Statistical and Structural Aspects of Unbalanced Optimal Transport Barycenters

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
“Doctor rerum naturalium”
der Georg-August-Universität Göttingen

im Promotionsprogramm
“Mathematical Sciences (SMS)”
der Georg-August University School of Science (GAUSS)

vorgelegt von
Florian Heinemann
aus Wolfenbüttel

Göttingen, 2022

Thesis Committee:

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

Dr. Yoav Zemel
Statistical Laboratory, University of Cambridge

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

Members of the Examination Board:

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

Second Reviewer:
Prof. Dr. Bernhard Schmitzer
Institute of Computer Science, University of Göttingen

Further Members of the Examination Board:

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

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

Prof. Dr. Anja Sturm
Institute for Mathematical Stochastics, University of Göttingen

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

Date of the Oral Examination: September 12, 2022

"Wisdom is the offspring of suffering and time."

Izaro Phrecius – Path of Exile

Preface

Optimal transport (OT) has seen a stellar rise in interest and relevance in the past two decades. Its origins go back to Gaspard Monge in the 18th century with its modern formulation devised by Leonid Kantorovich in the first half of the last century. However, the twenty-first century came with fascinating new theoretical insights for OT, a deeper understanding of its statistical properties and massive advancements in computational methods driven by the growing machine learning communities. These advancements were underlined by the two Fields medallists Cedric Villani and Alessio Figalli in this area of research.
Following the rise of OT, a better understanding of more sophisticated concepts in relation to OT, such as dynamic formulations, geometric analysis, variational formulations such as Fréchet means (also known as barycenters) as well as the curvature and geodesic structure of the space OT defines has been achieved. As these ideas forged their path into modern data analysis and the natural sciences, the questions of estimation and computation grew in relevance. These statistical questions have spawned a beautiful collection of research treating deviation bounds, minimax-rates, central limit theorems and much more in a large variety of settings. The computational research has created a gigantic tool box of methods and algorithms which tackle OT and its variational problems efficiently. The most popular one is the idea of entropy-regularised OT, where a surrogate problem based on an added entropic penalty term is solved instead. The ease of implementation as well as parallelisation made the entropy approach a great tool for machine learning applications and greatly helped OT to popularity in this field.

More recently, severe limitation of OT have started to surface. Two key factors prevent it from becoming a standard tool in general data science applications. The first one is the fact that the era of big data and steadily improving measurement techniques in the natural sciences produce large scale data which is still out of reach for even modern state-of-the-art OT solvers. This problem is compounded for other quantities, such as barycenters, derived from OT, as their computation usually involves solving a large number of individual OT problems. The second limitation which prevents the

reasonable application of OT in several areas is that vanilla OT is only defined between measures of equal total mass intensity (usually probability measures). Foregoing this assumption requires to choose a way of treating mass differences between the measures. This has given birth to a new field of unbalanced optimal transport (UOT) concepts which have also revived some earlier ideas conceived in the last century, partially even by Leonid Kantorovich himself. However, at this point in time UOT lacks the maturity of its balanced counterparts and a significant number of questions remain open.

At the heart of this thesis lies the goal to advance research on OT to allow it to become a standard tool in modern data analysis. To achieve this, this thesis provides contributions to the research on both aforementioned limitations. OT barycenters in particular show promising features for a wide range of geometric data analysis problems, but are hindered by the significant computational burden of solving it. Randomised algorithms allow for fast, approximate computations of large scale problems on personal computers without the need to make use of high-performance-computing facilities. This enables data exploration pipelines where potentially interesting results can be found in an initial investigation at limited approximation quality. Datasets which have been identified as interesting in this manner, can then still be treated at higher accuracy with the appropriate resources. Simultaneously, the same results also provide an understanding of the error which occurs when the population quantities are unknown and have to be estimated from data. This provides some groundwork for statistical inference based on OT barycenters. This thesis treats suitable deviation bounds for OT barycenters in this context.
However, while the geometrical properties of the OT barycenters are superior to those of linear means, there are still scenarios where they fail to provide any meaningful geometric insight on a given dataset. Barycenters based on UOT improve on this by offering greater structural flexibility and robustness compared to their balanced counterparts. Though, reasonable data analysis based on a specific notion of UOT requires a good understanding of its structural properties. Additionally, due to the unbalanced nature of the setting, statistical modelling in this context requires alternative approaches. This work makes advances on both of these matters for a specific version of UOT.

This thesis is based on the key results from the three articles Heinemann et al. [2022b], Heinemann et al. [2021] and Heinemann et al. [2022a] found in the Addenda listed as A, B and C, respectively. Chapter 1 provides a general overview over the theory of OT. It introduces the OT problem and its related Wasserstein barycenter problem. Starting with the basic theory of linear programming, it also gives a short overview of a range of computational methods for OT problems and particularly the Wasserstein

barycenter problem. Chapter 2 is based on the results in Heinemann et al. [2022b]. It presents non-asymptotic statistical deviation bounds on the optimal value of the Wasserstein barycenter problem as well as the setwise distance between the set of population level barycenters and its empirical counterparts. It also introduces a randomised algorithm which takes advantage of the specific structure of the problem for empirical measures. The statistical deviation bounds then allow to control the error of the randomised method, allowing a controlled trade-off between computational complexity and runtime. Chapter 3 contains the results of Heinemann et al. [2021] and Heinemann et al. [2022a]. The first part of the chapter treats a specific version of UOT and the (p, C)-Kantorovich-Rubinstein distance (KRD) based on it. Structural and geometric properties of the distance as well as the transport plan in dependence on the penalty C are derived. For ultrametric trees a closed-form solution of the (p, C)-KRD is established. Barycenters with respect to the (p, C)-KRD are considered and a detailed analysis of their properties and in particular their support set is provided. These barycenters are compared to their Wasserstein counterparts as well as to alternative notions of UOT barycenters. Notably, there is a clear geometric connection between the penalty C and the properties of corresponding UOT plans or barycenters. This easy and intuitive understanding allows to adapt the (p, C)-KRD to specific datasets and highlights it as a prime candidate for data analysis. To build on this, the second part of Chapter 3 deals with statistical modelling and deviation bounds for the (p, C)-KRD. Since there is no unique, well-defined approach to sample from general measures, a framework which allows to analyse a range of potential statistical models is introduced. Within this context three specific models motivated by computational approaches and tasks in microscopy are analysed. For these models sharp non-asymptotic deviation bounds for the (p, C)-KRD and its barycenter are provided. In particular, these bounds again allow for a trade-off between approximation quality and computational complexity in randomised algorithms for both quantities.

Own Contributions

- Heinemann et al. [2022b] (Addendum A) was written jointly with **Y. Zemel** and **A. Munk**. **Y. Zemel** and I contributed equally to the theoretical results presented in the publication. The algorithmic considerations and simulation study are mostly my own contribution. **A. Munk** provided the initial idea for the work as well as many helpful comments and suggestions.

- Heinemann et al. [2021] (Addendum B) was written jointly with **M. Klatt** and **A. Munk**. **M. Klatt** and I contributed equally to the theoretical results presented in the publication. The numerical analysis of the barycenters is mostly my own contribution. **A. Munk** provided many helpful comments and suggestions.

- Heinemann et al. [2022a] (Addendum C) was written jointly with **M. Klatt** and **A. Munk**. **M. Klatt** and I contributed equally to the theoretical statements. The subsequent simulation studies are mostly my own contribution. **A. Munk** provided the ideas for the considered statistical models as well as many helpful comments and suggestions.

Acknowledgements

My supervisor **Axel Munk** has my absolute gratitude for introducing me to the fantastic and ever-expanding world of optimal transport and for the opportunities to be first a student assistant in his group and later his Ph.D. student. His ideas have always provided interesting direction of research and his guidance has shaped my approach to mathematics. I am also extremely grateful to **Yoav Zemel** for his cooperation, guidance and excellent mathematical knowledge. Even though he had his responsibilities as a father of two young children which were at home with him for larger parts of our joint work, due to the pandemic, he still always found time to answer my messages and be it at midnight. I also want to thank **Dominic Schuhmacher** for his thoughtful comments and discussions. He always found time for me, even when I inevitably went over our originally proposed meeting timeframe.

I am particularly thankful to **Marcel Klatt** for his close collaboration, mathematical curiosity, countless discussions and structured approach to mathematics. He has served as my role model for a well-structured and organised approach to my studies, since when I was still a Bachelor student and I greatly enjoyed our joint work in studies and research.

I believe a productive, friendly and supportive working environment to be critical for successful work and progress. Unfortunately, I spent nearly half of my PhD working from home, due to the global pandemic. This is not how I imagined my time as a PhD student, however, I would still like to thank all members of the institute of mathematical stochastics for the great atmosphere and insightful discussions during the time I was actually there. Special thanks in this regard go to **Christoph Weitkamp**, **Thomas Staudt**, **Shayan Hundrieser** and **Giacomo Nies**.

Thanks to the **Research Training Group 2088** for the financial support and the workshops and colloquia which were a great venue for new mathematical insights and fruitful discussions.

I also want to take this opportunity to thank **Christian Böhm** for resolving any technical issues I have been struggling with and to apologise for all the times I caused server issues with my simulations in the past few years.

I found sport to be a key factor in balancing out the long hours at the desk during my studies. A special thanks goes to all my team members in the **Jugger community in Göttingen** throughout the years.

Studying mathematics and in particular pursuing a Ph.D. in it can be a daunting task. This is a long journey one should not take alone. I would like thank the friends who have accompanied me on this journey from the first year to the last: **Germaine Ahrend, Leonard Aue, Erik Bertok, Marie Gutberlet, Tobias Heinrich, Marius Herold, Constantin Hilbrunner, Ricarda Kuntze, Alexander Monecke, Fenna Müller, Johann Priehs, Marike Schwickardi, David Seck, Julian Soltau** and **Hannah Strauch**.

Finally, I want to thank my beloved girlfriend **Lisa Harmsen** for everything. I would not have made it this far without your never-ending support and encouragement. You are my shoulder to lean on and my sunshine on a rainy day. Wherever you are is my home.

List of Symbols

$\mathbb{N}$	Set of natural numbers without zero
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of positive real numbers
$\mathbb{R}^d$	Euclidean space of dimension d
$\mathbb{Z}$	Set of integers
$\|\cdot\|_p$	Norm of order p on $\mathbb{R}^d$
$\mathcal{P}(\mathbb{R}^d)$	Set of probability measures on $\mathbb{R}^d$
$T\#\mu$	Pushforward measure of μ under T
$\mathrm{supp}(\mu)$	Support of the measure μ
$\Pi(\mu, \nu)$	Set of couplings of μ and ν
$\mathrm{W}_p(\mu, \nu)$	The p-th order Wasserstein distance between μ and ν
$F_p(\mu)$	Value of the p-Fréchet functional at μ
$\Pi(\mu_1, \ldots, \mu_J)$	Set of multi-couplings of $\mu_1, \ldots, \mu_J$
ρ_i	The i-th coordinate projection
$A \bigotimes B$	Kronecker product of matrices A and B
$\mathrm{diam}(\mathcal{X})$	Diameter of the space $\mathcal{X}$
$\mathbb{M}(\mu)$	Total mass intensity of μ
$\mathrm{UOT}_{p,C}(\mu, \nu)$	Cost of unbalanced transport between μ and ν
$\mathrm{KR}_{p,C}(\mu, \nu)$	The (p, C)-Kantorovich-Rubinstein distance between μ and ν
$\mathcal{L}(P)$	The length of a path P within a graph
$\mathrm{TV}(\mu, \nu)$	Total variation distance between μ and ν
$\mathcal{M}_+(\mathcal{Y})$	Set of positive measures on $\mathcal{Y}$
$\mathrm{med}(x_1, \ldots, x_J)$	Median of $x_1, \ldots, x_J$
$a \wedge b$	Minimum of a and b
$X \sim \mu$	A random variable X sampled from μ
$Ber(s)$	Bernoulli distribution with success probability s
$Poi(\lambda)$	Poisson distribution with expectation λ
$\mathcal{N}(\mathcal{Y}, \varepsilon)$	The ε-covering number of $\mathcal{Y}$
d_p	The p-distance on $\mathbb{R}^d$

Contents

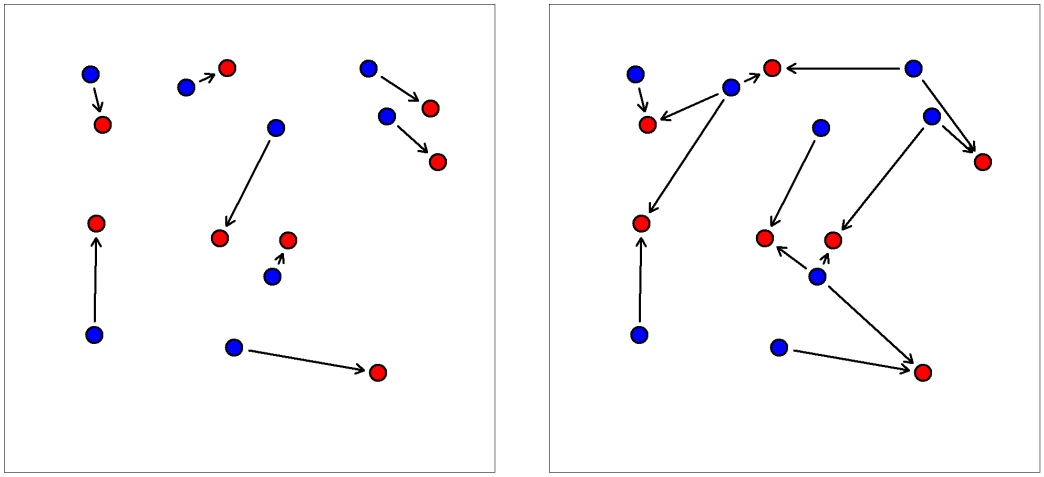
CHAPTER 1

Introduction

The earliest known formulation of OT goes back to the French mathematician Gaspard Monge [Monge, 1781]. He was interested in determining a cost optimal way of transporting soil, which had been extracted from the ground in certain locations to construction sites at different locations where this material was needed. Formally, consider two probability measures $\mu^1, \mu^2 \in \mathcal{P}(\mathbb{R}^d)$ which encode the amount of extracted soil available and the required soil, respectively, and a *cost* function $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ specifying the transport cost of a unit of mass between two given locations. The aim is to find a function $T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ minimising

$$\int_{\mathbb{R}^d} c(x, T(x)) d\mu^1(x) \tag{1.1}$$

among all functions T in the set $\{T : \mathbb{R}^d \longrightarrow \mathbb{R}^d \mid T\#\mu^1 = \mu^2\}$. The elements of this set are called *transport maps* and any minimiser in (1.1) is referred to as an *OT map*. The condition $T\#\mu^1 = \mu^2$ ensures that the entirety of the soil is transported, as well as that all demand of soil is satisfied. In particular, for measures of different total intensities the set of OT maps is empty. While this approach already encapsulates the essential concepts at the heart of OT, it might be ill-posed in many, even quite simple, examples. To illustrate this issue consider Figure 1.1. In part a), both measures have eight support points with mass $1/8$ at each location. Consequently, it is possible to find the OT map by picking the most cost-efficient pairwise assignment between the points. However, when considering part b), this approach fails. Here, one measure still has eight support points with mass $1/8$, but the other one has seven support points with mass $1/7$, respectively. It is immediately clear that in this scenario, the set of transport maps is empty. Though, of course it is still possible to transport mass between the two measures. The key ingredient which is missing is the ability to split mass, i.e. for a given location $x \in \text{supp}(\mu^1)$, to be allowed to send parts of its mass to different locations in $\text{supp}(\mu^2)$. This does not provide an OT map, however, it leads to finding an



(a) (b)

Figure 1.1: The OT plans between finitely supported measures (displayed in red and blue, respectively) in $[0,1]^2$ with respect to squared Euclidean cost. **(a):** Both measures have eight support points with mass $1/8$ each. **(b)** The red points have mass $1/7$ while the blue ones have mass $1/8$.

OT plan which gives rise to the more flexible OT formulation first considered by Leonid Kantorovich [Kantorovich, 1942].

