GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Contributions to the Theory of Statistical Optimal Transport

Dissertation

for the award of the degree
"Doctor rerum naturalium" (Dr. rer. nat.)
of the Georg-August University Göttingen

within the doctoral program
Mathematical Sciences
of the Georg-August University School of Science (GAUSS)

submitted by
Thomas Staudt
from Creußen

Göttingen, 2022

Thesis Committee

Prof. Dr. Axel Munk
Institute for Mathematical Stochastics, University of Göttingen

Prof. Dr. Dominic Schuhmacher
Institute for Mathematical Stochastics, University of Göttingen

Dr. Yoav Zemel
Statistical Laboratory, University of Cambridge

Examination Board

Reviewer

Prof. Dr. Axel Munk
Institute for Mathematical Stochastics, University of Göttingen

Second reviewer

Prof. Dr. Dominic Schuhmacher
Institute for Mathematical Stochastics, University of Göttingen

Further Members of the Examination Board

Prof. Dr. Stefan Halverscheid
Mathematics Institute, University of Göttingen

Prof. Dr. Dorothea Bahns
Mathematics Institute, University of Göttingen

Prof. Dr. Max Wardetzky
Institute for Numerical and Applied Mathematics, University of Göttingen

Prof. Dr. Ulf Fiebig
Institute for Mathematical Stochastics, University of Göttingen

Date of the Oral Examination

June 08, 2022

Preface

"If you are not confused, you are not thinking."
attributed to Albert Einstein

The trajectory that brought me to this point did certainly not feel linear or easily foreseeable. After graduating from high school, I was quite confident that my future would lie in the world of physics, while mathematics merely seemed like a useful but rather lifeless tool. The decision to pursue a physics bachelor program in Göttingen was then caused by a combination of accidental but fortunate events. In fact, it was nearly independent of the grand legacy that this university sustains.

Since then, the way I view and understand mathematics and physics, or science in general, has fundamentally shifted. I remember days in my first semester when I would be genuinely thrilled – shocked even – if our professor misspelled the tenth or eleventh digit of an important constant, like the speed of light. "Is this type of knowledge not an essential part of being a scientist?," the naive me back then wondered. By now, I certainly would not have noticed as well. Nor would I care too much. After all, being a proper scientist is much better characterized by a mode of thinking than a body of raw knowledge.

In my physics education, acquiring the proper mode of thinking turned out to be a painful exercise at times. The path to insight and understanding was often guided by unspoken or hazy rules and arguments, frequently referred to as "intuition" by teaching staff overwhelmed with sceptical "why?" questions of the students. Indeed, cultivating physical intuition was in many ways tremendously helpful and rewarding. I was repeatedly amazed when intense engagement and grappling with a subject of study would slowly turn perceived arbitrariness, absurdity, and contradiction into something like common sense, opening a door to elegant arguments and efficient reasoning. The confusion experienced in this process, however, often felt profound and discouraging.

I was delighted when I discovered that mathematics can help lift at least some of this confusion. Many rules that seem arbitrary from a naive student's perspective (e.g., why are equations with partial derivatives in classical thermodynamics dependent on which state variables are held constant, while equations with differentials are not?) follow a consistent logic that is revealed as soon as the actual mathematical objects and their relation to the physical theory are understood (manifolds in the case of thermodynamics: partial derivatives are chart-dependent while differentials are not). Since physics lectures are inclined to de-emphasize or hide these objects in favor of putting more focus on the resulting computation rules (the terms “groups” and “representation theory” would not be prominent in my first quantum mechanics lectures, even though their influence lingered everywhere), it became a habit of mine to search for “X for mathematicians” whenever a new topic X arose. Even though I would often be overchallenged by the texts I encountered, I was attracted by the structured deductive reasoning. This was an important trigger for me to more profoundly engage in mathematics.

It was only in the second year of my mathematics master program that I seriously came into contact with mathematical statistics. After earlier encounters with statistics and data science had been very applied (I confess I searched for the term “statistics for mathematicians” more than once), the more rigorous treatment of statistical modeling resonated as a sweet spot between mathematical abstraction, applicability in the sciences, and practical utility. In the time that followed at the Institute of Mathematical Stochastics under the mentorship of Axel Munk – first for my master thesis and then as a PhD student – I worked on various distinct topics and collaborated with fellow mathematicians and also lab scientists.

In a first phase, my research focused on the statistical modeling of super resolution microscopy and fluorescent molecules. Together with Timo Aspelmeier and a group of physicists around Alex Egner at the Institute for Nanophotonics in Göttingen, we established and analyzed a two-timescale hidden Markov model to statistically model the intensity traces that can be derived from data in coordinate-stochastic modes of super-resolution microscopy (like PALM or STORM). Through maximum likelihood inference, our model can be used to estimate the number of fluorescent molecules that contribute to a given trace, making it possible to quantify the molecule density in super-resolution images. One specialty of our method is that it requires little preprocessing of data (compared to other counting techniques), since it incorporates the whole imaging process from photon generation in the fluorescent dye to detection in the camera. Furthermore, the modeling is fine enough to detect when the assumption of statistical independence between fluorescent dyes is violated: experiments showed that our model fitted data notably better for samples where the distance between dyes is large than when they are placed within interaction range. Out of this work, a publication detailing the statistical model has emerged (Staudt et al. 2020), and a further manuscript on experimental results is in preparation. In addition, I helped prepare a chapter on the statistics of nanoscale photonic imaging for the summary book of the SFB 755 (Munk et al. 2020).

My work on counting molecules in microscopy has also served as an entry point into another project, in collaboration with Laura Fee Nern and Johannes Schmidt-Hieber, which investigated the estimation of the binomial parameter n (number of independent Bernoulli trials) if the success probability p of each trial is unknown

(Schmidt-Hieber et al. 2021). We were especially interested in the difficult asymptotic regime where p goes to zero as the number k of independent observations grows to infinity. By minimax estimation theory, we were able to show that k must at least grow with $(n/p)^2$ if n is to be estimated consistently. Furthermore, in several asymptotic settings of interest, the Bayesian estimators we considered could be shown to be consistent if k grows at least slightly (i.e., logarithmically) faster than $(n/p)^2$. Numerical work supporting our theory has been published as well (Schneider et al. 2019).

In parallel to the aforementioned projects, I came into contact with another topic that would heavily influence – and eventually dominate – my research interests: the theory of *optimal transport* and its applications in statistics. It is easy to be intrigued by this intuitive and attractive optimization problem, which starts from a very innocent premise (how to best transform one distribution of mass into another), but features enormous richness from a multitude of perspectives. Besides its fascinating and deep theory (related to aspects of optimization, combinatorics, probability theory, partial differential equations, and differential geometry, besides others), its scope for meaningful application is vast and touches diverse disciplines where transportation and matching are of importance. To name just a few, this includes economic modeling and image analysis, as well as various branches of data science, physics, and biology. In these applications, data is often incomplete and only (random) samples of a larger population are available. Thus, by necessity, the study of *statistical* optimal transport assumes a prominent role.

My work on optimal transport, in collaborations with Shayan Hundrieser, Thomas Giacomo Nies, and Marcel Klatt, has resulted in both novel statistical insights as well as a number of previously uncharted deterministic findings. The latter complement the statistical theory in essential ways, but they also provide mathematical value in their own right. In Hundrieser et al. 2022b, we establish convergence rates if the optimal transport cost between two probability measures is estimated by the empirical measures (the *empirical* optimal transport cost). If the two measures are different, our approach admits the surprising conclusion that the convergence rate is determined by the *simpler* of the two measure (think of a lower intrinsic dimension). This property, which we refer to as *lower complexity adaptation (LCA)*, opens the door to a plethora of subsequent questions concerning statistical optimal transport. For example, our work in Hundrieser et al. 2022a, where a unified approach to distributional limit theorems for the empirical optimal transport cost is established, profits from the LCA observation in requiring that only one of the involved measures has a support with sufficiently low dimension. In this context, we also put forward a precise characterization that clarifies when dual solutions of the optimal transport problem are constant. Together with the uniqueness of said dual solutions, which is the premier object of interest in Staudt et al. 2022, this provides a thorough understanding of situations in which the distributional limits are Gaussian or degenerate. Crucially, we for the first time derive dual uniqueness statements in continuous settings with measures that have disconnected support, considerably broadening the conditions where uniqueness is understood to hold. Finally, we have also studied a statistically attractive setting that benefits from the LCA property due to an inherent asymmetry in the two involved probability measures. In Nies et al. 2021, the statistical dependency between two random variables is quantified by comparing their joint distribution (which may have a low intrinsic dimension) to the product of their marginal distributions (which may have a higher intrinsic dimension) via optimal transport. Among several desirable properties of the resulting dependency measure, which we named the *transport dependency*, a key observation is that it is maximized if and only if the two random variables are related by a Lipschitz function in a deterministic fashion. This structural property distinguishes our concept from other measures of dependency, which are often either maximized for a much smaller class of deterministic relations (like linear ones) or a much larger class (like all measurable ones).

Since this is a cumulative doctoral thesis, my contributions to the topic of statistical optimal transport are chiefly contained in the four research papers referenced in the paragraph above. They are also attached to this document as Contribution A to D, with some minor editorial adaptions. The introductory chapter attempts to provide a general overview over the broader themes and concepts that guide and connect my work, revolving around the struggle to overcome the curse of dimensionality in statistics. It stresses that the four papers A to D are not merely isolated efforts, but that they are part of a larger development in a field with a rich history and a promising outlook for the future. The second chapter is meant to distill and present a more detailed summary of my core contributions. For the convenience of the reader, it starts with a brief introduction to the topic of optimal transport and its dual formulation, which plays a fundamental role for most of my results. Afterwards, several pages are devoted to motivate and spell out the main results that I have been involved in while working on statistical optimal transport. The scope, significance, and limitations of my research are discussed, and leverage points for future work are pointed out. Technicalities and exhaustive citations are consciously de-emphasized in this part of the document, favoring a more digestible and readable exposition. Readers whose interest has been aroused and who intend to dig deeper are invited to find much more details in the dedicated research articles.

Acknowledgements

Working on this PhD project, I experienced the company of people that not only shaped my scientific work, but that also impacted me in broader ways. The most important academic influence has certainly been the mentorship and guidance of my supervisor *Axel Munk*. When I first contacted him for a master thesis project, little did I know about the journey that would ensue, and what I could eventually learn from him. Exploring ideas in endless discussions was as exciting as insightful, often only interrupted by the sudden realization that he is once again violently overstretching his schedule. He conveys a philosophy towards mathematical questions that I grew to be quite fond of, searching for the right composition of purpose, beauty, and practicability. The latter is exemplified by numerous collaborations with experimental research groups, which I thoroughly enjoyed: it forces the abstract to serve the tangible.

My second and third PhD thesis advisors, *Dominic Schuhmacher* and *Yoav Zemel*, supported and encouraged me in various ways. Lectures, questions, comments, and suggestions from their part deepened my understanding of my own research topics and helped me broadened my statistical horizon.

One of the main reasons that make me feel looking back at a successful and satisfying PhD time have been the inspiring collaborations I enjoyed at the IMS in Göttingen. Exchanging arguments and insights with *Giacomo, Shayan, Marcel, Christoph, Laura Fee, Frank,* and *Timo*, not rarely on a daily basis, was greatly stimulating and without a doubt crucial to my research. Other people from the institute, with whom I shared conferences, excursions, game nights, (sometimes odd) discussions, and plenty of mensa time (with even odder discussions still, but also too much tedious gossip about soccer), include *Housen, Miguel, Jonas, Slava, Rupsa, Anne, Markus, Robin, Stefan, Laura, Katharina, Carla, Claudia, Marco, Johannes, Florian, Peter, Benjamin*, and recently *Gilles*. The time here would also have been less interesting and productive without our system administrator in chief and fellow Linux enthusiast *Christian*.

Reaching back further into the past, I also feel genuinely indebted to *Jürgen Vollmer*, who supervised my physics master thesis. It was him who introduced me to the world of conferences, talks, and poster sessions, and he went to great lengths to help me grow as a scientist. In my very first semester in Göttingen, he was the focal point of what became a (very profane) group of friends, which made studying, learning, and overcoming the intimidation caused by the first math lectures that much more fun: *Rolf, Marcel, Sebastian, Philipp, Erik, Theo, Leon, Verena, Moritz,* and *Arthur.*

Outside of the world of science, a most precious companion and confidante that I can always rely on is my dear friend *Maria.* Over many years, she has shared plenty of bright as well as dark moments with me, and would be there whenever I needed support. Besides being a cherished source of delight in my life, she taught me a lot about what truly matters. The time we spent together feels invaluable to me. Words of gratitude also go out to my parents, *Jürgen* and †*Judith.* I more and more come to realize that a lot of what I took for granted throughout my life is the blossom of their support.

Finally, I am aware that progress is always sustained by countless helping hands, some of which are distinctly visible from the beginning, while others remain concealed or barely noted. Without a doubt, the list of people I ought to name is much longer than this text can do justice to. Thanks to all of you.

Contents

1 Introduction

"Nature is pleased with simplicity. And nature is no dummy."

Isaac Newton

It is a common observation that the difficulty of computational and statistical problems can rapidly increase when the dimension, or the number of involved variables, grows. Difficulty, in this context, might refer to the number of computational steps and the memory requirements of an algorithm, or to the number of observations necessary for reliable statistical inference. For example, the number of points necessary to sample the d-dimensional unit cube with a given spatial accuracy increases exponentially with the dimension d. Working in high dimensions, this means that essentially any practical amount of samples or data will be distributed sparsely within the space of interest.

Various statistical and computational techniques that suffer from this and similar effects – apparent victims being numerical integration and function estimation or approximation – have long been investigated. The urgency of this phenomenon, however, became particularly glaring in the computer age, where even an explosive growth of data gathering and computational resources could frequently not offset the difficulties presented by increasingly high dimensional data sets. After all, sampling a 100-dimensional cube on a grid with 10 points in each direction, resulting in 10^{100} grid points in total (by far more than the estimated number of protons in the observable universe), is firmly off limits. In fact, current technology would easily be pushed towards its boundary for values like $d = 15$ already.

To capture the phenomenon that many problems become disproportionally harder as the dimension increases, Bellman 1957 coined the term *curse of dimensionality* in the preface of his textbook on dynamic programming.[1] This was indeed a curse for him, as it imposed major obstacles to the practicability of his theory. In his subsequent book (Bellman 1961), he laments that the curse "casts the pall over our victory celebration" and is "what defeats us at the present time in much of our work." After

[1] Regarding the origin of the name "dynamic programming", Bellman states in his autobiography that "it's impossible to use the word dynamic in a pejorative sense", making dynamic programming "something not even a Congressman could object to", when talking about political factors to justify his research (Bellman 1984). These quotes were found in Eddy 2004.

decades of further research, however, the outlook on the practicability of dynamic programming has turned: techniques like hierarchical partitioning or (stochastic) approximation made it applicable to problems that were out of reach initially (Powell 2011), establishing dynamic programming as an essential tool in computer science and computational biology, where it now powers standard algorithms for gene sequence or protein structure alignment (Waterman 2018).

This type of success story of a methodology that was originally handicapped by the curse of dimensionality is not unique. In fact, concerted efforts of researchers to overcome the curse have lead to progress and even breakthroughs in countless contexts and applications in statistics, data science, and machine learning. A good specimen is the development of the field of *high dimensional statistics*, which was born out of the difficulties and pathologies that classical statistical methods tend to experience in high dimensional settings (Johnstone and Titterington 2009; Wainwright 2019). In classical frameworks, statistical analysis is commonly performed under the assumption of increasing amounts of data in a fixed dimension. Fundamental results that provide excellent approximations for applications where the sample size n is large compared to the dimension d can be much less useful, or even break down entirely, if d is too large compared to n. This already concerns the statistics of linear models, where crucial parts of the theory fail for $d > n$. It also affects areas like regression, matrix and covariance estimation, clustering, or classification, besides many others (Wainwright 2019). Usually, the situation appears to be outright hopeless: within the given models, the curse of dimensionality is inherent and cannot be avoided merely by picking better test statistics or a more sophisticated estimator, say, as minimax theory readily tells us.

What saves statistics from becoming irrelevant in high dimensional settings, then, is the fact that real-world data often obeys an internal logic that makes it considerably simpler than it appears at first. For instance, a data set consisting of megapixel-resolution greyscale images of a certain type of animal could be regarded as a point cloud in the space $[0, 1]^d$ with $d = 10^6$, and one could apply generic statistical methods to analyze data in a million dimensions. At the same time, it is intuitive that the actual space of images depicting the relevant information about the animal is only a tiny subspace: some parameters describing the body form, positioning, and skin texture, and some more values describing the camera configuration and the lighting might actually suffice for most applications. Furthermore, the information to be extracted from the data, like inference of the age or weight of the animal, might not even require high-resolution images. The simple act of subsampling can thus already lead to a significant reduction of the dimensionality without sacrificing expressiveness. Analogous observations apply to high dimensional datasets generated in many real-world applications as well as the sciences. Think of biomedical patient data, genome expression studies, climate modeling, or economic time series data, for example. All of them will – at least to some degree – follow an internal logic that might be surprisingly simple once uncovered.

The realization that even complicated data is usually equipped with a hidden *low dimensional structure* is the recipe for success of high dimensional statistics. From this perspective, the important task is to establish methodologies that either explicitly reconstruct, or at least implicitly exploit, relevant low dimensional characteristics in the data. These characteristics can come in many forms. For example, the idea to exploit (strong) correlations between variables has served as a motivation for shrinkage techniques like ridge regression (Hoerl and Kennard 1970). Relatedly, the assumption of sparsity – assuming that many of the parameters in a statistical model are actually zero, usually without detailed apriori knowledge which entries these are going to be – has been extremely fruitful, both in theory and practice. Methodologies developed in the vast literature on sparse statistical modeling include wavelet thresholding (Donoho et al. 1996), the lasso (Tibshirani 1996; Candes and Tao 2005; Donoho 2006b) in various flavors (Zou 2006; Meier et al. 2008), as well as techniques for variable selection (Candes and Tao 2007; Fan and Lv 2008; Bickel et al. 2009), compression (Donoho 2006a), or clustering (Elhamifar and Vidal 2013). Furthermore, diverse methods of (explicit) dimension reduction search to expose relevant low dimensional structures in data sets. While techniques like projection pursuit (Friedman and Tukey 1974), (sparse) principal component analysis (Johnstone 2001; Zou et al. 2006), or independent component analysis (Comon 1994) can be used to find linear subspaces best explaining certain characteristics of a data set, the idea that high dimensional data is often supported on low dimensional (but possibly non-linear) surfaces underpins ongoing developments in the theory of manifold learning (Hastie and Stuetzle 1989; Goldberg et al. 2008) and other nonlinear dimensionality reduction techniques (Lee and Verleysen 2007). To counteract the curse of dimensionality for regression models in nonparametric statistics, structural assumptions like smoothness or a special (e.g., additive or compositional) form of the regressor function have been salvaged (Stone 1980; Stone 1985; Schmidt-Hieber 2020).

How is all of this related to this thesis? The link is that the topic of *statistical optimal transport* is currently facing similar difficulties to the ones that many areas of statistic have faced in the past: it is severely handicapped by the curse of dimensionality, and its applicability for a variety of high dimensional tasks remains an open question at the present point in time.

Optimal transport has a rich history that dates back to the work of the engineer and mathematician Gaspard Monge, who in 1781 formalized the problem of finding the best way to move soil at a construction site. More than a century later only, the topic would be revisited, most influentially so by Kantorovich 1942. He framed the optimal transport problem in its modern formulation: given two probability measures μ and ν on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, what is the optimal way to *couple* μ and ν such that the resulting transport between them is as *cost-efficient* as possible? The different ways to couple μ and ν are listed in the set $\mathcal{C}(\mu, \nu)$, the set of *couplings*. If a coupling $\pi \in \mathcal{C}(\mu, \nu)$ is chosen, which is understood to be a probability measure on $\mathcal{X} \times \mathcal{Y}$, then the value $\pi(A \times B)$ signifies the amount of mass that is moved between a subset $A \subset \mathcal{X}$ and a subset $B \subset \mathcal{Y}$. Clearly, it has to hold that $\pi(A \times \mathcal{Y}) = \mu(A)$ and $\pi(\mathcal{X} \times B) = \nu(B)$ for this interpretation to work without loss or creation of mass. Which couplings are good ones? To answer this, a cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is introduced, which describes the effort that needs to be undertaken to move points in $\mathcal{X}$ to points in $\mathcal{Y}$ (or vice versa). The total cost for moving all of the mass according to the coupling π is found by integrating c against π, which naturally leads to the optimization task

$$\inf_{\pi \in \mathcal{C}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi(x,y) \tag{1.1}$$

when searching for the best coupling. Optimal couplings that attain the infimum in this problem, if they exist, are called *optimal transport plans*, and the optimal value in (1.1) is called the *optimal transport cost*. It will be denoted by $T_c(\mu, \nu)$ in the following. A more thorough introduction to the topic with more precise definitions is provided in Section 2.1.

The optimization problem (1.1) expresses a deeply fundamental question, and the answers to this question contain a lot of valuable information. Indeed, the optimal transport *plan* tells us which regions of μ are related to which regions of ν under a general notion of how expensive transport is. It even tells us how strong this relation is. In this sense, the optimal plan constitutes a natural matching procedure. The optimal transport *cost*, then, quantifies how different the measures μ and ν are. It provides a way to discriminate between probability measures in general contexts, and in a way that incorporates flexible criteria of similarity. If $\mathcal{X} = \mathcal{Y}$ and $c = d^p$ is the p-th power of a metric d on $\mathcal{X}$ for $p \geq 1$, then the *p-Wasserstein distance* $W_p(\mu, \nu) = T_c(\mu, \nu)^{1/p}$ is in fact a metric on the space of probability measures on $\mathcal{X}$ with finite p-th moment. In respecting the structure of the ground spaces, optimal transport distinguishes itself crucially from information-based notions of discrepancy between probability measures, like the Kullback-Leibler divergence.

The conceptual appeal of optimal transport has not remained without consequences: countless seminal research papers (by Kantorovich, Sudakov, Rachev, Rüschendorf, Brenier, McCann, Ambrosio, or Figalli, to cite just the tip of the iceberg) and thousands of pages of monographs (Rachev and Rüschendorf 1998; Villani 2008; Santambrogio 2015; Peyré and Cuturi 2019; Panaretos and Zemel 2020; Ambrosio et al. 2021) dive into great depth to study the discipline from analytical, probabilistic, geometric, statistical, computational, and other perspectives. Additionally, optimal transport has been the subject of a fields medal, and interest in its application to tackle scientific and engineering problems has rapidly developed over the past decades. Techniques that are based upon or inspired by optimal transport have, for example, been picked up in economics (Galichon 2016), biology (Schiebinger et al. 2019; Tameling et al. 2021), physics (Lohse et al. 2020; Pollard and Windischhofer 2022), computer vision (Ge et al. 2021), machine learning (Courty et al. 2017; Yurochkin et al. 2019), and deep learning (Arjovsky et al. 2017; Tolstikhin et al. 2017). This has contributed to a broadening of the research focus, and statistical issues – how do we best work with optimal transport tools in the presence of finite data, and which kind of guarantees do such techniques enjoy? – are becoming more relevant.

A central question in this context is how well the optimal transport value can be estimated on the basis of data. In the simplest case, we observe n independent and identically distributed (i.i.d.) samples $X_1, \ldots, X_n \sim \mu$ and $Y_1, \ldots, Y_n \sim \nu$ and want to understand how well this data can be used to estimate $T_c(\mu, \nu)$. If we assume no structure at all on μ and ν, then one obvious choice would be to base an estimator on the *empirical measures* $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}$, where δ_x denotes the point measure at x. This leads to estimators $\hat{T}_{c,n}$ of the form $T_c(\hat{\mu}_n, \nu)$, $T_c(\mu, \hat{\nu}_n)$, or $T_c(\hat{\mu}_n, \hat{\nu}_n)$, depending on which of the distributions are assumed to be known beforehand. We refer to the study of these estimators as *empirical optimal transport*.

What do we know about the empirical estimators? On the one hand, we know that they are *consistent*, meaning that $\hat{T}_{c,n}$ approaches $T_c(\mu, \nu)$ almost surely under mild moment and continuity assumptions as n tends towards ∞. On the other hand, we know that this convergence can be painfully slow. Early results by Dudley 1969 show asymptotic proportionality

$$\mathbb{E}[T_c(\hat{\mu}_n, \mu)] \sim n^{-1/d} \tag{1.2}$$

for the Euclidean cost function $c(x, y) = \|x - y\|$ if μ is compactly supported and absolute continuous with respect to the Lebesgue measure in $\mathbb{R}^d$ for $d \geq 3$. Regrettably, this means that the curse of dimensionality strikes with full force: the number of sample points required for reliable approximation of $T_c(\mu, \mu)$, which equals zero in this case, grows exponentially in d. And the situation is even worse. Namely, it is understood that the curse of dimensionality also haunts us if $\mu \neq \nu$, in which case the mean absolute error $\mathbb{E}\big[\big|\hat{T}_{c,n} - T_c(\mu, \nu)\big|\big]$ may only converge with the rate $n^{-1/d}$ in relevant settings as well (Manole and Niles-Weed 2021). Moreover, if no additional structural assumptions are placed on μ, then the rate $n^{-1/d}$ in (1.2) is *minimax optimal* (Singh and Póczos 2018), meaning that *no estimator* $\tilde{\mu}_n$ of μ that is based on the observations $X_1, \ldots, X_n$ can significantly outperform the empirical measure $\hat{\mu}_n$ in general. To add to the problem, optimal transport is accompanied by a high computational burden (Peyré and Cuturi 2019). Even though computational optimal transport is a highly active field of research and improvements are constantly being reported (Schmitzer 2016; Genevay et al. 2016; Jacobs and Léger 2020; Mai et al. 2021), the curse of dimensionality coupled with at least quadratic run times to assign n points to n points continue to plage the estimation of $T_c(\mu, \nu)$ in high-dimensional applications.

Fortunately, there are some signs for hope also. In fact, several advancements have been pushed to curb the curse. A first branch of these advancements exploits smoothness assumptions on the measures μ and ν, like smooth Lebesgue densities. In this case, better estimators $\tilde{\mu}_n$ and $\tilde{\nu}_n$ of μ and ν than the empirical ones are available; for example by means of kernel density estimation or wavelet estimation (Weed and Berthet 2019; Deb et al. 2021; Manole et al. 2021). Under sufficient smoothness guarantees, these estimators do *not* suffer from the curse of dimensionality. In a similar vein, the approach of *smooth optimal transport* enforces smoothness by considering the problem $T_c^\kappa(\mu,\nu) = T_c(\mu{*}\kappa, \nu{*}\kappa)$, where μ and ν may be arbitrary but are both convoluted with a smooth kernel κ, e.g., a Gaussian one (Goldfeld et al. 2020; Goldfeld and Greenewald 2020). Escaping the curse of dimensionality by these means, however, has a high price: the actual computation of these estimators is difficult and tricky, and no reliable and generically applicable implementation has as of yet been established. An alternative approach, which has sparked enormous interest recently, is to consider the optimal transport problem with an additional entropic penalty term on the transport plan (Cuturi 2013). Due to this smoothness-inducing regularization, *entropic optimal transport* does not suffer from the curse of dimensionality (Genevay et al. 2019). Furthermore, for a given regularization parameter, efficient iterative algorithms for its computation are available (Altschuler et al. 2017; Schmitzer 2019). Still, if the regularization is not chosen mildly enough, which can drastically affect numerical performance and stability, entropic regularization leads to blurry transport plans that may not resolve all details of the true optimal transport problem (Feydy et al. 2019).

Within the realm of empirical optimal transport, where we restrict ourselves to the empirical estimators $\hat{\mu}_n$ and $\hat{\nu}_n$ in the vanilla optimal transport problem, understanding has also advanced recently. For example, regularity of the cost function has been shown to improve on the $n^{-1/d}$ rate in (1.2). For $d \geq 5$, Manole and Niles-Weed 2021 established that costs of the form $c(x,y) = \|x-y\|^\alpha$ for $0 < \alpha \leq 2$ (or even more general α-Hölder smooth cost functions) essentially lead to minimax rates $n^{-\alpha/d}$ in relevant settings. In particular, this means that the impact of the dimension d is only halve as bad in case of cost functions that are at least twofold differentiable. Moreover, a vital lesson from the history of high dimensional statistics has been picked up: data in high dimensional spaces is usually equipped with a low-dimensional structure. Advances to exploit this have, for example, been proposed by Weed and Bach 2019, who clarify that the convergence of the Wasserstein distance $W_p(\hat{\mu}_n, \mu)$ for $p \geq 1$ towards zero is in fact governed by an *intrinsic dimension* s of μ, and not by the dimension d of the surrounding ambient space. Put in simplified terms: if μ is concentrated on an s-dimensional surface in $\mathbb{R}^d$, then we can expect convergence rates determined by s instead of d. Thus, empirical optimal transport exhibits the pleasant property of automatically exploiting low dimensional structures in the data, with no explicit adaptions necessary.

This is, finally, where my efforts enter the picture. One of the core realizations captured in Contribution A is that the adaption of empirical optimal transport to the intrinsic dimension of the data is even stronger than previously understood. Indeed, if the two measures μ and ν are different – for example concentrated on an s- and an r-dimensional surface, respectively – then it is the *minimum* of s and r that determines the rate of convergence of $\hat{T}_{c,n}$ to $T_c(\mu,\nu)$. In this sense, the quantity $\hat{T}_{c,n}$ adapts to whichever is the *simpler* measure between μ and ν. We call this fundamental property of empirical optimal transport *lower complexity adaptation*, or short *LCA*.

On the first look, the LCA phenomenon might not exactly seem intuitive. We, at least, were initially puzzled, since apparent arguments to upper bound the deviation of $\hat{T}_{c,n}$ from the true value tend to suggest otherwise. For instance, the triangle inequality of the p-Wasserstein distance implies

$$\left|W_p(\hat{\mu}_n, \hat{\nu}_n) - W_p(\mu, \nu)\right| \leq W_p(\hat{\mu}_n, \mu) + W_p(\hat{\nu}_n, \nu),$$

which leads to an upper bound with a *worse-case rate* if both μ and ν are estimated from data. Unfortunately, our proof strategy does not foster a strong geometric interpretation about why the LCA property is true. It is based on a rather technical observation about the complexity of certain function classes that are important in the dual formulation of optimal transport (see Section 2.1 below and Section 2.2 of Contribution A). Still, some intuition is offered by our work in Section 2.3 of Contribution A. It roughly suggests that finding an optimal assignment between two point clouds, one of which lives on a lower dimensional surface, benefits from the fact that the matching can essentially be performed in two acts: first, finding a suitable "projection" onto the surface, and only then assigning the data points within this lower dimensional space. The projection part, so the reasoning goes, is statistically less costly than the matching part, and thus the smaller dimension determines the statistical behavior. This explanation, however, should probably not be overburdened, since we can rigorously justify it only in case of linear subspaces and orthogonal projections.

Our work on the LCA principle brings forth a considerable list of implications and questions, both of conceptual and of technical nature. Most strikingly, it unlocks the door to a number of statistical techniques that rely on fast convergence. A particular beneficiary is the theory of central limit theorems for empirical optimal transport costs, around which Contribution B is focused. The core achievement of this publication is the derivation of limit laws of the form

$$\sqrt{n}\big(T_c(\hat{\mu}_n, \nu) - T_c(\mu, \nu)\big) \to \sup_{f \in S_c(\mu,\nu)} G(f) \tag{1.3}$$

under a number of technical conditions on the cost function c and the measures μ and ν. Here, G designates a Gaussian process that is indexed by so-called *Kantorovich potentials* $f \in S_c(\mu, \nu)$, which denote the dual solutions of the optimal transport problem $T_c(\mu, \nu)$ (see Section 2.1 below for a definition). Furthermore, if ν is estimated by $\hat{\nu}_n$, or both μ and ν are estimated by their empirical counterparts, similar limit laws can be derived. A crucial requisite for all of these laws is that the class of candidates for Kantorovich potentials (the class of *c-transforms*) behaves nicely, meaning that it has small *uniform covering numbers* (see Section 2.2 below for details). In essence, what this means is that limit laws of the form (1.3) – which were previously only known for discrete or 1-dimensional optimal transport – can hold up to dimension $d = 3$ for suitable (e.g., twice differentiable) cost functions. In fact, as a consequence of our work on the LCA property, it suffices if the *intrinsic* dimension s of *one of the two* measures μ or ν is smaller or equal to 3. This greatly extends the scope of settings for which central limit theorems for the empirical optimal transport cost are know to hold. Section 5 in Contribution B even provides negative results, indicating that limit laws of the form (1.3) can not exist if both of the measures have an intrinsic dimension ≥ 4.

My personal involvement with Contribution B has chiefly been the interpretation of the arising limit laws. In particular, to develop a comprehensive understanding of the right hand side of (1.3), two fundamental questions materialize: first, when are Kantorovich potentials *unique*, implying *Gaussian* limit distributions. Secondly, when are unique Kantorovich potentials additionally *constant*, implying *degenerate* limit laws (since $G(f)$ equals zero for constant potentials f). Both of these questions concern deterministic properties of the dual optimal transport problem, on which the general literature compiled in the past decades is remarkably rich. However, while the question of uniqueness of primal solutions has been targeted via duality theory in many publications, a systematic inquiry into the uniqueness of dual solutions themselves was missing. Synthesizing and considerably generalizing scattered results on aspects of the uniqueness of Kantorovich potentials, Contribution C draws a comprehensive picture which asserts that dual uniqueness holds much more broadly than formerly anticipated. For example, it was commonly accepted that uniqueness of Kantorovich potentials in continuous settings seems to rely on the connectedness of the support of the measures μ and ν, even for differentiable cost functions. Otherwise, trivial counter-examples can be constructed. Our work clarifies that these counter-examples rely on a special distribution of mass, which makes the optimal transport problem break down into several sub-problems between subsets of connected components. Without this specific symmetry in the mass distribution, Kantorovich potentials remain unique. In this sense, non-uniqueness is the *exception* rather than the rule, and we commonly expect Gaussian limits in (1.3), at least for differentiable costs. Degenerate limits, on the other hand, are closer examined in Section 4 of Contribution B. The core result follows a simple geometric logic: constant potentials $f \in S_c(\mu, \nu)$ exist *if and only if* the optimal transport from ν to μ coincides with a *c-projection.* The latter means that when mass is transported between $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ by the optimal transport plan, then $c(x, y) = \inf_{x'} c(x', y)$ has to hold, where the infimum ranges over the support of μ. Therefore, the degeneracy of limit laws of the empirical optimal transport cost necessitates a very special geometric setup. More details on the results mentioned in this paragraph are provided in Section 2.3 below.

By its very nature, the LCA principle plays a particular role when there is an inevitable asymmetry between the probability measures to be compared via optimal transport. Many such setups are conceivable, like the attempt of dimensionality reduction or feature extraction by matching a high dimensional distribution with a low dimensional one. In Contribution D, we consider the problem of measuring dependency, where a similar asymmetry can arise. If two random variables $X \sim \mu$ and $Y \sim \nu$ follow a joint distribution γ, then quantifying their dependency effectively means assigning a number to γ. This number should ideally be high if X and Y are related deterministically, i.e., $Y = f(X)$ for a suitable function f, and low if X and Y are statistically independent. We pursue the strategy to compare γ to the product distribution of μ and ν via defining the *transport dependency*

$$\tau_c(\gamma) = T_c(\gamma, \mu \otimes \nu). \tag{1.4}$$

Since $(X, Y) \sim \mu \otimes \nu$ would correspond to statistical independence of X and Y, this approach quantifies the discrepancy of γ from independence. A core achievement of Contribution D is the precise characterization of distributions γ that maximize τ_c under suitable additive cost functions. Loosely speaking, we show that τ_c is maximized (for given μ and ν) if and only if X and Y are related by Lipschitz functions, whose modulus can be manipulated by adapting the cost function. This results in a principled but flexible dependency measure with many desirable properties. Crucially, lower complexity adaptation applies: it will be the intrinsic dimension of γ and not the (potentially twice as high) dimension of $\mu \otimes \nu$ that dictates the statistical properties of empirical estimators. As discussed in Section 2.4 below, this insight may pave the way to computationally simpler estimators of $\tau_c(\gamma)$ that have the same statistical efficiency as the (computationally very costly) plug-in approach $\tau_c(\hat{\gamma}_n)$.

Of course, our work on the lower complexity adaptation and its consequences are by no means exhaustive or completed. Many questions are left open for future research. On the technical side, for example, the arguments currently rely on bounded costs and compact spaces. There are no fundamental reasons, however, why it should not be possible to extend these statements to unbounded scenarios as long as a proper interplay between the tail of the distributions and the costs is ensured. Furthermore, it is as of yet unclear which notion of intrinsic dimension of a measure is best suited for a general formulation of the LCA property. More importantly, it would be desirable to understand the actual scope of the LCA principle. In particular, what about estimators that are not the empirical ones, or observations that are not i.i.d.? Under which conditions can we also find lower complexity adaption in these cases, extending LCA to a genuine property of *statistical* optimal transport, and not just *empirical* optimal transport? What about smoothness? Is there maybe an LCA principle related to smoothness as well, and smoothness of a single of the two measures μ and ν may be enough to alleviate the curse of dimensionality? Structural insight along these and similar lines will be vital for the development of statistical techniques that explicitly exploit the automatic adaption of optimal transport to low dimensions.

If we direct our eyes beyond the LCA phenomenon, broader questions still linger. After all, even the potentially smaller dimensionality associated to the simpler measure might be too much to handle in real-world applications. So while the work presented in this doctoral thesis attempts to provide a valuable window into the nature of high-dimensional statistical optimal transport, we are, as of yet, only at the beginning of forming a comprehensive understanding. In the future, we might have to pick up further lessons from the study of high dimensional statistics, and, for example, systematically integrate sparsity assumptions directly into the toolbox of optimal transport. First work in this directions has already been pushed (Forrow et al. 2019). Also, it will be crucial to see if the theoretical benefits enjoyed under smoothness assumptions will actually spill over into the practical world, ultimately providing us with computationally feasible, reliable, and scalable estimators. Eventually, we will also have to develop a more systematic understanding of properties that make optimal transport attractive and practically useful even in settings where the application of asymptotic statistical theory appears demanding or is even out of reach. This, for example, concerns independence testing via the transport dependency: even in finite-sample regimes of high bias, where the curse of dimensionality hits hard, we find that there is still *much* information to be gathered from the empirical optimal transport cost (and plan), leading to high discriminative power of permutation tests.

There is no doubt in my mind that statistical optimal transport and associated topics will remain to be exciting and active research disciplines in the coming years and probably decades. There is much left to learn and uncover. And with computational barriers – which are contemporary still a major nuisance – being lifted step by step, the adoption of methodologies based on or inspired by optimal transport will only accelerate. After all, the basic premise of matching two distributions of objects in the best way possible is truly ubiquitous in natural and scientific matters.

2 Contributions to (Statistical) Optimal Transport

The goals of this chapter are twofold. First, to provide a summary representation of the core findings on optimal transport that I have worked on in the course of my PhD time. Second, to clarify my own contributions to the research papers A to D, which are the joint product of multiple authors each.

The first section contains a brief introduction into the topic of optimal transport with a focus on duality, and serves to prepare the notation and concepts referred to by the rest of the chapter. All of the facts presented here are standard in the theory of optimal transport and can be found in the first chapters of the books by Villani 2008 and Santambrogio 2015. The remaining sections form an attempt to lay out the most important insights stemming from my work in a way that is digestible for non-experts. In doing so, I focus on conceptual points and largely refrain from emphasizing technical aspects, for which the corresponding research papers can be consulted.

2.1 Optimal Transport and Duality

Let $\mathcal{X}$ and $\mathcal{Y}$ be separable and completely metrizable topological spaces. These spaces are also called Polish. We equip them with their Borel σ-algebra and denote the set of probability measures (or probability distributions) on them via $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$. For given $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, let

$$\mathcal{C}(\mu, \nu) = \big\{ \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \,|\, \pi(\cdot \times \mathcal{Y}) = \mu, \pi(\mathcal{X} \times \cdot) = \nu \big\}$$

denote the set of *couplings* of μ and ν, which form a convex subset of $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Under the topology of weak convergence of probability measures, $\mathcal{C}(\mu, \nu)$ is compact as a result of Prokhorov's theorem. Couplings are also denoted as *transport plans* since $\pi(A \times B)$ can be interpreted as the amount of mass transported between Borel sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$. In the context of random variables $X \sim \mu$ and $Y \sim \nu$, couplings $\pi \in \mathcal{C}(\mu, \nu)$ arise as *joint distributions* of X and Y if the latter obey $(X, Y) \sim \pi$.

If $\pi \in \mathcal{C}(\mu, \nu)$ is concentrated on the graph of a measurable map $t \colon \mathcal{X} \to \mathcal{Y}$, we call π a *deterministic coupling* and the map t the corresponding *transport map*. In this case, it holds that $t_\#\mu = \nu$ and $(\mathrm{id}, t)_\#\mu = \pi$, where $t_\#\mu(B) = \mu\big(t^{-1}(B)\big)$ for any Borel set $B \subset \mathcal{Y}$ denotes the pushforward measure and $(\mathrm{id}, t)(x) = (x, t(x))$ lifts $x \in \mathcal{X}$ to the graph of t. Analogous language is used for transport maps $t\colon \mathcal{Y} \to \mathcal{X}$ in the reverse direction.

Let $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable map, which we refer to as *(ground) cost function.* Given two marginal distributions μ and ν, the core aspiration of optimal transport theory is to find couplings that minimize the total cost $\pi c = \int c \,\mathrm{d}\pi$ associated to transforming μ into ν under the plan π and cost c. The value attained in this linear optimization problem is called the *optimal transport cost*,

$$T_c(\mu, \nu) = \inf_{\pi \in \mathcal{C}(\mu,\nu)} \pi c, \tag{2.1}$$

and minimizers of this problem are called *optimal transport plans.* For general cost functions, optimal transport plans do not have to exist. However, they do exist for lower semicontinuous c, for which the map $\pi \mapsto \pi c$ is always lower semicontinuous with respect to the weak convergence of probability measures. The existence of *optimal transport maps*, on the other hand, which was the original problem formulated by Monge 1781, is a much more delicate issue and depends on various technical assumptions. Closely related is the question of *unique* optimal transport plans, which is mainly studied in settings in which optimal transport maps exist (see the references in Contribution C for details).

As a linear program, the optimal transport problem (2.1) admits a dual formulation. Strong duality holds under very general conditions, e.g., for measurable non-negative cost functions (Beiglböck and Schachermayer 2011), and is a valuable asset for many analytical insights into optimal transport theory. In its vanilla formulation, the dual optimal transport problem is given by

$$\sup_{f \oplus g \le c} \mu f + \nu g,$$

where $(f \oplus g)(x, y) = f(x) + g(y)$ and the supremum runs over all integrable functions $f \in L^1(\mu)$ and $g \in L^1(\nu)$. An alternative formulation of the dual problem exploits that the function g can be replaced by the largest function f^c that satisfies $f \oplus f^c \le c$. This function is determined by $f^c(y) = \inf_{x \in \mathcal{X}} c(x, y) - f(x)$ and denoted as *c-transform* of f. The dual formulation then reads

$$\sup_{f \in L^1(\mu)} \mu f + \nu f^c. \tag{2.2}$$

This argument can be repeated once more, and f can be replaced by the largest function f^{cc} such that $f^{cc} \oplus f^c \le c$, defined by $f^{cc}(x) = \inf_{y \in \mathcal{Y}} c(x, y) - f^c(y)$. In particular, it is possible to restrict the supremum in (2.2) to the class of *c-concave functions* on $\mathcal{X}$, which are defined as

$$S_c = \big\{ g^c \,\big|\, g\colon \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}, g \neq -\infty \big\}, \tag{2.3}$$

where only the constant function $g = -\infty$ is excluded. Since constant shifts in f do not change the value of $\mu f + \nu f^c$, one can, if desired, restrict the function class in (2.2) even more, e.g., by demanding that $f(x_0) = 0$ for a given anchor point $x_0 \in \mathcal{X}$.

In general, it is not guaranteed that c-concave functions are measurable, which can lead to technical intricacies in the formulas presented above. For continuous cost functions c, however, they are always upper semicontinuous and thus also measurable. Under additional regularity conditions, such as $(\mu \otimes \nu)c < \infty$ for non-negative costs, it is known that c-concave functions that attain the supremum in (2.2) always exist. These dual solutions of the optimal transport problem are called *Kantorovich potentials*, which we denote by $S_c(\mu, \nu)$ for given μ, ν, and c. If π denotes an optimal transport plan, and $\operatorname{supp} \pi \subset \mathcal{X} \times \mathcal{Y}$ is its support, then Kantorovich potentials are exactly the c-concave functions f for which

$$\operatorname{supp} \pi \subset \big\{(x, y) \,|\, f(x) + f^c(y) = c(x, y)\big\} \subset \mathcal{X} \times \mathcal{Y}.$$

Certain properties of the cost function carry over to the set of c-concave functions, and thus also to Kantorovich potentials. For example, if c is absolutely bounded by some value $C > 0$, then Kantorovich potentials can always be shifted such that f and f^c are absolutely bounded by $2C$. Furthermore, if the family $\{c(\cdot, y) \,|\, y \in \mathcal{Y}\}$ of partially evaluated cost functions is *equicontinuous*, then all functions $f \in S_c$ share the same modulus of continuity. In particular, if $c(\cdot, y)$ is L-Lipschitz or (α, L)-Hölder for $L > 0$ and $0 < \alpha \leq 1$, then so is each Kantorovich potential. Another inherited property concerns concavity in Euclidean spaces $\mathcal{X} = \mathbb{R}^d$: if there exists a $\lambda > 0$ such that $c(\cdot, y)$ is *λ-semiconcave* for each $y \in \mathcal{Y}$, meaning that $x \mapsto c(\cdot, y) - \lambda\|x\|^2$ is concave, then each $f \in S_c$ is λ-semiconcave as well. Note that the preferential treatment of f and μ in the presented formalism is arbitrary, and the special roles could also be assumed by $g = f^c$ and ν.

2.2 Lower Complexity Adaptation

This section summarizes the statistical theory developed in Contribution A, which I co-authored together with Shayan Hundrieser and Axel Munk. The contributions of Shayan Hundrieser and me were of comparable nature. We established the foundation of the research paper – the observation of lower complexity adaptation – during intense discussions and close collaboration. We both contributed in significant ways to the writing of the document, which was restructured and generalized several times, in part prompted by valuable insights and suggestions of Axel Munk. I conducted the accompanying simulations.

Let $X_1, \ldots, X_n \sim \mu$ and $Y_1, \ldots, Y_n \sim \nu$ be $n \in \mathbb{N}$ independent and identically distributed random variables with probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ on Polish spaces $\mathcal{X}$ and $\mathcal{Y}$. We call the measures

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad \text{and} \qquad \hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$$

the *empirical measures* with respect to μ and ν, where δ_x denotes the point mass on the point x. We are concerned with situations in which μ is unknown, ν is unknown, or both are unknown, so we consider any of the empirical plug-in estimators

$$\hat{T}_{c,n} \in \big\{T_c(\hat{\mu}_n, \nu), T_c(\mu, \hat{\nu}_n), T_c(\hat{\mu}_n, \hat{\nu}_n)\big\},$$

in order to estimate the true population cost $T_c(\mu, \nu)$. Our main results in Contribution A concern the behavior of the *mean absolute error*

$$\mathbb{E}\big[\big|\hat{T}_{c,n} - T_c(\mu, \nu)\big|\big] \tag{2.4}$$

as a function of n. In particular, we search for upper (and lower) bounds of this quantity, which depend on the properties of c, μ, and ν. By exploiting the dual formulation (2.2) together with the facts on c-concave functions presented in the previous section, one can establish that

$$\mathbb{E}\big[\big|\hat{T}_{c,n} - T_c(\mu, \nu)\big|\big] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}_c} \big|(\hat{\mu}_n - \mu)f\big|\right] + \mathbb{E}\left[\sup_{f^c \in \mathcal{F}_c^c} \big|(\hat{\nu}_n - \nu)f^c\big|\right], \tag{2.5}$$

where $\mathcal{F}_c \subset S_c$ is a suitable subset of c-concave functions that can replace $L^1(\mu)$ in problem (2.2), and $\mathcal{F}_c^c$ is the element wise c-transform of functions in $\mathcal{F}_c$. In our case, we assume c to be absolutely bounded, which means that $\mathcal{F}_c$ can be picked as a set of c-concave functions that are absolutely bounded as well.

The expectation of suprema over function classes on the right hand side of inequality (2.5) are classical objects of interest in empirical process theory (Vaart and Wellner 1996; Wainwright 2019). This theory provides suitable upper bound in terms of the *covering numbers* of a function class. We make use of the *uniform* covering numbers $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ of a class $\mathcal{F}$ of real valued functions on some space $\mathcal{X}$, which is defined as the minimal number m of functions $f_1, \ldots, f_m \colon \mathcal{X} \to \mathbb{R}$ necessary such that

$$\sup_{f \in \mathcal{F}} \min_{1 \leq i \leq m} \|f_i - f\|_\infty < \epsilon,$$

where $\|\cdot\|_\infty$ denotes the uniform norm. Then, one can show that for each $\delta > 0$

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_c} \big|(\hat{\mu}_n - \mu)f\big|\right] \lesssim \delta + \frac{1}{\sqrt{n}} \int_\delta^\infty \sqrt{\log \mathcal{N}\big(\epsilon, \mathcal{F}_c, \|\cdot\|_\infty\big)}\, \mathrm{d}\epsilon, \tag{2.6}$$

where $\lesssim$ denotes inequality up to a constant. An analogous bound holds for the supremum over $f^c \in \mathcal{F}_c^c$ as well. The central observation now, which forms the foundation for the LCA property, is that $\mathcal{N}(\epsilon, \mathcal{F}_c, \|\cdot\|_\infty)$ can be shown to be *equal* to $\mathcal{N}(\epsilon, \mathcal{F}_c^c, \|\cdot\|_\infty)$ for bounded cost functions. Therefore, it actually does not matter whether we control the complexity of $\mathcal{F}_c$ or the one of $\mathcal{F}_c^c$, even if we sample from both measures. Our main result, Theorem 1 in Contribution A, operates under the assumption that

$$\log \mathcal{N}\big(\epsilon, \mathcal{F}_c, \|\cdot\|_\infty\big) = \log \mathcal{N}\big(\epsilon, \mathcal{F}_c^c, \|\cdot\|_\infty\big) \lesssim \epsilon^k \tag{2.7}$$

for some $k > 0$, based on which (2.5) can be combined with (2.6) for suitable choices of δ to yield the upper bounds

$$\mathbb{E}\big[\big|\hat{T}_{c,n} - T_c(\mu,\nu)\big|\big] \lesssim \begin{cases} n^{-1/2} & \text{if } k < 2, \\ n^{-1/2}\log n & \text{if } k = 2, \\ n^{-1/k} & \text{if } k > 2. \end{cases} \tag{2.8}$$

Since we can always choose $\mathcal{X} = \operatorname{supp}\mu$ and $\mathcal{Y} = \operatorname{supp}\nu$, which makes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ classes of functions defined on the support of the measures μ and ν, this approach lets us derive upper bounds for (2.4) that only rely on the properties of c as well as the *simpler* of the supports of μ and ν.

The remainder of Contribution A is chiefly concerned with various settings of interest to which (2.8) can be applied in order to obtain novel upper bounds. In total, we derive bounds for *semidiscrete*, *Lipschitz*, *semiconcave*, and *Hölder-smooth* settings in Sections 3.1 to 3.4. In each of these cases, the uniform covering numbers are controlled by relating k in (2.7) to (a) the regularity and dimensionality of $\operatorname{supp}\mu$ or $\operatorname{supp}\nu$ (whichever turns out to provide better results), and (b) the regularity of c evaluated in either the first or second component. Put in conceptual terms, the *smoothness* $0 < \alpha \leq 2$ of the setting and the intrinsic *dimension* d of either μ or ν are related to k via $k = d/\alpha$. This leads to the general upper bounds

$$\mathbb{E}\big[\big|\hat{T}_{c,n} - T_c(\mu,\nu)\big|\big] \lesssim \begin{cases} n^{-1/2} & \text{if } d < 2\alpha, \\ n^{-1/2}\log n & \text{if } d = 2\alpha, \\ n^{-\alpha/d} & \text{if } d > 2\alpha, \end{cases} \tag{2.9}$$

which can actually be understood to cover the semidiscrete ($d = 0$), Lipschitz ($\alpha = 1$), semiconcave ($\alpha = 2$), or the general α-Hölder-smooth setting (up to $\alpha = 2$, i.e., the existence of Lipschitz first derivatives).

Of course, several technical aspects have to be fleshed out to arrive at well-defined meanings for the smoothness α and the dimension d in this context. First, to capture the notion of the intrinsic dimension d of a probability measure, say μ, we essentially expect that $\operatorname{supp}\mu$ is contained in the *union of the images* of finitely many mappings $g\colon \mathcal{U} \to \mathcal{X}$ from well-behaved, bounded patches $\mathcal{U} \subset \mathbb{R}^d$. Second, to arrive at a suitable notion of smoothness α, we assume that the *composition* $u \mapsto c(g(u), y)$ has suitable smoothness properties uniformly in $y \in \mathcal{Y}$. Due to a union bound (Lemma 2) and a composition bound (Lemma 3) for covering numbers derived in Contribution A, these two assumptions lend themselves well to control the uniform covering numbers of c-concave functions on $\operatorname{supp}\mu$. Special instances of this setting include compact d-dimensional smooth manfolds $\mathcal{X}$, where the cost function $c(\cdot, y)$ is once (for $\alpha = 1$) or twice (for $\alpha = 2$) continuously differentiable with bounds on the derivatives that are uniform in $y \in \mathcal{Y}$ (see Example 2 and 4 in Contribution A). The precise requirements for the different settings are formulated in Assumptions Lip, SC, and Hol in Contribution A.

Besides upper bounds derived from (2.8), additional work in Contribution A also concerns lower bounds for the mean absolute error of the empirical optimal transport cost. Here, we exploit that the LCA principle is also accessible from the primal formulation of optimal transport in settings with a particular geometry (see Section 2.3 in Contribution A). In such settings, we find that the rates in the upper bounds in (2.9) are essentially sharp (with a missing logarithmic term for $d = 2\alpha$), see Examples 3 and 5. Additionally, we have numerically verified that the convergence rates predicted by (2.9) conform well with the ones observed in simulations (Section 4 of Contribution A).

While the concept of lower complexity adaptation provides valuable insight into the nature of empirical optimal transport in high dimensional settings, there are some technical limitations in our work that are subject to future research. For example, our argumentation relies on bounded settings – bounded cost functions and measures with bounded support. Otherwise, the uniform covering numbers can not be controlled well or are not even finite. This restriction could potentially be addressed by using sample-dependent covering numbers (in which case the argumentation via the properties of c-transforms is not straightforward anymore, however), or by employing concentration-of-measure arguments (e.g., arguing that samples reside in compact subsets of the support with high probability), which would require explicit and tight control over the constants in many of the emerging inequalities. Furthermore, our proof strategy via empirical process theory only works under i.i.d. assumptions on the data. It is as of yet unknown if an LCA principle also holds for dependent data, or even for completely different estimators than the empirical ones. While first numerical evidence suggests that a moderate amount of dependency in the observed data does not corrupt the LCA property, exploring in how far this principle actually permeates statistical optimal transport theory in general is an exciting venue for further work.

2.3 Properties of Kantorovich Potentials

This section is concerned with an overview over the results on deterministic properties of Kantorovich potentials that I have established while working on Contribution B and C. Contribution B, on distributional limit laws of the empirical optimal transport cost, is joint work together with Shayan Hundrieser, Marcel Klatt, and Axel Munk. Besides some corrections and improvements of general nature, I have derived and formulated the results on constant Kantorovich potentials in Section 4 of the research paper. Contribution C, on the uniqueness of Kantorovich Potentials, is joint work together with Shayan Hundrieser and Axel Munk. I have largely prepared and formulated the document, while discussions with Axel Munk and especially Shayan Hundrieser were crucial to flesh out the core ideas.

As mentioned in Section 2.1, Kantorovich potentials are never truely unique in the strict sense of the word, since $f \in S_c(\mu, \nu)$ implies $f + a \in S_c(\mu, \nu)$ for any constant

$a \in \mathbb{R}$. Therefore, in the following, uniqueness always means uniqueness *up to constant shifts*. Moreover, it is often too restrictive to expect uniqueness on all of $\mathcal{X}$, and it is usually sufficient to work with μ-almost sure uniqueness. Similarly, we do not expect $f \in S_c(\mu, \nu)$ to be constant on all of $\mathcal{X}$ if we talk about constant Kantorovich potentials; we just assume it to be constant μ-almost surely. In Contribution B we also use the term *trivial* to refer to almost surely constant Kantorovich potentials.

The motivation to scrutinize the aforementioned properties of Kantorovich potentials (them being *unique* and them being *constant*) not only stems from general interest into structural aspects of optimal transport, but also from important questions arising in statistical optimal transport. For example, Contribution B establishes that distributional limits of the empirical optimal transport cost in diverse settings assume the form

$$\sqrt{n}\big(T_c(\hat{\mu}_n, \nu) - T_c(\mu, \nu)\big) \to \sup_{f \in S_c(\mu,\nu)} G(f), \tag{2.10a}$$

where the notation of Section 2.1 and 2.2 above is reused. Here, G denotes a centered Gaussian process indexed in $S_c(\mu, \nu)$, which is determined by the covariance structure

$$\mathbb{E}[G(f_1)G(f_2)] = \mu(f_1 f_2) - \mu f_1 \, \mu f_2$$

for $f_1, f_2 \in S_c(\mu, \nu)$. This in particular implies that the variance of $G(f)$ vanishes if and only if f is constant μ-almost surely. Note that validity of (2.10a) is only proven in Contribution B for settings with sufficient regularity of c as well as an intrinsic dimension $d \leq 3$ of at least one of the measures μ and ν (this exploits the LCA property, see Section 5.5 in Contribution B). Indeed, there is good reason to believe that limit laws of this form will fail if $d \geq 4$ due to a high bias of the empirical optimal transport cost (see Section 5.4 in Contribution B). Still, under a set of technical assumptions, related results by del Barrio and Loubes 2019 and del Barrio et al. 2021 prove Gaussian limits of the form

$$\sqrt{n}\big(T_c(\hat{\mu}_n, \nu) - \mathbb{E}[T_c(\hat{\mu}_n, \nu)]\big) \to G(f) \tag{2.10b}$$

in general dimensions, provided that there is an (up to constants) unique Kantorovich potential $f \in S_c(\mu, \nu)$. Notably, these limits do not center around the population value but the expectation of the empirical cost, and it again holds that $G(f)$ has expectation zero and variance $\mu f^2 - (\mu f)^2$. Analogous distributional limits as in (2.10) also hold if ν is estimated by its empirical measure $\hat{\nu}_n$, in which case the c-transformed Kantorovich potentials f^c enter the formalism. We conclude that *uniqueness* of Kantorovich potentials is intimately related to *Gaussianity* of the distributional limits of the empirical optimal transport cost. If the unique Kantorovich potentials are additionally *constant*, then the limits *degenerate* to a point measure at zero.

The usual argumentation to derive uniqueness of dual optimal transport solutions exploits the fact that any $f \in S_c(\mu, \nu)$ satisfies $f(x) + f^c(y) \leq c(x, y)$ with equality if (x, y) lies in the support of an optimal transport plan π. This implies that the mapping given by $x' \mapsto c(x', y) - f(x)$ has to attain its minimum at such $(x, y) \in \operatorname{supp} \pi$.

If several conditions align – the space $\mathcal{X}$ has a differentiable structure, the cost function is differentiable in the first component, and the Kantorovich potential is also differentiable – this characterization will essentially determine the *gradient* of f on the support of μ. And if the gradient of f is determined, then the potential f itself is (up to constants) determined on each connected component of the support (barring technicalities). Thus, if the support of μ is *well-behaved and connected*, then the regularity of the cost function, which affects the regularity of S_c as well, may directly lead to the uniqueness of Kantorovich potentials. Several results that follow this logic to establish statements on the uniqueness of Kantorovich potentials are available in the literature (Villani 2008, Remark 10.30; Santambrogio 2015, Proposition 7.18; del Barrio et al. 2021, Corollary 2.7). All of them, however, are (in part much) more restrictive in their assumptions than necessary.

One important pillar of Contribution C is a systematic reiteration of this reasoning, with some care taken to only work with assumptions that are actually needed to arrive at uniqueness. Out of this grew Theorem 2 in Contribution C, which applies to settings where $\mathcal{X}$ is a smooth manifold and the cost function c is locally Lipschitz, while $\mathcal{Y}$ can be arbitrary Polish. The technical difficulties that make the assumptions of this statement somewhat opaque are mainly related to controlling *transport towards infinity*, meaning that mass around some points in $\operatorname{supp}\mu$ can be transported out of any $\mathcal{Y}$-compactum. Around these points, there is no guarantee for sufficient regularity of c-concave functions, which could prevent the necessary differentiability of $f \in S_c(\mu, \nu)$. These problems, however, can be shown to be immaterial in a number of more specialized settings, like the one of compact spaces $\mathcal{Y}$ (Corollary 2 in Contribution C) or the one of geodesic spaces $\mathcal{X} = \mathcal{Y}$ with rapidly increasing costs (Section 4 in Contribution C). The conceptual takeaway from these results is that uniqueness in connected settings often holds for differentiable costs under mainly *topological* conditions on the support of μ, while the actual distribution of mass (at least in the interior of $\operatorname{supp}\mu$) does not matter.

The actual main advancement in Contribution C, however, concerns novel insights into the uniqueness of Kantorovich potentials in *disconnected* settings. Since it is easy to find instances where uniqueness fails due to a disconnected support – for example, if $\mu = \nu$ is the uniform measure on the union of two disjoint compact sets in $\mathbb{R}^d$ under squared Euclidean costs – it has been a common perception that uniqueness relies on the connectedness of the support of at least one of the measures μ and ν. That this cannot be the full story is anticipated by the theory of finite linear programming: if $\mathcal{X} = \{x_i \mid i \in I\}$ and $\mathcal{Y} = \{y_j \mid j \in J\}$ are finite discrete spaces with index sets I and J, it is well-known (Hung et al. 1986) that transportation problems have unique dual solutions whenever there are no two proper nonempty subsets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$ such that $\mu(A) = \nu(B)$. Intuitively, this condition rules out that the transport problem might be split up in two separate problems by decomposing $\mathcal{X}$ and $\mathcal{Y}$. These settings are referred to as *non-degenerate*. To lift these results to non-discrete settings with measures μ and ν on general Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, let

$$\operatorname{supp}\mu = \bigcup_{i\in I}\mathcal{X}_i \qquad \text{and} \qquad \operatorname{supp}\nu = \bigcup_{j\in J}\mathcal{Y}_j \tag{2.11}$$

denote the decomposition of the respective supports into their connected components. For simplicity, the index sets I and J are taken to be finite. Furthermore, let $\pi \in \mathcal{C}(\mu, \nu)$ be an optimal plan of the optimal transport problem between μ and ν under the cost function c. We denote the optimal transport problem between $\mu_i = \mu|_{\mathcal{X}_i}/\mu(\mathcal{X}_i)$ and $\nu_i = \pi(\mathcal{X}_i \times \cdot)/\mu(\mathcal{X}_i)$ under c as the *X_i-restricted problem* for given $i \in I$. Furthermore, we say that π is *non-degenerate* if there are no nonempty proper subsets $I' \subset I$ and $J' \subset J$ such that

$$\sum_{i\in I'}\mu(\mathcal{X}_i) = \sum_{i\in I', j\in J'}\pi(\mathcal{X}_i \times \mathcal{Y}_j) = \sum_{j\in J'}\nu(\mathcal{Y}_j). \tag{2.12}$$

What this effectively means is that it is impossible to split $\operatorname{supp}\mu$ and $\operatorname{supp}\nu$ into groups of connected components such that the optimal plan does not mix mass between the different groups. With these preparations, our uniqueness result for disconnected settings, Theorem 1 in Contribution C, can be summarized as follows: (a) if the optimal transport plan π is non-degenerate, (b) if the c-transformed Kantorovich potentials $S_c^c(\mu, \nu)$ are continuous, and (c) if the Kantorovich potentials of all $\mathcal{X}_i$-restricted problems are unique, then the potentials $S_c(\mu, \nu)$ of the full problem (and equivalently their c-transforms) are also unique.

Condition (a) might seem hard to verify, since it is concerned with properties of the (potentially unknown) optimal transport plan, but it can easily be ensured by enforcing that the measures μ and ν do not assign the same amount of mass to groups of connected components, analogous to the finite case. The continuity condition (b) can typically be shown to hold in interesting settings (see Section 2 and 4 of Contribution C). It can even be weakened, which is further discussed in the context of Theorem 1 in Contribution C. Finally, condition (c) relies on uniqueness results on the connected components of $\operatorname{supp}\mu$. As outlined above, such statements often hold under mild regularity conditions for c and topological requirements for $\mathcal{X}_i$. Furthermore, extensions to countable index sets I and J are, with some technical limitations, possible and further discussed in Contribution C. To our knowledge, this is the first result on the uniqueness of Kantorovich potentials in disconnected (but not finite) settings in the literature on optimal transport. The fundamental message is that non-uniqueness due to disconnected supports is typically the *exception*, caused by a specific symmetry in the mass distribution, rather than the rule.

All taken together, Contribution C provides a quite complete picture regarding the issue of Gaussian distributional limits in empirical optimal transport theory. It establishes that the main enemies of Gaussianity are (a) non-differentiabilities of the cost function, which is well-known for metric costs, and (b) disconnected supports with a peculiar mass distribution. If these factors do not apply, Gaussian limits can be expected. The remaining question is under which conditions these Gaussian limits have a vanishing variance, which occurs if the unique Kantorovich potentials are constant. Section 4 in Contribution B is dedicated to this problem and provides a compelling geometric answer.

Given a probability measure $\nu \in \mathcal{P}(\mathcal{Y})$, we say that $\mu \in \mathcal{P}(\mathcal{X})$ is a *ν-projected measure* with respect to a cost function c if there exists a coupling $\pi \in \mathcal{C}(\mu, \nu)$ such that

$$c(x, y) = \inf_{x' \in \operatorname{supp} \mu} c(x', y) \qquad \text{for all } (x, y) \in \operatorname{supp} \pi. \tag{2.13}$$

If the relation is reversed, we say that ν is a μ-projected measure. Expressed in words, a ν-projected measure is one that can be transported to μ by simply projecting each point in the support of ν to the support of μ, in the sense of finding the point with the smallest movement cost according to c. The coupling π in (2.13) is then automatically optimal. Clearly, such an arrangement between μ and ν implies severe restrictions on how the supports (and the mass distribution on the supports) can be situated relative to one another. For Euclidean costs, for example, μ cannot be ν-projected if their supports are disjoint and $\operatorname{supp} \mu$ has interior points, since only boundary points could be reached via projection (see Section 4 in Contribution B for more examples and illustrative figures).

In a certain way, projected arrangements like these make the optimal transport problem particularly simple: to find the optimal plan, each point can be projected independently, effectively eliminating the marginal constraint that makes optimal transport a hard problem in the first place. This simplicity also has consequences of statistical nature. Theorem 5 in Contribution B states that constant Kantorovich potentials $f \in S_c(\mu, \nu)$ exist *if and only if* μ is ν-projected. Likewise, constant c-transformed Kantorovich potentials $f^c \in S_c^c(\mu, \nu)$ exist if and only if ν is μ-projected. As a direct consequence, under the assumption of unique Kantorovich potentials, the limit laws for $T_c(\hat{\mu}_n, \nu)$ degenerate if and only if μ is ν-projected; the limit laws for $T_c(\mu, \hat{\nu}_n)$ degenerate if and only if ν is μ-projected; and the limit laws for $T_c(\hat{\mu}_n, \hat{\nu}_n)$ degenerate if and only if both measures are projected with respect to the other. This provides a complete characterization of degenerate limit laws in intuitive geometric terms.

While the work on uniqueness and triviality of Kantorovich potentials in Contribution B and Contribution C covers a lot of ground and offers a comprehensive image, some aspects still remain open and are subject to future work. On the technical side, this includes the question in how far Theorem 1 in Contribution C can be generalized to countable index sets I. Also, our work on uniqueness for connected settings currently (at least in part) relies on the assumption that no mass is placed on the boundary of the support. I suspect that this is not a genuine condition but a technical artefact that may be overcome with some additional care. Another point of interest is to also find *necessary* conditions for the uniqueness of Kantorovich potentials, i.e., to better characterize settings where the potentials are known to be ambiguous.

Especially if the ambiguity is due to disconnected supports, criteria along those lines are not well understood and only a few simple examples are accessible (see, e.g., Lemma 11 in Contribution C).

2.4 Quantifying Dependency

This section summarizes some of the main findings of my work on the transport dependency, a concept to measure statistical dependency via optimal transport. It is based on Contribution D, which is joint work together with Thomas Giacomo Nies and Axel Munk. Most of the core insights and results that the research paper features were developed in active discussion and collaboration with Thomas Giacomo Nies, while Axel Munk provided the initial idea and proposed important corrections. The text was largely prepared by myself and the included simulations are joint work.

Let X and Y be random variables on Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ with marginal distributions $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. Their joint distribution is denoted by $\gamma \in \mathcal{C}(\mu, \nu)$. For a given cost function $c\colon (\mathcal{X} \times \mathcal{Y})^2 \to \mathbb{R}$, the *transport dependency* is defined as

$$\tau_c(X, Y) = \tau_c(\gamma) = T_c(\gamma, \mu \otimes \nu), \tag{2.14}$$

which involves an optimal transport problem on the product space $\mathcal{X} \times \mathcal{Y}$. Here, $\mu \otimes \nu$ is the product measure of μ and ν. If γ equals $\mu \otimes \nu$, the two random variables X and Y are statistically independent. Thus, the transport dependency τ_c quantifies the discrepancy between the *actual* joint distribution γ and the *independent* case $\mu \otimes \nu$. In this sense, it is comparable to the *mutual information* of X and Y, a well-known quantifier of dependency, that is defined as the Kullback-Leibler divergence between γ and $\mu \otimes \nu$. For reasonable cost functions, like metrics on $\mathcal{X} \times \mathcal{Y}$, the relation $\tau_c(X, Y) = 0$ completely characterizes statistical independence. In contrast, a large value of $\tau_c(X, Y)$ intuitively suggests that X and Y should be strongly dependent.

This raises the question what "strong dependency" between random variables actually means. In general, this question has no easy or canonical answers. The mutual information, for instance, is maximized under deterministic relations $Y = \varphi(X)$ for *any* measurable function $\varphi\colon \mathcal{X} \to \mathcal{Y}$ or vice versa. According to this criterion, even very chaotic relations, for which prediction of Y based on X could be impossible in statistical settings with a finite amount of data, would express strong dependency. The transport dependency, on the other hand, enforces much more structure on relations that maximize it. In particular, via the choice of the cost function c, it can be adapted to take metric properties into account. Under costs of the form

$$c(x_1, y_1, x_1, y_2) = d_{\mathcal{X}}(x_1, x_2) + d_{\mathcal{Y}}(y_1, y_2), \tag{2.15}$$

where $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are metrics on $\mathcal{X}$ and $\mathcal{Y}$, one of the core insights of Contribution D states that $\tau_c(X, Y)$ is essentially maximized if and only if X and Y are deterministically related by Lipschitz functions. More precisely, we prove in Proposition 6 that the transport dependency is upper bounded by

$$\tau_c(X, Y) \leq \mathbb{E}[d_{\mathcal{Y}}(Y_1, Y_2)], \tag{2.16}$$

where $Y_1, Y_2 \sim \nu$ are i.i.d. copies of Y, and that this upper bound is attained *if and only if* there is a *1-Lipschitz function* $\varphi\colon (\mathcal{X}, d_{\mathcal{X}}) \to (\mathcal{Y}, d_{\mathcal{Y}})$ such that $Y = \varphi(X)$ holds (Theorem 7 and Corollary 1). Of course, the roles of X and Y in this statement can also be reversed with an equivalent upper bound based on $d_{\mathcal{X}}$.

This insight equips the transport dependency with a clean interpretation: a large value of $\tau_c(X, Y)$ hints at a deterministic relation between X and Y that is very structured, in the sense that perturbing the value of X will not change Y too much. Furthermore, the upper bound in (2.16) offers a natural way to normalize τ_c to obtain a *coefficient of dependency* with values in $[0, 1]$. For cost functions of the form

$$c(x_1, y_1, x_1, y_2) = \alpha\, d_{\mathcal{X}}(x_1, x_2) + d_{\mathcal{Y}}(y_1, y_2),$$

where $\alpha > 0$, we define the *α-transport correlation* $\rho_\alpha(X, Y) = \tau_c(X, Y)/\mathbb{E}[d_{\mathcal{Y}}(Y_1, Y_2)]$. It is equal to 0 if and only if X and Y are independent, and equals 1 if and only if $Y = \varphi(X)$ for an α-Lipschitz function φ. These and several other hallmark properties of ρ_α are derived in Section 5 of Contribution D; for example, it is invariant under isometric transformations of the spaces $\mathcal{X}$ and $\mathcal{Y}$. Also, we study two special cases that lead to particularly interesting dependency coefficients. For a special choice of α, which depends on the marginals μ and ν, the resulting coefficient $\rho_*(X, Y)$ is *symmetric* in X and Y. In particular, it attains the value 1 if and only if $Y = \varphi(X)$ for a multiple of an *isometry* φ. In the limit $\alpha \to \infty$, on the other hand, we obtain a dependency coefficient ρ_∞ that assumes maximal values for *any* measurable function $\varphi\colon \mathcal{X} \to \mathcal{Y}$. In this respect, it is closer to the properties of purely information-based concepts of dependency quantification, like the mutual information. The transport dependency is hence a very flexible tool and allows for the interpolation between *structured* (emphasizing isometric or Lipschitz relations) and *unstructured* (emphasizing measurable relations) dependency quantification.

In the presence of empirical data, the transport dependency and the derived coefficients can be estimated in a straightforward manner. If $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. samples drawn from $\gamma \in \mathcal{C}(\mu, \nu)$, then the plug-in approach

$$\tau_c(\hat{\gamma}_n) = T_c(\hat{\gamma}_n, \hat{\mu}_n \otimes \hat{\nu}_n) \tag{2.17}$$

where

$$\hat{\gamma}_n = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_i)} \qquad \text{and} \qquad \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \quad \hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i},$$

defines a natural estimator of $\tau_c(\hat{\gamma}_n)$. This estimator is consistent under suitable moment assumptions (Theorem 2 in Contribution D). Furthermore, we have established that $\mathbb{E}[|\tau_c(\hat{\gamma}_n) - \tau_c(\gamma)|] \leq 3\, T_c(\hat{\gamma}_n, \gamma)$ for cost functions like in (2.15). Since this means that the intrinsic dimension of γ determines the rate of convergence (and not the potentially higher dimension of $\mu \otimes \nu$), the estimator $\tau_c(\hat{\gamma}_n)$ in fact features *lower complexity adaption.* Furthermore, the numerical study in Contribution D shows that the transport dependency under this estimator compares generally very well to established methods when discerning dependency in permutation based independence tests. We are currently also working on the application of the transport dependency to gene expression data, where our generic approach seems to admit similar conclusions to much more specialized methods. The overall picture emerging from our work thus confirms that the transport dependency is a robust and versatile quantifier of dependency that can greatly benefit from the flexibility to work with arbitrary notions of costs or similarity.

Still, there are some critical drawbacks to the transport dependency, in particular regarding the plug-in estimator $\tau_c(\hat{\gamma}_n)$. Foremost, the measure $\hat{\mu}_n \otimes \hat{\nu}_n$ is supported on n^2 points, which quickly makes computation of $\tau_c(\hat{\gamma}_n)$ very expensive. This essentially limits its usage to problems with small to moderate values of n for many practical problems (see Section 6 in Contribution D). Combined with the curse of dimensionality, which makes $\tau_c(\hat{\gamma}_n)$ suffer from a high bias if the dimensionality of γ is large, this might pose a serious hurdle for the application of the transport dependency in statistical data science and machine learning applications. However, these apparent drawbacks also raise interesting questions. For example, our work in Contribution A indicates that alternative estimators of the form

$$\hat{\tau}_{c,n} = T_c\big(\hat{\gamma}_n, (\widehat{\mu \otimes \nu})_n\big),$$

where $\mu \otimes \nu$ is not estimated by $\hat{\mu}_n \otimes \hat{\nu}_n$, but by an estimator $(\widehat{\mu \otimes \nu})_n$ with fewer support points only, might be able to achieve the same convergence rates as $\tau_c(\hat{\gamma}_n)$ while being computationally much more viable. Simulations and theoretical results which confirm the thrust of this argument are currently being worked on. Noteworthy issues in this context include the questions (a) how to best estimate $\mu \otimes \nu$ based on the n given sample points (e.g., via random permutations or clustering), (b) if an LCA property can rigorously be established for the derived estimator $\hat{\tau}_{c,n}$, and (c) if even distributional limit laws in the sense of Contribution B might be feasible.

Bibliography

Altschuler, J., J. Niles-Weed, and P. Rigollet (2017). “Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Ambrosio, L., D. Semola, and E. Brué (2021). *Lectures on Optimal Transport*. Springer.

Arjovsky, M., S. Chintala, and L. Bottou (2017). “Wasserstein generative adversarial networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, pp. 214–223.

Beiglböck, M. and W. Schachermayer (2011). “Duality for Borel measurable cost functions”. *Transactions of the American Mathematical Society* 363.8, pp. 4203–4224.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Bellman, R. (1961). *Adaptive Control Processes. A Guided Tour*. Princeton University Press.

Bellman, R. (1984). *Eye Of The Hurricane*. World Scientific Publishing Company.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). “Simultaneous analysis of lasso and Dantzig selector”. *The Annals of Statistics* 37.4, pp. 1705–1732.

Candes, E. J. and T. Tao (2005). “Decoding by linear programming”. *IEEE Transactions on Information Theory* 51.12, pp. 4203–4215.

Candes, E. J. and T. Tao (2007). “The Dantzig selector: Statistical estimation when p is much larger than n”. *The Annals of Statistics* 35.6, pp. 2313–2351.

Comon, P. (1994). “Independent component analysis, a new concept?” *Signal processing* 36.3, pp. 287–314.

Courty, N., R. Flamary, A. Habrard, and A. Rakotomamonjy (2017). “Joint distribution optimal transportation for domain adaptation”. *Advances in Neural Information Processing Systems* 30.

Cuturi, M. (2013). “Sinkhorn distances: Lightspeed computation of optimal transport”. In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc.

Deb, N., P. Ghosal, and B. Sen (2021). “Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 29736–29753.

del Barrio, E., A. González-Sanz, and J.-M. Loubes (2021). “Central limit theorems for general transportation costs”. *Preprint arXiv:2102.06379*.

del Barrio, E. and J.-M. Loubes (2019). “Central limit theorems for empirical transportation cost in general dimension”. *The Annals of Probability* 47.2, pp. 926–951.

Donoho, D. L. (2006a). “Compressed sensing”. *IEEE Transactions on Information Theory* 52.4, pp. 1289–1306.

Donoho, D. L. (2006b). “For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution”. *Communications on Pure and Applied Mathematics* 59.6, pp. 797–829.

Donoho, D. L., I. M. Johnstone, G. Kerkyacharian, and D. Picard (1996). “Density estimation by wavelet thresholding”. *The Annals of Statistics*, pp. 508–539.

Dudley, R. M. (1969). “The speed of mean Glivenko-Cantelli convergence”. *The Annals of Mathematical Statistics* 40.1, pp. 40–50.

Eddy, S. R. (2004). “What is dynamic programming?” *Nature Biotechnology* 22.7, pp. 909–910.

Elhamifar, E. and R. Vidal (2013). “Sparse subspace clustering: Algorithm, theory, and applications”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.11, pp. 2765–2781.

Fan, J. and J. Lv (2008). “Sure independence screening for ultrahigh dimensional feature space”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5, pp. 849–911.

Feydy, J., T. Séjourné, F.-X. Vialard, S. Amari, A. Trouve, and G. Peyré (2019). “Interpolating between optimal transport and MMD using Sinkhorn divergences”. In: PMLR, pp. 2681–2690.

Forrow, A., J.-C. Hütter, M. Nitzan, P. Rigollet, G. Schiebinger, and J. Weed (2019). “Statistical optimal transport via factored couplings”. In: PMLR, pp. 2454–2465.

Friedman, J. H. and J. W. Tukey (1974). “A projection pursuit algorithm for exploratory data analysis”. *IEEE Transactions on Computers* 100.9, pp. 881–890.

Galichon, A. (2016). *Optimal Transport Methods in Economics*. Princeton University Press.

Ge, Z., S. Liu, Z. Li, O. Yoshie, and J. Sun (2021). “OTA: Optimal transport assignment for object detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 303–312.

Genevay, A., L. Chizat, F. Bach, M. Cuturi, and G. Peyré (2019). “Sample complexity of Sinkhorn divergences”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1574–1583.

Genevay, A., M. Cuturi, G. Peyré, and F. Bach (2016). “Stochastic optimization for large-scale optimal transport”. In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc.

Goldberg, Y., A. Zakai, D. Kushnir, and Y. Ritov (2008). “Manifold learning: The price of normalization”. *Journal of Machine Learning Research* 9.8.

Goldfeld, Z. and K. Greenewald (2020). “Gaussian-smoothed optimal transport: Metric structure and statistical efficiency”. In: PMLR, pp. 3327–3337.

Goldfeld, Z., K. Greenewald, J. Niles-Weed, and Y. Polyanskiy (2020). “Convergence of smoothed empirical measures with applications to entropy estimation”. *IEEE Transactions on Information Theory* 66.7, pp. 4368–4391.

Hastie, T. and W. Stuetzle (1989). “Principal curves”. *Journal of the American Statistical Association* 84.406, pp. 502–516.

Hoerl, A. E. and R. W. Kennard (1970). “Ridge regression: Biased estimation for nonorthogonal problems”. *Technometrics* 12.1, pp. 55–67.

Hundrieser, S., M. Klatt, T. Staudt, and A. Munk (2022a). “A unifying approach to distributional limits for empirical optimal transport”. *Preprint arXiv:2202.12790.*

Hundrieser, S., T. Staudt, and A. Munk (2022b). “Empirical optimal transport between different measures adapts to lower complexity”. *Preprint arXiv:2202.10434.*

Hung, M. S., W. O. Rom, and A. D. Waren (1986). “Degeneracy in transportation problems”. *Discrete Applied Mathematics* 13.2, pp. 223–237.

Jacobs, M. and F. Léger (2020). “A fast approach to optimal transport: The back-and-forth method”. *Numerische Mathematik* 146.3, pp. 513–544.

Johnstone, I. M. (2001). “On the distribution of the largest eigenvalue in principal components analysis”. *The Annals of Statistics* 29.2, pp. 295–327.

Johnstone, I. M. and D. M. Titterington (2009). “Statistical challenges of high-dimensional data”. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906, pp. 4237–4253.

Kantorovich, L. (1942). “On the translocation of masses”. *Doklady Akademii Nauk URSS* 37, pp. 7–8.

Lee, J. A. and M. Verleysen (2007). *Nonlinear Dimensionality Reduction.* Vol. 1. Springer.

