with} \quad E(\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \sum_{(x,y) \in X \times Y} c(x,y)\, \pi(x,y) + G(\nu) \tag{3.1.14}$$

there is a binary optimal segmentation ν and a deterministic optimal coupling $\pi \in \Pi(\mu, \nu)$ that induces a bijection between the template X and the foreground of Y, as indicated by ν.

By picking $c(x, y) = c_{\mathrm{LinGW},p}(\mu, G; x, y)$ this applies to $E_{\mathrm{LinGW,static}}$ (3.1.11).

Proof.

$$\inf_{\nu \in \mathrm{SegMeas}(Y,M)} E(\nu) \tag{3.1.15}$$

$$= \inf_{\nu \in \mathrm{SegMeas}(Y,M)} \inf_{\pi \in \Pi(\mu,\nu)} \sum_{(x,y) \in X \times Y} c(x,y)\, \pi(x,y) + \sum_{y \in Y} g(y)\, \nu(y) \tag{3.1.16}$$

Since ν can be reconstructed from π via marginalization, $\nu = \mathrm{pr}_{Y\sharp}\pi$, one can merge the nested optimization problem over ν and π. Let

$$\begin{aligned}\mathrm{SegCoupl}(Y,\mu) &= \bigcup_{\nu\in\mathrm{SegMeas}(Y,M)} \Pi(\mu,\nu) \\ &= \left\{\pi \in \mathrm{Meas}(X\times Y) \,:\, \mathrm{pr}_{X\sharp}\pi = \mu \wedge \mathrm{pr}_{Y\sharp}\pi \leq \mathcal{L}_Y\right\}. \quad (3.1.17)\end{aligned}$$

We can then write for the nested optimization above:

$$= \inf_{\pi\in\mathrm{SegCoupl}(Y,\mu)} \sum_{(x,y)\in X\times Y} c(x,y)\,\pi(x,y) + \sum_{y\in Y} g(y) \sum_{x\in X} \pi(x,y) \quad (3.1.18)$$

$$= \inf_{\pi\in\mathrm{SegCoupl}(Y,\mu)} \sum_{(x,y)\in X\times Y} \Big(c(x,y) + g(y)\Big)\,\pi(x,y) \quad (3.1.19)$$

Let the set $\hat{X} = X \sqcup \{1,\dots,|Y|-|X|\}$ be an extension of the set X by dummy elements such that it has the same cardinality as Y and let $\hat{\mu}$ be the counting measure on $\hat{X}$. It is obvious that any $\hat{\pi} \in \Pi(\hat{\mu},\mathcal{L}_Y)$ induces a $\pi' \in \mathrm{SegCoupl}(Y,\mu)$ by restriction from $\hat{X}\times Y$ to $X\times Y$ and that any $\pi' \in \mathrm{SegCoupl}(Y,\mu)$ can be written as the restriction of some $\hat{\pi} \in \Pi(\hat{\mu},\mathcal{L}_Y)$. One then can rewrite (3.1.19) as

$$= \inf_{\hat{\pi}\in\Pi(\hat{\mu},\mathcal{L}_Y)} \sum_{(x,y)\in(X\times Y)} \Big(c(x,y) + g(y)\Big)\,\hat{\pi}(x,y) \quad (3.1.20)$$

$$= \inf_{\hat{\pi}\in\Pi(\hat{\mu},\mathcal{L}_Y)} \sum_{(x,y)\in(\hat{X}\times Y)} \hat{c}(x,y)\,\hat{\pi}(x,y) \quad (3.1.21)$$

with

$$\hat{c} \in \mathbb{R}^{|\hat{X}|\times|Y|}, \quad \hat{c}(x,y) = \begin{cases} c(x,y) + g(y) & \text{if } x \in X, \\ 0 & \text{else} \end{cases}. \quad (3.1.22)$$

Note that there is a natural one-to-one correspondence between $\Pi(\hat{\mu},\mathcal{L}_Y)$ and $\mathrm{Stoch}_{|Y|}$. So problem (3.1.21) is equivalent to the linear programming formulation of the LAP (2.2.5). From Corollary 2.2.2 follows then the existence of binary optimizers $\hat{\pi}^*$ to (3.1.21) and through restriction to $X\times Y$ follows existence of a binary optimizer $\pi^* \in \mathrm{SegCoupl}(Y,\mu)$ to (3.1.19). Naturally, taking the Y marginal of π^* provides a corresponding binary optimal segmentation ν^*. The restriction of $\hat{\pi}^*$ to $X\times(\mathrm{spt}\,\nu^*)$ in Y is a permutation matrix that induces a bijection between the template and the segmented foreground region $\mathrm{spt}\,\nu^*$. □

Remark 3.1.7 (Merged Optimizations)**.** The merging of the nested optimization over ν and π in (3.1.16) to the joint linear program (3.1.19) shows how problem (3.1.11) can be solved and evaluated at the optimum in a single pass. The same reformulation applies for the computation of $c_{\mathrm{LinGW},p}(\mu,G;x,y)$ (3.1.5) for linear G.

We will also apply this merging in Sect. 4.2.

3.1.3.2 Inhomogeneous Appearance Models

The proof of Proposition 3.1.6 provides inspiration for a natural extension of the linear appearance term (3.1.13). In (3.1.19) the term $\sum_y g(y)\,\nu(y)$ is pulled into the linear matching cost. So instead of uniform coefficients $g(y)$ that specify homogeneously for each $y \in Y$ its affinity for being foreground, 'localized' coefficients $g(x, y)$ are conceivable that specify the affinity of a pixel y to a specific part of the template. The proof still holds if G is no longer a function of just ν but in fact of the full coupling π:

$$G(\pi) = \int_{X\times Y} g(x, y)\, d\pi(x, y) \tag{3.1.23}$$

This would, for example, allow different appearance models for different parts of the template, which is a quite natural: the head is expected to look different from the torso. We will therefore refer to such models as *inhomogeneous*.

A particular subset of such models can be constructed in the following manner: let $\mathcal{F}$ be some sort of feature space (this could be the real line, a discrete set of labels or a space of patch-based local signatures such as filter responses or gradient histograms) and let $c_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a matching similarity function thereon. We could then equip any element $x \in X$ with an expected feature f_x, every $y \in Y$ with an observed feature f_y and define the appearance based matching cost

$$g(x, y) = c_{\mathcal{F}}(f_x, f_y)\,. \tag{3.1.24}$$

Note that the inhomogeneous appearance model can not only be applied in the global final matching through the second term in $E(\nu)$ (3.1.11) but can also, through the computation of $c_{\mathrm{LinGW},p}(\mu, G; \cdot, \cdot)$, influence the first term.

We will in Section 3.1.5 demonstrate the benefit of this additional modelling power.

3.1.4 Radial Distribution Comparison

Through the relaxations presented in Sect. 3.1.2 the original non-convex functional (3.1.1) has been turned into a convex problem (3.1.11). For the particular choice of a linear appearance model (3.1.13) all involved sub-problems are linear programs (cf. Remark 3.1.7).

Still, even for the case that X, Y are super-pixel over-segmentations, the cardinality of X and Y is easily of the order 10^3, naïvely leaving us with about a million cost function values $c_{\mathrm{LinGW},p}(\mu, G; x, y)$ to compute, each of which requires the solution of an optimization problem with about a million variables.

In this thesis we will present several approaches to significantly reduce the naïve computational effort. One possibility is the reformulation of the static cost function $c_{\mathrm{LinGW},(}\mu, \nu; \cdot, \cdot)$ in terms of radial mass distributions (Proposition 3.1.9) which we will examine in more detail here and provide an extension

to $c_{\mathrm{LinGW},(}\mu, G; \cdot, \cdot)$ for linear G based on a feature space cost (3.1.24). Another possibility to effectively reduce dimensionality of the involved problems is the multi-scale approach presented in Chapter 5.

3.1.4.1 Push-forwards and Couplings

As in Sect. 3.1.3.1 we will now assume that X and Y are discrete and finite. We will need the following statement:

Proposition 3.1.8 (Push-forwards and couplings [SS13c, Prop. 4.5])**.** *Let $\mu \in \mathrm{Meas}(X)$, $\nu \in \mathrm{Meas}(Y)$ be two measures with equal total mass. For two maps $\phi_X : X \to S_X$ and $\phi_Y : Y \to S_Y$ that take X and Y to two discrete sets S_X, S_Y, denote by ϕ the product map $\phi(x, y) = \big(\phi_X(x), \phi_Y(y)\big)$. Then one finds*

$$\phi_\sharp \Pi(\mu, \nu) = \Pi(\phi_{X\sharp}\,\mu, \phi_{Y\sharp}\,\nu)\,. \tag{3.1.25}$$

That is, the push-forwards of the couplings are the couplings of the push-forwards.

Proof. For any $\pi \in \Pi(\mu, \nu)$ get

$$\begin{aligned}(\phi_\sharp \pi)(\sigma) &= \pi(\phi^{-1}(\sigma)) \geq 0\\ (\phi_\sharp \pi)(\sigma_{S_X} \times S_Y) &= \pi(\phi_X^{-1}(\sigma_{S_X}) \times Y)\\ &= \mu(\phi_X^{-1}(\sigma_{S_X})) = (\phi_{X\sharp}\mu)(\sigma_{S_X})\end{aligned}$$

and analogous

$$(\phi_\sharp \pi)(X \times \sigma_{S_Y}) = (\phi_{Y\sharp}\nu)(\sigma_{S_Y})$$

for all measurable $\sigma \subset S_X \times S_Y, \sigma_{S_X} \subset S_X, \sigma_{S_Y} \subset S_Y$. Thus $\phi_\sharp \Pi(\mu, \nu) \subset \Pi(\phi_{X\sharp}\mu, \phi_{Y\sharp}\nu)$.

We now show by construction for any $\rho \in \Pi(\phi_{X\sharp}\mu, \phi_{Y\sharp}\nu)$ the existence of some $\pi \in \Pi(\mu, \nu)$ such that $\rho = \phi_\sharp \pi$. For any element $(s_X, s_Y) \in S_X \times S_Y$ construct the pre-image measure

$$\pi_{(s_X, s_Y)}(x, y) = \begin{cases} 0 & \text{if } \rho(s_X, s_Y) = 0 \vee \\ & (s_X, s_Y) \neq \phi(x, y) \\ \frac{\mu(x)\,\nu(y)}{(\phi_{X\sharp}\mu)(s_X)\,(\phi_{Y\sharp}\nu)(s_Y)}\rho(s_X, s_Y) & \text{else} \end{cases}$$

where this element wise definition for each (x, y) is extended to all subsets of $X \times Y$ by

$$\pi_{(s_X, s_Y)}(\sigma) = \sum_{(x,y)\in\sigma} \pi_{(s_X, s_Y)}(x, y).$$

Now consider $\pi = \sum_{(s_X,s_Y)\in S_X\times S_Y} \pi_{(s_X,s_Y)}$: first verify that it is indeed contained in $\Pi(\mu,\nu)$:

$$\pi(\sigma) \geq 0$$

since $\pi(x,y) \geq 0$ for all (x,y). Furthermore

$$\begin{aligned}
\pi(\sigma_X \times Y) &= \sum_{\substack{x\in\sigma_X\\ y\in Y}} \sum_{\substack{s_X,s_Y:\\ \phi(x,y)=(s_X,s_Y),\\ \rho(s_X,s_Y)>0}} \frac{\mu(x)\,\nu(y)}{(\phi_{X\sharp}\mu)(s_X)\,(\phi_{Y\sharp}\nu)(s_Y)}\rho(s_X,s_Y) \\
&= \sum_{x\in\sigma_X} \sum_{\substack{s_Y:\\ \rho(\phi_X(x),s_Y)>0}} \frac{\mu(x)\,(\sum_{y:\phi_Y(y)=s_y}\nu(y))}{(\phi_{X\sharp}\mu)(\phi_X(x))\,(\phi_{Y\sharp}\nu)(s_Y)}\rho(\phi_X(x),s_Y) \\
&= \sum_{x\in\sigma_X} \sum_{\substack{s_Y:\\ \rho(\phi_X(x),s_Y)>0}} \frac{\mu(x)\,\nu(\phi_Y^{-1}(s_Y))}{(\phi_{X\sharp}\mu)(\phi_X(x))\,(\phi_{Y\sharp}\nu)(s_Y)}\rho(\phi_X(x),s_Y) \\
&= \sum_{x\in\sigma_X} \frac{\mu(x)}{(\phi_{X\sharp}\mu)(\phi_X(x))} \sum_{\substack{s_Y:\\ \rho(\phi_X(x),s_Y)>0}} \rho(\phi_X(x),s_Y) \\
&= \sum_{x\in\sigma_X} \mu(x) = \mu(\sigma_X)
\end{aligned}$$

and likewise

$$\pi(X\times\sigma_Y) = \nu(\sigma_Y)$$

for all measurable subsets $\sigma \subseteq X\times Y$, $\sigma_X \subseteq X$, $\sigma_Y \subseteq Y$.

Now check whether $\phi_\sharp\pi = \rho$:

$$\begin{aligned}
(\phi_\sharp\pi)(\sigma) = \pi(\phi^{-1}(\sigma)) &= \sum_{(x,y)\in\phi^{-1}(\sigma)} \pi(x,y) \\
&= \sum_{\substack{(x,y)\in\phi^{-1}(\sigma):\\ \rho(\phi(x,y))>0}} \frac{\mu(x)\,\nu(y)}{(\phi_{X\sharp}\mu)(\phi_X(x))\,(\phi_{Y\sharp}\nu)(\phi_Y(y))}\rho(\phi(x,y)) \\
&= \sum_{\substack{(s_X,s_Y)\in\sigma\\ \rho((s_X,s_Y))>0}} \sum_{(x,y)\in\phi^{-1}((s_X,s_Y))} \frac{\mu(x)\,\nu(y)}{(\phi_{X\sharp}\mu)(s_X)\,(\phi_{Y\sharp}\nu)(s_Y)}\rho(s_X,s_Y) \\
&= \sum_{\substack{(s_X,s_Y)\in\sigma\\ \rho((s_X,s_Y))>0}} \frac{(\sum_{x\in\phi_X^{-1}(s_X)}\mu(x))\,(\sum_{y\in\phi_Y^{-1}(s_Y)}\nu(y))}{(\phi_{X\sharp}\mu)(s_X)\,(\phi_{Y\sharp}\nu)(s_Y)}\rho(s_X,s_Y) \\
&= \sum_{\substack{(s_X,s_Y)\in\sigma\\ \rho((s_X,s_Y))>0}} \rho(s_X,s_Y) = \rho(\sigma).
\end{aligned}$$

Consequently any $\rho \in \Pi(\phi_{X\,\sharp}\mu, \phi_{Y\,\sharp}\nu)$ is also contained in $\phi_\sharp\Pi(\mu,\nu)$ and the two sets are equal. □

By setting $\phi_X = d_X(x,\cdot)$ and $\phi_Y = d_Y(y,\cdot)$ the next corollary follows.

Proposition 3.1.9 ([SS13c, Cor. 4.6])**.** *Let X, Y be discrete mm-spaces, fix some element $(x,y) \in X \times Y$ and let $\rho_X = \mathrm{RadDist}_X(x)$ and $\rho_Y = \mathrm{RadDist}_Y(y)$ (recall Definition 2.5.11). Then*

$$c_{\mathrm{LinGW},p}(\mu,\nu;x,y) = \inf_{\pi\in\Pi(\rho_X,\rho_Y)} \int_{\mathbb{R}_+^2} |l_1 - l_2|^p \, d\pi(l_1,l_2)\,. \tag{3.1.26}$$

Proof.

$$c_{\mathrm{LinGW},p}(\mu,\nu;x,y) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} |d_X(x,x') - d_Y(y,y')|^p \, d\pi(x',y') \tag{3.1.27}$$

$$= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^2} |l_X - l_Y|^p \, d(\phi_\sharp\,\pi)(l_X,l_Y) \tag{3.1.28}$$

where $\phi(x',y') = \big(d_X(x,x'), d_Y(y,y')\big)$. Then

$$= \inf_{\pi\in\phi_\sharp\Pi(\mu,\nu)} \int_{\mathbb{R}_+^2} |l_X - l_Y|^p \, d\pi(l_X,l_Y) \tag{3.1.29}$$

and now use Proposition 3.1.8 to find

$$= \inf_{\pi\in\Pi(\rho_X,\rho_Y)} \int_{\mathbb{R}_+^2} |l_X - l_Y|^p \, d\pi(l_X,l_Y) \tag{3.1.30}$$

with $\rho_X = \mathrm{RadDist}_X(x) = d_X(x,\cdot)_\sharp\,\mu$ and $\rho_Y = \mathrm{RadDist}_Y(y) = d_Y(y,\cdot)_\sharp\,\nu$. Note that although $\mathbb{R}_+$ is not discrete, Proposition 3.1.8 still applies since the support of ρ_X and ρ_Y is discrete. □

3.1.4.2 Extension to Static Cost Functions

Proposition 3.1.9 cannot be used to speed up the solution of (3.1.11). We need to find an equivalent result for the computation of $c_{\mathrm{LinGW},p}(\mu, G;\cdot,\cdot)$. We consider the linear inhomogeneous case based on a feature space and a corresponding similarity function, introduced in Sect. 3.1.3.2. This naturally covers the homogeneous linear appearance term as a special case.

Let $\mathcal{F}_X : X \to \mathcal{F}$, $x \mapsto f_x$ and $\mathcal{F}_Y : Y \to \mathcal{F}$, $y \mapsto f_y$ be the maps that take points in X and Y to their associated features in $\mathcal{F}$. For a fixed pair $(x,y) \in X \times Y$ let further $\phi_X : X \to \mathbb{R}_+ \times \mathcal{F}$, $x' \mapsto \big(d_X(x,x'), \mathcal{F}_X(x')\big)$ and ϕ_Y correspondingly. As in Proposition 3.1.8 let ϕ be the product map of ϕ_X and ϕ_Y.

We are then prepared for the following manipulations:

$$\begin{aligned} & c_{\mathrm{LinGW},p}(\mu, G; x, y) \\ = & \inf_{\nu \in \mathrm{SegMeas}(Y,M)} \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} \Big(|d_X(x,x') - d_Y(y,y')|^p + c_{\mathcal{F}}(f_{x'}, f_{y'}) \Big) \, d\pi(x', y') \end{aligned} \tag{3.1.31}$$

and by involving Proposition 3.1.8 with ϕ_X, ϕ_Y, ϕ as defined above

$$\begin{aligned} = & \inf_{\nu \in \mathrm{SegMeas}(Y,M)} \inf_{\pi \in \Pi(\phi_{X\sharp}\mu, \phi_{Y\sharp}\nu)} \\ & \int_{(\mathbb{R}_+ \times \mathcal{F})^2} \Big(|l_X - l_Y|^p + c_{\mathcal{F}}(f_X, f_Y) \Big) \, d\pi\big((l_X, f_X), (l_Y, f_Y)\big) \end{aligned} \tag{3.1.32}$$

Analogous to the proof for Proposition 3.1.6 we can summarize the two nested optimizations by introducing a set equivalent to $\mathrm{SegCoupl}(Y, \mu)$ (3.1.17):

$$\mathrm{SegCoupl}(\mathbb{R}_+ \times \mathcal{F}, \phi_{X\sharp}\mu) = \bigcup_{\nu \in \mathrm{SegMeas}(Y,M)} \Pi(\phi_{X\sharp}\mu, \phi_{Y\sharp}\nu) \, . \tag{3.1.33}$$

This set is given by

$$\begin{aligned} \mathrm{SegCoupl}(\mathbb{R}_+ \times \mathcal{F}, \phi_{X\sharp}\mu) = \Big\{ & \pi \in \mathrm{Meas}\left((\mathbb{R}_+ \times \mathcal{F})^2\right) : \\ & \pi\big(\sigma_X \times (\mathbb{R}_+ \times \mathcal{F})\big) = (\phi_{X\sharp}\mu)(\sigma_X) \; \forall \, \sigma_X \subset \mathbb{R}_+ \times \mathcal{F} \wedge \\ & \pi\big((\mathbb{R}_+ \times \mathcal{F}) \times \sigma_Y\big) \leq (\phi_{Y\sharp}\mathcal{L}_Y)(\sigma_Y) \; \forall \, \sigma_Y \subset \mathbb{R}_+ \times \mathcal{F} \Big\} \, . \end{aligned} \tag{3.1.34}$$

This definition looks quite bulky but is in its form completely equivalent to (3.1.17). Figure 3.1 illustrates the measure $\phi_{X\sharp}\mu$ for an example with $\mathcal{F} = \mathbb{R}$. Compared to $c_{\mathrm{LinGW},p}(\mu, \nu; x, y)$ instead of comparing only radial distributions, here one has to compare distributions over radius and features.

Of course using formula (3.1.32) and optimizing over (3.1.34) to compute the static cost $c_{\mathrm{LinGW},p}(\mu, G; x, y)$ pays off in terms of computational cost only if the dimension of $\mathcal{F}$ is low. Otherwise the discretization of $(\mathbb{R}_+ \times \mathcal{F})^2$ will not be possible with significantly less bins as $X \times Y$ would require.

We will revisit the issue of how to efficiently compute $c_{\mathrm{LinGW},p}(\mu, G; x, y)$ and consequently $D_{\mathrm{LinGW},p}(\mu, \nu, G)$ again in Chapter 5, then by means of a multi-scale approach.

3.1.5 Numerical Examples

In this section the potential of the proposed shape prior functional, based on the relaxed linearized Gromov-Wasserstein distance, combined with a linear appearance term (Sect. 3.1.3) for variational image segmentation is demonstrated.

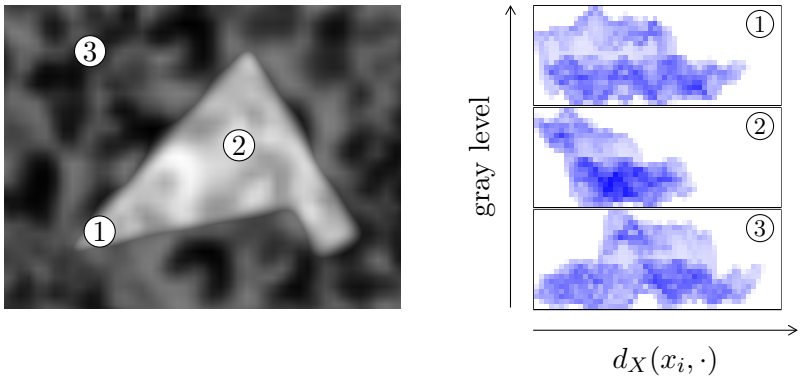


Figure 3.1: Illustration of radial mass distributions. **Left:** example image with gray level as feature. Three points $\{x_i\}_i$, $i = 1, 2, 3$ are marked. We assume a measure $\mathcal{L}_X$ is defined on the image, giving unit mass to every pixel. **Right:** densities of $(d_X(x_i, \cdot), \mathcal{F}_X)_\sharp \mathcal{L}_X$ on $\mathbb{R}_+ \times [0, 1]$ for the three marked points.

Some numerical examples are given to illustrate the favourable properties of the approach and also the limitations implied by the involved approximations.

For a linear appearance term, the exact approach (3.1.1), without approximations, corresponds to a quadratic assignment problem (Sect. 2.3). As X and Y describe the shape template and the test image, their cardinalities are far beyond what one could optimize exactly (Recall $|Y| \gg |X|$, Remark 3.1.5, the unequal cardinalities can formally be handled by introducing dummy nodes.). As a consequence, performing ground truth experiments, i.e. comparing (3.1.12c) to relaxation (3.1.12a), from the viewpoint of optimization is elusive. Therefore the numerical experiments can only investigate the extent to which invariant matching of metric measure spaces can be enforced by the presented convex relaxation approach to shape prior design.

Before presenting numerical results, the next section describes technical details of the implementation, in particular how computational effort can be reduced (including using the results presented in Sec. 3.1.4).

3.1.5.1 Implementation Details and Computational Complexity

The prior mm-spaces (X, d_X, μ) were created from binary images, depicting the template shapes. All pixels with value 0 were removed from the space. The remaining pixels were equipped with the Euclidean metric and μ was set to be the counting measure on these points. For a given experiment (Y, d_Y) represents the test-image grid with Euclidean metric. For most examples the coefficients g that define the appearance term (3.1.13) were constructed from the gray values of the test-image. That is, in those cases we take the feature space to be the real line $\mathcal{F} = \mathbb{R}$. To illustrate the generality of the approach, an example with a more elaborate feature space is also given.

To compute the cost function $c_{\mathrm{LinGW},p}(\mu, G; x, y)$ for some (x, y) one needs to compute a modified mass transport problem on $\mathbb{R}_+ \times \mathcal{F}$ (see Sect. 3.1.4.2). For this $\mathbb{R}_+$ was approximated by a set of equally sized bins over the range of $d_X(x, \cdot)$ and $d_Y(y, \cdot)$. In the case $\mathcal{F} = \mathbb{R}$ a set of bins on $\mathbb{R}$ over the range of relevant features is used. While only inflicting a small discretization error this reduced the involved problem sizes by several orders of magnitude. This method becomes particularly efficient when the affinity coefficients g are binary (e.g. ± 1, indicating unweighted preference for yes/no only). Then $\mathcal{F} = \{0, 1\}$ and one only needs two bins. Also, it is straightforward to parallelize the computation of $c_{\mathrm{LinGW},p}(\mu, G; x, y)$ for all (x, y).

For solving the global matching between X and Y we experimented with constraining the full coupling space $X \times Y$ to smaller subsets to keep the problem size low, while still solving the global problem. Consider the following modification to the partial assignment problem in Proposition 3.1.6 for a given cost function $c : X \times Y \to \mathbb{R}$, usually $c = c_{\mathrm{LinGW},p}(\mu, G; \cdot, \cdot)$. For any $x \in X$ include only a subset $Y_x \subset Y$ with the lowest assignment costs. Then, for every x, add an additional overflow element $y_{\mathrm{of},x}$ and a corresponding variable $\pi(x, y_{\mathrm{of},x})$ whose assignment cost $c(x, y_{\mathrm{of},x})$ is chosen such that $c(x, y_1) \leq c(x, y_{\mathrm{of},x}) \leq c(x, y_2)$ for all $y_1 \in Y_x$, $y_2 \in (Y \setminus Y_x)$. For each x the coupling value $\pi(x, y_{\mathrm{of},x})$ will be taken into account when computing the X-marginal constraint, but there will be no Y-constraints on any of the $y_{\mathrm{of},x}$. Then for any feasible coupling π in the original problem that is non-zero outside of the constrained coupling set, one can create a feasible coupling in the modified problem with non-zero overflow variables, which will yield a lower score. This implies that when solving the restricted partial assignment problem and one gets an optimizer where $\pi(x, y_{\mathrm{of},x}) = 0$ for all x then one knows to have found an optimizer for the original problem with the full coupling space $X \times Y$. As long as some of the overflow variables $\pi(x, y_{\mathrm{of},x})$ are non-zero, the corresponding sets Y_x have to be enlarged and one has to resolve. In "easy" problems this allowed to find global minimizers while considering only $< 5\%$ of the coupling space, "harder" problems were still generally $< 25\%$. In the special case $p = 1$ it is easy to show that both $c_{\mathrm{LinGW},1}(\mu, \nu; \cdot, \cdot)$ and $c_{\mathrm{LinGW},1}(\mu, G; \cdot, \cdot)$ are Lipschitz. Then one can estimate a suitable subset of $X \times Y$ by sub-sampling and lower bounds via the Lipschitz condition without scanning all possible pairs.

A more systematic approach to reducing the naïve computational complexity of the functional will be provided in Chapter 5 in terms of a multi-scale approach (see Figure 5.8).

In the presented experiments $|X|$ is of the order 10^3 and $|Y|$ up to the order of several 10^4. We have set $p = 2$ but we did not observe a substantial change of results for other values $p \geq 1$.

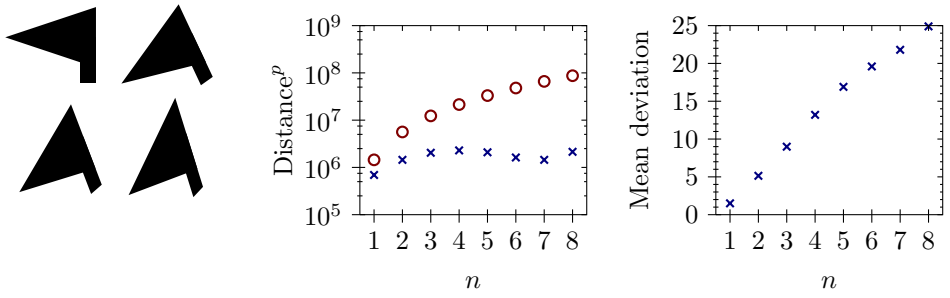
3.1.5.2 Approximation Quality $D_{\mathrm{GW},p} \to D_{\mathrm{LinGW},p}$

The purpose of the first experiment is to gain an insight into the quality of the relaxation $D_{\mathrm{GW},p} \to D_{\mathrm{LinGW},p}$, see Proposition 2.5.8. We take a simple shape, rotate it, distort it by non-isometric but mass preserving scalings with factors q^n, q^{-n} along the vertical and horizontal axis and then compute the optimal coupling according to $D_{\mathrm{LinGW},p}$ between the original and the distortion for various $n > 0$. As an estimate for ground truth we use the coupling induced by the distortion map $(r_1, r_2) \to (q^n \cdot r_1, q^{-n} \cdot r_2)$ and plug it into the Gromov-Wasserstein functional (2.5.10). This provides an upper bound to the true distance.

The results are summarized in Fig. 3.2. For low distortions one can see how $D_{\mathrm{LinGW},p}$ is a good measure for increasing non-isometry, although growing slower than the functional value of the distortion map. For high n the deviation becomes more significant as $D_{\mathrm{LinGW},p}$ decreases, while the upper bound grows further. Here one can assume that the distortion map is no longer the optimal assignment and thus the estimated "ground truth" is in fact too high. There is an additional subtlety in in this experiment: $D_{\mathrm{LinGW},p}$ was computed between two rasterizations of a vector graphic, one as is and one undergoing the distortion transformation. Thus even for $n = 0$ (applying only a rotation) the two resulting metric spaces would not be isometric due to different rasterization. When estimating the ground truth this rasterization cannot be taken into account, since it is unclear how to match the two rasterized graphics. It has thus been estimated on the vector graphics level. The fact that such problems appear even for such simple shapes is a clear indicator of how hard it is to solve the full quadratic problem.

3.1.5.3 Linear Appearance Term and Matching

We want to go beyond measuring the distance between two fixed shapes and perform shape optimization according to the two criteria shape and appearance. The affinity coefficients $g(y)$ for the linear appearance term G are generated from a grayscale image: First an imprint of the template shape is created somewhere in the initially empty test image and then noise is added in different levels. For low noise level the shape restoration works perfectly, without any prior knowledge about the location or the need for proper initialization. Increasing noise will cause an increasing number of assignments to become inaccurate, while remaining roughly correct. Eventually only the coarse location, but no longer the contours of the shape can be recovered (see Fig. 3.3). In Fig. 3.4 the influence of the global weighting parameters λ_1, λ_2 (cf. (3.1.12a)) is demonstrated: For high values of λ_2 the optimal ν is determined locally by the appearance coefficients, for low values the shape prior becomes more dominant and leads to a more accurate restoration of the original contours.



(a) (b) (c)

Figure 3.2: Relaxed linearization of $D_{\mathrm{GW},p}$: (a) top-left to bottom right: original shape and three distortions q^n for $n = 2, 4, 6$ and $q = 0.95$ (dimensions: $\approx 60 \cdot 70 \cdot (\text{length units})^2$). (b) $\times$: $D_{\mathrm{LinGW},p}$ between original and q^n-distortion, $\circ$: Gromov-Wasserstein functional evaluated for coupling induced by distortion map. (c) mean metric deviation between underlying distortion assignment and assignment computed by $D_{\mathrm{LinGW},p}$ (averaged over all assigned pairs). For small $n \leq 4$, $D_{\mathrm{LinGW},p}$ grows with increasing metric distortion, although slower than the estimated "true" Gromov-Wasserstein distance. For $n > 4$, $D_{\mathrm{LinGW},p}$ first starts to decrease a little, before eventually growing again. The assignment computed by $D_{\mathrm{LinGW},p}$ is (up to rasterization errors on the pixel level) identical to the underlying distortion transformation for $n = 1$, deviation grows with increasing non-isometry. From $n \leq 4$ we learn that $D_{\mathrm{LinGW},p}$ is a lower bound that grows with increasing level of non-isometry, which is a favourable property for the functional. For $n > 4$ presumingly the distortion map itself is no longer the best distance-preserving assignment between the two shapes and thus the estimated ground truth value is in fact too high (note how the triangle transforms from being horizontally elongated to vertically elongated). This is an illustration for the difficulty of obtaining ground truth data and the need for relaxations.

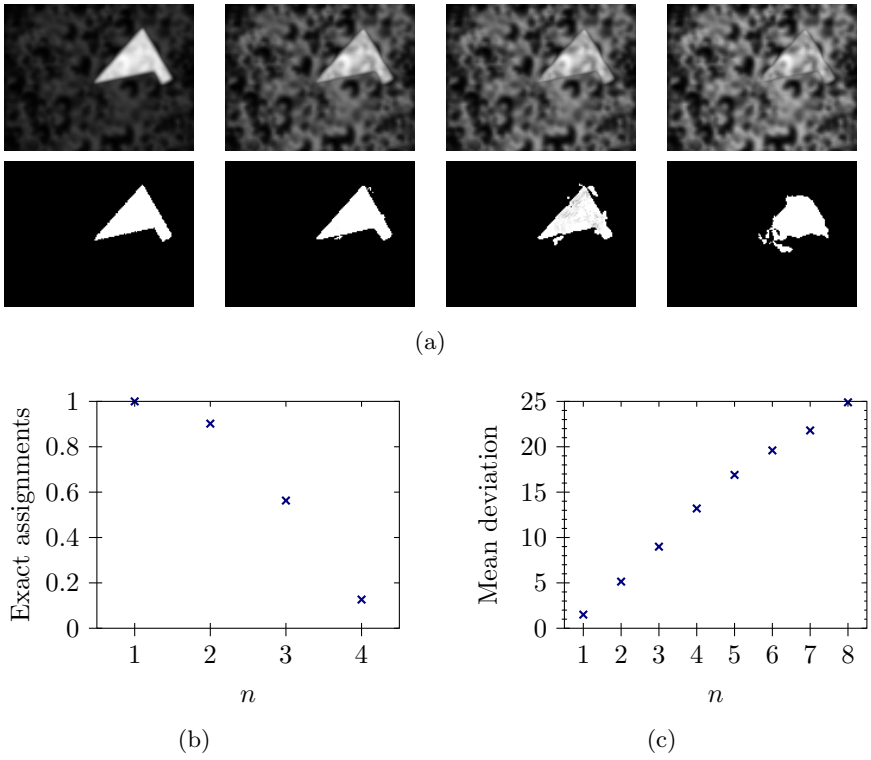
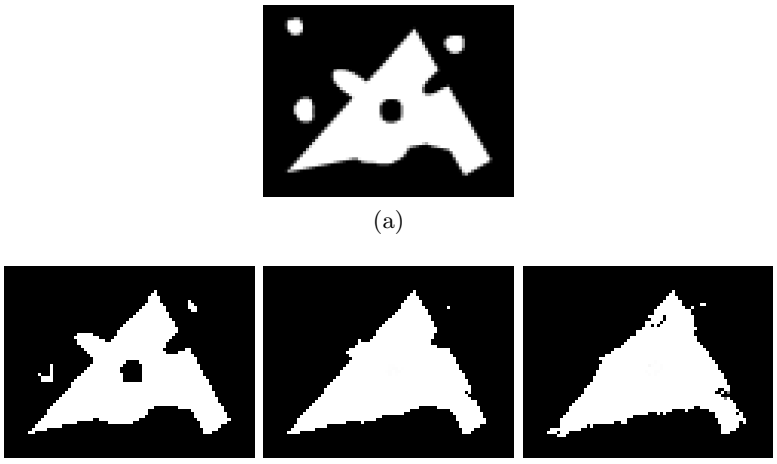


Figure 3.3: Appearance term and noise: (a) top row: appearance coefficients g for the image (bright indicating higher foreground-affinity), noise increasing from left to right. bottom row: corresponding optimal ν according to (3.1.12a), brightness indicating mass density. (X, d_X, μ) as in Fig. 3.2, Y-image dimensions: 160×120(length units)^2. (c) fraction of computed assignments that is closer than 3 pixels to the underlying transformation (c) mean metric deviation between true underlying transformation and assignment computed by (3.1.12a) (averaged over all assigned pairs). For low noise levels the appearance term in combination with the metric information can compensate for noisy appearance data and correctly restore the original shape. Although the location and orientation of the shape within Y is not known a priori, the isometry invariant approach can extract the correct transformation. With higher noise levels the number of assignments that is led astray increases, starting to erode the shape contours, although the majority remains correct. Finally, for very high noise levels the relaxation becomes too loose and hardly any of the assignments are exact. The rough location of the object is however still detected. Probably here the local Y-marginal estimation during the computation of $c_{\mathrm{LinGW},p}(\mu, G; \cdot, \cdot)$ is no longer powerful enough. For better results a more global approach would be required that exchanges information between different local estimates (see Sect. 3.2).



(a)

(b)

Figure 3.4: Influence of the global regularization parameters λ_1, λ_2: (a) appearance coefficients g, representing a transformed, distorted version of the original shape (see Fig. 3.2). (b) from left to right: optimal ν for $\lambda_2 = 10^4, 10^3, 10^2$, λ_1 is fixed to 1. With decreasing λ_2 the shape prior term becomes more influential and pushes towards restoration of the original shape. It should be noted here, that tiny holes or jagged contours in the optimal segmentation regions are not due to numerical instabilities of the optimization implementation but due to discretization artifacts. Sometimes, from the metric point of view, it is better to drop single pixels when matching two different rasterizations of the same shape. Spatial regularity of ν on the rasterization scale is not enforced by the used functional, so this does not increase the functional value.

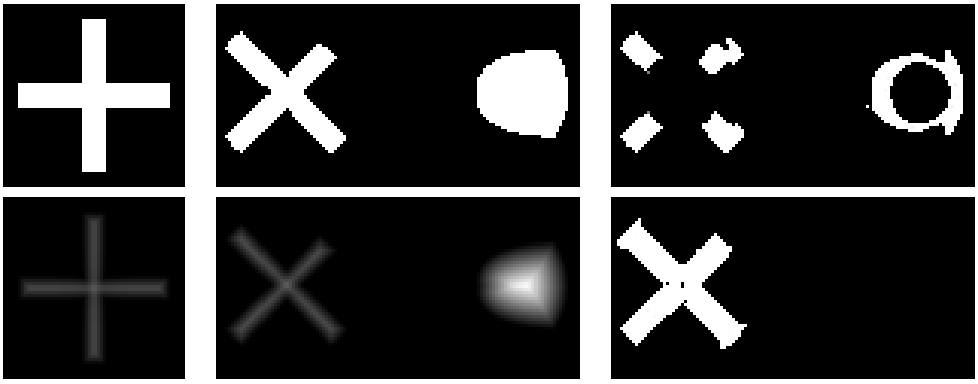
3.1.5.4 Pseudo-Local Features

When the appearance coefficients g are binary, for example ± 1, indicating unweighted preference for fore- or background, one can interpret the region $\{y \in Y : g(y) = -1\}$ as a noisy foreground proposal and extract additional information from this region. An example for this would be the distance transform (for a noise resistant alternative see [GGS+06]). In a scenario with strong appearance classifier one might assume that this preliminary foreground region already resembles the true sought-after region. Thus the distance transform might yield similar values in corresponding places of the template and the image and can therefore be used as a feature for an inhomogeneous appearance model as discussed in Sect. 3.1.3.2, with f_x, f_y being the values of the distance transforms, $\mathcal{F}$ the real line and $c_{\mathcal{F}}$ for example the absolute difference. These features are pseudo-local in the sense that the local value of the distance transform already encodes some information about its neighbourhood.

