universität wien

DISSERTATION | DOCTORAL THESIS

Titel | Title

Optimal transport and financial applications

verfasst von | submitted by

Dipl.-Ing. Lorenz Riess BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 796 605 405
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Mathematik
Betreut von | Supervisor:	Assoz. Prof. Dr. Julio Daniel Backhoff-Veraguas Univ.-Prof. Dipl.-Ing. Dr.techn. Mathias Beiglböck

Abstract

Optimal transport stands out as a mathematical field due to its rich interplay between theory and applications. Questions arising from practice – particularly in areas such as finance, economics, and data science – lead to new variants of the classical optimal transport framework. These theoretical extensions, in turn, enable further applications. This thesis is devoted to this exchange between theory and applications, with a particular focus on the use of optimal transport tools in financial contexts.

The first chapter of this thesis introduces a novel application of optimal transport in *financial regulation*. Motivated by the challenges faced by regulators of the financial industry in analyzing large amounts of credit data, we represent the credit portfolios of financial institutions as probability measures. Using generalized Wasserstein barycenters, we develop a clustering and reconstruction method for probability measures defined on different subspaces. This enables the construction of a metric structure over financial institutions, resulting in a *banking landscape* that can be used to identify clusters and detect outliers in the financial system.

The second chapter is devoted to the theory of weak optimal transport, an extension of classical optimal transport that allows for non-linear cost functions. We prove a *fundamental theorem of weak optimal transport*; in particular, we establish duality and a complementary slackness condition. This results in a wide range of applications, allowing us to recover the convex Kantorovich–Rubinstein formula, the Brenier–Strassen theorem, the structure theorem of entropic optimal transport, and to obtain new results concerning relaxed martingale optimal transport, as well as interpolations between Brenier–Strassen and martingale Benamou–Brenier.

The third chapter of this thesis connects weak martingale optimal transport to local volatility modeling in mathematical finance. By extending the classical change of numeraire transformation to weak martingale optimal transport, we relate Bass martingales to a new class of processes, namely *geometric Bass martingales*. This yields a local volatility model that is as close as possible to the Black–Scholes model, the benchmark model in the financial industry, while being perfectly calibrated to observed market prices.

Overall, this thesis illustrates how advances in optimal transport can be both applied to, and motivated by, practical problems in the financial sector, while simultaneously contributing to the theoretical development of the field.

Zusammenfassung

Optimaler Transport zeichnet sich als mathematisches Forschungsfeld durch ein besonders intensives Wechselspiel zwischen Theorie und Anwendung aus. Fragestellungen aus der Praxis – insbesondere aus Bereichen wie Finanzwesen, Ökonomie und Data Science – führen zu neuen Varianten des klassischen Transportproblems, während diese theoretischen Verallgemeinerungen ihrerseits neue Anwendungsfelder eröffnen. Diese Arbeit widmet sich diesem gegenseitigen Austausch mit einem besonderen Fokus auf der Anwendung von Methoden der optimalen Transporttheorie in der Finanzmathematik.

Das erste Kapitel dieser Arbeit stellt eine neuartige Anwendung optimalen Transports in der *Finanzmarktregulierung* vor. Motiviert durch Herausforderungen, vor denen Aufsichtsbehörden bei der Analyse großer Mengen an Kreditdaten stehen, interpretieren wir Kreditportfolios von Banken als Wahrscheinlichkeitsmaße. Mithilfe verallgemeinerter Wasserstein–Baryzentren entwickeln wir ein Clustering- und Rekonstruktionsverfahren für Wahrscheinlichkeitsmaße, die auf unterschiedlichen Teilräumen definiert sind. Dies ermöglicht die Definition eines Abstandsbegriffs zwischen Finanzinstituten und führt zu einer *Bankenlandschaft*, welche zur Identifikation von Clustern und Ausreißern im Finanzsystem genutzt werden kann.

Das zweite Kapitel widmet sich der Theorie des nichtlinearen optimalen Transports, einer Verallgemeinerung des klassischen optimalen Transportproblems, welche nichtlineare Kostenfunktionen erlaubt. Wir beweisen einen *fundamentalen Satz des nichtlinearen optimalen Transports*; insbesondere zeigen wir Dualität und eine Komplementaritätsbedingung. Dieses Resultat ermöglicht eine Vielzahl von Anwendungen: die konvexe Kantorovich–Rubinstein–Formel, der Satz von Brenier–Strassen und der Struktursatz des entropischen optimalen Transports folgen direkt. Zudem ergeben sich neue Resultate zu einem relaxierten Martingaltransportproblem sowie zu einer Interpolation zwischen Brenier–Strassen und Martingal–Benamou–Brenier.

Das dritte Kapitel dieser Arbeit verbindet nichtlinearen Martingaltransport mit lokalen Volatilitätsmodellen aus der Finanzmathematik. Eine Verallgemeinerung der *Numéraire-wechsel Transformation*, einer klassischen Methode der Finanzmathematik, auf das nichtlineare Martingaltransportproblem ermöglicht es, eine Verbindung zwischen Bass–Martingalen und einer neuen Klasse von Prozessen herzustellen: den *geometrischen Bass–Martingalen*. Dies führt zu einem lokalen Volatilitätsmodell, das dem in der Finanzpraxis etablierten Black–Scholes–Modell möglichst nahekommt, dabei jedoch eine exakte Kalibrierung an beobachtete Marktpreise erlaubt.

Insgesamt zeigt diese Arbeit, wie Fortschritte im Bereich des optimalen Transports sowohl durch praktische Fragestellungen motiviert sein können als auch zu deren Lösung beitragen und leistet darüber hinaus einen Beitrag zur theoretischen Weiterentwicklung des Gebiets.

Acknowledgements

This dissertation marks a major milestone in my life, so I want to thank the many people who enabled this journey and shared it with me.

I want to start by thanking my supervisors Mathias and Julio from the bottom of my heart. Mathias, I have idolized you since the first measure theory course you gave during my bachelor's – you taught me so much, both mathematically and on a personal level. You were always there for me, not just for academic, but also for personal problems. Julio, the same goes for you. With your calm attitude and by looking at problems from angles I could never think of, you always managed to calm me down whenever I was stressed again. Hopefully, I have picked up at least a bit of this huge skill of yours.

Next, I want to thank all my colleagues from university, whom I hope I can call friends, during my time in this research group – in particular Stefan, Gudi, Anne, Dani, Daniel, Xin, Ben, and Astrid.

Thanks to Johannes and Andi for enabling my first PhD project and for guiding me through my start at OeNB – I really look up to both of you.

I gratefully acknowledge the financial support from the Austrian Science Fund (FWF) under grants P35197, P36835-N, and Y0782, and from the Oesterreichische Nationalbank (OeNB) through project EATE II.

Thanks to Stefan, Roman and Luki. We met during the first weeks of our bachelors and you guys stuck with me until the very end – especially you, Stefan. My studies and this thesis would never have been possible without you guys. Thanks, Luki and Roman, for letting me share any PhD specific problems with you – talking to the two of you helped me immensely.

A major part in finishing my studies is due to my best friends. You guys were always there for me and helped me stay sane during these last years. Special thanks to Ralph for looking after me and letting me share literally anything with you, to Nico for always being there when something was burning, to Stefan for showing me that there is so much more to life than mathematics and to Nico for sharing an apartment with me for four years. Thank you Philipp, Beppe, Peter, Pascal, Jonah, Yvonne, Max, Naz, David, Nathalie, and Sebastian.

Finally, I want to thank my family, with whom I am truly blessed. Thanks, Martina, for always being there for me if needed – no matter when and where. Thanks, Jonas and Magdi, for letting me be the silly older brother – spending time with you makes me forget any problems and fuels me with so much energy. Thanks, Karo, I think there is no one in life who understands me as much as you do and knows how I feel by just looking at me. Thanks, Mum and Dad, for fostering my love for mathematics from the very beginning when I was still in kindergarten, and most importantly, for your unconditional love and (financial) support.

Contents

Introduction

Optimal transport has rapidly evolved over the past two decades and now plays a significant role not only in pure mathematics but also in a wide range of real-world applications, including finance, economics, and various business domains. During this evolution there has been an ongoing interplay between the mathematical theory and its applications. Problems arising in financial and economic contexts often inspire new directions in optimal transport, while these new directions, in turn, allow for new applications. In particular, this interplay has motivated several extensions of the classical optimal transport framework. For instance, entropy regularized optimal transport gained prominance in statistics and in the machine learning world [18, 24, 25, 26, 30, 42, 43], martingale optimal transport allows to derive model-independent bounds of option prices and perform robust hedging [16, 17, 29, 35], adapted optimal transport offers new tools for stochastic optimiziaton and the theory of stochastic processes [8, 9, 12, 13, 14], and weak optimal transport can be applied to problems in economics, finance and machine learning [1, 4, 6, 7, 10, 11, 15, 19, 23, 27, 34, 38, 39, 40].

This thesis aims to contribute to this two-way exchange between theory and applications, with a particular focus on financial applications of (weak) optimal transport.

In the first chapter we go from theory to applications by building on a recent extension of the Wasserstein barycenter [28] (an optimal transport concept used to average probability measures) to develop a clustering and reconstruction algorithm for probability measures with missing coordinates. This allows a *novel application of optimal transport in financial regulation.*

The second chapter addresses optimal transport problems with non-linear costs and establishes a *fundamental theorem for the weak optimal transport problem.* This leads to short unified derivations of known results as well as new applications.

In the third chapter of this thesis, we show how problems arising from financial applications can naturally lead to new theoretical results. While the Black-Scholes model remains the benchmark model in the financial industry, we explore an approach using Bass martingales to construct a local volatility model that is as close as possible to Black-Scholes (in an adapted optimal transport sense) while being perfectly calibrated to market data. Specifically, we extend a classical mathematical finance principle - the *change of numeraire* transformation - which was applied by Campi, Laachir and Martini to martingale optimal transport [22], to the setting of weak martingale optimal transport, in order to map Bass martingales to *geometric Bass martingales.*

Optimal Transport in Financial Regulation

The Geometry of Financial Institutions – Wasserstein Clustering of Financial Data

The optimal transport problem goes back to Gaspard Monge, who introduced it in [41] in 1781. Given probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ on Polish spaces $\mathcal{X}, \mathcal{Y}$ and a lower semi-continuous cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$, the problem is formulated as

$$\inf_{\substack{T:\mathcal{X}\to\mathcal{Y} \\ T_{\#}\mu=\nu}} \int_{\mathcal{X}} c(x, T(x))\, \mu(\mathrm{d}x). \tag{Monge}$$

Here and throughout, # denotes the push-forward operator of a measure by a measurable map. Specifically, for a measurable map $T : \mathcal{X} \to \mathcal{Y}$ and a measure μ, it defines the measure $T_{\#}\mu$ on $\mathcal{Y}$ via $T_{\#}\mu(B) := \mu(T^{-1}(B))$ for measurable $B \subset \mathcal{Y}$.

The Monge problem looks for the most cost-efficient way of moving the mass described by μ to the mass described by ν, where the cost of moving one particle x to $y = T(x)$ is $c(x, y)$. However, the problem is nonlinear in T and often no maps pushing μ towards ν exist, e.g. when $\mu = \delta_x$ and $\nu \neq \delta_y$ for all $y \in \mathcal{Y}$.

More than 150 years later, Leonid Kantorovich relaxed Monge's formulation in [37] by allowing for *couplings* rather than deterministic maps. The set of couplings is defined as

$$\Pi(\mu, \nu) := \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \mathrm{proj}_{\mathcal{X}\#}\pi = \mu, \mathrm{proj}_{\mathcal{Y}\#}\pi = \nu\},$$

i.e. the set of all probability measures on the product space $\mathcal{X} \times \mathcal{Y}$ whose marginals are μ and ν, respectively. This leads to the Kantorovich formulation of the optimal transport problem, namely

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi(\mathrm{d}x, \mathrm{d}y). \tag{OT}$$

Unlike the Monge problem (Monge), the Kantorovich formulation is linear in π and (OT) admits a solution under very mild conditions on the cost function (see e.g. [45, Theorem 1.3]). In the first chapter of this thesis, we use a particular case of the Kantorovich problem and consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x, y) = |x - y|^p$, for $p \in [1, \infty)$. Then, the value of the optimal transport problem gives rise to the *Wasserstein distance*

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p \, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p},$$

which is a metric on $\mathcal{P}_p(\mathbb{R}^d)$, the space of all probability measures with finite p-th moment. This metric induces a complete, separable metric space, the *Wasserstein space*.

The use of the Wasserstein distance, specifically with $p = 2$, in the first chapter of this thesis is motivated by a particular problem in the financial industry: regulators are responsible for the examination and supervision of a large number of financial institutions while the amount and granularity of reported data continues to increase. For instance, Oesterreichische Nationalbank (OeNB), the central bank of Austria, is responsible for more than 400 banks. It is infeasible to thoroughly examine each institution on a yearly basis. Therefore, it is crucial to identify banks which form clusters, i.e. are similar to each other, and to detect outliers, i.e. banks very different from most other banks. Concerning the increasing amount of data, banks in Austria must report their credits to the central bank if their size exceeds a certain threshold. While this granular credit data best reflects the credit portfolio of a bank, it makes it more challenging for OeNB to keep a comprehensive overview of the Austrian banking landscape.

In the first chapter we address this challenge by interpreting each bank as a probability measure on the space of credits. Each point in the support of a measure corresponds to a specific credit, with its location determined by the attributes of that credit. The Wasserstein distance can be used to measure similarity between the credit portfolios of two financial institutions.

Unfortunately, the situation is made more complicated by the fact that the reporting requirements vary across financial institutions depending on their size and loan types. As a result, not all institutions report the same attributes and the probability measures representing the banks may live on different subspaces. To overcome this, we use the generalized Wasserstein barycenter introduced by Delon, Gozlan and Saint-Dizier [28], which extends the classical notion of the Wasserstein barycenter by Agueh and Carlier [3]. The classical Wasserstein barycenter is a notion of averaging probability measures, specifically a Fréchet mean in the Wasserstein space, and the generalized Wasserstein barycenter adapts this to partially observed data.

To recall the generalized Wasserstein barycenter, let $\mu_1, \dots, \mu_n \in \mathcal{P}_2(\mathbb{R}^d)$ be probabilities, $P_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$ be linear maps and let $\lambda_1, \dots, \lambda_n \geq 0$ be weights summing to 1. Suppose we observe the transformed measures $\tilde{\mu}_i = P_{i\#}\mu_i$. Then, a generalized Wasserstein barycenter of $\tilde{\mu}_1, \dots, \tilde{\mu}_n$ with weights $\lambda_1, \dots, \lambda_n$ is a solution of

$$\min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i W_2^2(\tilde{\mu}_i, P_{i\#}\nu).$$

Given a bank's reported credit data, we interpret it as a probability measure under a bank-specific projection on the reported coordinates; this corresponds to $\tilde{\mu}_i$ and P_i. The generalized Wasserstein barycenter enables us to compute an average of such measures, thus representing the banks. Combined with the Wasserstein distance, this allows us to generalize the classical k-means clustering problem. Specifically, we propose a clustering algorithm for probability measures defined on potentially different subspaces. Once the clustering is done, we use the clusters to impute the missing information: non-fully observed probability measures are imputed by using the fully observed probability measures from the same cluster. This procedure embeds all the probability measures into a common space, allowing the computation of pairwise distances and the construction of a metric structure on these probability measures. Since these probability measures represent banks, we refer to this metric structure as *banking landscape*, cf. Figure 1.

The Fundamental Theorem of Weak Optimal Transport

As in the previous section, many applications of optimal transport are in the setting $\mathcal{X} = \mathcal{Y}$ with cost $c = d_{\mathcal{X}}$, the metric on $\mathcal{X}$. One such example is the *Kantorovich-Rubinstein* formula [36], which states that for $\mu, \nu \in \mathcal{P}_1(\mathcal{X})$ we have $W_1(\mu, \nu) = \max_{\psi:1-Lip} \int \psi \, \mathrm{d}\mu - \int \psi \, \mathrm{d}\nu$. Another example is Brenier's theorem [21], which asserts that for $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $c(x, y) = |x - y|^2$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with μ being absolutely continuous with respect to the Lebesgue measure, the Monge optimal transport problem (Monge) admits a unique solution characterized as the gradient of a convex function. Both of these results follow directly from the *fundamental theorem of optimal transport*, cf. [5, Theorem 1.13], which we recall here.

Theorem (Fundamental Theorem of OT)**.** *For $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ lower semi-continuous,*
(OT) *is attained and equals the dual problem*

$$\sup \left\{ \int f(x) \, \mu(\mathrm{d}x) + \int g(y) \, \nu(\mathrm{d}y) : f(x) + g(y) \leq c(x, y), f \in \mathcal{L}^1(\mu), g \in \mathcal{L}^1(\nu) \right\}.$$

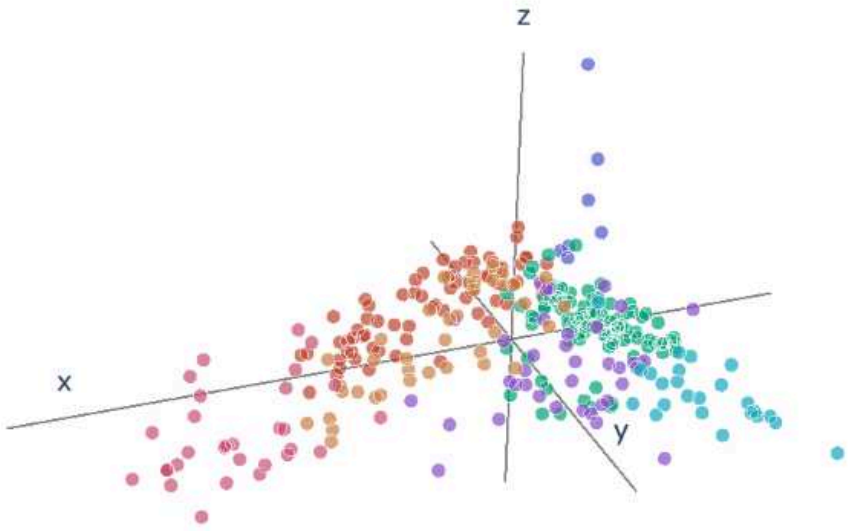

Figure 1: 3-d visualization of the Austrian Banking Landscape.
Interactive plot: `https://lorenzriess.github.io/TGOFI_landscape.html`

The dual problem is attained if c is upper bounded in the sense that

$$c(x,y) \leq a(x) + b(y), \quad \text{for some } a \in \mathcal{L}^1(\mu), b \in \mathcal{L}^1(\nu). \tag{b}$$

Candidates $\pi, (f,g)$ are optimal if and only if they satisfy the complementary slackness condition

$$c(x,y) = f(x) + g(y), \quad \pi - a.s.$$

However, many applications cannot be addressed by the classical optimal transport problem, since in (OT) the couplings enter the objective function linearly. To overcome this, Gozlan, Roberto, Samson and Tetali introduced the *weak optimal transport problem* in [33]. Given a measurable cost function $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to [0,\infty]$ that is lower semi-continuous and convex in its second argument, they define

$$\mathrm{WT}_C(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} C(x, \pi_x)\, \mu(\mathrm{d}x), \tag{WOT}$$

where $(\pi_x)_{x\in\mathcal{X}}$ denotes the disintegration of π with respect to μ.

In the second chapter of this thesis, we establish a fundamental theorem for the weak optimal transport problem. To achieve this, we adapt the boundedness condition (b) of classical optimal transport and additionally introduce a mild continuity condition. The boundedness condition for weak transport costs states

$$C(x,\rho) \leq a(x) + \int b(y)\, \rho(\mathrm{d}y) + \int h\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(y)\Big)\, \nu(\mathrm{d}y), \tag{B}$$

for some μ-integrable function a, ν-integrable function b and a convex, increasing function $h : [0,\infty) \to [0,\infty)$, where the right-hand side of (B) is defined as $+\infty$ if ρ is not absolutely continuous with respect to ν.

The continuity condition states that for any increasing and exhausting sequence of Borel sets $(Y_k)_k \subset \mathcal{Y}$, we have

$$C(x,\rho) \geq \limsup_k C(x, (\rho_{|Y_k})/\rho(Y_k)). \tag{C}$$

Theorem (Fundamental Theorem of WOT)**.** *Let $C : \mathcal{X} \times \mathcal{P}_p(\mathcal{Y}) \to [0,\infty]$ be measurable. Assume that $\rho \mapsto C(x,\rho)$ is convex and lower semi-continuous with respect to p-weak convergence on $\mathcal{P}_p(\mathcal{Y})$. Then $\mathrm{WT}_C(\mu,\nu)$ is attained and equals the dual problem*[1]

$$\mathrm{D}_C(\mu,\nu) := \sup\Big\{ \int f(x)\,\mu(\mathrm{d}x) + \int g(y)\,\nu(\mathrm{d}y) : \\ f(x) + \int g(y)\,\rho(\mathrm{d}y) \leq C(x,\rho),\ f \in \mathcal{L}^1(\mu),\ g \in \mathcal{L}^1(\nu)\Big\}.$$

Moreover, if C satisfies conditions (B) *and* (C)*, then the dual problem is attained. A pair of candidates $(\pi,(f,g))$ is optimal if and only if it satisfies the complementary slackness condition*

$$C(x,\pi_x) = f(x) + \int g(y)\,\pi_x(\mathrm{d}y), \quad \mu\text{-a.s.}$$

Just as the fundamental theorem of optimal transport leads directly to important results, also the fundamental theorem of weak optimal transport has a number of applications. In the second chapter of this thesis, we discuss several of them, three of which we recall below.

Convex Kantorovich–Rubinstein Analogously to the Kantorovich–Rubinstein formula which is a direct consequence of the fundamental theorem of optimal transport (as discussed above), the *convex Kantorovich–Rubinstein* formula [4, 32] can be seen as a direct consequence of the fundamental theorem of weak optimal transport. It states

$$\inf_{\pi\in\Pi(\mu,\nu)} \int |x - \mathrm{mean}(\pi_x)|\,\mu(\mathrm{d}x) = \max_{\psi \text{ convex, 1-Lip}} \int \psi(x)\,\mu(\mathrm{d}x) - \int \psi(y)\,\nu(\mathrm{d}y), \tag{1}$$

where $\mathrm{mean}(\rho) = \int y\,\rho(\mathrm{d}y)$ denotes the barycenter of ρ. The left-hand side of (1) equals zero if and only if there exists a coupling between μ and ν that is the law of a one-step martingale, i.e. a *martingale coupling*. We denote the set of all martingale couplings between μ and ν by

$$\mathscr{M}(\mu,\nu) := \{\pi \in \Pi(\mu,\nu) : \mathrm{mean}(\pi_x) = x,\ \mu\text{-a.s.}\}.$$

A famous result by Strassen [44] asserts that $\mathscr{M}(\mu,\nu) \neq \emptyset$ if and only if μ and ν are in convex order, i.e. $\int \psi\,\mathrm{d}\mu \leq \int \psi\,\mathrm{d}\nu$ for all convex functions ψ. Therefore, (1) can be seen as a quantitative generalization of Strassen's theorem: the left-hand side measures how close we are to finding a martingale coupling, while the right-hand side quantifies how far μ and ν are from being in convex order.

[1] Specifically we require the inequality in (1.6) for all $(x,\rho) \in \mathcal{X} \times \mathcal{P}_p(\mathcal{Y})$ with $g \in \mathcal{L}^1(\rho)$.

The Brenier–Strassen theorem Analogous to how Brenier's theorem follows from the fundamental theorem of optimal transport, the *Brenier–Strassen theorem*, established by Gozlan and Juillet [31], follows from the fundamental theorem of weak optimal transport. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the Brenier–Strassen theorem asserts that the weak transport problem

$$\inf_{\pi \in \Pi(\mu,\nu)} \int |x - \mathrm{mean}(\pi_x)|^2 \, \mu(\mathrm{d}x), \tag{2}$$

admits a μ-a.s. unique solution characterized by a map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T(x) = \mathrm{mean}(\pi_x)$ for every optimal π. Moreover, T is a Lipschitz function with Lipschitz constant 1 and is the gradient of a convex function. Optimizers of (2) satisfy

$$\pi \in \Pi(\mu,\nu) \text{ is optimal } \iff \pi_x = \kappa_{T(x)}, \ \mu\text{-a.s. for some } \kappa \in \mathscr{M}(T_{\#}\mu, \nu).$$

Entropic Optimal Transport Another variant of optimal transport is entropic optimal transport. While Monge-type couplings are supported on the graph of a function, entropic optimal transport seeks more regular couplings by adding an entropic penalty term. Specifically, for $\varepsilon > 0$, the entropic optimal transport problem is given by

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c \, \mathrm{d}\pi + \varepsilon H(\pi | \mu \otimes \nu), \tag{3}$$

where H denotes relative entropy. That is, $H(Q|R) = \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}R}\right) \mathrm{d}Q$ if Q, R are probability measures with Q being absolutely continuous with respect to R and $H(Q|R) = +\infty$ otherwise. Since the entropy can be rewritten as $H(\pi|\mu \otimes \nu) = \int H(\pi_x|\nu)(x) \, \mu(\mathrm{d}x)$, problem (3) is a weak optimal transport problem. Applying the fundamental theorem of weak optimal transport then recovers the main *structure theorem of entropic optimal transport*, which states that (3) admits a unique solution π characterized by

$$\frac{\mathrm{d}\pi}{\mathrm{d}\mu \otimes \nu}(x,y) = \exp\left(-\frac{c(x,y) + f(x) + g(y)}{\varepsilon}\right).$$

In this case, (f, g) are also the optimizers of the corresponding dual problem.

Local Volatility through Weak Martingale Transport

Change of Numeraire for Weak Martingale Transport

While the second chapter of this thesis focuses on non-linear costs in the transport problem, the third chapter is concerned with *weak martingale optimal transport problems*, where we additionally restrict the set of admissible couplings to martingale couplings. Throughout, we will work on the positive real line $\mathbb{R}_{++} = (0, \infty)$ and consider for $\mu, \nu \in \mathcal{P}_1(\mathbb{R}_{++})$ in convex order, i.e. $\mu \leq_c \nu$, the weak martingale transport problem

$$\inf_{\pi \in \mathscr{M}(\mu,\nu)} \int_{\mathbb{R}_{++}} C(x, \pi_x) \, \mu(\mathrm{d}x). \tag{WMOT}$$

In this third chapter, we generalize a classical concept from mathematical finance - the *change-of-numeraire* (CN) transformation - in order to relate weak martingale optimal

transport problems. Campi, Laachir, and Martini [22] previously established its applicability to martingale optimal transport by showing that (CN) maps a martingale coupling π between μ and ν to a martingale coupling $S(\pi)$ between $S(\mu)$ and $S(\nu)$. We extend this idea to the weak setting: that is, we map the solution of a weak martingale optimal transport problem to the solution of an altered weak martingale optimal transport problem. In the third chapter, we present several applications of this technique, in particular we establish a correspondence between *Bass martingales (stretched Brownian motion)* and *geometric Bass martingales (geometric stretched Brownian motion)*.

The classical change of numeraire transformation stems from Bayes' rule: given a one-step $\mathbb{P}$-martingale (X_0, X_1), the process $(Y_0, Y_1) := (1/X_0, 1/X_1)$ is a $\hat{\mathbb{P}}$-martingale under the measure $\hat{\mathbb{P}}$ defined via $\frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\mathbb{P}} := \frac{X_1}{\mathbb{E}_{\mathbb{P}}[X_1]}$, since

$$\mathbb{E}_{\hat{\mathbb{P}}}[Y_1|Y_0] = \frac{\mathbb{E}_{\mathbb{P}}[Y_1 X_1|Y_0]}{\mathbb{E}_{\mathbb{P}}[X_1|Y_0]} = \frac{\mathbb{E}_{\mathbb{P}}[1|Y_0]}{\mathbb{E}_{\mathbb{P}}[X_1|X_0]} = \frac{1}{X_0} = Y_0.$$

This motivates the definition of the *CN-transformation*. Let $I \in \{1, \{0,1\}, [0,1]\}$ denote a set of time indices, and write $(1/x)$ for the path $(1/x_t)_{t\in I}$. We define

$$S : \mathscr{M}_I \to \mathscr{M}_I, \quad S(\pi)(B) := \frac{\int x_1 \delta_{(1/x)}(B)\, \pi(\mathrm{d}x)}{\int x_1\, \pi(\mathrm{d}x)}, \quad \text{for measurable } B \subseteq \mathcal{D}_I,$$

where $\mathcal{D}_I \in \{\mathbb{R}_{++}, \mathbb{R}^2_{++}, \mathcal{D}([0,1])\}$, depending on I, and $\mathscr{M}_I$ denotes the set of all martingale laws on $\mathcal{D}_I$. We identify $\mathscr{M}_1 = \mathcal{P}_1(\mathbb{R}_{++})$ and denote by $\mathcal{D}([0,1])$ the set of càdlàg paths on $[0,1]$.

Before applying the CN-transformation, let us recall the notion of stretched Brownian motion. It is the process of maximal correlation with Brownian motion given initial and terminal laws $\mu \leq_c \nu$, i.e. the solution of

$$\sup\Big\{\mathbb{E}\Big[\int_0^1 \sigma_t\, \mathrm{d}t\Big] : \mathrm{d}X_t = \sigma_t \mathrm{d}B_t, X_0 \sim \mu, X_1 \sim \nu\Big\}. \tag{SBM}$$

There exists a unique-in-law solution to (SBM), cf. [10, Theorem 1.5] , which we denote by $\mathbb{Q}^{\mathrm{SBM},\mu,\nu}$.

The stretched Brownian motion is related to weak martingale transport since the optimal value of (SBM) coincides with the value of

$$\sup_{\pi\in\mathscr{M}_{\{0,1\}}(\mu,\nu)} \int_{\mathbb{R}_{++}} \mathrm{MCov}(\pi_x, \gamma_1)\, \mu(\mathrm{d}x), \tag{4}$$

which is a weak martingale optimal transport problem with cost function $C(x,\rho) = -\mathrm{MCov}(\rho, \gamma_1)$. Here, MCov denotes the maximum covariance, defined as $\mathrm{MCov}(p_1, p_2) := \sup_{q\in\Pi(p_1,p_2)} \int x_1 x_2\, q(\mathrm{d}x_1, \mathrm{d}x_2)$. The connection goes further: the projection of $\mathbb{Q}^{\mathrm{SBM},\mu,\nu}$ to $\{0,1\}$ is the unique optimizer of (4).

If liquidly traded European call options are available at a given maturity, a famous result by Breeden and Litzenberger [20] states that their prices determine the marginal distribution of the underlying asset under the risk-neutral measure at that maturity. If options are available at two maturities, the Breeden-Litzenberger theorem determines both

the initial and terminal laws μ_0 and μ_1 of the underlying. It is then a key modeling question which stochastic process to select to model the underlying asset given these marginals. The stretched Brownian motion is one such choice, since it is the stochastic process whose returns are maximally correlated with Brownian motion among all processes fitting the marginals.

In contrast, the benchmark model in mathematical finance is the Black-Scholes model, i.e. a geometric Brownian motion. This motivates the consideration of the *geometric stretched Brownian motion*, the process whose log-returns are maximally correlated with Brownian motion while still fitting the given marginals, i.e.

$$\sup\Big\{\mathbb{E}\Big[\int_0^1 \sigma_t \,\mathrm{d}t\Big] : \mathrm{d}Y_t = \sigma_t Y_t \mathrm{d}B_t, Y_0 \sim \mu, Y_1 \sim \nu\Big\}. \tag{GSBM}$$

We show that geometric Brownian motion also corresponds to a weak martingale optimal transport problem, namely

$$\sup_{\pi \in \mathscr{M}_{\{0,1\}}(\mu,\nu)} \int_{\mathbb{R}_{++}} x\mathrm{MCov}(S(\pi_x), \gamma_1)\, \mu(\mathrm{d}x).$$

Using the CN-transformation, we then establish a correspondence between stretched Brownian motion and its geometric counterpart. Specifically, for the unique-in-law solution $\mathbb{Q}^{\mathrm{gSBM},\mu,\nu}$ to (GSBM), we prove

$$\mathbb{Q}^{\mathrm{gSBM},\mu,\nu} = S(\mathbb{Q}^{\mathrm{SBM},S(\mu),S(\nu)}). \tag{5}$$

The identity (5) is also relevant in practice, since after fitting a model, it is often necessary to simulate paths for applications such as Monte Carlo pricing. Since efficient algorithms for simulating paths of stretched Brownian motion are available (cf. [2]), this transformation allows one to simulate paths of geometric stretched Brownian motion by converting simulated paths of its non-geometric counterpart.

References

[1] B. Acciaio, M. Beiglböck, and G. Pammer. Weak transport for non-convex costs and model-independence in a fixed-income market. *Math. Finance* 31.4 (2021), pp. 1423–1453.

[2] B. Acciaio, A. Marini, and G. Pammer. Calibration of the Bass Local Volatility model. *ArXiv e-prints* 2311.14567 (2023).

[3] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[4] J.-J. Alibert, G. Bouchitté, and T. Champion. A new class of costs for optimal transport planning. *European Journal of Applied Mathematics* 30.6 (2019), pp. 1229–1263.

[5] L. Ambrosio and N. Gigli. A user's guide to optimal transport. *Modelling and optimisation of flows on networks*. Vol. 2062. Lecture Notes in Math. Springer, Heidelberg, 2013, pp. 1–155.

[6] A. Asadulaev, A. Korotin, V. Egiazarian, P. Mokrov, and E. Burnaev. Neural Optimal Transport with General Cost Functionals. *The Twelfth International Conference on Learning Representations*. 2024.

[7] J. Backhoff-Veraguas and G. Pammer. Applications of weak transport theory. *Bernoulli* 28.1 (2022), pp. 370–394.

[8] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance Stoch.* 24.3 (2020), pp. 601–632.

[9] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and J. Wiesel. Estimating processes in adapted Wasserstein distance. *Ann. Appl. Probab.* 32.1 (2022), pp. 529–550.

[10] J. Backhoff-Veraguas, M. Beiglböck, M. Huesmann, and S. Källblad. Martingale Benamou-Brenier: a probabilistic perspective. *Ann. Probab.* 48.5 (2020), pp. 2258–2289.

[11] J. Backhoff-Veraguas, G. Loeper, and J. Obloj. Geometric Martingale Benamou-Brenier transport and geometric Bass martingales. *ArXiv e-prints* 2406.04016 (2024).

[12] D. Bartl, M. Beiglböck, and G. Pammer. The Wasserstein space of stochastic processes. *J. Eur. Math. Soc.* (2024). To appear. arXiv:2104.14245.

[13] D. Bartl, M. Beiglböck, G. Pammer, S. Schrott, and X. Zhang. The Wasserstein Space of Stochastic Processes in Continuous Time. *arXiv:2501.14135* (2025).

[14] D. Bartl and J. Wiesel. Sensitivity of Multiperiod Optimization Problems with Respect to the Adapted Wasserstein Distance. *SIAM J. Financ. Math.* 14.2 (2023), pp. 704–720.

[15] M. Beiglböck, B. Jourdain, W. Margheriti, and G. Pammer. Monotonicity and stability of the weak martingale optimal transport problem. *Annals of Applied Probability, to appear* (2024).

[16] M. Beiglböck and N. Juillet. On a problem of optimal transport under marginal martingale constraints. *Ann. Probab.* 44.1 (2016), pp. 42–106.

[17] M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent Bounds for Option Prices: A Mass Transport Approach. *Finance Stoch.* 17.3 (2013), pp. 477–501.

[18] J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electron. J. Stat.* 13.2 (2019), pp. 5120–5150.

[19] K. Bredies, E. Chenchene, and A. Hosseini. A hybrid proximal generalized conditional gradient method and application to total variation parameter learning (2022).

[20] D. T. Breeden and R. H. Litzenberger. Prices of State-contingent Claims Implicit in Option Prices. *The Journal of Business* 51.4 (1978), pp. 621–51.

[21] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.* 44.4 (1991), pp. 375–417.

[22] L. Campi, I. Laachir, and C. Martini. Change of numeraire in the two-marginals martingale transport problem. *Finance Stoch.* 21.2 (June 2017), pp. 471–486.

[23] A. Conze and P. Henry-Labordere. Bass Construction with Multi-Marginals: Lightspeed Computation in a New Local Volatility Model. en. *SSRN Electronic Journal* (2021).

[24] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[25] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2017), pp. 1853–1865.

[26] M. Cuturi and A. Doucet. Fast Computation of Wasserstein Barycenters. *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, June 2014, pp. 685–693.

[27] C. Daskalakis, A. Deckelbaum, and C. Tzamos. Strong Duality for a Multiple-Good Monopolist. *Econometrica* 85.3 (2017), pp. 735–767.

[28] J. Delon, N. Gozlan, and A. S. Dizier. Generalized Wasserstein barycenters between probability measures living on different subspaces. English. *Ann. Appl. Probab.* 33.6A (2023), pp. 4395–4423.

[29] Y. Dolinsky and H. M. Soner. Martingale optimal transport and robust hedging in continuous time. *Probab. Theory Relat. Fields* 160.1-2 (2014), pp. 391–427.

[30] A. Genevay, G. Peyre, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, Apr. 2018, pp. 1608–1617.

[31] N. Gozlan and N. Juillet. On a mixture of Brenier and Strassen theorems. *Proceedings of the London Mathematical Society* 120.3 (2020), pp. 434–463.

[32] N. Gozlan, C. Roberto, P.-M. Samson, Y. Shu, and P. Tetali. Characterization of a class of weak transport-entropy inequalities on the line. *Ann. Inst. Henri Poincaré Probab. Stat.* 54.3 (2018), pp. 1667–1693.

[33] N. Gozlan, C. Roberto, P.-M. Samson, and P. Tetali. Kantorovich duality for general transport costs and applications. *J. Funct. Anal.* 273.11 (2017), pp. 3327–3405.

[34] J. Guyon, R. Menegaux, and M. Nutz. Bounds for VIX futures given S&P 500 smiles. *Finance and Stochastics* 21.3 (2017), pp. 593–630.

[35] P. Henry-Labordère. Model-free hedging. A martingale optimal transport viewpoint. English. Chapman Hall/CRC Financ. Math. Ser. Boca Raton, FL: CRC Press, 2017. ISBN: 978-1-138-06223-8; 978-1-315-16174-7.

[36] L. Kantorovich and G. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ.* 13.7 (1958), pp. 52–59.

[37] L. Kantorovich. On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)* 37 (1942), pp. 199–201.

[38] A. Kolesov, P. Mokrov, I. Udovichenko, M. Gazdieva, G. Pammer, E. Burnaev, and A. Korotin. Estimating barycenters of distributions with neural optimal transport. English. *Proceedings of the 41st International Conference on Machine Learning*. 2024.

[39] A. Korotin, D. Selikhanovych, and E. Burnaev. Neural Optimal Transport. *The Eleventh International Conference on Learning Representations*. 2023.

[40] M. Kupper, M. Nendel, and A. Sgarabottolo. Risk measures based on weak optimal transport. *Quantitative Finance* (2024), pp. 1–18.

[41] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'académie Royale des Sciences de Paris* (1781).

[42] G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.

[43] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances. *ACM Transactions on Graphics* 34.4 (2015), 66:1–66:11.

[44] V. Strassen. The existence of probability measures with given marginals. *Ann. Math. Statist.* 36 (1965), pp. 423–439.

[45] C. Villani. Topics in Optimal Transportation. Vol. 58. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, 2003, pp. xvi+370. ISBN: 0-8218-3312-X.

Publication and Authorship Information

Article 1

The Geometry of Financial Institutions – Wasserstein Clustering of Financial Data

L. Riess, J. Backhoff, M. Beiglböck, J. Temme, A. Wolf

Under revision at Mathematics and Financial Economics.(Submission acknowledged April 4, 2025.)

Preprint available at: `https://arxiv.org/abs/2305.03565`

Contributions of the thesis-author:

The article originated from a practical problem introduced by J. Temme and A. Wolf. The core idea was developed jointly with J. Backhoff and M. Beiglböck. The paper was subsequently written by the author of the thesis, accompanied by regular discussions with J. Backhoff and M. Beiglböck. The proofs of the results as well as implementation of the algorithm were carried out by the author of the thesis.

Article 2

The Fundamental Theorem of Weak Optimal Transport

M. Beiglböck, G. Pammer, L. Riess, S. Schrott

Submitted to Probability Theory and Related Fields.(Submission acknowledged June 30, 2025.)

Preprint available at: `https://arxiv.org/abs/2501.16316`

Contributions of the thesis-author:

The article was worked out in close collaboration.

Article 3

Change of Numeraire for Weak Martingale Transport

M. Beiglböck, G. Pammer, L. Riess

Under revision at Stochastic Processes and their Applications.(Submission acknowledged July 31, 2024.)

Preprint available at: `https://arxiv.org/abs/2406.07523`

Contributions of the thesis-author:

The article was worked out in close collaboration.

The Geometry of Financial Institutions - Wasserstein Clustering of Financial Data

L. Riess[†,‡], J. Backhoff[*,†], M. Beiglböck[*,†], J. Temme[*,‡], A. Wolf[*,‡]

Abstract

Financial regulation requires the submission of diverse and often highly granular data from financial institutions to regulators. In turn, regulators face the challenge of condensing this data into a comprehensive map that captures the mutual similarity or distance between different institutions and identifies clusters or outliers based on features like size, credit portfolio, or business model. Additionally, missing data due to varying regulatory requirements for different types of institutions, can further complicate this task.

To address these challenges, we interpret the credit data of financial institutions as probability distributions whose respective distances can be assessed through optimal transport theory. Specifically, we propose a variant of Lloyd's algorithm that applies to probability distributions and uses generalized Wasserstein barycenters to construct a metric space. Our approach provides a solution for the mapping of the banking landscape, enabling regulators to identify clusters of financial institutions and assess their relative similarity or distance.

This research was funded in whole or in part by the Austrian Science Fund (FWF) under grants 10.55776/P36835, P35197, and Y782, and the Oesterreichische Nationalbank (OeNB) through project EATE II.

[*]Authors listed in alphabetical order
[†]University of Vienna
[‡]Oesterreichische Nationalbank

1 Introduction

The main contribution of this article is an algorithm that takes several (discrete) probability distributions in a high-dimensional space and clusters them, representing each cluster as a point in a metric space. In this space, the distance between points reflects how different the underlying distributions are. A key feature of the algorithm is its ability to handle missing data — even when entire coordinates are missing systematically from some distributions.

Our original motivation to devise this type of algorithm stems from a challenge faced by regulators of the financial industry. Data that are delivered from financial institutions to the regulator consist of various different formats, from highly aggregate data such as the total volume of the balance sheet, down to very granular data about individual credits described by characteristics such as volume, interest rate, etc. Two individual credits might then be considered as similar if those characteristics take similar numerical values, i.e. have small Euclidean distance when viewed as vectors in $\mathbb{R}^d$. To build a distance between financial institutions, one can view these institutions as probability distributions on the space of possible credits (see Section 7.2 below) and determine the respective distance as a Wasserstein distance. One can then interpret the landscape of all financial institutions as an ensemble of points in the Wasserstein space, susceptible to familiar methods of clustering, outlier detection, etc. A particular challenge which renders the problem more complicated is the "missingness of data". Data delivered by institutions may have missing values. More crucially, data may be missing systematically, as different institutions are required to deliver different data at varying levels of detail and granularity.

The algorithm we propose as a ramification simultaneously clusters probability measures with missing data and represents them as elements of a metric space. The principal structure follows the idea of Lloyd's algorithm for k-means clustering and combines it with the concept of generalized Wasserstein barycenters, recently introduced by Delon, Gozlan and Saint-Dizier [12]. A classical approach to dealing with missing data would be (e.g.) to impute from weighted nearest neighbors. However, such a type of imputation systematically skews results since probability distributions with less reported data tend to appear closer to other points which the imputation procedure is attempting to mimic. This type of bias may be undesirable, e.g. when one is trying to identify distributions that are "atypical", where "atypical" could specifically refer to the manner in which data are missing. We devise a particular way of soft imputation which accounts for a random element in filling up missing values, and in particular avoids the above mentioned bias.

It will be technically convenient to formulate the algorithm in a more general form in Section 3. That is, we cluster and perform soft imputation for points in an arbitrary metric space rather than a Wasserstein space of probability measures. This allows us to simplify the presentation and has the additional benefit that we obtain a version of classical Euclidean k-means clustering with missing data. This general perspective allows us in Section 7.1 to compare our method to existing solutions for reconstructing the metric arrangement of Euclidean points. Additionally, further simulation experiments can be found in the Supplementary Material, and our method is compared to others when viewed solely as a clustering method. In particular, it is compared to k-pod [9], another method that tackles the same problem in the Euclidean case, i.e. (3.8), meaning the clustering

problem without imputing a priori. Notably, our clustering algorithm outperforms k-pod consistently. As for the important case of probability distributions with missing data our contribution is detailed in Section 5 and complemented experimentally in Sections 7.2 and 7.3 with results using actual (anonymized) loan data reported by financial institutions to Oesterreichische Nationalbank, the central bank of Austria.

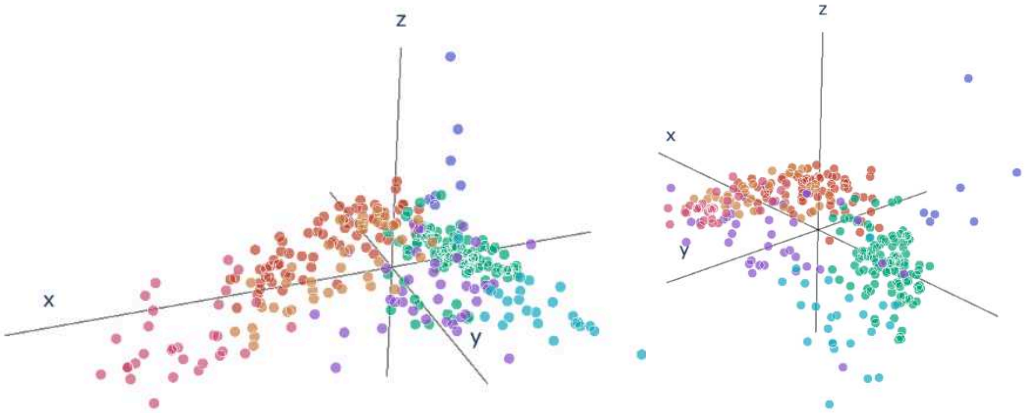

Figure 2: 3-d visualization of the Austrian Banking Landscape from two different perspectives.
Interactive plot: `https://lorenzriess.github.io/TGOFI_landscape.html`

Related Literature

Clustering distributions using Wasserstein distances has been explored in various works including [16, 19, 25, 36, 37]. This approach has also been applied in financial contexts, such as in [17], for the clustering of market regimes. In [31], Staib et al. proposed Wasserstein k-means++ as well as an initialization strategy, generalizing the classical k-means++ algorithm, cf. [6]. However, to the best of our knowledge, clustering distributions with missing coordinates has not been investigated before. To address this problem, we rely heavily on the concept of the generalized Wasserstein barycenter, introduced by Delon, Gozlan, and Saint-Dizier in [12], which extends the classical Wasserstein barycenter of Agueh and Carlier [1]. In the Euclidean case, we are aware of one method for the k-means problem with missing values which does not impute points beforehand. This method is called k-pod and was introduced by Chi et.al. in [9]. For computational aspects related to optimal transport and regularized Wasserstein distance, we refer to the work of Cuturi [10] and the book [27] by Cuturi and Peyré. Additionally, we use the `POT` package (Python package for optimal transport, see [13]) extensively for implementation purposes.

2 k-means Clustering in Metric Spaces: A Summary

Let $x_1, \dots, x_N$ be given points in a fixed metric space $(\mathcal{X}, d)$. The metric k-means problem is concerned with assigning the points to k clusters. The clusters hereby are governed by k

barycenters, i.e. an appropriate notion of average of the points in the corresponding cluster. k-means was first introduced by MacQueen in [23]. The problem can be formalized as

$$\min_{\substack{c_j \in \mathcal{X} \\ a \in [k]^N}} \sum_{i=1}^{N} d(x_i, c_{a_i})^2, \tag{KM}$$

where we use the notation $[n] := \{1, \dots, n\}$, $n \in \mathbb{N}$. In this formulation a denotes a vector of assignments, i.e. a_i indicates the cluster membership of data point x_i. The points $c_1, \dots, c_k$ serve as cluster barycenters. Furthermore, a notion of barycenter, typically a function of some of the data points, is needed. Then, problem (KM) can be tackled by the well-established Lloyd algorithm, cf. [22].

Following an initialization phase, two steps are iterated:

i) assignment step: given barycenters $c_j \in \mathcal{X}$, for each $i \in [N]$ pick

$$a_i \in \operatorname*{argmin}_{j \in [k]} d(x_i, c_j), \tag{2.1}$$

ii) barycenter step: given assignment a, update the barycenters, i.e. for each $j \in [k]$ pick

$$c_j \in \text{barycenter}(\{x_i : a_i = j\}). \tag{2.2}$$

Let us give two examples with particular choices of data space $\mathcal{X}$, metric d, and barycenter operation, which will be of interest in the sequel:

Example 2.1. Let $\mathcal{X} = \mathbb{R}^d$ and $d(x, y) = \|x - y\|_2$, the Euclidean space and distance. Then (KM) simplifies to the classical k-means problem in Euclidean space,

$$\min_{\substack{c_j \in \mathbb{R}^d \\ a \in [k]^N}} \sum_{i=1}^{N} \|x_i - c_{a_i}\|_2^2.$$

The barycenter function is the Euclidean mean/average.

Example 2.2 (Wasserstein k-means). Here $\mathcal{X} = \mathcal{P}_2(\mathbb{R}^d)$, the space of probability measures on $\mathbb{R}^d$ with finite second moments, and d is taken to be the Wasserstein distance W_2 (recalled in (4.1) below). We will henceforth write $\mu_i = x_i$ for $i \in [N]$. Then (KM) turns into

$$\min_{\substack{\nu_j \in \mathcal{P}_2(\mathbb{R}^d) \\ a \in [k]^N}} \sum_{i=1}^{N} W_2^2(\mu_i, \nu_{a_i}).$$

As barycenter one takes the Wasserstein barycenter, i.e. the barycenter updating step (2.2) reads

$$\nu_j \in \operatorname*{argmin}_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i : a_i = j} W_2^2(\mu_i, \nu). \tag{2.3}$$

Subsequently, we want to generalize the previous examples, in particular Example 2.2, to the case in which certain coordinates/marginals are missing for some observations, for an illustration see Figure 3 in Section 5.

