Numerical Schemes for Calculating the Discrete Wasserstein Distance

Benjamin Cabrera

Born 9th April 1991 in Groß-Gerau, Germany

February 26, 2016

Master's Thesis Mathematics

Advisor: Prof. Dr. Martin Rumpf

Mathematical Institute

Mathematisch-Naturwissenschaftliche Fakultät der

Rheinischen Friedrich-Wilhelms-Universität Bonn

Contents

1 Introduction

The goal of this thesis is to derive a feasible numerical scheme for computing time-discretized geodesics of a recently proposed new type of "Wasserstein" metric acting on discrete rather than on continuous spaces. Using an actual implementation of this scheme we aim to compute the geodesics for different problem instances and present the results to learn more about the underlying geometry this metric induces on discrete spaces.

The Wasserstein metric is closely tied to the concept of optimal transportation, a theory that has sparked a lot of interest from researchers around the world. A first version of the optimal transport problem was stated by Monge [Mon81] in 1781. He imagined it as the problem of transporting an entity (e.g. sand, dirt, etc.) from a start to an end configuration in optimal way with respect to some cost function describing how expensive the transport of a particle is. This problem was then generalized to a more well-posed problem in the more rigorous context of probability measures by Kantorovich [Kan42; Kan48] in 1948. In the following decades the topic was developed further and the Wasserstein distance was introduced as a special case of Kantorovich's problem [Was69] where the cost function is just an usual L^p-norm. The topic then regained a lot of traction around the end of the 20th century after some ground breaking papers by Otto, Jordan and Kinderlehrer [JKO98] and Benamou and Brenier [BB00] were published. Otto et al. researched gradient flows in the space of probability measures with respect to the Wasserstein distance and found a relation to particular partial differential equations that suggests that the Wasserstein metric is indeed a very good choice as a metric on probability measures. On the other hand Benamou and Brenier introduced a fluid mechanics view of the Wasserstein space. Later this led to an interpretation of the Wasserstein space as a Riemannian manifold and the Wasserstein distance as a Riemannian distance in this context. Today optimal transport is studied from the theoretical geometric point of view [Erb10; ASZ07; OS08; Gig10] as well as from the numerical side which tries to find numerical schemes for different applications [RPDB12; RPC10; BL15].

The problem we are focusing on in this thesis is the Wasserstein distance on discrete (finite) sets. We will see later that the original Wasserstein distance looses some of its important geometric properties when applied to a discrete setting. As a result recently different authors [Maa11; Mie11; CHLZ12] suggested to define a new metric for the discrete case heavily influenced by the original Benamou-Brenier formula for the Wasserstein distance on continuous spaces but still different in some critical aspects. In this thesis we solely focus on the version of the metric by Maas [Maa11]. Although it has been shown theoretically that the new metric has the desired properties little is still known about the actual shape of the induced space and especially the behaviour of the geodesics. The idea of this thesis is now to compute geodesics numerically, then review the results, and finally learn about the space and which properties it exhibits.

To this end we will adapt two of the numerical schemes that have been used for computing geodesics for the continuous Wasserstein distance to our discrete setting. On the one hand there is the Benamou-Brenier method described in [BB00] right after they suggested the flow formulation of the Wasserstein distance. On the other hand there is the more recent proximal splitting approach by Papadakis, Peyré and Oudet [PPO14]. Both require some differential operations to be carried out (partial integration, etc.) which is trivial in the continuous $\mathbb{R}^d$ case but has to be carried out with care in a discrete setting. However, doing so will lead us to a

feasible way of computing the geodesics for the discrete Wasserstein distance and we will be able to compare the results for different instances.

1.1 Outline

We begin in Chapter 2 by recalling the fundamental definitions and statements from optimal transportation theory. After that in Chapter 3 we introduce the mentioned new metric along with its general setting on the discrete space. In the subsequent chapter we describe our numerical approaches to the problem of computing the geodesics for the new metric. It turns out that one part of our algorithms will pose the most problems for us in that the computation is quite unstable. We deal with the theory behind this projection problem in Chapter 5. In the following chapter we explain some of the discretization and implementation details that were needed to turn the numerical schemes into working programs. In Chapter 7 we finally present the results that were attained using our computations. Also we try to start interpreting them and give some interesting insights into the geometry behind the new metric. Finally in Chapter 8 we summarize what we have done and point out some possible future improvements of our algorithms.

2 An Introduction to Optimal Transport

In this chapter we will give a brief introduction to optimal transportation theory, the Wasserstein distance and the associated Wasserstein space with its geometric properties. We will introduce the optimal transport models by Monge and Kantorovich as well as the alternative formulations by Benamou and Brenier which gives a more geometric interpretation of the underlying space and can be used to numerically compute the transportation distance. Finally we explain the geometric connection between some differential operators and the Wasserstein distance. For a proper introduction in the theory of optimal transport we refer to some of the excellent books by Villani [Vil09; Vil03] or for a more analytical perspective to the works of Ambrosio, Gigli and Savaré [AGS05]. Most of the definitions and theorems in this chapter are taken from one of these sources.

2.1 From Monge to Kantorovich

In 1781 Monge [Mon81] introduced a basic transportation problem which should be the earliest mention of such a problem in the context of optimal transportation. The problem can be pictured as a pile of sand that should be transported to another pile of the exact same volume. This transport should happen in an optimal way in the sense that the cost of transporting the sand should be minimal with respect to some metric, i.e. a grain of sand should not be transported further than necessary.

To formulate this problem more mathematically let $\mathcal{P}(\mathbb{R}^d)$ denote the set of all probability measures on the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ on $\mathbb{R}^d$. We encode the piles of sand at the beginning and at the end as probability distributions μ_A, $\mu_B \in \mathcal{P}(\mathbb{R}^d)$ and the cost of transporting a grain of sand as a function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We get the original Monge problem by asking for a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that the sum of all transport cost $c(x, T(x))$ is minimal. For a mathematically rigorous problem formulation we first introduce the push forward measure.

Definition 2.1. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be a Borel probability measure on $\mathbb{R}^d$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ Borel measurable. Then we define the push-forward $T_{\#}\mu$ as*

$$T_{\#}\mu(E) = \mu(T(E)) \quad \forall \mathcal{B}(\mathbb{R}^d).$$

Now we can define Monge's problem.

Definition 2.2 (Monge's problem)**.** *Let $\mu_A, \mu_B \in \mathcal{P}(\mathbb{R}^d)$. Then a solution to Monge's problem is given by a Borel measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ that realises*

$$\inf \left\{ \int_{\mathbb{R}^d} c(x, T(x)) \, \mathrm{d}\mu_A(x) : T_{\#}\mu_A = \mu_B \right\}. \tag{2.1}$$

The constraint $T_{\#}\mu_A = \mu_B$ enforces that mass is preserved during the transport.

Monge's problem formulation has some disadvantages. In particular no mass can be split up. In the example of transporting sand this might make sense because we see a grain of sand as the smallest atomic entity that can not be split up, however in general there might not exist a solution to Monge's problem. For example consider $\mu_A = \delta_{x_0}$ a Dirac measure and $\mu_B = \frac{1}{2}(\delta_{x_1} + \delta_{x_2})$ $(x_1 \neq x_2)$. Obviously for this example the push-forward condition can never be satisfied for any map T.

To really pose a well-defined problem Kantorovich [Kan42] [Kan48] proposed a more general setup. In the approach by Kantorovich instead of transport maps $T : \mathbb{R}^d \to \mathbb{R}^d$ he described the transport plan through couplings on the product space.

Definition 2.3 (Coupling)**.** *Let μ_1 and μ_2 be Borel probability measures on $\mathbb{R}^d$. Then we call a Borel probability measure $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ on the product space $\mathbb{R}^d \times \mathbb{R}^d$ a coupling of μ_1 and μ_2 if its marginals coincide with μ_1, resp. μ_2, i.e.*

$$p_{1\#}\pi = \mu_1 \quad \text{and} \quad p_{2\#}\pi = \mu_2,$$

where p_i is the projection on the i-th component. We denote the set of all couplings of μ_1, μ_2 as $\Pi(\mu_1, \mu_2)$.

Using these couplings we can define Kantorovich's problem.

Definition 2.4 (Kantorovich's problem)**.** *Let $\mu_A, \mu_B \in \mathcal{P}(\mathbb{R}^d)$. Then a solution to Kantorovich's problem is given by a coupling $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ which realises the following infimum.*

$$\inf \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \, \mathrm{d}\pi(x, y) : \pi \in \Pi(\mu_A, \mu_B) \right\}.$$

Remark: *Note that Kantorovich's problem is a generalization of Monge's problem. Indeed let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a minimizer of (2.1) then $\pi = (\mathrm{id} \times T)_{\#}\mu_A$ is a coupling in $\Pi(\mu_A, \mu_B)$ and*

$$\int_{\mathbb{R}^d} c(x, T(x)) \, \mathrm{d}\mu_A = \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \, \mathrm{d}((\mathrm{id} \times T)_{\#}\mu_A)(x, y).$$

In contrast to Monge's problem using the direct method of the calculus of variations one can show the existence of transport plans for all marginals μ_A, μ_B.

Lemma 2.5. *Let $\mu_A, \mu_B \in \mathcal{P}(\mathbb{R}^d)$ and let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \cup \infty$ be lower semi-continuous and bounded from below. Then the infimum of Kantorovich's problem is always attained by a transport plan $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$.*

These concepts of optimal transport can now be used to introduce a metric on the space of Borel probability measures in the sense that the distance between two measures should just be the transport cost for an optimal coupling. Using cost functions of the type $c(x, y) = |x - y|^p$, $p \in [1, \infty)$ the resulting metric is called the Wasserstein distance, named after Leonid Vaserstein who introduced the concept in 1969.

Definition 2.6 (Wasserstein distance)**.** *Denote by $\mathcal{P}_p(\mathbb{R}^d)$ the space of Borel probability measures on $\mathbb{R}^d$ with finite p-th moments, i.e. for all $\mu \in \mathcal{P}_p(\mathbb{R}^d)$*

$$\int_{\mathbb{R}^d} |x|^p \, \mathrm{d}\mu < \infty.$$

Let $\mu_A, \mu_B \in \mathcal{P}_p(\mathbb{R}^d)$ then we define the L^p-Wasserstein distance as

$$W_p(\mu_A, \mu_B)^p = \inf \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p \, \mathrm{d}\pi(x, y) : \pi \in \Pi(\mu_A, \mu_B) \right\}.$$

Lemma 2.7. *The L^p-Wasserstein distance is a metric on $\mathcal{P}_p(\mathbb{R}^d)$.*

It is of course legitimate to ask if this metric on the space of probability measures is in any sense meaningful. Indeed it turns out that the Wasserstein metric carries very natural properties. One might be its close relation to the so called *narrow convergence* which is nothing else than the weak*-convergence on $\mathcal{P}(\mathbb{R}^d)$.

Definition 2.8 (Narrow convergence)**.** *Let $\mu_n, \mu \in \mathcal{P}(\mathbb{R}^d)$, then we say μ_n converges **narrowly** to μ and write $\mu_n \rightharpoonup \mu$ if*

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} f \, \mathrm{d}\mu_n \to \int_{\mathbb{R}^d} f \, \mathrm{d}\mu \quad \forall f \in \mathcal{C}_b(\mathbb{R}^d).$$

One can show the following properties for the Wasserstein metric and its relation to narrow convergence.

Theorem 2.9 (Narrow lower semi-continuity of W_p)**.** *Let $\mu_n, \nu_n \in \mathcal{P}_p(\mathbb{R}^d)$ converging narrowly to $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Then*

$$\liminf_{n\to\infty} W_p(\mu_n, \nu_n) \geq W_p(\mu, \nu).$$

One can even show a more precise characterization of this relation.

Theorem 2.10. *Let $\mu_n, \mu \in \mathcal{P}_p(\mathbb{R}^d)$, then the following are equivalent*

1. *$W_p(\mu_n, \mu) \to 0$*
2. *$\mu_n \rightharpoonup \mu$ as $n \to \infty$ and*

$$\lim_{R\to\infty} \limsup_{k\to\infty} \int_{|x|\geq R} |x|^p \, \mathrm{d}\mu_k = 0.$$

This tells us that the Wasserstein metric is closely related to the narrow topology on probability measures. This topology however is very useful for working with probability measures. All this gives motivation for building a theory around the so called Wasserstein spaces $(\mathcal{P}_p(\mathbb{R}^d), W_p)$.

2.2 The Wasserstein distance in a Riemannian context

In 2000 Benamou and Brenier [BB00] introduced a reformulation of the Monge problem in the setting of fluid mechanics. The original purpose of this work was to rewrite the original Monge problem in a way such that one can find a numerical scheme for explicitly calculating the Wasserstein distance. However, with this new formulation the Wasserstein distance could be interpreted as a Riemannian distance, hence the Wasserstein space as a infinite dimensional Riemannian manifold. This interpretation has given rise to many new ideas around optimal transport and today this reformulation is widely accepted as a milestone in development of optimal transport theory. In this section we will follow this path and interpret the Wasserstein space as a Riemannian manifold. The rigorous statements are discussed in [AGS05] but here we assume some basic knowledge of the theory of Riemannian manifolds and also skip some of the more technical steps.

As Benamou and Brenier point out in [BB00] their idea of seeing optimal transport as a flow of mass in time was not completely new. Indeed already Monge had originally mentioned the problem in this way but then proposed his formulation as a way of getting rid of the time dependence. It turns out that at least for computational purposes the flow formulation is much more viable as it results in the problem being a much more simple convex space-time minimization with linear constraints instead of solving a highly nonlinear PDE (Monge-Ampère equation).

From now on we will solely focus one the case $p = 2$, i.e. the L^2-Wasserstein distance. This has several reasons, the main one being that because L^2 is a Hilbert space we don't have to deal with special dual spaces. We now want to look at curves in the space $\mathcal{P}_2(\mathbb{R}^d)$, i.e. functions that map a time $t \in [0,1]$ to $\mathcal{P}_2(\mathbb{R}^d)$ in a "continuous" way. However, the first thing we have to do is define what such curves in a metric space do really look like.

Definition 2.11 (Absolutely continuous curves and the metric derivative)**.** *Let (X,d) be a metric space. A curve $\gamma : (0,1) \to X$ is said to be **(locally) absolutely continuous** of order $p \in [1,\infty]$, if there exists $m \in L^p_{loc}((a,b))$ such that*

$$d(\gamma(s),\gamma(t)) \leq \int_s^t m(r)\ \mathrm{d}r \qquad \forall\ 0 < s \leq t < 1.$$

*The set of all such curves is denoted by $\mathcal{AC}^p((0,1);X)$. Further we define the **metric derivative** of γ for $t \in (a,b)$ as $|\gamma'| : (0,1) \to [0,\infty]$ with*

$$|\gamma'(t)| := \lim_{s \to t} \frac{d(\gamma(s),\gamma(t))}{|s-t|}$$

if the limit exists.

Remark: *For the Wasserstein space one can show the existence of the metric derivative for $\mathcal{L}^1$-a.e. $t \in (0,1)$.*

So the curves we are going to deal with in $\mathcal{P}_2(\mathbb{R}^d)$ will be absolutely continuous curves in the above sense. The next step towards a Riemannian setting for the Wasserstein space is a beautiful characterization of absolutely continuous curves in $\mathcal{P}_2(\mathbb{R}^d)$ as curves that satisfy some form of continuity equation.

Definition 2.12 (Continuity equation)**.** *Let $\mu : (0,1) \to \mathcal{P}_2(\mathbb{R}^d)$ and $v : (0,1) \times \mathbb{R}^d \to \mathbb{R}^d$ be such that*

$$\int_a^b \int_{\mathbb{R}^d} |v_t(x)| \, \mathrm{d}\mu_t \, \mathrm{d}t < \infty \qquad \forall \, 0 < a < b < 1.$$

Then μ is a solution to the continuity equation with respect to v, if $\partial_t \mu_t + \operatorname{div}(v_t \mu_t) = 0$ holds in $(0,1) \times \mathbb{R}^d$ in the sense of distributions, i.e. for all $\varphi \in C_c^\infty((0,1) \times \mathbb{R}^d)$

$$\int_0^1 \int_{\mathbb{R}^d} \Big(\partial_t \varphi(t,x) + (v_t(x), \nabla \varphi(t,x)) \Big) \, \mathrm{d}\mu_t(x) \, \mathrm{d}t = 0.$$

We denote the set of such pairs (μ, v) that satisfy the continuity equation by $\overline{\mathcal{CE}}$.

Before we rigorously state the relation between absolutely continuous curves in $\mathcal{P}_2(\mathbb{R}^d)$ and the continuity equation we introduce the **tangent velocity space**

$$\mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d) := \{ w \in L^2_\mu(\mathbb{R}^d) : \operatorname{div}(w\mu) = 0 \text{ distributionally } \}^\perp$$

At first this definition might seem a bit random because we skipped some steps of its derivation (which can be found in [AGS05]). However, the next theorem should shed some light on this.

Theorem 2.13. *Suppose we have given $\mu : (0,1) \to \mathcal{P}_2(\mathbb{R}^d)$. Then μ is absolutely continuous if and only if there exists a vector field $v : (0,1) \times \mathbb{R}^d \to \mathbb{R}^d$ with $v_t \in \mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d)$ for almost every $t \in (0,1)$ such that*

1. *$(\mu, v) \in \overline{\mathcal{CE}}$,*
2. *$\|v_t\|_{L^2_{\mu_t}(\mathbb{R}^d)} = |\mu'|(t)$ for almost every $t \in (0,1)$.*

This theorem is useful for defining a Riemannian metric on $\mathcal{P}_2(\mathbb{R}^d)$ because now we have a characterization of the formal derivative $\partial_t \mu_t$ of an absolutely continuous curve $(\mu_t)_t$ through a vector field v_t in a space which is not that abstract anymore. So using this indirect way we define the tangent space for the Riemannian metric as

$$\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d) := \{ s \in \mathbb{D}(\mathbb{R}^d) : \exists v \in \mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d) \text{ s.t. } s + \operatorname{div}(v\mu) = 0 \}.$$

Here $\mathbb{D}(\mathbb{R}^d)$ denotes the set of distributions on $\mathbb{R}^d$. On this tangent space we can finally define the metric tensor for $s_1, s_2 \in \mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$ as

$$g_\mu(s_1, s_2) := \int_{\mathbb{R}^d} (v_1(x), v_2(x)) \, \mathrm{d}\mu(x)$$

where v_1 and v_2 are the corresponding tangent velocity fields to s_1 and s_2 and $(\cdot, \cdot)$ is the standard euclidean inner product.

The famous Benamou-Brenier formulation of the L^2-Wasserstein distance is now nothing else than the Riemannian distance in the above setting.

Theorem 2.14 (Benamou-Brenier formula)**.** *Let $\mu_A, \mu_B \in \mathcal{P}_2(\mathbb{R}^d)$. Then*

$$\begin{aligned} W_2(\mu_A,\mu_B)^2 &= \inf\left\{\int_0^1 \|v_s\|_{L^2_{\mu_t}(\mathbb{R}^d)}\ \mathrm{d}s : t \mapsto \mu_t \in \mathcal{AC}^2((0,1);\mathcal{P}_2(\mathbb{R}^d)), \mu_0=\mu_A, \mu_1=\mu_B\right\} \\ &= \inf\left\{\int_0^1 \|v_s\|_{L^2_{\mu_t}(\mathbb{R}^d)}\ \mathrm{d}s : (\mu,v) \in \overline{\mathcal{CE}}, \mu_0=\mu_A, \mu_1=\mu_B\right\}. \end{aligned} \tag{2.2}$$

Remark: *For absolute continuous probability measures $\mu_A, \mu_B \in \mathcal{P}_2(\mathbb{R}^d)$, $\mu \ll \mathcal{L}^d$ one can show that the geodesic stays absolutely continuous almost everywhere. Then (2.2) can be written as*

$$W_2(\mu_A,\mu_B)^2 = \inf\left\{\int_0^1 \int_\Omega |v_t(s)|^2 \rho_t(s)\ \mathrm{d}s\ \mathrm{d}t : (\mu,v) \in \overline{\mathcal{CE}}, \mu_0=\mu_A, \mu_1=\mu_B\right\},$$

where $\rho_t \mathcal{L}^d = \mu_t$. In this case we can do a change of variable by replacing the velocity v with the momentum $m(t,x) = \rho(t,x)v(t,x)$. We introduce the action

$$\alpha(x,y) = \begin{cases} \frac{|x|^2}{y} & y > 0 \\ 0 & x = y = 0 \\ \infty & \text{otherwise} \end{cases}$$

and then rewrite

$$W_2(\rho_A,\rho_B)^2 = \inf\left\{\int_0^1 \int_\Omega \alpha(m(t,x),\rho(t,x))\ \mathrm{d}x\ \mathrm{d}t : (\rho,m) \in \overline{\mathcal{CE}}', \mu_0=\mu_A, \mu_1=\mu_B\right\}.$$

Here $\overline{\mathcal{CE}}'$ is the set where $\partial_t \rho + \operatorname{div} m = 0$ has to hold in the sense of distributions. In the future we will work with this formulation and refer to it as the Benamou-Brenier formulation.

2.3 Gradient-Flows with respect to the Wasserstein distance

In this final section on the foundations of the Wasserstein distance we want to quickly explain the relation between gradient flows of certain functions with respect to the Wasserstein metric and certain partial differential equations. This relation will once more justify our interest in the Wasserstein distance and especially mark a property that we want to carry over to the discrete case we are introducing in the next chapter. The property was first observed in a seminal work by Otto et al [JKO98]. For this section we assume some basic understandings of gradient flows, especially in context of subdifferentials.

We start with a definition of gradient flows in our context.

Definition 2.15 (Gradient flows in the Wasserstein space)**.** *Let $\mu \in \mathcal{AC}^2((0,1);\mathcal{P}_2(\mathbb{R}^d))$ with $(\mu_t)_{t\in(0,1)} \subset \mathcal{P}_2^a(\mathbb{R}^d)$ and corresponding tangent velocity field v. Let $\Phi : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,\infty]$ be a proper and lower semi-continuous functional. Then μ is called gradient flow for Φ, if*

$$-v_t \in \partial\Phi(\mu_t) \qquad \text{for a.e. } t \in (0,1).$$

Further we call $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ the initial value of μ, if

$$\lim_{t\downarrow 0} W_2(\mu_t,\mu_0) = 0.$$

Of course there are many different interesting choices for potential functionals Φ. We only look at the relative entropy as the resulting gradient flow is very interesting. We define the **relative entropy** by

$$\mathcal{H}(\mu) = \begin{cases} \int_{\mathbb{R}^d} \rho(x) \log(\rho(x)) \, \mathrm{d}x & \mu \ll \mathcal{L}^d \text{ and } \rho\mathcal{L}^d = \mu \\ \infty & \text{otherwise} \end{cases}$$

Now we can state the surprising result that relates the gradient flow of $\mathcal{H}$ with the heat equation.

Theorem 2.16. *Let $\mu : (0, \infty) \to \mathcal{P}_2(\mathbb{R}^d)$ be continuous with respect to the Wasserstein distance. Then μ is the gradient flow for $\mathcal{H}$, if and only if, for each $t > 0$, μ_t has a density ρ_t with respect to the Lebesgue measure such that $(\rho_t)_{t>0}$ is a weak solution to the heat equation*

$$\partial_t \rho - \Delta \rho = 0 \qquad \text{in } (0, \infty) \times \mathbb{R}^d$$

satisfying $\int_s^t \int_{\mathbb{R}^d} \frac{|\nabla \rho|^2}{\rho} \, \mathrm{d}x \, \mathrm{d}t < \infty$ for all $0 < s < t < \infty$.

This theorem is very interesting because it connects different objects from different parts of mathematics with each other in a very natural way. Gradient flows can be interpreted as flows that are always moving "downhill" the potential functional in the direction of the steepest descent. So in this setting one might have the idea that a gradient flow with respect to the entropy leads to some kind of diffusion process, smoothing out the initial configuration. However, that the Wasserstein metric is exactly the correct metric to get something as trivial as the heat equation speaks for its important position as a metric on probability measures.

So far we have only worked with probability measures in $\mathbb{R}^d$. However, we will see in the next chapter that many of the properties we derived do not hold if the underlying space is discrete.

3 A 'Wasserstein-like' Metric for Discrete Spaces

After we have recalled some basics of optimal transport and geometric properties of the L^2-Wasserstein distance in the previous chapter we now want to focus on the particular case where the underlying metric space is discrete. We will first see that unfortunately the geometric properties of the original L^2-Wasserstein distance do not carry over if the underlying space is discrete. Instead recently different authors have suggested alternative definitions of metrics for discrete spaces that do have similar geometric properties than the original L^2-Wasserstein distance on continuous spaces [CHLZ12; Maa11; Mie11]. Although these definitions represent very similar concepts we will in this thesis solely focus on the metric introduced in [Maa11] by Jan Maas.

In section 3.1 we will first look at the original Wasserstein distance on discrete spaces and why it looses its nice geometric properties. After that in section 3.2 we introduce the new general setting, the space involved and some basic properties of it. As a result in section 3.3 we can now rigorously define the new metric. Finally in section 3.4 we look at some duality result that will be the foundation of all numerical schemes discussed in the next chapter.

3.1 The original Wasserstein metric on discrete spaces

We have already learned that the L^2-Wasserstein metric introduces a particular geometry on the space it is applied to. First of all it is perfectly valid to take its definition and apply it to a discrete setting where the underlying space $\mathcal{X}$ is a finite set. However, when we recall the definition of the L^2-Wasserstein distance

$$W_2(\mu_A, \mu_B)^2 = \inf \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, \mathrm{d}\pi(x,y) : \pi \in \Pi(\mu_A, \mu_B) \right\}.$$

we can already spot some potential problems. It may be that a cost function like $c(x,y) = |x-y|^2$ causes problems on a discrete space where there are no continuous "connections" to transport mass over. Indeed - as will be shown in the following example from [Maa11] - there exist no constant speed geodesics between two different elements of $\mathcal{X}$.

Example. *Assume $\mathcal{X} = \{a, b\}$, i.e. the discrete space contains two elements. Then a probability distribution ρ on $\mathcal{X}$ can be characterized by just one parameter $\beta \in [-1, 1]$ by $\rho^\beta = \frac{1}{2}((1-\beta)\delta_a + (1+\beta)\delta_b)$. The mass at a and b is given by*

$$\rho^\beta(a) = \frac{1-\beta}{2}, \quad \rho^\beta(a) = \frac{1+\beta}{2}.$$

Note that because we are only dealing with Dirac measures we can explicitly calculate

$$W_2(\rho^\alpha, \rho^\beta)^2 = \sqrt{2|\beta - \alpha|}.$$

Now assume that $\left(\rho^{\beta(t)}\right)_{0 \le t \le 1}$ is a constant speed geodesic. This means that for $s, t \in [0,1]$ we have

$$\sqrt{2|\beta(t) - \beta(s)|} = W_2(\rho^{\beta(t)}, \rho^{\beta(s)}) = |t-s| W_2(\rho^{\beta(0)}, \rho^{\beta(1)}) = |t-s|\sqrt{2|\beta(0) - \beta(1)|}.$$

Thus we get that β is 2-Hölder and thus constant on $[0,1]$. As a result all constant speed geodesics are already constant.

This lack of non-constant geodesic tells us that the original Wasserstein metric is not the correct metric to study the evolution of probability measures on a discrete spaces. In the next section we will introduce a possible solution by defining a different metric with nevertheless similar properties to the original Wasserstein metric.

3.2 The discrete setting and basic Markov chain properties

In order to solve the problem we encountered in the last section we will define a new metric that uses more implicit features of the underlying space to introduce a better fitting geometry. However, in order to be able to define this new metric we have to provide more information about the discrete space in the form of a transition kernel and a stationary distribution.

Remark: *As already mentioned in the introduction of this chapter there have been other attempts in defining such a metric on a discrete space. However, all have used some additional concepts providing additional information for setting in which the metric is defined.*

From now on $\mathcal{X}$ will always be a finite set which will be the state space of a Markov chain with transition kernel Q. We recall the following properties of Markov chains.

Definition 3.1. *A mapping $Q : \mathcal{X} \times \mathcal{X} \to [0,1]$ is a **transition kernel** if $Q(x,y) \geq 0$ for all $x, y \in \mathcal{X}$ and $\sum_{y\in\mathcal{X}} Q(x,y) = 1$ for all x.*

*A Markov chain with transition kernel Q is called **irreducible** if one can get from each state to each other state with positive probability. Equivalently if we interpret Q as a quadratic matrix than the Markov chain is irreducible if there exists $n > 0$ such that $Q^n > 0$ meaning that each entry is greater than zero.*

*We call π a **stationary distribution** of the Markov chain if*

$$\sum_{x\in\mathcal{X}} \pi(x) = 1 \qquad \text{and} \qquad \pi(y) = \sum_{x\in\mathcal{X}} \pi(x)Q(x,y).$$

We know from basic Markov chain theory that if a chain on a finite space is irreducible there also exists a unique stationary distribution π.

Definition 3.2. *A Markov chain is called **reversible** if the **detailed balance condition***

$$Q(x,y)\pi(x) = Q(y,x)\pi(y) \tag{3.1}$$

holds for all $x, y \in \mathcal{X}$ where π is the stationary distribution.

Remark: *For a more natural understanding we will often refer to elements of $\mathcal{X}$ as **nodes** of a graph and pairs $(x,y) \in \mathcal{X} \times \mathcal{X}$ with $Q(x,y) > 0$ as **transitions** or **edges** of that graph.*

Lets assume that we are given a irreducible, reversible Markov chain with transition kernel Q and stationary distribution. The domain on which the new metric will then act is

$$\mathcal{P}(\mathcal{X}) := \left\{ \rho : \mathcal{X} \to \mathbb{R}_+ : \sum_{x\in\mathcal{X}} \pi(x)\rho(x) = 1 \right\},$$

the set of all probability densities on $\mathcal{X}$ with relative to the stationary π.