Lohse, L., M. Vassholz, M. Töpperwien, T. Jentschke, A. Bergamaschi, S. Chiriotti, and T. Salditt (2020). “Spectral μCT with an energy resolving and interpolating pixel detector”. *Optics Express* 28.7, pp. 9842–9859.

Mai, V. V., J. Lindbäck, and M. Johansson (2021). “A fast and accurate splitting method for optimal transport: Analysis and implementation”. *Preprint arXiv:2110.11738.*

Manole, T., S. Balakrishnan, J. Niles-Weed, and L. Wasserman (2021). “Plugin estimation of smooth optimal transport maps”. *Preprint arXiv:2107.12364.*

Manole, T. and J. Niles-Weed (2021). “Sharp convergence rates for empirical optimal transport with smooth costs”. *Preprint arXiv:2106.13181.*

Meier, L., S. Van De Geer, and P. Bühlmann (2008). “The group lasso for logistic regression”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 53–71.

Monge, G. (1781). “Mémoire sur la théorie des déblais et des remblais”. In: *Histoire de l’Académie Royale des Sciences de Paris*, pp. 666–704.

Munk, A., T. Staudt, and F. Werner (2020). “Statistical foundations of nanoscale photonic imaging”. In: *Nanoscale Photonic Imaging*. Ed. by T. Salditt, A. Egner, and D. R. Luke. Springer International Publishing, pp. 125–143.

Nies, T. G., T. Staudt, and A. Munk (2021). “Transport dependency: Optimal transport based dependency measures”. *Preprint arXiv:2105.02073.*

Panaretos, V. M. and Y. Zemel (2020). *An Invitation to Statistics in Wasserstein Space.* Springer Nature.

Peyré, G. and M. Cuturi (2019). “Computational optimal transport: With applications to data science”. *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607.

Pollard, C. and P. Windischhofer (2022). “Transport away your problems: Calibrating stochastic simulations with optimal transport”. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1027, p. 166119.

Powell, W. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality.* 2nd ed. Wiley.

Rachev, S. and L. Rüschendorf (1998). *Mass Transportation Problems: Volume I: Theory.* Springer.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.* Springer.

Schiebinger, G., J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. (2019). “Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming”. *Cell* 176.4, pp. 928–943.

Schmidt-Hieber, J. (2020). “Nonparametric regression using deep neural networks with ReLU activation function”. *The Annals of Statistics* 48.4, pp. 1875–1897.

Schmidt-Hieber, J., L. F. Schneider, T. Staudt, A. Krajina, T. Aspelmeier, and A. Munk (2021). “Posterior analysis of n in the binomial (n,p) problem with both parameters unknown – with applications to quantitative nanoscopy”. *The Annals of Statistics* 49.6, pp. 3534–3558.

Schmitzer, B. (2016). “A sparse multiscale algorithm for dense optimal transport”. *Journal of Mathematical Imaging and Vision* 56.2, pp. 238–259.

Schmitzer, B. (2019). “Stabilized sparse scaling algorithms for entropy regularized transport problems”. *SIAM Journal on Scientific Computing* 41.3, A1443–A1481.

Schneider, L. F., T. Staudt, and A. Munk (2019). “Posterior consistency in the binomial model with unknown parameters: A numerical study”. In: *Bayesian Statistics and New Generations.* Ed. by R. Argiento, D. Durante, and S. Wade. Springer International Publishing, pp. 35–42.

Singh, S. and B. Póczos (2018). “Minimax distribution estimation in Wasserstein distance”. *Preprint arXiv:1802.08855.*

Staudt, T., T. Aspelmeier, O. Laitenberger, C. Geisler, A. Egner, and A. Munk (2020). “Statistical molecule counting in super-resolution fluorescence microscopy: Towards quantitative nanoscopy”. *Statistical Science* 35.1, pp. 92–111.

Staudt, T., S. Hundrieser, and A. Munk (2022). “On the uniqueness of Kantorovich potentials”. *Preprint arXiv:2201.08316.*

Stone, C. J. (1985). “Additive regression and other nonparametric models”. *The Annals of Statistics* 13.2, pp. 689–705.

Stone, C. J. (1980). “Optimal rates of convergence for nonparametric estimators”. *The Annals of Statistics* 8.6, pp. 1348–1360.

Tameling, C., S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk (2021). “Colocalization for super-resolution microscopy via optimal transport”. *Nature Computational Science* 1.3, pp. 199–211.

Tibshirani, R. (1996). “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tolstikhin, I., O. Bousquet, S. Gelly, and B. Schoelkopf (2017). “Wasserstein autoencoders”. *Preprint arXiv:1711.01558.*

Vaart, A. W. van der and J. A. Wellner (1996). *Weak Convergence and Empirical Processes.* Springer New York.

Villani, C. (2008). *Optimal Transport: Old and New.* Springer.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press.

Waterman, M. S. (2018). *Introduction to Computational Biology.* Taylor & Francis Ltd. 448 pp.

Weed, J. and F. Bach (2019). “Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance”. *Bernoulli* 25.4A, pp. 2620–2648.

Weed, J. and Q. Berthet (2019). “Estimation of smooth densities in Wasserstein distance”. In: PMLR, pp. 3118–3119.

Yurochkin, M., S. Claici, E. Chien, F. Mirzazadeh, and J. M. Solomon (2019). “Hierarchical optimal transport for document representation”. *Advances in Neural Information Processing Systems* 32.

Zou, H. (2006). “The adaptive lasso and its oracle properties”. *Journal of the American Statistical Association* 101.476, pp. 1418–1429.

Zou, H., T. Hastie, and R. Tibshirani (2006). “Sparse principal component analysis”. *Journal of Computational and Graphical Statistics* 15.2, pp. 265–286.

A Empirical Optimal Transport between Different Measures Adapts to Lower Complexity

Shayan Hundrieser [*,†,‡]
s.hundrieser@math.uni-goettingen.de

Thomas Staudt [*,†,‡]
thomas.staudt@uni-goettingen.de

Axel Munk [†,‡,§]
munk@math.uni-goettingen.de

Abstract

The empirical optimal transport (OT) cost between two probability measures from random data is a fundamental quantity in transport based data analysis. In this work, we derive novel guarantees for its convergence rate when the involved measures are *different*, possibly supported on different spaces. Our central observation is that the statistical performance of the empirical OT cost is determined by the *less* complex measure, a phenomenon we refer to as *lower complexity adaptation* of empirical OT. For instance, under Lipschitz ground costs, we find that the empirical OT cost based on n observations converges at least with rate $n^{-1/d}$ to the population quantity if one of the two measures is concentrated on a d-dimensional manifold, while the other can be arbitrary. For semi-concave ground costs, we show that the upper bound for the rate improves to $n^{-2/d}$. Similarly, our theory establishes the general convergence rate $n^{-1/2}$ for semi-discrete OT. All of these results are valid in the two-sample case as well, meaning that the convergence rate is still governed by the simpler of the two measures. On a conceptual level, our findings therefore suggest that the curse of dimensionality only affects the estimation of the OT cost when *both* measures

[*] These authors contributed equally

[†] Institute for Mathematical Stochastics, University of Göttingen

[‡] Cluster of Excellence: Multiscale Bioimaging (MBExC), University Medical Center, Göttingen

[§] Max Planck Institute for Biophysical Chemistry, Göttingen

exhibit a high intrinsic dimension. Our proofs are based on the dual formulation of OT as a maximization over a suitable function class $\mathcal{F}_c$ and the observation that the c-transform of $\mathcal{F}_c$ under bounded costs has the same uniform metric entropy as $\mathcal{F}_c$ itself.

Keywords: Wasserstein distance, convergence rate, curse of dimensionality, metric entropy, semi-discrete, manifolds

MSC 2020 subject classification: primary 62R07, 62G20, 62G30, 49Q22; secondary 62E20, 62F35, 60B10

1 Introduction

The theory of optimal transport (OT) allows for an effective comparison of probability measures that is faithful to the geometry of the underlying ground space (see Rachev and Rüschendorf 1998a; Rachev and Rüschendorf 1998b; Villani 2003; Villani 2008; Santambrogio 2015 for comprehensive treatments). Origins of OT date back to the seminal work by Monge 1781 and its measure theoretic generalization by Kantorovich 1942; Kantorovich 1958, paving the way for a rich theory and many applications. With recent computational advances (for a survey see Bertsimas and Tsitsiklis 1997; Peyré and Cuturi 2019) OT based methodology is also quickly emerging as a useful tool for data analysis with diverse applications in statistics. This includes bootstrap and resampling (Bickel and Freedman 1981; Sommerfeld et al. 2019; Heinemann et al. 2020), goodness of fit testing (del Barrio et al. 1999; Hallin et al. 2021b), multivariate quantiles and ranks (Chernozhukov et al. 2017; Deb and Sen 2021; Hallin et al. 2021a) and general notions of dependency (Nies et al. 2021; Deb et al. 2021; Mordant and Segers 2022). For a recent survey see Panaretos and Zemel 2019. Further areas of application include machine learning (Arjovsky et al. 2017; Altschuler et al. 2017; Dvurechensky et al. 2018), and computational biology (Evans and Matsen 2012; Schiebinger et al. 2019; Tameling et al. 2021; Wang et al. 2021), among others.

Intuitively, OT aims to transform one probability measure into another one in the most cost-efficient way. For a general formulation, let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be probability measures on Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, and consider a measurable cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The *optimal transport cost* between μ and ν is defined as

$$T_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\pi(x, y), \tag{1}$$

where $\Pi(\mu, \nu)$ represents the set of all couplings between μ and ν, i.e., the probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginal distributions μ and ν. In statistical problems, the measure μ is typically unknown and only i.i.d. observations $X_1, \dots, X_n \sim \mu$, defining the empirical measure $\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$, are available. A standard approach to estimate $T(\mu, \nu)$ in this setting is by means of the *empirical optimal transport cost* $T_c(\hat{\mu}_n, \nu)$, whose convergence to the population value for increasing n has been the subject of numerous works. Most research in this context, of which we can only give a selective overview, is devoted to the analysis of the *Wasserstein distance* (cf. Mallows 1972; Shorack and Wellner 1986; Villani 2008) where $\mathcal{X} = \mathcal{Y}$ and the cost c in (1) corresponds to the p-th power of a metric d on $\mathcal{X}$. More specifically, for $\mu, \nu \in \mathcal{P}(\mathcal{X})$, the p-Wasserstein distance for $p \geq 1$ is defined by

$$W_p(\mu, \nu) := \left(T_{d^p}(\mu, \nu)\right)^{1/p},$$

which is a metric on the space of probability measures on $(\mathcal{X}, d)$ with finite p-th moment.

A first fundamental contribution for the analysis of the empirical Wasserstein distance $W_p(\hat{\mu}_n, \mu)$ in case of $p = 1$ was made by Dudley 1969 via metric entropy bounds, asserting[5] $\mathbb{E}\left[W_1(\hat{\mu}_n, \mu)\right] \lesssim n^{-1/d}$ for compactly supported probability measures μ on $\mathbb{R}^d$ with $d \geq 3$. In particular, if μ is absolutely continuous with respect to the Lebesgue measure, this upper bound is tight. Under similar conditions, Dobrić and Yukich 1995 derived almost sure limits of $n^{1/d} W_1(\hat{\mu}_n, \hat{\mu}_n')$ through explicit matching arguments for two independent empirical measures $\hat{\mu}_n$ and $\hat{\mu}_n'$ of a common distribution μ. Extensions to $p > 1$ in Polish metric spaces were obtained by Boissard and Le Gouic 2014 relying on covering arguments of the underlying ground space. For probability measures on Euclidean spaces with possibly unbounded support, Dereich et al. 2013 and Fournier and Guillin 2015 derived upper bounds on the p-th moment $\mathbb{E}\left[W_p^p(\hat{\mu}_n, \mu)\right]$ under certain moment assumptions by explicitly constructing a couplings between $\hat{\mu}_n$ and μ. For a compactly supported probability measure μ on $\mathbb{R}^d$, their main result implies for $n \geq 1$ that

$$\mathbb{E}\left[W_p(\hat{\mu}_n, \mu)\right] \leq \mathbb{E}\left[W_p^p(\hat{\mu}_n, \mu)\right]^{1/p} \lesssim r_{p,d}(n) := \begin{cases} n^{-1/2p} & \text{if } d < 2p, \\ n^{-1/2p} \log(n)^{1/p} & \text{if } d = 2p, \\ n^{-1/d} & \text{if } d > 2p. \end{cases} \tag{2}$$

This bound is known to be tight in several settings, e.g., for $d < 2p$ when μ is discretely supported and for $d > 2p$ when $\mu = \mathrm{Unif}[0,1]^d$ is the uniform distribution on the unit cube. For $d = 2p$, the differences between $\hat{\mu}_n$ and μ at multiple scales culminate in the proof of the upper bound to an additional logarithmic factor, however, it remains open whether it is of correct order. For instance, contributions by Ajtai et al. 1984 and Talagrand 1994 show for $d = 2$ and $p \geq 1$ that $\mathbb{E}\left[W_p(\hat{\mu}_n, \mu)\right] \asymp n^{-1/2} \log(n)^{1/2}$ if $\mu = \mathrm{Unif}[0,1]^2$ (see also Bobkov and Ledoux 2021 for an alternative proof) which improves (2) for $p = 1$ by an additional $\log(n)^{1/2}$ factor. Notably, the bounds in (2) are known to delimit the accuracy of *any* estimator $\tilde{\mu}_n$ of μ with respect to the Wasserstein distance in the high-dimensional regime. More precisely, without additional assumptions on μ, the rates in (2) are (up to logarithmic factors) minimax

[5] Throughout this work, we write $a_n \lesssim b_n$ for two non-negative real-valued sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all $n \in \mathbb{N}$. If $a_n \lesssim b_n \lesssim a_n$, we write $a_n \asymp b_n$.

optimal (Singh and Póczos 2018), which demonstrates that the estimation of measures in the Wasserstein distance severely suffers from the *curse of dimensionality.*

To overcome this issue, there has been increased interest in structural properties of μ that allow for improved convergence rates. For probability measures on a compact Polish space, Weed and Bach 2019 derived tight bounds in terms of a notion of intrinsic dimension of μ (the upper and lower Wasserstein dimension). In particular, if μ is compactly supported on $\mathbb{R}^d$ with upper Wasserstein dimension $s > 2p$, they established that $\mathbb{E}\left[W_p(\hat{\mu}_n, \mu)\right] \lesssim n^{-1/s}$. Moreover, for uniformly distributed μ on a compact connected Riemannian manifold of dimension $d \geq 3$, Ledoux 2019 derived the bound $\mathbb{E}\left[W_p(\hat{\mu}_n, \mu)\right] \asymp n^{-1/d}$, effectively improving upon (2) if $3 \leq d \leq 2p$. Faster convergence rates can also be obtained under smoothness assumptions on Lebesgue absolutely continuous measures by taking suitable wavelet or kernel density estimators (Weed and Berthet 2019; Deb et al. 2021; Manole et al. 2021), which exploit the smoothness explicitly in contrast to the vanilla empirical OT cost. Under a high degree of smoothness, they approach the population measure in Wasserstein distance nearly with the parametric rate $n^{-1/2}$ (instead of $n^{-1/d}$), but come with additional computational challenges (Vacher et al. 2021).

So far, we only discussed the situation when $\hat{\mu}_n$ is compared to μ. From a statistical perspective, however, it is of similar interest to investigate $W_p(\hat{\mu}_n, \nu)$ for a different measure ν. We refer to Munk and Czado 1998 and Sommerfeld and Munk 2018 for various applications, such as testing for relevant differences and confidence intervals for W_p. One way to transfer the rates from $W_p(\hat{\mu}_n, \mu)$ to $W_p(\hat{\mu}_n, \nu)$ is by means of the triangle inequality,

$$\left|W_p(\hat{\mu}_n, \nu) - W_p(\mu, \nu)\right| \leq W_p(\hat{\mu}_n, \mu). \tag{3a}$$

Hence, all of the previous bounds on $W_p(\hat{\mu}_n, \mu)$ immediately imply the same upper bounds for the convergence rate of $W_p(\hat{\mu}_n, \nu)$ towards $W_p(\mu, \nu)$ when μ and ν are *distinct* measures on a common metric space. In the two-sample case, when ν is additionally estimated by $\hat{\nu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ based on an i.i.d. sample $Y_1, \dots, Y_n \sim \nu$, the triangle inequality yields

$$|W_p(\hat{\mu}_n, \hat{\nu}_n) - W_p(\mu, \nu)| \leq W_p(\hat{\mu}_n, \mu) + W_p(\hat{\nu}_n, \nu), \tag{3b}$$

which implies the same upper bounds as in (2) (as well as all improvements described above) for compactly supported μ, ν on $\mathbb{R}^d$. Therefore, with $r_{p,d}(n)$ as in (2),

$$\mathbb{E}\left[\left|W_p(\hat{\mu}_n, \hat{\nu}_n) - W_p(\mu, \nu)\right|\right] \lesssim r_{p,d}(n). \tag{4}$$

These upper bounds match the minimax rates (up to logarithmic factors) among all estimators of $W_p(\mu, \nu)$ when no additional assumptions are placed on the measures (Liang 2019; Niles-Weed and Rigollet 2019). In particular, this suggests that estimation of the Wasserstein distance between (potentially different) two measures is (without additional assumptions) statistically as difficult as estimation of the underlying measure with respect to Wasserstein loss.

However, crucial to the minimax optimality of (4) is the fact that μ and ν can be chosen to be arbitrarily close. In fact, in case $\mu \neq \nu$ are sufficiently separated, faster convergence rates may occur. Indeed, for compactly supported μ, ν on $\mathbb{R}^d$, Chizat et al. 2020 employed the dual formulation of the squared 2-Wasserstein distance (with a similar strategy as for the 1-Wasserstein distance by Sriperumbudur et al. 2012) to derive the bound

$$\mathbb{E}\left[\left|W_2^2(\hat{\mu}_n, \hat{\nu}_n) - W_2^2(\mu, \nu)\right|\right] \lesssim r_{2,d}^2(n). \tag{5a}$$

If $\mu \neq \nu$ with $W_2(\mu, \nu) \geq \delta > 0$, this implies squared convergence rates

$$\mathbb{E}\left[\left|W_2(\hat{\mu}_n, \hat{\nu}_n) - W_2(\mu, \nu)\right|\right] \lesssim r_{2,d}^2(n)/\delta \tag{5b}$$

when compared to (4). For $d \geq 5$, these upper bounds were recently generalized by Manole and Niles-Weed 2021 to arbitrary $p \geq 1$, asserting the convergence rate $n^{-\min(p,2)/d}$ for the empirical p-Wasserstein distance. They also provided analogous bounds under convex Hölder smooth costs and proved their sharpness for certain instances as well as minimax rate optimality up to logarithmic factors.

Inspired by these developments, this work is dedicated to a comprehensive understanding of the statistical performance of the empirical OT cost when the underlying probability measures are not only different but may additionally be supported on distinct spaces, for example if $\mathcal{X}$ and $\mathcal{Y}$ are submanifolds of $\mathbb{R}^d$ with (possibly) different dimension. This setting is practically relevant, since the concentration of observations from a high-dimensional ambient space on a low dimensional subspace is a commonly encountered phenomenon, reflected by the popularity of nonlinear dimensionality reduction techniques like manifold learning (see, e.g., Talwalkar et al. 2008; Zhu et al. 2018). Based on the upper bound in (3), one is inclined to believe that the convergence rate is determined by the slower rate, i.e., by the measure with *higher* intrinsic dimension. However, the pivotal (and maybe unexpected) finding of this work is that the convergence rate is actually determined by the measure with *lower* intrinsic dimension. In this sense, empirical OT naturally adapts to measures with distinct complexity in the most favorable way, and estimating the population value is statistically no harder than estimating the *simpler* one of the measures μ and ν. We refer to this phenomenon of OT as *lower complexity adaptation* (LCA).

> Example: Consider $\mathcal{Y} = [0,1]^{d_2}$ for $d_2 \geq 1$ and let $\mathcal{X} \subset \mathcal{Y}$ be a convex subset with dimension $d_1 \leq d_2$. In Section 2.3, we establish that the optimal transportation of any $\nu \in \mathcal{P}(\mathcal{Y})$ to any $\mu \in \mathcal{P}(\mathcal{X})$ under squared Euclidean costs can be decomposed into two motions (see Figure 1): first an orthogonal projection onto the linear space spanned by $\mathcal{X}$, and then an OT assignment within that linear space. Since such a projection is statistically negligible when compared to an OT assignment, it follows for $W_2(\mu, \nu) \geq \delta > 0$ by (5) that
>
> $$\mathbb{E}\left[\left|W_2(\hat{\mu}_n, \hat{\nu}_n) - W_2(\mu, \nu)\right|\right] \lesssim r_{2,d_1}^2(n)/\delta, \tag{6}$$
>
> which is independent of d_2, reflecting the LCA principle.

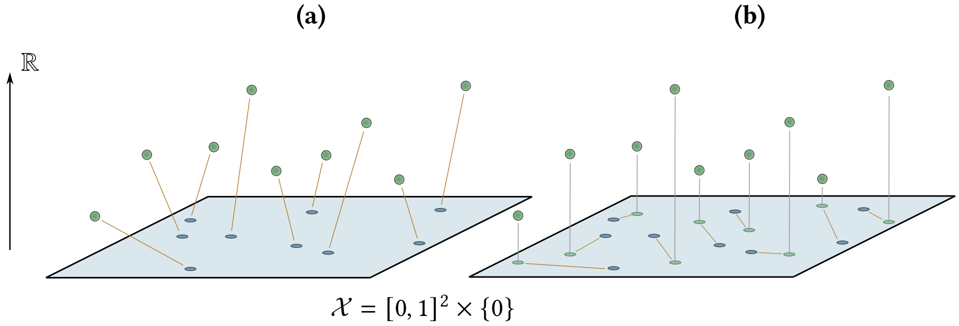


Figure 1: Optimal transport between two- and three-dimensional point clouds (blue and green) for squared Euclidean costs $\|x - y\|^2$. The optimal assignment in **(a)** is characterized by first projecting each point to the plane spanned by $\mathcal{X}$ before matching the data points, as depicted in **(b)**.

Our core contribution is to show that this phenomenon is a hallmark feature of empirical OT that far exceeds the scope of convex subsets and orthogonal projections. To formalize our main result, let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and consider a continuous bounded cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. In this setting, the OT cost enjoys a dual formulation (Villani 2008)

$$T_c(\mu,\nu) = \max_{f \in \mathcal{F}_c} \int_{\mathcal{X}} f \,\mathrm{d}\mu + \int_{\mathcal{Y}} f^c \mathrm{d}\nu,$$

where $\mathcal{F}_c$ is a suitable collection of uniformly bounded measurable functions on $\mathcal{X}$ (defined in Section 2.1) and $f^c(y) := \inf_{x\in\mathcal{X}} c(x,y) - f(x)$ denotes the c-transform of $f \in \mathcal{F}_c$. To investigate the empirical OT cost, we quantify the complexity of the class $\mathcal{F}_c$ and its c-transformed counterpart $\mathcal{F}_c^c = \{f^c \mid f \in \mathcal{F}_c\}$ in terms of their *uniform metric entropy*. The uniform metric entropy of a class $\mathcal{G}$ of real-valued functions on a set $\mathcal{Z}$ is defined as the logarithm of the covering number with respect to uniform norm $\|\cdot\|_\infty$, which is given for $\varepsilon > 0$ by

$$\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_\infty) := \inf\left\{ n \in \mathbb{N} \,\middle|\, \exists\, g_1, \dots, g_n\colon \mathcal{Z} \to \mathbb{R} \text{ with } \sup_{g\in\mathcal{G}} \min_{1\le i\le n} \|g - g_i\|_\infty \le \varepsilon \right\}.$$

A simple but crucial observation, which lies at the heart of this work, is that c-transformation with bounded costs is a Lipschitz operation under the uniform norm. Since $f^{cc} = f$ for all $f \in \mathcal{F}_c$, this in particular implies (Lemma 1)

$$\mathcal{N}(\varepsilon, \mathcal{F}_c^c, \|\cdot\|_\infty) = \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty). \tag{7}$$

This captures the LCA principle from the dual perspective: we only need to be able to control the complexity of either $\mathcal{F}_c$ or $\mathcal{F}_c^c$. Then, under the growth condition

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \lesssim \varepsilon^{-k}$$

for $\varepsilon > 0$ sufficiently small and a fixed $k > 0$, we prove for arbitrary $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ that any of the empirical estimators

$$\hat{T}_{c,n} \in \{T_c(\hat{\mu}_n, \nu), T_c(\mu, \hat{\nu}_n), T_c(\hat{\mu}_n, \hat{\nu}_n)\} \tag{8}$$

satisfies the upper bound (Theorem 1)

$$\mathbb{E}\left[\left|\hat{T}_{c,n} - T_c(\mu, \nu)\right|\right] \lesssim \begin{cases} n^{-1/2} & \text{if } k < 2, \\ n^{-1/2}\log(n) & \text{if } k = 2, \\ n^{-1/k} & \text{if } k > 2. \end{cases}$$

As we discuss in Section 3, suitable bounds for the uniform metric entropy of the function class $\mathcal{F}_c$ are typically determined by the space $\mathcal{X}$ and the regularity properties of the cost function. Since $\mathcal{X}$ can be chosen as the support of μ, these bounds can often be understood in terms of the intrinsic complexity of μ (or the one of ν, if it turns out to be lower). To explore the consequences of the LCA principle, we put special focus on the setting where the measure μ is supported on a space with *low intrinsic dimension* while ν may live on a general Polish space. For example, we obtain convergence rates for semi-discrete OT, where μ is supported on finitely many points only. In this setting, we find that the empirical estimator $\hat{T}_{c,n}$ always enjoys the parametric rate (Theorem 2)

$$\mathbb{E}\left[\left|\hat{T}_{c,n} - T_c(\mu, \nu)\right|\right] \lesssim n^{-1/2}.$$

This complements distributional limits for the one-sample estimator $\hat{T}_{c,n} = T_c(\hat{\mu}_n, \nu)$ by del Barrio et al. 2021. We also derive metric entropy bounds for $\mathcal{F}_c$ if $\mathcal{X}$ is given in terms of the image of sufficiently regular functions on sufficiently nice domains, where we exploit the Lipschitz continuity, semi-concavity, or Hölder continuity of the cost function to obtain novel theoretical guarantees for the convergence rate of $\hat{T}_{c,n}$ (Theorems 3, 4 and 5). For example, a special case of Theorem 4 states that the bound (6) for the 2-Wasserstein distance on $\mathbb{R}^d$ remains valid for compactly supported probability measures μ and ν if μ is concentrated on a d_1-dimensional $\mathcal{C}^2$ submanifold (Example 4(iii)).

In Section 4, we gather computational evidence for the LCA principle and present simulation results for various settings where the underlying measures have different intrinsic dimensions. In particular, we observe that the numerical findings are in line with the predictions of our theory. Section 5 concludes our work with a discussion and an outline of some open questions. Appendix A contains bounds on the uniform metric entropy of $\mathcal{F}_c$.

2 Lower Complexity Adaptation (LCA)

Throughout the manuscript, we work with absolutely bounded and continuous cost functions $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ on Polish spaces $\mathcal{X}$ and $\mathcal{Y}$. For convenience, we formulate our theory for costs whose range is restricted to the interval $[0,1]$. Since $T_{ac+b} = a \cdot T_c + b$ for any $a > 0$ and $b \in \mathbb{R}$, this is not a genuine restriction, and all of our results can easily be adapted to general costs that are absolutely bounded.

2.1 Duality and Complexity

In the following, we consider the dual formulation of the OT problem. This requires the notion of c-conjugacy under a given cost function $c : \mathcal{X} \times \mathcal{Y} \to [0,1]$ for non-empty sets $\mathcal{X}$ and $\mathcal{Y}$. The *c-transforms* of $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ are defined by

$$f^c(y) \coloneqq \inf_{x\in\mathcal{X}} c(x,y) - f(x) \qquad \text{and} \qquad g^c(x) \coloneqq \inf_{y\in\mathcal{Y}} c(x,y) - g(y). \tag{9}$$

A function $f : \mathcal{X} \to \mathbb{R}$ is called *c-concave* if there exists $g : \mathcal{Y} \to \mathbb{R}$ such that $f = g^c$. For Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a continuous cost function, any c-transform f^c or g^c is upper semi-continuous (as an infimum over continuous functions) and thus (Borel-)measurable. The following existence statement, which is tailored to bounded costs, shows that dual solutions of OT can always be assumed to be bounded and c-concave.

Theorem: Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and let $c : \mathcal{X}\times\mathcal{Y} \to [0,1]$ be continuous. Denote the class of feasible c-concave potentials by

$$\mathcal{F}_\mathrm{c} = \left\{ f : \mathcal{X} \to [-1,1] \,\middle|\, f \text{ is } c\text{-concave with } \|f^c\|_\infty \le 1 \right\}. \tag{10}$$

Then, for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, it holds that

$$T_\mathrm{c}(\mu,\nu) = \max_{f\in\mathcal{F}_\mathrm{c}} \int f \mathrm{d}\mu + \int f^c \mathrm{d}\nu. \tag{11}$$

Proof. Strong duality and the existence of maximizers in $\mathcal{F}_\mathrm{c}$ follow from Theorem 5.10(iii) in Villani 2008, where the bounds on $f \in \mathcal{F}_\mathrm{c}$ and f^c are detailed in step 4 of the proof. □

The properties of the feasible c-concave potentials $\mathcal{F}_\mathrm{c}$ strongly depend on the cost function c and the ground space $\mathcal{X}$. For example, if the family $\{c(\cdot,y) \,|\, y \in \mathcal{Y}\}$ of partially evaluated costs has a common modulus of continuity (with respect to a metric that metrizes $\mathcal{X}$), then all c-concave functions $f \in \mathcal{F}_\mathrm{c}$ are continuous with the same modulus. Hence, if $c(\cdot,y)$ is Lipschitz continuous uniform in $y \in \mathcal{Y}$, then each $f \in \mathcal{F}_\mathrm{c}$ is Lipschitz as well (Santambrogio 2015, Section 1.2). We denote the element-wise c-transform of the set $\mathcal{F}_\mathrm{c}$ by $\mathcal{F}_\mathrm{c}^c = \{f^c \,|\, f \in \mathcal{F}_\mathrm{c}\}$, which is by definition also uniformly bounded by one. A crucial observation is that the uniform metric entropy of the function class $\mathcal{F}_\mathrm{c}^c$ is bounded by the one of $\mathcal{F}_\mathrm{c}$, no matter how complex the space $\mathcal{Y}$ is or how badly the functions $c(x,\cdot)$ for fixed $x \in \mathcal{X}$ behave.

Lemma 1 (Complexity under c-transformation): Let $\mathcal{X}$ and $\mathcal{Y}$ be non-empty sets and $c : \mathcal{X} \times \mathcal{Y} \to [0,1]$. If $\mathcal{F}$ is a bounded function class on $\mathcal{X}$, it follows for $\varepsilon > 0$ that

$$\mathcal{N}(\varepsilon, \mathcal{F}^c, \|\cdot\|_\infty) \leq \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty).$$

Proof. If $N := \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) = \infty$, the claim is trivial, so assume $N < \infty$. Let $\{f_1, \dots, f_N\}$ be an ε-covering for $\mathcal{F}$ with respect to the uniform norm on $\mathcal{X}$. For $f \in \mathcal{F}$, consider f_i such that $\|f - f_i\|_\infty \leq \varepsilon$. Since f and c are both bounded, it follows that the c-transform f^c is bounded on $\mathcal{Y}$. For all $y \in \mathcal{Y}$, we obtain

$$f^c(y) = \inf_{x\in\mathcal{X}} c(x,y) - f(x) = \inf_{x\in\mathcal{X}} c(x,y) - f_i(x) + f_i(x) - f(x)$$
$$\begin{cases} \leq \inf_{x\in\mathcal{X}} c(x,y) - f_i(x) + \sup_{x\in\mathcal{X}} |f_i(x) - f(x)| \leq f_i^c(y) + \varepsilon \\ \geq \inf_{x\in\mathcal{X}} c(x,y) - f_i(x) - \sup_{x\in\mathcal{X}} |f_i(x) - f(x)| \geq f_i^c(y) - \varepsilon, \end{cases}$$

which implies $\sup_{y\in\mathcal{Y}} |f^c(y) - f_i^c(y)| \leq \varepsilon$. Thus, $\{f_1^c, \dots, f_N^c\}$ is an ε-covering of $\mathcal{F}^c$ with respect to the uniform norm. □

Since any feasible c-concave $f \in \mathcal{F}_c$ fulfills $f = f^{cc}$ Santambrogio 2015, Proposition 1.34, we conclude that the uniform metric entropies of the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are identical for any covering radius $\varepsilon > 0$, see (7). Hence, to control the complexity of both function classes simultaneously, it suffices to upper bound only one of them. In particular, in a concrete setting where different bounds for $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are available, their (ε-wise) minimum can be employed to derive an upper bound for the convergence rate in Theorem 1 below. Methods to control the uniform metric entropy of $\mathcal{F}_c$ and $\mathcal{F}_c^c$ typically exploit regularity properties of the underlying spaces $\mathcal{X}$ and $\mathcal{Y}$ as well as the ground cost c (cf. Section 3).

2.2 LCA: Dual Perspective

The observation that $\mathcal{F}_c$ and $\mathcal{F}_c^c$ have identical uniform metric entropies implies the following upper bound on the convergence of the empirical estimators $\hat{T}_{c,n}$ in (8), demonstrating the LCA principle. Notably, in the two-sample case $\hat{T}_{c,n} = T_c(\hat{\mu}_n, \hat{\nu}_n)$, the statement holds irregardless of the dependency structure between the empirical measures $\hat{\mu}_n$ and $\hat{\nu}_n$.

Theorem 1 (General LCA): Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and let $c : \mathcal{X} \times \mathcal{Y} \to [0,1]$ be continuous. Consider the function class $\mathcal{F}_c$ from (10) and assume that there exist $k > 0$, $K > 0$, and $\varepsilon_0 \in (0,1]$ such that the uniform metric entropy of $\mathcal{F}_c$ is bounded by

$$\log \mathcal{N}\big(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty\big) \leq K\varepsilon^{-k} \qquad \text{for } 0 < \varepsilon \leq \varepsilon_0. \tag{12}$$

Then, for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the empirical estimator $\hat{T}_{c,n}$ from (8) satisfies

$$\mathbb{E}\left[\left|\hat{T}_{c,n} - T_c(\mu,\nu)\right|\right] \lesssim \begin{cases} n^{-1/2} & \text{if } k < 2, \\ n^{-1/2}\log(n) & \text{if } k = 2, \\ n^{-1/k} & \text{if } k > 2, \end{cases} \tag{13}$$

where the implicit constant only depends on k, K, and ε_0.

When interpreting this statement, one should keep in mind that it is always possible to assume $\mathcal{X} = \operatorname{supp}(\mu)$ and $\mathcal{Y} = \operatorname{supp}(\nu)$ if this leads to improved bounds for the uniform metric entropy of $\mathcal{F}_\mathrm{c}$ (or equivalently $\mathcal{F}_\mathrm{c}^c$). In this sense, Theorem 1 can be tailored to take advantage of intrinsic properties of (the support of) μ (or ν). The proof employs arguments from empirical process theory and generalizes the technique of Sriperumbudur et al. 2012 and Chizat et al. 2020, where upper bounds for bounded convex sets $\mathcal{X} = \mathcal{Y} \subseteq \mathbb{R}^d$ and Euclidean costs $c(x,y) = \|x-y\|^p$ for $p = 1$ and $p = 2$ were derived.

Proof of Theorem 1. We first consider $\hat{T}_{c,n} = T_c(\hat{\mu}_n, \hat{\nu}_n)$. The definition of $\mathcal{F}_\mathrm{c}$ implies

$$\begin{aligned} T_c(\hat{\mu}_n,\hat{\nu}_n) - T_c(\mu,\nu) &= \max_{f\in\mathcal{F}_\mathrm{c}}\left(\int_\mathcal{X} f\mathrm{d}\hat{\mu}_n + \int_\mathcal{Y} f^c\mathrm{d}\hat{\nu}_n\right) - \max_{f\in\mathcal{F}_\mathrm{c}}\left(\int_\mathcal{X} f\mathrm{d}\mu + \int_\mathcal{Y} f^c\mathrm{d}\nu\right) \\ &\begin{cases} \leq \quad \sup_{f\in\mathcal{F}_\mathrm{c}} \int_\mathcal{X} f\,\mathrm{d}(\hat{\mu}_n - \mu) + \int_\mathcal{Y} f^c\mathrm{d}(\hat{\nu}_n - \nu), \\ \geq -\sup_{f\in\mathcal{F}_\mathrm{c}} \int_\mathcal{X} f\,\mathrm{d}(\mu - \hat{\mu}_n) + \int_\mathcal{Y} f^c\mathrm{d}(\nu - \hat{\nu}_n), \end{cases} \end{aligned}$$

and hence

$$|T_c(\hat{\mu}_n,\hat{\nu}_n) - T_c(\mu,\nu)| \leq \sup_{f\in\mathcal{F}_\mathrm{c}}\left|\int_\mathcal{X} f\,\mathrm{d}(\hat{\mu}_n - \mu)\right| + \sup_{f^c\in\mathcal{F}_\mathrm{c}^c}\left|\int_\mathcal{Y} f^c\mathrm{d}(\hat{\nu}_n - \nu)\right|. \tag{14}$$

Note by Lemma 1 that both $\mathcal{F}_\mathrm{c}$ and $\mathcal{F}_\mathrm{c}^c$ have finite uniform metric entropy for any $\varepsilon > 0$. Hence, both function classes contain subsets of at most countable cardinality that are dense in uniform norm, and the right hand side of (14) can thus be considered as a countable supremum and is therefore measurable. Further, recall that all elements in $\mathcal{F}_\mathrm{c}$ and $\mathcal{F}_\mathrm{c}^c$ are absolutely bounded by one. Taking the expectation and invoking symmetrization techniques (see Wainwright 2019, Proposition 4.11), we obtain

$$\begin{aligned} \mathbb{E}\left[\left|T_c(\hat{\mu}_n,\hat{\nu}_n) - T_c(\mu,\nu)\right|\right] &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}_\mathrm{c}}\left|\int_\mathcal{X} f\,\mathrm{d}(\hat{\mu}_n - \mu)\right|\right] + \mathbb{E}\left[\sup_{f^c\in\mathcal{F}_\mathrm{c}^c}\left|\int_\mathcal{Y} f^c\mathrm{d}(\hat{\nu}_n - \nu)\right|\right] \\ &\leq 2\big(\mathcal{R}_n(\mathcal{F}_\mathrm{c}) + \mathcal{R}_n(\mathcal{F}_\mathrm{c}^c)\big), \end{aligned}$$

where $\mathcal{R}_n(\mathcal{F}_\mathrm{c})$ and $\mathcal{R}_n(\mathcal{F}_\mathrm{c}^c)$ denote the Rademacher complexities of the function classes $\mathcal{F}_\mathrm{c}$ and $\mathcal{F}_\mathrm{c}^c$. The Rademacher complexity of $\mathcal{F}_\mathrm{c}$ is defined by

$$\mathcal{R}_n(\mathcal{F}_\mathrm{c}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}_\mathrm{c}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right|\right],$$

for i.i.d. $X_1, \ldots, X_n \sim \mu$ and independent i.i.d. Rademacher variables $\sigma_1, \ldots, \sigma_n \sim \mathrm{Unif}\{-1, 1\}$. This quantity is dominated by Dudley's entropy integral (see Luxburg and Bousquet 2004, Theorem 16),

$$\mathcal{R}_n(\mathcal{F}_\mathrm{c}) \leq \inf_{\delta \in [0,1]} \left(2\delta + \sqrt{32}\, n^{-1/2} \int_{\delta/4}^{1} \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}_\mathrm{c}, \|\cdot\|_\infty)}\, \mathrm{d}\varepsilon \right).$$

Let $\tilde{K} := \sqrt{32\, K}$. Since the covering number is a decreasing function in ε, a short calculation shows that the assumption $\log \mathcal{N}(\varepsilon, \mathcal{F}_\mathrm{c}, \|\cdot\|_\infty) \leq K\varepsilon^{-k}$ for $\varepsilon \leq \varepsilon_0$ implies that

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}_\mathrm{c}) &\leq \inf_{\delta \in [0,1]} \left(2\delta + \tilde{K} n^{-1/2} \int_{\delta/4}^{1} \min(\varepsilon, \varepsilon_0)^{-k/2} \mathrm{d}\varepsilon \right) \\ &\leq \begin{cases} \tilde{K} \left(\frac{\varepsilon_0^{1-k/2}}{1-k/2} + \frac{1-\varepsilon_0}{\varepsilon_0^{k/2}} \right) n^{-1/2} & \text{if } k < 2 \text{ for } \delta = 0, \\ \left(8 + \tilde{K} \log(\varepsilon_0) + \frac{\tilde{K}(1-\varepsilon_0)}{\varepsilon_0} \right) n^{-1/2} + \frac{\tilde{K}}{2} n^{-1/2} \log(n) & \text{if } k = 2 \text{ for } \delta = 4n^{-1/2}, \\ \tilde{K} \left(\frac{\varepsilon_0^{1-k/2}}{1-k/2} + \frac{1-\varepsilon_0}{\varepsilon_0^{k/2}} \right) n^{-1/2} + \left(8 + \frac{\tilde{K}}{k/2-1} \right) n^{-1/k} & \text{if } k > 2 \text{ for } \delta = 4n^{-1/k}, \end{cases} \end{aligned}$$

where we assume n large enough such that $\delta \leq \varepsilon_0$ for the respective choices. As $\mathcal{F}_\mathrm{c}$ and $\mathcal{F}_\mathrm{c}^c$ share the same covering number with respect to uniform norm (Lemma 1) and since all functions in $\mathcal{F}_\mathrm{c}^c$ are bounded in absolute value by one as well, it follows that $\mathcal{R}_n(\mathcal{F}_\mathrm{c}^c)$ can be bounded just like $\mathcal{R}_n(\mathcal{F}_\mathrm{c})$. Finally, for the other estimators $\hat{T}_{c,n}$ in (8), we obtain

$$\mathbb{E}\left[\left| T_c(\hat{\mu}_n, \nu) - T_c(\mu, \nu) \right| \right] \leq \mathcal{R}_n(\mathcal{F}_\mathrm{c}) \qquad \text{and} \qquad \mathbb{E}\left[\left| T_c(\mu, \hat{\nu}_n) - T_c(\mu, \nu) \right| \right] \leq \mathcal{R}_n(\mathcal{F}_\mathrm{c}^c)$$

in analogous fashion, which proves the bounds from (13) and finishes the proof. □

2.3 LCA: Primal Perspective

The proof of Theorem 1 relies on a technical observation about the nature of c-transforms and does not convey a geometric interpretation why empirical OT should follow the LCA principle. To provide some additional intuition, we next consider the LCA phenomenon from the primal perspective. Even though this approach yields less general results, its more explicit character has benefits, e.g., for establishing matching lower bounds.

Proposition 1 (Decomposition under additive costs): Let $\mathcal{X}$ be a Polish space and $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$ be the product of two Polish spaces. Let $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be continuous so that

$$c(x, y) = c_1(x, y_1) + c_2(y_2) \tag{15}$$

for all $x \in \mathcal{X}$ and $y = (y_1, y_2) \in \mathcal{Y}$ with continuous $c_1 : \mathcal{X} \times \mathcal{Y}_1 \to [0, 1]$ and $c_2 : \mathcal{Y}_2 \to [0, 1]$, and let $\mathfrak{p} : \mathcal{Y} \to \mathcal{Y}_1$ be the Cartesian projection to $\mathcal{Y}_1$. Then, for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$,

$$T_c(\mu, \nu) = T_{c_1}(\mu, \mathfrak{p}_\# \nu) + R_{c_2}(\nu), \tag{16}$$

where $R_{c_2}(\nu) := \int_{\mathcal{Y}} c_2(y_2)\, \mathrm{d}\nu(y_1, y_2)$.

For example, this statement can be applied in Euclidean spaces under l_p^p costs if $\mathcal{X} \subset \mathcal{Y}_1$. In this case, we find $c(x, y) = \|x - y_1\|_p^p + \|y_2\|_p^p$ (see also Figure 1 in the introduction), which satisifies condition (15). If $p = 2$, a relation of the form (16) clearly remains valid whenever $\mathcal{X}$ is contained in an affine linear subspace of $\mathcal{Y}$ and $\mathfrak{p}$ is the corresponding orthogonal projection.

Proof of Proposition 1. For any coupling $\pi \in \Pi(\mu, \nu)$, we consider the decomposition

$$\int_{\mathcal{X}\times\mathcal{Y}} c\, \mathrm{d}\pi = \int_{\mathcal{X}\times\mathcal{Y}} c_1\, \mathrm{d}\pi + \int_{\mathcal{X}\times\mathcal{Y}} c_2\, \mathrm{d}\pi.$$

The second term on the right is independent of π and equals $R_{c_2}(\nu) := \int_{\mathcal{Y}} c_2(y_2)\, \mathrm{d}\nu(y_1, y_2)$, while the first term can be rewritten in terms of $\tilde{\pi} = (\mathrm{id}, \mathfrak{p})_\# \pi \in \Pi(\mu, \mathfrak{p}_\# \nu)$ by a change of variables, such that $\int_{\mathcal{X}\times\mathcal{Y}} c_1\, \mathrm{d}\pi = \int_{\mathcal{X}\times\mathcal{Y}_1} c_1\, \mathrm{d}\tilde{\pi}$. Conversely, the gluing lemma (cf. Villani 2008, Chapter 1) implies that each $\tilde{\pi} \in \Pi(\mu, \mathfrak{p}_\# \nu)$ gives rise to a $\pi \in \Pi(\mu, \nu)$ for which $\int_{\mathcal{X}\times\mathcal{Y}} c_1\, \mathrm{d}\pi = \int_{\mathcal{X}\times\mathcal{Y}_1} c_1\, \mathrm{d}\tilde{\pi}$ holds as well. Therefore, we conclude that

$$\begin{aligned} T_c(\mu, \nu) &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c_1\, \mathrm{d}\pi + \int_{\mathcal{X}\times\mathcal{Y}} c_2\, \mathrm{d}\pi \\ &= \inf_{\tilde{\pi} \in \Pi(\mu, \mathfrak{p}_\# \nu)} \int_{\mathcal{X}\times\mathcal{Y}_1} c_1\, \mathrm{d}\tilde{\pi} + R_{c_2}(\nu) = T_{c_1}(\mu, \mathfrak{p}_\# \nu) + R_{c_2}(\nu). \end{aligned}$$

□

Since relation (16) in Proposition 1 holds for any pair of probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, we obtain

$$T_c(\hat{\mu}_n, \hat{\nu}_n) - T_c(\mu, \nu) = T_{c_1}(\hat{\mu}_n, \mathfrak{p}_\# \hat{\nu}_n) - T_{c_1}(\mu, \mathfrak{p}_\# \nu) + R_{c_2}(\hat{\nu}_n - \nu), \tag{17}$$

where $\hat{\nu}_n - \nu$ is understood as a signed measure. Thus, the statistical performance when estimating $T_c(\mu, \nu)$ via the empirical OT cost is governed by the (potentially much simpler) complexity of $T_{c_1}(\mu, \mathfrak{p}_\# \nu)$. By the reverse triangle inequality, it furthermore follows that

$$\mathbb{E}\left[|T_c(\hat{\mu}_n, \hat{\nu}_n) - T_c(\mu, \nu)|\right] \geq \mathbb{E}\left[\left||T_{c_1}(\hat{\mu}_n, \mathfrak{p}_\# \hat{\nu}_n) - T_{c_1}(\mu, \mathfrak{p}_\# \nu)| - |R_{c_2}(\hat{\nu}_n - \nu)|\right|\right], \tag{18}$$

where we note $\mathbb{E}\left[|R_{c_2}(\hat{\nu}_n - \nu)|\right] \asymp n^{-1/2}$ if $\sigma_{c_2}^2 := \mathrm{Var}_{Y\sim\nu}[c_2(Y)] > 0$. Let $\hat{f}_n \in \mathcal{F}_{c_1}$ denote an optimizer for (11) between $\hat{\mu}_n$ and $\mathfrak{p}_\# \nu$ under the cost function c_1. If the i.i.d. random variables $X_1, \dots, X_n \sim \mu$ and $Y_1, \dots, Y_n \sim \nu$ are independent, we find

$$\begin{aligned}
&\mathbb{E}\left[T_{c_1}(\hat\mu_n, \mathfrak{p}_\# \hat\nu_n) - T_{c_1}(\hat\mu_n, \mathfrak{p}_\# \nu)\right] \\
&\quad = \mathbb{E}\left[\max_{f \in \mathcal{F}_{c_1}} \left(\int_{\mathcal{X}} f \mathrm{d}\hat\mu_n + \int_{\mathcal{Y}_1} f^{c_1} \mathrm{d}\mathfrak{p}_\# \hat\nu_n\right) - \max_{f \in \mathcal{F}_{c_1}} \left(\int_{\mathcal{X}} f \mathrm{d}\hat\mu_n + \int_{\mathcal{Y}_1} f^{c_1} \mathrm{d}\mathfrak{p}_\# \nu\right)\right] \\
&\quad \geq \mathbb{E}_{X_1,\dots,X_n}\left[\mathbb{E}_{Y_1,\dots,Y_n}\left[\int_{\mathcal{Y}_1} \hat f_n^{c_1} \mathrm{d}(\mathfrak{p}_\# \hat\nu_n - \mathfrak{p}_\# \nu) \,\middle|\, X_1, \dots, X_n\right]\right] \\
&\quad = 0,
\end{aligned}$$

where independence is crucial for the final equality. In particular, this implies

$$\begin{aligned}
\mathbb{E}\left[|T_{c_1}(\hat\mu_n, \mathfrak{p}_\# \hat\nu_n) - T_{c_1}(\mu, \mathfrak{p}_\# \nu)|\right] &\geq \mathbb{E}\left[T_{c_1}(\hat\mu_n, \mathfrak{p}_\# \hat\nu_n) - T_{c_1}(\mu, \mathfrak{p}_\# \nu)\right] \\
&\geq \mathbb{E}\left[T_{c_1}(\hat\mu_n, \mathfrak{p}_\# \nu) - T_{c_1}(\mu, \mathfrak{p}_\# \nu)\right], \qquad (19)
\end{aligned}$$

which means that lower bounds for the two-sample setting can be obtained from lower bounds for the one-sample case. In Examples 3 and 5 of the subsequent section, we employ relation (18) and (19) to this end.

Remark 1 (Dependent empirical measures): Dependencies between the empirical measures $\hat\mu_n$ and $\hat\nu_n$ can lead to parametric convergence rates of order $n^{-1/2}$ irregardless of the underlying spaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$. For example, if $\mathcal{X} = \mathcal{Y}_1$ and $\mu = \mathfrak{p}_\# \nu$ with empirical measures related by $\hat\mu_n = \mathfrak{p}_\# \hat\nu_n$, it follows from (17) that

$$\mathbb{E}\left[\left|T_c(\hat\mu_n, \hat\nu_n) - T_c(\mu, \nu)\right|\right] = \mathbb{E}\left[\left|R_{c_2}(\hat\nu_n - \nu)\right|\right] \asymp n^{-1/2}$$

for any non-negative cost c_1 with $c_1(y_1, y_1) = 0$ for all $y_1 \in \mathcal{Y}_1$ if $\sigma_{c_2} > 0$.

3 Applications and Examples

The LCA principle, as formalized in Theorem 1, can readily be employed whenever suitable bounds on the uniform metric entropy of $\mathcal{F}_c$ are available. In the following, we consider a number of settings where well-known entropy bounds lead to novel results on the convergence rate of empirical OT. In order to efficiently exploit the properties of the space $\mathcal{X}$ in settings of low intrinsic dimensionality, the following observation is useful.

Lemma 2 (Union bound): Let $\mathcal{F}$ be a class of real valued functions on a set $\mathcal{X} = \bigcup_{i=1}^I \mathcal{X}_i$ for (not necessarily disjoint) subsets $\mathcal{X}_i \subset \mathcal{X}$ and $I \in \mathbb{N}$, and let $\mathcal{F}|_{\mathcal{X}_i} := \left\{f|_{\mathcal{X}_i} : \mathcal{X}_i \to \mathbb{R} \,|\, f \in \mathcal{F}\right\}$ be the class of functions restricted to $\mathcal{X}_i$ for all $i \in \{1, \dots, I\}$. Then, for each $\varepsilon > 0$,

$$\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \sum_{i=1}^I \log \mathcal{N}(\varepsilon, \mathcal{F}|_{\mathcal{X}_i}, \|\cdot\|_\infty).$$

Proof. Suppose that the right hand side is finite, otherwise the bound is trivial. Denote by $\mathcal{F}_i$ a minimal ε-covering of $\mathcal{F}|_{\mathcal{X}_i}$ with respect to the uniform norm. Let $\tilde{\mathcal{X}}_1 := \mathcal{X}_1$ and define $\tilde{\mathcal{X}}_i := \mathcal{X}_i \setminus \bigcup_{j=1}^{i-1} \mathcal{X}_j$ for $i \geq 2$. Then $\left\{ \sum_{i=1}^{I} f_i \cdot \mathbb{1}_{\tilde{\mathcal{X}}_i} \mid f_i \in \mathcal{F}_i \right\}$ is an ε-covering of $\mathcal{F}$ under the uniform norm with cardinality at most $\prod_{i=1}^{I} \mathcal{N}(\varepsilon, \mathcal{F}|_{\mathcal{X}_i}, \|\cdot\|_\infty)$. □

3.1 Semi-Discrete Optimal Transport

We first address the setting of *semi-discrete* OT, where $\mathcal{X} = \{x_1, \dots, x_I\}$ is a finite discrete space with $I \in \mathbb{N}$ elements. Structural and computational properties of semi-discrete OT have been investigated extensively, especially in Euclidean contexts (see, e.g., Aurenhammer et al. 1998; Mérigot 2011; Geiß et al. 2013; Hartmann and Schuhmacher 2020). For our purposes, consider a general Polish space $\mathcal{Y}$ and a continuous cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. By definition (10), the function class $\mathcal{F}_c$ is absolutely bounded by one and we find that

$$\mathcal{N}(\varepsilon, \mathcal{F}_c|_{\{x_i\}}, \|\cdot\|_\infty) \leq \lceil 1/\varepsilon \rceil \tag{20}$$

for any $\varepsilon > 0$ and $i \in \{1, \dots, I\}$. Hence, Lemma 2 implies that $\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \leq I \log \lceil 1/\varepsilon \rceil \lesssim I/\varepsilon$ and we can apply Theorem 1 to derive the following bound.

Theorem 2 (Semi-discrete LCA): Let $\mathcal{X} = \{x_1, \dots, x_I\}$ be a finite discrete space, $\mathcal{Y}$ a Polish space, and $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ continuous. Then, for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the empirical estimator $\hat{T}_{c,n}$ from (8) satisfies

$$\mathbb{E}\left[\left| \hat{T}_{c,n} - T_c(\mu, \nu) \right| \right] \lesssim n^{-1/2}. \tag{21}$$

This result is in line with recent findings by del Barrio et al. 2021, who derive a central limit theorem for the empirical semi-discrete OT cost. Their result allows for possibly unbounded costs, but it is limited to the one-sample estimator $\hat{T}_{c,n} = T_c(\hat{\mu}_n, \nu)$. According to Markov's inequality, Theorem 2 implies that the sequence of random variables $\sqrt{n}\big(\hat{T}_{c,n} - T_c(\mu, \nu)\big)$ is tight even for $\hat{T}_{c,n} = T_c(\mu, \hat{\nu}_n)$ or $T_c(\hat{\mu}_n, \hat{\nu}_n)$, which indicates that it might also be possible to derive limit distributions when the measure on the general space $\mathcal{Y}$ is estimated empirically. Moreover, Theorem 2 asserts novel bounds for the Wasserstein distance when one measure is supported on finitely many points only while the support of the other measure is bounded.

3.2 Optimal Transport under Lipschitz costs

Semi-discrete OT can be regarded as a special OT setting where one probability measure has intrinsic dimension zero. We now broaden this perspective to higher dimensions and consider parameterized spaces and surfaces. The effective dimension is then governed by the (possibly low-dimensional) domain of the parameterization, and not by the (possibly high-dimensional) ambient space. In this section, we work with an additional Lipschitz requirement for the cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ and impose the following condition on the Polish space $\mathcal{X}$. By rescaling, we may assume that the Lipschitz constant is equal to one.

Assumption (Lip): Suppose $\mathcal{X} = \bigcup_{i=1}^{I} g_i(\mathcal{U}_i)$ for $I \in \mathbb{N}$ connected metric spaces $(\mathcal{U}_i, d_i)$ and maps $g_i : \mathcal{U}_i \to \mathcal{X}$ so that $c(g_i(\cdot), y)$ is 1-Lipschitz with respect to d_i for all $y \in \mathcal{Y}$.

This setting captures a broad notion of generalized surfaces in an ambient space. Since the mappings g_i are not required to be injective, self-intersections are possible. Moreover, exploiting the (not necessarily disjoint) decomposition $\mathcal{X} = \bigcup_{i=1}^{I} \mathcal{X}_i$ with $\mathcal{X}_i := g_i(\mathcal{U}_i)$ for $i \in \{1, \dots, I\}$, it suffices by Lemma 2 to control the complexity of $\mathcal{F}_\mathrm{c}|_{\mathcal{X}_i}$ to bound the metric entropy of $\mathcal{F}_\mathrm{c}$. For this purpose, we note that the Lipschitz continuity of $c(g_i(\cdot), y)$ implies that $f \circ g_i$ for any c-concave potential $f \in \mathcal{F}_\mathrm{c}$ is Lipschitz as well. This relates the restricted function class $\mathcal{F}_\mathrm{c}|_{\mathcal{X}_i}$ to the class of bounded Lipschitz functions on $\mathcal{U}_i$. For the latter, metric entropy bounds in terms of the covering number of $\mathcal{U}_i$ are available (Kolmogorov and Tikhomirov 1961, Section 9). For $\varepsilon > 0$, the covering number of a metric space $(\mathcal{U}, d)$ is defined by

$$\mathcal{N}(\varepsilon, \mathcal{U}, d) := \inf \left\{ n \in \mathbb{N} \,\middle|\, \exists\, U_1, \dots, U_n \subseteq \mathcal{U} \text{ with } \operatorname{diam}(U_k) \le 2\varepsilon \text{ and } \mathcal{U} = \bigcup_{k=1}^{n} U_k \right\},$$

where $\operatorname{diam}(U) := \sup_{u,v \in U} d(u, v)$ denotes the diameter of a subset $U \subseteq \mathcal{U}$.

Theorem 3 (Lipschitz LCA): Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and let $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be continuous. If Assumption (Lip) holds and there exists $k > 0$ so that for all $i \in \{1, \dots, I\}$

$$\mathcal{N}(\varepsilon, \mathcal{U}_i, d_i) \lesssim \varepsilon^{-k} \qquad \text{for } \varepsilon > 0 \text{ sufficiently small,}$$

then, for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the empirical estimator $\hat{T}_{c,n}$ from (8) satisfies

$$\mathbb{E}\left[\left| \hat{T}_{c,n} - T_c(\mu, \nu) \right| \right] \lesssim \begin{cases} n^{-1/2} & \text{if } k < 2, \\ n^{-1/2} \log(n) & \text{if } k = 2, \\ n^{-1/k} & \text{if } k > 2. \end{cases} \tag{22}$$

Proof. Lemma 4 in Appendix A shows that the uniform metric entropy in this setting is bounded by $\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \lesssim \varepsilon^{-k}$. Applying Theorem 1 yields bound (22). □

Remark 2 (Disconnected domains): If the metric spaces $\mathcal{U}_1, \dots, \mathcal{U}_I$ in Assumption (Lip) consist of finitely many connected components, Theorem 3 remains valid at the price of a possibly larger constant. Moreover, if some $\mathcal{U}_i$ has infinitely many components, Lemma 4 ensures

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \lesssim \varepsilon^{-k} \log(\varepsilon^{-1}) \lesssim \varepsilon^{-k-\delta}$$

for any $\delta > 0$ (where the implicit constant depends on δ). Hence, Theorem 1 shows that bound (22) still holds when k is replaced by $k + \delta$.

We emphasize once more that no additional assumptions on the complexity of the Polish space $\mathcal{Y}$ are necessary. To highlight applications and noteworthy consequences of Theorem 3, we consider a number of examples.

Example 1 (Metric spaces with Lipschitz costs): Let $\mathcal{X}$ and $\mathcal{Y}$ be closed subsets of a Polish metric space $(\mathcal{Z}, d)$ and consider costs $c : \mathcal{Z}^2 \longrightarrow \mathbb{R}$ that are continuous and absolutely bounded on $\mathcal{X} \times \mathcal{Y}$. Furthermore, assume that $c(\cdot, y)$ is Lipschitz on $\mathcal{X}$ uniformly in $y \in \mathcal{Y}$, which, for example, holds for $c(x, y) = d^p(x, y)$ and $p \geq 1$ if $\mathcal{Y}$ is bounded. Then, Theorem 3 provides convergence rates whenever $\mathcal{N}(\varepsilon, \mathcal{X}, d) \lesssim \varepsilon^{-k}$. This condition holds if the *upper Minkowski-Bouligand dimension* of $\mathcal{X}$ (Mattila 1995, Section 5.3), defined by

$$\overline{\dim}_M(\mathcal{X}) := \limsup_{\varepsilon \searrow 0} \frac{\log \mathcal{N}(\varepsilon, \mathcal{X}, d)}{\log(\varepsilon^{-1})},$$

is strictly dominated by k. Note that non-integral values of $\overline{\dim}_M(\mathcal{X})$ are possible and that $\mathcal{X}$ does not necessarily have to be connected (Remark 2).

Example 2 (Euclidean spaces with locally Lipschitz costs): Consider a locally Lipschitz cost function $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ for $d \in \mathbb{N}$. This setting entails the choice $c(x, y) = \|x - y\|^p$ for the Euclidean norm $\|\cdot\|$, or the l_p^p-costs $c(x, y) = \sum_{i=1}^d |x_i - y_i|^p$, both of which are popular options for the Wasserstein distance of order $p \geq 1$. In the following examples, we implicitly assume that $\mathcal{X}$ and $\mathcal{Y}$ are Polish subsets of $\mathbb{R}^d$.

(i) *Bounded sets*: If $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ are bounded sets, then $c(\cdot, y)$ is Lipschitz continuous on $\mathcal{X}$ uniformly in $y \in \mathcal{Y}$ and $\mathcal{N}(\varepsilon, \mathcal{X}, \|\cdot\|) \lesssim \varepsilon^{-d}$. Since one can always enlarge $\mathcal{X}$ and $\mathcal{Y}$ to connected compact sets, Theorem 3 asserts that the convergence rate of the empirical OT cost is dominated by (22) with $k = d$.

(ii) *Surfaces*: Improved bounds can be derived if $\mathcal{X} = \bigcup_{i=1}^I g_i(\mathcal{U}_i)$ for Lipschitz maps $g_i : \mathcal{U}_i \longrightarrow \mathbb{R}^d$ on bounded and connected sets $\mathcal{U}_i \subseteq \mathbb{R}^s$ with $d > s \in \mathbb{N}$. In this setting, the partially evaluated cost $c\big(g_i(\cdot), y\big)$ is Lipschitz on $\mathcal{X}$ uniformly in $y \in \mathcal{Y}$ for bounded sets $\mathcal{Y}$, and it holds that $\mathcal{N}(\varepsilon, \mathcal{U}_i, \|\cdot\|) \lesssim \varepsilon^{-s}$. Hence, bound (22) is valid for $k = s$.

(iii) *Manifolds*: The examples outlined in (ii) include compact s-dimensional immersed $\mathcal{C}^1$ submanifolds of $\mathbb{R}^d$, possibly with boundary (Lee 2013). Due to compactness, a finite atlas of charts with connected and bounded codomains $\mathcal{U}_i$ can always be found.

We next show that the upper bounds in Theorem 3 can be complemented by matching lower bounds in settings where the primal approach to the LCA principle (see Proposition 1) is available.

Example 3 (Lower bounds under Lipschitz costs): Let $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$ be the product of two connected Polish spaces and let $\mathcal{X} = \mathcal{Y}_1$. Consider the costs $c(x, y) = d_1(x, y_1) + c_2(y_2)$, where d_1 is a Polish metric on $\mathcal{Y}_1$ such that $0 < \operatorname{diam}(\mathcal{Y}_1) \leq 1$ and $c_2 : \mathcal{Y}_2 \to [0, 1]$ is continuous. Then $c(\cdot, y)$ is Lipschitz with respect to d_1 uniformly in $y \in \mathcal{Y}$, so Assumption (Lip) is fulfilled. If $\mathcal{N}(\varepsilon, \mathcal{X}, d_1) \asymp \varepsilon^{-k}$ for some $k > 0$ as $\varepsilon \searrow 0$ and if the empirical measures $\hat{\mu}_n$ and $\hat{\nu}_n$ are independent, one can show that

$$\sup_{\mu,\nu} \mathbb{E}\left[|T_c(\hat{\mu}_n, \hat{\nu}_n) - T_c(\mu, \nu)|\right] \gtrsim \begin{cases} n^{-1/2} & \text{if } k \leq 2, \\ n^{-1/k} & \text{if } k > 2, \end{cases} \tag{23}$$

where the supremum is taken over $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. For $k \leq 2$, inequality (23) follows by selecting suitable discrete measures μ and ν (see Sommerfeld et al. 2019, Section 5). For $k > 2$, inequality (23) follows by (18) and (19) in conjunction with the minimax lower bounds for the 1-Wasserstein distance $W_1(\mu, \nu) = T_{d_1}(\mu, \nu)$ by Singh and Póczos 2018, Theorem 2

$$\sup_{\mu,\nu} \mathbb{E}\left[|T_{d_1}(\hat{\mu}_n, \mathfrak{p}_{\#}\nu) - T_{d_1}(\mu, \mathfrak{p}_{\#}\nu)|\right] \geq \sup_{\mu} \mathbb{E}\left[T_{d_1}(\hat{\mu}_n, \mu)\right] \gtrsim n^{-1/k}.$$

Hence, the upper bounds from (22) are sharp for $k \neq 2$ and sharp up to logarithmic terms in case of $k = 2$.

3.3 Optimal Transport under Semi-Concave Costs

It is known that better convergence rates than those offered by Theorem 3 can be obtained on Euclidean spaces for more regular cost functions (Manole and Niles-Weed 2021). In this section, we consider improvements for semi-concave costs (before we turn to Hölder smooth costs in Section 3.4). A function $f : U \to \mathbb{R}$ on a bounded, convex domain $U \subset \mathbb{R}^d$ for $d \in \mathbb{N}$ is called Λ-*semi-concave* with *modulus* $\Lambda \geq 0$ if the map

$$u \mapsto f(u) - \Lambda \|u\|^2$$

is concave, where $\|\cdot\|$ denotes the Euclidean norm (cf. Gangbo and McCann 1996). The uniform metric entropy of the class of bounded, Lipschitz, and semi-concave functions on U is of order $\varepsilon^{-d/2}$ (Bronshtein 1976; Guntuboyina and Sen 2013), compared to ε^{-d} without semi-concavity. Since boundedness, Lipschitz continuity, and semi-concavity are all inherited to $\mathcal{F}_c$ by the cost function, we impose the following conditions. For convenience, we assume the Lipschitz constant and the modulus of semi-concavity to be equal to one, since other constants can be accommodated by scaling the cost function.

Assumption (SC): Suppose $\mathcal{X} = \bigcup_{i=1}^{I} g_i(\mathcal{U}_i)$ for $I \in \mathbb{N}$ bounded, convex subsets $\mathcal{U}_i \subseteq \mathbb{R}^d$ and maps $g_i : \mathcal{U}_i \to \mathcal{X}$ so that $c(g_i(\cdot), y)$ is 1-Lipschitz and 1-semi-concave for all $y \in \mathcal{Y}$.

Similar to Assumption (Lip) in the previous section, Assumption (SC) enables the application of Lemma 2 to the union $\mathcal{X} = \bigcup_{i=1}^{I} \mathcal{X}_i$ with $\mathcal{X}_i = g_i(\mathcal{U}_i)$ for all $i \in \{1, \dots, I\}$. Combined with the uniform metric entropy bounds by Bronshtein 1976 and Guntuboyina and Sen 2013, this leads to the following result.

Theorem 4 (Semi-concave LCA): Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be continuous. If Assumption (SC) holds, then, for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the empirical estimator $\hat{T}_{c,n}$ from (8) satisfies

$$\mathbb{E}\left[\left|\hat{T}_{c,n} - T_c(\mu, \nu)\right|\right] \lesssim \begin{cases} n^{-1/2} & \text{if } d \leq 3, \\ n^{-1/2} \log(n) & \text{if } d = 4, \\ n^{-2/d} & \text{if } d \geq 5. \end{cases} \tag{24}$$

Proof. Lemma 5 in Appendix A shows that the uniform metric entropy in this setting is bounded by $\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \lesssim \varepsilon^{-d/2}$. This implies bound (24) via Theorem 1. □

Compared to Theorem 3, which would guarantee a convergence rate of $n^{-1/d}$ for $d \geq 5$ under Assumption (SC), Theorem 4 provides the considerably faster rate $n^{-2/d}$. To explore applications, we revisit the settings discussed in Example 2 and show how they fare under the stronger Assumption (SC). For this purpose, note that semi-concavity of a $\mathcal{C}^2$ function f on a convex domain is implied by the boundedness of the Eigenvalues of its Hessian. Indeed, if the Eigenvalues are bounded from above by $2\Lambda \geq 0$, then $u \mapsto f(u) - \Lambda \|u\|^2$ has a negative semi-definite Hessian and is therefore concave (see Luenberger 2003, Section 6.4, Proposition 5).

Example 4 (Euclidean spaces with $\mathcal{C}^2$ costs): Extending Example 2, consider a twice continuously differentiable cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for $d \in \mathbb{N}$. This setting includes $c(x, y) = \|x - y\|^p$ as well as $c(x, y) = \sum_{i=1}^{d} |x_i - y_i|^p$ for $p \geq 2$. We implicitly assume that $\mathcal{X}$ and $\mathcal{Y}$ are Polish subsets of $\mathbb{R}^d$ in the following.

(*i*) *Bounded sets*: Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ be bounded sets. Since $\mathcal{X}$ and $\mathcal{Y}$ can always be enlarged to convex compact sets, the functions $c(\cdot, y)$ are Lipschitz continuous and semi-concave on $\mathcal{X}$ uniformly in $y \in \mathcal{Y}$. Hence, Theorem 4 provides the upper bounds (24).

(*ii*) *Surfaces*: Improved bounds can be obtained if $\mathcal{X} = \bigcup_{i=1}^{I} g_i(\mathcal{U}_i)$ for $\mathcal{C}^2$ maps $g_i : \mathcal{U}_i \to \mathbb{R}^d$ with bounded second derivatives on open, bounded, and convex sets $\mathcal{U}_i \subset \mathbb{R}^s$ for $s < d$. In this setting, the partially evaluated cost $c(g_i(\cdot), y)$ is Lipschitz and semi-concave on $\mathcal{U}_i$ uniformly in $y \in \mathcal{Y}$ for bounded $\mathcal{Y}$. Hence, bound (24) holds with d replaced by s.

(*iii*) *Manifolds*: The setting described in (*ii*) includes compact s-dimensional immersed $\mathcal{C}^2$ submanifolds of $\mathbb{R}^d$ (Lee 2013). Compactness ensures the existence of an atlas with finitely many charts such that all involved maps have bounded second derivative and all co-domains of charts are convex and bounded.

We continue with a setting in which lower bounds for the empirical OT cost that match the upper bounds in Theorem 4 can be derived via Proposition 1.

Example 5 (Lower bounds under semi-concave costs): Let $1 \leq d_1 \leq d_2$ and consider the unit cubes $\mathcal{X} = [0,1]^{d_1}$ and $\mathcal{Y} = [0,1]^{d_2}$. If $\mathcal{X}$ is embedded into $\mathcal{Y}$ along the first d_1 coordinates, the squared Euclidean cost function amounts to

$$c(x,y) = \|x - y_1\|^2 + \|y_2\|^2 =: c_1(x,y_1) + c_2(y_2),$$

where $y = (y_1, y_2) \in [0,1]^{d_1+(d_2-d_1)}$. Up to scaling of the cost function, this setting satisfies Assumption (SC). For independent empirical measures $\hat{\mu}_n$ and $\hat{\nu}_n$, one can show that

$$\sup_{\mu,\nu} \mathbb{E}\left[|T_c(\hat{\mu}_n,\hat{\nu}_n) - T_c(\mu,\nu)|\right] \gtrsim \begin{cases} n^{-1/2} & \text{if } d_1 \leq 4, \\ n^{-2/d_1} & \text{if } d_1 \geq 5, \end{cases} \tag{25}$$

where $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ in the supremum. For $d_1 \leq 4$, this lower bound follows by selecting suitable discrete measures (see Sommerfeld et al. 2019, Section 5), and for $d_1 \geq 5$, it follows from inequality (18) in conjunction with Proposition 21 in Manole and Niles-Weed 2021 (which is also applicable for more general strictly convex cost functions). Overall, the upper bounds in (24) match the lower bounds (25) in case $d_1 \neq 4$ and are sharp up to logarithmic factors for $d_1 = 4$.

We want to highlight that the fast rates of Theorem 4 (compared to Theorem 3) can often be expected in the settings of Example 4 even if the cost function is not $\mathcal{C}^2$ on all of $\mathbb{R}^{2d}$. For example, if the set

$$\mathcal{D} := \left\{(x,y) \,\middle|\, c \text{ is not } \mathcal{C}^2 \text{ at } (x,y)\right\} \subset \mathbb{R}^{2d}$$

is strictly separated from $\Sigma := \mathrm{supp}(\mu) \times \mathrm{supp}(\nu) \subset \mathbb{R}^{2d}$, one can extend the restriction $c|_\Sigma$ to a $\mathcal{C}^2$ cost function on all of $\mathbb{R}^{2d}$ (by the extension theorem of Whitney 1934) without altering $\hat{T}_{c,n}$ or $T_c(\mu,\nu)$. If $c(x,y) = \|x-y\|^p$ for any $0 < p < 2$, we find $\mathcal{D} = \{(x,x) \mid x \in \mathbb{R}^d\}$, implying the fast convergence rates in (24) whenever the supports of μ and ν are strictly separated. Similar observations were pointed out by Manole and Niles-Weed 2021, Corollary 3(ii) under additional convexity assumptions. In contrast, considering the l_p^p-cost function $c(x,y) = \sum_{i=1}^d |x_i - y_i|^p$ for $0 < p < 2$, the set

$$\mathcal{D} = \left\{(x,y) \,\middle|\, x_i = y_i \text{ for some } i \in \{1,\ldots,d\}\right\} \subset \mathbb{R}^{2d} \tag{26}$$

is notably larger. Therefore, the p-Wasserstein distance based on the Euclidean norm may exhibit faster convergence rates than its l_p counterpart, e.g., if μ is a translation of ν along a coordinate axis.

We conjecture that the faster rates implied by Theorem 4 will occur under even more general circumstances. For example, in some settings it might suffice that $\mathcal{D}$ is negligible under the actual optimal transport, meaning that $\pi(\mathcal{D}) = 0$ for an OT plan between μ and ν. A proof of this claim, however, would likely require quantitative statements on the regularity of the cost function along the support of empirical OT plans and lies beyond the scope of this work. Still, we pick up on this hypothesis and observe some numerical evidence in Section 4.

3.4 Optimal Transport under Hölder Costs

In Section 3.2 and 3.3, we have shown that the rate of convergence of the empirical OT cost in $\mathbb{R}^d$ is bounded by $n^{-1/d}$ for Lipschitz and $n^{-2/d}$ for $\mathcal{C}^2$ costs (if $d \geq 5$). The recent work of Manole and Niles-Weed 2021 demonstrated that these results can be generalized to α-Hölder smooth costs for $0 < \alpha \leq 2$, deriving the rates $n^{-\alpha/d}$. In this section, we employ similar arguments to bound the uniform metric entropy of the class $\mathcal{F}_c$ by $\varepsilon^{-d/\alpha}$ (for any $d \in \mathbb{N}$ and $0 < \alpha \leq 2$) in settings resembling the ones of Assumption (Lip) and (SC).

We say that a function $f \colon U \to \mathbb{R}$ on a convex domain $U \subset \mathbb{R}^d$ is (α, Λ)-Hölder smooth for $0 < \alpha \leq 1$ and $\Lambda > 0$ if $\|f\|_\infty < \Lambda$ and

$$|f(x) - f(y)| \leq \Lambda \cdot \|x - y\|^\alpha.$$

Moreover, we say that f is (α, Λ)-Hölder smooth for $1 < \alpha \leq 2$ if $\|f\|_\infty < \Lambda$ and f is differentiable with $(\alpha - 1, \Lambda)$-Hölder smooth partial derivatives. If the convex domain U in this definition is not open, we assume the existence of a Hölder smooth function on an open subset of $\mathbb{R}^d$ containing U that coincides with f on U.

Assumption (Hol) : Let $\alpha \in (0, 2]$ and suppose that $\mathcal{X} = \bigcup_{i=1}^I g_i(\mathcal{U}_i)$ for $I \in \mathbb{N}$ compact, convex subsets $\mathcal{U}_i \subset \mathbb{R}^d$ and maps $g_i : \mathcal{U}_i \to \mathcal{X}$ so that $c(g_i(\cdot), y)$ is $(\alpha, 1)$-Hölder for all $y \in \mathcal{Y}$.

Theorem 5 (Hölder LCA): Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be continuous. If Assumption (Hol) holds, then, for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the empirical estimator $\hat{T}_{c,n}$ from (8) satisfies

$$\mathbb{E}\left[\left|\hat{T}_{c,n} - T_c(\mu, \nu)\right|\right] \lesssim \begin{cases} n^{-1/2} & \text{if } d < 2\alpha, \\ n^{-1/2} \log(n) & \text{if } d = 2\alpha, \\ n^{-\alpha/d} & \text{if } d > 2\alpha. \end{cases} \tag{27}$$

Proof. Lemma 6 in Appendix A shows that the uniform metric entropy in this setting is bounded by $\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \lesssim \varepsilon^{-d/\alpha}$. An application of Theorem 1 yields the claim. □

Theorem 5 can be used to derive upper bounds for Hölder smooth cost functions on Euclidean spaces analogous to Example 4. Moreover, it is again possible to derive lower bounds in certain situations. For instance, in the setting of Example 5, we can consider α-Hölder costs of the form $c(x,y) = \sum_{i=1}^{d_1} |x_i - y_{1i}|^\alpha + \sum_{i=d_1+1}^{d_2} |y_{2i}|^\alpha$ for $\alpha \in (0,2]$. Then, Assumption (Hol) is fulfilled (for $d = d_1$) and one can show that

$$\sup_{\mu,\nu} \mathbb{E}\left[|T_c(\hat{\mu}_n, \hat{\nu}_n) - T_c(\mu,\nu)|\right] \gtrsim \begin{cases} n^{-1/2} & \text{if } d_1 \leq 2\alpha, \\ n^{-\alpha/d_1} & \text{if } d_1 > 2\alpha, \end{cases}$$

where the supremum is taken over $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. Herein, the lower bound for $d_1 \leq 2\alpha$ follows by selecting discretely supported measures, whereas the regime $d_1 > 2\alpha$ is covered by Proposition 21 in Manole and Niles-Weed 2021. In particular, in case of $d_1 \neq 2\alpha$, the upper bound from Theorem 5 matches the lower bound, whereas for $d_1 = 2\alpha$ it is sharp only up to a logarithmic factor.

4 Simulations

In the previous sections, we investigated the convergence rate for the empirical OT cost in various settings, stressing than an intrinsic adaptation to the less complex measure governs asymptotic statistical properties. We now turn to the question if these asymptotic properties can already be observed in the finite sample regime accessible to numerical analysis. For this purpose, we fix probability measures μ and ν and approximate the mean absolute deviation

$$\Delta_n = \mathbb{E}\left[\left|\hat{T}_{c,n} - T_c(\mu,\nu)\right|\right]$$

for various values of n via Monte-Carlo simulations with 2000 independent repetitions.[6] Since the value of $T_c(\mu,\nu)$ has to be known with high accuracy for conclusive results, we are restricted to relatively simple settings where analytical approaches are feasible and the optimal transport cost or map can be computed explicitly. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are considered to be subsets of $\mathbb{R}^d$, with either l_1 and l_2^2 cost functions of the form

$$c_1(x,y) = \sum_{i=1}^{d} |x_i - y_i| \qquad \text{or} \qquad c_2(x,y) = \sum_{i=1}^{d} |x_i - y_i|^2.$$

In total, we look at the following choices for μ and ν. The intrinsic dimension of the former is denoted by d_1, and the one of the latter by d_2, where $d_1 \leq d_2 \leq d$. The r-dimensional unit-sphere is denoted by $\mathbb{S}^r \subset \mathbb{R}^{r+1}$.

[6] We employed the network-simplex based C++ solver by Bonneel et al. 2011 for the computation of the empirical OT cost. The full source code used to produce the data in this section can be found under `https://gitlab.gwdg.de/staudt1/lca`.

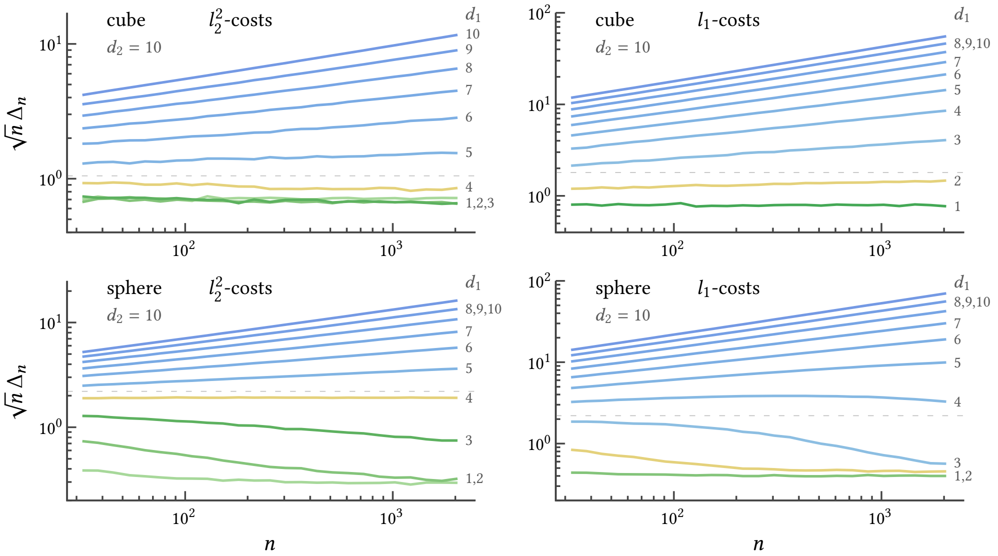


Figure 2: Simulations of the mean absolute deviation Δ_n in the *cube* and *sphere* settings. The different curves correspond to $1 \leq d_1 \leq 10$. Green lines mark the dimensions d_1 for which the upper bounds in Theorem 3 and Theorem 4 suggest $\sqrt{n}\Delta_n$ to be bounded, while yellow and blue lines enjoy no such guarantee.