When applying Lloyd's algorithm to a specific setup, one needs to specify a distance function and a notion of barycenter. In the case of a metric space it is natural to take a Fréchet mean for the barycenter operation, see [14]. In the case of missing coordinates, however, the data does not come from one metric space but from several different spaces, and hence several distances need to be considered simultaneously. In particular, it is also necessary to consider a generalized notion of barycenter. We will introduce an approach to this challenge in the next section and discuss our specific examples in more detail in Example 3.2 and Section 5, respectively.

3 Clustering Projected Elements of Metric Spaces

Suppose we are working in a metric space $(\mathcal{X}, d)$ and have points $x_1, \dots, x_N \in \mathcal{X}$ that we want to assign to k clusters. However, what we actually observe are the points $\tilde{x}_i := \varphi_i(x_i)$, where $\varphi_i : \mathcal{X} \to \mathcal{X}_i$ is a known map into another metric space $\mathcal{X}_i$. For instance, in the case $\mathcal{X} = \mathbb{R}^d$ the functions φ_i might be projections onto subspaces. Having observed $\tilde{x}_i \in \mathcal{X}_i$ define for $y \in \mathcal{X}$

$$\begin{aligned} d_i(\tilde{x}_i, y) &:= d(\varphi_i^{-1}(\tilde{x}_i), y) \\ &:= \inf_{x \in \mathcal{X} : \varphi_i(x) = \tilde{x}_i} d(x, y), \end{aligned}$$

which serves as a type of dissimilarity measure of a point $y \in \mathcal{X}$ in the "full" metric space to the observed point $\tilde{x}_i$.

3.1 Problem and Algorithm

Based on these notations, we introduce the problem

$$\min_{\substack{c_j \in \mathcal{X} \\ a \in [k]^N}} \sum_{i=1}^{N} d_i(\tilde{x}_i, c_{a_i})^2. \tag{3.1}$$

I.e. we want to find optimal cluster barycenters $c_1, \dots, c_k \in \mathcal{X}$, as well as a vector a that optimally assigns each observed point to a cluster. It is important to note that we are looking for cluster barycenters in the "full" space $\mathcal{X}$. This ensures that we are be able to compare them to each observed point $\tilde{x}_i$ using $d_i(\tilde{x}_i, \cdot)$.

We propose tackling (3.1) using the following two steps which are iterated in a Lloyd algorithm fashion after initializing the barycenters:

i) assignment step: given barycenters $c_j \in \mathcal{X}$, set

$$a_i \in \operatorname*{argmin}_{j \in [k]} d_i(\tilde{x}_i, c_j), \tag{3.2}$$

ii) barycenter step: given assignment a, update the barycenters, i.e. for each $j \in [k]$, set

$$c_j \in \operatorname*{argmin}_{y \in \mathcal{X}} \sum_{i:a_i=j} d_i(\tilde{x}_i, y)^2. \tag{3.3}$$

Let us note that step ii) does not necessarily admit minimizers. However, in our two applications – NA k-means (cf. Example 3.2) and NA Wasserstein k-means (cf. Section 5) – minimizers always exist. A more precise discussion of this technical point is given in Appendix A. We use the abbreviation "NA" for "Not Available". Thus, NA k-means and NA Wasserstein k-means refer to the respective algorithms of Examples 2.1 and 2.2 adapted to handle missing coordinates.

Concerning initialization, we (slightly) adapt the widely used k-means++ initialization algorithm introduced in [6]. In its original form, the initial cluster barycenters are selected from the observed points themselves. The algorithm begins by choosing the first barycenter uniformly at random from the observed points and then repeatedly choosing points at random with probability proportional to their squared distance from the already chosen barycenters until k barycenters are chosen. This approach is not directly applicable to our setting because not all points are fully observed, and thus not all pairwise distances can be computed. Therefore we apply the k-means++ initialization algorithm to the subset of fully observed points only, i.e. those points x_i with $\varphi_i = \mathrm{Id}_{\mathcal{X}}$.

3.2 Using Clustering for Imputation

After clustering we obtain a vector of assignments a and barycenters $c_1, \ldots, c_k$. We can use these to impute the not fully observed points. For this sake define

$$I_f := \{i : \varphi_i = \mathrm{Id}_{\mathcal{X}}\}, \quad I_m := [N] \setminus I_f, \tag{3.4}$$

i.e. respectively the set of indices of points being observed in $\mathcal{X}$ (i.e. fully observed data) and the set of indices of points that are only observed after some non-trivial map (i.e. with missing data). The final clusters are defined by letting for $j \in [k]$

$$C_j := \{i \in [N] : a_i = j\}.$$

We want to use the clusters to find, for a point $\tilde{x}_i$ with $i \in I_m$, a probability measure on $\mathcal{X}$, i.e. an element of $\mathcal{P}(\mathcal{X})$, which attempts to concentrate around the true (unobserved) point x_i. We define for $i \in I_m$ the set of indices we use for filling up the missing information of $\tilde{x}_i$, as

$$J_i := I_f \cap C_{a_i}.$$

This says that we want to use the "full" points in the same cluster as $\tilde{x}_i$ in order to compensate the incomplete information that we have about $\tilde{x}_i$. It can of course happen that $J_i = \emptyset$. In this case we use the corresponding cluster barycenter c_{a_i} to complement

the information about $\tilde{x}_i$. In the following we assume $J_i \neq \emptyset$. Taking now some x_ℓ with $\ell \in J_i$ we set

$$y_\ell^i \in \operatorname*{argmin}_{\substack{x \in \mathcal{X} \\ \varphi_i(x) = \tilde{x}_i}} d(x, x_\ell), \tag{3.5}$$

which will be one potential choice of "filling up". We introduce the shorthand $D_\ell^i := d(y_\ell^i, x_\ell) = d_i(\tilde{x}_i, x_\ell)$. In order to determine the weights of a probability measure, we set $p_\ell^i := f(D_\ell^i)$ with $f : [0, \infty) \to [0, \infty)$ being some positive decreasing function fixed in advance, e.g. $f(x) := \exp(-x^2)$. This way $p_\ell^i \geq 0$ and also $\sum_{\ell \in J_i} p_\ell^i = 1$ after possibly renormalizing the weights by a positive constant. Having obtained the weights, we define the probability measure which should represent a *randomly reconstructed* x_i as

$$\theta_i := \sum_{\ell \in J_i} p_\ell^i \delta_{y_\ell^i}. \tag{3.6}$$

To also embed the fully observed points, for $i \in I_f$ we set $\theta_i := \delta_{x_i}$, that is, the Dirac delta concentrated on x_i. For $i \in I_m$ with $J_i = \emptyset$, we use the corresponding cluster barycenter c_{a_i} and set θ_i to be the Dirac delta concentrated on some minimizer of $d(\varphi_i^{-1}(\tilde{x}_i), c_{a_i})$. Thus, we embed all observed points $\tilde{x}_1, \ldots, \tilde{x}_N$ in the same space $\mathcal{P}(\mathcal{X})$. In other words, we identify the possibly unobserved point x_i with a probability measure θ_i.

In order to compare the hitherto constructed probability measures, as we will need to do in Section 7, we define a (generalized) metric ρ on $\mathcal{P}(\mathcal{X})$. Readers primarily interested in clustering and imputation may skip this construction. The (generalized) metric ρ is defined via the cost induced by the *product* or *independent coupling*. That is, for $\mu, \nu \in \mathcal{P}(\mathcal{X})$ we set $\rho(\mu, \nu) := 0$ if $\mu = \nu$ and otherwise

$$\rho(\mu, \nu) := \int_{\mathcal{X}} \int_{\mathcal{X}} d(x, x') \, \mu(\mathrm{d}x) \, \nu(\mathrm{d}x'). \tag{3.7}$$

For completeness we provide a short lemma proving that ρ is indeed a (generalized) metric.

Lemma 3.1. *Let $(\mathcal{X}, d)$ be a metric space and define on $\mathcal{P}(\mathcal{X})$ the map $\rho : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, \infty]$ by*

$$\rho(\mu, \nu) := \begin{cases} \int_{\mathcal{X} \times \mathcal{X}} d(x, x') \, (\mu \otimes \nu)(\mathrm{d}x, \mathrm{d}x'), & \text{if } \mu \neq \nu \\ 0, & \text{if } \mu = \nu. \end{cases}$$

Then, ρ is a generalized metric on $\mathcal{P}(\mathcal{X})$.

Proof. Symmetry of the generalized metric is clear due to the symmetry of the underlying distance d. Concerning the triangle inequality take $\mu, \nu, \theta \in \mathcal{P}(\mathcal{X})$ which we assume to be different from each other (otherwise there is nothing to prove). Furthermore, take three independent random variables $X \sim \mu, Y \sim \nu, Z \sim \theta$. Then,

$$\begin{aligned} \rho(\mu, \theta) &= \int_{\mathcal{X} \times \mathcal{X}} d(x, x') \, (\mu \otimes \theta)(\mathrm{d}x, \mathrm{d}x') \\ &= \mathbb{E}[d(X, Z)] \\ &\leq \mathbb{E}[d(X, Y) + d(Y, Z)] = \rho(\mu, \nu) + \rho(\nu, \theta), \end{aligned}$$

which proves the triangle inequality for ρ. Concerning definiteness, if $\mu = \nu$, we have $\rho(\mu, \nu) = 0$ by definition.

Suppose now that

$$\int_{\mathcal{X}\times\mathcal{X}} d(x, x')\,(\mu \otimes \nu)(\mathrm{d}x, \mathrm{d}x') = 0.$$

This implies $d(x, x') = 0$ for $\mu \otimes \nu$-almost all (x, x'). Since d is a metric, we have $x = x'$, $\mu \otimes \nu$-almost surely. Thus, $\mu = \delta_x = \nu$ for some $x \in \mathcal{X}$. □

Using a Wasserstein distance or a similar notion of distance in (3.7), imputed points would be biased to be closer than "fully" observed points. The metric ρ is designed to avoid this type of bias. Indeed, our choice in (3.7) formalises the idea that missing values are not imputed in a deterministic sense but in a *random* or *soft* fashion, as indicated in the introduction. The distance between two randomly imputed values is then estimated as an *independent* average of distances, following the rationale that the imputation for one point does not inform the imputation for a different point.

Example 3.2 (NA k-means). Corresponding to classical k-means, i.e. Example 2.1, consider $\mathcal{X} = \mathbb{R}^d$. Let $\varphi_i = P_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$ be projections onto some of the coordinates. Then, $\tilde{x}_i = P_i(x_i)$ and (3.1) becomes

$$\min_{\substack{c_j \in \mathbb{R}^d \\ a \in [k]^N}} \sum_{i=1}^{N} \|\tilde{x}_i - P_i(c_{a_i})\|_2^2. \tag{3.8}$$

The cluster assignment iteration step assigns each point to the cluster whose barycenter is closest, considering only the known coordinates. The barycenter updating step is solved by

$$\Big(\sum_{i:a_i=j} P_i^T P_i \Big)^{-1} \sum_{i:a_i=j} P_i^T P_i(x_i),$$

which, in each coordinate, corresponds to the average of all the points in the cluster for which that coordinate is available. We detail in Section 5 how to deal with the case where the inverse does not exit. By imputing missing values as described above, we obtain measures in $\mathcal{P}_2(\mathbb{R}^d)$, of which we can then calculate pairwise distances using the metric ρ, defined in (3.7). Experiments using this method may be found in Section 7.1.

It is worth noting that [9] considers the same loss function, i.e. (3.8), but proposes a different algorithm for clustering Euclidean points with missing values, known as k-pod. In the Supplementary Material a comparison to their algorithm can be found when we view our method solely as clustering algorithm. Notably, our proposed method, NA k-means, consistently outperforms k-pod.

Before discussing our second example in Section 5, which generalizes Theorem 2.2, we first recall the necessary notions from optimal transport theory.

4 Wasserstein Distance and Generalized Wasserstein Barycenter: a Summary

Optimal Transport and Wasserstein Distance

Let μ, ν be probability measures on Polish spaces X, Y respectively. For a measurable map $T : X \to Y$ we use $_\#$ to denote the push-forward operator of measures (image measure), i.e. for a measurable set $B \subset Y$, we put $T_\#\mu(B) := \mu(T^{-1}(B))$. Denote by

$$\Pi(\mu, \nu) := \{\pi \in \mathcal{P}(X \times Y) : \mathrm{proj}_{X\#}\pi = \mu, \mathrm{proj}_{Y\#}\pi = \nu\}$$

the set of couplings of μ and ν, i.e. the set of all measures on the product space having μ and ν as marginals. The Kantorovich problem for a cost function $c : X \times Y \to \mathbb{R}_+$, introduced in [20], is

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X\times Y} c(x, y)\, \pi(\mathrm{d}x, \mathrm{d}y). \tag{KP}$$

We specialize to the case $X = Y = \mathbb{R}^d$ and $c(x, y) = \|x - y\|_p^p$ for $p \in [1, \infty)$. The Wasserstein distance W_p on the space $\mathcal{P}_p(\mathbb{R}^d)$ of probability measures with finite p-moment is then defined via

$$W_p(\mu, \nu)^p := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_p^p \, \pi(\mathrm{d}x, \mathrm{d}y). \tag{4.1}$$

Wasserstein Barycenter

A notion of averaging probability measures that has recently received significant attention is the concept of Wasserstein barycenters, introduced by Agueh and Carlier [1]. A Wasserstein barycenter of $\mu_1, \ldots, \mu_n \in \mathcal{P}_2(\mathbb{R}^d)$ with weights $\lambda_1, \ldots, \lambda_n \geq 0$ summing to 1, is a solution of

$$\inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i W_2^2(\mu_i, \nu).$$

Since we want to generalize the algorithm introduced in Sections 2-3 to the setting of probability measures, we require a variant of the Wasserstein barycenter that is still applicable when only some coordinates of the measures are known. A suitable concept, the *generalized Wasserstein barycenter*, was introduced by Julien, Gozlan and Saint-Dizier in [12]. To formally define it, let probability measures $\mu_1, \ldots, \mu_n \in \mathcal{P}_2(\mathbb{R}^d)$ be given along with linear maps $P_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$. In our intended applications, these maps correspond to projections onto some of the coordinates. A generalized Wasserstein barycenter of the push-forwarded measures $P_{1\#}\mu_1, \ldots, P_{n\#}\mu_n$ with associated weights $\lambda_1, \ldots, \lambda_n \geq 0$ summing to 1, is a solution of

$$\inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i W_2^2(P_{i\#}\mu_i, P_{i\#}\nu). \tag{4.2}$$

To solve the problem, it is useful to reformulate it as a classical Wasserstein barycenter problem. By associating each projection P_i with a matrix $P_i \in \mathbb{R}^{d_i \times d}$, we define $A := \sum_{i=1}^n \lambda_i P_i^T P_i$ and assume in the following that A is invertible. Next, we set $\bar{\mu}_i := (A^{-1/2} P_i^T)_{\#} \mu_i \in \mathcal{P}_2(\mathbb{R}^d)$ for $i \in [n]$, and consider the classical Wasserstein barycenter problem

$$\inf_{\bar{\nu} \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i W_2^2(\bar{\mu}_i, \bar{\nu}). \tag{4.3}$$

Then, ν is a solution to (4.2) if and only if $\bar{\mu} = A^{1/2}{}_{\#}\nu$ is a solution to (4.3), see Proposition 3.1 in [12]. Regarding computational aspects, especially in the discrete case, efficient algorithms for computing Wasserstein barycenters have already been developed; see e.g. [11], as well as [2], and are implemented in the Python package `POT` (see [13]).

5 NA WASSERSTEIN k-MEANS

Algorithm Description

We can now discuss the case of $\mathcal{X} = \mathcal{P}_2(\mathbb{R}^d)$ in detail, as we have established the two necessary operations — the Wasserstein distance and the generalized Wasserstein barycenter — for an algorithm as described in Section 3.

Suppose we want to cluster probability measures $\mu_1, \dots, \mu_N \in \mathcal{P}_2(\mathbb{R}^d)$ into k clusters but we only observe their push-forwards $P_{i\#}\mu_i \in \mathcal{P}_2(\mathbb{R}^{d_i})$ under projections $P_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$ for $i \in [N]$. Thus, in the language of Section 3, we have $\varphi_i(\cdot) := P_{i\#}(\cdot)$, and $\tilde{x}_i = P_{i\#}\mu_i =: \tilde{\mu}_i$. Problem (3.1) then reads as

$$\inf_{\substack{\nu_j \in \mathcal{P}_2(\mathbb{R}^d) \\ a \in [k]^N}} \sum_{i=1}^N W_2^2(\tilde{\mu}_i, P_{i\#}\nu_{a_i}).$$

We can tackle this problem by the iterations suggested in Section 3.1, which read as

i) update cluster assignments given barycenters ν_j, i.e. for each $i \in [N]$ set

$$a_i \in \operatorname*{argmin}_{j \in [k]} W_2(\tilde{\mu}_i, P_{i\#}\nu_j), \tag{5.1}$$

ii) update cluster barycenters given an assignment a, i.e. for each $j \in [k]$ set

$$\nu_j \in \operatorname*{argmin}_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i:a_i=j} W_2^2(\tilde{\mu}_i, P_{i\#}\nu).$$

Again, we aim to find an optimal assignment vector a that assigns each measure to one of the k clusters, as well as optimal barycenters, specifically generalized Wasserstein barycenters, $\nu_1, \dots, \nu_k$. Note that once again, we are looking for barycenters in the full space $\mathcal{P}_2(\mathbb{R}^d)$, i.e. having all coordinates.

It may happen that a cluster consists entirely of measures missing the same coordinate, i.e. for some $j \in [k]$, we have

$$\bigcap_{i:a_i=j} \ker P_i \neq \{0\}.$$

To avoid numerical problems in such cases, one can initialize the barycenters without missing coordinates and adapt step ii) to include the previous barycenter with a small weight. This ensures that all subsequent barycenters have no missing coordinates. Specifically, assume that at iteration t the barycenters are $\nu_j^{(t)} \in \mathcal{P}_2(\mathbb{R}^d)$ and a new assignment vector $a^{(t+1)} \in [k]^N$ has been computed. Given a weight $\lambda^{(t)} \in (0,1)$, set

$$\nu_j^{(t+1)} \in \operatorname*{argmin}_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} (1-\lambda^{(t)}) \sum_{i:a_i^{(t+1)}=j} W_2^2(\tilde{\mu}_i, P_{i\#}\nu) + \lambda^{(t)} W_2^2(\nu_j^{(t)}, \nu). \tag{5.2}$$

Experimentally we have observed that $\lambda^{(t)} := \frac{1}{(t+1)^{1/2}}$ works well. Having defined the two iterative steps, we now present the main computational contribution of this work: Algorithm 1. If no prior information is available, we propose initializing it as described in Section 3.1.

Algorithm 1 NA Wasserstein k-means

Input: N observed measures $\tilde{\mu}_i = P_{i\#}\mu_i$, number of clusters k, maximum number of iterations T, weighting schedule $(\lambda^{(t)})_{t=0}^T$
Result: barycenters ν_j, assignments a
Initialize $t \leftarrow 0$, $\nu_1^{(0)}, \dots, \nu_k^{(0)} \in \mathcal{P}_2(\mathbb{R}^d)$, and $a^{(0)} \in [k]^N$
while $t < T$ **do**
    for $i = 1$ to N **do**
        $a_i^{(t+1)} \leftarrow \operatorname*{argmin}_{j \in [k]} W_2^2\left(\tilde{\mu}_i, P_{i\#}\nu_j^{(t)}\right)$
    end for
    for $j = 1$ to k **do**
        choose $\nu_j^{(t+1)}$ according to (5.2) with weight $\lambda^{(t)}$
    end for
    if $a^{(t)} == a^{(t+1)}$ **then**
        break
    end if
    $t \leftarrow t + 1$
end while

In (6) we use the entry of the old assignment vector $a_i^{(t)}$ if it is a minimizer. This leads to a simple convergence result. For this sake let

$$L(\nu_1, \dots, \nu_K, a) := \sum_{i=1}^N W_2^2(P_{i\#}\mu_i, P_{i\#}\nu_{a_i}). \tag{5.3}$$

Theorem 5.1. *Given an initialization $\nu_1^{(0)}, \dots, \nu_k^{(0)} \in \mathcal{P}_2(\mathbb{R}^d)$, $a^{(0)} \in [k]^N$, Algorithm 1 strictly decreases L until it terminates after finitely many steps, for any choice of $(\lambda^{(t)})_{t\in\mathbb{N}}$ (even for $T = \infty$). If in every iteration each cluster has a measure with all coordinates, then Algorithm 1 with $\lambda^{(t)} = 0$ yields a local minimum of L.*

Remark 5.2 (Speed of convergence). It is known that the classical Euclidean k-means algorithm can require an exponential number of iterations in the worst-case. Specifically, there

exists a lower bound $2^{\Omega(N)}$ even in two dimensions, see [5, 35]. Since Algorithm 1 includes k-means as a special case (specifically, when applied to Dirac measures, i.e. Euclidean data, without missing values), this worst-case behavior also applies to our method. It cannot be worse than exponential, however, due to the trivial upper bound k^N of possible cluster assignments.

Nevertheless, in practice, k-means typically converges within a moderate number of iterations - often around 20 to 50 - when clustering a not-too-large number of objects, see [8, 15]. In our experiments, we observed a similar behavior for both NA Wasserstein k-means and NA k-means. On average, the number of iterations hovered around 30 iterations without exceeding 50. Specifically, in all experiments, we set the maximum number of iterations to 100 and this bound was never reached.

Proof of Theorem 5.1. Let us first show that the loss function decreases monotonically. Using the assignment step in the first and the barycenter updating step in the second inequality, we obtain

$$\begin{aligned}
(1-\lambda^{(t+1)})L(\nu_1^{(t)},\dots,\nu_k^{(t)},a^{(t)}) &= (1-\lambda^{(t+1)})\sum_{i=1}^N W_2^2(P_{i\#}\mu_i, P_{i\#}\nu^{(t)}_{a_i^{(t)}}) \\
&\overset{(\star)}{\geq} (1-\lambda^{(t+1)})\sum_{i=1}^N W_2^2(P_{i\#}\mu_i, P_{i\#}\nu^{(t)}_{a_i^{(t+1)}}) \\
= \sum_{j=1}^k \Big[(1-\lambda^{(t+1)}) \sum_{i:a_i^{(t+1)}=j} W_2^2(P_{i\#}\mu_i, P_{i\#}\nu_j^{(t)}) + \lambda^{(t+1)} \underbrace{W_2^2(\nu_j^{(t)},\nu_j^{(t)})}_{=0}\Big] \\
\geq \sum_{j=1}^k \Big[(1-\lambda^{(t+1)}) \sum_{i:a_i^{(t+1)}=j} W_2^2(P_{i\#}\mu_i, P_{i\#}\nu_j^{(t+1)}) + \lambda^{(t+1)} \underbrace{W_2^2(\nu_j^{(t)},\nu_j^{(t+1)})}_{\geq 0}\Big] \\
&\geq (1-\lambda^{(t+1)})\sum_{i=1}^N W_2^2(P_{i\#}\mu_i, P_{i\#}\nu^{(t+1)}_{a_i^{(t+1)}}) \\
&= (1-\lambda^{(t+1)})L(\nu_1^{(t+1)},\dots,\nu_k^{(t+1)},a^{(t+1)}).
\end{aligned}$$

Since $1-\lambda^{(t+1)} > 0$, this shows the monotonicity. In inequality $(\star)$, equality holds if and only if there is no change in the assignment step, i.e. if

$$a_i^{(t)} = a_i^{(t+1)},\ \forall i \in [N].$$

In this case the algorithm terminates. If there is some $i \in [N]$ such that $a_i^{(t)} \neq a_i^{(t+1)}$, then we have a strict inequality, since the assignment in Algorithm 1, cf. (6), is only updated when L decreases due to the new assignment.

Since only a finite number of possible assignments exist, due to the fact that we are clustering finitely many measures into finitely many clusters, the algorithm strictly decreases L and terminates after a finite number of steps, thus concluding the proof. □

Remark 5.3 (Convergence in abstract metric spaces). In the setting of abstract metric spaces, that is, Section 3.1, we can modify the assignment step as in Algorithm 1 – that

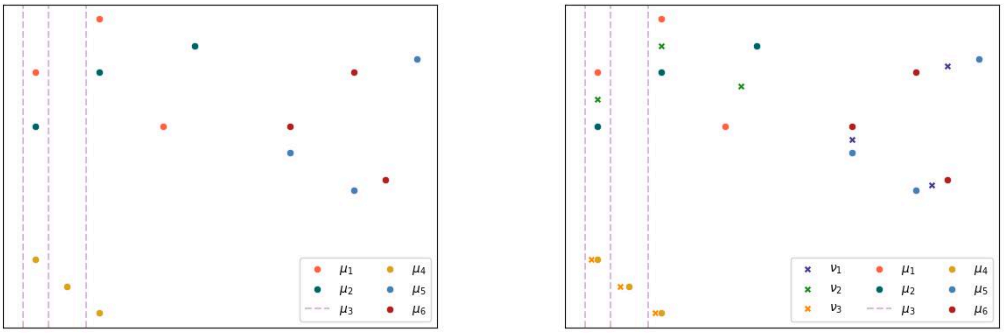

Figure 3: (left): 6 Measures on the Plane of Which One Misses the Vertical Coordinate of Its Points. (right) Clustering of the 6 Measures Into 3 Clusters.

is, assignments are only updated when the loss function strictly decreases. If we further assume that barycenters always exist, the convergence of the algorithm follows by the same arguments as in the previous proof.

A Toy Example

To illustrate the algorithm, consider six measures on the plane, one of which misses the vertical coordinate (indicated by vertical lines). Each measure consists of three support points with equal weight. The measures are visualized by different colors and depicted on the left of Figure 3. When clustering these measures into three clusters, the natural choice of clusters is evident. Applying the above algorithm indeed yields the results shown in Figure 3. To obtain these results, we calculated free support barycenters at each step of the algorithm with a fixed support size of three (cf. [3, 11]). For the blue and brown barycenters, each support point is precisely the average of the support points of the two measures in that cluster. For the pink barycenter the situation is slightly different. In the y coordinate it inherits the values from the red measure, while as in the x coordinate the support points are the averages of the x coordinates of a red and a rose support point, respectively.

Imputation via NA Wasserstein k-means

Next, we consider how the imputation procedure described in Section 3.2 looks, when $\mathcal{X} = \mathcal{P}_2(\mathbb{R}^d)$. Having clustered measures $P_{i\#}\mu_i \in \mathcal{P}_2(\mathbb{R}^{d_i})$, $i \in [N]$, we obtain generalized Wasserstein barycenters $\nu_1, \dots, \nu_k \in \mathcal{P}_2(\mathbb{R}^d)$ and a vector of assignments $a \in [k]^N$. For measures that are not fully observed, i.e. for $i \in I_m$ (see (3.4) for the definition), we will use the fully observed measures in the same cluster, i.e. μ_ℓ with $\ell \in J_i$, to impute $P_{i\#}\mu_i$ in a randomized fashion. Therefore, we have to interpret (3.5) in the current setting.

Specifically, for μ_ℓ with $\ell \in J_i$ we need to find

$$\eta_\ell^i \in \underset{\substack{\eta\in\mathcal{P}_2(\mathbb{R}^p)\\ P_{i\#}\eta=P_{i\#}\mu_i}}{\operatorname{argmin}} W_2^2(\eta,\mu_\ell). \tag{5.4}$$

To obtain such η_ℓ^i we will apply the following lemma.

Lemma 5.4. *Let $\mu_1 \in \mathcal{P}_2(\mathbb{R}^k), \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with $d > k$. Denote by ν_1 the projection on the first k coordinates and by $\pi_1^* \in \Pi(\mu_1,\nu_1)$ the optimal transport coupling between μ_1 and ν_1. Then, the measure*

$$\eta^*(\mathrm{d}x_1,\mathrm{d}x_2) := \int_{Y_1} \nu_{y_1}(\mathrm{d}x_2)\,\pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1), \tag{5.5}$$

solves

$$\min_{\substack{\eta\in\mathcal{P}_2(\mathbb{R}^d)\\ \eta_1=\mu_1}} W_2^2(\eta,\nu) = W_2^2(\mu_1,\nu_1). \tag{5.6}$$

Proof. Let us first note that for any $\eta \in \mathcal{P}_2(\mathbb{R}^d)$ with $\eta_1 = \mu_1$ there is the trivial bound

$$W_2^2(\eta,\nu) \geq W_2^2(\eta_1,\nu_1) = W_2^2(\mu_1,\nu_1). \tag{5.7}$$

We will show that η^*, as defined in (5.5) achieves this lower bound, which will imply its optimality. In order to do this, we define a coupling $\pi \in \Pi(\eta^*,\nu)$ having those costs. In the following we split a point $z \in \mathbb{R}^d$ like $z = (z_1, z_2) \in \mathbb{R}^k \times \mathbb{R}^{d-k}$ and use the notations X_1, X_2, Y_1, Y_2 to emphasize over which space we are integrating, even though $X_1 = Y_1 = \mathbb{R}^k$ and $X_2 = Y_2 = \mathbb{R}^{d-k}$. Let us define π by

$$\pi(\mathrm{d}x_1,\mathrm{d}y_1,\mathrm{d}x_2,\mathrm{d}y_2) := \pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1)\nu_{y_1}(\mathrm{d}y_2)\delta_{y_2}(\mathrm{d}x_2). \tag{5.8}$$

To see that this is actually a coupling of (η^*,ν), note that for the second marginal, we have

$$\begin{aligned}\int_{X_1\times X_2} \pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1)\,\nu_{y_1}(\mathrm{d}y_2)\,\delta_{y_2}(\mathrm{d}x_2) &= \int_{X_1}\int_{X_2} \delta_{y_2}(\mathrm{d}x_2)\,\pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1)\,\nu_{y_1}(\mathrm{d}y_2)\\ &= \int_{X_1} \pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1)\,\nu_{y_1}(\mathrm{d}y_2)\\ &= \nu_1(\mathrm{d}y_1)\nu_{y_1}(\mathrm{d}y_2) = \nu(\mathrm{d}y_1,\mathrm{d}y_2).\end{aligned}$$

For the first marginal, we have

$$\begin{aligned}\int_{Y_1\times Y_2} \pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1)\,\nu_{y_1}(\mathrm{d}y_2)\,\delta_{y_2}(\mathrm{d}x_2) &= \int_{Y_1}\int_{Y_2} \delta_{y_2}(\mathrm{d}x_2)\,\nu_{y_1}(\mathrm{d}y_2)\,\pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1)\\ &= \int_{Y_1} \nu_{y_1}(\mathrm{d}x_2)\,\pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1) = \eta^*(\mathrm{d}x_1,\mathrm{d}x_2).\end{aligned}$$

Next, let us calculate the coupling's cost,

$$\begin{aligned}\int_{X\times Y} |x-y|^2\,\pi(\mathrm{d}x,\mathrm{d}y) &= \int_{X_1\times Y_1} |x_1-y_1|^2\,\pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1) + \int_{X\times Y} |x_2-y_2|^2\,\pi(\mathrm{d}x,\mathrm{d}y)\\ &= W_2^2(\mu_1,\nu_1) + \int_{X\times Y} |x_2-y_2|^2\,\pi(\mathrm{d}x,\mathrm{d}y),\end{aligned}$$

and note for the second term,

$$\begin{aligned}\int_{X\times Y}|x_2-y_2|^2\,\pi(\mathrm{d}x,\mathrm{d}y) &= \int_{X_1\times Y_1}\int_{Y_2}\int_{X_2}|x_2-y_2|^2\,\delta_{y_2}(\mathrm{d}x_2)\,\nu_{y_1}(\mathrm{d}y_2)\,\pi_1^*(\mathrm{d}x_1,\mathrm{d}y_1)\\ &= 0.\end{aligned}$$

Therefore, $W_2(\eta^*,\nu)=W_2(\mu_1,\nu_1)$, showing that η^* solves (5.6). □

Remark 5.5. If π_1^* in Theorem 5.4 is of Monge-type, i.e. if there is a measurable map $T_1:\mathbb{R}^k\to\mathbb{R}^k$ such that $\pi_1^*=(\mathrm{id}_{\mathbb{R}^k},T_1)_{\#}\mu_1$, then (5.5) simplifies to

$$\eta^*(\mathrm{d}x_1,\mathrm{d}x_2)=\mu_1(\mathrm{d}x_1)\nu_{T_1(x_1)}(\mathrm{d}x_2). \tag{5.9}$$

Lemma 5.4 describes the form of η_ℓ^i in (5.4). To apply it, let $\pi_\ell^i\in\Pi(P_{i\#}\mu_i,P_{i\#}\mu_\ell)$ be the optimal coupling between $P_{i\#}\mu_i$ and $P_{i\#}\mu_\ell$. Additionally, let $\mathcal{I}_i\subset[d]$ denote the indices of the projection P_i, i.e. $P_i(x_1,\dots,x_d)=(x_j)_{j\in\mathcal{I}_i}$, and for the remaining indices define $\mathcal{I}_i^C:=[d]\setminus\mathcal{I}_i$. A solution to (5.4) then is

$$\eta_\ell^i(\mathrm{d}x_1,\dots,\mathrm{d}x_d):=\int_{\mathbb{R}^{|\mathcal{I}_i^C|}}\mu_{\ell(y_j)_{j\in\mathcal{I}_i}}((\mathrm{d}x_j)_{j\in\mathcal{I}_i^C})\,\pi_\ell^i((\mathrm{d}x_j,\mathrm{d}y_j)_{j\in\mathcal{I}_i}). \tag{5.10}$$

Having this, we can translate (3.6) to the current setting by using η_ℓ^i with $\ell\in J_i$, and define

$$\mathbb{P}_i:=\sum_{\ell\in J_i}p_\ell^i\boldsymbol{\delta}_{\eta_\ell^i} \tag{5.11}$$

to obtain a measure $\mathbb{P}_i\in\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$, i.e. a random measure. For fully observed measures, i.e. $i\in I_f$ we set $\mathbb{P}_i:=\boldsymbol{\delta}_{\mu_i}$. For $i\in I_m$ and $|J_i|>1$ we suggest the weights

$$p_\ell^i\propto\exp\big(-\frac{\lambda}{2\sigma_i^2}W_2^2(P_{i\#}\mu_i,P_{i\#}\mu_\ell)\big),\quad \ell\in J_i.$$

This reflects the idea that measures which are closer to the to be imputed measure in the known coordinates should receive more weight. Here, $\lambda>0$ is a tuning parameter that controls this weighting and the variance σ_i^2 is used to standardize the distances, i.e.

$$\sigma_i^2:=\frac{1}{|J_i|-1}\sum_{\ell\in J_i}W_2^2(P_{i\#}\mu_i,P_{i\#}\mu_\ell).$$

If $|J_i|=1$ it is clear what to do, as there is only one measure available for imputation and for $|J_i|=0$ we use the barycenter ν_{a_i} to impute $P_{i\#}\mu_i$.

After imputation we obtain random measures $\mathbb{P}_1,\dots,\mathbb{P}_N\in\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$. To calculate their pairwise distances, we use the metric ρ defined in (3.7). That is, for $\mathbb{P},\mathbb{Q}\in\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$, $\rho(\mathbb{P},\mathbb{Q})=0$, if $\mathbb{P}=\mathbb{Q}$, and for $\mathbb{P}\neq\mathbb{Q}$

$$\rho(\mathbb{P},\mathbb{Q})=\int_{\mathcal{P}_2(\mathbb{R}^d)}\int_{\mathcal{P}_2(\mathbb{R}^d)}W_2(\mu,\nu)\,\mathbb{P}(\mathrm{d}\mu)\,\mathbb{Q}(\mathrm{d}\nu). \tag{5.12}$$

6 Evaluation Method

Our main motivation for NA Wasserstein k-means was to develop a method for clustering and reconstructing measures that may have missing coordinates. To the best of our knowledge, this is the first method to address this problem, so we cannot compare it to existing methods. In this short section we propose an abstract methodology for evaluating the quality of a reconstruction method in the setting of metric (measure) spaces, based on the concept of Gromov-Wasserstein distance. As just explained this evaluation methodology will not be applied directly to our main problem of interest, since there is no alternative method to compare to. However, we will apply it to a related problem in Section 7.1, where the data consists of Euclidean points rather than measures. For the Euclidean case we can compare NA k-means, a special case of NA Wasserstein k-means, to existing alternative methods.

Suppose there are elements $x_1, \dots, x_N$ in a metric space $(\mathcal{X}, d_{\mathcal{X}})$, which we collect in the set $X := \{x_1, \dots, x_N\}$. The data we observe are these points but only via the maps $\varphi_i : \mathcal{X} \to \mathcal{X}_i$ defined in Section 3, i.e. we observe $\tilde{x}_i = \varphi_i(x_i)$. The observed points are collected in $(\tilde{x}_1, \dots, \tilde{x}_N) \in \mathcal{X}_{NA}$, where we define $\mathcal{X}_{NA} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$. This can be summarized by saying that we only observe the values of $(x_1, \dots, x_N)$ under the map $h : \mathcal{X}^N \to \mathcal{X}_{NA}$, in particular $h_i(x_1, \dots, x_N) = \varphi_i(x_i) = \tilde{x}_i$.

The goal is to reconstruct the metric structure of $X = \{x_1, \dots, x_N\}$, which is described by their respective pairwise distances, as accurately as possible. This means that we want to find a *good* metric space $(\mathcal{Y}, d_{\mathcal{Y}})$ and a reconstruction map $R : \mathcal{X}_{NA} \to \mathcal{Y}^N$. The reconstructed point corresponding to x_i then is $R_i(\tilde{x}_1, \dots, \tilde{x}_N) = y_i$.

Before establishing a way of comparing the reconstructed points $R((\tilde{x}_1, \dots, \tilde{x}_N)) = (y_1, \dots, y_N)$ to the original points X, we want to consider another piece of information, namely that some observations of X might be more important for us than others. This is motivated by our main intended application of reconstructing a banking landscape, in which a bank's importance may be linked to attributes such as its size, measured, for example, by its total assets. This follows the rationale that for a regulator it may be more important to reconstruct the characteristics of a very large bank than those of a smaller bank. A simple way to account for this is to assign weights proportional to the (or a function of) size of the bank. Formally, we define a probability measure μ_X on X by setting $\mu_X(\{x_i\}) := p_i \geq 0$, where we assume $\sum_{i=1}^N p_i = 1$. Naturally we define μ_Y, a probability measure on $\mathcal{Y}$, via $\mu_Y(\{y_i\}) = \mu_Y(R_i(\tilde{x}_1, \dots, \tilde{x}_N)) := p_i$.

Mémoli [24], and the related work of Sturm [32], introduced the *Gromov-Wasserstein distance* GW_2 in order to compare metric measure spaces. For two metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$, it is defined as

$$GW_2((\mathcal{X}, d_{\mathcal{X}}, \mu), (\mathcal{Y}, d_{\mathcal{Y}}, \nu))^2 := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2 \times \mathcal{Y}^2} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^2 \, \pi(\mathrm{d}x, \mathrm{d}y) \, \pi(\mathrm{d}x', \mathrm{d}y').$$

We will use the Gromov-Wasserstein distance to evaluate a reconstruction, i.e. we will use

$$GW_2\big((\mathcal{X}, d_{\mathcal{X}}, \mu_X), (\{R_1(\tilde{x}_1, \dots, \tilde{x}_N), \dots, R_N(\tilde{x}_1, \dots, \tilde{x}_N)\}, d_{\mathcal{Y}}, \nu_Y)\big) \tag{6.1}$$

as performance measure. Note, that in the case of reconstructing as described in Section 3.2, we have $\mathcal{Y} = \mathcal{P}(\mathcal{X})$. For algorithms to compute the Gromov-Wasserstein distance we refer to [24] and [28], which are implemented in Python optimal transport library `POT` (see [13]). We will use this implementation in Section 7.1. The curious reader is also referred to [4, 28, 34] for some machine learning applications of Gromov-Wasserstein distances and related metrics.

7 Experimental Results

7.1 Reconstructing Points from a Gaussian Mixture Model

We start by comparing our clustering / reconstruction method with existing imputation methods in the Euclidean case, i.e. $\mathcal{X} = \mathbb{R}^d$, and evaluate the results using the Gromov-Wasserstein distance, cf. (6.1), introduced in Section 6.

We simulate data from a Gaussian Mixture Model

$$\gamma := \sum_{j=1}^{k} \alpha_j \mathcal{N}(\mu_j, \Sigma_j), \tag{7.1}$$

with weights $\alpha_1, \ldots, \alpha_k \geq 0$ and $\sum_{j=1}^{k} \alpha_j = 1$. Additionally, to simulate the importance of points, i.e. $\mu_X(\{x_i\})$ from Section 6, we draw samples from a Lognormal(μ, σ^2) distribution and assign each observation a weight proportional to its sampled value. In our simulation study we fix the parameters $k = 5$, $d = 5$, and $N = 500$, meaning that we always sample 500 points. To simulate missing values we employ various missingness structures, corresponding to different choices of the map h from Section 6. Little and Rubin [21] classify missing data mechanisms into the following three categories:

- *MCAR* (missing completely at random): The probability of a missing value does not depend on any observed or unobserved values.
- *MAR* (missing at random): The probability of a missing value depends only on observed data, not on the missing data itself.
- *MNAR* (missing not at random): The probability of a missing value depends on the missing values themselves, even when accounting for observed data.

In order to achieve robust results we incorporate different combinations of all these missingness mechanisms in our simulation study, i.e. in our definition of h from Section 6.

Before describing the method used to create missing values, we first explain how the parameters for the data-generating process are chosen, i.e. $\alpha_1, \ldots, \alpha_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k$ for the Gaussian Mixture Model, and μ, σ^2 for the lognormal distribution governing the importance of sampled points. We set $\mu = 20, \sigma = 1.5$, as these parameters (empirically) model the total assets of a financial institution reasonably well. For the Gaussian Mixture Model, we aim to simulate the weights of the normal distributions, means and covariance matrices in a non-informative manner. Specifically, the weights α_j are sampled uniformly from the probability simplex, the means μ_j's are sampled uniformly from the cube $[-5, 5]^d$

and the covariance matrices Σ_j's are drawn from a Wishart distribution with parameters $(d, \mathrm{Id}_d/d)$. Note that with this choice of parameters for the Wishart distribution, we have $\mathbb{E}[\Sigma_j] = \mathrm{Id}_d$.

In order to create missing data, we aim to combine the types of missingness mechanisms described above. To achieve this, we choose weights $\beta^{MCAR}, \beta^{MAR}, \beta^{MNAR} \geq 0$ with $\beta^{MCAR} + \beta^{MAR} + \beta^{MNAR} = 1$. We then fix a proportion $p \in (0, 1)$ which determines the total fraction of missing values. Given $N = 500$ points sampled independently from γ, we compute for each dimension the $p\beta^{MNAR}$ quantile and proceed to create missing values as follows:

i) *MCAR*: set $100p\beta^{MCAR}\%$ values missing completely at random,

ii) *MNAR*: for each dimension set all values to missing which are below the corresponding $p\beta^{MNAR}$ quantile,

iii) *MAR*: for points where the first coordinate is observed and non-negative, the probability of missing values in the other coordinates is four times higher than for points where the first coordinate is observed and negative. In total, we create $100p\beta^{MAR}\%$ of missing values in the data following this rule.

This procedure defines the map h from Section 6. To generate the results in Table 1, we set $p = 0.15$. For $\beta^{MCAR}, \beta^{MAR}$, and β^{MNAR}, we consider seven different combinations by applying each type of missingness individually, in pairs, and all three types together.

For each choice of $\beta^{MCAR}, \beta^{MAR}, \beta^{MNAR}$, we sample the described parameters of the Gaussian Mixture Model 100 times. We generate data, and compute pairwise distances between data points. Then, we create missing values according to h and apply different imputation procedures presented below. If missing values are imputed by points, we compute pairwise distances between the imputed points. If the imputation is done via measures, as abstractly defined in (3.6), we use the distance ρ from Section 3.2 to compute pairwise distances. Finally, we calculate the corresponding Gromov-Wasserstein distance between original pairwise distances and the imputed pairwise distances, using the corresponding weights derived from the Lognormal(μ, σ^2) samples, i.e. we compute (6.1). In our experiment we look at the following imputation techniques:

- NA k-means: Our method in the case of Euclidean data, as outlined in Example 3.2.
- NA k-means-m: this corresponds to first applying NA k-means for clustering and imputation. After obtaining the measures $(\theta_i)_{i=1}^N$, we compute their expected values and use them as the imputed points.
- Mean imputation: Each missing value is replaced by the mean of the corresponding attribute.
- Median imputation: Each missing value is replaced by the median of the corresponding attribute.
- Multiple imputation: A method [7, 30] that imputes missing values by generating multiple possible values and combines results.

Table 1: Gromov Wasserstein Distance for Euclidean Simulation With $p = 0.15$.

β^{MCAR}	β^{MAR}	β^{MNAR}	**NA k-means.**	**NA k-means-m**	**mean imp.**	**median imp.**	**multiple imp.**	**KNN**	**LR**
1	0	0	0.525 ± 0.016	0.549± 0.021	1.28 ± 0.024	1.28 ± 0.024	0.724 ± 0.02	0.643 ± 0.014	0.734 ± 0.02
0	1	0	0.517 ± 0.012	0.574± 0.03	1.296 ± 0.034	1.296 ± 0.034	0.791 ± 0.029	0.619 ± 0.021	0.802 ± 0.033
0	0	1	1.534 ± 0.048	1.592± 0.052	2.316 ± 0.053	2.316 ± 0.053	1.719 ± 0.051	1.792 ± 0.047	1.743 ± 0.053
0.5	0.5	0	0.51 ± 0.017	0.513± 0.016	1.286 ± 0.029	1.286 ± 0.029	0.703 ± 0.019	0.627 ± 0.019	0.709 ± 0.021
0	0.5	0.5	1.035 ± 0.036	1.12± 0.04	1.86 ± 0.033	1.86 ± 0.033	1.306 ± 0.039	1.308 ± 0.034	1.316 ± 0.042
0.5	0	0.5	1.012 ± 0.03	1.106± 0.039	1.883 ± 0.033	1.883 ± 0.033	1.311 ± 0.038	1.347 ± 0.04	1.335 ± 0.038
0.333	0.333	0.333	0.872 ± 0.035	0.918± 0.037	1.702 ± 0.028	1.702 ± 0.028	1.126 ± 0.028	1.125 ± 0.028	1.145 ± 0.033

- KNN: K-nearest-neighbor imputation, where each missing value is replaced by a weighted average of points that are close in the coordinates which are not missing. We choose $K = 4$.

- LR: Missing values are imputed by regressing on observed values and using the predicted values as imputations.

For all methods not introduced in this article, i.e. all except NA k-means and NA k-means-m, we use the implementations from `scikit-learn`, [26].

In Table 1, we report the estimated Gromov Wasserstein Distances along with their standard errors, based on sampling 100 times for each combination of β^{MCAR}, β^{MAR} and β^{MNAR}.

We observe that, in the case of Euclidean points with missing values, the proposed method consistently outperforms classical imputation techniques when considering the Gromov-Wasserstein distance as evaluation measure. Specifically, whether we apply *NA k-means* directly or include an additional averaging step, i.e. *NA k-means-m*, both algorithms outperform the remaining five methods across all considered missing data scenarios, i.e. combinations of MCAR, MAR and MNAR. In the Supplementary Material we use the (adjusted) Rand index [18, 29] as evaluation method when treating NA k-means purely as a clustering algorithm, rather than as a method to reconstruct the metric structure of objects, i.e. as imputation method. There we also compare it to k-pod which is consistently outperformed by NA k-means (see Table 3). Moreover, we provide results for additional values of p, i.e. the total share of missing values in the observed data, to accompany and robustify our results.

7.2 Reconstructing Financial Institutions

As described in the introduction, the primary motivation for developing NA Wasserstein k-means comes from clustering financial institutions based on granular loan data they are obliged to report to the central bank. This problem is of particular interest to regulators, as it enables a data-driven assessment of similarities and differences between financial institutions. If prior beliefs about similarities exist, this method provides a way to confirm or challenge them. In particular, if financial institutions have already been grouped based on prior knowledge, our approach can serve as a validation tool, highlighting institutions whose cluster assignment deviates from the expected grouping and may therefore warrant further investigation.

Before presenting an experiment of applying NA Wasserstein k-means, note that there is another experiment justifying the usage of NA Wasserstein k-means in Section 7.3.

To demonstrate NA Wasserstein k-means in a practical example, we use real data from Oesterreichische Nationalbank, the central bank of Austria, and analyze loan data reported by 321 financial institutions in Austria. In total, our dataset consists of 129230 loans. Each loan is described by up to four attributes: interest rate, interest rate margin, probability of default and par value. In our analysis we apply a logarithmic transformation to par value and then standardize all four attributes across all loans. Additionally, we account for loan size by weighting each loan within a financial institution proportionally to the logarithm of its size. Among the 321 institutions, 265 reported all four attributes, 45 institutions reported three and 11 institutions reported only two. As a result, our observed data consists of 265 probability measures in $\mathcal{P}_2(\mathbb{R}^4)$, 45 probability measures in $\mathcal{P}_2(\mathbb{R}^3)$ and 11 probability measures in $\mathcal{P}_2(\mathbb{R}^2)$. Ideally, the complete dataset would consist of 321 probability measures in $\mathcal{P}_2(\mathbb{R}^4)$ but due to differences in reporting, missing values must be accounted for in practice. While meaningful imputation is already challenging for Euclidean data, it is even more difficult when dealing with measure-valued data.