The experiment presented in Fig. 3.5 has been specifically designed to demonstrate how the local estimation of ν during the computation of the cost $c_{\mathrm{LinGW},p}(\mu, G; \cdot, \cdot)$ via the simple homogeneous binary features can fail: for points near the center of the cross of the template the outer regions of the "blob" on the right of the input image appear more suitable than the center of the actually corresponding cross, where one "arm" has been shortened. By including the additional information encoded in the distance transform features this mismatches can be fixed.

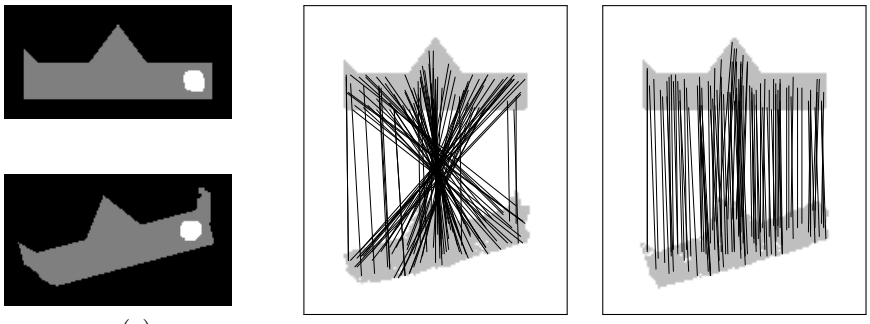
The setup of Fig. 3.5 is also well suited to discuss the implications of the convexity of the functional. A major advantage is the independence of initialization. An approach based on active contours would, if initialized around the blob, be stuck on the right hand side no matter how bad the matching cost will be. The contour could not leave the blob and move through an area without any indication of foreground (and thus without reasonable gradient information). The proposed approach does not suffer from this issue (up to the discussed level of confusion caused by approximations of the GW-functional).

The question then arises how the optimal coupling measure looks like if there are multiple (approximately) equivalent optimal solutions. Up to rasterization artifacts there is no preferred choice how to map the template cross onto the input: eight orientations (rotations, reflections) are equally possible. Each corresponding to one local extremum for an active contour approach that one would consider as valid solution. For the proposed approach such symmetries cause degeneration of the space of optimal couplings, making a whole facet of the feasible polytope extremal. Interior point methods then usually do not lead to integer solutions. Integer solutions exists and applying a simplex algorithm will produce one. Some may correspond to one of the eight possible assignments, some may be highly discontinuous (meaning that adjacent pixels are assigned to very different target pixels), but from the viewpoint of the functional they are all equivalent and the choice is arbitrary.



(a) (b) (c)

Figure 3.5: Breakdown of second Approximation step, enhancement by inhomogeneous appearance features: (a) top: template (white indicates mass), bottom: corresponding distance transform. (b) top: binary appearance coefficients g of query image (white indicates foreground), bottom: distance transform. (c) top: optimal ν according to (3.1.12a), bottom: optimal ν with distance transform as additional feature. In the input image, one arm of the cross has been shortened and the 'blob' on the right-hand side of the input has been designed to confuse the local ν estimation during the computation of $c_{\mathrm{LinGW},p}(\mu, G; \cdot, \cdot)$, thus causing faulty assignments. This demonstrates the limitations of the second approximation step. By including additional information like the distance transform as a feature this confusion can be resolved.



(a) (b) (c)

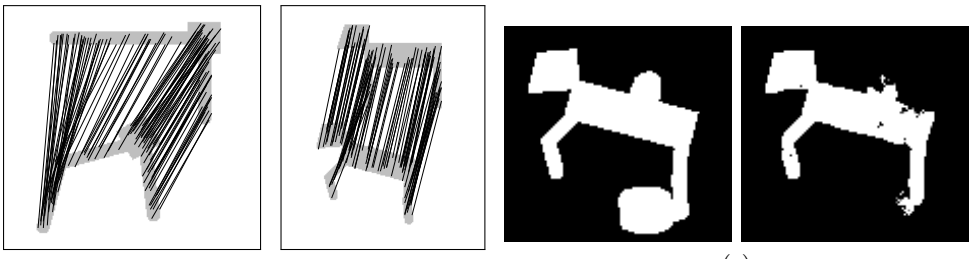
Figure 3.6: Inhomogeneous appearance term: (a) top: template with region specific features, black indicating background, gray → "body" and white → "eye"; bottom: input image with detected features. (b) optimal ν (gray shading bottom) and assignment (black lines, subsampling) of homogeneous appearance term, distinguishing only background ↔ fish (=body & eye). Due to the approximate mirror symmetry and the noise in the image features, front and back are confused (while still "correctly" reconstructing the edges according to the mix-up). (c) optimal ν and assignment with an appearance term that penalizes the matching $x \leftrightarrow y$ between different feature classes (see Sec. 3.1.3.2). The confusion between back and front is remedied.

3.1.5.5 Inhomogeneous Appearance Term

The inhomogeneous linear appearance term discussed in Sect. 3.1.3.2 can also be used to incorporate information about different characteristic appearances being associated with different regions of the sought-after shape. See for example Fig. 3.6: the shape itself is almost mirror-symmetric and in fact the noise was chosen such that the matching purely based on background ↔ "fish" confuses front and back of the schematic fish. Assume now from the underlying image data there is additional information available, like a dedicated detector for the eye. Then this can be exploited, leading to the desired effect. Also note that in both cases, corresponding to the assumed orientation of the fish, the appropriate shape is restored.

3.1.5.6 Geodesic Metric and Pose Invariance

A way to recognizing the same object in different poses is to equip shapes with the geodesic metric. In Figures 3.7a and 3.7b the geodesic metric is used to compute the optimal assignments between two pairs of objects in different poses via D_1. In Sect. 3.1.2.3 the problems were discussed that arise when one wants to port the concepts of intrinsic mm-space matching to image segmentation and faces the involved shape optimization task. The estimation of a static cost



(a) (b) (c)

Figure 3.7: Geodesic metric and pose invariance: (a) assignment between two different poses of an object, computed by $D_{\mathrm{LinGW},2}$. (b) assignment between two different poses of a schematic "horse" via $D_{\mathrm{LinGW},2}$. Both assignments correctly associate the corresponding parts of the objects. (c) left: modification of the lower horse from (b) by adding additional false positive foreground detections; right: optimal marginal ν for matching between the modified lower and the original upper horse via (3.1.12a). Excess detections are mostly removed at the correct locations.

function $c_{\mathrm{LinGW},p}(\mu, G; \cdot, \cdot)$ can in general not be performed in a straightforward fashion. Here we demonstrate the potential of the geodesic metric for a pose invariant shape prior functional in a restricted setup where the aforementioned difficulties can be avoided.

Consider binary appearance coefficients, as introduced earlier, and assume that all true foreground pixels are in fact also labelled as foreground by the appearance term. In addition some false positive detections are possible, i.e. regions in the test image that are wrongfully indicated to be foreground. The template shape and the apparent foreground region are then equipped with their respective geodesic metrics, efficiently computed by fast marching methods [Set95]. This requires that the false positive detections are rare enough to keep the geodesic metric of the underlying true foreground approximately unchanged. Figure 3.7c shows an image of binary appearance-coefficients with such superfluous false positive foreground labels and the computed optimal ν via (3.1.12a). Although the reconstruction is by no means perfect, the method still tends to neglect the false positive foreground-detections.

3.1.6 Conclusion

In the previous sections an isometry invariant shape prior functional, based on convex relaxation of the Gromov-Wasserstein distance was presented. Prior information on the shape was encoded in a template object. The representation of shapes through metric measure spaces is compatible with the representation as indicator functions and thus such a prior can in principle be combined with any convex segmentation functional. For the particular case of a linear appearance

model existence of binary optimal segmentations and deterministic bijective assignments between shape template and image foreground was established. It was indicated how a broad range of local image features can be incorporated into the approach.

We numerically illustrated the key property of *simultaneous convexity and isometry invariance* that yields globally optimal segmentations without need for initialization. The price for this is that the relaxation becomes too loose in scenarios with strong noise, resulting in spatially irregular segmentations and assignments. It was demonstrated how this can be alleviated to some extend by using more elaborate (psuedo-)local features.

In the next section the robustness of the prior to noise is increased by tightening the relaxation. A different approach to construct isometry invariant shape priors, based on mm-spaces, is presented in Chap. 4. We have already discussed the naïve computational complexity of the approach and how it can be reduced in various ways. A systematic method is presented in Chap. 5.

3.2 Assignment Regularization

In the last section we developed functional (3.1.12a) through convex relaxation of (3.1.12c). This reduced the computational complexity and enables application to image segmentation. But of course the simplifications do not come for free. In the relaxed functional every pair (x, y) locally estimates how the best marginal ν and coupling $\pi \in \Pi(\mu, \nu)$ would have to look like from their perspective, obtaining an assignment cost $c_{\mathrm{LinGW},p}(\mu, G; x, y)$ (3.1.5). At the time of the final assignment computation, based on $c_{\mathrm{LinGW},p}$, the only interaction between different $x \in X$ and $y \in Y$ is via the coupling constraints. This can be considered a waste of information, since all points have to find their correct assignment partners basically 'on their own'. Therefore, as we have seen, in difficult cases the local estimates can be inconsistent and the resulting global assignment can become quite spatially irregular.

To improve upon this, one must obviously increase the strength of interaction between points $x \in X$ and $y \in Y$ when choosing their assignment partners, such that information is propagated and flawed local estimates are combined to obtain a globally reasonable assignment.

In this section discuss two types of approaches to introduce additional interaction. The first type is based on *explicit* regularization. That is, during the final matching additional terms are added that tend to enforce local regularity of the assignment. Such an approach has already been briefly discussed in Sect. 2.4.4. The second type we dubbed *implicit*. The final matching will still be based on a linear assignment, but the cost function is computed by propagating information between local estimates.

It should be noted that the problem of assignment regularization is not specific to the approach in Sect. 3.1 but arises in many computer vision and

image processing applications.

3.2.1 Explicit and Implicit Regularization

3.2.1.1 Mean Assignment Regularization

Let us discuss application of the approach discussed in Sect. 2.4.4 to isometry invariant shape segmentation.

The originally proposed regularizer $J_p\big(\hat{V}(\cdot,\pi)\big)$ (2.4.54) is only invariant to translations. Assignments that correspond to rotations, flips or more general isometries will be penalized. Hence, for application to the isometry invariant shape matching proposed in Sect. 3.1 another choice is more appropriate, for example:

$$J_p\big(\hat{T}(\cdot,\pi)\big) = \sum_{(x,x')\in X^2} w(x,x') \max\Big\{\|\hat{T}(x,\pi) - \hat{T}(x',\pi)\| - d_X(x,x'), 0\Big\} \tag{3.2.1}$$

This regularizer is convex and blind w.r.t. isometries, in fact it is blind to any transformation T with $\|T(x) - T(x')\| \leq d_X(x,x')$, but when one chooses $w(x,x') > 0$ only for nearby pairs (x,x'), such that $d_X(x,x')$ is small, this can be considered a sufficiently close convex approximation.

The averaging in (2.4.51) implies that Y has an underlying vector space structure. This assumption is true in many applications but it limits the flexibility of the metric matching framework, which can be applied to more general metric spaces.

Remark 3.2.1. A major limitation of the approach (2.4.55) is the introduction of degenerate optimizers through the convex relaxation of the regularizer, i.e. the expression of T in terms of π. Consider a case where for a given $x \in X$ all $y \in Y$ which have a low cost $c(x,y)$, are unfavourable in terms of the vector field regularization. Since the regularizer 'sees' only the barycenter $\sum_{y\in Y} \pi(x,y)y$ of distributed mass, x can try to 'trick' the regularizer by distributing its mass over several $y \in Y$ with cheap costs such that the resulting mean assignment looks favourable (see Fig. 3.8 for illustration). This may not be relevant when the potential support Y is rather constrained. But in the image segmentation application, where every pixel is a potential transportation target, this can become an issue.

3.2.1.2 Probabilistic Assignment Regularization

The problem described in Remark 3.2.1 can be somewhat alleviated by not simply taking the mean $\hat{T}(x,\pi)$ in (2.4.51), but to interpret $\pi(x,\cdot)$ as a probabilistic assignment, i.e. $\pi(x,\cdot)$ is the probability distribution of a random

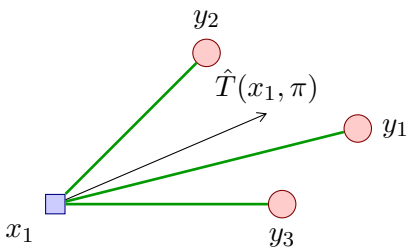


Figure 3.8: Mean assignment regularization: mass is transported from x_1 to y_1, y_2, y_3, leading to an 'apparent' mean mass destination $\hat{T}(x_1, \pi)$. Such apparent destinations may seem more favourable than deterministic assignments and cause non-deterministic degenerate optimal couplings π.

variable $T(x) \in Y$. given two adjacent $x, x' \in X$ (in the sense of the neighbourhood structure expressed in the weights w), one must then determine, which potential assignments for $T(x)$ and $T(x')$ co-occur and hence must be rated by a regularizer. This co-occurrence relation can be encoded in a coupling $\pi_{(x,x')} \in \Pi\big(\pi(x,\cdot), \pi(x',\cdot)\big)$ and for a given coupling the regularizer value can be given by a corresponding cost function $c_{(x,x')} : Y \times Y \to \mathbb{R}$ through

$$\sum_{(y,y') \in Y^2} c_{(x,x')}(y,y')\, \pi_{(x,x')}(y,y')\,. \tag{3.2.2}$$

For the isometry invariant matching application the obvious choice for $c_{(x,x')}$ is

$$c_{(x,x')}(y,y') = \Gamma(x,y,x',y')^p = \|d_X(x,x') - d_Y(y,y')\|^p\,, \text{ see (2.5.7)}. \tag{3.2.3}$$

Given two probabilistic assignments $\pi(x,\cdot)$, $\pi(x',\cdot)$ one can define their inflicted regularizer cost by finding the best co-occurrence coupling $\pi_{(x,x')} \in \Pi\big(\pi(x,\cdot), \pi(x',\cdot)\big)$ according to (3.2.2,3.2.3). This is in itself an optimal transport problem. Analogously to (2.4.55) one could define

$$E(\pi) = \sum_{(x,y) \in X \times Y} c(x,y)\, \pi(x,y) + \sum_{(x,x') \in X^2} w(x,x')\, J_{p,(x,x')}\big(\pi(x,\cdot), \pi(x',\cdot)\big) \tag{3.2.4a}$$

with

$$J_{p,(x,x')}(\pi_1,\pi_2) = \inf_{\pi_{(x,x')} \in \Pi(\pi_1,\pi_2)} \sum_{(y,y') \in Y^2} \Gamma(x,y,x',y')^p\, \pi_{(x,x')}(y,y')\,. \tag{3.2.4b}$$

Since the optimal transport cost is a convex function of the involved marginals, this functional is convex, too, and can therefore in principle be optimized 'easily'.

However, in Sect. 3.1.5 we were considering examples where the pure linear assignment part was already challenging from a computational point of view. The naïve complexity of (3.2.4) is considerably larger. The overall functional is no longer linear in π but contains nested optimization problems $J_{p,(x',x)}(\pi_1, \pi_2)$. Each of these terms is naïvely again an optimal transport problem on $Y \times Y$.

Additional insight into (3.2.4) can be gained by looking at it from the perspective of convex relaxation methods for graphical models. Let us for now drop the Y-marginal mass constraint on $\pi \in \mathrm{SegMeas}(Y, \mu)$. That is, we consider optimizing (3.2.4) over

$$\{\pi \in \mathrm{Meas}(X \times Y) : \ \mathrm{pr}_{X\sharp}\pi = \mu\}\,. \tag{3.2.5}$$

That is, multiple $x \in X$ may be assigned to the same $y \in Y$. This problem corresponds to the local polytope relaxation [WJ08, Sect. 4.1.1] of the following graphical model:

$$\min_{T:X\to Y} E'(T) \tag{3.2.6a}$$

with

$$E'(T) = \sum_x c\big(x, T(x)\big) + \sum_{(x,x')} w(x,x')\,\Gamma\big(x, T(x), x', T(x')\big) \tag{3.2.6b}$$

The local polytope relaxation is tight only for tree graphs, i.e. when the graph implied by w has no cycles [WJ08, Prop. 4.1]. The graphs corresponding to templates used in Sect. 3.1.5 however, have many cycles. In such a case, there may not exist binary optimizers π^* to (3.2.4) within the set (3.2.5) (cf. [WJ08, Example 4.1]). By reintroducing the Y-marginal constraint one can even construct cases where no binary optimizers exist for tree graphs. So even though (3.2.4) circumvents the 'barycenter-cheating' discussed in Sect. 3.2.1.1, regularization still invalidates the guarantee for existence of binary optimizers.

3.2.1.3 Implicit Regularization

The approaches discussed in the previous sections aimed at counter-balancing the relaxation steps (3.1.12c) → (3.1.12b) → (3.1.12a) by introducing an additional geometric regularizer in the final matching step.

Both ideas had two common problems: although being convex, the additional regularization term made optimization of the final assignment more difficult. And the guarantee for the existence of deterministic optimal π^*, given by Proposition 3.1.6 was lost.

In this section we will discuss an alternative approach: instead of adding explicit regularizers to the final assignment functional, we attempt to improve the estimation of the static cost function (Sect. 3.1.2.2). The final matching remains linear. The method applies to general QAP instances and will be

formulated in this context. Application as isometry invariant shape prior for segmentation in the setting of Sect. 3.1.3.1 is done by introducing dummy nodes to X, as done during the proof of Proposition 3.1.6.

The basic idea is the following: the estimated locally optimal coupling π^* for computation of $c_{\mathrm{LinGW},p}(\mu, G; x_1, y_1)$ for a given pair (x_1, y_1) might assign some pair (x_2, y_2), because it seems favourable from the perspective of (x_1, y_1), although the value of $c_{\mathrm{LinGW},p}(\mu, G; x_2, y_2)$ is in fact very high, indicating that the assignment is not very likely. We now try to create an improved re-estimation $c_{\mathrm{LinGW},p,2}(\mu, G; x_1, y_1)$, taking into account values of the first estimates $c_{\mathrm{LinGW},p}(\mu, G; \cdot, \cdot)$. We will refer to this updating as *cost function iteration*.

In section 3.2.2 we will discuss cost function iteration and its relation to the Gilmore-Lawler bound (Sect. 2.3.2.2). In particular we will show that repeating the iteration procedure converges. In Sect. 3.2.3.1 a reformulation of the QAP in terms of a diffusion process will be given. This may seem unnecessarily complicated at the time. However, in Sect. 3.2.3.2 we show how the lower bound given by the cost function iteration can be seen as a relaxation to this particular formulation. A brief experimental evaluation of this relaxation is given in Sect. 3.2.4.

3.2.2 Cost Function Iteration

Remark 3.2.2 (Conventions)**.** We consider the QAP (2.3.2) for set cardinality $|X| = |Y| = N$ and a general cost coefficient array $d \in \mathbb{R}^{N^4}$. For a more compact notation, we will not address entries of d via elements in X and Y, but by indices $1 \leq i, j, k, l \leq N$, as done in Sect. 2.3.

In the course of this section we will handle *cost functions* in $\mathbb{R}^{N^2}$, *mass distributions* in $\mathbb{R}^{N^2}$ and vectors in $\mathbb{R}^{N^4}$ which we will at times interpret as *matrices* $\mathbb{R}^{N^2} \to \mathbb{R}^{N^2}$. To keep notation efficient, we will sometimes no longer explicitly give all involved indices and sums but resort to compact matrix notation. That is, for a cost function $c \in \mathbb{R}^{N^2}$, a mass distribution $\pi \in \mathbb{R}^{N^2}$, and a matrix $A \in \mathbb{R}^{N^4}$ we might write:

$$
\begin{aligned}
(c\,A)_{kl} &= \sum_{ij} c_{ij}\, A_{ijkl} \\
(A\,x)_{ij} &= \sum_{kl} A_{ijkl}\, \pi_{kl} \\
c\,\pi &= \sum_{ij} c_{ij}\, \pi_{ij} \\
c\,A\,\pi &= \sum_{ijkl} c_{ij}\, A_{ijkl}\, \pi_{kl}
\end{aligned}
$$

It will be convention in this article that cost functions are regarded as N^2-dimensional row vectors and mass distributions as column vectors.

The matrix multiplication convention will not always be applicable and explicit index / sum notation will then be chosen.

In Sect. 2.3.2.2 we have already introduced the Gilmore-Lawler bound to the QAP (2.3.5):

$$\mathrm{GLB} = \min_{\pi \in \mathrm{Perm}_N} \sum_{i,j=1}^{N} (c_0)_{ij}\, \pi_{ij} \tag{3.2.7}$$

with

$$(c_0)_{ij} = \min_{\pi \in \mathrm{Perm}_N \,:\, \pi_{ij}=1} \sum_{k,l=1}^{N} d_{ijkl}\, \pi_{kl} \tag{3.2.8}$$

Within this section we will refer to objects like c_0 as *cost functions*. The minimal linear assignment cost associated with a cost function c will be denoted by

$$D(c) = \min_{\pi \in \mathrm{Perm}_N} \sum_{i,j=1}^{N} c_{ij}\, \pi_{ij}\,, \tag{3.2.9}$$

i.e. the GLB is given by $D(c_0)$. Relative to (2.3.5b) we have added the subscript 0 to c_0 since we now try to estimate an updated c_1, based on c_0. Consider the following definition:

Definition 3.2.3 (Minimizing Cost Function Iteration)**.** For a given $c \in \mathbb{R}^{N^2}$ we define the minimizing cost function iteration $\mathcal{R} : \mathbb{R}^{N^2} \to \mathbb{R}^{N^2}$ by

$$\mathcal{R}(c)_{ij} = \min_{\pi \in \mathrm{Perm}_N \,:\, \pi_{ij}=1} \sum_{k,l=1}^{N} (\sigma\, d_{ijkl} + \tau\, c_{kl})\, \pi_{kl} \tag{3.2.10}$$

where $\sigma \in]0, 1]$ is a free parameter and

$$\tau = (1 - \sigma)/N\,. \tag{3.2.11}$$

Let $c_1 = \mathcal{R}(c_0)$, i.e.

$$(c_1)_{ij} = \min_{\pi \in \mathrm{Perm}_N \,:\, \pi_{ij}=1} \sum_{k,l=1}^{N} \left(\sigma\, d_{ijkl} + \tau\, (c_0)_{kl}\right) \pi_{kl}\,. \tag{3.2.12}$$

This is similar to the GLB computation, but instead of only using the coefficients d, we mix in some contribution by c_0, controlled by the parameter σ. So if a constellation is considered very unlikely by c_0, this will affect the new local estimates. The relation between the weights σ and τ has been carefully chosen to ensure that $D(c_1)$ will still be a lower bound to the QAP but potentially tighter as $D(c_0)$. In fact it can be shown:

Proposition 3.2.4.

$$D(c_0) \leq D(c_1) \leq \min_{\pi \in \mathrm{Perm}_N} \sum_{i,j,k,l=1}^{N} d_{ijkl}\, \pi_{ij}\, \pi_{kl} \tag{3.2.13}$$

Proof. Consider the first inequality:

$$\begin{aligned} D(c_1) &= \min_{\pi \in \mathrm{Perm}_N} \sum_{i,j=1}^{N} \left(\min_{\pi' \in \mathrm{Perm}_N \,:\, \pi'_{ij}=1} \sum_{kl} (\sigma\, d_{ijkl} + \tau\, (c_0)_{kl}) \pi'_{kl} \right) \pi_{ij} \\ &\geq \min_{\pi \in \mathrm{Perm}_N} \sum_{i,j=1}^{N} \left(\left(\min_{\pi' \in \mathrm{Perm}_N \,:\, \pi'_{ij}=1} \sum_{kl} \sigma\, d_{ijkl}\, \pi'_{kl} \right) \right. \\ &\qquad \left. + \left(\min_{\pi' \in \mathrm{Perm}_N \,:\, \pi'_{ij}=1} \tau\, (c_0)_{kl} \pi'_{kl} \right) \right) \pi_{ij} \\ &\geq \min_{\pi \in \mathrm{Perm}_N} \sum_{i,j=1}^{N} \sigma\, (c_0)_{ij}\, \pi_{ij} + \min_{\pi \in \mathrm{Perm}_N} \sum_{i,j=1}^{N} \tau\, D(c_0)|_{\mathrm{fix}:ij}\, \pi_{ij} \end{aligned}$$

where by $D(c_0)|_{\mathrm{fix}:ij}$ we denote the optimal linear assignment cost (3.2.9) under the restriction $\pi_{ij} = 1$. By picking for π in the second term an optimizer of $D(c_0)$, one finds

$$= (\sigma + \tau \cdot N) D(c_0) = D(c_0)$$

The second inequality is proven by writing down the full nested expression for $D(c_1)$, obtaining an upper bound by fixing all involved π-optimizations to be the same variable and summing up the terms. □

This re-estimation can in principle be repeated arbitrarily often, defining a sequence of cost functions $c_n = \mathcal{R}^n(c_0)$ (where $\mathcal{R}^n$ denotes repeated application of $\mathcal{R}$). Analogous to the proof above it can be established that $D(c_0) \leq D(c_n)$ and $D(c_n)$ will always be a lower bound to the QAP objective value. We will not detail this but instead investigate where the sequence c_n is heading. One can show:

Proposition 3.2.5 (Fixed-point of minimizing cost function iteration)**.** *The minimizing cost function iteration converges to a unique fixed-point which only depends on the cost array d and $\sigma \in]0, 1]$.*

The proof follows from the proof of Proposition 3.2.9, which we give a little further down.

Remark 3.2.6. For $\sigma = 1$ the map $\mathcal{R}$ will always return c_0, regardless of its argument. Thus for $\sigma = 1$, $c_0 = \mathcal{R}(c_0)$ is the unique fixed-point.

The following of this section is aimed at interpreting the minimizing cost function iteration, its fixed-point and the relation between the fixed-points of different parameters σ. For these investigations we need to establish some auxiliary concepts.

Definition 3.2.7 (Assignment field)**.** A vector $a \in \{0,1\}^{N^4}$ is called an *assignment field* if

(i) $a_{ij} \in \{0,1\}^{N^2}$ is in Perm_N for all $1 \leq i,j \leq N$,

(ii) $a_{ijij} = 1$ for all $1 \leq i,j \leq N$.

That is, a is a set of permutation matrices, one for every assignment $i \leftrightarrow j$, and for each assignment pair (i,j) the corresponding permutation matrix assigns $i \leftrightarrow j$. We denote the set of assignment fields by $\mathcal{A}$.

An example for an assignment field is given by the set of minimizers for computing the entries of the cost functions c_n: for every entry $(c_n)_{ij}$ one seeks an optimal permutation $\pi^* \in \mathrm{Perm}_N$ with $\pi_{ij} = 1$. Conversely, every assignment field $a \in \mathcal{A}$ induces a cost function iteration, where one simply replaces minimization in (3.2.10) by the fixed permutation a_{ij}. We define:

Definition 3.2.8 (Pre-specified cost function iteration)**.** For a given assignment field $a \in \mathcal{A}$ let analogous to Definition 3.2.3

$$\mathcal{R}_a(c)_{ij} = \sum_{k,l=1}^{N} (\sigma\, d_{ijkl} + \tau\, c_{kl})\, a_{ijkl}\,. \tag{3.2.14}$$

Proposition 3.2.9 (Fixed-point of pre-specified cost function iteration)**.** *The pre-specified cost function iteration converges to a unique fixed-point which only depends on the cost array d, $\sigma \in]0,1]$ and the assignment field a.*

Proof. We will show that $\mathcal{R}$ and $\mathcal{R}_a$ for any $a \in \mathcal{A}$ are contractions. Existence and uniqueness of the fixed-points follows then from the Banach fixed-point theorem (e.g. [BC11, Thm. 1.48]). Equip the space $\mathbb{R}^{N^2}$ with the max-metric $\delta(f,g) = \max_{1\leq i,j\leq N} |f_{ij} - g_{ij}|$, turning it into a complete metric space. For two $f, g \in \mathbb{R}^{N^2}$ consider now

$$\begin{aligned}
\mathcal{R}_a(f)_{ij} &= \sum_{k,l=1}^{N} (\sigma\, d_{ijkl} + \tau\, f_{kl})\, a_{ijkl} \\
&\leq \sum_{k,l=1}^{N} \big(\sigma\, d_{ijkl} + \tau\, (g_{kl} + \delta(f,g))\big)\, a_{ijkl} \\
&= \mathcal{R}_a(g)_{ij} + \tau\, N\, \delta(f,g)
\end{aligned}$$

Analogous

$$\mathcal{R}_a(g)_{ij} \leq \mathcal{R}_a(f)_{ij} + \tau\, N\, \delta(f,g)$$

and therefore

$$|\mathcal{R}_a(f)_{ij} - \mathcal{R}_a(g)_{ij}| \leq (1-\sigma)\, \delta(f,g)$$

and eventually

$$\delta\big(\mathcal{R}_a(f), \mathcal{R}_a(g)\big) \leq (1-\sigma)\, \delta(f,g)\,.$$

Since by assumption $\sigma > 0$ this proves Proposition 3.2.9. The reasoning above still holds when minimizing over $\mathcal{A}$, so Proposition 3.2.5 follows. □

It is easy to see that

$$\mathcal{R}(c)_{ij} = \min_{a\in\mathcal{A}} \mathcal{R}_a(c)_{ij} \tag{3.2.15}$$

and since in an assignment field all permutation matrices can be chosen independently (i.e. minimized independently), there is for every c some $a \in \mathcal{A}$ such that $\mathcal{R}(c) = \mathcal{R}_a(c)$. This is particularly true for the fixed-point $\hat{c}$ of $\mathcal{R}$, i.e. there is some $\hat{a} \in \mathcal{A}$ such that

$$\hat{c} = \mathcal{R}(\hat{c}) = \mathcal{R}_{\hat{a}}(\hat{c})\,.$$

So $\hat{c}$ is also the fixed-point of $\mathcal{R}_{\hat{a}}$ and together with (3.2.15) one immediately finds:

Proposition 3.2.10. *The linear assignment cost of the minimizing cost function iteration is equal to the minimal linear assignment cost of the pre-specified cost function iteration, minimized over $a \in \mathcal{A}$:*

$$\min_{a\in\mathcal{A}} D\big(\mathcal{R}_a(c)\big) = D\big(\mathcal{R}(c)\big) \tag{3.2.16}$$

In the next section we will investigate the role of the parameter σ. Yet, first a final remark on cost function iterations regarding computational efficiency.

Remark 3.2.11 ("Diagonal" and "non-diagonal" iteration)**.** Given the intuitive motivation for the minimizing cost function iteration in terms of using $(c_0)_{kl}$ to help judging whether the assignment $k \leftrightarrow l$ is reasonable when computing $(c_1)_{ij}$ in addition to d_{ijkl}, it seems strange, that in the computation of $(c_1)_{ij}$ the value $(c_0)_{ij}$ is also included by a constant factor, since $\pi_{ij} = 1$ is fixed when finding the minimum for $\mathcal{R}(c_0)_{ij}$. It would seem more natural, to define the minimizing cost function iteration in the following fashion:

$$\tilde{\mathcal{R}}(c)_{ij} = \min_{\pi\in\mathrm{Perm}_N\colon \pi_{ij}=1} \left(\sum_{k,l=1}^{N} \sigma\, d_{ijkl}\, \pi_{kl} + \sum_{k,l=1\colon k\neq i, l\neq j}^{N} \tilde{\tau}\, c_{kl}\, \pi_{kl} \right) \tag{3.2.17}$$

with $\tilde{\tau} = (1-\sigma)/(N-1)$. Here, the "old" information $(c_0)_{ij}$ is not dragged along and enters the computation of the "new" information $(c_1)_{ij}$ at the same location. Only old information from other locations is used. We refer to this variant as "non-diagonal", since there is no transmission from $(c_0)_{ij}$ to $(c_1)_{ij}$.

For this modified iteration both the minimizing and the pre-specified variant can be defined as above and one can establish similar results: convergence to unique fixed-points and the implication of a QAP bound which is at least as tight as GLB.

Let c_∞ be the fixed-point of $\tilde{\mathcal{R}}$ for some given assignment field a and $\sigma \in]0,1]$:

$$(c_\infty)_{ij} = \tilde{\mathcal{R}}(c_\infty)_{ij} = \min_{\pi \in \mathrm{Perm}_N \,:\, \pi_{ij}=1} \left(\sum_{kl} \sigma\, d_{ijkl}\, \pi_{kl} + \sum_{k,l=1\,:\, k\neq i, l\neq j}^{N} \tilde{\tau}\, (c_\infty)_{kl}\, \pi_{kl} \right)$$

$$\Rightarrow \quad (1+\tilde{\tau})(c_\infty)_{ij} = \min_{\pi \in \mathrm{Perm}_N \,:\, \pi_{ij}=1} \left(\sum_{k,l=1}^{N} \sigma\, d_{ijkl}\, \pi_{kl} + \sum_{k,l=1}^{N} \tilde{\tau}\, (c_\infty)_{kl}\, \pi_{kl} \right)$$

$$\Rightarrow \quad (c_\infty)_{ij} = \min_{\pi \in \mathrm{Perm}_N \,:\, \pi_{ij}=1} \sum_{k,l=1}^{N} \left(\frac{\sigma}{1+\tilde{\tau}}\, d_{ijkl} + \frac{\tilde{\tau}}{1+\tilde{\tau}}\, (c_\infty)_{kl} \right) \pi_{kl}$$

$$\Rightarrow \quad (c_\infty)_{ij} = \min_{\pi \in \mathrm{Perm}_N \,:\, x_{ij}=1} \sum_{k,l=1}^{N} \left(\sigma'\, d_{ijkl} + \tau'\, (c_\infty)_{kl} \right) \pi_{kl} = (\mathcal{R}(c)|_{\sigma'})_{ij}$$

where

$$\sigma' = \sigma \frac{N-1}{N-\sigma}\,, \quad \tau' = \frac{1-\sigma'}{N}$$

and $\mathcal{R}(c)|_{\sigma'}$ denotes minimizing iteration with parameter σ'.

That is, the fixed-point c_∞ of the non-diagonal variant and parameter σ is equal to the fixed-point of the regular variant with parameter σ'. One should note that $\sigma' \in]0,1] \Leftrightarrow \sigma \in]0,1]$ and in particular $\sigma' \leq \sigma$. Thus, with the non-diagonal iteration we are finding the fixed-point of the diagonal iteration with a smaller parameter σ'. From the proofs on the existence of the fixed-points we can thus see, that although $\tilde{\mathcal{R}}|_\sigma$ has the same fixed-point as $\mathcal{R}|_{\sigma'}$, the bound on the Lipschitz constant (i.e. the minimal contraction rate) is lower for $\tilde{\mathcal{R}}|_\sigma$ (as one would expect, since the old value of $(c_n)_{ij}$ is not simply dragged along with a constant factor).

The same is naturally true for comparison of the pre-specified iteration variants.

In the following we will continue to use the "diagonal" iteration variants, because the notation and proofs are simpler. However it should be kept in mind that for actual implementation one should choose the "non-diagonal" variant due to potentially faster convergence towards the fixed-points (which is what we will be looking for in the end).

3.2.3 Diffusion Relaxation of QAP

3.2.3.1 Diffusion Formulation of QAP

So far we have introduced the minimizing cost function iteration as a way to obtain potentially tighter bounds to the QAP as the GLB does. It was shown that the sequence converges. We now want to learn more about the fixed-point and the role of the parameter $\sigma \in]0,1]$. For this we give a reformulation of the QAP in terms of a discrete diffusion process.

Definition 3.2.12 ((Left) stochastic matrix)**.** A quadratic matrix with non-negative coefficients whose columns sum to 1 is called (left) stochastic. A stochastic matrix $A \in \mathbb{R}^{n\times n}$ is mass preserving:

$$\sum_{i,j=1}^{n} A_{ij}\, v_j \sum_{i=1}^{n} v_i \quad \text{for all} \quad v \in \mathbb{R}^n$$

Repeated application of a stochastic matrix $A \in \mathbb{R}^{n\times n}$ to a non-negative vector p_0 can be interpreted as a discrete diffusion process on the node set $\{1,\ldots,n\}$ with initial particle distribution given by p_0. We will now consider a particular type of diffusion, namely on the set of assignments $\{1,\ldots,N\}^2$ with transition probabilities given by some assignment field a.

Definition 3.2.13 (Assignment diffusion)**.** Let a be an assignment field and from the coefficients of a construct a $N^2 \times N^2$ matrix A via $A_{ijkl} = a_{klij}/N$ (note the swap in the order of indices), where the first two indices reference the rows and the second pair enumerates the columns (and we can think of both rows and columns to be again composed of rows and columns). We refer to A as the *assignment diffusion matrix* corresponding to assignment field a and by construction A is stochastic.

For a given non-negative vector $p_0 \in \mathbb{R}^{N^2}$ (i.e. a vector that assigns mass to pairs $(i,j) \in \{1,\ldots,N\}^2$) the sequence $\{A^n\, p\}_{n=0}^{\infty}$ describes a discrete diffusion process on $\{1,\ldots,N\}^2$. At each time step the mass located at (i,j) is evenly distributed onto the assignment pairs given by $a_{ij} \in \mathrm{Perm}_N$, which implies that a fraction of $1/N$ of mass is kept at (i,j). We refer to this process as *assignment diffusion*.

More general, for a given non-negative vector p_0 and a probability $\sigma \in [0,1]$ we consider the *assignment diffusion* process with *reset probability* σ, described by

$$p_n = \sigma\, p_0 + (1-\sigma)\, A\, p_{n-1} \;\text{ for }\; n > 0\,. \tag{3.2.18}$$

That is, at each time step a fixed fraction σ of particles is reset to the initial distribution while the rest $(1-\sigma)$ undergoes regular assignment diffusion.

It is not a coincidence that we picked for the reset probability the same symbol σ that has been used in the definitions of the cost function iterations and we will later on show a tight connection between the two concepts.

Proposition 3.2.14 (Properties of assignment diffusion matrices)**.** *Assignment diffusion matrices have the following properties:*

(i) Mass preservation: $\sum_{i,j,k,l=1}^{N} A_{ijkl}\, p_{kl} = \sum_{i,j=1}^{N} p_{ij}$ *for all* $p \in \mathbb{R}^{N\times N}$.

(ii) Any eigenvalue λ *satisfies* $|\lambda| \leq 1$.

(iii) The only eigenvalues λ *with* $|\lambda| = 1$ *are* $\lambda = 1$. *For* $\lambda = 1$ *algebraic and geometric multiplicity coincide.*

This proposition is closely related to the Perron-Frobenius Theorem (see for example [GR01, Thm. 8.1.1]), but assignment diffusion matrices can in general be reducible.