We have introduced the Markov chains and transition matrices because on top of the finite set $\mathcal{X}$ they will determine how "connected" the space is and heavily influence the underlying geometry of the space. Because the new metric will be introduced as a flow formulation similar to the Benamou-Brenier formula we have to provide some discrete differential operators on the space so that we can later perform analogue computations similar to the continuous setting.

Definition 3.3. *Let $\varphi, \psi \in \mathbb{R}^{\mathcal{X}}$ and $\Phi, \Psi \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ be functions on the nodes, resp. edges. Then we define the* ***discrete gradient*** *by*

$$\nabla_\chi \psi(x,y) = \psi(x) - \psi(y),$$

and the ***discrete divergence*** *by*

$$\operatorname{div}_\chi \Psi(x) = \frac{1}{2}\sum_{y\in\mathcal{X}} Q(x,y)(\Psi(y,x) - \Psi(x,y)).$$

We define the ***discrete Laplace-operator*** *by*

$$\begin{aligned}\Delta_\chi \psi(x) = \operatorname{div}_\chi(\nabla_\chi \psi) &= \frac{1}{2}\sum_{y\in\mathcal{X}} Q(x,y)\left[(\psi(y) - \psi(x)) - (\psi(x) - \psi(y))\right] \\ &= \sum_{y\in\mathcal{X}} Q(x,y)\left[\psi(y) - \psi(x)\right] = ((Q - I)\psi)(x).\end{aligned}$$

Remark: *Note that the discrete divergence imitates the continuous divergence in that both can be interpreted as measuring for a vector field the difference of how much is flowing in and out a certain point in space.*

Definition 3.4. *Define the inner products*

$$\begin{aligned}\langle \varphi, \psi\rangle_\pi &= \sum_{x\in\mathcal{X}} \varphi(x)\psi(x)\pi(x), \\ \langle \Phi, \Psi\rangle_\pi &= \frac{1}{2}\sum_{x,y\in\mathcal{X}} \Phi(x,y)\Psi(x,y)Q(x,y)\pi(x).\end{aligned}$$

Remark: *Notation-wise we do not explicitly distinguish between the inner product of nodal functions and those on edges. However, which one to use should always be clear from the situation.*

Definition 3.5. *Define $\mathcal{H} = \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ and for all $v = (\rho_1, m_1), w = (\rho_2, m_2) \in \mathcal{H}$ set*

$$\langle v, w\rangle_{\mathcal{H}} = \langle \rho_1, \rho_2\rangle_\pi + \langle m_1, m_2\rangle_\pi.$$

Although the previous definitions were of course a matter of choice and might seem arbitrary the following property shows that our discrete setting is intrinsically well-defined.

Lemma 3.6 (Integration by parts)**.** *The following integration by parts formula holds*

$$\langle \nabla_\chi \varphi, \Psi \rangle_\pi = -\langle \varphi, \operatorname{div}_\chi \Psi \rangle_\pi.$$

Proof. Calculation yields

$$\begin{aligned}
\langle \nabla_\chi \varphi, \Psi \rangle_\pi &= \frac{1}{2} \sum_{x,y \in \mathcal{X}} (\nabla_\chi \varphi)(x,y) \Psi(x,y) Q(x,y) \pi(x) \\
&= \frac{1}{2} \sum_{x,y \in \mathcal{X}} [\varphi(x) - \varphi(y)] \, \Psi(x,y) Q(x,y) \pi(x) \\
&= \frac{1}{2} \sum_{x,y \in \mathcal{X}} \varphi(x) \Psi(x,y) Q(x,y) \pi(x) - \frac{1}{2} \sum_{x,y \in \mathcal{X}} \varphi(y) \Psi(x,y) Q(x,y) \pi(x) \\
&= \frac{1}{2} \sum_{x,y \in \mathcal{X}} \varphi(x) \Psi(x,y) Q(x,y) \pi(x) - \frac{1}{2} \sum_{x,y \in \mathcal{X}} \varphi(x) \Psi(y,x) \underbrace{K(y,x) \pi(y)}_{= Q(x,y) \pi(x)} \\
&= \frac{1}{2} \sum_{x \in \mathcal{X}} \pi(x) \varphi(x) \sum_{y \in \mathcal{X}} (\Psi(x,y) - \Psi(x,y)) Q(x,y) \\
&= -\frac{1}{2} \sum_{x \in \mathcal{X}} \varphi(x) (\operatorname{div}_\chi \Psi)(x) \pi(x) = -\langle \varphi, \operatorname{div}_\chi \Psi \rangle_\pi
\end{aligned}$$

where we used the detailed balance condition of the stationary distribution π in the fifth step. ■

3.3 Equivalent definitions of the metric

In this section we can now rigorously define the metric as described in [Maa11]. The idea of this definition is to use the Benamou-Brenier flow formulation of the original L^2-Wasserstein distance as a starting point and translate the used concepts into the discrete setting. This is exactly the reason why we defined inner products and differential operators on discrete sets in the previous section.

There will be however one major difference between the original Benamou-Brenier formulation and the definition of our metric. It turns out that in order to get the nice geometric properties we are looking for we have to make transportation costs from one node to another depend also on how much mass is already at the target and how much is still left at the source. This kind of behaviour will be enforced by introducing an additional mean term and we will start by defining its properties.

Definition 3.7. *We assume the* ***mean*** $\theta : [0,\infty) \times [0,\infty) \to [0,\infty)$ *to satisfy the following conditions*

(A1) θ *is continuous on* $[0,\infty) \times [0,\infty)$*,*

(A2) θ *is* C^∞ *on* $(0,\infty) \times (0,\infty)$*,*

(A3) $\theta(s,t) = \theta(t,s)$ *for* $s,t \geq 0$*,*

(A4) $\theta(s,t) > 0$ *for* $s,t > 0$*.*

(A5) $\theta(0,t) = 0$ *for all* $t \geq 0$

(A6) $\theta(r,t) \leq \theta(s,t)$ *for all* $0 \leq r \leq s$ *and* $t \geq 0$

(A7) $\theta(\lambda s, \lambda t) = \lambda\theta(s,t)$ *for* $\lambda > 0$ *and* $s,t \geq 0$

(A8) $\theta : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ *is concave*

Remark: *The θ fulfilling (A1) to (A8) we are going to use in this thesis is the so called logarithmic mean which we will rigorously introduce at the end of this section.*

At this point we want to note that there are two equivalent definitions of the metric. The first one was described in [Maa11] and represents the original Benamou-Brenier formula in the discrete setting. The second one is a reformulation analogue to the one mentioned in the remark after Theorem 2.14 which offers some advantages (e.g. linear continuity equation) and was described in [EM12]. We will only use the second formulation in this thesis because it allows a more natural numerical approach of the problem.

Definition 3.8 (Action)**.** *We introduce a function* $\alpha : \mathbb{R} \times [0,1] \times [0,1] \to \mathbb{R}$ *with*

$$\alpha(x,s,t) = \begin{cases} 0 & \theta(s,t) = 0 \text{ and } x = 0 \\ \frac{x^2}{\theta(s,t)} & \theta(s,t) \neq 0 \\ +\infty & \theta(s,t) = 0 \text{ and } x \neq 0. \end{cases}$$

Then we define the ***action*** *as*

$$\mathcal{A}(\rho, m) := \frac{1}{2} \sum_{x,y \in \mathcal{X}} \alpha(m_{x,y}, \rho(x), \rho(y)) Q(x,y) \pi(x).$$

Remark: *Note that the action essentially represents the integrand in the Benamou-Brenier formulation, i.e.* $\mathcal{A}(\rho, m) \approx \|\frac{m}{\bar{\rho}}\|_\pi^2$ *where* $\bar{\rho}(x,y) = \theta(\rho(x), \rho(y))$.

Next also similar to the flow formulation of the original Wasserstein distance we have to define what it means to satisfy the continuity equation on a graph.

Definition 3.9 (Discrete continuity equation)**.** *Let* $\rho \in [0,1] \times \mathbb{R}^{\mathcal{X}}$ *and* $m \in [0,1] \times \mathbb{R}^{\mathcal{X} \times \mathcal{X}}$*, then we say that* (ρ, m) *satisfy the continuity equation and write* $(\rho, m) \in \mathcal{CE}(\rho_A, \rho_B)$ *if*

1. $t \mapsto \rho(t, \cdot)$ *is continuous*
2. $\rho(0, \cdot) = \rho_A$, $\rho(1, \cdot) = \rho_B$
3. $\rho(t, \cdot) \in \mathcal{P}(\mathcal{X})$ *for all* $t \in [0,1]$
4. $m(\cdot, x, y) : [0,1] \to \mathbb{R}$ *is in* $L^1(0,1)$
5. *For all* $x \in \mathcal{X}$ *we have in the sense of distributions*

$$\partial_t \rho_t + \operatorname{div}_\chi m = 0$$

or respectively in the sense of distributions

$$\partial_t \rho_t(x) + \frac{1}{2} \sum_{x \in \mathcal{X}} (m_t(x,y) - m_t(y,x)) Q(x,y) = 0. \tag{3.2}$$

Beside the integrability constraints and the continuity of $\rho_t(x)$ in time one can see this definition as a constraint that satisfies that mass is preserved and no new mass is generated and as a result all mass that leaves a node has to flow over an edge to another node.

Remark: *Comparing this continuity equation with the previous one of the continuous case we observe that we can now assume continuity of $t \mapsto \rho(t,\cdot)$ instead of the absolute continuity before. This is the case because we are now working on a finite set where a converging sequence of continuous $\rho^{(t)}$ would immediately result in a continuous limit (because of uniform convergence). This was not the case for a continuous space.*

We can now give the actual definition of the new metric which we will from now on call the discrete Wasserstein metric.

Definition 3.10 (Discrete Wasserstein metric)**.** *We define the metric by*

$$\mathcal{W}(\rho_A, \rho_B)^2 = \inf \left\{ \int_0^1 \mathcal{A}(\rho_t, m_t) \, \mathrm{d}t : (\rho, m) \in \mathcal{CE}(\rho_A, \rho_B) \right\}. \tag{3.3}$$

Remark: *Notice that as pointed out in [GM12] if (ρ, m) satisfies the discrete continuity equation* (3.2) *the same holds true for the pair (ρ, m^{asym}) where $m_t^{asym}(x,y) = \frac{1}{2}(m_t(x,y) - m_t(y,x))$ is the anti-symmetrization of m. Also*

$$\mathcal{A}(\rho_t, m_t^{asym}) \leq \mathcal{A}(\rho_t, m_t)$$

and thus we just have to consider anti-symmetric m (i.e. $m_t(x,y) = -m(y,x)$ for all $x, y \in \mathcal{X}$) in the infimum of (3.3)*.*

The logarithmic mean and its properties

It turns out that in order to get that gradient flows of the entropy with respect to the new metric are heat flows we have to choose θ as the logarithmic mean. As we will use this function heavily and one might not be familiar with it we give a short summary of definition and properties of the logarithmic mean. Also note that the logarithmic mean introduces a high level of "nonlinearity" to the discrete Wasserstein distance. This will later be a main challenge when finding a feasible computational scheme.

Definition 3.11. *The logarithmic mean is defined as*

$$\theta(s,t) = \int_0^1 s^{1-r} t^r \, \mathrm{d}r = \lim_{(\xi,\eta) \to (s,t)} \frac{\eta - \xi}{\log(\eta) - \log(\xi)} = \begin{cases} 0, & \text{if } s = 0 \text{ or } t = 0 \\ s, & \text{if } s = t \\ \frac{t-s}{\log(t) - \log(s)} & \text{otherwise} \end{cases}.$$

One can easily calculate its derivatives.

Lemma 3.12. *The first partial derivatives of the logarithmic mean are given by*

$$\partial_1\theta(s,t) = \begin{cases} +\infty & \text{if } s = 0 \\ 0 & \text{if } t = 0 \\ 0.5 & \text{if } s = t \neq 0 \\ \frac{-s\log(t/s)-s+t}{s\log^2(t/s)} & \text{otherwise} \end{cases}, \quad \partial_2\theta(s,t) = \begin{cases} 0 & \text{if } s = 0 \\ +\infty & \text{if } t = 0 \\ 0.5 & \text{if } s = t \neq 0 \\ \frac{-t\log(t/s)+s-t}{t\log^2(t/s)} & \text{otherwise} \end{cases}$$

Also we immediately get the following properties.

Corollary 3.13. *The logarithmic mean satisfies assumptions (A1) to (A8) and additionally has the following properties for all $s, t > 0$*

(i) $\partial_1\theta(s,t) = \partial_1\theta(cs,ct)$ *and* $\partial_2\theta(s,t) = \partial_2\theta(cs,ct)$ *for all* $c > 0$

(ii) $\partial_1\theta(s,t) = \partial_2\theta(t,s)$

(iii) $s\partial_1\theta(s,t) + t\partial_2\theta(s,t) = \theta(s,t)$

(iv) $\partial_1\theta(u,v) + t\partial_2(u,v) \geq \theta(s,t)$ *for all* $u, v > 0$

Figures 3.1 and 3.2 should help get a better understanding of the logarithmic mean and its derivatives. Most importantly note that, as the name is already implying, the logarithmic mean has a logarithmic slope when approaching zero from one side. In particular this means that there is a singularity at zero which can be problematic when our numerical schemes need to find solutions that are close to zero.

Finally the following formal theorem is rigorously proved in [Maa11] which gives the justification for the non-trivial construction we did on the metric.

Theorem 3.14 (Heat flow is gradient flow of the entropy)**.** *Let θ be the logarithmic mean and let $\rho \in \mathcal{P}(\mathcal{X})$. Then the heat flow $t \mapsto e^{t\Delta}\rho$ is a gradient flow trajectory for the entropy*

$$\mathcal{H}(\rho) = \sum_{x\in\mathcal{X}} \pi_x \rho_x \log(\rho_x)$$

with respect to the discrete Wasserstein distance $\mathcal{W}$.

3.4 The dual of the action

For the computational schemes we will need the dual of $\mathcal{A}$ calculated in the next lemma.

Lemma 3.15 ([EM, Lemma 2.3])**.** *The Fenchel-dual of $\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is given by*

$$\mathcal{A}^*(\rho^*, m^*) = \sup_a \left\{ \frac{1}{8} \sum_{x,y\in\mathcal{X}} |m^*_{x,y}|^2 \theta(a_x, a_y) Q(x,y)\pi(x) + \sum_{x\in\mathcal{X}} a_x \rho^*_x \pi(x) \right\} = \mathcal{I}_{\mathcal{K}},$$

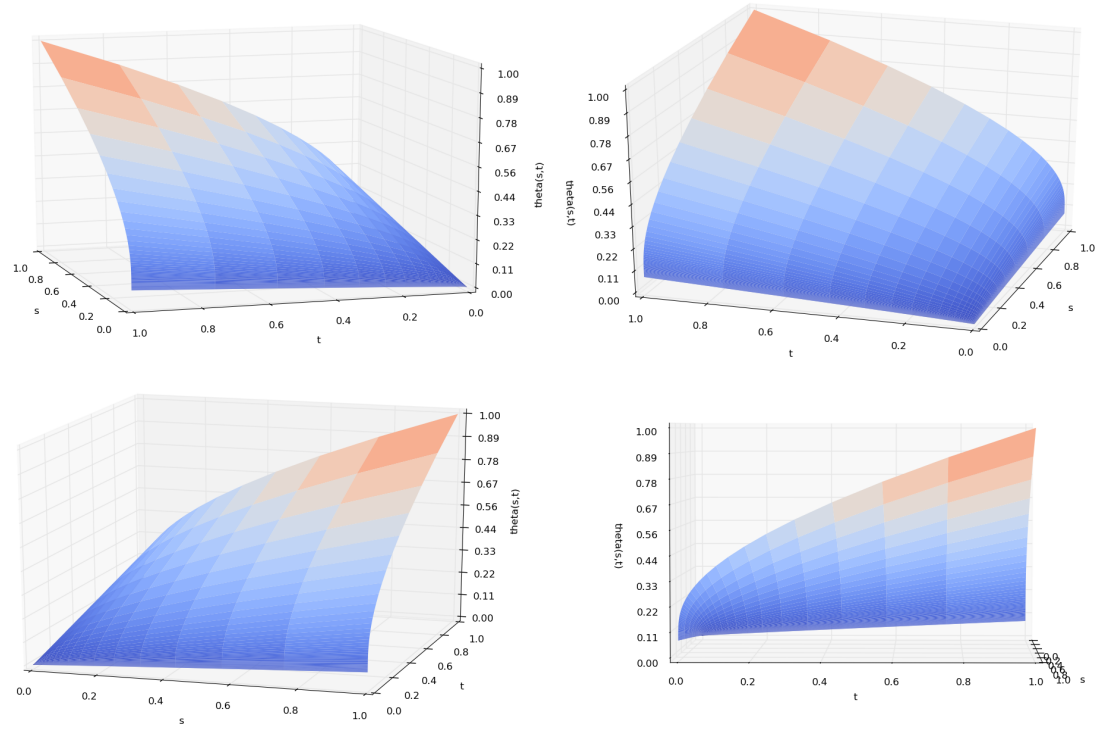


Figure 3.1: Some impressions of the logarithmic mean. Note the logarithmic slope near zero.

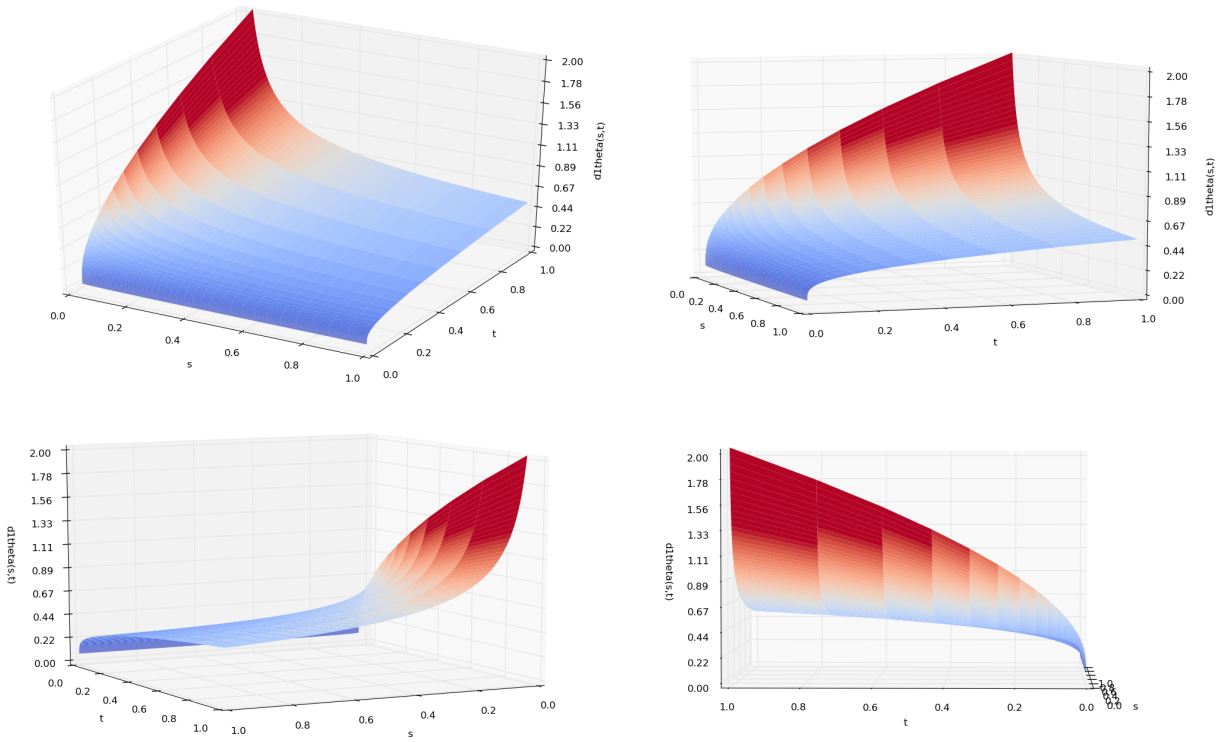


Figure 3.2: Some impressions of $\partial_1\theta$ (derivative in the first variable of the logarithmic mean).

where the indicator function $\mathcal{I}_K$ is given by

$$\mathcal{I}_{\mathcal{K}} = \begin{cases} 0, & (\rho^*, m^*) \in \mathcal{K} \\ +\infty, & else \end{cases}.$$

Here $\mathcal{K}$ denotes a convex set where $(\rho^*, m^*) \in \mathcal{K}$ if one of the following characterizations holds.

(K1) For all $a \in \mathbb{R}_+^{\mathcal{X}}$ we have

$$\sum_{x\in\mathcal{X}} a_x \rho_x^* \pi(x) + \frac{1}{8} \sum_{x,y\in\mathcal{X}} \theta(a_x, a_y)|m_{x,y}^*|^2 Q(x,y)\pi(x) \leq 0 \tag{3.4}$$

(K2) For all $x, y \in \mathcal{X}$ there exists $a_1, a_2 \in \mathbb{R}_+$ such that

$$\rho_x^* \leq -\frac{1}{4}\partial_1\theta(a_1, a_2)|m_{x,y}^*|^2 Q(x,y)$$
$$\rho_y^* \leq -\frac{1}{4}\partial_2\theta(a_1, a_2)|m_{x,y}^*|^2 Q(x,y)$$

(K3) It is sufficient (but not necessary) for $(\rho^*, m^*) \in \mathcal{K}$ that there exists $a \in \mathbb{R}_+^{\mathcal{X}}$ such that for all $x \in \mathcal{X}$

$$\rho_x^* + \frac{1}{4} \sum_{y\in\mathcal{X}} \partial_1\theta(a_x, a_y)|m_{x,y}^*|^2 Q(x,y) \leq 0.$$

Proof. We start by computing the Fenchel dual of $\mathcal{A}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.

$$\begin{aligned}
\mathcal{A}^*(\rho^*, m^*) &= \sup_{(a,w)} \left\{ \langle a, \rho^* \rangle_\pi + \langle w, m^* \rangle_\pi - \mathcal{A}(a,w) \right\} \\
&= \sup_{(a,w)} \left\{ \sum_{x\in\mathcal{X}} a_x \rho_x^* \pi(x) + \frac{1}{2} \sum_{x,y\in\mathcal{X}} w_{x,y} m_{x,y}^* Q(x,y)\pi(x) - \frac{1}{2} \sum_{x,y\in\mathcal{X}} \frac{|w_{x,y}|^2}{\theta(a_x,a_y)} Q(x,y)\pi(x) \right\} \\
&= \sup_{(a,w)} \left\{ -\frac{1}{2} \sum_{x,y\in\mathcal{X}} \left[\frac{|w_{x,y}|^2}{\theta(a_x,a_y)} - m_{x,y}^* w_{x,y} \right] Q(x,y)\pi(x) + \sum_{x\in\mathcal{X}} a_x \rho_x^* \pi(x) \right\} \\
&= \sup_{(a,w)} \left\{ \frac{1}{2} \sum_{x,y\in\mathcal{X}} \left(-\frac{1}{\theta(a_x,a_y)} \left[w_{x,y} - \frac{1}{2} m_{x,y}^* \theta(a_x,a_y) \right]^2 + \frac{1}{4}|m_{x,y}^*|^2 \theta(a_x,a_y) \right) Q(x,y)\pi(x) \right. \\
&\qquad \left. + \sum_{x\in\mathcal{X}} a_x \rho_x^* \pi(x) \right\}
\end{aligned}$$

Now note that since

$$-\frac{1}{\theta(a_x,a_y)} \left[w_{x,y} - \frac{1}{2} m_{x,y}^* \theta(a_x,a_y) \right]^2 \leq 0$$

it is optimal to choose $w_{x,y}$ such that it is exactly equal to zero. Then the supremum only runs over a and in total we get

$$\mathcal{A}^*(b,u) = \sup_a \left\{ \frac{1}{8} \sum_{x,y\in\mathcal{X}} |m_{x,y}^*|^2 \theta(a_x,a_y) Q(x,y)\pi(x) + \sum_{x\in\mathcal{X}} a_x \rho_x^* \pi(x) \right\}.$$

Because of (A7) and $|m_{x,y}^*|^2\theta(a_x,a_y)Q(x,y)\pi(x) \geq 0$ the quantity is positive homogeneous in a. Thus the value of the supremum is $+\infty$ unless (K1) holds.

We will now show the equivalence of (K1) and (K2). Suppose (K1) holds, then for fixed $x, y \in \mathcal{X}$ and $a_1, a_2 \in \mathbb{R}_+$ we set $a \in \mathbb{R}^{\mathcal{X}}$ as

$$a_z = \begin{cases} a_1 & \text{if } z = x \\ a_2 & \text{if } z = y \\ 0 & \text{otherwise} \end{cases}.$$

Testing (K1) with this a we get

$$a_1\rho_x^*\pi(x) + a_2\rho_y^*\pi(y) + \frac{1}{4}|m_{x,y}^*|^2\theta(a_1, a_2)Q(x, y)\pi(x) \leq 0. \tag{3.5}$$

Now we can use (iii) of Corollary 3.13 to split $\theta(a_1, a_2) = a_1\partial_1\theta(a_1, a_2) + a_2\partial_2\theta(a_1, a_2)$ and (3.5) becomes

$$a_1\rho_x^*\pi(x) + a_2\rho_y^*\pi(y) \leq -a_1\frac{1}{4}|m_{x,y}^*|^2\partial_1\theta(a_1, a_2)Q(x, y)\pi(x) - a_2\frac{1}{4}|m_{x,y}^*|^2\partial_2\theta(a_1, a_2)Q(x, y)\pi(x). \tag{3.6}$$

This condition obviously holds if

$$\begin{aligned} \rho_x^* &\leq -\frac{1}{4}|m_{x,y}^*|^2\partial_1\theta(a_1, a_2)Q(x, y) \\ \rho_y^* &\leq -\frac{1}{4}|m_{x,y}^*|^2\partial_2\theta(a_1, a_2)Q(x, y) \end{aligned} \tag{3.7}$$

which is what (K2) states. More the other direction move back from (3.7) to (3.6) and then use (iv) of Corollary 3.13 to make the step back to (3.5). Now summing over all x, y (3.5) implies (K1).

Finally we want to show that (K3) is sufficient for (K1). Calculating the derivative of (3.4) with respect to some a_z we get

$$\begin{aligned} &\partial_{a_z}\left(\sum_{x\in\mathcal{X}} a_x\rho_x^*\pi(x) + \frac{1}{8}\sum_{x,y\in\mathcal{X}}\theta(a_x, a_y)|m_{x,y}^*|^2Q(x, y)\pi(x)\right) \\ &= \rho_z^*\pi(z) + \frac{1}{8}\sum_{y\in\mathcal{X}}\partial_1\theta(a_z, a_y)|m_{z,y}^*|^2K(z, y)\pi(z) + \frac{1}{8}\sum_{x\in\mathcal{X}}\partial_2\theta(a_x, a_z)|m_{x,z}^*|^2K(x, z)\pi(x) \\ &= \rho_z^*\pi(z) + \frac{1}{8}\sum_{y\in\mathcal{X}}\partial_1\theta(a_z, a_y)|m_{y,z}^*|^2K(z, y)\pi(z) + \frac{1}{8}\sum_{x\in\mathcal{X}}\partial_2\theta(a_x, a_z)|m_{x,z}^*|^2K(z, x)\pi(z) \\ &= \rho_z^*\pi(z) + \frac{1}{4}\sum_{y\in\mathcal{X}}\partial_1\theta(a_z, a_y)|m_{y,z}^*|^2K(z, y)\pi(z) \end{aligned}$$

where we used the remark after Definition 3.10, the reversibility of the Markov chain (3.1) and the third property of Corollary 3.13. Now if there exists that an $a \in \mathbb{R}^{\mathcal{X}}$ such that (3.4) is strictly greater than zero because of the homogeneity in a there has to exists an $z \in \mathcal{X}$ for which we will get a positive derivative, i.e.

$$\rho_z^* + \frac{1}{4}\sum_{y\in\mathcal{X}}\partial_1\theta(a_z, a_y)|m_{y,z}^*|^2K(z, y) > 0.$$

This is a contradiction to (K3) and because we started with the negated (K1) we have shown (K3). ■

Remark: *By using (i) of Corollary 3.13 we can weaken the condition (K3) a bit. To be precise if there exists* $a \in \mathbb{R}^{\mathcal{X}}$ *such that*

$$\rho_x^* + \frac{1}{4} \sum_{y \in \mathcal{X}} \partial_1 \theta(a_x, a_y) |m_{x,y}^*|^2 Q(x,y) \leq 0 \qquad \forall x \in \mathcal{X}$$

then there exist infinitely many such a*, in particular also an* $a \in \mathbb{R}^{\mathcal{X}}$ *with* $\|a\|_1 = 1$.

4 Numerical Approaches for Calculating the Geodesic

The goal of this chapter, and this thesis in general, is to construct feasible numerical methods for computing the geodesics of the new discrete Wasserstein metric introduced in the previous chapter. We will use two different approaches as a starting point. On one hand in section 4.1 we will modify the augmented Lagrangian method proposed by Benamou and Brenier in 2000 [BB00] for the original Wasserstein distance to fit in our new setting. On the other hand in section 4.2 we will approach the problem by using a proximal splitting method inspired by [PPO14] where the authors applied this method to the original Wasserstein distance.