Since prior imputation is not necessary for NA Wasserstein k-means, we can apply it directly as described in Section 5 and cluster the financial institutions into seven groups. After clustering, we use the imputation method described in the same section to obtain random measures $\mathbb{P}_1, \dots, \mathbb{P}_{321}$ on $\mathbb{R}^4$, cf. Equation (5.4). Using the generalized distance ρ on $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^4))$, defined in (5.12), we compute pairwise distances between the (imputed) financial institutions. Once the pairwise distances are computed, it is possible to visualize the reconstructed financial institutions and clustering results. To create such a banking landscape, dimensionality reduction techniques may be applied to approximate pairwise distances in a lower-dimensional space. Here, we choose *Isomap* (cf. [33]) over other methods, such as multidimensional scaling, as it is better suited to recover nonlinear manifolds. Applying Isomap to the pairwise distance matrix $(\rho(\mathbb{P}_i, \mathbb{P}_j))_{i,j=1}^{321}$ maps the financial institutions to $\mathbb{R}^3$ while preserving the overall distance structure as closely as possible.

Figure 2 presents the result of this three-dimensional representation, i.e. a banking landscape with the corresponding clusters. Each point in the figure represents one of the 321 financial institutions, and each color describes a cluster. A regulatory analyst can now use this three-dimensional representation to reassess prior beliefs about the banking landscape or existing groupings of financial institutions and to identify institutions of particular interest. Additionally to prior beliefs, the visualization may help analysts to further their understanding which financial institutions are similar to each other and also identify financial institutions which behave significantly different to others with respect to their loan structure. Detecting potential outliers is of particular interest.

We also note that, having imputed the measures, i.e. "reconstructed" financial institutions, in particular having computed their pairwise distances $(\rho(\mathbb{P}_i, \mathbb{P}_j))_{i,j=1}^{321}$, any distance-based learning algorithm can be applied. Specifically, distance-based outlier algorithms can be used to identify institutions that deviate significantly from the rest.

To conclude this experiment, we emphasize that clustering probability measures directly using NA Wasserstein k-means can yield significantly different results compared

to clustering their aggregated data, such as their expectations, with methods like NA k-means. To illustrate this, we computed the expected values of the 321 probability measures considered in this section and applied NA k-means to the resulting points (with possibly missing values) in $\mathbb{R}^4$. Using the obtained cluster labels, we then computed generalized Wasserstein barycenters for these clusters. Notably, when these assignments and barycenters were used in the loss function Equation (5.3), the resulting value was 22.26% higher than when clustering the probability measures directly using Algorithm 1. Furthermore, comparing the labels of the two clusterings using the adjusted Rand index[1] yielded a value of 0.2058, indicating very poor agreement, as a value of 0 would correspond to the expected agreement of a random assignment. This highlights the advantage of clustering in the space of probability measures rather than relying on simple data aggregation.

7.3 Justifying NA Wasserstein k-means

To justify the use of NA Wasserstein k-means, we carry out another experiment. It is similar to the one in Section 7.1, where we artificially create missing values, but this time using *distributional data*. To the best of our knowledge, no existing method can cluster probability measures that are only observed as push-forwards under projections. Therefore, we perform an experiment on the real loan data from the previous section, simulating missing values completely at random.

Specifically, we consider 100 financial institutions from the previous section that reported all four attributes of their credits, meaning there were no missing values. For each of these institutions we sample 100 credits to reduce the support size and, consequently, the computational cost, as NA Wasserstein k-means will be applied multiple times to obtain reliable standard errors. The resulting 100 probability measures on $\mathbb{R}^4$ contain no missing values, allowing us to cluster them into five groups using Wasserstein k-means. The assignments from this clustering serve as the "ground truth" for this experiment.

Next, we create missing values completely at random: we choose a percentage $p \in (0, 1)$ and a number of affected dimensions $d_m \in \{1, 2\}$. Then, we randomly select $100p$ institutions and for each selected institution, we randomly choose d_m of the 4 dimensions and set them to NA. After simulating missing values in this way, we apply NA Wasserstein k-means and compare the resulting cluster assignments to the ground truth using the adjusted Rand index.[1]

For each combination of p and d_m, the simulation of missing values with subsequent clustering is repeated 100 times. The means and standard errors of the resulting adjusted Rand indices are presented in Table 2. They indicate that in most settings, the "ground truth" labels are well recovered. As expected, the adjusted Rand index decreases as p and d_m increase, i.e. as the proportion of missing data grows. Notably, when only one coordinate is missing for up to a quarter of the data, the adjusted Rand index remains

[1]The Rand index [29] is a similarity measure that evaluates the agreement between two clusterings by considering agreement or non-agreement of all pairwise assignments. It takes the value 1 for perfect agreement and 0 for complete disagreement. The adjusted Rand index [18] corrects the Rand index for chance by adjusting for the expected similarity under a random model. It ranges from -0.5 to 1, where 1 indicates perfect agreement, 0 corresponds to random labeling, and negative values suggest less agreement than expected by chance.

above 0.87. Even when two coordinates are missing, it only drops slightly below 0.8 if more than 20% of the measures are affected. We conclude that this experiment demonstrates the effectiveness of NA Wasserstein k-means as a clustering algorithm for probability measures that are only partially observed as push-forwards under projections.

Table 2: Adjusted Rand Index for Granular Data

	$p = 0.05$	$p = 0.1$	$p = 0.15$	$p = 0.2$	$p = 0.25$
$d_m = 1$	0.942 ± 0.006	0.919 ± 0.006	0.891 ± 0.006	0.87 ± 0.005	0.871 ± 0.006
$d_m = 2$	0.911 ± 0.006	0.881 ± 0.006	0.838 ± 0.007	0.821 ± 0.007	0.786 ± 0.007

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[2] P. C. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications* 441.2 (2016), pp. 744–762.

[3] P. C. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications* 441.2 (2016), pp. 744–762.

[4] D. Alvarez-Melis and T. S. Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013* (2018).

[5] D. Arthur and S. Vassilvitskii. How slow is the k-means method? *Proceedings of the Twenty-Second Annual Symposium on Computational Geometry*. SCG '06. Sedona, Arizona, USA: Association for Computing Machinery, 2006, pp. 144–153. ISBN: 1595933409.

[6] D. Arthur and S. Vassilvitskii. K-Means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. ISBN: 9780898716245.

[7] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research* 20.1 (2011), pp. 40–49.

[8] A. Broder, L. Garcia-Pueyo, V. Josifovski, S. Vassilvitskii, and S. Venkatesan. Scalable K-Means by ranked retrieval. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. WSDM '14. New York, New York, USA: Association for Computing Machinery, 2014, pp. 233–242. ISBN: 9781450323512.

[9] J. T. Chi, E. C. Chi, and R. G. Baraniuk. k-POD: A Method for k-Means Clustering of Missing Data. *The American Statistician* 70.1 (Jan. 2016), pp. 91–99.

[10] M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc., 2013.

[11] M. Cuturi and A. Doucet. Fast Computation of Wasserstein Barycenters. *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, June 2014, pp. 685–693.

[12] J. Delon, N. Gozlan, and A. S. Dizier. Generalized Wasserstein barycenters between probability measures living on different subspaces. English. *Ann. Appl. Probab.* 33.6A (2023), pp. 4395–4423.

[13] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8.

[14] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. fr. *Annales de l'institut Henri Poincaré* 10.4 (1948), pp. 215–310.

[15] S. Har-Peled and B. Sadri. How fast is the k-means method? English. *Algorithmica* 41.3 (2005), pp. 185–202.

[16] N. Ho, X. L. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel Clustering via Wasserstein Means. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1501–1509.

[17] B. Horvath, Z. Issa, and A. Muguruza. Clustering Market Regimes using the Wasserstein Distance. *arXiv:2110.11848* (2021).

[18] L. Hubert and P. Arabie. Comparing partitions. English. *Journal of Classification* 2 (1985), pp. 193–218.

[19] V. Huynh, N. Ho, N. Dam, X. Nguyen, M. Yurochkin, H. Bui, and D. Phung. On Efficient Multilevel Clustering via Wasserstein Distances. *Journal of Machine Learning Research* (2021).

[20] L. Kantorovich. On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)* 37 (1942), pp. 199–201.

[21] R. J. Little and D. B. Rubin. Statistical analysis with missing data. John Wiley & Sons, 2019.

[22] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137.

[23] J. MacQueen. Some methods for classification and analysis of multivariate observations. English. *Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297.* (1967).

[24] F. Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* 11 (2011), pp. 417–487.

[25] G. I. Papayiannis, G. N. Domazakis, D. Drivaliaris, S. Koukoulas, A. E. Tsekrekos, and A. N. Yannacopoulos. On clustering uncertain and structured data with Wasserstein barycenters and a geodesic criterion for the number of clusters. *Journal of Statistical Computation and Simulation* 91.13 (Mar. 2021), pp. 2569–2594.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[27] G. Peyré and M. Cuturi. Computational optimal transport. With applications to data sciences. English. *Found. Trends Mach. Learn.* 11.5-6 (2018), pp. 1–262.

[28] G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. *Proc. 33rd International Conference on Machine Learning.* Proc. 33rd International Conference on Machine Learning. New-York, United States, June 2016.

[29] W. M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* 66 (1971), pp. 846–850.

[30] D. B. Rubin. Multiple imputation for nonresponse in surveys. English. Reprint of the 1987 original. Hoboken, NJ: John Wiley & Sons, 2004. ISBN: 0-471-65574-0.

[31] M. Staib and S. Jegelka. Wasserstein k-means++ for Cloud Regime Histogram Clustering. *Proceedings of the Seventh International Workshop on Climate Informatics: CI 2017.* 2017.

[32] K.-T. Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *to appear in Memoirs AMS* (2020).

[33] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290.5500 (2000), pp. 2319–2323.

[34] V. Titouan, N. Courty, R. Tavenard, and R. Flamary. Optimal transport for structured data with application on graphs. *International Conference on Machine Learning.* PMLR. 2019, pp. 6275–6284.

[35] A. Vattani. k-means requires exponentially many iterations even in the plane. English. *Discrete Comput. Geom.* 45.4 (2011), pp. 596–616.

[36] I. Verdinelli and L. Wasserman. Hybrid Wasserstein distance and fast distribution clustering. *Electronic Journal of Statistics* 13.2 (2019), pp. 5088–5119.

[37] Y. Zhuang, X. Chen, and Y. Yang. Wasserstein K-means for clustering probability distributions. *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho. 2022.

A Existence of "barycenters"

In this section we discuss the existence of minimizers in the general barycenter updating step (3.3) of Section 3.1. Thus, let us consider a metric space $(\mathcal{X}, d)$ and finitely many continuous maps into other metric spaces $\varphi_i : \mathcal{X} \to \mathcal{X}_i$. We observe the image points $\tilde{x}_i = \varphi_i(x_i)$ for $i \in [n]$. The question is: under which conditions is

$$\inf_{y \in \mathcal{X}} \sum_{i=1}^{n} \lambda_i d(\varphi_i^{-1}(\tilde{x}_i), y)^2 \tag{A.1}$$

attained? Here, $\lambda_i > 0, \sum_{i=1}^n \lambda_i = 1$ denote some convex weights. Since in this general formulation the maps φ_i can be arbitrary, minimizers do not have to exist. However, if one of the maps φ_i is the identity on $\mathcal{X}$, we can guarantee existence under mild assumptions.

Lemma A.1. *Assume there exists a metrizable topology τ on $\mathcal{X}$ which is weaker than the topology induced by the metric d, and such that*

i) $\varphi_i^{-1}(\tilde{x}_i)$ is closed in τ for all $i \in [n]$,

ii) balls, i.e. $B_d(y, c) := \{x \in \mathcal{X} : d(x, y) \leq c\}$, are τ compact for all $y \in \mathcal{X}, c > 0$,

iii) $d(\cdot, \cdot)$ is lower semi continuous w.r.t. the product topology $\tau \times \tau$,

iv) there exists $j \in [n]$ such that $\varphi_j = Id_{\mathcal{X}}$.

Then (A.1) *admits a minimizer.*

Proof. Let us first prove that for $d(\varphi_i^{-1}(\tilde{x}_i), y)$ there exists $z \in \varphi_i^{-1}(\tilde{x}_i)$ such that

$$d(\varphi_i^{-1}(\tilde{x}_i), y) = \inf_{x \in \varphi_i^{-1}(\tilde{x}_i)} d(x, y) = d(z, y). \tag{A.2}$$

To this end, take a sequence $\{x_n\}_{n \geq 1} \subset \varphi_i^{-1}(\tilde{x}_i)$ such that

$$d(x_n, y) \searrow \inf_{x \in \varphi_i^{-1}(\tilde{x}_i)} d(x, y).$$

We then set $c := d(x_1, y)$, so $\{x_n\}_{n \geq 1} \subset B_d(y, c)$. Due to τ-compactness of $B_d(y, c)$ we obtain a subsequence $(x_{n_k})_{k \geq 1}$ such that $x_{n_k} \xrightarrow{\tau} z \in B_d(y, c)$. Since $\{x_{n_k}\}_{k \geq 1} \subset \varphi_i^{-1}(\tilde{x}_i)$ and $\varphi_i^{-1}(\tilde{x}_i)$ is closed in τ, we also have $z \in \varphi_i^{-1}(\tilde{x}_i)$. By lower semicontinuity of d w.r.t. $\tau \times \tau$ we obtain

$$d(z, y) \leq \liminf_{k \to \infty} d(x_{n_k}, y) = \inf_{x \in \varphi_i^{-1}(\tilde{x}_i)} d(x, y),$$

which proves the existence of $z \in \varphi_i^{-1}(\tilde{x}_i)$ such that $d(z, y) = d(\varphi_i^{-1}(\tilde{x}_i), y)$.

Let us next prove that for a set $A = \varphi_i^{-1}(\tilde{x}_i)$ the map $x \mapsto d(A, x)$ is almost lower semi continuous w.r.t. τ. Precisely, let us show that for a sequence $\{x_n\}_{\geq 1} \subset \mathcal{X}$ such that $x_n \xrightarrow{\tau} x \in \mathcal{X}$ and $d(x_n, x) \leq c$ we have

$$\liminf_{n \to \infty} d(A, x_n) \geq d(A, x). \tag{A.3}$$

Without loss of generality, assume that the left-hand side of (A.3) is finite, as otherwise there is nothing to prove. Therefore, we may assume the existence of a constant $c' > 0$ such that $d(A, x_n) \leq c'$. By (A.2), for $n \geq 1$ there exists $z_n \in A$ such that $d(A, x_n) = d(z_n, x_n)$. We then have

$$d(z_n, x) \leq d(z_n, x_n) + d(x_n, x) = d(A, x_n) + d(x_n, x) \leq c' + c.$$

By τ compactness of $B_d(x, c+c')$ and since A is closed in τ, there exists $z \in A \cap B_d(x, c+c')$ such that $z_n \xrightarrow{\tau} z$ up to switching to a subsequence. By lower semi continuity of d w.r.t. $\tau \times \tau$, i.e. assumption iii), we have

$$\liminf_{n\to\infty} d(A, x_n) = \liminf_{n\to\infty} d(z_n, x_n) \geq d(z, x) \geq d(A, x),$$

which proves (A.3).

To prove the existence of minimizers for (A.1) we set

$$V := \inf_{y\in\mathcal{X}} \sum_{i=1}^{n} \lambda_i d(\varphi_i^{-1}(\tilde{x}_i), y)^2$$

and take a sequence $\{y_n\}_{n\geq 1} \subset \mathcal{X}$ such that $\sum_{i=1}^{n} \lambda_i d(\varphi_i^{-1}(\tilde{x}_i), y_n)^2 \searrow V$. We assume without loss of generality that $d(\tilde{x}_j, y_n) \leq 2V$, since $\varphi_j = Id_{\mathcal{X}}$, so we have $\{y_n\}_{n\geq 1} \subset B_d(\tilde{x}_j, 2V)$. By τ-compactness of $B_d(\tilde{x}_j, 2V)$ we have the existence of a subsequence $\{y_{n_k}\}_{k\geq 1} \subset B_d(\tilde{x}_j, 2V)$ such that $y_{n_k} \xrightarrow{\tau} y^*$ for some $y^* \in B_d(\tilde{x}_j, 2V)$. We can now use the proved property (A.3) to obtain

$$\begin{aligned} V &= \liminf_{k\to\infty} \sum_{i=1}^{n} \lambda_i d(\varphi_i^{-1}(\tilde{x}_i), y_{n_k})^2 \\ &\geq \sum_{i=1}^{n} \lambda_i d(\varphi_i^{-1}(\tilde{x}_i), y^*)^2. \end{aligned}$$

Therefore, we have found $y^* \in \mathcal{X}$ such that

$$\sum_{i=1}^{n} \lambda_i d(\varphi_i^{-1}(\tilde{x}_i), y^*)^2 = \inf_{y\in\mathcal{X}} \sum_{i=1}^{n} \lambda_i d(\varphi_i^{-1}(\tilde{x}_i), y)^2,$$

finishing the proof. □

In the proof we also showed the existence of minimizers in (3.5), i.e. points used for "filling up", under the same assumptions. This is precisely (A.2).

Let us discuss the assumptions of Lemma A.1. Note that assumptions i), ii) and iii) are fulfilled in our two important examples of $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{X} = \mathcal{P}_2(\mathbb{R}^d)$. Indeed, for $\mathcal{X} = \mathbb{R}^d$ this is clear, whereas for $\mathcal{X} = \mathcal{P}_2(\mathbb{R}^d)$ the role of the metric τ is played by the metric induced by weak convergence of probability measures. For details we refer to [5, Chapter 7]. As for Assumption iv), this basically (in practical terms) means that for each cluster we need at least one fully observed data point. In our two examples, NA k-means and NA Wasserstein k-means, this would not even be necessary as in [2, Proposition 3.1] it is shown that existence of the generalized Wasserstein barycenter is always guaranteed as long as the maps φ_i are linear. Still, it might be numerically advantageous to have at least one fully observed point in each cluster or include the previous barycenter with a small weight, cf. (5.2).

Supplementary Material: Further Simulation Experiments

Let us here discuss an extended version of our experiment in Section 7.1 to evaluate our algorithm. In Section 7.1 we evaluated our method in the Euclidean setting, NA k-means, as an imputation method. However, without the imputation step it can also be viewed as a clustering algorithm. In the case of Euclidean data it can be viewed as a clustering algorithm for points with missing values. Thus, we can also compare it to classical k-means after imputing the missing data with more standard methods for imputing missing values as used in 7.1. To recall, the standard methods for imputing missing values we consider are mean imputation, median imputation, multiple imputation, K-nearest neighbor imputation, and imputation through linear regression. However, in the case of clustering Euclidean points with missing data we also consider the k-pod method which was introduced in [1]. These authors consider the same loss function as we do in Example 3.2, i.e. (3.8), however they propose a different algorithm. To perform the comparison, we use the same simulation from Section 7.1, i.e. sampling from a Gaussian Mixture Model with $k = 5$ clusters in 5 dimensions and different settings of missing values. We use precisely the same simulated data, and for each of those we apply classical k-means to cluster the points after imputing them with a standard imputation method. For our proposed method, i.e. NA k-means, and for the k-pod method, no imputation step is needed as these directly cluster points with missing values. To evaluate the clustering results we use the so-called Rand index, also known as the Rand score, introduced in [4], as well as the adjusted Rand index (c.f. adjusted Rand score), introduced in [3]. They both compare a baseline clustering (in our case, the labels from the normal distribution of the Gaussian Mixture Model to which an observation belongs) to a clustering of the data using an imputation and/or clustering algorithm. The Rand score takes values in $[0, 1]$, with 0 meaning nothing is clustered the same, and 1 meaning the clusterings coincide except for renaming of clusters. In particular, a higher Rand score corresponds to a better clustering method. The adjusted Rand score is similar to the Rand score but adjusted for chance. It takes value between -0.5 and 1 with 0 indicating a random clustering.

In Table 3 we report the corresponding mean Rand scores $\pm$ standard errors for the simulations of Section 7.1. In Table 4 there are the corresponding adjusted Rand scores.

Table 3: Rand Scores for Euclidean Simulation With $p = 0.15$.

β^{MCAR}	β^{MAR}	β^{MNAR}	**NA k-means**	**mean imp.**	**median imp.**	**multiple imp.**	**KNN**	**LR**	**k-pod**
1	0	0	0.901 ± 0.008	0.876 ± 0.008	0.864 ± 0.008	0.897 ± 0.008	0.896 ± 0.008	0.892 ± 0.008	0.855± 0.008
0	1	0	0.907 ± 0.008	0.874 ± 0.007	0.879 ± 0.007	0.894 ± 0.008	0.905 ± 0.008	0.899 ± 0.008	0.865± 0.007
0	0	1	0.861 ± 0.009	0.831 ± 0.007	0.836 ± 0.008	0.859 ± 0.008	0.857 ± 0.008	0.859 ± 0.008	0.838± 0.008
0.5	0.5	0	0.897 ± 0.008	0.879 ± 0.007	0.863 ± 0.007	0.907 ± 0.008	0.902 ± 0.008	0.893 ± 0.008	0.866± 0.007
0	0.5	0.5	0.884 ± 0.009	0.857 ± 0.007	0.855 ± 0.007	0.883 ± 0.008	0.879 ± 0.008	0.881 ± 0.008	0.856± 0.008
0.5	0	0.5	0.882 ± 0.008	0.857 ± 0.007	0.853 ± 0.007	0.882 ± 0.008	0.877 ± 0.008	0.888 ± 0.008	0.85± 0.007
0.333	0.333	0.333	0.884 ± 0.009	0.865 ± 0.007	0.864 ± 0.007	0.885 ± 0.008	0.884 ± 0.008	0.879 ± 0.009	0.851± 0.007

We can see that the introduced method NA k-means either performs best or lies within the standard error of the best performing method for each setting of missing values when considering it solely as clustering algorithm for points with missing data. The naive approaches of mean and median imputation are outperformed. Also, the k-pod method, which uses the same loss function (3.8) but a different algorithm than proposed here, is outperformed in each setting.

Table 4: Adjusted Rand Scores for Euclidean Simulation With $p = 0.15$.

β^{MCAR}	β^{MAR}	β^{MNAR}	**NA k-means**	**mean imp.**	**median imp.**	**multiple imp.**	**KNN**	**LR**	**k-pod**
1	0	0	0.768 ± 0.018	0.706 ± 0.016	0.68 ± 0.016	0.756 ± 0.017	0.756 ± 0.016	0.746 ± 0.017	0.661± 0.016
0	1	0	0.778 ± 0.017	0.699 ± 0.015	0.709 ± 0.015	0.752 ± 0.017	0.778 ± 0.017	0.763 ± 0.017	0.68± 0.016
0	0	1	0.669 ± 0.018	0.593 ± 0.015	0.608 ± 0.017	0.665 ± 0.017	0.661 ± 0.018	0.664 ± 0.017	0.617± 0.018
0.5	0.5	0	0.758 ± 0.018	0.71 ± 0.015	0.676 ± 0.015	0.78 ± 0.016	0.768 ± 0.017	0.75 ± 0.017	0.68± 0.016
0	0.5	0.5	0.727 ± 0.019	0.656 ± 0.016	0.654 ± 0.016	0.724 ± 0.017	0.714 ± 0.016	0.717 ± 0.017	0.658± 0.017
0.5	0	0.5	0.719 ± 0.018	0.656 ± 0.015	0.648 ± 0.015	0.718 ± 0.016	0.704 ± 0.017	0.735 ± 0.017	0.644± 0.015
0.333	0.333	0.333	0.727 ± 0.019	0.676 ± 0.015	0.676 ± 0.015	0.728 ± 0.017	0.726 ± 0.017	0.716 ± 0.018	0.644± 0.016

Generalizing our Results

In our simulation study, the amount of missing data is controlled by the parameter p, which represents the share of missing values. In Section 7.1 and the previous paragraph, this parameter was fixed at $p = 0.15$. However, to assess the robustness of our method, we now vary p across different values. In Table 5, Table 6 and Table 7 we present the corresponding results analogous to Table 1, Table 3 and Table 4, respectively, when letting $p \in \{0.1, 0.2, 0.25, 0.3\}$.

Table 5: Gromov Wasserstein Distance for Euclidean Simulation With Varying Share of Missing Values.

β^{MCAR}	β^{MAR}	β^{MNAR}	**NA k-means.**	**NA k-means-m**	**mean imp.**	**median imp.**	**multiple imp.**	**KNN**	**LR**
					$p = 0.1$				
1	0	0	0.447 ± 0.021	0.452 ± 0.02	1.021 ± 0.026	1.021 ± 0.026	0.625 ± 0.028	0.479 ± 0.019	0.618 ± 0.024
0	1	0	0.426 ± 0.018	0.427 ± 0.015	1.012 ± 0.028	1.012 ± 0.028	0.617 ± 0.025	0.445 ± 0.019	0.609 ± 0.023
0	0	1	1.171 ± 0.042	1.189 ± 0.041	1.833 ± 0.038	1.833 ± 0.038	1.418 ± 0.042	1.335 ± 0.038	1.391 ± 0.041
0.5	0.5	0	0.406 ± 0.01	0.453 ± 0.033	1.003 ± 0.024	1.003 ± 0.024	0.573 ± 0.017	0.446 ± 0.016	0.584 ± 0.02
0	0.5	0.5	0.874 ± 0.042	0.923 ± 0.039	1.462 ± 0.028	1.462 ± 0.028	1.057 ± 0.029	1.013 ± 0.033	1.045 ± 0.032
0.5	0	0.5	0.856 ± 0.036	0.876 ± 0.036	1.462 ± 0.026	1.462 ± 0.026	1.055 ± 0.031	0.989 ± 0.031	1.061 ± 0.033
0.333	0.333	0.333	0.764 ± 0.036	0.764 ± 0.03	1.33 ± 0.021	1.33 ± 0.021	0.908 ± 0.026	0.877 ± 0.032	0.914 ± 0.028
					$p = 0.2$				
1	0	0	0.605 ± 0.02	0.626 ± 0.021	1.547 ± 0.03	1.547 ± 0.03	0.857 ± 0.024	0.822 ± 0.016	0.879 ± 0.03
0	1	0	0.67 ± 0.036	0.697 ± 0.031	1.589 ± 0.039	1.589 ± 0.039	0.914 ± 0.026	0.778 ± 0.023	0.907 ± 0.026
0	0	1	1.884 ± 0.055	1.918 ± 0.055	2.757 ± 0.062	2.757 ± 0.062	2.068 ± 0.061	2.204 ± 0.054	2.062 ± 0.061
0.5	0.5	0	0.601 ± 0.032	0.618 ± 0.032	1.502 ± 0.03	1.502 ± 0.03	0.851 ± 0.023	0.779 ± 0.02	0.847 ± 0.023
0	0.5	0.5	1.208 ± 0.036	1.286 ± 0.041	2.201 ± 0.036	2.201 ± 0.036	1.515 ± 0.04	1.612 ± 0.036	1.493 ± 0.038
0.5	0	0.5	1.24 ± 0.043	1.3 ± 0.04	2.206 ± 0.037	2.206 ± 0.037	1.513 ± 0.043	1.644 ± 0.04	1.511 ± 0.043
0.333	0.333	0.333	1.006 ± 0.034	1.057 ± 0.036	2.041 ± 0.034	2.041 ± 0.034	1.314 ± 0.041	1.396 ± 0.037	1.33 ± 0.043
					$p = 0.25$				
1	0	0	0.714 ± 0.025	0.73 ± 0.025	1.783 ± 0.033	1.783 ± 0.033	0.977 ± 0.022	1.023 ± 0.027	0.987 ± 0.023
0	1	0	0.783 ± 0.035	0.818 ± 0.031	1.826 ± 0.044	1.826 ± 0.044	1.062 ± 0.028	0.893 ± 0.027	1.081 ± 0.033
0	0	1	2.279 ± 0.06	2.324 ± 0.061	3.177 ± 0.069	3.177 ± 0.069	2.414 ± 0.07	2.593 ± 0.059	2.409 ± 0.069
0.5	0.5	0	0.695 ± 0.026	0.676 ± 0.021	1.742 ± 0.034	1.742 ± 0.034	0.981 ± 0.031	0.922 ± 0.026	0.994 ± 0.033
0	0.5	0.5	1.481 ± 0.044	1.534 ± 0.045	2.544 ± 0.047	2.544 ± 0.047	1.759 ± 0.048	1.92 ± 0.045	1.744 ± 0.046
0.5	0	0.5	1.397 ± 0.038	1.48 ± 0.043	2.574 ± 0.048	2.574 ± 0.048	1.743 ± 0.048	1.929 ± 0.041	1.741 ± 0.047
0.333	0.333	0.333	1.201 ± 0.035	1.313 ± 0.055	2.343 ± 0.039	2.343 ± 0.039	1.501 ± 0.038	1.638 ± 0.035	1.493 ± 0.039
					$p = 0.3$				
1	0	0	0.758 ± 0.028	0.8 ± 0.024	2.009 ± 0.035	2.009 ± 0.035	1.095 ± 0.026	1.099 ± 0.018	1.085 ± 0.022
0	1	0	0.935 ± 0.032	1.013 ± 0.041	2.071 ± 0.05	2.071 ± 0.05	1.256 ± 0.033	1.015 ± 0.028	1.239 ± 0.033
0	0	1	2.719 ± 0.069	2.78 ± 0.069	3.569 ± 0.073	3.569 ± 0.073	2.756 ± 0.074	2.988 ± 0.063	2.757 ± 0.074
0.5	0.5	0	0.816 ± 0.038	0.819 ± 0.033	1.955 ± 0.039	1.955 ± 0.039	1.076 ± 0.025	1.059 ± 0.02	1.082 ± 0.025
0	0.5	0.5	1.585 ± 0.045	1.635 ± 0.045	2.846 ± 0.051	2.846 ± 0.051	1.883 ± 0.045	2.105 ± 0.046	1.898 ± 0.047
0.5	0	0.5	1.6 ± 0.047	1.673 ± 0.047	2.858 ± 0.051	2.858 ± 0.051	1.92 ± 0.05	2.147 ± 0.044	1.92 ± 0.05
0.333	0.333	0.333	1.351 ± 0.043	1.393 ± 0.04	2.656 ± 0.044	2.656 ± 0.044	1.662 ± 0.045	1.838 ± 0.038	1.647 ± 0.04

From Table 5 we conclude that the proposed method NA k-means outperforms all the other considered methods independent of the parameters needed for the missing value generation, when considering the Gromov Wasserstein distance as evaluation measure.

Table 6: Rand Scores for Euclidean Simulation With Varying Shares of Missing Values.

β^{MCAR}	β^{MAR}	β^{MNAR}	**NA k-means**	**mean imp.**	**median imp.**	**multiple imp.**	**KNN**	**LR**	**k-pod**
					$p = 0.1$				
1	0	0	0.906 ± 0.008	0.887 ± 0.008	0.886 ± 0.008	0.897 ± 0.008	0.911 ± 0.008	0.904 ± 0.008	0.875± 0.007
0	1	0	0.902 ± 0.008	0.893 ± 0.007	0.888 ± 0.007	0.906 ± 0.007	0.903 ± 0.008	0.906 ± 0.008	0.881± 0.007
0	0	1	0.881 ± 0.009	0.864 ± 0.008	0.858 ± 0.008	0.88 ± 0.008	0.876 ± 0.008	0.881 ± 0.008	0.873± 0.007
0.5	0.5	0	0.905 ± 0.008	0.888 ± 0.007	0.883 ± 0.007	0.906 ± 0.008	0.912 ± 0.008	0.901 ± 0.008	0.877± 0.007
0	0.5	0.5	0.896 ± 0.009	0.874 ± 0.007	0.875 ± 0.007	0.885 ± 0.008	0.897 ± 0.008	0.895 ± 0.008	0.875± 0.008
0.5	0	0.5	0.892 ± 0.009	0.878 ± 0.008	0.869 ± 0.007	0.891 ± 0.008	0.894 ± 0.008	0.898 ± 0.008	0.875± 0.008
0.333	0.333	0.333	0.896 ± 0.009	0.876 ± 0.008	0.886 ± 0.007	0.897 ± 0.009	0.905 ± 0.008	0.905 ± 0.008	0.883± 0.007
					$p = 0.2$				
1	0	0	0.889 ± 0.009	0.845 ± 0.007	0.842 ± 0.007	0.88 ± 0.008	0.879 ± 0.008	0.878 ± 0.008	0.828± 0.006
0	1	0	0.884 ± 0.008	0.862 ± 0.007	0.853 ± 0.007	0.888 ± 0.007	0.883 ± 0.008	0.888 ± 0.008	0.834± 0.007
0	0	1	0.837 ± 0.009	0.803 ± 0.008	0.808 ± 0.007	0.847 ± 0.008	0.821 ± 0.009	0.844 ± 0.009	0.796± 0.01
0.5	0.5	0	0.898 ± 0.008	0.865 ± 0.007	0.848 ± 0.007	0.89 ± 0.008	0.895 ± 0.007	0.891 ± 0.008	0.835± 0.007
0	0.5	0.5	0.857 ± 0.009	0.828 ± 0.008	0.835 ± 0.008	0.871 ± 0.009	0.853 ± 0.008	0.868 ± 0.009	0.821± 0.008
0.5	0	0.5	0.87 ± 0.008	0.834 ± 0.007	0.832 ± 0.007	0.865 ± 0.007	0.859 ± 0.008	0.866 ± 0.008	0.826± 0.007
0.333	0.333	0.333	0.882 ± 0.008	0.835 ± 0.007	0.841 ± 0.007	0.88 ± 0.008	0.867 ± 0.008	0.879 ± 0.008	0.833± 0.007
					$p = 0.25$				
1	0	0	0.875 ± 0.009	0.824 ± 0.007	0.82 ± 0.006	0.867 ± 0.008	0.85 ± 0.008	0.867 ± 0.008	0.802± 0.007
0	1	0	0.873 ± 0.008	0.839 ± 0.007	0.836 ± 0.007	0.873 ± 0.007	0.873 ± 0.008	0.874 ± 0.008	0.817± 0.008
0	0	1	0.815 ± 0.009	0.793 ± 0.008	0.793 ± 0.009	0.821 ± 0.009	0.785 ± 0.008	0.826 ± 0.009	0.77± 0.01
0.5	0.5	0	0.891 ± 0.008	0.845 ± 0.007	0.841 ± 0.007	0.88 ± 0.007	0.87 ± 0.008	0.881 ± 0.008	0.817± 0.007
0	0.5	0.5	0.851 ± 0.009	0.805 ± 0.007	0.805 ± 0.007	0.854 ± 0.008	0.829 ± 0.008	0.856 ± 0.008	0.794± 0.008
0.5	0	0.5	0.855 ± 0.008	0.81 ± 0.007	0.812 ± 0.007	0.855 ± 0.009	0.827 ± 0.008	0.856 ± 0.008	0.796± 0.008
0.333	0.333	0.333	0.862 ± 0.009	0.822 ± 0.007	0.825 ± 0.007	0.864 ± 0.008	0.843 ± 0.008	0.864 ± 0.009	0.806± 0.008
					$p = 0.3$				
1	0	0	0.87 ± 0.009	0.803 ± 0.007	0.8 ± 0.006	0.856 ± 0.008	0.849 ± 0.008	0.856 ± 0.008	0.765± 0.008
0	1	0	0.864 ± 0.008	0.822 ± 0.007	0.828 ± 0.007	0.862 ± 0.007	0.863 ± 0.008	0.865 ± 0.007	0.787± 0.007
0	0	1	0.802 ± 0.009	0.781 ± 0.008	0.777 ± 0.008	0.81 ± 0.009	0.755 ± 0.007	0.806 ± 0.008	0.754± 0.009
0.5	0.5	0	0.883 ± 0.008	0.825 ± 0.007	0.828 ± 0.006	0.868 ± 0.008	0.861 ± 0.007	0.871 ± 0.007	0.793± 0.006
0	0.5	0.5	0.848 ± 0.009	0.798 ± 0.007	0.789 ± 0.007	0.846 ± 0.008	0.811 ± 0.007	0.844 ± 0.008	0.779± 0.008
0.5	0	0.5	0.845 ± 0.008	0.792 ± 0.007	0.798 ± 0.007	0.84 ± 0.007	0.812 ± 0.007	0.847 ± 0.007	0.778± 0.008
0.333	0.333	0.333	0.862 ± 0.008	0.803 ± 0.006	0.798 ± 0.007	0.853 ± 0.008	0.833 ± 0.007	0.847 ± 0.007	0.78± 0.008

From Tables 6 and 7 we can see that also w.r.t. the (adjusted) Rand score, when NA k-means is solely viewed as clustering procedure for Euclidean points with missing values, it competes with the best performing imputation + clustering techniques. Notably, k-pod is outperformed consistently.

Table 7: Adjusted Rand Scores for Euclidean Simulation With Varying Share of Missing Values.

β^{MCAR}	β^{MAR}	β^{MNAR}	**NA k-means**	**mean imp.**	**median imp.**	**multiple imp.**	**KNN**	**LR**	**k-pod**
					$p = 0.1$				
1	0	0	0.779 ± 0.017	0.733 ± 0.017	0.732 ± 0.016	0.758 ± 0.016	0.791 ± 0.017	0.774 ± 0.016	0.704± 0.015
0	1	0	0.768 ± 0.017	0.748 ± 0.016	0.734 ± 0.015	0.779 ± 0.016	0.774 ± 0.017	0.778 ± 0.017	0.716± 0.016
0	0	1	0.72 ± 0.018	0.676 ± 0.017	0.662 ± 0.016	0.717 ± 0.017	0.707 ± 0.017	0.719 ± 0.018	0.698± 0.016
0.5	0.5	0	0.776 ± 0.017	0.734 ± 0.016	0.722 ± 0.015	0.779 ± 0.016	0.793 ± 0.018	0.767 ± 0.017	0.711± 0.015
0	0.5	0.5	0.757 ± 0.019	0.699 ± 0.015	0.705 ± 0.016	0.732 ± 0.018	0.757 ± 0.017	0.754 ± 0.017	0.705± 0.017
0.5	0	0.5	0.744 ± 0.018	0.712 ± 0.016	0.687 ± 0.015	0.743 ± 0.017	0.752 ± 0.018	0.76 ± 0.018	0.705± 0.018
0.333	0.333	0.333	0.755 ± 0.018	0.703 ± 0.016	0.73 ± 0.014	0.76 ± 0.018	0.776 ± 0.017	0.776 ± 0.018	0.724± 0.016
					$p = 0.2$				
1	0	0	0.741 ± 0.018	0.629 ± 0.014	0.622 ± 0.014	0.72 ± 0.016	0.718 ± 0.016	0.714 ± 0.017	0.593± 0.015
0	1	0	0.728 ± 0.017	0.665 ± 0.015	0.649 ± 0.014	0.736 ± 0.015	0.723 ± 0.016	0.733 ± 0.016	0.608± 0.015
0	0	1	0.614 ± 0.019	0.528 ± 0.016	0.537 ± 0.016	0.634 ± 0.017	0.572 ± 0.018	0.629 ± 0.018	0.521± 0.02
0.5	0.5	0	0.759 ± 0.017	0.674 ± 0.015	0.634 ± 0.014	0.739 ± 0.017	0.75 ± 0.015	0.742 ± 0.017	0.608± 0.016
0	0.5	0.5	0.664 ± 0.018	0.592 ± 0.016	0.605 ± 0.017	0.697 ± 0.018	0.652 ± 0.016	0.691 ± 0.017	0.576± 0.018
0.5	0	0.5	0.691 ± 0.017	0.598 ± 0.015	0.596 ± 0.014	0.677 ± 0.016	0.663 ± 0.016	0.682 ± 0.017	0.588± 0.016
0.333	0.333	0.333	0.719 ± 0.017	0.6 ± 0.014	0.616 ± 0.015	0.714 ± 0.016	0.683 ± 0.017	0.71 ± 0.016	0.598± 0.015
					$p = 0.25$				
1	0	0	0.709 ± 0.018	0.583 ± 0.014	0.574 ± 0.014	0.689 ± 0.016	0.646 ± 0.016	0.688 ± 0.016	0.538± 0.015
0	1	0	0.698 ± 0.016	0.607 ± 0.015	0.606 ± 0.015	0.695 ± 0.016	0.699 ± 0.016	0.701 ± 0.016	0.566± 0.017
0	0	1	0.558 ± 0.019	0.505 ± 0.016	0.508 ± 0.02	0.574 ± 0.018	0.488 ± 0.017	0.583 ± 0.019	0.466± 0.02
0.5	0.5	0	0.742 ± 0.017	0.628 ± 0.015	0.616 ± 0.015	0.713 ± 0.016	0.695 ± 0.016	0.719 ± 0.016	0.567± 0.015
0	0.5	0.5	0.645 ± 0.018	0.53 ± 0.015	0.532 ± 0.015	0.651 ± 0.018	0.59 ± 0.016	0.656 ± 0.017	0.506± 0.018
0.5	0	0.5	0.656 ± 0.017	0.544 ± 0.015	0.552 ± 0.016	0.657 ± 0.018	0.587 ± 0.016	0.658 ± 0.018	0.524± 0.018
0.333	0.333	0.333	0.674 ± 0.018	0.573 ± 0.015	0.579 ± 0.016	0.677 ± 0.017	0.627 ± 0.016	0.678 ± 0.018	0.54± 0.017
					$p = 0.3$				
1	0	0	0.695 ± 0.017	0.528 ± 0.013	0.525 ± 0.014	0.661 ± 0.014	0.641 ± 0.015	0.662 ± 0.015	0.462± 0.014
0	1	0	0.674 ± 0.016	0.571 ± 0.014	0.588 ± 0.015	0.667 ± 0.014	0.672 ± 0.016	0.674 ± 0.015	0.503± 0.016
0	0	1	0.522 ± 0.019	0.474 ± 0.015	0.467 ± 0.018	0.541 ± 0.019	0.407 ± 0.015	0.532 ± 0.018	0.425± 0.018
0.5	0.5	0	0.722 ± 0.016	0.577 ± 0.014	0.588 ± 0.015	0.687 ± 0.016	0.669 ± 0.015	0.691 ± 0.015	0.51± 0.014
0	0.5	0.5	0.634 ± 0.018	0.507 ± 0.014	0.491 ± 0.015	0.629 ± 0.016	0.543 ± 0.014	0.626 ± 0.017	0.478± 0.017
0.5	0	0.5	0.624 ± 0.017	0.492 ± 0.014	0.515 ± 0.017	0.609 ± 0.016	0.542 ± 0.015	0.628 ± 0.017	0.469± 0.017
0.333	0.333	0.333	0.667 ± 0.017	0.519 ± 0.013	0.511 ± 0.015	0.646 ± 0.016	0.594 ± 0.015	0.629 ± 0.015	0.474± 0.016

The Fundamental Theorem of Weak Optimal Transport

M. Beiglböck[†], G. Pammer[‡], L. Riess[†], S. Schrott[†]

Abstract

The fundamental theorem of classical optimal transport establishes strong duality and characterizes optimizers through a complementary slackness condition. Milestones such as Brenier's theorem and the Kantorovich-Rubinstein formula are direct consequences.

In this paper, we generalize this result to non-linear cost functions, thereby establishing a fundamental theorem for the weak optimal transport problem introduced by Gozlan, Roberto, Samson, and Tetali. As applications we provide concise derivations of the Brenier–Strassen theorem, the convex Kantorovich–Rubinstein formula and the structure theorem of entropic optimal transport. We also extend Strassen's theorem in the direction of Gangbo–McCann's transport problem for convex costs. Moreover, we determine the optimizers for a new family of transport problems which contains the Brenier–Strassen, the martingale Benamou–Brenier and the entropic martingale transport problem as extreme cases.

keywords: weak optimal transport, Gangbo-McCann theorem, dual attainment, martingale optimal transport, entropic optimal transport, adapted weak topology

[†]University of Vienna
[‡]TU Graz

1 Introduction

Optimal transport (OT) provides a powerful theory with applications in numerous fields. However, there is a range of problems that cannot be addressed within this framework due to non-linearities in the cost function. This motivated Gozlan, Roberto, Samson and Tetali [51] to introduce the more encompassing framework of *weak optimal transport* (WOT), see also the closely related works of Aliberti, Bouchitte and Champion [5] as well as Bowles and Ghoussoub [32]. Our aim is to generalize the basic structure theorem of transport—often referred to as the "fundamental theorem of optimal transport" (see e.g. Theorem 1.13 in Ambrosio–Gigli [6])—to the WOT setting. To prepare for this extension and its applications, we first recall the classical framework.

Given a Polish metric space Z, we denote by $\mathcal{P}(Z)$ the set of probabilities on Z, equipped with weak convergence and by $\mathcal{P}_p(Z)$, $p \in [1,\infty)$ probabilities with finite p-moment with p-weak convergence. Throughout we consider Polish probability spaces (X,μ), (Y,ν). We write $\Pi(\mu,\nu)$ for the set of all *couplings* or *transport plans*, that is, probabilities on $X \times Y$ which have X-marginal μ and Y-marginal ν. As customary, we assume that the '*cost function*' $c : X \times Y \to [0,\infty]$ is lsc, implying attainment for the (primal) *Monge-Kantorovich transport problem*

$$\mathrm{T}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\,\pi(dx,dy). \tag{1.1}$$

In key applications we are often in the situation that $X = Y$ and c relates to the metric d_X of X: E.g. *Kantorovich-Rubinstein* [56] asserts that for $\mu,\nu \in \mathcal{P}_1(X)$ and $c = d_X$ we have $\mathrm{T}_{d_X}(\mu,\nu) = \max_{\psi\ 1\text{-Lip}} \nu(\psi) - \mu(\psi)$.

Likewise *Brenier's theorem* [35] asserts that for $X = Y = \mathbb{R}^d$, the squared distance cost $c(x,y) = |x-y|^2$ and $\mu,\nu \in \mathcal{P}_2(\mu,\nu)$, $\mu \ll \mathrm{Leb}$ the transport problem admits a unique solution. Moreover, the optimizer is characterized as the only admissible transport plan supported by the graph of a convex function.

Both results are consequences of the basic structure theorem of optimal transport:

Theorem 1.1 (Fundamental Theorem of OT)**.** *For $c : X \times Y \to [0,\infty]$ lsc, the primal problem $\mathrm{T}_c(\mu,\nu)$ is attained and equals the dual problem*

$$\mathrm{D}_c(\mu,\nu) := \sup\{\mu(f) + \nu(g) : f(x) + g(y) \le c(x,y),\ f \in \mathcal{L}^1(\mu),\ g \in \mathcal{L}^1(\nu)\}. \tag{1.2}$$

The dual problem is attained if c is upper bounded in the sense that

$$c(x,y) \le a(x) + b(y) \quad \text{for some } a \in \mathcal{L}^1(\mu),\ b \in \mathcal{L}^1(\nu). \tag{A}$$

Candidates $\pi, (f,g)$ are optimal if and only if they satisfy the complementary slackness condition

$$c(x,y) = f(x) + g(y), \quad \pi\text{-a.s.} \tag{1.3}$$

Given a measurable *cost function* $C : X \times \mathcal{P}_p(Y) \to [0,\infty]$ which is lsc and convex in the second argument, and denoting the μ-disintegration of π by $(\pi_x)_{x\in X}$, the weak

transport problem is

$$\mathrm{WT}_C(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int C(x,\pi_x)\,\mu(dx). \tag{1.4}$$

The classical transport problem corresponds to the case where $\rho \mapsto C(x,\rho)$ is even linear, i.e.

$$C(x,\rho) = \int c(x,y)\,\rho(dy). \tag{1.5}$$

To establish the desired counterpart of Theorem 1.1 we need the *boundedness condition*

$$C(x,\rho) \le a(x) + \rho(b) + \int h\Big(\frac{d\rho}{d\nu}\Big)d\nu, \tag{B}$$

for some $a \in \mathcal{L}^1(\mu), b \in \mathcal{L}^1(\nu)$ and a convex increasing function $h : [0,\infty) \to [0,\infty)$, where the right-hand side of (B) is understood as $+\infty$ if ρ is not absolutely continuous w.r.t. ν. Apparently (A) implies (B) if C is of transport type as in (1.5). Some boundedness assumption is necessary to obtain dual attainment in optimal transport, see e.g. [22, Examples 4.4, 4.5], and clearly these examples also pertain to weak transport.

We also need that for increasing sequences $(Y_k)_k$ of Borel sets with $\bigcup_k Y_k = Y$

$$C(x,\rho) \ge \limsup_k C(x,(\rho|_{Y_k})/\rho(Y_k)). \tag{C}$$

This is as a mild *continuity condition*, trivially satisfied in the classical setting (1.5) and our intended applications. However, see Example 2.12 for the necessity of some continuity condition.

With these preparations, we can now state our main theorem:

Theorem 1.2 (Fundamental Theorem of WOT)**.** *If $C : X \times \mathcal{P}_p(Y) \to [0,\infty]$ is measurable with $\rho \mapsto C(x,\rho)$ being convex and lsc for the p-weak convergence on $\mathcal{P}_p(Y)$, then $\mathrm{WT}_C(\mu,\nu)$ is attained and equals the dual problem*[2]

$$\mathrm{D}_C(\mu,\nu) := \sup\{\mu(f)+\nu(g) : f(x)+\rho(g) \le C(x,\rho),\ f\in\mathcal{L}^1(\mu),\ g\in\mathcal{L}^1(\nu)\}. \tag{1.6}$$

Moreover, if C satisfies conditions (B) and (C), then the dual problem is attained. A pair of candidates $(\pi,(f,g))$ is optimal if and only if it satisfies the complementary slackness condition

$$C(x,\pi_x) = f(x) + \pi_x(g), \quad \mu\text{-a.s.} \tag{1.7}$$

Starting with the work of Gozlan, Roberto, Samson, and Tetali [51], primal existence and duality for weak transport costs have been established in the literature under different assumptions, cf. [5, 15, 32]. The formulation we give in Theorem 1.2 is slightly stronger than previously known results in order to include also EOT in its usual generality. From

[2]Specifically we require the inequality in (1.6) for all $(x,\rho) \in X \times \mathcal{P}_p(Y)$ with $g \in \mathcal{L}^1(\rho)$.

our perspective, the main novelty of our result lies in establishing attainment for the dual problem.