Proof. Property (i) follows from A begin stochastic (Def. 3.2.12). Since A is non-negative we also find $\|A\,p\|_1 \leq \|p\|_1$. From the equivalence of the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ in finite dimensional vector spaces then follows (ii). Conversely, were there no eigenvalue λ with $|\lambda| = 1$, then any sequence $\{A^n\,p\}_{n=0}^{\infty}$ would converge to the 0-vector, violating the mass preservation property for non-negative p. Hence, at least one such eigenvalue must exist. We will now show that it must be $\lambda = 1$. Since $A_{ijij} = 1/N$ we can write

$$A = \frac{1}{N}\,\mathrm{id} + \frac{N-1}{N} B$$

with a mass-preserving non-negative matrix B. Thus, by the reasoning above, B may only have eigenvalues λ_B with $|\lambda_B| \leq 1$. By this decomposition of A we find for any eigenvector v of A with eigenvalue λ:

$$\lambda\, v = A\, v = \left(\frac{1}{N} + \frac{N-1}{N} B\right) v$$

That is, it is an eigenvector of B with eigenvalue λ_B with the relation

$$\lambda = \frac{1}{N} + \left(1 - \frac{1}{N}\right) \lambda_B$$

The only λ with $|\lambda| = 1$ that can satisfy this equation for some $|\lambda_B| \leq 1$ is $\lambda = 1$. Were the algebraic multiplicity of 1 not equal to its geometric multiplicity one could construct sequences that violate the mass-preservation property by using the Jordan normal form. □

One finds as a corollary:

Corollary 3.2.15. *For some* $a \in \mathcal{A}$ *let* A *be the corresponding assignment diffusion matrix. The sequence* $\{A^n\}_{n=0}^{\infty}$ *converges.*

We can then establish the following facts on assignment diffusion:

Proposition 3.2.16 (Properties of Assignment Diffusion)**.** *The assignment diffusion process with reset probability* $\sigma \in [0,1]$ *and initial/reset distribution* p_0 *satisfies the following properties:*

- *There exists a non-oscillating limit distribution towards which the sequence of successive distributions* $\{p_n\}_{n=0}^{\infty}$ *is converging.*
- *The limit distribution is a continuous function of* σ *for* $\sigma \in [0,1]$*.*

Proof of Proposition 3.2.16. For some σ denote by $p_\infty(\sigma)$ the corresponding limit distribution. We start by considering the case $\sigma \in]0,1]$. For the n-th iterated distribution one obtains

$$p_n = \sum_{k=0}^{n-1} \sigma\,(1-\sigma)^k\, A^k\, p_0 + (1-\sigma)^n\, A^n\, p_0\,.$$

Since A is stochastic, the total mass of $A^n\, p_0$ is that of p_0 and thus finite. Also $A^n\, p_0$ is non-negative, if p_0 is. Therefore, the second term will vanish in the limit $n \to \infty$. The first term is a geometric matrix series and yields

$$p_\infty(\sigma > 0) = \sigma\,(\mathrm{id} - (1-\sigma)\, A)^{-1}\, p_0$$

in the limit. Thus the sequence of distributions converges towards a well defined unique non-oscillating limit.

In the case $\sigma = 0$ the geometric sequence vanishes but the second term does not. By virtue of Corollary 3.2.15 we have that this term converges to a non-oscillating limit:

$$p_\infty(\sigma = 0) = \lim_{n\to\infty} A^n p_0$$

This implies that $p_\infty(0)$ is an eigenvector of the limit matrix $\lim_{n\to\infty} A^n$ with eigenvalue 1.

The limit distribution is obviously a continuous function of σ for $\sigma \in]0,1]$. It remains to be established, that it is continuous at $\sigma = 0$ which is equivalent to showing that the sequence $p_\infty(\sigma_m)$ for any positive sequence of σ_m-values that converges towards 0 as $m \to \infty$ converges towards the limit distribution $p_\infty(0)$.

Let $\{\sigma_m\}_{m=0}^{\infty}$ be some positive sequence $\sigma_m \in]0,1]$ that converges to 0. Let then s_m be the largest integer such that

$$\sigma_m \leq 1 - \left(1 - \frac{1}{s_m}\right)^{1/s_m}\,.$$

Since $\lim_{m\to\infty} \sigma_m = 0$ one finds $\lim_{m\to\infty} s_m = \infty$. By construction one has

$$(1-\sigma_m)^{s_m} \geq 1 - 1/s_m \to 1 \quad \text{as} \quad m \to \infty\,.$$

We write:

$$\begin{aligned} p_\infty(\sigma_m) &= \sum_{k=0}^{\infty} \sigma_m \, (1-\sigma_m)^k \, A^k \, p_0 \\ &= \underbrace{\sum_{k=0}^{s_m-1} \sigma_m \, (1-\sigma_m)^k \, A^k \, p_0}_{\mathbf{I}} + \\ &\quad \underbrace{\left(\sum_{k=0}^{\infty} \sigma_m \, (1-\sigma_m)^k \, A^k \right)}_{\mathbf{II}} \underbrace{(1-\sigma_m)^{s_m}}_{\mathbf{III}} \underbrace{A^{s_m} \, p_0}_{\mathbf{IV}} \end{aligned}$$

The norm $\|A^k \, p_0\|_1$ of all vectors in the first term can be bounded by $\|p_0\|_1$. The sum of all coefficients $\sum_{k=0}^{s_m-1} \sigma_m \, (1-\sigma_m)^k$ can be shown to tend to 0 for $m \to \infty$, thus the first term vanishes in the limit. The third term converges to 1 by construction. The fourth term converges to $p_\infty(0)$. For finite m we write $A^{s_m} \, p_0 = p_\infty(0) + \Delta p$ where $\|\Delta p\|_1 \to 0$ as $m \to \infty$. Since A is stochastic we find $\|A^k \, \Delta p\|_1 \leq \|\Delta p\|_1$ for all k. $p_\infty(0)$ will "pass through" all the A^k in term two unaltered, the coefficients $\sigma_m \, (1-\sigma_m)^k$ sum up to 1. The norm of the total contribution of Δp with any term in two will be bounded by $\|\Delta p\|_1$ and thus the product of Δp and term two will vanish in the limit and we find

$$\lim_{m \to \infty} p_\infty(\sigma_m) = p_\infty(0) \, . \qquad \square$$

During a diffusion step $p \to Y \, p$, we can imagine that an amount of mass $Y_{klij} \, p_{ij} = N^{-1} \, y_{ijkl} \, p_{ij}$ travels along an edge from (i,j) to (k,l). If we associate with each of these edges a cost d_{ijkl} per an amount N^{-1} of mass, we can see, that this diffusion step inflicts a cost of $\sum_{ijkl} d_{ijkl} \, y_{ijkl} \, p_{ij}$.

We now give an alternative definition of the QAP in terms of assignment diffusions and the diffusion cost inflicted during each iteration in the equilibrium state.

Definition 3.2.17 (Diffusion QAP)**.** For an array of costs $d \in \mathbb{R}^{N^4}$ the diffusion QAP is given by:

$$\min_{\pi \in \mathrm{Perm}_N, a \in \mathcal{A}} \sum_{ijkl} d_{ijkl} \, a_{ijkl} \, (p_\infty)_{ij} \tag{3.2.19}$$

subject to

$$\pi_{ij} = 1 \Rightarrow a_{ij} = \pi \tag{3.2.20}$$

where p_∞ is the unique limit distribution of the assignment diffusion process defined by assignment field y and initial distribution $p_0 = x$. This holds for any reset probability $\sigma \in [0,1]$.

Proposition 3.2.18 (Equivalence of QAP and Diffusion QAP)**.** *The diffusion QAP with cost coefficients d has the same optimal value as the corresponding QAP. If π is an optimal assignment in the diffusion QAP, it is also an optimal assignment in the QAP.*

Proof. Consider for the initial distribution $p_0 = \pi$ one iteration of the assignment diffusion for some feasible pair π, a:

$$\begin{aligned} p_1 &= \sigma\,\pi + (1-\sigma)\,A\,\pi \\ (p_1)_{ij} &= \sigma\,\pi_{ij} + (1-\sigma)\,N^{-1} \sum_{k,l=1}^{N} a_{klij}\,\pi_{kl} \\ &= \sigma\,\pi_{ij} + (1-\sigma)\,N^{-1} \sum_{k,l=1\colon\,\pi_{kl}=1}^{N} a_{klij} \end{aligned}$$

(and by virtue of constraint (3.2.20) for feasible pairs π, a)

$$\begin{aligned} &= \sigma\,\pi_{ij} + (1-\sigma)\,N^{-1} \sum_{k,l=1\colon\,\pi_{kl}=1}^{N} \pi_{ij} \\ &= \pi_{ij} \left(\sigma + (1-\sigma)\,N^{-1} \sum_{k,l=1\colon\,\pi_{kl}=1}^{N} 1 \right) \\ &= \pi_{ij} \end{aligned}$$

Thus we have $p_1 = \pi$ and thus any $p_n = \pi$ and the limit $p_\infty = \pi$, the diffusion is "degenerate" and already set up in its stationary state. The corresponding diffusion QAP score then is

$$\sum_{i,j,k,l=1}^{N} d_{ijkl}\,a_{ijkl}\,(p_\infty)_{ij} = \sum_{i,j,k,l=1\colon\,\pi_{ij}=1}^{N} d_{ijkl}\,\pi_{kl}\,\pi_{ij} = \sum_{i,j,k,l=1}^{N} d_{ijkl}\,\pi_{ij}\,\pi_{kl}$$

which is nothing but the usual QAP score for assignment π. Therefore both problems are equivalent.

□

The formulation of the QAP in terms of assignment diffusion seems impractically complicated: the introduced diffusion framework seems rather useless, since the initially chosen distribution equals the equilibrium distribution and nothing really is changed by diffusion. The motivation behind this is to obtain a new family of lower bounds by dropping the constraint (3.2.20) from the diffusion QAP.

3.2.3.2 Diffusion Relaxation

We will now introduce the diffusion relaxation of the QAP and relate it to the previously presented concept of cost function iterations.

Definition 3.2.19 (Diffusion relaxation of QAP)**.** For an array of costs $d \in \mathbb{R}^{N^4}$ the diffusion relaxation of the QAP is given by:

$$\min_{\pi \in \mathrm{Perm}_N, a \in \mathcal{A}} \sum_{i,j,k,l=1}^{N} d_{ijkl}\, a_{ijkl}\, (p_\infty)_{ij} \tag{3.2.21}$$

where p_∞ is the unique limit distribution of the assignment diffusion process defined by assignment field a and initial distribution $p_0 = \pi$.

The diffusion relaxation of the QAP obviously gives a lower bound, as the problem is identical to the diffusion formulation of the QAP without constraint (3.2.20).

Proposition 3.2.20 (Equivalence of assignment diffusion and cost function iteration)**.** *Let $\pi \in \mathrm{Perm}_N$ be a fixed assignment and $a \in \mathcal{A}$ a fixed assignment field. Also let c_0 be some arbitrary (bounded) initial cost function. Let some $\sigma \in]0, 1]$ be both the free parameter of the pre-specified cost function iteration as well as the reset probability in the assignment diffusion process. The initial (and reset) distribution of the assignment diffusion process shall be given by π. Then in the limit of $n \to \infty$ pre-specified cost function / diffusion iterations the linear cost associated with the iterated cost function and the assignment π equals the diffusion cost associated with iterations of the iterated mass distributions. More precisely:*

$$\lim_{n\to\infty} \sum_{i,j=1}^{N} \pi_{ij}\, (c_n)_{ij} = \lim_{n\to\infty} \sum_{i,j,k,l=1}^{N} d_{ijkl}\, a_{ijkl}\, (p_n)_{ij} \tag{3.2.22}$$

where c_n is the n-th pre specified cost function iteration of c_0 with assignment field a and parameter σ and p_n is the n-th iteration of the assignment diffusion with assignment field a, reset probability σ and initial (and reset) distribution π.

Proof. Given some assignment $\pi \in \mathrm{Perm}_N$, assignment field $a \in \mathcal{A}$, some parameter $\sigma \in]0, 1]$ and an initial cost function c_0 (for now not necessarily the one used in the GLB) we define recursively for $n > 0$

$$c_n = \mathcal{R}_a(c_{n-1})\,.$$

Explicitly

$$\begin{aligned}(c_n)_{ij} &= \sum_{k,l=1}^{N} (\sigma\, d_{ijkl} + \tau\, (c_{n-1})_{kl})\, a_{ijkl} \\ &= \sum_{k,l=1}^{N} (\sigma\, d_{ijkl}\, a_{ijkl} + (1-\sigma)\, (c_{n-1})_{kl}\, a_{ijkl}/N)\end{aligned}$$

which we denote in short as

$$c_n = \sigma\, d_a + (1-\sigma)\, c_{n-1}\, A$$

where $d_a \in \mathbb{R}^{N^2}$, $(d_a)_{ij} = \sum_{k,l=1}^{N} d_{ijkl}\, a_{ijkl}$ and we interpret all cost functions as N^2 dimensional row vectors and $c_{n-1}\, A$ denotes multiplication of the row vector c_{n-1} with the diffusion matrix A, $A_{ijkl} = a_{klij}/N$, from the left.

In this short notation one can easily resolve the recursion and give a direct expression for c_n:

$$c_n = \sum_{k=0}^{n-1} \sigma\, (1-\sigma)^k\, d_a\, A^k + (1-\sigma)^n\, c_0\, A^n$$

One then obtains for the linear assignment cost of assignment π (which we will regard as a column vector) simply by

$$c_n\, \pi = \sum_{k=0}^{n-1} \sigma\, (1-\sigma)^k\, d_a\, A^k\, \pi + (1-\sigma)^n\, c_0\, A^n\, \pi\,.$$

The second term is $(1-\sigma)^n c_0\, A^n\, \pi$, where A is a stochastic matrix. Thus its absolute value is bounded by $(1-\sigma)^n\, C\, N$, where C is a bound of the absolute value of entries of c_0. For $\sigma \in]0,1]$ this term tends to 0 as $n \to \infty$. By the same reasoning the sum in the first term converges for $n \to \infty$ (geometric series) and for the limit we obtain

$$\lim_{n\to\infty} c_n\, \pi = \sum_{k=0}^{\infty} \sigma\, (1-\sigma)^k\, d_a\, A^k\, \pi\,.$$

Now consider the assignment diffusion determined by assignment field a, initial distribution $p_0 = \pi$ and reset probability σ. For $n > 0$ we recursively define

$$p_n = \sigma\, \pi + (1-\sigma)\, A\, p_{n-1}\,.$$

Again, one can easily resolve the recursion and obtain

$$p_n = \sum_{k=0}^{n-1} \sigma\, (1-\sigma)^k\, A^k\, \pi + (1-\sigma)^n\, A^n\, p_0\,.$$

The second term is equal to $(1-\sigma)^n A^n p_0$. Since $p_0 = \pi$ is bounded in mass and L_1-norm, A is stochastic and $\sigma \in]0,1]$ this term vanishes in the limit $n \to \infty$. Again, the sum in the first term is a geometric series. For the diffusion step cost in the limit we then find

$$\lim_{n\to\infty} \sum_{ijkl} d_{ijkl}\, a_{ijkl}\, (p_n)_{ij} = \lim_{n\to\infty} d_a\, p_n = \sum_{k=0}^{\infty} \sigma\,(1-\sigma)^k\, d_a\, A^k\, \pi\,. \qquad \square$$

Proposition 3.2.21 (Reset Probability and Tightness of Diffusion Bound)**.** *The diffusion step cost in the equilibrium, minimized with respect to initial distribution π and assignment field a is a non-increasing function of the reset probability $\sigma \in]0,1]$.*

That is, the diffusion relaxation bound of the QAP is non-decreasing for decreasing σ. Before we can prove Proposition 3.2.21 it is convenient to establish the following lemma:

Lemma 3.2.22. *If $\pi \in \mathbb{R}^{N^2}$ is a doubly stochastic matrix and A is a assignment diffusion matrix, then $A\,\pi$ is doubly stochastic, where π is interpreted as column vector during multiplication with A. Also $\sigma(\mathrm{id} - (1-\sigma)\,A)^{-1}\pi$ is doubly stochastic for $\sigma \in]0,1]$.*

Proof. Let π be doubly stochastic. Then obviously $A\,\pi$ is non-negative. Further one finds

$$\sum_{j=1}^{N} (A\,\pi)_{ij} = \sum_{j,k,l=1}^{N} 1/N\, a_{klij}\, \pi_{kl} = \sum_{k,l=1}^{N} 1/N\, \pi_{kl} = 1$$

and

$$\sum_{i=1}^{N} (A\,\pi)_{ij} = \sum_{i,k,l=1}^{N} 1/N\, a_{klij}\, \pi_{kl} = \sum_{k,l=1}^{N} 1/N\, \pi_{kl} = 1$$

where we have used that a_{kl} is a permutation matrix for any (k,l). This proves the first claim.

Consider now

$$\sigma(\mathrm{id} - (1-\sigma)\,A)^{-1}\pi = \sigma \sum_{k=0}^{\infty} (1-\sigma)^k A^k\, \pi\,.$$

By virtue of the first part of the lemma any term $A^k\,\pi$ is double stochastic. The coefficients $\sigma(1-\sigma)^k$ are non-negative and sum up to 1 over $k = 0, \ldots, \infty$. Thus the whole term is a limit of convex combinations of doubly stochastic matrices, which makes it a double stochastic matrix itself (since the set of doubly stochastic matrices is closed). $\square$

Proof of Proposition 3.2.21. Let $0 < \sigma_1 < \sigma_2 < 1$. Consider now a fixed assignment field a, with assignment distribution matrix A, $A_{ijkl} = a_{klij}/N$, and initial distributions $\pi_{1,2}$ (which for now need not be assignments). We now look at the assignment diffusion, described by the iteration $p_{i,n} = \sigma_i \, \pi_i + (1 - \sigma_i) \, A \, p_{i,n-1}$ for $i = 1, 2$ (see Definition 3.2.13). Then we obtain as limit distributions (the no-reset term vanishes, since $\sigma_i > 0$):

$$p_{i,\infty} = \sum_{k=0}^{\infty} \sigma_i \, (1 - \sigma_i)^k \, A^k \, \pi_i = \sigma_i \, \left(\mathrm{id} - (1 - \sigma_i) \, A\right)^{-1} \pi_i$$

where we used the geometric series formula for matrices. Recall that in $(1-\sigma_i) \, A$ the matrix A is stochastic (i.e. absolute value of eigenvalues at most 1) and $1 - \sigma_i < 1$, thus the sum converges and the inverse exists.

Let now

$$\pi_2 = \frac{\sigma_1}{\sigma_2} (\mathrm{id} - (1 - \sigma_2) \, A) \, (\mathrm{id} - (1 - \sigma_1) \, A)^{-1} \pi_1 \tag{3.2.23}$$

which implies that $p_{1,\infty} = p_{2,\infty}$. The expression for π_2 can be rewritten as follows:

$$\begin{aligned}
\pi_2 &= \frac{\sigma_1}{\sigma_2} (\mathrm{id} - (1 - \sigma_2) \, A) \, \left(\sum_{k=0}^{\infty} (1 - \sigma_1)^k \, A^k \right) \pi_1 \\
&= \frac{\sigma_1}{\sigma_2} \, \left(\sum_{k=0}^{\infty} (1 - \sigma_1)^k \, A^k - \frac{1 - \sigma_2}{1 - \sigma_1} \sum_{k=1}^{\infty} (1 - \sigma_1)^k \, A^k \right) x_1 \\
&= \frac{\sigma_1}{\sigma_2} \, \left(\left(1 - \frac{1 - \sigma_2}{1 - \sigma_1} \right) \left(\sum_{k=0}^{\infty} (1 - \sigma_1)^k \, A^k \right) + \frac{1 - \sigma_2}{1 - \sigma_1} \, \mathrm{id} \right) \pi_1 \\
&= \frac{\sigma_1}{\sigma_2} \, \left(\left(1 - \frac{1 - \sigma_2}{1 - \sigma_1} \right) (\mathrm{id} - (1 - \sigma_1) \, A)^{-1} + \frac{1 - \sigma_2}{1 - \sigma_1} \, \mathrm{id} \right) \pi_1 \\
&= \left(\frac{\sigma_2 - \sigma_1}{\sigma_2 (1 - \sigma_1)} \right) \sigma_1 \, (\mathrm{id} - (1 - \sigma_1) \, A)^{-1} \, x_1 + \frac{\sigma_1 \, (1 - \sigma_2)}{\sigma_2 \, (1 - \sigma_1)} \pi_1
\end{aligned}$$

If π_1 is doubly stochastic, then by virtue of Lemma 3.2.22 so is

$$\sigma_1 \, (\mathrm{id} - (1 - \sigma_1) \, A)^{-1} \, \pi_1 \, .$$

Furthermore, for $\sigma_2 > \sigma_1$ the coefficients $(\sigma_2 - \sigma_1)/(\sigma_2 \, (1 - \sigma_1))$ and $(\sigma_1 \, (1 - \sigma_2))/(\sigma_2 \, (1 - \sigma_1))$ are non-negative and sum up to 1, so also π_2 is doubly stochastic. Note, that this argument is not possible if $\sigma_2 < \sigma_1$, so this construction only works one way.

Consider now the following optimization problems:

$$\min_{\pi_n \in \mathrm{Perm}_N} \sum_{i,j,k,l=1}^{N} d_{ijkl} \, a_{ijkl} \, (p_{n,\infty})_{ij}$$

for $n = 1, 2$ and some fixed assignment field a. For fixed a the limit distribution $p_{n,\infty}$ is linear in π, so due to the famous Birkhoff–von Neumann theorem (Theorem 2.2.1) the optimization of π can be carried out over the set of doubly stochastic matrices. Then, as we have shown, by construction according to (3.2.23) we can for any doubly stochastic π_1 construct an π_2 which is also doubly stochastic and yields the same limit distribution. Thus, we can conclude that

$$\min_{\pi_1 \in \mathrm{Perm}_N} \sum_{i,j,k,l=1}^{N} d_{ijkl}\, a_{ijkl}\, (p_{1,\infty})_{ij} \geq \min_{\pi_2 \in \mathrm{Perm}_N} \sum_{i,j,k,l=1}^{N} d_{ijkl}\, a_{ijkl}\, (p_{2,\infty})_{ij}\,.$$

This inequality will also hold after additional minimization over $a \in \mathcal{A}$ on both sides, thus the proof is established. □

That is, the diffusion relaxation bound of the QAP is non-decreasing for decreasing σ.

With the diffusion relaxation of the QAP we have introduced a new family of lower bounds for the QAP, depending on the reset probability parameter σ. The well known GLB is contained as special case $\sigma = 1$. For decreasing σ we have shown that the bound does never decrease. Per se it appears to be a difficult problem, how the diffusion relaxation of the QAP should be solved: the limit distribution p_∞ depends on the initial distribution x and the assignment field y in a highly non-linear fashion, thus optimization seems unfeasible. However, by having shown the equivalence of assignment diffusion to pre-specified cost function iteration, and by having shown that the cost function fixed-point with minimal cost can be found by successively performing minimizing cost function iterations, we have presented a simple way to solve the diffusion relaxation of the QAP. The following corollary combines the results provided by the single propositions.

Corollary 3.2.23 (Minimizing cost function iteration and diffusion relaxation of the QAP)**.** *Combining the results from Propositions 3.2.20 and 3.2.10 we can see for some $\sigma \in]0, 1]$, that the minimizing cost function iteration yields the fixed-point which has minimal associated linear assignment cost and which is thus yields the optimum of the diffusion relaxation of the QAP with reset probability σ.*

In particular for $\sigma = 1$ the diffusion relaxation bound is equal to the Gilmore-Lawler bound and by virtue of Proposition 3.2.21 the diffusion relaxation bound value is non-decreasing with decreasing σ. That means, the GLB is contained in the presented family of bounds as the weakest one.

Due to the continuity of the equilibrium distribution for $\sigma \in [0, 1]$ (Proposition 3.2.16) and therefore also of the diffusion relaxation QAP value, the minimal linear assignment cost associated with the minimizing cost function iteration fixed-point is a continuous function for $\sigma \in]0, 1]$ and converges for $\sigma \to 0$.

Remark 3.2.24 (Cost function iteration and dual QAP relaxation framework)**.** As briefly mentioned in Remark 2.3.6, the GLB can be seen as a special case of a framework of approximate solvers for the dual of a relaxed QAP formulation. Within this framework one adds Lagrangian multipliers for the constraints dropped in the linear formulation (2.3.4) and alternates between recomputing the GLB (taking into account the multipliers) and suitable updates of the multipliers. This seems similar to the proposed cost function iteration. However, we could not find a decomposition of the changes made to the standard GLB by our iterations, that can be interpreted as Lagrangian multipliers.

3.2.4 Experiments

In this section we provide a brief experimental study of the potential of the cost function iteration. Due to the naïve computational effort (see Remark 3.2.25 below) we will only study small examples, which nevertheless give a good overview over its weaknesses and strengths.

3.2.4.1 Implementation Details

Implementation of the minimizing cost function iteration, Def. 3.2.3, is not hard. So in principle one can determine the value for the diffusion relaxation for any $\sigma \in]0,1]$ by iteration. For small $\sigma \to 0$ the contraction ratio $1-\sigma \to 1$ and thus convergence can become very slow. In practice we started iterating with $\sigma = 1$, i.e. we computed the GLB and then gradually decreased σ by a constant factor. Before every decrement we checked via the a-posteriori estimate [BC11, Thm. 1.48]

$$\delta(c_n, c_\infty) \leq \delta(c_n, c_{n-1}) \frac{1-\sigma}{\sigma} \tag{3.2.24}$$

where δ denotes the max-metric, whether the iteration was sufficiently close to the fixed-point.

Just as the GLB, the cost function iteration can also provide a coarse upper bound to the QAP objective value by evaluating the QAP functional on the optimal linear matching. There is no guarantee that this upper bound decreases with decreasing σ, but in practice it usually does and we can always pick the smallest upper bound obtained so far.

Remark 3.2.25 (Computational complexity of cost function iteration)**.** Every step of the cost function is computationally equivalent to computing the GLB, hence the cost is $\mathcal{O}(N^5)$ per iteration (Remark 2.3.5). This constrains practical application to relatively small problems. In principle, application of the multi-scale algorithm presented in Chap. 5 is possible, but sophisticated data structures are required to handle the on-demand computation of entries of different cost functions c_n on different scales at different times. In this section we remain within the regime that is accessible via naïve computations.

3.2.4.2 Point Cloud Matching

As a first scenario we consider isometric matching of random point clouds in 2D. A first point cloud is randomly sampled. It is then subjected to an isometric transformation and distortion through noise to generate a second point cloud. Via the pairwise metric fidelity matching cost $\Gamma(x, y, x', y')^2 = |d_X(x, x') - d_Y(y, y')|^2$ (2.5.7), discussed in the context of the Gromov-Hausdorff / Wasserstein distances, we then try to find the optimal near-isometry between the two point clouds via cost function iteration. In a variant, additional clutter points are added to the second point cloud and the objective is to find the best near-isometric partial matching. Figure 3.9 illustrates the obtained results. In both cases we see, that the upper bound and the gap between upper and lower bound could be significantly decreased in most cases. Since the upper bound is very sensitive to irregular assignments we can thus deduce that the idea of implicit regularization does indeed work in most cases.

3.2.4.3 QAPLIB

The QAPLIB [BKR97] is a set of problem instances for the QAP, intended to provide a comparison benchmark for various (approximate) solution approaches. We applied the cost function iteration to several instances to compare the obtained bound with other methods. Unfortunately the results are quite disappointing. The GLB could hardly ever be increased by more than 2% and the corresponding upper bound was usually not decreased by more than 10%. Other approximation methods found in the literature can perform significantly better. Some details are given in Table 3.1.

3.2.4.4 Image Segmentation with Shape Prior

The intention behind the development of the cost function iteration was to improve the results of the shape prior based on the relaxed linearized Gromov-Wasserstein distance, as presented in Sect. 3.1. Figure 3.10 illustrates results obtained on an example with linear appearance model and different noise levels. As on the QAPLIB instances, the lower bounds are not increased substantially. But we see that this is because the linear appearance component is quite large compared to the minimally possible quadratic geometrical assignment costs. On the lowest noise level the basic approach from Sect. 3.1, equivalent to the GLB, already performs quite well, providing a relatively smooth assignment. The iteration is able to regularize the remaining outliers and to reduce the gap between upper and lower bound to about 5%. On the intermediate noise level the basic approach already produces several irregularities, most of which are correctly smoothed through iterations. On the highest noise level the basic approach produces a very irregular assignment which can no longer be improved through iterations. Only the coarse location of the triangle is detected.

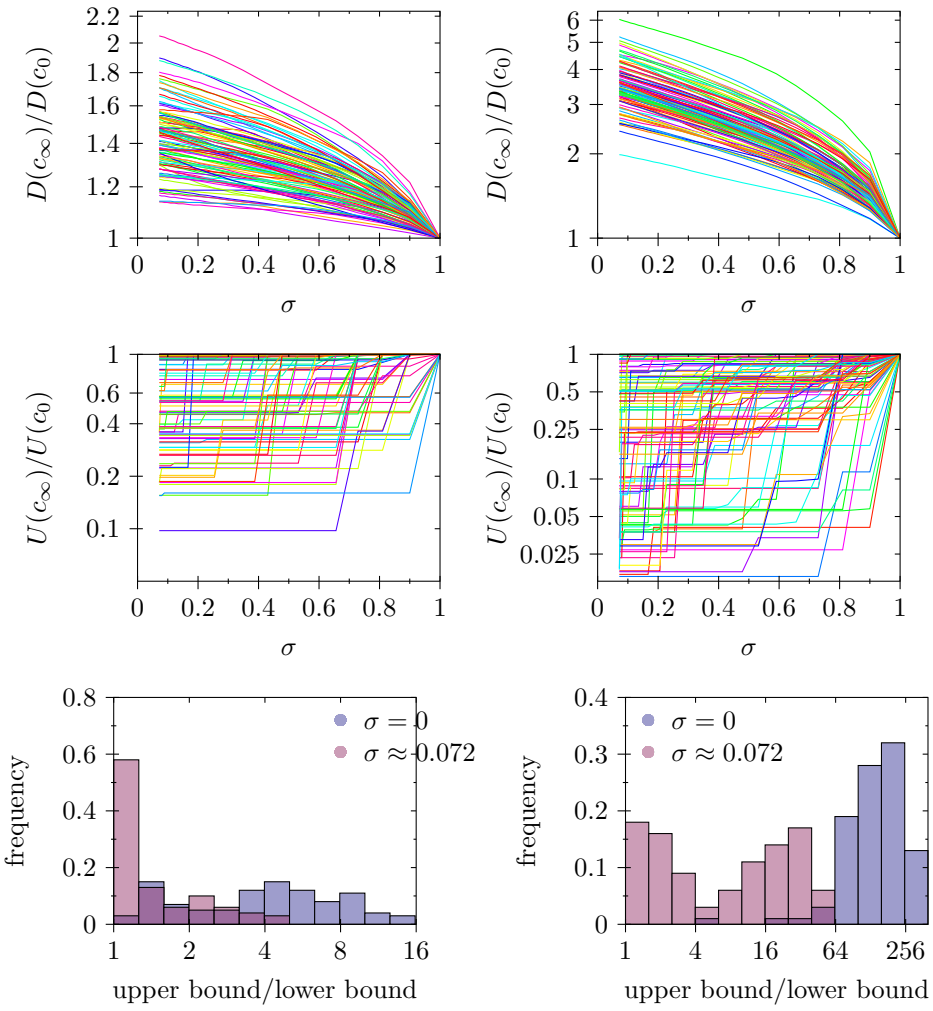


Figure 3.9: Performance of cost function iteration on isometric point cloud matching (see text). **Left column:** 10 points per cloud, **right column:** 10 additional clutter points in cloud 2. Data shown for 100 random iterations. **First row:** relative growth of the QAP diffusion bound with decreasing σ as compared to GLB ($\sigma = 1$). **Second row:** relative decrement of the best upper bound obtained so far, relative to GLB. **Third row:** histograms over upper/lower bound ratios for GLB ($\sigma = 1$) and $\sigma \approx 0.072$. We observe a consistent increase in the lower bound, a decrease in the upper bound. With cost function iteration most upper / lower bound ratios are significantly reduced and often brought close to 1, i.e. the problem is solved in good approximation.

Had16	3.72E3	3.36E3	4.02E3	3.38E3	3.89E3	3.72E3
Had18	5.36E3	4.78E3	5.80E3	4.82E3	5.80E3	5.36E3
Had20	6.92E3	6.17E3	7.62E3	6.22E3	7.46E3	6.92E3
Kra30a	8.89E4	6.84E4	1.16E5	6.93E4	1.11E5	8.67E4
Kra30b	9.14E4	6.91E4	1.26E5	6.99E4	1.11E5	8.77E4
Nug12	5.78E2	4.93E2	8.50E2	4.94E2	6.78E2	5.78E2
Nug15	1.15E3	9.63E2	1.44E3	9.67E2	1.30E3	1.15E3
Nug18	1.93E3	1.55E3	2.40E3	1.57E3	2.28E3	1.93E3
Nug20	2.57E3	2.06E3	3.13E3	2.07E3	2.89E3	2.57E3
Nug22	3.60E3	2.48E3	4.59E3	2.52E3	4.24E3	3.59E3
Nug30	6.12E3	4.54E3	7.34E3	4.56E3	7.26E3	5.93E3
Rou15	3.54E5	2.99E5	4.17E5	3.05E5	3.98E5	3.54E5
Rou20	7.26E5	5.00E5	8.20E5	6.10E5	8.20E5	7.25E5
Tai20a	7.03E5	5.81E5	8.78E5	5.90E5	8.56E5	7.03E5
Tai25a	1.17E6	9.62E5	1.46E6	9.73E5	1.42E6	1.11E6
Tai30a	1.82E6	1.50E6	2.16E6	1.52E6	2.13E6	1.71E6
Tho30	1.50E5	9.06E4	1.92E5	9.15E4	1.87E5	1.43E5

Table 3.1: Cost function iteration on QAPLIB problems: **First column:** name of instance (number gives problem size). **Second column:** optimal value. **Third column:** lower and upper bound provided by GLB. **Fourth column:** lower and upper bound provided by cost function iteration at $\sigma \approx 0.072$. **Fifth column:** best bound provided by either [BV06, AGHH07, Zhu07]. A more detailed table can be found at [HA10]. Unfortunately the cost function iteration cannot improve significantly upon the GLB. Other approximate methods prove to be much more efficient.

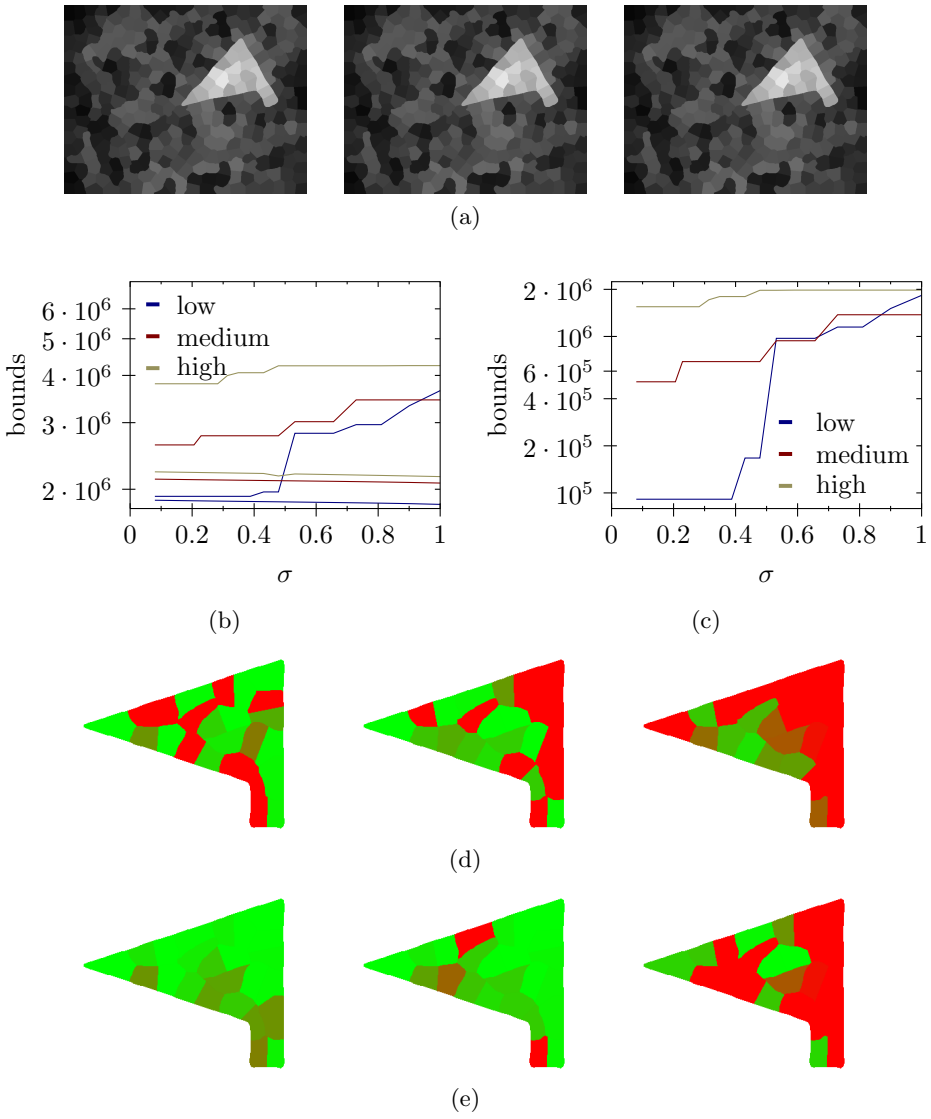


Figure 3.10: Cost function iteration applied to shape segmentation with relaxed linearized Gromov-Wasserstein distance (Sect. 3.1): (a) super-pixel input images with three different noise levels (low, middle, high), gray-level indicates foreground-affinity. (b) upper and lower bounds produced by cost function iteration. (c) upper bound on the geometric component of the assignment costs. (d) deviations of assignments of template super-pixels from ground-truth assignment using the GLB. (e) deviations of assignments of template super-pixels from ground-truth assignment after cost function iteration at $\sigma \approx 0.080$. It can be seen that the cost function iteration increases the robustness to noise and produces more regular assignments at moderate noise levels. At high noise levels the iteration can no longer improve the already quite distorted initial assignments.

3.2.5 Conclusion

We have discussed a method for tightening the relaxation applied in Sect. 3.1, that applies to general instances of the quadratic assignment problem, by implicit regularization of the assignment. Compared to explicit methods, discussed in the beginning, this has the advantage that the overall problem naturally decomposes into smaller instances and that a deterministic optimal assignment always exists. Further elegant theoretical properties, such as convergence of the cost function iteration and improvement of the bound through adjusting the parameter σ have been established.

But the naïve computational complexity of the iterations is prohibitive for large problem instances. As briefly discussed, this will require further development of the methods presented in Chap. 5, to be applicable to this problem. Also, it was demonstrated experimentally on the QAPLIB that the method is not competitive with state-of-the-art bounds for the QAP. Thus, this concept requires further research and development. Numerical results on an example of isometry invariant object segmentation however illustrated its potential.

Chapter 4

Wasserstein Modes

4.1 Shape Measures

In Section 2.1 indicator functions have been introduced to describe binary image segmentations. For combination with shape priors we choose to interpret such functions as densities of corresponding measures (Def. 2.1.2). As discussed in Section 1.2.2 it is rather difficult to model shape variations directly in the indicator function representation, whereas it is rather straight-forward on contours. In this section we will show that, under sufficient regularity conditions, the set of segmentation measures, considered as a submanifold of the 'manifold' of measures (Sect. 2.4.2) is in fact diffeomorphic to the manifold of simple closed contours (Sect. 2.6.1). We can then use this equivalence to construct shape priors that combine the advantages of the indicator and the contour representation (Sect. 4.2).

4.1.1 Shape Measures and Contour Lifting

4.1.1.1 Shape Measures

We now introduce the set of shape measures which are measures whose density is given by indicator functions with sufficiently regular support. All measures are normalized to have unit-mass to keep them comparable through optimal transport.

Definition 4.1.1 (Shape measure [SS13a, Def. 3.1])**.** A measure μ is called a shape measure, if there is an open set Ω, $0 < |\Omega| < \infty$ with a connected C^∞-boundary such that for all measurable $A \subset \mathbb{R}^2$

$$\mu(A) = |\Omega|^{-1} \cdot |A \cap \Omega| \,. \tag{4.1.1}$$

Integration w.r.t. μ can then be written as

$$\int_A \phi \, d\mu = |\Omega|^{-1} \int_{A\cap\Omega} \phi \, dx \,. \tag{4.1.2}$$

Denote by $\operatorname{dens}(\mu)$ the corresponding density function w.r.t. the Lebesgue measure. Identifying functions that are equal a.e., the density is unique and given by $\operatorname{dens}(\mu) = |\Omega|^{-1}\operatorname{ind}_\Omega$ where $\operatorname{ind}_\Omega$ denotes the indicator function of Ω. We can then write

$$\int_A \phi \, d\mu = \int_A \phi \, \operatorname{dens}(\mu) \, dx \,. \tag{4.1.3}$$

Denote the set of shape measures by $\mathcal{SP}$.