(i) *Cube*: We choose $\mu = \mathrm{Unif}(\mathcal{X})$ for $\mathcal{X} = [0,1]^{d_1} \times \{0\}^{d_2-d_1}$ and $\nu = \mathrm{Unif}(\mathcal{Y})$ for $\mathcal{Y} = [0,1]^{d_2}$. As the one-sample estimates $T_c(\hat{\mu}_n, \nu)$ and $T_c(\mu, \hat{\nu}_n)$ are computationally infeasible, we employ the two-sample estimates $\hat{T}_{c,n} = T_c(\hat{\mu}_n, \hat{\nu}_n)$ for up to $n = 2^{11} = 2048$. The value of $T_c(\mu, \nu)$ is calculated analytically.

(ii) *Sphere*: We choose $\mu = \mathrm{Unif}(\mathcal{X})$ for $\mathcal{X} = \mathbb{S}^{d_1} \times \{0\}^{d_2-d_1}$ and $\nu = \mathrm{Unif}(\mathcal{Y})$ for $\mathcal{Y} = \mathbb{S}^{d_2}$. The two-sample estimate $\hat{T}_{c,n} = T_c(\hat{\mu}_n, \hat{\nu}_n)$ is used for up to $n = 2^{11} = 2048$ and $T_c(\mu, \nu)$ is approximated numerically (the optimal transport map between ν to μ can be established due to the symmetry of the setting).

(iii) *Semi-discrete*: We choose $\nu = \mathrm{Unif}(\mathcal{Y})$ for $\mathcal{Y} = [0,1]^d$ and set $\mu = \mathfrak{p}_{\#}\nu$, where $\mathfrak{p}(y) = \mathrm{argmin}_{x \in \mathcal{X}}\, c(x, y)$ denotes the c-projection onto the finite set $\mathcal{X} = \{x_i\}_{i=1}^{I} \subset [0,1]^d$ with $I \in \mathbb{N}$. Consequently, $\mu(\{x_i\})$ equals the fraction of the volume of $[0,1]^d$ that lies closest to x_i. The positions x_i have been fixed once for each pair (I, d) by drawing them uniformly in $[0,1]^d$. The one-sample estimator $\hat{T}_{c,n} = T_c(\mu, \hat{\nu}_n)$ is used for up to $n = 2^{15} = 32768$ and $T_c(\mu, \nu)$ is approximated numerically (based on the observation that $\mathfrak{p}$ is the optimal transport map between ν and μ).

A first set of simulation results in the *cube* and *sphere* settings with smooth l_2^2 and non-smooth l_1 costs for fixed $d_2 = 10$ can be seen in Figure 2. As anticipated by the LCA principle, the smaller dimension d_1 appears to dictate the convergence rate of Δ_n towards zero as n becomes large. For smooth costs, Δ_n seems to converge with

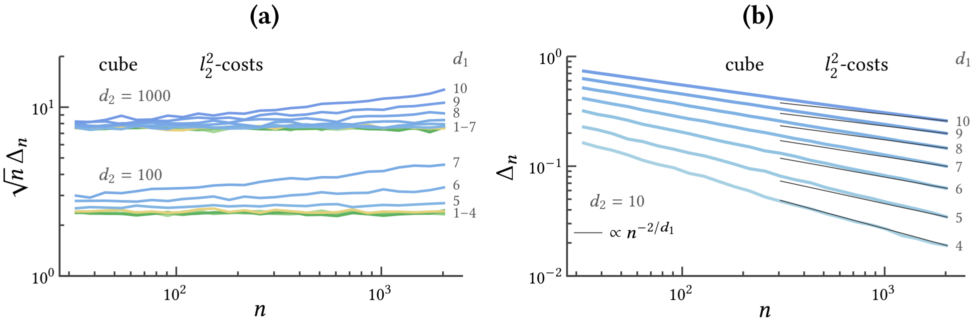


Figure 3: Additional simulations of the mean absolute deviation Δ_n in the *cube* setting. **(a)** shows analogous simulations to Figure 2 for higher dimensions $d_2 = 100$ and $d_2 = 1000$. **(b)** contrasts Δ_n to the power law behavior n^{-2/d_1} (black lines) which is expected asymptotically from Theorem 4 and Example 5.

the rate $1/\sqrt{n}$ for $d_1 \leq 3$ (and even in the critical case $d_1 = 4$, which is in line with results by Ledoux 2019), while the convergence for $d_1 \geq 5$ is perceivably slower in both settings. This is in good agreement with the upper bounds (24) established by Theorem 4 for $\mathcal{C}^2$ cost functions. For non-smooth l_1 costs, in contrast, the *cube* setting again exhibits the behavior to be expected if the bounds (22) of Theorem 3 were sharp (i.e, only $d_1 = 1$ leads to a clear $n^{-1/2}$ convergence), but the results for the *sphere* setting somewhat resemble the ones for smooth costs. This salient difference might be explained along the lines discussed in the context of equation (26). In fact, if π denotes an optimal transport plan for $T_{c_1}(\mu, \nu)$, then it is straightforward to see that the cost function c_1 is *not* differentiable π-almost surely in the *cube* setting (since all optimal movement of mass leaves the first d_1 coordinates unchanged), while it *is* differentiable π-almost surely in the *sphere* setting (almost all mass is moved such that each coordinate changes).

To understand the influence of a different choice of d_2, we also conducted analogous simulations with $d_2 = 100$ and $d_2 = 1000$. While the basic conclusions remained unchanged and the LCA principle could be confirmed (see Figure 3a), the statistical fluctuation of the Monte-Carlo estimate of Δ_n increased with increasing d_2 and larger sample sizes n were typically needed to observe linearity of Δ_n in the presented log-log plots. In this regard, Figure 3b indicates for $d_2 = 10$ in the *cube* setting that maximal sample sizes of $n = 2^{11} = 2048$ might not suffice to confidently discern the actual asymptotic convergence rates n^{-2/d_1} for $d_1 \geq 5$ (see Examples 4 and 5).

Finally, we turn to the results obtained in the *semidiscrete* setting. According to upper bound (21) established in Theorem 2, we anticipate asymptotically a convergence rate of order $n^{-1/2}$ for Δ_n independent of the choice of the cost c, the cardinality I of $\mathcal{X}$, and the dimension d_2 of the ambient space. Figure 4 confirms this expectation under smooth l_2^2 cost. Indeed, in simulations with $I = 50$ (and higher), the convergence of Δ_n seems to be quicker than $n^{-1/2}$ at first, but eventually slows down for sufficiently large n. Comparable results were observed for l_1 costs as well.

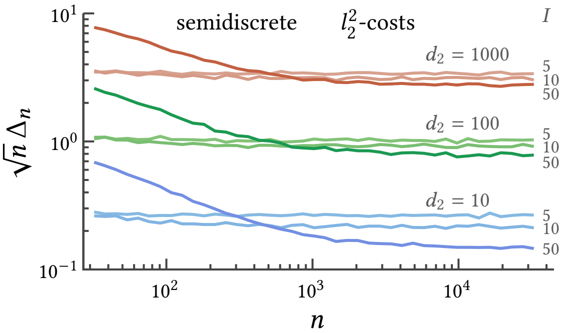


Figure 4: Simulations of the mean absolute deviation Δ_n in the *semidiscrete* setting. The three curves per value of d_2 correspond to $I = 5, 10, 50$, respectively. Since the computational burden was significantly reduced compared to the *cube* and *sphere* settings, the range of n could be increased substantially. Only this increased range makes the flattening of the curves for $I = 50$ apparent.

5 Discussion

In this work, we have established novel statistical guarantees for the empirical OT cost between *different* probability measures, showing that the mean convergence rate is governed by the less complex measure. In a broader sense, the LCA phenomenon suggests that the curse of dimensionality only affects the estimation of the OT cost when *both* probability measure exhibit high intrinsic dimensions – an observation with possibly significant repercussions for OT based data analysis applications, as the empirical OT functional automatically adapts to the complexity of the simpler measure and not to the ambient space. In particular, our theory can also be applied for the popular Wasserstein distance $W_p(\mu, \nu) = T_{d^p}(\mu, \nu)^{1/p}$ with $p \geq 1$, since

$$\mathbb{E}\left[\left|\hat{W}_{p,n} - W_p(\mu, \nu)\right|\right] \asymp \mathbb{E}\left[\left|\hat{T}_{d^p,n} - T_{d^p}(\mu, \nu)\right|\right]$$

for fixed measures $\mu \neq \nu$ on compact metric spaces, where $\hat{W}_{p,n}^p = \hat{T}_{d^p,n}$.

Several extensions of our theory seem to be natural targets for future research. First, our arguments crucially rely on *uniform* metric entropy bounds, which essentially restricts our results to bounded costs and spaces. A generalization to measure-dependent notions of metric entropy, or a technique to properly exploit the concentration of measures, might serve to overcome these limitations. In a similar vein, it might be possible to adapt the LCA principle to more general notions of dimensionality. While our theory already includes general compact Lipschitz and $\mathcal{C}^2$ surfaces in $\mathbb{R}^d$, and even metric spaces with finite Minkowski-Bouligand dimension, it is as of yet unclear if general extensions to the Hausdorff dimension Mattila 1995, Chapter 4 or the (concentration-dependent) Wasserstein dimension (Weed and Bach 2019) are viable.

Another interesting problem is to find non-trivial settings where the upper bounds in Theorem 1 fail to provide sharp rates. While it is easy to find simple examples where the rates suggested by Theorem 3 and 4 are not sharp, the bottleneck for these examples is typically a suboptimal bound for the uniform metric entropy of $\mathcal{F}_c$ when applying Theorem 1. For instance, additive costs of the form $c(x, y) = c_1(x) + c_2(y)$ always lead to the parametric convergence rate $n^{-1/2}$, which is not necessarily captured by Theorem 3 and 4. Adapting Theorem 1 to these specific costs, however, yields the correct rate for non-constant costs.

We finally stress that our proof technique explicitly relies on empirical measure based estimators under i.i.d. observations. It would be interesting to analyze whether the same (or even faster) convergence rates can be verified for other estimators or for dependent observations. In particular, it remains an open question to what extent estimators leveraging smoothness properties of the underlying measures, e.g., when μ and ν are measures on Riemannian manifolds with sufficiently regular densities with respect to the canonical volume forms, also obey the LCA principle.

Acknowledgements

S. Hundrieser and T. Staudt gratefully acknowledge support from the DFG RTG 2088 and A. Munk of the DFG CRC 1456. S. Hundrieser and T. Staudt were in part funded by the DFG under Germany's Excellence Strategy - EXC 2067/1- 390729940. S. Hundrieser was also in part funded by the DFG RTG 2088.

Bibliography

Ajtai, M., J. Komlós, and G. Tusnády (1984). "On optimal matchings". *Combinatorica* 4.4, pp. 259–264.

Altschuler, J., J. Niles-Weed, and P. Rigollet (2017). "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems.* Vol. 30. Curran Associates, Inc.

Arjovsky, M., S. Chintala, and L. Bottou (2017). "Wasserstein generative adversarial networks". In: *Proceedings of the 34th International Conference on Machine Learning.* PMLR, pp. 214–223.

Aurenhammer, F., F. Hoffmann, and B. Aronov (1998). "Minkowski-type theorems and least-squares clustering". *Algorithmica* 20.1, pp. 61–76.

Bertsimas, D. and J. Tsitsiklis (1997). *Introduction to Linear Optimization.* Athena Scientific.

Bickel, P. J. and D. A. Freedman (1981). "Some Asymptotic Theory for the Bootstrap". *The Annals of Statistics* 9.6, pp. 1196–1217.

Bobkov, S. and M. Ledoux (2021). "A simple Fourier analytic proof of the AKT optimal matching theorem". *The Annals of Applied Probability* 31.6, pp. 2567–2584.

Boissard, E. and T. Le Gouic (2014). "On the mean speed of convergence of empirical and occupation measures in Wasserstein distance". *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 50.2, pp. 539–563.

Bonneel, N., M. van de Panne, S. Paris, and W. Heidrich (2011). "Displacement interpolation using Lagrangian mass transport". In: *Proceedings of the 2011 SIGGRAPH Asia Conference*. Association for Computing Machinery.

Bronshtein, E. M. (1976). "ε-entropy of convex sets and functions". *Siberian Mathematical Journal* 17.3, pp. 393–398.

Chernozhukov, V., A. Galichon, M. Hallin, and M. Henry (2017). "Monge–Kantorovich depth, quantiles, ranks and signs". *The Annals of Statistics* 45.1, pp. 223–256.

Chizat, L., P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré (2020). "Faster Wasserstein distance estimation with the Sinkhorn divergence". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc.

Deb, N., P. Ghosal, and B. Sen (2021). "Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc.

Deb, N. and B. Sen (2021). "Multivariate rank-based distribution-free nonparametric testing using measure transportation". *Journal of the American Statistical Association*, pp. 1–16.

del Barrio, E., J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez (1999). "Tests of goodness of fit based on the L_2-Wasserstein distance". *The Annals of Statistics* 27.4, pp. 1230–1239.

del Barrio, E., A. González-Sanz, and J.-M. Loubes (2021). "A central limit theorem for semidiscrete Wasserstein distances". *Preprint arXiv:2105.11721*.

Dereich, S., M. Scheutzow, and R. Schottstedt (2013). "Constructive quantization: Approximation by empirical measures". *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 49.4, pp. 1183–1203.

Dobrić, V. and J. E. Yukich (1995). "Asymptotics for transportation cost in high dimensions". *Journal of Theoretical Probability* 8.1, pp. 97–118.

Dragomirescu, F. and C. Ivan (1992). "The smallest convex extensions of a convex function". *Optimization* 24.3-4, pp. 193–206.

Dudley, R. M. (1969). "The Speed of Mean Glivenko-Cantelli Convergence". *The Annals of Mathematical Statistics* 40.1, pp. 40–50.

Dvurechensky, P., A. Gasnikov, and A. Kroshnin (2018). "Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. PMLR, pp. 1367–1376.

Evans, S. N. and F. A. Matsen (2012). "The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, pp. 569–592.

Fournier, N. and A. Guillin (2015). "On the rate of convergence in Wasserstein distance of the empirical measure". *Probability Theory and Related Fields* 162.3, pp. 707–738.

Gangbo, W. and R. J. McCann (1996). "The geometry of optimal transportation". *Acta Mathematica* 177.2, pp. 113–161.

Geiß, D., R. Klein, R. Penninger, and G. Rote (2013). “Optimally solving a transportation problem using Voronoi diagrams”. *Computational Geometry* 46.8, pp. 1009–1016.

Guntuboyina, A. and B. Sen (2013). “Covering numbers for convex functions”. *IEEE Transactions on Information Theory* 59.4, pp. 1957–1965.

Hallin, M., E. del Barrio, J. Cuesta-Albertos, and C. Matrán (2021a). “Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach”. *The Annals of Statistics* 49.2, pp. 1139–1165.

Hallin, M., G. Mordant, and J. Segers (2021b). “Multivariate goodness-of-fit tests based on Wasserstein distance”. *Electronic Journal of Statistics* 15.1, pp. 1328–1371.

Hartmann, V. and D. Schuhmacher (2020). “Semi-discrete optimal transport: a solution procedure for the unsquared Euclidean distance case”. *Mathematical Methods of Operations Research* 92, pp. 133–163.

Heinemann, F., A. Munk, and Y. Zemel (2020). “Randomised Wasserstein barycenter computation: Resampling with statistical guarantees”. *Preprint arXiv:2012.06397.*

Kantorovich, L. (1958). “On the translocation of masses”. *Management Science* 5.1, pp. 1–4.

Kantorovich, L. (1942). “On the translocation of masses”. *Doklady Akademii Nauk URSS* 37, pp. 7–8.

Kolmogorov, A. and V. M. Tikhomirov (1961). “ε-Entropy and ε-Capacity of sets in functional spaces”. In: *Twelve Papers on Algebra and Real Functions.* American Mathematical Society, pp. 277–364.

Ledoux, M. (2019). “On Optimal Matching of Gaussian Samples”. *Journal of Mathematical Sciences* 238, pp. 495–522.

Lee, J. M. (2013). *Introduction to smooth manifolds.* Springer.

Liang, T. (2019). “On the minimax optimality of estimating the Wasserstein metric”. *Preprint arXiv:1908.10324.*

Luenberger, D. (2003). *Linear and Nonlinear Programming: Second Edition.* Springer US.

Luxburg, U. von and O. Bousquet (2004). “Distance-based classification with Lipschitz functions.” *Journal of Machine Learning Research* 5.Jun, pp. 669–695.

Mallows, C. L. (1972). “A note on asymptotic joint normality”. *The Annals of Mathematical Statistics* 43.2, pp. 508–515.

Manole, T., S. Balakrishnan, J. Niles-Weed, and L. Wasserman (2021). “Plugin estimation of smooth optimal transport maps”. *Preprint arXiv:2107.12364.*

Manole, T. and J. Niles-Weed (2021). “Sharp convergence rates for empirical optimal transport with smooth costs”. *Preprint arXiv:2106.13181.*

Mattila, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability.* Cambridge University Press.

McShane, E. J. (1934). “Extension of range of functions”. *Bulletin of the American Mathematical Society* 40.12, pp. 837–842.

Mérigot, Q. (2011). “A multiscale approach to optimal transport”. In: *Computer Graphics Forum.* Vol. 30. 5. Wiley Online Library, pp. 1583–1592.

Monge, G. (1781). “Mémoire sur la théorie des déblais et des remblais”. In: *Histoire de l’Académie Royale des Sciences de Paris*, pp. 666–704.

Mordant, G. and J. Segers (2022). “Measuring dependence between random vectors via optimal transport”. *Journal of Multivariate Analysis* 189, p. 104912.

Munk, A. and C. Czado (1998). “Nonparametric validation of similar distributions and assessment of goodness of fit”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1, pp. 223–241.

Nies, T. G., T. Staudt, and A. Munk (2021). “Transport Dependency: Optimal Transport Based Dependency Measures”. *Preprint arXiv:2105.02073*.

Niles-Weed, J. and P. Rigollet (2019). “Estimation of Wasserstein distances in the spiked transport model”. *Preprint arXiv:1909.07513*.

Panaretos, V. M. and Y. Zemel (2019). “Statistical aspects of Wasserstein distances”. *Annual Review of Statistics and Its Application* 6, pp. 405–431.

Peyré, G. and M. Cuturi (2019). “Computational optimal transport: With applications to data science”. *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607.

Rachev, S. and L. Rüschendorf (1998a). *Mass transportation problems: Volume I: Theory*. Springer.

Rachev, S. and L. Rüschendorf (1998b). *Mass transportation problems: Volume II: Applications*. Springer.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians: Calculus of variations, PDEs, and modeling*. Springer.

Schiebinger, G., J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander (2019). “Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming”. *Cell* 176.4, 928–943.e22.

Shorack, G. and J. Wellner (1986). *Empirical Processes with Applications to Statistics*. Wiley.

Singh, S. and B. Póczos (2018). “Minimax distribution estimation in Wasserstein distance”. *Preprint arXiv:1802.08855*.

Sommerfeld, M. and A. Munk (2018). “Inference for empirical Wasserstein distances on finite spaces”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1, pp. 219–238.

Sommerfeld, M., J. Schrieber, Y. Zemel, and A. Munk (2019). “Optimal transport: Fast probabilistic approximation with exact solvers”. *Journal of Machine Learning Research* 20, pp. 1–23.

Sriperumbudur, B. K., K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet (2012). “On the empirical estimation of integral probability metrics”. *Electronic Journal of Statistics* 6, pp. 1550–1599.

Stein, E. M. (1971). *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press.

Talagrand, M. (1994). “Matching theorems and empirical discrepancy computations using majorizing measures”. *Journal of the American Mathematical Society* 7.2, pp. 455–537.

Talwalkar, A., S. Kumar, and H. Rowley (2008). “Large-scale manifold learning”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8.

Tameling, C., S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk (2021). “Colocalization for super-resolution microscopy via optimal transport”. *Nature Computational Science* 1.3, pp. 199–211.

Vacher, A., B. Muzellec, A. Rudi, F. Bach, and F.-X. Vialard (2021). “A dimension-free computational upper-bound for smooth optimal transport estimation”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Vol. 134. PMLR, pp. 4143–4173.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wang, S., T. T. Cai, and H. Li (2021). “Optimal estimation of Wasserstein distance on a tree with an application to microbiome studies”. *Journal of the American Statistical Association* 116.535, pp. 1237–1253.

Weed, J. and F. Bach (2019). “Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance”. *Bernoulli* 25.4A, pp. 2620–2648.

Weed, J. and Q. Berthet (2019). “Estimation of smooth densities in Wasserstein distance”. In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Vol. 99. PMLR, pp. 3118–3119.

Whitney, H. (1934). “Analytic extensions of differentiable functions defined in closed sets”. *Transactions of the American Mathematical Society* 36.1, pp. 63–89.

Zhu, B., J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen (2018). “Image reconstruction by domain-transform manifold learning”. *Nature* 555.7697, pp. 487–492.

A Bounds on the Uniform Metric Entropy

In this appendix, we establish various upper bounds for the uniform metric entropy of the function class $\mathcal{F}_c$ defined by equation (10). To cover the settings introduced in Section 3, the following observation about uniform covering numbers under composition is useful.

> Lemma 3 (Composition bound): Let $g : \mathcal{U} \to \mathcal{V}$ be a surjective map between sets $\mathcal{U}$ and $\mathcal{V}$, and let $\mathcal{F}$ be a real-valued function class on $\mathcal{V}$. For any $\varepsilon > 0$, the class $\mathcal{F} \circ g := \{f \circ g \mid f \in \mathcal{F}\}$ satisfies
>
> $$\mathcal{N}\big(\varepsilon, \mathcal{F}, \|\cdot\|_{\infty,\mathcal{V}}\big) \le \mathcal{N}\big(\varepsilon, \mathcal{F} \circ g, \|\cdot\|_{\infty,\mathcal{U}}\big).$$

Proof. Assume that $N := \mathcal{N}\big(\varepsilon, \mathcal{F} \circ g, \|\cdot\|_{\infty,\mathcal{U}}\big)$ is finite, otherwise the inequality is trivial. Let $\{\tilde{f}_1, \dots, \tilde{f}_N\}$ be an ε-covering of $\mathcal{F} \circ g$ and let

$$f_i(v) := \sup_{u \in g^{-1}(v)} \tilde{f}_i(u).$$

For any $f \in \mathcal{F}$, there is an $\tilde{f}_i$ such that $|f \circ g - \tilde{f}_i| \le \varepsilon$ on $\mathcal{U}$. By definition of f_i, this implies $f - f_i \le \varepsilon$ and $f - f_i \ge -\varepsilon$ on $\mathcal{V}$, which shows that $\{f_1, \dots, f_N\}$ is an ε-covering of $\mathcal{F}$. □

We now provide upper bounds on the metric entropy of $\mathcal{F}_c$ under the respective assumptions (Lip), (SC), and (Hol) . Due to the union bound (Lemma 2) in conjunction with the composition bound (Lemma 3), it is in all three cases sufficient to bound

$$\log \mathcal{N}\big(\varepsilon, \mathcal{F}_c \circ g_i, \|\cdot\|_{\infty,\mathcal{U}_i}\big)$$

for all $i \in \{1, \dots, I\}$ and sufficiently small $\varepsilon > 0$. For convenience, we suppress the index i in the following proofs and work with generic $g := g_i$ and $\mathcal{U} := \mathcal{U}_i$, as well as $d := d_i$ for Assumption (Lip).

> Lemma 4 (Metric entropy under Lipschitz costs): Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and let $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be continuous so that Assumption (Lip) is fulfilled. Then, for any $\varepsilon > 0$,
>
> $$\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \lesssim \sum_{i=1}^{I} \big(\mathcal{N}(\varepsilon/4, \mathcal{U}_i, d_i) + \log(\varepsilon^{-1})\big) \tag{28a}$$
>
> Moreover, without the connectedness assumption on $\mathcal{U}_i$ in (Lip), it holds that
>
> $$\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \lesssim \sum_{i=1}^{I} \mathcal{N}(\varepsilon/4, \mathcal{U}_i, d_i) \log(\varepsilon^{-1}). \tag{28b}$$
>
> The implicit constants in (28) are universal.

Proof. By Assumption (Lip), it holds that $c(g(\cdot), y)$ is 1-Lipschitz for each $y \in \mathcal{Y}$. Hence, the class $\mathcal{F}_\mathrm{c} \circ g$ is contained in $\mathrm{BL}_1(\mathcal{U}, d)$, which denotes the 1-Lipschitz functions on $(\mathcal{U}, d)$ that are absolutely bounded by one (Santambrogio 2015, Section 1.2). For connected $\mathcal{U}$, their uniform metric entropy is bounded by (Kolmogorov and Tikhomirov 1961, Section 9)

$$\mathcal{N}\big(\varepsilon, \mathrm{BL}_1(\mathcal{U}, d), \|\cdot\|_{\infty,\mathcal{U}}\big) \le (2\,\lceil 2/\varepsilon \rceil + 1)\ 2^{\mathcal{N}(\varepsilon/4, \mathcal{U}, d)}, \tag{29a}$$

while general metric spaces only permit the bound

$$\mathcal{N}\big(\varepsilon, \mathrm{BL}_1(\mathcal{U}, d), \|\cdot\|_{\infty,\mathcal{U}}\big) \le (2\,\lceil 2/\varepsilon \rceil + 1)^{\mathcal{N}(\varepsilon/4, \mathcal{U}, d)}\,. \tag{29b}$$

This implies claim (28). Note that (29a) is a variation of equation (238) in Kolmogorov and Tikhomirov 1961, which only proves the stated bound for connected subsets of a *centrable* metric space (with some improvements, e.g., $\varepsilon/4$ can be relaxed to $s\varepsilon/(s+1)$ for $s \in \mathbb{N}$ at the cost of a possibly worse constant). However, with minor adaptions, the proof also works without requiring centrability for $\varepsilon/4$. □

Lemma 5 (Metric entropy under semi-concave costs): Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and let $c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be continuous so that Assumption (SC) is fulfilled. Then, for $\varepsilon > 0$ sufficiently small,

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_\mathrm{c}, \|\cdot\|_\infty) \lesssim I\varepsilon^{-d/2}, \tag{30}$$

where the implicit constant depends on the sets $\mathcal{U}_1, \ldots, \mathcal{U}_I \subset \mathbb{R}^d$.

Proof. Let $s := \dim\big(\mathrm{span}(\mathcal{U} - u)\big) \le d$ for an arbitrary $u \in \mathcal{U}$. If $s = 0$, the metric entropy of $\mathcal{F}_c \circ g$ is bounded as in (20), so we consider $s \ge 1$. By translation and rotation, we may w.l.o.g. assume that $\mathcal{U}$ is a bounded convex subset of $\mathbb{R}^s$ that contains the origin. By Assumption (SC) and the properties of the c-transform, any $f \in \mathcal{F}_\mathrm{c}$ is absolutely bounded by one, and the composition $f \circ g$ is 1-Lipschitz and 1-semi-concave on $\mathcal{U}$. Thus, the function $u \mapsto f \circ g(u) - \|u\|^2$ is concave, L-Lipschitz with $L := 1 + 2\,\mathrm{diam}(\mathcal{U})$, and absolutely bounded by $1 + \mathrm{diam}(\mathcal{U})^2$. According to Dragomirescu and Ivan 1992, Theorem 1 and Remark 2(ii) there exists a concave extension $\tilde{f}$ of this function to $\mathbb{R}^s$ with identical Lipschitz-modulus. If $\mathcal{D} \subset \mathbb{R}^s$ denotes a bounded closed cube that contains $\mathcal{U}$, then $\tilde{f}$ is absolutely bounded on $\mathcal{D}$ by $B := 1 + \mathrm{diam}(\mathcal{U})^2 + L\,\mathrm{diam}(\mathcal{D})$. Denoting the class of concave functions on $\mathcal{D}$ that are absolutely bounded by B and L-Lipschitz by $\mathrm{C}_{B,L}(\mathcal{D})$, we conclude for small $\varepsilon > 0$

$$\begin{aligned}\mathcal{N}\big(\varepsilon, \mathcal{F}_\mathrm{c} \circ g, \|\cdot\|_{\infty,\mathcal{U}}\big) &= \mathcal{N}\big(\varepsilon, \mathcal{F}_\mathrm{c} \circ g - \|\cdot\|^2, \|\cdot\|_{\infty,\mathcal{U}}\big)\\ &\le \mathcal{N}\big(\varepsilon, \mathrm{C}_{B,L}(\mathcal{D}), \|\cdot\|_{\infty,\mathcal{D}}\big) \lesssim \varepsilon^{-s/2} \le \varepsilon^{-d/2},\end{aligned}$$

where we used uniform metric entropy bounds for the class $C_{B,L}(\mathcal{D})$ provided in Bronshtein 1976 and Guntuboyina and Sen 2013. The implicit constants depend on B, L, and $\mathcal{D}$, which in turn depend on $\mathcal{U}$. □

Lemma 6 (Metric entropy under Hölder costs): Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and let $c : \mathcal{X} \times \mathcal{Y} \to [0,1]$ be continuous so that Assumption (Hol) is fulfilled for some $\alpha \in (0,2]$. Then, for $\varepsilon > 0$ sufficiently small,

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_{\mathrm{c}}, \|\cdot\|_\infty) \lesssim I\varepsilon^{-d/\alpha}, \tag{31}$$

where the implicit constant depends on α and the sets $\mathcal{U}_1, \dots, \mathcal{U}_I \subset \mathbb{R}^d$.

Proof. We consider $\alpha \in (0,1]$ first. An $(\alpha, 1)$-Hölder function with respect to the Euclidean norm $\|\cdot\|$ is a 1-Lipschitz function with respect to the metric induced by $\|\cdot\|^\alpha$. It follows by Assumption (Hol) that $\mathcal{F}_{\mathrm{c}} \circ g \subseteq \mathrm{BL}_1(\mathcal{U}, \|\cdot\|^\alpha)$ (see the proof of Lemma 4). Furthermore, each function in $\mathrm{BL}_1(\mathcal{U}, \|\cdot\|^\alpha)$ can be extended to an element in $\mathrm{BL}_1(\mathcal{D}, \|\cdot\|^\alpha)$, where $\mathcal{D} \subset \mathbb{R}^d$ is bounded, connected, and contains $\mathcal{U}$ (McShane 1934, Corollary 2). Thus, noting that $\mathcal{N}(\varepsilon, \mathcal{D}, \|\cdot\|^\alpha) = \mathcal{N}(\varepsilon^{1/\alpha}, \mathcal{D}, \|\cdot\|) \lesssim \varepsilon^{-d/\alpha}$ and employing (29a), we find

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_{\mathrm{c}} \circ g, \|\cdot\|_{\infty,\mathcal{U}}) \le \log \mathcal{N}\big(\varepsilon, \mathrm{BL}_1(\mathcal{D}, \|\cdot\|^\alpha), \|\cdot\|_{\infty,\mathcal{D}}\big) \lesssim \varepsilon^{-d/\alpha}.$$

For $\alpha \in (1,2]$, we apply Lemma 7 to $c(g(\cdot), y)$ for each $y \in \mathcal{Y}$ separately to define a collection of smoothed, approximated cost functions $c_\sigma \colon \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ for $\sigma \in (0,1]$, where $\mathcal{D} \subseteq \mathbb{R}^d$ contains $\mathcal{U}$ and is convex, open, and bounded. Furthermore, there is $K > 0$ so that the functions c_σ satisfy, for all $y \in \mathcal{Y}$,

$$\|c(g(\cdot), y) - c_\sigma(\cdot, y)\|_{\infty,\mathcal{U}} \le K\sigma^\alpha \quad \text{and} \quad \|c_\sigma(\cdot, y)\|_{\mathcal{C}^2(\mathcal{D})} \le K\sigma^{\alpha-2} =: \Gamma_\sigma, \tag{32}$$

where the $\mathcal{C}^2(\mathcal{D})$-norm of a twice continuously differentiable function $f \colon \mathcal{D} \to \mathbb{R}$ is

$$\|f\|_{\mathcal{C}^2(\mathcal{D})} := \max_{|\beta| \le 2} \left\| D^\beta f \right\|_{\infty,\mathcal{D}}, \quad \text{where} \quad D^\beta f = \partial^{|\beta|} f / \partial x_1^{\beta_1} \cdots \partial x_d^{\beta_d} \text{ for } \beta \in \mathbb{N}_0^d.$$

Note that a function with $\|f\|_{\mathcal{C}^2(\mathcal{D})} \le \Gamma$ for $\Gamma > 0$ is absolutely bounded by Γ, Γ-Lipschitz, and $d\Gamma$-semi-concave (since the Eigenvalues of its Hessian are bounded by $d \cdot \Gamma$). For each $f \in \mathcal{F}_{\mathrm{c}}$, we define $f_\sigma \colon \mathcal{D} \to \mathbb{R}, u \mapsto \inf_{y \in \mathcal{Y}} c_\sigma(u, y) - f^c(y)$. Due to $f = f^{cc}$ (Santambrogio 2015, Proposition 1.34) combined with the first inequality in (32), we conclude $|f \circ g - f_\sigma| \le K\sigma^\alpha$ on $\mathcal{U}$. For $\sigma(\varepsilon) := (\varepsilon/2K)^{1/\alpha}$, this implies $|f \circ g - f_{\sigma(\varepsilon)}| \le \varepsilon/2$ on $\mathcal{U}$. Consequently, defining $\mathcal{F}_{c,\sigma} := \{f_\sigma | f \in \mathcal{F}_{\mathrm{c}}\}$,

$$\mathcal{N}\big(\varepsilon, \mathcal{F} \circ g, \|\cdot\|_{\infty,\mathcal{U}}\big) \le \mathcal{N}\big(\varepsilon/2, \mathcal{F}_{c,\sigma(\varepsilon)}, \|\cdot\|_{\infty,\mathcal{U}}\big) \le \mathcal{N}\big(\varepsilon/2, \mathcal{F}_{c,\sigma(\varepsilon)}, \|\cdot\|_{\infty,\mathcal{D}}\big).$$

Since the functions $c_\sigma(\cdot, y)/d\Gamma_\sigma$ are bounded by one, 1-Lipschitz, and 1-semi-concave, we can apply the metric entropy bounds derived in the proof of Lemma 5 to conclude

$$\log \mathcal{N}\left(\frac{\varepsilon}{2}, \mathcal{F}_{\sigma(\varepsilon)}, \|\cdot\|_{\infty,\mathcal{D}}\right) = \log \mathcal{N}\left(\frac{\varepsilon}{2d\Gamma_{\sigma(\varepsilon)}}, \frac{\mathcal{F}_{\sigma(\varepsilon)}}{d\Gamma_{\sigma(\varepsilon)}}, \|\cdot\|_{\infty,\mathcal{D}}\right) \lesssim \left(\frac{\varepsilon}{\Gamma_{\sigma(\varepsilon)}}\right)^{-d/2} \asymp \varepsilon^{-d/\alpha},$$

where the constants depend on α and $\mathcal{D}$, which in turn depends on $\mathcal{U}$. □

Lemma 7: Let $\mathcal{D} \subset \mathbb{R}^d$ be bounded, convex, and open, and let $\mathcal{U} \subset \mathcal{D}$ be a compact and convex subset. Then, there exists $K > 0$ such that for any $(\alpha, 1)$-Hölder function h on $\mathcal{U}$ with $1 < \alpha \leq 2$ there is a collection of smooth functions $h_\sigma \colon \mathcal{D} \to \mathbb{R}$ such that

$$\|h - h_\sigma\|_{\infty,\mathcal{U}} \leq K\sigma^\alpha \qquad \text{and} \qquad \|h_\sigma\|_{C^2(\mathcal{D})} \leq K\sigma^{\alpha-2} \qquad \text{for } \sigma \in (0,1]. \quad (33)$$

Proof. Recall the definition of (α, Λ)-Hölder smooth functions for $1 < \alpha \leq 2$ and $\Lambda > 0$ from Section 3.4, and let $u, u_0 \in \mathcal{U}$. If we denote $z := u - u_0$, then the mean value theorem asserts the existence of $t \in [0,1]$ such that $h(u) = h(u_0) + \langle \nabla h(u_0 + tz), z\rangle$. This implies

$$h(u) = h(u_0) + \langle \nabla h(u_0), z\rangle + R_{u_0}(u), \quad \text{where} \quad R_{u_0}(u) = \langle \nabla h(u_0 + tz) - \nabla h(u_0), z\rangle.$$

Due to the $(\alpha - 1, 1)$-Hölder smoothness of the partial derivatives of h, we find

$$|R_{u_0}(u)| \leq \|\nabla h(u_0 + tz) - \nabla h(u_0)\|\|z\| \leq \sqrt{d}\,\|u - u_0\|^\alpha. \quad (34)$$

This shows that the function h is an element of the class $\mathrm{Lip}(\alpha, \mathcal{U})$ defined in Stein 1971, Chapter VI, Section 3. By Theorem 4 in the same reference, the function h admits an extension $\tilde{h}$ to $\mathbb{R}^d$ that is (α, K')-Hölder on $\mathbb{R}^d$ for some $K' > 0$ (which is independent of $\mathcal{U}$ and h). For an even and smooth mollifier $M : \mathbb{R}^d \to [0, \infty)$ supported on the unit ball B_1, we define $M_\sigma := \sigma^{-d} M(\,\cdot\,/\sigma)$, which is supported on the ball with radius $\sigma \in (0,1]$, and set

$$h_\sigma \colon \mathcal{D} \to \mathbb{R}, \quad u \mapsto \big(\tilde{h} * M_\sigma\big)(u) = \int \tilde{h}(u - z) M_\sigma(z)\,\mathrm{d}z,$$

where integration is over $\mathbb{R}^d$ (i.e., effectively over the support of M_σ). The desired properties (33) now follow analogously to the proof of Lemma 8 of Manole and Niles-Weed 2021. For completeness, we include the arguments here. We first observe

$$\tilde{h}(u) = \tilde{h}(u_0) + \langle \nabla\tilde{h}(u_0), u - u_0\rangle + \tilde{R}_{u_0}(u), \quad \text{where} \quad |\tilde{R}_{u_0}(u)| \leq \sqrt{d}K'\|u - u_0\|^\alpha,$$

for any $u, u_0 \in \mathcal{D}$, which can be derived analogously to (34). For the first bound in (33), we note that M is even, implying $\int z_i\, M_\sigma(z)\,\mathrm{d}z = 0$ for all $1 \leq i \leq d$. By expanding $\tilde{h}(u - z)$ around $u \in \mathcal{U}$ for $\|z\| \leq \sigma$, it follows that

$$\begin{aligned} |h_\sigma(u) - h(u)| &= \left|\int \Big(\tilde{h}(u - z) - \tilde{h}(u)\Big) M_\sigma(z)\,\mathrm{d}z\right| \\ &\leq \int \Big|\tilde{R}_u(u - z)\Big| M_\sigma(z)\,\mathrm{d}z \\ &\leq \sqrt{d}K'\sigma^\alpha. \end{aligned}$$

For the second inequality in (33), we fix some $u_0 \in \mathcal{D}$ and observe for any $u \in \mathcal{D}$ that

$$\begin{aligned}
h_\sigma(u) &= \int \tilde{h}(u - z) M_\sigma(z) \,\mathrm{d}z \\
&= \int \left(\tilde{h}(u_0) + \langle \nabla \tilde{h}(u_0), u - z - u_0 \rangle + \tilde{R}_{u_0}(u - z)\right) M_\sigma(z) \,\mathrm{d}z \\
&=: A_1 + \langle A_2, u \rangle + A_3(u),
\end{aligned}$$

where $A_1 \in \mathbb{R}$, $A_2 \in \mathbb{R}^d$, and $A_3(u) = \int \tilde{R}_{u_0}(z) M_\sigma(u - z) \,\mathrm{d}z$ (after a change of variables). For $\beta \in \mathbb{N}_0^d$ with $|\beta| = 2$, we evaluate $D^\beta h_\sigma$ at u_0. Exchanging differentiation and integration in the first inequality, and employing substitution in the final one, we observe

$$\begin{aligned}
\left|D^\beta h_\sigma(u_0)\right| = \left|D^\beta A_3(u_0)\right| &\leq \sigma^{-d-2} \int \left|\tilde{R}_{u_0}(z)\right| \left|D^\beta M\left(\frac{u_0 - z}{\sigma}\right)\right| \,\mathrm{d}z \\
&\leq \sqrt{d} K' \sigma^{-d-2} \int \|u_0 - z\|^\alpha \left|D^\beta M\left(\frac{u_0 - z}{\sigma}\right)\right| \,\mathrm{d}z \\
&\leq \sqrt{d} K' \sigma^{\alpha-2} \int \|z\|^\alpha |D^\beta M(z)| \,\mathrm{d}z \\
&= K'' \sigma^{\alpha-2}
\end{aligned}$$

for some $0 < K'' < \infty$. Since this holds for any $u_0 \in \mathcal{D}$ with K'' independent of u_0 and σ, we conclude $\|D^\beta h_\sigma\|_{\infty,\mathcal{D}} \leq K'' \sigma^{\alpha-2}$. Analogous inequalities for $|\beta| < 2$ follow from the fact that $\mathcal{D}$ is convex and bounded, so $D^\beta h_\sigma$ can be bounded in terms of the second derivatives of h_σ and the diameter of $\mathcal{D}$. □

B A Unifying Approach to Distributional Limits for Empirical Optimal Transport

Shayan Hundrieser [*], [†]
s.hundrieser@math.uni-goettingen.de

Marcel Klatt [*]
mklatt@mathematik.uni-goettingen.de

Thomas Staudt [*], [†]
thomas.staudt@uni-goettingen.de

Axel Munk [*], [†], [‡]
munk@math.uni-goettingen.de

Abstract

We provide a unifying approach to *central limit type theorems* for empirical optimal transport (OT). In general, the limit distributions are characterized as suprema of Gaussian processes. We explicitly characterize when the limit distribution is centered normal or degenerates to a Dirac measure. Moreover, in contrast to recent contributions on distributional limit laws for empirical OT on Euclidean spaces which require centering around its expectation, the distributional limits obtained here are centered around the *population* quantity, which is well-suited for statistical applications.

At the heart of our theory is Kantorovich duality representing OT as a supremum over a function class $\mathcal{F}_c$ for an underlying sufficiently regular cost function c. In this regard, OT is considered as a functional defined on $\ell^\infty(\mathcal{F}_c)$ the Banach space of bounded functionals from $\mathcal{F}_c$ to $\mathbb{R}$ and equipped with uniform norm. We prove the OT functional to be *Hadamard directional differentiable* and conclude distributional

[*] Institute for Mathematical Stochastics, University of Göttingen
[†] Cluster of Excellence: Multiscale Bioimaging (MBExC), University Medical Center, Göttingen
[‡] Max Planck Institute for Biophysical Chemistry, Göttingen

convergence via a *functional delta method* that necessitates weak convergence of an underlying empirical process in $\ell^\infty(\mathcal{F}_c)$. The latter can be dealt with *empirical process theory* and requires $\mathcal{F}_c$ to be a *Donsker* class. We give sufficient conditions depending on the dimension of the ground space, the underlying cost function and the probability measures under consideration to guarantee the Donsker property. Overall, our approach reveals a noteworthy trade-off inherent in central limit theorems for empirical OT: Kantorovich duality requires $\mathcal{F}_c$ to be sufficiently rich, while the empirical processes only converges weakly if $\mathcal{F}_c$ is not too complex.

Keywords: Central limit theorem, optimal transport, Wasserstein distance, regularity theory, empirical processes, bootstrap, Kantorovich potential

MSC 2020 subject classification: primary 60B12, 60F05, 60G15, 62E20, 62F40; secondary 90C08, 90C31

1 Introduction

Comparing probability distributions is a fundamental task in statistics, probability theory, machine learning, data analysis and related fields. From this viewpoint, in addition to longstanding mathematical interest, optimal transport (OT) based metrics have recently gained increasing attention for data analysis as well. A major reason is that OT metrics and related similarity measures not only allow comparing general probability distributions, but can also be designed to respect the metric structure of the underlying ground space. This often results in visually appealing and well interpretable outcomes which together with recent computational progress explains the advancement of OT based data analysis throughout various disciplines, ranging from economics (Galichon 2016) to statistics (Panaretos and Zemel 2019), machine learning (Peyré and Cuturi 2019), signal and image processing (Bonneel et al. 2011; Kolouri et al. 2017) and biology (Schiebinger et al. 2019; Tameling et al. 2021), among others.

In the following, we consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and denote the set of probability measures thereon by $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$, respectively. Given some non-negative measurable cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the OT cost between $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ is defined as

$$\mathrm{OT}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\pi(x, y), \tag{1}$$

where set $\Pi(\mu, \nu)$ denotes the collection of probability measures on $\mathcal{X} \times \mathcal{Y}$ such that their marginal distributions coincide with μ and ν, respectively. Any optimizer $\pi \in \Pi(\mu, \nu)$ for (1) is termed OT plan. Essentially, OT in (1) comprises the challenge to transform the measure μ into the measure ν in a cost optimal way. Under mild assumptions on the cost function, OT_c in (1) enjoys for any pair of measures $\mu \in$

$\mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ the dual formulation

$$\mathrm{OT}_c(\mu,\nu) = \sup_{f\in\mathcal{F}_c} \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} f^c(y)\,\mathrm{d}\nu(y), \tag{2}$$

formally known as *Kantorovich duality.* The function class $\mathcal{F}_c$ depends on the underlying cost function and $f^c(y) = \inf_{x\in\mathcal{X}} c(x,y) - f(x)$ is the *c-conjugate* of $f \in \mathcal{F}_c$. Any function f attaining the supremum in (2) is termed *Kantorovich potential* and the set

$$S_c(\mu,\nu) := \left\{ f \in \mathcal{F}_c \mid \mathrm{OT}_c(\mu,\nu) = \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} f^c(y)\,\mathrm{d}\nu(y) \right\} \tag{3}$$

denotes the collection of all Kantorovich potentials. We refer to the monographs by Rachev and Rüschendorf 1998a, Rachev and Rüschendorf 1998b, Villani 2003, Villani 2008, and Santambrogio 2015 for comprehensive treatment, and to Section 2 for further details. In a statistical application the probability measures μ and ν are estimated from data. For the moment we assume ν to be known and that we have access to realizations of independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n \sim \mu$. The corresponding *empirical measure* $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ thus serves as a proxy for μ. The *population quantity* $\mathrm{OT}_c(\mu,\nu)$ is then estimated by the empirical OT cost $\mathrm{OT}_c(\hat{\mu}_n,\nu)$, posing questions about its statistical performance.

To this end, distributional limits for the (properly standardized) empirical OT cost are fundamental as they capture the asymptotic fluctuation around their population quantities. In the following, we denote such results as *central limits theorems* (CLTs).[4] Available results in the literature can be broadly distinguished between the case that $\mu = \nu$ (the null hypothesis in the context of statistical testing) and $\mu \neq \nu$ (corresponding to the alternative when testing $\mu = \nu$). A well studied setting in this regard is the p-th order *Wasserstein distance*[5] $\mathrm{OT}_p^{1/p}(\mu,\nu)$ on $\mathbb{R}^d$ that arises by choosing the cost $c(x,y) := \|x-y\|^p$ and probability measures with finite p-th moments in (1). First analyses have been devoted to the real line ($d = 1$) for which the p-th order Wasserstein distance is equal the L^p distance between the quantile functions of the measures. Under the null $\mu = \nu$ and $p = 1$, early contributions by del Barrio et al. 1999 (see also Mason 2016) provide necessary and sufficient conditions on the probability measures such that the random quantity $\sqrt{n}\mathrm{OT}_1(\hat{\mu}_n,\mu)$ weakly converges towards an integral of a suitable Brownian bridge. A weak limit for $p = 2$ is obtained by del Barrio et al. 2005. The regime $p \in (1,2)$ was analyzed only recently by Berthet and Fort 2019. For $p > 2$, CLTs are not immediately available but the work by

[4] Historically, the term CLT meant to describe the asymptotic distribution of a *sum* of random variables (Le Cam 1986). Indeed, the empirical OT cost is the sum of costs between random points weighted according to an OT plan.

[5] To alleviate notation, we write $\mathrm{OT}_p(\mu,\nu)$ for probability measures $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ supported on a Euclidean space and cost function equal to $c(x,y) = \|x-y\|^p$ for some $p \geq 1$. The Wasserstein distance is then equal to $\mathrm{OT}_p^{1/p}(\mu,\nu)$ and commonly denoted by $W_p(\mu,\nu)$. Analogously, the set of Kantorovich potentials in (3) is denoted by $S_p(\mu,\nu)$ in this case.

Bobkov and Ledoux 2019 indicates that similar results can be obtained from general quantile process theory (Csörgö and Horváth 1993). Under the alternative $\mu \neq \nu$ for $p \geq 1$, the random quantity $\sqrt{n}(\mathrm{OT}_p(\hat{\mu}_n, \nu) - \mathrm{OT}_p(\mu, \nu))$ often asymptotically follows a centered Gaussian distribution (Munk and Czado 1998; del Barrio et al. 2019; Berthet and Fort 2019; Berthet et al. 2020). Similar in spirit are recent contributions by Hundrieser et al. 2021 for measures supported on the circle.

While the previous works all benefit from the representation of the OT plan as a quantile coupling when the ground space is totally ordered, the general situation is much more complicated. A unifying analysis for discrete metric spaces $\mathcal{X} = \{x_1, x_2, \dots\}$ and a metric based cost function $c(x_i, x_j) = d^p(x_i, x_j)$ for $p \geq 1$ is given by Sommerfeld and Munk 2018 and Tameling et al. 2019, which could be seen as a starting point for this paper. For arbitrary measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ the limiting random variable is specifically characterized applying the functional delta method in terms of a supremum over (infinite dimensional) Gaussian random vectors $\mathbb{G}_\mu \sim \mathcal{N}(0, \Sigma(\mu))$, namely

$$\sqrt{n}\,(\mathrm{OT}_c(\hat{\mu}_n, \nu) - \mathrm{OT}_c(\mu, \nu)) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \left\langle \mathbb{G}_\mu, f \right\rangle, \tag{4}$$

where the supremum is taken over $S_c(\mu, \nu)$, the set of optimal Kantorovich potentials in (3). Parallel and independently to our work, del Barrio et al. 2022 extended this to semi-discrete OT for which (4) remains valid even if the discrete measure ν is replaced by a general probability measure supported on some Polish space, e.g., absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$ (see Section 5).

Beyond the countable, one-dimensional and semi-discrete case, the asymptotic distributional behavior for empirical OT becomes much more involved. Already for $d = 2$, precise CLTs remain elusive as highlighted by Ajtai et al. 1984 (see also Talagrand 1994; Bobkov and Ledoux 2021) who showed for the uniform distribution μ on the unit square that $\mathrm{OT}_1(\hat{\mu}_n, \mu) \asymp (\log(n)/n)^{1/2}$ with high probability. For higher dimensions $d \geq 3$, it is well known that any absolutely continuous measure μ with compact support on $\mathbb{R}^d$ fulfills $\mathrm{OT}_1(\hat{\mu}_n, \mu) \asymp n^{-1/d}$ (Dudley 1969; Dobrić and Yukich 1995) which implies $\sqrt{n}\mathrm{OT}_1(\hat{\mu}_n, \mu)$ to diverge. Indeed, for general $p \geq 1$ the literature on the convergence of empirical OT is vast and we therefore only give a selective view on recent papers biased towards our main results. Overall, slow convergence rates in the high-dimensional regime seem inevitable as the Wasserstein distances of any order $p \geq 1$ suffers from the *curse of dimensionality* $\mathbb{E}[\mathrm{OT}_p(\hat{\mu}_n, \mu)] \asymp n^{-p/d}$ whenever $d > 2p$ and μ is Lebesgue absolutely continuous (Boissard and Le Gouic 2014; Fournier and Guillin 2015; Weed and Bach 2019). This demonstrates the scaling rate $\sqrt{n}$ in (4) to be of wrong order for high dimensional (Euclidean) spaces. However, when centering with the *empirical expectation*, concentration results demonstrate the random quantity $\sqrt{n}$ $(\mathrm{OT}_2(\hat{\mu}_n, \nu) - \mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n, \nu)])$ to be tight (Weed and Bach 2019; Chizat et al. 2020). A fundamental step further has been taken by del Barrio and Loubes 2019 who obtain for probability measures with $4 + \delta$ finite moments

$(\delta > 0)$ and a positive density in the interior of their convex support that

$$\sqrt{n}\,(\mathrm{OT}_2(\hat{\mu}_n, \nu) - \mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n, \nu)]) \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}\big(0, \mathrm{Var}_{X\sim\mu}[f(X)]\big). \tag{5}$$

Here and in the following, $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The asymptotic variance $\mathrm{Var}_{X\sim\mu}[f(X)]$ in (5) is equal to the variance of the random variable $f(X)$ with $X \sim \mu$ and f the *unique* Kantorovich potential for (2). Under the null $\mu = \nu$, the asymptotic variance is equal to zero and the CLT in (5) degenerates, in contrast to the alternative $\mu \neq \nu$ that usually leads to a non-degenerate normal limit law. The approach by del Barrio and Loubes 2019 relies on approximating the empirical OT cost via (2) as a linear functional involving a unique Kantorovich potential for which a CLT immediately follows. Based on the Efron-Stein variance inequality this functional is shown to serve as a good L^2-approximizer for the random quantity $\sqrt{n}\ (\mathrm{OT}_2(\hat{\mu}_n, \nu) - \mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n, \nu)])$ which yields the conclusion. These results have recently been extended by del Barrio et al. 2021 to more general convex cost functions, including a CLT similar to (5) for OT_p under $p > 1$ provided the probability measures have moments of order $2p$. Notably, the regularity conditions on the measures impose the Kantorovich potential to be unique and it remains unclear how to generalize their approach under non-uniqueness. More crucially, the statement requires centering around the empirical expectation which hinders its immediate use for statistical applications. It might be tempting to replace $\mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n, \nu)]$ by its population counterpart $\mathrm{OT}_2(\mu, \nu)$ in (5). On the real line $(d = 1)$ and under sufficient regularity assumptions, this is possible (del Barrio et al. 2019). However, it remains a delicate issue for $d \geq 2$. By an observation of Manole and Niles-Weed 2021, Proof of Proposition 21, if μ and ν are uniform measures on two different balls of equal radius the bias is lower bounded by $\mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n, \nu)] - \mathrm{OT}_2(\mu, \nu) \gtrsim n^{-2/d}$. This demonstrates the replacement of the centering with the population quantity in (5) to be invalid for $d \geq 5$ (the special case $d = 4$ is further addressed in Section 5.4). Nevertheless, employing a different estimator may allow under additional assumptions for faster convergence rates of the bias in high dimensions. Indeed, for probability measures μ, ν on the unit cube $[0, 1]^d$ with sufficiently smooth densities Manole et al. 2021 propose a suitable wavelet estimator $\tilde{\mu}_n$ and prove the bias to be of order $|\mathbb{E}\,[\mathrm{OT}_2(\tilde{\mu}_n, \nu)] - \mathrm{OT}_2(\mu, \nu)| = o(n^{-1/2})$. Combined with a strategy as outlined by del Barrio and Loubes 2019, the random quantity $\sqrt{n}(\mathrm{OT}_2(\tilde{\mu}_n, \nu) - \mathrm{OT}_2(\mu, \nu))$ is shown to asymptotically follow a centered Gaussian distribution analogous to (5), which also degenerates under the null $\mu = \nu$. Despite this being an interesting result, CLTs for empirical OT costs based on general cost functions and centered around the population quantity remain largely open.

In this work, we provide a unifying approach to obtain CLTs for empirical OT that include certain aforementioned settings but go far beyond. In particular, our approach does not rely on discrete spaces, neither on explicit formulas involving quantiles $(d = 1)$ nor unique Kantorovich potentials $(d \geq 2)$. Our limit laws are reminiscent of the discrete case (4) but hold for considerably more general cost functions and probability measures. The limiting random variables are characterized as suprema of Gaussian processes $\mathbb{G}_\mu$ in $\ell^\infty(\mathcal{F}_c)$ and indexed in Kantorovich potentials from $S_c(\mu,\nu)$ in (3). The CLTs are centered around the *population quantity.* Our main result (Theorem 1) states that under certain assumptions

$$\sqrt{n}\left(\mathrm{OT}_c(\hat{\mu}_n,\nu)-\mathrm{OT}_c(\mu,\nu)\right)\xrightarrow{\mathcal{D}}\sup_{f\in S_c(\mu,\nu)}\mathbb{G}_\mu(f). \tag{6}$$

The distributional limit law (6) is valid whenever $\mathcal{F}_c$ in (7) (below) is μ-Donsker and

(C) $c\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$ is continuous and bounded by $\|c\|_\infty := \sup_{x,y}|c(x,y)|<\infty$.

holds true combined with one of the following two assumptions:

(S1) $\mathcal{X}$ and $\mathcal{Y}$ are locally compact and $\{c(\cdot,y)\mid y\in\mathcal{Y}\}$ and $\{c(x,\cdot)\mid x\in\mathcal{X}\}$ are equicontinuous[6] on $\mathcal{X}$ and $\mathcal{Y}$, respectively.

(S2) $\mathcal{X}$ is compact and $\{c(\cdot,y)\mid y\in\mathcal{Y}\}$ is equicontinuous[6] on $\mathcal{X}$.

Under Assumption **(C)** the function class $\mathcal{F}_c$ used in (2) can be chosen as a uniformly bounded class of c-concave functions on $\mathcal{X}$ (Villani 2003, Remark 1.13)

$$\mathcal{F}_c := \left\{f\colon \mathcal{X}\to\mathbb{R}\mid \exists g\colon\mathcal{Y}\to\mathbb{R},\ -\|c\|_\infty\le g\le 0,\ f=\inf_{y\in\mathcal{Y}} c(\cdot,y)-g(y)\right\}. \tag{7}$$

Moreover, $\mathbb{G}_\mu$ in (6) is a tight centered Gaussian process and represents the weak limit of the empirical process $\sqrt{n}(\hat{\mu}_n-\mu)$ in $\ell^\infty(\mathcal{F}_c)$, the Banach space of bounded functionals from $\mathcal{F}_c$ to $\mathbb{R}$ equipped with uniform norm. The covariance of $\mathbb{G}_\mu$ specifically depends on μ and is equal to

$$\mathbb{E}\left[\mathbb{G}_\mu(f_1)\mathbb{G}_\mu(f_2)\right]=\int_{\mathcal{X}} f_1(x)f_2(x)\,\mathrm{d}\mu(x)-\int_{\mathcal{X}} f_1(x)\,\mathrm{d}\mu(x)\int_{\mathcal{X}} f_2(x)\,\mathrm{d}\mu(x) \tag{8}$$

for two functions $f_1,f_2\in\mathcal{F}_c$. If instead ν is estimated by its empirical version, then under the same conditions on the cost and the spaces, i.e., Assumption **(C)** combined with **(S1)** or **(S2)**, and if $\mathcal{F}_c^c$, the set of all c-conjugate functions from $\mathcal{F}_c$, is ν-Donsker, our CLT reads as

$$\sqrt{n}\left(\mathrm{OT}_c(\mu,\hat{\nu}_n)-\mathrm{OT}_c(\mu,\nu)\right)\xrightarrow{\mathcal{D}}\sup_{f\in S_c(\mu,\nu)}\mathbb{G}_\nu(f^c), \tag{9}$$

where $\mathbb{G}_\nu$ is a centered Gaussian process in the Banach space $\ell^\infty\left(\mathcal{F}_c^c\right)$ with similar covariance as (8) corresponding to the weak limit of the empirical process $\sqrt{n}(\hat{\nu}_n-\nu)$.

Our proof technique relies on Kantorovich duality that represents the $\mathrm{OT}_c(\cdot,\cdot)$ cost as a functional from $\ell^\infty(\mathcal{F}_c)\times\ell^\infty\left(\mathcal{F}_c^c\right)$ to $\mathbb{R}$. We take upon this approach in Section 2

[6] Equicontinuity refers to a common modulus of continuity for the function classes $\{c(\cdot,y)\mid y\in\mathcal{Y}\}$ and $\{c(x,\cdot)\mid x\in\mathcal{X}\}$ with respect to a continuous metric on $\mathcal{X}$ and $\mathcal{Y}$.

and prove that $\mathrm{OT}_c(\cdot,\cdot)$ is Hadamard directionally differentiable. An application of a general functional delta method (Dümbgen 1993; Römisch 2004) allows concluding our main results on CLTs for empirical OT. This necessitates the empirical process $\sqrt{n}(\hat{\mu}_n - \mu)$ to converge weakly to some tight random element $\mathbb{G}_\mu$ in $\ell^\infty(\mathcal{F}_c)$. The latter can be dealt with *empirical process theory* and requires the function class $\mathcal{F}_c$ to be *μ-Donsker*. Our results naturally extend to the two sample case where both probability measures are estimated by their empirical versions, simultaneously. We further characterize asymptotic *normality*, i.e., when the supremum in (6) and (9) is over a singleton (Theorem 3 in Section 3) and *degeneracy* (Theorem 4 in Section 4), i.e., when the limit law is equal to a Dirac measure which results from unique and trivial (almost surely constant) Kantorovich potentials, respectively. We emphasize that even if $\mu = \nu$, the CLTs might be non-degenerate, e.g., if μ has disconnected support, as for the discrete case in (4), where limit laws usually do not degenerate to a Dirac measure at zero (Tameling et al. 2019).

In light of the general CLT statement in (6), we discuss concrete settings for which the assumptions on the cost are satisfied and $\mathcal{F}_c$ is μ-Donsker (Section 5). The latter manifests itself in the complexity of $\mathcal{F}_c$ measured in terms of covering numbers (Van der Vaart and Wellner 1996) as well as tail conditions on the measure μ. The covering numbers depend on properties of the cost function and the underlying dimension of the ground space (see also Gangbo and McCann 1996; Chizat et al. 2020; Hundrieser et al. 2022). The simplest case appears on finite spaces where $\mathcal{F}_c$ is even *universal* Donsker. On countable discrete spaces it requires the popular *Borisov-Dudley-Durst* summability condition on the measure μ. This is in line with the aforementioned results by Sommerfeld and Munk 2018 and Tameling et al. 2019 (Corollary 4). Beyond discrete spaces, we employ well-known results in empirical process theory. More precisely and tailored to Euclidean spaces $\mathbb{R}^d$ with $d \leq 3$, if the cost satisfies certain regularity conditions, then $\mathcal{F}_c$ is μ-Donsker provided the measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfies

$$\sum_{k\in\mathbb{Z}^d} \sqrt{\mu\left([k, k+1)\right)} < \infty.$$

This implies CLTs for the empirical OT cost on the real line with general cost functions (Theorem 6) but also novel statements for dimension $d = 2, 3$ (Theorem 7). Moreover, in Section 5.4 we show for $\mu = \mathrm{Unif}([0,1]^d)$ that the μ-Donsker property of $\mathcal{F}_c$ is already violated under smooth costs $c(x, y) = \|x - y\|^2$ for $d \geq 4$ and that CLTs as in (6) cannot hold for $d \geq 5$ (for $d = 4$ such CLTs are still possible though and we give an example). In view of this observation, our approach is exhaustive in terms of the dimension. Nevertheless, provided only one of the probability measures is supported on low-dimensional compact smooth submanifold of $\mathbb{R}^d$ we highlight that the Donsker property is still fulfilled and that the CLTs in (6) remain valid (Theorem 9). This is related to recent findings by Hundrieser et al. 2022 discovering the *lower complexity adaptation* phenomenon of empirical OT which states that the convergence rate of the empirical OT cost under different population measures adapts to the measure with lower-dimensional support. Moreover, based on this observation, we also derive a CLT for the empirical OT cost in the semi-discrete framework (Theorem 8). We postpone further technical proofs and auxiliary results on Kantorovich potentials to Appendix A.

Notation. The set of non-negative real numbers is $\mathbb{R}_+$. For $a, b \in \mathbb{R}$ the inequality $a \lesssim_\kappa b$ means that a is larger than b up to a constant depending on κ. In certain instances we omit the κ. If $a \lesssim b \lesssim a$, we write $a \asymp b$. The class of real-valued continuous functions defined on a metric space $\mathcal{X}$ is denoted by $C(\mathcal{X})$. A real-valued function f defined on some convex subset $A \subset \mathcal{X}$ is λ-semi-concave if there exists some constant $\lambda > 0$ such that $f(x) - \lambda \|x\|_2^2$ is concave. Furthermore, f is said to be (α, L)-Hölder continuous if there exist positive constants $\alpha \in (0, 1]$ and $L > 0$ such that $|f(x) - f(x')| \leq L\|x - x'\|^\alpha$ for all x, x' in the domain of f. If $\alpha = 1$, then f is L-Lipschitz. For a function class $\mathcal{F}$ on $\mathcal{X}$ denote by $\|f - g\|_\infty$ the uniform norm $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Let $\ell^\infty(\mathcal{F})$ be the Banach space of real-valued bounded functionals on $\mathcal{F}$ with respect to uniform norm $\|\varphi\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\varphi(f)|$. For $(\mathcal{F}, d)$ a subset of some metric space with metric d and $\varepsilon > 0$, the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, d)$ is the minimal number of balls $\{g \mid d(f, g) < \varepsilon\}$ of radius ε such that their union contains $\mathcal{F}$. The metric entropy of $\mathcal{F}$ is the logarithm of the covering number $\log\big(\mathcal{N}(\varepsilon, \mathcal{F}, d)\big)$.

2 Central Limit Theorems for Empirical Optimal Transport

Throughout this section, we consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a non-negative cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ such that Assumption **(C)** is fulfilled. Under this assumption the OT cost in (1) is finite for any two probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and enjoys Kantorovich duality (2). The set $\mathcal{F}_c$, over which the dual is optimized, is the collection of uniformly bounded *c-concave* functions on $\mathcal{X}$ defined as in (7) (for details see Villani 2003, Remark 1.13 and the Appendix). The supremum in (2) over $\mathcal{F}_c$ is attained, i.e., there exists a Kantorovich potential $f \in \mathcal{F}_c$, where we recall the set $S_c(\mu, \nu)$ in (3) of all Kantorovich potentials. Notably, as the cost function is continuous, any function $f \in \mathcal{F}_c$ and its c-conjugate f^c are upper semi-continuous and thus measurable on the Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively.

More general structural properties for c-concave functions $\mathcal{F}_c$ and its c-conjugate class $\mathcal{F}_c^c := \{f^c \mid f \in \mathcal{F}_c\}$ are intrinsically linked to the cost function and the underlying Polish space. Hence, to guarantee (minimal) regularity properties of the set of Kantorovich potentials we impose Assumption **(S1)** or **(S2)**. Under either of these conditions, the Kantorovich potentials $S_c(\mu, \nu)$ and $S_c^c(\mu, \nu)$ are continuous as well (Lemma 1). Moreover, under compactness or local compactness of the spaces $\mathcal{X}$ and $\mathcal{Y}$ as well as the equicontinuity condition of the cost function, the classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are by the *Arzelà-Ascoli* theorem relatively compact with respect to uniform convergence on compact sets (see proof of Theorem 1 and in particular Step 3).

Examples of locally compact spaces are Euclidean spaces but also encompass finite and countable spaces equipped with discrete topology. Assumptions **(C)** and **(S1)** are then fulfilled for the Euclidean case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ if $c(x, y) = h(x - y)$ for some bounded Lipschitz function $h \colon \mathbb{R}^d \to \mathbb{R}_+$ and for discrete spaces $\mathcal{X}, \mathcal{Y}$ if the cost function is uniformly bounded. Furthermore and for compact subsets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, Assumptions **(C)** and **(S2)** are fulfilled for cost $c(x, y) = h(x - y)$ for some (possibly unbounded) Lipschitz function $h \colon \mathbb{R}^d \to \mathbb{R}_+$. In particular, this includes the costs $c(x, y) = \|x - y\|^p$ with $p \geq 1$. Other important settings are detailed in Section 5.

2.1 Main Result

In the following, the empirical process $\sqrt{n}(\hat{\mu}_n - \mu)$ is considered as a random element in the Banach space $\ell^\infty(\mathcal{F}_c)$ equipped with uniform norm.

Definition 1 (Van der Vaart and Wellner 1996): A function class $\mathcal{F}$ of measurable, pointwise bounded functions on $\mathcal{X}$ is called *μ-Donsker*, if the empirical process $\sqrt{n}(\hat{\mu}_n - \mu)$ weakly converges to a tight random element $\mathbb{G}_\mu$ in $\ell^\infty(\mathcal{F})$. Furthermore, $\mathcal{F}$ is *universal Donsker* if for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ the function class $\mathcal{F}$ is μ-Donsker.

If $\mathcal{F}_c$ is μ-Donsker, the weak limit $\mathbb{G}_\mu$ is a mean-zero Gaussian process indexed over the function class $\mathcal{F}_c$ with covariance for $f_1, f_2 \in \mathcal{F}_c$ as in (8). We now state our main result on the CLTs for the empirical OT cost.

Theorem 1: Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption **(C)** combined with **(S1)** or **(S2)**. For two probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$, i.i.d. random variables $X_1, \dots, X_n \sim \mu$ and independent to that i.i.d. random variables $Y_1, \dots, Y_m \sim \nu$ with $n, m \in \mathbb{N}$ denote their respective empirical measures by $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_m := \frac{1}{m}\sum_{i=1}^m \delta_{Y_i}$.

(i) (One-sample from μ) Suppose that $\mathcal{F}_c$ is μ-Donsker. Then, for $n \to \infty$,

$$\sqrt{n}\left(\mathrm{OT}_c(\hat{\mu}_n, \nu) - \mathrm{OT}_c(\mu, \nu)\right) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \mathbb{G}_\mu(f). \tag{10a}$$

(ii) (One-sample from ν) Suppose that $\mathcal{F}_c^c$ is ν-Donsker. Then, for $m \to \infty$,

$$\sqrt{m}\left(\mathrm{OT}_c(\mu, \hat{\nu}_m) - \mathrm{OT}_c(\mu, \nu)\right) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \mathbb{G}_\nu(f^c). \tag{10b}$$

(iii) (Two-sample) Suppose that $\mathcal{F}_c$ is μ-Donsker and that $\mathcal{F}_c^c$ is ν-Donsker. Then, for $n, m \to \infty$ with $m/(n+m) \to \delta \in (0, 1)$,

$$\begin{aligned}&\sqrt{\frac{nm}{n+m}}\left(\mathrm{OT}_c(\hat{\mu}_n, \hat{\nu}_m) - \mathrm{OT}_c(\mu, \nu)\right)\\ &\qquad\xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \left(\sqrt{\delta}\,\mathbb{G}_\mu(f) + \sqrt{1-\delta}\,\mathbb{G}_\nu(f^c)\right).\end{aligned} \tag{10c}$$

Proof. Our proof is based on the Hadamard directional differentiability of the OT cost on the set of probability measures $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ with respect to the topology induced by $\ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$ (see the subsequent Theorem 2). Under settings (i) and (ii) it holds for $n \to \infty$ and $m \to \infty$ that

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathbb{G}_\mu \text{ in } \ell^\infty(\mathcal{F}_c), \quad \sqrt{m}(\hat{\nu}_m - \nu) \xrightarrow{\mathcal{D}} \mathbb{G}_\nu \text{ in } \ell^\infty(\mathcal{F}_c^c),$$

respectively. For setting (iii) it holds by Van der Vaart and Wellner 1996, Example 1.4.6, that the empirical processes converge jointly

$$\sqrt{\frac{nm}{n+m}}(\hat{\mu}_n - \mu, \hat{\nu}_m - \nu) \xrightarrow{\mathcal{D}} \left(\sqrt{\delta}\mathbb{G}_\mu, \sqrt{1-\delta}\mathbb{G}_\nu\right) \text{ in } \ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$$

with $m/(n+m) \to \delta \in (0,1)$. Moreover, Theorem 2 below asserts that $\mathrm{OT}_c \colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ is Hadamard directionally differentiable at (μ, ν) with respect to $\ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$ tangentially to $\mathcal{P}(\tilde{\mathcal{X}}) \times \mathcal{P}(\tilde{\mathcal{Y}})$ with $\tilde{\mathcal{X}} = \mathrm{supp}(\mu)$ and $\tilde{\mathcal{Y}} = \mathrm{supp}(\nu)$ and $\mathcal{T}_\mu(\mathcal{P}(\tilde{\mathcal{X}})) = \mathrm{Cl}\{\frac{\mu'-\mu}{t} \text{ for } t > 0, \mu' \in \mathcal{P}(\tilde{\mathcal{X}})\}$ and analogously for $\mathcal{T}_\nu(\mathcal{P}(\tilde{\mathcal{Y}}))$. Portmanteau's Theorem for closed sets (Van der Vaart and Wellner 1996, Theorem 1.3.4 (iii)) proves that

$$\mathbb{P}\left(\mathbb{G}_\mu \in \mathcal{T}_\mu(\mathcal{P}(\tilde{\mathcal{X}}))\right) \geq \limsup_{n\to\infty} \mathbb{P}\left(\sqrt{n}(\hat{\mu}_n - \mu) \in \mathcal{T}_\mu(\mathcal{P}(\tilde{\mathcal{X}}))\right) = 1 \tag{11}$$

since $\mathrm{supp}(\mu_n) \subseteq \mathrm{supp}(\mu)$, and analogously for $\mathbb{G}_\nu$. Then, the statements in Theorem 1 follow from the functional delta method (Römisch 2004, Theorem 1). □

We investigate specific settings that yield novel CLTs in Section 5. At this stage, we like to highlight that Theorem 1 links CLTs for the empirical OT cost to the study of the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ being Donsker, respectively.

Remark 1: Even if Assumptions **(S1)** and **(S2)** fail to hold, the first step in the proof of Theorem 2 still asserts Hadamard directional differentiability of the OT cost. However, in this case the derivative rather depends on the set of ε-approximate optimizer $S_c(\mu, \nu, \varepsilon) := \{f \in \mathcal{F}_c \mid \mu(f) + \nu(f^c) \geq \mathrm{OT}_c(\mu,\nu) - \varepsilon\}$ instead of the set of Kantorovich potentials $S_c(\mu, \nu)$. Hence, employing a functional delta method (Römisch 2004, Theorem 1), we obtain that, as long as $\mathcal{F}_c$ is μ-Donsker, a CLT of the form

$$\sqrt{n}\Big(\mathrm{OT}_c(\hat{\mu}_n, \nu) - \mathrm{OT}_c(\mu, \nu)\Big) \xrightarrow{\mathcal{D}} \lim_{\varepsilon \searrow 0} \sup_{f \in S_c(\mu,\nu,\varepsilon)} \mathbb{G}_\mu(f)$$

is valid. Assumptions **(S1)** and **(S2)** serve to simplify the limit distribution obtained in Theorem 1. For more details we refer to Section 2.2.

Remark 2 (Bootstrap consistency): It is well-known that the naive n-out-of-n bootstrap fails to be consistent if the functional is not linearly Hadamard differentiable (Dümbgen 1993; Fang and Santos 2019). Instead, Hadamard directional differentiability (Dümbgen 1993, Proposition 2) asserts consistency of the k-out-of-n bootstrap for $k = o(n)$ which has immediate consequences for the approximation of quantiles for the empirical OT cost. We formalize this principle exemplary for the one-sample case. For $n, k \in \mathbb{N}$, consider i.i.d. samples $X_1, \dots, X_n \sim \mu$ with empirical measure $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and consider i.i.d. bootstrap samples $X_1^*, \dots, X_k^* \sim \hat{\mu}_n$ with corresponding (bootstrap) empirical measure $\hat{\mu}_{n,k}^* = \frac{1}{k}\sum_{i=1}^k \delta_{X_i^*}$. Then, it follows for $n, k \to \infty$ with $k = o(n)$ that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{R})} \Bigg| \mathbb{E}\left[h\left(\sqrt{k}(\mathrm{OT}_c(\hat{\mu}_{n,k}^*, \nu) - \mathrm{OT}_c(\hat{\mu}_n, \nu))\right) \Big| X_1, \dots, X_n \right] \\ - \mathbb{E}\left[h\left(\sqrt{n}(\mathrm{OT}_c(\hat{\mu}_n, \nu) - \mathrm{OT}_c(\mu, \nu))\right)\right] \Bigg| \xrightarrow{\mathbb{P}} 0.$$

Herein, $\xrightarrow{\mathbb{P}}$ denotes convergence in outer probability (Van der Vaart and Wellner 1996) and $\mathrm{BL}_1(\mathbb{R})$ is the set of real-valued functions on $\mathbb{R}$ that are absolutely bounded by one and Lipschitz with modulus one.

2.2 Hadamard Directional Differentiability

Based on Kantorovich duality (2), we consider the OT cost as an optimization problem over the set of c-concave functions mapping from the set of probability measures $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ to $\mathbb{R}$. For a pair of probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, we prove Hadamard directional differentiability of the OT cost at (μ, ν) with respect to the Banach spaces $\ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$ equipped with uniform norm. For a definition of this notion of differentiability, we refer to Römisch 2004. Moreover, since the support of empirical measures is always contained in the support of the underlying measure, we focus in the following only on differentiability *tangentially* to the set of probability measures whose support is contained in the support of μ and ν, respectively. This means that only those perturbations for μ and ν are considered which do not lead to an enlargement of the support.

Theorem 2 (Hadamard directional differentiability of OT cost): Suppose that for two Polish spaces $\mathcal{X}, \mathcal{Y}$ and the cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ the Assumption **(C)** combined with **(S1)** or **(S2)** are satisfied. Consider the function class $\mathcal{F}_c$ in (7) and two probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ with their respective support $\tilde{\mathcal{X}} := \mathrm{supp}(\mu)$ and $\tilde{\mathcal{Y}} := \mathrm{supp}(\nu)$. Then, the OT cost functional

$$\mathrm{OT}_c\colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \subseteq \ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c) \to \mathbb{R}, \\ (\mu, \nu) \mapsto \sup_{f \in \mathcal{F}_c} (\mu(f) + \nu(f^c)) \qquad (12)$$

is Hadamard directionally differentiable at (μ, ν) tangentially to the set of probability measures $\mathcal{P}(\tilde{\mathcal{X}}) \times \mathcal{P}(\tilde{\mathcal{Y}})$. The derivative is equal to

$$\begin{aligned}\mathcal{D}^H_{|(\mu,\nu)}\mathrm{OT}_c \colon \mathcal{T}_\mu\big(\mathcal{P}(\tilde{\mathcal{X}})\big) \times \mathcal{T}_\nu\big(\mathcal{P}(\tilde{\mathcal{Y}})\big) &\to \mathbb{R}, \\ (\Delta_\mu, \Delta_\nu) &\mapsto \sup_{f \in S_c(\mu,\nu)} \big(\Delta_\mu(f) + \Delta_\nu(f^c)\big).\end{aligned}$$

Herein, the contingent Bouligand cone $\mathcal{T}_\mu\big(\mathcal{P}(\tilde{\mathcal{X}})\big)$ to $\mathcal{P}(\tilde{\mathcal{X}})$ at μ is given by the topological closure $\mathrm{Cl}\big\{\frac{\mu'-\mu}{t} \text{ for } t > 0,\ \mu' \in \mathcal{P}(\tilde{\mathcal{X}})\big\} \subseteq \ell^\infty(\mathcal{F}_c)$ and analogously for $\mathcal{T}_\nu\big(\mathcal{P}(\tilde{\mathcal{Y}})\big)$.

Proof. The proof is inspired by Römisch 2004, Proposition 1, and Cárcamo et al. 2020, Theorem 2.1 and Corollary 2.2. Compared to their setting, the proof here is specifically tailored to the OT cost in (2) and exploits properties of the function class $\mathcal{F}_c$. We divide the proof into four steps. The first two essentially prove the Hadamard directional differentiability and simplify the representation of the derivative, whereas the last two are concerned with the convergence results for Kantorovich potentials.

Step 1. Hadamard directional differentiability. Let $(t_n)_{n\in\mathbb{N}}$ be a positive sequence with $t_n \searrow 0$ and take sequences $(\Delta_{\mu,n}, \Delta_{\nu,n}) \in \ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$ such that for all $n \in \mathbb{N}$ holds

$$\mu_n := \mu + t_n\Delta_{\mu,n} \in \mathcal{P}(\tilde{\mathcal{X}}), \quad \nu_n := \nu + t_n\Delta_{\nu,n} \in \mathcal{P}(\tilde{\mathcal{Y}}),$$

with $(\Delta_{\mu,n}, \Delta_{\nu,n}) \to (\Delta_\mu, \Delta_\nu) \in \mathcal{T}_\mu\big(\mathcal{P}(\tilde{\mathcal{X}})\big) \times \mathcal{T}_\nu\big(\mathcal{P}(\tilde{\mathcal{Y}})\big)$ for $n \to \infty$ in the space $\ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$. The representation of the Bouligand cone as a topological closure follows by an observation of Römisch 2004 since $\mathcal{P}(\tilde{\mathcal{X}})$ and $\mathcal{P}(\tilde{\mathcal{Y}})$ are convex sets. For μ, μ_n and ν, ν_n considered as bounded functionals on $\mathcal{F}_c$ and $\mathcal{F}_c^c$, respectively, it follows along the lines of Römisch 2004, Proposition 1 that the OT cost is Hadamard directionally differentiable with

$$\mathcal{D}^H_{|(\mu,\nu)}(\Delta_\mu, \Delta_\nu) = \lim_{n\to\infty} \frac{1}{t_n}\left(\mathrm{OT}_c(\mu_n, \nu_n) - \mathrm{OT}_c(\mu, \nu)\right) = \lim_{\varepsilon \searrow 0} \sup_{f \in S_c(\mu,\nu,\varepsilon)} (\Delta_\mu(f) + \Delta_\nu(f^c)),$$

where $S_c(\mu, \nu, \varepsilon)$ denotes the set of ε-approximizers

$$S_c(\mu, \nu, \varepsilon) := \{f \in \mathcal{F}_c \mid \mu(f) + \nu(f^c) \geq \mathrm{OT}_c(\mu, \nu) - \varepsilon\}.$$

Step 2. Simplifying the derivative. It remains to show that

$$\lim_{\varepsilon \searrow 0} \sup_{f \in S_c(\mu,\nu,\varepsilon)} \big(\Delta_\mu(f) + \Delta_\nu(f^c)\big) = \sup_{f \in S_c(\mu,\nu)} (\Delta_\mu(f) + \Delta_\nu(f^c)). \tag{13}$$

The left hand side in (13) is greater or equal to the right hand side since $S_c(\mu, \nu) \subseteq S_c(\mu, \nu, \varepsilon)$ for all $\varepsilon > 0$. For the converse, take a positive decreasing sequence $(\varepsilon_n)_{n\in\mathbb{N}}$

with $\varepsilon_n \searrow 0$ and a sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}_c$ with $f_n \in S_c(\mu, \nu, \varepsilon_n)$ for which

$$\sup_{f\in S_c(\mu,\nu,\varepsilon_n)} (\Delta_\mu(f) + \Delta_\nu(f^c)) - \varepsilon_n \leq \Delta_\mu(f_n) + \Delta_\nu(f_n^c).$$

As we prove below there exists a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ such that f_{n_k} and $f_{n_k}^c$ converge pointwise for $k \longrightarrow \infty$ on $\operatorname{supp}(\mu)$ and $\operatorname{supp}(\nu)$ to functions $h + a$ and $h^c - a$, respectively, for $h \in S_c(\mu, \nu)$ and $a \in \mathbb{R}$. Once we show that

$$\lim_{k\to\infty} (\Delta_\mu(f_{n_k}) + \Delta_\nu(f_{n_k}^c)) = (\Delta_\mu(h) + \Delta_\nu(h^c)), \tag{14}$$

equality (13) follows from

$$\begin{aligned}\lim_{\varepsilon\searrow 0} \sup_{f\in S_c(\mu,\nu,\varepsilon)} (\Delta_\mu(f) + \Delta_\nu(f^c)) &\leq \lim_{k\to\infty} \Delta_\mu(f_{n_k}) + \Delta_\nu(f_{n_k}^c) \\ &= (\Delta_\mu(h) + \Delta_\nu(h^c)) \leq \sup_{f\in S_c(\mu,\nu)} (\Delta_\mu(f) + \Delta_\nu(f^c)).\end{aligned}$$

To verify (14), let $\delta > 0$ and select $M \in \mathbb{N}$ such that $\left\|\Delta_\mu - \Delta_{\mu,M}\right\|_{\mathcal{F}_c} < \delta/4$. Pointwise convergence of f_{n_k} on $\operatorname{supp}(\mu)$ to $h + a$ combined with the uniform bound on $\mathcal{F}_c$ asserts by dominated convergence for μ_M, μ by $\operatorname{supp}(\mu_M) \subseteq \operatorname{supp}(\mu)$ existence of $K \in \mathbb{N}$ with

$$\left|\mu_M(f_{n_k} - (h + a))\right| + \left|\mu(f_{n_k} - (h + a))\right| < \delta t_M/2 \quad \forall k \geq K.$$

Hence, for all $k \geq K$ it follows that

$$\begin{aligned}|\Delta_\mu(f_{n_k}) - \Delta_\mu(h)| &\leq 2\left\|\Delta_\mu - \Delta_{\mu,M}\right\|_{\mathcal{F}_c} + |\Delta_{\mu,M}(f_{n_k}) - \Delta_{\mu,M}(h)| \\ &= 2\left\|\Delta_\mu - \Delta_{\mu,M}\right\|_{\mathcal{F}_c} + t_M^{-1}\left|(\mu_M - \mu)(f_{n_k} - h)\right| \\ &= 2\left\|\Delta_\mu - \Delta_{\mu,M}\right\|_{\mathcal{F}_c} + t_M^{-1}\left|(\mu_M - \mu)(f_{n_k} - (h - a))\right| < \delta,\end{aligned}$$

where we use in the second equality the definition of $\Delta_{\mu,M}$ and in the last equality that $(\mu - \mu_M)(a) = 0$. Repeating the argument for $|\Delta_\nu(f_{n_k}^c) - \Delta_\nu(h^c)|$ yields (14).