Note that analogous to the classical transport problem we can rewrite (1.6) as

$$\begin{aligned} \mathrm{D}_C(\mu,\nu) &= \sup\{\mu(g^C) + \nu(g) : g^C \in \mathcal{L}^1(\mu),\, g \in \mathcal{L}^1(\nu)\} \\ \text{where } g^C(x) &:= \inf\{C(x,\rho) - \rho(g) : \rho \in \mathcal{P}_p(Y) \text{ s.t. } g \in \mathcal{L}^1(\rho)\}. \end{aligned}$$

In fact, the core of utilizing Theorem 1.2 is the computation of the C-transform g^C for chosen instances of C.

1.1 Applications of the Fundamental Theorem of WOT

Next we give an overview of our applications of Theorem 1.2 which consist in shortened, unified derivations of old results as well as of new results following the spirit of Kantorovich-Rubinstein / Brenier's theorem.

1.1.1 Convex Kantorovich–Rubinstein

Given $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$, [5, 48] derive the following *convex Kantorovich–Rubinstein formula*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int |x - \mathrm{mean}(\pi_x)|\, \mu(dx) = \max_{\psi \text{ convex, 1-Lip}} \mu(\psi) - \nu(\psi). \tag{1.8}$$

A transport plan π is called a martingale coupling if it is the law of a two-step martingale, that is, if μ-a.s. $\mathrm{mean}(\pi_x) = x$. A classical result of Strassen [79] asserts that the set $\Pi_{\mathrm{M}}(\mu,\nu)$ of martingale couplings is non-empty if and only if μ, ν are in convex order $\leq_c$, i.e. $\mu(\psi) \leq \nu(\psi)$ for all convex ψ. The convex Kantorovich-Rubinstein formula can be seen as quantitative extension of Strassen's theorem: The left-hand side of (1.8) measures how close we are to finding a martingale coupling of μ and ν, while the right-hand side measures by how much μ, ν violate the convex order condition.

1.1.2 The Brenier–Strassen theorem

Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, Gozlan–Julliet [49] establish a Brenier–Strassen theorem for the weak transport problem

$$\inf_{\pi \in \Pi(\mu,\nu)} \int |x - \mathrm{mean}(\pi_x)|^2\, \mu(dx). \tag{1.9}$$

It asserts that there exists a μ-a.s. unique map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that for every optimizer $\pi \in \Pi(\mu,\nu)$, $T(x) = \mathrm{mean}(\pi_x)$. Moreover T is 1-Lipschitz, the gradient of a convex function and optimizers of (1.9) are characterized by

$$\pi \in \Pi(\mu,\nu) \text{ is optimal } \iff \pi_x = \kappa_{T(x)},\ \mu\text{-a.s. for some } \kappa \in \Pi_{\mathrm{M}}(T_{\#}\mu, \nu). \tag{1.10}$$

We emphasize that in contrast to Brenier's theorem, no regularity of the measure μ is required beyond the existence of finite second moment. In [49], the Brenier–Strassen theorem is established based on duality for compactly supported probabilities and using approximation arguments to pass to the general case, see also [15].

1.1.3 A Gangbo–McCann–Strassen theorem

In Theorem 3.1 below, we establish a weak transport version of Gangbo–McCann's result [44] which recovers the convex Kantorovich–Rubinstein formula as well as the Brenier–Strassen theorem as special cases. Specifically we consider for convex $\vartheta : \mathbb{R}^d \to [0, \infty)$ the problem

$$\inf_{\pi \in \Pi(\mu,\nu)} \int \vartheta(x - \text{mean}(\pi_x))\, \mu(dx). \tag{1.11}$$

We show that if ϑ is strictly convex, there exists a unique T such that for π optimal $T(x) = \text{mean}(\pi_x)$, μ-a.s. Moreover the set of optimizers is characterized as in (1.10).

On the other hand, if ϑ is positively homogeneous, (i.e. a support function), we obtain an extended convex Kantorovich–Rubinstein formula. Beyond (1.8) it implies for instance an increasing convex Kantorovich–Rubinstein formula, giving a quantitative version of Strassen's result for sub-martingales / the increasing convex order (see Corollary 3.5).

We note that barycentric costs can be seen as a relaxation of the martingale condition which appears naturally from mathematical finance considerations. Here the martingale condition for transport plans corresponds to a strict No-Arbitrage paradigm, this is a robust version of the 'fundamental theorem of asset pricing', e.g. [1, 55, 84]. Based on (1.8) and (1.11) we obtain quantitative extensions: The level of arbitrage that can be locked in under a given trading restriction is reflected by a corresponding relaxation of the martingale condition, see Section 3.2.

1.1.4 Entropic optimal transport

Given probabilities μ, ν on Polish spaces X, Y, a measurable cost function $c : X \times Y \to [0, \infty)$ which is bounded by the sum of two integrable functions, and $\varepsilon > 0$, the entropic transport problem reads

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c\, d\pi + \varepsilon H(\pi | \mu \otimes \nu), \tag{1.12}$$

where H denotes relative entropy, i.e. $H(Q|R) = \int \log\left(\frac{dQ}{dR}\right) dQ$ if Q, R are probability measures with $Q \ll R$ and $H(Q|R) = \infty$ otherwise. See [62, 70] for surveys on entropic optimal transport (EOT) and the closely related Schrödinger problem.

Observing that $H(\pi|\mu \otimes \nu) = \int H(\pi_x|\nu)\, \mu(dx)$, we notice that (1.12) is a weak transport problem. Based on Theorem 1.2 we recover the main structure theorem of EOT: (1.12) admits a unique solution π and among all couplings of μ and ν the optimizer is characterized by

$$\frac{d\pi}{d\mu \otimes \nu}(x, y) = \exp\left(-\frac{c(x, y) + f(x) + g(y)}{\varepsilon}\right). \tag{1.13}$$

In this case (f, g) is optimal in the dual problem (4.3).

Along the same lines we also prove the structure theorem for transport costs regularized by general convex functions, thus recovering a result of [67], see Section 4.2.

1.1.5 Relaxation of martingale transport

Weak martingale optimal transport (WMOT) differs from classical transport in that one minimizes over martingale couplings $\Pi_{\mathrm{M}}(\mu,\nu)$ for $\mu,\nu\in\mathcal{P}_p(\mathbb{R}^d)$. This restriction can be incorporated by considering weak transport costs of the form

$$C(x,\rho)=\begin{cases}\widehat{C}(\rho) & \text{mean}(\rho)=x\\ \infty & \text{else.}\end{cases}$$

Here C is lsc and convex in the second argument provided that $\widehat{C}:\mathcal{P}_p(\mathbb{R}^d)\to[0,\infty]$ is lsc and convex on all fibers $\{\rho\in\mathcal{P}_p(\mathbb{R}^d):\text{mean}(\rho)=x\}, x\in\mathbb{R}^d$. However, C does not satisfy the boundedness assumption (B). In fact, dual attainment for martingale transport fails even in very regular settings (e.g. [23, Section 4.3]). Positive results are only available for $d=1$ and under strong assumptions ([25, 26]). In consequence, important questions such as the entropic martingale transport and the martingale Benamou–Brenier problem are only partially understood.

In Section 5 we show that these challenges can be overcome by applying Theorem 1.2 to a suitable relaxation. In particular, if $\widehat{C}$ satisfies appropriate counterparts to (B) and (C), we establish dual attainment for WMOT, which is relaxed using a term $\beta\mathcal{W}_2^2$, where $\beta>0$:

$$\min_{\eta\in\mathcal{P}_2(\mathbb{R}^d)}\Big(\beta\mathcal{W}_2^2(\mu,\eta)+\min_{\kappa\in\Pi_{\mathrm{M}}(\eta,\nu)}\int C(x,\kappa_x)\,\mu(dx)\Big)=\max_{g\in\mathcal{L}^1(\nu)}\mu(\tfrac{1}{\beta}|.|^2\square g^C)+\nu(g).\quad(1.14)$$

We note that while the dual uses the classical infimal convolution / Yosida regularization $(\frac{1}{\beta}|.|^2\square f)(x)=\inf_{y\in\mathbb{R}^d}\frac{1}{\beta}|x-y|^2+f(y)$, the primal side is an infimal convolution with $\beta\mathcal{W}_2^2$.

We continue with discussing applications of (1.14) to martingale transport:

1.1.6 From Brenier–Strassen to martingale Benamou–Brenier

Finding 'natural' martingale transport plans between given marginals is an important problem in mathematical finance, e.g. to derive robust bounds or to obtain calibrated models. Recently, particular attention (see [3, 13, 16, 27, 38, 52, 64] among others) has gone to *Bass martingales*, i.e. martingales of the form

$$M_t=\mathsf{E}[\nabla v(B_1)|B_t],\qquad(1.15)$$

for some convex $v:\mathbb{R}^d\to\mathbb{R}$ and a Brownian motion $(B_t)_{t\in[0,1]}$ with possibly random B_0. For sufficiently regular $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d),\mu\leq_c\nu$, there is a (unique) Bass martingale with $M_0\sim\mu, M_1\sim\nu$ which is further characterized as the solution to the martingale Benamou–Brenier problem, see [12, 13].[3] The key to this result is to recast the optimization problem as the WMOT problem

$$\sup_{\pi\in\Pi_{\mathrm{M}}(\mu,\nu)}\int\mathrm{MCov}(\pi_x,\gamma)\,\mu(dx),\ \text{where MCov}(\rho,\gamma):=\sup_{q\in\Pi(\rho,\gamma)}\int y\cdot z\,q(dy,dz)$$

[3] Alternatively, the Bass martingale is the martingale which is closest to Brownian motion in adapted Wasserstein distance subject to the given marginal constraints.

and γ is the standard Gaussian.

Given $\alpha, \beta > 0$, we consider its 'barycentric relaxation'

$$\inf_{\pi \in \Pi(\mu,\nu)} \int \beta |x - \mathrm{mean}(\pi_x)|^2 - \alpha \mathrm{MCov}(\pi_x, \gamma)\, \mu(dx), \tag{1.16}$$

which has as extreme cases Brenier–Strassen (for $\alpha \to 0$) and the martingale Benamou–Brenier problem (for $\beta \to \infty$).

One can rewrite (1.16) as a relaxed WMOT problem so that dual attainment (1.14) is applicable. In particular, we obtain that there exist unique $T : \mathbb{R}^d \to \mathbb{R}^d$ and $\kappa \in \Pi_{\mathrm{M}}(T_{\#}\mu, \nu)$ such that $\pi_x := \kappa_{T(x)}$ is the unique optimizer of (1.16). Moreover, T is $\frac{1}{\beta}$-Lipschitz, the gradient of a convex function and $\kappa = \mathrm{law}(M_0, M_1)$ for a Bass martingale M.

We underline that this behavior is more regular than either of the extremes: The optimizer in Brenier–Strassen is not unique, while Bass martingales exist only under additional assumptions on $\mu \leq_c \nu$ in the classical MOT case. In Section 5.1 we give a precise description of the optimizer to (1.16) and further extensions.

1.1.7 Relaxing martingale EOT

We close this section with another example that becomes more regular by considering a relaxation of the martingale transport as in (1.14). Specifically, we consider the problem

$$\min_{\eta \in \mathcal{P}_2(\mathbb{R}^d)} \Big(\beta \mathcal{W}_2^2(\mu, \eta) + \min_{\kappa \in \Pi_{\mathrm{M}}(\eta,\nu)} \int c\, d\kappa + \varepsilon H(\kappa | \eta \otimes \nu) \Big) \tag{1.17}$$

where $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ is lsc and $\beta, \varepsilon > 0$.

Here the usual martingale transport would correspond to the case $\beta \to \infty$, where (1.17) is the natural entropic regularization of the martingale transport problem. However, due to the intricacies of MOT duality, this problem appears challenging and has so far been solved only for $d = 1$, $c \equiv 0$ and regularity conditions for μ, ν, see [72]. In contrast (1.17) is more tractable due to strong duality in (1.14): If c is Lipschitz and μ is absolutely continuous, we show there exist unique solutions η, κ; moreover, for suitable functions f, g, Δ, the density of κ is of Gibbs type

$$\frac{d\kappa}{\eta \otimes \nu}(\bar{x}, y) = \exp\left(-\frac{c(\bar{x}, y) + f(\bar{x}) + g(y) + \Delta(\bar{x})(y - \bar{x})}{\varepsilon} \right). \tag{1.18}$$

1.2 Duality and dual attainment without convexity assumptions

Many known problems which fall in the weak transport framework exhibit a cost function such that $\rho \mapsto C(x, \rho)$ is convex and lsc. Moreover, in sufficiently regular settings, it is possible to investigate the weak transport problem for non-convex costs by considering the respective convex hulls. Specifically, assume that μ is continuous, C is jointly continuous with bounded p-growth and define $\bar{C}$ such that $\bar{C}(x, \cdot)$ is the convex lsc hull of $C(x, \cdot)$ for every $x \in X$. Then [2, Theorem 3.9] asserts that

$$\inf_{\pi \in \Pi(\mu,\nu)} \int C(x, \pi_x)\, \mu(dx) = \inf_{\pi \in \Pi(\mu,\nu)} \int \bar{C}(x, \pi_x)\, \mu(dx). \tag{1.19}$$

In particular, we recover the duality relation as well as dual attainment in this case. However, while the right-hand side of (1.19) is of course attained, the original primal problem on the left-hand side is not in general.

In some settings it is important to have primal attainment, as well as dual attainment without imposing convexity or continuity of μ and C. For instance this is the case for the relaxed entropic MOT problem, in mathematical finance in connection to robust hedging of American options and for the model independent pricing problem in a fixed income market, see [2]. Here the solution is to consider a suitably enriched space of transport plans, we discuss this in Section 2.5. Using this relaxed formulation of the primal problem, we obtain a general fundamental theorem of WOT without imposing convexity of $\rho \mapsto C(x, \rho)$, see Theorem 2.17.

1.3 Related literature

Weak transportation costs appeared in the works of Marton [68, 69] and Talagrand [80, 81], where the authors studied the concentration of measure phenomenon via transport entropy inequalities. Motivated by questions of concentration and curvature properties of discrete measures, Gozlan, Roberto, Samson, and Tetali [51] provided the general definition of weak transport costs and studied their basic properties. In particular, the Kantorovich duality is obtained for costs which are convex in the second argument and satisfy some additional mild regularity conditions. Subsequently, Alibert, Bouchitte and Champion [5] as well as Bowles and Ghoussoub [32] introduced closely related setups, specifically [5] also establishes primal existence and duality on compact state spaces. Choné, Gozlan and Kramarz [36] relax weak transport further to include also transports by unnormalized kernels, see also [76].

Based on using the adapted weak topology on the set of couplings, [15] extends existence and duality to the level of generality familiar from classical transport, that is lsc, $[0, \infty]$-valued costs on abstract Polish spaces. Existence and duality results beyond costs given convexity assumptions are given in [2], see also Sections 1.2 and 2.5 below. In analogy to classical c-cyclical monotonicity [13, 15, 49] develop the notion of *C-monotonicity* as a necessary criterion for optimality in weak transport. The article [9] establishes sufficiency of this criterion as well as stability of WOT w.r.t. its marginals.

Besides the original motivation of geometric inequalities, WOT contains a variety of further problems that fall outside the class of classical optimal transport. Particular attention has gone to barycentric transport costs [4, 9, 15, 43, 49, 50, 75, 77, 78]. As noted in [37], WOT covers entropic optimal transport EOT (see e.g. [62, 70]). It includes martingale optimal transport MOT (see e.g. [21, 54]), semi-martingale optimal transport SMOT (see e.g. [28, 82]) and the optimal Skorokhod embedding problem with non-trivial starting law ([66]). WOT has been used to define so-called shadow (sub-)martingales ([19, 24]). It appears as important tool in the investigation of Bass martingales (e.g. [13, 27]) and the recent probabilistic proof of the Caffarelli contraction theorem [42]. Beyond its applications to martingale transport problems, WOT includes a number of further problems in economics and finance, we mention optimal mechanism design [40], optimal transport planning [5], stability of pricing and hedging [10], model-independence in fixed income markets [2], risk measures [61], and robust pricing of VIX futures [20,

39, 53]. Furthermore, the WOT framework has been applied to machine learning tasks such as unpaired domain translation problems, generative modeling, and the learning of Wasserstein barycenters, see [7, 33, 58, 60].

2 Fundamental theorem of weak optimal transport

The aim of this section is to prove the fundamental theorem of weak optimal transport. As already in the case of classical optimal transport (on non-compact spaces), we cannot expect dual attainment when maximizing only over continuous functions. For this reason, we need to introduce the dual problem in the correct generality. We denote by $\mathcal{L}^1(\rho)$ the set of $[-\infty,\infty]$-valued Borel functions that are ρ-integrable. We call (f,g) admissible if $f \in \mathcal{L}^1(\mu)$ and $g \in \mathcal{L}^1(\nu)$ are such that

$$f(x) + \rho(g) \leq C(x,\rho) \tag{2.1}$$

for all $(x,\rho) \in X \times \mathcal{P}_p(Y)$ with $g \in \mathcal{L}^1(\rho)$. The set of all admissible pairs is denoted by $\Phi_C(\mu,\nu)$. Note that since (2.1) has to hold pointwise, we do not identify f and g with their respective L^1-equivalence classes. The dual problem is then given by

$$\mathrm{D}_C(\mu,\nu) := \sup_{(f,g)\in\Phi_C(\mu,\nu)} \mu(f) + \nu(g). \tag{DP}$$

We recall that the C-transform for weak transport costs is given by

$$g^C(x) := \inf\{C(x,\rho) - \rho(g) : \rho \in \mathcal{P}_p(Y) \text{ s.t. } g \in \mathcal{L}^1(\rho)\},$$

with the convention that the infimum over the empty set is $+\infty$. Note that g^C is universally measurable, provided that C and g are measurable, see e.g. [29, Proposition 7.47].

To set the stage for the main result of this section, we introduce the following assumption under which we will work in most parts of this section.

Assumption 2.1. *Let $p \in [1,\infty)$, $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}_p(Y)$. Assume that the cost function $C : X \times \mathcal{P}_p(Y) \to (-\infty,\infty]$ is measurable, $\rho \mapsto C(x,\rho)$ is convex and lsc for the p-weak convergence on $\mathcal{P}_p(Y)$, and lower bounded in the sense that $C(x,\rho) \geq a_\ell(x) + \rho(b_\ell)$ for some $a_\ell \in \mathcal{L}^1(\mu), b_\ell \in \mathcal{L}^1(\nu)$. Moreover, we assume that $\mathrm{WT}_C(\mu,\nu) < \infty$.*

Theorem 2.2. *Under Theorem 2.1, we have the following:*

(i) *(primal attainment) $\mathrm{WT}_C(\mu,\nu)$ is attained.*

(ii) *(duality) It holds $\mathrm{WT}_C(\mu,\nu) = \mathrm{D}_C(\mu,\nu)$ and*

$$\mathrm{D}_C(\mu,\nu) = \sup\{\mu(g^C) + \nu(g) : g \in \mathcal{L}^1(\nu)\}. \tag{2.2}$$

(iii) *(dual attainment) If C satisfies conditions (B) and (C), then $\mathrm{D}_C(\mu,\nu)$ is attained.*

(iv) *(complementary slackness) A pair of candidates $(\pi,(f,g))$ is optimal if and only if*

$$C(x,\pi_x) = f(x) + \pi_x(g), \quad \mu\text{-a.s.} \tag{2.3}$$

In this case, $f(x) = g^C(x)$ μ-a.s.

Note that this is a slightly more general version of Theorem 1.2 stated in the introduction.

Proof. The primal attainment result and duality is proved in Theorem 2.6 below. Dual attainment is shown in Theorem 2.8 below. Finally, the complementary slackness result follows from Theorem 2.7 below. □

Remark 2.3. For the no-duality-gap result it is sufficient to work with bounded Borel functions f, g. Then there is no potential ambiguity in how to defined admissibility, namely $f(x) + \rho(g) \leq C(x, \rho)$ for all pairs $(x, \rho) \in X \times \mathcal{P}(Y)$.

However, for dual attainment, it is crucial to allow for unbounded functions which may also take the value $-\infty$. Allowing for such functions requires to deal with the fact that $f(x) + \rho(g)$ is not always well-defined, e.g. if $\rho(g^+)$ and $\rho(g^-)$ are both infinite or if $f(x) = -\infty$ and $\rho(g) = +\infty$.

There a several ways to deal with this, we chose to require the inequality $f(x) + \rho(g) \leq C(x, \rho)$ only for $\rho \in \mathcal{P}(Y)$ such that $\rho(g)$ is well-defined and finite. The continuity condition (C) for the cost implies that this choice for which pairs (x, ρ) we require the inequality $f(x) + \rho(g) \leq C(x, \rho)$ is not sensitive: One could require it for even more pairs (x, ρ) (e.g. whenever the left hand side makes sense in $[-\infty, +\infty]$) or for less pairs (e.g. only for ρ such that $g \in \mathcal{L}^\infty(\rho)$).

2.1 Primal attainment and duality

The C-transform is a key tool for deriving structural properties of (dual) optimizers to the weak transport problem. Before proving duality, we derive some basic properties of the C-transform.

Lemma 2.4. *Let $C : X \times \mathcal{P}_p(Y) \to [-\infty, +\infty]$ be Borel and suppose that $\mathrm{WT}_C(\mu, \nu) < \infty$. The C-transform has the following properties:*

(i) For every $(f, g) \in \Phi_C(\mu, \nu)$, we have $(g^C, g) \in \Phi_C(\mu, \nu)$ and $g^C \geq f$.

(ii) We can restrict the dual problem to admissible pairs of the form (g^C, g), i.e.,

$$\mathrm{D}_C(\mu, \nu) = \sup\{\mu(g^C) + \nu(g) : g \in \mathcal{L}^1(\nu)\}$$

(iii) If (f, g) is a dual optimizer, so is (g^C, g). In this case, $f = g^C$ μ-a.s.

Proof. To prove ((i)), let $(f, g) \in \Phi_C(\mu, \nu)$. Fix $x \in X$. For $\rho \in \mathcal{P}_p(Y)$ such that $g \in \mathcal{L}^1(\rho)$, we have $f(x) \leq C(x, \rho) - \rho(g)$. Taking the infimum over all such ρ yields $f(x) \leq g^C(x)$.

Next, we need to show that $g^C \in \mathcal{L}^1(\mu)$. Since $\mathrm{WT}_C(\mu, \nu) < \infty$, there exists a coupling $\pi \in \Pi(\mu, \nu)$ with $\int C(x, \pi_x)\, \mu(dx) < \infty$. As $g \in \mathcal{L}^1(\nu)$, we have

$$\infty > \int |g(y)|\, \nu(dy) = \int_X \int_Y |g(y)|\, \pi_x(dy)\, \mu(dx),$$

which implies $g \in \mathcal{L}^1(\pi_x)$ μ-a.s. Hence,

$$f(x) \leq g^C(x) \leq C(x, \pi_x) - \pi_x(g) \tag{2.4}$$

for μ-a.e. x. As the function $x \mapsto C(x, \pi_x) - \pi_x(g)$ is μ-integrable, this yields the claim.

The assertion ((ii)) follows directly from ((i)).

To show ((iii)), let (f, g) be a dual optimizer. Note that by ((i)) the pair (g^C, g) is admissible and $\mu(f)+\nu(g) \leq \mu(g^C)+\nu(g)$. Since (f, g) is optimal, so is (g^C, g) and we have $\mu(f) = \mu(g^C)$. Hence, as $g^C \geq f$ and $\mu(f) = \mu(g^C)$ we conclude that $f = g^C$ μ-a.s. □

Lemma 2.5. *Let $C : X \times \mathcal{P}_p(Y) \to [-\infty, +\infty]$ be Borel. We have $\mathrm{D}_C(\mu, \nu) \leq \mathrm{WT}_C(\mu, \nu)$.*

Proof. Wlog assume that $\mathrm{WT}_C(\mu, \nu) < +\infty$. Let $(f, g) \in \Phi_C(\mu, \nu)$ and $\pi \in \Pi(\mu, \nu)$ with finite cost. As in the proof of Lemma 2.4 ((i)) we find that μ-a.s.

$$f(x) \leq C(x, \pi_x) - \pi_x(g).$$

Integrating both sides w.r.t. μ and rearranging terms yield

$$\mu(f) + \nu(g) \leq \int C(x, \pi_x)\, \mu(dx).$$

Taking the infimum over $\pi \in \Pi(\mu, \nu)$ with finite cost and the supremum over $(f, g) \in \Phi_C(\mu, \nu)$ yields the claim. □

The first step of the proof of Theorem 2.2 is to derive the no duality gap result in Theorem 2.6. Note that this extends [15, Theorem 3.1] by requiring only that C is convex lsc in the second component but not jointly lsc. As mentioned above, this is useful to derive the structure theorem for entropic optimal transport in its usual generality in Section 4.

Theorem 2.6. *Under Theorem 2.1, we have primal attainment of $\mathrm{WT}_C(\mu, \nu)$ and*

$$\mathrm{WT}_C(\mu, \nu) = \mathrm{D}_C(\mu, \nu) = \sup\{\mu(g^C) + \nu(g) : g \in \mathcal{L}^1(\nu)\}.$$

Proof. Since the lower bound is merely Borel measurable rather than continuous or upper semicontinuous, some preliminary considerations are required before presenting the core argument. First, note that we may wlog assume $b_\ell \leq 0$. Next, write d_Y for the (complete, separable) metric on Y which is compatible with the Polish topology on Y, denoted by τ_Y. By [57, Theorem 13.11] there exists a Polish topology $\tilde{\tau}_Y \supseteq \tau_Y$ such that b_ℓ is $\tilde{\tau}_Y$-continuous and the Borel σ-algebra of τ_Y and $\tilde{\tau}_Y$ coincide. Let $\tilde{d}_Y$ be a bounded, complete, separable metric on Y that is compatible with $\tilde{\tau}_Y$. Define the metric $\hat{d}_Y(y_1, y_2) := \max(d_Y(y_1, y_2), \tilde{d}_Y(y_1, y_2))$, which is complete, separable and compatible with $\tilde{\tau}_Y$. Denote the associated p-Wasserstein distance by $\hat{\mathcal{W}}_p$. Since $\mathcal{W}_p \leq \hat{\mathcal{W}}_p$ and as b_ℓ is continuous and non-positive, we find that

$$\tilde{C}(x, \rho) := C(x, \rho) - a_\ell(x) - \rho(b_\ell),$$

is measurable, non-negative and $\rho \mapsto \tilde{C}(x, \rho)$ is convex and $\hat{\mathcal{W}}_p$-lsc. We proceed to show the assertion for the cost $\tilde{C}$ and remark that then the claim also follows for the cost C.

We endow $X \times Y$ with the metric $d((x, y), (x', y')) = ((d_X(x, x') \wedge 1)^p + \hat{d}_Y(y, y')^p)^{\frac{1}{p}}$. Consider

$$\Pi := \{\pi \in \mathcal{P}_p(X \times Y) : \pi(dx \times Y) = \mu(dx)\},$$

and observe that the usual weak convergence and the weak convergence that arises from testing against Caratheodory functions[4] coincide on Π by [20, Lemma A.10]. Hence, [20, Proposition A.12 (d)] yields that

$$\Pi \ni \pi \mapsto I_{\tilde{C}}[\pi] := \int \tilde{C}(x, \pi_x)\, \mu(dx)$$

is lsc for the p-weak convergence on $\mathcal{P}_p(X \times Y)$. In particular, if $(\nu_n)_{n\in\mathbb{N}}$ converges to ν in $\hat{\mathcal{W}}_p$ and $(\pi_n)_{n\in\mathbb{N}}$ is a sequence of couplings with $\pi_n \in \Pi(\mu, \nu_n)$, the latter sequence is even relatively compact in $\mathcal{P}_p(X \times Y)$. Consequently, by potentially passing to a subsequence, we may assume that $\pi_n \to \pi$ in p-convergence on $\mathcal{P}_p(X \times Y)$ where $\pi \in \Pi(\mu, \nu)$. Assuming that π_n was $\frac{1}{n}$-optimal for $\mathrm{WT}_{\tilde{C}}(\mu, \nu_n)$ we obtain that

$$\liminf_{n\to\infty} \mathrm{WT}_{\tilde{C}}(\mu, \nu_n) = \liminf_{n\to\infty} I_{\tilde{C}}[\pi_n] \geq I_{\tilde{C}}[\pi] \geq \mathrm{WT}_{\tilde{C}}(\mu, \nu).$$

Hence, $\nu \mapsto \mathrm{WT}_{\tilde{C}}(\mu, \nu)$ is $\hat{\mathcal{W}}_p$-lsc. By compactness of $\Pi(\mu, \nu)$ this yields that $\mathrm{WT}_{\tilde{C}}(\mu, \nu)$ is attained. Moreover, the lower semicontinuity of $\nu \mapsto \mathrm{WT}_{\tilde{C}}(\mu, \nu)$, allows us to follow line by line [14, Proof of Theorem 3.1] and obtain

$$\mathrm{WT}_{\tilde{C}}(\mu, \nu) = \sup \mu(g^{\tilde{C}}) + \nu(g),$$

where the supremum is taken over all $g \in C((Y, \hat{d}_Y))$ with at most p-growth that are bounded from above. In particular, $g \in \mathcal{L}^1(\nu)$ and $\mu(g^{\tilde{C}})$ is well-defined with value in $[-\infty, +\infty)$. Now, the assertion follows by the first part of the proof. □

2.2 Complementary slackness

Proposition 2.7. *Suppose that Theorem 2.1 is satisfied. Then $\pi \in \Pi(\mu, \nu)$ and $(f, g) \in \Phi_C(\mu, \nu)$ are both optimal if and only if they satisfy the complementary slackness condition*

$$C(x, \pi_x) = f(x) + \pi_x(g), \quad \mu\text{-a.s.} \tag{2.5}$$

Proof. Let $\pi \in \Pi(\mu, \nu)$ and $(f, g) \in \Phi_C(\mu, \nu)$ both be optimal. As $g \in \mathcal{L}^1(\nu)$, we have $g \in \mathcal{L}^1(\pi_x)$ μ-a.s. By admissibility of (f, g) we have $f(x) + \pi_x(g) \leq C(x, \pi_x)$ μ-a.s. Since π and (f, g) are both optimal, it follows

$$\int C(x, \pi_x) - f(x) - \pi_x(g)\, \mu(dx) = 0.$$

Hence, $f(x) + \pi_x(g) = C(x, \pi_x)$ μ-a.s.

Suppose now that $\pi \in \Pi(\mu, \nu)$ and $(f, g) \in \Phi_C(\mu, \nu)$ satisfy (2.5). By Theorem 2.5 we have

$$\mu(f) + \nu(g) \leq \mathrm{WT}_C(\mu, \nu) \text{ and } \mathrm{D}_C(\mu, \nu) \leq \int C(x, \pi_x)\, \mu(dx).$$

Hence, the assertion follows as $\int C(x, \pi_x)\, \mu(dx) = \mu(f) + \nu(g)$. □

[4] A Caratheodory function is a function $h : X \times Y \to \mathbb{R}$ such that $h(x, \cdot)$ is continuous for every $x \in X$ and $h(\cdot, y)$ is Borel for every $y \in Y$. Caratheodory functions are jointly Borel.

2.3 Dual attainment

The last major step in the proof of the fundamental theorem of weak optimal transport is the existence of dual optimizers.

Theorem 2.8. *Suppose that $C : X \times \mathcal{P}_p(Y) \to \mathbb{R}$ satisfies conditions* (B) *and* (C). *Then there is a dual optimizer $(f, g) \in \Phi_C(\mu, \nu)$.*

The proof of Theorem 2.8 uses an argument based on Komlós' lemma. In order to apply this technique, we first need to observe that the set of admissible dual potentials is convex.

Lemma 2.9. *Suppose that C satisfies condition* (C). *Then $\Phi_C(\mu, \nu)$ is convex.*

Proof. Let $(f_1, g_1), (f_2, g_2) \in \Phi_C(\mu, \nu)$ and $t \in [0, 1]$ be given. Write $f = (1-t)f_1 + tf_2$ and $g = (1-t)g_1 + tg_2$. In order to show that $(f, g) \in \Phi_C(\mu, \nu)$, pick $x \in X$ and $\rho \in \mathcal{P}(Y)$ such that $g \in \mathcal{L}^1(\rho)$. Note that this does in general not imply that $g_1, g_2 \in \mathcal{L}^1(\rho)$ but it implies that $g_1, g_2 > -\infty$ ρ-a.s. Hence, the sets $Y_n := \{|g_1| \le n\} \cap \{|g_2| \le n\}$ satisfy $\rho(Y_n) \to 1$. Using dominated convergence, $g_1, g_2 \in \mathcal{L}^1(\frac{1}{\rho(Y_n)}\rho|_{Y_n})$ and condition (C) we find

$$\begin{aligned} f(x) + \rho(g) &= \lim_n f(x) + \frac{1}{\rho(Y_n)}\rho|_{Y_n}(g) \\ &\le \limsup_n (1-t)\Big(f_1(x) + \frac{1}{\rho(Y_n)}\rho|_{Y_n}(g_1)\Big) + \limsup_n t\Big(f_2(x) + \frac{1}{\rho(Y_n)}\rho|_{Y_n}(g_2)\Big) \\ &\le \limsup_n C\Big(x, \frac{1}{\rho(Y_n)}\rho|_{Y_n}\Big) \le C(x, \rho). \end{aligned}$$

□

Lemma 2.10. *Let $h : [0, \infty) \to [0, \infty)$ be convex, increasing and super-coercive, $b \in \mathcal{L}^1(\nu)$ and $\alpha \in \mathbb{R}$. Then there exist constants K_1, K_2 (depending on α, b and h) such that the following holds: If $g \in \mathcal{L}^1(\nu)$ satisfies $\int g \, d\nu = 0$ and*

$$\int (g - b)\xi \, d\nu \le \alpha + \int h(\xi) \, d\nu \tag{2.6}$$

for every $\xi \in \mathcal{L}^\infty(\nu)$ with $\xi \ge 0$ and $\int \xi \, d\nu = 1$, then we have

$$\int h^*((g - b)^+ - K_1) \, d\nu \le K_2. \tag{2.7}$$

Proof. We assume that ν has no atoms. This is without loss of generality and will simplify the notation in the proof. Further, we can assume wlog that $\alpha = 0$ and $\int b \, d\nu = 0$. Otherwise replace h with $h + \alpha_+ + (\int b \, d\nu)_+$. As $(h + \alpha_+ + (\int b \, d\nu)_+)^* = h^* - \alpha_+ - (\int b \, d\nu)_+$, this replacement just causes a change of the constants K_1 and K_2 in (2.7). Moreover, we introduce the shorthand notation $\tilde{g} := g - b$.

The first step is to derive a bound on $\int \tilde{g}^+ \, d\nu$. To that end, we distinguish two cases depending on the value of $\beta := \nu(\tilde{g} \ge 0)$: In the case that $\beta \ge \frac{1}{3}$, we choose the test function $\xi = \beta^{-1} 1_{\{\tilde{g} \ge 0\}}$ in (2.6). Using that h is increasing we find

$$\int \tilde{g}^+ \, d\nu = \beta \int \tilde{g}\xi \, d\nu \le \beta \int h(\xi) \, d\nu = \beta(1 - \beta)h(0) + \beta^2 h(1/\beta) \le h(3).$$

To tackle the case $\beta < \frac{1}{3}$, choose $t < 0$ such that the set $A := \{\tilde{g} \in (t, 0)\}$ satisfies $\nu(A) = \frac{1}{3}$. Further, write $B := \{\tilde{g} < 0\} \setminus A$ and note that $\nu(B) \geq 1 - \beta - \frac{1}{3} \geq \frac{1}{3} = \nu(A)$. As $\tilde{g} \leq t$ on B, $\tilde{g} \geq t$ on A and $\int \tilde{g}\, d\nu = 0$, we find

$$\int \tilde{g}^+ \, d\nu = -\int_{A\cup B} \tilde{g}\, d\nu \geq -2 \int_A \tilde{g}\, d\nu.$$

By using the test function $\xi = (\beta + \frac{1}{3})^{-1} 1_{B^c}$ in (2.6) we derive

$$\begin{aligned}\int \tilde{g}^+ \, d\nu &\leq 2\left(\int \tilde{g}^+ \, d\nu + \int_A \tilde{g}\, d\nu\right) = 2\left(\beta + \frac{1}{3}\right) \int \tilde{g}\xi \, d\nu \\ &\leq 2\left(\beta + \frac{1}{3}\right) \int h(\xi)\, d\nu \leq 2\left(\beta + \frac{1}{3}\right) h(3) \leq 2h(3).\end{aligned}$$

Combining both cases and using again that $\int \tilde{g}\, d\nu = 0$, we find

$$\|g^-\|_{L^1(\nu)} = \|g^+\|_{L^1(\nu)} \leq 2h(3). \tag{2.8}$$

In the next step, we show that for every non-negative, bounded Borel function ξ

$$\int \tilde{g}^+ \xi \, d\nu \leq \int h(\xi) + h(0)\xi \, d\nu + 3h(3). \tag{2.9}$$

To this end, we distinguish two cases:

Case 1: If $\|\xi 1_{\{\tilde{g}\geq 0\}}\|_{L^1(\nu)} \geq 1$, we set $\tilde{\xi} = \|\xi 1_{\{\tilde{g}\geq 0\}}\|^{-1}_{L^1(\nu)} \xi 1_{\{\tilde{g}\geq 0\}}$. By first applying (2.6) with the test function $\tilde{\xi}$ and then using that h is increasing, convex, and non-negative, we find

$$\begin{aligned}\int \tilde{g}^+ \xi \, d\nu &= \|\xi 1_{\{\tilde{g}\geq 0\}}\|_{L^1(\nu)} \int \tilde{g}\tilde{\xi} \, d\nu \leq \|\xi 1_{\{\tilde{g}\geq 0\}}\|_{L^1(\nu)} \int h(\tilde{\xi})\, d\nu \\ &\leq \|\xi 1_{\{\tilde{g}\geq 0\}}\|_{L^1(\nu)} h(0) + \int h(\xi)\, d\nu \leq \int h(\xi) + h(0)\xi \, d\nu.\end{aligned}$$

Case 2: If $\|\xi 1_{\{\tilde{g}\geq 0\}}\|_{L^1(\nu)} < 1$, there is $r \in (0, 1]$ such that $\tilde{\xi} := (\xi 1_{\{\tilde{g}\geq 0\}}) \vee r$ satisfies $\|\tilde{\xi}\|_{L^1(\nu)} = 1$. Applying (2.6) with the test function $\tilde{\xi}$, equation (2.8), and the fact that h is increasing convex, yields the estimate

$$\int \tilde{g}^+ \xi \, d\nu \leq \int \tilde{g}\tilde{\xi} \, d\nu + r\|\tilde{g}^-\|_{L^1(\nu)} \leq \int h(\tilde{\xi})\, d\nu + 2rh(3) \leq \int h(\xi)\, d\nu + 3h(3).$$

This completes the proof of (2.9).

The next step is to use (2.9) to derive a bound on $\int h^*(\tilde{g}^+ - h(0))\, d\nu$. Formally by exchanging integration with supremum, we have

$$\begin{aligned}\int h^*(\tilde{g}^+ - h(0))\, d\nu &= \int \sup_{z\in[0,\infty)} \tilde{g}^+(y)z - h(0)z - h(z)\, \nu(dy) \\ &= \sup_{\xi\geq 0 \text{ Borel}} \int \tilde{g}^+(y)\xi(y) - h(0)\xi(y) - h(\xi(y))\, \nu(dy) \leq 3h(3).\end{aligned}$$

In order to rigorously justify this calculation, let g be a selector of ∂h^*, i.e., $g(t) \in \partial h^*(t)$ for every $t \in [0,\infty)$. This selector is well-defined (as h is super-coercive, $\mathrm{dom}(h^*) = [0,\infty)$ and thus $\partial h^*(t) \neq \emptyset$ for every t), monotone and in particular Borel. Moreover, we have $tg(t) = h(g(t)) + h^*(t)$. For every $n \in \mathbb{N}$, let $\xi_n = g((\tilde{g}^+ - h(0)) \wedge n)$ and note that $0 \leq \xi_n \leq g(n)$, so we can apply (2.9) and get

$$\int h^*((\tilde{g}^+ - h(0)) \wedge n)\, d\nu = \int ((\tilde{g}^+ - h(0)) \wedge n)\xi_n - h(\xi_n)\, d\nu \leq 3h(3).$$

The integral $\int [h^*(\tilde{g}^+ - h(0))]^-\, d\nu$ is well-defined, because $\tilde{g}^+ \in \mathcal{L}^1(\nu)$ and h^* is convex and thus bounded from below by an affine function. As h^* is increasing (cf. Section A.1), we can therefore use monotone convergence to obtain that $\int h^*(\tilde{g}^+ - h(0))\, d\nu \leq 3h(3)$. □

Proof of Theorem 2.8. Let $(f_n, g_n)_n$ be a maximizing sequence for the dual problem, that is, $(f_n, g_n) \in \Phi_C(\mu,\nu)$ and $\mu(f_n) + \nu(g_n) \nearrow \mathrm{D}_C(\mu,\nu)$. Replacing (f_n, g_n) by $(f_n + c, g_n - c)$ where $c \in \mathbb{R}$ leaves value and admissibility unchanged. Therefore, we can assume that $\nu(g_n) = 0$. By applying the admissibility condition (2.1) with $\rho = \nu$, we find the upper bound

$$f_n(x) \leq C(x,\nu) \leq a(x) + \int b\, d\nu + h(1).$$

Hence, the sequence $(f_n)_n$ is both bounded from above by an integrable function and satisfies $\mu(f_n) \to \mathrm{D}_C(\mu,\nu)$. This yields that $\sup_n \|f_n\|_{L^1(\mu)} < \infty$.

For this reason, we can apply Komlós' lemma [59]. It yields the existence of a sequence of convex combinations $(\hat{f}_n)_n$ of $(f_n)_n$ and $f \in \mathcal{L}^1(\mu)$ such that $\hat{f}_n \to f$ μ-a.s. We write $(\hat{g}_n)_n$ for the sequence of convex combinations of $(g_n)_n$ that result from using the same weights as for $(\hat{f}_n)_n$. By Theorem 2.9, $(\hat{f}_n, \hat{g}_n)$ is admissible as well. For notational simplicity, we rename $(\hat{f}_n, \hat{g}_n)$ to (f_n, g_n).

Next, fix some $x_0 \in X$ for which $f_n(x_0) \to f(x_0)$. As (f_n, g_n) is admissible, we get

$$\int g_n\, d\rho \leq C(x_0,\rho) - f_n(x_0) \leq a(x_0) - f_n(x_0) + \int b\, d\rho + \int h\Big(\frac{d\rho}{d\nu}\Big)\, d\nu,$$

for every $\rho \in \mathcal{P}_p(Y)$ with $\|\frac{d\rho}{d\nu}\|_{L^\infty(\nu)} < \infty$. As $(f_n(x_0))_n$ is convergent, there is $\alpha \in \mathbb{R}$ such that

$$\int g_n - b\, d\rho \leq \alpha + \int h\Big(\frac{d\rho}{d\nu}\Big)\, d\nu. \tag{2.10}$$

Hence, for every bounded $\xi \geq 0$ with $\int \xi\, d\nu = 1$, we have

$$\int (g_n - b)\xi\, d\nu \leq \alpha + \int h(\xi)\, d\nu. \tag{2.11}$$

In case that h is not supercoercive, we may replace h by $h + |\cdot|^2$. By Lemma 2.10, there exist constants K_1, K_2 such that $\int h^*((g_n - b)^+ - K_1)\, d\nu \leq K_2$ for every $n \in \mathbb{N}$. By Theorem A.1, h^* is super-coercive, thus the de La Vallée Poussin criterion for uniform integrability implies that the sequences $((g_n - b)^+ - K_1)_n$ and, in particular, $(g_n^+)_n$ are uniformly integrable.

As $\int g_n \, d\nu = 0$, the sequence $(g_n)_n$ is bounded in $L^1(\nu)$. Therefore, Komlós' lemma guarantees the existence of $g \in \mathcal{L}^1(\nu)$ and a sequence $(\hat{g}_n)_n$ consisting of forward convex combinations of $(g_n)_n$ such that $\hat{g}_n \to g$ ν-a.s. As before, we write $(\hat{f}_n)_n$ for the sequence of convex combinations of $(f_n)_n$ obtained using the same weights as for $(\hat{g}_n)_n$. To simplify notation, we again rename the sequences $(\hat{g}_n)_n$ and $(\hat{f}_n)_n$ as $(g_n)_n$ and $(f_n)_n$, respectively.Using Fatou's Lemma we obtain

$$\int f \, d\mu + \int g \, d\nu \geq \limsup_n \int f_n \, d\mu + \limsup_n \int g_n \, d\nu = \mathrm{D}_C(\mu, \nu). \tag{2.12}$$

It remains to show that the pair (f, g) is admissible. By Egorov's theorem, there is an increasing sequence $(K_N)_N$ of compact subsets of Y such that $\nu(\bigcup_N K_N) = 1$ and $g_n \to g$ uniformly on K_N for every $N \in \mathbb{N}$. We redefine f to be $-\infty$ on the μ-null set where $(f_n)_n$ does not converge to f. Similarly, we change g to $-\infty$ on the ν-null set $(\bigcup_N K_N)^C$.

To conclude that (f, g) is an optimal dual pair, it suffices to show that (f, g) is admissible. Specifically, we need to show that for $(x, \rho) \in X \times \mathcal{P}_p(Y)$ satisfying $f(x) \in \mathbb{R}$ and $g \in \mathcal{L}^1(\rho)$ we have

$$f(x) + \rho(g) \leq C(x, \rho).$$

First suppose that there is an $N \in \mathbb{N}$ such that $\rho(K_N) = 1$. As $g \in \mathcal{L}^1(\nu)$ and $g_n \to g$ uniformly on a set of full measure, we have $g_n \in \mathcal{L}^1(\nu)$ and

$$f(x) + \rho(g) = \lim_n f_n(x) + \rho(g_n) \leq C(x, \rho).$$

Suppose next that $\rho(\bigcup_N K_N) = 1$. Define $\rho_N := \frac{1}{\rho(K_N)} \rho_{|K_N}$ and note that $g \in \mathcal{L}^1(\rho_N)$. With the same reasoning as in the previous case, we have $f(x) + \rho_N(g) \leq C(x, \rho_N)$. By the dominated convergence theorem, we have $\rho_N(g) \to \rho(g)$. Invoking the continuity property (C) yields

$$f(x) + \rho(g) = \lim_N f(x) + \rho_N(g) \leq \limsup_N C(x, \rho_N) \leq C(x, \rho).$$

In the case that $\rho(\bigcup_N K_N) < 1$, we have $g \notin \mathcal{L}^1(\rho)$, so there is nothing to show. □

Remark 2.11. If the cost function C satisfies the stronger growth condition $C(x, \rho) \leq a(x) + \rho(b)$, for some $a \in \mathcal{L}^1(\mu)$, $b \in \mathcal{L}^1(\nu)$, then the argument for the uniform integrability of the maximizing sequence $(f_n, g_n)_n$ in the proof of Theorem 2.8 simplifies, namely there exists a constant α such that $f_n \leq \alpha + a$ and $g_n \leq \alpha + b$. In particular, there are dual maximizers (f, g) which satisfy $f \leq \alpha + a$ and $g \leq \alpha + b$. ◇

Example 2.12 (Necessity of assumption (C)). Even in the case of a one element space X and compact Y we find a bounded, convex and lsc cost function that does not satisfy assumption (C) and for which we have no dual attainment. Let $X = \{0\}, Y = [0, 1]$ and take $\sigma \in \mathcal{P}([0, 1])$ with $\sigma \ll \lambda$ and $\|\frac{d\sigma}{d\lambda}\|_\infty = \infty$, where λ denotes the Lebesgue measure. Consider the cost function

$$C(0, \rho) := C(\rho) := 1 - \frac{1}{\|\frac{d\sigma}{d\rho}\|_\infty}, \tag{2.13}$$

where we convene that $\|\frac{d\sigma}{d\rho}\|_\infty = \infty$ if σ is not absolutely continuous w.r.t. ρ. Observe that C does not satisfy (C). Indeed, consider Borel sets $Y_n \subset [0,1]$ with $\sigma(Y_n) < 1$ and $\bigcup_n Y_n = [0,1]$. As σ is not absolutely continuous w.r.t. σ_n, we have $C(\frac{1}{\sigma(Y_n)}\sigma_{|Y_n}) = 1 > 0 = C(\sigma)$, showing that C does not satisfy Assumption (C).

Assume for the sake of contradiction that we have dual attainment, i.e., there is $g : [0,1] \to [-\infty, \infty)$ satisfying $\int g\, d\lambda = 1 = \mathrm{WT}_C(\delta_0, \lambda)$ and $\int g\, d\rho \leq C(\rho)$ for all $\rho \in \mathcal{P}([0,1])$ with $g \in \mathcal{L}^1(\rho)$. Choosing $\rho = \delta_y$ implies $g(y) \leq C(\delta_y) = 1$ for every $y \in [0,1]$ with $g(y) > -\infty$ and, thus, $g \leq 1$. Next, choosing $\rho = \sigma$ shows that $\int g\, d\sigma \leq C(\sigma) = 0$. Consider the set $A := \{-\infty < g < 1\}$. On the one hand, $\int g\, d\lambda = 1$ as well as $g \leq 1$ imply $\lambda(A) = 0$. On the other hand, $\int g\, d\sigma \leq 0$ implies $\sigma(A) > 0$. This is a contradiction to $\sigma \ll \lambda$. Therefore, there are no dual optimizers which shows that Assumption (C) cannot be omitted in Theorem 1.2. ◇

We conclude this part with a result that guarantees the existence of particularly regular dual maximizers, provided that C is Lipschitz in its second argument.