Remark 4.1.2. We require that Ω is of class C^∞ and has a connected boundary to obtain compatibility with the contour description of shapes as introduced in Sect. 2.6.1. Within this class of regular shape measures one can however describe a more general class of shapes by metric completion in the sense of [DZ11, Chap. 3, Thm. 3.1]. The analogous step in the context of contours is briefly discussed in [MM06].

In analogy to Definition 2.4.14 we now introduce a corresponding tangent space for shape measures.

Definition 4.1.3 (Shape tangent space [SS13a, Def. 3.3])**.** For a shape measure μ the shape tangent space $\operatorname{STan}(\mu)$ at μ is defined as

$$\operatorname{STan}(\mu) = \{\nabla u : \; u \in \mathcal{D} \wedge \Delta u = \text{const in } \operatorname{spt}(\mu)\} \,. \tag{4.1.4}$$

Compared to the tangent space for conventional optimal transport, Definition 2.4.14, two modifications have been made: an additional constraint $\Delta u = \text{const}$ is introduced and no completion w.r.t. $L^2(\mu)$ is made.

The first modification ensures that vectors in $\operatorname{STan}(\mu_t)$ correspond to deformations that, to first order, keep the density of μ_t constant within its support. Consider for a given shape measure μ and some $\alpha \in \operatorname{STan}(\mu)$ the path $\mu_t = (\mathrm{id} + t \cdot \alpha)_\sharp \mu$. The density of μ_t can be determined from the density of μ and Jacobian of $\mathrm{id} + t \cdot \alpha$. A brief calculation shows that near $t = 0$

$$\det J_{\mathrm{id} + t\cdot\alpha} = 1 + t \cdot \operatorname{div} \alpha + \mathcal{O}(t^2) \,. \tag{4.1.5}$$

Since $\operatorname{div} \alpha = \text{const}$ on $\operatorname{spt} \mu$ for $\alpha \in \operatorname{STan}(\mu)$, we see that deformation keeps the density of μ_t homogeneous to first order.

The second change ensures that shape measure trajectories with tangents in $\operatorname{STan}(\mu_t)$ retain a C^∞ boundary during evolution. The following theorem properly establishes the relation between $\operatorname{STan}(\mu)$ and absolutely continuous paths of shape measures.

Theorem 4.1.4 (Paths with Tangents in $\operatorname{STan}(\mu_t)$ are Absolutely Continuous Shape Measure Paths [SS13a, Thm. 3.4])**.** *Given a measure path $t \mapsto \mu_t$ and a flow field path $t \mapsto \alpha_t$, $t \in [0, 1]$ such that*

(i) μ_t is a shape measure at $t = 0$,

(ii) $\alpha_t \in \mathrm{STan}(\mu_t)$,

(iii) μ_t *and* α_t *satisfy the continuity equation,*

(iv) there is an open bounded set $\hat{\Omega}$ *such that* $\operatorname{spt}\mu_t \subset \hat{\Omega}$ *for all* $t \in [0,1]$,

(v) $\alpha \in \mathcal{X}_p(\hat{\Omega})$ *for any positive integer* p, *see* (2.6.11),

then μ_t *is an absolutely continuous shape measure path .*

Remark 4.1.5. Note hat in condition (ii) we demand $\alpha_t \in \mathrm{STan}(\mu_t)$ without knowing whether $\mu_t \in \mathcal{SP}$. However, the definition of $\mathrm{STan}(\mu_t)$ is formally valid also for this case.

Proof. Since $\alpha \in \mathcal{X}_p(\hat{\Omega})$, by virtue of Theorem 2.6.9 the family of maps φ_t, $t \in [0,1]$ defined by

$$\partial_t \varphi_t = \alpha_t \circ \varphi_t, \qquad \varphi_0 = \mathrm{id}$$

is a family of C^∞-diffeomorphisms on $\hat{\Omega}$. One has $J_{\varphi^{-1}} = (J_\varphi \circ \varphi^{-1})^{-1}$. Therefore, one finds

$$0 = \partial_t \left(\varphi_t \circ \varphi_t^{-1}\right) = (\partial_t \varphi_t) \circ \varphi_t^{-1} + (J_{\varphi_t} \circ \varphi_t^{-1}) \partial_t(\varphi_t^{-1}) \tag{4.1.6}$$

and then

$$\partial_t(\varphi_t^{-1}) = -(J_{\varphi_t} \circ \varphi_t^{-1})^{-1}\big((\partial_t \varphi_t) \circ \varphi_t^{-1}\big) = -J_{\varphi_t^{-1}}\, \alpha_t\,. \tag{4.1.7}$$

Consider now the following integral and its time derivative for a test function $\phi \in \mathcal{D}$:

$$\begin{aligned} &\frac{d}{dt}\int \phi\, d(\varphi_{t\sharp}^{-1}\mu_t) = \frac{d}{dt}\int \phi \circ \varphi_t^{-1}\, d\mu_t \\ &= \int \langle (\nabla\phi) \circ \varphi_t^{-1}, \underbrace{\partial_t \varphi_t^{-1}}_{-J_{\varphi_t^{-1}}\alpha_t} \rangle\, d\mu_t + \int \underbrace{\langle \nabla(\phi \circ \varphi_t^{-1}), \alpha_t \rangle}_{\langle (\nabla\phi)\circ\varphi_t^{-1}, J_{\varphi_t^{-1}}\alpha_t\rangle}\, d\mu_t = 0 \end{aligned} \tag{4.1.8}$$

Where we applied the chain rule in the first term and the continuity equation in the second. Thus we have by integration

$$\int \phi\, d(\varphi_{t_1}^{-1}{}_\sharp \mu_{t_1}) = \int \phi\, d(\varphi_{t_2}^{-1}{}_\sharp \mu_{t_2}) \tag{4.1.9}$$

for all $t_1, t_2 \in [0,1]$. From [Hör90, Thm. 1.2.5] we know that if

$$\int f\,\phi\, dx = \int g\,\phi\, dx \qquad \text{for all} \quad \phi \in \mathcal{D}$$

for locally integrable f, g then $f = g$ almost everywhere. Let f and g be the density functions of $(\varphi_{t_i}^{-1}{}_\sharp \mu_{t_i})$ for $i = 1, 2$. These exist since the density

functions of μ_{t_i} exist and $\varphi_{t_i}^{-1}$ are diffeomorphisms. Then we can conclude that these density functions agree a.e. and hence the measures are in fact identical.

We thus have

$$\varphi_{t\,\sharp}^{-1}\mu_t = \varphi_{0\,\sharp}^{-1}\mu_0 = \mu_0 \tag{4.1.10}$$

and by conjugation of the push-forward with φ_t find

$$\mu_t = \varphi_{t\,\sharp}\mu_0\,. \tag{4.1.11}$$

Now check the Jacobian determinant of φ_t: recall for a differentiable family of matrices

$$\frac{d}{dt}\det(A_t) = \det(A_t)\ \mathrm{tr}(A_t^{-1}\,\partial_t A_t)\,. \tag{4.1.12}$$

From Theorem 2.6.9 we have that the Jacobian of φ_t satisfies

$$\partial_t J_{\varphi_t} = (J_{\alpha_t}\circ\varphi_t)J_{\varphi_t} \tag{4.1.13}$$

thus we find

$$\begin{aligned}\frac{\partial}{\partial t}\det(J_{\varphi_t}) &= \det(J_{\varphi_t})\,\mathrm{tr}(J_{\varphi_t}^{-1}\,\partial_t J_{\varphi_t}) &(4.1.14)\\ &= \det(J_{\varphi_t})\,\mathrm{tr}(J_{\varphi_t}^{-1}\,(J_{\alpha_t}\circ\varphi_t)J_{\varphi_t}) &(4.1.15)\\ &= \det(J_{\varphi_t})(\mathrm{div}\,\alpha_t)\circ\varphi_t &(4.1.16)\end{aligned}$$

where we have interpreted the vector field α_t as a map and denote its Jacobian accordingly by J_{α_t}. Since for any $x\in\mathrm{spt}(\mu_0)$ the path $\varphi_t(x)$ over t always lies within the support of μ_t for all $t\in[0,1]$ and since $\mathrm{div}\,\alpha_t = \text{const}$ throughout $\mathrm{spt}\,\mu_t$, the temporal derivative of the determinant of the Jacobian of φ_t is spatially constant. Since $J_{\varphi_0} = J_{\mathrm{id}} = 1$, i.e. $\det(J_{\varphi_0}) = 1$, one finds $\det J_{\varphi_t}$ is spatially constant at all times within $\mathrm{spt}(\mu_0)$. Since μ_0 is a shape measure, it has a density function, with constant value within $\mathrm{spt}(\mu_0)$ and zero elsewhere, i.e. a rescaled indicator function. One can then, through the push-forward via φ_t find density functions for other $t\in[0,1]$. Since $\det J_{\varphi_t}$ is constant within $\mathrm{spt}(\mu_0)$ one easily finds, that the density functions of μ_t is also a rescaled indicator function. Since φ_t is a C^∞-diffeomorphism it preserves simple connectedness and C^∞-smoothness of the boundary of $\mathrm{spt}(\mu_0)$. Therefore μ_t must be a shape measure at all times. Absolute continuity of the path μ_t is given by the assumption $\alpha\in\mathcal{X}_p(\hat{\Omega})$, from which absolute integrability with respect to $L^2(\mu_t)$ follows. □

Remark 4.1.6. Note that so far the term tangent space is only used in a sense of analogy, in the way that [AG13] discusses the weak Riemannian structure of Meas.

4.1.1.2 Lifting of Contours

Every contour $c \in \mathrm{Emb}$ has a well-defined interior $\Omega(c)$ of class C^∞. We formally define the map that takes c to the shape measure associated with $\Omega(c)$.

Definition 4.1.7 (Lifting of contours [SS13a, Def. 3.7])**.** For a C^∞-embedding $c : S^1 \to \mathbb{R}^2$ that parametrizes the boundary of some open, simply connected domain $\Omega(c)$ the corresponding shape measure $F(c)$ is given by

$$\big(F(c)\big)(A) = |\Omega(c)|^{-1} \cdot |A \cap \Omega(c)| \,. \tag{4.1.17a}$$

As in 4.1.1, integration w.r.t. $F(c)$ is given by

$$\int_A \phi \, dF(c) = |\Omega(c)|^{-1} \int_{A \cap \Omega(c)} \phi \, dx \,. \tag{4.1.17b}$$

It is evident that if two contours are related by some C^∞-diffeomorphism $\varphi : S^1 \to S^1$ such that $c_1 = c_2 \circ \varphi$ then $F(c_1) = F(c_2)$. Vice versa, if two contours c_1 and c_2 both parametrize the boundary of some shape measure μ, then there is a C^∞-diffeomorphism φ such that $c_1 = c_2 \circ \varphi$. Therefore we have:

Proposition 4.1.8 ([SS13a, Prop. 3.8])**.** *The map*

$$F_{\mathrm{B}} : \mathrm{B} \to \mathcal{SP} \,, \qquad [c] \mapsto F(c) \tag{4.1.18}$$

is a bijection between the quotient manifold B *and the space of shape measures* $\mathcal{SP}$*.*

4.1.1.3 Lifting of Contour Tangent Vectors

Let $c \in \mathrm{Emb}$ describe a shape in the contour representation and let $\mu = F(c)$ be the corresponding shape measure representation. Let $\alpha \in \mathrm{STan}(\mu)$ describe to first order a deformation of μ. The information encoded in α can also be encoded in some normal deformation $a \in H_c\mathrm{Emb}$ of the contour c. This raises the question how the two descriptions for deformation are related. As it turns out α is already completely determined by its behaviour on the boundary of $\mathrm{spt}(\mu)$. This can be used to define maps that convert between a and α.

Definition 4.1.9 (Lifting of contour tangent vectors [SS13a, Def. 3.9])**.** For a contour c and a normal deformation field $a \in H_c\mathrm{Emb}$, see Remark 2.6.4, we define the lifting

$$f_c : H_c\mathrm{Emb} \to \mathrm{STan}\big(F(c)\big), \qquad a \mapsto \alpha = f_c(a) \tag{4.1.19}$$

from c onto $F(c)$ as the gradient of the extended unique solution (up to constant shifts) of the Neumann problem

$$\Delta u = S \qquad \text{in } \Omega(c) \tag{4.1.20a}$$

$$\langle n, \nabla u \rangle = a \circ c^{-1} \qquad \text{on } \partial\Omega(c) \tag{4.1.20b}$$

where n is the outward pointing unit normal vector on $\partial\Omega(c)$ and

$$S = |\Omega(c)|^{-1} \int_{\partial\Omega(c)} a \circ c^{-1} \, ds \,. \tag{4.1.20c}$$

Proof. By virtue of Proposition 2.6.15 the solution u to the PDE is unique (up to constants, which we can fix arbitrarily) and sufficiently smooth, i.e. in $C^\infty(\overline{\Omega})$. We can then specify any well-designed extension method that maps $C^\infty(\overline{\Omega})$ to $\mathcal{D}$ to extend u. Since the extended u is in $\mathcal{D}$ and its Laplacian is constant within $\Omega(c)$, we find that $f_c(a) = \nabla u \in \mathrm{STan}\big(F(c)\big)$. □

Some examples for the lifting of contour tangent vectors to the shape measure tangent space are illustrated in Fig. 4.1.

Remark 4.1.10. The extension of the solution u to (4.1.20) is formally necessary such that $\alpha = \nabla u$ is contained in $\mathrm{STan}(F(c))$. For a unique solution u (up to the constant shift) there are many valid extensions, all of them however coincide on $\Omega(c)$. Hence, from now on we will identify functions in $\mathrm{STan}(\mu)$ that coincide on $\mathrm{spt}(\mu)$.

With this identification rule, for a fixed c, the map f_c is a bijection between $H_c\mathrm{Emb}$ and $\mathrm{STan}\big(F(c)\big)$, with inverse given by

$$f_c^{-1}(\alpha)(\theta) = \langle \alpha \circ c(\theta), n_c(\theta) \rangle_{\mathbb{R}^2} \text{ for } \theta \in S^1 \,, \tag{4.1.21}$$

that is taking the normal component on the restriction of α to the boundary.

In analogy to Proposition 4.1.8 we then find:

Proposition 4.1.11 ([SS13a, Prop. 3.11])**.** *The map*

$$f_{\mathrm{B}[c]} : T_{[c]}\mathrm{B} \to \mathrm{STan}\big(F_{\mathrm{B}}([c])\big), \qquad [a] \mapsto f_c(a) \tag{4.1.22}$$

is a bijection between the tangent space $T_{[c]}\mathrm{B}$ on the quotient manifold and the shape measure tangent space $\mathrm{STan}\big(F_{\mathrm{B}}([c])\big)$ at the shape measure obtained by lifting the footpoint contour.

Proof. Keep in mind the identification rule in Remark 4.1.10 and the resulting bijectivity through (4.1.21). Further, let $c_1 \sim c_2$ be two contours, related by some $\varphi \in \mathrm{Diff}$, i.e. $c_2 = c_1 \circ \varphi$, i.e. $F(c_1) = F(c_2)$, and let a_1, a_2 be two respective normal deformation fields. Then obviously $f_{c_1}(a_1) = f_{c_2}(a_2)$ if and only if $a_2 = a_1 \circ \varphi$, that is when $(c_1, a_1) \sim (c_2, a_2)$ in the sense of Proposition 2.6.5. Then, by the representation property from Proposition 2.6.5 the proposition follows. □

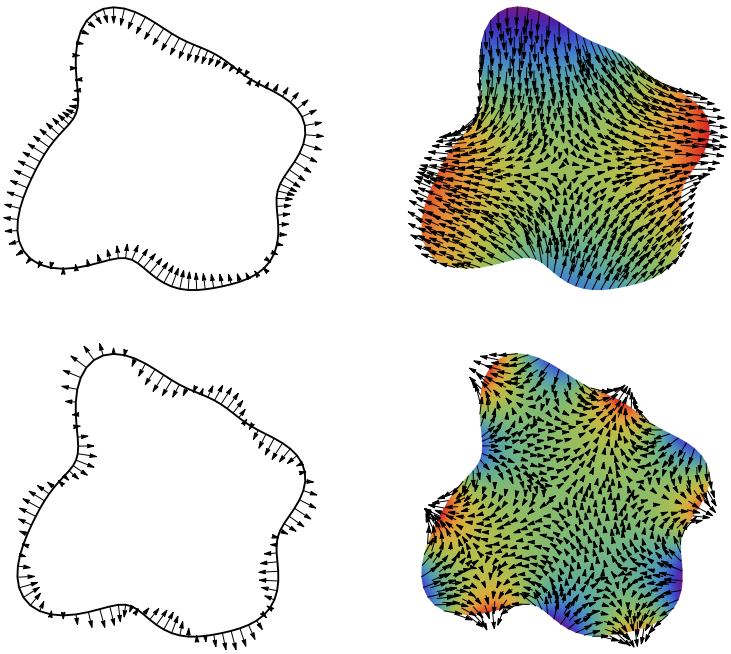


Figure 4.1: Lifting of tangential deformation fields. Left column: two different normal deformation fields on a given contour. Right column: lifted flow-fields with constant divergence on the corresponding shape measure. Color shading indicates the potential function that solves the involved Neumann problem. In the first example the contour deformation has a low frequency and the lifted flow-field has large amplitudes throughout the interior. In the high-frequency example in the second row, the lifted flow-field has non-vanishing amplitude only near the boundary. Note that the lifted flow-field is not normal to the contour at the shape boundary.

So far we have established that there is a map f_c that takes the (horizontal part of the) tangent space $H_c\mathrm{Emb}$ to $\mathrm{STan}\big(F(c)\big)$. Let c_t be a path of contours and let $\partial_t c_t$ be the tangent vectors. We need yet to check that $f_{c_t}(\partial_t c_t)$ is tangent to $F(c_t)$ in the sense of the continuity equation (2.4.23).

Theorem 4.1.12 (Commutation of deformation and lifting [SS13a, Thm. 3.12])**.** *Given a contour path c_t which is C^1 in time, with normal temporal deformation a_t, the following commutation relation holds for all test functions $\phi \in \mathcal{D}$:*

$$\frac{d}{dt}\int \phi \, dF(c_t) = \int \langle \nabla\phi, f_{c_t}(a_t)\rangle \, dF(c_t) \tag{4.1.23}$$

The implication is that the measure path $F(c_t)$ generated by lifting c_t satisfies the continuity equation (2.4.23) together with the flow-field $f(a_t)$ generated

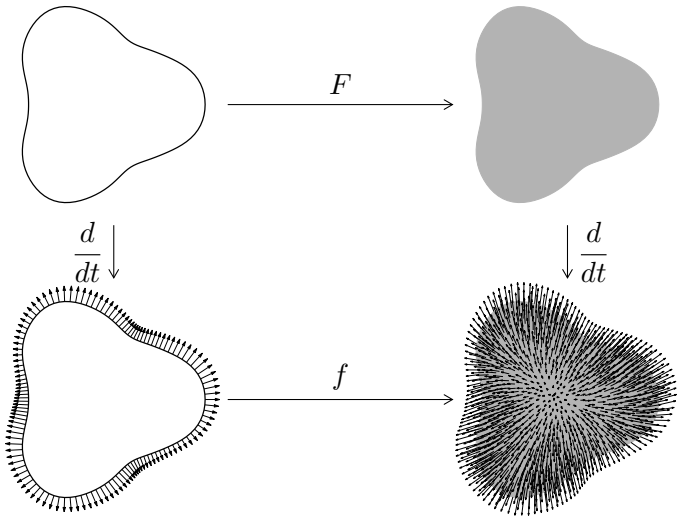


Figure 4.2: Shapes can be represented by parametrized contours (top left) or probability measures with constant density in the interior and zero elsewhere (top right). Let F be the map that takes contours to measures. Taking the time derivative of a path of contours will yield a normal deformation field (bottom left), on a path of measures it will yield a flow-field according to the continuity equation in optimal transport (bottom right). In this section we will discuss a map f that takes contour deformation fields to measure deformation fields, such that the diagram above commutes (Theorem 4.1.12).

by lifting the tangent path a_t. In analogy to (2.4.23) we can write for (4.1.23):

$$\frac{d}{dt}F(c_t) = -\nabla\left(f_{c_t}(a_t)\, F(c_t)\right) \tag{4.1.24}$$

This corresponds to the following commutation diagram (cf. Fig. 4.2):

$$\begin{array}{ccc} c_t & \xrightarrow{\text{time derivative}} & a_t = \frac{d}{dt}c_t \\ \downarrow \text{lift} & & \downarrow \text{lift} \\ F(c_t) & \xrightarrow{\text{time derivative}} & f_{c_t}(\frac{d}{dt}c_t) \equiv \frac{d}{dt}F(c_t) \end{array}$$

Proof. From Lemma 2.6.8 we have:

$$\frac{d}{dt}\int_{\Omega(c_t)} \phi\, dx = \int_{\partial\Omega(c_t)} a_t \circ c_t^{-1}\, \phi\, ds \tag{4.1.25}$$

Then one finds:

$$\frac{d}{dt}\int \phi\, dF(c_t) \tag{4.1.26}$$

$$=\frac{d}{dt}\,|\Omega(c_t)|^{-1}\int_{\Omega(c_t)}\phi\, dx \tag{4.1.27}$$

$$=\left(\frac{d}{dt}\,|\Omega(c_t)|^{-1}\right)\int_{\Omega(c_t)}\phi(x)\, dx + |\Omega(c_t)|^{-1}\left(\frac{d}{dt}\int_{\Omega(c_t)}\phi\, dx\right) \tag{4.1.28}$$

$$=-\,|\Omega(c_t)|^{-2}\int_{\partial\Omega(c_t)} a_t\circ c_t^{-1}\, ds\cdot\int_{\Omega(c_t)}\phi\, dx + |\Omega(c_t)|^{-1}\int_{\partial\Omega(c_t)} a_t\circ c_t^{-1}\,\phi\, ds \tag{4.1.29}$$

$$=-\,|\Omega(c_t)|^{-2}\int_{\partial\Omega(c_t)} a_t\circ c_t^{-1}\, ds\cdot\int_{\Omega(c_t)}\phi\, dx + |\Omega(c_t)|^{-1}\int_{\Omega(c_t)}\operatorname{div}\left(f_{c_t}(a_t)\,\phi\right) dx \tag{4.1.30}$$

(in the second term the properties of the lifting $f_{c_t}(a_t)$ and the divergence theorem were used)

$$=-\,|\Omega(c_t)|^{-2}\int_{\partial\Omega(c_t)} a_t\circ c_t^{-1}\, ds\cdot\int_{\Omega(c_t)}\phi\, dx + |\Omega(c_t)|^{-1}\int_{\Omega(c_t)}\underbrace{\left(\operatorname{div} f_{c_t}(a_t)\right)}_{|\Omega(c_t)|^{-1}\int_{\partial\Omega(c_t)} a_t\circ c_t^{-1}\, ds}\phi\, dx + |\Omega(c_t)|^{-1}\int_{\Omega(c_t)}\langle f_{c_t}(a_t),\nabla\phi\rangle\, dx \tag{4.1.31}$$

($\operatorname{div} f(a_t)$ is determined by $f(a_t)=\nabla u$, where u is the solution to (4.1.20))

$$=\int\langle f_{c_t}(a_t),\nabla\phi\rangle\, dF(c_t)$$ □

This means, the region deformations encoded in a_t and $f_{c_t}(a_t)$ in their respective representations coincide.

4.1.2 Equivalence of Shape Measures and Contours

4.1.2.1 Paths in Contour and Shape Measure Description

In the assumptions to Theorem 4.1.4 we have established a regularity class of paths of shape measures. We will now show that smooth paths on Emb transform into such paths via lifting by F and vice versa that 'de-lifting' of such paths on $\mathcal{SP}$ will result in smooth paths on Emb.

Proposition 4.1.13 (Contours → Measures [SS13a, Prop. 3.13])**.** *Let* $[0,1]\ni t\mapsto c_t\in\mathrm{Emb}$ *be a* C^∞*-family of contours. Then the shape measure path generated by lifting,* $\mu_t=F(c_t)$ *together with* $\alpha_t=f_{c_t}(\partial_t c_t)$ *satisfies the conditions of Theorem 4.1.4.*

Proof. Conditions (i) and (ii) follow immediately from $\mu_t = F(c_t) \in \mathcal{SP}$ and $\alpha_t = f_{c_t}(\partial_t c_t)$. Condition (iii) is implied by Theorem 4.1.12.

Condition (iv) is established as follows: let $\Omega(c_t)$ be the interior of the region enclosed by contour c_t. Let further, for any Ω

$$d(x, \Omega) = \inf\{\|x - y\| : \ y \in \Omega\} \tag{4.1.32}$$

and

$$\Omega(c_t, \varepsilon) = \left\{x \in \mathbb{R}^2 : \ d\big(x, \Omega(c_t)\big) < \varepsilon\right\} . \tag{4.1.33}$$

Each set $\Omega(c_t, \varepsilon)$ is open and bounded. The topology on Emb guarantees that for any $t \in [0, 1]$ and $\varepsilon > 0$ there is some $\delta > 0$ such that

$$c_{t'}(S^1) \subset \Omega(c_t, \varepsilon) \quad \text{for all} \quad t' \in \tau = [t - \delta, t + \delta] \cap [0, 1] .$$

Pick then a set of (t_i, ε_i) such that the corresponding intervals τ_i cover $[0, 1]$. Since $[0, 1]$ is compact, there must be a finite subcovering. We assume that $\{(t_i, \varepsilon_i)\}_{i=1}^m$ induces such a covering. Then we find

$$\operatorname{spt} \mu_t = \overline{\Omega(c_t)} \subset \hat{\Omega} = \bigcup_{i=1}^{m} \Omega(c_{t_i}, \varepsilon_i) \tag{4.1.34}$$

for all $t \in [0, 1]$ where $\hat{\Omega}$ is open and bounded.

Let us finally turn to the last condition (v). First extend the normal boundary deformations $\partial_t c_t$ to a flow-field $[0, 1] \ni t \mapsto \beta_t \in C^\infty(\hat{\Omega}, \mathbb{R}^2)$, for example as outlined in [DZ11, Chap. 4, Sect. 3.3.2]. This construction can be designed such that $\beta \in \mathcal{X}_p(\hat{\Omega})$ for any integer $p \geq 0$. Let φ_t be the family of diffeomorphisms induced by β, according to Theorem 2.6.9. Note, that φ_t is in general not volume preserving.

The weak solution to Poisson's equation, describing the tangent vector lifting at time t is given by the minimizer w.r.t. $u \in H^1(\Omega(c_t))/\mathbb{R}$ of (2.6.24) with the following parameters

$$E(u, \Omega(c_t), f_t, g_t) = \frac{1}{2} \int_{\Omega(c_t)} \|\nabla u\|^2 \, dx + \int_{\Omega(c_t)} f_t \, u \, dx - \int_{\partial\Omega(c_t)} g_t \, u \, ds \tag{4.1.35}$$

with

$$g_t = \langle \partial_t c_t, n_{c_t} \rangle \quad \text{and} \quad f_t = \frac{1}{|\Omega(c_t)|} \int_{\partial\Omega(c_t)} g_t \, ds . \tag{4.1.36}$$

By means of function space parametrization [DZ11, Chap. 10, Sect. 2.2] we can express $H^1(\Omega(c_t))/\mathbb{R}$ in terms of $H^1(\Omega(c_0))$ and φ_t:

$$H^1(\Omega(c_t))/\mathbb{R} = \left\{u \circ \varphi_t^{-1} : \ u \in H^1(\Omega(c_0))/\mathbb{R}\right\} \tag{4.1.37}$$

For some $u \circ \varphi_t^{-1} \in H^1(\Omega(c_t))/\mathbb{R}$ we then find

$$E(u \circ \varphi_t^{-1}, \Omega(c_t), f_t, g_t) = \frac{1}{2} \int_{\Omega(c_t)} \|\nabla(u \circ \varphi_t^{-1})\|^2 \, dx + \int_{\Omega(c_t)} f_t \, (u \circ \varphi_t^{-1}) \, dx - \int_{\partial\Omega(c_t)} g_t \, (u \circ \varphi_t^{-1}) \, ds \tag{4.1.38}$$

$$= \frac{1}{2} \int_{\Omega(c_0)} \langle A_t \nabla u, \nabla u \rangle dx + \int_{\Omega(c_0)} \tilde{f}_t \, u \, dx - \int_{\partial\Omega(c_0)} \tilde{g}_t \, u \, ds \tag{4.1.39}$$

$$= E(u, \Omega(c_0), A_t, \tilde{f}_t, \tilde{g}_t) \tag{4.1.40}$$

with

$$A_t = |\det J_{\varphi_t}| \cdot \left(J_{\varphi_t}^{-1}\right) \left(J_{\varphi_t}^{-1}\right)^\top \tag{4.1.41}$$

and

$$\tilde{f}_t = |\det J_{\varphi_t}| \cdot (f \circ \varphi_t) \quad \text{and} \quad \tilde{g}_t = |\det J_{\varphi_t|_{\partial\Omega(c_0)}}| \cdot (g \circ \varphi_t) \tag{4.1.42}$$

where $\varphi_t|_{\partial\Omega(c_0)}$ denotes the restriction of φ_t to the sub-manifold $\partial\Omega(c_0)$ and $J_{\varphi_t|_{\partial\Omega(c_0)}}$ is the Jacobian of this restriction.

Since φ_t is continuous in t w.r.t. the topology of uniform convergence in all its derivatives (Lemma 2.6.10) there is a $t > 0$ such that the matrix $A_{t'}$ will be positive-definite with bounds $0 < \lambda < \Lambda$ such that $A_{t'}$ satisfies (2.6.26) for all $t' \in [0, t[$. $\tilde{f}_{t'}$ and $\tilde{g}_{t'}$ are always $C^\infty\big(\Omega(c_0)\big)$ and continuous in time w.r.t. the supremum norm in any derivative. Also, the map $u \to u \circ \varphi_{t'}^{-1}$ is continuous w.r.t. any Sobolev norm between the connected spaces.

Hence, by virtue of the discussion in Remark 2.6.14 the minimizers $u_{t'}$ to the functionals $E(\cdot, \Omega(c_{t'}), f_{t'}, g_{t'})$ for $t' \in [0, t[$ have a uniformly bounded Sobolev norm $\|u_{t'}\|_{W^{m,p}(\Omega)/\mathbb{R}}$ for any positive integer m and $1 < p < \infty$. Repeating this construction at different times until the whole interval $[0, 1]$ is covered by finitely many 'starting points', one can extend the uniform bound to $[0, 1]$. By the embedding theorem and by taking the gradient $\alpha_t = \nabla u_t$ it follows then that $\alpha \in \mathcal{X}_p(\hat{\Omega})$ for any non-negative integer p. □

And similarly for the opposite direction:

Proposition 4.1.14 (Measures → Contours [SS13a, Prop. 3.14])**.** *Let (μ_t, α_t) be a pair of shape measure and flow field paths satisfying the conditions for Theorem 4.1.4. Then there is a smooth path $[0, 1] \ni t \mapsto c_t \in \mathrm{Emb}$ such that $F(c_t) = \mu_t$.*

Proof. Let c_0 be a contour that parametrizes the boundary of the region given by μ_0. Then, as in the proof for Theorem 4.1.4, integrate α_t into a family of

C^∞-diffeomorphisms φ_t. As the push-forward of μ_0 under φ_t yields μ_t we can deduce that $c_t = \varphi_t \circ c_0$ parametrizes the boundary of μ_t, i.e. $F(c_t) = \mu_t$. Since φ_t is a C^∞-diffeomorphism at all times, c_t will be a C^∞-embedding $S^1 \to \mathbb{R}^2$ at all times, hence it will really be a path in Emb. Recall from Proposition 2.6.6: since Emb is an open submanifold of the space $C^\infty(S^1, \mathbb{R}^2)$, we show continuity of the path there, continuity in Emb then follows. Convergence in $C^\infty(S^1, \mathbb{R}^2)$ is verified by uniform convergence on S^1 in all derivatives separately.

In analogy to [You10, Lemma 8.3] it is easy to proof by induction that for any non-negative integer n

$$c_t^{(n)} = \partial_\theta^n(\varphi_t \circ c_0) \tag{4.1.43}$$

$$= \sum_{I:|I|\leq n} \big((\partial_I \varphi_t) \circ c_0\big)\, B_{I,n}(c_0) \tag{4.1.44}$$

where $B_{I,n}(c_0)$ is a linear combination of terms $(c_0^{(n_1)})_{i_1}\,(c_0^{(n_2)})_{i_2} \dots (c_0^{(n_q)})_{i_q}$ such that the tuple $I = (i_1, i_2, \dots, i_q)$ and $n = \sum_{r=1}^q n_r$. By virtue of Lemma 2.6.10 $\partial_I \varphi_{t_k} \to \partial_I \varphi_t$ uniformly as $t_k \to t$ for any multi-index I. Hence, uniform convergence $c_{t_k}^{(n)} \to c_k^{(n)}$ is implied. □

4.1.2.2 Shape Measures as a Manifold

We have now established bijections between B and $\mathcal{SP}$ and between the deformations $T_{[c]}\mathrm{B}$ and $\mathrm{STan}\big(F_\mathrm{B}([c])\big)$. Further, we have shown how regular paths in contour and measure descriptions transform into each other. In this section we will formally establish that the set of shape measures $\mathcal{SP}$ is a manifold, diffeomorphic to B.

Equip $\mathcal{SP}$ with the topology induced by the Wasserstein metric W. Then it is easy to see that F_B is continuous but F_B^{-1} is not.

Proposition 4.1.15 (Continuity of F_B)**.** *Equip the set of shape measures with the topology induced by the Wasserstein metric W. Then the map F_B is continuous.*

Proof. Let $\{[c_n]\}_n$ be a sequence in B converging to some $[c]$, where by $[\cdot]$ we denote the equivalence class of reparametrizations of a given element of B. Hence, there is a sequence of contours $\{c_n\}_n$ and a contour c in Emb with $c_n \in [c_n], c \in [c]$ such that $c_n \to c$ in Emb (Proposition 2.6.6).

Since $c_n \to c$ uniformly, there exists for any $\varepsilon > 0$ some $n(\varepsilon) \in \mathbb{N}$ such that for $n > n(\varepsilon)$ the contour c_n in $\mathbb{R}^2$ lies completely within a tube of thickness ε (both inwards and outwards) around the contour c. Anything within the inner boundary of the tube lies within both $\Omega(c_n)$ and $\Omega(c)$ and anything beyond the outer boundary is in neither of the two sets. The area of the tube goes to 0 as $\varepsilon \to 0$. Therefore also $|\Omega(c_n)| \to |\Omega(c)|$. Hence, we find for any test function

$\phi \in \mathcal{D}$:

$$\int \phi \, dF(c_n) \to \int \phi \, dF(c) \tag{4.1.45}$$

The measures $F(c_n)$ as well as the measure $F(c)$ have support limited to the union of $\Omega(c)$ and the aforementioned tube, which is bounded. Therefore convergence w.r.t. test functions corresponds to the notion of narrow convergence [AG13, Sect. 1.1] and we can also conclude that the second order moments of $F(c_n)$ converge towards the second moments of $F(c)$. Hence, by virtue of [AG13, Thm. 2.7] we have $W\big(F(c_n), F(c)\big) \to 0$. □

Proposition 4.1.16. *F_{B}^{-1} is not continuous.*

Proof. For sufficiently small $\lambda > 0$, consider the sequence of contours c_n and the contour c in Emb given by

$$c_n(\theta) = \big(1 + (\lambda/n)\sin(n \cdot \theta)\big) \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}, \qquad c(\theta) = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix} \tag{4.1.46}$$

We have $c_n \to c$ uniformly, but not its derivatives. Hence, as in the reasoning for Proposition 4.1.15, we can conclude that $F(c_n) \to F(c)$ in the Wasserstein topology, but we have $c_n \not\to c$ and also $[c_n] \not\to [c]$ on B. Hence, F_{B}^{-1} is not continuous. □

We see from this example, that convergence in the optimal transport sense is only concerned with convergence of the regions towards each other, regardless of the boundary or even higher order regularity as required on Emb. For this reason additional assumptions on the regularity of α were necessary in Theorem 4.1.4 to be able to transform paths back and forth between contour and measure description.

However, if we equip $\mathcal{SP}$ with the topology induced by F_{B}, then by definition F_{B} and also F_{B}^{-1} are continuous, thus constituting a homeomorphism between the two sets. Then $\mathcal{SP}$ inherits the manifold property from B. Let ψ_i, ψ_j be any two charts mapping overlapping open sets $U_i, U_j \subset \mathrm{B}$ into the modelling space. Then $\psi_i \circ F_{\mathrm{B}}^{-1}$ will be a chart on $\mathcal{SP}$. The corresponding chart change $(\psi_i \circ F_{\mathrm{B}}^{-1}) \circ (\psi_j \circ F_{\mathrm{B}}^{-1})^{-1}$ remains differentiable as the lifting onto $\mathcal{SP}$ cancels. Likewise, the map

$$(\psi_j \circ F_{\mathrm{B}}) \circ F_{\mathrm{B}}^{-1} \circ \psi_i^{-1} = \mathrm{id}\,,$$

that takes the modelling space of a chart on B onto the modelling space of a corresponding chart on $\mathcal{SP}$ is trivial and thus differentiable. Hence the two manifolds are actually diffeomorphic.

By virtue of Proposition 4.1.11 we can represent the tangent space on $\mathcal{SP}$ at μ by $\mathrm{STan}(\mu)$. And due to Theorem 4.1.12 we have that such tangent vectors naturally represent directional derivatives of functions on $\mathcal{SP}$ that are given

by region integrals over test functions. Evaluation is given by the continuity equation (2.4.23).

The diffeomorphism between the contour manifold B and the set of shape measures establishes that the shape measure representation is a formally equivalent way of describing shapes. The tangent space $\mathrm{STan}(\mu)$ gives a linear structure to describe deformations that is just as powerful as $T_{[c]}\mathrm{B}$ in terms of shape analysis and modelling. In addition, every shape is uniquely represented in the shape measure description, whereas one has to handle parametrization ambiguities in the contour representation. Shape measures are therefore an elegant way for shape representation in image segmentation tasks.

4.1.3 A Riemannian Metric on the Manifold of Shape Measures

4.1.3.1 Metric Structure of the Tangent Space

In Sect. 2.4.2 we have discussed the analogy of Meas, metrized by W_2, to a Riemannian manifold with metric tensor (2.4.31). In the last Section we have formally established, that the set of shape measures $\mathcal{SP}$ can be viewed as an infinite dimensional manifold, diffeomorphic to B, albeit with a topology which is not compatible with the metric topology induced by W. Nevertheless, since $\mathcal{SP} \subset \mathrm{Meas}$ and since the tangent space w.r.t. $\mathcal{SP}$, Definition 4.1.3, is a subset of the tangent space w.r.t. Meas, Definition 2.4.14, $\mathrm{STan}(\mu) \subset \mathrm{Tan}(\mu)$ for $\mu \in \mathcal{SP}$, it suggests itself to informally view the shape measures as a submanifold of all measures and to equip $\mathcal{SP}$ with the Riemannian metric that is induced by 'restricting' the metric tensor on Meas to the 'submanifold'. A prominent example of how such treatments can yield valuable insights, is the Otto calculus and its success in the context of interpreting partial differential equations as gradient flows (see for example [Vil09, Chap. 15]).

This will yield a new type of metric on the contour manifold B, as opposed to contour-based metrics, for example discussed in [MM06, MM07, SMSY11]. Formally one can find an expression for the new metric inner product by pull-back through F. One would find a non-local integral involving the kernel for the PDE (4.1.20). In this article we study the new metric directly in the measure representation where the inner product is local.