Step 3. Existence of converging subsequences. We prove existence of a subsequence of (f_n, f_n^c) that converges uniformly on compact sets to a pair of continuous functions (f, g). Uniform convergence on compact sets of continuous functions on $\mathcal{X}$ is induced by the compact-open topology which is metrizable since $\mathcal{X}$ is a locally compact Polish space (McCoy and Ntantu 1988, page 68). We show that $\mathcal{F}_c$ is relatively compact in the compact-open topology by means of a general version of the Arzelà-Ascoli theorem.

> Fact 1 (McCoy and Ntantu 1988, Theorem 3.2.6): If $\mathcal{X}$ is locally compact, then a set of continuous real-valued functions on $\mathcal{X}$ is compact in the compact-open topology if and only if it is closed, pointwise bounded and equicontinuous.

Since $\{c(\cdot, y) \mid y \in \mathcal{Y}\}$ is equicontinuous, so is $\mathcal{F}_c$ and its closure $\mathrm{Cl}(\mathcal{F}_c)$ in the compact-open topology. Moreover, since $\mathcal{F}_c$ is uniformly bounded by $\|c\|_\infty$, Fact 1 asserts the existence of a subsequence of f_n that converges uniformly on compact sets to a continuous function $f \in \mathrm{Cl}(\mathcal{F}_c)$. By restricting to a subsequence, we assume that f_n uniformly converges on compact sets to f. Under Assumption **(S2)**, this asserts that f_n uniformly converges on $\mathcal{X}$ to f and thus f_n^c uniformly converges on $\mathcal{Y}$ to $g := f^c$. Under Assumption **(S1)**, we instead repeat the above argument and assume by local compactness of $\mathcal{Y}$ and equicontinuity of $\{c(x, \cdot) \mid x \in \mathcal{X}\}$ that f_n^c uniformly converges on compact sets of $\mathcal{Y}$ to some continuous function $g \in \mathrm{Cl}(\mathcal{F}_c^c)$.

Step 4. Structural properties of limits with general OT theory. It remains to show that there exists a c-concave function $h \in S_c(\mu, \nu)$ and some $a \in \mathbb{R}$ such that $f = h + a$ on $\mathrm{supp}(\mu)$ and $g = h^c - a$ on $\mathrm{supp}(\nu)$. For this purpose, note that uniform convergence on compact sets implies pointwise convergence. Since f_n, f, f_n^c, g are all absolutely bounded by $\|c\|_\infty$, we find by dominated convergence that

$$\mu(f) + \nu(g) = \lim_{n\to\infty} \mu(f_n) + \nu(f_n^c) \geq \lim_{n\to\infty} \mathrm{OT}_c(\mu, \nu) - \varepsilon_n = \mathrm{OT}_c(\mu, \nu).$$

Moreover, for $(x, y) \in \mathcal{X} \times \mathcal{Y}$ we find that

$$f(x) + g(y) = \lim_{n\to\infty} f_n(x) + \lim_{n\to\infty} f_n^c(y) = \lim_{n\to\infty} f_n(x) + f_n^c(y) \leq c(x, y), \tag{15}$$

which asserts by the dual formulation for the OT cost (Villani 2008, Theorem 5.9) that $\mu(f) + \nu(g) = \mathrm{OT}_c(\mu, \nu)$. Upon defining $\tilde{h} := g^c$, we therefore obtain from (15) the inequalities $f \leq \tilde{h}$ on $\mathcal{X}$ and $g \leq \tilde{h}^c$ on $\mathcal{Y}$. Hence, it holds that

$$\mathrm{OT}_c(\mu, \nu) = \mu(f) + \nu(g) \leq \mu(\tilde{h}) + \nu(g) \leq \mu(\tilde{h}) + \nu(\tilde{h}^c) \leq \mathrm{OT}_c(\mu, \nu),$$

where the last inequality follows from $\tilde{h}(x) + \tilde{h}^c(y) \leq c(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We thus conclude that $f = \tilde{h}$ holds μ-almost surely and $g = \tilde{h}^c$ holds ν-almost surely.

Under Assumption **(S1)** it follows from step three that both f and $\tilde{h}$ as well as g and $\tilde{h}^c$ are continuous on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Thus, it holds (deterministically) that $f = \tilde{h}$ on $\mathrm{supp}(\mu)$ and $g = \tilde{h}^c$ on $\mathrm{supp}(\nu)$. Likewise, under Assumption **(S2)** it follows that f and $\tilde{h}$ are continuous on $\mathcal{X}$ which yields $f = \tilde{h}$ on $\mathrm{supp}(\mu)$. Further, from step three we know under **(S2)** that $g = f^c$ on $\mathcal{Y}$, i.e., g is c-concave and hence $g = g^{cc} = \tilde{h}^c$ on $\mathcal{Y}$ by Santambrogio 2015, Proposition 1.34. Finally, we note by Villani 2003, Remark 1.13, that any c-concave Kantorovich potential $\tilde{h}$ can be suitably shifted by a constant $a \in \mathbb{R}$ such that $0 \leq \tilde{h}(x) - a \leq \|c\|_\infty$ for all $x \in \mathcal{X}$ and $-\|c\|_\infty \leq (\tilde{h} - a)^c(y) = \tilde{h}^c(y) + a \leq 0$ for all $y \in \mathcal{Y}$. In particular, the function $h := \tilde{h} - a$ lies in $S_c(\mu, \nu)$ and fulfills the asserted properties. □

Remark 3: In the proof of Theorem 2, we cannot guarantee that the c-concave function $\tilde{h} = g^c$ lies in $S_c(\mu, \nu) \subseteq \mathcal{F}_c$ and therefore have to shift it by a suitable constant a. This is due to the inequalities $-\|c\|_\infty \leq g \leq 0$ being possibly violated.

3 Normal Limits under Unique Kantorovich Potentials

The set of Kantorovich potentials $S_c(\mu, \nu)$ in (3) and its c-conjugates $S_c^c(\mu, \nu)$ play a defining role for the limiting random variables in Theorem 1. A particular setting of interest arises if Kantorovich potentials are uniquely determined. However, since any $f \in S_c(\mu, \nu)$ can be shifted arbitrarily and $f + a$ for $a \in \mathbb{R}$ is still a c-concave function that solves (2), the set $S_c(\mu, \nu)$ can only be a singleton up to additive constants. Furthermore, it suffices if uniqueness holds almost surely.

> Definition 2 (Unique Kantorovich potentials): The Kantorovich potentials $S_c(\mu, \nu)$ in (3) are said to be *unique* if $f_1 - f_2$ for all $f_1, f_2 \in S_c(\mu, \nu)$ is constant μ-almost surely.

Notably, one can show that uniqueness of Kantorovich potentials $S_c(\mu, \nu)$ with respect to μ is in fact equivalent to uniqueness of the c-conjugate Kantorovich potentials $S_c^c(\mu, \nu)$ with respect to ν (Staudt et al. 2021, Lemma 5). Under this form of uniqueness, the limit random variables for the empirical OT cost simplify and follow a centered normal distribution.

> Theorem 3 (Normal limits): Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption **(C)** combined with **(S1)** or **(S2)**. Assume that the Kantorovich potentials $S_c(\mu, \nu)$ for $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ are unique and let $f \in S_c(\mu, \nu)$. Then, for empirical measures $\hat{\mu}_n, \hat{\nu}_m$ the following CLT is valid.

(i) (One-sample from μ) Suppose that $\mathcal{F}_c$ is μ-Donsker. Then, for $n \to \infty$,

$$\sqrt{n}\,(\mathrm{OT}_c(\hat{\mu}_n, \nu) - \mathrm{OT}_c(\mu, \nu)) \xrightarrow{\mathcal{D}} \mathbb{G}_\mu(f) \sim \mathcal{N}(0, \mathrm{Var}_{X\sim\mu}[f(X)]). \tag{16a}$$

(ii) (One-sample from ν) Suppose that $\mathcal{F}_c^c$ is ν-Donsker. Then, for $n \to \infty$,

$$\sqrt{m}\,(\mathrm{OT}_c(\mu, \hat{\nu}_m) - \mathrm{OT}_c(\mu, \nu)) \xrightarrow{\mathcal{D}} \mathbb{G}_\nu(f^c) \sim \mathcal{N}(0, \mathrm{Var}_{Y\sim\nu}[f^c(Y)]). \tag{16b}$$

(iii) (Two-sample) Suppose that $\mathcal{F}_c$ is μ-Donsker and that $\mathcal{F}_c^c$ is ν-Donsker. Then, for $n, m \to \infty$ with $m/(n+m) \to \delta \in (0,1)$,

$$\begin{aligned}\sqrt{\frac{nm}{n+m}}\,(\mathrm{OT}_c(\hat{\mu}_n, \hat{\nu}_m) - \mathrm{OT}_c(\mu, \nu)) &\xrightarrow{\mathcal{D}} \sqrt{\delta}\,\mathbb{G}_\mu(f) + \sqrt{1-\delta}\,\mathbb{G}_\nu(f^c) \\ &\sim \mathcal{N}\left(0, \delta\mathrm{Var}_{X\sim\mu}[f(X)] + (1-\delta)\mathrm{Var}_{Y\sim\nu}[f^c(Y)]\right).\end{aligned} \tag{16c}$$

Proof. Due to Staudt et al. 2021, Lemma 5 and the continuity of Kantorovich potentials in this setting (Lemma 1), we find that the set of Kantorovich potentials $S_c(\mu, \nu)$ and its c-conjugate counterparts $S_c^c(\mu, \nu)$ are deterministically unique (up to constant shifts) on the support of μ and ν, respectively. Since any $\Delta_\mu \in \mathcal{T}_\mu(\mathcal{P}(\tilde{\mathcal{X}}))$ fulfills by step two of the proof for Theorem 2 the (deterministic) equality $\Delta_\mu(f) = \Delta_\mu(f + a)$ for $f \in \mathcal{F}_c$ and $a \in \mathbb{R}$ with $f + a \in \mathcal{F}_c$, and likewise for $\Delta_\nu \in \mathcal{T}_\nu(\mathcal{P}(\tilde{\mathcal{Y}}))$, uniqueness of Kantorovich potentials implies linearity of the directional Hadamard derivative. The functional delta method (Römisch 2004) thus implies the weak limit for the empirical OT cost to be centered normal with variance as stated in (16). □

A useful statistical application of Theorem 3 arises when the variance $\mathrm{Var}_{X\sim\mu}[f(X)]$ in Theorem 3 can consistently be estimated from i.i.d. data. Indeed, this holds under Assumption **(C)** combined with **(S1)** or **(S2)**, leading to the following pivotal limit law.

Corollary 1 (Pivotal limit law): If the unique Kantorovich potential $f \in S_c(\mu, \nu)$ in Theorem 3 is not constant μ-almost surely, then, for $n \longrightarrow \infty$ and any $f_n \in S_c(\hat{\mu}_n, \nu)$,

$$\sqrt{n}\,\frac{\mathrm{OT}_c(\hat{\mu}_n, \nu) - \mathrm{OT}_c(\mu, \nu)}{\sqrt{\mathrm{Var}_{X\sim\hat{\mu}_n}[f_n(X)]}} \xrightarrow{\mathcal{D}} \frac{\mathbb{G}_\mu(f)}{\sqrt{\mathrm{Var}_{X\sim\mu}[f(X)]}} \sim \mathcal{N}(0, 1)\,. \tag{17}$$

Analogous statements hold for the weak limits in (16b) if $f^c \in S_c^c(\mu, \nu)$ is not constant ν-almost surely and in (16c) if f or f^c is not constant μ-respectively ν-almost surely.

Proof of Corollary 1. We only show the claim for (17), the corresponding pivotal limits for (16b) and (16c) follow analogously. In view of Theorem 3 and Slutzky's lemma it suffices to show for $f_n \in S_c(\hat{\mu}_n, \nu)$ that $\mathrm{Var}_{X\sim\hat{\mu}_n}[f_n(X)]$ converges almost surely for $n \longrightarrow \infty$ to $\mathrm{Var}_{X\sim\mu}[f(X)] > 0$ for $f \in S_c(\mu, \nu)$. Note that $S_c(\hat{\mu}_n, \nu) \subseteq S_c(\mu, \nu, 2\,\|\hat{\mu}_n - \mu\|_{\mathcal{F}_c})$ where $\|\hat{\mu}_n - \mu\|_{\mathcal{F}_c}$ tends to zero almost surely since $\mathcal{F}_c$ is μ-Donsker. Hence, by steps three and four of the proof for Theorem 2, it follows that there exists a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ such that $(f_{n_k}, f_{n_k}^c)$ converges pointwise on $\mathrm{supp}(\mu) \times \mathrm{supp}(\nu)$ for $k \longrightarrow \infty$ to $(h + a, h^c - a)$ for some $h \in S_c(\mu, \nu)$ and some $a \in \mathbb{R}$. Since $\mathcal{F}_c$ is uniformly bounded by $\|c\|_\infty$, and since $\mathcal{F}_c$ as well as the element-wise squared function class $\mathcal{F}_c^2 = \{f^2 \colon f \in \mathcal{F}_c\}$ are both μ-Donsker (Van der Vaart and Wellner 1996, Theorem 2.10.6), we conclude that $\mathrm{Var}_{X\sim\hat{\mu}_{n_k}}[f_{n_k}(X)] \longrightarrow \mathrm{Var}_{X\sim\mu}[h(X)]$ almost surely for $k \longrightarrow \infty$. Finally, by almost sure uniqueness of Kantorovich potentials it holds that $\mathrm{Var}_{X\sim\mu}[h(X)] = \mathrm{Var}_{X\sim\mu}[f(X)]$. □

Uniqueness of Kantorovich potentials The recent work by Staudt et al. 2021 provides a detailed account on uniqueness guarantees of Kantorovich potentials. The gist of their article is that uniqueness in settings with differentiable costs will commonly hold, and that non-differentiabilities, erratic mass placement, or specific degeneracies in disconnected spaces have to be assumed for uniqueness to fail.

In the connected setting, one relies on the fact that the gradient of a Kantorovich potential at some point x is determined by the gradient of $c(\cdot, y)$ at x for any $(x, y) \in \operatorname{supp}(\pi)$, where π is an OT plan, i.e., an optimizer of (1). The connectedness of the support of μ combined with the uniqueness of the gradient then imply uniqueness of the potential. This approach was employed by Staudt et al. 2021 in case of probability measures with possibly unbounded support that assign negligible mass to the boundary of their support and differentiable cost functions that grow sufficiently rapidly. Notably, this encompasses the strictly convex costs considered by Gangbo and McCann 1996, for which uniqueness guarantees were also derived by del Barrio et al. 2021, Theorem 2.4 and Bernton et al. 2021, Theorem B.2. In discrete settings, the theory of linear programming shows Kantorovich potentials to be unique if the measures μ and ν are non-degenerate, which (loosely speaking) means that the OT problem cannot be divided into proper sub-problems. Staudt et al. 2021 show that similar results hold for probability measures with disconnected support on Polish spaces.

4 Degenerate Limits under Trivial Kantorovich Potentials

Another important special case for our CLTs emerges if Kantorovich potentials are not only unique but also constant. The asymptotic variance in Theorem 3 is then equal to zero and the limit law degenerates. As for uniqueness, it suffices if Kantorovich potentials are constant in an almost sure sense.

Definition 3 (Trivial Kantorovich potentials): A Kantorovich potential $f \in S_c(\mu, \nu)$ is called *trivial* if it is constant μ-almost surely. The set $S_c(\mu, \nu)$ is said to be trivial if all of its elements are trivial.

The same definition applies for conjugated potentials f^c and the set $S_c^c(\mu, \nu)$ with respect to ν. In contrast to uniqueness of Kantorovich potentials, triviality of $S_c(\mu, \nu)$ does not imply the triviality of $S_c^c(\mu, \nu)$, since a μ-almost surely constant f can (and often will) have a c-conjugate f^c that is not ν-almost surely constant (see also Remark 5 below). Simple examples with trivial Kantorovich potentials occur if one of the measures μ or ν is a Dirac measure or if the cost function is itself constant. More general and interesting settings will be discussed below.

Theorem 4 (Degenerate limits): Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption **(C)** combined with **(S1)** or **(S2)**. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be probability measures.

(i) (One-sample from μ) Suppose that $\mathcal{F}_c$ is μ-Donsker. Then, the limit law

(10a) degenerates to a Dirac measure if and only if $S_c(\mu,\nu)$ is trivial.

(ii) (One-sample from ν) Suppose that $\mathcal{F}_c^c$ is ν-Donsker. Then, the limit law (10b) degenerates to a Dirac measure if and only if $S_c^c(\mu,\nu)$ is trivial.

(iii) (Two-sample) Suppose that $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are μ- and ν-Donsker, respectively. Then, the limit law (10c) degenerates to a Dirac measure if and only if $S_c(\mu,\nu)$ and $S_c^c(\mu,\nu)$ are both trivial.

Proof. We only state the proof for setting (i), the remaining cases follow analogously. If $S_c(\mu,\nu)$ is trivial, then the limit distribution degenerates by Theorem 3(i) to a Dirac measure at zero. In case $S_c(\mu,\nu)$ is not trivial, there exists a Kantorovich potential $f \in S_c(\mu,\nu)$ with $\mathrm{Var}_{X\sim\mu}[f(X)] > 0$. As the limit distribution for the one-sample case from μ stochastically dominates $\mathcal{N}(0,\mathrm{Var}_{X\sim\mu}[f(X)])$ it does not degenerate to a Dirac measure. □

The existence of trivial Kantorovich potentials $f \in S_c(\mu,\nu)$ is intimately related to the existence of transport plans that act as projections onto the support of μ. To highlight the underlying geometric interpretation, we introduce the following notion of projected measures.

Definition 4 (Projected measures): Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be probability measures. We say that μ is a *ν-projected measure* (with respect to c) and write $\mu \in P_c(\nu)$ if there exists a coupling $\pi \in \Pi(\mu,\nu)$ such that

$$c(x,y) = \inf_{x'\in\mathrm{supp}(\mu)} c(x',y) \qquad \text{for all } (x,y) \in \mathrm{supp}(\pi). \tag{18}$$

Analogous definitions apply for μ-projected measures, which we denote by $\nu \in P_c(\mu)$. It can easily be verified that any coupling π satisfying (18) solves the OT problem in (1) between μ and ν. In fact, under continuous costs, it holds that (see Appendix A for a proof)

$$\mu \in P_c(\nu) \qquad \Longleftrightarrow \qquad \mathrm{OT}_c(\mu,\nu) = \int_{\mathcal{Y}} \inf_{x\in\mathrm{supp}(\mu)} c(x,y)\,\mathrm{d}\nu(y), \tag{19}$$

which can equivalently be used to characterize ν-projected measures.

Theorem 5 (Existence of trivial Kantorovich potentials): Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption **(C)** combined with **(S1)** or **(S2)**, and let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be probability measures. Then, trivial Kantorovich potentials $f \in S_c(\mu,\nu)$ exist if and only if $\mu \in P_c(\nu)$.

Remark 4: In settings where we can assume unique Kantorovich potentials, which is often not a restrictive condition (see Section 3), Theorem 5 in con-

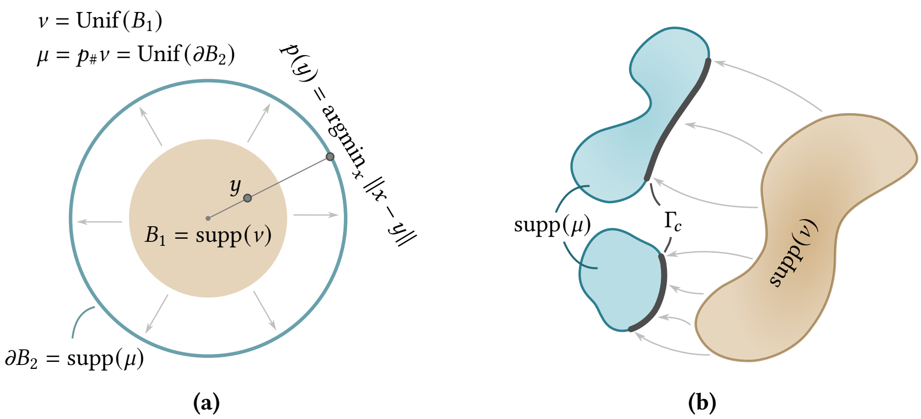

$\nu = \mathrm{Unif}(B_1)$
$\mu = p_{\#}\nu = \mathrm{Unif}(\partial B_2)$
$p(y) = \mathrm{argmin}_x \|x - y\|$
y
$B_1 = \mathrm{supp}(\nu)$
$\partial B_2 = \mathrm{supp}(\mu)$
(a)
$\mathrm{supp}(\mu)$
Γ_c
$\mathrm{supp}(\nu)$
(b)

Figure 1: Trivial Kantorovich potentials. In **(a)**, μ is equal to the projection of ν onto the set $\tilde{\mathcal{X}} = \partial B_2$, the boundary of an Euclidean ball. According to Corollary 2 (ii), there exists a Kantorovich potential f that is constant on ∂B_2. Furthermore, due to Staudt et al. 2021, Corollary 2, the constant potential is also unique. In **(b)**, the projection $\Gamma_c = \Gamma_c(\mu, \nu)$ of all points $y \in \mathrm{supp}(\nu)$ onto the support of μ under Euclidean costs is marked by black segments that are part of the boundary of $\mathrm{supp}(\mu)$. Since $\mathrm{supp}(\mu)$ is not contained in Γ_c, Corollary 3 (iii) states that there cannot be a Kantorovich potential f that is constant on $\mathrm{supp}(\mu)$.

junction with Theorem 4 states that the one-sample limit laws (10a) and (10b) degenerate if and only if $\mu \in P_c(\nu)$ or $\nu \in P_c(\mu)$, respectively. Furthermore the two-sample limit law (10c) degenerates if and only if $\mu \in P_c(\nu)$ and $\nu \in P_c(\mu)$ (see Remark 5 below).

Intuitively, a ν-projected measure is any measure that can be obtained from ν by projecting all points of $\mathrm{supp}(\nu)$ to some subset of $\mathcal{X}$ according to the cost c. For example, $\mu \in P_c(\nu)$ always holds if $\mu = p_{\#}\nu$ for a measurable map $p\colon \mathrm{supp}(\nu) \to \mathcal{X}$ that obeys

$$p(y) \in \operatorname*{argmin}_{x \in \mathrm{supp}(\mu)} c(x, y) \tag{20}$$

for each $y \in \mathrm{supp}(\nu)$. We denote maps p that satisfy (20) for given μ and ν as *c-projections*. One example of a c-projection is illustrated in Figure 1(a). Based on Theorem 5, we next formulate several necessary and sufficient criteria for the existence of trivial Kantorovich potentials, all of which have an intuitive geometric interpretation.

Corollary 2: Under the assumptions of Theorem 5, the set $S_c(\mu, \nu)$ contains trivial Kantorovich potentials if one of the following conditions is satisfied.

(i) $\mu = \nu$ and $c(x, x) = 0$ holds for all $x \in \mathcal{X}$,

(ii) $\mu = p_{\#}\nu$ for a c-projection $p : \mathrm{supp}(\nu) \to \tilde{\mathcal{X}}$ onto a closed set $\tilde{\mathcal{X}} \subset \mathcal{X}$.

Corollary 3: Under the assumptions of Theorem 5, the set $S_c(\mu, \nu)$ does not contain trivial Kantorovich potentials if one of the following conditions is satisfied.

(i) $\mu \neq \nu$ while $\mathrm{supp}(\nu) \subset \mathrm{supp}(\mu)$ with $c(x, x) = 0$ and $c(x, y) > 0$ for all $x \neq y$,

(ii) $\mathrm{int}(\mathrm{supp}(\mu)) \not\subseteq \mathrm{supp}(\nu)$ for subsets $\mathcal{X}, \mathcal{Y} \subset V$ of a normed linear space $(V, \|\cdot\|)$ with cost $c(x, y) = h(\|x - y\|)$ for strictly increasing h,

(iii) $\mathrm{supp}(\mu)$ is not equal to the topological closure $\Gamma_c(\mu, \nu) \subset \mathcal{X}$ of the set

$$\bigcup_{y \in \mathrm{supp}(\nu)} \operatorname*{argmin}_{x \in \mathrm{supp}(\mu)} c(x, y).$$

In each of these settings, the limit law (10a) in Theorem 1 is non-degenerate.

Exchanging the roles of μ and ν, Theorem 5 as well as Corollaries 2 and 3 can equally be applied to f^c. Examples illustrating Corollary 2 (ii) and Corollary 3 (iii) in Euclidean settings are provided in Figure 1. In particular, Figure 1(a) highlights that there is a crucial difference between demanding constant Kantorovich potentials on all of $\mathcal{X}$ and only demanding them to be constant almost surely. Indeed, if $f \in S_c(\mu, \nu)$ was constant on the whole space $\mathcal{X}$, then $f^c \in S_c^c(\mu, \nu)$ would be constant as well, which is by Theorem 5 not the case in Figure 1(a). This setting also serves as an example where $\mu \neq \nu$ and where every Kantorovich potential $f \in S_c(\mu, \nu)$ is trivial while f^c is not.

Remark 5 (Bi-triviality): OT problems where both $f \in S_c(\mu, \nu)$ and its conjugate f^c are trivial have a special underlying geometry. If f is constant on $\mathrm{supp}(\mu)$ and f^c is constant on $\mathrm{supp}(\nu)$, then the optimality condition $f(x) + f^c(y) = c(x, y)$ for points (x, y) on the support of any optimal transport plan π implies that c is constant on $\mathrm{supp}(\pi)$, i.e., all points are transported with the same cost. For an example, consider radially symmetric costs and uniform distributions μ and ν on centered Euclidean spheres of positive radius. For different radii, the measures are distinct and both Kantorovich potentials are trivial.

5 Examples

Elaborating the main result in Theorem 1, we focus in this section on concrete settings. The general setup requires Assumption **(C)** combined with **(S1)** or **(S2)** to hold. It remains to verify the Donsker properties for the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ which then leads to CLTs for the empirical OT cost.

5.1 Countable Discrete Spaces

Our general theory leads to CLTs for the OT cost on countable discrete spaces $\mathcal{X}$ and $\mathcal{Y}$ equipped with discrete topology. More precisely, for a non-negative cost function $c\colon \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ bounded by some constant $\|c\|_\infty < \infty$, Assumptions **(C)** and **(S1)** hold trivially. To prove that the function class $\mathcal{F}_c$ is μ-Donsker, we define $\mathcal{F}_c \mathbb{1}_x$ as the restriction of $\mathcal{F}_c$ to a fixed element $x \in \mathcal{X}$. Notably, each element in the latter function class is bounded by $\|c\|_\infty$ and we obtain

$$\mathbb{E}\left[\|\sqrt{n}\,(\hat{\mu}_n - \mu)\,\|_{\mathcal{F}_c \mathbb{1}_x}\right] \lesssim_{\|c\|_\infty} \sqrt{\mu(x)}.$$

According to Van der Vaart and Wellner 1996, Theorem 2.10.24, we deduce that the function class $\mathcal{F}_c$ is μ-Donsker if

$$\sum_{x \in \mathcal{X}} \sqrt{\mu(x)} < \infty \tag{21}$$

which is the celebrated *Borisov-Dudley-Durst* condition (Dudley 2014). Similarly, the function class $\mathcal{F}_c^c$ is ν-Donsker if $\sum_{y \in \mathcal{Y}} \sqrt{\nu(y)} < \infty$.

Corollary 4 (Countable discrete spaces, Tameling et al. 2019): Let $\mathcal{X}$ and $\mathcal{Y}$ be countable discrete spaces and $c\colon \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ a bounded cost function. Consider probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. If μ fulfills the Borisov-Dudley-Durst condition (21), then the CLT in (10a) is valid. If ν fulfills (21), then (10b) holds. In case both μ and ν fulfill (21), then (10c) holds.

5.2 The One-dimensional Euclidean Space

From Theorem 1 we immediately derive CLTs for the empirical OT cost between probability measures supported on the real line $\mathbb{R}$ and sufficiently regular cost function. The proofs require the notion of covering and metric entropy for a real-valued function class $\mathcal{F}$ defined on $\mathcal{X}$ as introduced in the notation. Based on metric entropy bounds, empirical process theory provides tools to assess if a given function class is μ-Donsker or even universal Donsker (Van der Vaart and Wellner 1996, Section 2.5). We first focus on the real line.

Theorem 6 ($d = 1$): Consider the Euclidean space $\mathbb{R}$ with cost $c\colon \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_+$ assumed to be bounded and (α, L)-Hölder for $\alpha \in (1/2, 1]$ and $L \geq 0$, i.e.,

$$|c(x, y) - c(x', y')| \leq L\left(|x - x'|^\alpha + |y - y'|^\alpha\right) \quad \forall x, x', y, y' \in \mathbb{R}. \tag{22}$$

If the probability measure $\mu \in \mathcal{P}(\mathbb{R})$ fulfills

$$\sum_{k \in \mathbb{Z}} \sqrt{\mu\left([k, k+1)\right)} < \infty, \tag{23}$$

then the CLT in (10a) is valid. If $\nu \in \mathcal{P}(\mathbb{R})$ fulfills (23), then (10b) holds. In case both μ and ν fulfill (23), then (10c) holds.

Proof. By the assumptions imposed on the cost and since the setting focuses on the real line $\mathbb{R}$, the Assumptions **(C)** and **(S1)** hold. Based on Theorem 1, it remains to prove the Donsker properties of the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ in this case. By Lemma 2 (i), the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are both contained in the class of uniformly bounded (α, L)-Hölder functions on $\mathbb{R}$. According to Van der Vaart and Wellner 1996, Example 2.10.25, with $M_k = L$ for all $k \in \mathbb{Z}$ this class is μ-Donsker if (23) is fulfilled. □

Example 1 (Costs $c(x, y) = h(x-y)$, $d = 1$): To demonstrate a few consequences of Theorem 6, suppose the cost is equal to $c(x, y) = h(x - y)$ on $\mathbb{R}$ for a suitable function $h\colon \mathbb{R} \longrightarrow \mathbb{R}_+$. For instance, under a bounded (α, L)-Hölder function $h\colon \mathbb{R} \longrightarrow \mathbb{R}_+$ with $\alpha > 1/2$ and $L > 0$ condition (22) is satisfied. Hence for $\mu, \nu \in \mathcal{P}(\mathbb{R})$ that fulfill condition (23) Theorem 6 yields CLTs for the empirical OT cost. This setting encompasses, e.g., thresholded costs $c_{p,T}(x, y) = \min(|x - y|^p, T)$ for $p > \alpha$ and $T > 0$. Notably, for $p = T = 1$, the function class $\mathcal{F}_c$ coincides by Kantorovich-Rubinstein duality with $\mathrm{BL}_1(\mathbb{R})$, the class of 1-Lipschitz functions uniformly bounded by one, for which condition (23) is necessary and sufficient in order to be μ-Donsker (Giné and Zinn 1986, Theorem 1). For this case it holds for $n \longrightarrow \infty$ that

$$\sqrt{n}\mathrm{OT}_{c_{1,1}}(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} \sup_{f \in \mathrm{BL}_1(\mathbb{R})} \mathbb{G}_\mu(f),$$

where the limit distribution only degenerates if μ is a Dirac measure.

When the probability measures μ, ν are compactly supported, the summability condition (23) is trivially fulfilled and by restricting to the support of the probability measures it suffices that h is only *locally* α-Hölder for $\alpha \in (1/2, 1]$. This provides for costs $c(x, y) = |x - y|^p$ with $p > 1/2$ novel CLTs for the empirical OT cost. In particular, if the support of μ is disconnected, then Staudt et al. 2021, Lemma 11 assert existence of non-trivial potentials for $S_c(\mu, \mu)$ which implies the resulting limit distribution of $\sqrt{n}\mathrm{OT}_c(\hat{\mu}_n, \mu)$ to be non-degenerate (Theorem 4). In contrast, if $\mathrm{supp}(\mu)$ is the closure of an open connected set for $p > 1$ Kantorovich potentials $S_c(\mu, \mu)$ are unique (Staudt et al. 2021, Corollary 2) and trivial and thus the respective limit distribution degenerates, indicating a faster convergence rate (see e.g. for $p = 2$ the CLT by del Barrio et al. 2005 which requires additional regularity assumptions on a density of μ and scaling by n).

Example 2 (Kantorovich-Rubinstein duality): For Euclidean costs $c(x, y) = |x - y|$ and a compactly supported probability measure μ set $\mathcal{X} = \mathcal{Y} = \mathrm{supp}(\mu) \subseteq \mathbb{R}$. Then, the set $\mathcal{F}_c = S_1(\mu, \mu) = \mathrm{Lip}_1(\mathcal{X})$ consists of 1-Lipschitz functions on $\mathcal{X}$ which are absolutely bounded by the diameter of $\mathcal{X}$. In particular, we obtain for $n \longrightarrow \infty$ that

$$\sqrt{n}\mathrm{OT}_1(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} \sup_{f \in \mathrm{Lip}(\mathcal{X})} \mathbb{G}_\mu(f).$$

The approach presented here is based on the dual formulation of the OT cost. Using the primal perspective, which provides for $p = 1$ an explicit formula for $\mathrm{OT}_1(\cdot,\cdot)$, del Barrio et al. 1999 derived a CLT that also holds for μ with non-compact support (see also Mason 2016). If F_μ denotes the cumulative distribution function of μ such that

$$\int_{-\infty}^{\infty} \sqrt{F_\mu(t)(1 - F_\mu(t))}\,\mathrm{d}t < \infty,$$

then, for $\mathbb{B}_\mu(t) = \mathbb{B}(F_\mu(t))$ with $\mathbb{B}(t)$ a standard Brownian bridge and $n \to \infty$, it holds

$$\sqrt{n}\mathrm{OT}_1(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} \int_{-\infty}^{\infty} |\mathbb{B}_\mu(t)|\,\mathrm{d}t. \tag{24}$$

Indeed, by suitably coupling the Gaussian processes $\mathbb{G}_\mu$ and $\mathbb{B}_\mu$ and approximating the respective random element in terms of an unsigned measure, an application of Fubini's theorem shows for compactly supported μ with $\mathcal{X} = \mathrm{supp}(\mu)$ that

$$\int_{-\infty}^{\infty} |\mathbb{B}_\mu(t)|\,\mathrm{d}t \overset{\mathcal{D}}{=} \sup_{f \in \mathrm{Lip}(\mathcal{X})} \mathbb{G}_\mu(f).$$

5.3 The Two- and Three-dimensional Euclidean Space

Theorem 1 also characterizes the limit law for the empirical OT cost beyond the real line. For Euclidean spaces with dimension $d = 2$ or $d = 3$ we obtain the following novel results.

Theorem 7 ($d = 2, 3$): Consider the Euclidean space $\mathbb{R}^d$ for $d = 2$ or $d = 3$ with cost $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ assumed to be bounded and L-Lipschitz[a]. Further, suppose there exists some $\Lambda > 0$ such that for all $k \in \mathbb{Z}^d$ there exist $x_k, y_k \in [k, k+1)$ such that

$$\begin{aligned} &c(\cdot, y) - \Lambda \,\|\cdot - x_k\|_2^2 \text{ is concave on } [k, k+1) \text{ for all } y \in \mathbb{R}^d,\\ &c(x, \cdot) - \Lambda \,\|\cdot - y_k\|_2^2 \text{ is concave on } [k, k+1) \text{ for all } x \in \mathbb{R}^d. \end{aligned} \tag{25}$$

If the probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ fulfills

$$\sum_{k \in \mathbb{Z}^d} \sqrt{\mu\left([k, k+1)\right)} < \infty, \tag{26}$$

then the CLT in (10a) is valid. If $\nu \in \mathcal{P}(\mathbb{R}^d)$ fulfills (26), then (10b) holds. In case both μ and ν fulfill (26), then (10c) holds.

[a] This refers to the cost function being $(1, L)$-Hölder, recall (22).

Proof. By the assumptions imposed on the cost and since the setting focuses on the Euclidean space $\mathbb{R}^d$, the Assumptions **(C)** and **(S1)** hold. Based on the main Theorem 1, it remains to prove the Donsker properties of the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ in this case. For a convex bounded set $\Omega \subseteq \mathbb{R}^d$ and constants $K, L > 0$, let $\mathcal{C}_{K,L}(\Omega)$ be the class of concave L-Lipschitz functions which are absolutely bounded by K. In order to prove that $\mathcal{F}_c$ is μ-Donsker if (26) holds, we employ Van der Vaart and Wellner 1996, Theorem 2.10.24. To this end, consider a partition $\mathbb{R}^d = \bigcup_{k\in\mathbb{Z}^d}[k, k+1)$ and define the function class $\mathcal{F}_{c,k} := \mathcal{F}_c \mathbb{1}_{[k,k+1)}$. We first verify that each class $\mathcal{F}_{c,k}$ is μ-Donsker. First note for $k \in \mathbb{Z}^d$ and any element $f \in \mathcal{F}_c$ that $f(\cdot) - \Lambda \|\cdot - x_k\|_2^2$ is bounded, Lipschitz and concave on $[k, k+1)$ (see Lemma 2 (ii)). More precisely, for any $f \in \mathcal{F}_c$ and since $x_k \in [k, k+1)$ it follows

$$\left(f(\cdot) - \Lambda \|\cdot - x_k\|_2^2\right)\Big|_{[k,k+1)} \in \mathcal{C}_{\kappa,l}\left([k, k+1)\right)$$

with $\kappa := (\|c\|_\infty + \Lambda d)$ and $l := (L + 2\Lambda d)$. According to Bronshtein 1976, Theorem 6[7], we conclude for $\varepsilon > 0$ sufficiently small

$$\log\left(\mathcal{N}(\varepsilon, \mathcal{F}_{c,k}, \|\cdot\|_\infty)\right) \leq \log\Big(\mathcal{N}(\varepsilon, \mathcal{C}_{\kappa,l}([k, k+1)), \|\cdot\|_{\infty,[k,k+1)})\Big) \lesssim_{\kappa,l,d} \varepsilon^{-d/2}.$$

Note that the function class $\mathcal{F}_{c,k}$ has envelope function $F_{c,k}(\cdot) := \|c\|_\infty \mathbb{1}_{[k,k+1)}(\cdot)$. Furthermore, an $\varepsilon\|c\|_\infty$-covering $\{f_1, \dots, f_N\}$ of $\mathcal{F}_{c,k}$ with respect to $\|\cdot\|_\infty$ defines for any finitely supported probability measure $\gamma \in \mathcal{P}(\mathbb{R}^d)$ with $\left\|F_{c,k}\right\|_{2,\gamma} > 0$ an $\varepsilon\left\|F_{c,k}\right\|_{2,\gamma}$-covering for $\mathcal{F}_{c,k}$ with respect to $\|\cdot\|_{2,\gamma}$. Indeed, for $f \in \mathcal{F}_{c,k}$ pick f_i such that $\|f - f_i\|_\infty < \varepsilon\|c\|_\infty$ which yields

$$\|f - f_i\|_{2,\gamma} \leq \sqrt{\int_{[k,k+1)} \varepsilon^2 \|c\|_\infty^2 \, \mathrm{d}\gamma} = \varepsilon\|c\|_\infty \sqrt{\gamma([k, k+1))} = \varepsilon\left\|F_{c,k}\right\|_{2,\gamma}.$$

Hence, we conclude for any finitely supported $\gamma \in \mathcal{P}(\mathbb{R}^d)$ with $\left\|F_{c,k}\right\|_{2,\gamma} > 0$ and sufficiently small ε that

$$\log\Big(\mathcal{N}(\varepsilon\left\|F_{c,k}\right\|_{2,\gamma}, \mathcal{F}_{c,k}, \|\cdot\|_{2,\gamma})\Big) \leq \log\left(\mathcal{N}(\varepsilon\|c\|_\infty, \mathcal{F}_{c,k}, \|\cdot\|_\infty)\right) \lesssim_{\kappa,l,d} \varepsilon^{-d/2}.$$

After taking square roots the latter bound is integrable around zero for $d \leq 3$ which yields the μ-Donsker property for $\mathcal{F}_{c,k}$ (Van der Vaart and Wellner 1996, Theorem 2.5.2). The μ-Donsker property of the whole $\mathcal{F}_c$ now follows if

$$\sup_{n\in\mathbb{N}} \sum_{k\in\mathbb{Z}^d} \mathbb{E}\left[\left\|\sqrt{n}(\hat{\mu}_n - \mu)\right\|_{\mathcal{F}_{c,k}}\right] < \infty. \tag{27}$$

[7] The work by Bronshtein 1976 in fact only provides metric entropy bounds for *convex* bounded Lipschitz functions on a cube but of course they remain valid for concave functions.

By standard chaining arguments each individual summand can be bounded by

$$\begin{aligned}\mathbb{E}\left[\left\|\sqrt{n}(\hat{\mu}_n-\mu)\right\|_{\mathcal{F}_{c,k}}\right] &\leq \int_0^1 \sup_{\gamma} \sqrt{1+\log\left(\mathcal{N}(\varepsilon \left\|F_{c,k}\right\|_{2,\gamma}, \mathcal{F}_{c,k}, \|\cdot\|_{2,\gamma})\right)}\, \mathrm{d}\varepsilon \left\|F_{c,k}\right\|_{2,\mu} \\ &\leq \int_0^1 \sqrt{1+\log\left(\mathcal{N}(\varepsilon \|c\|_\infty, \mathcal{F}_{c,k}, \|\cdot\|_\infty)\right)}\, \mathrm{d}\varepsilon \left\|F_{c,k}\right\|_{2,\mu} \\ &\lesssim_{\kappa,l,d} \int_0^1 \varepsilon^{-d/4}\, \mathrm{d}\varepsilon \left\|F_{c,k}\right\|_{2,\mu} \lesssim_d \left\|F_{c,k}\right\|_{2,\mu} = \|c\|_\infty \sqrt{\mu([k,k+1))}.\end{aligned}$$

Herein, the supremum runs over all finitely supported probability measures over $\mathbb{R}^d$ which leads to the claimed upper bound by our previous arguments. Summing over $k \in \mathbb{Z}^d$ and provided μ fulfills (26) yields (27). □

The summability constraints (23) and (26) are reminiscent of the Borisov-Dudley-Durst condition (21). Indeed, they naturally appear by partitioning the Euclidean space $\mathbb{R}^d = \bigcup_{k\in\mathbb{Z}^d}[k,k+1)$ and controlling the empirical process indexed over the respective function class restricted to individual partitions (Van der Vaart and Wellner 1996, Theorem 2.10.24). For the latter, the proof of Theorem 7 exploits well-known metric entropy bounds for the class of α-Hölder and concave, Lipschitz functions, respectively. Notably, crucial to CLTs for dimension $d = 2, 3$ is condition (25) that enables suitable upper bounds for the metric entropy of $\mathcal{F}_c$ and $\mathcal{F}_c^c$.

Remark 6 (On the assumptions): A few words regarding the required assumptions for the CLTs of this subsection are in order.

(i) The partition of $\mathbb{R}^d = \bigcup_{k\in\mathbb{Z}^d}[k,k+1)$ by regular cubes is arbitrary and any partition of convex, bounded sets $I_k \subset \mathbb{R}^d$ with non-empty interior such that $\sup_k \operatorname{diam}(I_k) < \infty$ serves to derive the same conclusion.

(ii) Condition (25) is fulfilled if the cost function is twice continuously differentiable in both components with a uniform bound $K > 0$ on the Eigenvalues of its Hessian. In this setting the bound from (25) is valid for $\Lambda = K/2$.

(iii) The summability constraints (23) and (26) are well-known in the context of empirical process theory (Van der Vaart and Wellner 1996, Section 2.10.4). A sufficient condition is given in terms of finite moments $\mathbb{E}\left[\|X\|_\infty^{2d+\delta}\right] < \infty$ for some $\delta > 0$ since

$$\begin{aligned}\sum_{k\in\mathbb{Z}^d} \sqrt{\mu([k,k+1))} &\lesssim 2^d \sum_{n=1}^{\infty} \sqrt{n^{2d-2}\mathbb{P}(\|X\|_\infty \geq n)} \\ &\leq 2^d \sqrt{\mathbb{E}\left[\|X\|_\infty^{2d+\delta}\right]} \sum_{n=1}^{\infty} n^{-(1+\delta/2)},\end{aligned}$$

where the latter inequality follows by Markov's inequality. Notably, for compactly supported measures the condition is vacuous.

Example 3 (Costs $c(x,y) = h(x-y)$, $d = 2,3$): Let us provide a few consequences of Theorem 7 by taking costs $c(x,y) = h(x-y)$ for some suitable function $h\colon \mathbb{R}^d \to \mathbb{R}_+$ with $d \in \{2,3\}$. If h is a bounded and twice continuously differentiable function with uniformly bounded first and second derivatives, then the required conditions on the cost in Theorem 7 are fulfilled. Moreover, since the pointwise minimum of bounded, Lipschitz, semi-concave functions with bounded modulus also exhibits these properties, we find that Theorem 7 also covers thresholded costs $c_{p,T}(x,y) = \min(\|x-y\|^p, T)$ for $p \geq 2$ and $T > 0$. Costs for the canonical flat torus $\tilde{c}(x,y) = \min_{z \in \mathbb{Z}^d} h(x-y-z)$, which have been considered by González-Delgado et al. 2021 for $h(x) = \|x\|^2$, also fulfill these conditions. For these cases, Theorem 7 provides novel CLTs for $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ if the summability constraint (26) is satisfied. Moreover, if μ and ν are both compactly supported, condition (26) is vacuous and it suffices that h is twice continuously differentiable, hence, our theory encompasses $c(x,y) = \|x-y\|^p$ with $p \geq 2$.
In view of Section 4 the resulting limit distributions typically do not degenerate for $\mu \neq \nu$ since triviality of Kantorovich potentials is linked to the underlying geometry of the corresponding measures' supports (Theorem 5) and appears to be limited to exotic settings. In contrast, under $\mu = \nu$ constant potentials do exist if $h(0) = 0$ (Corollary 2) and degeneracy of the distributional limits depends on whether the Kantorovich potentials are unique. Indeed, if $\mathrm{supp}(\mu)$ is the closure of a connected open set, then uniqueness holds (Staudt et al. 2021, Corollary 2) and the convergence rate of the empirical OT cost is strictly faster than $n^{-1/2}$. Notably, this in line with results by Ajtai et al. 1984 and Ledoux 2019 stating that the uniform distribution $\mu = \mathrm{Unif}([0,1]^d)$ fulfills $\mathbb{E}\left[\mathrm{OT}_p(\hat{\mu}_n, \nu)\right] = o(n^{-1/2})$ for $d \leq 3$ and $p \geq 2$. However, if $\mathrm{supp}(\mu)$ is disconnected and those components are cost-separated[a], then non-trivial Kantorovich potentials also exist (Staudt et al. 2021, Lemma 11) leading to non-degenerate limit laws.

[a] For costs $c(x,y)\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ with $c(x,x) = 0$ two subsets $\mathcal{X}_1, \mathcal{X}_2 \subseteq \mathcal{X}$ are said to be cost-separated if $\inf_{x \in \mathcal{X}_1, y \in \mathcal{X}_2} \min(c(x,y), c(y,x)) > 0$.

5.4 The d-dimensional Euclidean Space for $d \geq 4$

Beyond the low dimensional setting $d \leq 3$ treated so far, CLTs centered by the population quantity cannot hold in generality for $d \geq 5$ and remain a delicate issue for $d = 4$. For instance, under squared Euclidean costs $c(x,y) = \|x-y\|^2$ and $\mu = \mathrm{Unif}([a, a+1])$, $\nu = \mathrm{Unif}([b, b+1])$ for $a, b \in \mathbb{R}^d$ the OT plan between μ and ν is given by $(\mathrm{id}, \mathrm{id} + b - a)_{\#}\mu$ which implies according to Manole et al. 2021, Theorem 6 that

$$\mathbb{E}\left[\mathrm{OT}_2(\hat{\mu}_n, \nu)\right] - \mathrm{OT}_2(\mu, \nu) = \mathbb{E}\left[\mathrm{OT}_2(\hat{\mu}_n, \mu)\right].$$

For $d \geq 3$ it is known by Ledoux 2019 that $\mathbb{E}\left[\mathrm{OT}_2(\hat{\mu}_n, \mu)\right] \asymp n^{-2/d}$. Recalling the CLT by del Barrio and Loubes 2019 in (5) this implies the random sequence

$$\begin{aligned}&\sqrt{n}(\mathrm{OT}_2(\hat{\mu}_n, \nu) - \mathrm{OT}_2(\mu, \nu))\\ &\qquad = \sqrt{n}(\mathrm{OT}_2(\hat{\mu}_n, \nu) - \mathbb{E}\left[\mathrm{OT}_2(\hat{\mu}_n, \nu)\right]) + \sqrt{n}\mathbb{E}\left[\mathrm{OT}_2(\hat{\mu}_n, \mu)\right]\end{aligned}$$

to be tight for $d = 4$ and to diverge almost surely to ∞ for $d \geq 5$. In conjunction with Goldman and Trevisan 2021, Theorem 6.2 who prove $\lim_{n\to\infty} \sqrt{n}\mathbb{E}\left[\mathrm{OT}_2(\hat{\mu}_n, \mu)\right] = K$ under $d = 4$ for some positive constant $K > 0$, it thus follows from (5) for $n \longrightarrow \infty$ that

$$\sqrt{n}(\mathrm{OT}_2(\hat{\mu}_n, \nu) - \mathrm{OT}_2(\mu, \nu)) \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(K, \mathrm{Var}_{X\sim\mu}[f(X)]), \tag{28}$$

where $f \in S_c(\mu, \nu)$ is a Kantorovich potential between μ and ν. Notably, in this setting the Kantorovich potential is unique (recall Definition 2) and also trivial if and only if $\mu = \nu$ (or equivalently if $a = b$).

In view of Theorem 3 the previous example also highlights that the function class $\mathcal{F}_c$ is not μ-Donsker for $d \geq 4$, since otherwise a centered tight normal limit would result. In particular, the CLT in (28) resembles the first asymptotic distributional limit for the empirical squared 2-Wasserstein distance for $d = 4$ where the centering is given by the population quantity. An exception concerning the Donsker property of $\mathcal{F}_c$ in the high-dimensional regime occurs for different probability measures if one of them is supported on a sufficiently low dimensional space as will be detailed in the next subsection.

5.5 Empirical Optimal Transport under Lower Complexity Adaptation

In this section, we highlight our CLTs for empirical OT in view of the recently discovered *lower complexity adaptation* principle (Hundrieser et al. 2022). It states that statistical rates to estimate the empirical OT cost between two different probability measures μ and ν are driven by the less complex measure, e.g., the one with lower dimensional support. In light of this principle, we emphasize that our CLTs extend beyond the low-dimensional Euclidean case provided that at least one measure has some low-dimensional compact support. The main result relies on the observation that the uniform metric entropies for $\mathcal{F}_c$ and $\mathcal{F}_c^c$ coincide in such settings (Hundrieser et al. 2022, Lemma 2.1). Hence, under suitable bounds on the uniform metric entropy for only one of the function classes $\mathcal{F}_c$ or $\mathcal{F}_c^c$, it follows that *both* are universal Donsker (Definition 1).

The following result makes use of this observation and is specifically tailored to settings where $\mathcal{X}$ has low intrinsic dimension which allows the complexity of $\mathcal{F}_c$ to be suitably controlled such that it is universal Donsker. To formalize this, we consider the setting where $\mathcal{X}$ is a finite set or a compact submanifold of $\mathbb{R}^d$ (see Lee 2013 for comprehensive treatment) with sufficiently small intrinsic dimension. Notably, the first setting covers semi-discrete OT (Aurenhammer et al. 1998; Mérigot

2011; Hartmann and Schuhmacher 2020), i.e., where one of the probability measures is assumed to be finitely supported. Then, under suitable assumptions on the cost function the uniform metric entropy of $\mathcal{F}_c$ is suitably bounded which enables CLTs for the empirical OT cost.

Theorem 8 (Semi-discrete): Let $\mathcal{X}$ be a finite set equipped with discrete topology, let $\mathcal{Y}$ be Polish space and consider a bounded and continuous cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. Then, for arbitrary $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ the weak limits from (10) hold.

Remark 7 (Limit distribution for semi-discrete OT): We comment on structural properties of the limit distribution underlying Theorem 8. When the support of ν is connected (and the cost function is continuous), then Kantorovich potentials are unique (Staudt et al. 2021, Example 3). Hence, in this setting the corresponding weak limit is always centered normal. Even if the support of ν is not connected, under a suitable non-degeneracy condition of the OT plan Kantorovich potentials still remain unique. Hence, under uniqueness of Kantorovich potentials, Theorem 5 serves as a sharp statement for the degeneracy of the limit distribution (Remark 4). In particular, this shows $S_c(\mu, \nu)$ to be trivial only under certain geometrical configurations, whereas $S_c^c(\mu, \nu)$ to be generically non-trivial.
Let us also point out that, parallel and independently to this work, del Barrio et al. 2022 recently also obtained a CLT for the empirical OT cost in the semi-discrete framework for unbounded cost functions. In their setting, our triviality statements on Kantorovich potentials (Theorem 5) and degeneracy results for the corresponding limit laws also apply since for semi-discrete OT the Kantorovich potentials are always continuous (as a finite minimum over continuous functions).

Theorem 9 (Manifolds): Let $\mathcal{X}$ be an s-dimensional smooth compact submanifold of $\mathbb{R}^d$ with $s \leq \min(3, d)$, let $\mathcal{Y} \subseteq \mathbb{R}^d$ be compact and consider a continuous cost function $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. Suppose one of the following two settings.

(i) $s = 1$ and c is locally α-Hölder for $\alpha \in (1/2, 1]$.

(ii) $s = 2$ or $s = 3$ and c is twice continuously differentiable.

Then, for arbitrary $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ the weak limits from (10) hold.

Proofs for Theorems 8 and 9. Note that Assumptions **(C)** and **(S2)** are fulfilled for both settings. By Section 3.1, Lemma A.4 and Lemma A.3 in Hundrieser et al. 2022 for all three cases, respectively, it follows for $\varepsilon > 0$ sufficiently small that

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \lesssim_{\mathcal{X},c} \begin{cases} \log\left(\lceil \varepsilon^{-1} \rceil\right) & \text{for Theorem 8,} \\ \varepsilon^{-s/\alpha} & \text{for setting } (i) \text{ in Theorem 9,} \\ \varepsilon^{-s/2} & \text{for setting } (ii) \text{ in Theorem 9.} \end{cases}$$

Moreover, by Hundrieser et al. 2022, Lemma 2.1 it holds for any $\varepsilon > 0$ that

$$\mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) = \mathcal{N}(\varepsilon, \mathcal{F}_c^c, \|\cdot\|_\infty),$$

which implies that identical bounds on the uniform metric entropy of $\mathcal{F}_c^c$ hold for all settings. Since the square root of the uniform metric entropy is for all settings integrable with respect to $\varepsilon > 0$ near zero, it follows by Van der Vaart and Wellner 1996, Theorem 2.5.2 that both $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are universal Donsker. We thus conclude from Theorem 1 the CLTs for the empirical OT cost (10) for arbitrary probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. □

We emphasize that in contrast to previous CLTs from previous Subsections, no summability conditions are necessary in Theorems 8 and 9 for μ and ν. Additionally, let us point out that Hundrieser et al. 2022 also provide uniform metric entropy bounds for $\mathcal{F}_c$ under more general ground spaces $\mathcal{X}$, such as metric spaces or parametrized surfaces with low intrinsic dimension, as well as α-Hölder costs of smoothness degree $\alpha \in (1, 2]$. As long as the integrability condition by Van der Vaart and Wellner 1996, Theorem 2.5.2 for $\varepsilon > 0$ near zero is fulfilled, these bounds can also be employed for the derivation of CLTs of the empirical OT cost in even more general settings.

Remark 8 (Wasserstein distance in high-dimensional spaces): As a consequence of Theorems 8 and 9, we obtain CLTs for the empirical Wasserstein distance, i.e., for the Euclidean cost function $c(x, y) = \|x - y\|^p$ for $p \geq 1$ even beyond $d \leq 3$ as long as both μ and ν have bounded support and one of them is sufficiently low dimensional. More precisely, if μ is supported on a finite set or on an s-dimensional compact submanifold with $s = 1$ and $p \geq 1$, or in case $s \in \{2, 3\}$ and $p \geq 2$, our asymptotic results from (10) remain valid. Notably, for the latter case the asymptotic results still hold for any $p \in [1, 2)$ if $\mathrm{supp}(\nu)$ is disjoint from $\mathrm{supp}(\mu)$. Indeed, in this setting one can extend the cost function $c|_\Sigma$ for $\Sigma = \mathrm{supp}(\mu) \times \mathrm{supp}(\nu)$ in a smooth manner to $\mathbb{R}^{2d}$, e.g., by the extension theorem of Whitney 1934, without altering the population and empirical OT cost.

In particular, under $s < d$ it follows by Corollary 3 (ii) that $S_c^c(\mu, \nu)$, i.e., the set of Kantorovich potentials corresponding to ν is not trivial if $\mathrm{supp}(\nu)$ has non-empty interior. Hence, for s sufficiently small when replacing the measure ν by its empirical measure $\hat{\nu}_m$ the limit laws do not degenerate. When instead replacing μ by $\hat{\mu}_n$, the limit law could indeed degenerate although this is rather the exception than the rule (see Theorem 5). We recall Figure 1(a), for an example where all Kantorovich potentials $S_c(\mu, \nu)$ are trivial, whereas $S_c^c(\mu, \nu)$ is not.

6 Acknowledgement

S. Hundrieser, M. Klatt, and T. Staudt gratefully acknowledge support from the DFG RTG 2088 and DFG EXC 2067/1- 390729940, and A. Munk of DFG CRC 1456.

Bibliography

Ajtai, M., J. Komlós, and G. Tusnády (1984). “On optimal matchings”. *Combinatorica* 4.4, pp. 259–264.

Aurenhammer, F., F. Hoffmann, and B. Aronov (1998). “Minkowski-type theorems and least-squares clustering”. *Algorithmica* 20.1, pp. 61–76.

Bernton, E., P. Ghosal, and M. Nutz (2021). “Entropic optimal transport: Geometry and large deviations”. *Preprint arXiv:2102.04397.*

Berthet, P. and J.-C. Fort (2019). “Weak convergence of empirical Wasserstein type distances”. *Preprint arXiv:1911.02389.*

Berthet, P., J.-C. Fort, and T. Klein (2020). “A central limit theorem for Wasserstein type distances between two distinct univariate distributions”. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* 56.2, pp. 954–982.

Bobkov, S. and M. Ledoux (2019). *One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances.* American Mathematical Society.

Bobkov, S. G. and M. Ledoux (2021). “A simple Fourier analytic proof of the AKT optimal matching theorem”. *The Annals of Applied Probability* 31.6, pp. 2567–2584.

Boissard, E. and T. Le Gouic (2014). “On the mean speed of convergence of empirical and occupation measures in Wasserstein distance”. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* 50.2, pp. 539–563.

Bonneel, N., M. Van De Panne, S. Paris, and W. Heidrich (2011). “Displacement interpolation using Lagrangian mass transport”. In: *Proceedings of the 2011 SIGGRAPH Asia Conference*, pp. 1–12.

Bronshtein, E. M. (1976). “ε-entropy of convex sets and functions”. *Siberian Mathematical Journal* 17.3, pp. 393–398.

Cárcamo, J., A. Cuevas, and L.-A. Rodríguez (2020). “Directional differentiability for supremum-type functionals: Statistical applications”. *Bernoulli* 26.3, pp. 2143–2175.

Chizat, L., P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré (2020). “Faster Wasserstein distance estimation with the Sinkhorn divergence”. *Advances in Neural Information Processing Systems* 33, pp. 2257–2269.

Csörgö, M. and L. Horváth (1993). *Weighted Approximations in Probability and Statistics.* John Wiley & Sons.

del Barrio, E., E. Giné, and C. Matrán (1999). "Central limit theorems for the Wasserstein distance between the empirical and the true distributions". *The Annals of Probability* 27.2, pp. 1009–1071.

del Barrio, E., E. Giné, and F. Utzet (2005). "Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances". *Bernoulli* 11.1, pp. 131–189.

del Barrio, E., A. González-Sanz, and J.-M. Loubes (2021). "Central limit theorems for general transportation costs". *Preprint arXiv:2102.06379.*

del Barrio, E., A. González-Sanz, and J.-M. Loubes (2022). "Central limit theorems for semidiscrete Wasserstein distances". *Preprint arXiv:2202.06380.*

del Barrio, E., P. Gordaliza, and J.-M. Loubes (2019). "A central limit theorem for L_p transportation cost on the real line with application to fairness assessment in machine learning". *Information and Inference: A Journal of the IMA* 8.4, pp. 817–849.

del Barrio, E. and J.-M. Loubes (2019). "Central limit theorems for empirical transportation cost in general dimension". *The Annals of Probability* 47.2, pp. 926–951.

Dobrić, V. and J. E. Yukich (1995). "Asymptotics for transportation cost in high dimensions". *Journal of Theoretical Probability* 8.1, pp. 97–118.

Dudley, R. M. (1969). "The speed of mean Glivenko-Cantelli convergence". *The Annals of Mathematical Statistics* 40.1, pp. 40–50.

Dudley, R. M. (2014). *Uniform Central Limit Theorems.* Cambridge University Press.

Dümbgen, L. (1993). "On nondifferentiable functions and the bootstrap". *Probability Theory and Related Fields* 95.1, pp. 125–140.

Fang, Z. and A. Santos (2019). "Inference on directionally differentiable functions". *The Review of Economic Studies* 86.1, pp. 377–412.

Fournier, N. and A. Guillin (2015). "On the rate of convergence in Wasserstein distance of the empirical measure". *Probability Theory and Related Fields* 162.3, pp. 707–738.

Galichon, A. (2016). *Optimal Transport Methods in Economics.* Princeton University Press.

Gangbo, W. and R. J. McCann (1996). "The geometry of optimal transportation". *Acta Mathematica* 177.2, pp. 113–161.

Giné, E. and J. Zinn (1986). "Empirical processes indexed by Lipschitz functions". *The Annals of Probability* 14.4, pp. 1329–1338.

Goldman, M. and D. Trevisan (2021). "Convergence of asymptotic costs for random Euclidean matching problems". *Probability and Mathematical Physics* 2.2, pp. 341–362.

González-Delgado, J., A. González-Sanz, J. Cortés, and P. Neuvial (2021). "Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology". *Preprint arXiv:2108.00165.*

Hartmann, V. and D. Schuhmacher (2020). "Semi-discrete optimal transport: A solution procedure for the unsquared Euclidean distance case". *Mathematical Methods of Operations Research* 92, pp. 133–163.

Hundrieser, S., M. Klatt, and A. Munk (2021). "The statistics of circular optimal transport". In: *Directional Statistics for Innovative Applications*. To appear [*Preprint arXiv:2103.15426*]. Springer.

Hundrieser, S., T. Staudt, and A. Munk (2022). "Empirical optimal transport between different measures adapts to lower complexity". *Preprint arXiv 2022.10434*.

Kolouri, S., S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde (2017). "Optimal mass transport: Signal processing and machine-learning applications". *IEEE Signal Processing Magazine* 34.4, pp. 43–59.

Le Cam, L. (1986). "The Central Limit Theorem Around 1935". *Statistical Science* 1.1, pp. 78–91.

Ledoux, M. (2019). "On optimal matching of Gaussian samples". *Journal of Mathematical Sciences* 238, pp. 495–522.

Lee, J. M. (2013). *Introduction to Smooth Manifolds*. Springer.

Manole, T., S. Balakrishnan, J. Niles-Weed, and L. Wasserman (2021). "Plugin estimation of smooth optimal transport maps". *Preprint arXiv:2107.12364*.

Manole, T. and J. Niles-Weed (2021). "Sharp convergence rates for empirical optimal transport with smooth costs". *Preprint arXiv:2106.13181*.

Mason, D. M. (2016). "A weighted approximation approach to the study of the empirical Wasserstein distance". In: *High Dimensional Probability VII*. Springer, pp. 137–154.

McCoy, R. and I. Ntantu (1988). *Topological Properties of Spaces of Continuous Functions*. Springer.

Mérigot, Q. (2011). "A multiscale approach to optimal transport". In: *Computer Graphics Forum*. Vol. 30. 5. Wiley Online Library, pp. 1583–1592.

Munk, A. and C. Czado (1998). "Nonparametric validation of similar distributions and assessment of goodness of fit". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1, pp. 223–241.

Panaretos, V. M. and Y. Zemel (2019). "Statistical aspects of Wasserstein distances". *Annual Review of Statistics and its Application* 6, pp. 405–431.

Peyré, G. and M. Cuturi (2019). "Computational optimal transport: With applications to data science". *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607.

Rachev, S. and L. Rüschendorf (1998a). *Mass Transportation Problems: Volume I: Theory*. Springer.

Rachev, S. and L. Rüschendorf (1998b). *Mass Transportation Problems: Volume II: Applications*. Springer.

Römisch, W. (2004). "Delta method, infinite dimensional". In: *Encyclopedia of Statistical Sciences*. Vol. 16. Wiley, pp. 1575–1583.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer.

Schiebinger, G., J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. (2019). "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming". *Cell* 176.4, pp. 928–943.

Sommerfeld, M. and A. Munk (2018). “Inference for empirical Wasserstein distances on finite spaces”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1, pp. 219–238.

Staudt, T., S. Hundrieser, and A. Munk (2021). “On the uniqueness of Kantorovich potentials”. *Preprint arXiv:2201.08316*.

Talagrand, M. (1994). “Matching theorems and empirical discrepancy computations using majorizing measures”. *Journal of the American Mathematical Society* 7.2, pp. 455–537.

Tameling, C., M. Sommerfeld, and A. Munk (2019). “Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications”. *The Annals of Applied Probability* 29.5, pp. 2744–2781.

Tameling, C., S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk (2021). “Colocalization for super-resolution microscopy via optimal transport”. *Nature Computational Science* 1.3, pp. 199–211.

Van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer.

Weed, J. and F. Bach (2019). “Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance”. *Bernoulli* 25.4A, pp. 2620–2648.

Whitney, H. (1934). “Analytic extensions of differentiable functions defined in closed sets”. *Transactions of the American Mathematical Society* 36.1, pp. 63–89.

A Appendix

Omitted proofs for Section 4

Proof of Equation (19). We prove this equivalence under the assumption of a continuous cost function c and a finite value $\mathrm{OT}_c(\mu, \nu) < \infty$. We set $h = \inf_{x' \in \mathrm{supp}(\mu)} c(x', \cdot)$, which is upper semi-continuous as an infimum of continuous functions.

If $\mu \in P_c(\nu)$ with a corresponding coupling $\pi \in \Pi(\mu, \nu)$ in (18), then it follows that

$$\mathrm{OT}_c(\mu, \nu) \leq \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \,\mathrm{d}\pi(x, y) = \int_{\mathcal{Y}} h(y) \,\mathrm{d}\nu(y) \leq \mathrm{OT}_c(\mu, \nu),$$

where the final step relies on the fact that $h(y) \leq c(x, y)$ for all $(x, y) \in \mathrm{supp}(\mu) \times \mathrm{supp}(\nu)$. In particular, π is an OT plan. Conversely, let the right hand side of (19) be satisfied with an OT plan $\pi \in \Pi(\mu, \nu)$. Then

$$0 \leq \mathrm{OT}_c(\mu, \nu) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \,\mathrm{d}\pi(x, y) = \int_{\mathcal{Y}} h(y) \,\mathrm{d}\pi(x, y) < \infty.$$

Since $h(y) \leq c(x, y)$, this equation implies that equality of the integrands holds π-almost surely, and thus at least in a dense subset $A \subset \mathrm{supp}(\pi)$. For any $(x, y) \in \mathrm{supp}(\pi)$, choose a converging sequence $(x_n, y_n)_n \subset A$. Then, via continuity of c and semi-continuity of h,

$$h(y) \leq c(x, y) = \lim_{n \to \infty} c(x_n, y_n) = \lim_{n \to \infty} h(y_n) \leq h(y),$$

establishing $h(y) = c(x, y)$ for all $(x, y) \in \mathrm{supp}(\pi)$. This implies $\mu \in P_c(\nu)$. □

Proof of Theorem 5. According to Lemma 3 in Staudt et al. 2021, trivial Kantorovich potentials exist if and only if trivial Kantorovich potentials exist in the formally restricted problem where $\mathcal{X}$ is replaced by the support of μ and $\mathcal{Y}$ is replaced by the support of ν. We can therefore assume $\mathcal{X} = \mathrm{supp}(\mu)$ and $\mathcal{Y} = \mathrm{supp}(\nu)$.

Let f be a trivial Kantorovich potential and $\pi \in \Pi(\mu, \nu)$ be an OT plan. By Lemma 1, the Kantorovich potential f is continuous and we can assume that $f \equiv a$ on $\mathrm{supp}(\mu)$ for some $a \in \mathbb{R}$. Then, for each $(x, y) \in \mathrm{supp}(\pi)$, it holds that $f^c(y) = c(x, y) - a$. Therefore,

$$c(x, y) = f^c(y) + a = \inf_{x' \in \mathrm{supp}(\mu)} c(x', y),$$

which establishes relation (18) and shows $\mu \in P_c(\nu)$. Conversely, assume that (18) holds for some $\pi \in \Pi(\mu, \nu)$. Setting $f \equiv 0$, we observe that

$$f(x) + f^c(y) = \inf_{x' \in \mathrm{supp}(\mu)} c(x', y) = c(x, y) \qquad \text{for all } (x, y) \in \mathrm{supp}(\pi).$$

According to standard OT theory (Santambrogio 2015), this equality also holds if f is replaced by the c-concave function $f^{cc} \geq f$, which implies that $f^{cc} \in S_c(\mu, \nu)$

is a Kantorovich potential (up to a suitable additive constant). Since $f^{cc}(x) = c(x, y) - f^c(y) = f(x) = 0$ on the set $\{x \mid (x, y) \in \operatorname{supp}(\pi)\}$, which has full μ-measure, f^{cc} is trivial. □

Proof of Corollary 2. Under condition (i), we find that $\pi = (\mathrm{id}, \mathrm{id})_{\#}\mu \in \Pi(\mu, \nu)$ satisfies (18), implying $\mu \in P_c(\nu)$. If condition (ii) holds, the right hand side of (19) can easily be established since $\pi = (p, \mathrm{id})_{\#}\nu \in \Pi(\mu, \nu)$ is optimal. In both cases, the claim then follows by Theorem 5. □

Proof of Corollary 3. Due to Theorem 5, it is in all cases sufficient to show that $\mu \notin P_c(\nu)$. Under condition (i), we find $\mathrm{OT}_c(\mu, \nu) > 0$. At the same time, $\operatorname{supp}(\nu) \subset \operatorname{supp}(\mu)$ implies that the equality on the right hand side of (19) cannot hold.

We next note that $A_\pi := \{x \mid (x, y) \in \operatorname{supp}(\pi)\}$ is dense in $\operatorname{supp}(\mu)$ for any $\pi \in \Pi(\mu, \nu)$. Under condition (ii), the set $U := \operatorname{int}(\operatorname{supp}(\mu)) \setminus \operatorname{supp}(\nu)$ is non-empty and open, since $\operatorname{supp}(\nu)$ is closed. Therefore, we find $x \in U \cap A_\pi$ and $y \in \operatorname{supp}(\nu)$ with $(x, y) \in \operatorname{supp}(\pi)$. Let $u = x - y$. As U is open, there exists $\varepsilon > 0$ small enough that $x' := x - \varepsilon\, u \in U \subset \operatorname{supp}(\mu)$. Then, employing the strict monotonicity of h,

$$c(x', y) = h(\|x' - y\|) = h\big((1 - \varepsilon)\|x - y\|\big) < h(\|x - y\|) = c(x, y)$$

for $(x, y) \in \operatorname{supp}(\pi)$ and $x' \in \operatorname{supp}(\mu)$, so $\mu \notin P_c(\nu)$ follows by Definition 4.

Under condition (iii), let $U := \operatorname{supp}(\mu) \setminus \Gamma_c(\mu, \nu)$. For any coupling $\pi \in \Pi(\mu, \nu)$, we use the density of A_π in $\operatorname{supp}(\mu)$ to find $x \in U \cap A_\pi$, implying the existence of $y \in \operatorname{supp}(\nu)$ with $(x, y) \in \operatorname{supp}(\pi)$. We conclude $c(x, y) > \inf_{x' \in \operatorname{supp}(\mu)} c(x', y)$, since x would otherwise be an element of $\Gamma_c(\mu, \nu)$. By Definition 4, $\mu \notin P_c(\nu)$ follows. □

Regularity of Kantorovich Potentials

Assumptions imposed on the cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ translate to properties on the function class $\mathcal{F}_c$ and its c-conjugate $\mathcal{F}_c^c$. A simple observation is uniform boundedness of $\mathcal{F}_c$ if the cost is non-negative and bounded. Indeed, for any $f \in \mathcal{F}_c$ it holds that

$$-\|c\|_\infty \leq \inf_{y \in \mathcal{Y}} c(x, y) - \|c\|_\infty \leq f(x) \leq \inf_{y \in \mathcal{Y}} c(x, y) \leq \|c\|_\infty$$

and analogously for any $f \in \mathcal{F}_c^c$. We summarize additional findings in this regard in the following two statements (see also Gangbo and McCann 1996; Villani 2008; Santambrogio 2015; Staudt et al. 2021 for further details).

Lemma 1 (Continuity of Kantorovich potentials): Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption **(C)** combined with **(S1)** or **(S2)**. Then, Kantorovich potentials $S_c(\mu, \nu)$ and $S_c^c(\mu, \nu)$ are continuous on the supports μ and ν, respectively.

Proof. Equicontinuity of the partially evaluated costs implies continuity of Kantorovich potentials since the modulus of continuity of a function class is preserved under pointwise infima or suprema (see Santambrogio 2015, Section 1.2, for details). This implies continuity of Kantorovich potentials on both $\mathcal{X}$ and $\mathcal{Y}$ under Assumption **(S1)**, whereas under Assumption **(S2)** this only implies continuity on $\mathcal{X}$. Moreover, under Assumption **(S2)** the space $\mathcal{X}$ is compact, hence continuity of the costs asserts by Staudt et al. 2021, Lemma 2 continuity of Kantorovich potentials on the support of ν. □

As particular instances where structural properties of the cost function are inherited to Kantorovich potentials, we focus on the setting of Hölder smoothness and semi-concavity.

Lemma 2: Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces.

(i) If for some $\alpha \in (0,1]$ and $L > 0$ the cost $c(\cdot, y)$ is (α, L)-Hölder as a function in x for all $y \in \mathcal{Y}$, then any $f \in \mathcal{F}_c$ is (α, L)-Hölder continuous.

(ii) If for some $\lambda > 0$ the cost $c(\cdot, y)$ is Λ-semi-concave as a function in x for all $y \in \mathcal{Y}$, then any $f \in \mathcal{F}_c$ is λ-semi-concave.

Note that reversing the roles of x and y leads to analogous results for the c-conjugate function class $\mathcal{F}_c^c$ if $\{c(x, \cdot) \mid x \in \mathcal{X}\}$ is uniformly Hölder smooth or semi-concave.

Proof. Suppose $c(\cdot, y)$ is (α, L)-Hölder continuous for any $y \in \mathcal{Y}$. Then, it follows that

$$f(x) = \inf_{y' \in \mathcal{Y}} c(x, y') - g(y') \leq c(x, y) - g(y) \leq c(x', y) - g(y) + L\|x - x'\|^{\alpha}.$$

Taking the infimum on the right hand side with respect to y yields

$$f(x) \leq f(x') + L\|x - x'\|^{\alpha}.$$

Changing the roles of x and x' proves that $|f(x) - f(x')| \leq L\|x - x\|^{\alpha}$ for any $f \in \mathcal{F}_c$. Suppose that $c(\cdot, y)$ is Λ-semi-concave, i.e., $c(\cdot, y) - \Lambda\| \cdot \|^2$ is concave for all $y \in \mathcal{Y}$ as a function of x. It then follows that

$$\begin{aligned}
&f(tx + (1-t)x') - \Lambda\|tx + (1-t)x'\|^2 \\
&\geq t\left(\inf_{y \in \mathcal{Y}} c(x, y) - \Lambda\|x\|^2 - g(y)\right) + (1-t)\left(\inf_{y \in \mathcal{Y}} c(x', y) - \Lambda\|x'\|^2 - g(y)\right) \\
&= t\left(f(x) - \lambda\|x\|^2\right) + (1-t)\left(f(x') - \Lambda\|x'\|^2\right).
\end{aligned}$$

Hence, any $f \in \mathcal{F}_c$ is itself a Λ-semi-concave function. □

C On the Uniqueness of Kantorovich Potentials

Thomas Staudt [*,†]
thomas.staudt@uni-goettingen.de

Shayan Hundrieser [*,†]
s.hundrieser@math.uni-goettingen.de

Axel Munk [*,†,‡]
munk@math.uni-goettingen.de

Abstract

Kantorovich potentials denote the dual solutions of the renowned optimal transportation problem. Uniqueness of these solutions is relevant from both a theoretical and an algorithmic point of view, and has recently emerged as a necessary condition for asymptotic results in the context of statistical and entropic optimal transport. In this work, we challenge the common perception that uniqueness in continuous settings is reliant on the connectedness of the support of at least one of the involved measures, and we provide mild sufficient conditions for uniqueness even when both measures have disconnected support. Since our main finding builds upon the uniqueness of Kantorovich potentials on connected components, we revisit the corresponding arguments and provide generalizations of well-known results. Several auxiliary findings regarding the continuity of Kantorovich potentials, for example in geodesic spaces, are established along the way.

Keywords: Kantorovich potentials, uniqueness, optimal transport, duality theory, regularity

MSC 2020 Subject Classification: primary 49Q22, 46N60, 90C08; secondary 28A35, 49Q20, 62E20

[*] Institute for Mathematical Stochastics, University of Göttingen

[†] Cluster of Excellence: Multiscale Bioimaging (MBExC), University Medical Center, Göttingen

[‡] Max Planck Institute for Biophysical Chemistry, Göttingen

1 Introduction

Optimal transport theory addresses the question how mass can be moved from a source to a target distribution in the most cost-efficient way. While the history of this mathematical quest is long and rich, dating back to Monge 1781 and then Kantorovich 1942, who framed the theory in its modern formulation, it has drawn an enormous amount of attention in the past decades. In-depth monographs cover analytical, probabilistic, and geometric (Rachev and Rüschendorf 1998; Villani 2008; Santambrogio 2015; Ambrosio et al. 2021; Figalli and Glaudo 2021) as well as computational (Peyré and Cuturi 2019) and statistical (Panaretos and Zemel 2020) perspectives, while countless applications span from economics (Galichon 2016) and biology (Schiebinger et al. 2019; Tameling et al. 2021) to machine learning (Arjovsky et al. 2017).

For given probability distributions $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ on measurable spaces X and Y, the optimal transport problem is to find the minimal transportation cost

$$T_c(\mu, \nu) = \inf_{\pi \in \mathcal{C}(\mu,\nu)} \int c \, \mathrm{d}\pi, \tag{1a}$$

where $c\colon X \times Y \longrightarrow \mathbb{R}_+$ denotes a non-negative measurable *cost function* that quantifies the effort of moving one unit of mass between elements in X and Y, and where $\mathcal{C}(\mu, \nu) \subset \mathcal{P}(X \times Y)$ is the set of probability measures with marginal distributions μ and ν. Any solution π to (1a) is called an *optimal transport plan*. Conditions under which optimal transport plans exist are well-known, see for example Villani 2008 for the general framework of Polish spaces and lower-semicontinuous cost functions. Under these assumptions, the optimal transport problem (1a) also admits a dual formulation that reliably serves as a fertile ground for investigating structural properties of T_c. In fact, it holds that

$$T_c(\mu, \nu) = \sup_{f \in L_1(\mu)} \int f \, \mathrm{d}\mu + \int f^c \, \mathrm{d}\nu, \tag{1b}$$

where f^c, the *c-transform* of f, denotes the largest function satisfying $f(x) + f^c(y) \leq c(x, y)$ for all $x \in X$ and $y \in Y$. Specific optimizers f for the dual problem – namely, those which can be written as the c-transform of a function g on Y – are called *Kantorovich potentials*, which exist under mild conditions (Villani 2008, Theorem 5.10).