4.1 Benamou-Brenier's Augmented-Lagrangian Method

When in 2000 Benamou and Brenier published there now famous paper [BB00] it contained the first practical numerical scheme for computing an arbitrary L^2-Wasserstein distance. Since then the Benamou-Brenier algorithm has been the first practical method for computing the Wasserstein distance we will also apply it to the discrete Wasserstein distance.

Because the discrete Wasserstein distance is defined by an Benamou-Brenier type formula we do not have to reformulate the problem but can instead directly go for the numerical scheme. The idea of the Benamou-Brenier approach is to incorporate the continuity equation constraint directly into the equation such that we get a saddle point problem. After that we can calculate the optimality conditions of this saddle point problem and use an alternating descent method to approach the saddle point. We start by deriving the saddle point problem in our new discrete setting.

Theorem 4.1. *We can reformulate the discrete Wasserstein metric as*

$$\mathcal{W}(\rho_A, \rho_B)^2 = \sup_{\rho, m} \inf_{\xi, \rho^*, m^*} L[\rho, m, \rho^*, m^*, \xi]$$

where the Lagrangian L is defined by

$$\begin{aligned} L[\rho, m, \rho^*, m^*, \xi] = \mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*, m^*) + \langle \xi_0, \rho_A \rangle_\pi - \langle \xi_1, \rho_B \rangle_\pi \\ + \int_0^1 \Big(\langle \partial_t \xi_t - \rho_t^*, \rho_t \rangle_\pi + \langle \nabla_\chi \xi_t - m_t^*, m_t \rangle_\pi \Big) \, \mathrm{d}t \end{aligned}$$

and the infimum and supremum run over all $\rho, \rho^* \in L^2([0,1], \mathbb{R}^{\mathcal{X}})$, $m, m^* \in L^2([0,1], \mathbb{R}^{\mathcal{X}\times\mathcal{X}})$ and $\xi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$.

Proof. By introducing a Lagrange multiplier $\xi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$ for the continuity equation constraint we can rewrite

$$\mathcal{W}(\rho_A, \rho_B)^2 = \inf_{(\rho,m)\in[0,1]\times\mathcal{H}} \sup_{\xi} \Big\{ \underbrace{\int_0^1 \mathcal{A}(\rho_t, m_t) \, \mathrm{d}t}_{\mathcal{E}_{\mathrm{Trans}}} + \underbrace{\int_0^1 \langle \xi_t, \partial_t \rho_t + \mathrm{div}_\chi \, m_t \rangle_\pi \, \mathrm{d}t}_{\mathcal{E}_{\mathrm{CE}}} \Big\}.$$

The continuity equation is enforced because, since $\mathcal{E}_{\mathrm{CE}}$ is homogeneous in ξ, if $\partial_t \rho_t + \operatorname{div} m_t$ is not equal to zero the inner supremum will be $+\infty$. However, this is a case that is ignored by the outer infimum.
Now we calculate

$$\begin{aligned}
\mathcal{E}_{\mathrm{CE}} &= \int_0^1 \langle \xi_t, \partial_t \rho_t \rangle_\pi \, \mathrm{d}t + \int_0^1 \langle \xi_t, \operatorname{div}_\chi m_t \rangle_\pi \, \mathrm{d}t \\
&= \int_0^1 \sum_{x \in \mathcal{X}} \xi_t(x) \partial_t \rho_t(x) \pi(x) \, \mathrm{d}t + \int_0^1 \langle \xi_t, \operatorname{div}_\chi m_t \rangle_\pi \, \mathrm{d}t \\
&= \sum_{x \in \mathcal{X}} \pi(x) \int_0^1 \xi_t(x) \partial_t \rho_t(x) \, \mathrm{d}t + \int_0^1 \langle \xi_t, \operatorname{div}_\chi m_t \rangle_\pi \, \mathrm{d}t \\
&= \sum_{x \in \mathcal{X}} \left[\xi_1(x)\rho_1(x) - \xi_0(x)\rho_0(x) - \int_0^1 \partial_t \xi_t(x) \rho_t(x) \, \mathrm{d}t \right] \pi(x) + \int_0^1 \langle \xi_t, \operatorname{div}_\chi m_t \rangle_\pi \, \mathrm{d}t \\
&= \langle \xi_1, \rho_B \rangle_\pi - \langle \xi_0, \rho_A \rangle_\pi - \int_0^1 \langle \partial_t \xi_t, \rho_t \rangle_\pi \, \mathrm{d}t - \int_0^1 \langle \nabla_\chi \xi_t, m_t \rangle_\pi \, \mathrm{d}t
\end{aligned}$$

where we used both, the common integration by parts in t and the integration by parts formula in x of Lemma 3.6. Note that formally for this computation we assumed that $t \mapsto (\rho_t, m_t)$ is differentiable (otherwise $\partial_t \rho$ would not be defined). However, because the continuity equation of Definition 3.9 only asks for $\partial_t \rho_t + \operatorname{div}_\chi m = 0$ to hold in the sense of distributions $\mathcal{E}_{\mathrm{CE}}$ should actually only be used in integrated form with ρ and m in L^2.

Similar to the continuous L^2-Wasserstein case in [BB00] we now use the convexity of the action $\mathcal{A}$ by replacing $\mathcal{A}^{**} = \mathcal{A}$ to rewrite the Lagrangian. From basic convex analysis we know that for convex, lower semi-continuous functionals in a reflexive space ($\mathcal{H}$ is a Hilbert space) we have

$$\mathcal{A}(\rho, m) = \mathcal{A}^{**}(\rho, m) = \sup_{(\rho^*, m^*) \in \mathcal{H}} \langle \rho^*, \rho \rangle_\pi + \langle m^*, m \rangle_\pi - \mathcal{I}_{\mathcal{K}}(\rho^*, m^*).$$

Now we replace the $\mathcal{A}(\rho, m)$ of the $\mathcal{E}_{\mathrm{Trans}}$ functional by $\mathcal{A}^{**}(\rho, m)$ and get

$$\begin{aligned}
\mathcal{E}_{\mathrm{Trans}} &= \int_0^1 \left[\sup_{(\rho_t^*, m_t^*) \in \mathcal{H}} \langle \rho_t^*, \rho_t \rangle_\pi + \langle m_t^*, m_t \rangle_\pi - \mathcal{I}_{\mathcal{K}}(\rho_t^*, m_t^*) \right] \mathrm{d}t \\
&= \sup_{(\rho^*, m^*) \in [0,1] \times \mathcal{H}} \left[-\mathcal{I}_{[0,1] \times \mathcal{K}}(\rho^*, m^*) + \int_0^1 \left(\langle \rho_t^*, \rho_t \rangle_\pi + \langle m_t^*, m_t \rangle_\pi \right) \mathrm{d}t \right] \\
&= \sup_{(\rho^*, m^*) \in [0,1] \times \mathcal{K}} \left[\int_0^1 \left(\langle \rho_t^*, \rho_t \rangle_\pi + \langle m_t^*, m_t \rangle_\pi \right) \mathrm{d}t \right]
\end{aligned}$$

Note that we did not yet justify why we can pull the supremum out of the integral. This is due to the continuity of $\mathcal{A}(\rho, m)$ and continuity of $t \mapsto \rho_t$ that is assumed for the continuity equation in Definition 3.9. These continuities do not allow cases where $\mathcal{I}_{\mathcal{K}}(\rho_t^*, m_t^*) = \infty$ for just a null set of t's. Thus we can pull $\mathcal{I}_{\mathcal{K}}(\rho_t^*, m_t^*)$ out of the integral as $\mathcal{I}_{[0,1] \times \mathcal{K}}(\rho^*, m^*)$ and because of the linearity of $\langle \cdot, \cdot \rangle_\pi$ we can pull out the supremum. Also because later in practice we are working with time-discretized versions of the functions the integral will decouple in time and we can optimize for the supremum in each timestep separately.

Now plugging all this in, we can rewrite the saddle point problem as

$$
\begin{aligned}
&-\inf_{\rho,m}\sup_{\xi}\left\{\mathcal{E}_{\text{Trans}}+\mathcal{E}_{\text{CE}}\right\} = \sup_{\rho,m}\inf_{\xi}\left\{-\mathcal{E}_{\text{Trans}}-\mathcal{E}_{\text{CE}}\right\}\\
&= \sup_{\rho,m}\inf_{\xi}\Bigg\{-\sup_{\rho^*,m^*}\left[-\mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*,m^*)+\int_0^1\left(\langle\rho_t^*,\rho_t\rangle_\pi+\langle m_t^*,m_t\rangle_\pi\right)\,\mathrm{d}t\right]\\
&\qquad\qquad-\left[\langle\xi_1,\rho_B\rangle_\pi-\langle\xi_0,\rho_A\rangle_\pi-\int_0^1\langle\partial_t\xi_t,\rho_t\rangle_\pi\,\mathrm{d}t-\int_0^1\langle\nabla_\chi\xi_t,m_t\rangle_\pi\,\mathrm{d}t\right]\Bigg\}\\
&= \sup_{\rho,m}\inf_{\xi,\rho^*,m^*}\Bigg\{\mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*,m^*)-\int_0^1\left(\langle\rho_t^*,\rho_t\rangle_\pi+\langle m_t^*,m_t\rangle_\pi\right)\,\mathrm{d}t\\
&\qquad\qquad-\left[\langle\xi_1,\rho_B\rangle_\pi-\langle\xi_0,\rho_A\rangle_\pi-\int_0^1\langle\partial_t\xi_t,\rho_t\rangle_\pi\,\mathrm{d}t-\int_0^1\langle\nabla_\chi\xi_t,m_t\rangle_\pi\,\mathrm{d}t\right]\Bigg\}\\
&= \sup_{\rho,m}\inf_{\xi,\rho^*,m^*}\Bigg\{\mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*,m^*)+\langle\xi_0,\rho_A\rangle_\pi-\langle\xi_1,\rho_B\rangle_\pi\\
&\qquad\qquad-\int_0^1\left(\langle\rho_t^*-\partial_t\xi_t,\rho_t\rangle_\pi+\langle m_t^*-\nabla_\chi\xi_t,m_t\rangle_\pi\right)\,\mathrm{d}t\Bigg\}\\
&= \sup_{\rho,m}\inf_{\xi,\rho^*,m^*}\Bigg\{\mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*,m^*)+\langle\xi_0,\rho_A\rangle_\pi-\langle\xi_1,\rho_B\rangle_\pi\\
&\qquad\qquad+\int_0^1\left(\langle\partial_t\xi_t-\rho_t^*,\rho_t\rangle_\pi+\langle\nabla_\chi\xi_t-m_t^*,m_t\rangle_\pi\right)\,\mathrm{d}t\Bigg\}
\end{aligned}
$$

■

So far we have not looked into how to solve the saddle point problem we derived in the previous theorem. As already mentioned we want to calculate the optimality conditions for the problem and then do an alternating gradient descent to find the saddle point. However, to this end we need to compute derivatives of the Lagrangian L. In analogy to the Benamou-Brenier paper [BB00] we will look at the *augmented Lagrangian* which gives us a regularized version of L for which the optimality conditions can be computed numerically. The augmented Lagrangian is derived by observing that part of the Lagrangian L, namely

$$\int_0^1\left(\langle\partial_t\xi_t-\rho_t^*,\rho_t\rangle_\pi+\langle\nabla_\chi\xi_t-m_t^*,m_t\rangle_\pi\right)\,\mathrm{d}t,$$

can be interpreted in a way where ρ, m are actually the Lagrange multipliers of the constraints

$$\partial_t\xi_t-\rho_t^*=0,\qquad\nabla_\chi\xi_t-m_t^*=0.$$

Since these constraints force $\partial_t\xi_t-\rho_t^*$ resp. $\nabla_\chi\xi_t-m_t^*$ to be zero we can add

$$\frac{r}{2}\|\partial_t\xi_t-\rho_t^*\|_\pi^2+\frac{r}{2}\|\nabla_\chi\xi_t-m_t^*\|_\pi^2$$

without changing the possible solutions of the original saddle point problem. The factor r in front of the norms is just an additional parameter which will turn out to be step length of the gradient descent and which we could tune to get better convergence properties. However, it will most of the time just be set to 1. In this way we define the augmented Lagrangian as

$$
\begin{aligned}
L_r[\rho,m,\rho^*,m^*,\xi] &:= L[\rho,m,\rho^*,m^*,\xi]+\frac{r}{2}\int_0^1\left(\|\partial_t\xi_t-\rho_t^*\|_\pi^2+\|\nabla_\chi\xi_t-m_t^*\|_\pi^2\right)\,\mathrm{d}t\\
&=\mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*,m^*)+\langle\xi_0,\rho_A\rangle_\pi-\langle\xi_1,\rho_B\rangle_\pi+\int_0^1\left(\langle\partial_t\xi_t-\rho_t^*,\rho_t\rangle_\pi+\langle\nabla_\chi\xi_t-m_t^*,m_t\rangle_\pi\right)\,\mathrm{d}t\\
&\quad+\frac{r}{2}\int_0^1\left(\langle\partial_t\xi_t-\rho_t^*,\partial_t\xi_t-\rho_t^*\rangle_\pi+\langle\nabla_\chi\xi_t-m_t^*,\nabla_\chi\xi_t-m_t^*\rangle_\pi\right)\,\mathrm{d}t.
\end{aligned}
$$

We can simplify

$$\begin{aligned}
&\Big(\langle \partial_t \xi_t - \rho_t^*, \rho_t\rangle_\pi + \langle \nabla_\chi \xi_t - m_t^*, m_t\rangle_\pi\Big) + \frac{r}{2}\Big(\langle \partial_t \xi_t - \rho_t^*, \partial_t \xi_t - \rho_t^*\rangle_\pi + \langle \nabla_\chi \xi_t - m_t^*, \nabla_\chi \xi_t - m_t^*\rangle_\pi\Big)\\
&= \frac{r}{2}\Big(\langle \partial_t \xi_t - \rho_t^* + \frac{\rho_t}{r}, \partial_t \xi_t - \rho_t^* + \frac{\rho_t}{r}\rangle_\pi + \langle \nabla_\chi \xi_t - m_t^* + \frac{m_t}{r}, \nabla_\chi \xi_t - m_t^* + \frac{m_t}{r}\rangle_\pi\Big)\\
&\quad - \frac{1}{2}\langle \frac{\rho_t}{r}, \frac{\rho_t}{r}\rangle_\pi - \frac{1}{2}\langle \frac{m_t}{r}, \frac{m_t}{r}\rangle_\pi\\
&= \frac{r}{2}\Big(\|\partial_t \xi_t - \rho_t^* + \frac{\rho_t}{r}\|_\pi^2 + \|\nabla_\chi \xi_t - m_t^* + \frac{m_t}{r}\|_\pi^2\Big) - \frac{1}{2}\langle \frac{\rho_t}{r}, \frac{\rho_t}{r}\rangle_\pi - \frac{1}{2}\langle \frac{m_t}{r}, \frac{m_t}{r}\rangle_\pi.
\end{aligned}$$

In total we get

$$\begin{aligned}
L_r[\rho, m, \rho^*, m^*, \xi] = \mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*, m^*) + \langle \xi_0, \rho_A\rangle_\pi - \langle \xi_1, \rho_B\rangle_\pi\\
+ \frac{r}{2}\int_0^1 \Big(\|\partial_t \xi_t + \frac{\rho_t}{r} - \rho_t^*\|_\pi^2 + \|\nabla_\chi \xi_t + \frac{m_t}{r} - m_t^*\|_\pi^2\Big)\,\mathrm{d}t\\
- \frac{1}{2r^2}\int_0^1 \big(\|\rho_t\|_\pi^2 + \|m_t\|_\pi^2\big)\,\mathrm{d}t.
\end{aligned} \tag{4.1}$$

We have now converted the original minimization problem with the continuity equation constraint into a saddle point problem. For solving this problem we will now alternatingly solve the optimality conditions in ξ, ρ^*, m^* and ρ, m which can be seen as an alternating gradient descent. In total we get the following algorithm which we will abbreviate by `(BB)`.

Algorithm 1 (Alternating gradient descent of Augmented-Lagrangian `(BB)`)**.** Lets assume we already have the values (ρ^n, m^n) for iteration n of the algorithm (if $n = 0$ pick an initialization for (ρ^0, m^0)). Then the next iteration consists of the following three steps:

(A) Find $\xi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$ such that

$$L_r[\rho^n, m^n, \rho^*, m^*, \xi] \le L_r[\rho^n, m^n, \rho^*, m^*, \tilde{\xi}] \quad \forall \tilde{\xi} \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$$

(B) Find $\rho^* \in L^2([0,1], \mathbb{R}^{\mathcal{X}}), m^* \in L^2([0,1], \mathbb{R}^{\mathcal{X}\times\mathcal{X}})$ such that for all $\tilde{\rho}^* \in L^2([0,1], \mathbb{R}^{\mathcal{X}}), \tilde{m}^* \in L^2([0,1], \mathbb{R}^{\mathcal{X}\times\mathcal{X}})$

$$L_r[\rho^n, m^n, \rho^*, m^*, \xi] \le L_r[\rho^n, m^n, \tilde{\rho}^*, \tilde{m}^*, \tilde{\xi}]$$

(C) Update

$$\begin{aligned}
\rho^{n+1} &= \rho^n + r(\partial_t \xi - \rho^*),\\
m^{n+1} &= m^n + r(\nabla_\chi \xi - m^*).
\end{aligned}$$

In the next subsections we will explain how steps A and B can be computed.

4.1.1 Step A: Solving an elliptic problem

Step A is the optimization of L_r in ξ. Fortunately L_r is differentiable in ξ and we can just compute the derivative.

Lemma 4.2. *Let ρ, m, ρ^*, m^* be fixed. Let $\xi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$ be a minimizer in the sense that*

$$L_r[\rho, m, \rho^*, m^*, \xi] \le L_r[\rho, m, \rho^*, m^*, \tilde{\xi}] \quad \forall \tilde{\xi} \in H^1([0,1], \mathbb{R}^{\mathcal{X}}). \tag{4.2}$$

Then the following optimality condition holds for all $\varphi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$

$$r \int_0^1 \left(\langle \partial_t \xi_t, \partial_t \varphi_t \rangle_\pi + \langle \nabla_\chi \xi_t, \nabla_\chi \varphi_t \rangle_\pi \right) \mathrm{d}t = \langle \varphi_1, \rho_B \rangle_\pi - \langle \varphi_0, \rho_A \rangle_\pi - \int_0^1 \left(\langle \partial_t \varphi_t, \rho_t - r\rho_t^* \rangle_\pi + \langle \nabla_\chi \varphi_t, m_t - r m_t^* \rangle_\pi \right) \mathrm{d}t. \tag{4.3}$$

Proof. We calculate the Euler-Lagrange equation

$$\begin{aligned}
&\partial_\xi L_r[\rho, m, \rho^*, m^*, \xi](\varphi) \\
&= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} L_r[\rho, m, \rho^*, m^*, \xi + \varepsilon\varphi]\Big|_{\varepsilon=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \Big[\langle \xi_0 + \varepsilon\varphi_0, \rho_A \rangle_\pi - \langle \xi_1 + \varepsilon\varphi_1, \rho_B \rangle_\pi \Big]_{\varepsilon=0} \\
&\quad + \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \left[\frac{r}{2} \int_0^1 \left(\| \partial_t(\xi_t + \varepsilon\varphi_t) + \frac{\rho_t}{r} - \rho_t^* \|_\pi^2 + \| \nabla_\chi(\xi_t + \varepsilon\varphi_t) + \frac{m_t}{r} - m_t^* \|_\pi^2 \right) \mathrm{d}t \right]_{\varepsilon=0}.
\end{aligned}$$

For the boundary term we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \Big[\langle \xi_0 + \varepsilon\varphi_0, \rho_A \rangle_\pi - \langle \xi_1 + \varepsilon\varphi_1, \rho_B \rangle_\pi \Big]_{\varepsilon=0} &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \Big[\langle \xi_0, \rho_A \rangle_\pi + \varepsilon \langle \varphi_0, \rho_A \rangle_\pi - \langle \xi_1, \rho_B \rangle_\pi - \varepsilon \langle \varphi_1, \rho_B \rangle_\pi \Big]_{\varepsilon=0} \\
&= \langle \varphi_0, \rho_A \rangle_\pi - \langle \varphi_1, \rho_B \rangle_\pi.
\end{aligned}$$

The remainder yields

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \left[\frac{r}{2} \int_0^1 \left(\| \partial_t(\xi_t + \varepsilon\varphi_t) + \frac{\rho_t}{r} - \rho_t^* \|_\pi^2 + \| \nabla_\chi(\xi_t + \varepsilon\varphi_t) + \frac{m_t}{r} - m_t^* \|_\pi^2 \right) \mathrm{d}t \right]_{\varepsilon=0} \\
&= \int_0^1 \langle \partial_t \varphi_t, \partial_t \xi_t + \frac{\rho_t}{r} - \rho_t^* \rangle_\pi \, \mathrm{d}t + \int_0^1 \langle \nabla_\chi \varphi_t, \nabla_\chi \xi_t + \frac{m_t}{r} - m_t^* \rangle_\pi \, \mathrm{d}t \\
&= r \int_0^1 \left(\langle \partial_t \varphi_t, \partial_t \xi_t \rangle_\pi + \langle \nabla_\chi \varphi_t, \nabla_\chi \xi \rangle_\pi \right) \mathrm{d}t + \int_0^1 \left(\langle \partial_t \varphi_t, \rho_t - r\rho_t^* \rangle_\pi \, \mathrm{d}t + \langle \nabla_\chi \varphi_t, m_t - r m_t^* \rangle_\pi \right) \mathrm{d}t
\end{aligned}$$

■

Careful observers will note that (4.3) looks like the weak formulation of an elliptic boundary value problem. Indeed if we assume more regularity we can even transform this problem in a typical strong formulation containing the discrete differential operators.

Lemma 4.3. *Suppose additionally* $\xi \in C^2([0,1], \mathbb{R}^{\mathcal{X}}), \rho \in C^1([0,1], \mathbb{R}^{\mathcal{X}}), m \in C^1([0,1], \mathbb{R}^{\mathcal{X} \times \mathcal{X}})$. *Then the optimality condition in Lemma 4.2 is the weak formulation of the elliptic boundary value problem*

$$-r\left(\partial_{tt}\xi + \Delta_\chi \xi\right) = \partial_t\left(\rho_t - r\rho_t^*\right) + \mathrm{div}_\chi\left(m_t - r m_t^*\right)$$

with Neumann boundary values in time

$$\begin{aligned}
\partial_t \xi_0(x) &= \rho_0^*(x) \quad \forall x \in \mathcal{X}, \\
\partial_t \xi_1(x) &= \rho_1^*(x) \quad \forall x \in \mathcal{X}.
\end{aligned}$$

Proof. Using both integration by parts formulas we get for the left hand side of (4.3)

$$\begin{aligned}
&r\int_0^1 \big(\langle\partial_t\xi_t,\partial_t\varphi_t\rangle_\pi + \langle\nabla_\chi\xi_t,\nabla_\chi\varphi_t\rangle_\pi\big)\,\mathrm{d}t\\
&= r\big(\langle\partial_t\xi_1,\varphi_1\rangle_\pi - \langle\partial_t\xi_0,\varphi_0\rangle_\pi\big) - r\int_0^1 \big(\langle\partial_{tt}\xi_t,\varphi_t\rangle_\pi + \langle\mathrm{div}_\chi(\nabla_\chi\xi_t),\varphi_t\rangle_\pi\big)\,\mathrm{d}t\\
&= r\big(\langle\partial_t\xi_1,\varphi_1\rangle_\pi - \langle\partial_t\xi_0,\varphi_0\rangle_\pi\big) - r\int_0^1 \langle\partial_{tt}\xi_t + \Delta_\chi\xi_t,\varphi_t\rangle_\pi\,\mathrm{d}t
\end{aligned}$$

and for the right hand side

$$\begin{aligned}
&\langle\varphi_1,\rho_B\rangle_\pi - \langle\varphi_0,\rho_A\rangle_\pi - \int_0^1 \big(\langle\partial_t\varphi_t,\rho_t - r\rho_t^*\rangle_\pi + \langle\nabla_\chi\varphi_t,m_t - rm_t^*\rangle_\pi\big)\,\mathrm{d}t\\
&= \langle\varphi_1,\rho_B\rangle_\pi - \langle\varphi_0,\rho_A\rangle_\pi - \langle\varphi_1,\rho_B - r\rho_1^*\rangle_\pi + \langle\varphi_0,\rho_A - r\rho_0^*\rangle_\pi\\
&\quad + \int_0^1 \big(\langle\varphi_t,\partial_t(\rho_t - r\rho_t^*)\rangle_\pi + \langle\varphi_t,\mathrm{div}_\chi(m_t - rm_t^*)\rangle_\pi\big)\,\mathrm{d}t\\
&= +\langle\varphi_1,r\rho_1^*\rangle_\pi - \langle\varphi_0,r\rho_0^*\rangle_\pi + \int_0^1 \big(\langle\varphi_t,\partial_t(\rho_t - r\rho_t^*)\rangle_\pi + \langle\varphi_t,\mathrm{div}_\chi(m_t - rm_t^*)\rangle_\pi\big)\,\mathrm{d}t\\
&= r\big(\langle\varphi_0,\rho_0^*\rangle_\pi - \langle\varphi_1,\rho_1^*\rangle_\pi\big) + \int_0^1 \big(\langle\varphi_t,\partial_t(\rho_t - r\rho_t^*) + \mathrm{div}_\chi(m_t - rm_t^*)\rangle_\pi\big)\,\mathrm{d}t
\end{aligned}$$

Thus (4.3) is equivalent to

$$\begin{aligned}
&\int_0^1 \langle r\partial_{tt}\hat{\xi}_t + r\Delta_\chi\hat{\xi}_t + \partial_t(\rho_t - r\rho_t^*) + \mathrm{div}_\chi(m_t - rm_t^*),\varphi_t\rangle_\pi\,\mathrm{d}t\\
&= r\big(\langle\varphi_0,\rho_0^* - \partial_t\xi_0\rangle_\pi - \langle\varphi_1,\rho_1^* - \partial_t\xi_1\rangle_\pi\big).
\end{aligned}$$

This means that for all

$$\varphi \in \{\psi : [0,1]\times\mathcal{X}\to\mathbb{R} \mid \psi(\cdot,x)\in\mathcal{C}_0^\infty\ \forall x\in\mathcal{X}\}$$

we get

$$\int_0^1 \langle r\partial_{tt}\hat{\xi}_t + r\Delta_\chi\hat{\xi}_t + \partial_t(\rho_t - r\rho_t^*) + \mathrm{div}_\chi(m_t - rm_t^*),\varphi_t\rangle_\pi\,\mathrm{d}t = 0.$$

Now using the fundamental lemma of the calculus of variations we get

$$-r\big(\partial_{tt}\xi + \Delta_\chi\xi\big) = \partial_t\big(\rho_t - r\rho_t^*\big) + \mathrm{div}_\chi\left(m_t - rm_t^*\right)$$

and as boundary conditions in time

$$\begin{aligned}
\partial_t\xi_0(x) &= \rho_0^*(x) \quad \forall x\in\mathcal{X},\\
\partial_t\xi_1(x) &= \rho_1^*(x) \quad \forall x\in\mathcal{X}.
\end{aligned}$$

■

As a conclusion Step A of the Benamou-Brenier approach (which is finding $\hat{\xi}$ that fulfils (4.2)) means just solving an elliptic Poisson-type problem. We will describe how this can be done using finite elements in chapter 6.

4.1.2 Step B: Projecting on a convex set

In this section we will discuss how we can perform step B of (BB). Recall that step B means optimizing in the dual variables, i.e. finding $\tilde{\rho}^* \in L^2([0,1], \mathbb{R}^{\mathcal{X}}), \tilde{m}^* \in L^2([0,1], \mathbb{R}^{\mathcal{X}\times\mathcal{X}})$ such that

$$L_r[\rho^n, m^n, \rho^*, m^*, \xi] \leq L_r[\rho^n, m^n, \tilde{\rho}^*, \tilde{m}^*, \xi].$$

For this step we cannot differentiate with respect to ρ^* or m^*. However, we observe on (4.1) that only the few summands actually contribute to the optimization in ρ^* and m^*. To be more precise we have

$$\begin{aligned}
&\operatorname*{arg\,min}_{(\rho^*,m^*)\in[0,1]\times\mathcal{H}} L_r[\rho, m, \rho^*, m^*, \xi] \\
&= \operatorname*{arg\,min}_{(\rho^*,m^*)\in[0,1]\times\mathcal{H}} \left\{ \mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*, m^*) + \int_0^1 \left(\|\partial_t \xi_t + \frac{\rho_t}{r} - \rho_t^*\|_\pi^2 + \|\nabla_\chi \xi_t + \frac{m_t}{r} - m_t^*\|_\pi^2 \right) \mathrm{d}t \right\} \\
&= \operatorname*{arg\,min}_{(\rho^*,m^*)\in[0,1]\times\mathcal{K}} \int_0^1 \left(\|\partial_t \xi_t + \frac{\rho_t}{r} - \rho_t^*\|_\pi^2 + \|\nabla_\chi \xi_t + \frac{m_t}{r} - m_t^*\|_\pi^2 \right) \mathrm{d}t.
\end{aligned}$$

We will later see in section 6.3 that using certain finite element discretizations the minimization of the integral decouples in time and we can perform the minimization for each timestep $t \in [0,1]$ separately. This means for now we can forget about the time coordinate and just figure out, how to solve the problem for one timestep, i.e. for fixed t and $\xi, (\rho, m)$ solve

$$\operatorname*{arg\,min}_{(\rho^*,m^*)\in\mathcal{K}} \left(\|\partial_t \xi_t + \frac{\rho_t}{r} - \rho_t^*\|_\pi^2 + \|\nabla_\chi \xi_t + \frac{m_t}{r} - m_t^*\|_\pi^2 \right) = \mathrm{proj}_{\mathcal{K}} \left(\partial_t \xi_t + \frac{\rho_t}{r}, \nabla_\chi \xi_t + \frac{m_t}{r} \right).$$

Thus we need to be able to compute the projection on $\mathcal{K}$ for solving step B of (BB). Because this is also needed in the other numerical approach that will follow in the next section and because calculating the projection turns out to be more difficult than in the original Benamou-Brenier case we will discuss the projection algorithm separately in chapter 5.