We start by analyzing the metric structure on $\mathrm{STan}(\mu)$. First note that the equivalence classes of tangent vectors, induced by the pseudo-metric (2.4.31) (two vectors being equivalent if they have zero distance), are just those described in Remark 4.1.10.

We now consider various subspaces of $\mathrm{STan}(\mu)$.

Translation. Let μ be some shape measure and $\alpha = v$ be a flow field that is constant in space for some $v \in \mathbb{R}^2$. Such fields span a two dimensional subspace of $\mathrm{STan}(\mu)$. Then $\mu_t = (\mathrm{id} + t \cdot \alpha)_\sharp \mu$ is for every t just the translation of μ by

the vector $t \cdot v$. One finds for any test function $\phi \in \mathcal{D}$

$$\begin{aligned}\frac{d}{dt}\int \phi\, d\mu_t\Big|_{t=0} &= \frac{d}{dt}\int \phi\, d(\mathrm{id} + t\cdot\alpha)_\sharp\mu\Big|_{t=0} \\ &= \frac{d}{dt}\int \phi\big(x + t\cdot\alpha(x)\big)\, d\mu(x)\Big|_{t=0} = \int \langle \nabla\phi, \alpha\rangle d\mu\,, \end{aligned} \tag{4.1.47}$$

i.e. $(\mu_t, \alpha_t = \alpha)$ satisfy the continuity equation.

Strictly, for $\alpha_t = \alpha$ to be within $\mathrm{STan}(\mu_t)$ at any time, we need to smoothly truncate it, such that its support is compact. We will assume that such a truncation has been applied, but at such a large radius that at all times $t \in [0,1]$ we have $\alpha = v$ on the support of μ_t.

Let now $\alpha \in \mathrm{STan}(\mu)$ be a flow-field that is orthogonal to any translation flow field w.r.t. the Riemannian inner product. That is

$$0 = \int \langle \alpha(x), v\rangle_{\mathbb{R}^2} d\mu \qquad \text{for all } v \in \mathbb{R}^2\,.$$

We then find

$$0 = \left\langle \int \alpha(x)\, d\mu, v \right\rangle_{\mathbb{R}^2} \qquad \text{for all } v \in \mathbb{R}^2, \text{ and thus} \qquad 0 = \int \alpha(x)\, d\mu\,.$$

From this follows after a brief calculation

$$0 = \frac{d}{dt}\int x\, d(\mathrm{id} + t\,\alpha)_\sharp\mu(x)\Big|_{t=0}$$

where we need to smoothly truncate the function $x \mapsto x$ beyond the support of μ to turn it into a test function. We then see that any tangent vector that is locally orthogonal to any translation field keeps the center of mass of μ unchanged.

Scale. Assume now, we fix some tangent vector $\alpha_{\mathrm{scale}} \in \mathrm{STan}(\mu)$ with divergence 1 in $\operatorname{spt}\mu$, that is orthogonal to the translation fields. We refer to α_{scale} as *scale component*. Then we can uniquely decompose any given tangent vector α into the following components:

$$\alpha = \alpha_{\mathrm{trans}} + \lambda\cdot\alpha_{\mathrm{scale}} + \alpha_{\mathrm{def}} \tag{4.1.48}$$

where α_{trans} is a *translation component* as discussed above, λ is given by $\operatorname{div}\alpha$ and α_{def} is a divergence-free residual, orthogonal to the translation component, we will refer to as *deformation component*. We now discuss, how a scale component can be determined which is orthogonal to all divergence-free flow fields, this includes the translation component and the residual α_{def}, such that the

decomposition (4.1.48) is an orthogonal one. Demand for any α with $\operatorname{div}\alpha = 0$ that

$$0 = \int \langle \alpha_{\text{scale}}, \alpha \rangle d\mu = \int_{\Omega} \langle \alpha_{\text{scale}}, \alpha \rangle dx \tag{4.1.49}$$

where $\Omega = \operatorname{spt}(\mu)$ and we can neglect the normalization factor $|\Omega|^{-1}$, since the integral vanishes. We then take $\alpha_{\text{scale}} = \nabla u_{\text{scale}}$, $u_{\text{scale}} \in \mathcal{D}$, and find

$$0 = \int_{\Omega} \langle \nabla u_{\text{scale}}, \alpha \rangle dx = \int_{\Omega} \nabla(u_{\text{scale}} \cdot \alpha) dx$$

where the second equality holds since $\operatorname{div}\alpha = 0$. Then by the divergence theorem

$$0 = \int_{\partial\Omega} u_{\text{scale}} \langle n, \alpha \rangle_{\mathbb{R}^2} ds \tag{4.1.50}$$

where n is the outward pointing unit-normal on $\partial\Omega$.

If u_{scale} were non-constant on $\partial\Omega$, one could locate a region where u_{scale} is above average (w.r.t. the boundary length as weight) and one, where u_{scale} is below average. One could then choose some smooth normal components for a field α, $\langle n, \alpha \rangle_{\mathbb{R}^2}$, that have some influx in the above-average region and a corresponding outward flux in the below average region with zero net flux. For this normal component (4.1.50) would be non-zero. This normal component could then be lifted to a complete divergence-free flow-field α by virtue of Definition 4.1.9, yielding a contradiction to assumption (4.1.49). Thus we can conclude that u_{scale} must be constant on $\partial\Omega$. We choose to set $u_{\text{scale}} = 0$ on $\partial\Omega$.

To obtain a valid u_{scale} throughout Ω, one can solve the following Dirichlet problem:

$$\Delta u_{\text{scale}} = 1 \qquad \text{in } \Omega \tag{4.1.51a}$$

$$u_{\text{scale}} = 0 \qquad \text{on } \partial\Omega \tag{4.1.51b}$$

In analogy to Proposition 2.6.15 this problem has a unique solution in $C^{\infty}(\overline{\Omega})$ and thus (after extension onto $\mathcal{D}$) induces a unique scale component $\alpha_{\text{scale}} = \nabla u_{\text{scale}}$ which is orthogonal to all divergence-free modes. The effect on shapes when moving along the scale-component on the manifold of shape measures is illustrated in Fig. 4.3.

4.1.3.2 Geodesic Equation on Shape Measures

We have recalled in Sect. 2.4.2 some results from [Lot08] about the manifold Meas^{∞} of absolutely continuous measures with smooth density functions. Points in $\mathcal{SP}$ and Meas can be approximated to arbitrary precision by points

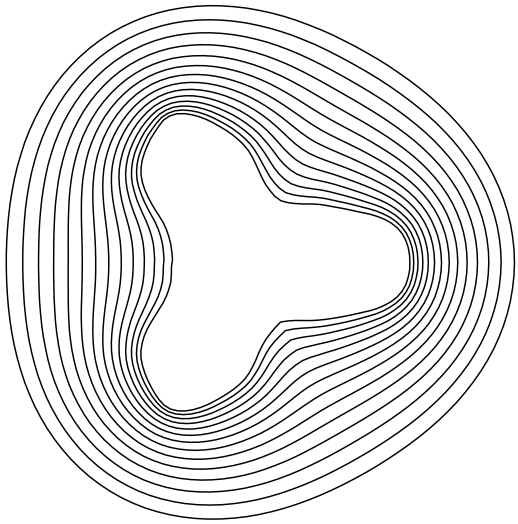


Figure 4.3: Moving along the scale-component: the contours correspond to different shape measures along a path in $\mathcal{SP}$ that is locally tangent to the scale-component (Sect. 4.1.3.1). As one moves to larger scales, small details are increasingly smoothened. Towards smaller scales they become more emphasized.

in Meas^∞ as measured by W. Thus, encouraged also by Remark 2.4.18, in this section we will pretend, expressions (2.4.34) and (2.4.35) were also valid for sufficiently smooth tangent vectors on Meas. From these we will want to find equivalent expressions on $\mathcal{SP}$. *We emphasize that this is not a rigorously justified analysis.* It is yet a worthwhile excursion as one might gain some intuition on the new metric structure of the space of shape measures, which, as shown, is a new metric on the contour manifold B.

Geodesics on Meas**.** Let us first have a look at geodesics on Meas. For an initial measure $\mu \in \mathrm{Meas}$ and an initial tangent vector $\alpha \in \mathrm{Tan}(\mu)$, the solution to the geodesic equation for regular optimal transport, (2.4.37), is given by

$$\mu_t = (\mathrm{id} + t \cdot \alpha)_\sharp \mu \,. \tag{4.1.52}$$

That is, every infinitesimal 'mass particle' in μ_t moves along a straight line, direction and velocity determined by α at $t = 0$. Once this flow-field has been chosen, no interaction between 'mass particles' is necessary, which is why the corresponding geodesic equation (2.4.37) is local in u_t.

Let now $\mu \in \mathcal{SP}$ and $\alpha \in \mathrm{STan}(\mu)$. Then from the discussion around (4.1.5) we know that to first order μ_t as in (4.1.52) has homogeneous density within its support. However, let us check the geodesic equation (2.4.37) for the potential function u_t of α_t, where α_t is the temporal evolution of $\alpha_0 = \alpha$ along the

geodesic. Applying the Laplacian to both sides (assuming for now sufficient regularity), we find at $t = 0$

$$\partial_t \, \Delta u_t|_{t=0} = -\frac{1}{2}\Delta\|\nabla u_t\|^2\Big|_{t=0} = \sum_{i,j=1}^{2} (\partial_i \partial_j u)^2 \tag{4.1.53}$$

which is the Frobenius norm of the Hessian of u. So, if the Hessian is not spatially constant, we find that μ_t will leave the subset $\mathcal{SP}$. Hence, for geodesics on $\mathcal{SP}$, 'mass particles' will not always be allowed to simply move along straight lines. They will need to make sure, that their joint density remains spatially constant. Hence, there is need for another equation of evolution, which we will now informally try to motivate.

Projection. Recall the following result from differential geometry in finite dimensions: Let M, N be Riemannian manifolds, let N be a submanifold of M. Let $x \in N \subset M$, let $a \in T_x N \subset T_x M$ and let b be a vector field on M with $b(x') \in T_{x'} N$ for all $x' \in N$. Then b can be turned into a vector field on N by restriction. Denote by $\nabla_M\big(b, (x, a)\big)$ the covariant derivative of b at point x w.r.t. direction a, and likewise for other parameters. Then

$$\nabla_N\big(b, (x, a)\big) = \mathrm{Proj}_{T_x N}\Big(\nabla_M\big(b, (x, a)\big)\Big) \tag{4.1.54}$$

where projection is w.r.t. the Riemannian inner product.

Next, let us find the projection map $\mathrm{Proj}_{\mathrm{STan}(\mu)}$. For a given shape measure $\mu \in \mathcal{SP}$, let $u \in \mathcal{D}$, so $\nabla u \in \mathrm{Tan}(\mu)$. Our goal is now to find $\hat{u} \in \mathcal{D}$ such that $\nabla \hat{u} = \mathrm{Proj}_{\mathrm{STan}(\mu)}(\nabla u)$. In that case $\nabla(u - \hat{u})$ is orthogonal to any vector in $\mathrm{STan}(\mu)$. Let $u_\perp$ be the unique solution to the following Dirichlet problem:

$$\Delta \, u_\perp = \Delta u \qquad \text{in } \Omega \tag{4.1.55a}$$

$$u_\perp = 0 \qquad \text{on } \partial\Omega \tag{4.1.55b}$$

Again, we find $u_\perp \in C^\infty(\overline{\Omega})$ and can suitably extend to $\mathcal{D}$. Recall the discussion on the scale component in Sect. 4.1.3.1 to find that $\nabla u_\perp$ is perpendicular to any divergence-free vector in $\mathrm{STan}(\mu)$ w.r.t. the inner product (2.4.31). Further, the vector $\nabla(u - u_\perp)$ lies in $\mathrm{STan}(\mu)$. Thus, all that remains to be done is, to orthogonalize w.r.t. the scale component $\nabla u_{\mathrm{scale}}$ as introduced in Sect. 4.1.3.1, which spans the only direction of $\mathrm{STan}(\mu)$ which has non-zero divergence. Thus, begin with the ansatz

$$\hat{u} = u - u_\perp + \lambda \cdot u_{\mathrm{scale}} \tag{4.1.56}$$

and determine λ such that $\nabla(u - \hat{u}) \perp \nabla u_{\mathrm{scale}}$ w.r.t. (2.4.31).

Geodesic Equation. Now we put together the pieces: Combining (2.4.36) and (4.1.54) to obtain the covariant derivative of u_t in $\mathcal{SP}$ along itself, and setting this to zero, we find:

$$0 = \mathrm{Proj}_{\mathrm{STan}(\mu_t)} \left(\nabla \left(\frac{1}{2} \|\nabla u_t\|^2 + \partial_t u_t \right) \right) \tag{4.1.57}$$

Since the projection is linear, we can separately apply it to the $\|\nabla u_t\|^2$ and to the $\partial_t u_t$ terms. Further, since $\nabla u_t \in \mathrm{STan}(\mu_t)$, we have that $\nabla \partial_t u_t \in \mathrm{STan}(\mu_t)$ since also the divergence of the temporal derivative must be spatially constant. Hence, the projection of the second term is redundant and we can write:

$$0 = \mathrm{Proj}_{\mathrm{STan}(\mu_t)} \left(\nabla \frac{1}{2} \|\nabla u_t\|^2 \right) + \nabla \partial_t u_t \tag{4.1.58}$$

Since $\mathrm{Proj}_{\mathrm{STan}(\mu)}$ is a non-local operation, the new geodesic equation is non-local, in contrast to (2.4.37). This non-locality is necessary to keep the density of μ_t spatially constant along the path.

4.1.3.3 Geodesics on the Manifold of Shape Measures

We now discuss some particular solutions to (4.1.58). Let $\mu \in \mathcal{SP}$ be some shape measure and the initial tangent vector $\alpha_0 = \alpha_{\mathrm{trans}} = v$ be a spatially constant translation mode, as discussed in Sect. 4.1.3.1. The geodesic in Meas through μ, tangent to α_0 is given by $\mu_t = (\mathrm{id} + t \cdot v)_\sharp \mu$. This is the translation of μ by the vector $t \cdot v$. Obviously this is a path in $\mathcal{SP}$. Since $\mathcal{SP} \subset \mathrm{Meas}$, it must therefore also be a geodesic in $\mathcal{SP}$.

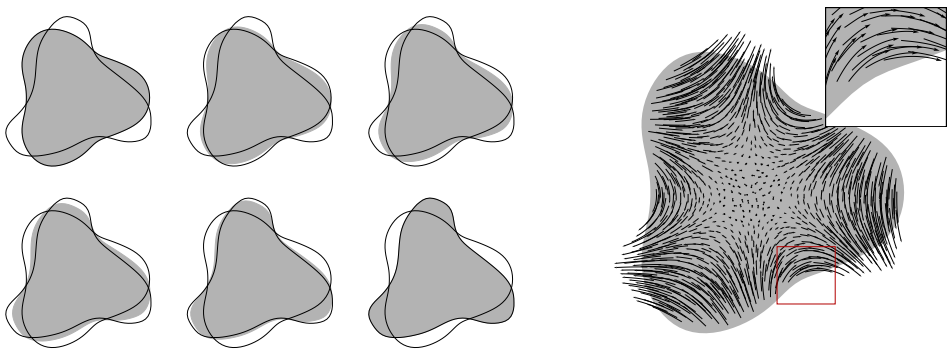
This can be verified explicitly: we have $\alpha_0 = \nabla u_0$ for $u_0 = \langle x, v \rangle_{\mathbb{R}^2}$ and consequently find $\nabla \|\nabla u_0\|^2 = 0 \in \mathrm{STan}(\mu)$. Hence, the projection will change nothing and we find $\partial_t \alpha_t|_{t=0} = 0$. One can thus see that $\alpha_t = \alpha_0$ is in fact a solution to (4.1.58).

Consider further the initial tangent vector $\alpha_0(x) = x$. This corresponds to resizing the original shape. A possible potential function is given by $u(x) = \|x\|^2/2$. One can check that the induced optimal transport geodesic $\mu_t = (\mathrm{id} + t \cdot \alpha_0)_\sharp \mu$ lies within $\mathcal{SP}$, hence by the same reasoning as with the translations, it must therefore also be a geodesic on the shape measures.

A numerical solution to the geodesic equation where the projection is important is illustrated in Fig. 4.4.

4.1.4 Conclusion

In this section the set of shape measures as a 'submanifold' of $\mathcal{W}_2(\mathbb{R}^2)$ was introduced and it was established that this set, equipped with a suitable topology, is in fact diffeomorphic to the manifold B of closed contours modulo reparametrization. The metric induced on this manifold by optimal transport



(a) (b)

Figure 4.4: Numerically computed geodesic on $\mathcal{SP}$. (a) Top left to bottom right: geodesic between two shape measures (initial and final shape indicated by contours for orientation). Unlike the natural linear structure on measures (cf. Fig. 1.1) this gives a meaningful interpolation between the two shapes. (b) Trajectories of 'mass particles' in the measure (with close-up). Unlike in conventional optimal transport, particles do not all travel on straight lines. This is necessary such that the intermediate measure will always represent a shape.

was investigated and a candidate for the geodesic equation was discussed. A more theoretical study of the geodesic equation of this metric is a natural open question.

The diffeomorphism established that modelling shapes via shape measures is equivalent to modelling on contour manifolds and thus provides a way of combining the advantages of the respective representations (cf. Sect. 1.2.2). A shape prior based on this equivalence is presented in the next Section.

4.2 Wasserstein Modes

Based on the description of *shape measures* as a manifold we will now construct another shape prior functional for object segmentation.

Instead of using an inherently isometry invariant regularizer as in Sect. 3.1, we start in Sect. 4.2.1.1 with a basic functional, where the shape prior is given by the 2-Wasserstein distance of the current segmentation measure to a template. Among other limitations this is of course by no means invariant under any kind of geometric transformation. To overcome these constraints, additional degrees of freedom are added, to manipulate the template measure in a tangent space approximation to the previously introduced manifold of shape measures (Sect. 4.2.1.2). These degrees of freedom can then be used to obtain (approximate) invariance under Euclidean isometries and scale transformations and to model non-isometric deformations that one can learn from training data

(Sects. 4.2.1.3 and 4.2.1.4). So, as opposed to the prior discussed in Sect. 3.1 this is an extrinsic way to obtain isometry invariance (cf. discussion in Sect. 2.5.2.3). Some considerations on modelling the image background to improve results are given in Sect. 4.2.1.5.

The overall resulting functional unsurprisingly turns out to be non-convex, but non-convexity is constrained to a low-dimensional variable. A locally optimal alternating optimization method, a globally optimal branch-and-bound scheme and a fast graph-cut relaxation are discussed in Sect. 4.2.2. Some numerical examples are given in 4.2.3.

4.2.1 Shape Regularization with Wasserstein Distances

4.2.1.1 Setup and Basic Functional

The setup will be analogous to Sect. 3.1.1.1: Let Y describe the image domain in which we want to locate and match the object. Let $\mathcal{L}_Y$ be the Lebesgue measure on Y. A segmentation proposal will be encoded by some measure $\nu \in \mathrm{SegMeas}(Y, M)$ (see (2.1.8)).

As in Sect. 3.1.3.2, to handle local image data we introduce a suitable *feature space* $\mathcal{F}$. Depending on the image this may be the corresponding color space. It may however also be a more elaborate space spanned by small image patches or local filter responses. We then assume that any point $y \in Y$ is equipped with some $f_y \in \mathcal{F}$ which we refer to as the *observed feature*. We can thus consider every pixel to be a point in the enhanced space $Y \times \mathcal{F}$ with coordinates (y, f_y).

For regularization with optimal transport we again need to provide a *template*. Let X be a set whose geometry will model the shape of the sought after object and let μ be again the Lebesgue measure thereon. Recall that the constant M in the feasible set for ν will be the mass of μ:

$$M = \mu(X) \tag{4.2.1}$$

Additionally, we describe the *appearance* of the template by associating to all elements $x \in X$ corresponding $f_x \in \mathcal{F}$, the *expected features*.

We assume that both the template X and the image domain Y are embedded into $\mathbb{R}^2$. The squared Euclidean distance $\|x - y\|^2$ for $x \in X$ and $y \in Y$ then provides a geometric matching cost for points:

$$c_{\mathrm{geo}}(x, y) = \|x - y\|^2 \tag{4.2.2}$$

Moreover, we pick some function $c_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ which models the matching cost on the feature space. Possible choices for $c_{\mathcal{F}}$ are for example a (squared) metric, or a Bayesian log-likelihood $c_{\mathcal{F}}(f_x, f_y)$ for observing a noisy feature f_y when expecting feature f_x.

Combining this, we can construct a functional for rating the plausibility of a segmentation proposal $\nu \in \mathrm{SegMeas}(Y, M)$:

$$E(\nu) = \frac{1}{2} \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} \Big(c_{\mathrm{geo}}(x, y) + c_{\mathcal{F}}(f_x, f_y) \Big) \, d\pi(x, y) + G(\nu) \tag{4.2.3}$$

The first term is the minimal matching cost between the segmentation region and the template via optimal transport with a cost function that combines the geometry and appearance. The second term can contain other typical components of a segmentation functional, for example a local boundary regularizer (cf. Sect. 2.1). The functional is illustrated in Fig. 4.5a.

Remark 4.2.1 (Generality of functional)**.** This setup provides the same generality as functional Sect. 3.1.1.1, see Remark 3.1.1. The introduction of a general feature space allows to handle all sorts of features associated to elements in X and Y, cf. Sect. 3.1.3.2.

Limitations of the Basic Functional. Functional (4.2.3) has three major shortcomings for the application of object segmentation and shape matching, related to the choice of the embedding $X \to \mathbb{R}^2$:

(i) The location and orientation of the sought-after object are often unknown beforehand. Hence, a segmentation method should be invariant under Euclidean isometries, which is clearly violated by picking an arbitrary embedding $X \to \mathbb{R}^2$. If μ and ν were fixed measures in $\mathrm{Meas}(\mathbb{R}^2)$ with equal mass, then the optimal coupling for $W_2(\mu, \nu)$ would be invariant under translation (up to an adjustment of the coordinates according to the translation, of course). However, since in this application ν is not fixed and the cost function is enhanced by an appearance component, this quasi-invariance cannot be exploited. Also, there is no similar invariance w.r.t. rotation.

(ii) Any non-isometric deformation between template foreground and the object will be uniformly penalized by the geometric part of the corresponding optimal transport cost. No information on more or less common deformations can be encoded.

(iii) Since the mass M of μ, related to the size of the template X, equals the mass of ν, this determines the size of the foreground object in Y. Hence, the presented functionals imply that one must know the scale of the sought-after object beforehand. This is not possible in all applications.

In the next sections we will discuss how to overcome these obstacles. By making the embedding $X \to \mathbb{R}^2$ flexible, the resulting functionals become fit for isometry invariance, can handle prior information on more or less common non-isometric deformations and can dynamically adjust the object scale.

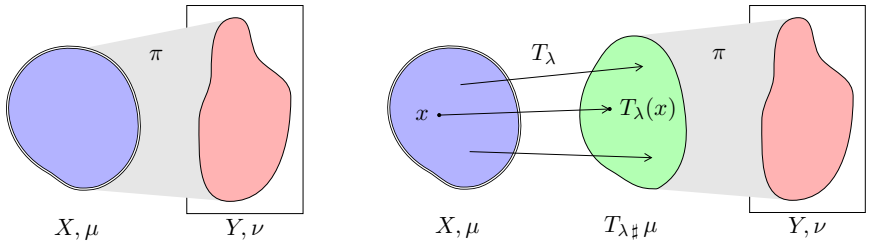


Figure 4.5: Illustration of functionals $E_0(\nu)$, eq. (4.2.3), and $E_1(\lambda, \nu)$, eq. (4.2.5): (a) The segmentation in Y is described by measure ν which is regularized by the Wasserstein distance to a template measure μ, living on X. This simple approach introduces strong bias, depending on the relative location of X and Y, and lacks the ability to explicitly model typical object deformations. (b) In the enhanced functional the template measure μ is deformed by the map T_λ, resulting in the push-forward $T_{\lambda\sharp}\mu$. The segmentation ν is then regularized by its Wasserstein distance to $T_{\lambda\sharp}\mu$. The corresponding optimal coupling π gives a registration between the foreground part of the image and the deformed template.

4.2.1.2 Wasserstein Modes

To overcome the limitations listed in Sect. 4.2.1.1 we will allow X to move and be deformed within $\mathbb{R}^2$. We choose the following family of embeddings:

$$T_\lambda : X \to \mathbb{R}^2, \qquad T_\lambda(x) = x + \sum_{i=1}^{n} \lambda_i \cdot t_i(x) \tag{4.2.4}$$

The transformation is parametrized by the coefficients $\lambda \in \mathbb{R}^n$. This linear decomposition will allow enough flexibility for modelling while keeping the resulting functionals simple. We refer to the basis maps $\{t_i\}_{i=1}^n$ as *modes*. Including the coefficients λ as degrees of freedom into (4.2.3) yields:

$$E(\lambda, \nu) = \frac{1}{2} \inf_{\pi \in \Pi(\mu,\nu)} \int_{X\times Y} \Big(c_{\text{geo}}\big(T_\lambda(x), y\big) + c_{\mathcal{F}}(f_x, f_y) \Big)\, d\pi(x, y) + F(\lambda) + G(\nu) \tag{4.2.5}$$

The function F can be used to introduce statistical knowledge on the distribution of the coefficients λ. The enhanced functional is illustrated in Fig. 4.5b.

Functional (4.2.5) is generally non-convex. For fixed λ it is convex in ν. For fixed ν and a fixed coupling π in the optimal transport term it is convex in λ if transformations are of the form (4.2.4) and F is convex. Joint non-convexity does not come as a surprise. It is in fact easy to see that a meaningful isometry invariant segmentation functional with explicitly modelled transformations is bound to be non-convex (Fig. 4.6).

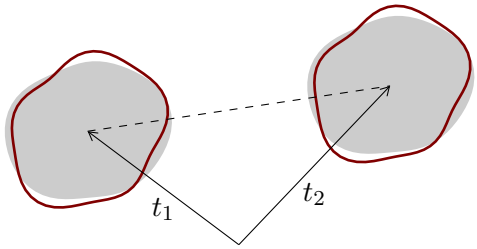


Figure 4.6: Explicit transformation variables and non-convexity. Gray shading indicates 'foreground features'. Placing the template (red contour) at t_1 or t_2 yields equally good hypotheses. Were the prior functional convex in the translation variable, any point along the line $(1-\lambda)\cdot t_1 + \lambda \cdot t_2$ for $\lambda \in [0,1]$ would yield an at least equally good proposal, which is clearly unreasonable.

Remark 4.2.2 (Eliminating ν)**.** For optimization of (4.2.5) assume we first eliminate the high-dimensional variable ν through minimization (which corresponds to solving a convex problem). One is then left with:

$$E_1(\lambda) = \inf_{\nu \in \mathrm{SegMeas}(Y,M)} E(\lambda, \nu) \tag{4.2.6}$$

This is in general non-convex, but the dimensionality of λ is typically very low (of the order of 10). We can thus still hope to find globally optimal solutions by means of non-convex optimization. We will present a corresponding branch and bound scheme in Sect. 4.2.2.2.

Remark 4.2.3 (Modelling transformations in feature space)**.** When the feature space $\mathcal{F}$ has an appropriate linear structure a natural generalization of (4.2.4) is to not only model geometric transformations of X but also of the expected features f_x. In analogy to (4.2.4) consider

$$\hat{T}_\lambda : X \to \mathbb{R}^2 \times \mathcal{F}, \qquad \hat{T}_\lambda(x) = (x, f_x) + \sum_{i=1}^{n} \lambda_i \cdot \hat{t}_i(x) \tag{4.2.7}$$

where $\hat{T}_0(x) = (x, f_x)$ returns the original position and expected feature of a point. The modes $\hat{t}_i : X \to \mathbb{R}^2 \times \mathcal{F}$ can then be used to alter both the geometry of X as well as its appearance.

This will be useful when the appearance of a sought-after object is known to be subject to variations or when a feature is affected by geometric transformations: for example the expected response to an oriented local filter will need to be changed when the object is rotated. The corresponding generalized functional is

$$E_{\mathcal{F}}(\lambda, \nu) = \frac{1}{2} \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} \hat{c}\big(\hat{T}_\lambda(x), (y, f_y)\big)\, d\pi(x,y) + F(\lambda) + G(\nu) \tag{4.2.8}$$

with

$$\hat{c} : (\mathbb{R}^2 \times \mathcal{F})^2 \to \mathbb{R}, \qquad \hat{c}\big((x', f'_x), (y, f_y)\big) = c_{\text{geo}}(x, y) + c_{\mathcal{F}}(f'_x, f_y)\,. \tag{4.2.9}$$

Further study of this generalization is however beyond the scope of this thesis.

As pointed out, for a meaningful template μ should be the Lebesgue measure on X with constant density 1. We need to take into account how transforming X through T_λ effectively influences μ. Applying T_λ to X will result in the deformed measure $T_{\lambda\sharp}\,\mu$ on $T_\lambda(X)$. We have established earlier (Sect. 4.1.1.1) that for keeping the density of $T_{\lambda\sharp}\,\mu$ constant to first order we must have $\operatorname{div} t_i = \text{const}$. Moreover, non-zero divergence will lead to a density which is not 1. Hence, $T_{\lambda\sharp}\,\mu$ must be rescaled accordingly, which will change its total mass and thus influence the corresponding feasible set $\text{SegMeas}(Y, M)$ for ν. This will require some additional care during optimization. All constant-divergence modes can be decomposed into zero-divergence modes plus an additional 'scale mode' (see Sect. 4.2.1.3). We will thus see to it that all but one mode will have zero-divergence and handle the scale mode with particular care (Sect. 4.2.2).

We can thus see that introducing suitable Wasserstein modes corresponds to working in a *tangent space approximation* to the manifold of shape measures where the original template μ is the footpoint.

4.2.1.3 Geometric Invariance

The framework provided by transformations (4.2.4) and functional (4.2.5) allows to introduce geometric invariance into the segmentation / matching approach. In this section we will consider translations, (approximate) rotations and scale transformations. Scale transformations will play a special role as they change the mass of the template.

The transformations will be modelled with the generators of the corresponding (local) Lie group in $\mathbb{R}^2$. Likewise invariance w.r.t. transformation Lie groups could be introduced into matching functionals on other manifolds.

Translation and Rotation. If one chooses modes

$$t_{\text{t1}}(x) = (1, 0)^\top, \qquad t_{\text{t2}}(x) = (0, 1)^\top \tag{4.2.10}$$

the corresponding coefficients $\lambda_{\text{t1,t2}}$ parametrize translations of the template. Further, let $R(\phi)$ be the 2-dimensional rotation matrix by angle ϕ. Then the mode

$$t_{\text{r}}(x) = \frac{d}{d\phi} R(\phi)\Big|_{\phi=0} x = (-x_2, x_1)^\top \tag{4.2.11}$$

will approximately rotate the template. This first order expansion works satisfactory for angles up to about $\pm 30°$.

For explicit invariance under translations and rotations the modelling function F in (4.2.5) should be constant w.r.t. the coefficients $\lambda_{\mathrm{t}1}, \lambda_{\mathrm{t}2}$ and λ_{r}.

Scale. The size of X and μ determines the size of the object within the image. In many applications the scale is not known beforehand, thus dynamical resizing of the template during the search is desirable. With slight extensions the framework of transformations can be employed to introduce as a scale-mode into the approach. Let

$$t_{\mathrm{s}}(x) = x\,. \tag{4.2.12}$$

By the change of variable formula the density of $T_{\lambda\sharp}\,\mu$ is given by

$$\mathrm{dens}\left(T_{\lambda\sharp}\,\mu\right)\left(T_\lambda(x)\right) = \mathrm{dens}(\mu)(x) \cdot \left(\det J_{T_\lambda}(x)\right)^{-1}. \tag{4.2.13}$$

By plugging in the scale mode t_{s} and ignoring other modes, which due to zero divergence do not contribute to first order, we find in 2 dimensions:

$$= (1+\lambda_{\mathrm{s}})^{-2} \tag{4.2.14}$$

Thus, introducing a scale mode into (4.2.5) yields

$$\begin{aligned} E_{\mathrm{s}}(\lambda,\nu) = & \frac{1}{2\,(1+\lambda_{\mathrm{s}})^2} \inf_{\pi\in\Pi\left((1+\lambda_{\mathrm{s}})^2\cdot\mu,\nu\right)} \\ & \int_{X\times Y} \left(c_{\mathrm{geo}}\big(T_\lambda(x),y\big) + c_{\mathcal{F}}(f_x,f_y)\right)\, d\pi(x,y) \\ & + F(\lambda) + G(\nu) \end{aligned} \tag{4.2.15}$$

where we have scaled μ by the appropriate factor in the feasible set for π and we have normalized the first term by a factor of $(1+\lambda_{\mathrm{s}})^{-2}$ to make the term scale invariant. Depending on whether scale invariance is desired the terms $F(\lambda)$ and $G(\nu)$ may need to be rescaled appropriately, too. The feasible set for ν in E_{s} is $\mathrm{SegMeas}\left(Y, (1+\lambda_{\mathrm{s}})^2 \cdot M\right)$.

While the modes for translation and rotation leave the area of the template unaltered, statistical deformation modes that we learn from sample data will in general have non-zero divergence. Handling changes in mass will require some extra care during optimization. Therefore we will decompose such modes into a divergence-free part and a contribution of the scale-component.

4.2.1.4 Statistical Deformations

In this section we describe how modes $t_{\mathrm{d}i}$ can be learned that model class-typical shape variations from a set of training samples. These modes can then be used

to allow X to be deformed in the learned way, to prefer known deformations over unknown deformations during the segmentation process.

Let $\{\mu_i\}_{i=1}^m$ be a set of training segmentations, given as indicator-measures: the support of μ_i marks the foreground of the corresponding segmentation. Assume that all μ_i have the same mass (we normalize accordingly, see prev. Section). We arbitrarily choose μ_1 to be the reference segmentation and compute the optimal transport couplings $\{\pi_{1,i}\}_{i=1}^k$ between the reference and the other segmentations, *optimized over rotation*. It is easy to see that the relative translation with smallest cost is the one where the centers of mass coincide [WSB+12]. Note that the optimal coupling $\pi_{1,1}$ simply transports mass from all pixels onto themselves. By virtue of Theorem 2.4.8 in a continuous setting an optimal transport map would exist. Granted sufficient numerical resolution, the discrete optimal coupling is almost deterministic and one can extract an approximate assignment field as in Sect. 2.4.4. The relative transportation maps that underlie the optimal couplings $\pi_{1,i}$ are then elements of the tangent space of the manifold of measures at μ_1. As in [WSB+12], we can perform a classical principal component analysis (PCA) on the set of tangent vectors to obtain the *mean deformation* t_m, a set of *principal deformation modes* $\{t_{\mathrm{d}i}\}_i$ and corresponding parameters Σ_i that give the data variance along each mode. From this we can construct a Gaussian prior on the non-isometric deformation by choosing

$$F(\lambda) = \frac{1}{2}\lambda_\mathrm{d}^\top \Sigma^{-1}\, \lambda_\mathrm{d} \tag{4.2.16}$$

where $\Sigma^{-1} = \mathrm{diag}(\{\Sigma_i^{-1}\}_i)$ and λ_d is the sub-vector of λ that describes deformations.

The choice of a reference template is arbitrary and rather heuristic. But computing the barycenter on the manifold of shape measures is yet an unsolved problem. Even in unrestricted (i.e. no constant density constraint) Wasserstein spaces this is a difficult problem [AC11]. In [WSB+12] simply the L^2-mean of the densities was taken, which is however not applicable here due to the constraint of binary densities. We consider the presented choice sufficient for a proof of concept. The arbitrary choice of the reference measure is somewhat alleviated by the PCA, where the mean of all observed transportation maps is determined. During segmentation the template μ will then be generated by the reference template μ_1, shifted by this mean.

It should be noted that through the diffeomorphism between shape measures and the manifold of closed contours, one can in principle 'outsource' the learning problem to the contour representation and lift the obtained mean and deformation modes, as discussed in Defs. 4.1.7 and 4.1.9. One can then draw from the theory on contours referred to in Sect. 1.2.2. However, we consider this to be beyond the scope of this thesis.

4.2.1.5 Background Modelling

The feasible set for (4.2.5) is $\mathrm{SegMeas}(Y, M)$. A feasible ν will be matched to the template measure μ according to geometry and appearance. Therefore the optimal ν will select the mass of $\mathcal{L}_Y$ on Y that fits best. This will be the object / the foreground. The rest of the mass of $\mathcal{L}_Y$, which is ignored, will be considered to be 'assigned to background'. This 'selection of the best matching mass' is a typical approach in the literature [CG99, RTG00]. This assumption is however problematic in some cases: Consider looking for an object of particular shape and particular appearance. In the image let there be another, much bigger object with same appearance but different shape. Then any 'immersion' of the small template into the large second object will yield a very low value for functional (4.2.5). But of course these are not the configurations that we are looking for.

This issue can be solved by explicitly modelling the background around the foreground object with the template. From now on, refer to the original template that describes the foreground by X_F. Then add to the template an additional layer X_B of thickness $r_B > 0$ around the foreground region (Fig. 4.7a). μ_F will be the Lebesgue measure on X_F, μ_B on X_B and $\mu_F + \mu_B$ on $X_F \cup X_B$. Elements $x \in X_B$ will also be associated with features that model the expected appearance of the image background. One optimizes functional (4.2.5) with the enhanced template. The optimal image foreground will then be the region that the mass from μ_F is transported to.

'Immersion' of the template into another, larger object will now inflict a large cost, as the background part X_B will be a bad match for the surrounding image region (Fig. 4.7b).

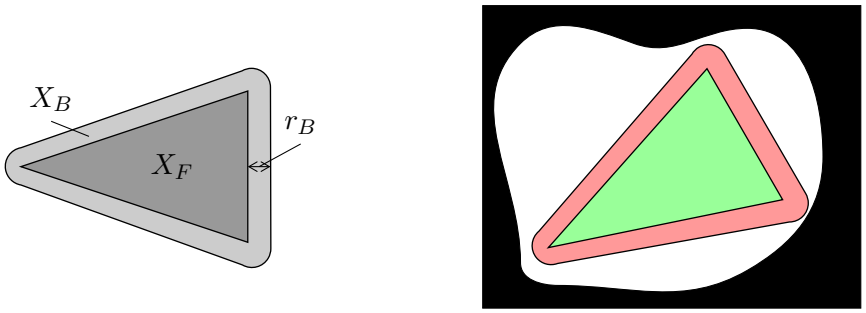
4.2.2 Optimization

4.2.2.1 Alternating Optimization

Functional (4.2.5) is generally non-convex. It is convex in ν for fixed λ and it is convex in λ under suitable conditions (see Sect. 4.2.1.2). Based on this, an alternating optimization scheme is conceivable for divergence-free modes. This has also been proposed in [CG99, Sect. 3.2.1]. We require the following reformulation of (4.2.6):

Remark 4.2.4 (Coupling reformulation)**.** Computing (4.2.6) involves a nested optimization problem over $\nu \in \mathrm{SegMeas}(Y, M)$ and then $\pi \in \Pi(\mu, \nu)$. Given a coupling $\pi \in \Pi(\mu, \nu)$ the marginal ν can be reconstructed via projection: $\nu = \mathrm{pr}_{Y\sharp}\pi$. This allows to reformulate optimization of (4.2.6) directly in terms of couplings. Let

$$\hat{E}(\lambda, \pi) = \frac{1}{2} \int_{X \times Y} \Big(c_{\mathrm{geo}}\big(T_\lambda(x), y\big) + c_{\mathcal{F}}(f_x, f_y) \Big) \, d\pi(x, y) + F(\lambda) + G(\mathrm{pr}_{Y\sharp}\pi) \tag{4.2.17}$$



(a) (b)

Figure 4.7: Explicit background modelling: (a) around the original template X_F that models the foreground, a layer X_B of thickness $r_B > 0$ is added, that models the background surrounding the object. (b) let white indicate features typical for foreground. Without explicit background modelling, 'immersing' the template into the white region yields a low score, while it is actually not a good match from the viewpoint of shape. Through adding X_B such faulty detections can be suppressed.

and let the feasible set for π in $\hat{E}$ be

$$\begin{aligned}\mathrm{SegCoupl}(Y,\mu) &= \bigcup_{\nu\in\mathrm{SegMeas}(Y,M)} \Pi(\mu,\nu)\\ &= \left\{\pi\in\mathrm{Meas}(X\times Y)\,:\ \mathrm{pr}_{X\sharp}\pi=\mu\wedge\mathrm{pr}_{Y\sharp}\pi\leq\mathcal{L}_Y\right\}. \quad (4.2.18)\end{aligned}$$

Then for fixed λ one has by construction

$$\inf_{\nu\in\mathrm{SegMeas}(Y,M)} E(\lambda,\nu) = \inf_{\pi\in\mathrm{SegCoupl}(Y,\mu)} \hat{E}(\lambda,\pi) \quad (4.2.19)$$

and for any optimizer π^* of $\hat{E}$ the marginal $\mathrm{pr}_{Y\sharp}\pi^*$ is an optimizer of E.