While our efforts in this work focus on Kantorovich potentials, most of the foundational research on the optimal transport problem (1) has targeted properties of the primal solutions. Significant advances, which provided sufficient conditions for optimal plans to be concentrated on the graph of a uniquely determined function (the *optimal transport map*), have been achieved in Euclidean spaces (Smith and Knott 1987; Cuesta and Matrán 1989; Brenier 1991; Gangbo and McCann 1996), on manifolds (McCann 2001; Figalli 2007; Villani 2008; Figalli and Gigli 2011), and more recently also in more general metric spaces (Bertrand 2008; Gigli et al. 2012; Ambrosio and Rajala 2014). Strong regularity properties of optimal transport maps under squared Euclidean costs have first been established by the seminal work of Caffarelli (1990,1991,1992) for probability measures with bounded convex support (with recent extensions to unbounded settings by Cordero-Erausquin and Figalli 2019). Further insights were obtained by Ma et al. 2005 and Loeper 2009 for C^4-costs that satisfy a certain differential inequality (the *Ma-Trudinger-Wang condition*). Later, De Philippis and Figalli 2015 demonstrated regularity of optimal maps outside of "bad sets" of measure zero under more general conditions. A related line of research is devoted to the analysis of optimal transport plans that are not necessarily induced by a transport map. Uniqueness results for the primal solution in this context were, for example, obtained by Ahmad et al. 2011 or McCann and Rifford 2016, and more recently by Moameni and Rifford 2020.

Many of the techniques employed to characterize the primal solutions of (1a) crucially depend on the duality theory (1b). In fact, the gradients of dual solutions are intimately related to optimal transport maps (see Villani 2008, Chapter 10, for an in-depth treatment). Still, dual solutions are commonly not studied as objects of interest in their own right, and certain properties, such as the uniqueness of Kantorovich potentials, have received considerably less attention when compared to their primal counterparts. Recent developments, however, have emphasized the utility of dual uniqueness. For example, in the context of statistical optimal transport, uniqueness of Kantorovich potentials ensures a Gaussian limit distribution for the empirical optimal transport cost (Sommerfeld and Munk 2018; del Barrio and Loubes 2019; Tameling et al. 2019; del Barrio et al. 2021a; del Barrio et al. 2021b). Furthermore, recent results on the convergence of entropically regularized optimal transport to its vanilla counterpart as the regularization tends to zero utilize dual uniqueness as a critical assumption as well (Altschuler et al. 2021; Bercu and Bigot 2021; Bernton et al. 2021; Nutz and Wiesel 2021). On a more general note, uniqueness is also required for the meaningful and efficient computation of optimal transport gradient flows in $\mathcal{P}(X)$, since Kantorovich potentials coincide with the subgradients of the functional $\mu \mapsto T_c(\mu, \nu)$ (see Santambrogio 2015, Section 7.2).

For continuous measures in Euclidean spaces, several sufficient criteria for unique Kantorovich potentials are available. The involved arguments are well-known from the above mentioned literature on optimal transport maps, and depend on (i) sufficient local regularity of dual solutions f (e.g., local Lipschitz continuity) and (ii) exploiting that the gradient of f is (where it exists) determined by the cost function and an (arbitrary) optimal transport plan. Notable results that adopt this strategy include Proposition 7.18 in Santambrogio 2015, which is applicable in compact settings, or Appendix B of Bernton et al. 2021 and Corollary 2.7 of del Barrio et al. 2021b, both relying on regularity properties of dual solutions derived by Gangbo and McCann 1996 for a certain family of strictly convex costs. Meanwhile, Remark 10.30 in Villani 2008 sketches a general argument for uniqueness of Kantorovich potentials on Riemannian manifolds. A commonality of these (and, to our knowledge, all other related) results in this vein is the requirement that the support of at least one probability measure μ or ν is connected. This requirement is usually taken as self-evident, since the standard proof technique – concluding uniqueness of a function (up to constants) from uniqueness of its gradients – naturally cannot bridge separated connected components. In fact, it is easy to construct trivial counter examples, like $X = Y = [0, 1] \cup [2, 3]$ with uniformly distributed $\mu = \nu$ on X, where uniqueness of the Kantorovich potentials is bound to fail for a wide range of cost functions (see Lemma 11 in Appendix A).

At the same time, on finite spaces, dual uniqueness results for transportation problems have long been established via methods from finite linear programming (Klee and Witzgall 1968; Hung et al. 1986). Even though these settings naturally involve disconnected supports, they still feature unique Kantorovich potentials whenever the measures μ and ν are *non-degenerate*, meaning that all proper subsets $X' \subset X$ and $Y' \subset Y$ are assigned different masses $\mu(X') \neq \nu(Y')$. The main conceptual contribution of our work is the formulation of an analogon of this observation accessible to the continuous world. This is realized by lifting the uniqueness of dual solutions on connected components to their uniqueness on the full space, a strategy that works in general Polish spaces. The following statement is an informal version of our central result (Theorem 1).

> Theorem (Informal): Assume that μ and ν are non-degenerate (in a suitable sense) and that each optimal potential f^c is continuous. If the optimal transport problems restricted to the connected components of the support of μ have unique Kantorovich potentials, then the full optimal transport problem has unique Kantorovich potentials as well.

Clearly, this statement is mainly useful if combined with dual uniqueness results for measures with connected support. This motivates us to revisit the common proof strategy for uniqueness in the connected setting and present a formulation (Theorem 2) that is more general than the ones we have cited above. Theorem 2 covers settings where X is a smooth manifold and Y is allowed to be a generic Polish space, and we carefully discuss which properties of μ, ν, and c are actually necessary for the argumentation. Particular scenarios where the assumptions of Theorem 2 can easily be checked include settings where the space Y is compact (Corollary 2) or where the cost function satisfies certain growth and regularity conditions outside of compact sets (Corollary 3 and Section 4). In fact, we will learn that the requirements on μ are primarily of topological nature (i.e., concerning the shape of its support), while the actual distribution of mass can often be quite arbitrary. For example, in the setting $X = Y = [0, 1] \cup [2, 3]$ mentioned earlier, combining Theorem 1 and Corollary 2 establishes that Kantorovich potentials are unique if μ and ν are supported on all of X and satisfy $\mu([0, 1]) \neq \nu([0, 1])$. This holds for general differentiable costs without further assumptions on μ or ν such as the existence of a Lebesgue density. In this sense, failure of uniqueness due to disconnected supports is typically an exception caused by a specific symmetry, and not the rule.

Outline. The notion of c-concavity is introduced and discussed in Section 2. Kantorovich potentials are then defined as c-concave solutions of the optimal transport problem in its dual formulation (1b). We proceed to discuss the regularity of such potentials, with a focus on the connection between continuity and transport towards infinity. Some technical results that cope with restrictions of the base spaces X and Y are emphasized as well. Section 3 opens with a clarification of equivalent ways to define almost surely unique Kantorovich potentials. Afterwards, our main results on uniqueness of Kantorovich potentials for probability measures with disconnected support (Theorem 1 and Corollary 1) are presented and discussed, including its consequences for the semi-discrete setting and countably discrete spaces. We then turn to uniqueness statements for measures with connected support on smooth manifolds (Theorem 2, Corollary 2, and Corollary 3). Section 4 contributes some findings on continuity properties of Kantorovich potentials for fast-growing cost functions (particularly for geodesic metric spaces, Theorem 3), revealing that discontinuities are often confined to the boundary of the support. Finally, Section 5 contains the proof of Theorem 1 and Appendix A documents auxiliary observations and arguments that have been omitted in the main text.

Notation. Throughout the manuscript, X and Y denote Polish spaces, i.e., completely metrizable and separable topological spaces. For $A \subset X$, we write $\mathrm{int}(A)$ for its interior, $\mathrm{cl}(A)$ for its closure, and $\partial A = \mathrm{cl}(A) \setminus \mathrm{int}(A)$ for its boundary. The Cartesian projection from a product of spaces to a component X is denoted by p_X. Real-valued functions f and g on spaces X and Y can be lifted to $X \times Y$ via the operation $(x, y) \mapsto f(x) + g(y)$, which we denote by $f \oplus g$. The set of Borel probability measures, or distributions, on a Polish space X are called $\mathcal{P}(X)$. The support of a probability distribution $\mu \in \mathcal{P}(X)$, which is the smallest closed set $A \subset X$ such that $\mu(A) = 1$, is denoted by $\mathrm{supp}\,\mu$. We write $\mu \otimes \nu \in \mathcal{P}(X \times Y)$ to denote the product of probability measures on Polish spaces X and Y. Integration $\int f \,\mathrm{d}\mu$ of a real-valued function f on X is abbreviated by juxtaposition μf.

If M is a smooth manifold (without boundary), we call $f\colon M \to \mathbb{R}$ locally Lipschitz if $f \circ \varphi^{-1}$ is locally Lipschitz for every chart φ of an atlas of M. Similarly, we call f locally semiconcave if $f \circ \varphi^{-1}$ is locally semiconcave for every chart φ of an atlas of M. By this we mean that each point in $\mathrm{range}\,\varphi$ admits $\lambda > 0$ and a convex neighbourhood $V \subset \mathrm{range}\,\varphi$ such that

$$v \mapsto f\big(\varphi^{-1}(v)\big) - \lambda\|v\|^2 \tag{2}$$

is concave on V. A family of functions $f_y\colon M \to \mathbb{R}$ for $y \in Y$ is called *locally Lipschitz (or locally semiconcave) uniformly in y* if f_y is locally Lipschitz (or locally semiconcave) with neighborhoods and constants that do not depend on y. Similarly, the functions f_y are *locally Lipschitz locally uniformly in y* if the functions f_y are locally Lipschitz uniformly in $y \in K$ for each compact set $K \subset Y$. We furthermore say that a Borel set A has *full Lebesgue measure in charts of M* if $\mathrm{range}\,\varphi \setminus \varphi\big(A \cap \mathrm{domain}\,\varphi\big)$ is a Lebesgue null set for each chart φ of an atlas of M.

2 Kantorovich Potentials

Let $c\colon X \times Y \longrightarrow \mathbb{R}_+$ be a non-negative cost function that compares elements of Polish spaces X and Y. As laid out comprehensively in Villani 2008 or Santambrogio 2015, a central part of the duality theory of optimal transport is the notion of c-conjugacy. For any $g\colon Y \longrightarrow \mathbb{R} \cup \{-\infty\}$, its associated *$c$-transform* is defined via

$$g^c\colon X \longrightarrow \mathbb{R} \cup \{-\infty\}, \qquad g^c(x) = \inf_{y \in Y} c(x,y) - g(y). \tag{3a}$$

Any function $f\colon X \longrightarrow \mathbb{R} \cup \{-\infty\}$ that coincides with g^c for some $g\colon Y \longrightarrow \mathbb{R} \cup \{-\infty\}$ and that is not equal $-\infty$ everywhere is called *c-concave on X*. The set of all functions that are c-concave on X is denoted by S_c. Since the roles of f and g can easily be exchanged in these definitions, we also write

$$f^c\colon Y \longrightarrow \mathbb{R} \cup \{-\infty\}, \qquad f^c(y) = \inf_{x \in X} c(x,y) - f(x) \tag{3b}$$

for the c-transform of a function $f\colon X \longrightarrow \mathbb{R} \cup \{-\infty\}$. Any $g\colon Y \longrightarrow \mathbb{R} \cup \{-\infty\}$ that originates from a c-transform and that is not equal $-\infty$ everywhere is called *c-concave on Y*. Since $f = f^{cc}$ and $g = g^{cc}$ for any c-concave f or g (see Santambrogio 2015, Proposition 1.34), the set S_c^c of pointwise c-transformed elements of S_c equals the set of functions that are c-concave on Y. Under continuity of the cost function c, all functions in S_c and S_c^c are upper-semicontinuous and thus Borel measurable. Note that our notation accentuates the asymmetry in the operations (3a) and (3b) less explicitly than Santambrogio 2015, who denotes (3a) as $\bar{c}$-transform, or Villani 2008, who picks a different sign convention and contrasts c-concavity to c-convexity.

For any given $f\colon X \longrightarrow \mathbb{R} \cup \{-\infty\}$, the c-transform $g = f^c$ designates the largest function that satisfies $f \oplus g \leq c$. The set of points in $X \times Y$ where equality holds is denoted as *c-subdifferential of f*, and we write

$$\partial_c f = \big\{(x,y) \in X \times Y \,\big|\, f(x) + f^c(y) = c(x,y)\big\}. \tag{4}$$

This set is closed when c is continuous and f is upper-semicontinuous (so in particular when f is c-concave). If f is a solution of the dual optimal transport problem (1b), it is clear that any optimal transport plan has to be concentrated on $\partial_c f$. The following statement, which is a special case of Theorem 5.10 (ii) in Villani 2008, establishes that (generalized) c-concave dual solutions exist under mild conditions.

> Theorem (Existence of optimal solutions): Let X and Y be Polish and $c\colon X \times Y \longrightarrow \mathbb{R}_+$ continuous. For any $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ with $T_c(\mu,\nu) < \infty$, there exists an optimal transport plan $\pi \in \mathcal{C}(\mu,\nu)$ and a c-concave function $f \in S_c$ such that
>
> $$T_c(\mu,\nu) = \pi c = \pi\big(f \oplus f^c\big). \tag{5}$$

We emphasize that the function f in this statement does not have to be μ-integrable, nor does f^c have to be ν-integrable. Ensuring integrability requires further conditions

(Villani 2008, Remark 5.14), for instance $(\mu \otimes \nu)\, c < \infty$. Only then can f be viewed as a dual optimizer of (1b) in the strict sense

$$T_c(\mu, \nu) = \pi c = \mu f + \nu f^c.$$

For our ends, however, the more general solutions provided by (5) are sufficient. We call these solutions *(generalized) Kantorovich potentials*, and we write $f \in S_c(\mu, \nu) \subset S_c$ or $f^c \in S_c^c(\mu, \nu) \subset S_c^c$ to emphasize their dependence on μ and ν. We stress that any $f \in S_c(\mu, \nu)$ satisfies (5) for *all* optimal transport plans π, so f does not favor a particular primal solution (Beiglböck and Schachermayer 2011, Lemma 1.1).

Note that the existence of solutions as well as duality statements for optimal transportation problems have also been established for non-continuous cost functions (Villani 2008; Beiglböck and Schachermayer 2011) or more general spaces (Rüschendorf 2007). Two major advantages of working with continuous costs are the closedness of the c-subdifferential $\partial_c f$ for any $f \in S_c(\mu, \nu)$ and the (related) upper-semicontinuity of c-conjugate functions. The former implies

$$\operatorname{supp} \pi \subset \partial_c f$$

for any optimal transport plan π, a property which we will often resort to, while the latter is needed for continuity results and permits sidestepping measurability issues.

Regularity

Due to their nature as c-concave functions, Kantorovich potentials inherit certain regularity properties from the cost function c. For example, if c is concave in its first argument, then $f \in S_c$ is also concave as an infimum over concave functions. Similarly, if the family $\{c(\cdot, y) \mid y \in Y\}$ of partially evaluated costs is (locally) equicontinuous, then f shares the respective (local) modulus of continuity (Santambrogio 2015, Section 1.2). Imposing conditions of this form is hence a convenient way to guarantee continuity of c-concave functions, which are in general only upper-semicontinuous (for continuous costs). In the following, we introduce tools that give us a more fine-grained control over the continuity of Kantorovich potentials. We begin with continuity along sequences in $\partial_c f$.

Lemma 1: Let X and Y be Polish, $c\colon X \times Y \to \mathbb{R}_+$ continuous, and $f \in S_c$. If $(x_n, y_n)_{n\in\mathbb{N}}$ is a sequence in $\partial_c f$ that converges to $(x, y) \in \partial_c f$, then $f(x_n) \to f(x)$ and $f^c(y_n) \to f^c(y)$ as $n \to \infty$.

Proof. Both f and f^c are upper-semicontinuous. Since (x_n, y_n) and (x, y) are elements in $\partial_c f$,

$$f^c(y) \geq \limsup_{n\to\infty} f^c(y_n) \geq c(x, y) - \limsup_{n\to\infty} f(x_n) \geq c(x, y) - f(x) = f^c(y).$$

Therefore, $\limsup_{n\to\infty} f(x_n) = f(x)$ and $\limsup_{n\to\infty} f^c(y_n) = f^c(y)$ has to hold. □

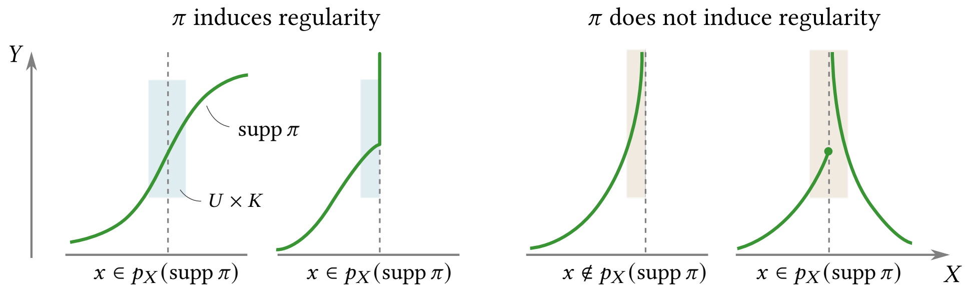


Figure 1: Transport plans and induced regularity. The two plans on the left induce regularity at $x \in \operatorname{supp}(\mu)$ (dashed line), since a relatively open neighborhood $U \subset \operatorname{supp}(\mu)$ and a compact set $K \subset Y$ can be found such that condition (6) is satisfied. The two plans on the right fail to induce regularity. Note that $x \in p_X(\operatorname{supp}\pi)$ is possible even if regularity is not induced at the point x (rightmost sketch).

We next show that Kantorovich potentials can only be discontinuous at points that are "sent to infinity" by all optimal transport plans. To put this more precisely, given a relatively open set $U \subset \operatorname{supp}\mu$, we say that a transport plan $\pi \in \mathcal{C}(\mu, \nu)$ *induces regularity on* U if there exists a compactum $K \subset Y$ such that

$$p_X(\operatorname{supp}\pi) \cap U = p_X\big(\operatorname{supp}\pi \cap (U \times K)\big), \tag{6}$$

where p_X denotes the coordinate projection onto X. We also say that π *induces regularity at* $x \in \operatorname{supp}\mu$, if there is a (relatively) open neighborhood $U \subset \operatorname{supp}\mu$ of x such that π induces regularity on U (see Figure 1). Similar definitions with reversed roles are deployed for subsets or points in $\operatorname{supp}\nu$, to which all of the following statements can be adjusted as well.

Lemma 2: Let X and Y be Polish, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \longrightarrow \mathbb{R}_+$ continuous with $T_c(\mu, \nu) < \infty$. Let $\pi \in \mathcal{C}(\mu, \nu)$ be optimal and $f \in S_c(\mu, \nu)$. If π induces regularity on $U \subset \operatorname{supp}\mu$ with respect to the compact set $K \subset Y$, then $U \subset p_X(\operatorname{supp}\pi)$ and

$$f(x) = \inf_{y \in K} c(x, y) - f^c(y) \tag{7}$$

for all $x \in U$. In particular, $f|_{\operatorname{supp}\mu}$ is continuous on U.

Proof. Let π induce regularity on the relatively open set $U \subset \operatorname{supp}\mu$ with respect to the compact set $K \subset Y$ as defined in (6). Restricted to the domain $X \times K$, the projection p_X is a closed map. Condition (6) thus establishes that $A = p_X(\operatorname{supp}\pi) \cap U = p_X\big(\operatorname{supp}\pi \cap (X \times K)\big) \cap U$ is relatively closed in U. Since $p_X(\operatorname{supp}\pi)$ is dense in $\operatorname{supp}\mu$, any point in U is a limit point of A. Therefore, $U = A \subset p_X(\operatorname{supp}\pi)$. Furthermore, each $x \in U$ admits a partner $y \in K$ such that $(x, y) \in \operatorname{supp}\pi \subset \partial_c f$. This establishes (7), since

$$f(x) = c(x, y) - f^c(y) = \inf_{y' \in K} c(x, y') - f^c(y').$$

To show continuity of $f|_{\operatorname{supp}\mu}$ on U, it is enough to show that each sequence $(x_n)_{n\in\mathbb{N}} \subset \operatorname{supp}\mu$ converging to $x \in U$ has a subsequence attaining the limit $f(x)$. Since U is (relatively) open, we can assume $(x_n)_n \subset U$. Thus, each x_n admits $y_n \in K$ such that $(x_n, y_n) \in \operatorname{supp}\pi$. After taking a suitable subsequence, we may assume that $y_n \to y \in K$ due to the compactness of K. Thus, $(x_n, y_n) \to (x, y)$ as $n \to \infty$. Since $\operatorname{supp}\pi \subset \partial_c f$ is closed in $X \times Y$, it contains (x, y), and so Lemma 1 can be applied to establish $\lim_{n\to\infty} f(x_n) = f(x)$. □

Remark 1 (Projections and measurability): We commonly formulate our results in terms of the projected sets $p_X(\operatorname{supp}\pi) \subset \operatorname{supp}\mu$ and $p_Y(\operatorname{supp}\pi) \subset \operatorname{supp}\nu$, where $\pi \in \mathcal{C}(\mu, \nu)$ denotes an (arbitrary) optimal transport plan. Note that the inclusions can be strict, for which case Lemma 2 establishes the relation

$$\operatorname{supp}\mu \setminus p_X(\operatorname{supp}\pi) \subset \big\{x \in \operatorname{supp}\mu \,\big|\, \pi \text{ does not induce regularity at } x\big\}$$

and vice versa for ν. Projected sets of the form $p_X(\operatorname{supp}\pi)$ or $p_Y(\operatorname{supp}\pi)$ are not Borel measurable in general, but they are analytic and thus contain Borel subsets of full μ- or ν-measure (see Lemma 9 in Appendix A for references). We usually prefer the explicit formulation via these sets (instead of writing "almost surely") to emphasize that the domain of a property does not depend on the choice of a specific Kantorovich potential.

Some consequences of Lemma 2 deserve to be highlighted. First, the Kantorovich potentials $S_c(\mu, \nu)$ are always continuous if the space Y is compact. If Y is not compact, the intuition fostered by Lemma 2 is that discontinuities can only occur at points from whose immediate vicinity some mass is sent towards infinity, in the sense that this mass leaves any compactum in Y. Relation (7) is particularly useful for transferring properties of the cost function to Kantorovich potentials, like a modulus of continuity of $c(\cdot, y)$ that holds only *locally* in y. This observation becomes crucial for Theorem 2 in Section 3.

Examples where induced regularity fails at some points can easily be found, and include settings where μ is compactly supported but ν is not. At the offending points, f can turn out to be both continuous or discontinuous, depending on the regularity of the cost function as well as the specific behavior of μ and ν (this can already be observed for $X = Y = \mathbb{R}$). The following example anticipates that points of discontinuity are often restricted to the boundary of the support, a phenomenon that we more closely study in Section 4.

Example 1 (Continuity in geodesic spaces): Let $X = Y$ for a locally compact complete geodesic space (X, d) and consider a cost function of the form $c(x, y) = h(d(x, y))$ with convex and differentiable $h\colon \mathbb{R}_+ \to \mathbb{R}_+$. Then every Kantorovich potential $f \in S_c(\mu, \nu)$ is continuous on the interior of the support of μ and discontinuities at the boundary are only possible if $h'(a) \to \infty$ as $a \to \infty$. Proofs and more details are provided in Section 4.

We stress that regularity properties of Kantorovich potentials that go beyond mere continuity, like their degree of differentiability, are extensively studied in the literature on the optimal transport map (see Villani 2008, Chapter 10, for a detailed exposition). To mention a common argument in this context, one possible way to enforce twofold differentiability of $f \in S_c$ is to work with semiconcave cost functions.

> Example 2 (Continuity under semiconcavity): Let $X = \mathbb{R}^d$ for some $d \in \mathbb{N}$ with the Euclidean norm $\|\cdot\|$ and assume that the function class $\{c(\cdot, y) \,|\, y \in Y\}$ has uniformly bounded second derivatives. Then there exists $\lambda > 0$ such that $x \mapsto c(x, y) - \lambda\|x\|^2$ is concave for each $y \in Y$, which implies concavity of $x \mapsto f(x) - \lambda\|x\|^2$ for $f \in S_c(\mu, \nu)$ as well. In particular, f is continuous and Lebesgue-almost everywhere twice differentiable in the domain $\Omega = \operatorname{int}(\{x \,|\, f(x) > -\infty\}) \subset \mathbb{R}^d$, which is convex and contains the interior of $\operatorname{supp}\mu$. On the boundary of Ω, the potential may assume finite values or $-\infty$.

Restrictions

A valuable trait of optimal transport theory is that both primal and dual solutions behave consistently when the base spaces X and Y are restricted to subspaces. General results in this direction can be found in Villani 2008, Theorem 4.6 and Theorem 5.19. In the following, we stress some selected statements that complement our assertions on continuity and uniqueness of Kantorovich potentials. We begin with a technical observation that restrictions of Kantorovich potentials of the form $f|_{\operatorname{supp}\mu}$, as they appear in Lemma 2, can (almost) be understood as Kantorovich potentials of a suitably restricted problem. The proof is delegated to Appendix A.

> Lemma 3 (Restriction to sets of full mass): Let X and Y be Polish and $c\colon X \times Y \to \mathbb{R}_+$ continuous. Suppose $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ such that $T_c(\mu, \nu) < \infty$. Let $\pi \in \mathcal{C}(\mu, \nu)$ be an optimal plan and let $\tilde{c}$ denote the restriction of c to the set $\tilde{X} \times \tilde{Y}$, where $\tilde{X} \subset X$ and $\tilde{Y} \subset Y$ are Borel and Polish subspaces with $\mu(\tilde{X}) = \nu(\tilde{Y}) = 1$. Let $\tilde{\Gamma} = \operatorname{supp}\pi \cap (\tilde{X} \times \tilde{Y})$.
>
> (Restrict) Every $f \in S_c(\mu, \nu)$ admits $\tilde{f} \in S_{\tilde{c}}(\mu, \nu)$ that agrees with f on $p_X(\tilde{\Gamma})$.
>
> (Extend) Every $\tilde{f} \in S_{\tilde{c}}(\mu, \nu)$ admits $f \in S_c(\mu, \nu)$ that agrees with $\tilde{f}$ on $p_X(\tilde{\Gamma})$.
>
> In both cases, the conjugates f^c and $\tilde{f}^{\tilde{c}}$ agree on $p_Y(\tilde{\Gamma})$.

> Remark 2 (Ambiguity of extensions): Depending on the setting, there can be distinct ways of extending Kantorovich potentials from the support to the whole space. For example, if $X = Y$ are equal and c is a metric, then the c-concave functions are exactly the 1-Lipschitz functions with respect to c, and it is easy to see that $f^c = -f$ holds for any $f \in S_c$. In this situation, ambiguous extensions are common if $\operatorname{supp}\mu \cup \operatorname{supp}\nu$ does not cover the whole space X.

We next address the behavior of Kantorovich potentials when transportation is restricted to a part of X that does not necessarily occupy full μ-mass. Let $\pi \in \mathcal{C}(\mu, \nu)$

be an optimal plan and suppose that $\mu(\tilde{X}) > 0$ for some closed subset $\tilde{X} \subset X$. We denote the optimal transport problem between the probability measures $\mu_{\tilde{X}} = \mu|_{\tilde{X}}/\mu(\tilde{X})$ and $\nu_{\tilde{X}} = \pi(\tilde{X} \times \cdot)/\mu(\tilde{X})$ under the cost function $c_{\tilde{X}} = c|_{\tilde{X}\times Y}$ as the $\tilde{X}$-*restricted problem (with respect to π)*. As an application of Villani 2008, Theorem 5.19, we note that the restriction of Kantorovich potentials in the original problem yields Kantorovich potentials in the restricted problem.

Lemma 4 (Restriction to sets of partial mass): Let X and Y be Polish and $c\colon X \times Y \to \mathbb{R}_+$ continuous. Suppose $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ such that $T_c(\mu, \nu) < \infty$ and let $\tilde{X} \subset X$ be closed with $\mu(\tilde{X}) > 0$. Then any $f \in S_c(\mu, \nu)$ admits $\tilde{f} \in S_{c_{\tilde{X}}}(\mu_{\tilde{X}}, \nu_{\tilde{X}})$ that agrees with f on $p_X(\operatorname{supp}\pi) \cap \tilde{X}$.

As a consequence of Lemma 4, it is always possible to decompose Kantorovich potentials defined on disconnected spaces in a natural way. Indeed, if $\operatorname{supp}\mu = \bigcup_{i\in I} X_i$ is a countable partitioning into connected components (which are always closed) with $\mu(X_i) > 0$ for each $i \in I$, then any $f \in S_c(\mu, \nu)$ admits restricted potentials $f_i \in S_{c_{X_i}}(\mu_{X_i}, \nu_{X_i})$ such that

$$f = \sum_{i\in I} \mathbb{1}_{X_i} \cdot f_i \qquad \text{on } p_X(\operatorname{supp}\pi) \tag{8}$$

for any optimal $\pi \in \mathcal{C}(\mu, \nu)$. This simple but crucial observation lies at the heart of the uniqueness result for probability measures with disconnected support discussed in the next section.

3 Uniqueness

In a strict sense, Kantorovich potentials are never unique. Indeed, it is easy to see that $f \in S_c(\mu, \nu)$ implies $f + a \in S_c(\mu, \nu)$ for any $a \in \mathbb{R}$. Therefore, statements about uniqueness are generally only reasonable up to constant shifts. Besides this ambiguity, it is often too restrictive to require uniqueness to hold outside of the supports of the involved measures (see Remark 2 on ambiguous extensions). We will therefore focus on the notion of *almost surely unique Kantorovich potentials (up to constant shifts)*, by which we mean that $f_1 - f_2$ is μ-almost surely constant for all $f_1, f_2 \in S_c(\mu, \nu)$. Due to the regularizing nature of the c-transform, almost sure uniqueness of $S_c(\mu, \nu)$ is actually equivalent to almost sure uniqueness of $S_c^c(\mu, \nu)$.

Lemma 5: Let X and Y be Polish, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \to \mathbb{R}_+$ continuous such that $T_c(\mu, \nu) < \infty$. For any optimal transport plan π and any $f_1, f_2 \in S_c(\mu, \nu)$

1) $f_1 = f_2$ μ-almost surely iff $f_1 = f_2$ on $p_X(\operatorname{supp}\pi)$,

2) $f_1^c = f_2^c$ ν-almost surely iff $f_1^c = f_2^c$ on $p_Y(\operatorname{supp}\pi)$,

3) $f_1 = f_2$ on $p_X(\operatorname{supp}\pi)$ iff $f_1^c = f_2^c$ on $p_Y(\operatorname{supp}\pi)$.

Proof. We begin with the first assertion and assume that $f_1 = f_2$ holds on a Borel set $A \subset X$ with $\mu(A) = 1$. The set $B = \operatorname{supp}\pi \cap (A \times Y)$ is dense in $\operatorname{supp}\pi$, so that there is a convergent sequence in B to any $(x, y) \in \operatorname{supp}\pi$. Lemma 1 then asserts $f_1(x) = f_2(x)$ for all $x \in p_X(\operatorname{supp}\pi)$. Conversely, Lemma 9 in Appendix A shows that $p_X(\operatorname{supp}\pi)$ contains a Borel set of full μ-measure. Assertion 2 follows similarly. To show assertion 3, it is sufficient to observe $c(x, y) = f_1(x) + f_1^c(y) = f_2(x) + f_2^c(y)$ and thus $f_1(x) - f_2(x) = f_2^c(y) - f_1^c(y)$ for any $(x, y) \in \operatorname{supp}\pi$. □

Disconnected support

Our contributions regarding the uniqueness of Kantorovich potentials for measures with disconnected support are inspired by well-known results from the theory of finite linear programming. In a nutshell, we will show that unique Kantorovich potentials on the connected components of the support are sufficient to imply the uniqueness on the whole support, as long as we have continuous c-transformed potentials and so-called non-degenerate optimal plans. If

$$\operatorname{supp}\mu = \bigcup_{i\in I} X_i \qquad \text{and} \qquad \operatorname{supp}\nu = \bigcup_{j\in J} Y_j \tag{9}$$

are (at most countable) decompositions of the supports of μ and ν into connected components, then $\pi \in \mathcal{C}(\mu, \nu)$ is called *degenerate* if there exist subsets $I' \subset I$ and $J' \subset J$ such that

$$0 < \sum_{i\in I'} \mu(X_i) = \sum_{i\in I'} \sum_{j\in J'} \pi(X_i \times Y_j) = \sum_{j\in J'} \nu(Y_j) < 1. \tag{10}$$

This definition allows for the following sufficient criterion for non-degeneracy, which has the advantage that it can easily be checked on the basis of μ and ν alone.

Lemma 6: If all nonempty proper $I' \subset I$ and $J' \subset J$ satisfy $\sum_{i\in I'} \mu(X_i) \neq \sum_{j\in J'} \nu(Y_j)$, then no transport plan $\pi \in \mathcal{C}(\mu, \nu)$ is degenerate.

Under suitable conditions, non-degenerate optimal transport plans make it possible to uniquely link together Kantorovich potentials of the X_i-restricted transport problems (recall this notion from Section 2) to assert uniqueness of Kantorovich potentials on the full support. For the following result, note that $\mu(X_i) > 0$ for all $i \in I$ if I is finite, since each $X_i \subset \operatorname{supp}\mu$ is open in this case.

Theorem 1 (Uniqueness under disconnected support): Let X and Y be Polish, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \to \mathbb{R}_+$ continuous with $T_c(\mu, \nu) < \infty$. Assume decomposition (9) for I finite and J (at most) countable and assume that for all $f^c \in S_c^c(\mu, \nu)$ either

1) $f^c|_{\operatorname{supp}\nu}$ is continuous, or

2) $f^c|_{Y_j}$ is continuous and $\operatorname{supp}\nu|_{Y_j}$ is connected for all $j \in J$ with $\nu(Y_j) > 0$.

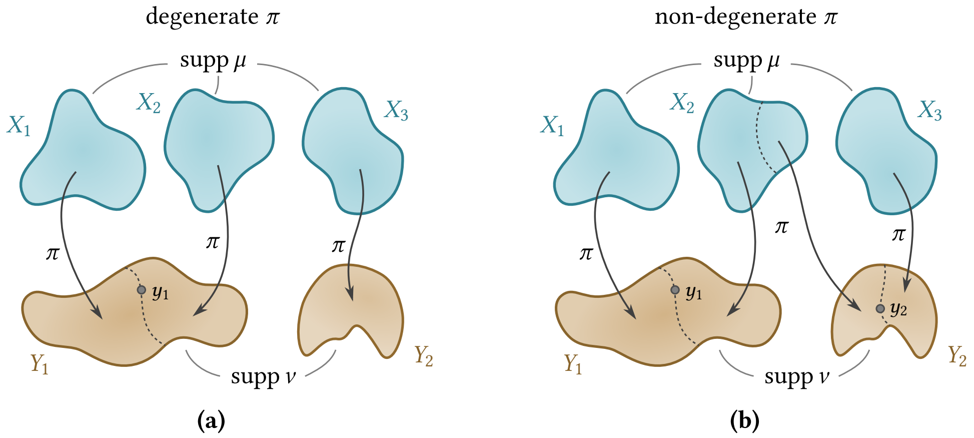


Figure 2: Uniqueness of Kantorovich potentials in disconnected spaces. In sketch **(a)**, transport between $X_1 \cup X_2$ and Y_1 is decoupled from transport between X_3 and Y_2. As the proof of Theorem 1 shows, the existence of a contact point y_1 links the (restricted) Kantorovich potentials $f_1 : X_1 \to \mathbb{R}$ and $f_2 : X_2 \to \mathbb{R}$. However, since $f_3 : X_3 \to \mathbb{R}$ is linked to neither f_1 nor f_2, uniqueness of the full Kantorovich potential $f : X \to \mathbb{R}$ is not guaranteed. In sketch **(b)**, f_1 is linked to f_2 via y_1 and f_2 is linked to f_3 via y_2. Therefore, uniqueness of f follows from uniqueness of the restricted Kantorovich potentials if f^c is continuous at y_1 and y_2.

If there exists a non-degenerate optimal transport plan $\pi \in \mathcal{C}(\mu, \nu)$ with respect to which the X_i-restricted Kantorovich potentials $S_{c_{X_i}}(\mu_{X_i}, \nu_{X_i})$ are almost surely unique for all $i \in I$, the Kantorovich potentials $S_c(\mu, \nu)$ are also almost surely unique.

Example 3 (Semi-discrete optimal transport): Let X be a finite set with $n \in \mathbb{N}$ elements and Y be a general Polish space. Questions concerning dual uniqueness in this setting, which is referred to as *semi-discrete optimal transport*, have recently been raised by Altschuler et al. 2021 and Bercu and Bigot 2021. Theorem 1 provides simple and general answers in this context, since the uniqueness of the X_i-restricted Kantorovich potentials and the continuity of all $f^c \in S_c^c$ turn out to be trivial (for continuous c). For example, if $\nu \in \mathcal{P}(Y)$ has connected support, Kantorovich potentials are *always* unique in the above sense for *any* $\mu \in \mathcal{P}(X)$. If the support of ν is disconnected with (at most) countably many components Y_j, uniqueness holds if the measures μ and ν are non-degenerate in the sense of Lemma 6. In particular, when μ is the uniform distribution on X, Kantorovich potentials are guaranteed to be unique unless ν assigns a multiple of mass $1/n$ to individual connected components of supp ν.

A sketch that assists in the interpretation of Theorem 1 is provided by Figure 2. At the heart of the proof lies observation (8), which ensures that each Kantorovich potential in $S_c(\mu,\nu)$ assumes the form

$$f_a = \sum_{i\in I} \mathbb{1}_{X_i} \cdot (f_i + a_i) \qquad \text{on } p_X(\operatorname{supp}\pi)$$

for some $a \in \mathbb{R}^{|I|}$, where representatives $f_i \in S_{c_{X_i}}(\mu_{X_i}, \nu_{X_i})$ of the X_i-restricted problems have been fixed. Not each choice of a leads to a viable optimal solution $f_a \in S_c(\mu,\nu)$, however. Due to the continuity of f_a^c, one can show that a value a_{i_1} uniquely determines a_{i_2} for $i_1, i_2 \in I$ if the masses transported from X_{i_1} and X_{i_2} to Y touch one another, meaning that their topological closures have a common *contact point* (see Section 5 for formal definitions). The non-degeneracy of π then ensures the existence of suitable contact points such that fixing a_{i_0} for an arbitrary $i_0 \in I$ actually determines the whole vector a, which in consequence implies dual uniqueness. The full proof is documented in Section 5, while the following paragraphs discuss the assumptions in Theorem 1.

Degeneracy. The uniqueness of Kantorovich potentials can break down if the condition of non-degeneracy of π is not satisfied. For a simple family of examples, consider non-negative continuous and symmetric costs with $c(x,x) = 0$ for $x \in X = Y$ in a setting where $\mu = \nu$ with $\operatorname{supp}(\mu) = X_1 \cup X_2$. If the components X_1 and X_2 are strictly cost separated,

$$\Delta = \inf_{x_1\in X_1, x_2\in X_2} c(x_1, x_2) > 0, \tag{11}$$

each optimal plan π satisfies $\pi(X_1 \times X_2) = \pi(X_2 \times X_1) = 0$. In particular, no mass is transported between X_1 and X_2. In this situation, any pair $a, b \in \mathbb{R}$ with $|a - b| \leq \Delta$ defines a Kantorovich potential $f_{a,b}$ via $f_{a,b} = a$ on X_1 and $f_{a,b} = b$ on X_2. A proof of this observation is provided in Appendix A, Lemma 11.

Continuity. Even in the presence of non-degenerate optimal transport plans, the uniqueness of Kantorovich potentials can in principle still break down due to discontinuities of the cost function or the potentials. For instance, the construction of non-unique potentials $f_{a,b}$ in Lemma 11 under a Δ-separation between components of a disconnected support can easily be carried over to cost functions that exhibit a jump by Δ between two disjoint subsets of $\operatorname{supp}\mu$, even if the support is connected.

For continuous costs, we discussed in Section 2 that the c-transformed Kantorovich potentials $f^c \in S_c^c(\mu,\nu)$ are always continuous when the family $\{c(x,\cdot) \mid x \in X\}$ of partially evaluated costs is (locally) equicontinuous, or when the space X is compact (Lemma 2). We also note that conditions 1 and 2 in Theorem 1 can actually be relaxed, and it would in both cases suffice to require continuity only at the finite number of contact points (implicitly) constructed in the proof. According to Lemma 2, this is for example guaranteed if π induces regularity at each contact point $y \in Y$. In settings with $\nu(\partial \operatorname{supp}\nu) = 0$, we can sometimes even drop the additional continuity assumption altogether: if the functions f^c are known to be continuous in the interior of $\operatorname{supp}\nu$, which is true for a wide range of superlinear costs (see Section 4, which in particular covers Example 1) or costs with uniformly bounded second derivatives (see Example 2), then we can apply Lemma 3 to transition to the restricted problem with $\tilde{X} = X$ and $\tilde{Y} = \mathrm{int}(\mathrm{supp}\, \nu) \subset Y$. This reformulation, where one only has to heed possible changes in decomposition (9) when replacing Y by $\tilde{Y}$ (which may affect the degeneracy of optimal transport plans), makes sure that suitable contact points y can always be found in the interior of $\mathrm{supp}\, \nu$.

Countable index sets. Due to topological complications in the proof, the decomposition of $\mathrm{supp}\, \mu$ in Theorem 1 is restricted to a finite index set I. Indeed, if mass of an infinite number of components X_i is transported to a single component Y_j, our proof technique cannot be used to establish the existence of suitable contact points. Two settings where this issue can easily be reconciled is when 1) all sets Y_j receive mass from finitely many X_i only, or when 2) each Y_i is a single point.

> Corollary 1: Under either of the following additional assumptions, Theorem 1 remains valid for countable index sets I, where uniqueness of the X_i-restricted Kantorovich potentials is only required for $i \in I$ with $\mu(X_i) > 0$.
>
> 1) Condition 2 in Theorem 1 holds and $\left|\{i \in I \mid \pi(X_i \times Y_j) > 0\}\right| < \infty$ for all $j \in J$.
> 2) Y_j consists of a single point for each $j \in J$ with $\nu(Y_j) > 0$ (then condition 2 in Theorem 1 is always satisfied).

In the special case of statement 2 of Corollary 1 where all components X_i are also single points, we conclude that non-degeneracy of the vectors $(\mu(X_i))_{i \in I}$ and $(\nu(Y_j))_{j \in J}$ as in Lemma 6 is already sufficient to imply uniqueness of Kantorovich potentials without further assumptions. This criterion has long been established for finite transportation problems via the theory of finite linear programming (Klee and Witzgall 1968; Hung et al. 1986), but we are not aware of a comparable result that covers probability measures with countable support. In particular, we note that Corollary 1 yields sufficient conditions for Gaussian distributional limits in the countable discrete settings studied by Tameling et al. 2019.

Connected support

One piece that is still missing to fully utilize Theorem 1 is a set of criteria for the uniqueness of Kantorovich potentials on the individual connected components of the support. In Euclidean settings, results in this regard are readily available, see for example Proposition 7.18 in Santambrogio 2015 (compactly supported measures) or more recently Appendix B of Bernton et al. 2021 and Corollary 2.7 of del Barrio et al. 2021b (possibly non-compactly supported measures). The necessary techniques for these statements have long been established (Brenier 1991; Gangbo and McCann 1996) and have been extended to the more general setting of manifolds (McCann 2001; Villani 2008; Fathi and Figalli 2010; Figalli and Gigli 2011). In the following, we briefly revisit the underlying arguments and spell out a general uniqueness result together with some of its consequences for probability measures with connected support. To our knowledge, no formal statement with comparable scope has yet been assembled in this form, even though the involved arguments are well known (see for example Villani 2008, Remark 10.30).

Let $\pi \in \mathcal{C}(\mu, \nu)$ denote an optimal transport plan under a continuous cost function c. Recalling the properties of the c-transform, any Kantorovich potential $f \in S_c(\mu, \nu)$ satisfies $f(x) + f^c(y) \leq c(x, y)$ for all $(x, y) \in X \times Y$ with equality if $(x, y) \in \partial_c f$. Fixing $(x, y) \in \operatorname{supp} \pi \subset \partial_c f$, this implies

$$x' \mapsto f(x') - c(x', y) \qquad \text{is minimal at } x' = x.$$

Therefore, if X is a smooth manifold (without boundary) and the functions f as well as $x' \mapsto c(x', y)$ are both differentiable at x, it has to hold that

$$\nabla f(x) = \nabla_x c(x, y), \tag{12}$$

by which we mean equality of the respective gradients in charts of X. This relation determines the derivatives of Kantorovich potentials in the set $p_X(\Gamma) \subset X$, where we define

$$\Gamma = \big\{(x, y) \in \operatorname{supp} \pi \,\big|\, \nabla_x c(x, y) \text{ exists}\big\}. \tag{13}$$

In order to conclude uniqueness of f up to constants from characterization (12), we will make use of the following auxiliary result. A proof is provided in Appendix A.

> Lemma 7: Let $M \subset X$ be an open and connected subset of a smooth manifold X and let $f_1, f_2 \colon M \to \mathbb{R}$ be locally Lipschitz. If $\nabla f_1 = \nabla f_2$ on a set that has full Lebesgue measure in charts of M, then $f_1 - f_2$ is constant on M.

At this point, several considerations have to be taken into account.

1. The region $M \subset X$ chosen for Lemma 7 should have full μ-measure, or it should at least be possible to uniquely recover the function f from $f|_M$ on a set with full μ-measure.

2. The Kantorovich potential f must be locally Lipschitz on M in order for Lemma 7 to be applicable. This also implies the existence of ∇f in a set of full Lebesgue measure in each chart (via Rademacher's theorem).

3. The cost function has to be sufficiently smooth in its first argument along the transport. More precisely, $p_X(\Gamma)$ must have full Lebesgue measure in charts of M. Otherwise, relation (12) would not determine the gradients of f suitably for application of Lemma 7.

One immediate conclusion of the first point is the necessity of the condition $\mathrm{cl}(M) = \operatorname{supp} \mu$, without which some mass would be out of reach from the region M controlled by the gradients. The second and third points stress that the cost function should be locally Lipschitz in its first argument, in a way that is inherited to S_c. In order to choose M properly for a general uniqueness statement, we recall the notion of induced regularity defined in Section 2 and set

$$\Sigma = \big\{x \in \operatorname{supp}\mu \,\big|\, \pi \text{ does not induce regularity at } x\big\}. \tag{14}$$

We know that this set is closed (see Lemma 10 in Appendix A), that $\operatorname{supp}\mu \setminus \Sigma \subset p_X(\operatorname{supp}\pi)$, and that Σ contains all points of discontinuity of $f|_{\operatorname{supp}\mu}$ for any $f \in S_c(\mu, \nu)$ (see Lemma 2).

Theorem 2 (Uniqueness under connected support): Let X be a smooth manifold, Y be Polish, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \longrightarrow \mathbb{R}_+$ continuous such that $c(\cdot, y)$ is locally Lipschitz locally uniformly in $y \in Y$ with $T_c(\mu, \nu) < \infty$. Let Γ and Σ be as in (13) and (14) for an optimal $\pi \in \mathcal{C}(\mu, \nu)$. If

1) $\mu(\operatorname{supp}\mu \setminus \Sigma) = 1$,
2) $M = \operatorname{int}(\operatorname{supp}\mu \setminus \Sigma)$ is connected with $\operatorname{cl}(M) = \operatorname{supp}\mu$,
3) $p_X(\Gamma)$ has full Lebesgue measure in charts of M,

then the Kantorovich potentials $S_c(\mu, \nu)$ are almost surely unique.

Remark 3 (Uniqueness of optimal transport maps): Solving equation (12) for $y \in Y$ is a standard method to construct and study optimal transport maps $t\colon X \longrightarrow Y$, see Villani 2008, Chapter 10. In this context, a natural requirement is the injectivity of $\nabla_x c(x, \cdot)$, denoted as the *twist condition*, which also implies that an optimal map is uniquely defined wherever (12) holds. To make sure that t is determined by (12) μ-almost surely, one usually imposes some form of regularity on c (such as local semiconcavity) and requires the probability measure μ to assign no mass to sets on which Kantorovich potentials may be non-differentiable (e.g., by assuming a Lebesgue density). Theorem 2 helps clarify to which extent similar assumptions on c and μ are necessary if we are only interested in uniqueness of the Kantorovich potentials, and not in the uniqueness of optimal maps.

Proof. According to Lemma 2, each point of M admits an open neighborhood U and a compactum $K \subset Y$ such that $f(x) = \inf_{y\in K} c(x, y) - f^c(y)$ for all $x \in U$ and $f \in S_c(\mu, \nu)$. Since $c(\cdot, y)$ is locally Lipschitz uniformly in $y \in K$ by assumption, we conclude that f is locally Lipschitz on M. Now, let $f_1, f_2 \in S_c(\mu, \nu)$ and let A be the subset of M where both functions are differentiable. Set $B = A \cap p_X(\Gamma)$, which is the set where the gradients of f_1 and f_2 have to coincide via (12). Due to Rademacher's Theorem (see, e.g., Federer 2014, Theorem 3.1.6) and assumption 3, the set B has full Lebesgue measure in each chart of M. We can thus apply Lemma 7 under assumption 2 and conclude that $f_1 = f_2$ (up to a constant) on M. Since any $f \in S_c(\mu, \nu)$ is continuous at each point in $\operatorname{supp}\mu \setminus \Sigma \subset \operatorname{cl}(M)$ via Lemma 2, the function f is uniquely determined on $\operatorname{supp}\mu \setminus \Sigma$ by its values on M. This shows that $f_1 = f_2$ (up to a constant) on $\operatorname{supp}\mu \setminus \Sigma$, which is a set of full μ-measure by condition 1. □

Without further context, the assumptions in this theorem may seem to be fairly opaque, as they rely on specific details of an optimal transport plan π, like the μ-measure of Σ and topological properties of $\operatorname{supp}\mu \setminus \Sigma$. In more specialized settings, however, the requirements of Theorem 2 can often be checked easily. As a first example, we assume Y to be compact. If c is differentiable in its first component, then all assumptions of Theorem 2 that depend on π are automatically satisfied and only conditions on the topology of $\operatorname{supp}\mu$ remain.

> Corollary 2: Let X be a smooth manifold, Y be compact Polish, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \longrightarrow \mathbb{R}_+$ continuous such that $c(\cdot, y)$ is differentiable and locally Lipschitz uniformly in $y \in Y$ with $T_c(\mu, \nu) < \infty$. If $M = \operatorname{int}(\operatorname{supp}\mu)$ is connected and $\operatorname{cl}(M) = \operatorname{supp}\mu$, then the Kantorovich potentials $S_c(\mu, \nu)$ are almost surely unique.

Proof. For compact Y, it holds by definition that $\Sigma = \emptyset$. Thus, conditions 1 and 2 of Theorem 2 are satisfied. Condition 3 is ensured since c is differentiable in the first component and we thus find $\Gamma = p_X(\operatorname{supp}\pi) = \operatorname{supp}\mu$ (since the projection p_X is a closed map for compact Y). □

If Y is not compact, uniqueness statements based on Theorem 2 hinge on the behavior of the cost function outside of Y-compacta. The simplest setting of this kind is the one where $c(\cdot, y)$ is (locally) Lipschitz uniformly in $y \in Y$. Then, a statement analogous to Corollary 2 is possible, where the only obstacle is to assert condition 3 of Theorem 2. To provide a convenient sufficient criterion to this end, we work with the assumption that

$$\lambda \text{ is absolutely continuous w.r.t. } \varphi_\#\mu \text{ on range } \varphi \tag{15}$$

for any chart φ of $M = \operatorname{int}(\operatorname{supp}\mu)$, where $\varphi_\#\mu := \mu \circ \varphi^{-1}$ corresponds to the push-forward measure of μ under φ and λ denotes the Lebesgue measure. Loosely speaking, this property states that the mass of μ can be placed quite arbitrarily on $\operatorname{supp}\mu$, as long as it contains a continuous component everywhere on its support. As an example where condition (15) fails for $X = \mathbb{R}$, consider a measure μ that is concentrated on the rational numbers $\mathbb{Q}$ but where $\operatorname{supp}\mu$ contains an open set.

> Corollary 3: Let X be a smooth manifold, Y Polish, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \longrightarrow \mathbb{R}_+$ continuous such that $c(\cdot, y)$ is differentiable and locally Lipschitz uniformly in $y \in Y$ with $T_c(\mu, \nu) < \infty$. If $M = \operatorname{int}(\operatorname{supp}\mu)$ is connected such that $\operatorname{cl}(M) = \operatorname{supp}\mu$ and condition (15) holds, then the Kantorovich potentials $S_c(\mu, \nu)$ are almost surely unique.

Proof. Since $c(\cdot, y)$ is assumed to be locally Lipschitz uniformly in $y \in Y$, every Kantorovich potential $f \in S_c(\mu, \nu)$ is locally Lipschitz on $\operatorname{supp}\mu$. Looking at the proof of Theorem 2, we furthermore note that the set Σ can actually be replaced by any other subset of X that contains all points at which some Kantorovich potential fails to be locally Lipschitz. Therefore, we may functionally assume $\Sigma = \emptyset$ in conditions 1 and 2 of Theorem 2, and only have to show that condition 3 holds with $M = \operatorname{int}(\operatorname{supp}\mu)$. Since $\Gamma = \operatorname{supp}\pi$ due to differentiability of c, we find a Borel set $A \subset p_X(\Gamma) \subset X$ that satisfies $\mu(A) = 1$ (Lemma 9). Hence, $\operatorname{range}\varphi \setminus \varphi(A \cap \operatorname{domain}\varphi)$ is a $\varphi_\#\mu$-null set for any chart φ of M. Due to condition (15), it is also a Lebesgue-null set and we conclude that $p_X(\Gamma)$ has full Lebesgue measure in charts of M. □

Remark 4 (local semiconcavity): In Example 2, we pointed out that Kantorovich potentials can inherit semiconcavity from the cost function. In fact, it is possible to formulate Corollary 3 for costs where $c(\cdot, y)$ is locally semiconcave (instead of locally Lipschitz) uniformly in $y \in Y$, in the sense of equation (2). This alternative formulation, whose proof is documented in Appendix A, can be put to use in settings where Corollary 3 might not apply directly. For example, let X and Y be two (possibly distinct) affine subspaces of $\mathbb{R}^d$ equipped with an atlas of linear charts. Then the squared Euclidean cost function $c(x, y) = \|x - y\|^2$ is differentiable and locally semiconcave (but *not* necessarily locally Lipschitz) in x uniformly in y. Consequently, the Kantorovich potentials in this setting are unique under the mild conditions on μ imposed by Corollary 3. If $\operatorname{supp}\mu$ and $\operatorname{supp}\nu$ are separated sets, the same holds for Euclidean cost functions $c(x, y) = \|x - y\|^p$ with $0 < p \leq 2$.

One question largely unaddressed by the previous results is how Theorem 2 fares in the general setting that $c(\cdot, y)$ is actually no more than locally Lipschitz *locally* uniformly in $y \in Y$, which is typically the case for rapidly growing cost functions. In the next section, we show that such cost functions often confine the set Σ to the boundary of $\operatorname{supp}\mu$, which makes the application of Theorem 2 particularly simple: condition 1 collapses into $\mu(\partial\operatorname{supp}\mu) = 0$ and condition 2 only relies on the topology of $\operatorname{supp}\mu$. Furthermore, condition 3 is always satisfied when c is differentiable in the first component, since then $\Gamma = \operatorname{supp}\pi$ and $M \subset \operatorname{supp}\mu \setminus \Sigma \subset p_X(\Gamma)$ via Lemma 2.

4 Interior regularity

We now investigate conditions on the cost function under which each optimal transport plan π induces regularity in the interior of the support of μ. In other words, we search for criteria that ensure

$$\Sigma \subset \partial\operatorname{supp}\mu, \tag{16}$$

where $\Sigma \subset \operatorname{supp}\mu$ was defined in (14) and denotes the set of points where induced regularity fails. This property is of interest for both Theorem 1 (applied to ν in place of μ) and Theorem 2, as it guarantees continuity of Kantorovich potentials in the interior of the support. Instead of working with definition (6) of induced regularity directly, we will show the slightly stronger result that only boundary points can be "sent towards infinity", which implies (16). To achieve this, the cost function has to behave in a certain way if y leaves all compacta in Y. For convenience, we write $y_n \longrightarrow \infty$ to denote that each compact set in Y contains only a finite number of elements of the sequence $(y_n)_{n\in\mathbb{N}} \subset Y$. Despite the suggestive notation, we want to point out that $y_n \longrightarrow \infty$ does not necessarily imply that y_n leaves all bounded sets if Y is not a proper metric space. We also define the *region of dominated cost*

$$C(x, y) = \left\{x' \in X \,\middle|\, c(x', y) \leq c(x, y)\right\} \tag{17}$$

for any $(x, y) \in X \times Y$. In Euclidean settings, for example, under costs $c(x, y) = \|x - y\|^p$ for $x, y \in \mathbb{R}^d$ and $p \geq 1$, the set $C(x, y)$ is the closed ball centered at y with radius $\|x - y\|$. As we will see next, the geometry of $C(x, y)$ as $y \longrightarrow \infty$ shapes the region where Kantorovich potentials do not attain finite values.

Lemma 8 (Interior regularity): Let X and Y be Polish, $c\colon X \times Y \longrightarrow \mathbb{R}_+$ be continuous, and $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ with $T_c(\mu, \nu) < \infty$. Let $\pi \in \mathcal{C}(\mu, \nu)$ be optimal and $x \in \operatorname{supp}\mu$ such that there exists $(x_n, y_n)_{n\in\mathbb{N}} \subset \operatorname{supp}\pi$ with $x_n \longrightarrow x$ and $y_n \longrightarrow \infty$ as $n \longrightarrow \infty$. If

1) there is $(\tilde{x}_n)_n \subset X$ converging to x and $\lim_{n\to\infty} c(\tilde{x}_n, y_n) - c(x_n, y_n) = -\infty$,

then all $f \in S_c(\mu, \nu)$ assume the value $-\infty$ on $C_\infty = \limsup_{n\to\infty} C(\tilde{x}_n, y_n)$. If additionally

2) C_∞ contains an open subset U that touches x, meaning $x \in \operatorname{cl}(U)$,

then $x \in \partial \operatorname{supp}\mu$.

Proof. Let $x' \in C_\infty$. After a suitable subsequence has been taken, we may assume that $x' \in C(\tilde{x}_n, y_n)$ for all $n \in \mathbb{N}$. For $f \in S_c(\mu, \nu)$, note $f(x_n) = c(x_n, y_n) - f^c(y_n)$ and observe

$$\begin{aligned} f(x') &\leq c(x', y_n) - f^c(y_n) \\ &\leq c(\tilde{x}_n, y_n) - f^c(y_n) \\ &= c(\tilde{x}_n, y_n) - c(x_n, y_n) + f(x_n) \longrightarrow -\infty \end{aligned}$$

due to condition 1 and the upper-semicontinuity of f, which implies $\sup_{n\in\mathbb{N}} f(x_n) < \infty$. This shows the first claim. To show the second claim, we lead $x \in \operatorname{int}(\operatorname{supp}\mu) \cap \operatorname{cl}(U)$ to a contradiction: by density of $p_X(\operatorname{supp}\pi)$ in $\operatorname{supp}\mu$, such an x would imply $p_X(\operatorname{supp}\pi) \cap U \neq \emptyset$. Thus, there would exist $(x', y') \in \operatorname{supp}\pi \subset \partial_c f$ with $-\infty = f(x') = c(x', y') - f^c(y') > -\infty$. □

The assumptions in Lemma 8 deserve some more context and discussion. First, note that any $x \in \Sigma$ admits a suitable sequence $(x_n, y_n)_n \subset \operatorname{supp}\pi$ with $x_n \longrightarrow x$ and

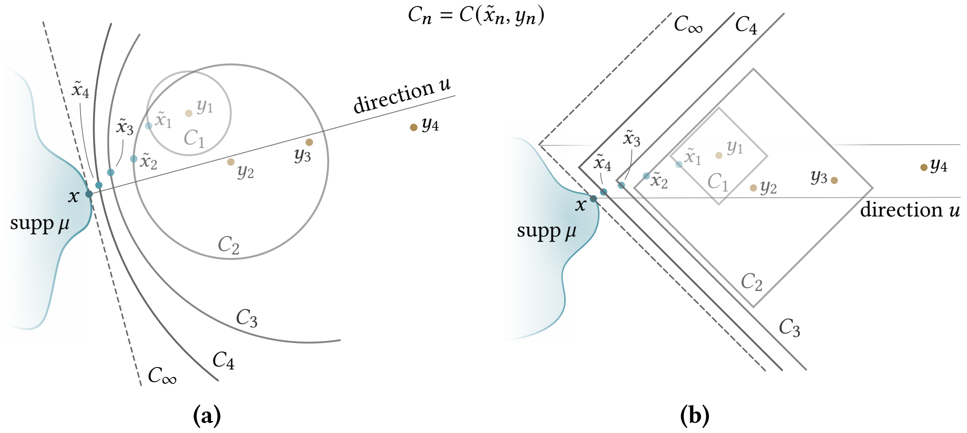


Figure 3: Asymptotic regions of dominated cost. Both figures depict the sets $C(\tilde{x}_n, y_n)$ as defined in (17) for a sequence $(\tilde{x}_n, y_n)_{n\in\mathbb{N}}$ with $\tilde{x}_n \longrightarrow x$ and $y_n \longrightarrow \infty$ as $n \longrightarrow \infty$, where $u = \lim_{n\to\infty}(y_n - \tilde{x}_n)/\|y_n - \tilde{x}_n\|$. In sketch **(a)**, the cost function $c(x, y) = h\big(\|x - y\|\big)$ for the Euclidean norm and some strictly increasing h is chosen, while sketch **(b)** shows a similar setting for costs based on the l_1 norm. In both examples, condition 1 of Lemma 8 is always satisfied if h is differentiable and $h'(a) \longrightarrow \infty$ as $a \longrightarrow \infty$ (see Theorem 3). Therefore, all Kantorovich potentials assume the value $-\infty$ on the region C_∞ if mass is transported from x towards infinity in direction u. For illustration, **(b)** shows the rather exceptional case where C_∞ is *not* a half space, which (in this example) can only happen if u is aligned with one of the coordinate axes. More precisely, the depicted shape is only possible if the vertical coordinates of y_n converge to the vertical coordinate of the apex of C_∞.

$y \longrightarrow \infty$, meaning that Lemma 8 can be applied to all points at which π fails to induce regularity. This is a direct implication of definition (6). Furthermore, we observe that conditions 1 and 2 in the presented formulation of Lemma 8 rely not only on the cost function c, but also on the support of π (via the points x_n and y_n). Typically, however, these two conditions can be shown to hold for *any* sequences $x_n \longrightarrow x \in X$ and $y_n \longrightarrow \infty$, which makes them an assumption on the cost function only. In fact, condition 1 is implied by property $(\mathrm{H}_\infty)_2$ in Villani 2008, Chapter 10, which can be viewed as a condition on the growth behavior of the cost function as $y_n \longrightarrow \infty$.

Remark 5: Lemma 8 shows that optimal transport towards infinity places conditions on the set $\{f = -\infty\} \subset X$ for $f \in S_c(\mu, \nu)$, at least for rapidly growing cost functions. In a certain sense, this relation can be reversed. If $f(x) = \inf_{y\in Y} c(x, y) - f^c(y) = -\infty$ for any $x \in X$ (not necessarily in supp μ), it follows that there exists $(y_n)_{n\in\mathbb{N}}$ with $c(x, y_n) - f^c(y_n) \longrightarrow -\infty$ as $n \longrightarrow \infty$. Due to the upper-semicontinuity of f^c, this implies $y_n \longrightarrow \infty$. Moreover, it is straightforward to see that f has to assume the value $-\infty$ on the set $C_\infty = \limsup_{n\to\infty} C(x, y_n)$. Therefore, if condition 2 of Lemma 8 holds, this set has to be disjoint from the interior of supp μ.

Example 4: Let $X = Y = \mathbb{R}^d$ with squared Euclidean costs $c(x, y) = \|x - y\|^2$. Given any sequence $x_n \longrightarrow x \in \mathbb{R}^d$ and $y_n \longrightarrow \infty$, the choice $\tilde{x}_n = x_n + (y_n -$

$x_n)/\|y_n - x_n\|^{3/2}$ provides a perturbation of x_n that satisfies condition 1 in Lemma 8. To see that condition 2 is also satisfied, note that the asymptotic set C_∞ will contain an open half-space anchored at the point x (see Figure 3a for an illustration). The (inwards pointing) normal direction is given by an (arbitrary) limit point u of the directions $u_n = (y_n - \tilde{x}_n)/\|y_n - \tilde{x}_n\|$. Such a limit exists, since the unit sphere in $\mathbb{R}^d$ is compact. Therefore, Lemma 8 lets us conclude (16) and provides additional insights about the set $\{f = -\infty\}$ and its relation to the direction u of transport towards infinity. Indeed, the fact that C_∞ always contains half-spaces (which is also true for all other l_p costs for $p > 1$, but not necessarily for $p = 1$, see Figure 3b) implies that the interior of the convex hull of $\operatorname{supp}\mu$ is contained in the set $\{f < \infty\}$ (see Remark 5).

We stress that the reasoning in this example can easily be extended to other costs on $\mathbb{R}^d$ and even non-Euclidean spaces. For instance, Gangbo and McCann 1996 work with cost functions of the form $c(x, y) = g(x-y)$ on $\mathbb{R}^d$ for a convex and superlinear g, which automatically implies condition 1 of Lemma 8. They also consider a geometric monotonicity condition on g, which ensures that, for y large, the sets $C(x, y)$ contain broad cones with apex x (implying condition 2). Similarly, the validity of (16) can be exposed in geodesic spaces as well.

Theorem 3 (Interior regularity in geodesic spaces): Let $X = Y$ be a locally compact complete geodesic space with metric d and let $c\colon X^2 \to \mathbb{R}_+$ be of the form $c(x, y) = h\big(d(x, y)\big)$, where $h\colon \mathbb{R}_+ \to \mathbb{R}_+$ is differentiable with $\lim_{a\to\infty} h'(a) = \infty$. Then

$$\Sigma \subset \partial \operatorname{supp}\mu$$

for any $\mu, \nu \in \mathcal{P}(X)$ with $T_c(\mu, \nu) < \infty$, where Σ is defined in (14) for $\pi \in \mathcal{C}(\mu, \nu)$ optimal.

Proof. Let $x_n \to x \in X$ and $y_n \to \infty$. Since X is a proper metric space (e.g., by the Hopf-Rinow theorem as stated in Bridson and Haefliger 2013, Proposition 3.7), each closed ball in X is compact. This implies $r_n = d(x_n, y_n) \to \infty$. Next, let $g\colon \mathbb{R}_+ \to \mathbb{R}$ be a function that satisfies $g \le h'$ and $g(a) \to \infty$ as $a \to \infty$. For example, one can pick $g(a) = \inf_{b\ge a} h'(b)$. We also let $\gamma_{x'x''}\colon [0, d(x', x'')] \to X$ denote a geodesic connecting x' to x'' in X.

To show condition 1 of Lemma 8, let $\tilde{x}_n = \gamma_{x_n y_n}(t_n)$ for $0 < t_n < 1$, meaning that x_n is pushed towards y_n along a geodesic to generate perturbations $\tilde{x}_n$. The amount t_n by which x_n is pushed is chosen to satisfy $t_n \to 0$ and $t_n\, g(r_n - 1) \to \infty$ as $n \to \infty$. Since $d(x, \tilde{x}_n) \le d(x, x_n) + t_n$, the points $\tilde{x}_n$ indeed converge to x. We also note that $d(\tilde{x}_n, y_n) = r_n - t_n$ and observe, as $n \to \infty$,

$$c(\tilde{x}_n, y_n) - c(x_n, y_n) = h(r_n - t_n) - h(r_n) \le -t_n\, g(r_n - 1) \to -\infty.$$

To verify condition 2, let $u_n = \gamma_{\tilde{x}_n y_n}(1)$ and pick a limit point $u \in X$ of this sequence.

Such a point exists, since the points u_n are bounded and X is proper. By selecting a suitable subsequence, we may assume $u_n \to u$. Let U be the open unit ball at u. We will show $U \subset \limsup C(\tilde{x}_n, y_n)$ and $x \in \mathrm{cl}(U)$. The latter is evident, since $d(x, u) \leq d(x, \tilde{x}_n) + d(\tilde{x}_n, u_n) + d(u_n, u) \to 1$ as $n \to \infty$. To see the former, let $x' \in U$ and note that $d(x', u_n) < 1 = d(\tilde{x}_n, u_n)$ for large n. Then

$$d(x', y_n) \leq d(x', u_n) + d(u_n, y_n) < d(\tilde{x}_n, u_n) + d(u_n, y_n) = d(\tilde{x}_n, y_n).$$

For n large enough such that h can be assumed to be increasing, this implies $h\big(d(x', y_n)\big) \leq h\big(d(\tilde{x}_n, y_n)\big)$ and thus $x' \in C(\tilde{x}_n, y_n)$. □

The proof above shows that the asymptotic region C_∞ in Lemma 8 at the very least contains the unit ball touching x centered at a suitable $u \in X$. Of course, this approach can also be extended to balls of arbitrary radius $r > 1$, which provides additional insight about the geometry of C_∞. Like in Example 4, a central argument is the compactness of closed balls, which guarantees the existence of asymptotic directions u of transport towards infinity.

5 Proofs of the main results

In the following, we formulate the proofs of the uniqueness statements for Kantorovich potentials under disconnected support, Theorem 1 and Corollary 1. For reasons of exposition, we start with Theorem 1 under continuity assumption 2, before we document the adjustments necessary to prove the theorem under the alternative assumption 1, which requires a slightly different strategy. Afterwards, we provide the arguments to extend Theorem 1 to countable I as claimed in Corollary 1. For notational convenience, we denote the topological closure of a set A by $\overline{A}$ instead of $\mathrm{cl}(A)$ in this section.

Proof of Theorem 1 under assumption 2. Recall decomposition (9) of the support of μ and ν into connected components $(X_i)_{i\in I}$ and $(Y_j)_{j\in J}$ for finite I and countable J. Since we consider the second condition of Theorem 1 first, we can assume for each $f^c \in S_c^c(\mu, \nu)$ and $j \in J$ with $\nu(Y_j) > 0$ that $f^c|_{Y_j}$ is continuous and that the set $\tilde{Y}_j = \operatorname{supp} \nu|_{Y_j}$ is connected. As I is finite, each X_i is open in $\operatorname{supp} \mu$ and consequently satisfies $\mu(X_i) > 0$. Therefore, the X_i-restricted optimal transport problem is well defined and we can fix a representative $f_i \in S_{c_{X_i}}(\mu_{X_i}, \nu_{X_i})$ for each $i \in I$. The uniqueness assumption in Theorem 1 together with Lemma 5 implies that f_i is uniquely determined on $p_X(\operatorname{supp} \pi) \cap X_i$ (up to an additive offset), so its actual choice does not matter to us. Applying Lemma 4, we conclude that each $f \in S_c(\mu, \nu)$ can be assigned a unique *offset vector* $a = (a_i)_{i\in I} \in \mathbb{R}^{|I|}$ with components $a_i = f(x_i) - f_i(x_i)$, where the point $x_i \in p_X(\operatorname{supp} \pi) \cap X_i$ can be chosen arbitrarily. We suggestively write $f = f_a$ if $f \in S_c^c(\mu, \nu)$ has offset vector a and emphasize that the equality

$$f_a = \sum_{i\in I} \mathbb{1}_{X_i} \cdot (f_i + a_i) \tag{18}$$

holds on $p_X(\operatorname{supp}\pi)$. Clearly, two Kantorovich potentials in $S_c(\mu,\nu)$ have identical offset vectors if and only if they coincide on $p_X(\operatorname{supp}\pi)$. Therefore, almost sure uniqueness of $S_c(\mu,\nu)$ follows if we can show that there is only a single feasible offset vector a (up to an additive constant that is the same in each component). To formalize this idea, we divide the support of π into closed disjoint pieces $\Gamma_{ij} = \operatorname{supp}\pi \cap (X_i \times Y_j)$, and we say that two indices i_1 and i_2 in I are *linked* if there exists a *contact index* $j \in J$ with $\nu(Y_j) > 0$ and a *contact point* $y \in Y_j$ such that

$$y \in \overline{p_Y(\Gamma_{i_1 j})} \cap \overline{p_Y(\Gamma_{i_2 j})}. \tag{19}$$

Intuitively, two indices in I are linked if the masses transported from X_{i_1} and X_{i_2} to Y_j touch one another at a common point $y \in Y_j$. In a first step, we establish that

$$i_1 \text{ and } i_2 \text{ are linked} \qquad \text{implies} \qquad a_{i_1} - a_{i_2} \text{ is fixed}$$

under the adopted continuity assumptions on $S_c^c(\mu,\nu)$, where the right hand side indicates that the difference $a_{i_1} - a_{i_2}$ has to be the same for all feasible offset vectors. In a second step, we then show that non-degeneracy of the optimal plan π guarantees that there are enough contact points to connect all indices in I, which will conclude the proof.

Step 1. Let i_1 and i_2 be indices in I that are linked through a contact point $y \in Y_j$ for $j \in J$. According to (19), there are sequences $(x_n, y_n)_n \subset \Gamma_{i_1 j}$ and $(x'_n, y'_n)_n \subset \Gamma_{i_2 j}$ such that $y_n \longrightarrow y$ and $y'_n \longrightarrow y$ in Y_j as $n \longrightarrow \infty$. Since $f_a \oplus f_a^c = c$ on $\operatorname{supp}\pi$ as well as $f_a = f_{i_1} + a_{i_1}$ and $f_a = f_{i_2} + a_{i_2}$ on $p_X(\Gamma_{i_1 j})$ respectively $p_X(\Gamma_{i_2 j})$ due to relation (18), we find

$$a_{i_1} - a_{i_2} = \big(c(x_n, y_n) - f_{i_1}(x_n) - f_a^c(y_n)\big) - \big(c(x'_n, y'_n) - f_{i_2}(x'_n) - f_a^c(y'_n)\big)$$

for all $n \in \mathbb{N}$. Exploiting the continuity of $f_a^c|_{Y_j}$ at the contact point $y \in Y_j$, we thus obtain

$$\begin{aligned} a_{i_1} - a_{i_2} &= \lim_{n\to\infty} \big(c(x_n, y_n) - f_{i_1}(x_n) - f_a^c(y_n)\big) - \big(c(x'_n, y'_n) - f_{i_2}(x'_n) - f_a^c(y'_n)\big) \\ &= \lim_{n\to\infty} c(x_n, y_n) - c(x'_n, y'_n) - f_{i_1}(x_n) + f_{i_2}(x'_n). \end{aligned}$$

Crucially, the limit in the second line exists and does not depend on a anymore. It only depends on the cost function c, the restricted potentials f_i, and the sets Γ_{ij} (determined by π), whose topologies decide the contact point y and the involved sequences. Hence, knowing the value of a_{i_1} determines the one of a_{i_2} and vice versa.

Step 2. It is left to show that all indices are linked, at least indirectly, such that the vector a is in fact determined by fixing a single component. To do so, we consider an arbitrary decomposition $I = I_1 \cup I_2$ of the index set I into a disjoint union of non-empty subsets, and show that there always exist $i_1 \in I_1$ and $i_2 \in I_2$ that are linked. First, define

$$J_1 = \big\{ j \in J \,\big|\, \pi(\Gamma_{ij}) > 0 \text{ for some } i \in I_1 \big\} \tag{20}$$

and analogously J_2. Intuitively, an index j is in J_1 (or J_2) if π transports mass between Y_j and some component X_i with $i \in I_1$ (or $i \in I_2$). We note that the sets J_1 and J_2 cannot be disjoint: if they were, then π would transport all mass in $\bigcup_{i \in I_1} X_i$ to $\bigcup_{j \in J_1} Y_j$ and vice versa, contradicting the condition of non-degeneracy. Formally, this follows from $0 < \sum_{i \in I_1} \mu(X_i) < 1$ and

$$\begin{aligned}
\sum_{i \in I_1} \mu(X_i) &= \sum_{i \in I_1} \pi\big(\operatorname{supp} \pi \cap (X_i \times Y)\big) \\
\text{(definition of } J_1\text{)} \quad &= \sum_{i \in I_1} \sum_{j \in J_1} \pi\left(\Gamma_{ij}\right) \\
(J_1 \text{ and } J_2 \text{ disjoint}) \quad &= \sum_{j \in J_1} \pi\big(\operatorname{supp} \pi \cap (X \cap Y_j)\big) = \sum_{j \in J_1} \nu(Y_j),
\end{aligned}$$

the former of which holds since I_1 is nonempty and a proper subset of I. Therefore, J_1 and J_2 are not disjoint and we find some $j \in J_1 \cap J_2$. By definition of J_1 and J_2, this implies that the two sets

$$B_1 = \bigcup_{i \in I_1} \overline{p_Y(\Gamma_{ij})} \subset Y_j \qquad \text{and} \qquad B_2 = \bigcup_{i \in I_2} \overline{p_Y(\Gamma_{ij})} \subset Y_j \tag{21}$$

have positive ν-mass and are thus non-empty. Since $\pi\big(\bigcup_{i \in I} \Gamma_{ij}\big) = \pi(X \times Y_j) = \nu(Y_j) > 0$, we can apply (a suitably restricted version of) Lemma 9 to conclude that $p_Y\big(\bigcup_{i \in I} \Gamma_{ij}\big) = \bigcup_{i \in I} p_Y(\Gamma_{ij})$ contains a subset that is dense in $\tilde{Y}_j = \operatorname{supp} \nu|_{Y_j}$. Thus, we observe

$$\tilde{Y}_j \subset \overline{\bigcup_{i \in I} p_Y(\Gamma_{ij})} = \overline{\bigcup_{i \in I_1} p_Y(\Gamma_{ij})} \cup \overline{\bigcup_{i \in I_2} p_Y(\Gamma_{ij})} = B_1 \cup B_2, \tag{22}$$

where the last equality hinges on the fact that B_1 and B_2 are closed (this is where we need the assumption that I is finite). Since $\nu(B_k) = \nu|_{Y_j}(B_k) > 0$, we find that $\tilde{B}_k = B_k \cap \tilde{Y}_j$ is non-empty for $k \in \{1, 2\}$. Together with the connectedness of $\tilde{Y}_j = \tilde{B}_1 \cup \tilde{B}_2$, this implies that $\tilde{B}_1$ and $\tilde{B}_2$ are not disjoint (since closed disjoint sets can be separated by open neighborhoods in metric spaces) and the intersection $\tilde{B}_1 \cap \tilde{B}_2$ hence contains at least one element $y \in Y_j$. In particular, there also exist $i_1 \in I_1$ and $i_2 \in I_2$ such that

$$y \in \overline{p_Y(\Gamma_{i_1 j})} \cap \overline{p_Y(\Gamma_{i_2 j})},$$

which means that i_1 and i_2 are linked with contact index j and contact point y. We have thus shown that any proper decomposition $I = I_1 \cup I_2$ admits links between the components I_1 and I_2, implying that all indices in I can be connected by a chain of links. As discussed above, this makes the Kantorovich potentials $f_a \in S_c(\mu, \nu)$ almost surely unique and finishes the proof of Theorem 1 under continuity assumption 2. □

Proof of Theorem 1 under assumption 1. The preceding proof has to be adapted to some degree if we work with the slightly stronger continuity requirement that

$f^c|_{\operatorname{supp}\nu}$ is continuous for each $f^c \in S_c^c(\mu,\nu)$, but in turn do not require any topological features of $\operatorname{supp}\nu|_{Y_j}$. The main difference is that we now allow a contact point $y \in \operatorname{supp}\nu$ to be reached along sequences that hop through different components Y_j (while j was considered fixed for such sequences before). Thus, we let $\Gamma_i = \operatorname{supp}\pi \cap (X_i \times Y) = \bigcup_{j\in J}\Gamma_{ij}$ and this time define $i_1, i_2 \in I$ to be linked if there exists a contact point $y \in \operatorname{supp}\nu$ such that

$$y \in \overline{p_Y(\Gamma_{i_1})} \cap \overline{p_Y(\Gamma_{i_2})}, \tag{23}$$

which replaces definition (19). In particular, we do not care about the contact index anymore. It is now easy to check that continuity of $f_a^c|_{\operatorname{supp}\nu}$ is sufficient for **step 1** of the proof above to work as before, and we find that $a_{i_1} - a_{i_2}$ is fixed if i_1 and i_2 are linked in the sense of (23).

For **step 2**, we choose the same approach as above and again exploit the non-degeneracy of π to find a suitable index $j \in J_1 \cap J_2$ with J_1 and J_2 defined as in (20). Then, however, we define the sets

$$B_1 = Y_j \cap \bigcup_{i\in I_1}\overline{p_Y(\Gamma_i)} \qquad \text{and} \qquad B_2 = Y_j \cap \bigcup_{i\in I_2}\overline{p_Y(\Gamma_i)}$$

somewhat differently, which is better aligned with (23). These sets are again closed (use that I is finite) and have positive ν-mass (follows from the definition of J_1 and J_2). Furthermore, $p_Y(\operatorname{supp}\pi) = p_Y(\bigcup_{i\in I}\Gamma_i)$ is dense in Y_j, which leads to $Y_j = B_1 \cup B_2$ along similar lines as in equation (22). Connectedness of Y_j thus shows that $B_1 \cap B_2$ cannot be empty, from which the existence of $y \in Y_j$ as well as $i_1 \in I_1$ and $i_2 \in I_2$ that satisfy (23) follows. By the same argument as before, the claim of the theorem is established. □

Proof of Corollary 1. There are two issues that arise in the proof of Theorem 1 when I is allowed to be countable. The first is that some components X_i might now have a μ-measure of zero, for which the notion of the X_i-restricted transport problem ceases to make sense. This can be reconciled by replacing the index set I by $I_+ = \{i \in I \mid \mu(X_i) > 0\}$ throughout the proof. Then, representation (18) of f_a only works on the set $p_X(\operatorname{supp}\pi) \cap \bigcup_{i\in I_+} X_i$, which is, however, sufficient for almost sure uniqueness.

The second issue concerns the sets B_1 and B_2 constructed in equation (21). For countable I, these sets do in general not have to be closed, which would invalidate the ensuing argumentation. For the two settings described in Corollary 1, however, this can easily be fixed. First, if the index set $I^j = \{i \in I \mid \pi(X_i \times Y_j) > 0\}$ has finite cardinality, then we may as well work with the alternative sets

$$B_1 = \bigcup_{i\in I_1\cap I^j}\overline{p_Y(\Gamma_{ij})} \subset Y_j \qquad \text{and} \qquad B_2 = \bigcup_{i\in I_2\cap I^j}\overline{p_Y(\Gamma_{ij})} \subset Y_j,$$

for which the remainder of the proof works just as before. Since the unions are finite, these sets are closed. Secondly, if $Y_j = \{y_j\}$ for $y_j \in Y$ consists of a single point only, then noting that B_1 and B_2 in (21) are both non-empty already establishes $B_1 = B_2 = Y_j$, directly yielding the desired contact point $y = y_j$. In this case, the continuity of $f^c|_{Y_j}$ and the connectedness of supp $\nu|_{Y_j}$ are trivially true. □

Acknowledgements

Thomas Staudt and Shayan Hundrieser acknowledge support of DFG RTG 2088 and Axel Munk of DFG CRC 1456. Axel Munk is supported by, and Thomas Staudt and Shayan Hundrieser were in part funded by, the DFG under Germany's Excellence Strategy EXC 2067/1-390729940.

Bibliography

Ahmad, N., H. K. Kim, and R. J. McCann (2011). "Optimal transportation, topology and uniqueness". *Bulletin of Mathematical Sciences* 1.1, pp. 13–32.