1.1 Optimal Transport and the Wasserstein Distance

The so called Kantorovich relaxation can be formulated as follows. Let $(\mathcal{X}, d)$ be a metric space and let $\mu^1, \mu^2 \in \mathcal{P}(\mathcal{X})$ be probability measures on $\mathcal{X}$. Define the set of *couplings* between μ^1 and μ^2 as

$$\Pi(\mu^1, \mu^2) = \{\pi \in \mathcal{P}(\mathcal{X}^2) \mid \pi(A, \mathcal{X}) = \mu^1(A), \pi(\mathcal{X}, B) = \mu^2(B) \text{ for } A, B \in \mathcal{B}(\mathcal{X})^1\}.$$

Instead of finding an OT map, the goal is now to find an optimal coupling between μ^1 and μ^2, i.e. to find a coupling π in

$$\underset{\pi \in \Pi(\mu^1, \mu^2)}{\arg\min} \int_{\mathcal{X}^2} c(x, y) d\pi(x, y). \tag{1.2}$$

To distinguish the two problems, (1.1) is referred to as the *Monge problem* and (1.2) is referred to as the *Kantorovich problem*.
In the following and throughout all of this thesis the focus is on OT between finitely

[1] Here, $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra on $(\mathcal{X}, d)$.

supported measures. However, for reasons which will later become apparent when discussing OT-barycenters, it is still assumed that there exists a general (usually connected and geodesic), metric *ambient space* $(\mathcal{Y}, d)$[2] which contains the finite supports of the considered measures. Thus, let $\mu^1 = \sum_{k=1}^{M_1} a_k^1 \delta_{x_k^1}$ and $\mu^2 = \sum_{k=1}^{M_2} a_k^2 \delta_{x_k^2}$ be finitely supported probability measures on $\mathcal{Y}$. Their supports are denoted by $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Each coupling between μ^1 and μ^2 can be identified with a $M_1 \times M_2$ matrix and the set of all couplings between μ^1 and μ^2 is denoted as

$$\{\pi \in \mathbb{R}^{M_1 \times M_2} \mid \pi \mathbf{1}_{M_1} = a^1, \pi^T \mathbf{1}_{M_2} = a^2\}.$$

Correspondingly, the cost function can also be seen as a *cost matrix* $C \in \mathbb{R}^{M_1 \times M_2}$ where $C_{kl} = c(x_k^1, x_l^2)$. Using this, the Kantorovich problem can be rewritten as

$$\min_{\pi \in \Pi(\mu^1, \mu^2)} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} C_{kl} \pi_{kl}. \tag{1.3}$$

Often, the cost matrix is defined in terms of the power of a distance d, i.e. $C_{kl} = d(x_k^1, x_l^2)^p$ for some $p \geq 1$. A key observation is that this choice of cost allows to define a distance on the space of probability measures. More precisely, let $p \geq 1$, then the *p-Wasserstein distance*[3] W_p between μ^1 and μ^2 is defined as

$$\mathrm{W}_p(\mu^1, \mu^2) = \left(\min_{\pi \in \Pi(\mu^1, \mu^2)} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} d(x_k^1, x_l^2)^p \pi_{kl} \right)^{\frac{1}{p}}. \tag{1.4}$$

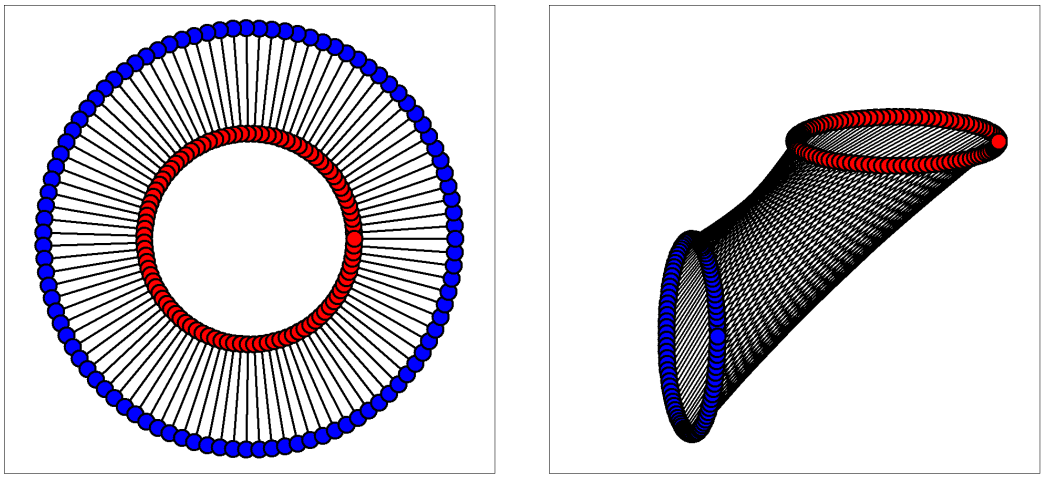
1.2 Wasserstein Geodesics

The solutions of (1.3) (or (1.4) respectively) can not only be understood in a measure theoretical sense as optimal couplings between the two measures, but can also be considered as an OT plan describing a cost optimal way to transform μ^1 into μ^2. The entry π_{kl} then specifies the amount of mass which should be moved from x_k^1 to x_l^2 according to the plan π. Illustrations of such plans are found in Figure 1.1 and Figure 1.2. Based on this, it would be natural to ask how such a transport or transformation from one measure to the other is realised. This question leads to the notion of *Wasserstein geodesics*.

Let $(\nu_t)_{t \in [0,1]}$ be a *continuous curve* in $\mathcal{P}(\mathcal{Y})$, i.e. the map $t \mapsto \nu_t$ from $([0,1], |\cdot|)$ to

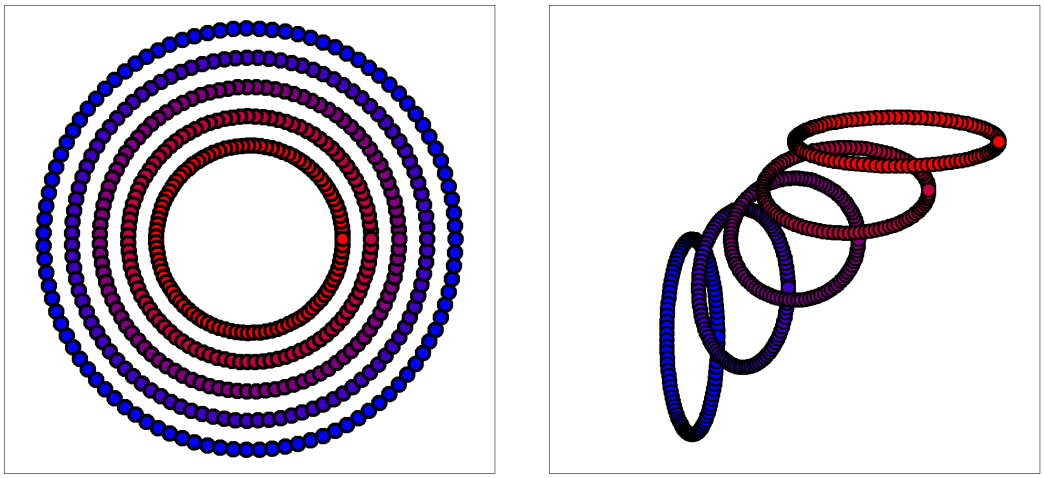
[2] The distance on $\mathcal{X}$ is assumed to be the restriction of the distance on $\mathcal{Y}$ to $\mathcal{X}$.

[3] For a rigorous treatment and proof that this is indeed a distance see Villani [2009].



(a) (b)

Figure 1.2: The OT plans with respect to W_2 between finitely supported measures (displayed in red and blue, respectively) in $([0,1]^2, d_2)$. **(a)** The measures are supported on a two nested ellipses discretised onto 100 points, respectively. Each measure has mass $1/100$ at each location. **(b)** The two measures supported on two different ellipses discretised onto 100 points, respectively. Each measure has mass $1/100$ at each location.



(a) (b)

Figure 1.3: **(a)** A W_2-geodesic between the two measures in Figure 1.2(a) at the time points $t = 0, 0.25, 0.5, 0.75, 1$. The colour of a point indicates to which time point it belongs (from blue for $t = 0$ to red for $t = 1$). **(b)** A W_2-geodesic between the two measures in Figure 1.2(b) at the time points $t = 0, 0.25, 0.5, 0.75, 1$. The colour of a point indicates to which time point it belongs (from blue for $t = 0$ to red for $t = 1$).

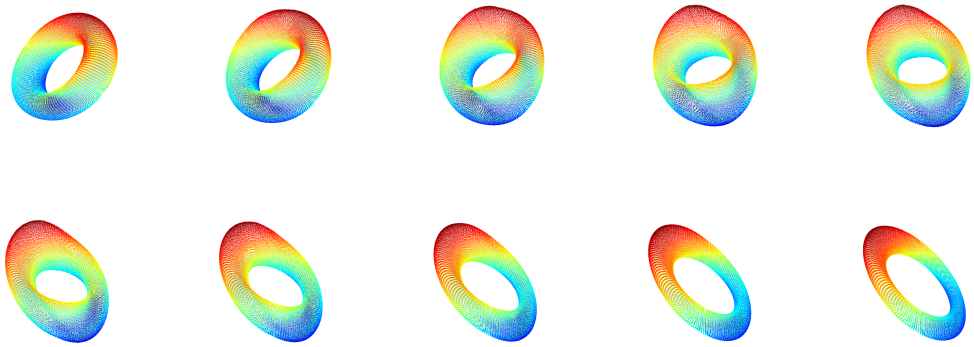


Figure 1.4: A W_2-geodesic between two measures supported on different tori in $([0,1]^3, d_2)$ at the time points $t = 0, 1/9, 2/9, \ldots, 1$ from top-left to bottom-right. Each support point has mass $1/10000$. The colour coding only conveys the three dimensional shape. It is not indicative of any mass of the measures.

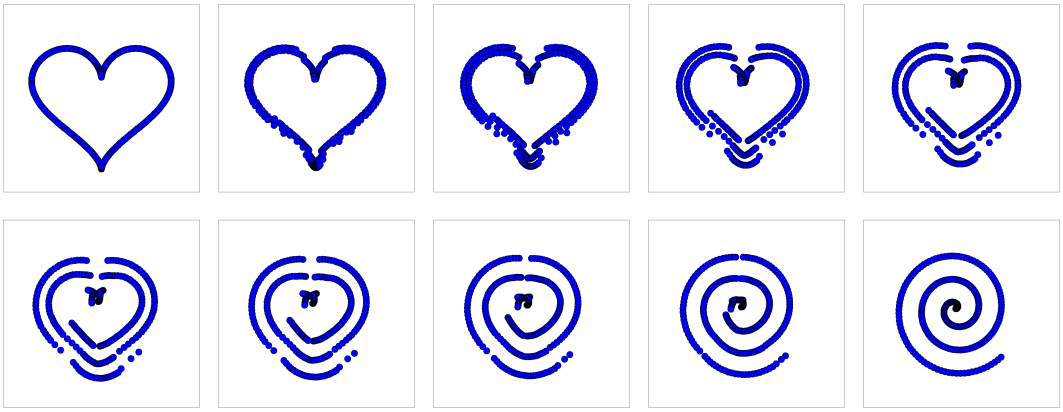


Figure 1.5: A W_2-geodesic between two measures supported on a heart and a spiral in $([0,1]^2, d_2)$ at the time points $t = 0, 1/9, 2/9, \ldots, 1$ from top-left to bottom-right. Each support point has mass $1/1000$.

$(\mathcal{P}(\mathcal{Y}), W_p)$ is continuous. It is called a *(constant speed) geodesic* w.r.t. W_p between μ^1 and μ^2, if $\nu_0 = \mu^1$, $\nu_1 = \mu^2$ and for all $t, s \in [0,1]$ it holds

$$W_p(\nu_t, \nu_s) = |t - s| W_p(\nu_0, \nu_1). \tag{1.5}$$

Intuitively, the curve $(\nu_t)_{t\in[0,1]}$ describes a *locally shortest path* between μ^1 and μ^2 with respect to the geometry of the metric space $(\mathcal{P}(\mathcal{Y}), W_p)$. A specific time point ν_t can be understood as a W_p-interpolation of μ^1 and μ^2 with weight $1 - t$ for μ^1 and t for μ^2. Recall the example OT plan from Figure 1.2(b). Following the trajectory of the corresponding geodesic (see Figure 1.3(b)), it can be seen that the ellipse is being squeezed together, while it simultaneously moves from bottom-left to the top-right.

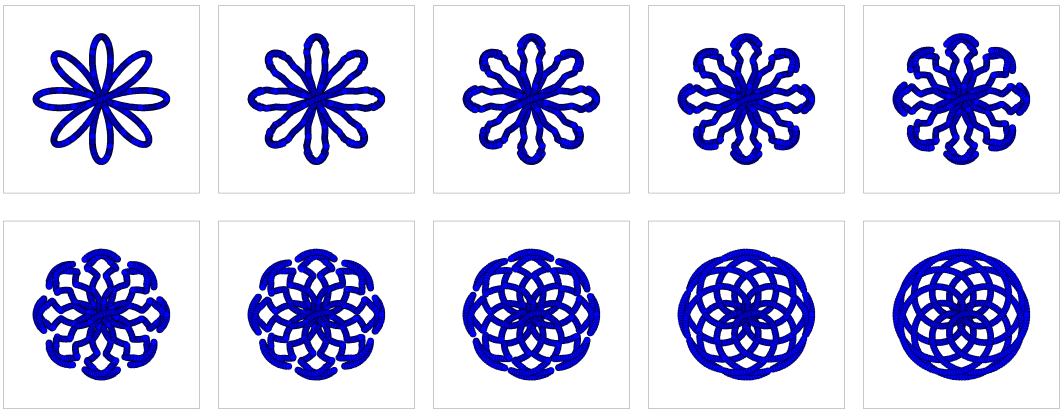
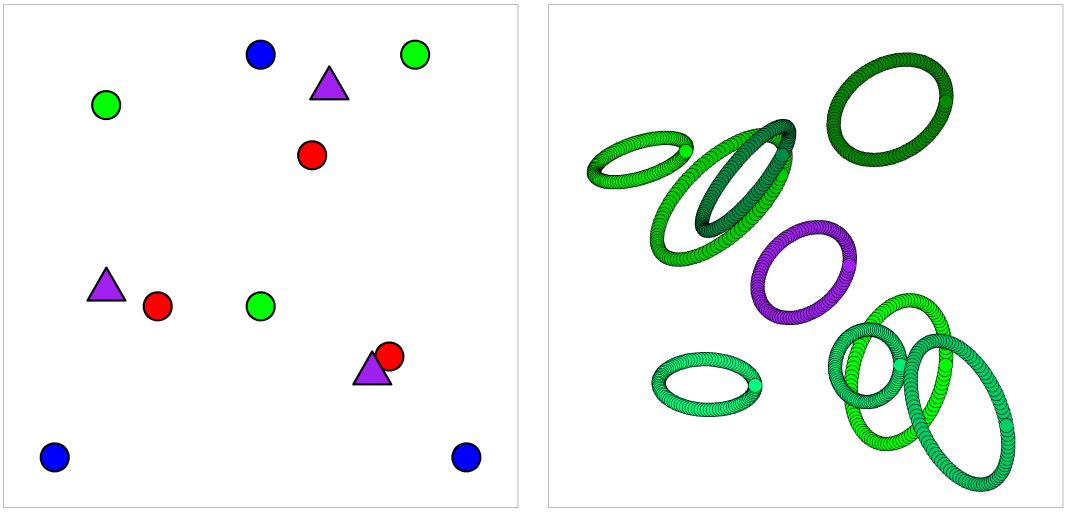


Figure 1.6: A W_2-geodesic between two measures supported on a flower petal and a mosaic in $([0,1]^2, d_2)$ at the time points $t = 0, 1/9, 2/9, \ldots, 1$ from top-left to bottom-right. Each support point has mass $1/1000$.

Around the middle of the curve, it is a near perfect circle with its center being the mean of the centers of the two ellipses being interpolated. Similarly, for OT plan in Figure 1.2(a) it can be seen in Figure 1.3(a) that on the trajectory of the corresponding geodesic, the outer ellipse shrinks until it reaches the size of the inner one.
It should be noted, that, even when computing OT between two measures whose support sets belong to the same class of geometric objects (e.g. triangles, stars, tori), the time points induced by their respective geodesics do not necessarily belong to this class. For this it suffices to consider measures supported on discretized tori in $\mathbb{R}^3$ in Figure 1.4. Here, while each point of the geodesic still roughly resembles a torus, the transport still creates bulges and dents in the support structure. It should also be noted, that in general the support might "break up" into smaller parts instead of remaining as one, essentially connected, structure during the whole transformation. This can be seen in Figure 1.5, where the heart "splits" into multiple parts which only join together at the end of the geodesic to form different parts of the spiral. Though, such splits do not necessarily occur even when transporting between noticeably different shapes such as the flower and mosaic in Figure 1.6. Here, the leafs of the flower start to deform until they form the mosaic structure of the target.