Functional $\hat{E}(\lambda,\pi)$ is separately convex in λ and π for transformations of the form (4.2.4) and convex F. For some initial λ^1 consider the following sequence:

$$\pi^{k+1} \in \mathrm{argmin}_{\pi\in\mathrm{SegCoupl}(Y,\mu)}\, \hat{E}(\lambda^k,\pi) \quad (4.2.20a)$$
$$\lambda^{k+1} \in \mathrm{argmin}_{\lambda\in\mathbb{R}^n}\, \hat{E}(\lambda,\pi^{k+1}) \quad (4.2.20b)$$

Proposition 4.2.5. *The sequence of energies* $\hat{E}(\pi^k,\lambda^k) \to \hat{E}(\pi^{k+1},\lambda^k) \to \hat{E}(\pi^{k+1},\lambda^{k+1}) \to \ldots$ *is non-increasing and converges.*

Proof. Since π^k is a feasible point for computing π^{k+1}, one has $\hat{E}(\pi^{k+1},\lambda^k) \leq \hat{E}(\pi^k,\lambda^k)$. Likewise λ^k is feasible when determining λ^{k+1} so $\hat{E}(\nu^{k+1},\lambda^{k+1}) \leq \hat{E}(\nu^{k+1},\lambda^k)$. Hence, the sequence of energies is non-increasing. As $\hat{E}$ is bounded from below, the sequence must converge. □

Unfortunately this cannot be extended to modes with non-zero divergence: as changing λ_s changes the feasible set $\mathrm{SegCoupl}\left(Y, (1+\lambda_s)^2 \cdot \mu\right)$ for π. Thus π^k need not be feasible for the problem that determines π^{k+1} and the sequence of energies created may be increasing. We will provide a workaround for this in the next section (Remark 4.2.10).

The alternating scheme (4.2.20) is fast and tends to converge after few iterations. But obviously it need not converge to a global optimum and the result depends on the initialization λ^1. Therefore, similar to contour based segmentation functionals it must be applied with care. In practice (cf. Sect. 4.2.3) application to 'large' transformations, e.g. translations and rotations, works only if a good initial guess is available. On the other hand it achieves decent results on smaller transformations, as most statistically learned deformations are.

4.2.2.2 Globally Optimal Branch and Bound

For handling large displacement transformations, one needs a global optimization scheme. As discussed in Remark 4.2.2, for fixed λ we can eliminate ν by a separate convex optimization. One obtains (4.2.6):

$$\begin{aligned} E_1(\lambda) &= \inf_{\nu \in \mathrm{SegMeas}(Y,M)} E(\lambda, \nu) \\ &= \inf_{\nu \in \mathrm{SegMeas}(Y,M)} \frac{1}{2} \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} \left(c_{\mathrm{geo}}\big(T_\lambda(x), y\big) + c_{\mathcal{F}}(f_x, f_y) \right) d\pi(x,y) \\ &\qquad + F(\lambda) + G(\nu) \end{aligned} \tag{4.2.21}$$

This function is in general non-convex but low dimensional. We thus can hope to find a non-convex global optimization scheme.

Given Remark 4.2.4 $E_1(\lambda)$ can be written as

$$\begin{aligned} E_1(\lambda) = \inf_{\pi \in \mathrm{SegCoupl}(Y,\mu)} \frac{1}{2} \int_{X \times Y} \left(c_{\mathrm{geo}}\big(T_\lambda(x), y\big) + c_{\mathcal{F}}(f_x, f_y) \right) d\pi(x,y) + \\ F(\lambda) + G(\mathrm{pr}_{Y\sharp}\pi)\,. \end{aligned} \tag{4.2.22}$$

If G is zero, then by inserting suitable dummy nodes, computing $E_1(\lambda)$ can be written as an optimal transport problem for which efficient solvers are available.

In this section we will consider a hierarchical branch and bound approach. We will compute lower bounds for E_1 on whole intervals of λ-configurations for successively refined intervals. Let $\Lambda \subset \mathbb{R}^n$ be a set of λ-values. We assume for now all modes have zero divergence. For such subsets define

$$\begin{aligned} E_2(\Lambda) = \inf_{\pi \in \mathrm{SegCoupl}(Y,\mu)} \frac{1}{2} \int_{X \times Y} \left(\left(\inf_{\lambda \in \Lambda} c_{\mathrm{geo}}\big(T_\lambda(x), y\big) \right) + c_{\mathcal{F}}(f_x, f_y) \right) d\pi(x,y) \\ + \inf_{\lambda \in \Lambda} F(\lambda) + G(\mathrm{pr}_{Y\sharp}\pi) \end{aligned} \tag{4.2.23}$$

where have again merged optimizations as above. This is an adaptive convex relaxation of $E_1(\lambda)$ over Λ. The relaxation becomes tighter as the set becomes smaller. For application in a branch and bound scheme following properties are required:

Proposition 4.2.6 ([SS13d, Prop. 1])**.** *The functional E_2 has the following properties:*

(i) $E_2(\Lambda) \leq E_1(\lambda)\ \forall\ \lambda \in \Lambda$,

(ii) $\lim_{\Lambda \to \{\lambda_0\}} E_2(\Lambda) = E_1(\lambda_0)$,

(iii) $\Lambda_1 \subset \Lambda_2 \Rightarrow E_2(\Lambda_1) \geq E_2(\Lambda_2)$.

Proof. Property (i): For any $\lambda \in \Lambda$ obviously

$$\inf_{\lambda' \in \Lambda} c_{\text{geo}}\big(T_{\lambda'}(x), y\big) \leq c_{\text{geo}}\big(T_{\lambda}(x), y\big) \qquad \text{and} \qquad \inf_{\lambda' \in \Lambda} F(\lambda') \leq F(\lambda)\,. \quad (4.2.24)$$

So for any fixed $\pi \in \text{SegCoupl}(Y, \mu)$ (fixing for now the minimization in (4.2.22, 4.2.23)) have $E_2(\Lambda) \leq E_1(\lambda)$. Consequently this inequality will also hold after minimization of π.

For the limit property (ii) note that the functions $c_{\text{geo}}\big(T_{\lambda}(x), y\big)$ and $F(\lambda)$ are continuous functions of λ. Hence, when $\Lambda \to \{\lambda_0\}$ all involved minimizations will converge towards the respective function values at λ_0 and E_2 converges as desired.

For the hierarchical bound property (iii) note that for fixed π in (4.2.23) minimization over the larger set λ_2 will never yield the larger result for all occurrences of λ. This relation will then also hold after minimization. □

`refine`**(L):**

(1) Find the element $(\Lambda_{i^*}, b_{i^*}) \in L$ with the smallest lower bound b_{i^*}.

(2) Let $\texttt{subdiv}(\Lambda) = \{\Lambda_{i^*,j}\}_j$ be a subdivision of the interval Λ_{i^*} into smaller intervals.

(3) Compute $b_{i^*,j} = E_2(\Lambda_{i^*,j})$ for all $\Lambda_{i^*,j} \in \texttt{subdiv}(\Lambda)$.

(4) Remove (Λ_{i^*}, b_{i^*}) from L and add $\{(\Lambda_{i^*,j}, b_{i^*,j})\}_j$ for $\Lambda_{i^*,j} \in \texttt{subdiv}(\Lambda)$.

Algorithm 1: Refinement procedure for hierarchical branch & bound.

With the aid of E_2 one can then construct a branch and bound scheme for optimization of E_1. Let $L = \{(\Lambda_i, b_i)\}_i$ be a finite list of λ-parameter intervals Λ_i and lower bounds b_i on E_1 on these respective intervals. For such a list

consider the refinement procedure given in Algorithm 1. It allows the following statement:

Proposition 4.2.7 ([SS13d, Prop. 2])**.** *Let L be a list of finite length. Let the subdivision in* `refine` *be such that any interval will be split into a finite number of smaller intervals, and that any two points will eventually be separated by successive subdivision.* $\texttt{subdiv}(\{\lambda_0\}) = \{\{\lambda_0\}\}$*. Then repeated application of* `refine` *to the list L will generate an adaptive piecewise underestimator of E_1 throughout the union of the intervals Λ appearing in L. The sequence of smallest lower bounds will converge to the global minimum of E_1.*

Proof. Obviously the sequence of smallest lower bounds is non-decreasing and never greater than the minimum of E_1 throughout the considered region (see Proposition 4.2.6 (iii) and (i)). So it must converge to a value which is at most this minimum. Assume that $\{\Lambda_i\}_i$ is a sequence with $\Lambda_{i+1} \in \texttt{subdiv}(\Lambda_i)$ such that $E_2(\Lambda_i)$ is a subsequence of the smallest lowest bounds of L (there must be such a sequence since L is finite). Since `subdiv` will eventually separate any two distinct points, this sequence must converge to a singleton $\{\lambda_0\}$ and the corresponding subsequence of smallest lowest bounds converges to $E_2(\{\lambda_0\}) = E_1(\lambda_0)$. Since the sequence of smallest lowest bounds converges, and the limit is at most the minimum of E_1, $E_1(\lambda_0)$ must be the minimum. □

In practice we start with a coarse grid of hypercubes covering the space of reasonable λ-parameters (e.g. translation throughout the image, rotation within bounds where the approximation is valid and the deformation-coefficients in ranges according to the statistical model) and the respective E_2-bounds. Any hypercube with the smallest bound will then be subdivided into equally sized smaller hypercubes, leading to an adaptive 2^n-tree cover on the considered parameter range.

The refinement is stopped, when the interval with the lowest bound has edge lengths that correspond to an uncertainty in $T_\lambda(x)$ which is in the range of the discretization of X and Y. Further refinement would only reveal structure determined by rasterization effects.

Remark 4.2.8 (Combining hierarchical and alternating optimization)**.** The optimum of E_1 w.r.t. modes that have large displacements (such as translation and rotation) tends to be rather distinct, i.e. there is a small, distinct basin around the optimal position. The hierarchical optimization scheme then tends to be working rather efficiently.

On the other hand, modes that model smaller, local displacements (often those learned from training samples), often have broad, shallow basins around the optimal value. The branch and bound scheme can then take longer to converge.

Therefore it suggests itself to combine the two optimization schemes: the hierarchical approach is used to determine a good initial guess for translation,

rotation and a coarse estimate for the smaller modes. For this the alternating scheme is not applicable due to the non-convexity. But once the broad basin around the global optimum is located, the branch and bound scheme may become inefficient. Conversely, using the estimate of the hierarchical scheme as initialization, we can then expect that the alternating method will give reasonable results.

Scale Mode. In the presence of a scale mode one can define $E_{\mathrm{s},1}$ and $E_{\mathrm{s},2}$ equivalent to E_1 and E_2 with slight adaptations.

$$\begin{aligned} E_{\mathrm{s},1}(\lambda) &= \inf_{\nu \in \mathrm{SegMeas}\left(Y,(1+\lambda_\mathrm{s})^2 \cdot M\right)} E_\mathrm{s}(\lambda, \nu) \\ &= \inf_{\pi \in \mathrm{SegCoupl}\left(Y,(1+\lambda_\mathrm{s})^2 \cdot \mu\right)} \frac{1}{2\,(1+\lambda_\mathrm{s})^2} \\ &\qquad \int_{X \times Y} \left(c_\mathrm{geo}\big(T_\lambda(x), y\big) + c_\mathcal{F}(f_x, f_y) \right) d\pi(x,y) + F(\lambda) + G(\mathrm{pr}_{Y\sharp}\pi) \end{aligned} \tag{4.2.25}$$

where in the second line we have merged the nested optimization over ν and π, see Remark 4.2.4. To obtain $E_{\mathrm{s},2}(\Lambda)$ all occurrences of λ will again be replaced by local optimizations over Λ. To handle the dependency of the feasible set on λ_s consider the following set:

$$\mathrm{SegCoupl}(Y, \mu_1, \mu_2) = \left\{ \pi \in \mathrm{Meas}(X \times Y) : \ \mu_1 \leq \mathrm{pr}_{X\sharp}\pi \leq \mu_2 \wedge \mathrm{pr}_{Y\sharp}\pi \leq \mathcal{L}_Y \right\} \tag{4.2.26}$$

Obviously $\mathrm{SegCoupl}\left(Y,(1+\lambda_\mathrm{s})^2 \cdot \mu\right) \subset \mathrm{SegCoupl}\left(Y,(1+\lambda_{\mathrm{s,l}})^2 \cdot \mu, (1+\lambda_{\mathrm{s,u}})^2 \cdot \mu\right)$ as long as $\lambda_{\mathrm{s,l}} \leq \lambda_\mathrm{s} \leq \lambda_{\mathrm{s,u}}$. Then a possible definition of $E_{\mathrm{s},2}$ equivalent to (4.2.23) is

$$\begin{aligned} E_{\mathrm{s,2a}}(\Lambda) = \inf_{\pi \in \mathrm{SegCoupl}\left(Y,(1+\lambda_{\mathrm{s,l}})^2 \cdot \mu, (1+\lambda_{\mathrm{s,u}})^2 \cdot \mu\right)} & \left(\min_{\lambda_\mathrm{s} \in [\lambda_{\mathrm{s,l}}, \lambda_{\mathrm{s,u}}]} \frac{1}{2\,(1+\lambda_\mathrm{s})^2} \right) \\ & \int_{X \times Y} \left(\left(\inf_{\lambda \in \Lambda} c_\mathrm{geo}\big(T_\lambda(x), y\big) \right) + c_\mathcal{F}(f_x, f_y) \right) d\pi(x,y) + \\ & \inf_{\lambda \in \Lambda} F(\lambda) + G(\mathrm{pr}_{Y\sharp}\pi) \end{aligned} \tag{4.2.27}$$

where $\lambda_{\mathrm{s,l}}$ and $\lambda_{\mathrm{s,u}}$ are the largest lower and smallest upper bound on λ_s in Λ. It is easy to see that $E_{\mathrm{s,2a}}$ satisfies Proposition 4.2.6 w.r.t. $E_{\mathrm{s},1}$. The proof is analogous.

If G is zero the definition of $E_{\mathrm{s,2a}}$ can be improved upon. Consider the following Lemma:

Lemma 4.2.9. *For some cost function c and $m > 0$ let*

$$f(m) = \inf_{\pi \in \mathrm{SegCoupl}(Y, m\cdot\mu)} \int_{X\times Y} c(x,y)\, d\pi(x,y)\,. \tag{4.2.28}$$

Then $f(m_2)/m_2 \geq f(m_1)/m_1$ for $m_2 > m_1$.

Proof. Assume $f(m_2) < (m_2/m_1)\cdot f(m_1)$ for $m_2 > m_1$ and let π_2^* be an optimizer for $f(m_2)$. Then $(m_1/m_2)\cdot\pi_2^*$ is feasible for computation of $f(m_1)$ and one has

$$\frac{m_1}{m_2}\int_{X\times Y} c(x,y)\, d\pi_2^*(x,y) = \frac{m_1}{m_2} f(m_2) < f(m_1) \tag{4.2.29}$$

which is a contradiction. □

With the aid of Lemma 4.2.9 one then finds that the following is a suitable variant of $E_{\mathrm{s},2}$:

$$E_{\mathrm{s},2\mathrm{b}}(\Lambda) = \inf_{\pi\in\mathrm{SegCoupl}\left(Y,(1+\lambda_{\mathrm{s},\mathrm{l}})^2\cdot\mu\right)} \frac{1}{2\,(1+\lambda_{\mathrm{s},\mathrm{l}})^2} \int_{X\times Y} \left(\left(\inf_{\lambda\in\Lambda} c_{\mathrm{geo}}\big(T_\lambda(x), y\big) \right) + c_{\mathcal{F}}(f_x, f_y) \right) d\pi(x,y) + \inf_{\lambda\in\Lambda} F(\lambda) \tag{4.2.30}$$

The advantages over $E_{\mathrm{s},2\mathrm{a}}$ are a tighter scaling factor and a simpler feasible set for the optimal transport term.

Remark 4.2.10 (Scale mode and alternating optimization)**.** The alternating optimization scheme presented in Sect. 4.2.2.1 only works with zero-divergence modes. The hierarchical optimization scheme can be used to extend this to the scale mode. The non-scale coefficients are determined by separate optimization as before, see (4.2.20b). The new coefficient $\lambda_{\mathrm{s}}^{k+1}$ and ν^{k+1} are jointly determined by global hierarchical optimization, while keeping the other mode coefficients fixed (this replaces (4.2.20a)). This hierarchical scheme will only go over one degree of freedom and thus be relatively quick. Again one finds a non-increasing sequence that must eventually converge.

4.2.2.3 Graph Cut Relaxation

Both alternating and hierarchical optimization require solving a lot of optimal transport problems. Even with efficient solvers this will quickly become computationally expensive as the size of X and Y or the number of modes increases. If G is non-zero then usually even more so because dedicated optimal transport solvers can no longer be applied directly to compute $E_1(\lambda)$. Therefore, in this section we present a mass-constraint relaxation that, for suitable choice of

G, turns computation of $E_1(\lambda)$ into a min-cut problem. These can be solved extremely fast with dedicated algorithms and therefore the relaxation yields a huge speed-up.

Throughout this section let X and Y be discrete sets, e.g. pixels or super-pixels. The Lebesgue measure on Y is approximated by

$$\mathcal{L}_Y(\sigma) = \sum_{y \in \sigma} m_y \tag{4.2.31}$$

for subsets $\sigma \subset Y$, where m_y is the area of super-pixel y. Any feasible $\nu \in \mathrm{SegMeas}(Y, M)$ can then be expressed as

$$\nu(\sigma) = \sum_{y \in \sigma} m_y \, u_\nu(y) \tag{4.2.32}$$

for all $\sigma \subset Y$ with some function $u_\nu \in Y \to [0,1]$. Let G be a total-variation-like local boundary regularizer of ν, expressed in terms of u_ν:

$$G(\nu) = \sum_{(y,y') \in \mathcal{G}} a_{y,y'} \cdot |u_\nu(y) - u_\nu(y')| \tag{4.2.33}$$

where $\mathcal{G}$ is the set of super-pixel neighbours and $a_{y,y'}$ is a weight that models the likelihood of a boundary between neighbours y and y'. Such weights can be constructed from feature dissimilarity in y, y', from edge detectors and from the length of the boundary.

We now relax the template-marginal constraint from the coupling set $\Pi(\mu, \nu)$ and allow ν to have arbitrary mass. So the feasible set of ν will be

$$\mathrm{SegMeas}(Y) = \left\{ \nu \in \mathrm{Meas}(Y) \,:\, 0 \leq \nu \leq \mathcal{L}_Y \right\}. \tag{4.2.34}$$

This is (2.1.8) without the mass constraint. The 'couplings' π will be taken from the set

$$\hat{\Pi}(\nu) = \left\{ \pi \in \mathrm{Meas}(X \times Y) \,:\, \mathrm{pr}_{Y\sharp}\pi = \nu \right\}. \tag{4.2.35}$$

Merging optimizations (see Remark 4.2.4) yields the feasible set

$$\mathrm{SegCoupl}(Y) = \left\{ \pi \in \mathrm{Meas}(X \times Y) \,:\, \mathrm{pr}_{Y\sharp}\pi \leq \mathcal{L}_Y \right\}. \tag{4.2.36}$$

The relaxed equivalent of E_1 (4.2.22) that we consider in this section is

$$E_{\mathrm{r},1}(\lambda) = \inf_{\pi \in \mathrm{SegCoupl}(Y)} \frac{1}{2} \int_{X \times Y} \Big(c_{\mathrm{geo}}\big(T_\lambda(x), y\big) + c_{\mathcal{F}}(f_x, f_y) \Big) \, d\pi(x,y) + F(\lambda) + G(\mathrm{pr}_{Y\sharp}\pi)\,. \tag{4.2.37}$$

Let π^* be an optimizer of $E_{\mathrm{r},1}(\lambda)$ for some configuration λ. If $(\mathrm{pr}_{Y\sharp}\pi^*)(y) > 0$ for some $y \in Y$, this mass will come from the cheapest $x \in X$ for this y, since

there is no longer any constraint on the mass on X. The linear matching in the first term simplifies to a nearest neighbour matching for each $y \in Y$. This implies that the minimization in (4.2.37) over $\pi \in \mathrm{SegCoupl}(Y)$ can be simplified to a minimization over $\nu \in \mathrm{SegMeas}(Y)$. Therefore (4.2.37) is equivalent to

$$E_{\mathrm{r},1}(\lambda) = \inf_{\nu \in \mathrm{SegMeas}(Y)} \frac{1}{2} \sum_{y \in Y} c_{\min}(y, \lambda)\, \nu(y) + F(\lambda) + G(\nu) \tag{4.2.38}$$

with

$$c_{\min}(y, \lambda) = \min_{x \in X} c_{\mathrm{geo}}\big(T_\lambda(x), y\big) + c_{\mathcal{F}}(f_x, f_y)\,. \tag{4.2.39}$$

We express now ν in terms of u_ν, see (4.2.32), and plug in the form of the regularizer G (4.2.33). This yields

$$\begin{aligned} E_{\mathrm{r},1}(\lambda) = \inf_{u: Y \to [0,1]} \frac{1}{2} \sum_{y \in Y} c_{\min}(y, \lambda) \cdot m_y \cdot u(y) + F(\lambda) + \\ \sum_{(y,y') \in \mathcal{G}} a_{y,y'} \cdot |u(y) - u(y')|\,. \end{aligned} \tag{4.2.40}$$

This is a convex formulation of the max-flow / min-cut problem with nodes Y and edges $\mathcal{G}$. The flow between $y \in Y$ and the source is given by $c_{\min}(y, \lambda) \cdot m_y$ and the capacities of the edges by $a_{y,y'}$. This problem can be solved very efficiently by dedicated algorithms, see for example [BK04].

Remark 4.2.11 (Optimization of $E_{\mathrm{r},1}$)**.** Both the alternating method and the hierarchical scheme, Sections 4.2.2.1 and 4.2.2.2, can be applied directly to the optimization of $E_{\mathrm{r},1}$. The sequence equivalent to (4.2.20) will provide a non-increasing converging sequence of energies. Since the dependence of the feasible set on the mass of μ has disappeared, it can also be extended to the scale mode. Also, handling the scale mode in the hierarchical scheme is simplified, since the feasible set for π is constant over λ_{s}.

Remark 4.2.12 (Background modelling in $E_{\mathrm{r},1}$)**.** $E_{\mathrm{r},1}$ can also be adapted to background modelling as introduced in Section 4.2.1.5. Since then $\nu = \mathcal{L}_Y$, the regularizer G must be modified to be meaningful in this context. Recall $X = X_F \cup X_B$. For any $\pi \in \mathrm{SegCoupl}(Y)$ let $\pi|_{X_F}$ be the restriction of π onto $X_F \times Y$. Introduce then $\nu_F = \mathrm{pr}_{Y\sharp}\pi|_{X_F} \in \mathrm{SegMeas}(Y)$ and correspondingly $\nu_B = \nu - \nu_F$. The measure ν_F marks out the foreground segmentation, ν_B corresponds to the background. So naturally $G(\nu)$ in (4.2.38) must be replaced by $G(\nu_F)$. As previously, consider some $y \in Y$ for which $(\mathrm{pr}_{Y\sharp}\pi)(y) > 0$. If the mass comes from some $x \in X_F$, it will come from the cheapest there. If it comes from some $x \in X_B$, it will come from the cheapest x there. The remaining problem is only to decide, which mass comes from X_F and which

from X_B and which is simply being ignored (which corresponds to background, too). Analogous to (4.2.39) we define

$$c_{\min,F}(y,\lambda) = \min_{x\in X_F} c_{\text{geo}}(x,y) + c_{\mathcal{F}}(f_x,f_y)\,, \tag{4.2.41}$$

$$c_{\min,B}(y,\lambda) = \min\{\min_{x\in X_B} c_{\text{geo}}(x,y) + c_{\mathcal{F}}(f_x,f_y), 0\} \tag{4.2.42}$$

where the dependence of c_{geo} on λ should be kept in mind. The equivalent to (4.2.38) is then

$$E_{\text{r,b,1}}(\lambda) = \inf_{\substack{\nu_F,\nu_B\in\text{SegMeas}(Y):\\ \nu_F+\nu_B=\nu}} \frac{1}{2}\left(\sum_{y\in Y} c_{\min,F}(y,\lambda)\,\nu_F(y) + \sum_{y\in Y} c_{\min,B}(y,\lambda)\,\nu_B(y)\right) + F(\lambda) + G(\nu_F)\,. \tag{4.2.43}$$

Let $u_F : Y \to [0,1]$ be the function representing ν_F analogous to (4.2.32). The corresponding function u_B for ν_B is given by $u_B = 1 - u_F$. In terms of u_F (4.2.43) can be rewritten as

$$E_{\text{r,b,1}}(\lambda) = \inf_{u_F:Y\to[0,1]} \frac{1}{2}\left(\sum_{y\in Y} c_{\min,F}(y,\lambda)\cdot m_y\cdot u_F(y) + \sum_{y\in Y} c_{\min,B}(y,\lambda)\cdot m_y\cdot\big(1-u_F(y)\big)\right) + F(\lambda) + \sum_{(y,y')\in\mathcal{G}} a_{y,y'}\cdot|u_F(y)-u_F(y')|\,. \tag{4.2.44}$$

Again, this is a max-flow problem where $c_{\min,F}(y,\lambda)\cdot m_y$ are the flows from the nodes to the source, $c_{\min,B}(y,\lambda)\cdot m_y$ are the flows from the nodes to the sink and edge capacities are again given by $a_{y,y'}$.

Functional (4.2.44) can be interpreted as a binary Markov random field (MRF) with labels fore- and background ($u_F = 1, 0$) and latent object configuration variables λ. Such enhanced MRFs have been used in [KTZ05] with the latent variables describing layered pictorial structures and in [YV13] with graph-based shape models. Optimization of a general class of such models via branch and bound has been discussed in [LBR08].

4.2.3 Experiments

4.2.3.1 Implementation Details

Computation of the E_2-lower bound requires local optimization w.r.t. λ for the cost function entries of the optimal transport term. Given the linear decomposition of T_λ these are low-dimensional constrained quadratic programs that can quickly be solved. For a given λ-interval Λ the locally minimized cost function $\min_{\lambda\in\Lambda}\|T_\lambda(x)-y\|^2$ has low values where λ values in Λ allow $T_\lambda(x)$ to be close

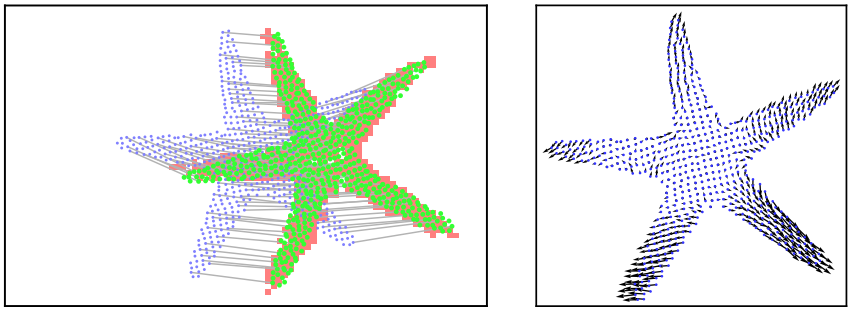


Figure 4.8: Left: Illustration of the approach on example data, analogous to Fig. 4.5b. Small blue dots indicate the arbitrary position of X relative to the image Y (bounding box). Large green dots give the position of $T_\lambda(X)$, the map is indicated by gray lines. The optimal segmentation ν is given by the red shaded region in the image. As intended, the *modes* model the Euclidean isometries (the true object position is not known beforehand and is not relevant for the result), and the major deformations. The Wasserstein distance handles the remaining degrees of freedom, guided by the local data term. Right: The deformation of X by the non-Euclidean modes. Length and relative orientation of the arms are adjusted.

to y and rises quickly elsewhere. Exploiting this, we only consider a sparse subset of the full product space $X \times Y$ to speed up computation. To ensure that we still obtain the global optimum, we add overflow variables (cf. Sect. 3.1.5.1). As long as no mass is transported onto these dummy variables, the global optimum is attained. Otherwise, more coupling combinations need to be added.

4.2.3.2 Homogeneous Appearance Model

For learning of the object class 'starfish' we used about 20 ground truth segmentations. We took the first four principal components as modes, capturing about 70% of the variance in the training set. Together with translation and rotation this amounts to seven modes to be optimized over.

To the test images we applied a simple local color model, trained on seeds, to obtain a homogeneous appearance model expressed in terms of linear affinity coefficients (cf. Sect. 3.1.13). Note that we specifically chose test images with inhomogeneously colored foreground objects and insufficient seeds for color model training, to obtain coefficients on which a purely local segmentation would fail and the benefit of shape-modelling can be demonstrated.

Fig. 4.8 illustrates the approach for a typical example. Position and pose of the sought-after object are correctly estimated by the modes, *independent of the position of X, i.e. without requiring a good initial guess.* Figure 4.9 gives original image, affinity coefficients g and the resulting segmentation for the example in Fig. 4.8 in column 1 and for further examples. The segmentations in

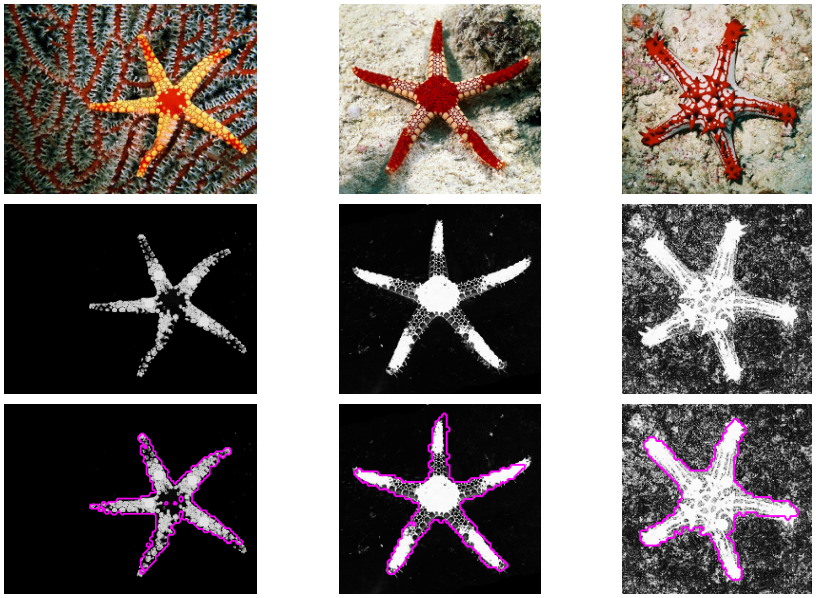


Figure 4.9: Segmentation results with starfish-prior. **First row:** original images. **Second row:** affinity coefficients g, based on a primitive local color model. There is false-positive clutter, foreground parts are poorly detected or missing. **Third row:** optimal segmentations, based on joint segmentation and matching. The objects are correctly located, clutter is ignored, missing parts are 'filled in'. Different variants are segmented with the same prior, due to statistical deformation modelling with modes.
On a small scale fluctuations may appear, although the underlying matching is smooth (cp. Fig. 4.8). These could be handled by enhancing the functional G to locally regularize the boundary of the segmentation.

Fig. 4.9 sometimes exhibit small holes or fluctuations along the boundary, even though the underlying object position and pose are very accurately determined (see Fig. 4.8) and the computed matching is smooth. These irregularities on the pixel level are induced by noise in G and could be removed by local regularization of the boundary of ν (e.g. total variation). As long as such an extension yields a convex functional G, it is still compatible with our approach. To make the acting of the presented Wasserstein-regularization as transparent as possible, however, we chose to omit such fine-tuning.

4.2.3.3 Scale Invariance and Representation Flexibility.

In this section we demonstrate two further important properties of our approach: scale invariance and flexibility in application. Due to the general formulation of optimal transport, adaptation to super-pixels is straightforward, which facilitates application to large images. In a discrete implementation Y

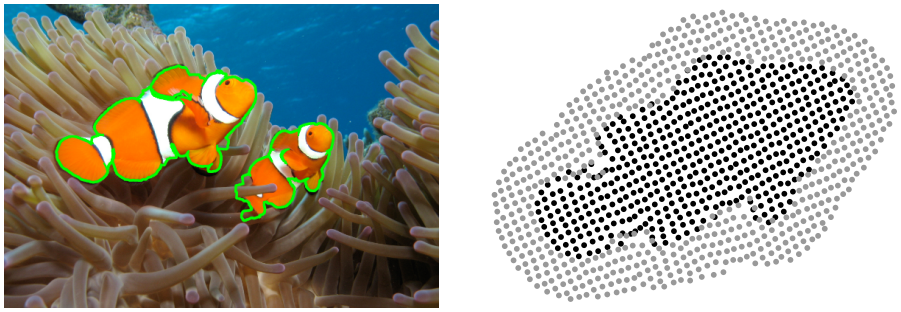


Figure 4.10: Scale invariance and super-pixels. **Left:** Foreground-affinities for G of a super-pixel over-segmentation of an image with two fish of different size, both of which induce strong local minima of our approach. By artificially breaking scale invariance through modelling the scale coefficient λ_s to be Gaussian, one can choose which one shall be segmented by setting the mean scale. Other than that, the setup was absolutely identical. One obtains similar segmentation results with a free-moving scale coefficient, if one, in turn, erases one of the fish from the G-coefficients. **Right:** Template X for fish-experiment. To prevent that the small fish is simply immersed in the big one, one must explicitly model a region of background around the fish, by reversing the affinity coefficients for this region of X. Black (gray) dots indicate fore-(back-)ground. See Sect. 4.2.1.5 for details.

need not be a regular grid (pixel-level) in $\mathbb{R}^2$, but can be any set of points.

We illustrate both aspects in Fig. 4.10. Our approach, equipped with a prior trained on fish, is run on an image with a large and a small fish. We demonstrate scale invariance by actually artificially breaking it: by modelling the scale-coefficient λ_s to be Gaussian, through the choice of the mean scale $\lambda_{s,m}$ we can trigger which of the two fish is segmented, while the wrong sized one is ignored. Except for the mean scale, no modifications in the approach were made.

The same experiment with the graph-cut relaxation is illustrated in Fig. 4.11.

4.2.3.4 Inhomogeneous Appearance Model

The framework proposed in Sect. 4.2.1.2 can also handle inhomogeneous appearance models where different parts of the template expect different features. The UIUC car dataset (see e.g. [AAR04]) provides a set of gray level side views of cars. Foreground (the cars) and background do not appear to be separable by a homogeneous appearance model. Only by taking into account the relative spatial location of particular appearance features one can locate the cars.

From each image we extracted three simple local features: the local color and the response to horizontally and vertically aligned small Gaussian edge filters. Then we used the registrations between the training samples that was

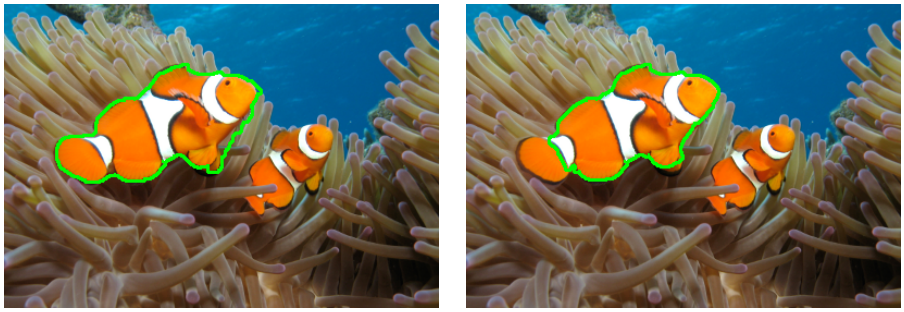


Figure 4.11: Graph cut relaxation: The experiment illustrated in Fig. 4.10 is repeated with the graph cut relaxation (**left:** large, **right:** small). While the segmentation outlines are more regular, as expected, due to the lack of the marginal constraint, the small fish is no longer correctly located.

computed during learning of the statistical deformations to learn for each part of the template a separate appearance classifier, based on the features that were assigned to that particular part. While the rows of the appearance cost function $c_{\mathcal{F}}(x_i, \cdot)$ are not very informative, the prior functional correctly locates the cars by combining the information of all template parts. The experiment is illustrated in Fig. 4.12.

4.2.4 Conclusion

A shape prior based on shape measures was presented. The manifold structure allows uniform modelling of geometric transformations and statistical variations. Although the resulting overall functional is non-convex, global optimality can be obtained by a hierarchical branch & bound scheme. Also an alternating local optimization scheme and an efficient graph-cut relaxation were discussed.

Compared to the relaxed Gromov-Wasserstein distance, as used in Sect. 3.1, this method has several notable differences: while the former relied on an intrinsic notion of invariance, here the template is explicitly embedded into the image plane. Thus, to obtain invariance, one must optimize over a suitable class of embeddings, resulting in a non-convex overall functional. In return, since for any fixed embedding one computes an optimal transport problem with the squared Euclidean distance as a part of the cost function, the obtained matchings tend to be more spatially regular as in Sect. 3.1, without the need for additional regularization. Also, explicitly keeping track of the location and pose of the object via the mode parameters in the tangent space allows to model non-isometric statistical variations of the template. This would be difficult to include into the relaxed Gromov-Wasserstein approach, since the statistical variation would also have to be estimated locally and independent, further loosening the relaxation.

The usefulness of the shape measure perspective was demonstrated numerically in the previous section. Open questions are the extension of the shape

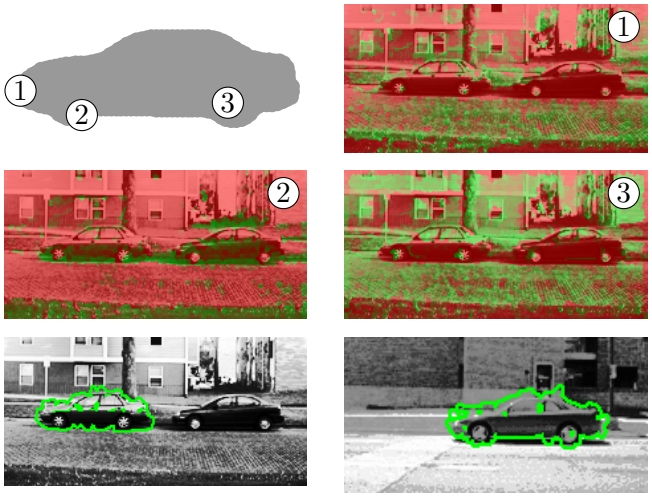


Figure 4.12: Wasserstein modes and inhomogeneous appearance model: For each template pixel a local appearance model was learned. **Top left:** template with three selected (super-)pixels $\{x_i\}_i$. **Top right, middle:** costs $c_{\mathcal{F}}(x_i, \cdot)$ for the three selected pixels. **Bottom:** two optimal segmentation examples. Although the appearance costs of single pixels are not very informative, the complete template with the information about relative spatial location of features can detect the cars. The segmentation outlines are irregular on a small scale because no local regularizer was employed.

model beyond the linear tangent space approximation through modes, the description of local appearance features and geometry by a unified manifold model and the application of shape measures to other applications such as object tracking.