Corollary 2.13. *Suppose that Theorem 2.1 is satisfied with $p = 1$ and that $C(x, \cdot)$ is L-Lipschitz w.r.t. $\mathcal{W}_1$-distance. Then we have*

$$\mathrm{WT}_C(\mu,\nu) = \max_{\substack{(f,g)\in\Phi_C(\mu,\nu)\\ g\ L\text{-Lipschitz}}} \mu(f) + \nu(g).$$

Proof. Let $(f,g) \in \Phi_C(\mu,\nu)$ and define the L-Lipschitz function $\tilde{g}$ by

$$\tilde{g}(y) = \sup_{z\in Y} g(z) - L d_Y(y,z).$$

Fix $\varepsilon > 0$ and pick a measurable map $T : Y \to Y$ such that $\tilde{g}(y) \leq g(T(y)) - L d_Y(y, T(y)) + \varepsilon$. Note that $g \leq \tilde{g}$. Therefore, the claim follows immediately from Theorem 2.8 if we can show that $(f, \tilde{g})$ is also admissible.[5] To this end, fix $(x,\rho) \in X \times \mathcal{P}_1(Y)$ and write $\tilde{\rho} = T_\# \rho$. Using admissibility of (f,g) we get

$$f(x) + \rho(\tilde{g}) \leq f(x) + \tilde{\rho}(g) - L\mathcal{W}_1(\rho,\tilde{\rho}) + \varepsilon \leq C(x,\tilde{\rho}) - L\mathcal{W}_1(\rho,\tilde{\rho}) + \varepsilon \leq C(x,\rho) + \varepsilon.$$

As $\varepsilon > 0$ was arbitrary, $(f,\tilde{g})$ is admissible. □

2.4 C-monotonicity of optimal couplings

The dual attainment result allows us to easily recover, under the current set of assumptions, the result from [14, Theorem 5.3] that C-monotonicity is a necessary optimality criterion.

Definition 2.14. *A set $\Gamma \subset X \times \mathcal{P}(Y)$ is called C-monotone if for every $(x_1,\rho_1),\dots,(x_n,\rho_n) \in \Gamma$ and all $\tilde{\rho}_1,\dots,\tilde{\rho}_n \in \mathcal{P}(Y)$ with $\sum_{i=1}^n \rho_i = \sum_{i=1}^n \tilde{\rho}_i$ it holds*

$$\sum_{i=1}^n C(x_i,\rho_i) \leq \sum_{i=1}^n C(x_i,\tilde{\rho}_i).$$

A coupling $\pi \in \Pi(\mu,\nu)$ is called C-monotone if $\mu(\{x \in X : (x,\pi_x) \in \Gamma\}) = 1$.

[5] Indeed in the present Lipschitz case we could avoid Theorem 2.8 by invoking an Arzelà–Ascoli argument.

Corollary 2.15. *Suppose that Theorem 2.1,* (B) *and* (C) *are satisfied. Then every optimal* $\pi \in \Pi(\mu,\nu)$ *is* C*-monotone.*

Proof. Let π be a primal and (f,g) a dual optimizer. Set

$$\Gamma := \{(x,\rho) \in X \times \mathcal{P}_p(Y) : C(x,\rho) = f(x) + \rho(g)\}.$$

The complementary slackness criterion (see Theorem 2.7) yields $\mu(\{x \in X : (x,\pi_x) \in \Gamma\}) = 1$. To see that Γ is C-monotone, let $(x_1,\rho_1),\dots,(x_n,\rho_n) \in \Gamma$ and let $\tilde{\rho}_1,\dots,\tilde{\rho}_n \in \mathcal{P}_p(Y)$ be competitors satisfying $\sum_{i=1}^n \rho_i = \sum_{i=1}^n \tilde{\rho}_i$. Then

$$\sum_{i=1}^n C(x_i,\rho_i) = \sum_{i=1}^n f(x_i) + \rho_i(g) = \sum_{i=1}^n f(x_i) + \tilde{\rho}_i(g) \le \sum_{i=1}^n C(x_i,\tilde{\rho}_i). \qquad \square$$

We notice that C-monotonicity requires relatively strong assumptions in order to be a sufficient condition for optimality. Indeed, even if we are in the classical "linear" OT case $C(x,\rho) = \int c(x,y)\,\rho(dy)$, C-monotonicity does not imply c-cyclical monotonicity see [8, Example 5.2]. Moreover, while C-monotonicity as a necessary condition for optimality is used for instance in [13, 49], we are not aware of applications of C-monotonicity as a sufficient condition in WOT and do not pursue it further in this paper.

2.5 Non convex costs and duality for relaxed WOT

As noted in the introduction, duality can fail if C is not convex in the second argument, even in otherwise very regular situations:

Example 2.16. Let $(X,\mu) = (\{0\},\delta_0)$ and $(Y,\nu) = (\{-1,+1\},(\delta_{-1}+\delta_{+1})/2)$. Consider the cost $C(0, m\delta_{-1} + (1-m)\delta_{+1}) = 1 - |2m-1|$. Then the primal value is 1 while the dual value is 0. $\diamond$

A possible remedy is to suitably enrich the space of admissible transport plans, by allowing for initial randomization. This can be done either in measure theoretic or in probabilistic language:

1. *Probabilistically*, the primal weak transport problem can be formulated as

$$\mathrm{WT}_C(\mu,\nu) = \inf_{X_1\sim\mu, X_2\sim\nu} \mathsf{E}[C(X_1, \mathrm{law}(X_2|X_1))]. \tag{2.14}$$

   The right-hand side of (2.14) can be relaxed by allowing to take the minimum over all adapted processes $(X_t)_{t=1}^2$ defined on some stochastic basis $(\Omega,\mathcal{F},\mathbb{P},(\mathcal{F}_t)_{t=1}^2)$, where $\mathcal{F}_1$ could be larger than $\sigma(X_1)$, i.e.

$$\overline{\mathrm{WT}}_C(\mu,\nu) := \inf_{\substack{(\Omega,\mathcal{F},\mathbb{P},(\mathcal{F}_t)_{t=1}^2,(X_t)_{t=1}^2),\\ X_1\sim\mu,\, X_2\sim\nu}} \mathsf{E}[C(X_1, \mathrm{law}(X_2|\mathcal{F}_1))].$$

2. For the *measure theoretic* formulation, we introduce the set of *lifted transport plans*. Specifically, for $\mu \in \mathcal{P}_p(X), \nu \in \mathcal{P}_p(Y)$, we write $\Lambda(\mu,\nu)$ for the collection of $P \in \mathcal{P}_p(X \times \mathcal{P}_p(Y))$ for which the first marginal is μ (i.e. $\mathrm{pr}_{X\#}P = \mu$) and the intensity

of the second marginal is ν (i.e. $\int \rho(g)\,(\mathrm{pr}_{\mathcal{P}(Y)\#}P)(d\rho) = \nu(g)$ for every $g \in C_b(Y)$). Using this notion, the relaxed weak optimal transport problem can be written as

$$\overline{\mathrm{WT}}_C(\mu,\nu) = \inf_{Q\in\Lambda(\mu,\nu)} \int C(x,\rho)\,Q(dx,d\rho). \tag{2.15}$$

Transport plans can be identified with lifted transport plans using the map

$$J : \Pi(\mu,\nu) \to \Lambda(\mu,\nu), \quad J(\pi) := (x \mapsto (x,\pi_x))_{\#}\mu. \tag{2.16}$$

As every transport plan π induces a lifted plan $J(\pi)$, we always have

$$\overline{\mathrm{WT}}_C(\mu,\nu) \le \mathrm{WT}_C(\mu,\nu). \tag{2.17}$$

Note that lifted transport plans of the form $J(\pi)$ are of Monge type in the sense that they are concentrated on the graph of a function. Hence, $\overline{\mathrm{WT}}_C(\mu,\nu)$ can be considered as a relaxation of $\mathrm{WT}_C(\mu,\nu)$ in the same way as the Kantorovich formulation relaxes the Monge formulation of the transport problem. In particular, it was established in [14, Lemma 2.1] that in the standard setting, where $C(x,\cdot)$ is convex, the value of the weak transport problem and its relaxation coincide, i.e.

$$\overline{\mathrm{WT}}_C(\mu,\nu) = \mathrm{WT}_C(\mu,\nu). \tag{2.18}$$

The relaxed WOT problem is of particular interest in the setting of non-convex cost introduced in Section 1.2. A major reason for this is that primal attainment is guaranteed for $\overline{\mathrm{WT}}_C(\mu,\nu)$, while failing in general for $\mathrm{WT}_C(\mu,\nu)$ in the non-convex case. Further note that the inequality (2.17) is in general strict. However, it was shown in [2, Proposition 3.8] that if $\rho \mapsto C(x,\rho)$ is lsc

$$\overline{\mathrm{WT}}_C(\mu,\nu) = \mathrm{WT}_{\overline{C}}(\mu,\nu), \tag{2.19}$$

where $\overline{C}$ is defined such that $\overline{C}(x,\cdot)$ is the lsc convex hull of $C(x,\cdot)$ for every $x \in X$. Furthermore, [2, Theorem 3.9] states that if μ has no atoms and C is continous with bounded p-growth, then

$$\overline{\mathrm{WT}}_C(\mu,\nu) = \overline{\mathrm{WT}}_{\overline{C}}(\mu,\nu) = \mathrm{WT}_C(\mu,\nu). \tag{2.20}$$

We recall that the initial example of this section showed that duality for the non-relaxed WOT problem fails. However, it is true for its relaxed version, as was shown in [14, Theorem 3.1][6]. We obtain the following variant of the fundamental theorem of WOT.

Theorem 2.17. *Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}_p(Y)$ and let $C : X \times \mathcal{P}_p(Y) \to (-\infty,\infty]$ be measurable and $\rho \mapsto C(x,\rho)$ lsc for the p-weak convergence on $\mathcal{P}_p(Y)$. Suppose that there are $a_\ell \in \mathcal{L}^1(\mu), b_\ell \in \mathcal{L}^1(\nu)$ such that $C(x,\rho) \ge a_\ell(x) + \rho(b_\ell)$, then:*

[6] We note that there it was shown for cost functions that are jointly lsc. The generalization of this proof to cost functions that are jointly Borel and lsc in the second argument follows line by line as in the proof of Theorem 2.6.

(i) *(primal attainment)* $\overline{\mathrm{WT}}_C(\mu,\nu)$ *is attained,*

(ii) *(duality) we have* $\overline{\mathrm{WT}}_C(\mu,\nu) = \mathrm{D}_C(\mu,\nu)$ *where*

$$\mathrm{D}_C(\mu,\nu) = \sup\{\mu(g^C) + \nu(g) : g \in \mathcal{L}^1(\nu) \text{ s.t. } g^C \in \mathcal{L}^1(\mu)\}. \tag{2.21}$$

(iii) *(dual attainment) If* C *satisfies conditions* (B) *and* (C)*, then* $D_C(\mu,\nu)$ *is attained.*

(iv) *(complementary slackness) If* $\overline{\mathrm{WT}}_C(\mu,\nu) < \infty$*, a pair of candidates* $(\pi,(f,g))$ *is optimal if and only if*

$$C(x,\rho) = f(x) + \rho(g), \quad P(dx,d\rho)\text{-a.s.} \tag{2.22}$$

In this case, $f(x) = g^C(x)$ μ*-a.s.*

Remark 2.18. We note that already in the setting of the non-relaxed WOT with convex cost, the duality (2.21) is used to establish the WOT duality (on general Polish spaces). This is because one equips the set of couplings $\Pi(\mu,\nu)$ with the adapted weak topology (see e.g. [11]) which is stronger than the usual weak topology but not compact. Both the set of filtered processes and the set $\Lambda(\mu,\nu)$ are compactifications of $\Pi(\mu,\nu)$. It follows from [17] that the two viewpoints are equivalent. ◇

3 Barycentric costs

A particularly relevant class of cost functions consists of those only depending on the barycenter of the measure ρ. More specifically, we consider the case $X = Y = \mathbb{R}^d$ with cost functions $C : X \times \mathcal{P}_1(Y) \to \mathbb{R}$ of the form

$$C(x,\rho) = \vartheta(x - \mathrm{mean}(\rho)),$$

where $\vartheta : \mathbb{R}^d \to \mathbb{R}$ is convex and $\mathrm{mean}(\rho) = \int y\,\rho(dy)$. Then the weak optimal transport problem with cost C reads as

$$\mathrm{BT}_\vartheta(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \int \vartheta(x - \mathrm{mean}(\pi_x))\,\mu(dx). \tag{3.1}$$

Recall that two probability measures $\eta,\rho \in \mathcal{P}_1(\mathbb{R}^d)$ are said to be in convex order, denoted by $\eta \le_c \rho$ if $\int f\,d\eta \le \int f\,d\rho$ for every convex function $f : \mathbb{R}^d \to \mathbb{R}$. It is easy to observe that (3.1) is closely related to the problem of projecting μ onto $\{\eta : \eta \le_c \nu\}$ w.r.t. the value of the classical transport problem with cost $\vartheta(x-y)$, i.e.

$$\mathrm{BT}_\vartheta(\mu,\nu) = \min_{\eta\le_c\nu} \mathrm{T}_\vartheta(\mu,\eta). \tag{3.2}$$

Specifically, if π is an optimizer for $\mathrm{BT}_\vartheta(\mu,\nu)$, then $\eta := (x \mapsto \mathrm{mean}(\pi_x))_{\#}\mu$ is optimal in the minimum on the right-hand side of (3.2) and $\xi := (x \mapsto (x, \mathrm{mean}(\pi_x)))_{\#}\mu \in \Pi(\mu,\eta)$ is an optimal coupling between μ and η for the cost $c(x,y) = \vartheta(x-y)$. Conversely, if $\eta \le_c \nu$ and $\xi \in \Pi(\mu,\eta)$ are both optimal and κ is any martingale coupling from η to ν,

then the conditionally independent gluing of ξ and κ is optimal for $\mathrm{BT}_\vartheta(\mu,\nu)$. Note that this argument also shows that if η is optimal in (3.2), then there is an optimal Monge coupling between μ and η (without any regularity assumption on μ).

In contrast to the previous sections, we change the sign convention of the dual potential according to the rule

$$(\phi,\psi) := (f,-g).$$

This is because g turns out to be concave in the setting of this section. As we will often use techniques from convex analysis, it significantly simplifies the notation to work with the convex function $\psi = -g$. We state the main result of this section.

Theorem 3.1. *Let $\vartheta : \mathbb{R}^d \to \mathbb{R}$ be convex and assume that there exist $a \in \mathcal{L}^1(\mu)$ and $b \in \mathcal{L}^1(\nu)$ such that $\vartheta(x-y) \le a(x)+b(y)$ for all $x,y \in \mathbb{R}^d$.*

(i) *The dual problem of* (3.1) *is given by*

$$\sup_{\psi \text{ convex, lsc}} \mu(\vartheta\square\psi) - \nu(\psi). \tag{3.3}$$

We have strong duality, primal attainment and dual attainment.

(ii) *We have the following complementary slackness criterion: The coupling $\pi \in \Pi(\mu,\nu)$ and the convex function $\psi \in \mathcal{L}^1(\nu)$ are both optimal if and only if for μ-a.e. x*

$$\vartheta\square\psi(x) = \vartheta(x - \mathrm{mean}(\pi_x)) + \pi_x(\psi).$$

(iii) *If $\pi \in \Pi(\mu,\nu)$ is optimal for $\mathrm{BT}_\vartheta(\mu,\nu)$ and ψ is a dual optimizer, writing $\phi := \vartheta\square\psi$, we have $\mathrm{mean}(\pi_x) \in \partial^\vartheta\phi(x)$ for μ-a.s.*

(iv) *Suppose that ϑ is strictly convex. If $\pi,\tilde\pi \in \Pi(\mu,\nu)$ are both optimal for $\mathrm{BT}_\vartheta(\mu,\nu)$, then $\mathrm{mean}(\pi_x) = \mathrm{mean}(\tilde\pi_x)$ μ-a.s. Moreover, the optimal $\eta \le_c \nu$ and the optimal coupling $\gamma \in \Pi(\mu,\eta)$ mentioned in* (3.2) *are both unique.*

(v) *Suppose that ϑ is strictly convex, differentiable and supercoercive. If ψ and π are optimal, then $\phi := \vartheta\square\psi \in C^1(\mathbb{R}^d)$ and $\mathrm{mean}(\pi_x) = x - \nabla\vartheta^*(\nabla(\phi(x)))$ μ-a.s.*

Note that the growth condition on ϑ is automatically satisfied if μ,ν have bounded support, or more generally, if μ,ν have finite p-th moments and ϑ does not grow faster than $|x|^p$. By abuse of notation, we write $\partial^\vartheta\phi(x)$ to denote the c_ϑ-subdifferential of ϕ at x, where $c_\vartheta(x,y) = \vartheta(x-y)$, that is, the set

$$\partial^\vartheta\phi(x) = \{y \in \mathbb{R}^d : \phi(x) + \vartheta(x-y) \le \phi(z) + \vartheta(z-y),\ \forall z \in \mathbb{R}^d\}.$$

3.1 Applications of Theorem 3.1

Note that Theorem 3.1 is an extension of the Brenier–Strassen theorem in a similar way as the Gangbo–McCann theorem [44, Theorem 1.2] is an extension of Brenier's theorem. We note that (3.2) was already established by Gozlan and Julliet [49] for general convex ϑ; moreover, they establish the existence of dual maximizers in the case of compactly supported measures.

Remark 3.2. We note that applying Theorem 3.1((i)) to a convex function $\vartheta : \mathbb{R}^d \to \mathbb{R}$ which attains its unique minimum in zero, implies Strassen's theorem [79]. To see this assume wlog that $\vartheta \geq 0$ and $\vartheta(0) = 0$. As $\vartheta \square \psi \leq \psi$, we find

$$0 \leq \mathrm{BT}_\vartheta(\mu,\nu) = \sup_{\psi \text{ convex, lsc}} \mu(\vartheta\square\psi) - \nu(\psi) \leq \sup_{\psi \text{ convex, lsc}} \mu(\psi) - \nu(\psi). \tag{3.4}$$

Suppose that the measures $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$ are in convex order, then we have by (3.4) that $\mathrm{BT}_\vartheta(\mu,\nu) = 0$. Hence, there is $\pi \in \Pi(\mu,\nu)$ such that $\int \vartheta(x - \mathrm{mean}(\pi_x))\,\mu(dx) = 0$. As ϑ has its unique minimum in 0, we have $\mathrm{mean}(\pi_x) = x$ μ-a.s., i.e. π is a martingale coupling. ◇

Example 3.3 ($\vartheta(x) = |x|^p$; p-Wasserstein projection in convex order). A particularly relevant special case of Theorem 3.1 is $\vartheta(x) = |x|^p$. Then, for $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the problem (3.2) reads as

$$\inf\{\mathcal{W}_p^p(\mu,\eta) : \eta \leq_c \nu\}, \tag{3.5}$$

which is precisely finding the metric projection of μ onto the set $\{\eta \in \mathcal{P}_p(\mathbb{R}^d) : \eta \leq_c \nu\}$ in the p-Wasserstein space $(\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p)$. Theorem 3.1 states existence, uniqueness (if $p > 1$) of solutions to this projection problem and gives a description of this solution. ◇

Example 3.4 ($\vartheta(x) = \frac{1}{2}|x|^2$; Brenier–Strassen theorem). In this case, we recover precisely the Brenier–Strassen theorem of Gozlan and Julliet [49].

By Theorem 3.1, there exist a primal optimizer $\pi \in \Pi(\mu,\nu)$ and a dual optimizer ψ. Then the function $\phi := \frac{1}{2}|\cdot|^2 \square \psi$ in the dual problem (3.3) is the Moreau envelope of ψ (with parameter 1). Moreover, there is a unique $\eta \in \mathcal{P}_2(\mathbb{R}^d)$ which solves the projection problem (3.5) for $p = 2$ and the optimal coupling in $\mathcal{W}_2^2(\mu,\eta)$ is unique and induced by the Monge map $T(x) = \mathrm{mean}(\pi_x)$. By the complementary slackness condition (Theorem 3.1((ii))), we find

$$\frac{1}{2}|\cdot|^2\square\psi(x) \leq \frac{1}{2}|x - T(x)|^2 + \psi(T(x)) \leq \frac{1}{2}|x - T(x)|^2 + \pi_x(\psi) = \frac{1}{2}|\cdot|^2\square\psi(x). \tag{3.6}$$

Hence, all inequalities in (3.6) are in fact equalities. Note that (3.6) is then the defining equation of the proximal map of ψ (see Section A.1), i.e. $T = \mathrm{Prox}_\psi$. It is well-known that the proximal map is differentiable with 1-Lipschitz gradient.[7] ◇

Recall that a support function is a convex function $\vartheta : \mathbb{R}^d \to (-\infty, \infty]$ that satisfies $\vartheta(tx) = t\vartheta(x)$ for every $x \in \mathbb{R}^d$ and $t \geq 0$. Note that support functions are subadditive, i.e., $\vartheta(x+y) \leq \vartheta(x) + \vartheta(y)$. In the case of support functions, the duality simplifies in the following way.

Corollary 3.5 (Convex Kantorovich–Rubinstein for support functions)**.** *Let $\vartheta : \mathbb{R}^d \to \mathbb{R}$ be a support function. Then*

$$\mathrm{BT}_\vartheta(\mu,\nu) = \max\left\{\mu(\phi) - \nu(\phi) : \phi \text{ convex},\ \phi(x_1) - \phi(x_2) \leq \vartheta(x_1 - x_2)\right\}. \tag{3.7}$$

[7] Note that $T = \mathrm{Prox}_\psi$ coincides with the transport map stated in Theorem 3.1((v)). Indeed, [18, Proposition 12.29] asserts that $\mathrm{Prox}_\psi = \mathrm{Id} - \nabla(\frac{1}{2}|\cdot|^2 \square \psi)$. As $\vartheta = \frac{1}{2}|\cdot|^2$, we have $\nabla\vartheta^* = \mathrm{Id}$ and hence $\mathrm{Prox}_\psi(x) = x - \nabla\vartheta^*(\nabla(\phi(x)))$.

Note that the condition $\phi(x_1) - \phi(x_2) \le \vartheta(x_1 - x_2)$ for all $x_1, x_2 \in \mathbb{R}^d$ characterizes those convex functions ϕ such that there is a convex function ψ with $\phi = \vartheta \square \psi$, see Theorem A.3. Hence, we can equivalently write the assertion as

$$\mathrm{BT}_\vartheta(\mu,\nu) = \max\left\{\mu(\vartheta\square\psi) - \nu(\vartheta\square\psi) : \psi \text{ convex}\right\}. \tag{3.8}$$

Proof. As support functions have linear growth, the growth condition in Theorem 3.1 is automatically satisfied, and we have duality with the problem (3.3) and attainment. By subadditivity of ϑ, we have $\vartheta\square\vartheta\square\psi = \vartheta\square\psi$, see Theorem A.3. As $\vartheta(0) = 0$, we have $\vartheta\square\psi \le \psi$. These two observations show that replacing a convex function ψ with $\vartheta\square\psi$ increases the value of the dual functional. Hence, we can restrict the maximum to functions ϕ of the form $\phi = \vartheta\square\psi$. This implies the claim by again using that $\vartheta\square\vartheta\square\psi = \vartheta\square\psi$. □

Example 3.6 ($\vartheta(x) = \|x\|$; convex Kantorovich–Rubinstein). Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. For $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$, we have

$$\min_{\eta \le_c \nu} \mathcal{W}_1(\mu,\eta) = \min_{\pi\in\Pi(\mu,\nu)} \int \|x - \mathrm{mean}(\pi_x)\|\,\mu(dx) = \max_{\substack{\phi \text{ convex},\\ \|\cdot\|\text{-1-Lipschitz}}} \mu(\phi) - \nu(\phi). \tag{3.9}$$

This follows from (3.2) and Theorem 3.5 because every norm is a support function and the condition $\phi(x_1) - \phi(x_2) \le \|x_1 - x_2\|$ precisely states that ϕ is 1-Lipschitz w.r.t. $\|\cdot\|$. ⋄

Example 3.7 ($\vartheta(x) = x_+$; increasing convex Kantorovich–Rubinstein). Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$. Then we have

$$\min_{\pi\in\Pi(\mu,\nu)} \int (x - \mathrm{mean}(\pi_x))_+\,\mu(dx) = \max_{\substack{\phi \text{ increasing convex},\\ \text{1-Lipschitz}}} \mu(\phi) - \nu(\phi).$$

This follows from Theorem 3.5 noting that the condition $\phi(x_1) - \phi(x_2) \le (x_1 - x_2)_+$ precisely states that ϕ is 1-Lipschitz and increasing. ⋄

Example 3.8 (Multidimensional increasing convex Kantorovich–Rubinstein). Let $\le$ be the coordinatewise order on $\mathbb{R}^d$, i.e. $x \le y$ if $x_i \le y_i$ for every $i \in \{1, \dots, d\}$. Then we have

$$\min_{\pi\in\Pi(\mu,\nu)} \int \max_{i=1,\dots,d} (x_i - \mathrm{mean}(\pi_x)_i)_+\,\mu(dx) = \max_{\substack{\phi \text{ convex, } \|\cdot\|_1\text{-1-Lipschitz}\\ \text{increasing w.r.t. } \le}} \mu(\phi) - \nu(\phi).$$

More generally, let $\le$ be a partial order on $\mathbb{R}^d$ that is compatible with the linear structure and continuous (i.e. if $x_n \to x$ and $x_n \ge 0$, then $x \ge 0$) and let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Then

$$\min_{\pi\in\Pi(\mu,\nu)} \int \mathrm{dist}_{\|\cdot\|}(x - \mathrm{mean}(\pi_x), -P)\,\mu(dx) = \max_{\substack{\phi \text{ convex, } \|\cdot\|\text{-1-Lipschitz,}\\ \text{increasing w.r.t. } \le}} \mu(\phi) - \nu(\phi),$$

where $P := \{y \in \mathbb{R}^d : y \ge 0\}$ is the positive cone for the order $\le$ and $\mathrm{dist}_{\|\cdot\|}(\cdot, -P)$ denotes the distance to $-P$ w.r.t. the norm $\|\cdot\|$.

To see this, we apply (3.8) with the function $\vartheta(x) = \mathrm{dist}_{\|\cdot\|}(x, -P)$. Then it remains to observe that a convex function ϕ is $\|\cdot\|$-1-Lipschitz and $\leq$-increasing if and only if it is of the form $\phi = \vartheta \square \psi$ for some convex function ψ. For this, note that

$$\vartheta(x) = \mathrm{dist}_{\|\cdot\|}(x, -P) = \inf_{y \in -P} \|x - y\| = \|\cdot\| \square \chi_{(-P)}(x).$$

Using this, Theorem A.4 implies that $\phi = \vartheta \square \psi$ for some convex function ψ if and only if there is a convex function ψ_1 such that $\phi = \|\cdot\| \square \psi_1$ and there is a convex function ψ_2 such that $\phi = \chi_{(-P)} \square \psi_2$, where χ denotes the convex indicator, see Section A.2. The first condition is equivalent to ϕ being 1-Lipschitz w.r.t. $\|\cdot\|$ and the second condition is equivalent to ϕ being increasing w.r.t. $\leq$. $\diamond$

3.2 Mathematical finance interpretation of convex Kantorovich-Rubinstein

Theorem 3.5 admits a natural financial interpretation in terms of model-independent arbitrage under trading restrictions. To provide a concise formulation we recall that there is a one-to-one correspondence between support functions $\vartheta : \mathbb{R}^d \to \mathbb{R}$ and compact convex subsets of $\mathbb{R}^d$. The support function of a compact convex set $K \subset \mathbb{R}^d$ is given by

$$\vartheta_K(x) = \sup_{y \in K} x \cdot y. \tag{3.10}$$

In this case $\partial \vartheta_K(0) = K$. As $\vartheta = \vartheta_{\partial\vartheta(0)}$ and $\partial\vartheta(0)$ is compact convex, every support function arises as in (3.10).

We then have the following result which is essentially a reformulation of Theorem 3.5.

Corollary 3.9. *Let $K \subseteq \mathbb{R}^d$ be compact and convex. Then we have*

$$\begin{aligned} &\min_{\pi \in \Pi(\mu,\nu)} \int \vartheta_K(\mathrm{mean}(\pi_x) - x)\, \mu(dx) = \\ &\quad \max\{w : \exists f_1, f_2, \Delta;\ w \leq (f_1(x_1) - \mu(f_1)) + (f_2(x_2) - \nu(f_2)) + \Delta(x_1) \cdot (x_2 - x_1)\}, \end{aligned} \tag{3.11}$$

where Δ is measurable with values in K and f, g are Lipschitz.

Corollary 3.9 can be seen as a robust quantitative FTAP ('fundmental theorem of asset pricing') for a two time-step market X_1, X_2 consisting of d assets. In mathematical finance terms, f_1, f_2 are vanilla options with maturities $t = 1$ and $t = 2$, resp. which can be bought at prices $\mu(f_1)$ and $\nu(f_2)$, resp.; this is based on the famous Breeden–Litzenberger observation [34] that prices of liquidly traded call / put options can be expressed as distributions of the underlying X under 'pricing measures'. Then

$$S_{f_1,f_2,\Delta}(x_1, x_2) := (f_1(x_1) - \mu(f_1)) + (f_2(x_2) - \nu(f_2)) + \Delta(x_1) \cdot (x_2 - x_1)$$

represents the gains / losses from self-financing trading, with static positions $f_1(X_1), f_2(X_2)$ and holding $\Delta(X_1)$ assets from time $t = 1$ to time $t = 2$. The right-hand side of (3.11) is the maximal risk less profit ("arbitrage") an investor can achieve by trading under the restriction $\Delta(X_1) \in K$.

To illustrate Corollary 3.9, we consider first the case $d = 1$ and $K = [-1, 1]$ where $\vartheta_K(x) = |x|$. If the market satisfies the strict No Arbitrage condition, the maximal risk less profit equals 0 and (3.11) recovers the existence of a martingale measure π consistent with μ, ν. More generally, if the maximal risk less profit under the trading restriction $|\Delta(X_1)| \leq 1$ has level ℓ, (3.11) yields the existence of a consistent 'ℓ-almost' martingale pricing measure.

To consider another example, let again $d = 1$ and $K = [0, 1]$ such that $\vartheta_K(x) = (x)_+$. Financially, $\Delta(x) \geq 0$ means that no short-selling is allowed. Corollary 3.9 then connects the maximal level of arbitrage per stock to the existence of an 'almost' super-martingale measure.

Proof of Theorem 3.9. First note that the right-hand side of (3.11) is equal to $\max \mu(f_1) + \nu(f_2)$, where the maximum is taken over all (f_1, f_2, Δ) such that $\Delta(x_1) \in K$ and $f_1(x_1) + f_2(x_2) + \Delta(x_1) \cdot (x_2 - x_1) \geq 0$. By Theorem 3.1((i)) and (3.8) applied with $\vartheta := \vartheta_{-K}$, it suffices to show that

$$\max_{\zeta \text{ convex}} \mu(\vartheta \square \zeta) - \nu(\vartheta \square \zeta) \leq \max_{\substack{(f_1,f_2,\Delta) \text{ s.t. } \Delta(x_1)\in K, \\ f_1(x_1)+f_2(x_2)+\Delta(x_1)\cdot(x_2-x_1)\geq 0}} \mu(f_1)+\nu(f_2) \leq \max_{\substack{\phi(x_1)-\rho(\psi)\leq \\ \vartheta(x-\text{mean}(\rho))}} \mu(\phi)-\nu(\psi).$$

To see the first inequality, let ζ be optimal and define $f_1 := -(\vartheta \square \zeta)$, $f_2 := \vartheta \square \zeta$ and let Δ be a selector of $-\partial(\vartheta \square \zeta)$. Then the sub-differential inequality $\vartheta \square \zeta(x_2) \geq \vartheta \square \zeta(x_1) + H(x_1) \cdot (x_2 - x_1)$ yields that $f_1(x_1) + f_2(x_2) + \Delta(x_1) \cdot (x_2 - x_1) \geq 0$. As $\partial(\vartheta \square \zeta)(x_1) \subseteq -K$ (see Theorem A.2), we have $\Delta(x_1) \in K$.

To see the second inequality, let (f_1, f_2, Δ) be optimal and set $\phi := -f_1$, $\psi := f_2$. Then integrating the admissibility condition for (f_1, f_2, Δ) for fixed x_1 w.r.t. $\rho(dx_2)$ yields

$$\phi(x_1) \leq -\Delta(x_1) \cdot (x_1 - \text{mean}(\rho)) + \rho(\psi) \leq \vartheta(x_1 - \text{mean}(\rho)) + \rho(\psi),$$

where the second inequality is the subdifferential inequality for ϑ in the point 0, noting that $-\Delta(x_1) \in -K = \partial\vartheta(0)$. □

3.3 Proof of Theorem 3.1

We aim to prove Theorem 3.1 by invoking the fundamental theorem of weak optimal transport. As the first step, we calculate the C-transform.

Lemma 3.10. *Let $\vartheta : \mathbb{R}^d \to \mathbb{R}$ be convex, $C(x, \rho) = \vartheta(x - \text{mean}(\rho))$ and let $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Then we have*

$$(-\psi)^C = \vartheta \square \psi^{**}. \tag{3.12}$$

If ϑ is supercoercive, we have $\psi^C(x) > -\infty$ for every $x \in \mathbb{R}^d$.

Note that (3.12) is to be understood as pointwise equality in $\mathbb{R} \cup \{-\infty, +\infty\}$. For example, if ψ has no affine minorant, then both sides of (3.12) are constant $-\infty$.

Proof. Using the definition of the C-transform, we find

$$(-\psi)^C(x) = \inf_{\rho\in\mathcal{P}_1(\mathbb{R}^d)} C(x,\rho) + \rho(\psi) = \inf_{y\in\mathbb{R}^d} \vartheta(x-y) + \inf_{\mathrm{mean}(\rho)=y} \rho(\psi) = \vartheta\square\widehat{\psi}(x), \quad (3.13)$$

where $\widehat{\psi}(y) = \inf_{\mathrm{mean}(\rho)=y} \rho(\psi)$ is the convex envelope of ψ. We need to distinguish three cases: If $\widehat{\psi}$ is a proper convex function, then the bi-conjugate ψ^{**} is its lsc hull and we have $\vartheta\square\widehat{\psi} = \vartheta\square\psi^{**}$ (see (A.5) in Section A.1) and we can conclude (3.12). If $\widehat{\psi} \equiv +\infty$, then $\psi \equiv +\infty$ as well and it easy to see that both sides of (3.12) are constant $+\infty$. If there is an y such that $\widehat{\psi}(y) = -\infty$, then $\vartheta\square\widehat{\psi} \equiv -\infty$ and (3.13) shows that $(-\psi)^C \equiv -\infty$. As $\psi^{**} \le \widehat{\psi}$, also $\vartheta\square\psi^{**} = -\infty$, so both sides of (3.12) are $-\infty$. □

Proof of Theorem 3.1((i)) *and* ((ii)). We first check that C satisfies the assumptions of the fundamental theorem of weak optimal transport (Theorem 2.2). First note that $C : \mathbb{R}^d \times \mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R}$ is continuous because ϑ is continuous and $\mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R}^d : \rho \mapsto \mathrm{mean}(\rho)$ is continuous. The convexity of C in the second argument follows from the convexity of ϑ. Moreover, C satisfies the growth condition because

$$C(x,\rho) = \vartheta(x - \mathrm{mean}(\rho)) \le \int \vartheta(x-y)\,\rho(dy) \le a(x) + \rho(b).$$

Since ϑ is convex and thus lower bounded by an affine function, C satisfies the lower bound, i.e., $\vartheta(z) \ge r + v\cdot z$ for some $r \in \mathbb{R}$ and $v \in \mathbb{R}^d$, showing that $C(x,\rho) \ge r - |v\cdot x| - \int |y\cdot v|\,\rho(dy)$.

Note that $(-\psi)^C = \vartheta\square\psi^{**}$ by Theorem 3.10. Hence, Theorem 2.2 yields primal and dual existence and strong duality for the dual problem

$$\sup\left\{\mu(\vartheta\square\psi^{**}) - \nu(\psi); \psi \in \mathcal{L}^1(\nu)\right\}. \quad (3.14)$$

As $(-\psi)^C = (-\psi^{**})^C$ and $\psi \ge \psi^{**}$, we can restrict the supremum in (3.14) to lsc convex functions.

Note that the complementary slackness condition, cf. Theorem 3.1 ((ii)), applied to a pair of the form $((-\psi)^C, \psi)$ yields ((ii)). □

Proof of Theorem 3.1((iii)). Let ψ and π be optimal. Using the definition of the infimal convolution, the convexity of ψ, and complementary slackness (Theorem 3.1 ((ii))), we get

$$\vartheta\square\psi(x) \le \vartheta(x - \mathrm{mean}(\pi_x)) + \psi(\mathrm{mean}(\pi_x)) \le \vartheta(x - \mathrm{mean}(\pi_x)) + \pi_x(\psi) = \vartheta\square\psi(x).$$

Hence, all of these inequalities are, in fact, equalities, i.e., $\mathrm{mean}(\pi_x)$ is a minimizer in the definition of $\vartheta\square\psi(x)$. It is easy to see that this is equivalent to $\mathrm{mean}(\pi_x) \in \partial^\vartheta\phi(x)$, where we write $\phi = \vartheta\square\psi$ and $c(x,y) = \vartheta(x-y)$. To see this, recall that

$$\partial^\vartheta\phi(x) = \{y \in \mathbb{R}^d : \phi(v) \le \phi(x) + c(v,y) - c(x,y) \text{ for all } v \in \mathbb{R}^d\},$$

and note that, for every $v \in \mathbb{R}^d$, we have

$$\begin{aligned}\vartheta\square\psi(v) &\le \vartheta(v - \mathrm{mean}(\pi_x)) + \psi(\mathrm{mean}(\pi_x)) \\ &= \vartheta\square\psi(x) + \vartheta(v - \mathrm{mean}(\pi_x)) - \vartheta(x - \mathrm{mean}(\pi_x)).\end{aligned}$$

□

For the proof of Theorem 3.1((iv)) we need the following observation.

Lemma 3.11. *Suppose that $\vartheta : \mathbb{R}^d \to \mathbb{R}$ is strictly convex, η is optimal in $\min_{\eta \leq_c \nu} \mathrm{T}_c(\mu, \eta)$, and $\xi \in \Pi(\mu, \eta)$ is optimal. Then ξ is of Monge type.*

Proof. Write $T(x) = \mathrm{mean}(\xi_x)$. We define the measure $\tilde{\eta} := T_{\#}\mu$ and the coupling $\tilde{\xi} := (\mathrm{id}, T)_{\#}\mu \in \Pi(\mu, \tilde{\eta})$. Using Jensen's inequality we find for every convex function $f : \mathbb{R}^d \to \mathbb{R}$

$$\int f \, d\tilde{\eta} = \int f(\mathrm{mean}(\xi_x)) \, \mu(dx) \leq \iint f(y) \, \xi_x(dy) \, \mu(dx) = \int f \, d\eta.$$

Hence, $\tilde{\eta} \leq_c \eta \leq_c \nu$, so $\tilde{\eta}$ is admissible for the problem $\min_{\eta \leq_c \nu} \mathrm{T}_c(\mu, \eta)$. As η is optimal for this problem and $\xi \in \Pi(\mu, \eta)$ is optimal, we get by Jensen's inequality

$$\int \vartheta(x - y) \, \xi_x(dy) \, \mu(dx) \leq \mathrm{T}_c(\mu, \tilde{\eta}) \leq \int \vartheta(x - T(x)) \, \mu(dx) \leq \iint \vartheta(x - y) \, \xi_x(dy) \, \mu(dx).$$

As ϑ is strictly convex, this yields that for μ-a.e. x we have $y = T(x)$, ξ_x-a.s. Hence, $\xi = (\mathrm{id}, T)_{\#}\mu$. □

Proof of Theorem 3.1((iv)). Let $\pi^1, \pi^2 \in \Pi(\mu, \nu)$ be optimal and let ψ be a dual optimizer. By the consideration in the proof of claim ((iii)), both $\mathrm{mean}(\pi^1_x)$ and $\mathrm{mean}(\pi^2_x)$ are minimizers of $\vartheta \square \psi(x)$. As ϑ is strictly convex, this minimizer is unique, so $\mathrm{mean}(\pi^1_x) = \mathrm{mean}(\pi^2_x)$.

To show the second uniqueness claim, let $\eta^i \leq_c \nu$ and $\xi^i \in \Pi(\mu, \eta^i)$ be optimal for $i \in \{1, 2\}$. By Theorem 3.11, there are maps T^i such that $\xi^i = (\mathrm{id}, T^i)_{\#}\mu$ and by Strassen's theorem (see also Theorem 3.2) there are martingale couplings κ^i from η^i to ν. Then the couplings $\pi^i(dx, dz) := \kappa^i_{T^i(x)}(dz)\mu(dx)$ are optimal for $T_\vartheta(\mu|\nu)$. By the previous considerations, we find $T^1(x) = \mathrm{mean}(\pi^1_x) = \mathrm{mean}(\pi^2_x) = T^2(x)$ μ-a.s. and hence $\xi^1 = \xi^2$. In particular, their second marginals η^1 and η^2 have to coincide. □

Proof of Theorem 3.1((v)). We write $\phi := \vartheta \square \psi$. Note that $\phi \in C(\mathbb{R}^d)$ because it is the infimal convolution of a lsc convex function with a differentiable super-coercive convex function, see e.g. [18, Corollary 18.8] and [31, Theorem 2.2.2]. Using the differentiability of ϑ and ψ, we find (see e.g. [44, Lemma 3.1])

$$y \in \partial^\vartheta \phi(x) \iff \nabla\phi(x) = \nabla\vartheta(x - y) \iff x - y = \nabla\vartheta^*(\nabla\phi(x)).$$

For the last equivalence note that ϑ^* is differentiable as ϑ is strictly convex, cf. Section A.2. □

4 Regularized optimal transport

The aim of this section is to outline how (entropic) regularized optimal transport is covered by weak optimal transport and to derive the structure results for these problems from the fundamental theorem of weak optimal transport.

4.1 Entropic optimal transport

The entropic transport problem is given by

$$\mathrm{ET}_{c,\varepsilon}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int c\,d\pi + \varepsilon H(\pi|\mu\otimes\nu), \tag{4.1}$$

where $\varepsilon > 0$, and H denotes relative entropy of π w.r.t. the product measure $\mu\otimes\nu$, i.e.

$$H(\pi|\mu\otimes\nu) = \int \log\left(\frac{d\pi}{d\mu\otimes\nu}\right) d\pi = \int \frac{d\pi}{d\mu\otimes\nu}\log\left(\frac{d\pi}{d\mu\otimes\nu}\right) d\mu\otimes\nu.$$

Further, we assume that the cost function $c : X\times Y\to\mathbb{R}$ is Borel, lower bounded and there exist $a\in\mathcal{L}^1(\mu), b\in\mathcal{L}^1(\nu)$ such that $c(x,y)\le a(x)+b(y)$. Problem (4.1) can be regarded as a weak transport problem with cost $C : X\times\mathcal{P}(Y)\to(-\infty,\infty]$ defined as

$$C(x,\rho) = \int c(x,\cdot)\,d\rho + \varepsilon\int \frac{d\rho}{d\nu}\log\left(\frac{d\rho}{d\nu}\right) d\nu, \tag{4.2}$$

and $+\infty$ if ρ is not absolutely continuous w.r.t. ν. Applying the fundamental theorem of weak optimal transport guarantees existence of dual optimizers and a complementary slackness criterion for optimality.

Proposition 4.1. *The weak optimal transport dual problem for* (4.1) *is given by*

$$\sup\left\{\mu(f)+\nu(g)\,\middle|\,(f,g)\in\mathcal{L}^1(\mu)\times\mathcal{L}^1(\nu) : f(x)+\rho(g)\le\int c(x,\cdot)\,d\rho+\varepsilon H(\rho|\nu)\right\}. \tag{4.3}$$

We have strong duality, primal attainment and dual attainment. Moreover, for $\pi\in\Pi(\mu,\nu)$ and an admissible pair (f,g) the following are equivalent:

(i) *π is a primal optimizer and (f,g) is a dual optimizer.*

(ii) *We have for μ-a.e. x*

$$\pi_x(c(x,\cdot)) + \varepsilon H(\pi_x|\nu) = f(x)+\pi_x(g). \tag{4.4}$$

Proof. To simplify the notation in the proof, we set $\varepsilon = 1$. We need to check that the cost function (4.2) satisfies the assumptions of the fundamental theorem of weak optimal transport (Theorem 2.2) which then yields the claim. C satisfies the growth condition (B) with the super-coercive increasing convex function $h(t) = t\log(t)_+ + 1$. It is clear that C is Borel.

To see that $\rho\mapsto C(x,\rho)$ is convex and lsc w.r.t. the weak topology, we write it as a relative entropy and use that this is known to be convex and lsc (see e.g. [70, Lemma 1.3]). Indeed, writing $\alpha_x := \int e^{-c(x,\cdot)}d\nu$ and $\nu^x := \alpha_x^{-1}e^{-c(x,\cdot)}\nu$, we find that $C(x,\rho) = H(\rho|\nu^x) - \log(\alpha_x)$.

To check that C satisfies condition (C), let $(Y_n)_n$ be an increasing sequence of Borel subsets of Y satisfying $\bigcup_n Y_n = Y$. Writing $\rho_n = \frac{1}{\rho(Y_n)}\rho|_{Y_n}$, we have for every $x\in X$ and

$\rho \in \mathcal{P}(Y)$ with $\rho(Y_n) \to 1$ that $\frac{d\rho_n}{d\nu} \to \frac{d\rho}{d\nu}$ ν-a.s. As $\frac{d\rho_n}{d\nu} \leq \frac{1}{\rho(Y_1)} \frac{d\rho}{d\nu}$ for every $n \in \mathbb{N}$ and as c and $t \mapsto t\log(t)$ are lower bounded, dominated convergence yields

$$\begin{aligned} C(x,\rho) &= \int c(x,\cdot)\frac{d\rho}{d\nu} + \frac{d\rho}{d\nu}\log\Big(\frac{d\rho}{d\nu}\Big)\,d\nu \\ &= \lim_n \int c(x,\cdot)\frac{d\rho_n}{d\nu} + \frac{d\rho_n}{d\nu}\log\Big(\frac{d\rho_n}{d\nu}\Big)\,d\nu = \lim_n C(x,\rho_n), \end{aligned}$$

showing that C satisfies the continuity assumption (C). □

Next, we derive the classical structure theorem of entropic optimal transport (see e.g. [63, Theorem 2.9], [70, Theorem 4.2]) from the complementary slackness condition (4.4).

Theorem 4.2. *Let $\pi \in \Pi(\mu,\nu)$. Then π is optimal for $\mathrm{ET}_{c,\varepsilon}(\mu,\nu)$ if and only if there exist measurable $f \in \mathcal{L}^1(\mu)$ and $g \in \mathcal{L}^1(\nu)$ such that*

$$\frac{d\pi}{d\mu \otimes \nu} = \exp\left(\frac{-c + f \oplus g}{\varepsilon}\right). \tag{4.5}$$

In this case (f,g) is optimal in the dual problem (4.3).

Proof. By replacing C with $\frac{1}{\varepsilon}C$ and then scaling the respective dual potentials by ε, we assume wlog that $\varepsilon = 1$. We first show that if $\pi \in \Pi(\mu,\nu)$ is optimal and (f,g) is optimal, then (4.5) holds. By possibly changing the optimizers on nullsets, we assume wlog that (4.4) holds for every $x \in X$. We fix $x \in X$ and aim to show that

$$\frac{d\pi_x}{d\nu} = \exp(f(x) + g - c(x,\cdot)). \tag{4.6}$$

To that end, let $\eta \in \mathcal{P}(Y)$ and set $\rho_t := (1-t)\pi_x + t\eta$. As ρ_t is a probability measure for every $t \in [0,1]$, the admissibility condition for (f,g) yields

$$f(x) + \int g\,d\rho_t \leq \int c(x,\cdot)\,d\rho_t + H(\rho_t|\nu).$$

Subtracting (4.4) from this yields the following inequality (which is an equality for $t = 0$ because $\rho_0 = \pi_x$)

$$0 \leq \int c(x,\cdot) - g\,d(\rho_t - \pi_x) + \int h\Big(\frac{d\rho_t}{d\nu}\Big) - h\Big(\frac{d\pi_x}{d\nu}\Big)\,d\nu,$$

where $h(t) = t\log(t) - t$. Note that $h'(t) = \log(t)$. Taking the right derivative at $t = 0$ yields

$$\begin{aligned} 0 &\leq \frac{d}{dt}\Big|_{t=0^+} \int (c(x,\cdot) - g)\,d(\rho_t - \pi_x) + \int h\Big(\frac{d\rho_t}{d\nu}\Big) - h\Big(\frac{d\pi_x}{d\nu}\Big)\,d\nu \\ &= \int (c(x,\cdot) - g)\,d(\eta - \pi_x) + \int \frac{d}{dt}\Big|_{t=0^+} h\Big(\frac{d\rho_t}{d\nu}\Big)\,d\nu \\ &= \int c(x,\cdot) - g + \log\Big(\frac{d\pi_x}{d\nu}\Big)\,d(\eta - \pi_x). \end{aligned} \tag{4.7}$$

Recall that (4.4) states that

$$f(x) = \int c(x,\cdot) - g + \log\Big(\frac{d\pi_x}{d\nu}\Big)\, d\pi_x. \tag{4.8}$$

Using this, inequality (4.7) simplifies to

$$0 \leq \int c(x,\cdot) - f(x) - g + \log\Big(\frac{d\pi_x}{d\nu}\Big)\, d\eta. \tag{4.9}$$

Applying (4.9) to every $\eta \in \mathcal{P}(Y)$ that is absolute continuous w.r.t. ν with bounded density, yields the ν-a.s. inequality

$$0 \leq c(x,\cdot) - f(x) - g + \log\Big(\frac{d\pi_x}{d\nu}\Big). \tag{4.10}$$

Note that (4.8) implies that we have equality when integrating (4.10) w.r.t. π_x. Hence, (4.10) is in fact a π_x-a.s. equality. Rearranging terms and applying the exponential function yields (4.6). This finishes the proof of (4.5).