Notably, in these examples 2-Wasserstein geodesics for the space $([0,1]^2, d_2)$ are considered, i.e. the cost of the corresponding OT problem is the squared euclidean distance. This is a common choice for the cost as the transport induced by it promotes satisfying demand in local neighbourhoods since large scale transports are penalised heavily by the quadratic cost term. Additionally, the resulting transports and corresponding geodesics tend to coincide with human intuition for a deformation between the source and target



(a) (b)

Figure 1.7: **(a)** Three measures on $([0,1]^2, d_2)$ with mass $1/3$ at each location (displayed by red, green or blue circles, respectively) and their 2-Wasserstein barycenter displayed by purple triangles with mass $1/3$ each. **(b)** Eight measures supported on different ellipses in $([0,1]^2, d_2)$ displayed in different shades of green. Each location has a mass of $1/100$. The 2-Wasserstein barycenter of these measures is displayed in purple and has mass $1/100$ at each support point.

measure. This allows for geometrical meaningful data analysis in this context.

1.3 Wasserstein Barycenters

If the transportation process is stopped in the middle, i.e. $\nu_{0.5}$ is considered, then this yields a notion of a p-Wasserstein midpoint between μ^1 and μ^2. This object could then be understood as the mean of these two measures with respect to the W_p geometry. Notably, this definition of a mean remains limited to sets of size two. However, it can be generalised to sets of J measures $\mu^1, \ldots, \mu^J \in \mathcal{P}(\mathcal{Y})$ by considering the following alternative characterisation of $\nu_{0.5}$:

$$\nu_{0.5} \in \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathcal{Y})} W_p^p(\mu^1, \nu) + W_p^p(\mu^2, \nu). \tag{1.6}$$

This can be seen easily by noting that since ν is a geodesic it holds

$$W_p^p(\mu^1, \nu_{0.5}) + \mathrm{W}_p^p(\mu^2, \nu_{0.5}) = 2^{1-p}\mathrm{W}_p^p(\mu^1, \mu^2) \tag{1.7}$$

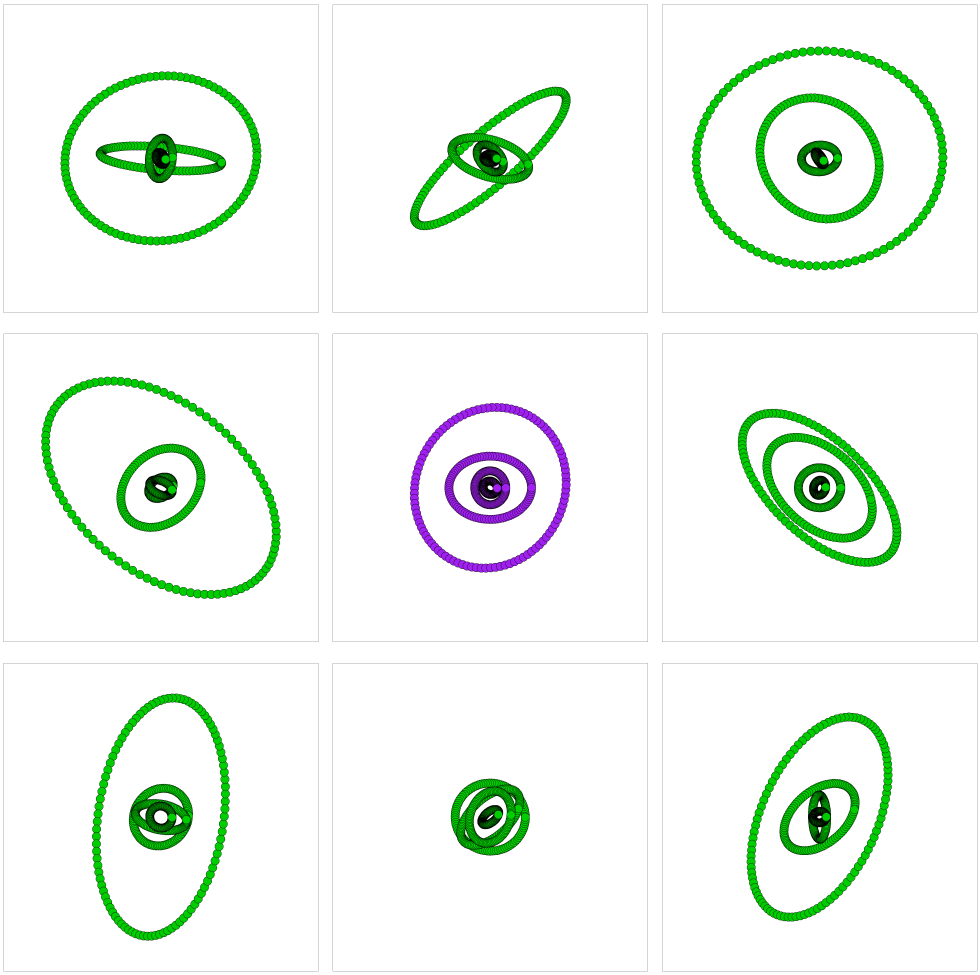


Figure 1.8: Eight measures supported on nested ellipses in $([0,1]^2, d_2)$ in green. Each support point has mass $1/400$. In the center the 2-Wasserstein barycenter with mass $1/400$ at each location is displayed in purple.

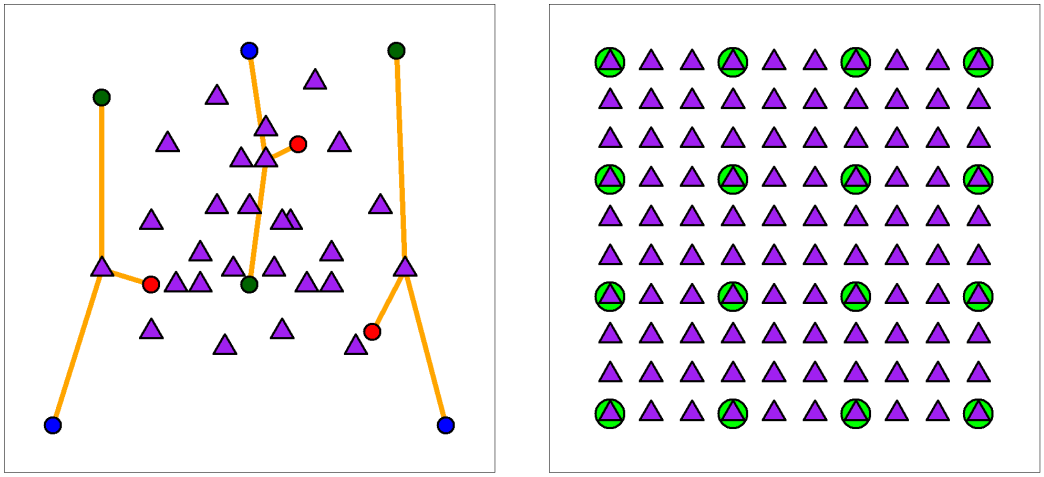
and for any measure $\rho \in \mathcal{P}(\mathcal{Y})$ it holds (due to the triangle inequality and the basic inequality $(a+b)^p \leq 2^{p-1}(a^p + b^p)$ for all $a, b \in \mathbb{R}_+$) that

$$\mathrm{W}_p^p(\mu^1, \mu^2) \leq (\mathrm{W}_p(\mu^1, \rho) + \mathrm{W}_p(\mu^2, \rho))^p \leq 2^{p-1}(\mathrm{W}_p^p(\mu^1, \rho) + \mathrm{W}_p^p(\mu^2, \rho)). \tag{1.8}$$

Hence, it holds for any $\rho \in \mathcal{P}(\mathcal{Y})$ that $\mathrm{W}_p^p(\mu^1, \rho) + \mathrm{W}_p^p(\mu^2, \rho) \geq 2^{1-p}\mathrm{W}_p^p(\mu^1, \mu^2)$, thus, due to the equality in (1.7), $\nu_{0.5}$ is a minimiser of (1.6).

The formulation in (1.6) easily generalises to more than two measures. Let $\mu^1, \ldots, \mu^J \in \mathcal{P}(\mathcal{Y})$ be measures with finite support sets $\mathcal{X}_1, \ldots, \mathcal{X}_J$ and respective cardinalities $M_1, \ldots, M_J$. Let $F : \mathcal{P}(\mathcal{Y}) \longrightarrow \mathbb{R}$ be given by

$$F_p(\mu) = \frac{1}{J} \sum_{i=1}^{J} \mathrm{W}_p^p(\mu, \mu^i). \tag{1.9}$$



(a) (b)

Figure 1.9: **(a)** The support sets of three measures in $([0,1]^2, d_2)$ displayed by red, green and blue circles. The corresponding centroid set in (1.11) is displayed by purple triangles. The orange lines indicate which centroid points have been constructed from which support points for three exemplary points. **(b)** The support sets of four measures in $([0,1]^2, d_2)$ supported on the same equidistant 4×4 grid. The corresponding centroid set in (1.11) is given by a 10×10 grid and displayed by purple triangles.

The function F_p is referred to as *p-Fréchet functional* and any minimiser within $\mathcal{P}(\mathcal{Y})$ is called a *p-Wasserstein barycenter*. Some examples are found in Figure 1.7 and Figure 1.8. Note, that such barycenters exist under mild conditions [Le Gouic and Loubes, 2017], but in general are not unique[4]. The weights of $1/J$ at each summand can be easily generalised to an arbitrary sum of positive values $\lambda_1, \ldots, \lambda_J$, summing to 1. This has no significant impact on any of the results discussed later, thus it is avoided for brevity and simplicity of presentation in this thesis.

Critically, even though the space $\mathcal{Y}$ is not necessarily finite, any p-Wasserstein barycenter of the finitely supported measures $\mu^1, \ldots, \mu^J$ has a finite support. In particular, define the *centroid set*[5]

$$C = \left\{ \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{J} d^p(y, x_i) \;\middle|\; x_i \in \mathcal{X}_i,\ \forall i = 1, \ldots, J \right\}, \tag{1.10}$$

[4]Considering e.g. $\mu^1 = (\delta_{(0,-1)} + \delta_{(0,1)})$ and $\mu^2 = (\delta_{(-1,0)} + \delta_{(1,0)})$ it can quickly be seen that any convex combination of $(\delta_{(-0.5,0.5)} + \delta_{(0.5,-0.5)})$ and $(\delta_{(-0.5,-0.5)} + \delta_{(0.5,0.5)})$ is a 2-Wasserstein barycenter of these two measures.

[5]Note, that there might be multiple sets fulfilling this definition, since the argmin in (1.10) is not necessarily unique.

then it holds for any p-barycenter $\mu^\star$ of $\mu^1, \dots, \mu^J$ that $\operatorname{supp}(\mu^\star) \subset C$ [Le Gouic and Loubes, 2017]. If $\mathcal{Y} = \mathbb{R}^d$, d is the Euclidean distance and $p = 2$, then the centroid set simplifies to

$$C = \left\{ \frac{1}{J} \sum_{i=1}^{J} x_i \;\middle|\; x_i \in \mathcal{X}_i,\ \forall i = 1, \dots, J \right\}. \tag{1.11}$$

Two key consequences of this property are that

$$\inf_{\mu \in \mathcal{P}(\mathcal{Y})} F_p(\mu) = \inf_{\mu \in \mathcal{P}(C)} F_p(\mu)$$

and that for any p-Wasserstein barycenter $\mu^\star$ of $\mu^1, \dots, \mu^J$ it holds

$$\mu^\star = \sum_{x \in C} a_x \delta_x.$$

For an illustration of the structure of the centroid set, see Figure 1.9 (a). Note, that (1.11) implies that if all J measures are supported on an equidistant $G_1 \times \cdots \times G_d$ grid in $[0, 1]^d$, then the corresponding centroid set is simply an equidistant $J(G_1 - 1) + 1 \times \cdots \times J(G_d - 1) + 1$ grid in $[0, 1]^d$ (compare Figure 1.9 (b)). Thus, any 2-Wasserstein barycenter in this setting is supported on a (roughly) J-times finer grid. This simplification is based on the fact that due to the identical support sets and structure of the grids, many N-tupels of points construct the same centroid points, thus reducing the overall size of the set. A notable insight from this is the observation that while the centroid set might be of the size of the product of the support sizes in the worst case, in practice it is often significantly smaller. Furthermore, there always exists [Anderes et al., 2016] a barycenter which support size is a most linear in the support sizes of the underlying measures. Specifically, there exists a p-Wasserstein barycenter $\mu^\star$ of $\mu^1, \dots, \mu^J$, such that $|\operatorname{supp}(\mu^\star)| \leq \sum_{i=1}^{J} M_i - J + 1$. For comparison, the size of the centroid set scales at the worst case as $\prod M_i$, which is exponential in the individual support sizes. Hence, there exists barycenters which are only supported on a relatively small subset of C. However, determining these point remains a difficult combinatorial problem. In fact, the problem of finding the sparsest p-Wasserstein barycenter is NP-hard [Altschuler and Boix-Adsera, 2022].

1.4 Multi-Marginal Optimal Transport

Closely related to the notion of p-Wasserstein barycenters is the notion of *multi-marginal optimal transport (MMOT)*. Consider measures $\mu^1, \dots, \mu^J \in \mathcal{P}(\mathcal{Y})$ and a cost function

$c : \mathcal{Y}^J \to \mathbb{R}$. Define $\Pi(\mu^1, \dots, \mu^J) := \left\{\pi \in \mathcal{P}(\mathcal{Y}^J) \mid \rho_i \# \pi = \mu^i\right\}$. The elements of $\Pi(\mu^1, \dots, \mu^J)$ are called *multi-couplings* of $\mu^1, \dots, \mu^J$. The goal of MMOT is to find an optimal multi-coupling with respect to the cost c, i.e.

$$\pi^\star \in \operatorname*{arg\,min}_{\pi \in \Pi(\mu^1, \dots, \mu^J)} \int_{\mathcal{Y}^J} c(x_1, \dots, x_J) d\pi(x_1, \dots, x_J). \tag{1.12}$$

Using the fact that all J measures have finite support sets $X_1, \dots, X_J$, this can be rewritten as

$$\pi^\star \in \operatorname*{arg\,min}_{\pi \in \Pi(\mu^1, \dots, \mu^J)} \sum_{x_1 \in X_1} \cdots \sum_{x_J \in X_J} c(x_1, \dots, x_J) \pi(x_1, \dots, x_J). \tag{1.13}$$

Denote the objective function of (1.13) as G. For specific costs based on d^p, the MMOT problem and the p-Wasserstein barycenter problem are closely related. In particular, for any $p \geq 1$, let $T^{J,p}$ be the *barycentric application* given by

$$T^{J,p}(x_1, \dots, x_J) \in \operatorname*{arg\,min}_{y \in \mathcal{Y}} \sum_{i=1}^{J} d^p(y, x_i)$$

and

$$c(x_1, \dots, x_J) := \sum_{i=1}^{J} d^p(x_i, T^{J,p}(x_1, \dots, x_J)).$$

Thus, the cost of a J-tupel $x_1, \dots, x_J$ is precisely the optimal value of the Fréchet functional associated with these J points and for any optimal multi-coupling $\pi^\star$ and any p-Wasserstein barycenter $\mu^\star$ it holds $F(\mu^\star) = G(\pi^\star)$. Moreover, for any optimal multi-coupling $\pi^\star$, the pushforward measure $T^{J,p} \# \pi^\star$ is a p-Wasserstein barycenter of $\mu^1, \dots, \mu^J$ and for any p-barycenter $\mu^\star$, there exists an optimal multi-coupling $\pi^\star$, such that $\mu^\star = T^{J,p} \# \pi^\star$.[6] As they tend to be geometrically more interesting for many applications, this thesis usually treats the barycentric viewpoint instead of the multi-marginal formulation. However, the MMOT equivalence provides a powerful tool for establishing theoretical properties as well as for computational approaches. For proofs and a more detailed discussion of this equivalence see Chapter 3 of Panaretos and Zemel [2020].

[6] Note, that there might be several multicouplings which create the same barycenter in this manner, since a location x might be the d-barycenter of different points. More details are found in Section 3.1.

1.5 Linear Programs

To utilise OT in statistics or data analysis, it is is necessary to solve it numerically. Unfortunately, the OT problem does not have a closed form solution outside of very specific cases. Thus suitable algorithms to tackle this problem are necessary. Early works on OT [Tolstoi, 1930, Hitchcock, 1941, Kantorovich, 1942] considered a formulation which would later become known as a special case of a general *linear program (LP)*. A thorough treatment of general LP theory and algorithms is found in Bertsimas and Tsitsiklis [1997].