Altschuler, J. M., J. Niles-Weed, and A. J. Stromme (2021). "Asymptotics for semi-discrete entropic optimal transport". *Preprint arXiv:2106.11862.*

Ambrosio, L. and T. Rajala (2014). "Slopes of Kantorovich potentials and existence of optimal transport maps in metric measure spaces". *Annali di Matematica Pura ed Applicata* 193.1, pp. 71–87.

Ambrosio, L., D. Semola, and E. Brué (2021). *Lectures on Optimal Transport.* Springer.

Arjovsky, M., S. Chintala, and L. Bottou (2017). "Wasserstein generative adversarial networks". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, pp. 214–223.

Beiglböck, M. and W. Schachermayer (2011). "Duality for Borel measurable cost functions". *Transactions of the American Mathematical Society* 363.8, pp. 4203–4224.

Bercu, B. and J. Bigot (2021). "Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures". *The Annals of Statistics* 49.2, pp. 968–987.

Bernton, E., P. Ghosal, and M. Nutz (2021). "Entropic optimal transport: Geometry and large deviations". *Preprint arXiv:2102.04397.*

Bertrand, J. (2008). "Existence and uniqueness of optimal maps on Alexandrov spaces". *Advances in Mathematics* 219.3, pp. 838–851.

Brenier, Y. (1991). "Polar factorization and monotone rearrangement of vector-valued functions". *Communications on Pure and Applied Mathematics* 44.4, pp. 375–417.

Bridson, M. R. and A. Haefliger (2013). *Metric Spaces of Non-Positive Curvature.* Springer Science & Business Media.

Caffarelli, L. A. (1990). "A localization property of viscosity solutions to the Monge-Ampere equation and their strict convexity". *Annals of Mathematics* 131.1, pp. 129–134.

Caffarelli, L. A. (1991). "Some regularity properties of solutions of Monge Ampere equation". *Communications on Pure and Applied Mathematics* 44.8-9, pp. 965–969.

Caffarelli, L. A. (1992). “The regularity of mappings with a convex potential”. *Journal of the American Mathematical Society* 5.1, pp. 99–104.

Cordero-Erausquin, D. and A. Figalli (2019). “Regularity of monotone transport maps between unbounded domains”. *Discrete & Continuous Dynamical Systems* 39.12, pp. 7101–7112.

Cuesta, J. A. and C. Matrán (1989). “Notes on the Wasserstein metric in Hilbert spaces”. *The Annals of Probability*, pp. 1264–1276.

De Philippis, G. and A. Figalli (2015). “Partial regularity for optimal transport maps”. *Publications Mathématiques de l’IHÉS* 121.1, pp. 81–112.

del Barrio, E., A. González-Sanz, and J.-M. Loubes (2021a). “A central limit theorem for semidiscrete Wasserstein distances”. *Preprint arXiv:2105.11721.*

del Barrio, E., A. González-Sanz, and J.-M. Loubes (2021b). “Central limit theorems for general transportation costs”. *Preprint arXiv:2102.06379.*

del Barrio, E. and J.-M. Loubes (2019). “Central limit theorems for empirical transportation cost in general dimension”. *The Annals of Probability* 47.2, pp. 926–951.

Fathi, A. and A. Figalli (2010). “Optimal transportation on non-compact manifolds”. *Israel Journal of Mathematics* 175.1, pp. 1–59.

Federer, H. (2014). *Geometric Measure Theory.* Springer.

Figalli, A. (2007). “Existence, uniqueness, and regularity of optimal transport maps”. *SIAM Journal on Mathematical Analysis* 39.1, pp. 126–137.

Figalli, A. and N. Gigli (2011). “Local semiconvexity of Kantorovich potentials on non-compact manifolds”. *ESAIM: Control, Optimisation and Calculus of Variations* 17.3, pp. 648–653.

Figalli, A. and F. Glaudo (2021). *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows.* EMS Press.

Galichon, A. (2016). *Optimal Transport Methods in Economics.* Princeton University Press.

Gangbo, W. and R. J. McCann (1996). “The geometry of optimal transportation”. *Acta Mathematica* 177.2, pp. 113–161.

Gigli, N. et al. (2012). “Optimal maps in non branching spaces with Ricci curvature bounded from below”. *Geometric and Functional Analysis* 22.4, pp. 990–999.

Hung, M. S., W. O. Rom, and A. D. Waren (1986). “Degeneracy in transportation problems”. *Discrete Applied Mathematics* 13.2-3, pp. 223–237.

Kallenberg, O. (1997). *Foundations of Modern Probability.* Vol. 2. Springer.

Kanamori, A. (1995). “The emergence of descriptive set theory”. In: *From Dedekind to Gödel.* Springer, pp. 241–262.

Kantorovich, L. V. (1942). “On the translocation of masses”. *Doklady Akademii Nauk URSS* 37, pp. 7–8.

Kechris, A. (2012). *Classical Descriptive Set Theory.* Springer Science & Business Media.

Klee, V. and C. Witzgall (1968). “Facets and vertices of transportation polytopes”. *Mathematics of the Decision Sciences* 1, pp. 257–282.

Lebesgue, H. (1905). "Sur les fonctions représentables analytiquement". *Journal de Mathematiques Pures et Appliquees* 1, pp. 139–216.

Loeper, G. (2009). "On the regularity of solutions of optimal transportation problems". *Acta Mathematica* 202.2, pp. 241–283.

Ma, X.-N., N. S. Trudinger, and X.-J. Wang (2005). "Regularity of potential functions of the optimal transportation problem". *Archive for Rational Mechanics and Analysis* 177.2, pp. 151–183.

McCann, R. J. (2001). "Polar factorization of maps on Riemannian manifolds". *Geometric and Functional Analysis* 11.3, pp. 589–608.

McCann, R. J. and L. Rifford (2016). "The intrinsic dynamics of optimal transport". *Journal de l'École polytechnique—Mathématiques* 3, pp. 67–98.

Moameni, A. and L. Rifford (2020). "Uniquely minimizing costs for the Kantorovitch problem". *Annales de la Faculté des Sciences de Toulouse: Mathématiques* 29.3, pp. 507–563.

Monge, G. (1781). "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences de Paris*, pp. 666–704.

Nutz, M. and J. Wiesel (2021). "Entropic optimal transport: Convergence of potentials". *Preprint arXiv:2104.11720.*

Panaretos, V. M. and Y. Zemel (2020). *An Invitation to Statistics in Wasserstein Space.* Springer Nature.

Peyré, G. and M. Cuturi (2019). "Computational optimal transport: With applications to data science". *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607.

Qi, L. (1989). "The maximal normal operator space and integration of subdifferentials of nonconvex functions". *Nonlinear Analysis: Theory, Methods & Applications* 13.9, pp. 1003–1011.

Rachev, S. and L. Rüschendorf (1998). *Mass Transportation Problems: Volume I: Theory.* Springer.

Rockafellar, R. T. (2015). *Convex Analysis.* Princeton University Press.

Rüschendorf, L. (2007). "Monge-Kantorovich transportation problem and optimal couplings". *Jahresbericht der Deutschen Mathematiker Vereinigung* 109.3, pp. 113–137.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.* Springer.

Schiebinger, G., J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander (2019). "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming". *Cell* 176.4, 928–943.e22.

Smith, C. S. and M. Knott (1987). "Note on the optimal transportation of distributions". *Journal of Optimization Theory and Applications* 52.2, pp. 323–329.

Sommerfeld, M. and A. Munk (2018). "Inference for empirical Wasserstein distances on finite spaces". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1, pp. 219–238.

Tameling, C., M. Sommerfeld, and A. Munk (2019). “Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications”. *The Annals of Applied Probability* 29.5, pp. 2744–2781.

Tameling, C., S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk (2021). “Colocalization for super-resolution microscopy via optimal transport”. *Nature Computational Science* 1.3, pp. 199–211.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer.

A Auxiliary results and omitted proofs

In a brief assertion of his 1905 memoir, Henry Lebesgue famously sketched an erroneous proof claiming that the projection of Borel subsets of the plane $\mathbb{R}^2$ onto one of the coordinate axes is again Borel. The invalidity of this claim was uncovered in 1916 by Mikhail Suslin, which inspired the study of what is now called *analytic sets* or *Suslin sets* (Kanamori 1995). Placed into the context of our work, we learn that it can happen that projected sets of the form $p_X(A)$ and $p_Y(A)$ for Borel sets $A \subset X \times Y$ are not Borel again. However, these sets are analytic and thus *universally measurable*. In particular, they can be approximated from within and without by Borel sets whose difference is a null set. This settles potential measurability issues in a satisfactory manner.

> Lemma 9: Let X and Y be Polish and $\pi \in \mathcal{C}(\mu, \nu)$ be a transport plan between $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Suppose $A \subset X \times Y$ is Borel with $\pi(A) = 1$. Then there exist Borel sets
>
> $$A_\mu \subset p_X(A) \qquad \text{and} \qquad A_\nu \subset p_Y(A)$$
>
> such that $\mu(A_\mu) = \nu(A_\nu) = 1$. In particular, A_μ and A_ν are dense in $\operatorname{supp}\mu$ and $\operatorname{supp}\nu$.

Proof of Lemma 9. We only show the statement for μ since the one for ν follows equivalently. According to Kechris 2012, Exercise 14.3, the set $p_X(A) \subset X$ is *analytic*, i.e., the continuous image of a Polish space. By Kechris 2012, Theorem 21.10, every analytic set is *universally measurable* (see Definition 12.5 for the term *standard Borel space*), which implies that $p_X(A) = A_\mu \cup N$ where $A_\mu \subset X$ is Borel and N is a subset of a Borel μ-null set $M \subset X$ (see Kechris 2012, Section 17.A, for respective definitions). It is left to show that $\mu(A_\mu) = \mu(A_\mu \cup M) = 1$, which follows from observing that $A \subset p_X(A) \times Y \subset (A_\mu \cup M) \times Y$ and thus

$$\mu(A_\mu \cup M) = \pi\big((A_\mu \cup M) \times Y\big) \geq \pi(A) = 1. \qquad \square$$

Most of the time, we employ Lemma 9 with the choice $A = \operatorname{supp}\pi$ for an optimal transport plan π, making sure that properties on the sets $p_X(\operatorname{supp}\pi)$ and $p_Y(\operatorname{supp}\pi)$ are valid μ- and ν-almost surely. We next highlight a simple consequence of Lemma 2, relating the points $x \in \operatorname{supp}\mu$ with $x \notin p_X(\operatorname{supp}\pi)$ to the ones where π does not induce regularity.

> Lemma 10: Let X and Y be Polish, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \to \mathbb{R}_+$ continuous such that $T_c(\mu, \nu) < \infty$. Let $\pi \in \mathcal{C}(\mu, \nu)$ be an optimal transport plan. Then
>
> $$\operatorname{cl}\big(\operatorname{supp}\mu \backslash p_X(\operatorname{supp}\pi)\big) \subset \big\{x \in \operatorname{supp}\mu \,\big|\, \pi \text{ does not induce regularity at } x\big\} =: \Sigma$$
>
> and the set Σ is closed in $\operatorname{supp}\mu$.

Proof. To see that Σ is closed, assume π to induce regularity at $x \in \operatorname{supp} \mu$ with relatively open $U \subset \operatorname{supp}(\mu)$ and compact $K \subset Y$. Then π is also inducing regularity at any other $x' \in U$ with the same U and K, showing that $\operatorname{supp} \mu \setminus \Sigma$ is open. It now suffices to note that $\operatorname{supp} \mu \setminus \Sigma \subset p_X(\operatorname{supp} \pi)$ according to Lemma 2, which, together with closedness of Σ, implies the inclusion. □

We now turn to Lemma 3, which states that Kantorovich potentials behave as expected when restricted to subsets $\tilde{X} \subset X$ and $\tilde{Y} \subset Y$ with full μ- and ν-mass. The proof of this statement follows the reasoning behind Villani 2008, Lemma 5.18 and Theorem 5.19. Cases of particular interest to us are restrictions to the (interior of the) support of μ or ν (if the boundary carries no mass). Then the subsets $\tilde{X}$ and $\tilde{Y}$ are either closed or open, and so they are always Borel and Polish.

Proof of Lemma 3. By assumption, $\tilde{X} \subset X$, $\tilde{Y} \subset Y$, and $\tilde{X} \times \tilde{Y} \subset X \times Y$ are Borel and Polish subsets with $\mu(\tilde{X}) = \nu(\tilde{Y}) = 1$. Since the Borel σ-algebra of a restricted spaces coincides with the respective subspace σ-algebra (see, e.g., Kallenberg 1997, Lemma 1.6), it is easy to recognize that $T_c(\mu, \nu) = T_{\tilde{c}}(\mu, \nu)$ with equal optimal plans $\pi \in S(\mu, \nu)$, where we permissively identify the measures μ, ν, and π with their restrictions to the Borel sets $\tilde{X}$, $\tilde{Y}$, and $\tilde{X} \times \tilde{Y}$ of full mass.

Recall that $\tilde{\Gamma} = \operatorname{supp} \pi \cap (\tilde{X} \times \tilde{Y})$ and observe $\pi(\tilde{\Gamma}) = 1$, which implies that $\tilde{\Gamma}$ is dense in the support of π. We begin with the claim on restrictions. Let $f \in S_c(\mu, \nu)$ and define the $\tilde{c}$-concave function $\tilde{f} \colon \tilde{X} \to \mathbb{R} \cup \{-\infty\}$ via

$$\tilde{f}(x) = \big(f^c|_{\tilde{Y}}\big)^{\tilde{c}}(x) = \inf_{y \in \tilde{Y}} c(x, y) - f^c(y).$$

For any $(x_0, y_0) \in \tilde{\Gamma}$, we calculate

$$\begin{aligned}\tilde{f}(x_0) &= \inf_{y \in \tilde{Y}} c(x_0, y) - f^c(y) \le c(x_0, y_0) - f^c(y_0) = f(x_0),\\ \tilde{f}(x_0) &= \inf_{y \in \tilde{Y}} c(x_0, y) - \inf_{x \in X} c(x, y) + f(x) \ge f(x_0),\end{aligned}$$

which shows that $\tilde{f} = f$ on $p_X(\tilde{\Gamma})$. Similarly,

$$\begin{aligned}\tilde{f}^{\tilde{c}}(y_0) &= \inf_{x \in \tilde{X}} c(x, y_0) - \tilde{f}(x) \le c(x_0, y_0) - f(x_0) = f^c(y_0),\\ \tilde{f}^{\tilde{c}}(y_0) &= \inf_{x \in \tilde{X}} c(x, y_0) - \inf_{y \in \tilde{Y}} c(x, y) + f^c(y) \ge f^c(y_0),\end{aligned}$$

which asserts $\tilde{f}^{\tilde{c}} = f^c$ on $p_Y(\tilde{\Gamma})$. Since the set $\tilde{\Gamma}$ has full π-measure, it follows that $T_{\tilde{c}}(\mu, \nu) = T_c(\mu, \nu) = \pi\,(f \oplus f^c) = \pi\,(\tilde{f} \oplus \tilde{f}^{\tilde{c}})$. This shows $\tilde{f} \in S_{\tilde{c}}(\mu, \nu)$ and thus proves the first statement.

We next turn towards the claim on extending potentials, where we begin with $\tilde{f} \in S_{\tilde{c}}(\mu, \nu)$ and observe $\tilde{\Gamma} \subset \partial_{\tilde{c}} \tilde{f}$ (which is true since $\tilde{\Gamma}$ equals the support of the measure π restricted to $\tilde{X} \times \tilde{Y}$). We extend $\tilde{f}$ to a function $f : X \to \mathbb{R}$ on all of X via defining

$$f(x) = \inf_{y \in \tilde{Y}} c(x, y) - \tilde{f}^{\tilde{c}}(y).$$

This function is c-concave, since it is the c-transform of a function that equals $\tilde{f}^{\tilde{c}}$ on $\tilde{Y}$ and $-\infty$ on $Y \setminus \tilde{Y}$. Furthermore, since $\tilde{f}$ is $\tilde{c}$-concave, the definition of f directly shows that it coincides with $\tilde{f}$ ($= \tilde{f}^{\tilde{c}\tilde{c}}$) on $\tilde{X}$. For $(x_0, y_0) \in \tilde{\Gamma}$, we furthermore note that

$$\begin{aligned} f^c(y_0) &= \inf_{x \in X} c(x, y_0) - f(x) \leq c(x_0, y_0) - f(x_0) = \tilde{f}^{\tilde{c}}(y_0), \\ f^c(y_0) &= \inf_{x \in X} c(x, y_0) - \inf_{y \in \tilde{Y}} c(x, y) + \tilde{f}^{\tilde{c}}(y) \geq \tilde{f}^{\tilde{c}}(y_0), \end{aligned}$$

and thus $f^c = \tilde{f}^{\tilde{c}}$ on $p_Y(\tilde{\Gamma})$. The optimality of f is checked like above, yielding $f \in S_c(\mu, \nu)$. □

Note that the sets of consensus $p_X(\tilde{\Gamma})$ and $p_Y(\tilde{\Gamma})$ in Lemma 3 can be enlarged to the (potentially slightly bigger) sets $p_X(\operatorname{supp} \pi) \cap \tilde{X}$ and $p_Y(\operatorname{supp} \pi) \cap \tilde{Y}$ when restricting Kantorovich potentials. To prove this, the density of $\tilde{\Gamma}$ in $\operatorname{supp} \pi$ can be exploited in combination with Lemma 1. For extending a potential $\tilde{f}$, the proof above shows that it is always possible to pick an extension that agrees with $\tilde{f}$ on all of $\tilde{X}$, or alternatively to pick an extension whose c-transform agrees with $\tilde{f}^{\tilde{c}}$ on all of $\tilde{Y}$.

We next prove the claim stated in the context of equation (11), which provides an example where degeneracy of the optimal transport plan leads to non-unique Kantorovich potentials.

Lemma 11: Let $X = Y$ be Polish, $\mu = \nu \in \mathcal{P}(X)$ with $\operatorname{supp}(\mu) = X_1 \cup X_2$, and $c : X^2 \to \mathbb{R}_+$ be continuous and symmetric with $c(x, x) = 0$ for all $x \in X$. If

$$\Delta = \inf_{x_1 \in X_1, x_2 \in X_2} c(x_1, x_2) > 0,$$

then for all $a, b \in \mathbb{R}$ with $|a - b| \leq \Delta$, there exists $f_{a,b} \in S_c(\mu, \mu)$ such that $f_{a,b} = a$ on X_1 and $f_{a,b} = b$ on X_2.

Proof. It is apparent that $T_c(\mu, \mu) = 0$. For real numbers $a, b \in \mathbb{R}$ with $|a - b| \leq \Delta$, we define the map $g : X \to \mathbb{R} \cup \{-\infty\}$ via

$$g(x) = \begin{cases} -a & \text{if } x \in X_1, \\ -b & \text{if } x \in X_2, \\ -\infty & \text{if } x \notin X_1 \cup X_2. \end{cases}$$

For $x \in X_1$, the c-concave function $f_{a,b} \coloneqq g^c$ fulfills

$$f_{a,b}(x) = \inf_{y \in X} c(x,y) - g(y) = \inf_{y \in X} \begin{cases} a & \text{for } y = x \in X_1, \\ c(x,y) + a & \text{for } y \in X_1, \\ c(x,y) + b & \text{for } y \in X_2, \\ \infty & \text{for } y \notin X_1 \cup X_2. \end{cases}$$

Since $c \geq 0$ and $a \leq \Delta + b \leq c(x,y) + b$ for $x \in X_1, y \in X_2$, we find $f_{a,b} = a$ on X_1. Likewise, we also find $f_{a,b} = b$ on X_2. By a similar argument, it follows that $f^c_{a,b} \leq -a$ on X_1 and $f^c_{a,b} \leq -b$ on X_2. Due to the lower bound $g \leq g^{cc} = f^c_{a,b}$ (see Villani 2008, Proposition 5.8), we conclude that $f^c = -f$ on $X_1 \cup X_2$. This implies $T_c(\mu,\mu) = 0 = \pi\left(f_{a,b} \oplus f^c_{a,b}\right)$ for any optimal π, asserting that $f_{a,b}$ is indeed a Kantorovich potential. □

Next, we address the claim raised in Lemma 7, which states that (locally) Lipschitz continuous functions on connected manifolds coincide (up to constants) if their gradients coincide almost surely (in charts). This is a simple generalization of corresponding results on $\mathbb{R}^d$, which can, for example, be gathered from considerations in Qi 1989.

Proof of Lemma 7. Let $f_1, f_2 \colon M \to \mathbb{R}$ be locally Lipschitz on a d-dimensional smooth manifold M, and let $(\varphi_x)_{x \in M}$ be a family of charts with $x \in U_x = \operatorname{domain} \varphi_x$ for all $x \in M$. By translating, restricting, and rescaling the charts, we can assume that $\operatorname{range} \varphi_x = B_1 \subset \mathbb{R}^d$ is the unit ball and that $f_{i,x} = f_i \circ \varphi_x^{-1} \colon B_1 \to \mathbb{R}$ is Lipschitz for each $x \in M$ and $i \in \{1,2\}$. Since $\nabla f_{1,x} = \nabla f_{2,x}$ holds by assumption on a set with full Lebesgue measure, we can conclude (e.g., by formula (2) of Qi 1989) that $f_{1,x} - f_{2,x} = c_x$ on all of B_1, where $c_x \in \mathbb{R}$ is a constant. This implies $f_1|_{U_x} - f_2|_{U_x} = c_x$ as well. To see that c_x is actually independent of x, note that $c_{x'} = c_x$ holds for any $x' \in U_x$, since then $U_x \cap U_{x'} \neq \emptyset$. Thus, $x \mapsto c_x$ is a locally constant function. The connectedness of M implies that it is also constant on the whole space. To see this, just note that the set $V = \{x \in X \mid c_x = c_{x_0}\} \neq \emptyset$ and its complement are both open (for some fixed $x_0 \in X$), which makes V open and closed. Hence, $V = M$ and the claim $f_1 - f_2 = c_{x_0}$ is established. □

To conclude this appendix, we show that Corollary 3 is also true if $c(\cdot, y)$ is assumed to be locally semiconcave uniformly in $y \in Y$ (instead of locally Lipschitz uniformly in y), which is claimed in Remark 4.

Proof of Remark 4. We adhere to the general proof strategy of Theorem 2, but will provide additional details for some of the arguments. For every $x \in X$, let $\varphi_x \colon U_x \to V_x \subset \mathbb{R}^d$ denote a chart around x. By restricting and translating, we can assume that V_x is convex with $\varphi(x) = 0 \in V_x$. Recalling equation (2), we may also assume that $v \mapsto c\big(\varphi_x^{-1}(v), y\big) - \lambda_x \|v\|^2$ is concave on V_x for some $\lambda_x > 0$ and all $y \in Y$. Due to the nature of c-conjugates as infima, we find that the function $v \mapsto h_x(v) = f\big(\varphi_x^{-1}(v)\big) - \lambda_x\|v\|^2$ is concave as well for any $f \in S_c(\mu, \nu)$ (Rockafellar 2015, Theorem 5.5).

We first show that $f > -\infty$ on $M = \mathrm{int}(\mathrm{supp}\,\mu)$. Indeed, assume that there existed a point $x \in M$ with $f(x) = -\infty$. Then $h_x(0) = -\infty$, and, since the effective domain $h_x^{-1}(\mathbb{R}) \subset V_x$ is convex, it followed that $h_x^{-1}\big(\{-\infty\}\big)$ contained at least an open half-space (intersected with V_x) touching the origin. Hence, $f = -\infty$ would have to hold on an open subset of $M \subset \mathrm{supp}\,\mu$, which cannot be true, since $p_X(\mathrm{supp}\,\pi)$ is dense in $\mathrm{supp}\,\mu$ and $f > -\infty$ on $p_X(\mathrm{supp}\,\pi)$ due to $\mathrm{supp}\,\pi \subset \partial_c f$. Since (locally semi-)concave functions are locally Lipschitz in the interior of their effective domain (Rockafellar 2015, Theorem 10.4), we can conclude that each Kantorovich potential is locally Lipschitz on all of M.

Next, since $\Gamma = \mathrm{supp}\,\pi$, we find a Borel set $A \subset p_X(\Gamma) \subset X$ that satisfies $\mu(A) = 1$ (Lemma 9). Hence, $B = \mathrm{range}\,\varphi \setminus \varphi(A \cap \mathrm{domain}\,\varphi)$ is a $\varphi_\#\mu$-null set for any chart φ of M. Due to condition (15), it is also a Lebesgue-null set and we conclude that $p_X(\Gamma)$ has full Lebesgue measure in charts of M. We can now argue as in Theorem 2 to find that any two Kantorovich potentials $f_1, f_2 \in S_c(\mu, \nu)$ coincide on M. It remains to show that $f_1 = f_2$ holds on the boundary of $\mathrm{supp}\,\mu$ as well. Let $x \in \partial\,\mathrm{supp}\,\mu$ and let $(x_n)_{n\in\mathbb{N}} \subset M$ be a sequence converging to x. Then it holds that $\lim_{n\to\infty} f_i(x_n) = f_i(x)$ for $i \in \{1, 2\}$. This can be seen as follows: if $f_i(x) > -\infty$, then the limit holds since (locally semi-)concave functions are continuous on their effective domain (Rockafellar 2015, Theorem 10.1), and if $f_i(x) = -\infty$, then the limit holds due to upper-semicontinuity of f_i. Since $f_1(x_n) = f_2(x_n)$ for all n, this establishes equality of f_1 and f_2 on the full support. □

D Transport Dependency: Optimal Transport Based Dependency Measures

Thomas Giacomo Nies [*,†,‡]
thomas.nies@uni-goettingen.de

Thomas Staudt [*,†,‡]
thomas.staudt@uni-goettingen.de

Axel Munk[†,‡,§]
munk@math.uni-goettingen.de

Abstract

Finding meaningful ways to determine the dependency between two random variables ξ and ζ is a timeless statistical endeavor with vast practical relevance. In recent years, several concepts that aim to extend classical means (such as the Pearson correlation or rank-based coefficients like Spearman's ρ) to more general spaces have been introduced and popularized, a well-known example being the distance correlation. In this article, we propose and study an alternative framework for measuring statistical dependency, the *transport dependency* $\tau \geq 0$, which relies on the notion of optimal transport and is applicable in general Polish spaces. It can be estimated consistently via the corresponding empirical measure, is versatile and adaptable to various scenarios by proper choices of the cost function, and intrinsically respects metric properties of the ground spaces. Notably, statistical independence is characterized by $\tau = 0$, while large values of τ indicate highly regular relations between ξ and ζ. Indeed, for suitable base costs, τ is maximized if and only if ζ can be expressed as 1-Lipschitz function of ξ or vice versa. Based on sharp upper bounds, we exploit this characterization and define three distinct dependency coefficients

[*] These authors contributed equally
[†] Institute for Mathematical Stochastics, University of Göttingen
[‡] Cluster of Excellence: Multiscale Bioimaging (MBExC), University Medical Center, Göttingen
[§] Max Planck Institute for Biophysical Chemistry, Göttingen

(a-c) with values in $[0, 1]$, each of which emphasizes different functional relations. These *transport correlations* attain the value 1 if and only if $\zeta = \varphi(\xi)$, where φ is a) a Lipschitz function, b) a measurable function, c) a multiple of an isometry. The properties of coefficient c) make it comparable to the distance correlation, while coefficient b) is a limit case of a) that was recently studied independently by Wiesel 2021. Numerical results suggest that the transport dependency is a robust quantity that efficiently discerns structure from noise in simple settings, often out-performing other commonly applied coefficients of dependency.

Keywords: transport dependency, transport correlation, optimal transport, statistical dependence, mutual information, correlation, distance correlation

MSC 2020 Subject Classification: primary 62H20, 49Q22; secondary 62R20, 62G35, 60E15

1 Introduction

In this article, we explore a method to quantify the statistical dependence between two random variables ξ and ζ on Polish spaces X and Y via optimal transport. The core idea is to calculate the effort necessary to transform the joint distribution γ of ξ and ζ into the product $\mu \otimes \nu$ of their marginal distributions μ and ν. This motivates the definition of the *transport dependency*

$$\tau(\xi, \zeta) = \tau(\gamma) = T_c\big(\gamma, \mu \otimes \nu\big) = \inf_{\pi} \int c \, \mathrm{d}\pi, \tag{1}$$

where T_c is the optimal transport cost with (non-negative) base costs c on $X \times Y$. The infimum on the right is taken over the set of all couplings between γ and $\mu \otimes \nu$, i.e., distributions π with marginals γ and $\mu \otimes \nu$ (see Villani 2008 for a comprehensive treatment). If the cost function has benign properties, the transport dependency τ displays many traits that are attractive for a measure of statistical association. For example, if c is a metric (or a power thereof), then $\tau(\gamma) = 0$ if and only if $\gamma = \mu \otimes \nu$, which is the case of statistical independence of ξ and ζ. Figure 1a illustrates the intuition behind this concept.

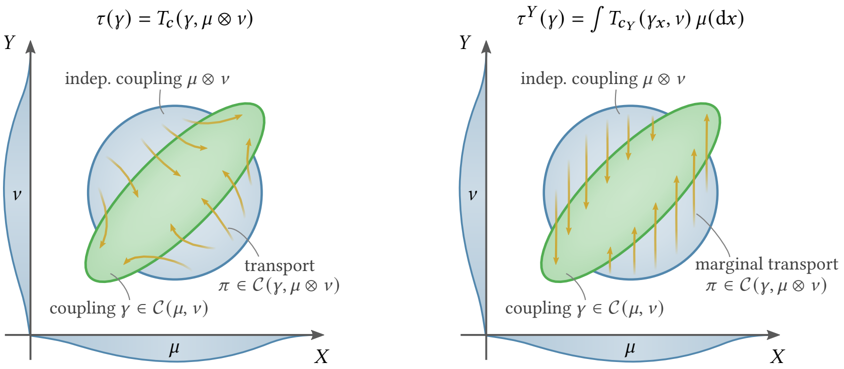
In a Euclidean setting, the general idea of applying an optimal transport distance between a coupling γ and the product $\mu \otimes \nu$ of its marginals has recently gained traction in both the statistics and the machine learning literature. For example, it was proposed under the names *Wasserstein dependence measure* and *Wasserstein total correlation* by Ozair et al. 2019 and Xiao and Wang 2019, who applied it to improve results during representation learning. The same ansatz also underlies a very recent article by Mordant and Segers 2022, who, parallel to our work, defined normalized coefficients of association (*Wasserstein dependency coefficients*) that are based on (1) under squared Euclidean costs. Since the authors divide $\tau(\gamma)$ by the supremum of $\tau(\tilde{\gamma})$ over all $\tilde{\gamma}$ with fixed marginals μ and ν, however, their coefficients are difficult to calculate and rely on a quasi-Gaussian approximation. On general Polish metric spaces, Móri and Székely 2020 introduced the *Earth mover's correlation*, which is also based on (1). While it fails to satisfy some desirable properties proposed in Móri and Székely 2019, it is grounded on an easily computable upper bound of (1) in terms of the diameters of the marginals μ and ν. As we will see, this bound is sharp for suitable cost structures on $X \times Y$ (which do *not* include the Euclidean squared distance). Indeed, the properties of couplings γ that attain this bound are a premier object of interest for us.

Independently from our work, another close relative of the transport dependency τ has recently been explored by Wiesel 2021, who proposes the association measure

$$\tau^Y(\xi, \zeta) = \tau^Y(\gamma) = \int T_{c_Y}(\gamma_x, \nu)\, \mu(\mathrm{d}x), \tag{2}$$

where $(\gamma_x)_{x \in X}$ denotes the disintegration of γ with respect to the first coordinate (i.e., γ_x is the law of ζ given $\xi = x$), and c_Y is (the power of) a metric on the space Y. For normalization to a coefficient in $[0, 1]$, Wiesel 2021 also uses the diameter of ν, which does not only upper bound $\tau(\gamma)$, but $\tau^Y(\gamma)$ as well. He further shows that this coefficient exhibits a list of desirable properties advanced by the work of Chatterjee 2021 (and earlier proposed by Dette et al. 2013), and suggests a method by which it can be estimated consistently. Even though (2) appears to address a different problem when compared to (1) – after all, the optimal transport in τ^Y is calculated only on Y and the metric structure of X plays no role – it is in fact a special case: if transport along the space X is forbidden by choosing costs c that assume the value ∞ for movements of mass that are not vertical (see Figure 1b), then τ reduces to the expression in (2). We therefore call $\tau^Y(\gamma)$ the *marginal transport dependency* of γ. In the course of our exploration, the marginal transport dependency will naturally emerge both as an upper bound and as a limiting case. We also want to mention that integrals of the form (2) have previously appeared in the context of generalization bounds for statistical learning problems (Zhang et al. 2018; Lopez and Jog 2018; Wang et al. 2019), and that a special case of τ^Y has been employed by Xiao and Wang 2019 in order to improve estimates of the dependency gap in representation learning tasks.

Mutual information. The idea of quantifying dependency as a measure of discrepancy between the joint distribution of ξ and ζ and the product distribution of their marginals reaches far back. In his landmark work, Shannon 1948 introduced the mutual information $M(\gamma) = D(\gamma \,|\, \mu \otimes \nu)$, where the Kullback-Leibler divergence D is used to compare γ to $\mu \otimes \nu$. The mutual information has become an indispensable tool for measuring the information content stored in the relation between random variables and has found application in feature selection (Estévez et al. 2009), image registration and alignment (Maes et al. 1997; Pluim et al. 2000; Viola and Wells III 1997), clustering (Kraskov et al. 2005), and independence testing (Berrett and Samworth 2019), besides others. Its immediate use for statistical data analysis, however, is complicated by several issues. For example, it is often inconvenient


$\tau(\gamma) = T_c(\gamma, \mu \otimes \nu)$
Y
indep. coupling $\mu \otimes \nu$
ν
transport
$\pi \in \mathcal{C}(\gamma, \mu \otimes \nu)$
coupling $\gamma \in \mathcal{C}(\mu, \nu)$
μ
X
$\tau^Y(\gamma) = \int T_{c_Y}(\gamma_x, \nu)\, \mu(\mathrm{d}x)$
Y
indep. coupling $\mu \otimes \nu$
ν
marginal transport
$\pi \in \mathcal{C}(\gamma, \mu \otimes \nu)$
coupling $\gamma \in \mathcal{C}(\mu, \nu)$
μ
X

(a) **(b)**

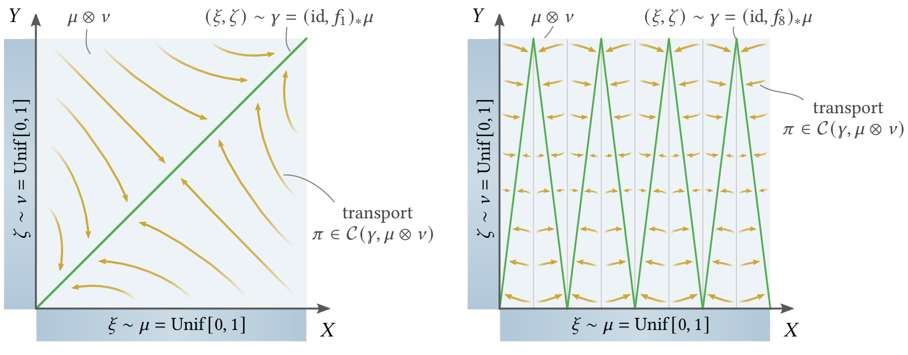
Figure 1: Illustration of the (marginal) transport dependency. Sketch **(a)** shows how mass could be moved in an optimal way when transforming $\mu \otimes \nu$ into γ (visualized on level sets). The closer γ is to $\mu \otimes \nu$, the less mass has to be transported and the smaller the value of $\tau(\gamma)$ will be. Sketch **(b)** displays the same situation as (a), but this time transport along the space X is forbidden and mass has to be transported vertically, which is clearly less optimal.

to estimate $M(\gamma)$ from data, as density estimates or binning / clustering methods are necessary and estimation suffers from the curse of dimensionality (Hall and Morton 1993; Paninski and Yajima 2008; Berrett et al. 2019). Furthermore, the mutual information itself does not respect topological or metric properties of the coupling γ, as rearrangements of ξ and ζ by measurable functions leave $M(\gamma)$ invariant. In this sense, the mutual information is not able to distinguish chaotic relations between ξ and ζ from well-behaved ones (see Figure 2). Still, the mutual information and its surrogates, such as the mutual information dimension (Sugiyama and Borgwardt 2013) or the maximal information coefficient (Reshef et al. 2011), have proven to be useful tools for detecting statistical dependency in data analysis.

Distance covariance. Another popular approach to measure dependency by comparing γ to $\mu \otimes \nu$ is the *distance covariance* proposed by Székely et al. 2007. If ξ and ζ are random vectors in $\mathbb{R}^r$ and $\mathbb{R}^q$ for $r, q \in \mathbb{N}$, the (Euclidean) distance covariance between ξ and ζ is a weighted L_2 distance between the joint characteristic function f_γ and the product of the individual characteristic functions f_μ and f_ν,

$$\operatorname{dcov}^2(\xi, \zeta) = \frac{1}{c_r c_q} \int \frac{\left|f_\gamma(t, s) - f_\mu(t) f_\nu(s)\right|^2}{\|t\|^{1+r} \|s\|^{1+q}} \, \lambda^r(\mathrm{d}t)\, \lambda^q(\mathrm{d}s), \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm, c_d is a constant only depending on the dimensions d, and λ^d denotes the Lebesgue measure in $\mathbb{R}^d$. Lyons 2013 later proposed a generalization of (3) to metric spaces (X, d_X) and (Y, d_Y). Subsequent work by Jakobsen 2017 established that these spaces have to be separable in order to provide a sound



(a) (b)

Figure 2: Marginal and joint distribution of random variables $\xi \sim \mathrm{Unif}\,[0, 1]$ and $\zeta = f_n(\xi) \sim \mathrm{Unif}\,[0, 1]$ for zigzag functions f_n with n linear segments for $n = 1$ in **(a)** and $n = 8$ in **(b)**. The drawn arrows illustrate how the optimal transport between $\mu \otimes \nu$ and γ could look like. Note that the mutual information and related concepts that only measure the information content do not distinguish between the two scenarios. In contrast, dependency measures that are aware of metric or topological properties, like the transport dependency τ or the distance covariance, assign a (much) lower degree of dependence to scenario (b). This discrepancy stresses an important point: should deterministic but chaotic relations between ξ and ζ maximize a measure for statistical dependency? After all, one may not be able to recover the relation in practice and distinguish it from noise if data is limited.

theory, see also Lyons 2018. The generalized distance covariance takes the form

$$\begin{aligned}\mathrm{dcov}^2(\xi, \zeta) = \mathbb{E}\big[d_X(\xi, \xi')\, d_Y(\zeta, \zeta')\big] + \mathbb{E}\big[d_X(\xi, \xi')\big]\, \mathbb{E}\big[d_Y(\zeta, \zeta')\big] \\ - 2\,\mathbb{E}\big[d_X(\xi, \xi')\, d_Y(\zeta, \zeta'')\big],\end{aligned} \tag{4}$$

where $(\xi', \zeta') \sim \gamma$ and $\zeta'' \sim \nu$ are independent copies of ξ and ζ. If the metric spaces X and Y are of strong negative type[5], one can show that $\mathrm{dcov}^2(\xi, \zeta)$ as defined above is indeed non-negative, and that it vanishes if and only if ξ and ζ are statistically independent (Lyons 2013; Jakobsen 2017). Together with fast computability on data and a well-understood limit theory, this renders the distance covariance a compelling instrument for non-parametric (i.e., without assuming a specific dependency model) independence testing (Yao et al. 2018; Castro-Prado and González-Manteiga 2020; Chakraborty and Zhang 2021) and related problems, like independent component analysis (Matteson and Tsay 2017). Moreover, the work of Sejdinovic et al. 2013 has established the equivalence of the distance covariance and the *Hilbert-Schmidt independence criterion* (Gretton et al. 2005). Corollary 26 in (Sejdinovic et al. 2013) additionally clarifies that the general distance covariance (4) can be understood as

[5] A separable metric space (X, d_X) is of *negative type* iff there is an isometric embedding φ of $(X, d_X^{1/2})$ into a separable Hilbert space. This condition asserts $\mathrm{dcov}^2 \geq 0$. If the mean embedding $\mu \mapsto \int \varphi \,\mathrm{d}\mu$ for probability measures $\mu \in \mathcal{P}(X)$ with finite first d_X-moment is additionally injective, the space (X, d_X) is of *strong negative type* and $\mathrm{dcov} = 0$ characterizes independence. Examples for spaces of negative type are L_p spaces for $1 \leq p \leq 2$, ultrametric spaces, and weighted trees (Meckes 2013, Theorem 3.6). Known counter examples are $\mathbb{R}^d$ with l_p norms for $p > 2$ (see the references in Lyons 2013).

the *maximum mean discrepancy* (MMD, see Gretton et al. 2012) between γ and $\mu \otimes \nu$ under a suitable kernel choice.

The distance covariance possesses a natural upper bound that only depends on the marginal measures μ and ν, namely $\mathrm{dcov}^2(\xi, \zeta) \leq \mathrm{dcov}(\xi, \xi) \cdot \mathrm{dcov}(\zeta, \zeta)$. Through normalization with this bound, a coefficient that assumes values in $[0, 1]$ can be defined. This coefficient, called the *distance correlation* dcor, can be used as a surrogate for classical coefficients of dependency, like the Pearson correlation, in the general setting of separable metric spaces (X, d_X) and (Y, d_Y) of strong negative type. It has the following two distinguishing characteristics (Lyons 2013):

- $\mathrm{dcor}(\xi, \zeta) = 0$ iff ξ and ζ are independent,
- $\mathrm{dcor}(\xi, \zeta) = 1$ iff there is a $\beta > 0$ and an isometry $\varphi \colon (X, \beta d_X) \to (Y, d_Y)$ with $\zeta = \varphi(\xi)$,

where φ only has to be defined μ-almost surely. This lends the distance covariance a neat and tangible interpretation: while the Pearson correlation measures the degree of *linear* functional dependency between random variables, the distance correlation measures a degree of *isometric* functional dependency (up to scalings). For the Euclidean case, this means that a value of $\mathrm{dcor}(\xi, \zeta) = 1$ is assumed if and only if $\zeta = \beta\,(A\xi + a)$, where A is an orthogonal matrix, a a shift vector, and $\beta > 0$ (at least if the support of μ contains an open set). Other relations between ξ and ζ, even if they are deterministic, result in smaller values $\mathrm{dcor}(\xi, \zeta) < 1$. Indeed, the more chaotic the relation becomes, the further away one is from an isometric dependency, and the lower the value of the distance correlation will typically be (see Figure 2). This draws a sharp distinction to other (non-parametric) concepts of dependency, like the mutual information or several recently proposed coefficients of association (Dette et al. 2013; Chatterjee 2021; Deb et al. 2020; Wiesel 2021), which assume maximal values for *any measurable* deterministic relation – and not only for suitably structured ones.

Other measures of association. On a more general note, when one looks past a number of classical concepts of correlation that are universally applied (like the Pearson correlation, Spearman's ρ, or Kendall's τ), the statistical literature on how to best measure dependency in applications quickly becomes immensely broad and scattered. Besides the approaches cited above, suggestions include maximal correlation coefficients (Gebelein 1941; Koyak 1987), rank or copula based methods (Schweizer and Wolff 1981; Marti et al. 2017), or various measures acting on the distribution of pairwise distances (Friedman and Rafsky 1983; Heller et al. 2013), to only mention a few (for a more complete survey of established methods, see for example Tjøstheim et al. 2018). Recently, optimal transport maps have been utilized to establish multivariate rank statistics that allow for (asymptotically) distribution-free tests of independence (Ghosal and Sen 2019; Shi et al. 2020; Shi et al. 2021).

A different class of data analysis and exploration techniques to be mentioned in this context are those that quantify how multiple data sets are *spatially* associated. A prominent example is Ripley's K function (Ripley 1976), for which new developments have recently been advanced (Amgad et al. 2015). Regarding functional data analysis, we refer to (Petersen and Müller 2019; Dubey and Müller 2020). For colocalization problems in cell microscopy, we mention Wang et al. 2017, who propose a statistically optimal colocalization metric based on Kendall's τ, and Tameling et al. 2021, who suggest certain surrogates of the optimal transport plan for quantifying colocalization.

Transport dependency. A primary reason for the widely scattered literature on this topic is that the notion of "dependency", and what exactly is meant and intended by it, eludes the reduction to a real number that suits all needs. One important demarcation line in this regard has already been stressed: do we aim to measure dependency in a purely stochastic sense (like with the mutual information), or do we also seek to impose structural conditions, like linearity (Pearson correlation), monotonicity (rank correlations), or metric compatibility (distance correlation)? Indeed, the theme of *shape restrictions* is central for recent efforts to find meaningful quantifiers of dependency (Cao and Bickel 2020; see also Guntuboyina, Sen, et al. 2018 for related work on shape-restricted regression).

In this article, we contribute to this topic by establishing the transport dependency as a principled tool that can flexibly bridge the gap between unstructured and structured dependency quantification. On an abstract level, the transport dependency $\tau(\gamma)$ defined in equation (1) features a series of attractive and intuitive properties. Under mild assumptions on the costs c, it is true that $\tau(\gamma) = 0$ characterizes independence (e.g., if c is a metric, see Theorem 1), and that $\tau(\gamma)$ can consistently be estimated by $\tau(\hat{\gamma}_n)$, where $\hat{\gamma}_n$ is the empirical measure (e.g., if c is continuous and obeys some moment conditions, see Theorem 2). Furthermore, the transport dependency does not explicitly rely on how ξ and ζ are embedded into the respective spaces X and Y, but only on the chosen cost structure on them. In particular, the value $\tau(\gamma)$ does not change under (marginal) transformations of ξ and ζ that leave the costs invariant (Proposition 2). The transport dependency is also a convex functional on the space of probability measures on $X \times Y$ with fixed marginals (Proposition 1). Thus, under convex contamination models commonly studied in robust estimation theory (Huber 1992; Horowitz and Manski 1995), polluting γ with a distribution γ' of smaller transport dependency $\tau(\gamma') < \tau(\gamma)$ and the same marginals will inevitably decrease the measured dependence,

$$\tau\big((1-t)\,\gamma + t\,\gamma'\big) < \tau(\gamma)$$

for all $t \in (0, 1]$. Similarly, if c is a translation invariant cost on vector spaces, then convolutions of γ with any (product) kernel κ cannot increase the transport dependency (see Theorem 4), meaning

$$\tau\big(\gamma * \kappa\big) \leq \tau(\gamma).$$

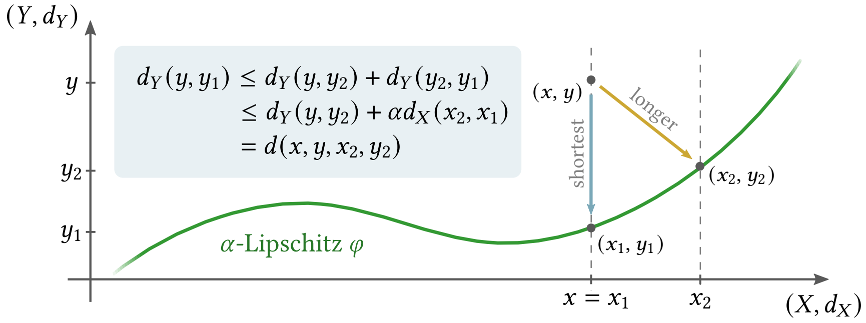

(Y, d_Y)
y
y_2
y_1
$d_Y(y, y_1) \leq d_Y(y, y_2) + d_Y(y_2, y_1)$
$\leq d_Y(y, y_2) + \alpha d_X(x_2, x_1)$
$= d(x, y, x_2, y_2)$
(x, y)
longer
shortest
(x_2, y_2)
(x_1, y_1)
α-Lipschitz φ
$x = x_1$
x_2
(X, d_X)

Figure 3: Projection onto the graph of an α-Lipschitz function $\varphi\colon X \longrightarrow Y$ under costs $c = d^p$ as in (6). By applying the triangle inequality of d_Y and the Lipschitz property $d_Y\big(\varphi(x_1), \varphi(x_2)\big) \leq \alpha d_X(x_1, x_2)$, one can see that the cost-optimal way to move any point $(x, y) \in X \times Y$ to the graph of φ is to simply shift it up- or downwards.

Since the convolution of measures γ and κ is equivalent to taking the sum of independent random variables $(\xi, \zeta) \sim \gamma$ and $(\xi', \zeta') \sim \kappa$, this implies that additive contaminations with independent noise never increase the value of τ. It is also possible to upper bound τ in instructive ways. If c_Y is a marginal cost on Y that bounds c suitably, we derive the inequality

$$\tau(\gamma) \leq \int c_Y \,\mathrm{d}(\nu \otimes \nu) = \mathbb{E}[c_Y(\zeta, \zeta')], \tag{5}$$

where $\zeta, \zeta' \sim \nu$ are independent (Proposition 6). This inequality, which already appeared Móri and Székely 2020, bounds τ in terms of the diameter of ν and is central for deriving dependency coefficients living in $[0, 1]$.

A particularly interesting theory unfolds when the structure of the costs c on the space $X \times Y$ is restricted to a specific form. If (X, d_X) and (Y, d_Y) are Polish metric spaces, we consider additive costs

$$c(x_1, y_1, x_2, y_2) = \big(\alpha \cdot d_X(x_1, x_2) + d_Y(y_1, y_2)\big)^p \tag{6}$$

for $\alpha, p > 0$, where $x_1, x_2 \in X$ and $y_1, y_2 \in Y$. The primary reason why costs of this form are fertile for our purposes is visualized in Figure 3: the price of moving any point (x, y) in $X \times Y$ to the graph of an α-Lipschitz function $\varphi\colon X \longrightarrow Y$, which satisfies

$$d_Y\big(\varphi(x_1), \varphi(x_2)\big) \leq \alpha \cdot d_X(x_1, x_2)$$

for all $x_1, x_2 \in X$, is minimized by vertical movements, i.e., by the projection $(x, y) \mapsto \big(x, \varphi(x)\big)$. Under suitable conditions, this (locally optimal) choice of movement corresponds to a valid transport plan that is globally optimal. As a consequence, we find that our upper bounds are sharp if and only if the support of γ coincides with the graph of an α-Lipschitz function. In other words, $\tau(\gamma)$ assumes upper bound (5) if and only if $\zeta = \varphi(\xi)$ for an α-Lipschitz map φ (Theorem 7 and Corollary 1).

Transport correlation. We use the special properties of the transport dependency under costs of the form (6) to propose a family $\rho_\alpha \in [0, 1]$ of coefficients for $\alpha > 0$ that are tuned to detect α-Lipschitz associations between ξ and ζ. The *α-transport correlation* is defined by

$$\rho_\alpha(\xi, \zeta) = \rho_\alpha(\gamma) = \left(\frac{\tau(\gamma)}{\int d_Y^p \, \mathrm{d}(\nu \otimes \nu)}\right)^{1/p},$$

where the scaling by the inverse of bound (5), which we assume to be positive, guarantees that $0 \leq \rho_\alpha \leq 1$. Two of the hallmark features of ρ_α are (Proposition 9)

- $\rho_\alpha(\gamma) = 0$ iff ξ and ζ are independent,
- $\rho_\alpha(\gamma) = 1$ iff there is an α-Lipschitz function $\varphi\colon (X, d_X) \to (Y, d_Y)$ with $\zeta = \varphi(\xi)$,

where φ only needs to be defined μ-almost surely. Therefore, we may view ρ_α as a general alternative to Pearsons's correlation coefficient that measures the degree of association not in terms of best approximation by linear functions, but by α-Lipschitz functions. Later, we will see that the idea behind ρ_α can fluently be extended to the limit $\alpha \to \infty$ (Theorem 6). In this case, the transport dependency τ is equal to the marginal transport dependency defined in (2), while upper bound (5) stays valid. One can then show that the resulting coefficient ρ_∞, which we name *marginal transport correlation*, satisfies (Proposition 10)

- $\rho_\infty(\gamma) = 0$ iff ξ and ζ are independent,
- $\rho_\infty(\gamma) = 1$ iff there is a measurable function $\varphi\colon X \to Y$ with $\zeta = \varphi(\xi)$,

where the equality $\zeta = \varphi(\xi)$ only has to hold μ-almost surely. In fact, the marginal transport correlation ρ_∞ is equal to the measure of association introduced by Wiesel 2021, who also recognized the above properties as essential.

Both the α-transport correlation and the marginal transport correlation are asymmetric and measure to what extend ζ can be understood as a function of ξ, not the other way around. If symmetry is desired, the coefficients can easily be adapted, for example by taking the maximum of $\rho_\alpha(\xi, \zeta)$ and $\rho_\alpha(\zeta, \xi)$. Another way to obtain a symmetric dependency coefficient consists in choosing $\alpha_*^p = \int d_Y^p \, \mathrm{d}(\nu \otimes \nu) / \int d_X^p \, \mathrm{d}(\mu \otimes \mu)$ and setting $\rho_* = \rho_{\alpha_*}$. This defines the *isometric transport correlation*, which is symmetric in ξ and ζ and has the following properties (Proposition 11):

- $\rho_*(\gamma) = 0$ iff ξ and ζ are independent,
- $\rho_*(\gamma) = 1$ iff there is a $\beta > 0$ and an isometry $\varphi\colon (X, \beta d_X) \to (Y, d_Y)$ with $\zeta = \varphi(\xi)$,

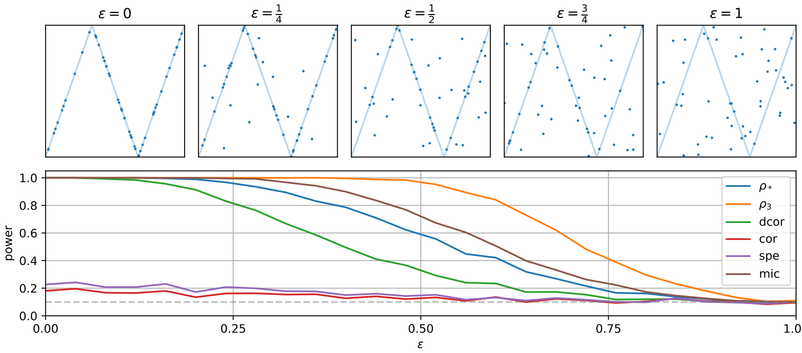


Figure 4: Power of independence tests under increasing levels of noise. The upper row depicts exemplary samples of size $n = 50$ at varying noise levels ϵ. The undisturbed samples ($\epsilon = 0$) are drawn from the graph of a 3-Lipschitz zigzag function. The figure on the bottom depicts the power of level 0.1 permutation tests for independence based on ρ_*, ρ_α for $\alpha = 3$, the distance correlation, the Pearson correlation, the Spearman correlation, and the maximal information coefficient. See Section 6 for more details on the setting.

where, again, φ only needs to be defined μ-almost surely (and β has to equal α_*). This shows that the coefficient ρ_* assumes its extremal values $\rho_*(\gamma) = 0$ and $\rho_*(\gamma) = 1$ for the same joint distributions γ as the distance correlation dcor. In particular ρ_* satisfies the “Four simple axioms for dependency measures” proposed by Móri and Székely 2019 and may be explored as a principled alternative to dcor for quantifying dependency in general Polish spaces. When we inspect commonalities of and differences between ρ_α, ρ_*, dcor, and a series of other dependency coefficients later on (Section 6), we find that their absolute values are often comparable in nature, but that their performance when testing for independence can differ considerably. Figure 4 conveys a first impression along this line, highlighting the interpretation of ρ_α as a coefficient well suited to detect α-Lipschitz relations between ξ and ζ.

The article is structured as follows. We begin by introducing our notation and the basic definitions of optimal transport and the transport dependency in Section 2. Section 3 revolves around establishing a number of general properties of τ as a functional on $\mathcal{P}(X \times Y)$. This entails convexity, invariance properties, continuity, the behavior under convolutions, as well as upper bounds. We also recover the marginal transport dependency as a limit case. In Section 4, we consider additive costs like (6) and focus on the distinguished role of contracting relations between random variables. The results in this section may be regarded as our main findings, since they lend the concept of transport dependency a clean interpretation. We make use of this interpretation in Section 5, where we introduce the transport correlation ρ_α and show that this family of coefficients features the properties summarized above. Section 6 concludes our work with a series of elementary numerical examples that hint at the potential of the transport dependency, comparing it to other measures of association. Proofs omitted in the main text in favor of a better reading flow are collected in Appendix A, and additional simulations can be found in Appendix B.

Discussion and outlook. Our work establishes the transport dependency as a versatile tool for the quantification of dependency between random variables. We particularly underscore that the transport correlation ρ_α provides a systematic way to interpolate between structured ($\alpha < \infty$) and unstructured ($\alpha = \infty$) dependency quantification through the selection of suitable cost functions. Moreover, the general flexibility of τ to work with any costs c, which do not even have to be based on metrics, could be attractive for a range of practical applications, as state of the art methods for comparing complex objects are often engineered in an ad-hoc fashion without strong mathematical guarantees. One prospect in this regard are nonparametric independence tests guided by arbitrary heuristics of similarity. In addition, probing the transport dependency of a data set under different costs c might expose structural properties of the data, which could render τ to become a handy tool for exploration.

Despite the apparent conceptual appeal, however, we do not want to conceal a major obstacle when the transport dependency is applied to data, which is its heavy computational burden. For the (naive) estimation of τ based on n points of data, an optimal transport problem of size $n \times n^2$ has to be solved. For generic exact optimal transport solvers, this leads to excessively long run times even for moderate problem sizes of, say, $n \approx 500$. By making use of an elaborate approximation method of optimal transport costs via entropy regularization (Cuturi 2013) proposed by Schmitzer 2019, we verify in our simulations (Section 6) that τ can accurately be calculated for $n = 1000$ in a matter of seconds to minutes on a modern GPU. Still, it is much more time consuming to compute τ than alternative coefficients of dependency, like the distance correlation. Since the application of optimal transport based methods and its surrogates to large datasets is an active research topic that steadily advances, however, it might only be a matter of time until current computational barriers are further mitigated and eventually overcome (for recent algorithmic progress, see for example Lin et al. 2019; Sommerfeld et al. 2019; Erbar et al. 2020; Jacobs and Léger 2020; Liu et al. 2021). Another potential hurdle for the application of τ on data sets is the bias that comes along with the empirical estimation of optimal transport costs, especially in high dimensions. Refining the theory of transport dependency to settings with structural regularity (like grid-based ground spaces) and devising specialized estimators might offer opportunity for improvements in this respect.

To conclude, we want to emphasize that this article should only be regarded as one step amongst many towards the goal of fully exploring the use of optimal transport as a statistical association measure between random variables. Indeed, as witnessed by the work of Wiesel 2021 and Mordant and Segers 2022 conducted in parallel to ours, this topic currently experiences a surge of attention, with many questions left open. In particular, the framework of transport dependency has yet to be vetted in real world applications to make sure that its promising performance is not limited to artificial settings. Furthermore, establishing a more dedicated statistical theory, including distributional and asymptotic properties or statistically efficient estimators, and a better sense of τ in high-dimensional settings remain an outstanding task that seems worth to be investigated in the future.

2 Optimal transport and transport dependency

In the following, we are always concerned with probability distributions on Polish spaces (which are separable and completely metrizable topological spaces) equipped with their respective Borel σ-algebra. On these spaces, analytic properties of optimal transport are well understood (Rachev and Rüschendorf 1998; Villani 2008; Ambrosio et al. 2008; Santambrogio 2015) and we have the general toolbox of disintegration, or conditioning, of measures at our disposal. See Kallenberg 2006 for a general treatment of this topic and Chang and Pollard 1997 for a view on conditioning through the lens of disintegration.

Notation. We denote the set of all probability measures on a Polish space X by $\mathcal{P}(X)$. The integration of a (Borel-)measurable function $f\colon X \to \mathbb{R}$ with respect to $\mu \in \mathcal{P}(Y)$ is flexibly denoted as $\int f(x)\,\mu(\mathrm{d}x)$, $\int f\,\mathrm{d}\mu$, or simply μf. Every measurable function $\varphi\colon X \to Y$ between Polish spaces induces a pushforward map $\varphi_*\colon \mathcal{P}(X) \to \mathcal{P}(Y)$ via $\mu \mapsto \varphi_*\mu = \mu \circ \varphi^{-1}$. We write $\mu_n \rightharpoonup \mu$ if a sequence $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(X)$ converges weakly (or in distribution) to μ in $\mathcal{P}(X)$. The product $X \times Y$ of two Polish spaces is again Polish if equipped with its product topology. We write p^X and p^Y for the Cartesian projections onto the spaces X and Y, and we let

$$(\varphi, \psi)(x) = \big(\varphi(x), \psi(x)\big) \qquad \text{and} \qquad (\varphi \times \psi)(x, y) = \big(\varphi(x), \psi(y)\big)$$

for suitable functions φ and ψ. The notation $\mathcal{C}(\mu, \nu) \subset \mathcal{P}(X \times Y)$ is used to denote the set of couplings (or transport plans) between measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Thus, $\gamma \in \mathcal{C}(\mu, \nu)$ is a joint distribution on $X \times Y$ with marginals $p^X_*\gamma = \mu$ and $p^Y_*\gamma = \nu$. We also write $\mathcal{C}(\mu, \cdot)$ or $\mathcal{C}(\cdot, \nu)$ for the subsets of $\mathcal{P}(X \times Y)$ where only one marginal distribution is fixed. If we want to condition $\gamma \in \mathcal{C}(\mu, \nu)$ on one of its components, we use the shorthand notation

$$\gamma(\mathrm{d}x, \mathrm{d}y) = \gamma(x, \mathrm{d}y)\,\mu(\mathrm{d}x)$$

to indicate a disintegration of γ along the space X, where $\gamma(x, \cdot) \in \mathcal{P}(Y)$ is a probability distribution for each $x \in X$. The family $\big(\gamma(x, \cdot)\big)_{x\in X}$ is a probability kernel (or Markov kernel or stochastic kernel), which means that the mapping $x \mapsto \gamma(x, A)$ is measurable for each Borel set $A \subset Y$. Analog notation will be used for conditioning on $y \in Y$ or if products of more than two spaces are considered.

Whenever convenient, we may express our arguments in terms of random elements instead of probability measures. Typically, we then refer to ξ and ζ as random elements on the spaces X and Y. Their joint law is usually $(\xi, \zeta) \sim \gamma \in \mathcal{C}(\mu, \nu)$. Note that $\gamma(x, \cdot)$ is the conditional distribution of ζ given $\xi = x$ and vice versa for $\gamma(\cdot, y)$, and that the pushforward $f_*\gamma$ equals the law of $f(\xi, \zeta)$ whenever $f\colon X \times Y \to Z$ is a measurable map into a Polish space.

We call a lower semi-continuous function $c\colon X \times X \to [0, \infty]$ a *cost function* on X if it is symmetric and vanishes on the diagonal, meaning $c(x_1, x_2) = c(x_2, x_1)$ and

$c(x, x) = 0$ for all $x, x_1, x_2 \in X$. Sometimes we require $c(x_1, x_2) > 0$ for $x_1 \neq x_2$, in which case we call the cost function *positive*. Finally, the *c-diameter* of a measure $\mu \in \mathcal{P}(X)$ is defined as

$$\operatorname{diam}_c \mu = \int c(x_1, x_2)\, \mathrm{d}\mu(x_1)\, \mathrm{d}\mu(x_2) = (\mu \otimes \mu)\, c, \tag{7}$$

where $\mu \otimes \mu$ denotes the product measure of μ with itself.

Optimal transport. The *optimal transport cost* T_c between measures μ and ν in $\mathcal{P}(X)$ for base costs c on a Polish space X is defined as

$$T_c(\mu, \nu) = \inf_{\pi \in \mathcal{C}(\mu, \nu)} \pi c. \tag{8}$$

Within the assumptions we work in, an *optimal transport plan* π^* that attains the infimum in equation (8) always exists if c is lower semi-continuous (Villani 2008, Theorem 4.1). In general, optimal transport plans need not be unique. If $c = d^p$ for a metric d on X with $p \geq 1$, the quantity $T_c(\mu, \nu)^{1/p}$ is called the *p-Wasserstein distance* and is a metric on the probability measures in $\mathcal{P}(X)$ that have finite p-th moments.

When a transport plan $\pi \in \mathcal{C}(\mu, \nu)$ is concentrated on the graph of a function $\varphi\colon X \longrightarrow X$, then φ is called *transport map* and satisfies $\varphi_*\mu = \nu$ and $(\mathrm{id}, \varphi)_*\mu = \pi$. Under certain conditions on μ, ν, and c, optimal transport plans π^* that minimize (8) correspond to optimal transport maps φ^* (see Villani 2008; Santambrogio 2015). For example, this always holds when μ has a Lebesgue density in $\mathbb{R}^d$ for $d \in \mathbb{N}$ and c is given by an l_p norm with $p > 1$.

The optimal transport problem (8) can alternatively be stated in its dual formulation,

$$T_c(\mu, \nu) = \sup_{f \oplus g \leq c} \mu f + \nu g, \tag{9}$$

where $(f \oplus g)(x_1, x_2) = f(x_1) + g(x_2)$ and where the supremum is taken over bounded and continuous functions f and g. This fact is commonly known as Kantorovich-duality, dating back to Kantorovich 1942. Like for the existence of transport plans, the cost function c being lower semi-continuous is sufficient for (8) and (9) to coincide (Villani 2008, Theorem 5.10). If c is a metric, (9) takes the particular form

$$T_c(\mu, \nu) = \sup_{f \in \mathrm{Lip}_1(X)} \mu f - \nu f, \tag{10}$$

where $\mathrm{Lip}_1(X)$ denotes all real valued 1-Lipschitz functions on X with respect to c. Note that this is a special case of an integral probability metric (Müller 1997; Sriperumbudur et al. 2012).

The literature on optimal transport is vast, deep, and fragmented over many disciplines, reaching from analysis over statistics and optimization to economics. Apart from the analytical work cited before, valuable resources are the recent monograph by Panaretos and Zemel 2020, which studies the Wasserstein space from a statistical perspective, and a book by Peyré and Cuturi 2019, which focuses on computational aspects of optimal transport.

Transport dependency. Let X and Y be Polish spaces, and let c be a cost function on the product space $X \times Y$. Recall that p^X and p^Y denote the Cartesian projections onto the spaces X and Y. We define the (c-)*transport dependency* $\tau_c \colon \mathcal{P}(X \times Y) \to [0, \infty]$ via

$$\tau(\gamma) = \tau_c(\gamma) = T_c\big(\gamma, p^X_* \gamma \otimes p^Y_* \gamma\big), \tag{11}$$

where we usually omit the cost function c in the subscript if it is apparent from the context. Each coupling $\gamma \in \mathcal{C}(\mu, \nu) \subset \mathcal{P}(X \times Y)$ can be understood as the joint distribution of random elements ξ and ζ with laws $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. In this case, the value $\tau(\xi, \zeta) = \tau(\gamma)$ quantifies how much the distribution of (ξ, ζ) deviates from the joint distribution $\mu \otimes \nu$ that ξ and ζ would have if they were independent. One can easily see that $\tau(\xi, \zeta) = 0$ in case of statistical independence. If c is a positive cost function, this criterion is even sufficient for independence of ξ and ζ.

Theorem 1 (independence): Let X and Y be Polish spaces and c a positive cost function on $X \times Y$. Then $\tau(\gamma) = 0$ if and only if $\gamma = \mu \otimes \nu$ for some $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$.

Proof. If $\gamma = \mu \otimes \nu$, picking the transport plan $\pi = (\mathrm{id}, \mathrm{id})_*(\mu \otimes \nu) \in \mathcal{C}(\gamma, \mu \otimes \nu)$ asserts that $0 \leq \tau(\gamma) \leq \pi c = 0$, where we used that $c \geq 0$ and that c vanishes on the diagonal. Conversely, if $\tau(\gamma) = 0$ for some $\gamma \in \mathcal{C}(\mu, \nu)$, we find $\pi^* c = 0$ for the optimal plan π^*, so $c = 0$ holds π^*-almost surely. Since c is positive, this implies $\pi^*\{(x, y, x, y) \mid x \in X, y \in Y\} = 1$. Consequently, it follows that $\pi^* = (\mathrm{id}, \mathrm{id})_*(\mu \otimes \nu) \in \mathcal{C}(\mu \otimes \nu, \mu \otimes \nu)$, establishing $\gamma = \mu \otimes \nu$. □

Example 1 (Multivariate Gaussian): Let (ξ, ζ) be a pair of real valued and normally distributed random vectors on $X \times Y = \mathbb{R}^{r+q}$ that follow a joint distribution $\gamma = \mathcal{N}(\eta, \Sigma)$ so that

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

with $\eta_1 \in \mathbb{R}^r$, $\eta_2 \in \mathbb{R}^q$, $\Sigma_{11} \in \mathbb{R}^{r \times r}$, $\Sigma_{22} \in \mathbb{R}^{q \times q}$, $\Sigma_{12} \in \mathbb{R}^{r \times q}$ and $\Sigma_{21} = \Sigma_{12}^T$. Since independence of normal random variables is characterized by zero correlation, the independent coupling is given by

$$\mu \otimes \nu = \mathcal{N}(\eta, \Sigma_{\mathrm{ind}}) \quad \text{with} \quad \Sigma_{\mathrm{ind}} = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.$$

As costs on the space $X \times Y$, we consider the squared Euclidean distance

$$c(x_1, y_1, x_2, y_2) = \|x_1 - x_2\|^2 + \|y_1 - y_2\|^2$$

for $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^r \times \mathbb{R}^q$. In this setting, evaluating the transport dependence of γ corresponds to computing the square of the 2-Wasserstein distance between two normal distributions. We obtain (Dowson and Landau 1982)

$$\tau(\gamma) = 2\,\mathrm{trace}(\Sigma_{11}) + 2\,\mathrm{trace}(\Sigma_{22}) - 2\,\mathrm{trace}\left(\begin{pmatrix} \Sigma_{11}^2 & \Sigma_{11}\Sigma_{12} \\ \Sigma_{22}\Sigma_{21} & \Sigma_{22}^2 \end{pmatrix}^{1/2}\right). \tag{12}$$

If ξ and ζ are univariate random variables, a more explicit formula can easily be derived. Let the covariance matrix be given by

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

with $\rho \in [0, 1)$ and $\sigma_1, \sigma_2 > 0$. Then, using an identity provided in (Levinger 1980), we can compute the square root of the 2×2 matrix appearing in (12) and obtain

$$\tau(\gamma) = 2\left(\sigma_1^2 + \sigma_2^2 - \sqrt{\sigma_1^4 + \sigma_2^4 + 2\sigma_1^2\sigma_2^2\sqrt{1-\rho^2}}\right). \tag{13}$$

If $\sigma_1 = \sigma_2 = \sigma$ for some $\sigma > 0$, this expression simplifies to

$$\tau(\gamma) = 2\sigma^2\left(2 - \sqrt{2 + 2\sqrt{1-\rho^2}}\right).$$

As to be expected, the transport dependency between ξ and ζ is a strictly increasing function of the correlation ρ^2. Its minimal value is 0 for $\rho = 0$ and its maximal value is $(4 - 2\sqrt{2})\,\sigma^2 \approx 1.2\sigma^2$ for $\rho = \pm 1$ (if $\sigma_1 = \sigma_2 = \sigma$). Note that the mutual information and the Euclidean distance covariance also have closed forms in the bivariate normal setting. These are given by $M(\gamma) = -\log(1-\rho^2)/2$ and

$$\mathrm{dcov}^2(\gamma) = \frac{4\sigma^2}{\pi}\left(\rho \arcsin\rho + \sqrt{1-\rho^2} - \rho\arcsin(\rho/2) - \sqrt{4-\rho^2} + 1\right),$$

see Gelfand 1959 and Székely et al. 2007. Figure 5 depicts how the values of τ compare to these measures of dependency as a function of ρ.

3 General properties

The transport dependency features several desirable traits for a measure of statistical association. In this section, we discuss a series of its general properties, including convexity, symmetry, continuity, and the behavior under convolutions. We also establish three distinct upper bounds, one of which corresponds to the special case of τ when movement of mass along the space X is forbidden. This naturally leads to the definition of the marginal transport dependency.

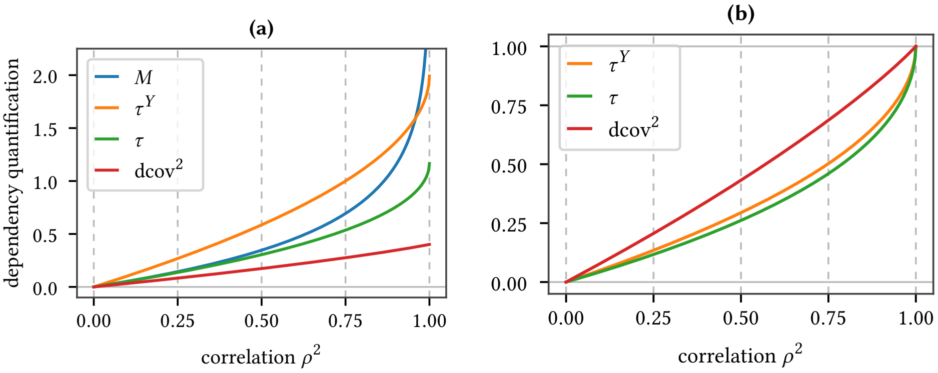


Figure 5: Behavior of selected dependency measures in the bivariate case $\xi, \zeta \sim \mathcal{N}(0,1)$ with $\operatorname{cov}(\xi, \zeta) = \rho$, as discussed in Example 1. Graph **(a)** shows the mutual information (which diverges as $\rho^2 \longrightarrow 1$), the Euclidean distance covariance (3), the marginal transport dependency (2), and the general transport dependency as function of ρ^2. Graph **(b)** shows the latter three quantities normalized such that their respective maximal value equals 1 for $\rho^2 = 1$. See the forthcoming Example 5 for a closed form of the marginal transport dependency τ^Y in this setting.

Convexity and invariance. The optimal transport cost T_c is convex in both of its arguments. As a consequence, the transport dependency τ can also be shown to be convex, albeit with some restrictions. Following the arguments in Villani 2008, Theorem 4.8, it is straightforward to establish convexity on subsets of $\mathcal{P}(X \times Y)$ that share (at least) one marginal.

> Proposition 1 (convexity): Let X and Y be Polish spaces and c a cost function on $X \times Y$. Fix marginal distributions $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then the transport dependency τ is convex when restricted to $\mathcal{C}(\mu, \cdot)$ or $\mathcal{C}(\cdot, \nu)$.

Proof. Let $\gamma_0, \gamma_1 \in \mathcal{C}(\mu, \cdot)$ with second marginals $\nu_i \in \mathcal{P}(Y)$ for $i \in \{0, 1\}$, and let π_i^* be an optimal transport plan between γ_i and $\mu \otimes \nu_i$ with respect to c. Then, defining $\gamma_t = (1-t)\gamma_0 + t\gamma_1$ for $t \in [0,1]$ (and similarly ν_t), we find $\gamma_t \in \mathcal{C}(\mu, \nu_t)$ and $\pi_t = (1-t)\pi_0^* + t\pi_1^* \in \mathcal{C}(\gamma_t, \mu \otimes \nu_t)$. Hence,

$$\tau(\gamma_t) = T_c(\gamma_t, \mu \otimes \nu_t) \leq \pi_t c = (1-t)\,\tau(\gamma_0) + t\,\tau(\gamma_1),$$

which establishes convexity on $\mathcal{C}(\mu, \cdot)$. The result on $\mathcal{C}(\cdot, \nu)$ follows analogously. □

Note that the marginal restrictions on the convexity in Proposition 1 are reasonable, since τ cannot be convex on the whole space $\mathcal{P}(X \times Y)$. In fact, if it were, then

$$0 \leq 2\,\tau\big(\delta_{z_1}/2 + \delta_{z_2}/2\big) \leq \tau(\delta_{z_1}) + \tau(\delta_{z_2}) = 0,$$

where δ_{z_i} denotes the point mass at z_i for points $z_1, z_2 \in X \times Y$. Pursuing this thought further, it is easy to see that τ would have to vanish on all empirical measures – which it clearly does not. The same argument holds for any other dependency measure as soon as it assigns the value 0 to point measures, which shows that convexity on all of $\mathcal{P}(X \times Y)$ is generally not desirable. On the contrary, it is intuitive to expect the amount of dependency inherent in $t\gamma_1 + (1-t)\gamma_2$ to exceed the one in γ_1 or γ_2 individually if the supports are sufficiently distinct.

Recalling that $\tau(\mu \otimes \nu) = 0$ for any $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, one interesting conclusion of Proposition 1 is that the map $t \mapsto \tau(\gamma_t)$ with $\gamma_t = (1-t)\gamma + t(\mu \otimes \nu)$ for $\gamma \in \mathcal{C}(\mu, \nu)$ and $t \in [0,1]$ is convex, monotonically decreasing, and that it satisfies

$$\tau(\gamma_t) \leq (1-t)\,\tau(\gamma). \tag{14}$$

Therefore, the contamination of γ with an independent contribution consistently decreases the transport dependency. When c is a metric, we even find equality

$$\tau(\gamma_t) = (1-t)\,\tau(\gamma)$$

for all $t \in [0,1]$. In this case, the optimal transport plan π_t^* mediating between γ_t and $\mu \otimes \nu$ will not move the mass $t(\mu \otimes \nu)$ at all. Only the transportation of $(1-t)\,\gamma$ to the remainder $(1-t)\,(\mu \otimes \nu)$ contributes to $\tau(\gamma_t)$, which can be validated via the dual formulation (10):

$$\tau(\gamma_t) = \sup\big(\gamma_t f - (\mu \otimes \nu) f\big) = (1-t)\,\sup\big(\gamma f - (\mu \otimes \nu) f\big) = (1-t)\,\tau(\gamma),$$

where both suprema are taken over $f \in \mathrm{Lip}_1(X \times Y)$ with respect to c.

Next, we shed some light on the symmetries of the transport dependency τ. They follow from a fundamental statement of invariance of the optimal transport cost (a proof is provided in Appendix A).

> Lemma 1: Let X and Y be Polish spaces and let c be a cost function on Y. Consider a measurable function $f\colon X \to Y$ and let c_f denote the induced cost function on X defined by $c_f(x, x') = c\big(f(x), f(x')\big)$. Then for any $\mu, \nu \in \mathcal{P}(X)$,
>
> $$T_c(f_*\mu, f_*\nu) = T_{c_f}(\mu, \nu).$$

Let us consider a measurable function f that maps $X \times Y$ into itself. In order to be able to apply Lemma 1 to $\tau(\gamma)$ for $\gamma \in \mathcal{C}(\mu, \nu) \subset \mathcal{P}(X \times Y)$, we have to require that $f_*(\mu \otimes \nu)$ is again a product measure of, say, distributions μ' and ν' in $\mathcal{P}(X)$ and $\mathcal{P}(Y)$. When $f_*\gamma$ has the same marginals μ' and ν', we may conclude

$$\tau_c(f_*\gamma) = \tau_{c_f}(\gamma).$$

If, in addition, applying f does not affect the costs c, meaning that $c = c_f$, we even obtain the invariance relation

$$\tau_c(f_*\gamma) = \tau_c(\gamma). \tag{15}$$

This observation is particularly potent if c is a function of marginal costs.

Proposition 2 (invariance): Let X and Y be Polish spaces and let c be a cost function on $X \times Y$ satisfying

$$c(x_1, y_1, x_2, y_2) = h\big(c_X(x_1, x_2), c_Y(y_1, y_2)\big) \tag{16}$$

for marginal costs c_X and c_Y on X and Y, and a measurable function $h\colon [0, \infty]^2 \to [0, \infty]$. If $f_X\colon X \to X$ and $f_Y\colon Y \to Y$ are measurable maps that leave c_X and c_Y invariant, then $f = f_X \times f_Y$ leaves c invariant and (15) holds for any $\gamma \in \mathcal{P}(X \times Y)$.

Proof. For $\gamma \in \mathcal{C}(\mu, \nu)$ with $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, it is sufficient to observe that $c_f = c$ as well as $f_*(\mu \otimes \nu) = f_{X*}\mu \otimes f_{Y*}\nu$ and $f_*\gamma \in \mathcal{C}(f_{X*}\mu, f_{Y*}\nu)$, which makes Lemma 1 applicable. □

If c_X and c_Y are (based on) metrics and ξ and ζ are random elements on X and Y, we conclude that applying isometries to either ξ or ζ does not change the value $\tau(\xi, \zeta)$. In other words, τ measures dependency in a way that is indifferent to isometric transformations on the margins, a property that is shared by the distance covariance (4). We also note that Proposition 2 can readily be generalized to cost preserving maps between distinct Polish spaces.

Example 2 (invariance in Euclidean space): Let ξ and ζ be random vectors in $X = \mathbb{R}^r$ and $Y = \mathbb{R}^q$. If c_X and c_Y denote the respective Euclidean metrics and c takes the form $h(c_X, c_Y)$, for example $c = c_X^2 + c_Y^2$ as in Example 1, Proposition 2 asserts that

$$\tau(\xi, \zeta) = \tau(A\,\xi + a, B\,\zeta + b)$$

for all orthogonal matrices $A \in \mathbb{R}^{r\times r}$, $B \in \mathbb{R}^{q\times q}$ as well as vectors $a \in \mathbb{R}^r$, $b \in \mathbb{R}^q$.

Continuity and convergence. The optimal transport cost T_c with lower semi-continuous base costs c is again lower semi-continuous with respect to the weak convergence of measures. This property carries over to the transport dependency.

Proposition 3 (lower semi-continuity): Let X and Y be Polish spaces and let c be a cost function on $X \times Y$. Then τ is lower semi-continuous with respect to weak convergence.

Proof. Let $(\gamma_n)_{n\in\mathbb{N}} \subset \mathcal{P}(X \times Y)$ be a sequence of probability measures that weakly converges to $\gamma \in \mathcal{C}(\mu, \nu)$ for $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. By the continuous mapping theorem, the respective marginals $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ weakly converge to μ and ν. According to Billingsley 2013, Theorem 2.8, this implies weak convergence of the product measures $\mu_n \otimes \nu_n$ to $\mu \otimes \nu$.

The result of the proposition now follows along the lines of the proof of the lower semi-continuity of the optimal transport cost (see for example Santambrogio 2015, Proposition 7.4, which requires only slight adaptations to make it work for sequences in both arguments of T_c). □

To approach proper continuity of τ, we will need a stronger set of assumptions, since the optimal transport cost T_c is in general not continuous under weak convergence, even for continuous costs c (cf. Santambrogio 2015, Proposition 7.4). Indeed, we have to sharpen the notion of convergence $\gamma_n \longrightarrow \gamma$ in $\mathcal{P}(X \times Y)$. To this end, we equip the Polish space X with a compatible metric d_X that (completely) metrizes its topology. For $p \geq 1$, we define the set of probability distributions with finite p-th moment by

$$\mathcal{P}_p(X) = \big\{\mu \in \mathcal{P}(X) \,\big|\, \mu\, d_X^p(\cdot, x_0) < \infty \text{ for some } x_0 \in X\big\},$$

and we say that a sequence $(\mu_n)_{n\in\mathbb{N}}$ in $\mathcal{P}_p(X)$ converges p-weakly to $\mu \in \mathcal{P}_p(X)$ if

$$\mu_n \rightharpoonup \mu \qquad \text{and} \qquad \mu_n\, d_X(\cdot, x_0)^p \longrightarrow \mu\, d_X(\cdot, x_0)^p \tag{17}$$

as $n \longrightarrow \infty$ for some $x_0 \in X$. It is a well known fact (going back to Mallows 1972) that the p-Wasserstein distance metrizes this particular form of convergence (see Villani 2008, Theorem 6.9 for a general proof), so (17) is equivalent to $T_{d_X^p}(\mu_n, \mu) \longrightarrow 0$ as $n \longrightarrow \infty$. Note that the anchor point x_0 in these definitions does not matter and can be replaced by any other point in X.