Conversely, assume that $\pi \in \Pi(\mu,\nu)$ and that (4.5) holds for some functions $(f,g) \in \mathcal{L}^1(\mu) \times \mathcal{L}^1(\nu)$. As the first marginal of π is μ, the disintegration of π w.r.t. μ satisfies (4.6). As the second marginal of π is ν, we have for μ-almost every x

$$f(x) = -\log\left(\int \exp(g - c(x,\cdot))\, d\nu\right).$$

Then a straightforward calculation shows that (4.4) holds true. Hence, we derive optimality of π and (f,g) from Propsition 4.1. □

Remark 4.3. In entropic optimal transport usually a different dual problem is used, namely

$$\sup_{u\in\mathcal{L}^1(\mu), v\in\mathcal{L}^1(\nu)} \int u\, d\mu + \int v\, d\nu - \int \exp(u+v-c)\, d\mu\otimes\nu + 1. \tag{4.11}$$

It is well known (see e.g. [70, Theorem 4.7]) that there is strong duality for this problem. Note that if (f,g) is admissible in (4.3), then it is also admissible for (4.11). Moreover, if (f,g) is optimal in (4.3), then (4.5) yields that $\int \exp(f+g-c)\, d\mu\otimes\nu = 1$, so (f,g) also yields the optimal value in (4.11). ◇

4.2 Regularization with a general convex function

In this section, we briefly sketch how regularization of optimal transport with a general convex function as studied in [30, 41, 67] fits into the framework of weak optimal transport. A particularly relevant case is quadratically regularized OT, in which case there are significantly more refined results, see [45, 46, 47, 65, 71].

Specifically, we consider the problem

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi + \int h\left(\frac{d\pi}{d\mu\otimes\nu}\right) d\mu\otimes\nu, \tag{4.12}$$

where $c : X \times Y \to \mathbb{R}$ is lsc, lower bounded and bounded from above by integrable functions, i.e. $c(x,y) \leq a(x) + b(y)$ for $a \in \mathcal{L}^1(\mu), b \in \mathcal{L}^1(\nu)$, and $h : \mathbb{R} \to \mathbb{R}$ is a convex function. The case of quadratically regularized OT corresponds to $h(t) = \frac{1}{2}t^2$ if $t \geq 0$ and $h(t) = +\infty$ if $t < 0$.

Note that (4.12) is a weak optimal transport problem with cost

$$C(x,\rho) = \int c(x,\cdot)\, d\rho + \int h\left(\frac{d\rho}{d\nu}\right) d\nu. \tag{4.13}$$

By using the same arguments as in the proof of Proposition 4.1, we find that the weak optimal transport dual problem for (4.12) is given by

$$\sup\left\{\mu(f) + \nu(g)\,;\, (f,g) \in \mathcal{L}^1(\mu) \times \mathcal{L}^1(\nu) : f(x) + \rho(g) \leq \int c(x,\cdot)\frac{d\rho}{d\nu} + h\Big(\frac{d\rho}{d\nu}\Big)\, d\nu\right\}. \tag{4.14}$$

Moreover, the fundamental theorem of weak optimal transport states that there is strong duality, primal attainment and dual attainment. Further, complementary slackness for this problem reads as: Given $\pi \in \Pi(\mu,\nu)$ and admissible pair (f,g) the following are equivalent:

1. π is a primal optimizer and (f,g) is a dual optimizer.
2. We have for μ-a.e. x

$$\int c(x,\cdot)\, d\pi_x + \int h\Big(\frac{d\pi_x}{d\nu}\Big)\, d\nu = f(x) + \int g\, d\pi_x.$$

Using the methods of the proof of Theorem 4.2, we find that if π and (f,g) are optimal, then

$$\frac{d\pi_x}{d\nu}(y) = (h^*)'(\alpha(x) + g(y) - c(x,y)), \tag{4.15}$$

where $\alpha(x)$ is chosen such that

$$\int (h^*)'(\alpha(x) + g(y) - c(x,y))\, \nu(dy) = 1. \tag{4.16}$$

Further note that if $h'(0) = -\infty$ (understood as right-derivative), then h^* is strictly increasing, hence $(h^*)' > 0$, showing that $\mathrm{spt}(\pi) = \mathrm{spt}(\mu \otimes \nu)$.

We close this section with a remark on the connection between the weak optimal transport dual (4.14) an the dual problem used in [67], which reads as

$$\sup_{(u,v)\in\mathcal{L}^1(\mu)\times\mathcal{L}^1(\nu)} \int u\, d\mu + \int v\, d\nu - \int h^*(u+v-c)\, d\mu\otimes\nu. \tag{4.17}$$

If (f,g) is optimal for (4.14), then $(u,v) := (\alpha, g)$, where α is defined according to (4.16), is optimal for problem (4.17). This can be seen by comparing (4.15) with the complementary slackness condition provided in [67, Theorem 3.6].

5 Relaxed weak martingale transport

A weak martingale transport problem is a weak transport problem of the form

$$\mathrm{WMT}_C(\mu,\nu) := \inf_{\pi\in\Pi_{\mathrm{M}}(\mu,\nu)} \int C(x,\pi_x)\,\mu(dx),$$

where $C : \mathbb{R}^d \times \mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R}$ is a cost function and $\Pi_{\mathrm{M}}(\mu,\nu)$ is the set of martingale transports from μ to ν. Note that WMT_C is a weak transport problem with cost of the form

$$C(x,\rho) = \begin{cases} \widehat{C}(\rho) & \mathrm{mean}(\rho) = x, \\ \infty & \text{else.} \end{cases} \tag{5.1}$$

The fundamental theorem of weak optimal transport guarantees primal existence and strong duality for this problem. However, the fact that $C(x,\rho) = +\infty$ whenever $\mathrm{mean}(\rho) \neq x$ means that the boundedness condition (B) is not satisfied. Therefore, Theorem 2.2 does not guarantee dual attainment; and, as already pointed out in the introduction, dual attainment for (weak) martingale transport in dimension $d > 1$ can even fail in very regular settings.

As a remedy for this, we relax the first marginal condition in the weak transport problem. Specifically, we consider the problem

$$\inf_{\eta\in\mathcal{P}_p(\mathbb{R}^d)} \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu), \tag{5.2}$$

where $\vartheta : \mathbb{R}^d \to \mathbb{R}$ is a convex function. In the case of $p = 2$ and $\vartheta(x) = |x|^2$ it can also be seen as Wasserstein–Yosida regularization of $\mathrm{WMT}_C(\cdot,\nu)$.

The main observation of this section is that problem (5.2) corresponds to the weak transport problem with cost

$$C_\vartheta(x,\rho) = \vartheta(x - \mathrm{mean}(\rho)) + \widehat{C}(\rho). \tag{5.3}$$

For $\vartheta = \chi_{\{0\}}$, we formally recover the cost C introduced in (5.1) and the relaxed problem (5.2) reduces to WMOT.

We observe that convexity of $\widehat{C}$ on fibers of the map $\rho \mapsto \mathrm{mean}(\rho)$, i.e. $\widehat{C}((1-t)\rho_1 + t\rho_2) \leq (1-t)\widehat{C}(\rho_1) + t\widehat{C}(\rho_2)$ whenever $\rho_1, \rho_2 \in \mathcal{P}_p(\mathbb{R}^d)$ satisfy $\mathrm{mean}(\rho_1) = \mathrm{mean}(\rho_2)$, implies convexity of the cost C, but is a too weak condition to guarantee convexity of C_ϑ.[8] Therefore, we need to invoke the theory of weak transport with non-convex cost as outlined in Section 2.5.

The aim of this section is to derive a 'fundamental theorem of relaxed WMOT' from the fundamental theorem of WOT with cost C_ϑ. For this, we need appropriate conditions on $\widehat{C}$ to guarantee that C_ϑ satisfies (B) and (C). For the boundedness condition (B), this

[8] Convexity of $\widehat{C}$ is sufficient to guarantee convexity of C_ϑ. Note however that the function $\widehat{C}$ arising in entropic martingale transport (see Section 5.2) is merely convex on the fibers of $\rho \mapsto \mathrm{mean}(\rho)$.

is the existence of a function $b \in \mathcal{L}^1(\nu)$ and an increasing function $h : [0,\infty) \to [0,\infty)$ such that

$$\widehat{C}(\rho) \le \rho(b) + \int h\Big(\frac{d\rho}{d\nu}\Big)\, d\nu. \tag{BM}$$

The analogue of the continuity condition is the following: For every increasing sequence $(Y_k)_k$ of Borel sets with $\bigcup_k Y_k = \mathbb{R}^d$ and every $\rho \in \mathcal{P}_p(\mathbb{R}^d)$ we have

$$\widehat{C}(\rho) \ge \limsup_k \widehat{C}(\tfrac{1}{\rho(Y_k)}\rho|_{Y_k}). \tag{CM}$$

Theorem 5.1. *Let $\mu,\nu \in \mathcal{P}_p(\mathbb{R}^d)$ and let $\widehat{C} : \mathcal{P}_p(\mathbb{R}^d) \to [0,\infty]$ be lsc and convex on the fibers of $\rho \mapsto \mathrm{mean}(\rho)$. Suppose that $\widehat{C}$ satisfies* (BM) *and* (CM)*. Further, let $\vartheta : \mathbb{R}^d \to [0,\infty)$ be a convex function satisfying $\vartheta(x-y) \le a(x) + b(y)$ for some $a \in \mathcal{L}^1(\mu)$ and $b \in \mathcal{L}^1(\nu)$. Then we have the following assertions:*

(i) *The problem* (5.2) *is equivalent to the relaxed WOT problem with cost C_ϑ, i.e.*

$$\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu) = \min_{\eta\in\mathcal{P}_p(\mathbb{R}^d)} \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu). \tag{5.4}$$

(ii) *We have strong duality and dual attainment, that is*

$$\min_{\eta\in\mathcal{P}_p(\mathbb{R}^d)} \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu) = \max_{g\in\mathcal{L}^1(\nu)} \mu(\vartheta\square g^C) + \nu(g). \tag{5.5}$$

(iii) *Both $\eta \in \mathcal{P}_p(\mathbb{R}^d)$ and $g \in \mathcal{L}^1(\nu)$ are optimal in* (5.5) *if and only if*

$$\mathrm{T}_\vartheta(\mu,\eta) = \mu(\vartheta\square g^C) - \eta(g^C), \tag{5.6}$$

$$\mathrm{WMT}_C(\eta,\nu) = \eta(g^C) + \nu(g). \tag{5.7}$$

(iv) *Both $P \in \Lambda(\mu,\nu)$ and $g \in \mathcal{L}^1(\nu)$ are optimal in* (5.4) *if and only if for P-a.e. (x,ρ)*

$$\mathrm{mean}(\rho) \in \partial^\vartheta(\vartheta\square g^C)(x), \tag{5.8}$$

$$\rho \in \arg\min\{\widehat{C}(q) - q(g) : q \in \mathcal{P}_1(\mathbb{R}^d),\, q(|g|) < \infty\}. \tag{5.9}$$

The interpretation of (5.4) is that the following viewpoints are equivalent: Relaxing the martingale constraints while the marginals remain fixed and keeping martingale constraint fixed while the first marginal condition is relaxed. For $p = 2$ and $\vartheta(x) = |x|^2$ the second point of view corresponds to a Wasserstein–Yosida regularization of $\mathrm{WMT}_C(\cdot,\nu)$.

Note that the conditions (5.6) and (5.7) precisely say that the pair $(\vartheta\square g^C, g^C)$ is a dual optimizer for the transport problem from μ to η and that (g^C, g) is a dual optimizer for the weak martingale optimal transport problem between η and ν.

We further note the connection between the transforms associated to the weak martingale transport problem and its regularized version. Specifically, we have

$$\begin{aligned} g^{C_\vartheta}(x) &= \inf_{m\in\mathbb{R}^d} \inf_{\mathrm{mean}(\rho)=m} \vartheta(x-m) + C(m,\rho) - \rho(g) \\ &= \inf_{m\in\mathbb{R}^d} \vartheta(x-m) + g^C(m) = \vartheta\square g^C(x). \end{aligned} \tag{5.10}$$

Before proving Theorem 5.1, we discuss two relevant instances where problem (5.2) even corresponds to non-relaxed WOT problem and allows us to strengthen the results from Theorem 5.1 ((i)).

Proposition 5.2. *In addition to the assumptions of Theorem 5.1 suppose one of the following:*

(a) *The cost is of the form $C(x,\rho) = C(\rho)$ for a convex lsc function $C : \mathcal{P}_p(\mathbb{R}^d) \to \mathbb{R}$.*

(b) *μ is absolutely continuous, and ϑ is of the form $\vartheta(x) = \bar{\vartheta}(|x|)$ for some increasing, strictly convex function $\bar{\vartheta} : [0,\infty) \to [0,\infty)$.*

Then (5.2) *is also equivalent to the non-relaxed WOT problem with cost C_ϑ, i.e.*

$$\mathrm{WT}_{C_\vartheta}(\mu,\nu) = \min_{\eta\in\mathcal{P}_p(\mathbb{R}^d)} \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu). \tag{5.11}$$

Moreover, there exists an optimizer $\pi \in \Pi(\mu,\nu)$ for $\mathrm{WT}_{C_\vartheta}(\mu,\nu)$.

If C is strictly convex in ρ, then π is unique and we have the following optimality condition: $\pi \in \Pi(\mu,\nu)$ is optimal in $\mathrm{WT}_{C_\vartheta}(\mu,\nu)$ if and only if

$$\begin{cases} \eta = (x \mapsto \mathrm{mean}(\pi_x))_{\#}\mu \text{ is optimal for (5.11)}, \\ \xi = (x \mapsto (x, \mathrm{mean}(\pi_x))_{\#}\mu \text{ is optimal for } \mathrm{T}_\vartheta(\mu,\eta), \\ \kappa \text{ is optimal for } \mathrm{WMT}_C(\eta,\nu), \\ \pi_x = \kappa_{\mathrm{mean}(\pi_x)}\ \mu\text{-a.s.} \end{cases} \tag{5.12}$$

We point out that both assumptions in Theorem 5.2 arise naturally. Case ((a)) arises in the relaxation of the martingale Benamou–Brenier problem (see Section 5.1) while ((b)) is natural in the context of regularized entropic martingale transport (see Section 5.2). There are similar results to (5.12) if C is not strictly convex and for the optimal $P \in \Lambda(\mu,\nu)$ in $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$, see Theorem 5.10 below.

Remark 5.3. The assumptions of Theorem 5.2 ((b)) are chosen to guarantee the existence of an optimal Monge coupling between μ and η. We point out this is also true for a more general class of strictly convex functions (see [44, Theorem 1.2]) and also for the function $\vartheta(x) = |x|$ (see e.g. [83, Theorem 2.50]). In these cases the assertions of Theorem 5.2 remain vaild (except for uniqueness of the optimal π, for which strict convexity of both C and ϑ is needed). ◇

5.1 Application to martingale Benamou–Brenier

The weak transport problem associated to martingale Benamou–Brenier is given by

$$\mathrm{MBB}(\mu,\nu) = \sup_{\pi\in\Pi_{\mathrm{M}}(\mu,\nu)} \int \mathrm{MCov}(\pi_x,\gamma)\,\mu(dx),$$

where

$$\mathrm{MCov}(\rho_1,\rho_2) = \sup_{q\in\Pi(\rho_1,\rho_2)} \int x\cdot y\, q(dx,dy).$$

In the case that one of the measures ρ_1, ρ_2 is centered, $\mathrm{MCov}(\rho_1, \rho_2)$ is the maximal covariance that a pair of random variables (X_1, X_2) with $\mathrm{law}(X_i) = \rho_i$ can admit. In this section, we study an interpolation between the martingale Benamou–Brenier problem and a barycentric transport problem, i.e.

$$\mathrm{MBB}_{\mathrm{reg}}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \beta\vartheta(x - \mathrm{mean}(\pi_x)) - \alpha \mathrm{MCov}(\pi_x, \gamma)\, \mu(dx), \tag{5.13}$$

where $\alpha, \beta \geq 0$ are real parameters, $\vartheta : \mathbb{R}^d \to \mathbb{R}$ is a convex function, and $\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ is a centered absolutely continuous measure, the most relevant example being the standard Gaussian.

Specifically, if $\vartheta : \mathbb{R}^d \to \mathbb{R}$ is a convex function with unique minimum in 0 (e.g. $\vartheta(x) = |x|$ or $\vartheta(x) = |x|^2$), $\alpha = 1$ and $\beta \to \infty$, then (5.13) can be seen as a martingale Benamou–Brenier problem with relaxed martingale constraint. Conversely, for $\beta = 1$ and small α, (5.13) can be seen as a regularization that enforces strict convexity, while the barycentric transport problem itself is non-strictly convex even when ϑ is strictly convex.

For notational simplicity we set $\alpha = \beta = 1$ from now on. Note that this is wlog by replacing γ with $\gamma^\alpha := (x \mapsto \alpha x)_{\#}\gamma$ and ϑ with $\vartheta^\beta(x) := \beta\vartheta(x)$. We write $\check{\gamma} := (x \mapsto -x)_{\#}\gamma$ for γ reflected at the origin. Applying Theorem 5.1 to $\mathrm{MBB}_{\mathrm{reg}}$ yields

Theorem 5.4. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ be centered and absolutely continuous, and let $\vartheta : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Assume there exist functions $a \in \mathcal{L}^1(\mu)$ and $b \in \mathcal{L}^1(\nu)$, both convex, such that $\vartheta \leq a \square b$.*

(i) *We have*

$$\mathrm{MBB}_{\mathrm{reg}}(\mu, \nu) = \min_{\eta \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{T}_\vartheta(\mu, \eta) - \mathrm{MBB}(\eta, \nu). \tag{5.14}$$

There exist a unique primal optimizer π for $\mathrm{MBB}_{\mathrm{reg}}(\mu, \nu)$ and a unique optimal η on the right-hand side. Moreover, π is optimal if and only if

$$\begin{cases} \eta = (x \mapsto \mathrm{mean}(\pi_x))_{\#}\mu \text{ is optimal for (5.14)}, \\ \xi = (x \mapsto (x, \mathrm{mean}(\pi_x))_{\#}\mu \text{ is optimal for } \mathrm{T}_\vartheta(\mu, \eta), \\ \kappa \text{ is optimal for } \mathrm{MBB}(\eta, \nu), \\ \pi_x = \kappa_{\mathrm{mean}(\pi_x)}\ \mu\text{-a.s.} \end{cases} \tag{5.15}$$

(ii) *We have strong duality and dual attainment:*

$$\mathrm{MBB}_{\mathrm{reg}}(\mu, \nu) = \max_{\psi \text{ convex, lsc}} \mu(\vartheta \square (\psi^* * \check{\gamma})^*) - \nu(\psi). \tag{5.16}$$

as well as primal attainment and dual attainment. Moreover, the primal optimizer is unique.

(iii) *Both $\eta \in \mathcal{P}_2(\mathbb{R}^d)$ and ψ are optimal in (5.14) and (5.16), resp. if and only if*

$$\mathrm{T}_\vartheta(\mu, \eta) = \mu(\vartheta \square (\psi^* * \gamma)^*) - \eta((\psi^* * \gamma)^*), \tag{5.17}$$

$$-\mathrm{MBB}(\eta, \nu) = \eta((\psi^* * \gamma)^*) - \nu(\psi). \tag{5.18}$$

Note that if $\gamma = N(0, \mathrm{id})$, then (5.18) is equivalent to $\kappa = \mathrm{law}(M_0, M_1)$ for a Bass martingale $M_t = \mathsf{E}[v(B_1)|B_t]$ where $v = \psi^*$, cf. (1.15) and [12, Theorem 1.4].

We also point out that for $\vartheta(x) = |x|^2$, (5.14) states that $\mathrm{MBB}_{\mathrm{reg}}(\cdot, \nu)$ is the Yosida regularization of $\mathrm{MBB}(\cdot, \nu)$ in the metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$.

Remark 5.5. The fact that it suffices to take the supremum over convex functions in the dual problem, see (5.16), follows a more general principle observed by Pramenković [73]: If the function $\widehat{C}$ is decreasing w.r.t. convex order (i.e. if $\rho_1 \leq_c \rho_2$, then $\widehat{C}(\rho_1) \geq \widehat{C}(\rho_2)$), the dual problem can be restricted to convex functions. ◇

In order to derive Theorem 5.4 from Theorem 5.1 we need to calculate the C-transform associated to the cost associated to MBB, i.e.

$$C(x, \rho) = \begin{cases} \mathrm{MCov}(\rho, \gamma) & \mathrm{mean}(\rho) = x, \\ +\infty & \mathrm{mean}(\rho) \neq x. \end{cases} \tag{5.19}$$

Proposition 5.6. *Let $\psi : \mathbb{R}^d \to (-\infty, \infty]$ and C be the cost defined in* (5.19)*. Then $(-\psi)^C$ is convex and*

$$((-\psi)^C)^{**} = (\psi^* * \check{\gamma})^*.$$

In particular, $(\psi^* * \check{\gamma})^*$ is proper if and only if $(-\psi)^C$ is proper.

Proof. Applying the definition of the C-transform to the cost (5.19) yields

$$(-\psi)^C(m) = \inf_{\substack{\rho \in \mathcal{P}_2(\mathbb{R}^d) \\ \mathrm{mean}(\rho) = m \\ \psi \in \mathcal{L}^1(\rho)}} \inf_{\pi \in \Pi(\rho, \gamma)} \int \psi(y) - y \cdot z \, \pi(dy, dz).$$

From this expression it is straightforward to see that $(-\psi)^C$ is convex. In order to calculate its lsc envelope, we fist calculate its convex conjugate. We have

$$((-\psi)^C)^*(x) = \sup \left\{ \int y \cdot (z + x) - \psi(y) \, \pi(dy, dz) : \pi \in \mathcal{P}(\mathbb{R}^{2d}), \psi \in \mathcal{L}^1(\mathrm{pr}_{1\#}\pi), \mathrm{pr}_{2\#}\pi = \gamma \right\}. \tag{5.20}$$

Next, we show that the right-hand side of (5.20) is equal to $\int \psi^*(z + x) \, \gamma(dz)$. If π is admissible in the right hand side of (5.20), we find using the Fenchel–Young inequality

$$\begin{aligned} \int y \cdot (z + x) - \psi(y) \, \pi(dy, dz) &= \iint y \cdot (z + x) - \psi(y) \, \pi_z(dy) \, \gamma(dz) \\ &\leq \iint \psi^*(z + x) \, \pi_z(dy) \, \gamma(dz) = \int \psi^*(z + x) \, \gamma(dz). \end{aligned}$$

To prove the converse inequality, for every $n \in \mathbb{N}$, let $h_n : \mathbb{R}^d \to \mathbb{R}^d$ be a Borel selector of almost optimizers in

$$\sup_{\substack{y \in \mathbb{R}^d \text{ s.t.} \\ |y| \leq n, \, |\psi(y)| \leq n}} y \cdot (z + x) - \psi(y),$$

i.e. $|h_n(z)| \leq n$, $|\psi(h_n(z))| \leq n$ and $h_n(z) \cdot (z+x) - \psi(h_n(z))$ equals the supremum up to $1/n$ if it is finite and $h_n(z) \cdot (z+x) - \psi(h_n(z)) \geq n$ if the supremum is $+\infty$. Writing $\pi_n := (z \mapsto (h_n(z), z))_* \gamma$ we find that right-hand side of (5.20) is bigger than

$$\begin{aligned}\limsup_n \int y \cdot (z+x) - \psi(y)\, \pi_n(dy, dz) &= \limsup_n \int \sup_{\substack{y \in \mathbb{R}^d \text{ s.t.} \\ |y| \leq n,\, |\psi(y)| \leq n}} y \cdot (z+x) - \psi(y)\, \gamma(dz) \\ &= \int \psi^*(z+x)\, \gamma(dz),\end{aligned}$$

where the second equality follows from monotone convergence. This allows us to conclude that for every $x \in \mathbb{R}^d$

$$((-\psi)^C)^*(x) = \int \psi^*(z+x)\, \gamma(dz) = (\psi^* * \check{\gamma})(x).$$

Taking the convex conjugate on both sides of this equality yields the claim. □

Proof of Theorem 5.4. We aim to derive Theorem 5.4 from Theorem 5.1 and Theorem 5.2, case ((a)). To that end, we first need to observe that the cost is strictly convex in the second argument. Clearly, $-\mathrm{MCov}(\cdot, \gamma)$ and the barycentric term are both convex.

For strict convexity, we show that $-\mathrm{MCov}(\cdot, \gamma)$ is strictly convex. This follows from Brenier's theorem: Suppose that ρ^1, ρ^2 are such that $\frac{1}{2}\mathrm{MCov}(\rho^1, \gamma) + \frac{1}{2}\mathrm{MCov}(\rho^2, \gamma) = \mathrm{MCov}(\frac{1}{2}(\rho^1 + \rho^2), \gamma)$. If $\xi^i \in \Pi(\gamma, \rho^i)$, $i \in \{1, 2\}$ are optimal in $\mathrm{MCov}(\gamma, \rho^i)$, then $\xi = \frac{1}{2}(\xi^1 + \xi^2)$ is optimal in $\mathrm{MCov}(\gamma, \frac{1}{2}(\rho^1 + \rho^2))$. As γ is absolutely continuous, Brenier's theorem yields that ξ is supported on the graph of a function. This implies $\xi^1 = \xi^2$ and hence $\rho^1 = \rho^2$.

Hence, the assumptions of Theorem 5.1 and Theorem 5.2, case ((a)), are satisfied. This yields all statements of Theorem 5.4, except for the precise structure of the dual problem stated in (ii).

To show this, note that Theorem 5.6 yields that the lsc hull of the C-transform of the martingale Benamou–Brenier cost is given by $(\psi^* * \check{\gamma})^*$. Using (5.10) and (A.5) we find that the C-transform associated to the $\mathrm{MBB}_{\mathrm{reg}}$ cost is given by

$$\vartheta\Box(\psi^* * \check{\gamma})^*.$$

The fact that this C-transform only depends on ψ^*, implies that we can restrict the dual problem to convex lsc functions. Indeed, replacing ψ with ψ^{**} increases the value of $\mu(\vartheta\Box(\psi^* * \check{\gamma})^*) - \nu(\psi)$ because $\psi^{**} \leq \psi$ and $\psi^{***} = \psi^*$. □

5.2 Application to entropic martingale transport

The final section is dedicated to the entropic martingale transport problem and its relaxation. More specifically, given $\eta, \nu \in \mathcal{P}_1(\mathbb{R}^d)$, a measurable cost $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, and the regularization parameter $\varepsilon > 0$, the entropic martingale transport problem reads as

$$\mathrm{EMT}_{c,\varepsilon}(\eta, \nu) = \inf_{\kappa \in \Pi_{\mathrm{M}}(\eta, \nu)} \int c(m, y)\, \kappa(dm, dy) + \varepsilon H(\kappa | \eta \otimes \nu). \tag{5.21}$$

This corresponds to the weak transport problem with cost $C : \mathbb{R}^d \times \mathcal{P}_1(\mathbb{R}^d) \to (-\infty, \infty]$ defined as in (5.1) where

$$\widehat{C}(\rho) = \int c(\mathrm{mean}(\rho), y)\, \rho(dy) + \varepsilon H(\rho|\nu). \tag{5.22}$$

In this section, we study the relaxation as proposed in (5.2) where we penalize deviation from the mean according to a convex function $\vartheta : \mathbb{R}^d \to \mathbb{R}$, i.e.

$$\inf_{\eta \in \mathcal{P}_1(\mathbb{R}^d)} \mathrm{T}_\vartheta(\mu, \eta) + \mathrm{EMT}_{c,\varepsilon}(\eta, \nu). \tag{5.23}$$

Throughout this section, we denote by C_ϑ the associated weak transport cost given by

$$C_\vartheta(x, \rho) = \vartheta(x - \mathrm{mean}(\rho)) + \int c(\mathrm{mean}(\rho), y)\, \rho(dy) + \varepsilon H(\rho|\nu). \tag{5.24}$$

Relaxed entropic martingale transport is an instance of a relaxed weak martingale transport problem. Under the natural assumptions of the next proposition, the structural results of Theorem 5.1 also hold in the current setting.

Proposition 5.7. *Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, let $\vartheta : \mathbb{R}^d \to [0, \infty)$ be convex, and let $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ be lsc. Suppose that there are functions $a \in \mathcal{L}^1(\mu)$ and $b \in \mathcal{L}^1(\nu)$ such that*

$$\vartheta(x - y) \le a(x) + b(y) \quad \text{and} \quad \int c(\mathrm{mean}(\rho), y)\, \rho(dy) \le \rho(b).$$

Then $\widehat{C}$ given in (5.22) *satisfies the assumptions of Theorem 5.1. In particular, the conclusions of Theorem 5.1 hold in the current setting for the relaxed entropic martingale transport problem.*

Proof. It remains to check that $\widehat{C}$ satisfies the assumptions of Theorem 5.1. Since c is non-negative and lsc, the same holds true for the map $\pi \mapsto \pi(c)$ for $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. Furthermore, the map $\rho \mapsto \delta_{\mathrm{mean}(\rho)} \otimes \rho$ is continuous from $\mathcal{P}_1(\mathbb{R}^d)$ to $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. Hence, $\rho \mapsto \int c(\mathrm{mean}(\rho), y)\, \rho(dy)$ is lsc on $\mathcal{P}_1(\mathbb{R}^d)$ as concatenation of these maps. As the relative entropy is lsc on $\mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$, we conclude that

$$\widehat{C}(\rho) = \int c(\mathrm{mean}(\rho), y)\, \rho(dy) + \varepsilon H(\rho|\nu)$$

is lsc on $\mathcal{P}_1(\mathbb{R}^d)$. Letting $h(x) = \varepsilon x \log(x)$, we have that $\widehat{C}$ satisfies (BM). Finally, the continuity property (CM) follows by the same argument as in the proof of Theorem 4.1. □

When we assume mild regularity properties of the cost c and ϑ, we are able to derive a more precise structure theorem that incorporates structural aspects of the entropic (martingale) optimal transport, namely that optimal martingale couplings are of Gibbs type.

Theorem 5.8. *In addition to the assumptions of Theorem 5.7 suppose that c is Lipschitz, μ is absolutely continuous and ϑ is of the form $\vartheta(x) = \bar{\vartheta}(|x|)$ for some increasing, strictly convex function $\bar{\vartheta} : [0,\infty) \to [0,\infty)$. Then $\mathrm{WT}_{C_\vartheta}(\mu,\nu) = \overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$ and both problems have a unique minimizer.*

Moreover, $\pi \in \Pi(\mu,\nu)$ and $g \in \mathcal{L}^1(\nu)$ are primal and dual optimizers of $\mathrm{WT}_{C_\vartheta}(\mu,\nu)$ if and only if there is $\Delta : \mathbb{R}^d \to \mathbb{R}^d$ measurable such that

$$\mathrm{mean}(\pi_x) \in \partial^\vartheta(\vartheta \square g^C)(x) \quad \mu\text{-a.s.}, \tag{5.25}$$

$$\frac{d\kappa}{d\eta \otimes \nu}(m,y) = \exp\Big(\frac{g^C(m) + g(y) + \Delta(m)\cdot(y-m) - c(m,y)}{\varepsilon}\Big), \tag{5.26}$$

where $\kappa = ((x,y) \mapsto (\mathrm{mean}(\pi_x), y))_\# \pi$ and $\eta = (x \mapsto \mathrm{mean}(\pi_x))_\# \mu$.

Proof. First, note that $\widehat{C}$ is strictly convex on the fibers $\{\rho \in \mathcal{P}_1(\mathbb{R}^d) : \mathrm{mean}(\rho) = m\}$ for $m \in \mathbb{R}^d$. Thanks to Theorem 5.7 and the additional assumptions on μ and ϑ, we can invoke Theorem 5.2. It remains to show the characterization of optimality given in (5.25) and (5.26).

To this end, recall from (2.16) that $J(\pi) := (x \mapsto (x, \pi_x))_\# \mu$ Theorem 5.1 ((iv)) asserts that $\pi \in \Pi(\mu,\nu)$ and $g \in \mathcal{L}^1(\nu)$ are optimal if and only if for $J(\pi)$-a.e. (x,ρ) we have (5.8) and (5.9). Clearly, (5.8) holds $J(\pi)$-almost surely if and only if (5.25) holds. Hence, it suffices to show that $\kappa = ((x,y) \mapsto (\mathrm{mean}(\pi_x), y))_\# \pi \in \Pi_{\mathrm{M}}(\eta,\nu)$ is optimal for $\mathrm{EMT}_{c,\varepsilon}(\eta,\nu)$ if and only if there is $\Delta : \mathbb{R}^d \to \mathbb{R}^d$ measurable such that κ is given by (5.26).

On the one hand, if κ is as in (5.26), then we have that for η-a.e. m

$$\kappa_m \in \arg\min\{\rho(c(m,\cdot) - g) + \varepsilon H(\rho|\nu) - \Delta(m)\cdot(\mathrm{mean}(\rho) - m) : \rho \in \mathcal{P}_1(\mathbb{R}^d),\, \rho(|g|) < \infty\},$$

because $\varepsilon H(\rho|\kappa_m)$ and $\rho(c(m,\cdot) - g) + \varepsilon H(\rho|\nu) - \Delta(m)\cdot(\mathrm{mean}(\rho) - m)$ differ as a function of m just by a constant. Since $\mathrm{mean}(\kappa_m) = m$ it follows now directly that for η-a.e. m

$$\kappa_m \in \arg\min\{\rho(c(m,\cdot) - g) + \varepsilon H(\rho|\nu) : \rho \in \mathcal{P}_1(\mathbb{R}^d),\, \mathrm{mean}(\rho) = m,\, \rho(|g|) < \infty\}.$$

Hence, $g^C = \kappa_m(c(m,\cdot) - g) + \varepsilon H(\kappa_m|\nu)$ for η-a.e. m and we conclude by Theorem 2.7 that κ is the optimizer of $\mathrm{EMT}_{c,\varepsilon}(\eta,\nu)$.

On the other hand, if $\pi \in \Pi(\mu,\nu)$ and $g \in \mathcal{L}^1(\nu)$ are primal and dual optimizers of $\mathrm{WT}_{C_\vartheta}(\mu,\nu)$, then $\kappa = ((x,y) \mapsto (\mathrm{mean}(\pi_x), y))_\# \pi \in \Pi_{\mathrm{M}}(\eta,\nu)$ and $g \in \mathcal{L}^1(\nu)$ are primal and dual optimizers of $\mathrm{EMT}(\eta,\nu)$. Wlog we assume that ν is not a Dirac measure. In order to find Δ, we define the auxiliary function

$$F(x,m) := \inf\{\rho(c(x,\cdot) - g) + \varepsilon H(\rho|\nu) : \mathrm{mean}(\rho) = m,\, \rho(|g|) < \infty\}.$$

From the definition, we see that $F(x,\cdot)$ is convex, finite on $D := \mathrm{relint}(\mathrm{co}(\mathrm{supp}(\nu)))$ (the relative interior of the convex hull of the support of ν), and $F(x,x) = g^C(x)$. It follows from the Arsenin–Kunugui selection theorem (see [57, Theorem 18.8]) that there exists a measurable map $\Delta : D \to \mathbb{R}^d$ such that $\Delta(x) \in \partial F(x,\cdot)(x)$, where we extend Δ to $\mathbb{R}^d$ by setting it 0 outside of D. We have for every $x \in D$ and $\rho \in \mathcal{P}_1(\mathbb{R}^d)$ with $\mathrm{mean}(\rho) = m$ and $\rho(|g|) < \infty$ that

$$F(x,x) + \Delta(x)\cdot(m - x) \le F(x,m) \le \rho(c(x,\cdot) - g) + \varepsilon H(\rho|\nu),$$

where we have η-a.s. equality for $\rho = \kappa_x$, since $\eta(D) = 1$ by Theorem 5.9. We can proceed as in the first part of the proof of Theorem 4.2, which yields that

$$\frac{d\kappa}{d\eta \otimes \nu} = \exp\Big(\frac{g^C(m) + g(y) + \Delta(m) \cdot (y - m) - c(m, y)}{\varepsilon}\Big),$$

and completes the proof. □

Lemma 5.9. *Under the assumptions of Theorem 5.7, the optimizer $P \in \Lambda(\mu, \nu)$ for $\overline{\mathrm{WT}}_{C_\vartheta}(\mu, \nu)$ satisfies $\rho \ll \nu$ and $\nu \ll \rho$ for P-a.e. (x, ρ). In particular, the measure $((x, \rho) \mapsto \mathrm{mean}(\rho))_{\#}P$ is concentrated on the relative interior of* $\mathrm{co}(\mathrm{supp}(\nu))$.

Proof. Since the reasoning in Theorem 2.15 also works for $P \in \Lambda(\mu, \nu)$ that are optimal for $\overline{\mathrm{WT}}_{C_\vartheta}(\mu, \nu)$, we find that P is concentrated on a C-monotone set $\Gamma \subseteq \mathbb{R}^d \times \mathcal{P}_1(\mathbb{R}^d)$. We claim that $\rho_0 \ll \rho_1$ and $\rho_1 \ll \rho_0$. As the intensity of the second marginal of P is ν, this yields that ρ is equivalent to ν P-a.s. and the assertion of the lemma readily follows.

To this end, fix $(x_i, \rho_i) \in \Gamma$ with $m_i = \mathrm{mean}(\rho_i)$, $H(\rho_i|\nu) < \infty$ and $b \in \mathcal{L}^1(\rho_i)$, $i \in \{0, 1\}$. For the sake of contradiction, suppose that ρ_0 and ρ_1 are not equivalent. Next, we define a curve of measures by

$$\rho_t := \rho_0 + t(\rho_1 - \rho_0),$$

where $t \in [0, 1]$ and set $m_t = \mathrm{mean}(\rho_t)$. By C-monotonicity of Γ, we find the inequality

$$C_\vartheta(x_0, \rho_0) + C_\vartheta(x_1, \rho_1) \leq C_\vartheta(x_0, \rho_t) + C_\vartheta(x_1, \rho_{1-t}). \tag{5.27}$$

We subtract the left-hand side from the right-hand side, divide by t and send $t \searrow 0$. Observe

$$\limsup_{t \searrow 0} \frac{1}{t}\Big(\big|\vartheta(x_0 - m_t) - \vartheta(x_0 - m_0)\big| + \big|\vartheta(x_1 - m_{1-t}) - \vartheta(x_1 - m_1)\big|\Big) < \infty, \tag{5.28}$$

since ϑ is a real-valued convex function. Since $c(\cdot, y)$ is non-negative and Lipschitz for some constant $L \geq 0$, we find that $c(m_t, \cdot) \leq c(m_0, \cdot) \wedge c(m_1, \cdot) + L|m_1 - m_0|$ and

$$\begin{aligned}\limsup_{t \searrow 0} \frac{1}{t}\Big(\int c(m_t, \cdot)\, d\rho_t - \int c(m_0, \cdot)\, d\rho_0\Big) \\ \leq \limsup_{t \searrow 0} \frac{1}{t}\Big(\int c(m_t, \cdot) - c(m_0, \cdot)\, d\rho_0\Big) + \limsup_{t \searrow 0} \int c(m_t, \cdot) d(\rho_1 - \rho_0) \\ \leq L + \rho_1(b) + L|m_1 - m_0| < \infty. \end{aligned} \tag{5.29}$$

Similarly, we find

$$\limsup_{t \searrow 0} \frac{1}{t}\Big(\int c(m_{1-t}, \cdot)\, d\rho_{1-t} - \int c(m_1, \cdot)\, d\rho_1\Big) < \infty.$$

Finally, we deal with the entropy terms in C_ϑ. For $x, y \in \mathbb{R}$ we have the inequality $\log(y)y - \log(x)x \geq (\log(x) + 1)(y - x)$, which shows that $\log\left(\frac{d\rho_t}{d\nu}\right) \vee 0 \in \mathcal{L}^1(\rho_0 + \rho_1)$. On the other hand, $\log(\frac{d\rho_t}{d\nu}) \geq \log(\frac{d\rho_0}{d\nu}) + \log(1 - t)$. Combining these two inequalities yields

$$H(\rho_0|\nu) + \log(1 - t) \leq \int \log\Big(\frac{d\rho_t}{d\nu}\Big) d\rho_0 \leq H(\rho_0|\nu),$$

and $\log(\frac{d\rho_t}{d\nu}) \in \mathcal{L}^1(\rho_0)$ for $t \in [0,1)$. Hence,

$$\begin{aligned} H(\rho_t|\nu) - H(\rho_0|\nu) &= \int \log\Big(\frac{d\rho_t}{d\nu}\Big) - \log\Big(\frac{d\rho_0}{d\nu}\Big)\, d\rho_0 + t\int \log\Big(\frac{d\rho_t}{d\nu}\Big)\, d(\rho_1 - \rho_0) \\ &\leq t\int \log\Big(\frac{d\rho_t}{d\nu}\Big)\, d(\rho_1 - \rho_0) \end{aligned}$$

where we used that $(\log(y) - \log(x))x \leq y - x$. Dividing both sides by t, we find

$$\limsup_{t\searrow 0} \Big(H(\rho_t|\nu) - H(\rho_0|\nu)\Big) \leq \int \log\Big(\frac{d\rho_0}{d\nu}\Big)\, d(\rho_1 - \rho_0),$$

by dominated convergence. We note that the right-hand side is $-\infty$ if ρ_0 is not absolutely continuous w.r.t. ρ_1. Similarly, we find that

$$\limsup_{t\searrow 0} H(\rho_{1-t}|\nu) - H(\rho_0|\nu) \leq \int \log\Big(\frac{d\rho_1}{d\nu}\Big)\, d(\rho_0 - \rho_1),$$

where the right-hand side is $-\infty$ if ρ_1 is not absolutely continuous w.r.t. ρ_0. Summarizing, using (5.27) we find the contradiction

$$0 \leq \limsup_{t\searrow 0} \frac{1}{t}\Big(C_\vartheta(x_0,\rho_t) + C_\vartheta(x_1,\rho_{1-t}) - C_\vartheta(x_0,\rho_0) - C_\vartheta(x_1,\rho_1)\Big) \leq -\infty,$$

if $\rho_0 \not\ll \rho_1$ or $\rho_1 \not\ll \rho_0$. This concludes the proof. □

5.3 Proof of Theorem 5.1 and Proposition 5.2

Proof of Theorem 5.1 ((i)). We first show that the left-hand side of (5.4) is bigger than the right-hand side. To this end, let $P \in \Lambda(\mu,\nu)$. We then write $\eta = ((x,\rho) \mapsto \mathrm{mean}(\rho))_\# P$ and define the couplings $\xi = ((x,\rho) \mapsto (x, \mathrm{mean}(\rho)))_\# P$ and $\kappa(dm,dy) = \int \delta_{\mathrm{mean}(\rho)}(dm)\rho(dy)\, P(d\rho)$. Observe that $\xi \in \Pi(\mu,\eta)$ and $\kappa \in \Pi_{\mathrm{M}}(\eta,\nu)$. Using convexity of C, we find

$$\begin{aligned} \int C_\vartheta\, dP &= \int \vartheta(x-m)\, d\xi + \int C(\mathrm{mean}(\rho),\rho)\, P(dx,d\rho) \\ &\geq \int \vartheta(x-m)\, d\xi + \int C(m,\kappa_m)\, \eta(dm) \\ &\geq \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu). \end{aligned}$$

For the reverse inequality, let $\eta \in \mathcal{P}_p(\mathbb{R}^d)$ and suppose that $\eta \leq_c \nu$ (otherwise there is nothing to prove as $\mathrm{WMT}_C(\eta,\nu) = +\infty$ in this case). Let $\xi \in \Pi(\mu,\eta)$ and $\kappa \in \Pi_{\mathrm{M}}(\eta,\nu)$ be optimal for $\mathrm{T}_\vartheta(\mu,\eta)$ and $\mathrm{WMT}_C(\eta,\nu)$ respectively. We set $P = ((x,m) \mapsto (x,\kappa_m))_\#\xi$ and observe that $P \in \Lambda(\mu,\nu)$. Using P, we estimate

$$\begin{aligned} \mathrm{WT}_{C_\vartheta}(\mu,\nu) \leq \int C_\vartheta\, dP &= \int h(x-m) + C(m,\kappa_m)\, \xi(dx,dm) \\ &= \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu). \end{aligned}$$

□

By going through the constructions given in the previous proof and in particular checking the equality cases of the estimates, it is straightforward to derive

Corollary 5.10. *Let $P \in \Lambda(\mu,\nu)$ and consider the following conditions:*

(a) $\eta = ((x,\rho) \mapsto \mathrm{mean}(\rho))_{\#}P$ *is optimal for* (5.4).

(b) $\xi = ((x,\rho) \mapsto (x, \mathrm{mean}(\rho)))_{\#}P$ *is optimal for* $\mathrm{T}_\vartheta(\mu,\eta)$.

(c) κ *is optimal for* $\mathrm{WMT}_C(\eta,\nu)$.

(d) $\rho = \kappa_{\mathrm{mean}(\rho)}$ P*-a.s.*

We then have:

(i) *If P satisfies (a)–(d) it is optimal in* $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$.

(ii) *If P is optimal in* $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$*, it satisfies (a)–(c).*

(iii) *If P is optimal in* $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$ *and $C(x,\cdot)$ is strictly convex, P satisfies (a)–(d).*

(iv) *There exists an optimizer P for* $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$ *that satisfies (a)–(d).*

Proof of Theorem 5.1 ((ii)) *and* ((iii)). We first observe that C_ϑ satisfies the assumptions of the fundamental theorem of WOT for non-convex cost (Theorem 2.17). It is easy to check that C satisfying (CM) implies that C_ϑ satisfies (C). Moreover, C_ϑ satisfies the growth condition as (BM) yields

$$C_\vartheta(x,\rho) \le \int \vartheta(x-y)\,\rho(dy) + \rho(b) + \int h\Big(\frac{d\rho}{d\nu}\Big)\,d\nu \le a(x) + \rho(2b) + \int h\Big(\frac{d\rho}{d\nu}\Big)\,d\nu.$$

Therefore, Theorem 2.17 yields duality and dual attainment for the weak transport problem with cost C_ϑ. Noting that the C_ϑ-transform is first applying the C-transform and then inf-convolving with ϑ (cf. (5.10)), we obtain (5.5).

Next, we prove the optimality criterion ((iii)). Suppose that (5.6) and (5.7) are satisfied. We then have

$$\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu) \ge \mu(\vartheta\square g^C) + \nu(g) = \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu) \ge \overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu),$$

where the first inequality follows because $(\vartheta\square g^C, g)$ is an admissible dual pair, the equality is obtained by adding (5.6) and (5.7) and the last inequality follows from (5.4). Hence, all of these inequalities are in fact equalities, i.e. η is optimal in (5.4) and g is optimal in the dual problem of $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$.

Conversely, assume that $g \in \mathcal{L}^1(\nu)$ and $\eta \in \mathcal{P}_p(\mathbb{R}^d)$ are optimal. Then we find

$$\mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu) = \overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu) = \mu(\vartheta\square g^C) - \eta(g^C) + \eta(g^C) + \nu(g). \tag{5.30}$$

As $(\vartheta\square g^C, g^C)$ and (g^C, g) are admissible in the respective dual problems,

$$\mu(\vartheta\square g^C) \le \mathrm{T}_\vartheta(\mu,\eta) \quad \text{and} \quad \eta(g^C) + \nu(g) \le \mathrm{WMT}_C(\eta,\nu).$$

By (5.30), these inequalities are in fact equalities. □

Proof of Theorem 5.2, Case ((a)). The key observation is that the assumption $C(x,\rho) = C(\rho)$ with C convex guarantees that C_ϑ is convex in the second argument. (5.11) follows from (5.4) because $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu) = \mathrm{WT}_{C_\vartheta}(\mu,\nu)$, cf. (2.18). Moreover, Theorem 2.2 ((i)) guarantees the existence of an optimizer $\pi \in \Pi(\mu,\nu)$.

Now suppose that $C(x,\cdot)$ is strictly convex. Then $C_\vartheta(x,\cdot)$ is strictly convex as well and therefore the optimal π is unique.

It remains to prove the optimality criterion (5.12). Using strict convexity of $C(x,\cdot)$, Theorem 5.10 and the fact that $J(\pi) := (x \mapsto (x,\pi_x))_\#\mu$ is optimal if π is optimal (see Section 2.5), yield the following: π is optimal for $\mathrm{WT}_{C_\vartheta}(\mu,\nu)$ if and only if $J(\pi)$ is optimal for $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$ if and only if $J(\pi)$ satisfies the conditions (a)–(d) in Theorem 5.10. It is straightforward to check that (5.12) is equivalent to the conditions (a)–(d) if $P = J(\pi)$. □

Proof of Theorem 5.2, Case ((b)). As $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu) \le \mathrm{WT}_{C_\vartheta}(\mu,\nu)$, equation (5.4) yields that

$$\mathrm{WT}_{C_\vartheta}(\mu,\nu) \ge \min_{\eta\in\mathcal{P}_p(\mathbb{R}^d)} \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu).$$

Let $\eta \in \mathcal{P}_p(\mathbb{R}^d)$, $\xi \in \Pi(\mu,\eta)$ and $\kappa \in \Pi_{\mathrm{M}}(\eta,\nu)$ all be optimal. As the assumptions of the Gangbo–McCann theorem [44, Theorem 1.2] are satisfied, $\xi = (\mathrm{id},T)_\#\mu$ for some Borel map T. We set $\pi(dx,dy) = \kappa_{T(x)}(dy)$. Observe that $\pi \in \Pi(\mu,\nu)$ and $\mathrm{mean}(\pi_x) = \mathrm{mean}(\kappa_{T(x)}) = T(x)$.

$$\begin{aligned}\mathrm{WT}_{C_\vartheta}(\mu,\nu) &\le \int \vartheta(x-\mathrm{mean}(\pi_x)) + C(\mathrm{mean}(\pi_x),\pi_x)\,\mu(dx)\\ &\le \int \vartheta(x-T(x))\,\mu(dx) + \int C(m,\kappa_m)\,\eta(dm)\\ &= \mathrm{T}_\vartheta(\mu,\eta) + \mathrm{WMT}_C(\eta,\nu).\end{aligned}$$

This proves (5.4) and attainment for $\mathrm{WT}_{C_\vartheta}(\mu,\nu)$.