4.2 A proximal splitting approach

In the following section we will explore another approach for calculating the geodesics based on a so called proximal splitting method, namely the Douglas-Rachford algorithm. In the first subsection we will introduce the general algorithm for generic convex functions. Afterwards we apply it to our particular problem of calculating the geodesics of the discrete Wasserstein distance.

4.2.1 The general Douglas-Rachford algorithm

Most of the basic definitions and properties in this subsection are taken from the monograph [EB92].
We start by introducing the a central object to proximal splitting methods - the proximal operator.

Definition 4.4. *Let H be a Hilbert space and let $f : H \to \mathbb{R}$ be convex, proper and lower semi-continuous. We define the* ***proximal operator*** *by*

$$\mathrm{prox}_f(x) = \operatorname*{arg\,min}_y \left\{ f(y) + \frac{1}{2}\|x - y\|^2 \right\}.$$

This proximal operator maps a point $x \in H$ to another point $y \in H$ that has the best trade-off between function value at y and distance from x. It has several nice properties but in particular there is the following relationship between the proximal mapping and the (sub-)differential of f. A proof for the following Lemma can be found for example in [PB14, Section 3.2].

Lemma 4.5. *The following holds*

$$x = \text{prox}_{\lambda f}(z) \quad \Leftrightarrow \quad z \in x + \lambda f(x) \quad \Leftrightarrow \quad (z - x) \in \lambda \partial f(x).$$

Equivalently we have

$$\text{prox}_{\lambda f} = (\mathbf{1} + \lambda \partial f)^{-1}. \tag{4.4}$$

These properties of the proximal operator are particularly interesting because they introduce a natural way to search for minimizers of f. The most basic approach could be a fixpoint iteration like

$$x^{(n+1)} = \text{prox}_{\lambda f}(x^{(n)}).$$

Because whenever we have $x = \text{prox}_{\lambda f}(x)$ Lemma 4.5 tells us that $x - x = 0 \in \lambda \partial f(x)$ and we found a minimizer. However, often it is quite hard to compute the proximal mapping of a some function f. This is where proximal splitting algorithms come into play.

Suppose we are looking at energies that can be written as a sum of smaller parts. Proximal splitting methods can now be used if it is hard to calculate the proximal operator of the whole energy but quite possible to calculate it for the summands independently. Alternating calculation of the resolvants (see (4.4)) yields a type of fixpoint iteration which - under some mild conditions - converges to the solution of the optimization problem.

The Douglas-Rachford algorithm is one particular (often used) proximal splitting method. We will introduce the basic algorithm and some reasoning for the convergence, however for more detailed analysis we refer to [EB92] and [CP09].

Algorithm 2 (Douglas-Rachford splitting `(DR)`)**.** Suppose we want to solve the problem

$$f_1(x) + f_2(x) \to \min$$

where f_1, f_2 are closed, lower semi-continuous and convex functions and $z^{(0)}$ is an arbitrary start value in the domains of f_1 and f_2. Let $\lambda > 0$ and $\alpha \in (0, 2)$ be fixed. Then each iteration of the Douglas-Rachford algorithm consists of the following three steps:

(1) $x^{(k)} = \text{prox}_{\lambda f_2}(z^{(k-1)})$

(2) $y^{(k)} = \text{prox}_{\lambda f_1}(2x^{(k)} - z^{(k-1)})$

(3) $z^{(k)} = z^{(k-1)} + \alpha(y^{(k)} - x^{(k)})$

Remark: *Note that the general Douglas-Rachford algorithm can not only be used for minimization problems but for generally finding the root of a maximal monotone operator. (see [EB92])*

Theorem 4.6. *Let $(x^{(k)}, y^{(k)}, z^{(k)})$ be a fixpoint of the Douglas-Rachford scheme then $x^{(k)}$ is a minimizer of $f_1 + f_2$.*

Proof. Consider a fixpoint iteration for the following function

$$\begin{aligned} F(z) &= z + \alpha(\text{prox}_{\lambda f_1}(2\text{prox}_{\lambda f_2}(z) - z) - \text{prox}_{\lambda f_2}(z)) \\ &= z + \alpha(\text{prox}_{\lambda f_1}(2x - z) - x) \end{aligned}$$

where we denote $x = \text{prox}_{\lambda f_2}(z)$. Assume we are at a fixpoint i.e. $F(z) = z$, then the following equation is satisfied

$$\text{prox}_{\lambda f_1}(2x - z) = x = \text{prox}_{\lambda f_2}(z).$$

Using Lemma 4.5 this is equivalent to

$$2x - z - x = x - z \in \lambda \partial f_1(x) \quad \text{and} \quad z - x \in \lambda \partial f_2(x).$$

Adding these up we get

$$0 \in \lambda \partial f_1(x) + \lambda \partial f_2(x) = \lambda \partial (f_1 + f_2)(x).$$

■

4.2.2 Application to the discrete Wasserstein distance

After we have introduced the general notion of proximal splitting methods in the previous subsection we can now focus on our problem of computing the geodesics for the discrete Wasserstein distance. While we added the continuity equation constraint in the Benamou-Brenier case through an Lagrange multiplier we added the constraint by using an indicator function, i.e.

$$\inf \left\{ \int_0^1 \mathcal{A}(\rho_t, m_t) \, \mathrm{d}t : (\rho, m) \in \mathcal{CE}(\rho_A, \rho_B) \right\} = \inf_{(\rho,m) \in [0,1] \times \mathcal{H}} \int_0^1 \mathcal{A}(\rho_t, m_t) \, \mathrm{d}t + \mathcal{I}_{\mathcal{CE}(\rho_A, \rho_B)}(\rho, m).$$

In the proximal splitting context we introduced above we set

$$f_1(\rho, m) = \mathcal{E}_{\text{Trans}}[\rho, m] \quad \text{and} \quad f_2(\rho, m) = \mathcal{I}_{\mathcal{CE}(\rho_A, \rho_B)}(\rho, m).$$

What we need for `(DR)` to work is to be able to compute prox_{f_1} and prox_{f_2}. We first start with the continuity equation, i.e. prox_{f_2}. First note that for indicator function on a convex set C we have

$$\text{prox}_{\mathcal{I}_C}(z) = \text{proj}_C(z).$$

Thus proximal mappings of indicator functions are projections on the set characterized by the indicator function. As a result computing prox_{f_2} in our context means projecting on the the set of $\mathcal{CE}(\rho_A, \rho_B)$. The following Lemma introduces a way to compute this projection.

Lemma 4.7. *Let $\rho \in L^2([0,1], \mathbb{R}^{\mathcal{X}}), m \in L^2([0,1], \mathbb{R}^{\mathcal{X} \times \mathcal{X}})$ be given and let $\xi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$ be such that for all $\varphi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$*

$$\int_0^1 \left(\langle \partial_t \xi_t, \partial_t \varphi_t \rangle_\pi + \langle \nabla_\chi \xi_t, \nabla_\chi \varphi_t \rangle_\pi \right) \mathrm{d}t = \langle \varphi_1, \rho_B \rangle_\pi - \langle \varphi_0, \rho_A \rangle_\pi - \int_0^1 \left(\langle \rho_t, \partial_t \varphi_t \rangle_\pi + \langle m_t, \nabla_\chi \varphi_t \rangle_\pi \right) \mathrm{d}t. \tag{4.5}$$

Then $\tilde{\rho} = \rho + \partial_t \xi$ and $\tilde{m} = m + \nabla_\chi \xi$ satisfy

$$(\tilde{\rho}, \tilde{m}) = \text{proj}_{\mathcal{CE}(\rho_A, \rho_B)}(\rho, m).$$

Remark: *Note that* (4.5) *is very similar to the elliptic problem in* (4.3).

Proof. In order to compute $\operatorname{proj}_{\mathcal{CE}(\rho_A,\rho_B)}(\rho, m)$ we do the following reformulation as a saddle point problem using Lagrange multipliers

$$\begin{aligned}
\operatorname{proj}_{\mathcal{CE}}(\rho_A,\rho_B)(\rho,m) &= \operatorname*{arg\,min}_{(\tilde\rho,\tilde m)\in\mathcal{CE}(\rho_A,\rho_B)} \left[\frac{1}{2}\int_0^1 (\|\rho_t-\tilde\rho_t\|_\pi + \|m_t-\tilde m_t\|_\pi)\ \mathrm{d}t\right] \\
&= \operatorname*{arg\,min}_{\tilde\rho,\tilde m}\sup_{\xi}\left[\frac{1}{2}\int_0^1 (\|\rho_t-\tilde\rho_t\|_\pi + \|m_t-\tilde m_t\|_\pi)\ \mathrm{d}t + \int_0^1 \langle \xi, \partial_t\tilde\rho_t + \operatorname{div}_\chi \tilde m_t\rangle_\pi\ \mathrm{d}t\right] \\
&= \operatorname*{arg\,min}_{\tilde\rho,\tilde m}\sup_{\xi}\left[\frac{1}{2}\int_0^1 (\|\rho_t-\tilde\rho_t\|_\pi + \|m_t-\tilde m_t\|_\pi)\ \mathrm{d}t\right. \\
&\qquad\left. + \langle\xi_1,\rho_B\rangle_\pi - \langle\xi_0,\rho_A\rangle_\pi - \int_0^1\langle\partial_t\xi,\tilde\rho_t\rangle_\pi\ \mathrm{d}t - \int_0^1\langle\nabla_\chi\xi_t,\tilde m_t\rangle_\pi\ \mathrm{d}t\right] \\
&=: \operatorname*{arg\,min}_{\tilde\rho,\tilde m}\sup_{\xi} L_p[\tilde\rho,\tilde m,\xi].
\end{aligned}$$

Computing the optimality conditions we get

$$0 = \partial_\xi L_p[\tilde\rho,\tilde m,\xi](\varphi) = \langle\varphi_1,\rho_B\rangle_\pi - \langle\varphi_0,\rho_A\rangle_\pi - \int_0^1\langle\partial_t\varphi_t,\tilde\rho_t\rangle_\pi\ \mathrm{d}t - \int_0^1\langle\nabla_\chi\varphi_t,\tilde m_t\rangle_\pi\ \mathrm{d}t \quad \forall\varphi, \tag{4.6}$$

$$\begin{aligned}
0 = \partial_{\tilde\rho} L_p[\tilde\rho,\tilde m,\xi](\hat\rho) &= -\int_0^1\langle\rho_t-\tilde\rho_t,\hat\rho_t\rangle_\pi\ \mathrm{d}t - \int_0^1\langle\partial_t\xi_t,\hat\rho_t\rangle_\pi\ \mathrm{d}t \\
&= -\int_0^1\langle(\rho_t+\partial_t\xi_t)-\tilde\rho_t,\hat\rho_t\rangle_\pi\ \mathrm{d}t \qquad \forall\hat\rho,
\end{aligned} \tag{4.7}$$

$$\begin{aligned}
0 = \partial_{\tilde m} L_p[\tilde\rho,\tilde m,\xi](\hat m) &= -\int_0^1\langle m_t-\tilde m_t,\hat m_t\rangle_\pi\ \mathrm{d}t - \int_0^1\langle\nabla_\chi\xi_t,\hat m_t\rangle_\pi\ \mathrm{d}t \\
&= -\int_0^1\langle(m_t+\nabla_\chi\xi_t)-\tilde m_t,\hat m_t\rangle_\pi\ \mathrm{d}t \qquad \forall\hat m.
\end{aligned} \tag{4.8}$$

Now testing (4.7) with $\hat\rho = \partial_t\varphi$ yields

$$0 = -\int_0^1\langle\rho_t-\tilde\rho_t,\partial_t\varphi_t\rangle_\pi - \int_0^1\langle\partial_t\xi_t,\partial_t\varphi_t\rangle_\pi\ \mathrm{d}t, \tag{4.9}$$

and testing (4.8) with $\hat m = \nabla_\chi\varphi$ yields

$$0 = -\int_0^1\langle m_t-\tilde m_t,\nabla_\chi\varphi_t\rangle_\pi\ \mathrm{d}t - \int_0^1\langle\nabla_\chi\xi_t,\nabla_\chi\varphi_t\rangle_\pi\ \mathrm{d}t. \tag{4.10}$$

Adding (4.6), (4.9) and (4.10) yields

$$0 = \int_0^1(\langle\partial_t\xi_t,\partial_t\varphi_t\rangle_\pi + \langle\nabla_\chi\xi_t,\nabla_\chi\varphi_t\rangle_\pi)\ \mathrm{d}t + \langle\varphi_0,\xi_0\rangle_\pi - \langle\varphi_1,\rho_1\rangle_\pi + \int_0^1(\langle\rho_t,\partial_t\varphi_t\rangle_\pi + \langle m_t,\nabla_\chi\varphi\rangle_\pi)\ \mathrm{d}t \tag{4.11}$$

which is exactly what we wanted.
Note that because of (4.7) and (4.8) the projection $(\tilde\rho,\tilde m)$ can be recovered from a solution ξ of (4.11) by

$$\tilde\rho = \rho + \partial_t\xi, \qquad \tilde m = m + \nabla_\chi\xi.$$

This finishes the proof. ∎

Thus we have seen that we can calculate prox_{f_2} by solving an elliptic problem very similar to the elliptic problem in the (BB) algorithm. Next we look at computing prox_{f_1}.

To calculate prox_{f_1} we will use the following statement.

Proposition 4.8 (Moreau decomposition)**.** *If $f : H \to \mathbb{R} \cup \{\infty\}$ is convex, lower semi-continuous then for all $\lambda > 0$*

$$x = \text{prox}_{\lambda f}(x) + \lambda \text{prox}_{\frac{1}{\lambda} f^*}\left(\frac{1}{\lambda} x\right)$$

where f^* is the Fenchel dual of f.

The Moreau composition can be seen as a generalization of orthogonal decomposition induced by a subspace. A proof can be found for example in [PB14, Section 2.5].

We introduced the Moreau decomposition because directly computing $\text{prox}_{f_1} = \text{prox}_{\mathcal{E}_{\text{Trans}}}$ is very difficult. However, computing the Fenchel dual of $\mathcal{E}_{\text{Trans}}$ with respect to $\int_0^1 \langle \cdot, \cdot \rangle_{\mathcal{H}} \, \mathrm{d}t$ we get

$$\begin{aligned}
\mathcal{E}^*_{\text{Trans}}(\rho^*, m^*) &= \sup_{(\rho,m)\in[0,1]\times\mathcal{H}} \left[\int_0^1 \left(\langle \rho_t^*, \rho_t \rangle_\pi + \langle m_t^*, m_t \rangle_\pi \right) \mathrm{d}t - \int_0^1 \mathcal{A}(\rho_t, m_t) \, \mathrm{d}t \right] \\
&= \int_0^1 \sup_{(\rho,m)\in[0,1]\times\mathcal{H}} \left(\langle \rho_t^*, \rho_t \rangle_\pi + \langle m_t^*, m_t \rangle_\pi - \mathcal{A}(\rho_t, m_t) \right) \mathrm{d}t \\
&= \int_0^1 \mathcal{A}^*(\rho_t^*, m_t^*) \, \mathrm{d}t = \mathcal{I}_{[0,1]\times\mathcal{K}}(\rho^*, m^*).
\end{aligned}$$

where we used again the continuity of $\mathcal{A}$ to pull the supremum inside the integral. As a result this means that computing $\text{prox}_{f_1^*}$ boils down to projecting on $\mathcal{K}$ which is something that we also need to be able to do for (BB). Thus in practice we will compute prox_{f_1} by using the Moreau decomposition, i.e. with $\lambda = 1$

$$\begin{aligned}
\text{prox}_{\mathcal{E}_{\text{Trans}}}(\rho, m) &= (\rho, m) - \text{prox}_{\mathcal{E}^*_{\text{Trans}}}(\rho, m) \\
&= (\rho, m) - \text{proj}_{[0,1]\times\mathcal{I}_\mathcal{K}}(\rho, m)
\end{aligned}$$

Again note that because of the time discretization we will use later on the projection decouples in time and each timestep has to be projected separately on $\mathcal{K}$.

In total we have in this section derived another feasible scheme for computing the geodesics. At the heart of this method lie the same computations that were also necessary for (BB), namely solving an elliptic problem and computing projections on $\mathcal{K}$. (DR) can therefore be quickly implemented if there is already a working implementation of (BB) - and the other way around.

In the next chapter we will look in detail on how to compute the projections on $\mathcal{K}$. It turns out that this step will be much more difficult than solving the elliptic problem or computing the projection for the original Wasserstein distance.

5 Computing the Projection on K

We have already seen in section 4.1.2 for the Augmented Lagrangian approach and step two of the proximal splitting approach that for calculating the geodesic we will need to be able to compute the projection on the convex set $\mathcal{K}$ for each timestep separately. Recall that this set $\mathcal{K}$ was defined by the indicator function that was the dual of $\mathcal{A}$ calculated in Lemma 3.15.

Such a convex set already appeared in [BB00] during the Augmented Lagrangian approach for the original L^2-Wasserstein distance. However, in this case calculating the projection was fairly easy because it decoupled also in space. As a result in practice the projection could be calculated for each timestep and for each finite element separately. Thus the problem reduced to calculating the projection of tuples of the form $(\rho^*, m^*) \in \mathbb{R} \times \mathbb{R}^d$ to a set

$$\{(a, b) \in \mathbb{R} \times \mathbb{R}^d \text{ s.t. } a + \frac{|b|^2}{2} \leq 0\}.$$

This can be easily done using for example a simple Newton method of a one-dimensional function.

As we have already seen in Lemma 3.15 the convex set in the case of the new metric is much more complex. We recall the following characterization of the set as one of the three provided in Lemma 3.15 (for the ease of notation we omit the $*$ for the variables). A tuple (ρ, m) is in $\mathcal{K}$ if there exists $a \in \mathbb{R}_+^{\mathcal{X}}$ such that

$$\rho_x + \frac{1}{4} \sum_{y \in \mathcal{X}} \partial_1 \theta(a_x, a_y) |m_{x,y}|^2 K(x, y) \leq 0 \qquad \forall x \in \mathcal{X}. \tag{5.1}$$

It's immediately obvious that this characterization of a convex set is more complex in the sense that the conditions are very highly coupled with each other. Each equation compares the mass at some node with a sum of the weighted flow over all outgoing edges. These weights come from a potential $a \in \mathbb{R}^{\mathcal{X}}$ also defined on nodes and plugged into a highly nonlinear function $\partial_1 \theta$. Because this setting is more complex calculating the projection on $\mathcal{K}$ is not easy anymore. Even checking if some (ρ, m) is in $\mathcal{K}$ is non-trivial because in order to show this one has to compute a "witness"-potential $a \in \mathbb{R}^{\mathcal{X}}$ which satisfies all the equations of (5.1).

In the following Lemma we will rewrite the problem of calculating the projection on $\mathcal{K}$ to a saddle point problem. For it we can compute the optimality conditions and get an explicit system of equations that characterizes projections on $\mathcal{K}$. These can later be used in an gradient descent or Newton-type algorithm to numerically compute the projected values.

Lemma 5.1. *Let $(\rho, m) \in \mathcal{H}$ and let $(\tilde{\rho}, \tilde{m}) \in \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X} \times \mathcal{X}}$ be its projection on the convex set $\mathcal{K}$ defined in Lemma 3.15. Then the following optimality conditions hold*

$$\begin{aligned}
0 &= a_x - \rho_x + \tilde{\rho}_x && \forall x \in \mathcal{X} && \text{(I)} \\
0 &= \left(\tfrac{1}{4}\theta(a_x, a_y) + 1\right) \tilde{m}_{x,y} - m_{x,y} && \forall (x, y) : Q(x, y) > 0 && \text{(II)} \\
0 &= \tfrac{1}{4} \textstyle\sum_{y \in \mathcal{X}} Q(x, y) |\tilde{m}_{x,y}|^2 \partial_1 \theta(a_x, a_y) + \tilde{\rho}_x && \forall x \in \mathcal{X} && \text{(III)}
\end{aligned}$$

Proof. We turn the problem into a saddle point problem by noticing that

$$\begin{aligned}\operatorname{proj}_{\mathcal{K}}(\rho, m) &= \operatorname{prox}_{\mathcal{I}_{\mathcal{K}}}(\rho, m)\\ &= \arg\min_{\tilde\rho,\tilde m}\left[\mathcal{A}^*(\tilde\rho,\tilde m) + \frac{1}{2}\|\rho-\tilde\rho\|_\pi^2 + \frac{1}{2}\|m-\tilde m\|_\pi^2\right]\\ &= \arg\min_{\tilde\rho,\tilde m}\ \sup_a S(\tilde\rho,\tilde m, a).\end{aligned}$$

where

$$S(\tilde\rho,\tilde m,a) = \frac{1}{8}\sum_{x,y\in\mathcal{X}} |\tilde m_{x,y}|^2\theta(a_x,a_y)Q(x,y)\pi(x) + \langle a,\tilde\rho\rangle_\pi + \frac{1}{2}\|\rho-\tilde\rho\|_\pi^2 + \frac{1}{2}\|m-\tilde m\|_\pi^2$$

Now we derive the optimality conditions for this saddle point problem. We start with with the $\tilde\rho$-variable and compute the derivative in direction $\hat\rho$

$$\partial_{\tilde\rho} S[\tilde\rho,\tilde m,a](\hat\rho) = \langle\hat\rho,a\rangle_\pi - \langle\hat\rho,\rho-\tilde\rho\rangle_\pi = \langle\hat\rho, a-\rho+\tilde\rho\rangle_\pi.$$

If we test $0 = \partial_{\tilde\rho} S[\tilde\rho,\tilde m,a](\hat\rho)$ against unit vectors of the form $\hat\rho = (0,\dots,0,1,0,\dots,0)^T$ we get the following linear equations

$$0 = (a_x - \rho_x + \tilde\rho_x)\underbrace{\pi(x)}_{>0} \quad \forall x\in\mathcal{X} \quad\Leftrightarrow\quad 0 = a_x - \rho_x + \tilde\rho_x \quad \forall x\in\mathcal{X}.$$

This gives (I). Now we do the same for the $\tilde m$-variable.

$$\begin{aligned}\partial_{\tilde m} S[\tilde\rho,\tilde m,a](\hat m) &= \sum_{x,y\in\mathcal{X}} \frac{1}{8}\tilde m_{x,y}\hat m_{x,y}\theta(a_x,a_y)Q(x,y)\pi(x) - \langle\hat m, m-\tilde m\rangle_\pi\\ &= \sum_{x,y\in\mathcal{X}} \hat m_{x,y}\left(\frac{1}{8}\tilde m_{x,y}\theta(a_x,a_y) - \frac{1}{2}m_{x,y} + \frac{1}{2}\tilde m_{x,y}\right)Q(x,y)\pi(x)\\ &= \frac{1}{2}\sum_{x,y\in\mathcal{X}} \hat m_{x,y}\left(\frac{1}{4}\tilde m_{x,y}\theta(a_x,a_y) - m_{x,y} + \tilde m_{x,y}\right)Q(x,y)\pi(x)\end{aligned}$$

Again if we test against

$$\hat m^{(x,y)}_{r,s} = \begin{cases}1 & (r,s) = (x_0,y_0)\\ 0 & \text{otherwise}\end{cases} \qquad \forall (x_0,y_0)\in\mathcal{X}\times\mathcal{X},$$

then the optimality condition $0 = \partial_{\tilde m} S[\tilde\rho,\tilde m,a](\hat m)$ dissolves to

$$\begin{aligned}&0 = \frac{1}{2}\left[\frac{1}{4}\tilde m_{x,y}\theta(a_x,a_y) - m_{x,y} + \tilde m_{x,y}\right]Q(x,y)\underbrace{\pi(x)}_{>0} \quad \forall (x,y)\in\mathcal{X}\times\mathcal{X}\\ \iff\quad &0 = \frac{1}{4}\tilde m_{x,y}\theta(a_x,a_y) - m_{x,y} + \tilde m_{x,y} \quad \forall (x,y)\in\mathcal{X}\times\mathcal{X} : Q(x,y) > 0\\ \iff\quad &0 = \left(\frac{1}{4}\theta(a_x,a_y) + 1\right)\tilde m_{x,y} - m_{x,y} \quad \forall (x,y)\in\mathcal{X}\times\mathcal{X} : Q(x,y) > 0,\end{aligned}$$

which gives (II). For (III) calculate

$$\begin{aligned}\partial_a S[\tilde\rho,\tilde m,a](\hat a) &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\sum_{x,y\in\mathcal{X}} \frac{1}{8}|\tilde m_{x,y}|^2\theta(a_x+\varepsilon\hat a_x, a_y+\varepsilon\hat a_y)Q(x,y)\pi(x)\Big|_{\varepsilon=0} + \langle\hat a,\tilde\rho\rangle_\pi\\ &= \sum_{x,y\in\mathcal{X}} \frac{1}{8}|\tilde m_{x,y}|^2\left[\hat a_x\partial_1\theta(a_x,a_y) + \hat a_y\partial_2\theta(a_x,a_y)\right]Q(x,y)\pi(x) + \langle\hat a,\tilde\rho\rangle_\pi \quad \forall\hat a\in\mathbb{R}^{\mathcal{X}}\end{aligned}$$

Again we test against

$$\hat{a}_x^{(z)} = \begin{cases} 1 & x = z \\ 0 & \text{otherwise} \end{cases}.$$

then notice that we get

$$\hat{a}_x \partial_1\theta(a_x, a_y) + \hat{a}_y \partial_2\theta(a_x, a_y) = \begin{cases} \partial_1\theta(a_z, a_y) & x = z \\ \partial_2\theta(a_x, a_z) & y = z \\ 0 & x \neq z \text{ and } y \neq z. \end{cases}.$$

Therefore $0 = \partial_a S[\tilde{\rho}, \tilde{m}, a](\hat{a}^{(z)})$ becomes

$$\begin{aligned} 0 &= \sum_{y\in\mathcal{X}} \frac{1}{8}|\tilde{m}_{z,y}|^2 \partial_1\theta(a_z, a_y) Q(z,y)\pi(z) \\ &\quad + \sum_{x\in\mathcal{X}} \frac{1}{8}|\tilde{m}_{x,z}|^2 \partial_2\theta(a_x, a_z) \underbrace{Q(x,z)\pi(x)}_{Q(z,x)\pi(z)} + \tilde{\rho}_z\pi(z) \\ &= \frac{1}{8}\pi(z) \sum_{y\in\mathcal{X}} Q(z,y) \left[|\tilde{m}_{z,y}|^2 \partial_1\theta(a_z, a_y) + |\tilde{m}_{y,z}|^2 \partial_2\theta(a_y, z)\right] + \tilde{\rho}_z\pi(z) \end{aligned}$$

Since $\pi(z) > 0$ for all $z \in \mathcal{X}$ this is equivalent to

$$0 = \frac{1}{8} \sum_{y\in\mathcal{X}} Q(z,y) \left[|\tilde{m}_{z,y}|^2 \partial_1\theta(a_z, a_y) + |\tilde{m}_{y,z}|^2 \partial_2\theta(a_y, a_z)\right] + \tilde{\rho}_z. \tag{5.2}$$

Without loss of generality we can assume that $m_{x,y} = -m_{y,x}$ for all $x \neq y$ at all times. This is true because in our setting $m_{x,y}$ represents the instantaneous flow over the edge (x, y) and will therefore always be anti-symmetric. As a result (5.2) simply becomes

$$0 = \frac{1}{8} \sum_{y\in\mathcal{X}} Q(z,y)|\tilde{m}_{z,y}|^2 \left[\partial_1\theta(a_z, a_y) + \partial_2\theta(a_y, a_z)\right] + \tilde{\rho}_z.$$

Now we apply 3.13 (iii) and get

$$0 = \frac{1}{4} \sum_{y\in\mathcal{X}} Q(z,y)|\tilde{m}_{z,y}|^2 \partial_1\theta(a_z, a_y) + \tilde{\rho}_z.$$

Doing so for all $z \in \mathcal{X}$ gives (III). ■

We have now arrived at a state where where we broke the problem of projecting on $\mathcal{K}$ down to solving a set of equations. However, we will see that that complexity of the convex set carries over to the set of equations in the sense that the resulting root-finding problem is not very stable. Again this is mostly due to the dependence on the nonlinear logarithmic mean θ resp. $\partial_1\theta$.

Because the problem is very unstable we want to have an additional way of checking whether we computed a projection correctly. Thus next we will derive some condition that can easily be checked in order to see that the result of a computation is indeed the projection and not some other solution of the system of equations. First of all to check if the result is in $\mathcal{K}$ is trivial because we can just check the inequalities of the characterizations of $\mathcal{K}$ given in Lemma 3.15. However, being in $\mathcal{K}$ (or on boundary to be more precise) is of course not enough for being a projection, it is also necessary that the connection $\rho - \tilde{\rho}$ between projection $\tilde{\rho}$ and the point outside $\mathcal{K}$ that was projected ρ is "perpendicular" on the boundary of $\mathcal{K}$. A generalization of this

idea is behind the following Lemma. Note that checking this condition is fairly simple because one only has to do one evaluation of $\mathcal{A}$ and compute some inner products $\langle \cdot, \cdot \rangle_\pi$.