Chapter 5

Hierarchical Optimal Transport

5.1 Problem and Overview

5.1.1 Observations on Optimal Transport Solvers

As discussed in Sect. 1.2.4 optimal transport and the linear assignment problem are important tools in machine learning and image processing. We have now seen that they are an elementary part of the work in this thesis. Some available solution methods have already been mentioned (Sect. 1.2.5). The continuous Monge formulation is however not applicable for our problems, since the cost functions are too general. Also, this would limit the applicability of the mm-space representation to absolutely continuous measures in $\mathbb{R}^n$, which rules out point clouds and super-pixels. The linear programming formulation or the combinatorial solvers are, when applied naïvely, too inefficient to solve large problems.

We have already presented some ideas to speed up computation: the radial mass distribution formulation (Sect. 3.1.4) or the heuristic cost function thresholding and iteration (Sect. 3.1.5.1). These methods have tight limitations however: the radial distribution approach becomes inefficient for complex feature spaces. In the 'thresholding and iteration' approach there is no upper bound on the number of iterations. Potentially one has to solve a long sequence of large problems, with only a few variables added at each step. This can turn out to be more expensive than a single exhaustive run.

A problem with combinatorial solvers is that they become slow on large, dense problems, since they do not exploit any information on the cost function structure. However, a priori all our problems are dense. Sometimes one can use additional information on the problem to exclude certain assignment pairs beforehand, thus reducing the problem size. But such methods will usually be very application specific. In general it is hard to find criteria for ruling out

1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0
5	5	4	3	2	1

Figure 5.1: Local and global structure of cost function: the table provides a cost function for a small example instance of the LAP with $N = 6$. The unique optimal assignment is highlighted in green. The cost function has a distinct 'valley' close to the diagonal and grows when moving away from this valley. It appears very 'regular'. Still one of the highest values of the cost function is part of the optimal assignment. This illustrates how difficult it is to find a priori rules for reliable exclusion of assignment pairs.

pairs that are both easy to check and can exclude a substantial fraction of assignments. Sometimes, assignments that appear very unlikely from a local perspective, can be part of the optimal assignment due to the global structure of the cost function (Fig. 5.1).

5.1.2 Chapter Overview

In this chapter we present a sparse/dense hybrid variant of the auction algorithm, based on a hierarchical representation of the optimal transport problem (Sect. 5.2.2). The algorithm can exploit any heuristic for exclusion of pairs, even if some exclusion-proposals may be wrong in the end. Wrongfully excluded pairs are detected by a hierarchical consistency check for violated dual constraints, thus guaranteeing global optimality of the final matching (Sect. 5.2.3). We will see that the efficiency of this scheme depends on the regularity of underlying cost function. The hierarchical structure lends itself to construct good sparse estimates for the initialization of the hybrid algorithm by a multi-scale scheme (Sect. 5.2.4). This scheme is very general and will in practice work on most problems. To obtain solving times that are competitive with other algorithms one needs to apply the technique of ε-scaling which can be naturally combined with our multi-scale scheme (Sect. 5.3.2). A complexity analysis shows that the worst case overhead of the hierarchical checks compared to the standard auction algorithm is bounded (Sect. 5.3.1). The algorithm is then experimentally compared to the standard auction algorithm in Sect. 5.4. We show that on 'typical' problems there is a significant gain in the average performance that increases with the problem size. We empirically investigate how the speed-up relates to a suitable measure of regularity.

Throughout this Chapter we consider discrete problems.

5.2 Multi-scale Auction Algorithm

5.2.1 Motivation

In the dual problem of the LAP and OT every cost function entry corresponds to one constraint. During updates of the dual variables one needs to make sure that no constraint is violated. In the auction algorithm this corresponds to the scans in (2.2.12) and (2.2.13) during the bidding phase. Of course, for large, dense problems checking all constraints takes a long time.

In most practical problems, the sets X and Y are equipped with some additional structure and notion of *closeness* or *similarity*, which also represents itself in the cost function. If x and y are close to x' and y' respectively, then we expect $|c(x,y) - c(x',y')|$ to be somehow bounded. Additionally, throughout execution of the algorithm, we hope that $|\alpha(x) - \alpha(x')|$ and $|\beta(y) - \beta'(y)|$ will be bounded. The appropriate notions of closeness and boundedness depend on the application. Some intuition for this is given in Sects. 5.2.4.2 and 5.4.

The idea underlying the sparse/dense hybrid auction algorithm is as follows: the algorithm operates with a sparse subset $\mathcal{N}' \subset \mathcal{N}$ of all potential pairings (2.2.1), thus requiring much fewer scans in (2.2.12) and (2.2.13). Before the bids are submitted however, it is checked, whether a constraint in $\mathcal{N} \setminus \mathcal{N}'$ was overlooked. This is not done for every constraint separately. Instead, by exploiting the regularity of c, α and β we will try to rule out constraint violations in whole groups of variables at once, thus reducing the number of required queries.

5.2.2 Hierarchical Structures

5.2.2.1 Hierarchical Partitions

To describe which elements in X and Y are close, we will use *partitions*.

Definition 5.2.1 (Partition)**.** For a finite set X a partition $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ is a set of mutually exclusive jointly exhaustive subsets of X, i.e.

$$a_i \subset X, \quad a_i \cap a_j = \emptyset \text{ for } i \neq j \quad \text{and} \quad \bigcup_{i=1}^{n} a_i = X\,. \tag{5.2.1}$$

The sets a_i are referred to as *cells* of the partition.

We will use partitions of X and Y where elements within each cell are considered close. Of course, whether two elements are close or not depends on the scale of the context. Therefore, we will not only use one partition, but a whole set of increasingly coarser partitions, a *hierarchical partition* [SS13b, Sect. 4].

Definition 5.2.2 (Hierarchical partition)**.** Let X be a finite set. A sequence $\mathcal{A} = \{\mathcal{A}_k\}_{k=1}^{K}$ of partitions of X is called a hierarchical partition of X if

(i) $\mathcal{A}_1 = \{\{x\} : x \in X\}$ is the partition of X into singletons,

(ii) for every $1 < k \leq K$ the cells of partition $\mathcal{A}_k$ are unions of cells of partition $\mathcal{A}_{k-1}$.

We call K the depth of the hierarchical partition. A hierarchical partition induces a directed tree graph with vertex set

$$V(\mathcal{A}) = \bigcup_{k=1}^{K} \mathcal{A}_k \,. \tag{5.2.2}$$

For $1 \leq k < K$ we say that $a \in \mathcal{A}_k$ is a child of $a' \in \mathcal{A}_{k+1}$ (and a' is the parent of a) and write $a \in \mathrm{ch}(a')$, $a' = \mathrm{pa}(a)$ when $a \subset a'$. Note that if $|\mathcal{A}_K| > 1$ then the graph is not connected.

Given a hierarchical partition $\mathcal{A} = \{\mathcal{A}_k\}_{k=1}^K$ on X, on scale $1 \leq k \leq K$ elements within the same cells of partition $\mathcal{A}_k$ are considered close.

5.2.2.2 Hierarchical Constraints

Consider now a given a linear assignment problem or an instance of optimal transport between sets X and Y. We first choose a suitable notion of closeness and construct corresponding hierarchical partitions $\mathcal{A} = \{\mathcal{A}_k\}_{k=1}^K$ and $\mathcal{B} = \{\mathcal{B}_k\}_{k=1}^K$ on X and Y (w.l.o.g. fixing them to have the same depth). Then we use these to define hierarchical extensions of the dual variables and the cost function:

Definition 5.2.3 (Hierarchical dual variables)**.** The hierarchical extension of dual variable α is given by

$$\hat{\alpha}(a) = \max_{x \in a} \alpha(x) = \begin{cases} \alpha(x) & \text{if } a = \{x\} \in \mathcal{A}_1 \text{ for some } x \\ \max_{a' \in \mathrm{ch}(a)} \hat{\alpha}(a') & \text{if } a \in \mathcal{A}_k \text{ for some } k > 1 \end{cases} \tag{5.2.3}$$

for every $a \in V(\mathcal{A})$ (5.2.2). Analogously we construct $\hat{\beta}$ from β and $\mathcal{B}$.

Similarly we define a hierarchical extension $\hat{c}$ of the cost function c over $V(\mathcal{A}) \times V(\mathcal{B})$:

Definition 5.2.4 (Hierarchical cost function)**.**

$$\hat{c}(a, b) = \min_{(x,y) \in a \times b} c(x, y) \tag{5.2.4}$$

for $(a, b) \in V(\mathcal{A}) \times V(\mathcal{B})$. In practice we only use values where a and b are cells of partitions at the same scale k.

The hierarchical dual variables and the hierarchical cost function on the smallest scale $k = 1$ correspond to the regular dual variables and cost function. Now consider the set of hierarchical dual constraints.

Definition 5.2.5 (Hierarchical dual constraints)**.** For some $1 \leq k \leq K$ we define the *dual constraints at scale* k by

$$\hat{\alpha}(a) + \hat{\beta}(b) \leq \hat{c}(a,b) \quad \text{for all} \quad (a,b) \in \mathcal{A}_k \times \mathcal{B}_k \tag{5.2.5}$$

For $a' \in \mathrm{ch}(a)$ and $b' \in \mathrm{ch}(b)$ one has by definition $\hat{\alpha}(a') \leq \hat{\alpha}(a)$, $\hat{\beta}(b') \leq \hat{\beta}(b)$ and $\hat{c}(a',b') \geq \hat{c}(a,b)$. Consequently, when the hierarchical dual constraint for $(a,b) \in \mathcal{A}_k \times \mathcal{B}_k$ for $k > 1$ holds, we know immediately that any hierarchical constraint for $(a',b') \in \mathrm{ch}(a) \times \mathrm{ch}(b)$ at scale $k-1$ and consequently all other constraints further down the graph on smaller scales hold. The hierarchical constraints at scale $k = 1$ are the regular dual constraints.

We will use this structure to verify validity of all dual constraints by trying to rule out as many as possible on large scales with few checks and only descending in scale if necessary. The requirement that elements within the same partition cell at a given generation should be close, will ensure that the dual constraints at scale k will not be a lot tighter than at scale $k-1$.

5.2.3 Hierarchical Consistence Phase

Based on the hierarchical data structures introduced in the last section we are now ready to devise a sparse/dense hybrid variant of the auction algorithm. Consider a feasible linear assignment or optimal transport problem between X and Y with cost function c. Let $\mathcal{N}$ be the set of neighbours as introduced in (2.2.1). We consider dense problems with $|\mathcal{N}| = \mathcal{O}(|X| \cdot |Y|)$. Let $\mathcal{N}' \subset \mathcal{N}$ be a sparse subset of neighbours, possibly given by some sparsification heuristic, but potentially faulty in the sense that pairings $(x,y) \notin \mathcal{N}'$ may actually be part of the globally optimal assignment (or in the support of the globally optimal coupling).

When one runs the auction algorithm with $\mathcal{N}'$ instead of $\mathcal{N}$ the bidding phase can be completed much faster, since the sets $\mathcal{N}'(x)$ and $\mathcal{N}'(y)$ to scan over for determining $\alpha(x)$ in (2.2.12) and $\mathcal{R}(x)$ in (2.4.40) are much smaller. At the same time however, since we are not considering all possible pairs in $\mathcal{N}$, the $\alpha(x)$ and $\alpha'(x)$ that we are determining or the candidates in $\mathcal{R}(x)$ might be larger than allowed by the dual constraints (2.2.7). Thus, before accepting any bids, we need to check that no relevant constraints have been missed. So we run the bidding and the assignment phase of the auction algorithm just as presented in Sects. 2.2.2.4 and 2.4.3.3, but in between the following hierarchical consistency check phase is executed:

Hierarchical Consistency Check Phase. Let $\hat{\alpha}$ be the hierarchical extension (Def. 5.2.3) of α' from (2.2.13) in the LAP case and from (2.4.44, 2.4.45) in the OT case. Let $\hat{\beta}$ be the hierarchical extension of the dual variable β in the LAP case and of (2.4.38) in the OT case. $\hat{c}$ is the hierarchical extension of c (Def. 5.2.4).

Then for $k = K$ check whether the hierarchical constraints $\hat{\alpha}(a) + \hat{\beta}(b) \leq \hat{c}(a, b)$ for all $(a, b) \in \mathcal{A}_K \times \mathcal{B}_K$ hold. For every violated constraint at (a, b) check whether $\hat{\alpha}(a') + \hat{\beta}(b') \leq \hat{c}(a', b')$ hold for all $(a', b') \in \mathrm{ch}(a) \times \mathrm{ch}(b)$. Recursively refine the checking on all violated constraints until either all constraints are satisfied or some violated constraints are found at scale $k = 1$. The pairs (x, y) where constraints are violated either correspond to $\alpha(x)$ and the corresponding minimizer y in (2.2.12) or to entries in $\mathcal{R}(x)$ before the position m (or m') that determines $\alpha'(x)$ in (2.4.40), or the pair $(x, y) \notin \mathcal{N}'$ and it was overlooked. In the latter case, add (x, y) to $\mathcal{N}'$ and list element x for rebidding. After completion of the consistency check phase, rerun the bidding phase for all listed x.

With this additional consistency check phase, one finds:

Proposition 5.2.6. *The sparse / dense hybrid auction algorithm with an hierarchical consistency phase between the bidding and assignment phase, initialized with a non-maximal neighbourhood $\mathcal{N}'$ converges to a globally optimal assignment / coupling under the same conditions as the regular auction algorithms.*

Proof. Note first that the auction algorithm for the assignment problem can be considered a special case of the optimal transport variant where $\alpha(x)$ (2.2.12) is always the smallest entry in $\mathcal{R}(x)$ and the position m/m' that determines $\alpha'(x)$ (2.4.44,2.4.45) is always 2, yielding $\alpha'(x)$ as specified by (2.2.13). We will henceforth only consider the optimal transport variant.

Consider the following observations on the algorithm:

(i) During the bidding stage, elements of the ordered list $\mathcal{R}(x)$, (2.4.40), that appear beyond position m/m', see (2.4.44,2.4.45), depending on which case is relevant, do not affect the process of the algorithm.

(ii) The algorithm converges regardless of the arbitrary ordering of elements of equal size in $\mathcal{R}(x)$.

By combinig (i) and (ii) we can easily see, that any variant of the auction algorithm that can reproduce the list $\mathcal{R}(x)$, (2.4.40), up to position m/m' (we will from now on simply write m, implying the correct case), and up to irrelevant reorderings of elements of equal size, has the same convergence properties as the full auction algorithm.

Let $\mathcal{N}$ be the complete neighbourhood set of the full problem and let $\mathcal{N}'$ be the reduced set that the hybrid variant is working with at a given iteration.

Assume the hybrid variant has completed the bidding stage and in the consistency check phase no new candidates for some element x were added. Let $\alpha'(x) = c(x, y_m) - \beta(x'_m, y_m)$ be the m-th entry of $\mathcal{R}(x)$. By the hierarchical structure of the dual constraints at all scales, we know that $\alpha'(x) \leq c(x, y) - \beta(y)$ for all $y \in \mathcal{N}(x) \setminus \mathcal{N}'(x)$ and because $\beta(y) \geq \beta(x', y)$ for all x' with $\pi(x', y) > 0$ we thus have that $\alpha'(x) \leq c(x, y) - \beta(x', y)$ for all pairs (x', y) that

correspond to potential candidates in $\mathcal{R}(x)$, (2.4.40), that have been skipped due to the reduced neighbourhood set $\mathcal{N}'$. All these skipped potential elements would therefore have appeared beyond position m (up to potentially reordering elements of equal size) and can thus be considered irrelevant for the process of the algorithm.

Now assume, during the consistency check phase a set of violated constraints $\mathcal{V} \subset \mathcal{N}(x) \setminus \mathcal{N}'(x)$ was found, i.e. $c(x,y) - \beta(y) < \alpha'(x)$ for $y \in \mathcal{V}$. After extending $\mathcal{N}'_{\text{new}}(x) = \mathcal{N}'(x) \cup \mathcal{V}$ and recomputing the bidding steps, one finds that the new $\alpha'_{\text{new}}(x) \leq \alpha'(x)$. However, for the set $\mathcal{N}(y) \setminus \mathcal{N}'_{\text{new}}(x)$ we know that $c(x,y) - \beta(y) \geq \alpha'(x) \geq \alpha'_{\text{new}}(x)$ (since those are the constraints that were not violated in the first consistency check) and thus a second consistency check would always be successful. Hence, by the reasoning above, no relevant candidates have been missed.

It may happen that the list $\mathcal{R}(x)$ is too short to find a valid integer m. This can happen for two reasons: either the maximal neighbourhood $\mathcal{N}$ is as small as feasibility allows and x needs all its neighbours $\mathcal{N}(x)$ to ship its mass to. In this case the value for $\alpha'(x)$ can formally be set to ∞ which will result in a bid of $-\infty$ which basically removes x and all its neighbours from the rest of the assignment problem. In the other case, the reduced neighbourhood $\mathcal{N}'(x)$ has been chosen too small. Handling this also with setting $\alpha'(x) = \infty$ will result in all hierarchical consistency checks being violated for that x and having to set $\mathcal{N}'_{\text{new}}(x) = \mathcal{N}(x)$. This can of course be very inefficient and sabotage the idea of reducing the number of active variables in $\mathcal{N}'$. Hence, upon initialization of the hybrid algorithm one should see to it that the sets $\mathcal{N}'(x)$ are sufficiently large. □

Given a 'reasonable' initialization $\mathcal{N}'$, that may be flawed, we can thus efficiently solve linear assignment and optimal transport problems to global optimality, with the aid of the hierarchical consistency checks.

It should be noted that the modifications preserve the parallel structure of the algorithm. Bidding and assigning work as before and the tree structure of the hierarchical constraints allows for distribution of the hierarchical consistency checks onto multiple processors.

In the next section we will describe how the hierarchical structure can be used to come up with a good initial guess for $\mathcal{N}'$ in the first place.

5.2.4 Multi-scale Sparse Initialization Scheme

How does one obtain a good initial guess for $\mathcal{N}'$? In Sect. 3.1.5.1 a heuristic was applied where $\mathcal{N}'(x)$ was chosen to be the subset of Y with the smallest cost function entries $c(x,y)$. While this worked well for that particular application it may in general not be very efficient, see Fig. 5.2. We will now present a very general scheme to generate proposals for initial sets $\mathcal{N}'$ by using the hierarchical structure that we introduced for the hybrid auction algorithm.

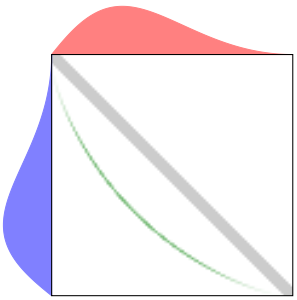


Figure 5.2: Naïve sparse initialization: a simple 1-dimensional optimal transport problem is illustrated. The blue and red shaded regions indicate the involved measures. As cost function the squared Euclidean distance is used. One could try to initialize the hybrid algorithm with the region of lowest cost function values, this region is given by the gray shading. However, in this example, due to the inhomogeneous densities, the support of the optimal coupling (green) lies almost completely outside of this region. This shows that more sophisticated initialization schemes are required.

5.2.4.1 Hierarchical bounds

Let μ and ν be the marginals in the OT problem that we consider (counting measures in the LAP case). After dual variables and the cost function we will now also introduce a hierarchical set of marginals.

Definition 5.2.7 (Hierarchical marginals)**.** For $1 \leq k \leq K$ the measure $\hat{\mu}_k \in \mathrm{Meas}(\mathcal{A}_k)$ is given by

$$\hat{\mu}_k(\sigma) = \sum_{a \in \sigma} \mu(a) \tag{5.2.6}$$

for all subsets $\sigma \subset \mathcal{A}_k$. One finds that $\hat{\mu}_1$ is basically identical to μ, only that $\hat{\mu}_1$ is formally defined on the set of singletons of X. By a slight abuse of notation we identify the two measures. By introducing the functions

$$p_{X,k} : \mathcal{A}_k \to \mathcal{A}_{k+1}, \quad a \mapsto \mathrm{pa}(a) \tag{5.2.7}$$

for $1 \leq k < K$ one easily sees that $\hat{\mu}_{k+1} = {p_{X,k}}_\sharp \hat{\mu}_k$. Similarly define $\hat{\nu}$ and the functions $p_{Y,k}$. Let p_k be the product map of $p_{X,k}$ and $p_{Y,k}$.

This induces a hierarchical sequence of optimal transport problems. For $1 \leq k \leq K$ let

$$D_k = \inf_{\pi \in \Pi(\hat{\mu}_k, \hat{\nu}_k)} \sum_{(a,b) \in \mathcal{A}_k \times \mathcal{B}_k} \hat{c}(a,b)\, \pi(a,b)\,. \tag{5.2.8}$$

We refer to problem (5.2.8) for some k as the problem at scale k. Computing D_1 corresponds to solving the original full problem. The coarser the scale k,

the smaller are the cardinalities of $\mathcal{A}_k$ and $\mathcal{B}_k$, thus the problem of computing D_k becomes smaller. We will now investigate how the values D_k at different scales relate to each other and whether they can be used to approximate or bound D_1. For $1 < k \leq K$ let

$$\begin{aligned}\Delta c_k &= \max_{(a,b)\in\mathcal{A}_k\times\mathcal{B}_k} \left(\left(\max_{(a',b')\in \mathrm{ch}(a)\times\mathrm{ch}(b)} \hat{c}(a',b') \right) - \left(\min_{(a',b')\in \mathrm{ch}(a)\times\mathrm{ch}(b)} \hat{c}(a',b') \right) \right) \\ &= \max_{(a,b)\in\mathcal{A}_k\times\mathcal{B}_k} \left(\left(\max_{(a',b')\in \mathrm{ch}(a)\times\mathrm{ch}(b)} \hat{c}(a',b') \right) - \hat{c}(a,b) \right) \end{aligned} \tag{5.2.9}$$

be the maximal variation of the hierarchical cost function within products of partition cells of a given scale. We then find

Proposition 5.2.8. *For $1 \leq k < K$ one has*

$$D_{k+1} \leq D_k \leq D_{k+1} + M \cdot \Delta c_{k+1} \tag{5.2.10}$$

where M is the total mass of μ and ν.

Proof. We start with the first inequality:

$$\begin{aligned} D_k &= \inf_{\pi\in\Pi(\hat{\mu}_k,\hat{\nu}_k)} \sum_{(a,b)\in\mathcal{A}_k\times\mathcal{B}_k} \underbrace{\hat{c}(a,b)}_{\geq \hat{c}\left(p_{X,k}(a),p_{Y,k}(b)\right)} \pi(a,b) \\ &\geq \inf_{\pi\in p_{k\sharp}\Pi(\hat{\mu}_k,\hat{\nu}_k)} \sum_{(a',b')\in\mathcal{A}_{k+1}\times\mathcal{B}_{k+1}} \hat{c}(a',b')\,\pi(a',b') \end{aligned}$$

and by virtue of Proposition 3.1.8 with $\phi_X = p_{X,k}$ and $\phi_Y = p_{Y,k}$

$$= \inf_{\pi\in\Pi(p_{X,k\sharp}\hat{\mu}_k, p_{Y,k\sharp}\hat{\nu}_k)} \sum_{(a',b')\in\mathcal{A}_{k+1}\times\mathcal{B}_{k+1}} \hat{c}(a',b')\,\pi(a',b') = D_{k+1}\,.$$

The second inequality follows analogously but instead of bounding

$$\hat{c}(a,b) \geq \hat{c}\big(p_{X,k}(a), p_{Y,k}(b)\big)$$

we bound

$$\hat{c}(a,b) \leq \hat{c}\big(p_{X,k}(a), p_{Y,k}(b)\big) + \Delta c_{k+1}\,. \qquad \square$$

Remark 5.2.9 (Hierarchical computation of bounds)**.** If the bounds Δc_k are sufficiently small this allows an immediate application: when one uses optimal transport for a classification task and needs to determine the nearest neighbours of a query measure in a set of reference samples, one could potentially first compute some bounds D_k for $k > 1$ which is computationally cheaper and only compute the more expensive D_1 for the most promising candidates.

5.2.4.2 Successive Refinement

Assume we have solved the optimal transport problem at some scale $k > 1$ with the auction algorithm with some parameter $\varepsilon_k > 0$ satisfying optimality condition (2.4.39). Let π_k^* be the optimal coupling and (α_k^*, β_k^*) the dual variables obtained upon convergence. We are now looking for a guess for the sparse neighbourhood $\mathcal{N}'_{k-1}$, a coupling $\pi_{\text{init},k-1}$ and a dual variable $\beta_{\text{init},k-1}$ to initialize the hybrid auction algorithm at scale $k-1$.

Proposition 5.2.8 implies existence of a feasible coupling $\hat{\pi}_{k-1}^*$ for the problem at scale $k-1$ which is suboptimal by at most $M \cdot \Delta c_k$. The proof of Proposition 3.1.8 provides an explicit construction of some $\hat{\pi}_{k-1}^*$ such that

$$\operatorname{spt}(\pi_k^*) \subset p_{k-1}^{-1}\big(\operatorname{spt}(\pi_k^*)\big)\,. \tag{5.2.11}$$

It seems therefore reasonable to use this set as an initial guess for $\mathcal{N}'_{k-1}$.

Alternatively, by the ε-complimentary slackness condition (2.2.11) we have

$$\hat{\pi}_{k-1}^*(a,b) > 0 \quad \Rightarrow \quad \alpha_k^*(a) + \beta_k^*(b) \geq \hat{c}(a,b) - \varepsilon_k \tag{5.2.12}$$

for $(a,b) \in \mathcal{A}_k \times \mathcal{B}_k$. So we propose:

$$\mathcal{N}'_{k-1} = \big\{(a,b) \in \mathcal{A}_{k-1} \times \mathcal{B}_{k-1} \,:\, \alpha_k^*\big(p_{X,k}(a)\big) + \beta_k^*\big(p_{Y,k}(b)\big) \geq \hat{c}(a,b) - \varepsilon_k\big\} \tag{5.2.13}$$

This set contains $\operatorname{spt}(\hat{\pi}_{k-1}^*)$ as is easy to check. In the context of the last paragraph of the proof of Proposition 5.2.6 we found that this initialization rule behaves more stable than picking $\operatorname{spt}(\hat{\pi}_{k-1}^*)$.

Additionally we must carefully pick the initial state variables $\pi_{\text{init},k-1}$ and $\beta_{\text{init},k-1}$. Since the new slackness parameter $\varepsilon_{k-1} < \varepsilon_k$ must be decreased when going down in scale to keep the optimality condition (2.4.39) satisfied, it is easiest to initialize $\pi_{\text{init},k-1}$ with the empty coupling such that ε-complimentary slackness (2.2.11) is trivially met (see also Remark 2.2.4). For the dual variable we propose

$$\beta_{\text{init},k-1}(b) = \beta_k^*\big(p_{Y,k-1}(b)\big)\,. \tag{5.2.14}$$

Proposition 5.2.10. *$\hat{\pi}_{k-1}^*$, $\beta_{\text{init},k-1}$ and the $\alpha_{\text{init},k-1}$ implied via (2.2.7) satisfy ε-complimentary slackness for $\varepsilon = \varepsilon_k + \Delta c_k$.*

Proof. First find for $a \in \mathcal{A}_{k-1}$

$$\begin{aligned}\alpha_{\text{init},k-1}(a) &= \min_{b \in \mathcal{B}_{k-1}} \hat{c}(a,b) - \beta_{\text{init},k-1}(b) \\ &\geq \min_{b \in \mathcal{B}_{k-1}} \hat{c}\big(p_{X,k-1}(a), p_{Y,k-1}(b)\big) - \beta_k^*\big(p_{Y,k-1}(b)\big) = \alpha_k^*\big(p_{X,k-1}(a)\big)\end{aligned}$$

And then

$$\begin{aligned}&\hat{\pi}_{k-1}^*(a,b) > 0 \overset{(5.2.11)}{\Rightarrow} \pi_k^*\big(p_{X,k-1}(a), p_{Y,k-1}(b)\big) > 0 \Rightarrow \\ &\alpha_k^*\big(p_{X,k-1}(a)\big) + \beta_k^*\big(p_{Y,k-1}(b)\big) \geq \hat{c}\big(p_{X,k-1}(a), p_{Y,k-1}(b)\big) - \varepsilon_k \Rightarrow \\ &\alpha_{\text{init},k-1}(a) + \beta_{\text{init},k-1}(b) \geq \hat{c}(a,b) - \varepsilon_k - \Delta c_k\end{aligned}$$

□

Corollary 5.2.11. *$\beta_{\text{init},k-1}$ and the $\alpha_{\text{init},k-1}$ implied via (2.2.7) are suboptimal for the problem at scale $k-1$ by at most $M \cdot (\varepsilon_k + \Delta c_k)$.*

Proof.

$$\begin{aligned}
&\sum_{a \in \mathcal{A}_{k-1}} \hat{\mu}_{k-1}(a)\, \alpha_{\text{init},k-1}(a) + \sum_{b \in \mathcal{B}_{k-1}} \hat{\nu}_{k-1}(b)\, \beta_{\text{init},k-1}(b) \\
=& \sum_{(a,b) \in \mathcal{A}_{k-1} \times \mathcal{B}_{k-1}} \hat{\pi}^*_{k-1}(a,b) \big(\alpha_{\text{init},k-1}(a) + \beta_{\text{init},k-1}(b)\big) \\
\geq& \sum_{(a,b) \in \mathcal{A}_{k-1} \times \mathcal{B}_{k-1}} \hat{\pi}^*_{k-1}(a,b) \big(\hat{c}(a,b) - \varepsilon_k - \Delta c_k\big) \\
\geq& D_{k-1} - M \cdot (\varepsilon_k + \Delta c_k)
\end{aligned}$$

□

This method of successive refinement can be used to recursively determine good initializations for large problems and solve them efficiently with the hybrid auction algorithm. One starts by solving the problem D_k at some scale which is coarse enough so that it can be solved densely. Then one uses the found optimal dual variables to initialize D_{k-1} and solves this with the hybrid variant. This cycle is repeated until D_1 is solved. We will in Sect. 5.4 provide numerical experiments that demonstrate the efficiency of this scheme.

Remark 5.2.12 (Some intuition on the efficiency of the hybrid variant and the multi-scale scheme)**.** Although it is quite difficult to predict the average computational complexity of the presented modified auction algorithm, this section provides some insight on why it may be more efficient on practical problems.

Assume the problem were solved at scale $k > 1$ and we initialize the problem at scale $k-1$ with $\beta_{\text{init},k-1}$ as specified by (5.2.14). We know from Proposition 5.2.10 that there is a feasible coupling $\hat{\pi}^*_{k-1}$ such that $\beta_{\text{init},k-1}$ and the corresponding $\alpha_{\text{init},k-1}$ satisfy ε-complimentary slackness with $\varepsilon = \varepsilon_k + \Delta c_k$. We can thus assume, that changes in the dual variables of the auction algorithm will happen on this scale.

From the previously solved, coarser scales, we can hope that 'irrelevant' dual constraints, corresponding to unlikely assignments, have a slack w.r.t. $\alpha_{\text{init},k-1}$ and $\beta_{\text{init},k-1}$ on the order of some $\Delta c_{k'}$ for $k' > k$.

If the problem at hand is such that Δc_k is decreasing quickly as we move to finer scales, i.e. the cost function varies less and less within the increasingly smaller partition cells, then these 'far off' constraints will in fact no longer be relevant for the execution of the algorithm.

Of course this is only a very rough sketch of what might be going on during the algorithm and the actual course of events depends on much more details than just the ratio of Δc_k between different scales. We will see however, in the experimental section, that there is in fact a relation between this ratio and the observed speed-up.

5.3 Complexity Analysis

5.3.1 Modified Bid Cost

We now compare the computational complexity of the sparse / dense hybrid variant together with the multi-scale initialization scheme relative to the standard auction algorithm as discussed in Sect. 2.2.2.

The worst case analysis in 2.2.2.5 was decomposed into estimating the cost per bid and the maximum number of bids. In this section we will estimate the worst case cost per bid for the hybrid variant. In the next section we discuss the maximum number of bids, how it can be reduced by ε-scaling and how ε-scaling can be combined with the multi-scale initialization scheme.

Consider a dense problem with $\mathcal{N} = X \times Y$, $|X| = |Y| = N$. The hierarchical partitions $\mathcal{A}$ and $\mathcal{B}$ will have K scales. Let b be a bound on the number of finer cells that any hierarchical partition cell in $\mathcal{B}$ will be decomposed to when going down to the next smaller scale and let B be the sum of the number of cells over all partitions in $\mathcal{B}$. In a worst case scenario the 'sparse' neighbourhood will be maximally extended, i.e. $\mathcal{N}' = \mathcal{N}$ and all the hierarchical consistency checks will have to be tried. Let us sum of the involved costs per bid:

(i) Scanning the cost function row to determine $\alpha(x)$ and $\alpha'(x)$ requires $\mathcal{O}(N)$ steps.

(ii) Updating $\hat{\alpha}'$: the values of $\alpha'(x)$ are non-decreasing throughout the algorithm. Since $\hat{\alpha}'$ is generated by maximizing over the cells of the next finer partition, an update after one value has been changed at the smallest scale needs only $\mathcal{O}(K)$ steps, one per scale.

(iii) Checking of all hierarchical constraints related to the bidding element x requires at most $\mathcal{O}(B)$ queries.

(iv) Potentially, after the hierarchical constraint check, bids have to be recomputed. This requires at most a constant factor of 2 of extra work which we will neglect.

(v) In the standard variant acceptance of a bid costs $\mathcal{O}(1)$ per bid: simply changing β accordingly. In the hybrid variant the hierarchical extension $\hat{\beta}$ must be updated. Since β is decreasing upon bidding, this requires at most $\mathcal{O}(K \cdot b)$ steps.

This yields a bound on the total cost per bid of

$$\mathcal{O}\big(N + K \cdot (1 + b) + B\big)\,. \tag{5.3.1}$$

Let us now consider a more particular hierarchical partition structure: assume

$|\mathcal{B}_{k-1}|/|\mathcal{B}_k| = b > 1$. For octrees one has for example $b = 8$. Then

$$B = \sum_{k=1}^{K} |\mathcal{B}_k| = N \sum_{k=0}^{K-1} b^{-k} < \frac{N}{1 - 1/b} \,. \tag{5.3.2}$$

Typically $K, b \ll N$. Then the first and last term in (5.3.1) dominate and one has

$$\mathcal{O}\left(N \cdot \left(1 + \frac{1}{1 - 1/b}\right)\right) . \tag{5.3.3}$$

This is the worst case overhead when 'everything that could go wrong goes wrong'. We see that it does not change the qualitative cost dependency on N but only adds a factor which is, w.r.t. N, constant. Note that for the optimal transport variant the worst case cost for scanning all bid candidates is $\mathcal{O}(N^2)$ (see Remark 2.4.19). Then the additional hierarchy consistency checks are not affecting the dominating term. In Sect. 5.4 we will demonstrate that on practical problems not 'everything goes wrong' and that the actual average cost per bid is lower.

The cost (5.3.3) also assumes that every bid is submitted on its own. When multiple bids are submitted in parallel the additional costs per bid due to the hierarchical consistency checks is smaller. This is why the hierarchical consistency phase fits better to the parallel variant of the auction algorithm.

5.3.2 Epsilon-Scaling

As in Sect. 2.2.2.6 it is advisable to use ε-scaling to reduce the total number of bids. Since both ε-scaling and the multi-scale initialization scheme attempt to solve the problem incremental from coarse to fine the two schemes should be adjusted to each other. We propose to use adapted ε-scaling to solve each of the scaled optimal transport problems (5.2.8).

As in Sect. 2.2.2.6 let θ be a parameter that regulates the reduction of ε. Let δc_k be the smallest absolute difference between two distinct cost function values at scale $1 \leq k \leq K$. This is the natural generalization of δc from Sect. 2.2.2.3 to coarser scales. Let

$$C_K = \left(\max_{(a,b) \in mcA_K \times \mathcal{B}_K} \hat{c}(a, b)\right) - \left(\min_{(a,b) \in mcA_K \times \mathcal{B}_K} \hat{c}(a, b)\right) \tag{5.3.4}$$

be the equivalent for C (2.2.17) at scale K and for $1 \leq k < K$ let

$$C_k = \Delta c_{k+1} + \delta c_{k+1} / \min\{|\mathcal{A}_{k+1}|, |\mathcal{B}_{k+1}|\} \,, \tag{5.3.5}$$

the usefulness of which we will see below (recall the definition of Δc_k from (5.2.9)). Also, let $N_k = |\mathcal{B}_k|$ for $1 \leq 1 \leq K$.

For solving D_K at the coarsest scale we assume no prior knowledge and thus choose first $\varepsilon_{K,0} = C_K/\theta$. The algorithm is successively run with $\varepsilon_{K,i} = \varepsilon_{K,i-1}/\theta$ until $\varepsilon_{K,i} < \delta c_K / \min\{|\mathcal{A}_K|, |\mathcal{B}_K\}$ which guarantees global optimality at scale K. In analogy to Sect. 2.2.2.6 this gives a bound on the number of bids of

$$\mathcal{O}\left(N_K^2 \log(N_K \cdot C_K/\delta c_K)\right) . \tag{5.3.6}$$

As in Sect. 5.2.4.2 let π_K^* and (α_K^*, β_K^*) be the primal and dual variable of the auction algorithm upon completing scale K. As then, one can use Proposition 3.1.8 to construct a coupling $\hat{\pi}_{K-1}^*$ for the problem at scale $K-1$ and dual variables $(\alpha_{\text{init},K-1}, \beta_{\text{init},K-1})$ via (5.2.14) and (2.2.7) such that they satisfy ε-complimentary slackness (2.2.11) for $\varepsilon = C_{K-1}$. So by using ε-scaling again, the scale $K-1$ can be solved by at most

$$\mathcal{O}\left(N_{K-1}^2 \log(N_{K-1} \cdot C_{K-1}/\delta c_{K-1})\right) \tag{5.3.7}$$

bids. This cycle can then be repeated for successively finer scales.

Remark 5.3.1 (Speed-up without sparsification through multi-scale)**.** Experiments show that the total number of bids is not reduced significantly through combining multi-scale solving and ε-scaling as compared to regular ε-scaling without multi-scale solving. However, at coarser scales bids are cheaper for the standard auction algorithm because the cardinalities of the problem are smaller. One could therefore attempt to speed up the standard auction algorithm without any sparsification by trying to get good initial dual variables through solving the coarse problems first. This will also be evaluated in the experimental section.

5.4 Experiments

5.4.1 Implementation Details and Test Problems

For evaluation of the effectiveness of the proposed hybrid-variant and multi-scale scheme we implemented the auction algorithm in `c++` with sparse data structures. The hybrid variant is based on the same implementation, extended by the hierarchical consistency phase, to obtain a meaningful basis for performance comparison. All computations were performed on standard desktop computers.

We used the multi-scale initialization scheme as described in Sect. 5.2.4 and scale-adjusted ε-scaling as described in Sect. 5.3.2. Note that at every iteration of ε-scaling we reset the sparse neighbourhood $\mathcal{N}_k'$ to the initial setting (5.2.13). We do this because relatively large ε-values at early stages of the scaling may cause large steps of the dual variables that cause $\mathcal{N}_k'$ to grow faster than required

for the later stage of the scaling with only small ε-values. On the standard variant we used ε-scaling as described in Sect. 2.2.2.6.

For experimental comparison we have devised several different test-scenarios that model a range of practical OT problems:

- **(H2D)** Assignment between two point clouds that are uniformly sampled from $[0,1]^2$, based on the squared Euclidean distance as cost function.
- **(H3D)** As above, but in three dimensions.
- **(H2D-1)** As (H2D), but with the non-squared Euclidean distance as cost function.
- **(I2D)** As (H2D), but with an inhomogeneous sampling density, given by gray-level images.
- **(Grid)** Continuous 2D densities, approximated by a regular grid on the unit square.
- **(Mesh)** Densities on the surface of a 3D mesh, equipped with its squared geodesic distance as cost function (Fig. 5.3).