On products $X \times Y$ of two Polish metric spaces (X, d_X) and (Y, d_Y), there are many different ways to choose a compatible metric. For simplicity, we pick

$$d(x_1, y_1, x_2, y_2) = d_X(x_1, x_2) + d_Y(y_1, y_2) \tag{18}$$

for $(x_1, y_1), (x_2, y_2) \in X \times Y$ in the following statement, even though any equivalent metric, like $d = \max(d_X, d_Y)$, would work as well. The proof can be found in Appendix A.

> Proposition 4 (continuity): Let (X, d_X) and (Y, d_Y) be Polish metric spaces and let c be a continuous cost function on $X \times Y$ bounded by $c \leq d^p$ for $p \geq 1$ and d as in (18). If the sequence $(\gamma_n)_{n\in\mathbb{N}}$ converges p-weakly to γ in $\mathcal{P}_p(X \times Y)$, then $\tau(\gamma_n) \longrightarrow \tau(\gamma)$.

One significant consequence of Proposition 4 is the guarantee of consistency when empirically estimating $\tau(\gamma)$. Let $(\xi_i, \zeta_i) \sim \gamma$ for $i \in \{1, \ldots, n\}$ be n independent and identically distributed random variables. We denote the empirical measure associated to γ by

$$\hat{\gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(\xi_i, \zeta_i)}, \tag{19}$$

which is a random element in $\mathcal{P}(X \times Y)$. The next result shows that we can simply plug $\hat{\gamma}_n$ in τ to obtain a strongly consistent estimator of $\tau(\gamma)$.

Theorem 2 (consistency): Let (X, d_X) and (Y, d_Y) be Polish metric spaces and c a continuous cost function on $X \times Y$ bounded by $c \leq d^p$ for some $p \geq 1$ and d as in (18). If $\gamma \in \mathcal{P}_p(X \times Y)$, then almost surely

$$\lim_{n\to\infty} \tau(\hat{\gamma}_n) = \tau(\gamma).$$

Proof. We know that $\hat{\gamma}_n \rightharpoonup \gamma$ almost surely as $n \to \infty$ (Varadarajan 1958). Furthermore, for any anchor point $(x_0, y_0) \in X \times Y$, the strong law of large numbers implies that the convergence

$$\hat{\gamma}_n\, d\big(\cdot, (x_0, y_0)\big)^p = \frac{1}{n}\sum_{i=1}^{n} d\big((\xi_i, \zeta_i), (x_0, y_0)\big)^p \to \gamma\, d\big(\cdot, (x_0, y_0)\big)^p < \infty$$

holds almost surely as $n \to \infty$. This makes Proposition 4 applicable and the assertion follows. □

Later on, when we consider the transport correlation coefficients ρ_α in Section 5, we face costs that are normalized in a way that depends on γ. In order to show that continuity and consistency are preserved in this setting, we need a slight extension of our previous results to uniformly perturbed cost functions. We write $\|\cdot\|_\infty$ to denote the sup-norm of a real valued function and use the convention $0/0 = 1$ in the following statement.

Proposition 5 (varying costs): Let X and Y be Polish spaces and let c and c_n be cost functions on $X \times Y$ that satisfy $\|c/c_n - 1\|_\infty \to 0$ as $n \to \infty$. Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}(X \times Y)$ and $\gamma \in \mathcal{P}(X \times Y)$ such that $\tau_c(\gamma) < \infty$. Then

$$\lim_{n\to\infty} \tau_c(\gamma_n) = \tau_c(\gamma) \qquad \text{implies} \qquad \lim_{n\to\infty} \tau_{c_n}(\gamma_n) = \tau_c(\gamma). \tag{20}$$

Proof. We set $a_n = \max\big(\|c/c_n - 1\|_\infty, \|c_n/c - 1\|_\infty\big)$ and observe $a_n \to 0$ as $n \to \infty$ by the assumption of uniform convergence. Applying Lemma 4 provided in Appendix A, we can control the deviation of τ_{c_n} from τ_c and find, as $n \to \infty$,

$$\begin{aligned} |\tau_{c_n}(\gamma_n) - \tau_c(\gamma)| &\leq |\tau_{c_n}(\gamma_n) - \tau_c(\gamma_n)| + |\tau_c(\gamma_n) - \tau_c(\gamma)| \\ &\leq a_n(1 + a_n)\,\tau_c(\gamma_n) + |\tau_c(\gamma_n) - \tau_c(\gamma)| \to 0. \end{aligned}$$

□

We also want to present an alternative continuity statement with explicit bounds in case that c is *equal* to the power of a metric d on $X \times Y$. This time, we need nothing more than the triangle inequality, so d may in fact be a pseudo-metric.

Theorem 3: Let X and Y be Polish spaces and let $c = d^p$ for a (pseudo-)metric d on $X \times Y$. Assume that $\gamma, \gamma' \in \mathcal{P}(X \times Y)$ with $\tau(\gamma), \tau(\gamma') < \infty$. Then, for any

$p \geq 1$,

$$\left|\tau(\gamma)^{1/p} - \tau(\gamma')^{1/p}\right| \leq T_c(\gamma, \gamma')^{1/p} + T_c(\mu \otimes \nu, \mu' \otimes \nu')^{1/p}.$$

Proof. Since d satisfies the triangle inequality, so does $\rho = T_c^{1/p}$ (see Villani 2008, Definition 6.1). Furthermore, the stated inequality is trivial if the right hand side is ∞. Thus, we can assume $\rho(\gamma, \gamma') < \infty$ and $\rho(\mu \otimes \nu, \mu' \otimes \nu') < \infty$. Then also $\rho(\gamma', \mu \otimes \nu) \leq \rho(\gamma', \mu' \otimes \nu') + \rho(\mu \otimes \nu, \mu' \otimes \nu') < \infty$. Now, the result follows from applying the reverse triangle inequality:

$$\begin{aligned}\left|\tau(\gamma)^{1/p} - \tau(\gamma')^{1/p}\right| &= \left|\rho(\gamma, \mu \otimes \nu) - \rho(\gamma', \mu' \otimes \nu')\right| \\ &\leq \left|\rho(\gamma, \mu \otimes \nu) - \rho(\gamma', \mu \otimes \nu)\right| + \left|\rho(\gamma', \mu \otimes \nu) - \rho(\gamma', \mu' \otimes \nu')\right| \\ &\leq \rho(\gamma, \gamma') + \rho(\mu \otimes \nu, \mu' \otimes \nu'). \qquad \square\end{aligned}$$

If the metric d on $X \times Y$ in the statement above is given in terms of marginal metrics d_X and d_Y on X and Y, like $d = d_X + d_Y$, then one can apply the triangle inequality and upper bound

$$\begin{aligned}T_c(\mu \otimes \nu, \mu' \otimes \nu')^{1/p} &\leq T_c(\mu \otimes \nu, \mu' \otimes \nu)^{1/p} + T_c(\mu' \otimes \nu, \mu' \otimes \nu')^{1/p} \\ &\leq T_{c_X}(\mu, \mu')^{1/p} + T_{c_Y}(\nu, \nu')^{1/p} \\ &\leq 2\, T_c(\gamma, \gamma')^{1/p} \qquad (21)\end{aligned}$$

for $c_X = d_X^p$ and $c_Y = d_Y^p$. In this situation, Theorem 3 implies that $\tau^{1/p}$ is Lipschitz continuous with respect to the p-Wasserstein distance, which can be used to deduce rates of convergence of the empirically estimated transport dependency $\tau(\hat{\gamma}_n)$. More specifically, combining Theorem 3 and observation (21) yields

$$\left|\tau(\hat{\gamma}_n)^{1/p} - \tau(\gamma)^{1/p}\right| \leq 3\, T_c(\hat{\gamma}_n, \gamma)^{1/p}. \qquad (22)$$

Any result on convergence rates $T_c(\hat{\gamma}_n, \gamma)^{1/p} \longrightarrow 0$ can therefore be used to derive analog rates for $\tau(\hat{\gamma}_n)^{1/p} \longrightarrow \tau(\gamma)^{1/p}$. One example is a bound on the p-Wasserstein distance for compact spaces derived by Weed and Bach 2019. For any s greater than the *upper Wasserstein dimension* of γ (as defined in their work), their results imply

$$\mathbb{E}\left[\left|\tau(\hat{\gamma}_n)^{1/p} - \tau(\gamma)^{1/p}\right|\right] \leq K\, n^{-1/s},$$

where $K > 0$ is a constant that does not depend on n. Other recent convergence result that can be applied in the context of (22) include distributional limits on countable spaces (Tameling et al. 2019) and finite sample convergence rates on unbounded spaces (Lei 2020).

Convolutions. In many statistical applications, the sum of independent random elements plays a distinguished role. Regression models, for example, assume a relation of the form $\zeta = \varphi(\xi)$ for random variables ζ and ξ determined by a function

$\varphi\colon X \longrightarrow Y$ between two vector spaces. Instead of observing samples of ξ and ζ directly, however, only polluted versions $\xi + \epsilon_X$ and $\zeta + \epsilon_Y$ are accessible, where ϵ_X and ϵ_Y are considered to be noise contributions independent of ξ and ζ. In this case, the distribution μ of ξ and the actually observed distribution $\tilde{\mu}$ of $\xi + \epsilon_X$ are related by a convolution:

$$\tilde{\mu} = \mu * \kappa_X, \qquad \text{where} \qquad (\mu * \kappa_X)(A) = \int \mathbb{1}_A(x_1 + x_2)\,(\mu \otimes \kappa_X)(\mathrm{d}x_1, \mathrm{d}x_2)$$

for any Borel-measurable set $A \subset X$, and where κ_X denotes the law of ϵ_X. An analog relation holds for $\zeta \sim \nu$ with noise distribution κ_Y, and the joint distributions of (ξ, ζ) and $(\xi + \epsilon_X, \zeta + \epsilon_Y)$ are connected by

$$\tilde{\gamma} = \gamma * \kappa, \qquad \text{where} \qquad \kappa = \kappa_X \otimes \kappa_Y.$$

In the following, we investigate how the transport dependency $\tau(\gamma * \kappa)$ is related to the (noise-free) value $\tau(\gamma)$. As preparation, we take a look at the effect of convolutions on optimal transport costs if the base cost is translation invariant. We work in Polish vector spaces, by which we mean topological vector spaces that are Polish; examples include separable Banach spaces. A proof of the following statement is provided in Appendix A.

Lemma 2: Let X be a Polish vector space and $c(x_1, x_2) = h(x_1 - x_2)$ for $x_1, x_2 \in X$ a translation invariant cost function on X. For any probability measures μ, ν, and κ in $\mathcal{P}(X)$,

$$T_c(\mu * \kappa, \nu * \kappa) \leq T_c(\mu, \nu) \qquad \text{and} \qquad T_c(\mu, \mu * \kappa) \leq \kappa h.$$

When trying to apply Lemma 2 to the transport dependency $\tau(\gamma) = T_c(\gamma, \mu \otimes \nu)$ for $\gamma \in \mathcal{C}(\mu, \nu)$ in a sensible manner, we have to make sure that the convolution of $\mu \otimes \nu$ with a kernel $\kappa \in \mathcal{P}(X \times Y)$ is again a product distribution. As long as we restrict ourselves to product kernels of the form $\kappa = \kappa_X \otimes \kappa_Y$ with $\kappa_X \in \mathcal{P}(X)$ and $\kappa_Y \in \mathcal{P}(Y)$, this condition is always satisfied and Lemma 2 can be put to use.

Theorem 4 (convolution): Let X and Y be Polish vector spaces and let c be a cost function on $X \times Y$ that satisfies $c(x_1, y_1, x_2, y_2) = h(x_1 - x_2, y_1 - y_2)$ for $(x_1, y_1), (x_2, y_2) \in X \times Y$ and some $h\colon X \times Y \longrightarrow [0, \infty]$. For any $\gamma \in \mathcal{P}(X \times Y)$ and $\kappa = \kappa_X \otimes \kappa_Y$ with $\kappa_X \in \mathcal{P}(X)$ and $\kappa_Y \in \mathcal{P}(Y)$, it holds that

$$\tau(\gamma * \kappa) \leq \tau(\gamma).$$

If additionally $c = d^p$ and $\tau(\gamma) < \infty$ for a (pseudo-)metric d on $X \times Y$ with $p \geq 1$, then

$$\tau(\gamma)^{1/p} - \tau(\gamma * \kappa)^{1/p} \leq 2(\kappa h)^{1/p}.$$

Proof. One can easily check that $\gamma * \kappa \in \mathcal{C}(\mu * \kappa_X, \nu * \kappa_Y)$ and $(\mu \otimes \nu) * \kappa = (\mu * \kappa_X) \otimes (\nu * \kappa_Y)$. Therefore, $\tau(\gamma * \kappa) = T_c\big(\gamma * \kappa, (\mu \otimes \nu) * \kappa\big)$ and Lemma 2 can be applied, which yields the first result. For the second result, we recall Theorem 3 and use the second part of Lemma 2 to find

$$\tau(\gamma)^{1/p} - \tau(\gamma * \kappa)^{1/p} \leq T_c(\gamma, \gamma * \kappa)^{1/p} + T_c\big(\mu \otimes \nu, (\mu \otimes \nu) * \kappa\big)^{1/p} \leq 2(\kappa h)^{1/p},$$

which finishes the proof. □

Theorem 4 unveils a fundamental property of the transport dependency: if a coupling γ is blurred by the convolution with a product kernel, then τ never increases. Intuitively, this is a desirable trait. In common regression settings with Gaussian noise, for example, it guarantees that τ monotonically decreases with increasing standard deviation of the noise. At the same time, the second part of Theorem 4 lets us control by how much we decrease the transport dependency at most.

Example 3 (Gaussian additive noise): Let $X = \mathbb{R}^r$ and $Y = \mathbb{R}^q$ and consider squared Euclidean costs of the form $c(x_1, y_1, x_2, y_2) = h(x_1 - x_2, y_1 - y_2) = \|x_1 - x_2\|^2 + \|y_1 - y_2\|^2$, where $x_1, x_2 \in X$ and $y_1, y_2 \in Y$. Under Gaussian noise contributions $\kappa_X \sim \mathcal{N}(0, \Sigma_X)$ and $\kappa_Y \sim \mathcal{N}(0, \Sigma_Y)$, where Σ_X and Σ_Y are covariance matrices in $\mathbb{R}^{r\times r}$ and $\mathbb{R}^{q\times q}$, Theorem 4 states that

$$\tau(\gamma)^{1/2} \geq \tau(\gamma * \kappa)^{1/2} \geq \tau(\gamma)^{1/2} - 2\big(\operatorname{trace}\Sigma_X + \operatorname{trace}\Sigma_Y\big)^{1/2}$$

for any $\gamma \in \mathcal{P}(X \times Y)$ with $\tau(\gamma) < \infty$.

Example 4 (Manhattan-type costs): If $\big(X, \|\cdot\|_X\big)$ and $\big(Y, \|\cdot\|_Y\big)$ are separable Banach spaces and the costs c are given by $h = \|\cdot\|_X + \|\cdot\|_Y$, application of Theorem 4 yields

$$\tau(\gamma) \geq \tau(\gamma * \kappa) \geq \tau(\gamma) - 2\left(\int \|x\|_X\, \kappa_X(\mathrm{d}x) + \int \|y\|_Y\, \kappa_Y(\mathrm{d}y)\right)$$

for any $\gamma \in \mathcal{P}(X \times Y)$, $\kappa_X \in \mathcal{P}(X)$, and $\kappa_Y \in \mathcal{P}(Y)$, as long as $\tau(\gamma) < \infty$ holds.

The role of convolutions in the theory of optimal transport has recently been studied in the context of *smoothed optimal transport*, where $T_c^\kappa(\mu, \nu) = T_c(\mu * \kappa, \nu * \kappa)$ for a suitable kernel κ, often Gaussian, is used as an approximation of T_c (Goldfeld et al. 2020; Goldfeld and Greenewald 2020; Chen and Niles-Weed 2021). This smoothed transport cost has a series of desirable properties that distinguish it from vanilla optimal transport. Most notably, it does not suffer from the curse of dimensionality when estimated from empirical data, which makes it a promising candidate for applications.

Upper bounds. Our next goal is to derive upper bounds of the transport dependency. We take the common route of bounding the infimum in definition (8) of the optimal transport problem by explicitly constructing transport plans π with feasible marginals. The overall idea behind our constructions is to restrict the transport in $X \times Y$ to the fibers $X \times \{y\}$ and $\{x\} \times Y$ for $x \in X$ and $y \in Y$. For this reason, we often have to deal with costs of the form $c(x, y_1, x, y_2)$ or $c(x_1, y, x_2, y)$, and we will thus assume that the cost function c is controlled by suitable marginal costs c_X and c_Y on X and Y via

$$c_X(x_1, x_2) \geq \sup_{y \in Y} c(x_1, y, x_2, y) \qquad \text{and} \qquad c_Y(y_1, y_2) \geq \sup_{x \in X} c(x, y_1, x, y_2) \tag{23}$$

for all $(x_1, y_1), (x_2, y_2) \in X \times Y$. To gain some intuition for the (in total three) upper bounds we propose in the following, Figure 6 can be consulted. We start with a simple but fundamental inequality that controls $\tau(\gamma)$ in terms of the diameters of the marginal distributions of γ.

Proposition 6: Let $\gamma \in \mathcal{C}(\mu, \nu)$ for $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ in Polish spaces X and Y. If the cost function c on $X \times Y$ satisfies the marginal bounds (23), then

$$\tau(\gamma) \leq \operatorname{diam}_{c_Y} \nu. \tag{24}$$

By symmetry, the inequality $\tau(\gamma) \leq \operatorname{diam}_{c_X} \mu$ holds as well.

Proof. In order to bound T_c from above, we explicitly construct a transport plan π_1 that mediates between γ and $\mu \otimes \nu$. We choose it in such a way that no transport within X takes place. To this end, let $r(x_1, y_1, x_2, y_2) = (x_1, y_1, x_1, y_2)$ be a mapping from $X \times Y$ to itself. We set $\pi_1 = r_*(\gamma \otimes \gamma)$, which can easily be checked to be an element in $\mathcal{C}(\gamma, \mu \otimes \nu)$. Therefore,

$$T_c(\gamma, \mu \otimes \nu) \leq \pi_1 c = \int c(x_1, y_1, x_1, y_2)\, \gamma(\mathrm{d}x_1, \mathrm{d}y_1)\, \gamma(\mathrm{d}x_2, \mathrm{d}y_2) \leq (\nu \otimes \nu)\, c_Y, \tag{25}$$

where we made use of condition (23) to bound c by c_Y. This establishes $\tau(\gamma) \leq \operatorname{diam}_{c_Y} \nu$. □

The preceding arguments can equivalently be formulated in the language of random variables. When $(\xi, \zeta) \sim \gamma$ and $(\xi', \zeta') \sim \mu \otimes \nu$, the construction of π_1 above corresponds to the dependency structure $\xi = \xi'$, while ζ and ζ' remain independent. This prevents any transport within the space X. To clarify the transport along Y, we can condition on $\xi = x$ and slightly reformulate equation (25) as

$$\tau(\gamma) \leq \int \left(\int c_Y(y_1, y_2)\, \gamma(x, \mathrm{d}y_1)\, \nu(\mathrm{d}y_2) \right) \mu(\mathrm{d}x). \tag{26}$$

The inner integral on the right hand side is the transport cost between $\gamma(x, \cdot)$ and ν along the product plan $\pi_x = \gamma(x, \cdot) \otimes \nu$. Thus, for each fixed x, the plan π_1 evenly smears the measure ν into $\gamma(x, \cdot)$. If ζ is a function of ξ, meaning $\zeta = \varphi(\xi)$ for some

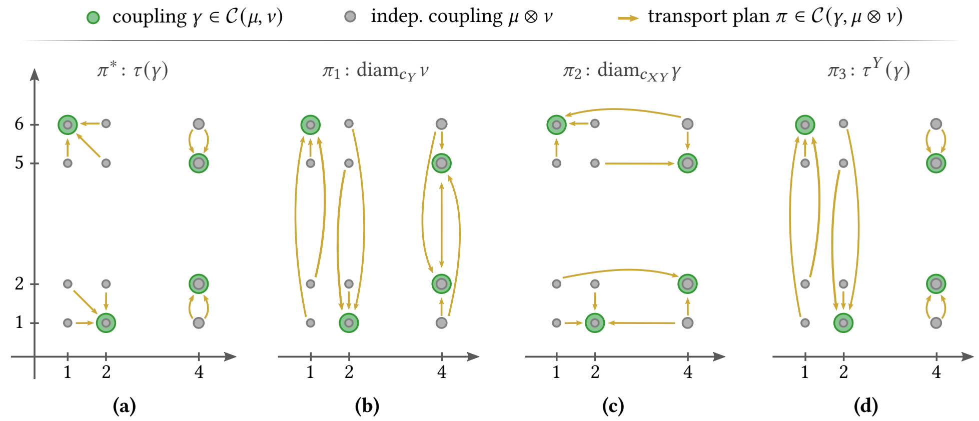

● coupling $\gamma \in \mathcal{C}(\mu, \nu)$
● indep. coupling $\mu \otimes \nu$
→ transport plan $\pi \in \mathcal{C}(\gamma, \mu \otimes \nu)$
$\pi^* : \tau(\gamma)$
$\pi_1 : \mathrm{diam}_{c_Y} \nu$
$\pi_2 : \mathrm{diam}_{c_{XY}} \gamma$
$\pi_3 : \tau^Y(\gamma)$
6
5
2
1
1
2
4
(a)
(b)
(c)
(d)

Figure 6: Several different transport plans between $\gamma \in \mathcal{C}(\mu \otimes \nu)$ and $\mu \otimes \nu$ for $\gamma = \mathrm{Unif}\{(1,6),(2,1),(4,2),(4,5)\}$. Each arrow corresponds to the movement of $1/16$ parts of mass. Graph **(a)** shows the optimal transport plan π^* under l_2 (or l_1) costs on $X \times Y = \mathbb{R}^2$, while the plans π_1 to π_3 in **(b-d)** correspond to the three upper bounds established in Proposition 6 to 8. The example was chosen such that all bounds are different, and we find $\pi^* c < \pi_2 c < \pi_3 c < \pi_1 c$. The difference between π_1 and π_3 is that the vertical movement of mass on the fiber $x = 4$ is optimal in case of π_3, while it is clearly not for π_1. The plan π_2 corresponds to the best assignment if transport is restricted to vertical and horizontal movements.

measurable map $\varphi\colon X \longrightarrow Y$, this implies that π_x moves all mass from ν to the single point $\varphi(x)$. Consequently, the vertical projection

$$t(x, y) = \big(x, \varphi(x)\big) \tag{27}$$

is the transport map that corresponds to π_1. Indeed, one can check that the two identities $t_*(\mu \otimes \nu) = \gamma$ and $(t, \mathrm{id})_*(\mu \otimes \nu) = \pi_1$ hold in this setting.

While the upper bound $\mathrm{diam}_{c_Y} \nu$ will play an important role later on (see Section 4 and 5), it is apparent that the plan π_1 leading to Proposition 6 is quite crude in general and can likely be improved. The following two results do so in different ways.

Proposition 7: Let $\gamma \in \mathcal{C}(\mu, \nu)$ for $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ in Polish spaces X and Y. If the cost function c on $X \times Y$ satisfies the marginal bounds (23), then

$$\tau(\gamma) \le \mathrm{diam}_{c_{XY}} \gamma \le \min\big(\mathrm{diam}_{c_X} \mu, \mathrm{diam}_{c_Y} \nu\big), \tag{28}$$

where $c_{XY}\colon (X \times Y)^2 \longrightarrow [0, \infty]$ is a cost function defined by

$$c_{XY}(x_1, y_1, x_2, y_2) = \min\big(c_X(x_1, x_2), c_Y(y_1, y_2)\big). \tag{29}$$

The idea of the proof is straightforward and similar to Proposition 6, in that we construct a feasible transport plan $\pi_2 \in \mathcal{C}(\gamma, \mu \otimes \nu)$ by vertical or horizontal movements. This time, however, we do not globally enforce vertical or horizontal movements, but we locally move mass along Y if $c_Y \le c_X$ and along X otherwise. In Appendix A, we show that this idea indeed leads to a valid transport plan that attains the upper bound in (28). Note that it is crucial that both c_X and c_Y are symmetric, as there is no guarantee that the constructed plan π_2 has the correct marginals otherwise. Furthermore, our arguments explicitly rely on the product structure of $\mu \otimes \nu$, as well as on the fact that μ and ν are the marginals of γ. As far as we can see, there is thus no way to generalize the upper bounds in Proposition 6 and 7 to more generic optimal transport costs $T_c(\gamma, \gamma')$ with $\gamma' \neq \mu \otimes \nu$.

Analog to (27), we can again state an explicit transport map for the plan π_2 in case of (suitable) deterministic couplings. If $\gamma = (\mathrm{id}, \varphi)_* \mu$ is induced by an invertible and bimeasurable function φ, one can show that $\pi_2 = (t, \mathrm{id})_*(\mu \otimes \nu)$ for the mapping

$$t(x,y) = \begin{cases} \big(x, \varphi(x)\big) & \text{if} \quad c_Y\big(y, \varphi(x)\big) \leq c_X\big(\varphi^{-1}(y), x\big), \\ \big(\varphi^{-1}(y), y\big) & \text{else.} \end{cases} \tag{30}$$

Marginal transport dependency. Finally, we present a third upper bound of the transport dependency, which corresponds to the *marginal transport dependency*[6]

$$\tau^Y(\gamma) = \tau^Y_{c_Y}(\gamma) = \int T_{c_Y}\big(\gamma(x, \cdot), \nu\big)\, \mu(\mathrm{d}x) \tag{31}$$

that was already mentioned in the introduction. Again, we typically suppress the dependence on c_Y in our notation if the costs are apparent from the context.

> Proposition 8: Let $\gamma \in \mathcal{C}(\mu, \nu)$ for $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ in Polish spaces X and Y. If the cost function c on $X \times Y$ satisfies the marginal bounds (23) for a continuous c_Y, then
>
> $$\tau(\gamma) \leq \tau^Y(\gamma) \leq \mathrm{diam}_{c_Y} \nu. \tag{32}$$
>
> An analog inequality with the roles of (X, μ) and (Y, ν) reversed holds if c_X is continuous.

This result is a natural refinement of Proposition 6: instead of choosing the conditional plans π_x between $\gamma(x, \cdot)$ and ν as product measures, as it is done in (26), they are chosen optimally. Details can be found in Appendix A.

Besides its function as an upper bound, the marginal transport dependency is interesting as a measure of association in its own right. Since τ^Y disregards the cost structure on X, it is a promising candidate to quantify dependency in asymmetric settings where points in X cannot be moved from one to another in a meaningful way, for instance if X is categorical. One of the key characteristic of $\tau^Y(\gamma)$ is that it equals the upper bound $\mathrm{diam}_{c_Y} \nu$ if and only if γ is induced by a function from X to Y (see Appendix A for a proof).

[6] Note that we have to assume that the mapping $x \mapsto T_{c_Y}\big(\gamma(x, \cdot), \nu\big)$ is measurable for this definition to make sense. If c_Y is continuous, measurability is guaranteed by Villani 2008, Corollary 5.22.

Theorem 5 (maximal marginal transport dependency): Let X and Y be Polish spaces and c_Y a positive continuous cost function on Y. For any $\gamma \in \mathcal{C}(\mu, \nu)$ with $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $\operatorname{diam}_{c_Y} \nu < \infty$, the equality

$$\tau^Y(\gamma) = \operatorname{diam}_{c_Y} \nu$$

holds if and only if $\gamma = (\mathrm{id}, \varphi)_* \mu$ for a measurable function $\varphi \colon X \to Y$.

It is worth pointing out that the marginal transport dependency defined in equation (31) is in fact a special case of the general transport dependency τ under costs of the form

$$c_\infty(x_1, y_1, x_2, y_2) = \begin{cases} c_Y(y_1, y_2) & \text{if } x_1 = x_2, \\ \infty & \text{else.} \end{cases}$$

This indicates that τ^Y arises as limit case if movements within the space X become prohibitively expensive, such that all transport eventually withdraws to the fibers $\{x\} \times Y$ for $x \in X$. The following result, which is proved in Appendix A, confirms that this intuition is accurate.

Theorem 6 (marginal transport dependency as limit): Let X and Y be Polish spaces. For $\alpha > 0$, let c_α be a cost function on $X \times Y$ that satisfies

$$c_\alpha(x_1, y_1, x_2, y_2) \begin{cases} = \ c_Y(y_1, y_2) & \text{if } x_1 = x_2, \\ \geq \ \max\big(\alpha\, c_X(x_1, x_2), c_Y(y_1, y_2)\big) & \text{else,} \end{cases}$$

where c_Y is a continuous cost function on Y and c_X a positive cost function on X. Then, for any $\gamma \in \mathcal{P}(X \times Y)$,

$$\lim_{\alpha \to \infty} \tau_{c_\alpha}(\gamma) = \tau_{c_\infty}(\gamma) = \tau^Y_{c_Y}(\gamma).$$

Example 5 (Multivariate Gaussian, part 2): We revisit the Gaussian setting of Example 1. Recall that $(\xi, \zeta) \sim \gamma = \mathcal{N}(\eta, \Sigma)$, where η is a mean vector and Σ a covariance matrix with blocks Σ_{11}, Σ_{12}, $\Sigma_{21} = \Sigma_{12}^T$ and Σ_{22}. This time, we work with costs of the form

$$c_\alpha(x_1, y_1, x_2, y_2) = \alpha\, \|x_1 - x_2\|^2 + \|y_1 - y_2\|^2 \tag{33}$$

for $\alpha > 0$. Since these costs can be interpreted as scaling ξ by the factor $\sqrt{\alpha}$ under the usual squared Euclidean distance c_1, we can employ the same arguments as in Example 1 and find expressions for τ_{c_α} via replacing Σ_{11} by $\alpha\, \Sigma_{11}$ and Σ_{12} by $\sqrt{\alpha}\, \Sigma_{12}$. Adapting equation (12) in this way yields

$$\tau_{c_\alpha}(\gamma) = 2\alpha \operatorname{trace}(\Sigma_{11}) + 2 \operatorname{trace}(\Sigma_{22}) - 2 \operatorname{trace}\left(\begin{pmatrix} \alpha^2 \Sigma_{11}^2 & \alpha^{3/2}\, \Sigma_{11}\Sigma_{12} \\ \alpha^{1/2}\, \Sigma_{22}\Sigma_{21} & \Sigma_{22}^2 \end{pmatrix}^{1/2}\right).$$

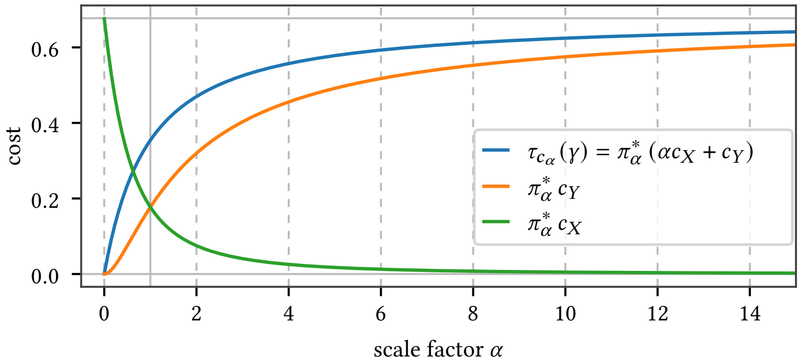


Figure 7: Transport dependency under increasing values of α in the setting of Example 5. The costs are given as weighted squared Euclidean distances of the form $c_\alpha = \alpha c_X + c_Y$, see equation (33). Shown is the actual transport dependency $\tau_{c_\alpha}(\gamma)$ in the bivariate normal case for $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.75$ (with optimal transport plan π_α^*), as well as the average vertical ($\pi_\alpha^* c_Y$) and horizontal ($\pi_\alpha^* c_X$) transport cost. For $\alpha = 0$, transport along X is free and thus governs π_0^*. For $\alpha \longrightarrow \infty$, movements along X become prohibitively expensive and all mass is eventually transported exclusively along Y (as predicted by Theorem 6).

According to Theorem 6, the right hand side converges to the marginal transport dependency $\tau_{c_Y}^Y(\gamma)$ as $\alpha \longrightarrow \infty$, where c_Y is the squared Euclidean distance. In the bivariate case (13), this limit can readily be calculated and reads

$$\begin{aligned}\tau_{c_Y}^Y(\gamma) &= \lim_{\alpha\to\infty} \tau_{c_\alpha}(\gamma) \\ &= \lim_{\alpha\to\infty} 2\left(\alpha\,\sigma_1^2 + \sigma_2^2 - \sqrt{\alpha^2\sigma_1^4 + \sigma_2^4 + 2\alpha\sigma_1^2\sigma_2^2\sqrt{1-\rho^2}}\right) \\ &= 2\sigma_2^2\left(1 - \sqrt{1-\rho^2}\right). \qquad (34)\end{aligned}$$

Figure 7 displays the average transport of mass occurring in this setting as a function of α.

Note that it is not straightforward to estimate the marginal transport dependency from empirical data $\hat{\gamma}_n$. Since the costs c_∞ are not continuous, we cannot apply Proposition 4 and thus cannot expect consistency of $\tau^Y(\hat{\gamma}_n)$. In fact, a similar obstacle that affects other quantifiers of unstructured dependency, like the mutual information, also arises here: in settings where μ is diffuse, empirical data $\hat{\gamma}_n$ with marginals $\hat{\mu}_n$ and $\hat{\nu}_n$ likely obeys functional relations $\hat{\gamma}_n = (\mathrm{id}, \hat{\varphi}_n)_*\hat{\mu}_n$, in which case $\tau^Y(\hat{\gamma}_n)$ is not informative (since it always equals $\mathrm{diam}_{c_Y}\hat{\nu}_n$ via Theorem 5).

A usual counter measure in this situation is to preprocess $\hat{\gamma}_n$. In Euclidean settings, for example, it is common to first estimate a smooth density or to bin the data. In both cases, one would typically pick a kernel size or bin width h_n that decreases to 0 as n grows. If h_n is chosen suitably for the setting at hand, one obtains a consistent estimator (see Tsybakov 2008). In a certain sense, the limiting procedure of $\alpha \to \infty$ in Theorem 6 has a similar effect: a finite $\alpha < \infty$ allows for some leeway along the space X when matching $\hat{\gamma}_n$ with $\hat{\mu}_n \otimes \hat{\nu}_n$, and this leeway becomes progressively smaller as α increases. Indeed, it is possible to find suitable sequences $\alpha_n \to \infty$ so that $\tau_{c_{\alpha_n}}(\hat{\gamma}_n)$ is a consistent estimator of $\tau^Y(\gamma)$ under mild assumptions. The following application of Theorem 6 serves as a proof of concept for this idea, even though the choice of α_n can likely be improved.

Example 6 (estimation of marginal transport dependency): Let (X, d_X) and (Y, d_Y) be Polish metric spaces, $c_Y = d_Y^p$, and $c_\alpha = (\alpha d_X + d_Y)^p$ for $p \geq 1$. Moreover, assume $\gamma \in \mathcal{P}(X \times Y)$ to have a finite p-moment with respect to the metric $d = d_X + d_Y$ (which in particular implies $\tau_{c_\alpha}(\gamma) < \infty$ for all $0 < \alpha \leq \infty$) and let $(\alpha_n)_{n \in \mathbb{N}}$ be a diverging sequence. We observe that

$$\mathbb{E}\left[\left|\tau_{c_{\alpha_n}}(\hat{\gamma}_n)^{1/p} - \tau_{c_Y}^Y(\gamma)^{1/p}\right|\right] \leq \\ \mathbb{E}\left[\left|\tau_{c_{\alpha_n}}(\hat{\gamma}_n)^{1/p} - \tau_{c_{\alpha_n}}(\gamma)^{1/p}\right|\right] + \left|\tau_{c_{\alpha_n}}(\gamma)^{1/p} - \tau_{c_Y}^Y(\gamma)^{1/p}\right|.$$

By Theorem 6, the second summand converges to zero as $n \to \infty$. The first summand can be controlled by Theorem 3 and bound (22), yielding

$$\mathbb{E}\left[\left|\tau_{c_{\alpha_n}}(\hat{\gamma}_n)^{1/p} - \tau_{c_{\alpha_n}}(\gamma)^{1/p}\right|\right] \leq 3\,\mathbb{E}\left[T_{c_{\alpha_n}}(\gamma, \hat{\gamma}_n)^{1/p}\right] \leq 3\,\alpha_n\,\mathbb{E}\left[T_{c_1}(\gamma, \hat{\gamma}_n)^{1/p}\right].$$

This reveals that $\tau_{c_{\alpha_n}}(\hat{\gamma}_n)$ is a consistent estimator of $\tau_{c_Y}^Y(\gamma)$ under the assumption

$$\alpha_n = o\left(\mathbb{E}\left[T_{c_1}(\gamma, \hat{\gamma}_n)^{1/p}\right]^{-1}\right).$$

One can therefore use the asymptotic convergence rates discussed in the context of equation (22) to find a suitable sequence $(\alpha_n)_{n \in \mathbb{N}}$.

Another route to consistently estimate the marginal transport dependency from data was recently proposed in Wiesel 2021, who, independently from our work, also considered expressions of the form (31). He derives convergence rates if $\tau^Y(\gamma)$ is estimated by $\tau^Y(\tilde{\gamma}_n)$ for a so-called adapted empirical measure $\tilde{\gamma}_n$. One example for such an adaption is the projection of the individual observations ξ_i and ζ_i to a grid in the unit cube (see Backhoff et al. 2022).

4 Contractions

Up to this point, we were concerned with universal properties of the transport dependency for generic cost functions. We now focus on a particular additive cost structure that enables us to characterize under which conditions the bounds of the previous section are sharp.

In the following, we equip the Polish space X with a (general) cost function k_X and the Polish space Y with a lower semi-continuous (pseudo-)metric d_Y. We consider costs of the form

$$c(x_1, y_1, x_2, y_2) = h\big(k_X(x_1, x_2) + d_Y(y_1, y_2)\big) \tag{35a}$$

for $(x_1, y_1), (x_2, y_2) \in X \times Y$, where $h\colon [0, \infty) \to [0, \infty)$ is a strictly increasing function that satisfies $h(0) = 0$. We also fix the marginal costs

$$c_X = h \circ k_X \qquad \text{and} \qquad c_Y = h \circ d_Y, \tag{35b}$$

which fulfill condition (23). Therefore, c_X and c_Y are suited for application in the previously derived upper bounds (Proposition 6 to 8).

To state our findings, we first define couplings that describe contractions between X and Y. For given d_Y and k_X as above, we say that $\gamma \in \mathcal{P}(X \times Y)$ is *contracting* (on its support) if

$$d_Y(y_1, y_2) \le k_X(x_1, x_2) \qquad \text{for all} \qquad (x_1, y_1), (x_2, y_2) \in \operatorname{supp}\gamma. \tag{36}$$

We also need a slightly weaker condition and say that γ is *almost surely contracting* if (36) holds $(\gamma \otimes \gamma)$-almost surely (but not necessarily on each single pair of points of the support). Our next result characterizes contracting couplings as those that maximize the transport dependency for fixed marginals and attain the upper bound in Proposition 6.

> Theorem 7 (contracting couplings): Let X and Y be Polish spaces, c be a cost function on $X \times Y$ of the form (35), and $\gamma \in \mathcal{C}(\cdot, \nu) \subset \mathcal{P}(X \times Y)$ for $\nu \in \mathcal{P}(Y)$. If γ is contracting,
>
> $$\tau(\gamma) = \operatorname{diam}_{c_Y} \nu. \tag{37}$$
>
> Conversely, if $\operatorname{diam}_{c_Y} \nu < \infty$ and (37) holds, then γ is almost surely contracting.

The second assertion in this statement is a direct consequence of the upper bound in Proposition 7. For the first claim, we use disintegration to insert a suitable auxiliary variable into the integrals πc for arbitrary $\pi \in \mathcal{C}(\gamma, \mu \otimes \nu)$, after which we can exploit $d_Y \le k_X$ as well as the triangle inequality of d_Y. This line of argumentation reveals that vertical movements are indeed globally optimal when transporting to a contracting coupling γ – an idea that was already anticipated in Figure 3 of the introduction. Details of the proof are documented in Appendix A.

Note that the distinction between “contracting” and “almost surely contracting” in Theorem 7 is necessary under the stated generality. Indeed, one can construct almost surely contracting couplings γ such that, e.g., $\tau(\gamma) = 0$ and $\operatorname{diam}_{c_Y} \nu > 0$. This can only be done, however, if the costs are not continuous. If d_Y and k_X are continuous, then every almost surely contracting coupling is also contracting on its support, and the two concepts coincide (see Lemma 7 in Appendix A). In this case, Theorem 7 implies the equivalence

$$\gamma \in \mathcal{C}(\cdot, \nu) \text{ is contracting} \qquad \Longleftrightarrow \qquad \tau(\gamma) = \operatorname{diam}_{c_Y} \nu.$$

Another crucial observation is that contracting couplings are, under mild assumptions, always deterministic. In fact, if d_Y is a proper metric and h is continuous, Proposition 8 and Theorem 5 show that any γ satisfying (37) is induced by a function. Furthermore, if a measurable function $\varphi\colon X \longrightarrow Y$ is *μ-almost surely contracting*, meaning that there is a Borel set $A \subset X$ with $\mu(A) = 1$ and

$$d_Y\big(\varphi(x_1), \varphi(x_2)\big) \le k_X(x_1, x_2) \qquad \text{for all} \qquad x_1, x_2 \in A, \tag{38}$$

then $\gamma = (\mathrm{id}, \varphi)_* \mu \in \mathcal{C}(\mu, \nu)$ is contracting and (37) holds under assumptions of continuity. The following corollary collects these observations into a useful statement.

Corollary 1 (contracting deterministic couplings): Let X and Y be Polish spaces and c be a cost function on $X \times Y$ of the form (35), where h and k_X are continuous and d_Y is a continuous metric. Let $\gamma \in \mathcal{C}(\mu, \nu)$ with $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $\operatorname{diam}_{c_Y} \nu < \infty$. Then

$$\tau(\gamma) = \operatorname{diam}_{c_Y} \nu$$

holds if and only if $\gamma = (\mathrm{id}, \varphi)_* \mu$ for a μ-almost surely contracting function $\varphi\colon X \longrightarrow Y$.

Proof of Corollary 1. Let $\varphi\colon X \longrightarrow Y$ be μ-almost surely contracting such that $\gamma = (\mathrm{id}, \varphi)_* \mu$. By a change of variables,

$$(\gamma \otimes \gamma)(d_Y \le k_X) = (\mu \otimes \mu)(d_\varphi \le k_X) = 1, \tag{39}$$

where $d_\varphi(x_1, x_2) = d_Y\big(\varphi(x_1), \varphi(x_2)\big)$ for any $x_1, x_2 \in X$. Therefore, γ is almost surely contracting. Since k_X and d_Y are continuous, we can apply Lemma 7 to conclude that γ is contracting on its support. By Theorem 7, $\tau(\gamma) = \operatorname{diam}_{c_Y} \nu$ follows.

For the reverse direction, we consult the upper bound in Proposition 8 to find that $\tau(\gamma) = \operatorname{diam}_{c_Y} \nu$ implies $\tau^Y(\gamma) = \operatorname{diam}_{c_Y} \nu$. Consequently, Theorem 5 establishes the existence of a measurable function $\varphi\colon X \longrightarrow Y$ with $\gamma = (\mathrm{id}, \varphi)_* \mu$. It is left to show that φ is μ-almost surely contracting. According to Lemma 7 and Theorem 7, γ is contracting on its support. We can therefore consider the set $A = (\mathrm{id}, \varphi)^{-1}(\operatorname{supp} \gamma)$ and conclude that both $\mu(A) = \gamma(\operatorname{supp} \gamma) = 1$ and

$$d_Y\big(\varphi(x_1), \varphi(x_2)\big) \le d_X(x_1, x_2) \qquad \text{for all} \qquad x_1, x_2 \in A,$$

since all tuples $\big(x, \varphi(x)\big) \in X \times Y$ for $x \in A$ are elements of the support of γ. □

A point that deserves emphasis is that Corollary 1 actually provides a characterization of Lipschitz and Hölder functions, as well as of isometries. If k_X is set to $\alpha \cdot d_X$ for a metric d_X on X, then condition (38) is equivalent to φ being α-Lipschitz μ-almost surely. Similarly, under the choice $k_X = \alpha \cdot d_X^\beta$ for $\beta \in (0, 1)$, condition (38) coincides with the β-Hölder criterion for φ (with constant α). Finally, the simultaneous equality

$$\tau(\gamma) = \operatorname{diam}_{c_X}\mu = \operatorname{diam}_{c_Y}\nu$$

for $k_X = d_X$ holds if and only if φ is μ-almost surely an isometry from (X, d_X) to (Y, d_Y). In this context, it might be interesting to note that a μ-almost sure contraction φ in the setting of Corollary 1 can always be uniformly extended to the full support of μ if (Y, d_Y) is a Polish metric space (see Lemma 8 in Appendix A). In other words: if d_Y completely metrizes the topological space Y, then we can always assume $A = \operatorname{supp}\mu$ in definition (38). In particular, the map φ in Corollary 1 can be chosen to be continuous on the support of μ.

To conclude this section, we want to highlight that the preceding results equip the transport dependency with a powerful interpretation as quantifier of structural dependence: the larger the transport dependency $\tau(\xi, \zeta)$ between two random variables ξ and ζ is, the more they have to be associated in a contracting manner. Intuitively, this means that the conditional law of $\zeta \mid \xi = x$ must behave well as a function of $x \in X$, judged in terms of k_X and d_Y. In fact, the highest possible degree of transport dependency for fixed marginals is (under continuity assumptions) only assumed if ξ and ζ are deterministically related by $\zeta = \varphi(\xi)$ for a contraction φ. Other deterministic relations between ζ and ξ, which exhibit rapid changes that break condition (38), are assigned a lower degree of dependency. This way of dependency quantification might often be desirable, especially in situations where quickly oscillating or chaotic relations between ξ and ζ practically cannot (or should not) be distinguished from actual noise.

5 Transport correlation

In this section, we introduce several coefficients of association that are based on the transport dependency. The central ingredient is upper bound (24) in Proposition 6, which can be used to scale τ to the interval $[0, 1]$ in a way that only depends on the marginal distributions. Without further assumptions, however, bound (24) is not necessarily sharp, and values close to 1 may be impossible (see Figure 8 for an illustrative example in this regard). In the following, we therefore focus on cost structures of the form (35a). In this setting, the conditions under which τ assumes its upper bounds are well understood (Theorem 7 and Corollary 1 in Section 4), which makes the resulting coefficients more expressive.

For the sake of clarity, we work in a slightly less general setting than in Section 4 and right away assume (X, d_X) and (Y, d_Y) to be Polish metric spaces, restricting to costs

$$c(x_1, y_1, x_2, y_2) = \big(\alpha \cdot d_X(x_1, x_2) + d_Y(y_1, y_2)\big)^p \tag{40}$$

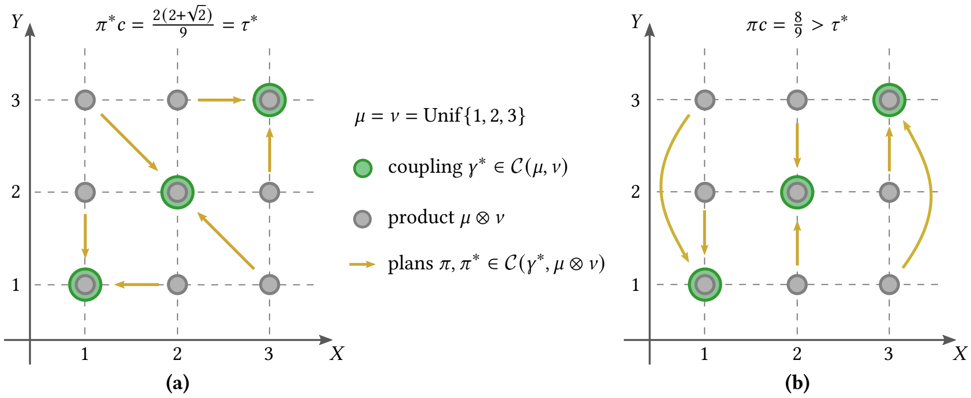

$\pi^* c = \frac{2(2+\sqrt{2})}{9} = \tau^*$
Y
3
2
1
1
2
3
X
(a)
$\mu = \nu = \mathrm{Unif}\{1, 2, 3\}$
coupling $\gamma^* \in \mathcal{C}(\mu, \nu)$
product $\mu \otimes \nu$
plans $\pi, \pi^* \in \mathcal{C}(\gamma^*, \mu \otimes \nu)$
$\pi c = \frac{8}{9} > \tau^*$
Y
3
2
1
1
2
3
X
(b)

Figure 8: Example for which the upper bounds (24), (28), and (32) are not sharp. The marginal distributions are $\mu = \nu = \mathrm{Unif}\{1, 2, 3\}$. **(a)** One can show that $\gamma^* = \mathrm{Unif}\{(1, 1), (2, 2), (3, 3)\}$ maximizes $\tau(\gamma)$ over $\gamma \in \mathcal{C}(\mu, \nu)$, assuming the value $\tau^* = 2(2 + \sqrt{2})/9$ if c is the Euclidean distance on $\mathbb{R}^2$. The visualized plan π^* is optimal. **(b)** All of the established upper bounds move mass along a plan π that is restricted to vertical or horizontal transports. The total transportation cost is $8/9 > \tau^*$. Note that the Euclidean distance on $X \times Y$ can *not* be expressed in the additive form (35a).

for $x_1, x_2 \in X$, $y_1, y_2 \in Y$, and $\alpha, p > 0$. In what follows, d_X, d_Y, and p are usually considered to be fixed, and we mainly explore the influence of α. We call a map $\varphi\colon X \longrightarrow Y$ a *dilatation* if there exists some $\beta > 0$ such that

$$d_Y\big(\varphi(x_1), \varphi(x_2)\big) = \beta\, d_X(x_1, x_2)$$

for all $x_1, x_2 \in X$. A dilatation can be thought of as an isometry from $(X, \beta d_X)$ to (Y, d_Y), i.e., an isometry up to the correct scaling. We also recall the notion of p-weak convergence on $\mathcal{P}_p(X \times Y)$, which was introduced and discussed in the context of continuity (Section 3).

For any $\gamma \in \mathcal{C}(\cdot, \nu)$ with $0 < \mathrm{diam}_{d_Y^p}\, \nu < \infty$, we define the *$\alpha$-transport correlation* via

$$\rho_\alpha(\gamma) = \left(\frac{\tau(\gamma)}{\mathrm{diam}_{d_Y^p}\, \nu}\right)^{1/p}. \tag{41}$$

This dependency coefficient has a series of interesting properties, most of which are direct consequences of our previous findings. For detailed proofs, see Appendix A.

Proposition 9 (α-transport correlation): Let (X, d_X) and (Y, d_Y) be Polish metric spaces. For any $\gamma \in \mathcal{C}(\mu, \nu)$ with $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $0 < \mathrm{diam}_{d_Y^p}\, \nu < \infty$, the α-transport correlation defined in (41) for $0 < \alpha < \infty$ has the following properties:

1. $\rho_\alpha(\gamma) = 0$ iff $\gamma = \mu \otimes \nu$,
2. $\rho_\alpha(\gamma) = 1$ iff $\gamma = (\mathrm{id}, \varphi)_* \mu$ for an α-Lipschitz function $\varphi : X \to Y$,

3. $\rho_\alpha(\gamma) = \rho_\alpha(\gamma')$ if $\gamma' = (f_X, f_Y)_* \gamma$ for isometries $f_X \colon X \to X$ and $f_Y \colon Y \to Y$,
4. $\gamma \mapsto \rho_\alpha(\gamma)^p$ is convex when restricted to $\mathcal{C}(\cdot, \nu) \subset \mathcal{P}(X \times Y)$ for fixed ν,
5. $\rho_{\alpha_n}(\gamma_n) \to \rho_\alpha(\gamma)$ as $n \to \infty$ if $(\gamma_n)_{n \in \mathbb{N}}$ converges p-weakly to γ and $\alpha_n \to \alpha$,
6. $\alpha \mapsto \rho_\alpha(\gamma)^p$ is monotonically increasing for all $p > 0$ and concave if $p \leq 1$,

where the functions φ, f_X, and f_Y only have to be defined μ- or ν-almost surely.

Properties 1 and 2 lend the transport correlation a distinctive interpretation as dependency coefficient that identifies Lipschitz relations between random variables ξ and ζ. This makes ρ_α a flexible tool for quantifying the degree of association between ξ and ζ, especially when we have a clear picture of how structured a deterministic relation φ with $\zeta = \varphi(\xi)$ should at least be in order to be distinguished from noise. Properties 3 to 5 state that ρ_α benefits from the general properties of the transport dependency τ. For example, property 5 implies that $\rho_\alpha(\gamma)$ can be estimated consistently just by plugging the empirical measure $\hat{\gamma}_n$ in definition (41).

Based on our earlier findings in Theorem 5 and 6, it is possible to extend $\rho_\alpha(\gamma)$ to $\alpha = \infty$ as a special case. Effectively, this means employing the marginal transport dependency (31) in equation (41), which leads us to define the ∞-*transport correlation* or *marginal transport correlation*

$$\rho_\infty(\gamma) = \left(\frac{\tau^Y_{d_Y^p}(\gamma)}{\operatorname{diam}_{d_Y^p} \nu} \right)^{1/p}. \tag{42}$$

Proposition 10 (marginal transport correlation): Let X and (Y, d_Y) be Polish (metric) spaces. For any $\gamma \in \mathcal{C}(\mu, \nu)$ with $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $0 < \operatorname{diam}_{d_Y^p} \nu < \infty$, the marginal transport correlation defined in (42) has the following properties:

1. $\rho_\infty(\gamma) = 0$ iff $\gamma = \mu \otimes \nu$,
2. $\rho_\infty(\gamma) = 1$ iff $\gamma = (\mathrm{id}, \varphi)_* \mu$ for a measurable function $\varphi : X \to Y$,
3. $\rho_\infty(\gamma) = \rho_\infty(\gamma')$ if $\gamma' = (f_X, f_Y)_* \gamma$ for a measurable injection $f_X \colon X \to X$ and a dilatation $f_Y \colon Y \to Y$,
4. $\gamma \mapsto \rho_\infty(\gamma)^p$ is convex when restricted to $\mathcal{C}(\cdot, \nu) \subset \mathcal{P}(X \times Y)$ for fixed ν,

where the functions φ, f_X, and f_Y only have to be defined μ- or ν-almost surely.

The differences in Proposition 9 and 10 reflect that ρ_α progressively loses its sense for the metrical structure of X when α is increased. In the limit $\alpha = \infty$, the Lipschitz restrictions appearing in Proposition 9 are dissolved and only conditions of measurability remain.

Since we normalize with the diameter of ν instead of the one of μ in definition (41), the transport correlation ρ_α singles out relations in the direction $X \longrightarrow Y$. One possibility to compensate for this asymmetry is to adapt the value of α by choosing

$$\alpha_* = \left(\operatorname{diam}_{d_Y^p} \nu / \operatorname{diam}_{d_X^p} \mu\right)^{1/p}. \tag{43}$$

The coefficient $\rho_* = \rho_{\alpha_*}$, which we call *isometric transport correlation*, takes the form

$$\rho_*(\gamma) = \left(\tau_{c_*}(\gamma)\right)^{1/p}, \tag{44}$$

where

$$c_*(x_1, y_1, x_2, y_2) = \left(\frac{d_X(x_1, x_2)}{\left(\operatorname{diam}_{d_X^p} \mu\right)^{1/p}} + \frac{d_Y(y_1, y_2)}{\left(\operatorname{diam}_{d_Y^p} \nu\right)^{1/p}} \right)^p$$

for $x_1, x_2 \in X$ and $y_1, y_2 \in Y$. We can interpret (44) as first normalizing the metric measure spaces (X, d_X, μ) and (Y, d_Y, ν) by their p-diameters before calculating the transport dependency.

Proposition 11 (isometric transport correlation): Let (X, d_X) and (Y, d_Y) be Polish metric spaces. For any $\gamma \in \mathcal{C}(\mu, \nu)$ with $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ such that $0 < \alpha_* < \infty$ in (43) is well defined, the isometric transport correlation (44) has the following properties:

1. $\rho_*(\gamma) = 0$ iff $\gamma = \mu \otimes \nu$,
2. $\rho_*(\gamma) = 1$ iff $\gamma = (\mathrm{id}, \varphi)_*\mu$ or $\gamma = (\psi, \mathrm{id})_*\nu$ for dilatations $\varphi : X \longrightarrow Y$ or $\psi : Y \longrightarrow X$,
3. $\rho_*(\gamma) = \rho_*(\gamma')$ if $\gamma' = (f_X, f_Y)_*\gamma$ for dilatations $f_X\colon X \longrightarrow X$ and $f_Y\colon Y \longrightarrow Y$,
4. $\gamma \longmapsto \rho_*(\gamma)^p$ is convex restricted to $\mathcal{C}(\mu, \nu) \subset \mathcal{P}(X \times Y)$ for fixed μ and ν,
5. $\rho_*(\gamma_n) \longrightarrow \rho_*(\gamma)$ as $n \longrightarrow \infty$ if $(\gamma_n)_{n\in\mathbb{N}}$ converges p-weakly to γ,
6. $\rho_*(\gamma) = \rho_*(\gamma')$ if $\gamma' = f_*\gamma$ for the symmetry map $f(x, y) = (y, x)$,

where the functions φ, ψ, f_X, and f_Y only have to be defined μ- or ν-almost surely.

Note that the dilatation φ in property 2 above actually has to be an isometry from $(X, \alpha_* d_X)$ to (Y, d_Y), i.e., the correct scaling is given by $\beta = \alpha_*$ such that $\operatorname{diam}_{(\alpha_* d_X)^p} \mu = \operatorname{diam}_{d_Y^p} \nu$.

To conclude, we want to mention another symmetric coefficient of association which can be built on top of the transport dependency. Instead of dividing by the diameter of ν in (41), we divide by the minimum of the diameters of ν and μ. Setting $\alpha = 1$ in (40), this results in the coefficient

$$\left(\frac{\tau_c(\gamma)}{\min\left(\operatorname{diam}_{d_X^p} \mu, \operatorname{diam}_{d_Y^p} \nu\right)} \right)^{1/p}. \tag{45}$$

This coefficient has already been introduced as *Earth mover's correlation* and is studied in Móri and Székely 2020 for $p = 1$. It enjoys similar properties to the isometric transport correlation. In Conjecture 3 of Móri and Székely 2020, it is even hypothesized that the expression in (45) might equal 1 if and only if $\gamma = (\mathrm{id}, \phi)_*\mu$ for a dilatation $\phi : X \longrightarrow Y$. In general, however, Corollary 1 shows that this is not correct and that the Earth mover's coefficient assumes the value 1 if and only if γ describes a 1-Lipschitz function from X to Y *or* a 1-Lipschitz function from Y to X.

6 Applications

To assess the actual usefulness of the transport dependency for applications, we next investigate the performance of the coefficients ρ_α and ρ_* on empirical data. In particular, we conduct a series of benchmarks that illuminate basic properties and demonstrate commonalities and differences to other commonly used coefficients of association. In what follows, we restrict our attention to the choice $p = 1$ in the cost structure (40); additional simulations that cover $p = 1/2$ and $p = 2$ and that adopt alternative dependency models can be found in Appendix B.

Setting. We consider Euclidean spaces $X = \mathbb{R}^r$ and $Y = \mathbb{R}^q$ for $r, q \in \mathbb{N}$, and equip them with their respective Euclidean metrics d_X and d_Y. We focus on joint distributions $\gamma \in \mathcal{P}(\mathbb{R}^r \times \mathbb{R}^q)$ that are typically either given deterministically via $\gamma = (\mathrm{id}, \varphi)_*\mu$ for $\varphi\colon [0,1]^r \longrightarrow [0,1]^q$ and $\mu = \mathrm{Unif}[0,1]^r$ (i.e., concentrated on the graph of a function), or by placing (uniform) mass on more general shapes in $[0,1]^{r+q}$. The number of samples independently drawn from γ for the purpose of empirical estimation is denoted by $n \in \mathbb{N}$, and we write $\hat{\gamma}_n$ to refer to the corresponding empirical measure, see equation (19). To study the influence of statistical noise, we consider convex contamination models of the form

$$\gamma^\epsilon = (1-\epsilon)\,\gamma + \epsilon\,(\mu \otimes \nu), \tag{46}$$

where $\epsilon \in [0,1]$ denotes the noise level distorting $\gamma \in \mathcal{C}(\mu,\nu)$. Observations $(x, y) \in \mathbb{R}^r \times \mathbb{R}^q$ sampled from γ^ϵ are with probability $(1-\epsilon)$ randomly drawn from γ and with probability ϵ randomly drawn from $\mu \otimes \nu$. Further simulations that use additive Gaussian noise instead are provided in Appendix B.

Our numerical study compares different dependency coefficients that assume values in $[0,1]$. Besides the isometric and α-Lipschitz transport correlations ρ_* and ρ_α for $\alpha > 0$, which were introduced in the previous section, we consider the following commonly applied quantities:

cor the Pearson correlation. It is only applicable if $r = q = 1$. Since it assumes values in $[-1,1]$, we always report its absolute value.

spe the Spearman rank correlation coefficient. It is only applicable if $r = q = 1$ and is comparable to Kendall's τ, another popular rank based correlation coefficient.

dcor the Euclidean distance correlation (Székely et al. 2007), based on the distance covariance defined in equation (3). It is applicable for all $r, q \in \mathbb{N}$. In its generalized form (4), it is also applicable in generic (separable) metric spaces of (strong) negative type. As discussed previously, several of its properties make it comparable to ρ_*.

mic the maximal information coefficient (Reshef et al. 2011). It is only applicable if $r = q = 1$. For its estimation, we use the estimator mic_e (Reshef et al. 2016), which we compute via the tools provided by Albanese et al. 2012. The two algorithmic parameters c and α were set to 5 and 0.75, respectively (as recommended in Albanese et al. 2018).

For each of these coefficients, generically called ρ for the moment, we are interested in several features. Apart from the actual value of $\rho(\gamma)$, which signifies the amount of dependency attributed to γ, we look at the variance and bias of $\rho(\hat{\gamma}_n)$ as an estimator of $\rho(\gamma)$ when data is limited. To check how well the coefficients are able to distinguish structure from noise, we also include the results of permutation tests for independence (see Lehmann and Romano 2006, Section 15.2 or Janssen and Pauls 2003 for background on permutation based tests).

The ρ-based permutation test we employ works as follows: for given data $z = (x_i, y_i)_{i=1}^n \in (X \times Y)^n$, assumed to be sampled from $\gamma^{\otimes n}$, we write $\rho(z)$ for the empirical estimate of $\rho(\gamma)$ based on z. Furthermore, we denote $z_\sigma = (x_i, y_{\sigma(i)})_{i=1}^n$, where σ is a permutation of n elements. For the test, we randomly select m permutations $\sigma_1, \ldots, \sigma_m$ and reject the null hypothesis that z is sampled from an independent coupling $\gamma = \mu \otimes \nu$ if

$$\left|\left\{i : \rho(z_{\sigma_i}) > \rho(z)\right\}\right| \leq k \tag{47}$$

for some $k \in \{0, \ldots, m\}$. Since a permutation of the second components does not affect the distribution of $\gamma^{\otimes n}$ if $\gamma = \mu \otimes \nu$ is a product coupling, this leads to a level $(k+1)/(m+1)$ test. In all of our applications, we choose k and m such that this level is ≤ 0.1. Note that the power of the permutation test will usually increase if m is increased while the level is held constant.

Computation. One issue that seriously affects our numerical work is the computational effort when solving optimal transport problems. In order to calculate the transport dependency based on n data points, we are faced with finding an optimal coupling with $n \times n^2$ entries. For generic algorithms, for example the network simplex, this implies a runtime of $O(n^5)$ (up to logarithmic factors; see Peyré and Cuturi 2019 for more details on computational aspects). For this reason, we have to choose moderate sample sizes, like $n = 50$, when the transport correlation is evaluated repeatedly (i.e., thousands of times) for Monte Carlo simulations. We also restrict the number of permutations to $m = 29$ when testing under the transport correlation. For all other coefficients, which can be calculated much faster, $m = 999$ is chosen.

In settings with $n \leq 100$, we make use of solvers based on the network simplex provided by the python optimal transport (POT) package (Flamary and Courty 2017). To still be able to capture the behavior of ρ_α and ρ_* for larger values of n, we additionally employ our own implementation of an approximation scheme recently proposed by Schmitzer 2019. It is based on the Sinkhorn algorithm for entropically regularized optimal transport (Cuturi 2013) and operates by successively decreasing the regularization constant η until a suitable approximation of the non-regularized problem is obtained. While η is scaled down, in our case from $\eta = 10^{-1}$ to $\eta = 10^{-3}$, the increasing sparsity of the transport plan (after entries with negligible values $\leq 10^{-15}$ are removed) can be exploited to boost performance by using sparsity-optimized data structures. On systems equipped with a modern GPU, we can thereby calculate the transport dependency for generic costs accurately for sample sizes up to $n = 1000$ (as is done in Figure 13) within a couple of seconds to minutes. Note, however, that we do not make use of the multi-resolution ansatz described by Schmitzer 2019, which might open the door to still larger sample sizes. Another way to ameliorate computational barriers is to restrict to data on grids (see, e.g., Schrieber et al. 2016 for benchmarks of different optimal transport algorithms applied to images) or to utilize resampling techniques (Sommerfeld et al. 2019).

Recognizing shapes. We begin with an investigation of the behavior of the afore mentioned coefficients in two dimensional settings, meaning $r = q = 1$. Figure 9 contains box plots depicting the coefficients' performance on simple geometries, like lines or circles, for $n = 50$ samples. As a point of reference, we also include uniform noise on $[0, 1]^2$.

To showcase the Lipschitz-selectivity of ρ_α, we chose to include the coefficient ρ_3 for $\alpha = 3$. Recall that this implies $\rho_3(\gamma) = 1$ whenever γ is concentrated on the graph of a function whose slope is at most 3. Figure 9 confirms this property of ρ_3, which is the only coefficient to assign maximal dependency to the zigzag function with slope 3 and the polynomial. On the zigzag function with 5 segments (and thus slope 5), it has already decreased to about 0.7. In this context, the mic, which generally achieves high values on all geometries, performs notably well. It is also the coefficient that most clearly distinguishes the pretzel example from noise. Another noteworthy observation is that ρ_* and dcor indeed behave comparably, especially on functional relations. On non-functional patterns, like the circle or the cross, ρ_* assumes somewhat higher values than dcor. At the same time, the bias of ρ_* on independent noise (for which values of 0 are expected in the limit of large n) is slightly higher than the one of dcor, albeit with a smaller variance. Finally, as to be expected, the Pearson and Spearman correlation coefficients do a poor job at discerning non-monotonic structures. In case of the circle, for example, the values of these coefficients are systematically lower than when confronted with random noise.

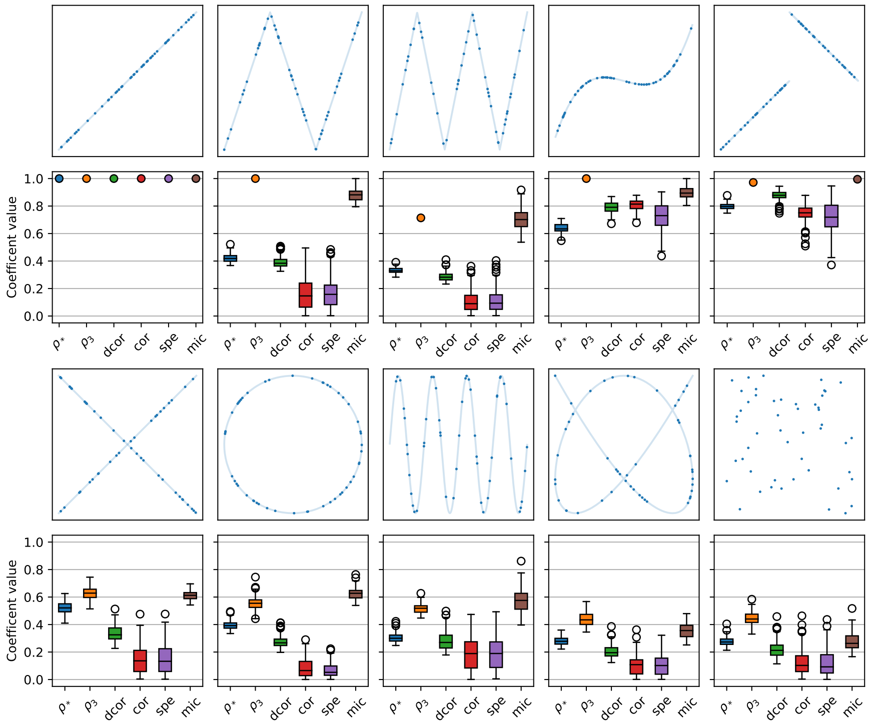


Figure 9: Comparison of several dependency coefficients on two dimensional geometries. As reference, independent $\mathrm{Unif}[0,1]^2$ noise has also been simulated (bottom right). For each geometry, the scatter plot on top displays an exemplary sample of size $n = 50$. The box plots below summarize the values of the empirically estimated coefficients based on 100 such samples. For the sake of visualization, boxes in the box plot are replaced by single dots if their extent would be smaller than 0.1.

Behavior under noise. Our next simulations concern the performance under convex noise models γ^ϵ as defined in equation (46). Figure 10 and 11 illustrate the coefficients' empirical estimates and their power when used for independence testing in case that γ is deterministically given in terms of the identity (Figure 10) or the zigzag function with maximal slope 5 (Figure 11). In Appendix B, this type of comparison can be found for other distributions considered in Figure 9 as well, and we also present results under additive Gaussian noise.

In case of the identity, all coefficients seem to behave roughly similar, especially for small noise levels. Under pure noise ($\epsilon = 1$), for which the coefficients should attain the value 0 in the limit of large n, the transport correlation exhibits a comparably large bias at a relatively small variance. This trend of high biases becomes even more serious in higher dimensions and is further investigated below (Figure 13). Regarding the test performance, the power curves in Figure 10 reveal that all coefficients except ρ_* and mic perform comparably. The power of ρ_* is consistently higher than its competitors', while mic performs notably worse.

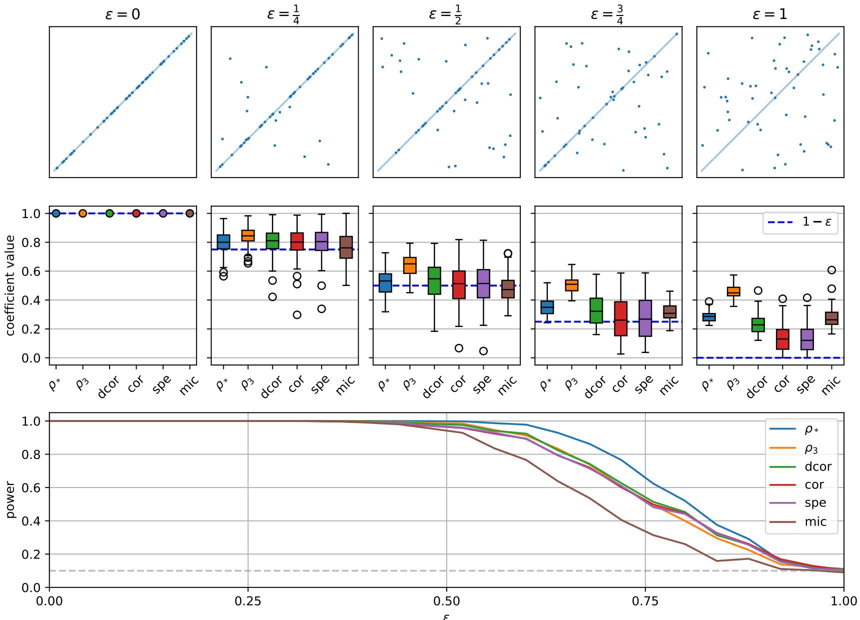


Figure 10: Dependency coefficients applied to increasingly noisy datasets according to the convex noise model in (46), where γ is given in terms of the identity on $[0, 1]$. The scatter plots on the top show exemplary samples drawn from γ^ϵ of size $n = 50$. The box plots are based on 100 such samples. The power curves on the bottom display the results of the permutation tests described in (47). To estimate the power, 1000 tests were conducted per value of ϵ for each coefficient. The significance level (dashed line) of these tests is 10%.

The picture changes substantially for the zigzag example in Figure 11. Due to the absence of monotonicity, cor and spe are not able to distinguish data points originated from the zigzag function from the ones coming from the independence coupling of its marginals. The coefficient dcor and, to a lesser extend, ρ_* also assume lower values and can only partially discern the dependency structure under noise. Meanwhile, the coefficients mic and specifically ρ_3 lie systematically higher and are still able to recognize dependency under high noise levels.

Adaptability. In Figure 4 of the introduction, we already noted that ρ_3 performs very well on a 3-Lipschitz functional relation. Additionally, Figure 11 testifies that ρ_3 also outperforms all other considered coefficients when applied to a 5-Lipschitz relation. It stands to reason, however, that ρ_5 would do even better than ρ_3 in this setting. To further examine the claim that adapting α to the slopes inherent in γ improves the performance, we conducted a series of numerical experiments with functions of different maximal slope and varying α. The findings are displayed in Figure 12, where the power of ρ_α-based permutation tests is plotted as a function of α (for fixed noise levels $\epsilon = 0.75$). The results clearly suggest that choosing a value of α close to the Lipschitz constant of the actual functional relation in γ can significantly improve the test performance. If α is chosen too small or too large, the power systematically declines.

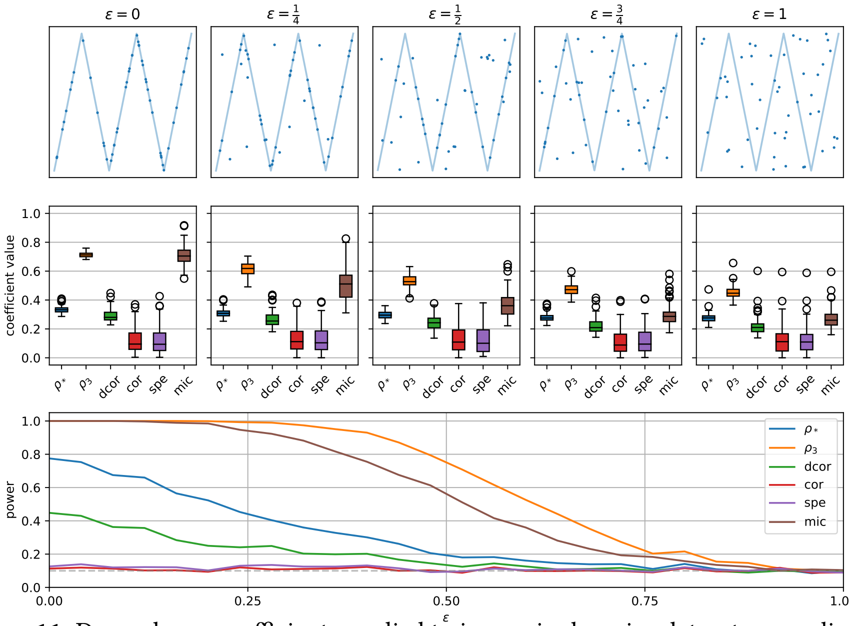


Figure 11: Dependency coefficients applied to increasingly noisy datasets according to the convex noise model in (46), where γ is given in terms of a zigzag function with maximal slope 5. See Figure 10 for a description of the individual graphs.

Bias and higher dimensions. Another prevalent trend in the numerical results so far is the notable bias of ρ_* and ρ_α on independent noise (i.e., for $\epsilon = 1$), where we expect a value of 0 for $n \to \infty$. In fact, empirical estimators of optimal transport distances are known to be susceptible to a certain degree of bias, especially in high dimensions. In Figure 13, we therefore take a look at the empirical estimation of ρ_* and compare it to dcor in settings of different dimensions for sample sizes n running from 10 to 1000. We observe that the bias of ρ_* and dcor seem comparable for $r = q = 1$. If the dimensions are chosen higher, the bias increases much quicker for ρ_* than for dcor. At the same time, the variance of the ρ_* estimates is (in part much) smaller for all choices of dimensions. Indeed, in case of $r = q = 5$ and $n = 1000$, the estimated standard deviation of ρ_* is multiple times smaller than the one of dcor, even though the bias is substantial ($\rho_* \approx 0.72$, while dcor ≈ 0.13).

A high bias in itself does not necessarily mean that ρ_* is blind to dependencies in high dimensions, however. Figure 14 reveals that simple linear structures with noise of the form (46) are still recognized somewhat better via ρ_* than via dcor for $n = 50$, particularly in the setting $r = q = 5$. To examine a more involved example, we also look at spherical dependencies, where γ is the uniform distribution on the sphere $\mathbb{S}^{r+q-1} \subset \mathbb{R}^r \times \mathbb{R}^q$. In this case, the results are more unintuitive and prompt several questions. First of all, spherical dependencies are separated from noise (much) better by the transport correlation than by dcor in settings with $r, q \in \{1, 2\}$. At the same time, for $r = q = 5$, both the distance correlation and the transport correlation consistently exhibit test powers that are smaller than 0.1, meaning that a random

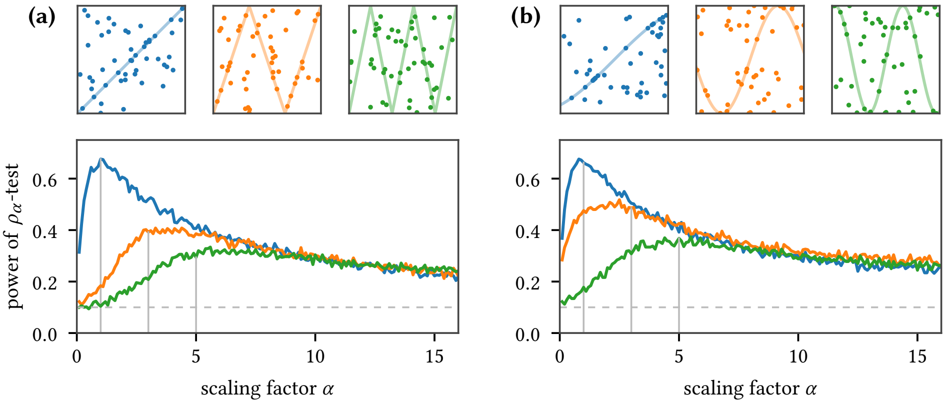


Figure 12: Influence of α on the test performance of ρ_α for linear **(a)** and sine **(b)** dependencies at a noise level of $\epsilon = 0.75$, see (46). The graphs on the top display exemplary samples together with the actual functional relations encoded in γ for different slopes. From left to right, the respective curves have maximal slopes of 1, 3, and 5 in both (a) and (b). The respective values of $\alpha \in \{1, 3, 5\}$ are marked as vertical lines in the power plot below. For each value of α, the power is estimated based on 500 samples of size $n = 50$.

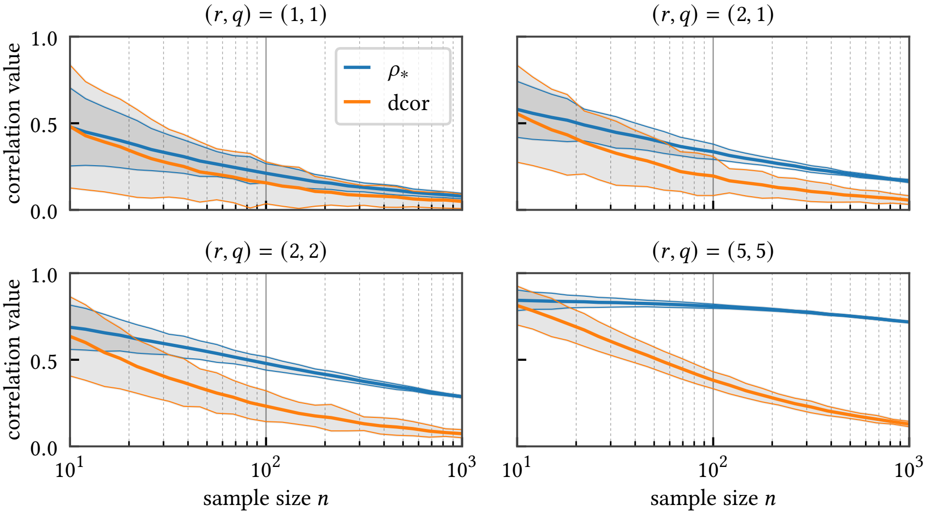


Figure 13: Empirical estimates of the isometric transport correlation and the distance correlation as a function of the sample size n in different dimensions r and q. The true distribution in the respective graphs is given by the product measure $\gamma = \mathrm{Unif}[0, 1]^r \otimes \mathrm{Unif}[0, 1]^q$. Therefore, $\rho_*(\gamma) = \mathrm{dcor}(\gamma) = 0$ is the (unbiased) value to be expected for $n \longrightarrow \infty$. The shaded error regions correspond to ± three times the estimated standard deviation. The number of samples employed for the estimation of the mean and variance was adaptively decreased from 500 for $n = 10$ to 20 for $n = 1000$.

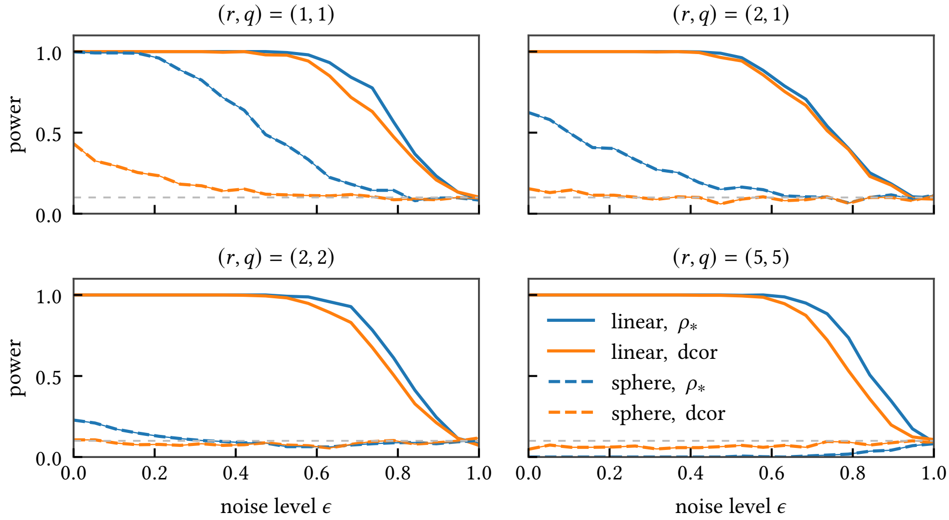


Figure 14: Test performance of the isometric transport correlation and the distance correlation in different dimensions r and q. Shown are results for both linear and spherical relations under the noise model (46). In the former case, γ is given by $(\mathrm{id}, \mathrm{id})_* \mathrm{Unif}[0, 1]^r$ if $r = q$ and by $(\mathrm{id}, \mathrm{pr}_1)_* \mathrm{Unif}[0, 1]^r$ in case of $r = 2$ and $q = 1$, where pr_1 denotes the projection on the first coordinate. In case of the sphere, γ is uniformly distributed on the surface $\mathbb{S}^{r+q-1} \subset \mathbb{R}^{r+q}$. For all power curves, 1000 samples of size $n = 50$ were used.

sample from the sphere regularly results in *lower* estimates of ρ_* and dcor than drawing from the corresponding marginals $\mu \otimes \nu$ would. This effect is particularly severe in case of the transport dependency. Even though this observation could (in part) be caused by the small sample size $n = 50$, which might not be sufficient for detecting non-trivial dependencies in higher dimensions, it demonstrates that the properties of the transport correlation in more complex settings remain an open issue that merit further investigation.