Linear Programming

Recalling (1.3), OT can be considered in the framework of LPs. Before discussing this reformulation, however, a brief description of general LPs is in order. The foundations for the modern theory of linear programming were laid in the 1940's. Following early findings on the OT problem the general formulation of a *linear program (LP)* was introduced by Dantzig [1948]. Other early notable contributions in this field include Koopmans [1951] and Ford and Fulkerson [1956].
The goal of linear programming is to find the solution of a (linearly constrained) optimisation problem posed with respect to a linear objective function. Let $c \in \mathbb{R}^L$ be the *cost vector* of the problem, $A \in \mathbb{R}^{K \times L}$ the *constraint matrix* of full row rank[7] and $b \in \mathbb{R}^K$ the *constraint vector*. For a LP in *standard form* the goal is to solve

$$\begin{aligned} \min_{x \in \mathbb{R}^L} \quad & c^T x \\ \text{s.t.} \quad & Ax = b, \quad x \geq 0. \end{aligned} \tag{1.14}$$

Duality

The minimisation problem in (1.14) is often referred to as the *primal LP*. One key result in LP theory is the existence of a related *dual LP* given by

$$\begin{aligned} \max_{x \in \mathbb{R}^L} \quad & b^T y \\ \text{s.t.} \quad & A^T y \leq c. \end{aligned} \tag{1.15}$$

In particular, duality theory connects the objective values of the primal and dual programs. For any feasible primal solution x and any feasible dual solution y, it holds $b^T y \leq c^T x$. This property is known as *weak duality*. The *strong duality* theorem states

[7] The row rank of a matrix is equal to the dimension of the linear space spanned by its rows.

that if (1.14) has an optimal solution $x^\star$, then (1.15) has an optimal solution $y^\star$ and it holds $c^T x^\star = b^T y^\star$. For a proof and more thorough treatment of LP duality refer to Chapter 4 of Bertsimas and Tsitsiklis [1997].

The Geometry of Linear Programming

Intuitively, the linear constraints in (1.14) define a *constraint polyhedron* $P = \{x \in \mathbb{R}_+^L \mid Ax = b\}$. A point $x \in P$ is called a *vertex* of P if there exists no non-zero vector $v \in \mathbb{R}^L$, such that $x + v, x - v \in P$. Notably, a point $x \in \mathbb{R}$ is a vertex of P if and only if there exist a set of indices $i_1, \dots, i_L$, such that the columns $A_{i_1}, \dots, A_{i_L}$ are linearly independent and $x_i = 0$ for $i \neq i_1, \dots, i_L$. In this sense, vertices are sparse point of $\mathcal{P}$. Since any LP which has a minimiser, can be shown to have one which is a vertex of its constraint polyhedron, this implies in particular the existence of *sparse solutions* to LPs. Note, that this sparsity property for the vertices only implies an upper bound on the number of non-zero components. There might be vertices which have less than L non-zero entries. Such vertices are referred to as *degenerate*. A polyhedron can be fully described in terms of its vertices, more precisely a non-empty, bounded polyhedron is the convex hull[8] of its vertices. As an immediate consequence, it can be seen, by linearity of the objective function, that each optimal solution is a convex combination of optimal vertices and the set of all optimal solutions is a polyhedron again. Its vertices are precisely the set of optimal vertices of P.

Integer LP

A related problem to classical LPs are integer LPs, where the set of feasible solutions is restricted to vectors with only integer components. While these problems in general differ significantly from usual LPs and require different methods to be solved numerically[9], there are specific conditions, which can ensure that a general LP has an integer solution without imposing any further restrictions on the constraint set.
A matrix $A \in \mathbb{R}^{K \times L}$ is called *totally unimodular (TU)* if any of its square sub-matrices

[8] The convex hull of vectors $x_1, \dots, x_R$ is the set of all convex combinations of these vectors.

[9] Classical approaches are for instance the *cutting plane* method, the *branch and bound* algorithm and the *branch and cut* algorithm combining the two former methods. For a detailed discussion of integer LPs see Section 10 of Bertsimas and Tsitsiklis [1997].

has determinant 1,−1 or 0. For a simple example consider the TU matrix

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{1.16}$$

Now, consider a LP as in (1.14), then if $b \in \mathbb{Z}^K$ and A is TU, then for any vertex $v \in P$ it holds $v \in \mathbb{Z}^L$ and therefore there exists an optimal solution $x^\star \in \mathbb{Z}^L$. This allows to ensure that a solution of a LP is integer, before actually solving it. It also removes the need to utilise integer LP solvers over standard LP solvers, which significantly reduces computational effort.

Optimal Transport

It remains to pose OT as a LP. Rewriting (1.4) as

$$\begin{aligned} \mathrm{W}_p^p(\mu^1, \mu^2) = \min_{\pi \in \mathbb{R}^{K \times L}} \quad & \sum_{k=1}^{K} \sum_{l=1}^{L} d(x_k^1, x_l^2)^p \pi_{kl} \\ \text{s.t.} \quad & \sum_{l=1}^{L} \pi_{kl} = a_k^1, \quad k = 1, \dots, K, \\ & \sum_{k=1}^{K} \pi_{kl} = a_l^2, \quad l = 1, \dots, L, \end{aligned} \tag{1.17}$$

it becomes obvious that it is possible to pose the p-Wasserstein distance as a LP by considering the transport plans as vectors instead of matrices. Let $D = (d(x_k^1, x_l^2)^p)_{kl}$ be the matrix containing the transport costs. The cost vector of the LP is then obtained by stacking the columns of D on top of each other from left to right, to obtain a vector of length LK. Correspondingly, it is possible to obtain an OT plan from the solution to the LP by columnwise filling the entries of the optimal solution vector into a $K \times L$ matrix. The constraint vector is given by $(a^1, a^2)^T$ and the constraint matrix is given by

$$A = \begin{pmatrix} I_K \bigotimes \mathbf{1}_L^T \\ \mathbf{1}_K^T \bigotimes I_L \end{pmatrix} \in \mathbb{R}^{(K+L) \times (KL)}, \tag{1.18}$$

where $\bigotimes$ is the Kronecker product. Notably, with this definition the row rank of A is $K + L - 1$, i.e. A is not of full rank. However, removing any arbitrary row from A yields a full rank matrix. This does not affect the solution of the problem. Assume w.l.o.g.

that the removed row is the one corresponding to a_L^2. Since both measures are assumed to have the same total mass intensity, it holds

$$a_L^2 = \sum_{k=1}^{K} a_k^1 - \sum_{l=1}^{L-1} a_l^2$$

and thus, the value of a_L^2 is implicitly specified by the $K + L - 1$ remaining constraints. Hence, the row can be removed without changing the solution to the problem.
Furthermore, the resulting constraint matrix A is TU[10]. Thus, if μ^1 and μ^2 have integer mass at all locations, there always exists an OT plan which also only has integer entries. In the special case, where for some $a > 0$ it holds $\mu^1(x), \mu^2(x) \in \{0, a\}$ for all $x \in \mathcal{Y}$, the marginal constraints imply in particular that any entry of the corresponding integer OT plan is in $\{0, a\}$ as well. Hence, there is an optimal plan which corresponds to a one-to-one matching of the support points of the two measures. This observation is sometimes referred to as the *Birkhoff-von Neumann theorem*[11].

Solving OT

Since OT can be posed as a LP, any algorithm of choice for solving LPs can be deployed to solve it. Though, naturally also more specialised methods, taking advantage of the specific structural properties of OT, have emerged. Prominent examples include the *network simplex*, the *transportation simplex* (see Bertsimas and Tsitsiklis [1997] for details on simplex methods), *dual ascent methods*, the most well-known being the *Hungarian algorithm* [Kuhn, 1955], the *auction algorithm* Bertsekas [1979], *dynamic formulations* based on the celebrated Benamou-Brenier formula [Benamou and Brenier, 2000] and (more recently) OT surrogate computations. Among the latter class of methods, the idea of *entropy penalised optimal transport* [Cuturi, 2013] has reached particular popularity in the past decade, due to its simplicity and ease of parallel implementation, which made it an attractive choice for many machine learning applications.

[10] For a range of different characterisations of TU matrices and a proof that this constraint matrix is TU, see Chapter III.1 of Wolsey and Nemhauser [1999].

[11] The Birkhoff-von Neumann theorem, in fact, provides the stronger assertion that the vertices of the polyhedron of doubly stochastic matrices, i.e. matrices with positive entries and row and column sums equal to one, are precisely the permutation matrices.

Wasserstein Barycenters as Linear Programs

Recall that for any p-Wasserstein barycenter $\mu^\star$ of $\mu^1, \dots, \mu^J$ it holds

$$\mu^\star = \sum_{x \in C} a_x \delta_x,$$

where C is the centroid set from (1.10). Hence, the p-Fréchet functional is a sum of Wasserstein distances between finitely supported measures and can be rewritten for any $\mu \in \mathcal{P}(C)$ as

$$F(\mu) = \frac{1}{N} \sum_{i=1}^{J} \min_{\pi^{(i)} \in \Pi(\mu, \mu^i)} \sum_{j=1}^{|C|} \sum_{k=1}^{M_i} \pi^{(i)}_{jk} c^i_{jk}. \tag{1.19}$$

where $c^i_{jk} = d^p(C_j, x^i_k)$ is the p-th power of the distance between the j-th point of C and the k-th point in the support of μ^i. Since μ is supported on the finite set C, it can be identified with a vector $a \in \mathbb{R}_+^{|C|}$ of weights on the centroid set. Rearranging the different minima and explicitly stating the marginal constraints yields

$$\begin{aligned}
\min_{\pi^{(1)}, \dots, \pi^{(J)}, a} \quad & \frac{1}{J} \sum_{i=1}^{J} \sum_{j=1}^{|C|} \sum_{k=1}^{M_i} \pi^{(i)}_{jk} c^i_{jk} \\
\text{subject to} \quad & \sum_{k=1}^{M_i} \pi^{(i)}_{jk} = a_j \quad \forall\, i = 1, \dots, J, \quad \forall\, j = 1, \dots, |C| \\
& \sum_{j=1}^{|C|} \pi^{(i)}_{jk} = a^i_k \quad \forall\, i = 1, \dots, J, \quad \forall k = 1, \dots, M_i \\
& \pi^{(i)}_{jk} \geq 0 \quad \forall i = 1, \dots, J \quad \forall\, j = 1, \dots, |C|, \quad \forall k = 1, \dots, M_i.
\end{aligned} \tag{1.20}$$

This problem can be intuitively understood as solving J OT problems linked together by a common marginal a, while also optimising over the weights of said marginal. As for the OT problem, this can be turned into a LP in standard form by vectorising the problem accordingly. It should be noted that the size of this LP grows rapidly in the support sizes of the measures. This minimisation problem has $|C| + \sum_{i=1}^{J} M_i$ equality constraints[12] and $|C| \left(1 + \sum_{i=1}^{J} M_i\right)$ variables ($|C|$ has a worst-case size scaling as $\prod_{i=1}^{J} M_i$).

One potential way of reducing the size of the considered LP is to consider the MMOT

[12] Not yet accounting for possible redundancies in the constraints.

instead of the barycenter problem. Rewriting (1.13) as an LP yields

$$\begin{aligned} &\underset{\pi \in \mathbb{R}_+^{M_1 \times \cdots \times M_J}}{\operatorname{argmin}} \sum_{(i_1,\ldots,i_J)\in I} c_{i_1,\ldots,i_J} \pi_{i_1,\ldots,i_J}, \\ &\text{s.t.} \sum_{(i_1,\ldots,i_J)\in I, i_r = s} \pi_{i_1,\ldots,i_J} = b_s^r \quad 1 \le s \le M_r,\ r = 1,\ldots,J, \end{aligned} \tag{1.21}$$

where $I = \{(i_1, \ldots, i_J) \in \mathbb{N}^J \mid 1 \le i_r \le M_r \ \forall\ r = 1, \ldots, J\}$. While this requires careful handling of the index set, (1.21) can also be vectorised into a LP in standard form. This LP has $\sum_{i=1}^J M_i$ equality constraints and $\prod_{i=1}^J M_i$ variables. By the same argument as for the OT case, one constraint can be removed for $J - 1$ of the measures, without changing the set of optimal solutions. Thus, the row rank of the constraint matrix is in fact only $\sum_{i=1}^J M_i - J + 1$ and the problem size can be reduced accordingly. Notably, this observation immediately recovers the, previously mentioned, sparsity bound for OT barycenters, as each optimal vertex of the LP has at most $\sum_{i=1}^J M_i - J + 1$ non-zero entries and therefore the pushforward under the barycentric application also yields a measure with a support size bounded from above by this value. Depending on the size of C, either (1.20) or (1.21) yields a more efficient approach to solve the problem. If in the definition of C only a relatively small number of constructed centroids coincide, i.e. $|C|$ is close to its maximal potential value of $\prod_{i=1}^J M_i$, then solving the MMOT problem instead of the OT barycenter problem is usually computationally more efficient.

1.6 Computing Wasserstein Barycenters

While the LP formulations in (1.20) and (1.21) theoretically allow to solve any OT barycenter problem between finitely supported measures, the complexity of the problem grows too rapidly. In particular, the scaling in the number of measures J is highly problematic. Assume $\mu^1, \ldots, \mu^J$ all have support size M, then the LP for the MMOT has M^J variables. To put this into perspective, this implies that for $J = 10$ measures supported on an equidistant 32×32 grid, the LP in (1.21) has over 10^{30} variables, which is already out of reach even for modern high-performance-computing systems[13]. Recalling Figure 1.9 (b), note that $|C| = 96721$ and thus the LP in (1.20) has over 10^9 variables, which is also already pushing the boundaries of what can be practically computed. Note, that the problem size in this example is still small, compared to potentially interesting data analysis applications, where each measure might correspond to a high resolution image of resolution 1024×1024 and where it is reasonable to

[13] For reference, the strongest super computer on the TOP500 list (for details on the list see Dongarra et al. [2003]) in June 2022 has about 10^{16} bytes of RAM.

consider hundreds or even thousands of images. This renders the LP approach infeasible for most interesting applications. Notably, this approach does still have its merits. For smaller-scale problems where it is feasible, it is guaranteed to provide an exact solution and not an approximate one. Even when infeasible, the LP approach sheds light on some structural properties of OT, such as the bound on the support size of the smallest barycenter.

The iterative method to approximate Wasserstein barycenter which brought them into broader interest for data analysis is due to Cuturi and Doucet [2014]. They propose to perform a subgradient method with respect to the weights of the barycenter a, supported on a fixed set X of size M. For $X = C$ this clearly computes a p-Wasserstein barycenter of $\mu^1, \dots, \mu^J$. Though, as discussed before, $|C|$ is usually infeasibly large. Hence, it can be replaced by a general finite set, usually an equidistant grid, to search for a minimiser of the p-Fréchet functional among the probability measures which are supported on X. This is referred to as the *fixed support barycenter problem.* An advantage of this approach is that it is immediately possible to interpret the resulting barycenter as an image again if the measures $\mu^1, \dots, \mu^J$ correspond to images[14]. However, choosing a sufficiently high resolution can still lead to an infeasible problem size in some situations. An alternative is to also optimise over the locations of the elements of X. This leads to a Llyod-type (compare the classical K-means algorithm in Lloyd [1982]) procedure alternating between optimising the weights a, for fixed locations X and then updating the positions X for fixed weights a. Unfortunately, this procedure is only guaranteed to converge to a local minimum, as optimising the positions X is a non-convex problem. However, in many practical applications this seems to perform reasonable well. Recalling the sparsity result for the OT barycenter, there is a barycenter with at most $\sum_{i=1}^{J} M_i - J + 1$ support points, hence if X is at least of this size, then it is theoretically possible to obtain an exact barycenter of the measures with this approach (at least if the method is initialised sufficiently close to an optimal solution).