We will need a special case of Lemma 4.5 which we state in the following corollary.

Corollary 5.2. *Let H be a Hilbert space, $\mathcal{K} \subset H$ a convex set, $p \in \mathcal{K}$ and $z \notin \mathcal{K}$ then if $p = \operatorname{proj}_{\mathcal{K}}(z)$ we have $(p - z) \in \partial \mathcal{I}_{\mathcal{K}}(p)$.*

We will also need the following fact from convex analysis (see for example [ABM06, Section 17]).

Lemma 5.3. *For every proper, closed, convex Φ,*

$$u^* \in \partial \Phi(u) \Leftrightarrow \Phi(u) + \Phi^*(u^*) - \langle u^*, u \rangle = 0.$$

Lemma 5.4. *Let $\mathcal{K}$ be the convex set of Lemma 3.15 and let $(\tilde{\rho}, \tilde{m}) = \operatorname{proj}_{\mathcal{K}}(\rho, m)$ the projection of any $(\rho, m) \notin \mathcal{K}$ on $\mathcal{K}$. Then*

$$\mathcal{A}(\tilde{\rho} - \rho, \tilde{m} - m) - \langle \tilde{\rho} - \rho, \rho \rangle_\pi - \langle \tilde{m} - m, \tilde{m} \rangle_\pi = 0.$$

Proof. From Corollary 5.2 we know because $\operatorname{prox}_{\mathcal{I}_{\mathcal{K}}}(z) = \operatorname{proj}_{\mathcal{K}}(z)$

$$(\tilde{\rho}, \tilde{m}) = \operatorname{proj}_{\mathcal{K}}(\rho, m) \Leftrightarrow (\tilde{\rho}, \tilde{m}) \in \mathcal{K} \wedge (\rho - \tilde{\rho}, m - \tilde{m}) \in \partial \mathcal{I}_{\mathcal{K}}(\tilde{\rho}, \tilde{m})$$

Now applying Lemma 5.4 where in this case $\Phi = \mathcal{A}^* = \mathcal{I}_{\mathcal{K}}$ and $\Phi^* = \mathcal{A}^{**} = \mathcal{A}$ yields the stated result. ∎

In order to compute a projection of some pair (ρ, m) on $\mathcal{K}$ we can solve the saddle point problem of Lemma 5.1 for $(\tilde{\rho}, \tilde{m})$. For doing so we will could use different methods. One possibility is an alternating gradient descent where we alternatingly update a, $\tilde{\rho}$ and $\tilde{m}$ according to (I), (II) and (III) until we get close to a solution. The advantage of this method is that it can be implemented fairly quickly, however this comes at the price of slow convergence. More advanced methods use also second derivatives of the saddle point problem. For example Newton's method has quadratic convergence for an area around the solution which is fast enough even for big instances. During the creation of this thesis different approaches were tried with varying success. We will go more into details of which methods work and are feasible to compute in chapter 6 about the implementation details.

All problems solving the optimality conditions of Lemma 5.1 are a result of the presence of the highly nonlinear logarithmic mean and its derivatives in the equations. In order to give an more explicit understanding of the arising problems we will now look at typical projection problem that we have to solve during our algorithms for calculating the geodesic.

Example. *Let $\mathcal{X} = \{a, b\}$ and*

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

thus in order to satisfy the conditions we get $\pi = (\frac{1}{2}, \frac{1}{2})$.

Recall that the convex set $\mathcal{K}$ is characterized by (5.1). Now because $\partial_1\theta(s,t) > 0$ for all $s,t \geq 0$ we know that for a pair $(\tilde{\rho},\tilde{m}) \in \mathcal{K}$ the $\tilde{\rho}$ needs to only have negative entries. After the algorithms for calculating the geodesics described in chapter 4 have already come close to its final state we would expect that inputs to our projection are already close to the resulting projected values, i.e. we do not have to project that far anymore. This also implies that after some iterations all projection inputs ρ should already be negative. Nevertheless it is common that for early iterations we get values like $\rho = (0.1, -1)$, $m = (0, -1, 1, 0)$ that should be projected on $\mathcal{K}$ (the representation of m as one-dimensional vector is explained in chapter 6 but also not really of importance right now).

In this case the approximate projection of (ρ, m) on $\mathcal{K}$ is given by $(\tilde{\rho},\tilde{m})$ where

$$\tilde{\rho} = (-0.0421303, -1.0016) \text{ and } \tilde{m} = (0, -0.984572, 0.984572, 0).$$

Notice how close the second component of $\tilde{\rho}$ is to the second component of ρ. This is not surprising because it is a projection, which should be the point in $\mathcal{K}$ closest to (ρ, m). However, in the sense of the optimality conditions of Lemma 5.1 this means that

$$a = \rho - \tilde{\rho} = (0.1421303, 0.0016)$$

can get very close to zero.

This a is used in the optimality conditions (II) and (III) which means that we plug in values into θ or $\partial_1\theta$ that are sometimes very close to zero. These functions however have singularities in zero. Also the logarithmic mean is only defined on $[0,\infty)^2$. This poses another problem for the numerical projection because whenever in an early iteration (of alternating gradient descent or Newton) we land in $\mathbb{R}^2 \backslash [0,\infty)^2$ our algorithm stops and has failed because since the logarithmic mean is not defined in this area there is no way to get back to the correct quadrant. The numerical instabilities of computing the projection can all be traced back to those and similar problems.

A possible solution for the numerical instabilities of the projection can be to extend the logarithmic mean to the whole $\mathbb{R} \times \mathbb{R}$. In the following we will describe two methods of doing so.

The easiest way to extend the logarithmic mean is to simply do a quadratic extension of the logarithm that is used so that the extended logarithmic mean is defined on $(-\infty,\infty)$ instead of just on $(0,\infty)$. For this we introduce a parameter $\varepsilon > 0$ and define

$$\log_\varepsilon(x) := \begin{cases} \log(x) & x \geq \varepsilon \\ a_\varepsilon x^2 + b_\varepsilon x + c_\varepsilon & x < \varepsilon \end{cases}.$$

Here the parameters $a_\varepsilon, b_\varepsilon$ and c_ε are implicitly given because the quadratic extension should be C^2 in order to be viable for applying Newton's method. This means we get the system of equations

$$\begin{aligned} \log(x) &= a_\varepsilon x^2 + b_\varepsilon x + c_\varepsilon \\ \log'(x) &= 2a_\varepsilon x + b_\varepsilon \\ \log''(x) &= c_\varepsilon \end{aligned}.$$

The resulting $\log_\varepsilon$ and θ_ε are displayed in Figures 5.2 and 5.3.

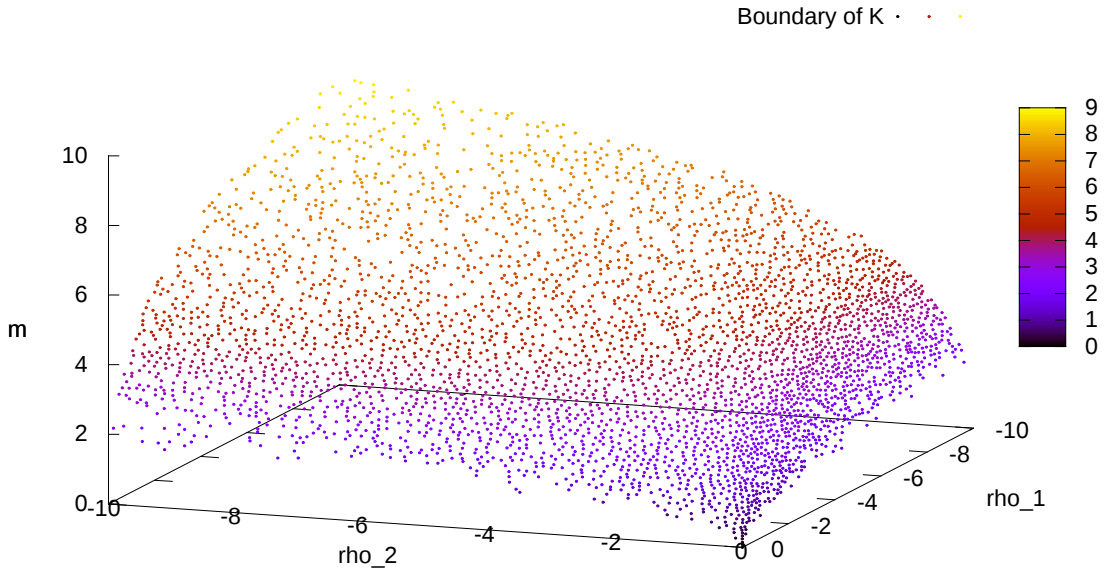


Figure 5.1: Point cloud of projections on $\mathcal{K}$ for the 2-point instance described in the example above. The x and y-axis represent ρ_1 and ρ_2, the z-axis represents the m-dimension. In the 2-point case the m dimension is effectively one-dimensional because since $Q(a,a) = Q(b,b) = 0$ we have $m_{a,a} = m_{b,b} = 0$ and also $m_{a,b} = -m_{a,b}$. Points were uniformly drawn from $[-10, 10]^3$ and then projected on $\mathcal{K}$ to generate the point cloud. The plot only shows points with $m \geq 0$. However, the set is actually symmetric with respect to the ρ_1-ρ_2-axis.

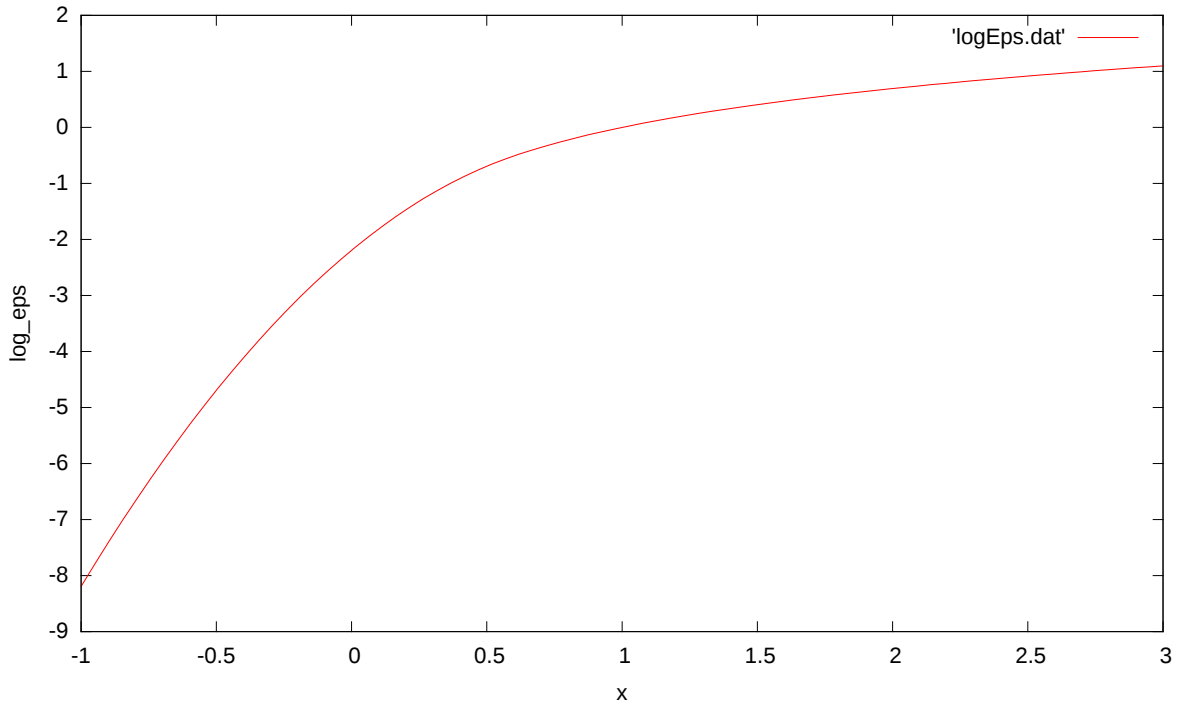


Figure 5.2: Quadratic extension of the logarithm ($\varepsilon = 0.5$). For this ε the parameters are given by $a_\varepsilon = -2$, $b_\varepsilon = 4$, $c_\varepsilon = -2.19315$.

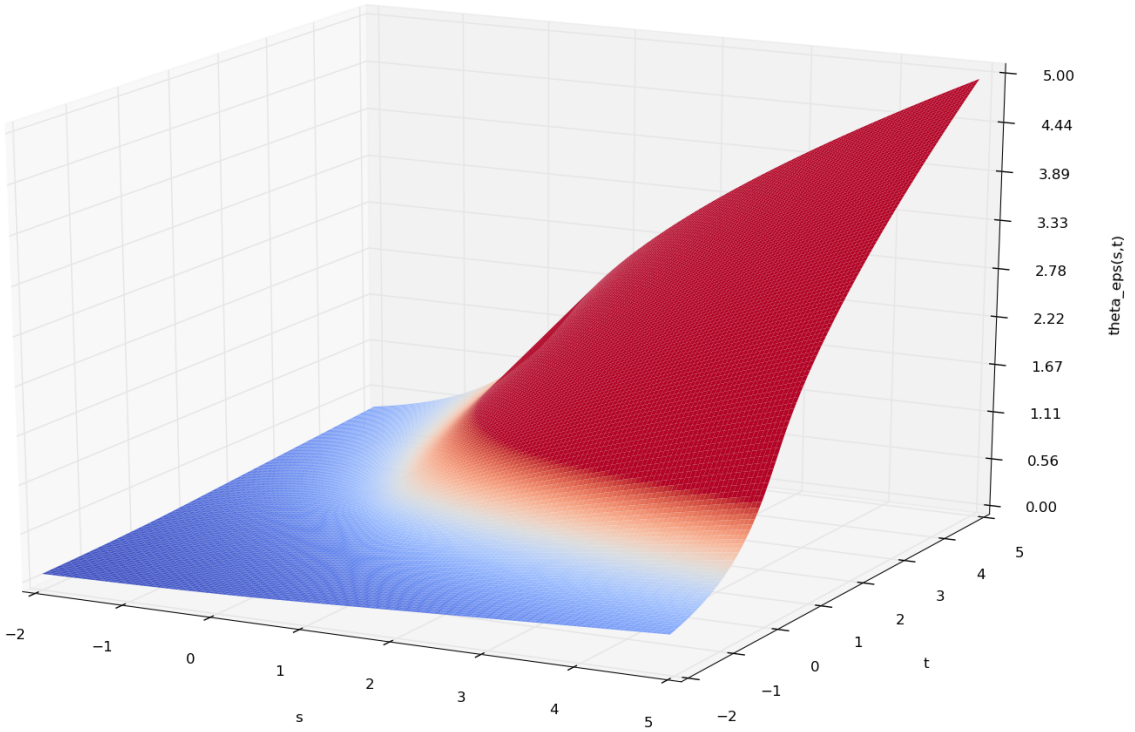


Figure 5.3: Extension of the logarithmic mean using the quadratic extension of the logarithm ($\varepsilon = 0.5$). Compare this with the original logarithmic mean in Figure 3.1.

6 Discretization and Implementation Details

After we introduced the theoretical concepts for our algorithms in chapter 4 and 5 we will now talk about some particular implementation details. We will not provide any code snippets or even pseudo-code, however we will explain all numerical choices that were made to reach a fully functional C++ - implementation so that it should be possible to reproduce the results that we will present in chapter 7.

We start by introducing some basic notions and the finite element spaces we will use for the discretization. Recall that a geodesic with respect to the discrete Wasserstein metric is encoded in a map $t \mapsto (\rho_t, m_t)$. Thus we have three dimensions to worry about: time, nodes and edges. Naturally nodes and edges are already discrete in our setting. As a result we only have to introduce a time-discretization and "tensorize" it with nodes and edges in order to fully describe a geodesic discretized in time and space. From now on

- N_T denotes the number of timesteps $\{t_0, \dots, t_{N_T-1}\}$ used for the discretization of the time dimension including $t = 0$ and $t = 1$,
- $N_{\mathcal{X}} := |\mathcal{X}|$ denotes the number of nodes and $M_{\mathcal{X}} := |\mathcal{X} \times \mathcal{X}| = |\mathcal{X}|^2$ the number of all possible edges,
- h denotes the mesh size used in the time discretization, i.e. $h = \frac{1}{N_T - 1}$.

Definition 6.1. *We introduce the following piece-wise affine and piece-wise constant finite element spaces*

$$
\begin{aligned}
V^1_{h,N} &:= \left\{ f : [0,1] \times \mathcal{X} \to \mathbb{R} \;\middle|\; \forall x \in \mathcal{X} : f(\cdot, x) \in \mathcal{S}^{1,0}([0,1]) \right\}, \\
V^0_{h,N} &:= \left\{ f : [0,1] \times \mathcal{X} \to \mathbb{R} \;\middle|\; \forall x \in \mathcal{X} : f(\cdot, x)|_{[t_i, t_{i+1}]} \in \Pi^1_0 \right\}, \\
V^1_{h,E} &:= \left\{ f : [0,1] \times (\mathcal{X} \times \mathcal{X}) \to \mathbb{R} \;\middle|\; \forall x, y \in \mathcal{X} : f(\cdot, x, y) \in \mathcal{S}^{1,0}([0,1]) \right\}, \\
V^0_{h,E} &:= \left\{ f : [0,1] \times (\mathcal{X} \times \mathcal{X}) \to \mathbb{R} \;\middle|\; \forall x, y \in \mathcal{X} : f(\cdot, x, y)|_{[t_i, t_{i+1}]} \in \Pi^1_0 \right\},
\end{aligned}
$$

where

$$
\mathcal{S}^{k,m}(\Omega) := \left\{ u \in C^m(\bar{\Omega}) \;\middle|\; u|_{[t_i, t_{i+1}]} \in \Pi^1_k \right\}
$$

and Π^d_k denotes d-variate polynomials of order lower or equal to k. We denote the number of elements in those spaces by

$$
\begin{aligned}
N^1_h &= N_T N_{\mathcal{X}} \text{ for } V^1_{h,N}, & N^0_h &= (N_T - 1) N_{\mathcal{X}} \text{ for } V^0_{h,N}, \\
M^1_h &= N_T M_{\mathcal{X}} \text{ for } V^1_{h,E}, & M^0_h &= (N_T - 1) M_{\mathcal{X}} \text{ for } V^0_{h,E}.
\end{aligned}
$$

The following f_{t_i} form a basis of $\mathcal{S}^{1,0}$ and g_{t_i} form a basis of piece-wise constant functions in

one variable,

$$f_{t_i}(t) = \begin{cases} 1 - \frac{1}{h}|t - t_i|, & |t - t_i| \leq h \\ 0, & \text{otherwise} \end{cases}, \qquad g_{t_i}(t) = \begin{cases} 1 & t \in [t_i, t_{i+1}] \\ 0, & \text{otherwise} \end{cases}.$$

In order to give an enumeration of the basis functions we define the bijective mappings

$$\sigma^1 : \{0, \dots, N_h^1 - 1\} \to \{0, \dots N_T - 1\} \times \mathcal{X}$$
$$\tau^1 : \{0, \dots, M_h^1 - 1\} \to \{0, \dots N_T - 1\} \times (\mathcal{X} \times \mathcal{X})$$

and analogue σ^0 and τ^0 for $V_{h,N}^0$ and $V_{h,E}^0$. With this enumeration we can easily construct the basis functions of all finite element spaces defined in Definition 6.1. For example we could denote a set of basis functions for $V_{h,N}^1$ by $\left\{\Psi_i^{1,h,N}\right\}_{i=0}^{N_h^1}$ where

$$(t, x) \mapsto \Psi_i^{1,h,N}(t, x) = \begin{cases} f_{\sigma_t^1(i)}(t) & \text{if } \sigma_x^1(i) = x \\ 0 & \text{otherwise} . \end{cases}$$

This can be done in a similar fashion for $V_{h,N}^0$, $V_{h,E}^1$ and $V_{h,E}^0$.

6.1 Solving the elliptic problem

Now that we have the finite element definitions in place we can start thinking about how to implement the actual algorithms. Recall that in step A of the Augmented Lagrangian method (BB) we need to compute a $\xi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$ for which we found out in Lemma 4.7 that $(\rho + \partial_t \xi, m + \nabla_\chi \xi)$ is just the projection of (ρ, m) on the set of such pairs that satisfy the continuity equation. This projection on the set of continuity equation also has to be done in the first step of the Douglas-Rachford splitting algorithm. So in both cases we need to be able to solve the weak formulation of an elliptic partial differential equation

$$\int_0^1 \left(\langle \partial_t \xi_t, \partial_t \varphi_t \rangle_\pi + \langle \nabla_\chi \xi_t, \nabla_\chi \varphi_t \rangle_\pi\right) \mathrm{d}t = \langle \varphi_1, \rho_B \rangle_\pi - \langle \varphi_0, \rho_A \rangle_\pi - \int_0^1 \left(\langle \rho_t, \partial_t \varphi_t \rangle_\pi + \langle m_t, \nabla_\chi \varphi_t \rangle_\pi\right) \mathrm{d}t. \tag{6.1}$$

for $\xi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$ with given $\rho \in L^2([0,1], \mathbb{R}^{\mathcal{X}})$ and $m \in L^2([0,1], \mathbb{R}^{\mathcal{X} \times \mathcal{X}})$. We define

$$a(\xi, \varphi) := \int_0^1 \left(\langle \partial_t \xi_t, \partial_t \varphi_t \rangle_\pi + \langle \nabla_x \xi_t, \nabla_x \varphi_t \rangle_\pi\right) \mathrm{d}t$$
$$l(\varphi) := \langle \varphi_1, \rho_B \rangle_\pi - \langle \varphi_0, \rho_A \rangle_\pi - \int_0^1 \left(\langle \rho_t, \partial_t \varphi_t \rangle_\pi + \langle m_t, \nabla_\chi \varphi_t \rangle_\pi\right) \mathrm{d}t.$$

Now (6.1) reads: find $\xi \in H^1([0,1], \mathbb{R}^{\mathcal{X}})$ such that

$$a(\xi, \varphi) = l(\varphi) \quad \forall \varphi \in H^1([0,1], \mathbb{R}^{\mathcal{X}}).$$

For a conform discretization of $H^1([0,1], \mathbb{R}^{\mathcal{X}})$ we choose the space of piece-wise affine functions $V_{h,N}^1$. Because we need no regularity on ρ and m we discretize them using the piece-wise constant spaces $V_{h,N}^0$ and $V_{h,E}^0$. This will later help us with the projection on $\mathcal{K}$. As a result solving the discrete elliptic problem means finding $\xi_h \in V_{h,N}^1$ such that

$$a(\xi_h, \varphi_h) = l(\varphi_h) \quad \forall \varphi_h \in V_{h,N}^1.$$

Since $a(\cdot, \cdot)$ is bilinear, this is equivalent to solving the following system of equations

$$a(\xi_h, \Psi_i) = l(\Psi_i), \quad i = 0, \dots, N - 1 \tag{6.2}$$

where $\{\Psi_i\}_{i=0}^{N-1}$ are basis functions of $V_{h,N}^1$. By definition ξ_h can be written as $\xi_h = \sum_{j=0}^{N} \bar{\xi}_j \Psi_j$ and again by bilinearity of a we get that (6.2) is equivalent to

$$A\bar{\xi} = b \quad \text{where} \quad a_{ij} := a(\Psi_i, \Psi_j),\ b_i := l(\Psi_i). \tag{6.3}$$

So solving the elliptic problem comes down to solving a linear system of equations with the so-called stiffness matrix A and the right-hand side b.

Remark: *Because it contains only derivatives the stiffness matrix is not invertible. So analogous to [BB00] we do not invert A but $A+\varepsilon M$ where $M = (m_{ij})_{i,j}$ is the so-called mass matrix given by*

$$m_{ij} = \int_0^1 \langle \Psi_i, \Psi_j \rangle_\pi \ \mathrm{d}t.$$

For small epsilon this does not change the solution ξ that much but makes the matrix invertible so that the problem is actually well posed.

Concerning the actual implementation of the elliptic problem solver in C++ there are very few things to consider. First of all we are using the EIGEN-library [1] for all linear algebra related computations. The number of basis functions grows quadratically in the number of nodes of the instance so the linear equation systems (6.3) can grow quickly. Because the stiffness matrix contains a lot of zeros we are using a special implementation of sparse matrices to store the stiffness matrix. Also, as is often the case with finite element methods, we are using a conjugate gradient method for solving the system of linear equations. With these considerations solving the elliptic problem for instances used in this thesis is very fast and should even be feasible for larger instances with node counts in the hundreds. The real bottleneck is the projection on $\mathcal{K}$ that is discussed in the next section.

6.2 Step B of (BB)

Recall that for step B of (BB) we have to be able to calculate

$$\operatorname*{arg\,min}_{(\rho^*, m^*) \in [0,1] \times \mathcal{K}} \int_0^1 \left(\|\partial_t \xi_t + \frac{\rho_t}{r} - \rho_t^*\|_\pi^2 + \|\nabla_\chi \xi_t + \frac{m_t}{r} - m_t^*\|_\pi^2 \right) \mathrm{d}t.$$

However, before we talk about the implementation of the projection itself we have to deal with another small problem. Because $\xi \in V_{h,N}^1$ we can naturally interpret $\partial_t \xi$ as an element of $V_{h,N}^0$. However, we have $\nabla_x \xi_t \in V_{h,E}^1$ whereas m_t and m_t^* are elements of $V_{h,E}^0$. Thus in practice we can not calculate for example $\nabla_\chi \xi_t + \frac{m_t}{r}$ because the summands live in different spaces.

We will solve this problem by considering a piece-wise constant interpretation of $\nabla_x \xi_t$ which does not change the minimizer of the problem. To this end define an operator $\mathcal{R} : V_{h,E}^1 \to V_{h,E}^0$ by

$$f \mapsto \mathcal{R} f(\cdot, x, y)|_{[t_i, t_{i+1}]} = \frac{f(t_i, x, y) + f(t_{i+1}, x, y)}{2} \quad \forall x, y \in \mathcal{X}, i = 0, \ldots, N_T - 1.$$

This means we map each linear slope to the mean of its two endpoints. Also $\mathcal{R}$ can be interpreted as the L^2-projection of the piece-wise affine function to its piece-wise constant pendant.

In the following lemma we prove that using $\mathcal{R}$ does indeed not change the minimizer in (6.4).