In the case that $C(x,\cdot)$ is strictly convex, the proof of uniqueness of π and (5.12) follow line by line as in the proof of Theorem 5.2, Case ((a)). □

Proof of Theorem 5.1 ((iv)). Suppose that $P \in \Lambda(\mu,\nu)$ and $g \in \mathcal{L}^1(\nu)$ are optimal. By Corollary 5.10, $\eta = ((x,\rho) \mapsto \mathrm{mean}(\rho))_\#P$ is optimal for (5.4),
$\xi = ((x,\rho) \mapsto (x,\mathrm{mean}(\rho)))_\#P$ is optimal for $\mathrm{T}_\vartheta(\mu,\eta)$, and κ is optimal for $\mathrm{WMT}_C(\eta,\nu)$. By Theorem 5.1 ((iii)), $(\vartheta\square g^C, g^C)$ is a dual optimizer for $\mathrm{T}_\vartheta(\mu,\eta)$ and (g^C, g) is a dual optimizer for $\mathrm{WMT}_C(\eta,\nu)$. Hence, (5.8) follows from the complementary slackness criterion for $\mathrm{T}_\vartheta(\mu,\eta)$ and (5.9) follows from the complementary slackness criterion for $\mathrm{WMT}_C(\eta,\nu)$.

Conversely, suppose that $P \in \Lambda(\mu,\nu)$ and $g \in \mathcal{L}^1(\nu)$ satisfy (5.8) and (5.9). Adding up these conditions, implies that $P(dx,d\rho)$-a.s.

$$(\vartheta\square g^C)(x) + \rho(g) = \widehat{C}(\rho) + \vartheta(x-\mathrm{mean}(\rho)).$$

This is precisely the complementary slackness condition for $\overline{\mathrm{WT}}_{C_\vartheta}(\mu,\nu)$. Hence, Theorem 2.17 ((iv)) yields optimality of P and g. □

A Auxiliary results from convex analysis

The aim of this appendix is to give a brief recap of results from convex analysis (for more details see e.g. [18, 31, 74]) and to derive a few auxiliary results that were used throughout.

A.1 Convex functions: subdifferential, conjugation and infimal convolution

A function $f : \mathbb{R}^d \to (-\infty, \infty]$ is convex if $f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$ for all $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$. The domain of a function f is $\mathrm{dom}(f) = \{x \in \mathbb{R}^d : f(x) < \infty\}$. Moreover, f is called proper if $\mathrm{dom}(f) \neq \emptyset$ and finite if $\mathrm{dom}(f) = \mathbb{R}^d$. Every convex function is continuous in the interior of its domain.

The convex conjugate (or Legendre transform) of a function $f : \mathbb{R}^d \to [-\infty, \infty]$ is defined as

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y\rangle - f(x)\}. \tag{A.1}$$

Note that $f^{**} \leq f$. If f is proper, then f^* is proper if and only if f has an affine minorant. In this case f^{**} is the largest lsc convex function that is smaller than f. The Fenchel–Moreau theorem states that $f = f^{**}$ if and only if f is convex and lsc.

The subdifferential of a convex function $f : \mathbb{R}^d \to (-\infty, \infty]$ at $x \in \mathbb{R}^d$ is

$$\partial f(x) = \{y \in \mathbb{R}^d : f(x') \geq f(x) + \langle x' - x, y\rangle \text{ for every } x' \in \mathbb{R}^d\}.$$

By the Hahn–Banach theorem, $\partial f(x) \neq \emptyset$ for every x in the interior of $\mathrm{dom}(f)$. The Fenchel–Young inequality states that $\langle x, y\rangle \leq f(x) + f^*(y)$ for every $x, y \in \mathbb{R}^d$. Moreover, $y \in \partial f(x)$ if and only if (x, y) is an equality case in the Fenchel–Young inequality if and only if $x \in \partial f^*(y)$. Further note that there is a Borel selector of ∂f (that is a Borel function $g : \{x \in \mathbb{R}^d : \partial f(x) \neq \emptyset\} \to \mathbb{R}^d$ such that $g(x) \in \partial f(x)$). This can be seen using the Arsenin–Kunugui selection theorem, see e.g. [57, Theorem 18.8].

Note that a convex function f is Frechet-differentiable in $x \in \mathrm{dom}(f)$ if and only if it is Gateaux-differentiable in x if and only if $\partial f(x)$ is a singleton, see e.g. [18, Corollary 17.34]. Moreover, if f is differentiable on an open set $U \subset \mathrm{dom}(f)$, then $f \in C^1(U)$, see e.g. [18, Corollary 17.34].

A convex function $f : \mathbb{R}^d \to (-\infty, \infty]$ is called supercoercive if

$$\lim_{|x| \to +\infty} \frac{f(x)}{|x|} = +\infty. \tag{A.2}$$

A convex function $f : \mathbb{R}^d \to (-\infty, \infty]$ is supercoercive if and only if f^* is finite, see e.g. [31, Proposition 3.5.4]. Given a finite supercoercive convex function f, we have that f is strictly convex if and only if $f^* \in C^1(\mathbb{R}^d)$, see e.g. [18, Corollary 18.12].

The infimal convolution of two proper functions $f, g : \mathbb{R}^d \to (-\infty, \infty]$ is

$$(f \square g)(x) := \inf_{y \in \mathbb{R}^d} \{f(x - y) + g(y)\}. \tag{A.3}$$

We then have

$$(f\square g)^* = f^* + g^* \qquad \text{and} \qquad (f+g)^* = f^*\square g^*, \tag{A.4}$$

where the second assertion is only guaranteed provided that f, g are convex lsc and that the intersection between $\mathrm{dom}(f)$ and the interior of $\mathrm{dom}(g)$ is not empty, see e.g. [18, Proposition 13.21, Theorem 15.3].

Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be convex and lsc. Then $f\square\frac{1}{2}|\cdot|^2$ is the Moreau–Yosida approximation of f. For $x \in \mathbb{R}^d$, the unique minimizer in the definition of $f\square\frac{1}{2}|\cdot|^2(x)$ is called $\mathrm{Prox}_f x$, i.e. $\frac{1}{2}|\cdot|^2\square f(x) = \frac{1}{2}|x - \mathrm{Prox}_f x|^2 + f(\mathrm{Prox}_f x)$. The map $x \mapsto \mathrm{Prox}_f x$ is called proximal operator of f. Both Prox_f and $\mathrm{id} - \mathrm{Prox}_f$ are 1-Lipschitz, see e.g. [18, Section 12.4].

If f, g are proper convex lsc, and one of them is finite, differentiable and supercoercive, then $f\square g \in C^1(\mathbb{R}^d)$, see e.g. [18, Corollary 18.8]. If $f : \mathbb{R}^d \to \mathbb{R}$ is continuous and $g : \mathbb{R}^d \to (-\infty, \infty]$ is convex, replacing it by its lsc envelope, does not affect their infimal convolution (see e.g. [18, Exercise 12.10]). Specifically, we have

$$f\square g = f\square g^{**}. \tag{A.5}$$

In the context of the growth condition (B), we often work with convex functions $h : [0, \infty) \to (-\infty, \infty]$. These functions fit into the framework by extending them with the value $+\infty$ to a function on the whole real line. It is easy to see from (A.1) that for such functions h, their convex conjugate h^* is increasing as supremum of increasing functions. In the same spirit as the equivalence between supercoercivity and finiteness of the convex conjugate, we have the following:

Lemma A.1. *A lsc convex function $h : [0, \infty) \to [0, \infty)$ satisfies $\lim_{x\to+\infty} h^*(x)/x = +\infty$ if and only if $h(x) < \infty$ for all sufficiently large x.*

Proof. If $\lim_{x\to+\infty} h^*(x)/x < +\infty$, there are $\alpha, \beta > 0$ such that $h^*(x) \le \chi_{[\alpha,\infty)}(x) + \beta x$. Taking the convex conjugation of this inequality yields, $h(x) = h^{**}(x) \ge -\alpha(x - \beta) + \chi_{(-\infty,\beta]}(x)$, contradicting $h(x) < \infty$ for all sufficiently large x.

Conversely, if $h(x) = +\infty$ for all $x \ge \gamma$, then $h(x) \ge \alpha + \beta x + \chi_{[\gamma,\infty)}$. By taking the convex conjugate, we have $h^*(x) \le -\alpha - \beta\gamma + \gamma x + \chi_{(-\infty,\beta]}(x)$ and hence $\lim_{x\to+\infty} h^*(x)/x \le \gamma < \infty$. □

A.2 Support functions, convex sets and cones

A support function is a convex function $h : \mathbb{R}^d \to (-\infty, \infty]$ that is positively homogeneous, i.e. $h(tx) = th(x)$ for every $x \in \mathbb{R}^d$ and $t \ge 0$. Note that support functions are subadditive, i.e., $h(x + y) \le h(x) + h(y)$.

There is a one-to-one correspondence between support functions and closed convex subsets of $\mathbb{R}^d$. The support function of a closed convex set $K \subset \mathbb{R}^d$ is given by

$$h_K(x) = \sup_{y\in K} x \cdot y. \tag{A.6}$$

Note that we have $\partial h_K(0) = K$. As $h = h_{\partial h(0)}$ and $\partial h(0)$ is convex, every support function arises as in (A.6). Moreover, h_K is finite if and only if K is compact.

The convex indicator of a convex set $C \subset \mathbb{R}^d$ is given by $\chi_K(x) = 0$ if $x \in K$ and $\chi_K(x) = +\infty$ if $x \notin K$. Note that χ_K is lsc if and only if K is closed. In this case we have $\chi_K^* = h_K$ and $h_K^* = \chi_K$. Using (A.4) it follows that $h_{K\cap L} = h_K \square h_L$ for $K, L \subset \mathbb{R}^d$ convex.

Lemma A.2. *Let $K \subset \mathbb{R}^d$ be a compact convex set and $\phi : \mathbb{R}^d \to (-\infty, \infty]$ be a convex function. Then $\partial(h_K \square \phi)(x) \subseteq K$ for every $x \in \mathbb{R}^d$.*

Proof. Let $x \in \mathbb{R}^d$ and $y \in \partial(h_K \square \phi)(x)$. Then the subdifferential inequality yields for every $z \in \mathbb{R}^d$

$$h_K \square \phi(z) \geq h_K \square \psi(x) + \langle y, z - x \rangle.$$

By considering both sides of the inequality as a function of z and applying the convex conjugate, we find that for every $w \in \mathbb{R}^d$

$$\chi_K(w) + \phi^*(w) \leq -h_K \square \psi(x) + \langle x, y \rangle + \chi_{\{y\}}(w).$$

As the right-hand side of this inequality yields a finite value at $w = y$, the left-hand side does as well. Hence, $\chi_K(y) < \infty$, i.e. $y \in K$. □

The following characterization of convex functions that arise as infimal convolution with a given support function is straightforward to prove.

Lemma A.3. *Let $\phi : \mathbb{R}^d \to (-\infty, \infty]$ be convex and $h : \mathbb{R}^d \to (-\infty, \infty]$ be a support function. Then the following are equivalent:*

(i) $\phi = h \square \phi$.

(ii) $\phi = h \square \psi$ *for some convex function* ψ.

(iii) $\phi(x_1) - \phi(x_2) \leq h(x_1 - x_2)$ *for all* $x_1, x_2 \in \mathbb{R}^d$.

From this we immediately derive

Corollary A.4. *Let $\phi : \mathbb{R}^d \to (-\infty, \infty]$ be convex and $h_1, h_2 : \mathbb{R}^d \to (-\infty, \infty]$ be support functions and write $h = h_1 \square h_2$. Then there is a convex function ψ such that $\phi = \psi \square h$ if and only if there are convex functions ψ_1, ψ_2 such that $\phi = h_1 \square \psi_1$ and $\phi = h_2 \square \psi_2$.*

A cone is a convex set $C \subset \mathbb{R}^d$ such that for every $t \geq 0$ and $x \in C$, we have $tx \in C$. A cone is called pointed if it contains no non-trivial subspace. There is a one-to-one correspondence between pointed cones and partial orders on $\mathbb{R}^d$ that are compatible with the linear structure.

Acknowledgment: This research was funded in whole or in part by the Austrian Science Fund (FWF) [doi: 10.55776/P34743, 10.55776/P36835-N and 10.55776/P35197]. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

References

[1] B. Acciaio, M. Beiglböck, F. Penkner, and W. Schachermayer. A model-free version of the fundamental theorem of asset pricing and the super-replication theorem. *Math. Finance* 26.2 (2016), pp. 233–251.

[2] B. Acciaio, M. Beiglböck, and G. Pammer. Weak transport for non-convex costs and model-independence in a fixed-income market. *Math. Finance* 31.4 (2021), pp. 1423–1453.

[3] B. Acciaio, A. Marini, and G. Pammer. Calibration of the Bass Local Volatility model. *ArXiv e-prints* 2311.14567 (2023).

[4] A. Alfonsi, J. Corbetta, and B. Jourdain. Sampling of probability measures in the convex order and approximation of Martingale Optimal Transport problems. *ArXiv e-prints* (Sept. 2017).

[5] J.-J. Alibert, G. Bouchitté, and T. Champion. A new class of costs for optimal transport planning. *European Journal of Applied Mathematics* 30.6 (2019), pp. 1229–1263.

[6] L. Ambrosio and N. Gigli. A user's guide to optimal transport. *Modelling and optimisation of flows on networks*. Vol. 2062. Lecture Notes in Math. Springer, Heidelberg, 2013, pp. 1–155.

[7] A. Asadulaev, A. Korotin, V. Egiazarian, P. Mokrov, and E. Burnaev. Neural Optimal Transport with General Cost Functionals. *The Twelfth International Conference on Learning Representations*. 2024.

[8] J. Backhoff-Veraguas and G. Pammer. Applications of weak transport theory. *Bernoulli* 28.1 (2022), pp. 370–394.

[9] J. Backhoff-Veraguas and G. Pammer. Stability of martingale optimal transport and weak optimal transport. *Ann. Appl. Probab.* 32.1 (2022), pp. 721–752.

[10] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance Stoch.* 24.3 (2020), pp. 601–632.

[11] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. All adapted topologies are equal. *Probab. Theory Relat. Fields* 178.3-4 (2020), pp. 1125–1172.

[12] J. Backhoff-Veraguas, M. Beiglböck, W. Schachermayer, and B. Tschiderer. The structure of martingale Benamou–Brenier in multiple dimensions. *ArXiv e-prints* 2306.11019 (2023).

[13] J. Backhoff-Veraguas, M. Beiglböck, M. Huesmann, and S. Källblad. Martingale Benamou-Brenier: a probabilistic perspective. *Ann. Probab.* 48.5 (2020), pp. 2258–2289.

[14] J. Backhoff-Veraguas, M. Beiglböck, and G. Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations* 58.6 (2019), pp. 1–28.

[15] J. Backhoff-Veraguas, M. Beiglböck, and G. Pammer. Weak monotone rearrangement on the line. *Electronic Communications in Probability* 25 (2020).

[16] J. Backhoff-Veraguas, G. Loeper, and J. Obloj. Geometric Martingale Benamou-Brenier transport and geometric Bass martingales. *ArXiv e-prints* 2406.04016 (2024).

[17] D. Bartl, M. Beiglböck, and G. Pammer. The Wasserstein space of stochastic processes. *J. Eur. Math. Soc.* (2024). To appear. arXiv:2104.14245.

[18] H. H. Bauschke and P. L. Combettes. Convex analysis and monotone operator theory in Hilbert spaces. Vol. 408. Springer, 2011.

[19] E. Bayraktar and D. Norgilas. Generalizing Super/Sub MOT using weak L^1 transport. *arXiv preprint* arXiv:2407.13002 (2024).

[20] M. Beiglböck, B. Jourdain, W. Margheriti, and G. Pammer. Monotonicity and stability of the weak martingale optimal transport problem. *Annals of Applied Probability, to appear* (2024).

[21] M. Beiglböck and N. Juillet. On a problem of optimal transport under marginal martingale constraints. *Ann. Probab.* 44.1 (2016), pp. 42–106.

[22] M. Beiglböck and W. Schachermayer. Duality for Borel measurable cost functions. *Trans. Amer. Math. Soc.* 363.8 (2011), pp. 4203–4224.

[23] M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent Bounds for Option Prices: A Mass Transport Approach. *Finance Stoch.* 17.3 (2013), pp. 477–501.

[24] M. Beiglböck and N. Juillet. Shadow couplings. *Trans. Amer. Math. Soc.* 374.7 (2021), pp. 4973–5002.

[25] M. Beiglböck, T. Lim, and J. Obłój. Dual attainment for the martingale transport problem. *Bernoulli* 25.3 (2019), pp. 1640–1658.

[26] M. Beiglböck, M. Nutz, and N. Touzi. Complete duality for martingale optimal transport on the line. *Ann. Probab.* 45.5 (2017), pp. 3038–3074.

[27] M. Beiglböck, G. Pammer, and L. Riess. Change of numeraire for weak martingale transport. *arXiv e-prints* 2406.07523 (2024).

[28] J.-D. Benamou, G. Chazareix, M. Hoffmann, G. Loeper, and F.-X. Vialard. Entropic Semi-Martingale Optimal Transport. *Preprint* (2024).

[29] D. P. Bertsekas and S. E. Shreve. Stochastic optimal control. Vol. 139. Mathematics in Science and Engineering. The discrete time case. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1978, pp. xiii+323. ISBN: 0-12-093260-1.

[30] M. Blondel, V. Seguy, and A. Rolet. Smooth and Sparse Optimal Transport. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, Apr. 2018, pp. 880–889.

[31] J. Borwein and J. Vanderwerff. Convex Functions: Constructions, Characterizations and Counterexamples. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2010. ISBN: 9781139811095.

[32] M. Bowles and N. Ghoussoub. A Theory of Transfers: Duality and convolution. *arXiv e-prints*, arXiv:1804.08563 (Apr. 2018), arXiv:1804.08563.

[33] K. Bredies, E. Chenchene, and A. Hosseini. A hybrid proximal generalized conditional gradient method and application to total variation parameter learning (2022).

[34] D. T. Breeden and R. H. Litzenberger. Prices of State-contingent Claims Implicit in Option Prices. *The Journal of Business* 51.4 (1978), pp. 621–51.

[35] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.* 44.4 (1991), pp. 375–417.

[36] P. Choné, N. Gozlan, and F. Kramarz. Weak Optimal Transport with Unnormalized Kernels. *SIAM Journal on Mathematical Analysis* 55.6 (2023), pp. 6039–6092.

[37] G. Conforti. A second order equation for Schrödinger bridges with applications to the hot gas experiment and entropic transportation cost. *Probability Theory and Related Fields* 174.1 (2019), pp. 1–47.

[38] A. Conze and P. Henry-Labordere. Bass Construction with Multi-Marginals: Lightspeed Computation in a New Local Volatility Model. en. *SSRN Electronic Journal* (2021).

[39] A. M. Cox and M. Vidmar. The structure of non-linear martingale optimal transport problems. *arXiv preprint* (2019).

[40] C. Daskalakis, A. Deckelbaum, and C. Tzamos. Strong Duality for a Multiple-Good Monopolist. *Econometrica* 85.3 (2017), pp. 735–767.

[41] A. Dessein, N. Papadakis, and J.-L. Rouas. Regularized Optimal Transport and the Rot Mover's Distance. *Journal of Machine Learning Research* 19.15 (2018), pp. 1–53.

[42] M. Fathi, N. Gozlan, and M. Prod'homme. A proof of the Caffarelli contraction theorem via entropic regularization. *Calculus of Variations and Partial Differential Equations* 59 (2020), pp. 1–18.

[43] M. Fathi and Y. Shu. Curvature and transport inequalities for Markov chains in discrete spaces. *Bernoulli* 24.1 (2018), pp. 672–698.

[44] W. Gangbo and R. McCann. The geometry of optimal transportation. *Acta Math.* 177.2 (1996), pp. 113–161.

[45] A. González-Sanz and M. Nutz. Quantitative Convergence of Quadratically Regularized Linear Programs. 2024.

[46] A. González-Sanz and M. Nutz. Sparsity of Quadratically Regularized Optimal Transport: Scalar Case. 2024.

[47] A. González-Sanz, M. Nutz, and A. R. Valdevenito. Monotonicity in Quadratically Regularized Linear Programs. 2024.

[48] N. Gozlan, C. Roberto, P.-M. Samson, Y. Shu, and P. Tetali. Characterization of a class of weak transport-entropy inequalities on the line. Preprint arXiv:1509.04202v2. 2015.

[49] N. Gozlan and N. Juillet. On a mixture of Brenier and Strassen theorems. *Proceedings of the London Mathematical Society* 120.3 (2020), pp. 434–463.

[50] N. Gozlan, C. Roberto, P.-M. Samson, Y. Shu, and P. Tetali. Characterization of a class of weak transport-entropy inequalities on the line. *Ann. Inst. Henri Poincaré Probab. Stat.* 54.3 (2018), pp. 1667–1693.

[51] N. Gozlan, C. Roberto, P.-M. Samson, and P. Tetali. Kantorovich duality for general transport costs and applications. *J. Funct. Anal.* 273.11 (2017), pp. 3327–3405.

[52] I. Guo, G. Loeper, and S. Wang. Local volatility calibration by optimal transport. *2017 MATRIX Annals*. Springer, 2019, pp. 51–64.

[53] J. Guyon, R. Menegaux, and M. Nutz. Bounds for VIX futures given S&P 500 smiles. *Finance and Stochastics* 21.3 (2017), pp. 593–630.

[54] P. Henry-Labordère and N. Touzi. An explicit martingale version of the one-dimensional Brenier theorem. *Finance Stoch.* 20.3 (2016), pp. 635–668.

[55] D. Hobson and A. Neuberger. Robust bounds for forward start options. *Math. Finance* 22.1 (2012). Ed. by I. Wiley-Blackwell Publishing, pp. 31–56.

[56] L. Kantorovich and G. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ.* 13.7 (1958), pp. 52–59.

[57] A. S. Kechris. Classical descriptive set theory. Vol. 156. Graduate Texts in Mathematics. New York: Springer-Verlag, 1995, pp. xviii+402. ISBN: 0-387-94374-9.

[58] A. Kolesov, P. Mokrov, I. Udovichenko, M. Gazdieva, G. Pammer, E. Burnaev, and A. Korotin. Estimating barycenters of distributions with neural optimal transport. English. *Proceedings of the 41st International Conference on Machine Learning*. 2024.

[59] J. Komlós. A generalization of a problem of Steinhaus. *Acta Math. Acad. Sci. Hungar.* 18 (1967), pp. 217–229.

[60] A. Korotin, D. Selikhanovych, and E. Burnaev. Neural Optimal Transport. *The Eleventh International Conference on Learning Representations*. 2023.

[61] M. Kupper, M. Nendel, and A. Sgarabottolo. Risk measures based on weak optimal transport. *Quantitative Finance* (2024), pp. 1–18.

[62] J. Lehec. Representation formula for the entropy and functional inequalities. *Ann. Inst. Henri Poincaré Probab. Stat.* 49.3 (2013), pp. 885–899.

[63] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.* 34.4 (2014), pp. 1533–1574.

[64] G. Loeper. Option pricing with linear market impact and nonlinear Black-Scholes equations. *Ann. Appl. Probab.* 28.5 (2018), pp. 2664–2726.

[65] D. A. Lorenz, P. Manns, and C. Meyer. Quadratically regularized optimal transport. English. *Appl. Math. Optim.* 83.3 (2021), pp. 1919–1949.

[66] M. Beiglböck, A. Cox, and M. Huesmann. Optimal Transport and Skorokhod Embedding. *Invent. Math.* 208.2 (2017), pp. 327–400.

[67] S. D. Marino and A. Gerolin. Optimal Transport losses and Sinkhorn algorithm with general convex regularization. 2020.

[68] K. Marton. A measure concentration inequality for contracting Markov chains. *Geometric & Functional Analysis GAFA* 6.3 (1996), pp. 556–571.

[69] K. Marton. Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration. *The Annals of Probability* 24.2 (1996), pp. 857–866.

[70] M. Nutz. Introduction to Entropic Optimal Transport. `https://www.math.columbia.edu/mnutz/docs/EOTlecturenotes.pdf` (2022).

[71] M. Nutz. Quadratically Regularized Optimal Transport: Existence and Multiplicity of Potentials. 2024.

[72] M. Nutz and J. Wiesel. On the Martingale Schrödinger Bridge between Two Distributions. *arXiv preprint* (2024).

[73] F. Pramenkovic. Imposing Convexity on Dual Formulations of Transport Problems. Diplomarbeit. University of Vienna, 2024.

[74] R. T. Rockafellar. Convex analysis. Princeton Landmarks in Mathematics. Reprint of the 1970 original, Princeton Paperbacks. Princeton University Press, Princeton, NJ, 1997, pp. xviii+451. ISBN: 0-691-01586-4.

[75] P.-M. Samson. Transport-entropy inequalities on locally acting groups of permutations. *Electron. J. Probab.* 22 (2017), Paper No. 62, 33.

[76] S. Schrott and D. Toneian. On Strassen's theorem for support functions. *Electronic Communications in Probability* 29 (2024), pp. 1–12.

[77] Y. Shu. From Hopf–Lax Formula to Optimal Weak Transfer Plan. *SIAM Journal on Mathematical Analysis* 52.3 (2020), pp. 3052–3072.

[78] Y. Shu. Hamilton-Jacobi equations on graph and applications. *Potential Anal.* 48.2 (2018), pp. 125–157.

[79] V. Strassen. The existence of probability measures with given marginals. *Ann. Math. Statist.* 36 (1965), pp. 423–439.

[80] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques* 81.1 (1995), pp. 73–205.

[81] M. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae* 126.3 (1996), pp. 505–563.

[82] X. Tan and N. Touzi. Optimal transportation under controlled stochastic dynamics. *Ann. Probab.* 41.5 (2013), pp. 3201–3240.

[83] C. Villani. Topics in Optimal Transportation. Vol. 58. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, 2003, pp. xvi+370. ISBN: 0-8218-3312-X.

[84] J. Wiesel and E. Zhang. A characterisation of convex order using the 2-Wasserstein distance. *arXiv preprint* (2022).

Change of numeraire for weak martingale transport

M. Beiglböck[†], G. Pammer[‡], and L. Riess[†]

Abstract

Change of numeraire is a classical tool in mathematical finance. Campi–Laachir–Martini [14] established its applicability to martingale optimal transport. We note that the results of [14] extend to the case of weak martingale transport. We apply this to shadow couplings (in the sense of [12]), continuous time martingale transport problems in the framework of Huesmann–Trevisan [21] and in particular to establish the correspondence between stretched Brownian motion with its geometric counterpart.[9]

[†]University of Vienna
[‡]ETH Zürich

[9]We emphasize that we learned about the geometric stretched Brownian motion gSBM (defined in PDE terms) in a presentation of Loeper [25] before our work on this topic started. We noticed that a change of numeraire transformation in the spirit of [14] allows for an alternative viewpoint in the weak optimal transport framework. We make our work public following the publication of Backhoff–Loeper–Obloj's work [7] on `arxiv.org`. The article [7] derives gSBM using PDE techniques as well as through an independent probabilistic approach which is close to the one we give in the present article.

1 Overview

While classical transport theory is concerned with the set $\Pi(\mu,\nu)$ of *couplings* or *transport plans* of probability measures μ,ν, the martingale variant restricts the problem to the set $\mathscr{M}_{\{0,1\}}(\mu,\nu)$ of *martingale couplings*, i.e. laws of two step martingales $(X_t)_{t\in\{0,1\}}$ satisfying $X_0\sim\mu, X_1\sim\nu$. The investigation of such martingale transport plans is primarily motivated by applications in mathematical finance. In this case the interest lies in probabilities μ,ν which are concentrated on the positive half line $\mathbb{R}_{++}=(0,\infty)$, have finite first moments and are in convex order. We make this a standing assumption throughout.

In this article we revisit the change of numeraire technique which was first used in the context of martingale optimal transport by Campi–Laachir–Martini [14]. This *CN-transformation* (for brevity) converts a martingal transport plan π between μ,ν into a martingale transport plan '$S(\pi)$' between altered marginals ('$S(\mu),S(\nu)$') changing cost functionals on the way. In particular this operation allows to transfer known solutions of martingale transport problems to solutions of altered problems.

In Section 1.1 we summarize the basic properties of the CN-transformation in the case of *weak martingale optimal transport* (wMOT). Weak versions of transport problems were introduced by Gozlan–Roberto–Samson–Tetali [17] and allow to consider non-linear cost functionals, see e.g. the survey [3] and [9] for an overview in the martingale context.

We then collect different applications:

In Section 2 we apply the CN-transformation to stretched Brownian motions (or 'Bass martingales'), [5, 6, 18, 26]. Stretched Brownian motion (SBM) is the process which has maximal correlation with Brownian motion, subject to having μ,ν as prescribed initial and terminal marginals. From a mathematical finance perspective it is natural to consider a geometric stretched Brownian motion (gSBM) where the log returns maximize the correlation with Brownian motion subject to the marginal constraints. We show that gSBM is the CN-transform of SBM and provide numerical simulations based on the known algorithm for SBM ([2, 15, 22]).

In Section 3 we consider the adapted Wasserstein distance $\mathcal{AW}$ which provides a metric between continuous martingales which takes the flow of information into account. It is known from [6] that stretched Brownian motion is the $\mathcal{AW}$-metric projection of Brownian motion onto the set of martingales with prescribed initial and terminal marginal. We introduce the geometric counterpart $g\mathcal{AW}$ of the adapted Wasserstein distance and establish that gSBM is a $g\mathcal{AW}$-metric projection.

In Section 4 we revisit continuous-time martingale transport problems in the spirit of Huesmann–Trevisan [21] and Guo–Loeper–Wang [18].

In Section 5 we consider the class of shadow couplings of martingale transport plans, see [12]. We find that the CN-transformation of a shadow coupling is again a shadow coupling but with 'inverted' source. In particular this implies that left-monotone martingale couplings are transformed into right-monotone martingale couplings (recovering a result of [14]) and that 'sunset-couplings' are invariant under the CN-transformation.

1.1 Martingale transport and the CN-transformation

We consider a set of time indices $I \subset [0,\infty)$ which is compact and denote its minimal element by o and its maximal element by T. Examples which we have in mind are $I = \{0,1\}, I = [0,1]$ and $I = \{t\}$ where $t \in [0,\infty)$. Throughout, we will denote for a probability η on $\mathbb{R}$ and a measurable function $f : \mathbb{R} \to \mathbb{R}$, the push-forward, by $f_{\#}\eta$. Given a function g on $\mathbb{R}$ we will write $\eta(g) := \int g\, d\eta$ whenever the integral exists. Additionally, we write $b(\eta)$ for the barycenter of the probability η.

Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R}_{++})$, that is, μ and ν are probabilities on $\mathbb{R}_{++}$ with finite first moments. We write $\mathscr{M}_I$ for the set of all laws of (cadlag) martingales $(X_t)_{t\in I}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and denote by $\mathscr{M}_I(\mu,\nu)$ the subset of martingale transport plans additionally satisfying $X_o \sim \mu, X_T \sim \nu$. A *cost function* is a measurable, $[0,\infty]$-valued function c defined on the *path space* $\mathcal{D}_I$ of cadlag $\mathbb{R}_{++}$-valued functions on I. The corresponding martingale transport problem consists in

$$\inf\Big\{ \int c\, d\pi : \pi \in \mathscr{M}_I(\mu,\nu)\Big\} = \inf\Big\{\mathbb{E}_{\mathbb{P}}[c(X)] : X \text{ mart., } \mathrm{law}_{\mathbb{P}}(X_o) = \mu, \mathrm{law}_{\mathbb{P}}(X_T) = \nu\Big\}.$$

Given a measurable lower bounded 'generalized' cost function $C : \mathcal{P}_1(\mathbb{R}_{++}) \to \mathbb{R} \cup \{\infty\}$, the weak martingale transport problem is

$$\begin{aligned}\inf\Big\{ \int C(\pi_x)\, \mu(dx) : \pi \in \mathscr{M}_{\{0,1\}}(\mu,\nu)\Big\}& \\ = \inf\Big\{\mathbb{E}_{\mathbb{P}}[C(\mathrm{law}_{\mathbb{P}}(X_1|X_0))] : X \text{ mart., } \mathrm{law}_{\mathbb{P}}(X_0) = \mu, \mathrm{law}_{\mathbb{P}}(X_1) = \nu\Big\},&\end{aligned}$$

where we use $(\pi_x)_x$ to denote the disintegration of π with respect to the first marginal. Usually cost functions are assumed to be lower semi-continuous and C is assumed to be convex which guarantees the existence of optimizers as well as weak duality, but this is not important at the current stage.

Next we discuss the classical change of numeraire transformation: Recall that *Bayes' rule* asserts that for integrable random variables D, Z on $(\Omega, \mathcal{F}, \mathbb{P})$, where $D > 0$ and $\mathcal{G} \subseteq \mathcal{F}$ and $\hat{\mathbb{P}} := \frac{D}{\mathbb{E}_{\mathbb{P}}[D]}\mathbb{P}$

$$\mathbb{E}_{\hat{\mathbb{P}}}[Z|\mathcal{G}] = \frac{\mathbb{E}_{\mathbb{P}}[ZD|\mathcal{G}]}{\mathbb{E}_{\mathbb{P}}[D|\mathcal{G}]}.$$

Given a path $x = (x_t)_{t\in I}$ we write $1/x$ for the path $(1/x_t)_{t\in I}$. For a $\mathbb{P}$-martingale $X = (X_t)_{t\in I}$, the process $Y := 1/X$ is a $\hat{\mathbb{P}}$-martingale where $\hat{\mathbb{P}} := \frac{X_T}{\mathbb{E}_{\mathbb{P}}[X_T]}\mathbb{P}$. This follows directly from Bayes rule since for $t \in I$

$$\mathbb{E}_{\hat{\mathbb{P}}}[Y_T|Y_t] = \frac{\mathbb{E}_{\mathbb{P}}[Y_T X_T|\mathcal{F}_t]}{\mathbb{E}_{\mathbb{P}}[X_T|\mathcal{F}_t]} = \frac{1}{X_t} = Y_t.$$

Importantly, $\mathrm{law}_{\hat{\mathbb{P}}}(Y)$ depends only on $\mathrm{law}_{\mathbb{P}}(X)$ since for $\pi := \mathrm{law}_{\mathbb{P}}(X)$ and measurable $B \subseteq \mathcal{D}(I)$

$$\mathrm{law}_{\hat{\mathbb{P}}}(Y)(B) = \mathbb{E}_{\mathbb{P}}\Big[\frac{X_T I_{\{1/X\in B\}}}{\mathbb{E}_{\mathbb{P}}[X_T]}\Big] = \frac{\int x_T \delta_{1/x}(B)\, \pi(dx)}{\int x_T\, \pi(dx)}.$$

We will call this transformation

$$S : \mathscr{M}_I \to \mathscr{M}_I, \quad S(\pi)(B) := \frac{\int x_T \delta_{1/x}(B)\, \pi(dx)}{\int x_T\, \pi(dx)}, \quad \text{for measurable } B \subseteq \mathcal{D}_I \tag{1.1}$$

the *CN*-transformation. We identify the sets $\mathscr{M}_{\{t\}}$ (where $t \in \mathbb{R}$) and the set $\mathcal{P}_1(\mathbb{R}_{++})$ of probabilities with finite first moments such that $S(\pi)$ is also defined for $\pi \in \mathcal{P}_1(\mathbb{R}_{++})$.

In the following two theorems we derive the basic properties of S. In the statement as well as in the proof we will use the representation $S(\pi) = \mathrm{law}_{\hat{\mathbb{P}}}(Y)$, where $\pi, X, Y, \mathbb{P}, \hat{\mathbb{P}}$ are as above. For $J \subseteq I$ we write $\mathrm{proj}^{\mathcal{D}_J}$ for the projection $\mathrm{proj}^{\mathcal{D}_J} : \mathcal{D}_I \to \mathcal{D}_J, (x_t)_{t\in I} \mapsto (x_t)_{t\in J}$.

Theorem 1.1. *The map S has the following properties:*

(i) S is an involution, i.e.

$$S(S(\pi)) = \pi \quad \textit{for } \pi \in \mathscr{M}_I. \tag{P1}$$

(ii) If $J \subseteq I$ is compact, in particular if $J = \{t\}$ for some $t \in I$, we have

$$\mathrm{proj}^{\mathcal{D}_J}_{\#}(S(\pi)) = S(\mathrm{proj}^{\mathcal{D}_J}_{\#}(\pi)) \quad \textit{for } \pi \in \mathscr{M}_I. \tag{P2}$$

(iii) For $\mu, \nu \in \mathcal{P}_1(\mathbb{R}_{++})$

$$S : \mathscr{M}_I(\mu, \nu) \to \mathscr{M}_I(S(\mu), S(\nu)) \tag{P3}$$

is a bijection.

(iv) The conditional distributions satisfy

$$\mathrm{law}_{\hat{\mathbb{P}}}(Y_T|Y_o) = S(\mathrm{law}_{\mathbb{P}}(X_T|X_o)). \tag{P4}$$

Proof. To show (P1), note that since $S(P) = \mathrm{law}_{\hat{\mathbb{P}}}(Y)$, we obtain $S(S(P)) = \mathrm{law}_{\bar{\mathbb{P}}}(Z)$ for $Z := 1/Y$ and $\bar{\mathbb{P}} := \frac{Y_T}{\mathbb{E}_{\hat{\mathbb{P}}}[Y_T]}\hat{\mathbb{P}}$. Since $Z = 1/(1/X) = X$ and

$$\bar{\mathbb{P}} = \frac{1/X_T}{\mathbb{E}_{\hat{\mathbb{P}}}[1/X_T]}\hat{\mathbb{P}} = \frac{1/X_T}{\mathbb{E}_{\mathbb{P}}[X_T/(X_T\mathbb{E}_{\mathbb{P}}[X_T])]}\frac{X_T}{\mathbb{E}_{\mathbb{P}}[X_T]}\mathbb{P} = \mathbb{P},$$

we obtain $S(S(\mathbb{P})) = \mathbb{P}$.

To show (P2), let $R := \max J$. For continuous bounded f on $\mathcal{D}_J$ we have by the tower property

$$\mathrm{proj}^{\mathcal{D}_J}_{\#}(S(\pi))(f) = \mathbb{E}_{\mathbb{P}}\Big[\frac{f((1/X_t)_{t\in J})X_T}{\mathbb{E}_{\mathbb{P}}[X_T]}\Big] = \mathbb{E}_{\mathbb{P}}\Big[\frac{f((1/X_t)_{t\in J})X_R}{\mathbb{E}_{\mathbb{P}}[X_R]}\Big] = (S(\mathrm{proj}^{\mathcal{D}_J}_{\#}\pi))(f).$$

Observe that combining (P1) and (P2) shows (P3). Therefore it remains to prove (P4). Note that given a continuous bounded function f on $\mathbb{R}_{++}$ we obtain

$$\begin{aligned}\mathrm{law}_{\hat{\mathbb{P}}}(Y_T|Y_o)(f) = \mathbb{E}_{\hat{\mathbb{P}}}[f(Y_T)|Y_o] &= \frac{\mathbb{E}_{\mathbb{P}}[X_T f(1/X_T)|X_o]}{\mathbb{E}_{\mathbb{P}}[X_T|X_o]} \\ &= \frac{\int x_T f(1/x_T)\, d\mathrm{law}_{\mathbb{P}}(X_T|X_o)(x_T)}{\int x_T\, d\mathrm{law}_{\mathbb{P}}(X_T|X_o)(x_T)} = S(\mathrm{law}_{\mathbb{P}}(X_T|X_o))(f),\end{aligned}$$

where we have used Bayes' rule and (1.1). □

Remark 1.2. In our opinion (P3) is a remarkable observation of [14]. In particular it directly suggests the connection to the martingale transport problem.

Remark 1.3. We can reformulate (P4) in terms of kernels which will be useful later on. Consider $\pi \in \mathscr{M}_{\{0,1\}}(\mu,\nu)$ and let $\mathrm{law}_{\mathbb{P}}((X_0,X_1)) = \pi$, then we directly obtain from (P3) and (P4) that

$$S(\pi)(dy_0,dy_1) = S(\mu)(dy_0)S(\pi_{1/y_0})(dy_1). \tag{1.2}$$

Remark 1.4. In martingale transport, the pair (μ,ν) is called *irreducible* if for all measurable $A,B \subseteq \mathbb{R}, \mu(A),\nu(B) > 0$ there exists $\pi \in \mathscr{M}_{\{0,1\}}(\mu,\nu)$ such that $\pi(A\times B) > 0$. This condition often allows for cleaner results e.g. in the context of dual attainment [13] or the description of optimizers [6]. It is a natural assumption in the mathematical finance context, where it corresponds to call prices being strictly increasing w.r.t. time to maturity. From Theorem 1.1 (P3) it follows that (μ,ν) is irreducible if and only if $(S(\mu),S(\nu))$ is irreducible.

In the present one-dimensional case it is straightforward to see that (μ,ν) can be decomposed into countably many irreducible components (cf. [10, Theorem A.4]) and, again by Theorem 1.1 (P3) it is plain that the irreducible components of (μ,ν) correspond to the ones of $(S(\mu),S(\nu))$.

The next theorem covers the relation of the CN transformation and martingale optimal transport problems, as well as weak martingale optimal transport problems. For this reason, we define for measurable $c : \mathcal{D}_I \to [0,\infty]$ the function $S^*(c)$ by

$$S^*(c)(x) := x_T c(1/x), \quad x \in \mathcal{D}_I. \tag{1.3}$$

Furthermore, for measurable $C : \mathcal{P}_1(\mathbb{R}_{++}) \to [0,\infty]$, set $S^*(C)(\rho) := b(\rho)C(S(\rho))$, $\rho \in \mathcal{P}_1(\mathbb{R}_{++})$.

Theorem 1.5. *For the martingale optimal transport problem we have*

$$\frac{1}{b(\mu)} \inf\Big\{ \int S^*(c)\, d\pi : \pi \in \mathscr{M}_I(\mu,\nu)\Big\} = \inf\Big\{ \int c\, d\hat{\pi} : \hat{\pi} \in \mathscr{M}_I(S(\mu),S(\nu))\Big\}, \tag{P5}$$

and $\pi \in \mathscr{M}_I(\mu,\nu)$ minimizes the left-hand side in (P5) *if and only if $S(\pi)$ minimizes the right-hand side. Similarly, for the weak martingale optimal transport problem we have*

$$\begin{aligned}\frac{1}{b(\mu)} \inf\Big\{ \int S^*(C)(\pi_x)\, \mu(dx) : \pi \in \mathscr{M}_{\{0,1\}}(\mu,\nu)\Big\}& \\ = \inf\Big\{ \int C(\hat{\pi}_y)\, S(\mu)(dy) : \hat{\pi} \in \mathscr{M}_{\{0,1\}}(S(\mu),S(\nu))\Big\},& \end{aligned} \tag{P6}$$

and $\pi \in \mathscr{M}_{\{0,1\}}(\mu,\nu)$ minimizes the left-hand side in (P6) *if and only if $S(\pi)$ minimizes the right-hand side.*

Proof. To show (P5), note that by Bayes' rule we have

$$\mathbb{E}_{\hat{\mathbb{P}}}[c(Y)] = \frac{\mathbb{E}_{\mathbb{P}}[X_T c(1/X)]}{\mathbb{E}_{\mathbb{P}}[X_T]} = \frac{1}{\mathbb{E}_{\mathbb{P}}[X_T]}\mathbb{E}_{\mathbb{P}}[S^*(c)(X)].$$

Similarly, to obtain (P6) we use (P4) to obtain

$$\mathbb{E}_{\hat{\mathbb{P}}}[C(\mathrm{law}_{\hat{\mathbb{P}}}(Y_1|Y_0))] = \mathbb{E}_{\mathbb{P}}\Big[\frac{X_1}{\mathbb{E}_{\mathbb{P}}[X_1]} C(S(\mathrm{law}_{\mathbb{P}}(X_1|X_0)))\Big] = \frac{1}{\mathbb{E}_{\mathbb{P}}[X_1]}\mathbb{E}_{\mathbb{P}}[S^*(C)(\mathrm{law}_{\mathbb{P}}(X_1|X_0))].$$

□

We remark that most of the above properties (P1)-(P6) were already established in [14, Lemma 2.3] while the novelty lies in also covering the weak martingale optimal transport setting.

2 Geometric stretched Brownian motion

2.1 gSBM as CN-transformation of SBM

If μ, ν have finite second moment, then the optimization problem

$$V^{\mathrm{SBM}}(\mu,\nu) := \sup\Big\{\mathbb{E}\Big[\int_0^1 \sigma_t\, dt\Big] : dX_t = \sigma_t\, dB_t, X_0 \sim \mu, X_1 \sim \nu\Big\} \tag{2.1}$$

has a unique-in-law solution $X^{\mathrm{SBM}} = X^{\mathrm{SBM},\mu,\nu}$. $X^{\mathrm{SBM},\mu,\nu}$ or, depending on the context, its law $\mathbb{Q}^{\mathrm{SBM},\mu,\nu}$ is called *stretched Brownian motion* from μ to ν. Stretched Brownian motion is a strong-Markov martingale ([6, Corollary 2.5]) which admits a precise structural description in the spirit of Brenier's theorem, see Section 2.2 below.

In the investigation of SBM it is key that $V^{\mathrm{SBM}}(\mu,\nu)$ is also the solution of the wMOT problem

$$V^{\mathrm{SBM}}(\mu,\nu) = \sup_{\pi \in \mathscr{M}_{\{0,1\}}(\mu,\nu)} \int \mathrm{MCov}(\pi_x, \gamma_1)\, \mu(dx), \tag{2.2}$$

where $\mathrm{MCov}(p_1,p_2) := \sup_{q\in\Pi(p_1,p_2)} \int x_1 x_2\, dq$ and γ_1 is the centered Gaussian with variance 1. Moreover, (2.2) has a unique solution $\pi^{\mathrm{SBM},\mu,\nu}$ and $(X_t^{\mathrm{SBM},\mu,\nu})_{t=0,1} \sim \pi^{\mathrm{SBM},\mu,\nu}$. On the other hand, it is also straightforward to reconstruct $X^{\mathrm{SBM},\mu,\nu}$ from $\pi^{\mathrm{SBM},\mu,\nu}$, see [6, Theorem 2.2].

Problem (2.1) can be interpreted as maximizing the correlation with Brownian motion subject to marginal constraints. From a mathematical finance perspective it appears equally natural to maximize the correlation of the *log return* with Brownian motion, e.g. to consider geometric stretched Brownian motions which solve

$$V^{\mathrm{gSBM}}(\mu,\nu) := \sup\Big\{\mathbb{E}\Big[\int_0^1 \sigma_t\, dt\Big] : dY_t = \sigma_t Y_t\, dB_t, Y_0 \sim \mu, Y_1 \sim \nu\Big\}. \tag{2.3}$$

In the main result of this section we show that (2.3) is equivalent to a wMOT (2.4) which is the CN-transformation of the wMOT-characterization of SBM in (2.2). Consequently gSBMs are precisely the (continuous-time) CN-transformations of SBMs.

Theorem 2.1. *Assume that $S(\nu)$ has finite second moment. Then*

$$V^{\mathrm{gSBM}}(\mu,\nu) = \sup_{\pi\in\mathscr{M}_{\{0,1\}}(\mu,\nu)} \int x\mathrm{MCov}(S(\pi_x),\gamma_1)\,\mu(dx) \quad \Big(= b(\mu)V^{\mathrm{SBM}}(S(\mu),S(\nu))\Big). \tag{2.4}$$

Problem (2.3) *has a unique-in-law solution $\mathbb{Q}^{\mathrm{gSBM},\mu,\nu}$ ('*geometric stretched Brownian motion*') and $\mathbb{Q}^{\mathrm{gSBM},\mu,\nu}$ is given by the continuous-time CN-transformation of the stretched Brownian motion from $S(\mu)$ to $S(\nu)$, i.e.,*

$$\mathbb{Q}^{\mathrm{gSBM},\mu,\nu} = S\left(\mathbb{Q}^{\mathrm{SBM},S(\mu),S(\nu)}\right). \tag{2.5}$$

2.2 Structure of gSBM and numerical simulation

To obtain numerical simulations of SBM and, in turn, gSBM one uses that SBM admits a precise structural description. Specifically, if B denotes Brownian motion started in a probability α and $f:\mathbb{R}\to\mathbb{R}$ is such that $f(B_1)$ has finite second moment, then the martingale

$$M_t^\alpha := \mathbb{E}[f(B_1)|B_t],\quad t\in[0,1]$$

is called *Bass martingale*.

Assume that $(\bar\mu,\bar\nu)$ is irreducible and that $\bar\nu$ has finite second moment. Then [6, Theorem 3.1], asserts that X is the stretched Brownian motion from $\bar\mu$ to $\bar\nu$ if and only if X is a Bass martingale. In this case the so called *Bass measure* α is unique up to translation.

In order to simulate paths from a gSBM between μ and ν we can, by Theorem 2.1, simulate paths from a SBM between $S(\mu)$ and $S(\nu)$ and convert them. Furthermore, we can assume the pair (μ,ν) to be irreducible, since otherwise we can simulate paths on each irreducible component.

Let us therefore recall how to find a Bass measure α for the SBM between $S(\mu)$ and $S(\nu)$, as well as how to simulate paths from a SBM. This is well described by the following diagram:

$$\begin{array}{ccc} Y_0\sim\mu & \xrightarrow{Y_t} & Y_1\sim\nu \\ \big\uparrow S & & \big\uparrow S \\ X_0\sim S(\mu) & \xrightarrow{X_t} & X_1\sim S(\nu) \\ {\scriptstyle T_{\alpha,S(\mu)}}\big\uparrow & & {\scriptstyle T_{\alpha*\gamma_1,S(\nu)}}\big\uparrow \\ B_0\sim\alpha & \xrightarrow{B_t} & B_1\sim\alpha*\gamma_1 \end{array} \tag{2.6}$$

Here, we denote by T_{θ_1,θ_2} the monotone transport map from θ_1 to θ_2 and by γ_t the centered Gaussian with variance t. Given a distribution $\alpha\in\mathcal{P}(\mathbb{R}_{++})$ and denoting by ϕ the density of the standard Gaussian, the diagram (2.6) commutes if and only if $T_{\alpha,S(\mu)} = \phi * T_{\alpha*\gamma_1,S(\nu)}$, i.e. α is the corresponding Bass measure. Hence, it remains to find a suitable α which can be achieved, for example, by using the fixed-point operator $\mathcal{A}:\mathrm{CDF}\to\mathrm{CDF}$, c.f. [2, 15, 22],

$$\mathcal{A}F := F_{S(\mu)}\circ(\phi*(Q_{S(\nu)}\circ(\phi*F))), \tag{2.7}$$

where we denote by F_θ the cdf of the measure θ and by Q_η the quantile function of the measure η.