However, Cuturi and Doucet [2014] do not deploy the previously described approaches directly. Instead, they turn to a specific type of surrogate OT distance, namely *entropy regularised optimal transport*. Here, an entropic penalty term scaled with a penalty parameter ε is added to the objective function of the OT, rendering the problem strictly convex. Following the previous work in Cuturi [2013], this problem can be solved using *Sinkhorn's algorithm* [Sinkhorn and Knopp, 1967]. This matrix scaling algorithm only relies on matrix-vector and matrix-matrix-multiplication, which allows for easy parallelisation. The ease of parallel GPU computing with this method, pushed this

[14] A grayscale image of resolution $G \times G$ can be considered to be a measure supported on an equidistant $G \times G$ grid on $([0,1]^2, d_2)$, where the mass intensity at each grid point is proportional to the grayscale intensity of the corresponding pixel.

approach into popularity, particularly within the machine learning communities. For $\varepsilon \to 0$ the solution to the regularised problem converges to a solution of the original one. However, for small values of ε, computations in the naive Sinkhorn algorithm tend to become numerically unstable. There have been modification to stabilise the algorithm for small ε [Schmitzer, 2019], but this stabilisation comes at a significant computational cost. Notably, there are applications where it might be desirable to consider a moderately sized regularisation instead of the smallest possible value, since the induced blurring effect can reduce the impact of potential discretisation artefacts within the data.

The subgradient approach to solve the fixed-support Wasserstein barycenter problem now requires solving J OT problems at each iteration. As solving OT is computationally costly, Cuturi and Doucet [2014] propose to replace all instances of OT in the algorithm with regularised OT and use this to approximate a true barycenter. Notably, this approach still relies on a reasonable choice of ε. Critically, there does not appear to be a universal method for choosing this parameter optimally for a given dataset. Hence, it requires tuning for given problems and can provide inaccurate results if set incorrectly. Still, entropy regularised OT enjoys great popularity and it has sparked a long line of research focused on its theoretical properties [for some examples see Carlier et al., 2017, Genevay et al., 2018, Feydy et al., 2019, Klatt et al., 2020] as well as improving the computational tools for solving it [Altschuler et al., 2017, Dvurechensky et al., 2018, Lin et al., 2019, Schmitzer, 2019]. One noteworthy extension of this approach is to apply the entropy regularisation directly to the barycenter problem instead of replacing each individual subgradient computation with it. This idea [Benamou et al., 2015] reduces run-time by orders of magnitude and can still be implemented in terms of Sinkhorn iterations. In particular, it can still be implemented efficiently in terms of matrix-matrix- and matrix-vector-multiplications. One downside however, is the fact that this method strictly requires all measures to have full support on a joint support set. This assumptions is usually satisfied for image datasets or other data supported on equidistant grids, but for measures with more general support structures, this method is not applicable. In these cases, the free support approach can be significantly more efficient. However, the naive Lloyd type approach does not offer any theoretical guarantees on the convergence to the true barycenter. The main issue behind this, is the fact that the p-Fréchet functional is convex in the weights of the measures for fixed locations, but not jointly convex in weights and locations. One alternative is to deploy a Frank-Wolfe algorithm instead. Luise et al. [2019] utilise regularity properties of the entropy regularised problem to obtain convergence rates for a free-support method in this context. Notably, this method does not require the specification of the support size or even an upper bound of the

support size at initialisation. Instead, the iterations can be initialised with a single Dirac measure and in each iterations the direction finding step of the Frank-Wolfe algorithm is shown to add (at most) one additional Dirac measure with a certain weight. However, this approach is still faced with the approximation error and stability issues inherent to entropy regularised OT. Attempting to avoid these downsides, there also has been approaches aiming to provide efficient solutions to the unregularised p-Wasserstein barycenter problem. Motivated by the results in Benamou et al. [2015] the idea in Xie et al. [2020] is conceptually similar. Their approach is build on an inexact proximal point method. This can be seen as a nested loop, where the inner loop is nearly identical to the iterations performed in the entropy-regularised context, while the outer loop modifies the problem in each iterations. They report good convergence properties to an exact barycenter given sufficient outer iterations for a small number of inner iterations. An alternative approach is to tackle the LP version of the problem directly. While naively using a standard LP solver, as discussed above, quickly becomes infeasible even at small data sizes, it is possible to modify a solver specifically for this problem class. The MAAIPM [Ge et al., 2019] method is a modification of a classical predictor-corrector interior point method (IPM) (see Nocedal and Wright [2006] for a general introduction into IPMs), which specifically exploits the structural properties of the p-Wasserstein barycenter problem. In particular, the problem can be decomposed into J related subproblems and a full solution can be constructed from the solutions to these subproblems. They can also be solved efficiently, due to their specific structure.
It should be stressed that the list of approaches above is by no means complete and the number of methods to approach p-Wasserstein barycenter problems is growing steadily. Though, the presented methods provide a good overview of the different avenues utilised to solve this problem.

1.7 Wasserstein Distance on Trees

While in general the computation of p-Wasserstein distances is a difficult ordeal, for measures supported on the leaves of ultra-metric trees, this problem does in fact have a closed-form solution. To discuss this formula some basic definitions for trees are recalled. A rooted, metric tree $\mathcal{T} = (V, E)$ is an undirected, circle-free graph endowed with a metric $d_{\mathcal{T}}$ possessing a designated element $\mathsf{r} \in V$ which is referred to as the *root of* $\mathcal{T}$. Each edge $e \in E$ is assigned a non-negative weight $w(e)$. Two nodes $\mathsf{v}, \mathsf{w} \in V$ are connected by a unique path denoted $\mathcal{P}(\mathsf{v}, \mathsf{w})$ either represented by a sequence of nodes or as a sequence of edges. The distance $d_{\mathcal{T}}(\mathsf{v}, \mathsf{w})$ is equal to the sum of the weights of those edges contained in $\mathcal{P}(\mathsf{v}, \mathsf{w})$. A *leaf* of $\mathcal{T}$ is any node, which is not the root, such

that its degree (number of edges attached to the node) is equal to one and the set of all leaf nodes is denoted as $L \subset V$. A node $\mathsf{v}^\star$ is termed *parent* of node v denoted by $\mathrm{par}(\mathsf{v}) = \mathsf{v}^\star$ if both are connected by a single edge but $\mathsf{v}^\star$ is closer to the root than v. The parent of the root node is set to $\mathrm{par}(\mathsf{r}) = \mathsf{r}$. For a node v its *children* are the elements of the set $C(\mathsf{v}) = \{\mathsf{w} \in V \mid \mathsf{v} \in \mathcal{P}(\mathsf{w}, \mathsf{r})\}$. Notice that with this definition v is a child of itself.

A rooted tree $\mathcal{T}$ with root r and endowed with a distance $d_{\mathcal{T}}$ is called *ultrametric* if all its leaf nodes have the same distance to r w.r.t. $d_{\mathcal{T}}$. Denote the set of leaf nodes of $\mathcal{T}$ by L. In particular, a tree is ultrametric if and only if there exists a *height function* $h\colon V \to \mathbb{R}_+$ with $d_{\mathcal{T}}(\mathsf{v}, \mathrm{par}(\mathsf{v})) = \left|h(\mathsf{v}) - h(\mathrm{par}(\mathsf{v}))\right|$ that is monotonically decreasing meaning that $h(\mathrm{par}(\mathsf{v})) \geq h(\mathsf{v})$ and such that $h(\mathsf{v}) = 0$ for $\mathsf{v} \in L$.

Let $\mathcal{T}$ be an ultrametric tree with height function h and measures μ^L, ν^L supported on the leaf nodes $L \subset V$. Then, Kloeckner [2015] showed that the p-Wasserstein distance between μ^L and ν^L has a closed form solution given by

$$\mathrm{W}_p^p(\mu^L, \nu^L) = 2^{p-1} \sum_{x \in V} h(par(x))^p - h(x)^p |\mu(C(x)) - \nu(C(x))|. \tag{1.22}$$

While the assumption of ultrametric trees is rather strict and there are little applications where it is reasonable (for an example on certain phylogentic trees see Gavryushkin and Drummond [2016]), it has powerful theoretical implications for controlling the statistical deviation of the p-Wasserstein distance (see Sommerfeld et al. [2019] and Chapter 2).

CHAPTER 2

Empirical Deviation Bounds for Wasserstein Barycenters

In practice, the population measures $\mu^1, \dots, \mu^J \in \mathcal{P}(\mathcal{Y})$ are not necessarily accessible, but only empirical versions $\hat{\mu}^1_{N_1}, \dots, \hat{\mu}^J_{N_J}$ generated from data are available. Specifically, for each $i = 1, \dots, J$ consider independent and identically distributed random variables $X^i_1, \dots, X^i_{N_i}$ and define

$$\mu^i_{N_i} = \frac{1}{N_i} \sum_{i=1}^{N_i} \delta_{X^i_{N_i}}.$$

This chapter summarises the results from Heinemann et al. [2022b][1] to treat upper bounds on the statistical error created by estimating a p-Wasserstein barycenter of the population measures by a p-Wasserstein barycenter of their empirical versions. This extends control on the expected error of the the plug-in estimator $\mathrm{W}_p(\hat{\mu}^1_N, \hat{\mu}^2_N)$ for $\mathrm{W}_p(\mu^1, \mu^2)$ and its approximation properties. Since by the reverse triangle and Jensen's inequality it holds

$$\mathbb{E}\left[|\mathrm{W}_p(\hat{\mu}^1_N, \hat{\mu}^2_N) - \mathrm{W}_p(\mu^1, \mu^2)|\right] \leq \mathbb{E}\left[\mathrm{W}^p_p(\mu^1, \hat{\mu}^1_N)\right]^{1/p} + \mathbb{E}\left[\mathrm{W}^p_p(\mu^2, \hat{\mu}^2_N)\right]^{1/p},$$

it is sufficient for this to understand the Wasserstein deviation of a single measure from its empirical measure to understand the deviation of the plug-in estimator[2]. It holds (see Sommerfeld et al. [2019] for the original result and Heinemann et al. [2022b] for the

[1] To streamline presentation results which have been taken from this work are therefore not referenced explicitly in the following section.

[2] In the naive upper bound based on the triangle inequality the potentially slower rate of the two measures dominates the speed of convergence. However, there are scenarios where the opposite holds true [Hundrieser et al., 2022]. This phenomenon is known as lower complexity adaptation. Since in this scenario both measures are finitely supported the upper bound has the same convergence rate as the left-hand side, however.

refined version given here)

$$\mathbb{E}\left[\mathrm{W}_p^p(\mu^1, \hat{\mu}_N^1)\right] \leq \frac{\mathcal{E}_p(\mathcal{X}_1)\mathrm{diam}(\mathcal{X}_1)^p}{N^{1/2}}, \tag{2.1}$$

where

$$\mathcal{E}_p(\mathcal{X}_1) = 2^{p-1} \inf_{q \geq 1, L \in \mathbb{N}} q^p \left(q^{(-L+1)p} \sqrt{M_1} + \left(\frac{q}{q-1}\right)^p \sum_{l=1}^{L} q^{-lp} \sqrt{|\mathcal{N}(\mathcal{X}_1, q^{-l}\mathrm{diam}(\mathcal{X}_1))|} \right).$$

Here, $\mathcal{N}(\mathcal{X}, \varepsilon)$ denotes an ε-covering of a space $\mathcal{X}$. It should be stressed, that this convergence rate does not exhibit a *curse of dimensionality* as common in many high-dimensional estimation problems. Instead the dimension (of the possible ambient space the support of μ^1 is embedded in) enters only as a constant. Intuitively, the constant $\mathcal{E}_p(\mathcal{X}_1)$ describes in some sense the complexity of the support of μ^1. In detail, this constant stems from the proof strategy which is based on an ultra-metric tree approximation. For each $l = 0, \ldots, L$ a $q^{-l}\mathrm{diam}(\mathcal{X}_1)$-covering of $\mathcal{X}_1$ is constructed, such that for $l = L$, each covering set contains at most one point from $\mathcal{X}_1$. Using this sequence of coverings, it is possible to construct an ultra-metric tree $(\mathcal{T}, d_{\mathcal{T}})$, which has the points of $\mathcal{X}_1$ as its leafs. In particular, this can be done in such a manner that for any two points $x, y \in \mathcal{X}_1$ it holds $d(x, y) \leq d_{\mathcal{T}}(x^{\mathcal{T}}, y^{\mathcal{T}})$.[3] Similarly, μ^1 can be considered as a measure $\mu^{1,\mathcal{T}}$ supported on the leafs of $\mathcal{T}$, such that it holds $\mathcal{W}_p^p(\mu^1, \hat{\mu}_N^1) \leq \mathcal{W}_p^p(\mu^{1,\mathcal{T}}, \hat{\mu}_N^{1,\mathcal{T}})$. By monotonicity of the expectation it suffices to control the empirical deviation for the tree-version of the measure. This greatly simplifies the problem as for ultra-metric trees their exist a closed form solution for the Wasserstein distance (recall Section 1.7). Using this, the bound in (2.1) can be derived. For the explicit details of the construction refer to the appendix of Sommerfeld et al. [2019] or Section 3.2 for a similar construction.

2.1 Deviation Bounds

Let F_p^N be the p-Fréchet functional of the empirical measures given by

$$F_p^N(\mu) = \sum_{i=1}^{J} \mathrm{W}_p^p(\mu, \hat{\mu}_{N_i}^i).$$

Any minimiser of F_p^N is referred to as an *empirical barycenter* of $\mu^1, \ldots, \mu^J$. Note, that neither the barycenter of the population measures nor the one of the empirical measures is necessarily unique. Thus, let $\mathbf{B}$ denote the set of barycenters of the population

[3] Here, $x^{\mathcal{T}}$ and $y^{\mathcal{T}}$ denote the points x and y as leafs of $\mathcal{T}$, respectively.

measures and $\hat{\mathbf{B}}$ the set of empirical barycenters. Let $\mu^\star \in \mathbf{B}$ and $\hat{\mu}^\star \in \hat{\mathbf{B}}$, then it holds for any $p \geq 1$ that

$$\mathbb{E}[|F^p(\mu^\star) - F^p(\hat{\mu}^\star)|] \leq \frac{2p \operatorname{diam}(\mathcal{Y})^p}{J} \sum_{i=1}^{J} \frac{\mathcal{E}_1(\mathcal{X}_i)}{\sqrt{N_i}}, \tag{2.2}$$

where $\mathcal{E}_1(\mathcal{X}_i)$ is the constant from (2.1). In particular, if $p = 2$ and $N_i = N$ for all $i = 1, \ldots, J$, then it holds

$$\mathbb{E}[|F^2(\mu^\star) - F^2(\hat{\mu}^\star)|] \leq \frac{4 \operatorname{diam}(\mathcal{Y})^2}{J\sqrt{N}} \sum_{i=1}^{J} \mathcal{E}_1(\mathcal{X}_i).$$

For any $i = 1, \ldots, J$, the constant $\mathcal{E}_p(\mathcal{X}_i)$ can be understood as a measure of the complexity of the set $\mathcal{X}_i$ in $\mathcal{Y}$ as it depends heavily on covering numbers of $\mathcal{X}_i$. If $\mathcal{X}_i$ has a complex structure, this value tends to be large, while for simple support structures this value is usually small. The upper bound in (2.2) scales as

$$\bar{\mathcal{E}} = \frac{1}{J} \sum_{i=1}^{J} \mathcal{E}_1(\mathcal{X}_i).$$

Thus, the size of the deviation bound for the Fréchet value of the empirical barycenter depends on the average complexity of the support sets of the population measures. This makes sense intuitively. If one of the population measures has a difficult support structure which makes it hard to estimate correctly, then its impact on the barycenter is harder to quantify and thus the estimation error for the barycenter should increase. However, if it is just one out of J measures for which estimation is difficult, then this should have little effect on the overall quality of the barycenter estimate if J is sufficiently large. Therefore, the dependence of the deviation bound on the average complexity of the supports of the population measures and thus in some sense on the average difficulty of estimating these measures, is reasonable. In particular, it seems unlikely that there could be any tight upper bound which is not impacted in some form by the difficulty of the individual estimation problems for the population measures. The second factor within the sum, namely the power of the diameter of $\mathcal{X}_i$ is also unlikely to be avoidable. Simply multiplying the locations of the support points of all measures by the same constant should clearly change the objective value of the problem by the respective power of this constant.

Finally, the rate of $N^{-\frac{1}{2}}$ is sharp, since the rate in the deviation bound of the barycenter can in general not be faster than for any individual population measure. For $J = 1$, the p-Wasserstein barycenter of μ^1 is μ^1 and the empirical barycenter is given by $\hat{\mu}_N^1$. In this

case, the error in the p-Fréchet functional is equal to $\mathrm{W}_p^p(\mu^1, \hat{\mu}_N^1)$, which is equivalent to the estimation of a single measure by its empirical counterpart. Since the $N^{-\frac{1}{2}}$ rate is already known to be optimal in this setting [Fournier and Guillin, 2015], it is also optimal for the estimation of the optimal Fréchet value.