Acknowledgements. Thomas Giacomo Nies acknowledges support of DFG CRC 1456. Thomas Staudt acknowledges support of DFG RTG 2088. Axel Munk is supported by, and Thomas Giacomo Nies and Thomas Staudt were in part funded by, the DFG under Germany's Excellence Strategy EXC 2067/1-390729940. We also want to thank Facundo Memoli and Yoav Zemel for insightful discussions in the initial phase of this project and Mathias Drton for helpful comments.

Bibliography

Albanese, D., M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello (2012). "Minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers". *Bioinformatics* 29.3, pp. 407–408.

Albanese, D., S. Riccadonna, C. Donati, and P. Franceschi (2018). "A practical tool for maximal information coefficient analysis". *GigaScience* 7.4.

Ambrosio, L., N. Gigli, and G. Savaré (2008). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Springer Science & Business Media.

Amgad, M., A. Itoh, and M. M. K. Tsui (2015). "Extending Ripley's K-function to quantify aggregation in 2-D grayscale images". *PloS one* 10.12, e0144404.

Backhoff, J., D. Bartl, M. Beiglböck, and J. Wiesel (2022). "Estimating processes in adapted Wasserstein distance". *The Annals of Applied Probability* 32.1, pp. 529–550.

Berrett, T. B. and R. J. Samworth (2019). "Nonparametric independence testing via mutual information". *Biometrika* 106.3, pp. 547–566.

Berrett, T. B., R. J. Samworth, and M. Yuan (2019). "Efficient multivariate entropy estimation via k-nearest neighbour distances". *Annals of Statistics* 47.1, pp. 288–318.

Billingsley, P. (2013). *Convergence of Probability Measures.* John Wiley & Sons.

Cao, S. and P. J. Bickel (2020). "Correlations with tailored extremal properties". *Preprint arXiv:2008.10177.*

Castro-Prado, F. and W. González-Manteiga (2020). "Nonparametric independence tests in metric spaces: What is known and what is not". *Preprint arXiv:2009.14150.*

Chakraborty, S. and X. Zhang (2021). "A new framework for distance and kernel-based metrics in high dimensions". *Electronic Journal of Statistics* 15.2, pp. 5455–5522.

Chang, J. T. and D. Pollard (1997). "Conditioning as disintegration". *Statistica Neerlandica* 51 (3), pp. 287–317.

Chatterjee, S. (2021). "A new coefficient of correlation". *Journal of the American Statistical Association* 116.536, pp. 2009–2022.

Chen, H.-B. and J. Niles-Weed (2021). "Asymptotics of smoothed Wasserstein distances". *Potential Analysis*, pp. 1–25.

Cuturi, M. (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems.* Vol. 26. Curran Associates, Inc.

Deb, N., P. Ghosal, and B. Sen (2020). "Measuring association on topological spaces using kernels and geometric graphs". *Preprint arXiv:2010.01768.*

Dette, H., K. F. Siburg, and P. A. Stoimenov (2013). "A copula-based non-parametric measure of regression dependence". *Scandinavian Journal of Statistics* 40.1, pp. 21–41.

Dowson, D. and B. Landau (1982). "The Fréchet distance between multivariate normal distributions". *Journal of Multivariate Analysis* 12.3, pp. 450–455.

Dubey, P. and H.-G. Müller (2020). "Functional models for time-varying random objects". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.2, pp. 275–327.

Erbar, M., M. Rumpf, B. Schmitzer, and S. Simon (2020). "Computation of optimal transport on discrete metric measure spaces". *Numerische Mathematik* 144.1, pp. 157–200.

Estévez, P. A., M. Tesmer, C. A. Perez, and J. M. Zurada (2009). “Normalized mutual information feature selection”. *IEEE Transactions on Neural Networks* 20.2, pp. 189–201.

Flamary, R. and N. Courty (2017). *POT Python Optimal Transport library.*

Friedman, J. H. and L. C. Rafsky (1983). “Graph-theoretic measures of multivariate association and prediction”. *The Annals of Statistics* 11.2, pp. 377–391.

Gebelein, H. (1941). “Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung”. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 21.6, pp. 364–379.

Gelfand, I. M. (1959). “Calculation of the amount of information about a random function contained in another such function”. In: *Eleven Papers on Analysis, Probability and Topology.* American Mathematical Society, pp. 199–246.

Ghosal, P. and B. Sen (2019). “Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing”. *Preprint arXiv:1905.05340.*

Goldfeld, Z. and K. Greenewald (2020). “Gaussian-smoothed optimal transport: Metric structure and statistical efficiency”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics.* PMLR, pp. 3327–3337.

Goldfeld, Z., K. Greenewald, J. Niles-Weed, and Y. Polyanskiy (2020). “Convergence of smoothed empirical measures with applications to entropy estimation”. *IEEE Transactions on Information Theory* 66.7, pp. 4368–4391.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). “A kernel two-sample test”. *The Journal of Machine Learning Research* 13.1, pp. 723–773.

Gretton, A., O. Bousquet, A. Smola, and B. Schölkopf (2005). “Measuring statistical dependence with Hilbert-Schmidt norms”. In: *International Conference on Algorithmic Learning Theory.* Springer, pp. 63–77.

Guntuboyina, A., B. Sen, et al. (2018). “Nonparametric shape-restricted regression”. *Statistical Science* 33.4, pp. 568–594.

Hall, P. and S. C. Morton (1993). “On the estimation of entropy”. *Annals of the Institute of Statistical Mathematics* 45.1, pp. 69–88.

Heller, R., Y. Heller, and M. Gorfine (2013). “A consistent multivariate test of association based on ranks of distances”. *Biometrika* 100.2, pp. 503–510.

Horowitz, J. L. and C. F. Manski (1995). “Identification and robustness with contaminated and corrupted data”. *Econometrica: Journal of the Econometric Society*, pp. 281–302.

Huber, P. J. (1992). “Robust estimation of a location parameter”. In: *Breakthroughs in Statistics.* Springer, pp. 492–518.

Jacobs, M. and F. Léger (2020). “A fast approach to optimal transport: The back-and-forth method”. *Numerische Mathematik* 146.3, pp. 513–544.

Jakobsen, M. E. (2017). “Distance covariance in metric spaces: Non-parametric independence testing in metric spaces”. *Preprint arXiv:1706.03490.*

Janssen, A. and T. Pauls (2003). “How do bootstrap and permutation tests work?” *The Annals of Statistics* 31.3, pp. 768–806.

Kallenberg, O. (2006). *Foundations of Modern Probability.* Springer Science & Business Media.

Kantorovich, L. V. (1942). “On the translocation of masses”. *Doklady Akademii Nauk URSS* 37, pp. 7–8.

Kechris, A. (2012). *Classical Descriptive Set Theory.* Vol. 156. Springer Science & Business Media.

Koyak, R. A. (1987). “On measuring internal dependence in a set of random variables”. *The Annals of Statistics*, pp. 1215–1228.

Kraskov, A., H. Stögbauer, R. G. Andrzejak, and P. Grassberger (2005). “Hierarchical clustering using mutual information”. *EPL (Europhysics Letters)* 70.2, p. 278.

Lehmann, E. L. and J. P. Romano (2006). *Testing Statistical Hypotheses.* Springer Science & Business Media.

Lei, J. (2020). “Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces”. *Bernoulli* 26.1, pp. 767–798.

Levinger, B. W. (1980). “The square root of a 2×2 matrix”. *Mathematics Magazine* 53.4, pp. 222–224.

Lin, T., N. Ho, and M. Jordan (2019). “On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms”. In: *Proceedings of the 36th International Conference on Machine Learning.* PMLR, pp. 3982–3991.

Liu, J., W. Yin, W. Li, and Y. T. Chow (2021). “Multilevel optimal transport: A fast approximation of Wasserstein-1 distances”. *SIAM Journal on Scientific Computing* 43.1, A193–A220.

Lopez, A. T. and V. Jog (2018). “Generalization error bounds using Wasserstein distances”. In: *2018 IEEE Information Theory Workshop (ITW).* IEEE.

Lyons, R. (2013). “Distance covariance in metric spaces”. *The Annals of Probability* 41.5, pp. 3284–3305.

Lyons, R. (2018). “Errata to “Distance covariance in metric spaces””. *The Annals of Probability* 46.4, pp. 2400–2405.

Maes, F., A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens (1997). “Multimodality image registration by maximization of mutual information”. *IEEE Transactions on Medical Imaging* 16.2, pp. 187–198.

Mallows, C. L. (1972). “A note on asymptotic joint normality”. *The Annals of Mathematical Statistics*, pp. 508–515.

Marti, G., S. Andler, F. Nielsen, and P. Donnat (2017). “Exploring and measuring non-linear correlations: Copulas, lightspeed transportation and clustering”. In: *NIPS 2016 Time Series Workshop.* PMLR, pp. 59–69.

Matteson, D. S. and R. S. Tsay (2017). “Independent component analysis via distance covariance”. *Journal of the American Statistical Association* 112 (518), pp. 623–637.

Meckes, M. W. (2013). “Positive definite metric spaces”. *Positivity* 17.3, pp. 733–757.

Mordant, G. and J. Segers (2022). “Measuring dependence between random vectors via optimal transport”. *Journal of Multivariate Analysis* 189, p. 104912.

Móri, T. F. and G. J. Székely (2019). “Four simple axioms of dependence measures”. *Metrika: International Journal for Theoretical and Applied Statistics* 82.1, pp. 1–16.

Móri, T. F. and G. J. Székely (2020). "The Earth Mover's Correlation". *Preprint arXiv:2009.04313*.

Müller, A. (1997). "Integral probability metrics and their generating classes of functions". *Advances in Applied Probability*, pp. 429–443.

Ozair, S., C. Lynch, Y. Bengio, A. Van den Oord, S. Levine, and P. Sermanet (2019). "Wasserstein dependency measure for representation learning". *Advances in Neural Information Processing Systems* 32.

Panaretos, V. M. and Y. Zemel (2020). *An Invitation to Statistics in Wasserstein Space*. Springer Nature.

Paninski, L. and M. Yajima (2008). "Undersmoothed kernel entropy estimators". *IEEE Transactions on Information Theory* 54.9, pp. 4384–4388.

Petersen, A. and H.-G. Müller (2019). "Wasserstein covariance for multiple random densities". *Biometrika* 106.2, pp. 339–351.

Peyré, G. and M. Cuturi (2019). "Computational optimal transport: With applications to data science". *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607.

Pluim, J. P., J. A. Maintz, and M. A. Viergever (2000). "Image registration by maximization of combined mutual information and gradient information". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 452–461.

Rachev, S. T. and L. Rüschendorf (1998). *Mass Transportation Problems: Volume I: Theory*. Vol. 1. Springer Science & Business Media.

Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti (2011). "Detecting novel associations in large data sets". *Science* 334 (6062), pp. 1518–1524.

Reshef, Y. A., D. N. Reshef, H. K. Finucane, P. C. Sabeti, and M. Mitzenmacher (2016). "Measuring dependence powerfully and equitably". *The Journal of Machine Learning Research* 17.1, pp. 7406–7468.

Ripley, B. D. (1976). "The second-order analysis of stationary point processes". *Journal of Applied Probability* 13.2, pp. 255–266.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs, and Modeling*. Springer.

Schmitzer, B. (2019). "Stabilized sparse scaling algorithms for entropy regularized transport problems". *SIAM Journal on Scientific Computing* 41.3, A1443–A1481.

Schrieber, J., D. Schuhmacher, and C. Gottschlich (2016). "Dotmark–a benchmark for discrete optimal transport". *IEEE Access* 5, pp. 271–282.

Schweizer, B. and E. F. Wolff (1981). "On nonparametric measures of dependence for random variables". *The Annals of Statistics* 9.4, pp. 879–885.

Sejdinovic, D., B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). "Equivalence of distance-based and RKHS-based statistics in hypothesis testing". *The Annals of Statistics*, pp. 2263–2291.

Shannon, C. E. (1948). "A mathematical theory of communication". *The Bell System Technical Journal* 27.3, pp. 379–423.

Shi, H., M. Drton, and F. Han (2020). "Distribution-free consistent independence tests via center-outward ranks and signs". *Journal of the American Statistical Association*, pp. 1–16.

Shi, H., M. Hallin, M. Drton, and F. Han (2021). "On universally consistent and fully distribution-free rank tests of vector independence". *Preprint arXiv:2007.02186.*

Sommerfeld, M., J. Schrieber, Y. Zemel, and A. Munk (2019). "Optimal transport: Fast probabilistic approximation with exact solvers". *Journal of Machine Learning Research* 20.105, pp. 1–23.

Sriperumbudur, B. K., K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet (2012). "On the empirical estimation of integral probability metrics". *Electronic Journal of Statistics* 6, pp. 1550–1599.

Sugiyama, M. and K. M. Borgwardt (2013). "Measuring statistical dependence via the mutual information dimension". In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.* AAAI Press, pp. 1692–1698.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). "Measuring and testing dependence by correlation of distances". *The Annals of Statistics* 35 (6), pp. 2769–2794.

Tameling, C., M. Sommerfeld, and A. Munk (2019). "Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications". *The Annals of Applied Probability* 29.5, pp. 2744–2781.

Tameling, C., S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk (2021). "Colocalization for super-resolution microscopy via optimal transport". *Nature Computational Science* 1.3, pp. 199–211.

Tjøstheim, D., H. Otneim, and B. Støve (2018). "Statistical dependence: Beyond Pearson's ρ". *Preprint arXiv:1809.10455.*

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation.* Springer Science & Business Media.

Varadarajan, V. S. (1958). "On the convergence of sample probability distributions". *Sankhyā: The Indian Journal of Statistics (1933-1960)* 19.1/2, pp. 23–26.

Villani, C. (2008). *Optimal Transport: Old and New.* Vol. 338. Springer Science & Business Media.

Viola, P. and W. M. Wells III (1997). "Alignment by maximization of mutual information". *International Journal of Computer Vision* 24.2, pp. 137–154.

Wang, H., M. Diaz, J. C. S. Santos Filho, and F. P. Calmon (2019). "An information-theoretic view of generalization via Wasserstein distance". In: *2019 IEEE International Symposium on Information Theory (ISIT).* IEEE, pp. 577–581.

Wang, S., E. T. Arena, K. W. Eliceiri, and M. Yuan (2017). "Automated and robust quantification of colocalization in dual-color fluorescence microscopy: A nonparametric statistical approach". *IEEE Transactions on Image Processing* 27.2, pp. 622–636.

Weed, J. and F. Bach (2019). "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance". *Bernoulli* 25.4A, pp. 2620–2648.

Wiesel, J. (2021). "Measuring association with Wasserstein distances". *Preprint arXiv:2102.00356.*

Xiao, Y. and W. Y. Wang (2019). "Disentangled representation learning with Wasserstein total correlation". *Preprint arXiv:1912.12818*.

Yao, S., X. Zhang, and X. Shao (2018). "Testing mutual independence in high dimension via distance covariance". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3, pp. 455–480.

Zhang, J., T. Liu, and D. Tao (2018). "An optimal transport view on generalization". *Preprint arXiv:1811.03270*.

A Omitted statements and proofs

This appendix provides detailed proofs for all statements that are not proven in the main text. As a principal strategy, we often employ disintegrations to construct or destruct transport plans. This is often not the only viable method, however, and several of our results could alternatively be tackled by other techniques, like exploiting the cyclical monotonicity of optimal transport plans (see Villani 2008).

Invariance, continuity, and convolutions. In the following, we provide proofs for Lemma 1 (invariance of optimal transport), Proposition 4 (continuity of the transport dependency), as well as Lemma 2 (properties of optimal transport under convolutions). We furthermore provide the auxiliary Lemma 4 that controls optimal transport costs under uniform changes of the base costs.

Proof of Lemma 1. By a change of variables, we can immediately establish $\pi c_f = (f \times f)_* \pi c$ for any $\pi \in \mathcal{C}(\mu, \nu)$, where $(f \times f)_* \pi \in \mathcal{C}(f_*\mu, f_*\nu)$. This shows

$$T_{c_f}(\mu, \nu) = \inf_{\pi \in \mathcal{C}(\mu,\nu)} \pi c_f = \inf_{\pi \in \mathcal{C}(\mu,\nu)} (f \times f)_* \pi c \geq \inf_{\tilde{\pi} \in \mathcal{C}(f_*\mu, f_*\nu)} \tilde{\pi} c = T_c(f_*\mu, f_*\nu).$$

To prove the reverse inequality, we fix some $\tilde{\pi} \in \mathcal{C}(f_*\mu, f_*\nu)$ and explicitly construct a measure $\pi \in \mathcal{C}(\mu, \nu)$ that satisfies $\pi c_f = \tilde{\pi} c$.

We construct π in two steps. First, we define an intermediate measure π' by the relation $\pi'(\mathrm{d}x_1, \mathrm{d}y_2) = \tilde{\pi}\big(f(x_1), \mathrm{d}y_2\big)\, \mu(\mathrm{d}x_1)$. In the second step, we set $\pi(\mathrm{d}x_1, \mathrm{d}x_2) = \pi'\big(\mathrm{d}x_1, f(x_2)\big)\, \nu(\mathrm{d}x_2)$. It is straightforward to check $\pi' \in \mathcal{C}(\mu, f_*\nu)$ and $\pi \in \mathcal{C}(\mu, \nu)$ by applying substitution and utilizing the properties of conditioning. In a similar vein, consecutive steps of substitution also show

$$\begin{aligned}
\pi c_f &= \int c_f(x_1, x_2)\, \pi(\mathrm{d}x_1, \mathrm{d}x_2) \\
&= \int c\big(f(x_1), f(x_2)\big)\, \pi'\big(\mathrm{d}x_1, f(x_2)\big)\, \nu(\mathrm{d}x_2) \\
&= \int c\big(f(x_1), y_2\big)\, \pi'(\mathrm{d}x_1, y_2)\, (f_*\nu)(\mathrm{d}y_2) \\
&= \int c\big(f(x_1), y_2\big)\, \pi'(\mathrm{d}x_1, \mathrm{d}y_2) \\
&= \int c\big(f(x_1), y_2\big)\, \tilde{\pi}\big(f(x_1), \mathrm{d}y_2\big)\, \mu(\mathrm{d}x_1) \\
&= \int c(y_1, y_2)\, \tilde{\pi}(y_1, \mathrm{d}y_2)\, (f_*\mu)(\mathrm{d}y_1) = \tilde{\pi} c.
\end{aligned}$$

This establishes $T_{c_f}(\mu, \nu) \leq T_c(f_*\mu, f_*\nu)$ and thus the equality of the two transport costs. Note that $T_c(f_*\mu, f_*\nu)$ and $T_{c_f}(\mu, \nu)$ do not have to be finite for this result to hold, as all integrands are non-negative. □

Proof of Proposition 4. In order to prove this claim, we first establish an auxiliary result.

Lemma 3: Let X be a Polish space and $f, g\colon X \longrightarrow [0, \infty)$ continuous functions with $g \leq f$. For a sequence $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(X)$ converging weakly to $\mu \in \mathcal{P}(X)$, it holds that

$$\lim_{n\to\infty} \mu_n f = \mu f < \infty \qquad \text{implies} \qquad \lim_{n\to\infty} \mu_n g = \mu g < \infty.$$

Proof. Assume $\mu_n f \longrightarrow \mu f < \infty$ as $n \longrightarrow \infty$. According to Ambrosio et al. 2008, Lemma 5.1.7, this implies that f is uniformly integrable with respect to $(\mu_n)_{n\in\mathbb{N}}$. Fix some $k > 0$. Since $g(x) > k$ implies $f(x) > k$ for all $x \in X$, we find $\mu_n\big(\mathbb{1}_{g>k}\, g\big) \leq \mu_n\big(\mathbb{1}_{f>k}\, f\big)$ for all n and conclude that g is also uniformly integrable with respect to $(\mu_n)_{n\in\mathbb{N}}$. This in turn asserts that $\mu_n g \longrightarrow \mu g$ as $n \longrightarrow \infty$. □

We now return to the proof of Proposition 4. We know that $\liminf_{n\to\infty} \tau(\gamma_n) \geq \tau(\gamma)$ due to the semi-continuity of τ (Proposition 3). Thus, it is enough to show $\limsup_{n\to\infty} \tau(\gamma_n) =: \bar{\tau} \leq \tau(\gamma)$. By passing to a (not explicitly named) subsequence, we can assume that $\tau(\gamma_n)$ converges to $\bar{\tau}$.

Let $\gamma_n \in \mathcal{C}(\mu_n, \nu_n)$ and $\gamma \in \mathcal{C}(\mu, \nu)$ for suitable marginal distributions $\mu, \mu_n \in \mathcal{P}(X)$ and $\nu, \nu_n \in \mathcal{P}(Y)$. Since $\gamma_n \rightharpoonup \gamma$ by assumption, we note that $\mu_n \rightharpoonup \mu$ and $\nu_n \rightharpoonup \nu$ as $n \longrightarrow \infty$ by the continuous mapping theorem. Like in the proof of Proposition 3, we conclude that $\mu_n \otimes \nu_n$ converges weakly to $\mu \otimes \nu$. Next, we fix a point $(x_0, y_0) \in X \times Y$. According to Lemma 3, $\mu_n\, d_X(\cdot, x_0)^p = \gamma_n\, d_X(\cdot, x_0)^p \longrightarrow \gamma\, d_X(\cdot, x_0)^p = \mu\, d_X(\cdot, x_0)^p$ as $n \longrightarrow \infty$, where we set $g(x, y) = d_X(x, x_0)^p$, $f(x, y) = d\big((x, y), (x_0, y_0)\big)^p$, and used that $\gamma_n f \longrightarrow \gamma f < \infty$ (by assumption). Therefore, μ_n converges p-weakly to μ, and the same holds for ν_n by an analog argument.

In the next step, we bound the metric d on $X \times Y$ by applying the triangle inequality and using that $\big(\sum_{i=1}^4 a_i\big)^p \leq \big(4 \max_i a_i\big)^p \leq 4^p \sum_{i=1}^4 a_i^p$ for numbers $a_i \geq 0$ with $1 \leq i \leq 4$. We find

$$\begin{aligned} d(x_1, y_1, x_2, y_2)^p &\leq \big(d_X(x_1, x_0) + d_X(x_2, x_0) + d_Y(y_1, y_0) + d_Y(y_2, y_0)\big)^p \\ &\leq 4^p \big(d_X(x_1, x_0)^p + d_X(x_2, x_0)^p + d_Y(y_1, y_0)^p + d_Y(y_2, y_0)^p\big) \\ &=: \delta(x_1, y_1, x_2, y_2). \end{aligned}$$

Let π_n^* denote an (arbitrary) optimal transport plan between γ_n and $\mu_n \otimes \nu_n$, meaning $\tau(\gamma_n) = \pi_n^* c$. Applying the previous inequality yields

$$\pi_n^*\, c \leq \pi_n^*\, d^p \leq \pi_n^*\, \delta = 2 \cdot 4^p \big(\mu_n\, d_X(\cdot, x_0)^p + \nu_n\, d_Y(\cdot, y_0)^p\big),$$

where the right hand side converges as $n \longrightarrow \infty$, since we have established that μ_n and ν_n converge p-weakly. As the cost c is continuous, we can use the stability result in (Villani 2008, Theorem 5.20) and conclude that there is a subsequence $\pi_{n_k}^*$ of optimal transport plans weakly converging to some optimal $\pi^* \in \mathcal{C}(\gamma, \mu \otimes \nu)$, such that $\tau(\gamma) = \pi^* c$. Exploiting that

$$\pi_{n_k}^* \delta \longrightarrow 2 \cdot 4^p \big(\mu \, d_X(\cdot, x_0)^p + \nu \, d_Y(\cdot, y_0)^p\big) = \pi^* \delta < \infty$$

as $k \longrightarrow \infty$, we can apply Lemma 3 with $f = \delta$ and $g = c$ to finally conclude

$$\bar{\tau} = \lim_{n \to \infty} \pi_n^* c = \lim_{k \to \infty} \pi_{n_k}^* c = \pi^* c = \tau(\gamma),$$

which completes the proof. □

Lemma 4: Let X be a Polish space and let c_1 and c_2 be cost functions on X that satisfy

$$\max\big(\|c_1/c_2 - 1\|_\infty, \|c_2/c_1 - 1\|_\infty\big) \leq a$$

under the convention $0/0 = 1$, where $\|\cdot\|_\infty$ is the sup norm and $a > 0$. Then for all $\mu, \nu \in \mathcal{P}(X)$ with $T_{c_2}(\mu, \nu) < \infty$, we find $T_{c_1}(\mu, \nu) \leq (1 + a) T_{c_2}(\mu, \nu) < \infty$ and

$$|T_{c_1}(\mu, \nu) - T_{c_2}(\mu, \nu)| \leq a \, \max\big(T_{c_1}(\mu, \nu), T_{c_2}(\mu, \nu)\big) \leq a(1 + a) \, T_{c_2}(\mu, \nu).$$

Proof of Lemma 4. We first note that $c_1 \leq (1 + a)c_2$, which implies $T_{c_1}(\mu, \nu) \leq (1 + a) T_{c_2}(\mu, \nu)$. Furthermore, if π_1^* and π_2^* denote optimal transport plans between μ and ν under the costs c_1 and c_2, then

$$T_{c_1}(\mu, \nu) - T_{c_2}(\mu, \nu) = \pi_1^* \, c_1 - \pi_2^* \, c_2 \leq \pi_2^* |c_1 - c_2| \leq a \, \pi_2^* \, c_2 = a \, T_{c_2}(\mu, \nu)$$

with an analog result for $T_{c_2}(\mu, \nu) - T_{c_1}(\mu, \nu)$, which establishes the claims of the lemma. □

Proof of Lemma 2. To prove the first inequality, we reach for the dual formulation (9) of optimal transport. For any continuous and bounded potential $f \colon X \longrightarrow \mathbb{R}$, we define $f_\kappa(x) = \kappa \, f(x + \cdot)$ for $x \in X$. It is easy to check that f_κ is again continuous and bounded. If g is another potential such that $f \oplus g \leq c$, we find

$$f_\kappa(x_1) + g_\kappa(x_2) = \kappa\big(f(x_1 + \cdot) + g(x_2 + \cdot)\big) \leq \int c(x_1 + y, x_2 + y) \, \kappa(\mathrm{d}y) = c(x_1, x_2)$$

for any $x_1, x_2 \in X$ due to the translation invariance of c. Therefore, $f_\kappa \oplus g_\kappa \leq c$. This implies

$$(\mu * \kappa) \, f + (\nu * \kappa) \, g = \mu f_\kappa + \nu g_\kappa \leq \sup \mu f' + \nu g' = T_c(\mu, \nu),$$

where the supremum is taken over continuous and bounded potentials f' and g' with $f' \otimes g' \leq c$. Since f and g were arbitrary, $T_c(\mu * \kappa, \nu * \kappa) \leq T_c(\mu, \nu)$ follows. Note that this arguments holds even if $T_c(\mu, \nu) = \infty$, so there are no restrictions on $\mu, \nu \in \mathcal{P}(X)$.

For the upper bound in the second result, we construct an explicit transport plan π between μ and $\mu * \kappa$. It is defined by $\pi f = \int f(x_1, x_1 + x_2)\, \mu(\mathrm{d}x_1)\kappa(\mathrm{d}x_2)$ for any measurable map $f\colon X \times X \longrightarrow [0, \infty)$. It is straightforward to check that $\pi \in \mathcal{C}(\mu, \mu * \kappa)$, and we can conclude

$$T_c(\mu, \mu * \kappa) \leq \pi c = \int c(x_1, x_1 + x_2)\, \mu(\mathrm{d}x_1)\, \kappa(\mathrm{d}x_2) = \kappa h,$$

where we made use of the translation invariance $c(x_1, x_2) = h(x_1 - x_2)$ and the symmetry $c(x_1, x_2) = c(x_2, x_1)$ of c for any $x_1, x_2 \in X$. Again, the argument stays valid even if $T_c(\mu, \mu * \kappa) = \infty$, so the results holds for any $\mu \in \mathcal{P}(X)$. □

Upper bounds and marginal transport dependency. The segment ahead establishes the upper bounds in Proposition 7 and 8, and contains proofs for Theorem 5 and 6 concerning the marginal transport dependency. While most of the guiding ideas are straightforward, several technical aspects have to be resolved.

Proof of Proposition 7. It is evident that c_{XY} is a cost function (non-negative, symmetric, lower semi-continuous). Furthermore, the second inequality in (28) follows trivially. To prove the first inequality, we construct a coupling $\pi_2 \in \mathcal{C}(\gamma, \mu \otimes \nu)$ that aims to prevent either horizontal or vertical movements, depending on which of the associated marginal costs is larger. We therefore define the set

$$S = \big\{(x_1, y_1, x_2, y_2) \,\big|\, c_X(x_1, x_2) \geq c_Y(y_1, y_2)\big\} \subset (X \times Y)^2,$$

and note that S is symmetric under exchanging (x_1, y_1) and (x_2, y_2) due to the symmetry of the costs c_X and c_Y. We write R to denote the complement of S, which is also symmetric. Next, we introduce the function $r\colon (X \times Y)^2 \longrightarrow (X \times Y)^2$ given by

$$r(x_1, y_1, x_2, y_2) = \begin{cases} (x_1, y_1, x_1, y_2) & \text{if } (x_1, y_1, x_2, y_2) \in S, \\ (x_2, y_2, x_1, y_2) & \text{else.} \end{cases}$$

The proof is completed once we show that the coupling defined by $\pi_2 = r_*(\gamma \otimes \gamma)$ has the correct marginals, meaning $\pi_2 \in \mathcal{C}(\gamma, \mu \otimes \nu)$, and that $\pi_2 c$, which is an upper bound for $\tau(\gamma)$, is in turn upper bounded by $\operatorname{diam}_{c_{XY}} \gamma$. The second marginal $\mu \otimes \nu$ is an immediate consequence of the definition of r and π_2. To check the first marginal, we consider an arbitrary positive and measurable function $f\colon X \times Y \longrightarrow \mathbb{R}$. Then, if q denotes the two first components of r,

$$\begin{aligned} \int f(x_1, y_1)\, \mathrm{d}\pi_2(x_1, y_1, x_2, y_2) &= \int f\big(q(x_1, y_1, x_2, y_2)\big)\, \mathrm{d}\gamma(x_1, y_1)\, \mathrm{d}\gamma(x_2, y_2) \\ &= \int \mathbb{1}_S(x_1, y_1, x_2, y_2)\, f(x_1, y_1)\, \mathrm{d}\gamma(x_1, y_2)\, \mathrm{d}\gamma(x_2, y_2) \\ &\quad + \int \mathbb{1}_R(x_1, y_1, x_2, y_2)\, f(x_2, y_2)\, \mathrm{d}\gamma(x_1, y_1)\, \mathrm{d}\gamma(x_2, y_2) \\ &= \int f(x_1, y_1)\, \mathrm{d}\gamma(x_1, y_1) = \gamma f, \end{aligned}$$

where we swapped the roles of (x_1, y_1) and (x_2, y_2) to establish the third equality. This is permissible due to the symmetry of S (and R). Similarly, we observe

$$\begin{aligned}\tau(\gamma) &\leq \pi_2 c \\ &= \int_S c(x_1, y_1, x_1, y_2)\, \mathrm{d}\gamma(x_1, y_1)\, \mathrm{d}\gamma(x_2, y_2) + \int_R c(x_2, y_2, x_1, y_2)\, \mathrm{d}\gamma(x_1, y_1)\, \mathrm{d}\gamma(x_2, y_2) \\ &\leq \int_S c_Y(y_1, y_2)\, \mathrm{d}\gamma(x_1, y_1)\, \mathrm{d}\gamma(x_2, y_2) + \int_R c_X(x_1, x_2)\, \mathrm{d}\gamma(x_1, y_1)\, \mathrm{d}\gamma(x_2, y_2) \\ &= \int \min\bigl(c_X(x_1, x_2), c_Y(y_1, y_2)\bigr)\, \mathrm{d}\gamma(x_1, y_1)\, \mathrm{d}\gamma(x_2, y_2) = (\gamma \otimes \gamma)\, c_{XY},\end{aligned}$$

where we bounded c by c_Y and c_X via condition (23). □

Proof of Proposition 8. In order to prove this result, we first introduce an alternative characterization of the measurability of probability kernels that is employed in Villani 2008.

> Lemma 5 (measurability of kernels): Let X be Polish and let $(\mu_\omega)_{\omega\in\Omega} \subset \mathcal{P}(X)$ be a family of probability measures indexed in a measurable space $(\Omega, \mathcal{F})$. Then
>
> $$\omega \mapsto \mu_\omega(A) \text{ is measurable for all Borel } A \subset X \quad \Longleftrightarrow \quad \omega \mapsto \mu_\omega \text{ is measurable,}$$
>
> where $\mathcal{P}(X)$ is equipped with the Borel σ-algebra with respect to the topology of weak convergence of measures.

Proof. Theorem 17.24 in Kechris 2012 asserts that the Borel σ-algebra of $\mathcal{P}(X)$ is generated by functions $r_A\colon \mathcal{P}(X) \to [0, 1]$ of the form $\nu \mapsto \nu(A)$ for Borel sets $A \subset X$. This means that $\mathcal{G} = \bigl\{r_A^{-1}(B) \bigm| B \subset [0, 1] \text{ Borel and } A \subset X \text{ Borel}\bigr\}$ is a generator of the Borel σ-algebra of $\mathcal{P}(X)$. In particular, each r_A is measurable.

Thus, if $\mu\colon \omega \mapsto \mu_\omega$ is measurable, then $\omega \mapsto (r_A \circ \mu)(\omega) = \mu_\omega(A)$ is also measurable as composition of measurable functions. Conversely, if $\omega \mapsto \mu_\omega(A)$ is measurable for each Borel set $A \subset X$, then $\mu^{-1}(G) \in \mathcal{F}$ for each $G \in \mathcal{G}$. Since $\mathcal{G}$ is a generator, this suffices to show that $\omega \mapsto \mu_\omega$ is measurable. □

We now return to the proof of Proposition 8, where we aim to improve the inner integral occurring in equation (26). To do so, we let $\pi_x^* \in \mathcal{C}\bigl(\gamma(x, \cdot), \mu\bigr)$ be an optimal transport plan with respect to the base costs c_Y for each $x \in X$. Corollary 5.22 in Villani 2008 together with the continuity of c_X and Lemma 5 above guarantee that π_x^* can be selected such that $(\pi_x^*)_{x\in X}$ is a probability kernel. We can thus define $\pi_3 \in \mathcal{P}(X \times Y)$ via

$$\mathrm{d}\pi_3(x_1, y_1, x_2, y_2) = \pi_{x_1}^*(\mathrm{d}y_1, \mathrm{d}y_2)\, \delta_{x_1}(\mathrm{d}x_2)\, \mu(\mathrm{d}x_1).$$

It can easily be checked that the marginals of π_3 match γ and $\mu \otimes \nu$. Using condition (23), we find

$$\tau(\gamma) \leq \pi_3 c \leq \int \left(\int c_Y(y_1, y_2)\, \pi_x^*(\mathrm{d}y_1, \mathrm{d}y_2) \right) \mu(\mathrm{d}x) = \int (\pi_x^*\, c_Y)\, \mu(\mathrm{d}x),$$

which shows the first inequality of the proposition. To assert the second inequality, one just has to note that $\pi_x^*\, c_Y \leq \big(\gamma(x,\cdot) \otimes \nu\big)\, c_Y$ for each $x \in X$ by construction of π_x^* as optimal plan. □

Proof of Theorem 5. We begin the proof by showing the following auxiliary statement.

> Lemma 6: Let X be a Polish space and c a positive continuous cost function on X. For $\mu, \nu \in \mathcal{P}(X)$ with $\operatorname{supp}\mu \subset \operatorname{supp}\nu$ and $T_c(\mu,\nu) < \infty$, it holds that
>
> $$T_c(\mu,\nu) = (\mu \otimes \nu)c \qquad \Longleftrightarrow \qquad \mu = \delta_x \text{ for some } x \in X.$$

Proof. The implication from right to left is trivial, since $\mu \otimes \nu$ is the only feasible coupling if μ is a point mass. To show the reverse direction, we assume that $\mu \neq \delta_x$ for any $x \in X$ and show that $T_c(\mu,\nu) = (\mu \otimes \nu)c$ is impossible. First, we pick two distinct points $x_1 \neq x_2$ from the support of μ. By assumption, these points also lie in the support of ν. To show that $\mu \otimes \nu$ cannot be an optimal transport plan, it is sufficient to show that $\operatorname{supp}(\mu \otimes \nu) = \operatorname{supp}\mu \times \operatorname{supp}\nu$ is not c-cyclically monotone (Villani 2008). This is easy to see, since for $(x_1, x_2), (x_2, x_1) \in \operatorname{supp}(\mu \otimes \nu)$, we find

$$c(x_1,x_2) + c(x_2,x_1) > c(x_1,x_1) + c(x_2,x_2) = 0$$

due to the positivity of c. □

Returning to the proof of Theorem 5, we note that $\tau^Y(\gamma)$ and $\operatorname{diam}_{c_Y}\nu$ are equal iff

$$\int T_{c_Y}\big(\gamma(x,\cdot),\nu\big)\, \mu(\mathrm{d}x) = \int \big(\gamma(x,\cdot) \otimes \nu\big)\, c_Y\, \mu(\mathrm{d}x). \tag{48}$$

Evidently, these two values are the same if $\gamma = (\mathrm{id},\varphi)_*\mu$, since this implies $\gamma(x,\cdot) = \delta_{\varphi(x)}$ for μ-almost all $x \in X$. So it only remains to show that equality in (48) implies $\gamma = (\mathrm{id},\varphi)_*\mu$ for some μ-almost surely defined measurable function $\varphi\colon X \to Y$.

Since the right hand side in (48) is by assumption finite and $0 \leq T_{c_Y}\big(\gamma(x,\cdot),\nu\big) \leq \big(\gamma(x,\cdot) \otimes \nu\big)c_Y$ holds for each $x \in X$, we find that equality in (48) can only hold if

$$T_{c_Y}\big(\gamma(x,\cdot),\nu\big) = \big(\gamma(x,\cdot) \otimes \nu\big)c_Y$$

for μ-almost all $x \in X$. As $\operatorname{supp}\gamma(x,\cdot) \subset \operatorname{supp}\nu$ also holds for μ-almost all $x \in X$ (see below), we can apply Lemma 6 and find that $\gamma(x,\cdot) = \delta_{\varphi(x)}$ for a μ-almost surely defined function φ. The measurability of φ follows from the measurability of the maps $x \mapsto \gamma(x, A)$ for all Borel sets $A \subset Y$, since $x \in \varphi^{-1}(A)$ is equivalent to $\gamma(x, A) = 1$ for all $x \in X$ for which φ is defined by the construction above.

A brief argument to see that $\operatorname{supp}\gamma(x, \cdot) \subset \operatorname{supp}\nu$ for μ-almost all $x \in X$ goes as follows: note that $\operatorname{supp}\gamma \subset \operatorname{supp}(\mu \otimes \nu) = \operatorname{supp}\mu \times \operatorname{supp}\nu$ and write

$$1 = \int \mathrm{d}\,\gamma = \int \mathbb{1}_{\operatorname{supp}\nu}(y)\,\gamma(\mathrm{d}x, \mathrm{d}y) = \int \mathbb{1}_{\operatorname{supp}\nu}(y)\,\gamma(x, \mathrm{d}y)\,\mu(\mathrm{d}x),$$

which proves that $\gamma(x, \operatorname{supp}\nu) = 1$ for μ-almost all $x \in X$. □

Proof of Theorem 6. We begin by defining a set that contains all vertical movements along the fibers $\{x\} \times Y$, given by $S = \{(x, y_1, x, y_2) \,|\, x \in X,\ y_1, y_2 \in Y\} \subset (X \times Y)^2$. Due to the definition of c_∞, it is evident that

$$\tau_{c_\infty}(\gamma) = \inf_{\substack{\pi \in \mathcal{C}(\gamma, \mu\otimes\nu) \\ \pi(S)=1}} \pi c_Y. \tag{49}$$

We use this characterization to show that the equality $\tau_{c_\infty}(\gamma) = \tau^Y_{c_Y}(\gamma)$ holds. First, we define $f\colon S \to X \times Y^2$ via $f(x, y_1, x, y_2) = (x, y_1, y_2)$. For each π that is feasible in the infimum in (49), we furthermore define $\pi_x = (f_*\pi)(x, \cdot, \cdot) \in \mathcal{P}(Y \times Y)$ for $x \in X$. One can check that $\pi_x \in \mathcal{C}\big(\gamma(x, \cdot), \nu\big)$ holds μ-almost surely due to the (almost sure) uniqueness property of disintegrations. If $\pi^*_x \in \mathcal{C}\big(\gamma(x, \cdot), \nu\big)$ are a measurable selection of optimal transport plans for the problem $T_{c_Y}\big(\gamma(x, \cdot), \nu\big)$ as in the proof of Proposition 8, we observe

$$\pi c_Y = \int_S c_Y\,\mathrm{d}\pi = \int c_Y\,\mathrm{d}(f_*\pi) = \int (\pi_x c_Y)\,\mu(\mathrm{d}x) \geq \int (\pi^*_x c_Y)\,\mu(\mathrm{d}x) = \tau^Y_{c_Y}(\gamma).$$

Taking the infimum on the left hand side and applying (49) implies $\tau_{c_\infty}(\gamma) \geq \tau^Y_{c_Y}(\gamma)$. Since we already know $\tau_{c_\infty}(\gamma) \leq \tau^Y_{c_Y}(\gamma)$ by Proposition 8, this proves equality.

It is left to show that $\tau_{c_\alpha}(\gamma) \to \tau_{c_\infty}(\gamma)$ as $\alpha \to \infty$. Since $c_\infty \geq c_\alpha$ for all $\alpha > 0$, we know that $\tau_{c_\infty}(\gamma) \geq \tau_{c_\alpha}(\gamma)$ and it is accordingly sufficient to show $\liminf_{\alpha\to\infty} \tau_{c_\alpha}(\gamma) = \tau_{c_\infty}(\gamma)$. We thus fix $\bar{\tau} = \liminf_{\alpha\to\infty} \tau_{c_\alpha}(\gamma)$ and consider a diverging sequence $(\alpha_n)_{n\in\mathbb{N}} \subset (0, \infty)$ such that $\tau_{c_{\alpha_n}}(\gamma) \to \bar{\tau}$. Due to Prokhorov's theorem and the existence of optimal transport plans, we can assume that there are couplings π^*_n and π_∞ in $\mathcal{C}(\gamma, \mu \otimes \nu)$ that satisfy $\tau_{c_{\alpha_n}}(\gamma) = \pi^*_n c_{\alpha_n}$ and $\pi^*_n \rightharpoonup \pi_\infty$. We use these properties to lead the assumption $\infty \geq \tau_{c_\infty}(\gamma) > \bar{\tau}$ to a contradiction. If this assumption were true, observe that for any $k > 0$,

$$\infty > \bar{\tau} = \lim_{n\to\infty} \pi^*_n c_{\alpha_n} \geq \liminf_{n\to\infty} \alpha_n \pi^*_n c_X \geq k\, \liminf_{n\to\infty} \pi^*_n c_X \geq k\,\pi_\infty c_X,$$

where the final inequality follows from the lower semi-continuity of the mapping $\pi \mapsto \pi h$ under weak convergence for lower semi-continuous integrands h that are bounded from below (Santambrogio 2015, Proposition 7.1). Since k is arbitrary, we conclude $\pi_\infty c_X = 0$, which implies $\pi_\infty(S) = 1$ due to the positivity of c_X. Employing representation (49) of $\tau_{c_\infty}(\gamma)$, we find the contradiction

$$\tau_{c_\infty}(\gamma) > \bar{\tau} = \lim_{n\to\infty} \pi_n^* c_{\alpha_n} \geq \liminf_{n\to\infty} \pi_n^* c_Y \geq \pi_\infty c_Y \geq \tau_{c_\infty}(\gamma).$$

This establishes $\bar{\tau} = \tau_{c_\infty}(\gamma)$ and finishes the proof. □

Contracting couplings and maps. We next provide proofs and statements that were omitted in our work on contractions in Section 4. This includes the proof of our main result, Theorem 7, as well as the formulation of two auxiliary statements (Lemma 7 and 8), which simplify Theorem 7 if sufficient regularity is imposed on the involved costs.

Proof of Theorem 7. We first show the second part. Assuming that $\tau(\gamma) = \operatorname{diam}_{c_Y} \nu < \infty$, we use the upper bound in Proposition 7 to conclude $\operatorname{diam}_{c_{XY}} \gamma = \operatorname{diam}_{c_Y} \nu$, which can be stated as

$$\int \min\big(c_X(x_1, x_2), c_Y(y_1, y_2)\big)\, \mathrm{d}(\gamma \otimes \gamma)(x_1, y_1, x_2, y_2) = \int c_Y(y_1, y_2)\, \mathrm{d}(\nu \otimes \nu)(y_1, y_2).$$

This implies $(\gamma \otimes \gamma)(c_Y \leq c_X) = 1$, since the left hand side in the equality above would otherwise be strictly smaller than the right hand side. Recalling that $c_X = h \circ k_X$ and $c_Y = h \circ d_Y$ for a strictly increasing h, we find $(\gamma \otimes \gamma)(d_Y \leq k_X) = 1$, meaning that γ is almost surely contracting.

To prove the first statement, we show that $\tau(\gamma) \geq \operatorname{diam}_{c_Y} \nu$, which is sufficient to assert equality due to Proposition 6. Let $\pi \in \mathcal{C}(\mu, \nu)$ be arbitrary. We define $\bar{\lambda} \in \mathcal{P}\big((X \times Y)^2 \times Y\big)$ via the relation $\mathrm{d}\bar{\lambda}(x_1, y_1, x_2, y, y_2) = \gamma(x_2, \mathrm{d}y)\, \mathrm{d}\pi(x_1, y_1, x_2, y_2)$. One can readily establish that integrating out $y_2 \in Y$ yields a measure $\lambda \in \mathcal{C}(\gamma, \gamma) \subset \mathcal{P}\big((X \times Y)^2\big)$. In particular, $\operatorname{supp} \lambda \subset \operatorname{supp} \gamma \times \operatorname{supp} \gamma$, which implies $\lambda(d_Y \leq k_X) = 1$ by the assumption that γ is contracting on its support. Combining this insight with the strict monotonicity of h and the triangle inequality for d_Y, we find

$$\begin{aligned}
\pi c &= \int h\big(k_X(x_1, x_2) + d_Y(y_1, y_2)\big)\, \mathrm{d}\bar{\lambda}(x_1, y_2, x_1, y, y_2) \\
&\geq \int h\big(d_Y(y_1, y) + d_Y(y_1, y_2)\big)\, \mathrm{d}\bar{\lambda}(x_1, y_2, x_1, y, y_2) \\
&\geq \int h\big(d_Y(y, y_2)\big)\, \mathrm{d}\bar{\lambda}(x_1, y_2, x_1, y, y_2) \\
&= \int c_Y(y, y_2)\, \mathrm{d}\nu(y)\, \mathrm{d}\nu(y_2) = \operatorname{diam}_{c_Y}(\nu).
\end{aligned}$$

One can check that the independence in the final line follows from the definition of $\bar{\lambda}$. Taking the infimum over $\pi \in \mathcal{C}(\gamma, \mu \otimes \nu)$ now yields the desired result. □

Lemma 7: Let X and Y be Polish spaces and let k_X and d_Y be continuous cost functions on X and Y. Then γ is contracting on its support iff it is almost surely contracting (with respect to k_X and d_Y).

Proof of Lemma 7. It is clear that a contracting coupling γ is almost surely contracting since the set $(X \times Y)^2 \setminus (\operatorname{supp} \gamma)^2$ is a null set for $\gamma \otimes \gamma$. Let γ therefore be almost surely contracting. If k_X and d_Y are continuous and there are $(x_1, y_1), (x_2, y_2) \in \operatorname{supp} \gamma$ with $d_Y(y_1, y_2) > k_X(x_1, x_2)$, then we also find an open neighbourhood $U \subset (X \times Y)^2$ of $(x_1, y_1, x_2, y_2) \in (\operatorname{supp} \gamma)^2 = \operatorname{supp} (\gamma \otimes \gamma)$ where this inequality is true. By the definition of the support, we conclude $(\gamma \otimes \gamma)(U) > 0$ and γ thus fails to be almost surely contracting. □

Lemma 8 (uniform extension): Let X be Polish and (Y, d_Y) a Polish metric space, and let $k_X \colon X \times X \longrightarrow [0, \infty)$ be a continuous cost function on X. Any function $\tilde{\varphi} \colon D \longrightarrow Y$ defined on a subset $D \subset X$ that satisfies

$$d_Y\big(\tilde{\varphi}(x_1), \tilde{\varphi}(x_2)\big) \leq k_X(x_1, x_2) \tag{50}$$

for all $x_1, x_2 \in D$ can uniquely be extended to a function $\varphi \colon \bar{D} \longrightarrow Y$ on the closure $\bar{D}$ of D that satisfies (50) for all $x_1, x_2 \in \bar{D}$. In particular, φ is continuous.

Proof of Lemma 8. Let $x \in \bar{D} \setminus D$ and $(x_n)_{n \in \mathbb{N}}$ be a sequence in D converging to x. Set $y_n = \tilde{\varphi}(x_n)$ for $n \in \mathbb{N}$ and observe that $d_Y(y_n, y_m) \leq k_X(x_n, x_m)$ for all $n, m \in \mathbb{N}$. In particular, the sequence $(y_n)_n$ is Cauchy: if it were not Cauchy, there would exist an $\epsilon > 0$ and values $n_r, m_r \geq r$ for each $r \in \mathbb{N}$ such that $d_Y(y_{n_r}, y_{m_r}) \geq \epsilon$. However, observing $\lim_{r \to \infty} k_X(x_{n_r}, x_{m_r}) \longrightarrow 0$ due to continuity of k_X leads this to a contradiction. Consequently, the sequence $(y_n)_{n \in \mathbb{N}}$ converges to a unique limit $y \in Y$ due to the completeness of (Y, d_Y).

The limit point y does not depend on the chosen sequence: if $(x'_n)_n$ is another sequence converging to x, and y' is the corresponding limit in Y, continuity of d_Y and k_X make sure that

$$0 \leq d_Y(y, y') = \lim_{n \to \infty} d_Y(y_n, y'_n) \leq \lim_{n \to \infty} k_X(x_n, x'_n) = 0.$$

Therefore, a well-defined extension of $\tilde{\varphi}$ to $\bar{D}$ exists. For any $x, x' \in \bar{D}$, this extension φ satisfies

$$d_Y\big(\varphi(x), \varphi(x')\big) = \lim_{n \to \infty} d_Y\big(\tilde{\varphi}(x_n), \tilde{\varphi}(x'_n)\big) \leq \lim_{n \to \infty} k_X(x_n, x'_n) = k_X(x, x')$$

as $n \longrightarrow \infty$ for any sequence $(x_n, x'_n)_{n \in \mathbb{N}} \subset D \times D$ that converges to $(x, x') \in \bar{D} \times \bar{D}$. □

Properties of the transport correlation. We next present proofs of Proposition 9 to 11 regarding basic properties of the transport correlation. Most of the claimed properties are direct consequences of previously established results. Additional arguments are mainly required for properties 5 and 6 in Proposition 9 and property 2 in Proposition 11. Recall from Section 5 that we focus on additive costs of the form

$$c = (\alpha d_X + d_Y)^p$$

on Polish metric spaces (X, d_X) and (Y, d_Y) for $\alpha, p > 0$.

Proof of Proposition 9. Properties 1 and 2, which characterize when $\rho_\alpha(\gamma)$ equals 0 and 1, follow from Theorem 1 in Section 2 and from Corollary 1 in Section 4. The invariance in property 3 is a consequence of Proposition 2. If γ is restricted to a set with fixed marginal $p^Y_* \gamma = \nu$, then $\operatorname{diam}_{d_Y^p} \nu$ is constant and convexity of $\gamma \mapsto \rho_\alpha(\gamma)^p = \tau(\gamma)/\operatorname{diam}_{d_Y^p} \nu$ is guaranteed by convexity of τ as stated in Proposition 1. This shows property 4.

We now turn to the continuity in property 5. Let ν_n and ν denote the respective second marginals of γ_n and γ. For convenience, we write $\delta_n = \operatorname{diam}_{d_Y^p} \nu_n = (\nu_n \otimes \nu_n)\, d_Y^p$. Our first goal is to show that p-weak convergence of γ_n to γ implies $\delta_n \longrightarrow \delta = \operatorname{diam}_{d_Y^p} \nu$ as $n \longrightarrow \infty$. To do so, we fix some $y_0 \in Y$ and define the function

$$f(y_1, y_2) = 2^p \big(d_Y(y_1, y_0)^p + d_Y(y_0, y_2)^p\big)$$

for $y_1, y_2 \in Y$. Noting that $(a + b)^p \leq 2^p(a^p + b^p)$ for $a, b \geq 0$, we find $d_Y^p \leq f$ by application of the triangle inequality. Consequently,

$$\delta_n = (\nu_n \otimes \nu_n)\, d_Y^p \leq (\nu_n \otimes \nu_n)\, f = 2^{p+1} \nu_n d_Y(\cdot, y_0)^p \longrightarrow 2^{p+1} \nu d_Y(\cdot, y_0)^p$$

as $n \longrightarrow \infty$. The convergence in this inequality follows from the fact that ν_n converges p-weakly if γ_n converges p-weakly, which was shown in the proof of Proposition 4. Reaching back to Lemma 3 (setting $g = d_Y^p$, $f = f$, and $\mu_n = \nu_n \otimes \nu_n$), we conclude that δ_n indeed converges to δ as $n \longrightarrow \infty$. Next, since $c = (\alpha d_X + d_Y)^p \leq \max(1, \alpha)^p (d_X + d_Y)^p$, we can apply Proposition 4 and find $\tau_c(\gamma_n) \longrightarrow \tau_c(\gamma)$. Setting $c_n = (\alpha_n d_X + d_Y)^p$, it is straightforward to see that $\|c/c_n - 1\|_\infty \longrightarrow 0$ due to $\alpha_n \longrightarrow \alpha > 0$ for $n \longrightarrow \infty$, which lets us use Proposition 5 to obtain

$$\lim_{n\to\infty} \rho_{\alpha_n}(\gamma_n)^p = \lim_{n\to\infty} \frac{\tau_{c_n}(\gamma_n)}{\delta_n} = \frac{\tau_c(\gamma)}{\delta} = \rho_\alpha(\gamma)^p.$$

Finally, the monotonicity of $\alpha \mapsto \rho_\alpha(\gamma)$ in property 6 is trivial, since c increases with α. The concavity of $\alpha \mapsto \rho_\alpha(\gamma)^p$ for $p \leq 1$ and fixed $\gamma \in \mathcal{C}(\mu, \nu)$ with marginals $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ follows from the fact that pointwise infima over concave functions are again concave. Indeed, we have that

$$\rho_\alpha(\gamma)^p = \frac{1}{\operatorname{diam}_{d_Y^p} \nu} \inf_{\pi \in \mathcal{C}(\gamma, \mu \otimes \nu)} \pi(\alpha d_X + d_Y)^p,$$

where the mapping $\alpha \mapsto \pi(\alpha d_X + d_Y)^p$ is concave for any fixed π if $p \leq 1$. □

Proof of Proposition 10. Due to Theorem 6, results established for τ under generic lower semi-continuous costs c also hold for the marginal transport dependency with $\alpha = \infty$. Therefore, properties 1, 3, and 4 follow from the general results stated in Theorem 1, Proposition 2, and Proposition 1. Note that we can allow dilatations f_Y in property 3 (instead of just isometries), since we normalize by $\operatorname{diam}_{d_Y^p} \nu$ in definition (42) of ρ_∞, which neutralizes the dilatation factor $\beta > 0$. Property 2 relies on the specific cost structure for $\alpha = \infty$ and was derived separately in Theorem 5. □

Proof of Proposition 11. Properties 1, 4, and 5 follow in the same way as in the proof of Proposition 9. The symmetry property 6 is trivial and directly visible from the definition of ρ_*. Property 3 is a consequence of the general invariance of the transport dependence under isometries (Proposition 2). We can extend this to dilatations f_X and f_Y since we divide by the respective diameters in definition (44) of ρ_*, which nullifies any scaling factors.

Regarding property 2, let $\gamma \in \mathcal{C}(\mu, \nu)$ for $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. We equip the spaces X and Y with the scaled metrics

$$\tilde{d}_X = d_X / (\operatorname{diam}_{d_X^p} \mu)^{1/p} \qquad \text{and} \qquad \tilde{d}_Y = d_Y / (\operatorname{diam}_{d_Y^p} \nu)^{1/p}.$$

According to Corollary 1 (combined with Lemma 8), a value of $\rho_*(\gamma) = 1$ is equivalent to there being a contraction φ from $(\operatorname{supp} \mu, \tilde{d}_X)$ to $(Y, \tilde{d}_Y)$ that satisfies $(\mathrm{id}, \varphi)_* \mu = \gamma$. This in particular means $\varphi_* \mu = \nu$. Defining $\tilde{d}_\varphi(x_1, x_2) = \tilde{d}_Y\big(\varphi(x_1), \varphi(x_2)\big)$ for any $x_1, x_2 \in \operatorname{supp} \mu$, we know that $\tilde{d}_\varphi \leq d_X$ since φ is a contraction. By a change of variables, we calculate

$$(\mu \otimes \mu)\, \tilde{d}_\varphi^p = (\nu \otimes \nu)\, \tilde{d}_Y^p = 1 = (\mu \otimes \mu)\, \tilde{d}_X^p$$

and assert $\tilde{d}_\varphi = \tilde{d}_X$ to hold $(\mu \otimes \mu)$-almost surely. Since $\tilde{d}_X$ and $\tilde{d}_\varphi$ are continuous, we can conclude that $\varphi \colon (\operatorname{supp} \mu, \tilde{d}_X) \to (Y, \tilde{d}_Y)$ is an isometry. Thus, φ is a dilatation under the metrics d_X and d_Y. Note that its dilatation factor β is uniquely given by α_* defined in equation (43). Due to the symmetry of the setting, the same arguments also hold for ψ instead of φ by exchanging the roles of X and Y. □

B Additional Simulations

This appendix contains a range of figures that supplement Section 6 and further illustrate the behaviour of the transport correlation on noisy datasets.

Convex noise. Like Figure 10 and 11 in Section 6, the upcoming Figures 15 to 17 compare ρ_*, ρ_3, and several other dependency coefficients under the convex noise model (46) for different geometries γ. One particularly noteworthy observation is that ρ_* has a higher (or at least the same) discriminative power compared to the distance correlation, Pearson correlation, and Spearman correlation for all geometries considered.

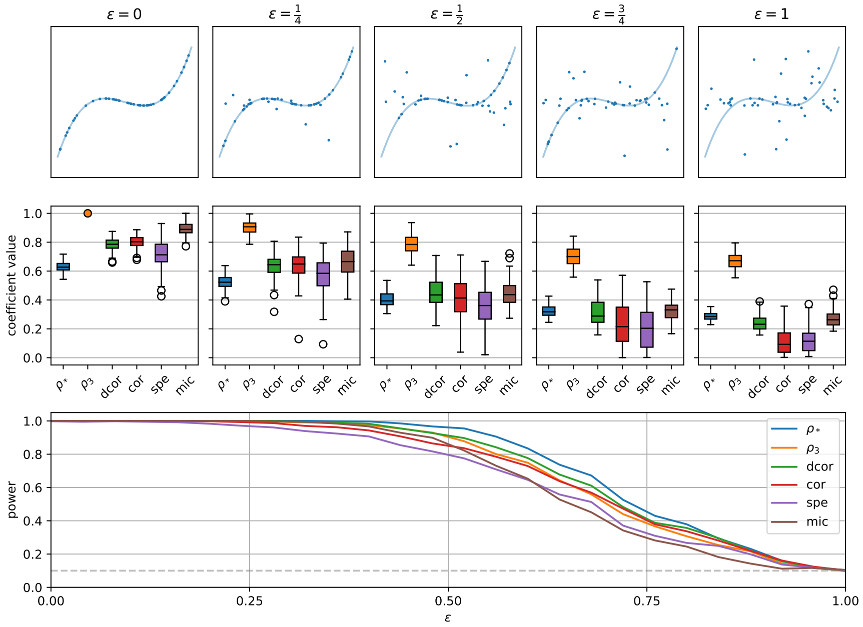


Figure 15

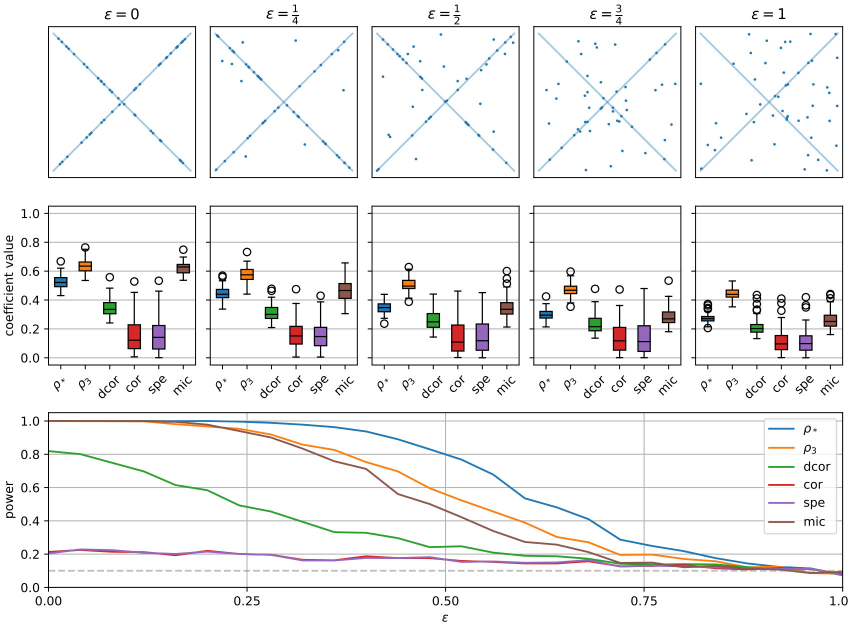


Figure 16

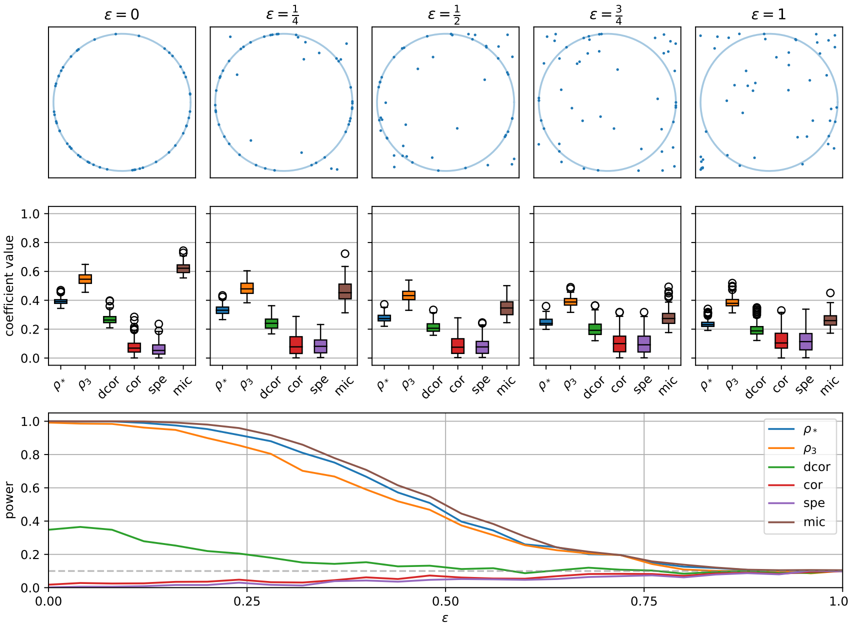


Figure 17

Gaussian additive noise. We repeated our previous simulations under a Gaussian additive noise model instead of convex noise (Figure 18 to Figure 22). For a given level of noise $\sigma > 0$ and a given base distribution γ, we consider the noisy relationships $(\xi, \zeta) \sim \gamma^\sigma$, where

$$\gamma^\sigma = \gamma * \kappa \qquad \text{with} \qquad \kappa = \mathcal{N}(0, \sigma\,\mathrm{Id}_2).$$

In this setting, we notice that the Pearson and Spearman correlations as well as the distance correlation have a higher power than ρ_* if γ exhibits a clear monotonic tendency. If, on the other hand, the linear correlation of the underlying distribution γ is low, then we observe similar trends as for the convex noise model.

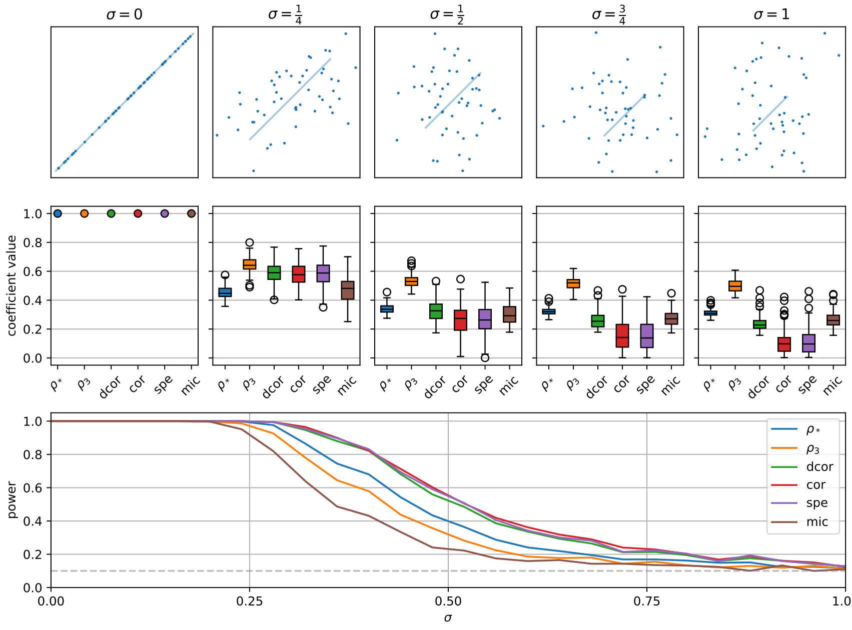


Figure 18

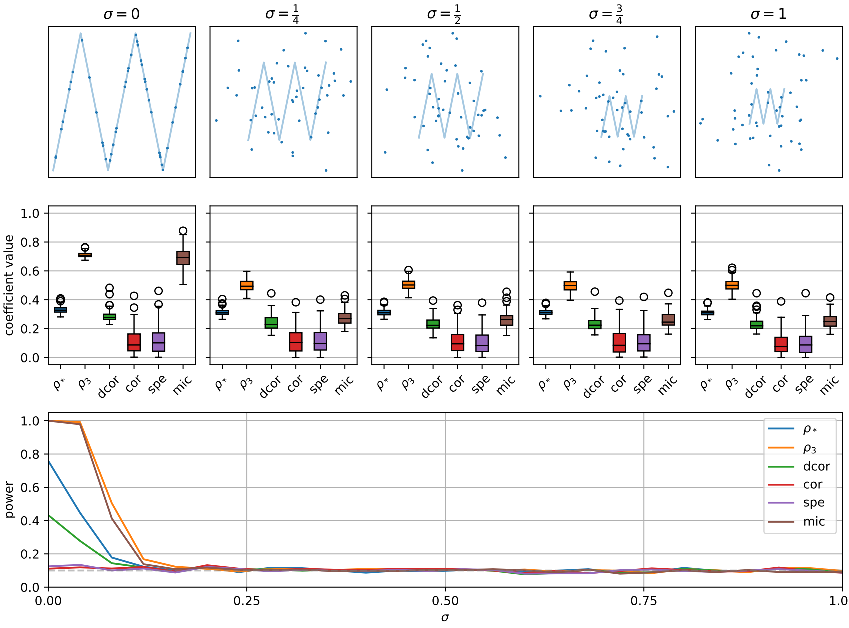


Figure 19

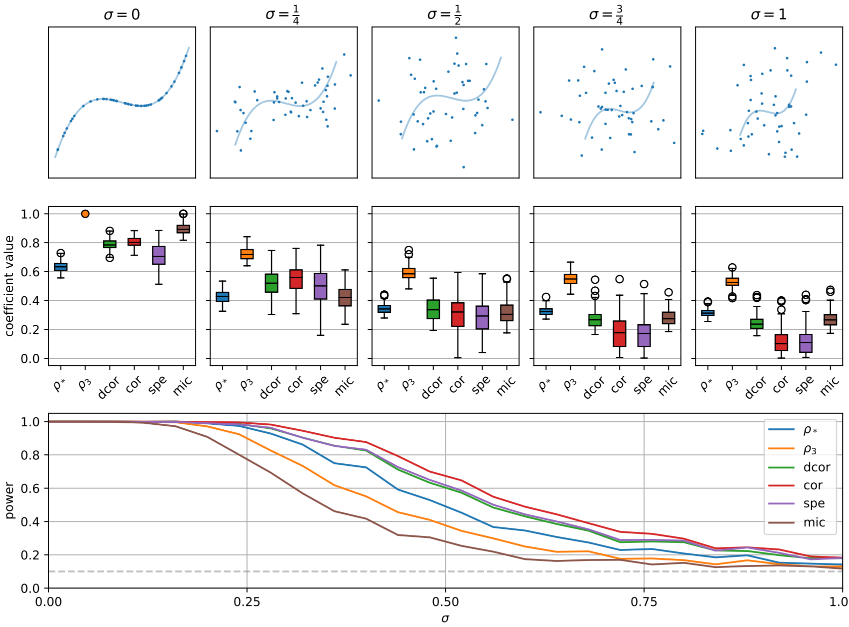


Figure 20

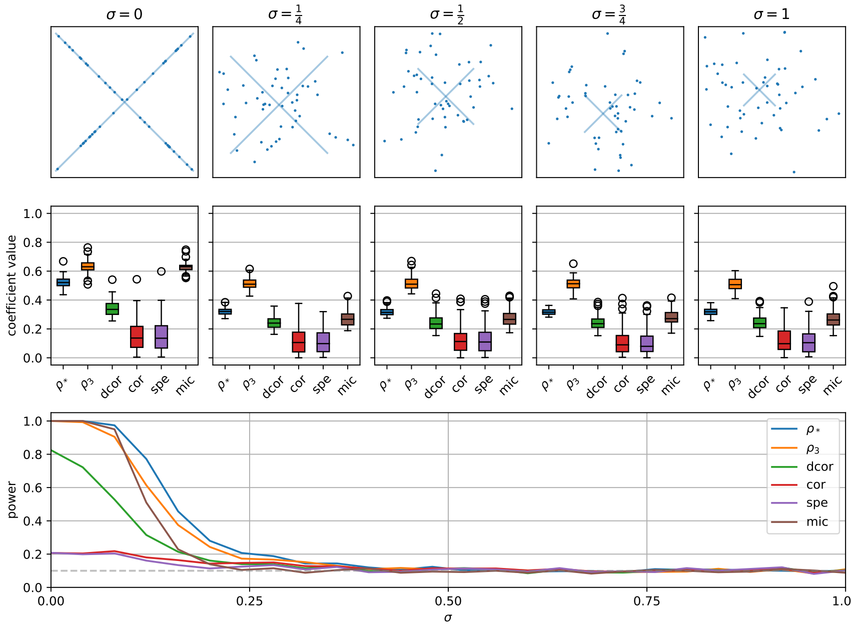


Figure 21

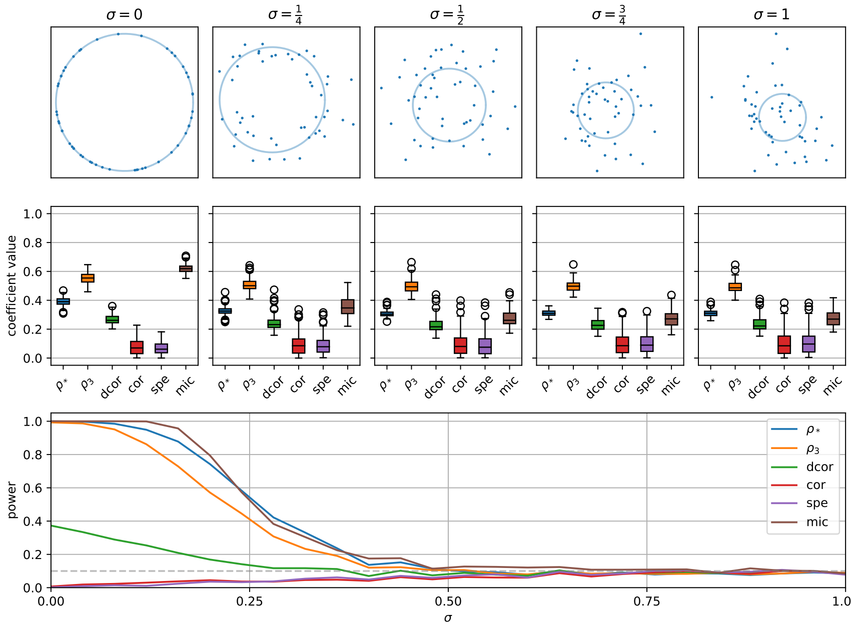


Figure 22

Influence of *p*. Finally, we use the same distributions and noise models as above to study the effect of the parameter p on the transport correlation coefficient ρ_*, see definition (44). From Figure 23 to Figure 27, we compare the values of ρ_* when the parameter p is set equal to 0.5, 1, and 2, respectively. The setting is the same as for the previous comparisons, the only difference being that just 9 (instead of 29) random permutations are used for the permutation tests. One observation that holds for all geometries γ is that the box plots of the estimates of τ are generally very similar for all p. In contrast, we observe that the parameter $p = 0.5$ often provides a (slightly) higher discriminative power against independence than the choices $p = 1$ or $p = 2$.

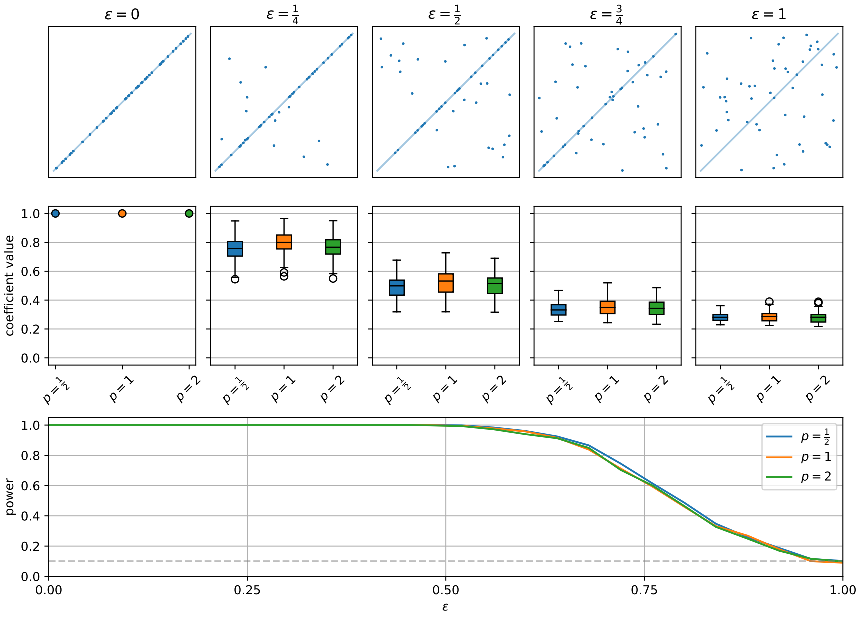


Figure 23

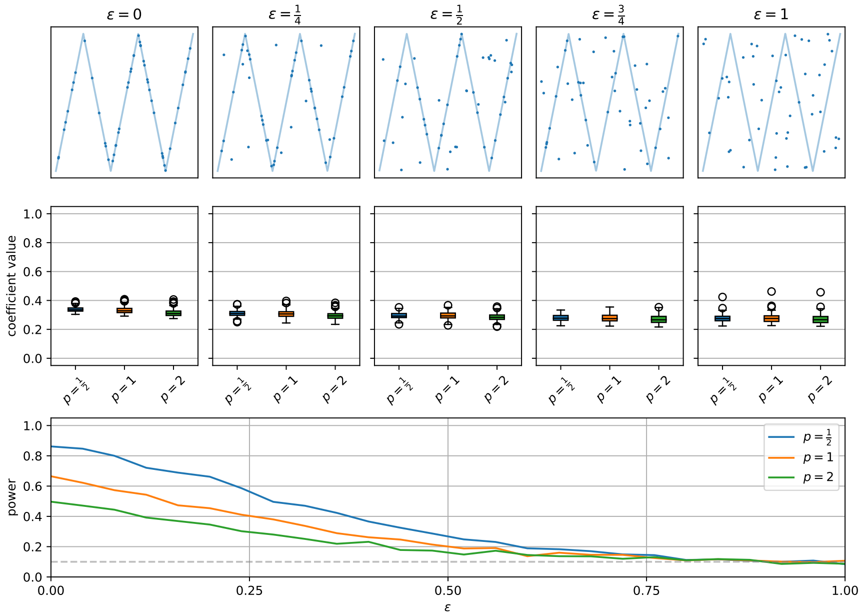


Figure 24

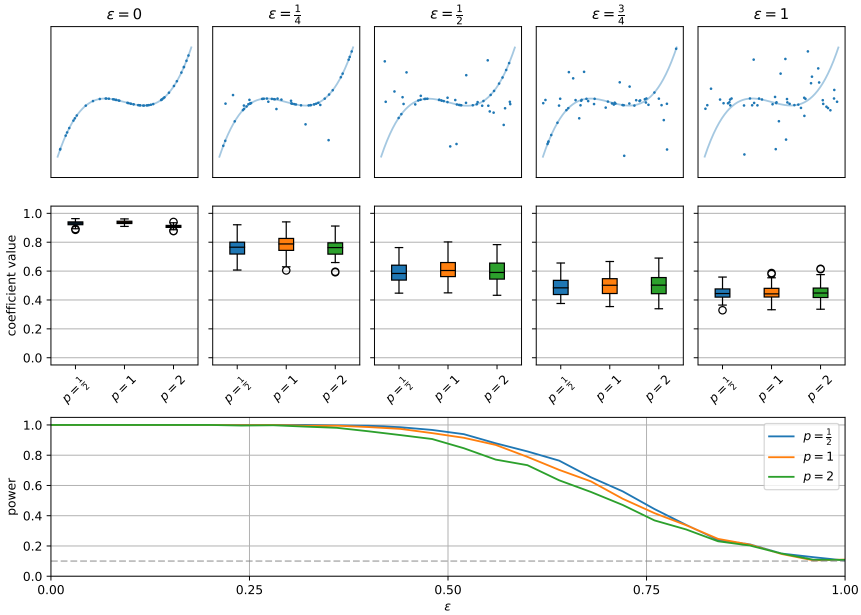


Figure 25

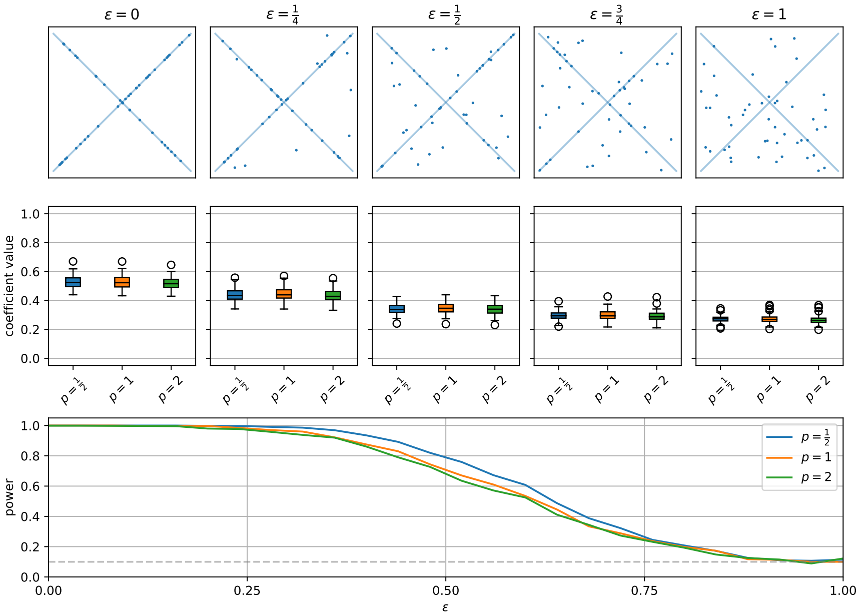


Figure 26

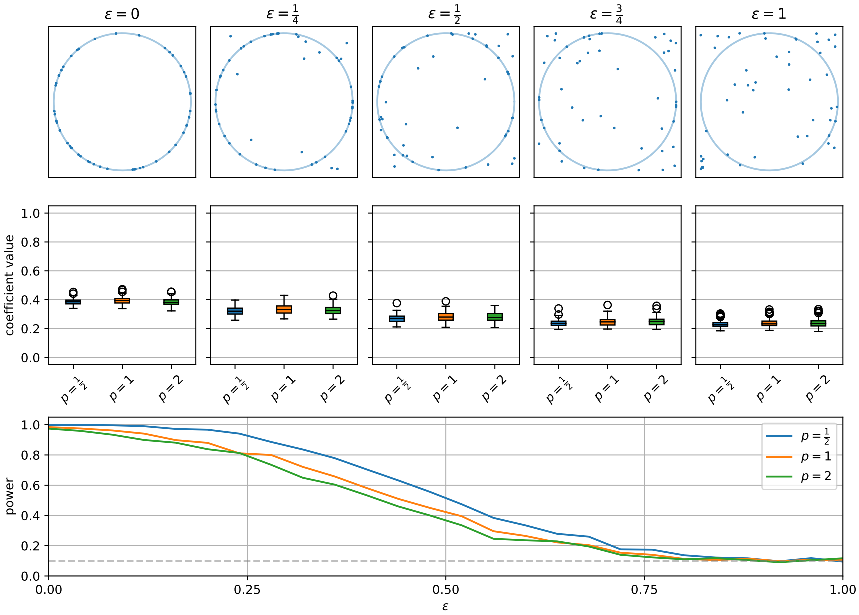


Figure 27