The operator $\mathcal{A}$ admits a (up to translations) unique fixed point if $S(\mu) \leq_c S(\nu)$ are in convex order, the pair $(S(\mu), S(\nu))$ is irreducible and certain regularity properties are satisfied, c.f. [2, Assumption 3.1] for details. Assume those are satisfied, then the fixed point F_α is the cdf of the starting law α of the Brownian martingale and the above diagram commutes. This enables us to propose an algorithm for simulating paths from a gSBM.

Algorithm 2 A pseudocode for sampling from gSBM

Require: μ and ν, starting cdf $F^{(0)}$
Calculate $S(\mu)$ and $S(\nu)$
Apply $\mathcal{A}$ starting with $F^{(0)}$ until converged leading α as starting cdf of SBM
Calculate OT map T_1 from $\alpha * \gamma_1$ to $S(\nu)$
Calculate maps for intermediate times $T_t := T_1 * \gamma_{1-t},\ t \in [0,1]$
Sample Brownian paths $(B_t)_t$
Transform Brownian paths to SBM $(X_t)_t = (T_t(B_t))_t$
Obtain samples of gSBM $(Y_t)_t$ from samples of SBM $(X_t)_t$ via acceptance-rejection sampling

2.3 Proofs

Before giving the proof of Theorem 2.1 we need a couple of preparations. We start by relating the costs in (2.3) and (2.4).

Lemma 2.2. *We have*

$$\sup\left\{\mathbb{E}\Big[\int_0^1 \sigma_t\,dt\Big] : Y_0 = b(\nu), Y_1 \sim \nu, dY_t = Y_t\sigma_t\,dB_t\right\} = b(\nu)\mathrm{MCov}(S(\nu),\gamma_1). \tag{2.8}$$

Proof. We will use a classical change of measure argument: We fix a stochastic basis $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_t)$ supporting a $\mathbb{P}$-Brownian motion B. In the first step, consider a positive $\mathbb{P}$-martingale X starting in $x_0 > 0$ with $dX_t = X_t\sigma_t\,dB_t$ and $\sigma = (\sigma_t)_t$ is $(\mathcal{F}_t)_t$-adapted. Since $x_0 = X_0 = \mathbb{E}_\mathbb{P}[X_1]$, the measure $\hat{\mathbb{P}}$ whose density is given by $\frac{d\hat{\mathbb{P}}}{d\mathbb{P}} = \frac{X_1}{x_0}$, is a probability measure equivalent to $\mathbb{P}$. By Girsanov's theorem [23, Theorem 16.19] the process $\hat{B} = (\hat{B}_t)_t$ defined by

$$\hat{B}_t := \int_0^t \frac{1}{X_s}\,d[X,B]_s - B_t = \int_0^t \frac{X_s\sigma_s}{X_s}\,ds - B_t = \int_0^t \sigma_s\,ds - B_t,$$

is a local $\hat{\mathbb{P}}$-martingale with $[\hat{B}]_t = t$, thus, a $\hat{\mathbb{P}}$-Brownian motion. As X is strictly positive, we can define $Y = (Y_t)_t$ by $Y_t = 1/X_t$, from where we deduce by Itô's formula the dynamics

$$dY_t = \frac{\sigma_t^2 X_t^2}{X_t^3}\,dt - \frac{\sigma_t X_t}{X_t^2}\,dB_t = \sigma_t Y_t(\sigma_t dt - dB_t) = \sigma_t Y_t\,d\hat{B}_t.$$

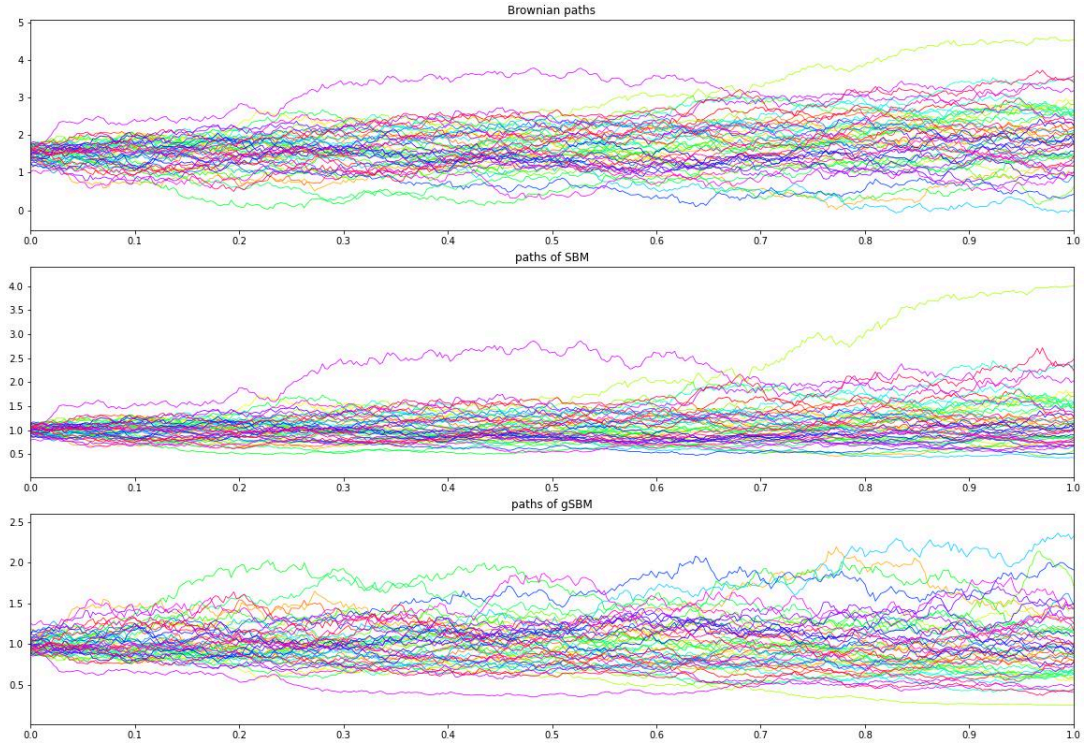

Figure 4: 50 paths of BM between α and $\alpha * \gamma_1$, of SBM between $S(\mu)$ and $S(\nu)$, as well of gSBM between μ and ν.

We observe that $\mathbb{E}_{\hat{\mathbb{P}}}[Y_1] = \frac{\mathbb{E}_{\mathbb{P}}[X_1 Y_1]}{x_0} = \frac{1}{x_0}$ and $\mathbb{E}_{\hat{\mathbb{P}}}[Y_1|\mathcal{F}_t] = \frac{\mathbb{E}_{\mathbb{P}}[X_1 Y_1|\mathcal{F}_t]}{\mathbb{E}_{\mathbb{P}}[X_1|\mathcal{F}_t]} = Y_t$, $\hat{\mathbb{P}}$-a.s. for $t \in [0,1]$. In particular, Y is a $\hat{\mathbb{P}}$-martingale. Therefore, we have

$$x_0 \mathbb{E}_{\hat{\mathbb{P}}}\Big[\int_0^1 \sigma_t \, dt\Big] = \mathbb{E}_{\mathbb{P}}\Big[X_1 \int_0^1 \sigma_t \, dt\Big] = \mathbb{E}_{\mathbb{P}}\Big[\int_0^1 X_t \sigma_t \, dt\Big] = \mathbb{E}_{\mathbb{P}}[X_1 B_1], \tag{2.9}$$

where we used first the tower property and that both X and B are $\mathbb{P}$-martingales.

Next, assume that additionally $x_0 = b(S(\nu)) = 1/b(\nu)$ and $X_1 \sim S(\nu)$ under $\mathbb{P}$. By Theorem 1.1 (P3) there is some $\pi \in \mathscr{M}_{[0,1]}(\delta_{1/x_0}, \nu)$ such that $X \sim S(\pi)$ under $\mathbb{P}$. By construction, we find that $Y \sim S(S(\pi)) = \pi$ under $\hat{\mathbb{P}}$.

Due to the preceding observation, we directly derive (2.8) from (2.9). □

Proof of Theorem 2.1. First we fix $\pi \in \mathscr{M}_{\{0,1\}}(\mu, \nu)$. Due to stability of optimal transport maps, it is well-known that the map $\mathcal{P}_2(\mathbb{R}) \ni \eta \mapsto Z^\eta \in L_2(\gamma_1)$ where $Z^\eta \sim \eta$ and $\mathbb{E}_{B_1 \sim \gamma_1}[Z^\eta B_1] = \mathrm{MCov}(\eta, \gamma_1)$, is $(\mathcal{W}_2, L_2)$-continuous where $\mathcal{W}_2$ denotes the Wasserstein-2 distance on $\mathcal{P}_2(\mathbb{R})$. Consider the probability space $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \mu \otimes \gamma_1)$ with the random variables $X_0(\omega_1, \omega_2) = \omega_1$ and $B_1(\omega_1, \omega_2) = \omega_2$ for $\omega = (\omega_1, \omega_2) \in \mathbb{R}^2$. It easily follows that there exists a random variable $X_1 \in L_2(\mu \otimes \gamma_1)$ where a.s. $\mathrm{law}(X_1|X_0) = \pi_{X_0}$ and

$$\mathbb{E}[X_1 B_1 | X_0] = \mathrm{MCov}(\pi_{X_0}, \gamma_1),$$

whence,

$$\mathbb{E}[X_1 B_1] = \int \mathrm{MCov}(\pi_{x_0}, \gamma_1) \, \mu(dx_0). \tag{2.10}$$

Using Lemma 2.2 and (2.10), we compute

$$\begin{aligned} V^{\mathrm{gSBM}}(\mu, \nu) &= \sup_{\pi \in \mathscr{M}_{\{0,1\}}(\mu,\nu)} \int x_0 \mathrm{MCov}(S(\pi_{x_0}), \gamma_1) \, \mu(dx_0) \qquad (2.11) \\ &= \sup_{\tilde{\pi} \in \mathscr{M}_{\{0,1\}}(S(\mu), S(\nu))} b(\mu) \int \mathrm{MCov}(\tilde{\pi}_{y_0}, \gamma_1) \, S(\mu)(dy_0) = b(\mu) V^{\mathrm{SBM}}(S(\mu), S(\nu)), \end{aligned}$$

where the second-to-last equality is due to Theorem 1.1 (P4) and Remark 1.3, and identifying $\tilde{\pi} = S(\pi)$. This yields the first assertion.

Moreover, by the same reasoning it follows that a martingale X (with $dX_t = X_t \sigma_t \, dB_t$ and law $\mathbb{Q}$) minimizes (2.3) if and only if $\pi^* := \mathrm{Law}(X_0, X_1) \in \mathscr{M}_{\{0,1\}}(\mu, \nu)$ is optimal for the right-hand side in (2.11) and $\mathbb{E}[\int_0^1 \sigma_t \, dt | X_0] = X_0 \mathrm{MCov}(S(\pi_{X_0}), \gamma_1)$ a.s. Thus, invoking Theorem 1.1, Theorem 1.5, and [6, Theorem 1.5] we find $S(\mathbb{Q}) = \mathbb{Q}^{\mathrm{SBM}, S(\mu), S(\nu)}$, which concludes the proof. □

3 The CN-transformation and adapted Wasserstein distance

3.1 (geometric) adapted Wasserstein distance

In order to recall the definition of the adapted Wasserstein distance, let us first define the set of bicausal couplings. For this sake, consider $C[0,1] \times C[0,1]$, and the canonical processes X, X', i.e. for $(\omega, \omega') \in C[0,1] \times C[0,1]$, consider for all $t \in [0,1]$

$$X_t(\omega, \omega') = \omega_t, \quad X'_t(\omega, \omega') = \omega'_t.$$

Given two measures $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{P}(C[0,1])$ we say that a coupling $\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ is causal from $\mathbb{Q}_1$ to $\mathbb{Q}_2$, if for $(X, X') \sim \pi$, it holds for all $t \in [0,1]$

$$\text{law}((X'_s)_{s\in[0,t]} \mid (X_s)_{s\in[0,1]}) = \text{law}((X'_s)_{s\in[0,t]} \mid (X_s)_{s\in[0,t]}). \tag{3.1}$$

This says that under π the conditional law of X' up to some time t given X can only depend on the past of X, not on the future. A coupling $\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ is then defined to be bicausal between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ if it is both causal from $\mathbb{Q}_1$ to $\mathbb{Q}_2$, as well as causal from $\mathbb{Q}_2$ to $\mathbb{Q}_1$. This means that (3.1) as well as for all $t \in [0,1]$

$$\text{law}((X_s)_{s\in[0,t]} \mid (X'_s)_{s\in[0,1]}) = \text{law}((X_s)_{s\in[0,t]} \mid (X'_s)_{s\in[0,t]}) \tag{3.2}$$

hold. The set of all bicausal couplings between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ is then denoted by $\Pi_{\text{bc}}(\mathbb{Q}_1, \mathbb{Q}_2)$. Having recalled the set of bicausal couplings, the adapted Wasserstein distance of martingale measures $\mathbb{Q}_1, \mathbb{Q}_2$ is defined as

$$\mathcal{AW}(\mathbb{Q}_1, \mathbb{Q}_2) := \inf_{\pi\in\Pi_{\text{bc}}(\mathbb{Q}_1,\mathbb{Q}_2)} \left(\int |x_1 - y_1|^2 \, \pi(dx, dy) \right)^{\frac{1}{2}}, \tag{3.3}$$

see [4].

Continuous time adapted Wasserstein distance is used to derive continuity of optimal stopping in [1] and to derive stability properties of pricing, hedging and utility maximization in [4]. Remarkably, it satisfies Talagrand type inequalities with respect to specific entropy, see [16]. There is a rich literature on adapted versions of the Wasserstein distance in discrete time, see e.g. [8, 27] and the references therein.

3.2 gSBM as metric projection

It is known that SBM can be defined as the metric projection of Brownian motion onto $\mathscr{M}^C_{[0,1]}(\mu, \nu)$, the set of continuous martingales starting in μ and terminating in ν, with respect to adapted Wasserstein distance.

Specifically, denoting by $\mathbb{W}$ the Wiener measure on $C[0,1]$, we have

Proposition 3.1. *The stretched Brownian motion $\mathbb{Q}^{\text{SBM},\mu,\nu}$ satisfies*

$$\mathbb{Q}^{\text{SBM},\mu,\nu} = \text{argmin}\{\mathcal{AW}(\mathbb{Q}, \mathbb{W}) : \mathbb{Q} \in \mathscr{M}^C_{[0,1]}(\mu, \nu)\}. \tag{3.4}$$

This is established in [6, Section 6].

We can give an analogous characterization of gSBM upon defining a suitable geometric adapted Wasserstein distance. For a martingale measure $\mathbb{Q}$ on $C([0,1]; \mathbb{R}_{++})$ we denote by $L\mathbb{Q}$ the law of the stochastic logarithm of $X \sim \mathbb{Q}$. We then set for martingale laws $\mathbb{Q}_1, \mathbb{Q}_2$ on $C([0,1]; \mathbb{R}_{++})$

$$g\mathcal{AW}(\mathbb{Q}_1, \mathbb{Q}_2) := \mathcal{AW}(L\mathbb{Q}_1, L\mathbb{Q}_2), \tag{3.5}$$

which is well-defined with values in $[0, \infty]$. We call $g\mathcal{AW}$ the *geometric adapted Wasserstein distance*.

Example 3.2. We calculate the geometric adapted Wasserstein distance between the laws of two geometric Brownian motions, i.e. for $\sigma, \sigma' \geq 0$ we consider

$$dX_t = \sigma X_t dB_t, \quad dX'_t = \sigma' X'_t dB'_t, \quad X_0 = X'_0 = 1,$$

and call their laws $\mathbb{Q}_1$ and $\mathbb{Q}_2$. As $g\mathcal{AW}(\mathbb{Q}_1, \mathbb{Q}_2) = \mathcal{AW}(L\mathbb{Q}_1, L\mathbb{Q}_2)$ note that for

$$dZ_t = \sigma dB_t, \quad dZ'_t = \sigma' dB'_t,$$

we have $Z \sim L\mathbb{Q}_1$ and $Z' \sim L\mathbb{Q}_2$. From this we obtain using [4, Example 3.4]

$$g\mathcal{AW}(\mathbb{Q}_1, \mathbb{Q}_2) = \mathcal{AW}(L\mathbb{Q}_1, L\mathbb{Q}_2) = |\sigma - \sigma'|.$$

That is, the geometric adapted Wasserstein distance of geometric Brownian motions is exactly the difference of the volatilities of the respective log-returns.

In analogy to SBM we define that the geometric stretched Brownian motion is the maximizer of

$$\mathrm{gSBM}_{\mathrm{corr}}(\mu, \nu) := \max\Big\{\mathbb{E}\Big[\int_0^1 \frac{\sigma_t}{X_t}\, dt\Big] : dX_t = \sigma_t dB_t,\ X_0 \sim \mu, X_1 \sim \nu\Big\}. \tag{3.6}$$

Note that we can also write this as in (2.3), i.e.

$$\mathrm{gSBM}_{\mathrm{corr}}(\mu, \nu) = \max\Big\{\mathbb{E}\Big[\int_0^1 \tilde{\sigma}_t\, dt\Big] : dX_t = \tilde{\sigma}_t X_t dB_t,\ X_0 \sim \mu, X_1 \sim \nu\Big\}.$$

Similarly, we will show below that (3.6) is equivalent to the minimization problem

$$\mathrm{gSBM}_{\mathrm{dist}}(\mu, \nu) := \min\Big\{\mathbb{E}\Big[\int_0^1 \Big|\frac{\sigma_t}{X_t} - 1\Big|^2 dt\Big] : dX_t = \sigma_t dB_t,\ X_0 \sim \mu, X_1 \sim \nu\Big\}. \tag{3.7}$$

Lemma 3.3. *The optimization problems* (3.6) *and* (3.7) *are equivalent in the sense that they have the same optimizers and*

$$gSBM_{dist}(\mu, \nu) = 1 - 2gSBM_{corr}(\mu, \nu) - 2\int \log(x_1)\, \nu(dx_1) + 2\int \log(x_0)\, \mu(dx_0).$$

Proof. Let $X = (X_t)_t$ be a $\mathbb{P}$-martingale satisfying the SDE $dX_t = \tilde{\sigma}_t X_t\, dB_t$ where B is a $\mathbb{P}$-Brownian motion. By [23, Theorem 23.8] the stochastic exponential is the unique solution to the aforementioned SDE, hence,

$$X_t = X_0 \exp\Big(\int_0^t \tilde{\sigma}_s\, dB_s - \frac{1}{2}\int_0^t \tilde{\sigma}_s^2\, ds\Big).$$

Consequently, we have

$$\log(X_1) - \log(X_0) = \int_0^1 \tilde{\sigma}_t\, dB_t - \frac{1}{2}\int_0^1 \tilde{\sigma}_t^2\, dt. \tag{3.8}$$

Using (3.8) we find

$$\begin{aligned}\mathbb{E}\Big[\int_0^1 (1-\tilde{\sigma}_t)^2\,dt\Big] &= 1 - 2\mathbb{E}\Big[\int_0^1 \tilde{\sigma}_t\,dt\Big] + \mathbb{E}\Big[\int_0^1 \tilde{\sigma}_t^2\,dt\Big] \\ &= 1 - 2\mathbb{E}\Big[\int_0^1 \tilde{\sigma}_t\,dt\Big] + 2\mathbb{E}[-\log(X_1) + \log(X_0)].\end{aligned}$$

We observe that the last expected value just depends on the initial and terminal distribution of X. Consequently, minimizing (3.7) over all continuous martingales X with $X_0 \sim \mu$ and $X_1 \sim \nu$ is equivalent to maximizing (3.6) over the same class. □

Write $\mathbb{W}_g$ for the law of geometric Brownian motion. Then we have

Theorem 3.4. *Assume that $S(\nu)$ has finite second moment. Then, $\mathbb{Q}^{gSBM,\mu,\nu}$ is the unique minimizer of*

$$\min\{g\mathcal{AW}(\mathbb{Q},\mathbb{W}_g) : \mathbb{Q} \in \mathscr{M}^C_{[0,1]}(\mu,\nu)\}. \tag{3.9}$$

Proof. Let first $dX_t = \sigma_t X_t dB_t$ with $X_0 \sim \mu, X_1 \sim \nu$ and $\sigma_t \geq 0$. Furthermore, denote $\mathbb{Q} = \text{law}(X)$. Note that

$$d\tilde{X}_t = \sigma_t dB_t, \quad \tilde{X}_t := (LX)_t \tag{3.10}$$

with LX denoting the stochastic logarithm of X and also $\text{law}(\tilde{X}) = L\mathbb{Q}$.

Let us now show that

$$\mathbb{E}\Big[\int_0^1 \sigma_t dt\Big] \geq \sup_{\pi \in \Pi_{\text{bc}}(L\mathbb{Q},\mathbb{W})} \mathbb{E}_\pi[\tilde{X}_1 W_1]. \tag{3.11}$$

We obtain this by estimating for a bicausal coupling $\pi \in \Pi_{\text{bc}}(L\mathbb{Q},\mathbb{W})$

$$\begin{aligned}\mathbb{E}_\pi[\tilde{X}_1 W_1] = \mathbb{E}_\pi[\langle \tilde{X}, W\rangle_1] &= \mathbb{E}_\pi\Big[\int_0^1 d\langle \tilde{X}, W\rangle_t\Big] = \mathbb{E}_\pi\Big[\int_0^1 \sigma_t d\langle B, W\rangle_t\Big] \\ &\leq \mathbb{E}_\pi\Big[\Big(\int_0^1 \sigma_t d\langle B\rangle_t \int_0^1 \sigma_t d\langle W\rangle_t\Big)^{1/2}\Big] \\ &= \mathbb{E}_\pi\Big[\int_0^1 \sigma_t dt\Big] = \mathbb{E}[\langle \tilde{X}, B\rangle_1],\end{aligned}$$

where the inequality is due to the Kunita-Watanabe inequality. We note that the last term does not depend on the coupling π anymore but only on the marginal law, i.e. $\text{law}(\tilde{X}) = L\mathbb{Q}$. This shows (3.11).

Recalling the definition of the adapted Wasserstein distance, we obtain

$$\mathcal{AW}(L\mathbb{Q},\mathbb{W})^2 = \inf_{\pi \in \Pi_{\text{bc}}(L\mathbb{Q},\mathbb{W})} \mathbb{E}_\pi[\langle \tilde{X} - W\rangle_1] = \mathbb{E}[\langle \tilde{X}\rangle_1] + 1 - 2\sup_{\pi \in \Pi_{\text{bc}}(L\mathbb{Q},\mathbb{W})} \mathbb{E}_\pi[\langle \tilde{X}, W\rangle_1]$$

Using (3.11) we get

$$g\mathcal{AW}(\mathbb{Q},\mathbb{W}_g)^2 = \mathcal{AW}(L\mathbb{Q},\mathbb{W})^2 \geq -2\mathbb{E}\Big[\int_0^1 \sigma_t dt\Big] + \mathbb{E}[\langle \tilde{X}\rangle_1] + 1. \tag{3.12}$$

By [6, Proposition 6.4] we have

$$\inf_{\tilde{\mathbb{Q}}\in\mathscr{M}^C_{[0,1]}(S(\mu),S(\nu))} \mathcal{AW}(\tilde{\mathbb{Q}},\mathbb{W})^2 = - \max_{\substack{d\tilde{X}_t=\sigma_t dB_t \\ \tilde{X}_0\sim S(\mu),\tilde{X}_1\sim S(\nu)}} 2\mathbb{E}\Big[\int_0^1 \sigma_t dt\Big] + \mathbb{E}[\langle \tilde{X}\rangle_1] + 1, \tag{3.13}$$

with attainment, i.e. there is an optimizer. We can now use the bijection from Theorem 1.1, (3.12) and (3.11) to arrive at

$$\inf_{\mathbb{Q}\in\mathscr{M}^C_{[0,1]}(\mu,\nu)} g\mathcal{AW}(\mathbb{Q},\mathbb{W}_g)^2 = - \max_{\substack{dX_t=\sigma_t X_t dB_t \\ X_0\sim\mu,X_1\sim\nu}} 2\mathbb{E}\Big[\int_0^1 \sigma_t dt\Big] + \mathbb{E}[\langle \tilde{X}\rangle_1] + 1. \tag{3.14}$$

Since, the geometric stretched Brownian motion is defined as the law of the (unique in law) maximizer of

$$\max\Big\{\mathbb{E}\Big[\int_0^1 \sigma_t dt\Big] : dX_t = \sigma_t X_t dB_t,\ X_0\sim\mu,\ X_1\sim\nu\Big\},$$

we get by the preceding observations that it is also the unique minimizer w.r.t. the geometric Wasserstein distance against the law of geometric Brownian motion, concluding the proof. □

4 CN-transformation for continuous-time MOT problems

In this section we investigate the CN-transformation for continuous martingales and linear cost functions, complementing Theorems 1.1 and 1.5. Let $h : [0,1]\times\mathbb{R}_{++}\times\mathbb{R}_+ \to \mathbb{R}$ be a measurable function. We consider the minimization problem

$$\inf \mathbb{E}\Big[\int_0^1 h\Big(t, X_t, \frac{\sigma_t}{X_t}\Big)\, dt\Big], \tag{4.1}$$

where the infimum is taken over all continuous martingales X with $X_0\sim\mu, X_1\sim\nu$ and absolutely continuous quadratic variation $d\langle X\rangle_t =: \sigma_t^2\, dt$.

Recall that $\mathscr{M}^{\mathrm{AC}}_{[0,1]}$ is precisely the set of laws of all martingales with absolutely continuous quadratic variation process. By [24] there is a measurable function $\sigma : [0,1]\times C[0,1] \to \mathbb{R}_+$ such that $\int_0^t \sigma_s^2(X)\, ds = \langle X\rangle_t$ π-almost surely, for every $\pi\in\mathscr{M}^{\mathrm{AC}}_{[0,1]}$. Then, (4.1) can be equivalently reformulated using laws

$$\mathcal{W}^h(\mu,\nu) := \inf_{\pi\in\mathscr{M}^{\mathrm{AC}}_{[0,1]}(\mu,\nu)} \int_{C[0,1]} c_h(x)\, \pi(dx), \tag{4.2}$$

where c_h is given by

$$c_h(x) := \int_0^1 h\Big(t, x_t, \frac{\sigma_t(x)}{x_t}\Big)\, dt.$$

In order to better understand what happens with (4.2) under change of numeraire, note that by (1.3) the transformed cost is

$$S^*(c_h)(x) = x_1 \int_0^1 h\Big(t, \frac{1}{x_t}, x_t\sigma_t\Big(\frac{1}{x}\Big)\Big)\, dt.$$

By Itô's formula we have that $\frac{\sigma_t(x)}{x_t} = x_t\sigma_t(1/x)$ π-almost surely for every $\pi \in \mathscr{M}^{\mathrm{AC}}_{[0,1]}$. Let us define

$$s^*(h)\Big(t, x_t, \frac{\sigma_t(x)}{x_t}\Big) := x_t h\Big(t, \frac{1}{x_t}, x_t\sigma_t\Big(\frac{1}{x}\Big)\Big), \tag{4.3}$$

so that we get

Theorem 4.1 (CN-Transformation in continuous time)**.** *For $\pi \in \mathscr{M}^{\mathrm{AC}}_{[0,1]}(\mu,\nu)$ we have*

$$\int_{C[0,1]}\int_0^1 s^*(h)\Big(t, x_t, \frac{\sigma_t(x)}{x_t}\Big)\,dt\,\pi(dx) = b(\mu)\int_{C[0,1]}\int_0^1 h\Big(t, y_t, \frac{\sigma_t(y)}{y_t}\Big)\,dt\,S(\pi)(dy), \tag{4.4}$$

given that one of the sides is well-defined. In particular, $\pi^* \in \mathscr{M}^{\mathrm{AC}}_{[0,1]}(\mu,\nu)$ is optimizer of $\mathcal{W}^{s^*(h)}(\mu,\nu)$ if and only if $S(\pi^*)$ is optimizer of $\mathcal{W}^h(S(\mu), S(\nu))$.

Proof. We use the same notation as in the proof of Theorems 1.1 and 1.5. Let $X \sim \pi \in \mathscr{M}^{\mathrm{AC}}_{[0,1]}(\mu,\nu)$ and $Y = \frac{1}{X}$. By Itô's formula we have (since $\inf_t X_t > 0$, a.s.)

$$dY_t = -\frac{dX_t}{X_t^2} + \frac{d\langle X\rangle_t}{X_t^3}.$$

Thus, the quadratic variation of Y satisfies $d\langle Y\rangle_t = \frac{d\langle X\rangle_t}{X_t^4}$, from where we derive the relation

$$\frac{\sigma_t^2(Y)}{Y_t^2} = \frac{d\langle Y\rangle_t}{Y_t^2} = \frac{d\langle X\rangle_t}{X_t^2} = \frac{\sigma_t^2(X)}{X_t^2}, \tag{4.5}$$

both, $\mathbb{P}$ and $\hat{\mathbb{P}}$-almost surely. Due to symmetry reasons, assume w.l.o.g. that $(t,x) \mapsto s^*(h)(t, x_t, \frac{\sigma_t(x)}{x_t})$ is $dt\otimes\pi$-integrable. With the preceding observations at hand, we obtain

$$\begin{aligned}
\int_{C[0,1]}\int_0^1 s^*(h)\Big(t, x_t, \frac{\sigma_t(x)}{x_t}\Big)\,dt\,\pi(dx) &= \mathbb{E}_{\mathbb{P}}\Big[\int_0^1 s^*(h)\Big(t, X_t, \frac{\sigma_t(X)}{X_t}\Big)\,dt\Big] \\
&= \mathbb{E}_{\mathbb{P}}\Big[\int_0^1 X_t h\Big(t, Y_t, \frac{\sigma_t(Y)}{Y_t}\Big)\,dt\Big] \\
&= \mathbb{E}_{\mathbb{P}}\Big[X_1\int_0^1 h\Big(t, Y_t, \frac{\sigma_t(Y)}{Y_t}\Big)\,dt\Big] \\
&= \mathbb{E}_{\hat{\mathbb{P}}}\Big[\int_0^1 h\Big(t, Y_t, \frac{\sigma_t(Y)}{Y_t}\Big)\,dt\Big]\mathbb{E}_{\mathbb{P}}[X_1] \\
&= b(\mu)\int_{C[0,1]}\int_0^1 h\Big(t, y_t, \frac{\sigma_t(y)}{y_t}\Big)\,dt\,S(\pi)(dy),
\end{aligned}$$

where the first and last equality are due to $X \sim \pi$ under $\mathbb{P}$ and $Y \sim S(\pi)$ under $\hat{\mathbb{P}}$. The second equality follows from (4.5) and the third is due to X being a martingale under $\mathbb{P}$. This concludes the proof. □

Recall the CN-transformation of the cost h as in (4.3). We close this section by highlighting two specific types of cost functions that go well together with the CN-transformation:

(T1) Assume that $h(t,x,\sigma) = \tilde{h}(t,x)\sigma$ and $\tilde{h}\colon [0,1]\times\mathbb{R}_{++}\to\mathbb{R}_{++}$ is Borel and bounded.

(T2) Assume that $h(t,x,\sigma) = xh(t,1/x,x^2\sigma)$ and h is bounded from below.

Indeed, for h satisfying (T1) we have $s^*(h)(t,x,\frac{\sigma}{x}) = \tilde{h}(t,\frac{1}{x})x^2\sigma = h(t,\frac{1}{x},x^2\sigma)$ and note that this type encompasses as a special case ($\tilde{h}\equiv -1$) the geometric stretched Brownian motion from Section 2.

Corollary 4.2. *Assume that h satisfies (T1). Then*

$$\inf_{\pi\in\mathscr{M}^{\mathrm{AC}}_{[0,1]}(\mu,\nu)}\int_{C[0,1]}\int_0^1 h\Big(t,x_t,\frac{\sigma_t(x)}{x_t}\Big)\,dt\,\pi(dx) \\ = b(\mu)\inf_{\hat{\pi}\in\mathscr{M}^{\mathrm{AC}}_{[0,1]}(S(\mu),S(\nu))}\int_{C[0,1]}\int_0^1 h(t,y_t,\sigma_t(y))\,dt\,\hat{\pi}(dy). \quad (4.6)$$

In particular, $\pi^*\in\mathscr{M}^{\mathrm{AC}}_{[0,1]}(\mu,\nu)$ minimizes the right-hand side of (4.6) *if and only if $S(\pi^*)$ minimizes the left-hand side of* (4.6).

Similarly, if h satisfies (T2) the optimization problems $\mathcal{W}^h$ and $\mathcal{W}^{s^*(h)}$ coincide, and we have $s^*(h)(t,x,\frac{\sigma}{x}) = h(t,x,\frac{\sigma}{x})$. Again, as a corollary of Theorems 1.1 and 1.5 we find

Corollary 4.3. *Assume that h satisifies (T2). Then*

$$\mathcal{W}^h(S(\mu),S(\nu)) = \mathcal{W}^h(\mu,\nu). \quad (4.7)$$

In particular, $\pi^*\in\mathscr{M}^{\mathrm{AC}}_{[0,1]}(\mu,\nu)$ is optimal for the left-hand side of (4.7) *if and only if $S(\pi^*)$ is optimal for the right-hand side of* (4.7).

5 Shadow couplings

Shadow couplings were introduced in [12] and represent a class of martingale couplings between given marginals μ,ν. These are parametrized through a so-called *source* $\hat{\mu}\in\mathcal{P}_1(\mathbb{R}_{++}\times[0,1])$ whose first marginal is μ and whose second marginal has no atoms. We write

$$\hat{\mathscr{M}}_{\{0\}} := \mathcal{P}_1(\mathbb{R}_{++}\times[0,1]) \quad\text{and}\quad \hat{\mathscr{M}}_{\{0,1\}} := \Big\{\pi\in\mathcal{P}_1(\mathbb{R}^2_{++}\times[0,1]) : b(\pi_{(x_0,u)}) = x_0,\ \pi\text{-a.s.}\Big\}.$$

The source measures serve as a parametrization of the set of shadow couplings, which we will now introduce. We define the set of lifted martingale transport plans

$$\hat{\mathscr{M}}_{\{0,1\}}(\hat{\mu},\nu) := \Big\{\pi\in\hat{\mathscr{M}}_{\{0,1\}} : ((x,u)\mapsto(x_0,u))_{\#}\pi = \hat{\mu}, ((x,u)\mapsto x_1)_{\#}\pi = \nu\Big\}.$$

Associated to every source $\hat{\mu}$, we can define a weak martingale transport problem

$$V^{\hat{\mu}} := \inf_{\pi\in\mathscr{M}_{\{0,1\}}(\mu,\nu)}\int C^{\hat{\mu}}(\pi_{x_0})\,\mu(dx_0), \quad (5.1)$$

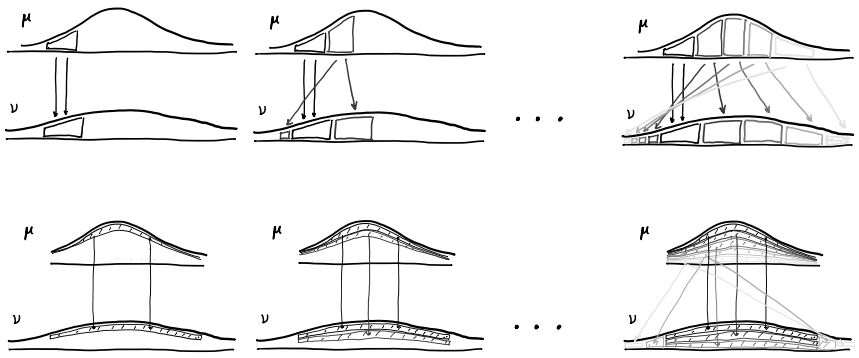

Figure 5: Depiction of the left-monotone martingale coupling and the sunset coupling, see [12].

where the cost function is given by

$$C^{\hat{\mu}}(\eta) := \inf_{\alpha \in \hat{\mathscr{M}}_{\{0,1\}}(\delta_m \otimes \hat{\mu}_m, \eta)} \int (1-u)\sqrt{1+x_1^2}\,\alpha(dx,du) \quad \text{and} \quad m := b(\eta). \tag{5.2}$$

Then the *shadow coupling* $\pi^{\hat{\mu}}$ between μ and ν with source $\hat{\mu}$ is the unique optimizer to (5.1), cf. [12, Proposition 6.2].

Remark 5.1. We collect the following facts on shadow couplings:

1. The function $(1-u)\sqrt{1+y^2}$ could be replaced by an arbitrary (sufficiently integrable) function $\phi(u)\psi(y)$, where ϕ is strictly decreasing and ψ is strictly convex, without changing the definition of the shadow $\pi^{\hat{\mu}}$.

2. Let $R : \mathbb{R}_{++} \times [0,1] \to \mathbb{R}_{++} \times [0,1] : (x,u) \mapsto (x, r(u))$, where $r : [0,1] \to [0,1]$ is increasing and such that $R_\#(\hat{\mu})$ has no atom. Then $\pi^{\hat{\mu}} = \pi^{R_\#(\hat{\mu})}$.

   In particular we would not lose generality by assuming that the second marginal of $\hat{\mu}$ is the Lebesgue measure. However the present choice is more convenient for the subsequent arguments.

3. If $\hat{\mu}$ is the monotone coupling of its marginals, $\pi^{\hat{\mu}}$ is the *left-monotone coupling* which is also studied in [10, 11, 19, 20]. Likewise if $\hat{\mu}$ is the comonotone coupling of its marginals, $\pi^{\hat{\mu}}$ is the *right-monotone coupling* (the mirrored version of the left-monotone coupling).

   If $\hat{\mu}$ is the product coupling of its marginals, then $\pi^{\hat{\mu}}$ is the so called *sunset coupling*.

In order to also cover this extended martingale transport setting, we define a transformation analogous to Section 1.1. Let $I = \{0,1\}$ or $I = \{0\}$, then we define *extended* CN-transformation

$$\hat{S}(\pi) : \hat{\mathscr{M}}_I \to \hat{\mathscr{M}}_I, \quad \hat{S}(\pi)(B) := \frac{\int x_T \delta_{(1/x,u)}(B)\,\pi(dx,du)}{\int x_T\,\pi(dx)}, \tag{5.3}$$

for Borel $B \subseteq \mathbb{R}_{++}^{|I|} \times [0,1]$.

In probabilistic terms $\hat{\mathscr{M}}_{\{0,1\}}$ consists of all $\pi = \mathrm{law}_{\mathbb{P}}(X_0, U, X_1)$ where U has a continuous distribution and $\mathbb{E}_{\mathbb{P}}[X_1|X_0, U] = X_0$. Writing $\hat{\mathbb{P}} = \frac{X_1}{\mathbb{E}X_1}\mathbb{P}$ we then have

$$\hat{S}(\pi)\mathrm{law}_{\hat{\mathbb{P}}} = \mathrm{law}_{\hat{\mathbb{P}}}(1/X_0, U, 1/X_1). \tag{5.4}$$

This allows us to state our main result concerning shadow couplings that is

Theorem 5.2. *Let $\pi^{\hat{\mu}} \in \mathscr{M}_{\{0,1\}}(\mu,\nu)$ be the shadow coupling with source $\hat{\mu}$. Then $S(\pi^{\hat{\mu}}) \in \mathscr{M}_{\{0,1\}}(S(\mu), S(\nu))$ is the shadow coupling with source $\hat{S}(\hat{\mu})$.*

As a particular consequence of Theorem 5.2 (and the definition of $\hat{S}(\hat{\mu})$) we obtain that the CN-transformation of the left-monotone martingale coupling is the right-monotone martingale coupling and vice versa. The CN-transformation of the sunset coupling is again the sunset coupling.

We need some preparations before giving the proof of Theorem 5.2. The extended CN-transformation admits the following counterpart to Theorem 1.1.

Proposition 5.3. *Let $I = \{0\}$ or $\{0,1\}$. The extended CN-transformation satisfies:*

(i) $\hat{S}$ is an involution, i.e.

$$\hat{S}(\hat{S}(\pi)) = \pi \quad \text{for } \pi \in \hat{\mathscr{M}}_I. \tag{$\hat{\mathrm{P}}$1}$$

(ii) The map

$$\hat{S} : \hat{\mathscr{M}}_{\{0,1\}}(\hat{\mu}, \nu) \to \hat{\mathscr{M}}_{\{0,1\}}(\hat{S}(\hat{\mu}), S(\nu)) \tag{$\hat{\mathrm{P}}$2}$$

is a bijection whenever $\hat{\mu} \in \mathcal{P}_1(\mathbb{R}_{++} \times [0,1])$ and $\nu \in \mathcal{P}_1(\mathbb{R}_{++})$.

(iii) For every $\pi \in \hat{\mathscr{M}}_{\{0,1\}}$ with $((x,u) \mapsto (x_0,u))_{\#}\pi = \hat{\mu}$, the disintegrations satisfy

$$\hat{S}(\pi)_{(y_0,u)} = S(\pi_{(x_0,u)}) \quad \text{and} \quad \hat{S}(\hat{\mu})_{y_0} = \hat{\mu}_{x_0}, \tag{$\hat{\mathrm{P}}$3}$$

for $x_0 = 1/y_0$ and $S(\pi)$-a.e. (y,u).

Proof. Using (5.4), the proof is completely analogous to Theorem 1.1. □

Lemma 5.4. *Let $\hat{\mu} \in \mathcal{P}_1(\mathbb{R}_{++} \times [0,1])$ with first marginal μ. Then we have for μ-a.e. x_0*

$$S^*(C^{\hat{\mu}})(\eta) = C^{\hat{S}(\hat{\mu})}(\eta) \quad \text{for every } \eta \in \mathcal{P}_1(\mathbb{R}_{++}) \text{ with } b(\eta) = x_0. \tag{5.5}$$

Proof. Observe that $f(x) := \sqrt{1+x^2}$ satisfies the functional equation $f(x) = xf(1/x)$. Let us next set $m := b(\eta)$. This means that for $\alpha \in \hat{\mathscr{M}}_{\{0,1\}}(\delta_m \otimes \hat{S}(\hat{\mu})_m, \eta)$ we have

$$m\int (1-u)f(y_1)\,\hat{S}(\alpha)(dy,du) = \int (1-u)x_1 f(1/x_1)\,\alpha(dx,du) = \int (1-u)f(x_1)\,\alpha(dx,du). \tag{5.6}$$

Further, it follows from ($\hat{\mathrm{P}}$2) and ($\hat{\mathrm{P}}$3) that for μ-a.e. $x_0 = 1/y_0$, $\delta_{y_0} \otimes \hat{\mu}_{x_0} = \delta_{y_0} \otimes \hat{S}(\hat{\mu})_{y_0}$ and

$$\hat{S} : \hat{\mathscr{M}}_{\{0,1\}}(\delta_{x_0} \otimes \hat{\mu}_{x_0}, \eta) \to \hat{\mathscr{M}}_{\{0,1\}}(\delta_{y_0} \otimes \hat{S}(\hat{\mu})_{y_0}, S(\eta)), \tag{5.7}$$

is bijective. Combining these two observations yields (for μ-a.e. $x_0 = 1/y_0$ and every $\eta \in \mathcal{P}_1(\mathbb{R}_{++})$ with $b(\eta) = x_0$)

$$\begin{aligned} S^*(C^{\hat{\mu}})(S(\eta)) &= x_0 C^{\hat{\mu}}(\eta) \\ &= x_0 \inf_{\alpha \in \hat{\mathscr{M}}_{\{0,1\}}(\delta_{x_0} \otimes \hat{\mu}_{x_0}, \eta)} \int (1-u) f(x_1)\, \alpha(dx, du) \\ &= x_0 \inf_{\alpha \in \hat{\mathscr{M}}_{\{0,1\}}(\delta_{y_0} \otimes \hat{S}(\hat{\mu})_{y_0}, S(\eta))} \int (1-u) f(y_1)\, \hat{S}(\alpha)(dy, du) \\ &= \inf_{\alpha \in \hat{\mathscr{M}}_{\{0,1\}}(\delta_{y_0} \otimes \hat{S}(\hat{\mu})_{y_0}, S(\eta))} \int (1-u) f(x_1)\, \alpha(dx, du) = C^{\hat{S}(\hat{\mu})}(S(\eta)), \end{aligned}$$

where the first equality is due to how S^* acts on weak transport cost functions, the third stems from (5.7), and the fourth follows from (5.6). □

Proof of Theorem 5.2. By Theorem 1.5, when $\pi^{\hat{\mu}}$ is minimizer of (5.1) then $S(\pi^{\hat{\mu}})$ is minimizer of

$$\inf_{\pi \in \mathscr{M}_{\{0,1\}}(S(\mu), S(\nu))} \int S^*(C^{\hat{\mu}})(\pi_{x_0})\, \mu(dx_0),$$

whence, by Lemma 5.4 this problem is equivalent to solving (5.1) between $S(\mu)$ and $S(\nu)$ with source $\hat{S}(\hat{\mu})$. This shows the assertion. □

Acknowledgements

M. Beiglböck and L. Riess gratefully acknowledge financial support through the FWF-projects Y0782 and P35197.

References

[1] B. Acciaio, J. Backhoff-Veraguas, and A. Zalashko. Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *Stoch. Proc. Appl.* 130.5 (2020), pp. 2918–2953.

[2] B. Acciaio, A. Marini, and G. Pammer. Calibration of the Bass Local Volatility model. *ArXiv e-prints* 2311.14567 (2023).

[3] J. Backhoff-Veraguas and G. Pammer. Applications of weak transport theory. *Bernoulli* 28.1 (2022), pp. 370–394.

[4] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance Stoch.* 24.3 (2020), pp. 601–632.

[5] J. Backhoff-Veraguas, M. Beiglböck, W. Schachermayer, and B. Tschiderer. The structure of martingale Benamou–Brenier in multiple dimensions. *ArXiv e-prints* 2306.11019 (2023).

[6] J. Backhoff-Veraguas, M. Beiglböck, M. Huesmann, and S. Källblad. Martingale Benamou-Brenier: a probabilistic perspective. *Ann. Probab.* 48.5 (2020), pp. 2258–2289.

[7] J. Backhoff-Veraguas, G. Loeper, and J. Obloj. Geometric Martingale Benamou-Brenier transport and geometric Bass martingales. *ArXiv e-prints* 2406.04016 (2024).

[8] D. Bartl, M. Beiglböck, and G. Pammer. The Wasserstein space of stochastic processes. *J. Eur. Math. Soc.* (2024). To appear. arXiv:2104.14245.

[9] M. Beiglböck, B. Jourdain, W. Margheriti, and G. Pammer. Monotonicity and stability of the weak martingale optimal transport problem. *Annals of Applied Probability, to appear* (2024).

[10] M. Beiglböck and N. Juillet. On a problem of optimal transport under marginal martingale constraints. *Ann. Probab.* 44.1 (2016), pp. 42–106.

[11] M. Beiglböck, P. Henry-Labordère, and N. Touzi. Monotone martingale transport plans and Skorokhod embedding. *Stochastic Process. Appl.* 127.9 (2017), pp. 3005–3013.

[12] M. Beiglböck and N. Juillet. Shadow couplings. *Trans. Amer. Math. Soc.* 374.7 (2021), pp. 4973–5002.

[13] M. Beiglböck, M. Nutz, and N. Touzi. Complete duality for martingale optimal transport on the line. *Ann. Probab.* 45.5 (2017), pp. 3038–3074.

[14] L. Campi, I. Laachir, and C. Martini. Change of numeraire in the two-marginals martingale transport problem. *Finance Stoch.* 21.2 (June 2017), pp. 471–486.

[15] A. Conze and P. Henry-Labordere. Bass Construction with Multi-Marginals: Lightspeed Computation in a New Local Volatility Model. en. *SSRN Electronic Journal* (2021).

[16] H. Föllmer. Optimal Couplings on Wiener Space and An Extension of Talagrand's Transport Inequality. *Stochastic Analysis, Filtering, and Stochastic Optimization: A Commemorative Volume to Honor Mark H. A. Davis's Contributions*. Cham: Springer International Publishing, 2022, pp. 147–175. ISBN: 978-3-030-98519-6.

[17] N. Gozlan, C. Roberto, P.-M. Samson, Y. Shu, and P. Tetali. Characterization of a class of weak transport-entropy inequalities on the line. Preprint arXiv:1509.04202v2. 2015.

[18] I. Guo, G. Loeper, and S. Wang. Local volatility calibration by optimal transport. *2017 MATRIX Annals*. Springer, 2019, pp. 51–64.

[19] P. Henry-Labordère and N. Touzi. An explicit martingale version of the one-dimensional Brenier theorem. *Finance Stoch.* 20.3 (2016), pp. 635–668.

[20] D. G. Hobson, D. Norgilas, et al. The left-curtain martingale coupling in the presence of atoms. *The Annals of Applied Probability* 29.3 (2019), pp. 1904–1928.

[21] M. Huesmann and D. Trevisan. A Benamou-Brenier formulation of martingale optimal transport. *Bernoulli* 25.4A (2019), pp. 2729–2757.

[22] B. Joseph, G. Loeper, and J. Obloj. The Measure Preserving Martingale Sinkhorn Algorithm. *arXiv e-prints* (2023).

[23] O. Kallenberg. Foundations of Modern Probability. Second. Probability and its Applications. New York: Springer-Verlag, 2002, pp. xx+638. ISBN: 0-387-95313-2.

[24] R. L. Karandikar. On the quadratic variation process of a continuous Martingale. *Illinois Journal of Mathematics* 27.2 (1983), pp. 178–181.

[25] G. Loeper. Black and Scholes, Legendre and Sinkhorn. Talk in "Vienna Seminar in mathematical finance and probability". Nov. 2023.

[26] G. Loeper. Option pricing with linear market impact and nonlinear Black-Scholes equations. *Ann. Appl. Probab.* 28.5 (2018), pp. 2664–2726.

[27] G. C. Pflug and A. Pichler. A distance for multistage stochastic optimization models. *SIAM J. Optim.* 22.1 (2012), pp. 1–23.