[1] http://eigen.tuxfamily.org

Lemma 6.2. *For $\xi \in V^1_{h,N}$ and $m, m^* \in V^0_{h,E}$ it holds*

$$\arg\min_{m^*} \int_0^1 \|\mathcal{R}(\nabla_x \xi)_t + \frac{m_t}{r} - m_t^*\|_\pi^2 \, \mathrm{d}t = \arg\min_{m^*} \int_0^1 \|\nabla_x \xi_t + \frac{m_t}{r} - m_t^*\|_\pi^2 \, \mathrm{d}t$$

Proof. We start by calculating

$$\begin{aligned}
\int_0^1 \|\mathcal{R}(\nabla_x \xi)_t + \frac{m_t}{r} - m_t^*\|_\pi^2 \, \mathrm{d}t &= \int_0^1 \langle \mathcal{R}(\nabla_x \xi)_t + \frac{m_t}{r} - m_t^*, \mathcal{R}(\nabla_x \xi)_t + \frac{m_t}{r} - m_t^* \rangle_\pi \, \mathrm{d}t \\
&= \int_0^1 \langle \mathcal{R}(\nabla_x \xi)_t, \mathcal{R}(\nabla_x \xi)_t \rangle_\pi \, \mathrm{d}t - 2 \int_0^1 \langle \mathcal{R}(\nabla_x \xi)_t, \frac{m_t}{r} - m_t^* \rangle_\pi \, \mathrm{d}t \\
&\quad + \int_0^1 \langle \frac{m_t}{r} - m_t^*, \frac{m_t}{r} - m_t^* \rangle_\pi \, \mathrm{d}t
\end{aligned}$$

Now because of what we will prove in the next lemma we have

$$\int_0^1 \langle \mathcal{R}(\nabla_x \xi)_t, \frac{m_t}{r} - m_t^* \rangle_\pi \, \mathrm{d}t = \int_0^1 \langle \nabla_x \xi_t, \frac{m_t}{r} - m_t^* \rangle_\pi \, \mathrm{d}t.$$

Continuing the calculation above we get

$$\begin{aligned}
\ldots &= \int_0^1 \langle \mathcal{R}(\nabla_x \xi)_t, \mathcal{R}(\nabla_x \xi)_t \rangle_\pi \, \mathrm{d}t - 2 \int_0^1 \langle \nabla_x \xi_t, \frac{m_t}{r} - m_t^* \rangle_\pi \, \mathrm{d}t + \int_0^1 \langle \frac{m_t}{r} - m_t^*, \frac{m_t}{r} - m_t^* \rangle_\pi \, \mathrm{d}t \\
&= \int_0^1 \|\mathcal{R}(\nabla_x \xi)_t\|_\pi^2 \, \mathrm{d}t - \int_0^1 \|\nabla_x \xi_t\|_\pi^2 \, \mathrm{d}t + \int_0^1 \|\nabla_x \xi_t + \frac{m_t}{r} - m_t^*\|_\pi^2 \, \mathrm{d}t.
\end{aligned}$$

Note that the first two integrals do not contain m^*. Thus they don't contribute to the arg min and the statement follows. ∎

Lemma 6.3. *Let $f \in \mathcal{S}^{1,0}([0,1])$, $g \in \{\varphi : [0,1] \to \mathbb{R} : \varphi|_{[t_i, t_{i+1}]} \in \Pi_0^1\}$, then*

$$\int_0^1 fg \, \mathrm{d}t = \int_0^1 (\mathcal{R}f) g \, \mathrm{d}t.$$

Proof.

$$\begin{aligned}
\int_0^1 fg \,\mathrm{d}t &= \sum_{i=0}^{N_T-1} \int_{t_i}^{t_{i+1}} fg \,\mathrm{d}t = \sum_{i=0}^{N_T-1} g_i \int_{t_i}^{t_{i+1}} m_i t + b_i \,\mathrm{d}t \\
&= \sum_{i=0}^{N_T-1} g_i \int_{t_i}^{t_{i+1}} \left(\frac{m_i}{2} t_{i+1}^2 + b_i t_{i+1} - \frac{m_i}{2} t_i^2 b_i t_i \right) \mathrm{d}t \\
&= \sum_{i=0}^{N_T-1} g_i \int_{t_i}^{t_{i+1}} \left(\frac{m_i}{2} \underbrace{(t_{i+1}^2 - t_i^2)}_{=(t_{i+1}-t_i)(t_{i+1}+t_i)} + b_i(t_{i+1} - t_i) \right) \mathrm{d}t \\
&= \sum_{i=0}^{N_T-1} g_i h \int_{t_i}^{t_{i+1}} \left(\frac{m_i}{2}(t_{i+1} + t_i) + b_i \right) \mathrm{d}t \\
&= \sum_{i=0}^{N_T-1} g_i h \int_{t_i}^{t_{i+1}} \left(\frac{f(t_i) + f(t_{i+1})}{2} \right) \mathrm{d}t = \sum_{i=0}^{N_T-1} g_i h (\mathcal{R}f)|_{[t_i, t_{i+1}]} \\
&= \int_0^1 g(\mathcal{R}f) \,\mathrm{d}t
\end{aligned}$$

■

6.3 Computing the projection on $\mathcal{K}$

After the small fix in the previous section we have now arrived at a point where step B of `(BB)` and step 2 of `(DR)` break down to the same computational problem, namely

$$\operatorname*{arg\,min}_{(\rho^*, m^*) \in [0,1] \times \mathcal{K}} \int_0^1 \left(\|\bar{\rho}_t - \rho_t^*\|_\pi^2 + \|\bar{m}_t - m_t^*\|_\pi^2 \right) \mathrm{d}t \tag{6.4}$$

where in the `(BB)` case we have $\bar{\rho} = \partial_t \xi_t + \frac{\rho_t}{r}$ and $\bar{m}_t = \mathcal{R}(\nabla_\chi \xi_t) + \frac{m}{r}$. This problem is effectively a projection of some $(\bar{\rho}, \bar{m}) \in [0,1] \times \mathcal{H}$ to the set $[0,1] \times \mathcal{K}$, i.e. for each timestep $t \in [0,1]$ we need $(\rho_t^*, m_t^*) \in \mathcal{K}$. Because $\bar{\rho}$ and $\bar{m}$ are given through a piece-wise constant discretization ($\bar{\rho} \in V_{h,N}^0$ and $\bar{m} \in V_{h,E}^0$) the integral in (6.4) becomes a sum and the minimization problem becomes

$$\begin{aligned}
&\operatorname*{arg\,min}_{(\rho^*, m^*) \in \{t_0, t_1, \ldots, t_N\} \times \mathcal{K}} h \sum_{t_i} \left(\|\bar{\rho}_{t_i} - \rho_{t_i}^*\|_\pi^2 + \|\bar{m}_{t_i} - m_{t_i}^*\|_\pi^2 \right) \\
&\qquad = h \sum_{t_i} \operatorname*{arg\,min}_{(\rho_{t_i}^*, m_{t_i}^*) \in \mathcal{K}} \left(\|\bar{\rho}_{t_i} - \rho_{t_i}^*\|_\pi^2 + \|\bar{m}_{t_i} - m_{t_i}^*\|_\pi^2 \right)
\end{aligned}$$

As a result the projection problem decouples in time and we can simply project on $\mathcal{K}$ for each timestep separately.

For computing a projection on $\mathcal{K}$ we use the optimality conditions derived in Lemma 5.1. In our implementation we initially started with an alternating gradient descent (i.e. alternatingly updating each equation until convergence). However, this method turned out to be rather slow and very unstable. As a result we moved to Newton's method with optimal step size control as described in [SW05]. To use it we also needed to compute the second derivative of the logarithmic mean which is the reason we need the extension of logarithmic mean to be C^2. With Newton's method we were able to compute many of the results that we will present in the following chapter. However, there seem to be still some improvements possible. For example we also tried a nonlinear solver directly implemented in the `EIGEN`-library. This implementation uses Powell's hybrid method described in [Pow70]. Using this method showed to be even more

stable and faster than just using Newton's method by itself.

Even with this advanced solvers however stability is still a huge issue when computing the projection on $\mathcal{K}$. By that we mean that the start values with which we initialize our algorithms determine whether we find a correct the projection or get trapped in some wrong solutions to the optimality conditions. To kind of skip this problem we use many different initial values for the algorithms until we find the correct projection. We tried different methods for guessing these start values. For example trying values close to the (ρ, m) that should be projected or trying start values that were successful for previous iterations of the grand numerical scheme. However, for later iterations of `(BB)` or `(DR)` it takes longer and longer to find working start values that converge to the correct projection. This behaviour is related to the issue mentioned in the example in Chapter 5. When the correct projected values are close to the values that should be projected then a becomes very close to zero which lets us run into the singularities of the logarithmic mean.

In this cases we used the extended logarithmic mean in the optimality conditions to guess some start values that can afterwards used as initializations of the real projection. Nevertheless there are still many cases where it just takes to long to find working start values. Some possible solutions that were not yet tried are described in Chapter 8.

7 Results

In this chapter we will present results from applying the both algorithms described in chapter 4 for different problem instances. We will go from the most basic instances for which there exist explicit formulas for calculating the Wasserstein distance to more complex examples. Instances are always introduced in a standardized way and labelled with a unique name for reference. After introducing an instance we will first discuss the expected behaviour for a particular instance and afterwards present plots and visualizations of the calculated geodesics. We can then compare expected and actual behaviour and try to answer questions about the geometry of that instance.

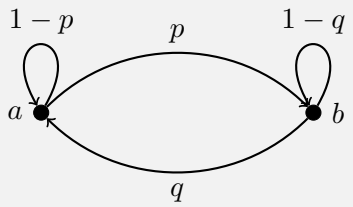
In the first section we will start with the most basic instance consisting only of two nodes. For this instance there exist explicit formulas that characterize the geodesic and we will first recall those and then compare our results with them. In the next section we will look at instances with three or four nodes and try to answer some general questions about how the mass is transported in discrete Wasserstein distance. Finally we will look at slightly bigger instances with interesting geometry, mainly circle, lines or grids.

Note that all results in the first four sections of this chapter were computed using the Douglas-Rachford algorithm. This is done because in practice we observe better convergence properties for it than for the Benamou-Brenier approach.

7.1 Comparison of analytical and approximate geodesic in the two-point Case

Right in the article [Maa11] where the discrete Wasserstein distance was introduced the author calculated a relatively easy closed form for the distance on finite sets with only two points. First lets characterize these instances.

Instance 1 (`2_Circle`)**.**



This most basic non-trivial instance consists of two nodes $\mathcal{X} = \{a, b\}$ connected by two edges. There are two degrees of freedom for the transition properties which we characterize by the variables $0 \leq p, q \leq 1$. Then Q is given by

$$Q = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

and the corresponding stationary distribution is $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$. In this instance we look at a full transport where all mass starts at a and is moved to b, i.e. $\rho_A = (1, 0)$, $\rho_B = (0, 1)$.

We will now quickly recap the results from [Maa11] which deal with this type of instance and give an explicit way of calculating the geodesic. First note that each probability density ρ with

respect to π is characterized by just one parameter β. To be more precise we denote

$$\rho^\beta(a) := \frac{p+q}{q}\frac{1-\beta}{2}, \quad \rho^\beta(b) := \frac{p+q}{p}\frac{1+\beta}{2}.$$

As a result we can describe the logarithmic mean between a and b through β

$$\hat{\rho}(\beta) := \theta(\rho^\beta(a), \rho^\beta(b)).$$

Using a characterization of the geodesic through Euler-Lagrange equations and the implicit function theorem the author of [Maa11] derives the following formula for the discrete Wasserstein distance.

Theorem 7.1 ([Maa11, Theorem 2.4])**.** *For $-1 \leq \alpha \leq \beta \leq 1$ we have*

$$\mathcal{W}(\rho^\alpha, \rho^\beta) = \frac{1}{2}\sqrt{\frac{1}{p}+\frac{1}{q}}\int_\alpha^\beta \frac{1}{\sqrt{\hat{\rho}(r)}}\,\mathrm{d}r \in [0,\infty].$$

Also doing some more computations one can get the following result.

Corollary 7.2. *If we also assume $p = q$ we can simplify $\hat{\rho}(\beta) = \frac{\beta}{\operatorname{arctanh}(\beta)}$ and get*

$$\mathcal{W}(\rho^\alpha, \rho^\beta) = \frac{1}{\sqrt{2p}}\int_\alpha^\beta \sqrt{\frac{\operatorname{arctanh}(r)}{r}}\,\mathrm{d}r.$$

Additionally by using that the Wasserstein distance is given by an integral one can derive the following ordinary differential equation which characterizes not only the distance but also geodesics with respect to the metric in the two-point case.

Proposition 7.3. *[[Maa11, Proposition 2.7]] Let $\rho_A, \rho_B \in \mathcal{P}(\mathcal{X})$. There exists a unique constant speed geodesic $\{\rho^{\gamma(t)}\}_{0\leq t\leq 1}$ in $\mathcal{P}(\mathcal{X})$ with $\rho^{\gamma(0)} = \rho_A$ and $\rho^{\gamma(1)} = \rho_B$. Moreover, the function γ belongs to $C^1([0,1],\mathbb{R})$ and satisfies the ordinary differential equation*

$$\gamma'(t) = 2w\sqrt{\frac{pq}{p+q}\hat{\rho}(\gamma(t))} \tag{7.1}$$

for $t \in [0,1]$, where $w := \operatorname{sgn}(\beta-\alpha)\mathcal{W}(\rho^\alpha, \rho^\beta)$.

Using Proposition 7.3 we have a way of computing the correct geodesic for the two-point instance. If we can solve (7.1) for γ we get the progression of mass in a and b. Fortunately at least for the case $p = q$ the ODE can be solved fairly easily by using an explicit Euler scheme. This means we choose a discretization level N_T and then recursively update

$$\gamma(n+1) = \gamma(n) + h * f(nh, \gamma(n)), \qquad n = 1, \ldots, N_T$$

where $h = \frac{1}{N_T}$ and $f(t,x) = 2w\sqrt{\frac{pq}{p+q}\hat{\rho}(x)}$.

Finally we can compare the geodesic from the ODE with our results. In this chapter we will always display the geodesics $t \mapsto (\rho_t, m_t)$ by plotting the progression $t \mapsto \rho_t(x)$ for all (only relevant) nodes $x \in \mathcal{X}$. The flow part m of a geodesic is only looked at if it is especially relevant in a particular instance. In Figure 7.1 we can see the progression of mass in node a for the `(2_Circle)` instance computed from the ODE and calculated from our Douglas-Rachford algorithm. We only see very small differences between the two slopes which do get even smaller if one goes for a finer time discretization and more iterations of the algorithm. A more detailed comparison between the number of timesteps in our algorithms and the resulting error compared to the correct ODE solution can be found in Figure 7.2.

These matching results from two completely different approaches make us confident that our algorithms are indeed working correctly and computing the correct geodesic. Thus we can now go on computing geodesics for cases where there is yet no explicit theory for.

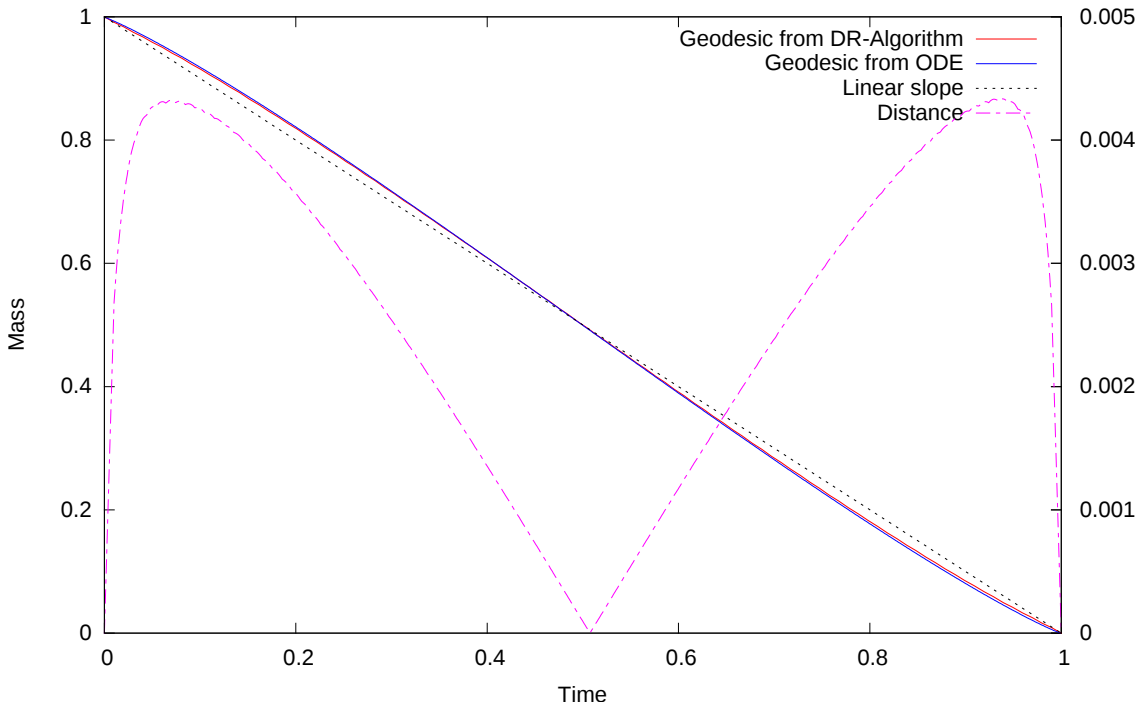


Figure 7.1: Comparison between the geodesic calculated using the ODE in [Maa11] and the result from the Douglas-Rachford algorithm. We plot the mass at node a over the time interval $[0, 1]$. Apparently the two slopes look identical, the distance is plotted with respect to the second y-axis on the right side.
— *Instance:* `(2_Circle)` with $p = q = 1$, *Algorithm:* `(DR)`, *Timesteps:* 100, $\mathcal{W}(\rho_A, \rho_B) \approx 1.13394$ —

7.2 Triangles and Squares

In this next section we are going to look at small instances with three or four nodes. The results are interesting even for this small instances because at the time writing there is yet no descriptive theory about the geodesics of the new discrete Wasserstein distance. Also these instance already cover some important questions. Some examples are:

- How strong is a longer path penalized compared to a shorter one?

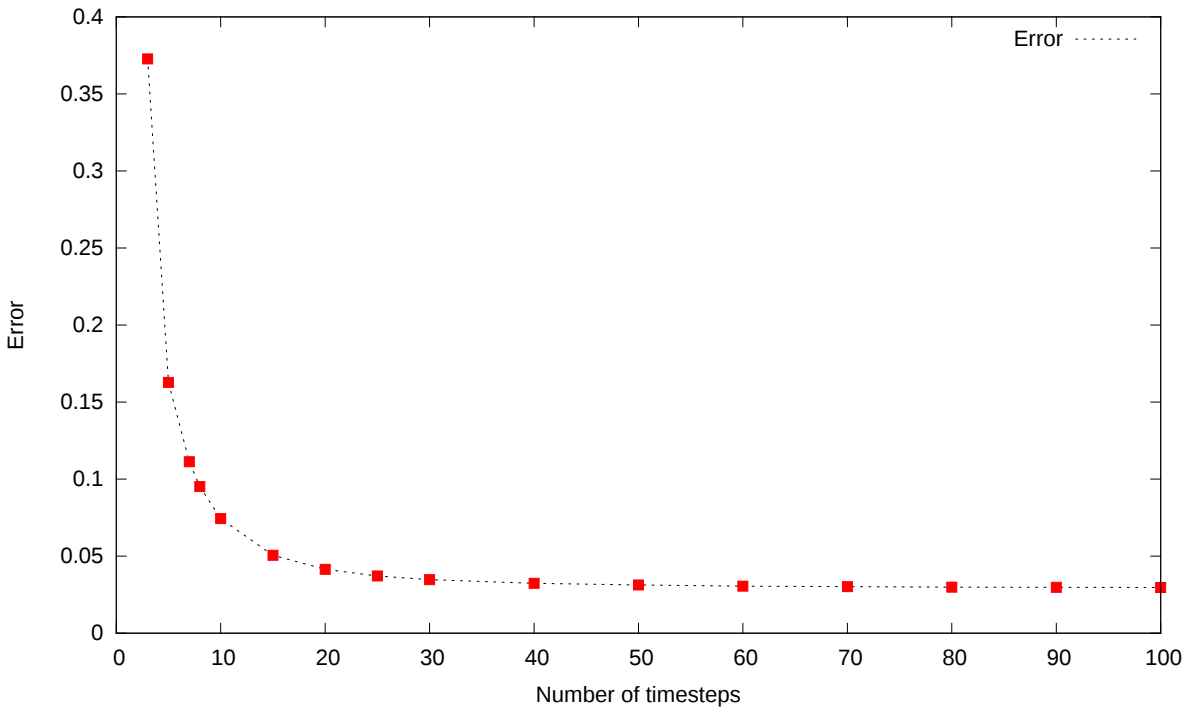
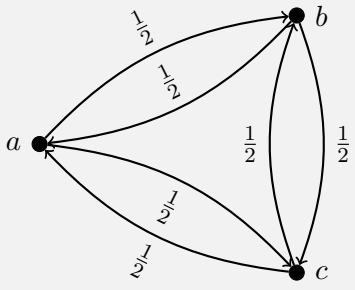


Figure 7.2: Approximation quality of the `(DR)`-algorithm where we set the number of timesteps N_T of our time-discretization in relation with the distance of the final result compared to the geodesic calculated using an explicit Euler's method on the ODE ($n = 1000$). The remaining distance even for a larger number of timesteps may be due to imperfections of the several numerical methods involved.
— *Instance:* `(2_Circle)`, *Algorithm:* `(DR)` —

- How much is the flow going to be stronger in the beginning when there is a bigger mass difference?

We start with very much the only interesting instance with three nodes.

Instance 2 (`3_Circle`).



This instance represent a fully uniformly connected triangle. Thus the transition matrix is given by

$$Q = \frac{1}{2}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

with stationary distribution $\pi = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Again we want to transport all mass from a to c, i.e. $\rho_A = (1, 0, 0)$, $\rho_B = (0, 0, 1)$.

The `(3_Circle)`-instance is interesting because on one hand there is a direct connection between node a and node c but on the other hand there is a longer (in the sense of the weights on the edges) connection that runs over b. Now the interesting questions are: Is there any mass taking the long way over b? and if so: How much mass does flow over b?

Numerical answers to this questions can be found in Figure 7.3. There we plot the geodesic in the same way we did for the `(2_Circle)`-instance. Using this plot it is immediately clear that

there is indeed flow over node b although the majority of the mass at a is directly transported to c. Another expected behaviour is the symmetry of the mass in node a compared to node c. This may not surprising as we are always looking at reversible Markov chains.

What is more interesting is the fact that the progression of mass in node a is not symmetric with respect to $t = \frac{1}{2}$. We can observe that in the beginning the flow of mass is very much like a linear slope. However, as more and more mass has already been transported the flow increases until at some point way beyond $t = \frac{1}{2}$ it slows down again. This behaviour can probably be explained by the use the logarithmic mean in the metric. It introduces a nonlinear dependence on how much mass is already at some point which is exactly what we are observing.

One might also note that the calculated discrete Wasserstein distance $\mathcal{W}(\rho_A, \rho_B)$ of `(2_Circle)` is lower than the one for `(3_Circle)`. This may seem strange because the `(3_Circle)`-instance only adds a new path compared to `(2_Circle)` but still offers a direct connection between nodes a and c. However, the difference lies in transition probabilities. While above for the `(2_Circle)`-instance we used $p = q = 1$ `(3_Circle)` has $\frac{1}{2}$ everywhere. This makes the transport more expensive.

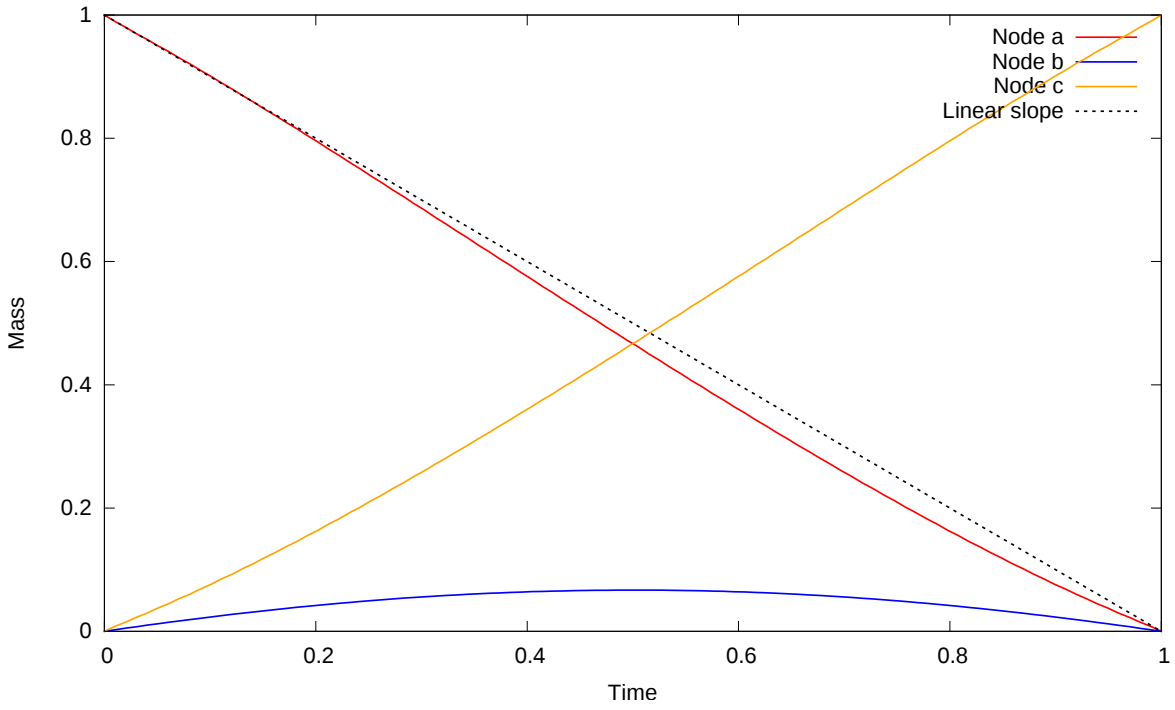
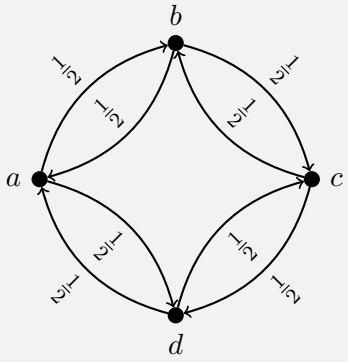


Figure 7.3: — *Instance:* `(3_Circle)`, *Algorithm:* `(DR)`, $\mathcal{W}(\rho_A, \rho_B) \approx 1.18999$ —

As there are not that many fundamentally different geometries we can build on three nodes we move to instances with four nodes.

Instance 3 (`4_Circle`)**.**



This instance is a circle with 4 nodes that are each connected to their two neighbours. The transition matrix is given by

$$Q = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

with stationary distribution $\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. To transport all mass from a to c (the node not directly connected to a) we set

$$\rho_A = (1, 0, 0, 0), \quad \rho_B = (0, 0, 1, 0).$$

Our interest with this graph lies in the fact that so far it is the first instance for which start- and end-node are not directly connected. For the intermediate node b and d we expect a similar behaviour as for node b of `(3_Circle)`. However, this time we know that all mass has to be transported over either b or d. The plots of the geodesic can be found in Figure 7.4.

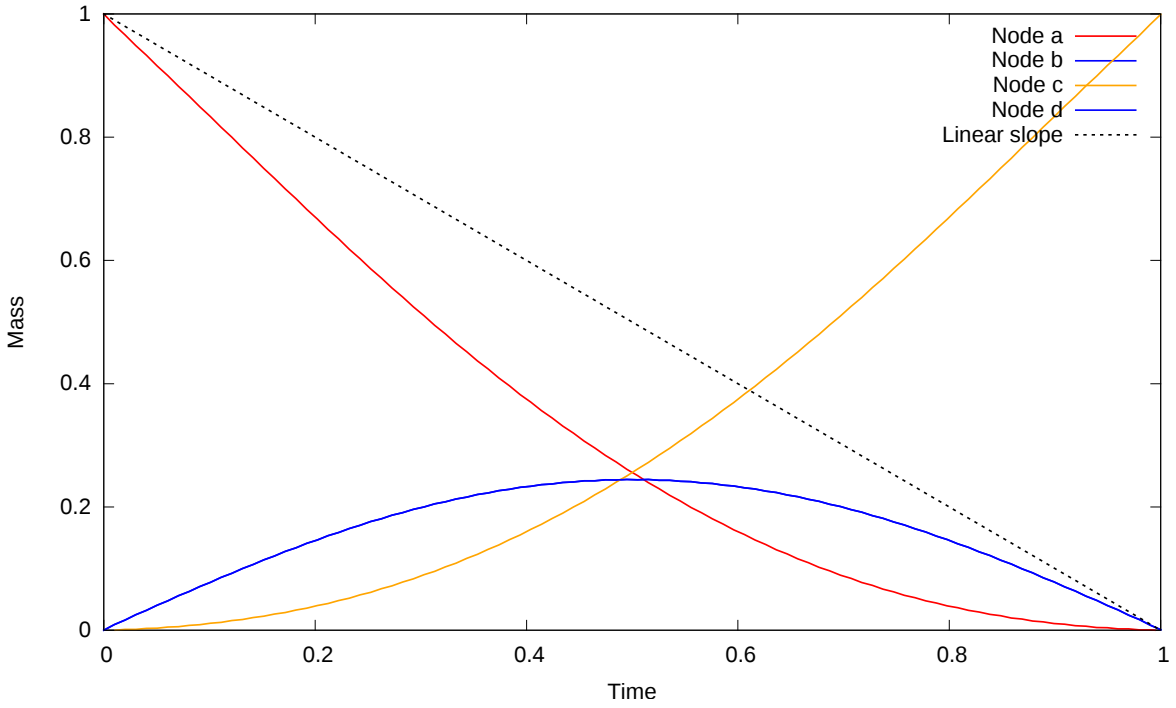


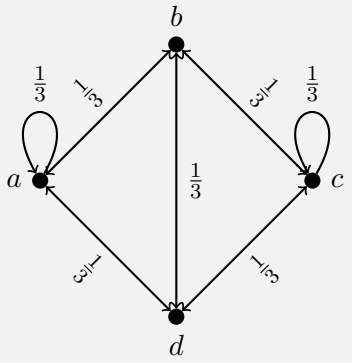
Figure 7.4: — *Instance:* `(4_Circle)`, *Algorithm:* `(DR)`, $\mathcal{W}(\rho_A, \rho_B) \approx 1.56855$ —

In a few pages we will introduce the `(3_Line)` instance which consists of three nodes where the last one is not connected with the first one. An interesting comparison can be done by using the geodesic calculated for `(3_Line)` in `(4_Circle)`. Since `(4_Circle)` basically contains two `(3_Line)` instances we can plug the solution of `(3_Line)` into one path connecting a and c of `(4_Circle)` and set the other path zero. We would expect that although this transport is feasible in that it fulfils the continuity equation it should result in an higher transportation

costs than the actual discrete Wasserstein distance for `(4_Circle)`. Doing the described confirms our expectations. The continuity equation is fulfilled for this artificial setup. However, the transportation cost using the geodesic of `(3_Line)` as one path is approximately 1.91279 where $\mathcal{W}(\rho_A, \rho_B) \approx 1.56855$ is the actual discrete Wasserstein distance.

Next we slightly modify `(4_Circle)` to check another theoretical condition with an actual numerical result.

Instance 4 (`4_Circle_Diag`)**.**



This instance is similar to `(4_Circle)` but additionally we add a diagonal which connects b with d. Now because we still want to have the uniform distribution as π we also add loops at the nodes a and c. The transition matrix is given by

$$Q = \frac{1}{3}\begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

with stationary distribution $\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. Again we want to transport all mass from a to c.

The thing that interests us the most about this instance is the question if there is any flow over the edge connecting b and d. We are expecting that there is no flow over this particular edge because the mass would be taking a very indirect road and it should lead to an higher overall distance in the Wasserstein metric. Now because we are looking for geodesics (the shortest connections with respect to the Wasserstein metric) we would expect no flow over the $b \to d$ edge. Looking at the calculated geodesic from our algorithm this assumption seems correct. For all timesteps $t \in \{0, \ldots, N_T\}$ we have $m_{b,d}(t) \approx 10^{-9}$ which is essentially zero taking the machine precision into account.