In all cases quadtrees (resp. octrees in 3D) were used as hierarchical partitions. To illustrate another potential advantage of the presented algorithm, we have devised an additional scenario:

- **(Impl)** As (H2D), but the hierarchical cost function $\hat{c}$ was not generated via explicit minimization over the relevant partition cells, as in Definition 5.2.4, but implicit lower bounds were computed directly from the quadtree structure. Obviously the whole hierarchical consistency scheme still works.

In some applications computing the cost function $c(x,y)$ is itself a costly procedure and for huge problems storing the complete cost function can require a lot of memory. This can be avoided by computing closed-form lower bounds for $\hat{c}$ as in the scenario (Impl) and only computing and storing fine-scale cost function entries on demand. Of course this will require appropriate dynamical data structures.

As a proof on concept we will also provide an example with the relaxed linearized Gromov-Wasserstein distance matching presented in Sect. 3.1. In this application computing entries $c(x,y)$ involves solving an optimal transport problem itself and thus a lot of computational effort can be spared if these entries are only computed on demand.

To introduce a reasonable lower bound to δc_k, which determines the smallest necessary ε-parameter, we rounded all cost function values to three decimal places.

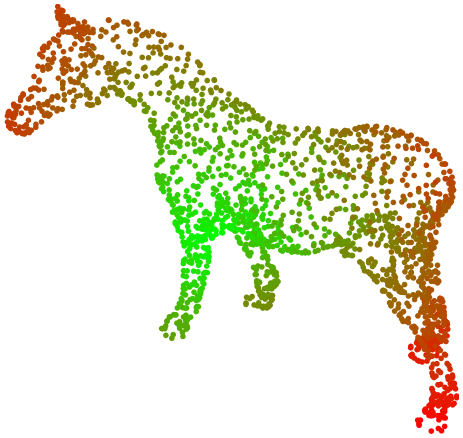


Figure 5.3: Illustration of scenario (Mesh): Two densities are given on the vertices of the mesh. The cost function is given by the squared geodesic distance between points, as induced by the mesh-structure. This is illustrated by the color of the vertices.

5.4.2 Coarse to Fine

Let us start the discussion of experimental results by going through the successive scales of the multi-scale scheme in detail. Figure 5.4 illustrates the comparison for each scale between the standard dense auction algorithm and the discussed acceleration schemes for an example of scenario (H2D). We see that when moving to finer scales, the relative sparsity of the hybrid variant increases and that the initial neighbourhood guess via the multi-scale scheme works well: less then 1% of all pairings need to be considered explicitly. As a result, the number of cost function queries (direct plus indirect via hierarchical checks) is reduced by about a factor of 20 relative to the number of queries in the dense variant. Eventually this results in a speed-up of about 5 in the presented example.

Concerning Remark 5.3.1 we see that the dense auction algorithm cannot be significantly accelerated through a multi-scale scheme. ε-scaling alone is just about as efficient.

5.4.3 Speed-up

It is clear that the number of hierarchical layers is critical for the speed-up that one obtains. In our scenarios we used 4×4 squares on the coarsest scale ($4 \times 4 \times 4$ cubes in 3D), a number of successively finer grids on the intermediate scales and the single points on the finest scale. Results of an investigation of the speed-up ratio, depending on the number of grid layers, for various scenarios, are plotted in Fig. 5.5 and listed in Table 5.1.

We see that if the finest grid level is too coarse, the multi-scale initialization

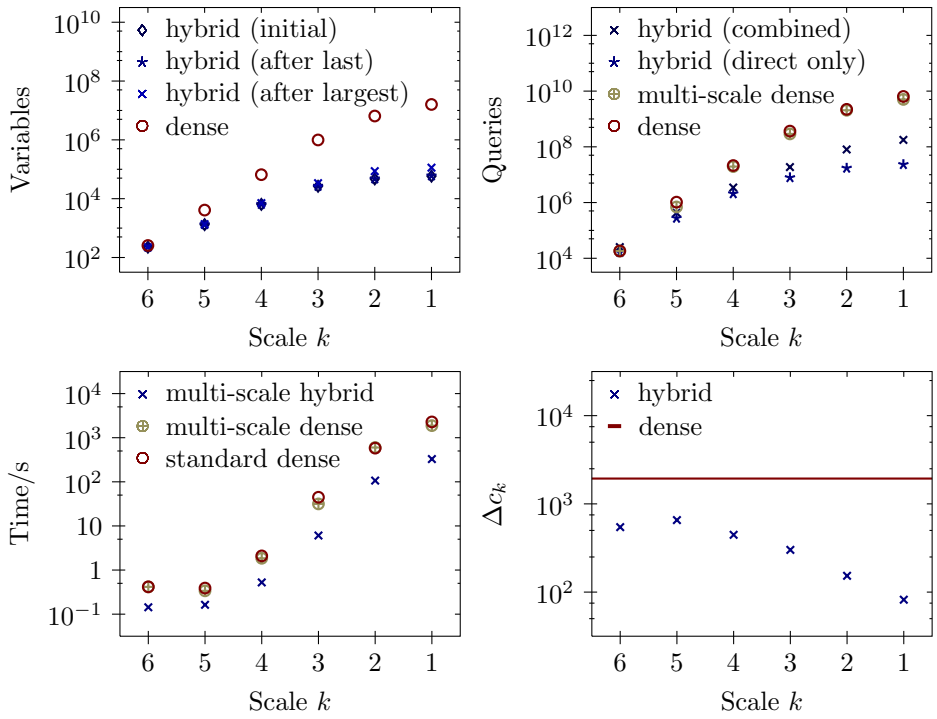


Figure 5.4: Illustration of solving a large LAP over multiple scales: a test problem of the class (H2D) with 4000 points in each cloud is solved over six scales, ranging from a 4×4 grid on each cloud (=256 coupling variables) to the fully resolved points. Shown are different relevant quantities for the multi-scale hybrid variant; for the dense variant with dual variable β initialized according to the multi-scale scheme; and the dense variant, trying to solve each scale problem from scratch. **Top left:** comparison of the problem sizes between the dense and the hybrid variant, upon initialization, upon termination during the last ε-scaling iteration and for the largest neighbourhood upon termination during any iteration of the ε-scaling scheme. We see that the initialization scheme works quite well in this example: the initial neighbourhood is increased by at most a factor of two and the increase during the last ε-scaling iteration is negligible. The relative level of sparsity increases with size. **Top right:** the number of constraint queries for the two dense runs (which are hardly distinguishable), the number of direct constraint queries of the hybrid variant and the combined number of direct queries plus consistency check phase scans. **Bottom left:** the time in seconds that each variant requires for solving a given scale. To compute the speed-up ratio one must sum up the times of the multi-scale approaches (hybrid and dense) and compare them to the time the dense, uninitialized variant needs on the finest scale. **Bottom right:** the values Δc_k (5.2.9) for the different scales and C (2.2.17) for the full problem.

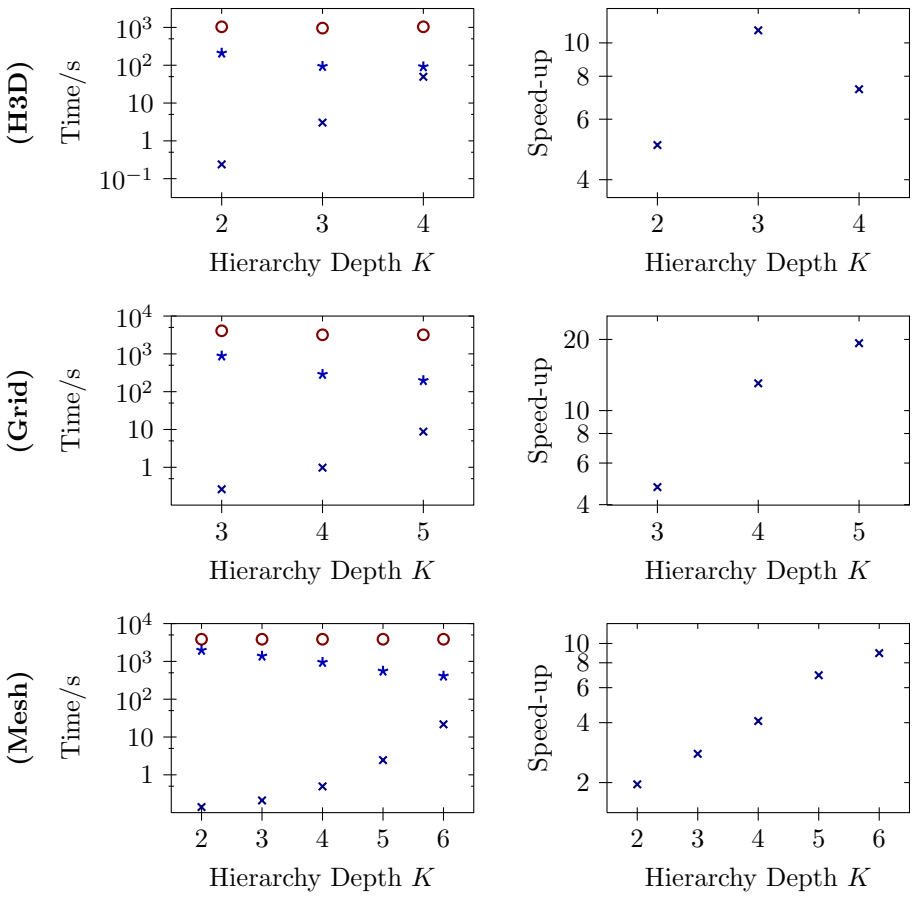


Figure 5.5: Dependency of the solving-time on the hierarchy depth for various test scenarios at $N = 4000$ points per cloud: **Left column:** Solving time of the multi-scale hybrid approach for all except the finest scale (×); solving time for the finest scale (*); and for comparison solving time of the standard dense approach on the same problem size (○). Slight variations in the dense times are due to statistical fluctuations. **Right column:** The ratio between the total solving time of the standard dense approach and the multi-scale hybrid approach. With increasing number of intermediate layers, the 'preparation' time for the final stage increases, but in return the time for the final stage itself decreases until saturation sets in. This determines the optimal number of intermediate levels.

Scenario	Scale Levels 2	3	4	5	6
(H2D)			8.02E-1	9.01E0	1.04E2
			3.62E2	3.15E2	3.26E2
			3.03E3	2.82E3	2.79E3
			9.05	9.18	6.42
(H3D)	2.39E-1	3.04E0	4.95E1		
	2.09E2	9.32E1	9.14E1		
	1.04E3	9.58E2	1.04E3		
	5.04	10.88	7.33		
(H2D-1)			1.09E0	9.33E0	4.17E1
			1.17E2	8.62E1	8.39E1
			1.22E3	1.21E3	9.59E2
			11.87	13.59	8.24
(I2D)			9.05E-1	1.26E1	
			2.37E2	3.98E2	
			2.05E3	2.88E3	
			8.86	7.43	
(Grid)		2.62E-1	9.78E-1	8.77E0	
		8.78E2	2.87E2	1.97E2	
		4.07E3	3.19E3	3.19E3	
		4.74	13.04	19.29	
(Mesh)	1.41E-1	2.10E-1	4.94E-1	2.45E0	2.16E1
	1.97E3	1.38E3	9.45E2	5.54E2	4.09E2
	3.85E3	3.85E3	3.85E3	3.85E3	3.85E3
	1.96	2.79	4.08	6.93	8.94
(Impl)			8.18E-1	3.96E0	3.39E1
			2.93E2	3.89E2	4.27E2
			2.81E3	3.30E3	3.01E3
			9.58	8.60	6.76

Table 5.1: Dependency of the solving time on the number of hierarchy layers for all test scenarios at $N = 4000$. **For each scenario** the **first row** gives the time the hybrid variant required for solving all but the finest scale. The **second row** gives the time the hybrid variant required for the finest scale. The **third row** gives the solving time of the dense standard algorithm. These values should be constant within each row. Variations are due to statistical fluctuations in generating random problems. The **fourth row** gives relative speed-up in total solving time from standard to hybrid variant. All times are given in seconds.

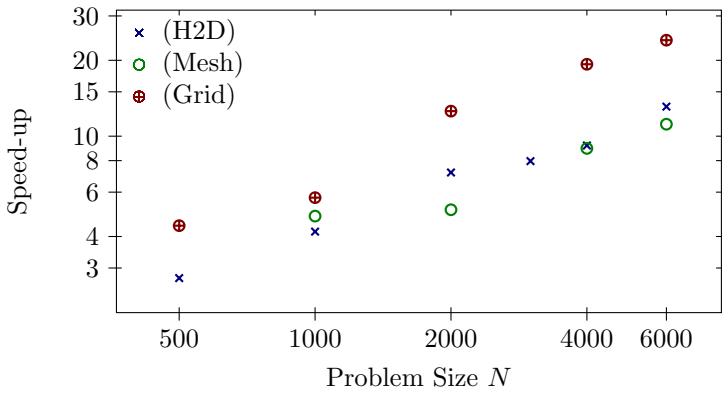


Figure 5.6: Average optimal relative speed-up (w.r.t. hierarchy depth) for different test-scenarios and for various problem sizes $N = |X| = |Y|$ of the multi-scale hybrid variant vs. the standard dense approach. For better overview only three scenarios are shown, the others essentially perform like (H2D). The hybrid variant significantly outperforms the standard algorithm in all scenarios. The gain increases with problem size and is at least of an order of magnitude at $N \approx 6000$ (grid-)points per side ($\approx 3.6 \cdot 10^7$ coupling variables).

for the single point level will not be very precise and the speed-up will be relatively small. By introducing more grid levels, the initialization for the final level will become better and the solving time at the final stage decreases. But when there are too many grid levels, solving them will take an increasing amount of time (since the cardinalities grow) whereas the gain on the last level will eventually saturate because the initialization does no longer improve significantly. Thus one has to determine the intermediate level of hierarchical layers at which the speed-up ratio is optimal.

In Figure 5.6 the optimal speed-up for the discussed scenarios and various problem sizes is shown. The acceleration consistently grows with problem size and is observed to be of about an order of magnitude for all scenarios at ≈ 6000 (grid-)points per side. In the particularly relevant scenario (Grid) the maximum observed speed-up is ≈ 24. For clarity the presented data is also given in Table 5.2.

In Remark 5.2.12 it was hypothesized that the hybrid auction algorithm will run more efficient, when the values Δc_k decrease quickly when moving to smaller scales, i.e. most of the 'structure' of the cost function is at coarse scales. Figure 5.7 provides some support of this notion: the relative number of queries of the dense variant and the hybrid variant at the finest scale is clearly related to the ratio $\Delta c_1 / C$. The smaller Δc_1, the better is the approximate solution obtained at scale $k = 2$ and thus the final scale can be solved more easily. Of course this is only a rule of thumb: different scenarios behave very different and the actual ratio depends on the detailed structure of the cost function. But the

Scenario	Problem Size N 500	1000	2000	3000	4000	6000
(H2D)	1.44E0	6.54E0	4.07E1	1.45E2	3.24E2	7.47E2
	3.63E0	2.59E1	2.74E2	1.06E3	2.82E3	9.56E3
	2.74	4.18	7.18	7.96	9.18	13.10
	1.13E7	2.03E7	1.13E8	3.89E8	2.45E8	5.04E8
	2.81E7	1.57E8	9.10E8	2.66E9	5.63E9	1.66E10
	2.49	7.68	8.03	6.82	23.15	32.79
(H3D)		3.95E0	1.51E1	4.41E1	9.63E1	3.08E2
		1.28E1	1.00E2	3.30E2	9.58E2	3.96E3
		3.35	6.91	8.12	10.88	12.70
		7.95E7	9.16E7	2.04E8	3.92E8	1.08E9
		8.59E7	4.18E8	1.07E9	2.35E9	6.34E9
		1.08	4.56	5.25	5.99	5.82
(H2D-1)	1.07E0	5.12E0	2.37E1	4.89E1	9.55E1	3.91E2
	2.49E0	1.82E1	1.55E2	3.80E2	1.21E3	3.50E3
	2.43	3.76	6.48	8.51	13.59	9.15
	1.18E7	2.36E7	1.34E8	1.01E8	2.01E8	6.11E8
	1.89E7	1.02E8	5.29E8	1.20E9	2.61E9	6.50E9
	1.62	4.34	4.04	12.02	13.13	10.91
(I2D)	1.39E0	5.94E0	5.89E1		2.38E2	9.96E2
	3.37E0	2.55E1	3.56E2		2.05E3	1.15E4
	2.63	4.40	6.62		8.86	11.59
	1.30E7	2.54E7	2.26E8		1.43E9	9.52E8
	2.84E7	1.60E8	1.03E9		4.79E9	1.57E10
	2.17	6.39	4.64		3.33	16.24
(Grid)	2.53E0	1.43E1	3.97E1		2.06E2	4.19E2
	1.07E1	7.96E1	4.49E2		3.19E3	9.75E3
	4.42	5.70	12.57		19.29	24.03
	2.10E7	3.49E7	1.42E8		3.43E8	7.04E8
	9.43E7	4.73E8	3.02E9		1.28E10	3.26E10
	4.45	13.55	20.69		36.99	45.71
(Mesh)		2.58E1	1.27E2		4.31E2	1.20E3
		1.25E2	6.50E2		3.85E3	1.34E4
		4.82	5.11		8.94	11.16
		6.04E8	2.06E9		5.90E9	1.68E10
		1.00E9	4.43E9		2.45E10	7.58E10
		1.66	2.15		4.15	4.50
(Impl)	1.69E0	8.05E0	5.61E1	1.31E2	2.93E2	4.42E2
	3.29E0	2.68E1	3.32E2	1.00E3	2.81E3	5.44E3
	2.03	3.53	6.63	7.91	9.58	12.19
	1.35E7	2.59E7	1.43E8	1.62E8	8.86E8	5.13E8
	2.71E7	1.60E8	9.91E8	2.65E9	5.66E9	1.43E10
	2.01	6.15	6.93	16.10	6.36	27.89

Table 5.2: Overview on speed-up and reduced constraint query numbers. **For each scenario** the **first three rows** give the mean runtime of the hybrid variant in seconds, the mean runtime of the dense variant in seconds and the mean speed-up ratio. The **second three rows** give the combined (hierarchical + finest level) mean number of constraint queries in the hybrid variant, the mean number of constraint queries in the dense variant and the mean reduction ratio.

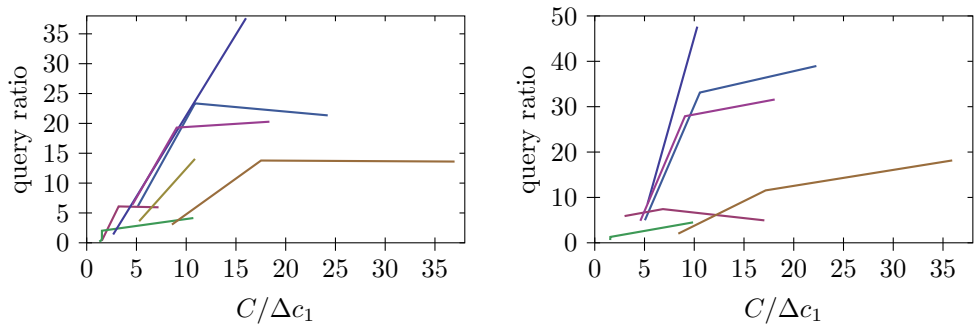


Figure 5.7: Reduction of query number depending on cost function regularity: Mean ratio of constraint queries of dense variant vs. hybrid variant on finest scale (regular and hierarchical) plotted against the mean ratio $C/\Delta c_1$ for different test scenarios. **Left:** $N = 4000$, **right:** $N = 6000$. **Legend:** ● (H2D), ● (H3D), ● (H2D-1), ● (I2D), ● (Grid), ● (Mesh), ● (Impl). There is a clear trend that the relative number of queries the hybrid variant requires decreases with increasing $C/\Delta c_1$.

general trend can clearly be seen.

5.4.4 Implicit Lower Bounds

We have already briefly studied the application of implicit lower bounds for the hierarchical cost function $\hat{c}$ in the scenario (Impl). The data in Table 5.2 shows that the multi-scale scheme still works well with the additional potential benefit of not having to compute all cost function entries $c(x, y)$ explicitly or store them in memory simultaneously.

This technique can be applied to shape segmentation with the relaxed linearized Gromov-Wasserstein distance, as presented in Sect. 3.1. Figure 5.8 illustrates results on an example instance. For this problem the number of required cost function entries compared to the dense naïve approach was only about 16% on the finest scale and additionally about 1% on coarser scales.

5.5 Conclusion

A sparse / dense hybrid variant of the auction algorithm was presented that only runs on a sparse subset of all potential dual constraints but ensures consistency with the dense problem through hierarchical checks for violated constraints. While the worst case complexity of the auction algorithm is only slightly increased by the proposed extension, it was shown in numerical experiments that the average runtime can be decreased significantly. The speed-up was shown to consistently increase with problem size.

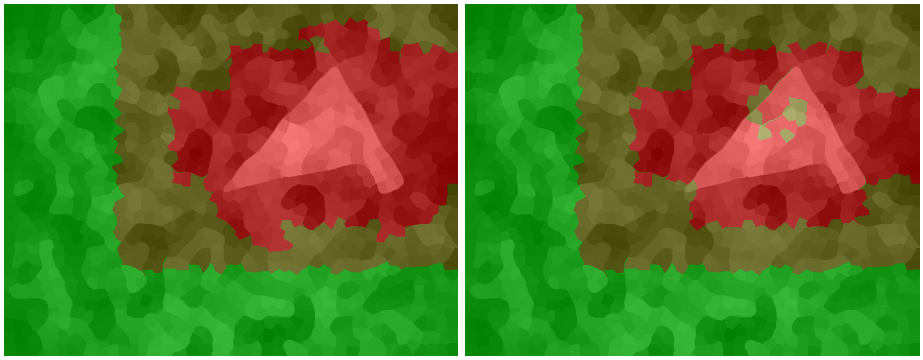


Figure 5.8: Illustration of applying the multi-scale scheme to the linearized Gromov-Wasserstein shape prior: Gray shading indicates foreground affinity (white ↔ foreground). The color coding shows the finest required resolution of the cost function (green indicating low resolution). **Left:** highest resolution among all template pixels. **Right:** highest resolution for a single template pixel. Points far from the true object location can be ruled out at coarse scales and do not require a finer resolution of the cost function. The rectangular structure of the boundaries between regions of different resolution are an artifact of the hierarchical quadtree partition structure.

Optimal transport was a central tool for the shape priors presented in Chapters 3 and 4, but is generally popular in image analysis and machine learning. Thus this Chapter is relevant beyond the topic of shape matching.

Development of hierarchical solvers for optimal transport is still in an early stadium. Several open questions naturally arise: Can the same hierarchical concepts be applied to different algorithms? Can a more precise relation between the regularity of the cost function and the observed speed-up be established? Furthermore, the development of an efficient implementation is important for practical application.

Chapter 6

Conclusion

Summary. With the work presented in this thesis we hope to have contributed to the development of shape priors that are both strong from a modelling as well as an optimization perspective. Two different approaches to achieve important isometry invariance have been presented, both based on the shape representation as metric measure spaces, thus allowing compatibility with the convex variational framework of image labelling.

In Chap. 3 an intrinsic variant, based on convex relaxation of the Gromov-Wasserstein distance was presented, making the computationally complex notion of metric shape matching applicable for the object segmentation problem. In cases with strong noise the relaxation proved to be too loose. Thus methods for counter balancing with more elaborate pseudo-local features, and in particular an approach to tightening the relaxation have been introduced.

Conversely, in Chap. 4 isometry invariance was obtained via explicitly optimizing over extrinsic isometries. For modelling of these transformations we employed the Riemannian structure of the 2-Wasserstein space. It was shown that restricting this structure to measures that model shapes, one can recover the manifold structure of contour embeddings, thus allowing to combine convex local matching with manifold based shape modelling with no need for online representation conversions. A shape prior based on a linearized tangent space approximation of the manifold was presented.

Finally, in Chap. 5 the computational side of the project was addressed systematically. A hierarchical sparse / dense hybrid variant of the auction algorithm for optimal transport and linear assignment problems was presented, that provided a significant speed-up on 'real world' problems, compared to the classical dense variant.

Future Work. Given the related literature on image labelling and shape matching, and the variety of tools and mathematical notions that were needed in this thesis, it becomes apparent that isometry invariant object segmentation is a challenging problem. It would be presumptuous to think it could be 'solved' just

like that. Of course the approaches presented in Chapters 3 to 5 are by no means finished. Open problems and further directions have already been pointed out at the respective places. We would like to emphasize three particular directions that seem most central to us:

- The optimization of assignments under geometric regularity constraints.
- Further investigation of the shape measure manifold: In particular the study of suitable metrics, geodesics and barycenters, to provide a stable basis for statistical analysis thereon. Also, the combination of matching based on geometry and appearance, as applied in Sect. 4.2, into a unified manifold model.
- The development of efficient corresponding numerical matching methods that heavily exploit the structure of 'good-natured' real world problems to significantly reduce the naïve combinatorial complexity.

The author hopes to be able to help tackling these questions in the future.

Bibliography

[AAR04] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 26(11):1475–1490, 2004.

[AC11] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

[ADN59] S. Agmon, A. Douglis, and L. Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general goundary conditions. *Comm. Pure Appl. Math.*, 12(4):623–727, 1959.

[AG13] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, volume 2062 of *Lect. Not. Math.*, pages 1–155. Springer, 2013.

[AGHH07] W. P. Adams, M. Guignard, P. M. Hahn, and W. Hightower. A level-2 reformulation-linearization technique bound for the quadratic assignment problem. *European Journal of Operational Research*, 180(3):983–996, 2007.

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. Birkhäuser Boston, 2nd edition, 2008.

[AMH11] S. Andrews, C. McIntosh, and G. Hamarneh. Convex multi-region probabilistic segmentation with shape prior in the isometric log-ratio transformation space. In *International Conference on Computer Vision (ICCV 2011)*, pages 2096 –2103, nov. 2011.

[BB97] J.-D. Benamou and Y. Brenier. A numerical method for the optimal time-continuous mass transport problem and related problems. In *Monge Ampère equation: applications to geometry and optimization*, volume 226 of *Contemp. Math.*, pages 1–11. Amer. Math. Soc., Providence, RI, 1997.

[BBK06] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Efficient computation of isometry-invariant distances between surfaces. *SIAM J. Sci. Comput.*, 28:1812–1836, 2006.

[BBK+10] A. M. Bronstein, M. M. Bronstein, R. Kimmel, M. Mahmoudi, and G. Sapiro. A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *Int. J. Comp. Vision*, 89:266–286, 2010.

[BC89] D. Bertsekas and D. Castanon. The auction algorithm for the transportation problem. *Annals of Operations Research*, 20:67–96, 1989.

[BC98] P. Blomgren and T. F. Chan. Color tv: total variation methods for restoration of vector-valued images. *IEEE Trans. Image Proc.*, 7(3):304–309, 1998.

[BC11] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 1 edition, 2011.

[BÇPP98] R. E. Burkard, E. Çela, P. M. Pardalos, and L. Pitsoulis. The quadratic assignment problem. In *Handbook of Combinatorial Optimization*, pages 241–238. Kluwer Acad. Publ., 1998.

[BE88] D. P. Bertsekas and J. Eckstein. Dual coordinate step methods for linear network flow problems. *Mathematical Programming, Series B*, 42:203–243, 1988.

[Ber79] D. P. Bertsekas. A distributed algorithm for the assignment problem. Technical report, Lab. for Information and Decision Systems Report, MIT, May 1979.

[Ber88] D. P. Bertsekas. The auction algorithm: A distributed relaxation method for the assignment problem. *Annals of Operations Research*, 14:105–123, 1988.

[BFH+13] B. Berkels, T. Fletcher, B. Heeren, M. Rumpf, and B. Wirth. Discrete geodesic regression in shape space. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR 2013)*, pages 108–122, 2013.

[BFS12] M. Burger, M. Franek, and C.-B. Schönlieb. Regularised regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress*, 3 2012.

[BK04] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in computer vision. *IEEE Trans. Patt. Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

[BKP10] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM J. Imaging Sci.*, 3(3):492–526, 2010.

[BKR97] R. Burkard, S. Karisch, and F. Rendl. QAPLIB - A Quadratic Assignment Problem Library. *J. Global Optim.*, 10:391–403, 1997.

[BMR13] M. Burger, J. Modersitzki, and L. Ruthotto. A hyperelastic regularization energy for image registration. *SIAM J. Sci. Comput.*, 35(1):B132–B148, 2013.

[Bre91] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[BV06] S. Burer and D. Vandenbussche. Solving lift-and-project relaxations of binary integer programs. *SIAM J. Optim.*, 16(3):726–750, 2006.

[BVZ01] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Patt. Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[CEN06] T. F. Chan, S. Esedoglu, and M. Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM J. Appl. Math.*, 66(5):1632–1648, 2006.

[CFK07] G. Charpiat, O. Faugeras, and R. Keriven. Shape statistics for image segmentation with prior. In *Computer Vision and Pattern Recognition (CVPR 2007)*, pages 1 –6, 2007.

[CG99] S. D. Cohen and L. J. Guibas. The earth mover's distance under transformation sets. In *International Conference on Computer Vision (ICCV 1999)*, pages 1076–1083, 1999.

[CGS10] G. Carlier, A. Galichon, and F. Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM J. Math. Anal.*, 41:2554–2576, 2010.

[CKS03] D. Cremers, T. Kohlberger, and C. Schnörr. Shape statistics in kernel space for variational image segmentation. *Patt. Recognition*, 36(9):1929–1943, 2003.

[CL06] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5 – 30, 2006.

[CRD07] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *Int. J. Comp. Vision*, 72(2):195–215, 2007.

[Cut13] M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. 2013. http://arxiv.org/abs/1306.0895.

[DHLM05] M. Desbrun, A. N. Hirani, M. Leok, and J. E. Marsden. Discrete exterior calculus. http://arxiv.org/abs/math/0508341, 2005.

[DZ11] M. C. Delfour and J.-P. Zolésio. *Shapes and geometries: analysis, differential calculus, and optimization*. Advances in Design and Control. SIAM Philadelphia, PA, USA, 2nd edition, 2011.

[Edm65] J. Edmonds. Maximum matching and a polyhedron with 0,1-vertices. *Journal of Research of the National Bureau of Standards*, 69B(1–2):125–130, 1965.

[EHW12] S. M. A. Eslami, N. Heess, and J. Winn. The shape Boltzmann machine: a strong model of object shape. In *CVPR*, 2012.

[EK03] A. Elad and R. Kimmel. On bending invariant signatures for surfaces. *IEEE Trans. Patt. Anal. Mach. Intell.*, 25(10):1285 – 1295, oct. 2003.

[FH05] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comp. Vision*, 61(5):55–79, 2005.

[FPPA13] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. http://arxiv.org/abs/1307.5551/, 2013. preprint.

[FXPA12] S. Ferradans, G.-S. Xia, G. Peyré, and J.-F. Aujol. Optimal transport mixing of Gaussian texture models. In *ICIP*, 2012.

[GC10] B. Goldluecke and D. Cremers. Convex relaxation for multilabel problems with product label spaces. In *European Conference on Computer Vision (ECCV 2010)*, 2010.

[GGS+06] L. Gorelick, M. Galun, E. Sharon, R. Basri, and A. Brandt. Shape representation and classification using the Poisson equation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 28(12):1991–2005, 2006.

[Gil62] P. C. Gilmore. Optimal and suboptimal algorithms for the quadratic assignment problem. *Journal of the Society for Industrial and Applied Mathematics*, 10(2):305–313, 1962.

[GL10] M. Grasmair and F. Lenzen. Anisotropic total variation filtering. *Applied Mathematics and Optimization*, 62(3):323–339, 2010.

[GM00] W. Gangbo and R. McCann. Shape recognition via Wasserstein distance. *Quart. Appl. Math.*, 58(4):705–737, 2000.

[GR86] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations*. Springer, 1986.

[GR01] C. Godsil and G. Royle. *Algebraic Graph Theory*. Springer, 2001.

[Gro07] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhäuser Boston, 2007.

[HA10] P. Hahn and M. Anjos. Qaplib home page. http://www.seas.upenn.edu/qaplib/, 2010.

[Hör90] L. Hörmander. *The Analysis of Linear Partial Differential Equations*. Springer, 2nd edition, 1990.

[HRWW12] B. Heeren, M. Rumpf, M. Wardetzky, and B. Wirth. Time-discrete geodesics in the space of shells. *Computer Graphics Forum*, 31(5):1755–1764, 2012.

[HZTA04] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *Int. J. Comp. Vision*, 60:225–240, December 2004.

[Kac66] M. Kac. Can one hear the shape of a drum? *The American Mathematical Monthly*, 73(4):1–23, 1966.

[KB57] T. C. Koopmans and M. J. Beckmann. Assignment problems and the location of economic activities. *Econometrica*, 25(1):53–76, 1957.

[KC11] M. Klodt and D. Cremers. A convex framework for image segmentation with moment constraints. In *International Conference on Computer Vision (ICCV 2011)*, pages 2236–2243, 2011.

[KÇCE99] S. E. Karisch, E. Çela, J. Clausen, and T. Espersen. A dual framework for lower bounds of the quadratic assignment problem based on linearization. *Computing*, 63:351–403, 1999.

[KM97] A. Kriegl and P. W. Michor. *The Convenient Setting of Global Analysis*, volume 53 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, 1997.

[Kol09] V. Kolmogorov. Blossom v: A new implementation of a minimum cost perfect matching algorithm. *Mathematical Programming Computation*, 1(1):43–67, 2009.

[KTZ05] M. P. Kumar, P. H. S. Torr, and A. Zisserman. OBJ CUT. In *Computer Vision and Pattern Recognition (CVPR 2005)*, volume 1, pages 18–25, 2005.

[Kuh55] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics*, 2:83–97, 1955.

[KV12] B. Korte and J. Vygen. *Combinatorial Optimization*. Springer, 5th edition, 2012.

[KZ04] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Patt. Anal. Mach. Intell.*, 26:147–159, 2004.

[Law63] E. L. Lawler. The quadratic assignment problem. *Management Science*, 9(4):586–599, 1963.

[LBR08] V. Lempitsky, A. Blake, and C. Rother. Image segmentation by branch-and-mincut. In *European Conference on Computer Vision (ECCV 2008)*, pages 15–29, 2008.

[Lot08] J. Lott. Some geometric calculations on Wasserstein space. *Comm. Math. Phys.*, 277:423–437, 2008.

[LS11] J. Lellmann and C. Schnörr. Continuous multiclass labeling approaches and algorithms. *SIAM J. Imaging Sci.*, 4(4):1049–1096, 2011.

[LVS10] X. Liu, O. Veksler, and J. Samarabandu. Order preserving moves for graph cut based optimization. *IEEE Trans. Patt. Anal. Mach. Intell.*, 32(7):1182–1196, 2010.

[Mém11] F. Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Found. Comp. Math.*, 11:417–487, 2011.

[Mér11] Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[MM06] P. W. Michor and D. Mumford. Riemannian geometries on spaces of plane curves. *Journal of the European Mathematical Society*, 8(1):1–48, 2006.

[MM07] P. W. Michor and D. Mumford. An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach. *Appl. Comput. Harmon. Anal.*, 23:74–113, 2007.

[MS05] F. Mémoli and G. Sapiro. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Found. Comp. Math.*, 5:313–347, July 2005.

[OFCD02] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions. *ACM Transactions on Graphics*, 21:807–832, 2002.

[Ott01] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.

[PCCB09] T. Pock, A. Chambolle, D. Cremers, and H. Bischof. A convex relaxation approach for computing minimal partitions. In *Computer Vision and Pattern Recognition (CVPR 2009)*, pages 810–817, 2009.

[PT13] O. Pele and B. Taskar. The tangent earth mover's distance. In *GSI*, 2013.

[PW08] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *European Conference on Computer Vision (ECCV 2008)*, 2008.

[PW09] O. Pele and W. Werman. Fast and robust Earth Mover's Distances. In *International Conference on Computer Vision (ICCV 2009)*, 2009.

[ROF92] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[RPC10] J. Rabin, G. Peyré, and L. D. Cohen. Geodesic shape retrieval via optimal mass transport. In *European Conference on Computer Vision (ECCV 2010)*, pages 771–784, 2010.

[RR98] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Volume I: Theory*, volume 1 of *Probability and its Applications*. Springer, 1998.

[RTG00] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comp. Vision*, 40(2):99–121, 2000.

[RW] M. Rumpf and B. Wirth. Variational time discretization of geodesic calculus. http://de.arxiv.org/abs/1210.2097.

[RWP06] M. Reuter, F.-E. Wolter, and N. Peinecke. Laplace-Beltrami spectra as "Shape-DNA" of surfaces and solids. *Computer-Aided Design*, 38(4):342–366, 2006.

[SC11] E. Strekalovskiy and D. Cremers. Generalized ordering constraints for multilabel optimization. In *International Conference on Computer Vision (ICCV 2011)*, 2011.

[Sch59] M. Schechter. General boundary value problems for elliptic partial differential equations. *Comm. Pure Appl. Math.*, 12(3):457–486, 1959.

[Sch03] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*, volume 24 of *Algorithms and Combinatorics*. Springer, 2003.

[Set95] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. In *Proc. Nat. Acad. Sci*, pages 1591–1595, 1995.

[SJ08] S. Shirdhonkar and D. W. Jacobs. Approximate earth mover's distance in linear time. In *Computer Vision and Pattern Recognition (CVPR 2008)*, 2008.

[SMSY11] G. Sundaramoorthi, A. Mennucci, S. Soatto, and A. Yezzi. A new geometric metric in the space of curves, and applications to tracking deforming objects by prediction and filtering. *SIAM J. Imaging Sci.*, 4(1):109–145, 2011.

[SS12] B. Schmitzer and C. Schnörr. Weakly convex coupling continuous cuts and shape priors. In *Scale Space and Variational Methods (SSVM 2011)*, pages 423–434, 2012.

[SS13a] B. Schmitzer and C. Schnörr. Contour manifolds and optimal transport. http://arxiv.org/abs/1309.2240, 2013. preprint.

[SS13b] B. Schmitzer and C. Schnörr. A hierarchical approach to optimal transport. In *Scale Space and Variational Methods (SSVM 2013)*, pages 452–464, 2013.

[SS13c] 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.

[SS13d] B. Schmitzer and C. Schnörr. Object segmentation by shape matching with Wasserstein modes. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR 2013)*, pages 123–136, 2013.

[Tro98] A. Trouvé. Diffeomorphisms groups and pattern matching in image analysis. *Int. J. Comp. Vision*, 28(3):213–221, 1998.

[TY05] A. Trouvé and L. Younes. Metamorphoses through lie group action. *Found. Comp. Math.*, 5(2):173–198, 2005.

[Vil03] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, 2003.

[Vil09] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

[WJ08] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[WSB+12] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comp. Vision*, 101:254–269, 2012.

[WSSC11] T. Windheuser, U. Schlickewei, F. R. Schmidt, and D. Cremers. Geometrically consistent elastic matching of 3D shapes: A linear programming solution. In *International Conference on Computer Vision (ICCV 2011)*, 2011.

[YMSM08] L. Younes, P. W. Michor, J. Shah, and D. Mumford. A metric on shape space with explicit geodesics. *Rend. Lincei Mat. Appl.*, 9:25–57, 2008.

[You10] L. Younes. *Shapes and Diffeomorphisms*, volume 171 of *Applied Mathematical Sciences*. Springer, 2010.

[YV13] B. Yangel and D. Vetrov. Learning a model for shape-constrained image segmentation from weakly labeled data. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR 2013)*, pages 137–150, 2013.

[Zhu07] Y.-R. Zhu. *Recent Advances and Challenges in Quadratic Assignment and Related Problems*. PhD thesis, University of Pennsylvania, 2007.