Extending the control on the error in the p-Fréchet functional to control over the p-Wasserstein distance between a barycenter of the population measures and a barycenter of the empirical measures is a more delicate issue. In particular, the fact that neither barycenter is necessarily unique requires additional care. Let $p \geq 1$ and assume that $N_i = N$ for all $i = 1, \ldots, J$, then, it holds

$$\mathbb{E}\left[\sup_{\hat{\mu}^\star \in \mathbf{B}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} W_p^p(\mu^\star, \hat{\mu}^\star)\right] \leq \frac{p\bar{\mathcal{E}}\,\mathrm{diam}(\mathcal{Y})^p}{V_P} N^{-\frac{1}{2}}, \tag{2.3}$$

where V_P is a strictly positive constant given by

$$V_P := V_P(\mu^1, \ldots, \mu^J) := (J+1)\mathrm{diam}(\mathcal{Y})^{-p} \min_{v \in V \setminus V^\star} \frac{c^T v - f^\star}{d_1(v, \mathcal{M})},$$

where V is the vertex set of the linear programming formulation of the barycenter problem (1.20), $V^\star$ is the subset of optimal vertices, c is the cost vector of the program, $f^\star$ is the optimal value, $\mathcal{M}$ is the set of minimisers of the problem (1.20), and $d_1(v, \mathcal{M}) = \inf_{x \in M} \|v - x\|_1$.

The sup-inf in (2.3) can be understood as follows: For each empirical p-Wasserstein barycenter its distance to the closest p-Wasserstein barycenter of the population measures is computed. The only major difference between the upper bounds in (2.2) and (2.3) is the factor V_P. This constant stems from the proof of this deviation bound and has a geometric interpretation. The constraint set of the LP formulation of the barycenter problem is a high dimensional polyhedron P. Within P there is a lower dimensional polyhedron $\mathcal{M}$, characterised by the set of vertices $V^\star$, which contains the optimal solutions to this problem. The constant V_P now depends on the direction of the slowest rate of ascent in the objective value within P from any vertex in $V^\star$. More precisely, it depends on the pair of vertices $v \in V \setminus V^\star$ and $v^\star \in V^\star$ which minimises

$$\frac{c^T v - c^T v^\star}{d_1(v, v^\star)}$$

among all possible pairs of such vertices. Notably, this minimum is always attained for a pair of vertices which are adjacent in P. This constant is related to the *sub-optimality*

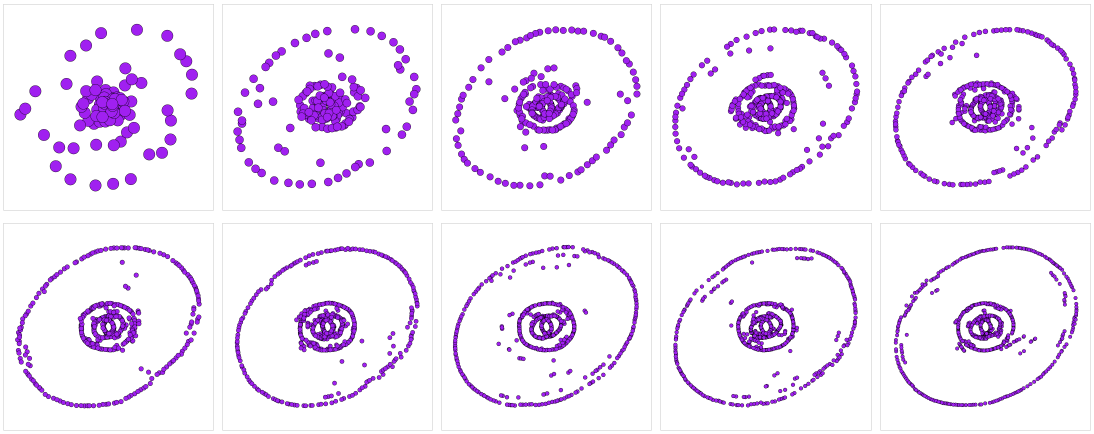


Figure 2.1: One realisation of an empirical 2-Wasserstein barycenter of eight probability measures for different sample sizes N. The underlying measures are supported on four nested ellipses in $([0,1]^2, d_2)$ discretised into 4000 support points with mass $1/4000$ each. From top-left to bottom-right the sample sizes are $100, 200, \ldots, 1000$.

gap[4] of the problem, but a major difference between the two objects is that the minimum in V_P depends on the distance between the two vertices. In particular, for this minimiser it does not suffice that the objective values of v and $v^\star$ are close, but their objective values must be close relative to their distance w.r.t. the d_1 distance in P. Thus, V_P becomes small (an thus the deviation bound large), if there exists a vertex v of P which is relatively far from the set of optimal solution, but has an objective value which is close to the optimal one. In this context it also becomes clear why the distance between a barycenter of the population measures and an empirical barycenter is likely to be large. An empirical barycenter might be close to v instead of $\mathcal{M}$, since there is a change in the right hand side of the constraints of the LP problem induced through the estimation of the measures. Now, by assumption the empirical barycenter is far from $\mathcal{M}$, even though its objective value is close to the optimal one. Thus, in this context the deviation bound in (2.3) should be larger.

The counterpart to this case is a problem where the optimal set $\mathcal{M}$ is well separated in the sense, that there are no vertices which have an objective value close to the optimal one relative to their distance to $\mathcal{M}$. In this case, V_P becomes large (and the constant in the deviation bound small) and a low objective value tends to imply proximity to the set $\mathcal{M}$.

The usefulness of statistical deviation bounds also extends beyond the setting where the population measures are unknown. Even if the population measures are available, their support size might be out of reach for any available solver for the OT problem.

[4]For a LP with cost vector c, vertex set V and optimal value $f^\star$ the sub-optimality gap is given by $\min_{v \in V \setminus V^\star} c^T v - f^\star$

One possible approach to reduce the size is to replace the population measures μ^1 and μ^2 by their empirical counterparts μ_N^1, μ_N^2. Then, the plug-in estimator $\mathrm{W}_p^p(\mu_N^1, \mu_N^2)$ can be used to obtain an approximation of $\mathrm{W}_p^p(\mu^1, \mu^2)$. Repeating this procedure multiple times and averaging over the resulting distances between the empirical measures can be used to increase the quality of the approximation and reduce the variability induced through the sampling. Statistical deviation bounds, such as (2.1), now allow to tune the runtime of the algorithm against the desired accuracy of the approximation. This approach of randomised computations for Wasserstein distances has become popular within the machine learning community where it is often known as *mini-batch OT* [Fatras et al., 2021]. The deviation bounds in (2.2) and (2.3) now allow to extend this idea of randomised OT computations from transport between pairs of measures to the barycenter of an arbitrary number of measures. Since p-Wasserstein barycenter computations are computationally demanding, the benefit of randomised computations is even more paramount. In particular, it enables fast, rough computations of barycenters on personal computers, instead of requiring dedicated high-performance computing facilities to compute barycenters at large datasizes. If a first, rough, computation indicates interesting results, it is then possible to increase the sample size to achieve improved accuracy of the approximation. Exemplary outcomes of this procedure for measures supported on nested ellipses and different sample sizes are found in Figure 2.1.
The randomised computation of 2-Wasserstein barycenters can be further accelerated with the use of a helpful heuristic. Recall that for two measures with M support points of mass one each by the Birkhoff-von Neumann theorem there exists an OT plan given by an optimal assignment between the two measures. In particular, the OT plan only has entries 1 and 0. From the initial definition of a barycenter between two measures as the mid point of a geodesic, it is immediate that the 2-Wasserstein barycenter is also a measure with mass 1 at M locations (namely the mid points of the geodesics in $\mathcal{Y}$ w.r.t. d between pairs of matched points). An extension of this result to $J > 2$ measures is not straightforward, however. Recalling the relation between the p-Wasserstein barycenter problem and the corresponding MMOT problem, the LP formulation of the MMOT problem in (1.21) can be considered. For $J = 2$ the constraint matrix of this LP is TU, thus by the consideration in Section 1.5 any vertex of the corresponding constraint polyhedron is integer. Since by standard LP theory there is always an optimal solution which is a vertex, this yields that there must be an integer solution.
Proofs of the Birkhoff-von Neumann theorem rely on all vertices being integer. For $J \geq 3$ and $M_1, M_2, M_3 \geq 3$, this does not hold, i.e. the constraint matrix is not TU [Lin et al., 2020] and consequently the vast majority of vertices of the polyhedron are not integer. Thus, the existing proofs of the Birkhoff-von Neumann theorem do not seem to

be adaptable to the case of more than two measures. In particular, they only rely on the constraints of the LP and not on the cost. Though, while MMOT admits non-integer solutions for certain costs, the polyhedron corresponding to the MMOT LP does have some integer vertices. In particular, this implies the existence of integer solutions for some costs. The key questions is under which conditions on the cost the existence of integer solutions can be guaranteed.

The result appears to hold in the geometrically important case of W_2-barycenters on $(\mathbb{R}^d, d_2)$. While it has not been verified theoretically even for this specific case, a large scale simulation study of over 10^8 examples did not yield a single example where this 2-Wasserstein barycenter problem did not admit an integer solution. Consequently, it seems reasonable to base a heuristic for the approximation of 2-Wasserstein barycenters on this empirical observation.

The previous overview of algorithms for Wasserstein barycenter computations are broadly distinguished between two settings: fixed-support and free-support methods. Fixed-support methods, where all measures are required to be supported on the same shared support set, are inevitably ill-suited for randomised computations. The main advantage of replacing the population measures with empirical ones is the reduction of problem size due to the artificially introduced sparsity. A method that requires all measures to be supported on the same equidistant grid, considers even measures supported on a sparse subset of this grid as measures on the whole set. This immediately negates any advantages the randomisation brings. Thus, approaches which can exploit the sparsity of the measures, such as free-support methods, are well-suited for this context. To avoid any regularisation and further surrogate usage, the subgradient method by Cuturi and Doucet [2014] (without entropy regularisation) is used. Their Lloyd-type approach is based upon alternating between optimising the positions and the weights of a barycenter candidate. Due to the sparsity bound on the support size of the barycenter this generally requires at least $\sum_i^J M_i - J + 1$ support points in the candidate to be able to approximate the true barycenter. Assuming the previous conjecture on the existence of integer solutions of barycenters with uniform weights is true, this bound can be improved. By considering each point of the empirical measure with mass k/N as k points with mass $1/N$ at the same location the conjecture implies that there exists an empirical 2-Wasserstein barycenter with mass $1/N$ at N (possibly non-distinct) locations. Thus, the number of points in the barycenter candidate can be reduced to N. Additionally, there is no longer any need for an alternating procedure. Since the weights of a barycenter are already known, it suffices to optimise the positions with respect to these weights once. This reduces the run-time of the algorithm by several orders of magnitude. Thus, the combined usage of randomised barycenter computations

to create artificial sparsity and enforce uniformity of the measure and the heuristic based on the uniformity conjecture reduces the runtime of the method by orders of magnitude. In particular, this allows for approximate computations at low computational cost on personal computers. This method is referred to as the *stochastic-uniform-approximation (SUA)*-method.

2.2 Discussion and Related Work

There is a large body of literature treating empirical deviation bounds for OT distances. An early entry in this line of work can be found in Dudley [1969] for the bounded Lipschitz metric which relates closely to the 1-Wasserstein distance. More recently, the topic has received renewed attention in more general frameworks. Using an explicit coupling construction Dereich et al. [2013] obtain more general deviation bounds for OT. Building on their ideas, Fournier and Guillin [2015] generalise their construction to obtain general deviation bounds for W_p^p for arbitrary measures with sufficient moment conditions on $\mathbb{R}^d$. With a related partition technique Weed and Bach [2019] establish sharp deviation bounds on compact, separable metric spaces. Notably, they also observe a difference in behaviour for convergence properties of the Wasserstein distances between small and large sample sizes. Extensions beyond the bounded case have also been achieved for unbounded Banach spaces and separable Hilbert spaces [Lei, 2020]. These general deviation bounds for empirical p-Wasserstein distances suffer from a curse of dimensionality, where the $N^{-\frac{1}{2}}$ rate in (2.1) is usually only achieved if the dimension D of the ground space is smaller than $2p$. For $D > 2p$ the convergence usually follows the slower $N^{-\frac{p}{d}}$ rate. For the critical case $D = 2p$, the corresponding rates contain an additional logarithmic factor in N.

The deviation bounds in (2.2) and (2.3) for the empirical p-Wasserstein barycenter recover the dimension free rate where the dimension of $\mathcal{Y}$ only enters through a constant which is independent of N. It should be pointed out, that replacing the role of (2.1) with one of the general upper bounds would allow to recreate an analog to the bound in (2.2) without being limited to finitely supported measures. However, since the proof of (2.3) relies on the specific characterisation of the barycenter of finitely supported measures, there is no straight-forward approach to replicate these results in those more general frameworks. Establishing an analog of (2.3) for general population measures requires a lower bound of the form

$$\frac{F^p(\mu) - F^p(\mu^\star)}{\mathrm{W}_p^p(\mu, \mu^\star)} \geq C > 0.$$

For $p = 1$, this is a reversed version of the usual Lipschitz condition

$$F^1(\mu) - F^1(\mu^\star) \geq C\mathrm{W}_1(\mu, \mu^\star), \tag{2.4}$$

for some $C > 0$. Note, that F^p is W_p Lipschitz, since

$$\begin{aligned}|F^p(\mu) - F^p(\nu)| &= \frac{1}{J}\left|\sum_{i=1}^{J} \mathrm{W}_p^p(\mu^i, \mu) - \mathrm{W}_p^p(\mu^i, \nu)\right| \\ &\leq p\mathrm{diam}(\mathcal{Y})^{p-1}\frac{1}{J}\sum_{i=1}^{J}|\mathrm{W}_p(\mu^i, \mu) - \mathrm{W}_p(\mu^i, \nu)| \leq p\mathrm{diam}(\mathcal{Y})^{p-1}\mathrm{W}_p(\mu, \nu).\end{aligned}$$

For more general $p > 1$ this relation to a Lipschitz condition is no longer valid, since W_p^p is generally not a metric, but it still has a similar interpretation. In particular, the constant C now takes the role of V_P in (2.3). It is equal to the direction in the Wasserstein space along which the value of F increases slowest relative to its distance to $\mu^\star$. In the proof of the deviation bound in the finitely supported setting the control of this ascend is achieved with the help of the LP formulation of the problem. Since the objective function there is linear, it is reasonable to expect a bound akin to (2.4) to hold. For general measures this remains open.

While the estimation of Wasserstein distances between empirical measures has received this large amount of attention, the estimation of the Wasserstein barycenter based on the Wasserstein barycenter of empirical measures is still only at its beginning. In particular, the two deviation bounds in (2.2) and (2.3) are novel and are the first of their kind. However, there is a related, but different, notion of Wasserstein barycenter for which estimation has been considered in the literature. Let $(\mathcal{Y}, d)$ be a metric space and let $\boldsymbol{\mu} \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{Y}))$ be a probability measure on the space of probability measures. Any minimiser of

$$\nu \mapsto \int_{\mathcal{P}(\mathcal{Y})} \mathrm{W}_p^p(\mu, \nu) d\boldsymbol{\mu} \tag{2.5}$$

is referred to as a *population p-Wasserstein barycenter*. Let $\mu^1, \ldots, \mu^N \sim \mu$ be an i.i.d. sample of probability measures drawn from μ. Then, the empirical version of μ is denoted as

$$\boldsymbol{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\mu^i} \tag{2.6}$$

and any minimiser of

$$\nu \mapsto \int_{\mathcal{P}(\mathcal{Y})} \mathrm{W}_p^p(\mu, \nu) d\boldsymbol{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} \mathrm{W}_p^p(\mu^i, \nu) \tag{2.7}$$

is referred to as an *empirical (population) p-Wasserstein barycenter*. This context differs significantly from the previously discussed deviation bounds. Here, the population level measure is a distribution on the space of probability distribution and the corresponding barycenter is the center of mass of this distribution with measure valued realisations. The empirical barycenter problem in this context then coincides with the population level problem in the earlier considerations. In the context of population p-Wasserstein barycenters results on convergence and deviation bounds usually consider the convergence of the solution of (2.7) to the solution of (2.5). This field of population Wasserstein barycenters has received significant attention in recent years.