Another thing we can do is use the results for `(4_Circle_Diag)` to learn about the influence the transition probabilities have on the shape of the geodesics. In Figure 7.5 we compare the geodesic of `(4_Circle)` to the one of `(4_Circle_Diag)`. We can observe quite a large difference, especially in the nodes b and d. Now of course `(4_Circle)` and `(4_Circle_Diag)` are fundamentally different in that there is an additional diagonal in `(4_Circle_Diag)`. But we have already seen that the additional edge is not used for any transportation so it should make no difference if the edge is there or not. The remaining difference between `(4_Circle)` and `(4_Circle_Diag)` however are the different transition probabilities for example from a to b where in `(4_Circle)` it is $\frac{1}{2}$ and in `(4_Circle_Diag)` is is $\frac{1}{3}$.

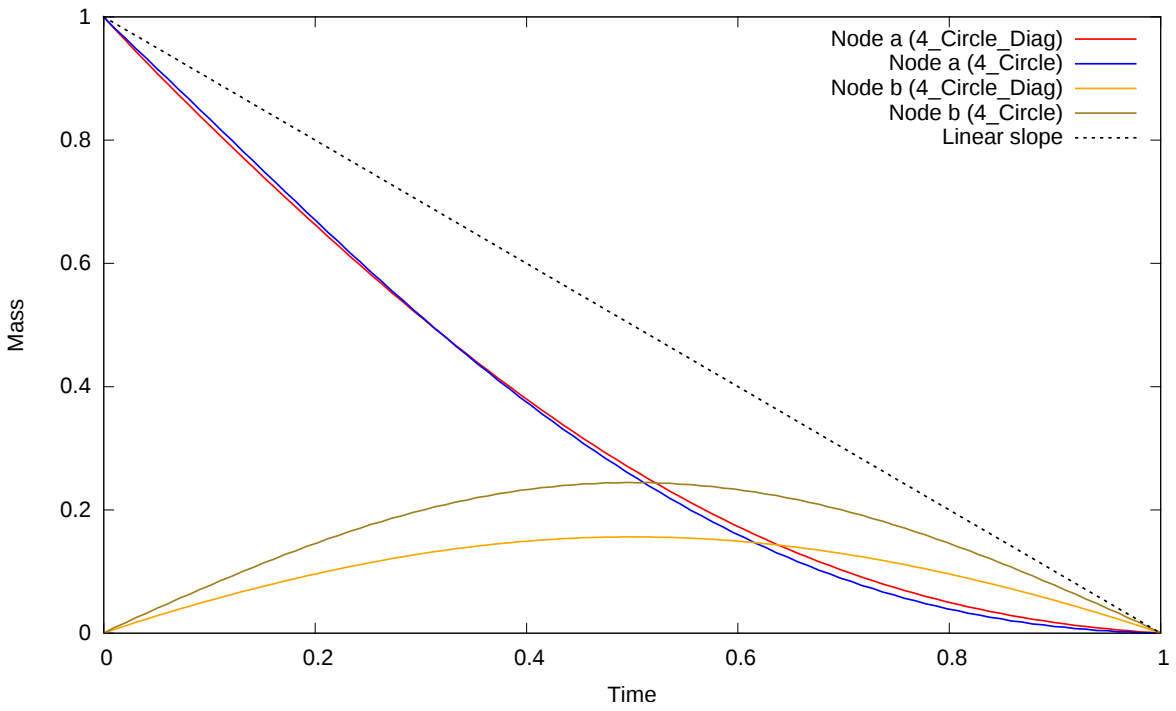
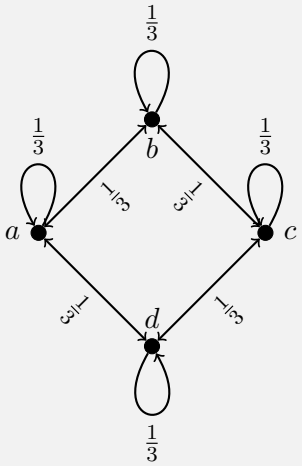


Figure 7.5: Comparison between the geodesics of `(4_Circle)` and `(4_Circle_Diag)`.
— *Instance:* `(4_Circle)` and `(4_Circle_Diag)`, *Algorithm:* `(DR)` —

In order to test the hypothesis that the different transition probabilities cause the differences we introduce a new instance.

Instance 5 (`4_Circle_Loops`)**.**



This instance is similar to `(4_Circle)` but each edge has a transition probability of $\frac{1}{3}$. The transition matrix is given by

$$Q = \frac{1}{3}\begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

with stationary distribution $\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ and all mass transported from a to c.

Now comparing the geodesics of `(4_Circle_Diag)` and `(4_Circle_Loops)` we do indeed observe that they are identical. Finally to get a really intuitive understanding of these geodesics we provide alternative visualization of them in Figure 7.6. In this visualization bigger points represent nodes with more mass, smaller points nodes with less mass. The same holds true for the edge and the current flow over them. However, note that there is a lower limit of how small a point or line can be so that nodes or edges with 0 mass / flow are still displayed.

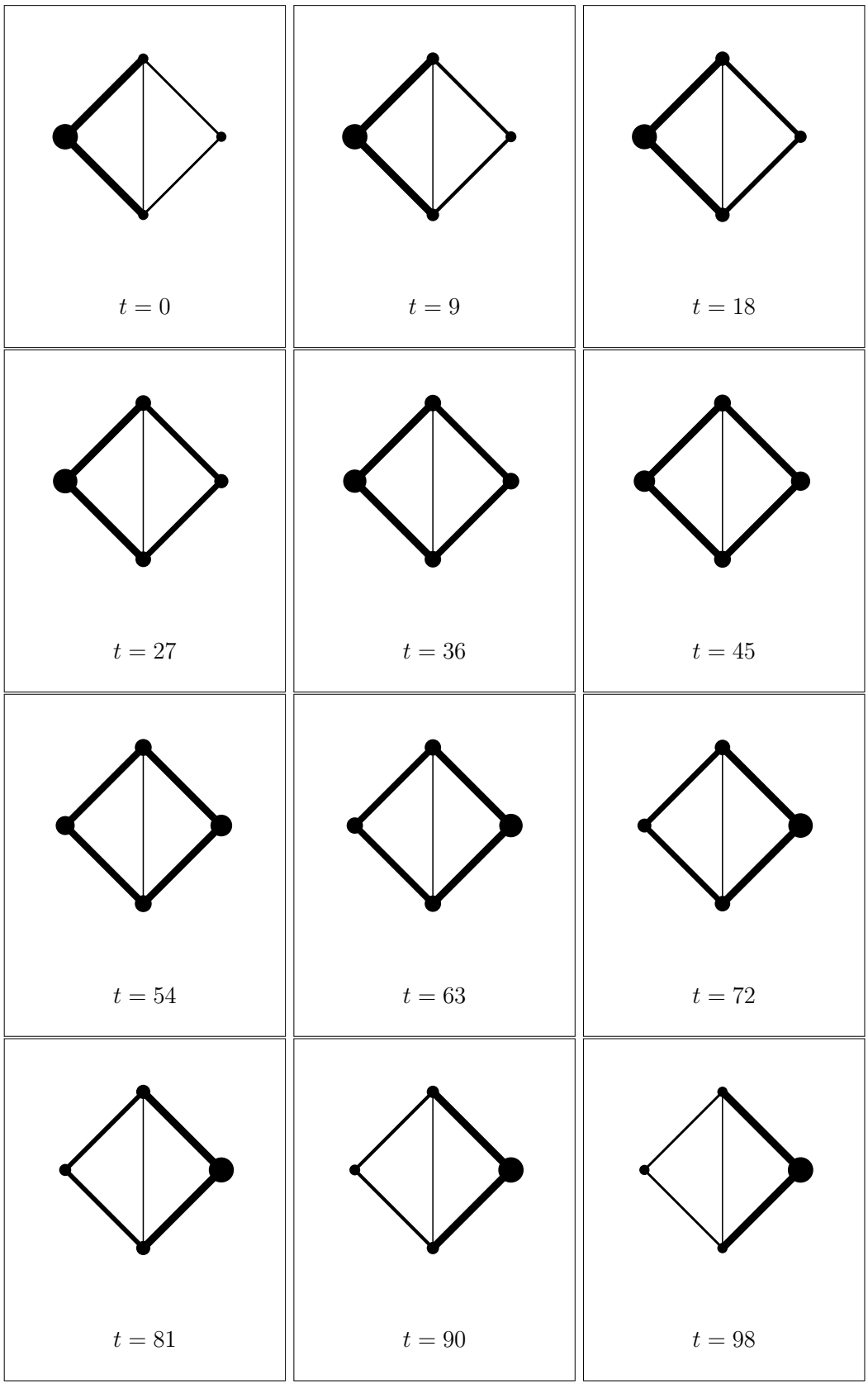
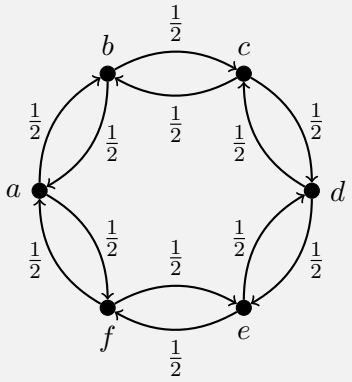


Figure 7.6: Visualization of the geodesic of `(4_Circle_Diag)`. The bigger a node the more mass it has. The same holds true for the flow on the edges.

7.3 Bigger circles, lines and grids

In this next section we are going to look into bigger instances with more than four nodes. They might already give us an understanding of some limiting behaviour for really large uniform instances. We start with the circles.

Instance 6 (`6_Circle`)**.**



This instance resembles a circle of 6 nodes where all mass is transported from a to d. The transition matrix is given by

$$Q = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

with stationary distribution $\pi = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$ and

$$\rho_A = (1,0,0,0,0,0), \quad \rho_B = (0,0,0,1,0,0).$$

In general such uniform circles with n nodes are labelled `(n_Circle)`.

The geodesic plot for `(6_Circle)` can be found in Figure 7.7, Figure 7.8 shows our graph visualization.

An observation we can make using the bigger circle instance is that it seems the more nodes there are between source and target node the faster the initial transport does go. One can see this by comparing the mass developments at node a in instances `(3_Circle)`, `(4_Circle)` and `(6_Circle)` with t close to zero. The more nodes the steeper the slope in the beginning. This also makes sense because when there are more nodes the mass can be distributed more evenly between them. In the discrete Wasserstein distance the transport cost depends also on how much there is already at the target node. So for more nodes the mass spreads more over the nodes thus the difference compared to the source stays higher and the transport happens faster.

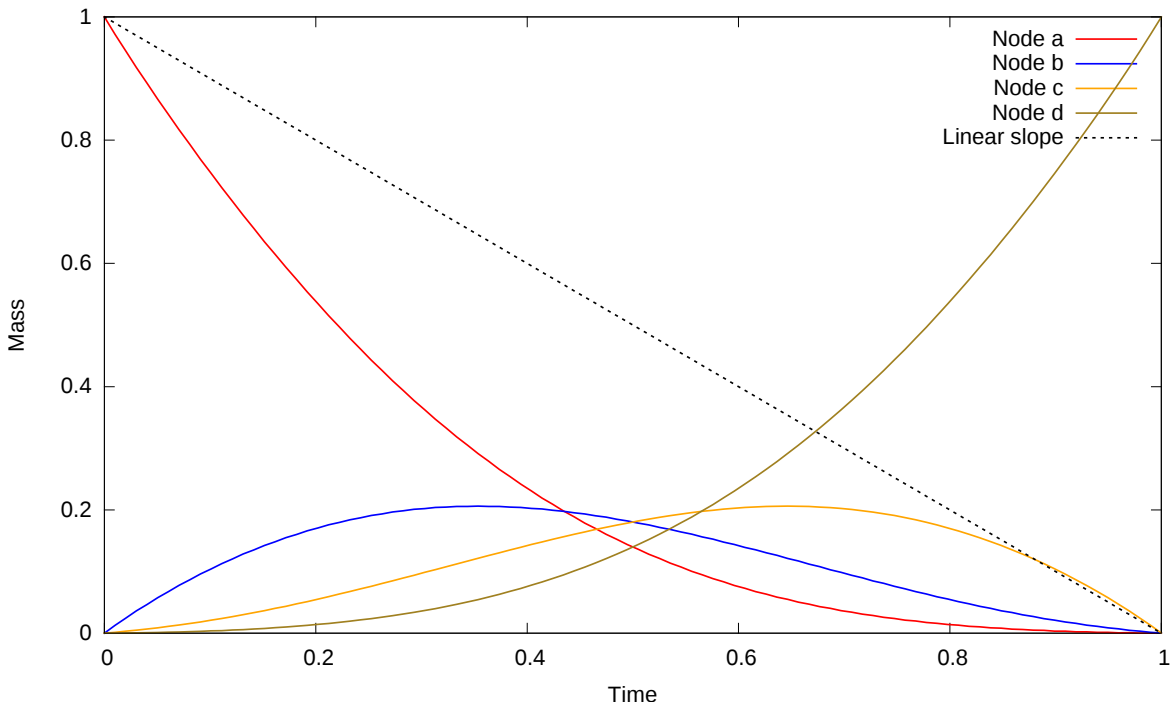
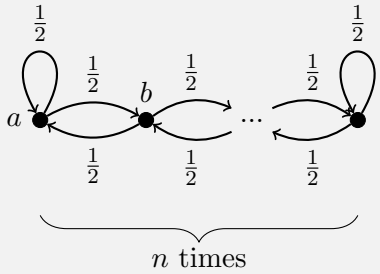


Figure 7.7: — *Instance:* `(6_Circle)`, *Algorithm:* `(DR)`, $\mathcal{W}(\rho_A, \rho_B) \approx 1.96517$ —

Next we introduce bigger line instances.

Instance 7 (`n_Line`)**.**



The `(n_Line)`-instances represent a line of n nodes where each node except the start and the end node are connected with their two neighbours. The transition matrix is given by

$$Q = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 \end{pmatrix}$$

with stationary distribution $\pi = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ and the transport from the start node to the end node is encoded in the boundary values

$$\rho_A = (1, 0, \ldots, 0), \quad \rho_B = (0, \ldots, 0, 1).$$

These line instances can be used to observe the behaviour of the discrete Wasserstein distance in the limit, i.e. for very large uniform graphs. It has been shown in [GM12] that the discrete Wasserstein distance for particular graphs converges to the continuous Wasserstein distance in a Gromov-Hausdorff sense. For our line instances this means that would expect behaviour similar to a Dirac measure that is transported with respect to the continuous Wasserstein distance along the real line or respectively the interval $[0, 1]$. However, the instances we were able to calculate efficiently for this thesis are rather small (around 7 to 9 nodes) so that the limiting behaviour can not really be observed yet. Nevertheless Figures 7.9, 7.10 and especially 7.11 show already

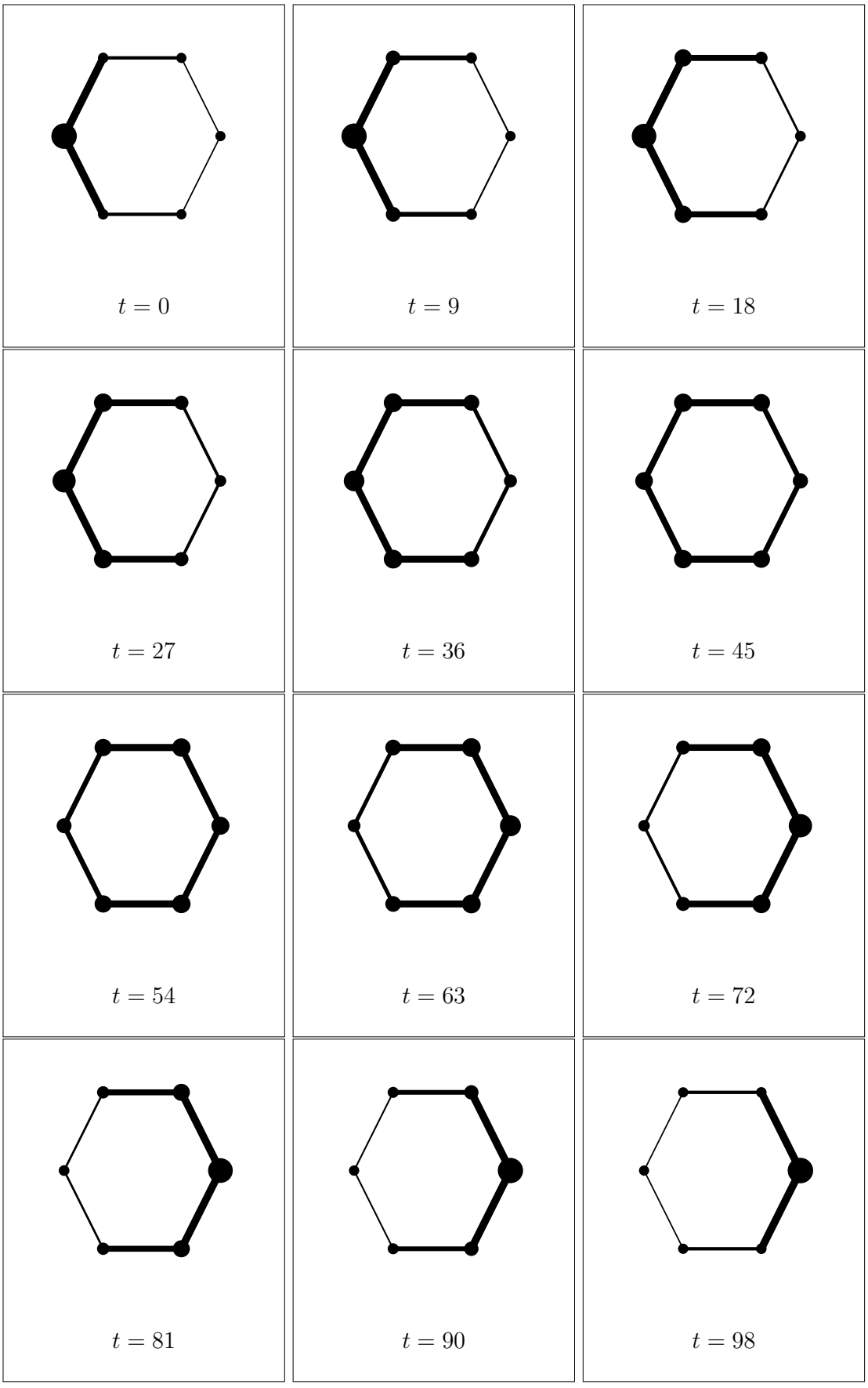
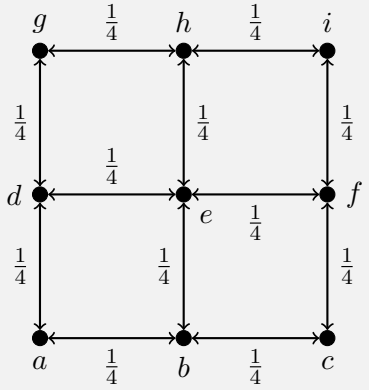


Figure 7.8: Visualization of the geodesic of `(6_Circle)`.

expected behaviour. As usual Figures 7.9 and 7.10 show the development of mass over time. We focus on the nodes a and b and observe that the more nodes there are in the line the faster the mass gets transported in the beginning. Also the more nodes become part of the line the earlier the mass in node b reaches its maximum. Figure 7.11 now shows a different type of visualization. Instead of displaying the mass over time for one node we show the whole mass distribution at all nodes for different timesteps. The behaviour we are observing seems correct as in the first timesteps most mass is concentrated around the start node and the later it gets the more mass at the end nodes. For the central timesteps the mass distributions have a bell shape that is moving along the line.

Instance 8 (`9_Grid`)**.**



Where the `(n_Line)`-instance can be seen as a discretization of a real interval $[0,1]$ the following instance is a very sparse approximation of $[0,1]^2$. The transition matrix is given by

$$Q = \frac{1}{4}\begin{pmatrix} 2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 2 \end{pmatrix}$$

and $\pi = \left(\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\right)$. We send all mass from a to i by setting

$$\rho_A = (1,0,0,0,0,0,0,0,0), \quad \rho_B = (0,0,0,0,0,0,0,0,1).$$

For simplicity we omitted the loops connecting a node with itself on the drawing.

For the geodesic plots of `(9_Grid)` we focus on the four nodes a, b, c and e because the other nodes have either the same or a symmetric slope to those nodes. We would expect node b to have its maximum of mass before $t = \frac{1}{2}$ because it is a node closer to node a. Nodes c, e and g should be symmetric with respect to $t = \frac{1}{2}$ because the instance is symmetric with respect to this diagonal axis of nodes. All this properties can be observed in Figure 7.12. What is a bit more surprising is the fact that the slopes in nodes c, e and g are identical. Because on the one hand all these nodes can be reached over two edges with weight $\frac{1}{4}$ but on the other hand the node e in the middle has two such ways connecting it with a while c and g only have one. Thus one might expect more flow over node e than over the two others. However, one has to keep in mind that e is also connected by two paths to i while the others have only one such connection. So there might flow more mass into node e but also by the same amount it might leave e. Because we are always plotting the mass on some node at some time we do not consider how much has already flown over this node. Node e certainly has more flow over it than c and g. Also the big picture of whats happening on `(9_Grid)` is probably best visualized in the graph visualization in Figure 7.13.

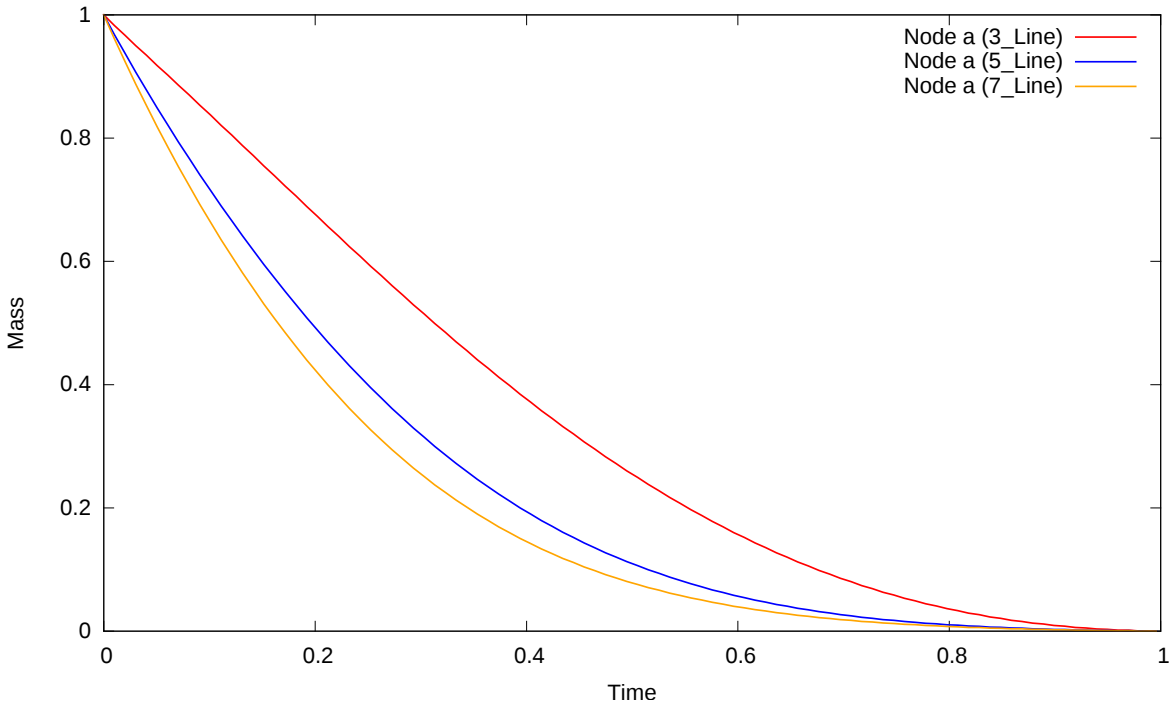


Figure 7.9: Comparison between the geodesics of different `(n_Line)` instances. We compare the mass in node a.
— *Instance:* `(n_Line)` with $n \in \{3, 5, 7\}$, *Algorithm:* `(DR)` —

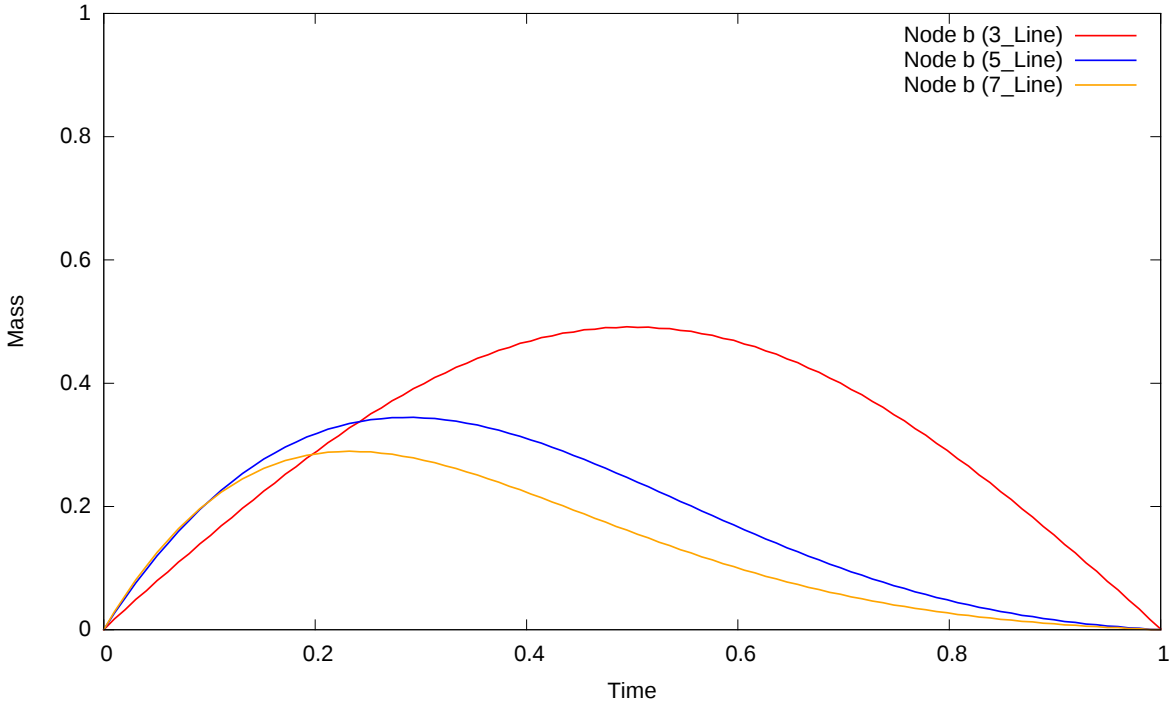


Figure 7.10: Comparison between the geodesics of different `(n_Line)` instances. We compare the mass in node b.
— *Instance:* `(n_Line)` with $n \in \{3, 5, 7\}$, *Algorithm:* `(DR)` —

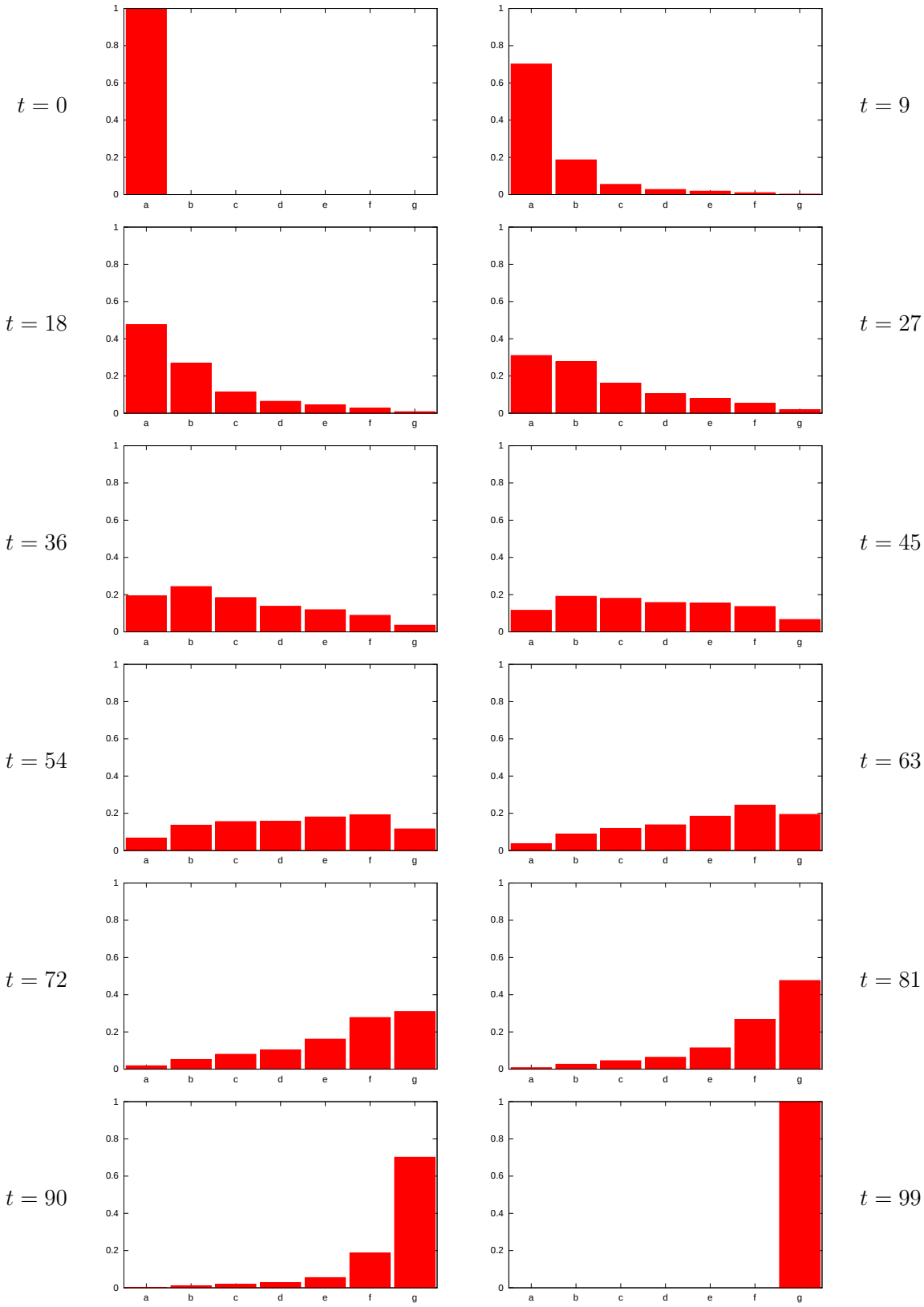


Figure 7.11: Visualization of the geodesic of `(7_Line)`. We display the whole mass configuration over time.

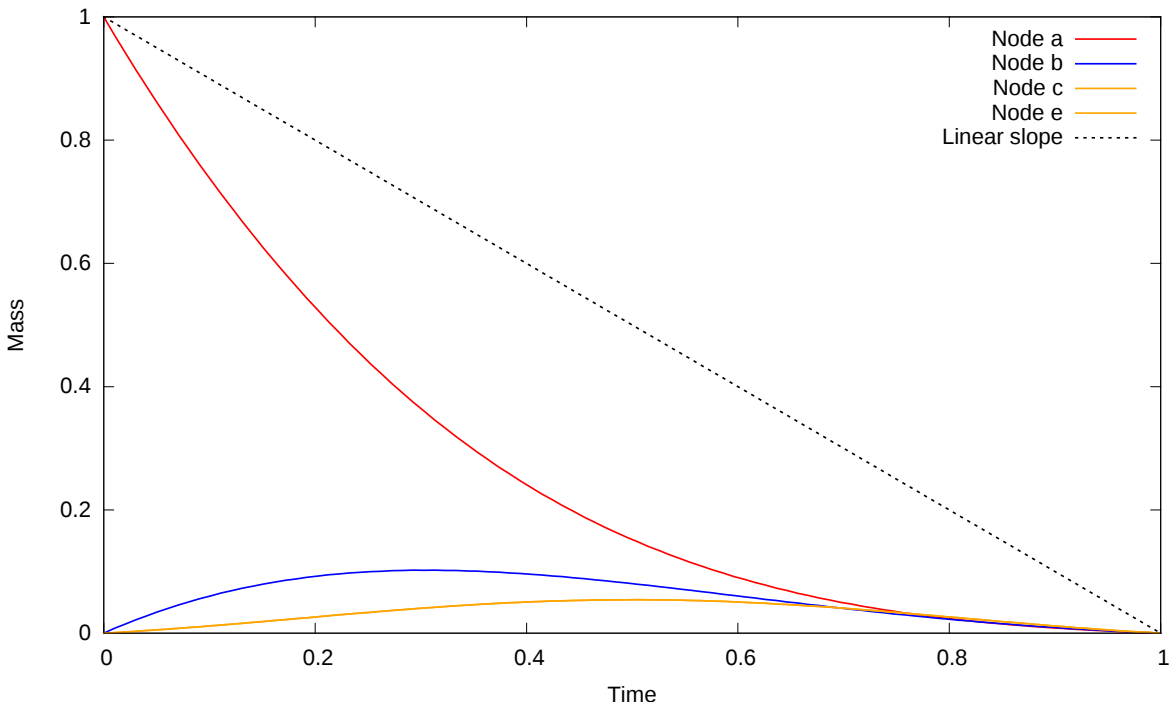
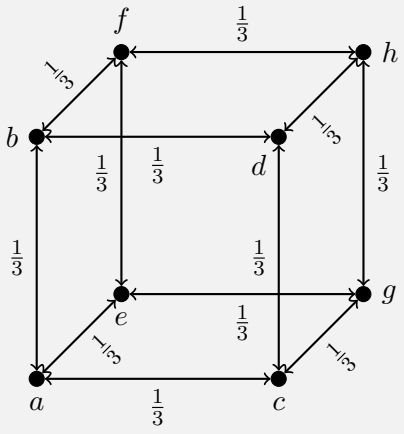


Figure 7.12: The geodesic of `(9_Grid)`. Note that node c and e have an identical slope because their distance from a is the same.
— *Instance:* `(9_Grid)`, *Algorithm:* `(DR)`, $\mathcal{W}(\rho_A, \rho_B) \approx 2.00994$ —

As a final instance we look at a cube in 3d.

Instance 9 (`8_3DCube`)**.**



This instance represents a cube in 3d. So each node is uniformly connected with three neighbours and we get as transition matrix

$$Q = \frac{1}{3}\begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

and as stationary distribution $\pi = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$. Again we send all mass from a to h.

Because of the symmetry of `(8_3dCube)` we focus on the nodes a, b and d. However, for the other nodes one can observe that for example c behaves the same as b and d the same as e and so on. The mass development for the geodesic can be found in Figure 7.14 and a graph visualization in Figure 7.15.

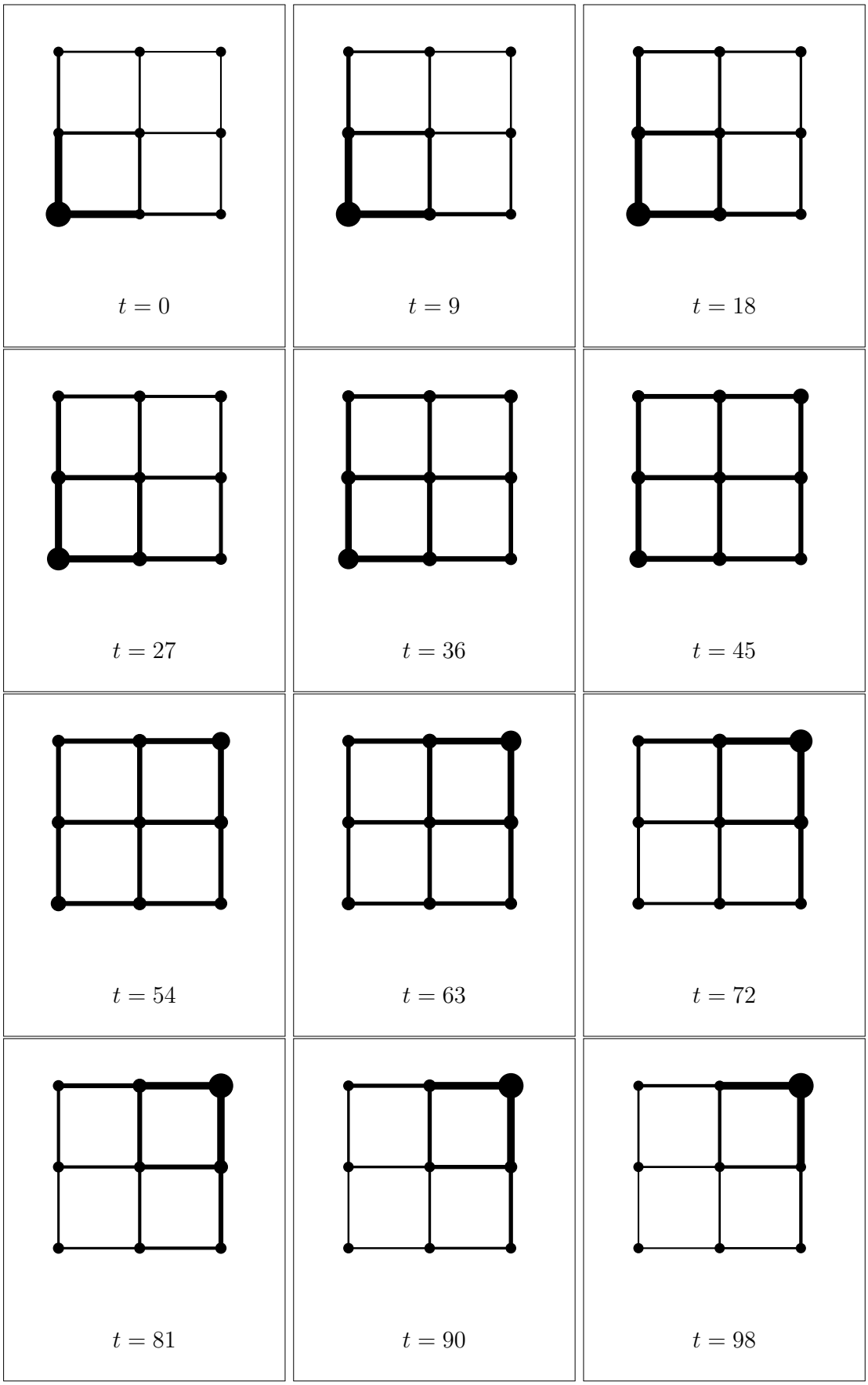


Figure 7.13: Visualization of the geodesic of `(9_Grid)`.

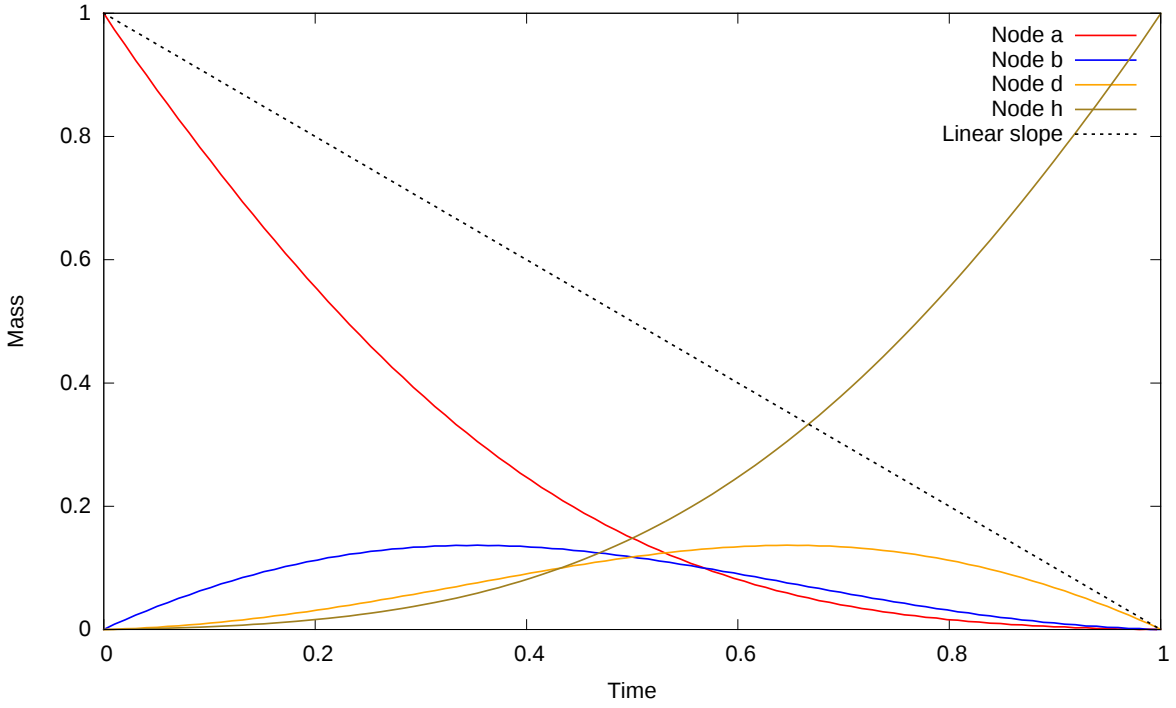


Figure 7.14: The geodesic of `(8_3dCube)`.
— *Instance:* `(8_3dCube)`, *Algorithm:* `(DR)`, $\mathcal{W}(\rho_A, \rho_B) \approx 1.67347$ —

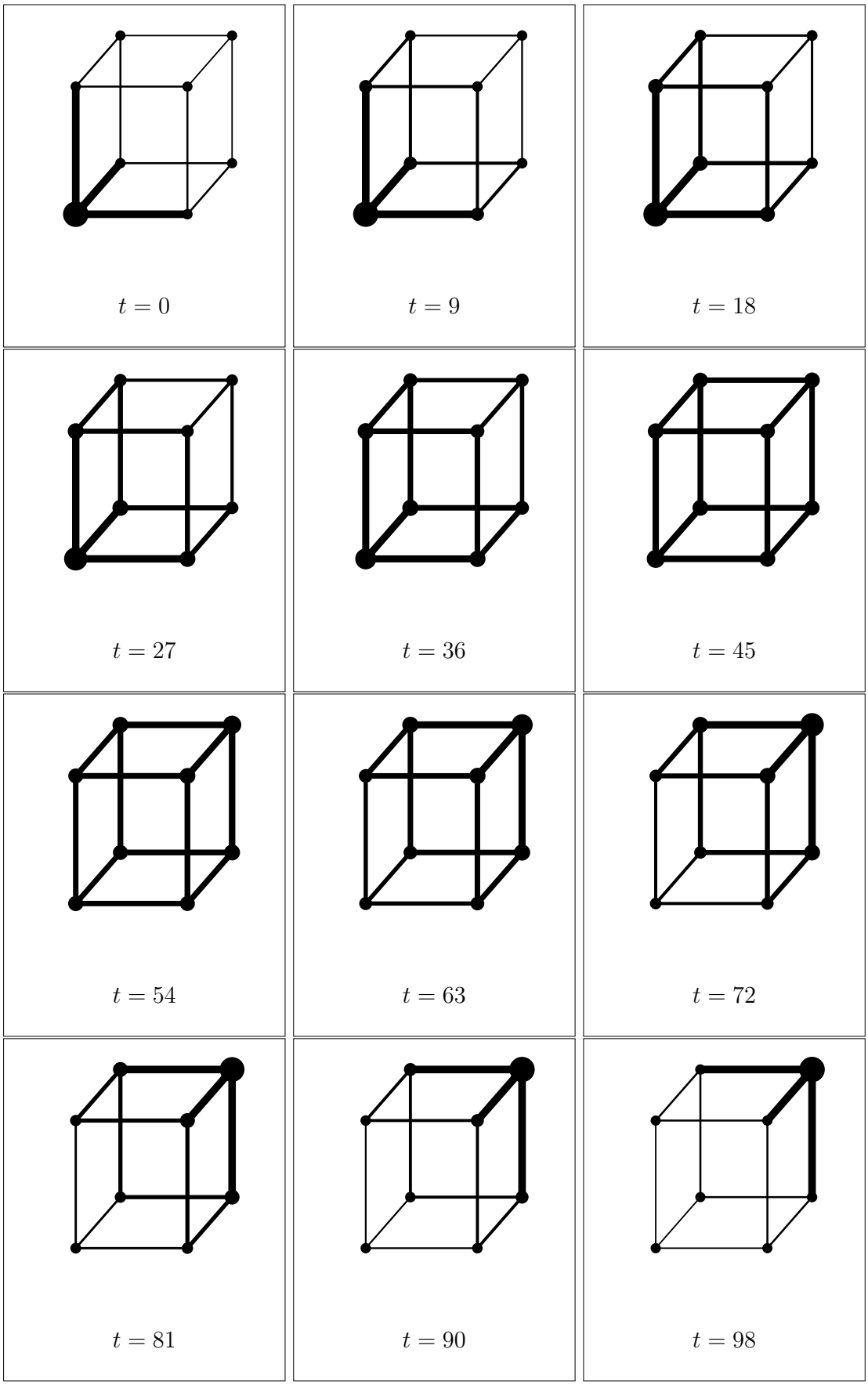


Figure 7.15: Visualization of the geodesic of `(8_3DCube)`.

8 Conclusions and Further Research

In summary we were able to derive two closely related numerical schemes for computing the geodesics of the discrete Wasserstein metric. Additionally we introduced a discretization of space and time that made the needed numerical approximations of the steps of our schemes possible. We have also implemented the algorithms in `C++` and are now left with working versions of `(BB)` and `(DR)`. As a result we were able to present results for smaller problem instances that give nevertheless ideas about the geometry of the metric.

Beside these advances there is still room for improvement. Our algorithms work fine for fairly small instances with under 10 nodes. However, the current implementations are not practical for example for the study of very large graphs. This is something that should probably be improved in further work on the topic. Because there don't seem to be any theoretical limitations apart from the pure size of instances the author is optimistic that there might in the future be further advances in the ability to compute the geodesic for larger instances.

However, to reach this goal one has to solve the stability problems with the projection on $\mathcal{K}$. To this end one might consider two different strategies in the future. On the one hand one might look at better ways to extend the logarithmic mean than the one we proposed in Chapter 5. For example the following approach might be more natural with respect to inherent properties of the logarithmic mean.

First note that the logarithmic mean is just a linear slope along lines $\{(s,t) \in \mathbb{R}^2_+ : \frac{s}{t} = \text{ const }\}$ that run through the origin $(0,0)$. This motivates us to transform the logarithmic mean into polar coordinates. We get

$$\theta_p(r,\varphi) = rh(\varphi) \quad \text{where} \quad h(\varphi) = \frac{\cos\varphi - \sin\varphi}{\log\left(\frac{\cos\varphi}{\sin\varphi}\right)}.$$

To get a better understanding of the angle dependence Figure 8.1 displays a plot of h.
We can clearly see how h is only defined on $(0, \frac{\pi}{2})$ and $(\pi, \frac{3\pi}{2})$ which represent quadrants one and three of the Cartesian coordinate system. In order to extend the logarithmic mean we have to extend h to the whole $[0, 2\pi]$. To this end we could cut the original h at ε distance around the singularities $0, \frac{\pi}{2}, \pi$ and $\frac{3\pi}{2}$ and then connect the parts with each other (for example with a linear slope or just zeroes). Finally in order to get C^2 regularity we convolute with a regular enough kernel. The result is a C^2 function that resembles the properties of h but is defined on $[0, 2\pi]$. Using this function we can define the extended logarithmic mean.

Our hope is that such an extension might be closer to the natural properties of the logarithmic mean and as a result we might get better convergence for the projection.

The other idea one might consider is to think of the projection problem more in graph theoretical than in a pure numerical context. There are for example some similarities between the projection problem and some flow problems on graphs. To this end recall that the most useful characterization of the convex set $\mathcal{K}$ was given as follows. A pair (ρ, m) is in $\mathcal{K}$ if there exists

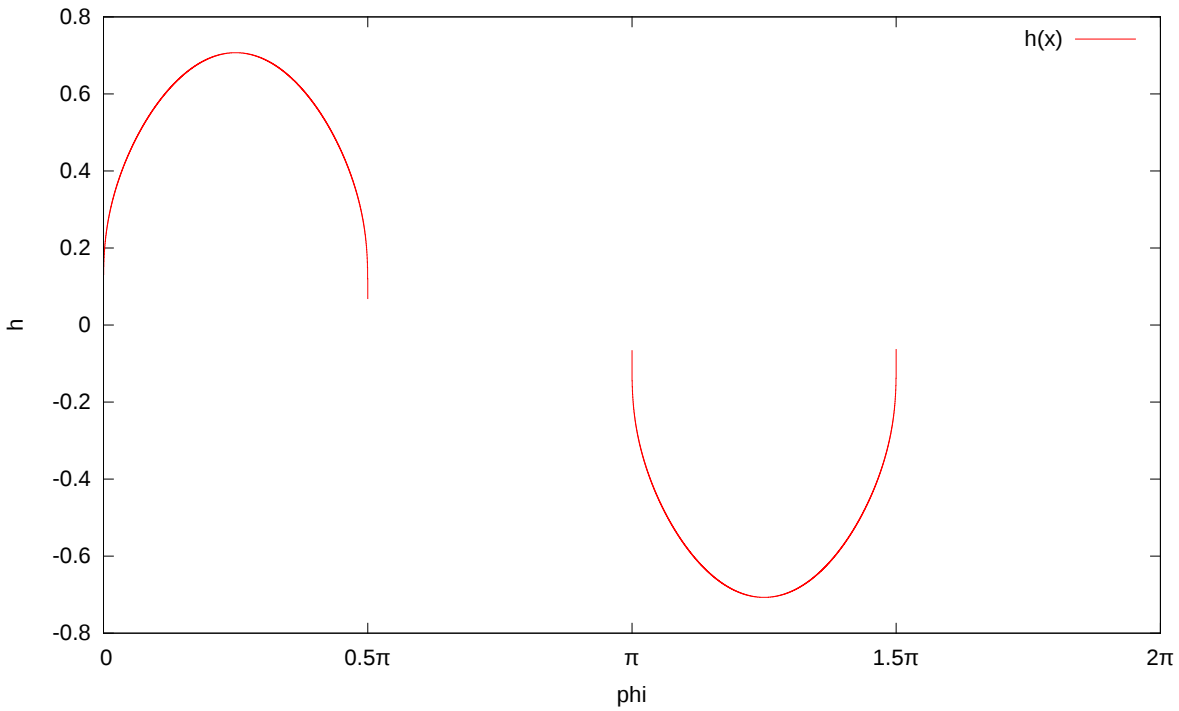


Figure 8.1: The function $h(\varphi)$ on the interval $[0, 2\pi]$. One can clearly see the singularities at $0, \frac{\pi}{2}, \pi$ and $\frac{3\pi}{2}$.

$a \in \mathbb{R}_+^{\mathcal{X}}$ such that for all $x \in \mathcal{X}$

$$\rho_x + \frac{1}{4} \sum_{y \in \mathcal{X}} \partial_1 \theta(a_x, a_y) |m_{x,y}|^2 Q(x,y) \leq 0.$$

Now we use the fact from Corollary 3.13 that $\partial_1 \theta(a_x, a_y)$ depends only on the ratio $\frac{a_x}{a_y}$. We denote $f(\frac{a_x}{a_y}) := \partial_1 \theta(a_x, a_y)$ and also $w_{x,y} = \frac{1}{4}|m_{x,y}|^2 Q(x,y)$. Note that f is actually a strictly monotone function. In total the conditions now looks like

$$\rho_x + \sum_{y \in \mathcal{X}} f\left(\frac{a_x}{a_y}\right) w_{x,y} \leq 0.$$

The next thing we can do is a change of variable where we replace $\frac{a_x}{a_y}$ by a variable $v_{x,y}$ defined on the edges instead of nodes. As a result we get

$$\rho_x + \sum_{y \in \mathcal{X}} f(v_{x,y}) w_{x,y} \leq 0.$$

However, we will have to add an additional constraint. To this end suppose that $v_{x,y}, v_{y,z}, v_{z,x}$ form an undirected circle in the graph. Then

$$v_{x,y} v_{y,z} v_{z,x} = \frac{a_x}{a_y} \frac{a_y}{a_z} \frac{a_z}{a_x} = 1$$

Thus for all circles C in the graph the edges $(v_{x,y})_{(x,y) \in C}$ have to satisfy the condition that $\Pi_{(x,y) \in C} v_{x,y} = 1$ or at least $\Pi_{(x,y) \in C} v_{x,y} \geq 1$ (because then we can choose smaller a_x that satisfy $= 1$). As a conclusion we have derived a different characterization of the convex set that is more from a graph theoretical point of view. This characterization might be beneficial to use in the projection algorithm. Maybe only for some instances that have for example no circles.

Beside the focus on improving the projection on $\mathcal{K}$ one might look into other computations that can be done on Riemannian manifold including

- Computing exponential or logarithmic maps using geodesics,
- Computing Riemannian barycenters,
- Computing geodesic splines.

One might also look more into the numerical properties of the algorithms presented in this thesis and prove convergence or results related to the approximation quality of the algorithms.

Bibliography

[ABM06] Attouch, H., Buttazzo, G., and Michaille, G. *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2006 (cit. on p. 40).

[AGS05] Ambrosio, L., Gigli, N., and Savaré, G. *Gradient Flows: In Metric Spaces And In The Space Of Probability Measures*. Lectures in math. Birkhäuser, 2005 (cit. on pp. 5, 8, 9).

[ASZ07] Ambrosio, Luigi, Savare, Giuseppe, and Zambotti, Lorenzo. *Existence and stability for Fokker-Planck equations with log-concave reference measure*. 2007 (cit. on p. 3).

[BB00] Benamou, Jean-David and Brenier, Yann. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". In: *Numer. Math.* 84 (2000), pp. 375–393 (cit. on pp. 3, 8, 25–27, 37, 47).

[BL15] Brauer, Christoph and Lorenz, Dirk. "Cartoon-Texture-Noise Decomposition with Transport Norms". In: *Scale Space and Variational Methods in Computer Vision - 5th International Conference, SSVM 2015, Lège-Cap Ferret, France, May 31 - June 4, 2015, Proceedings*. 2015, pp. 142–153 (cit. on p. 3).

[CHLZ12] Chow, Shui-Nee, Huang, Wen, Li, Yao, and Zhou, Haomin. "Fokker–Planck Equations for a Free Energy Functional or Markov Process on a Graph". English. In: *Archive for Rational Mechanics and Analysis* 203.3 (2012), pp. 969–1008 (cit. on pp. 3, 13).

[CP09] Combettes, P. L. and Pesquet, J.-C. "Proximal Splitting Methods in Signal Processing". In: *ArXiv e-prints* (Dec. 2009). arXiv: `0912.3522 [math.OC]` (cit. on p. 32).

[EB92] Eckstein, Jonathan and Bertsekas, Dimitri P. "On the Douglas-Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators". In: *Math. Program.* 55.3 (June 1992), pp. 293–318 (cit. on pp. 31, 32).

[EM] Erbar, Matthias and Maas, Jan. "Dual formulation for the discrete transport distance". English. In: () (cit. on p. 19).

[EM12] Erbar, Matthias and Maas, Jan. "Ricci Curvature of Finite Markov Chains via Convexity of the Entropy". English. In: *Archive for Rational Mechanics and Analysis* 206.3 (2012), pp. 997–1038 (cit. on p. 17).

[Erb10] Erbar, Matthias. "The heat equation on manifolds as a gradient flow in the Wasserstein space". In: *Ann. Inst. H. Poincaré Probab. Statist.* 46.1 (Feb. 2010), pp. 1–23 (cit. on p. 3).

[Gig10] Gigli, Nicola. "On the heat flow on metric measure spaces: existence, uniqueness and stability". English. In: *Calculus of Variations and Partial Differential Equations* 39.1-2 (2010), pp. 101–120 (cit. on p. 3).

[GM12] Gigli, N. and Maas, J. "Gromov-Hausdorff convergence of discrete transportation metrics". In: *ArXiv e-prints* (July 2012). arXiv: `1207.6501 [math.MG]` (cit. on pp. 18, 61).

[JKO98] Jordan, Richard, Kinderlehrer, David, and Otto, Felix. "The Variational Formulation of the Fokker-Planck Equation". In: *SIAM J. Math. Anal.* 29.1 (Jan. 1998), pp. 1–17 (cit. on pp. 3, 10).

[Kan42] Kantorovitch, Leonid Vital'evich. "On the translocation of masses." In: *Dokl. Akad. Nauk. USSR* 37 (1942), pp. 227–229 (cit. on pp. 3, 6).

[Kan48] Kantorovitch, Leonid Vital'evich. "On a problem of monge." In: *Uspekhi Mat. Nauk* 3 (1948), pp. 225–226 (cit. on pp. 3, 6).

[Maa11] Maas, Jan. "Gradient flows of the entropy for finite Markov chains". In: *Journal of Functional Analysis* 261.8 (2011), pp. 2250–2292 (cit. on pp. 3, 13, 16, 17, 19, 51–53).

[Mie11] Mielke, Alexander. "A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems". In: *Nonlinearity* 24.4 (2011), p. 1329 (cit. on pp. 3, 13).

[Mon81] Monge, Gaspard. "Mémoire sur la théorie des déblais et des remblais." In: *De l'Imprimerie* (1781) (cit. on pp. 3, 5).

[OS08] Ohta, Shin-ichi and Sturm, Karl-Theodor. "Heat flow on Finsler manifolds". In: *COMM. PURE APPL. MATH* (2008), pp. 1386–1433 (cit. on p. 3).

[PB14] Parikh, Neal and Boyd, Stephen. "Proximal Algorithms". In: *Found. Trends Optim.* 1.3 (Jan. 2014), pp. 127–239 (cit. on pp. 32, 35).

[Pow70] Powell, M. J. D. "A Hybrid Method for Nonlinear Equations". In: *Nonlinear Algebraic Equations* (1970) (cit. on p. 49).

[PPO14] Papadakis, N., Peyré, G., and Oudet, E. "Optimal Transport with Proximal Splitting". In: *SIAM Journal on Imaging Sciences* 7.1 (2014), pp. 212–238 (cit. on pp. 3, 25).

[RPC10] Rabin, Julien, Peyré, Gabriel, and Cohen, Laurent D. "Geodesic shape retrieval via optimal mass transport". In: *Computer Vision–ECCV 2010*. Springer Berlin Heidelberg, 2010, pp. 771–784 (cit. on p. 3).

[RPDB12] Rabin, Julien, Peyré, Gabriel, Delon, Julie, and Bernot, Marc. "Wasserstein barycenter and its application to texture mixing". In: *Scale Space and Variational Methods in Computer Vision*. Springer Berlin Heidelberg, 2012, pp. 435–446 (cit. on p. 3).

[SW05] Schaback, Robert and Wendland, Holger. *Numerische Mathematik*. Springer, 2005 (cit. on p. 49).

[Vil03] Villani, C. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003 (cit. on p. 5).

[Vil09] Villani, Cédric. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Berlin: Springer, 2009 (cit. on p. 5).

[Was69] Wasserstein, L.N. "Markov processes over denumerable products of spaces describing large systems of automata". In: *Probl. Inform. Transmission* 5 (1969), pp. 47–52 (cit. on p. 3).