Its modern analysis was popularised by the study of barycenters of measures on Hadamard spaces[5] [Sturm, 2003]. In this context, several variants of a law of large numbers have been established. Let $(\mathcal{Y}, d)$ be a Hadamard space, $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{Y})$ be a probability measure on $\mathcal{Y}$ and $X_1, \ldots, X_N \sim \boldsymbol{\mu}$ be i.i.d. samples from $\boldsymbol{\mu}$. Let $\hat{\boldsymbol{\mu}}_N = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ be the empirical version of μ based on these samples. Then, the solution of

$$y \mapsto \int_{\mathcal{Y}} d^p(x, y) d\boldsymbol{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} d^p(\mu^i, \nu) \tag{2.8}$$

converges to the unique minimiser of

$$y \mapsto \int_{\mathcal{Y}} d^p(x, y) d\boldsymbol{\mu} \tag{2.9}$$

at a rate of N^{-1}. Exploiting the fact that Hadamard spaces are uniquely geodesic, it is possible to derive alternative statements. Let $S_1 = X_1$ and define the sequence $(S_N)_{N \in \mathbb{N}}$ inductively by setting $S_N = V_{N-1}^N$ for $N \geq 2$, where $(V_t^N)_{[0,1]}$ is a constant speed geodesic w.r.t. d between S_{N-1} and X_N. If $\boldsymbol{\mu}$ has bounded support, then the sequence $(S_N)_{N \in \mathbb{N}}$ converges to the solution of (2.9) almost surely at a rate of N^{-1}. If $\boldsymbol{\mu}$ fulfils a second moment condition, the convergence still holds in probability. Notably, this alternative notions of a law of large numbers does not require any understanding of the behaviour of the empirical barycenter. The construction of the sequence S_N requires only geodesics on the space $(\mathcal{Y}, d)$, but it still converges to the population level barycenter of $\boldsymbol{\mu}$. From a computational point of view this also implies that only the abilities to generate samples

[5] A Hadamard space is a nonempty, complete metric space with globally non-positive curvature. For an overview over different notions of curvature consult Chapter 12.3 of Ambrosio et al. [2005].

from $\boldsymbol{\mu}$ and to compute geodesics between two elements of $(\mathcal{Y}, d)$ are required to approximate the population barycenter.

Unfortunately, the Wasserstein space has non-negative (and on most spaces even strictly positive) curvature[6] and hence Sturm's results are not applicable there. While existence of population barycenters and consistency of their empirical counterparts can be verified under very mild conditions [Le Gouic and Loubes, 2017], controlling the rate of this convergence seems to require stronger assumptions on the geometry of the space and the measure. The rate of convergence obtained in first consideration [Ahidar-Coutrix et al., 2020] on spaces of non-negative curvature suffers from the curse of dimensionality. A refined analysis in Le Gouic et al. [2022] is based on the geometrically interesting assumption of bi-extendable geodesics emanating from the population barycenter. A constant speed geodesic $(v_t)_{t\in[0,1]}$ between two points x, y in a metric space $(\mathcal{Y}, d)$ is set to be (λ_1, λ_2) bi-extendable (for some $\lambda_1 < 0$ and $\lambda_2 > 1$) if there exists another constant speed geodesic $\tilde{v} : [\lambda_1, \lambda_2] \to \mathcal{X}$ such that $\tilde{v}_{[0,1]} = v$. Under this assumption it is possible to recover the dimension-free N^{-1} convergence rate inherent in the strong law of large numbers. For the 2-Wasserstein space with a Hilbert space as the ground space, the abstract, general condition of bi-extendable geodesics can be replaced by a more intuitive sufficient condition. Assume that there exists $\mu_0 \in \mathcal{P}(\mathcal{Y})$ such that for all $\mu \in \mathrm{supp}(\boldsymbol{\mu})$ there exists a map $\alpha > 0$-strongly convex and $\beta > 0$ smooth map $\psi_{\mu_0,\mu}$ such that $\mu = (\nabla\psi_{\mu_0,\mu})\#\mu_0$. If $\beta - \alpha < 1$, then the empirical population barycenter $\mu_N^\star$ and the population barycenter $\mu^\star$ are unique and it holds

$$\mathbb{E}\left[\mathrm{W}_p^p(\mu^\star, \mu_N^\star)\right] \leq \frac{4\sigma^2}{(1-\beta-\alpha)^2 n},$$

where

$$\sigma^2 = \int_{\mathcal{P}(\mathcal{Y})} \mathrm{W}_p^p(\mu, \mu^\star) d\boldsymbol{\mu}. \tag{2.10}$$

Besides using empirical versions of the population measures μ^1 and μ^2 to reduce the problem size, it is also possible to consider deterministic approximations of these measures. One immediate advantage of the sampling approach, however, is the fact that the sampling can be done in negligible time (relative to the computation of the OT), while any (reasonable) deterministic approximation inevitably leads to an optimisation problem which yields the deterministic approximation as a solution. A natural choice might be to replace a measure μ^1 by its N-quantiser, i.e. the measure which minimises

[6] For a discussion of the curvature properties of the Wasserstein space consult Chapter 7 of Ambrosio et al. [2005].

$\mathrm{W}_p^p(\mu^1, \cdot)$ among all measures with a support size of at most N. However, this problem is difficult even for one dimensional measures [Graf and Luschgy, 2007]. A related approach is to consider a uniform N-quantiser of μ^1, i.e. a measure which minimises $\mathrm{W}_p^p(\mu^1, \cdot)$ among all measures with support size N and $\mu^1(y) \in \{0, 1/N\}$ for all $y \in \mathcal{Y}$ [Chevallier, 2018]. This uniform, deterministic approximation converges, in the worst case, at a rate of $log(N)/N$, which is faster than the $N^{-\frac{1}{2}}$ convergence rate for the empirical measures. Though, replacing the role of the deviation bound in (2.1) with respective deviation bounds for N-quantisers, allows to replicate the bound in (2.2) with the respective rate of convergence of the quantiser. However, the computation of such uniform quantisers still requires solving an involved optimisation problem. Moreover, in contrast to the empirical measures, the support of a quantiser (uniform or not) of μ^1 is not contained within the support of μ^1, the support of the barycenter of the quantised measures is in general also not contained within the centroid set C in (1.11). Therefore, it is not clear whether an analog of (2.3) holds true for the deterministic quantisers.

CHAPTER 3

The (p, C)-Kantorovich-Rubinstein Distance: Structure and Estimation

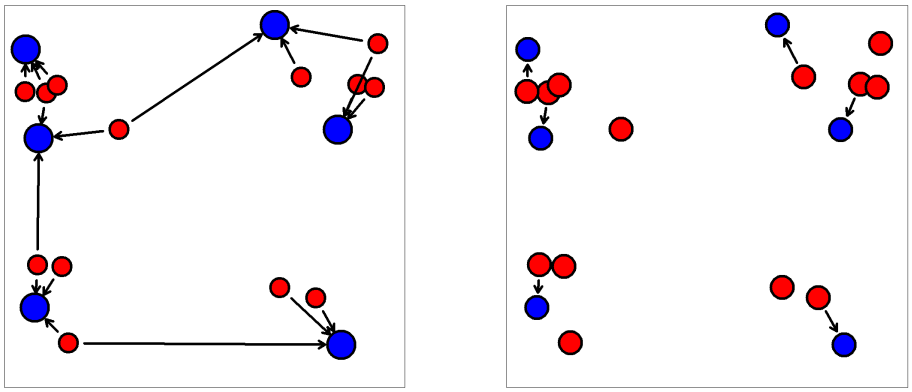
Despite its conceptual appeal and favourable geometric properties, OT still has one significant, inherent limitation. It is limited to measures of equal total intensities. Thus, OT based data analysis requires the normalisation of the measures' total intensities. However, this can destroy important geometric features such as the underlying stoichiometry. In Figure 3.1(a), the mass-splitting OT plan between two measures in $[0, 1]^2$ is displayed. Depending on the application, it might not make sense to allow for mass splitting or to normalise the measures. If, for instance, each support point in a measure corresponds to the location of a certain protein, then it might make sense to find close pairs of proteins, but it would be unreasonable to consider a third of a protein at some location. With the growing interest in OT tools for modern data analysis, this limitations quickly surfaced and a range of UOT formulations emerged to generalise the concepts to measures of arbitrary total intensity. This chapter first summarises the results of Heinemann et al. [2021] to present a detailed structural analysis of a specific model of UOT (in Figure 3.1(b) a possible UOT plan[1] based on this model of UOT is displayed) and then presents a first statistical analysis of this UOT model based on Heinemann et al. [2022a][2]

3.1 Structural Properties

Let $\mathcal{M}_+(\mathcal{Y})$ denote the set of a positive measures with finite mass on the space $\mathcal{Y}$ and define for any measure μ the total intensity function $\mathbb{M}(\mu) = \mu(\text{supp}(\mu))$. Define the set

[1] In general, UOT plans allow for mass-splitting, however, in the specific case of uniform, finitely supported measures it can be seen that there is always a non-mass splitting UOT plan.

[2] To streamline presentation results which have been taken from these two works are therefore not referenced explicitly in the following sections.



(a) (b)

Figure 3.1: Transport between two measures (blue and red) with uniform mass on their support points located in $[0,1]^2$. The ground cost is set to be the squared euclidean distance. **(a)** The measures have been normalised to probability measures (the blue points have mass $1/6$ and the red ones $1/13$). The OT plan between them is displayed in terms of its transport graph. **(b)** The UOT plan for $C = 0.2$ between the two unnormalised measures (all points have mass 1).

of *subcouplings* of two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{Y})$ as

$$\Pi_{\leq}(\mu,\nu) = \{\pi \in \mathcal{M}_+(\mathcal{Y}) \mid \pi(A,\mathcal{Y}) \leq \mu(A),\ \pi(\mathcal{Y},B) \leq \nu(B),\ \forall\, A,B \in \mathcal{B}(\mathcal{Y})\}\,.$$

For $p \geq 1$ and a parameter $C > 0$, the considered notion of UOT between two measures μ and ν is defined as

$$\mathrm{UOT}_{p,C}(\mu,\nu) := \min_{\pi\in\Pi_{\leq}(\mu,\nu)} \int_{\mathcal{Y}} d^p(x,y)d\pi + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\pi)\right). \tag{3.1}$$

If μ and ν have finite support, then this simplifies to

$$\mathrm{UOT}_{p,C}(\mu,\nu) := \min_{\pi\in\Pi_{\leq}(\mu,\nu)} \sum_{x,x'\in\mathcal{X}} d^p(x,x')\pi(x,x') + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\pi)\right). \tag{3.2}$$

Based upon this concept of UOT the (p,C)-*Kantorovich-Rubinstein distance* between two measures μ and ν is defined as

$$\mathrm{KR}_{p,C}(\mu,\nu) := \left(\mathrm{UOT}_{p,C}(\mu,\nu)\right)^{\frac{1}{p}}. \tag{3.3}$$

In some instances it is helpful to reformulate the penalty term to show its relation to the marginal constraints more clearly. For finitely supported measures, the penalty term

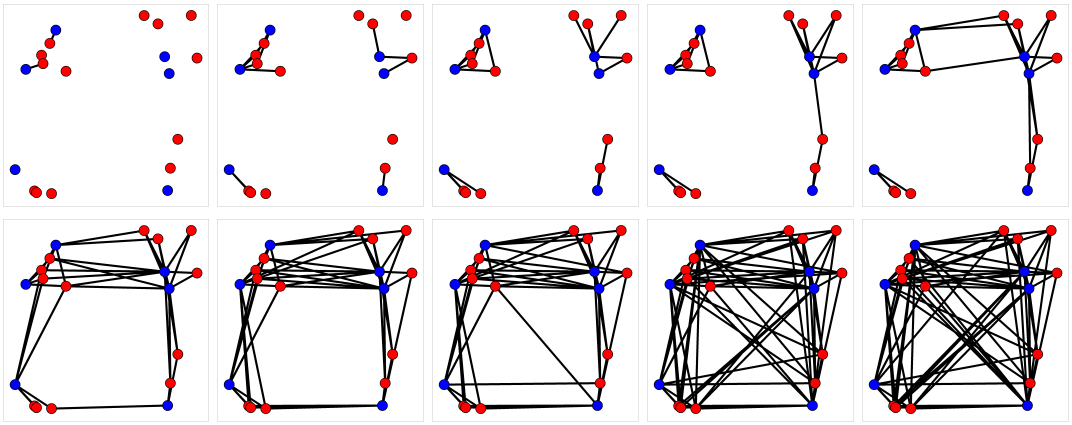


Figure 3.2: The two measures considered in Figure 3.1(b). The graph of available edges along which transport is possible according to (3.5) is displayed in black for different values of C. From top-left to bottom-right the values for C are $0.1, 0.2, \ldots, 1$.

can be equivalently replaced with

$$C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2}-\mathbb{M}(\pi)\right)=\frac{C^p}{2}\left(\sum_{x\in\mathcal{X}}(\mu(x)-\pi(x,\mathcal{X}))+\sum_{x'\in\mathcal{X}}(\nu(x')-\pi(\mathcal{X},x'))\right). \tag{3.4}$$

While in (3.2) the parameter $C > 0$ is seen to control the deviation of the total mass of π, the alternative representation in (3.4) demonstrates its marginal characterisation. Here, any mass deviation of the subcoupling from the original marginals is penalised with cost of $C^p/2$ per unit mass.

The first notable property of the (p, C)-KRD is that it defines a metric on the space of non-negative measures $\mathcal{M}_+(\mathcal{X})$ for any $C > 0$ and $p \geq 1$. Though it should be noted that this observation has already been made in specific instances, e.g., for $p = 1$ (Piccoli and Rossi [2014]) and uniform measures on point patterns (Müller et al. [2020]). The fact that the KRD is a distance is to be expected, given its similarity to the Wasserstein distance, but is stil novel in this generality.

The second notable property of the (p, C)-KRD concerns its dependence on the penalty parameter C. It can be seen that C governs the maximal scale at which any transport can be optimal. More precisely, if π_C is a solution to (3.2), then the length of any directed path P from the corresponding transport graph $G(\pi_C)$[3] is bounded by

$$\mathcal{L}(P) \leq C^p. \tag{3.5}$$

This implies, in particular, that if $d(x, x') > C$ then for any optimal solution of 3.2 it holds $\pi_C(x, x') = 0$. An illustration of the possible transports for varying C is found

[3]Let $\pi \in \Pi(\mu, \nu)$. Let $V = \mathcal{X}$ and $E = \{(x, x' \in \mathcal{X}^2 \mid \pi(x, x') > 0\}$. The directed graph given by (V, E) is referred to as transport graph corresponding to π.

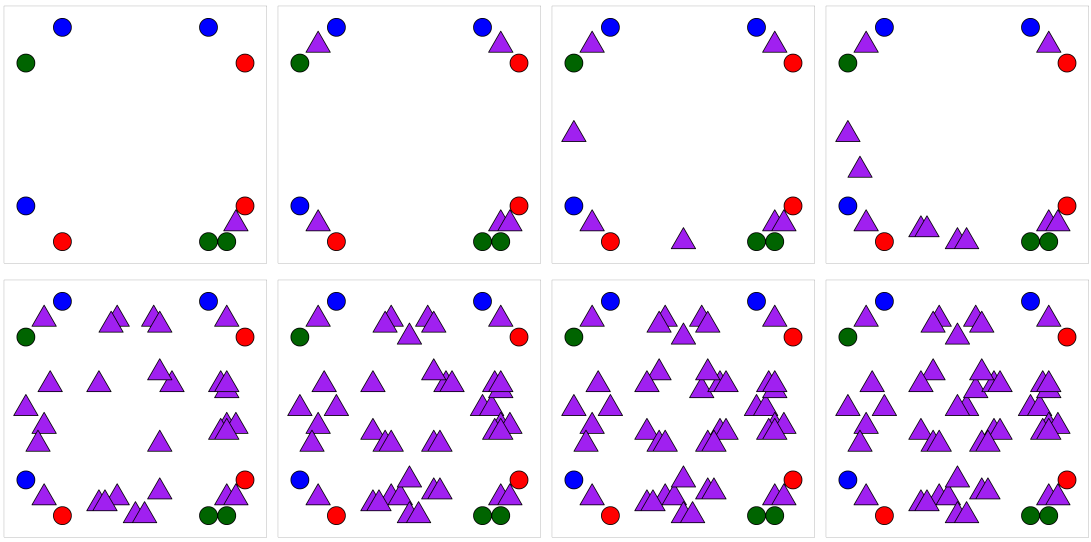


Figure 3.3: Three measures in $([0,1]^2, d_2)$ (with support points displayed as blue, green and red, circles respectively) and their restricted centroid set $C_{KR}(3,2,C)$ (displayed as purple triangles) for different values of C. From top-left to bottom-right the values of C are $0.2, 0.3, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2$. The sets for $C = 0.4, 0.5, 0.6$ have been omitted, since they are identical to the one for $C = 0.3$.

in Figure 3.2. For certain values of C, the KRD can be related to the total variation distance and the p-Wasserstein distance. If $C < \min_{x \neq x'} d(x, x')$, then any UOT plan $\pi^\star$ can only have positive mass on its diagonal. These entries have a cost of zero, though. Hence, the transport term of the KRD vanishes and only the penalty remains. This can be easily seen to be equal to $(C^p/2)\mathrm{TV}(\mu, \nu)$[4]. Thus, for sufficiently small values of $C > 0$ the KRD is proportional to the total variation distance. For $C > \mathrm{diam}(\mathcal{X})$, it is by construction cost-optimal to transport the maximal possible amount of mass. Thus, for any UOT plan $\pi^\star$ between μ and ν it holds $\mathbb{M}(\pi^\star) = \min(\mathbb{M}(\mu), \mathbb{M}(\nu))$. In particular, if $\mathbb{M}(\mu) = \mathbb{M}(\nu)$, then it also holds $\mathbb{M}(\mu) = \mathbb{M}(\pi^\star)$ and the penalty term vanishes. Thus, for measures of equal total mass intensity the (p, C)-KRD coincides with the p-Wasserstein distance for sufficiently large C. For intermediate values of C the (p, C)-KRD interpolates between the Wasserstein distance and the total variation distance. Note, that the (p, C)-KRD is increasing in C, i.e. for $C_1 < C_2$ it holds $\mathrm{KR}_{p,C_1}(\mu, \nu) \leq \mathrm{KR}_{p,C_2}(\mu, \nu)$.

While computing the KRD is challenging in general, for measures defined on ultrametric trees there exists, similarly as for the p-Wasserstein distance, a closed form solution. Consider an ultrametric tree $\mathcal{T}$ with leaf nodes L and height function $h\colon V \to$

[4]The *total variation distance* between two finitely supported measures $\mu^1, \mu^2 \in \mathcal{M}_+(\mathcal{X})$ is defined as $\mathrm{TV}(\mu^1, \mu^2) = \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$.

$\mathbb{R}_+$ inducing the tree metric $d_{\mathcal{T}}$. Define the set

$$\mathcal{R}(C) \coloneqq \left\{ \mathsf{v} \in V \,\middle|\, h(\mathsf{v}) \leq \frac{C}{2} < h(\mathrm{par}(\mathsf{v})) \right\} \tag{3.6}$$

with the convention that $\mathcal{R}(C) = \{\mathsf{r}\}$ if $\frac{C}{2} \geq h(\mathsf{r})$ and for a node $\mathsf{v} \in V$ set

$$\mu^L(C(\mathsf{v})) \coloneqq \sum_{\mathsf{w} \in C(v) \cap L} \mu^L(w),$$

where $C(\mathsf{v}))$ denotes the set of children of v in $\mathcal{T}$. Then, for any $p \geq 1$ and two measures $\mu^L, \nu^L \in \mathcal{M}_+(L)$ supported on the leaf nodes of $\mathcal{T}$ it holds that

$$\begin{aligned} &\mathrm{KR}^p_{d^p_{\mathcal{T}},C}\left(\mu^L, \nu^L\right) = \\ &\sum_{\mathsf{v} \in \mathcal{R}(C)} \left(2^{p-1} \sum_{\mathsf{w} \in C(\mathsf{v}) \setminus \{\mathsf{v}\}} \left((h(par(\mathsf{w}))^p - h(\mathsf{w})^p) \left| \mu^L(C(\mathsf{w})) - \nu^L(C(\mathsf{w})) \right| \right) \right. \\ &\left. + \left(\frac{C^p}{2} - 2^{p-1} h(\mathsf{v})^p \right) \left| \mu^L(C(\mathsf{v})) - \nu^L(C(\mathsf{v})) \right| \right). \end{aligned} \tag{3.7}$$

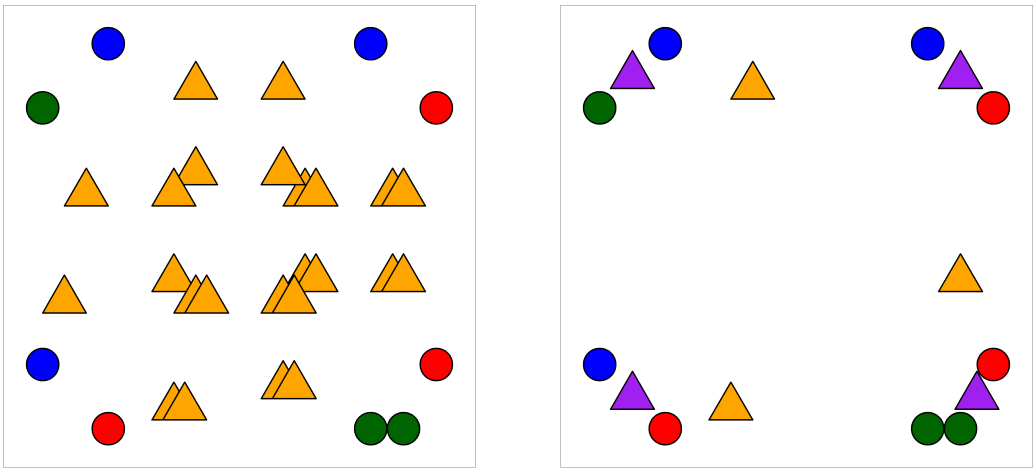
A notable novelty of this formula compared to the one for the Wasserstein distance on ultrametric trees (recall Section 1.7) is the fact this formula decomposes the optimisation problem into several subproblems. Due to the control of C on the maximal distance of transport in UOT plans, any edges with a weight over $C^p/2$ can be removed without changing the solution to the problem. This creates a forest of ultrametric trees on which the UOT problems can be solved independently. The UOT cost between two measures is then given by the sum of the costs over the forest. While this result is an interesting geometrical result for KRD in its own right, it is crucial in the statistical deviation bounds for the (p, C)-KRD in Section 3.2.

The KRD also allows to define a notion of a barycenter for a collection of measures as a generalisation of p-Wasserstein barycenter defined only for probability measures $\mu^1, \ldots, \mu^J \in \mathcal{P}(\mathcal{Y})$ as in (1.9). Any measure

$$\mu^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu) \coloneqq \frac{1}{J} \sum_{i=1}^{J} \mathrm{KR}^p_{p,C}(\mu^i, \mu) \tag{3.8}$$

is then said to be a (p, C)-*Kantorovich-Rubinstein barycenter* or (p, C)-*barycenter* for short. As for the p-Wasserstein barycenter this is straight-forward to generalise to arbitrary positive weights summing to one. As before this is omitted for brevity.

The p-Wasserstein barycenter of finitely supported measures has finite support itself, so the most natural first question which arises is whether this also holds for the (p, C)-



(a) (b)

Figure 3.4: The same measures as in Figure 3.3 considered as probability measures with mass $1/3$ at each support point. **(a)** The centroid set (displayed in orange) of the 2-Wasserstein barycenter problem corresponding to the three measures. **(b)** The 2-Wasserstein barycenter of the three measures (displayed in orange) and their $(2, 0.3)$-barycenter (displayed in purple). All displayed points of all measures have mass $1/3$.

barycenter. Recall the barycentric application $T^{L,p}$ given for any $p \geq 1$ and $L \in \mathbb{N}$ by

$$T^{L,p}(x_1, \ldots, x_J) \in \operatorname*{arg\,min}_{y \in \mathcal{Y}} \sum_{i=1}^{L} d^p(y, x_i).$$

Define the *full centroid set* of the measures $\mu^1, \ldots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ as

$$\begin{aligned} C_{KR}(J, p) = \Big\{ y \in \mathcal{Y} \mid \exists L \geq \lceil J/2 \rceil,\ \exists (i_1, \ldots, i_L) \subset \{1, \ldots, J\}, \\ x_1, \ldots, x_L :\ x_l \in \operatorname{supp}(\mu^{i_l}) \\ \forall l = 1, \ldots, L :\ y = T^{L,p}(x_1, \ldots, x_L) \Big\}, \end{aligned} \tag{3.9}$$

and the *restricted centroid set*

$$\begin{aligned} C_{KR}(J, p, C) = \Big\{ y = T^{L,p}(x_1, \ldots, x_L) \in C_{KR}(J, p) \mid \forall 1 \leq l \leq L : \\ d^p(x_l, y) \leq C^p;\ \sum_{l=1}^{L} d^p(x_l, y) \leq \frac{C^p(2L - J)}{2} \Big\}. \end{aligned} \tag{3.10}$$

An illustration of the restricted centroid set for different values of C is found in Figure 3.3. The corresponding Wasserstein centroid set for the same measures is shown in Figure 3.4(a) for reference. One critical caveat of the definition of this centroid sets is the fact that the barycentric application $T^{L,p}$ is in general not a map and the

respective minimiser is not necessarily unique. In particular, it is possible for multiple, even infinitely many, sets to fulfil the definitions of the centroid sets. To circumvent this issue, for each L-tupel one fixed representative of $T^{L,p}(x_1, \dots, x_L)$ is chosen for the construction of the centroid set $C_{KR}(J, p, C)$. All following statements regarding the centroid set are to be understood in the sense that there exists a choice of $C_{KR}(J, p, C)$ such that the statement holds true. To streamline presentation, this ambiguity is hidden in the following presentation.
Using the restricted centroid set it is possible to characterise the support of (p, C)-barycenters. It holds,

$$\min_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu) = \min_{\substack{\mu \in \mathcal{M}_+(\mathcal{Y}) \\ \operatorname{supp}(\mu) \subseteq C_{KR}(J,p,C)}} F_{p,C}(\mu). \tag{3.11}$$

Moreover, any (p, C)-barycenter $\mu^\star$ satisfies $\operatorname{supp}(\mu^\star) \subseteq C_{KR}(J, p, C)$. Thus, not only does it suffice to take the minimum over measures supported on the (finite) centroid set, but every (p, C)-barycenter is supported on this set. However, this results does not only guarantee a finite support for any barycenter, which is of critical importance for theoretical as well as practical matters, but it also provides an explicit geometrical characterisation of the possible support points of (p, C)-barycenters.
For the sake of simplicity let $p = 2$ and d the Euclidean distance. For $\mu^1, \dots, \mu^J \in \mathcal{P}(\mathcal{Y})$, recall that the support points of any 2-Wasserstein barycenter are contained in the set

$$\left\{ \frac{1}{J} \sum_{i=1}^{J} x_i \;\middle|\; x_i \in \mathcal{X}_i \right\}.$$

Thus, each support point of a barycenter $\mu^\star$ can be constructed by picking one specific point out of each of the J support sets of the measures, respectively, and computing the mean of these points. Now, for the $(2, C)$-barycenter (3.11) provides a similar characterisation. It holds

$$\begin{aligned} C_{KR}(J, 2, C) = \Bigg\{ y = \frac{1}{L} \sum_{k=1}^{L} x_{i_k} \;\Bigg|\; & \exists L \geq \lceil J/2 \rceil,\ \exists (i_1, \dots, i_L) \subset \{1, \dots, J\}, \\ & x_1, \dots, x_L :\ x_l \in \operatorname{supp}(\mu^{i_l}),\ d^2(x_l, y) \leq C^2 \\ & \forall l = 1, \dots, L;\ \sum_{l=1}^{L} d^2(x_l, y) \leq \frac{C^2(2L - J)}{2} \Bigg\}. \end{aligned} \tag{3.12}$$

In particular, this centroid set has far greater flexibility than the one for the 2-Wasserstein barycenter. While for the Wasserstein barycenter, for the construction of each support point, one point from each μ^i is necessary, for the $(2, C)$-barycenter it suffices to pick

any number of points L between $J/2$ and J, where each of the L points is in the support of a different μ^i. Obviously, it holds that $C \subset C_{KR}(J, p)$. The restricted centroid set is now constructed from the full centroid set, by removing any points which can not be optimal due to the control of C on the distances of any OT. The first inequality in the definition of $C_{KR}(J, p, C)$ ensures that transport between the centroid the locations it is constructed from is possible w.r.t. $\mathrm{KR}_{p,C}$. The second one ensures that if any mass is placed at a given centroid, the objective value can not be improved by removing this support point from the measure. Notably, for $2L < J$, this inequality is never fulfilled, which is essentially the reason for the $L \geq J/2$ restriction in the definition of the centroid set.

While the centroid set for the (p, C)-barycenter grows, in the worst case, as $\prod_{i=1}^{J}(M_i + 1)$ its support can be shown to be sparse within this set. In particular, there exists a (p, C)-barycenter which has a support size less or equal to $\sum_{i=1}^{J} M_i$. Notably, depending on the size of C and the structure of the support sets $\mathcal{X}_i$, the cardinality of the restricted centroid set might be smaller than the sum of the support sizes. Therefore, the support size of the barycenters can be bounded further in this setting.

On $(\mathbb{R}^d, d_2)$ there exist non-mass splitting OT plans w.r.t. W_2 between probability measures $\mu_1, \ldots, \mu_J$ and their 2-Wasserstein barycenter [Anderes et al., 2016]. Under some assumptions, the same holds true for the $(2, C)$-barycenter of $\mu_1, \ldots, \mu_J \in \mathcal{M}_+(\mathbb{R}^d)$. In particular, there exists UOT plans between the (p, C)-barycenter and the underlying measures that can be characterised by transport maps. More precisely, let $\mu^\star$ be a (p, C)-barycenter of $\mu_1, \ldots, \mu_J$, then there exist UOT plans π_i between $\mu^\star$ and μ^i for $i = 1, \ldots, J$, respectively, such that if $\pi_i(y, x) > 0$, then there exists $L \geq \lceil J/2 \rceil$, $x_l \in \mathrm{supp}(\mu^{i_l})$ for $l = 2, \ldots, L$, $(i_2, \ldots, i_L) \subset \{1, \ldots, J\}$ and $i_l \neq i$ for $l = 2, \ldots, L$ with $y = T^{L,p}(x, x_{i_2}, \ldots, x_{i_L})$, $\pi_j(y, x_j) > 0$ if $j \in \{i_2, \ldots, i_L\}$. Additionally, if for any $(x_1, \ldots, x_L) \in \mathcal{Y}^L$ it holds that

$$T^{L,p}(x_1, \ldots, x_L) = T^{L,p}(y_1, x_2, \ldots, x_L) \Leftrightarrow x_1 = y_1, \tag{3.13}$$

then $\pi_i(y, x) \in \{0, \mu^\star(y)\}$ for $i = 1, \ldots, J$. Since on $(\mathbb{R}^d, d_2)$ it holds $T^{L,2}(x_1, \ldots, x_L) = \frac{1}{L}\sum_{l=1}^{L} x_l$, the condition clearly holds on this space.

This property is useful for theoretical as well as practical purposes as it allows for a more explicit characterisation of an optimal solution as well as the optimal value. It is, for instance, used to derive an upper bound on the total mass intensity of any (p, C)-barycenter $\mu^\star$. While for any UOT plan between two measures it is clear by construction that the mass of the subcoupling is bounded by the minimum of total intensities of the two measures, the total mass intensity of a (p, C)-barycenter is not

necessarily bounded by the minimum of $\mathbb{M}(\mu^1) \ldots, \mathbb{M}(\mu^J)$. In particular, the mass of a (p, C)-barycenter is not even necessarily bounded by the maximum of these values. For an easy example of this consider

$$\mu^1 = \delta_{-1} + \delta_0, \quad \mu^2 = \delta_0 + \delta_1, \quad \mu^3 = \delta_{-1} + \delta_1.$$

The $(2, 0.01)$-barycenter of these measures is given by $\mu^\star = \delta_{-1} + \delta_0 + \delta_1$, thus it holds

$$\mathbb{M}(\mu^\star) = 3 > 2 = \mathbb{M}(\mu^1) = \mathbb{M}(\mu^2) = \mathbb{M}(\mu^3).$$

For a first, rough upper bound on the total mass intensity, note that if a (p, C)-barycenter $\mu^\star$ has mass larger than $\sum_{i=1}^J \mathbb{M}(\mu^i)$, then for any UOT plans $\pi^1, \ldots, \pi^J$ between $\mu^\star$ and μ^i, respectively, there exists a part $\mathbb{M}_0$ of the mass of $\mu^\star$ that is not transported to any of the μ^i, but is instead always destroyed. Thus, removing this mass from $\mu^\star$ decreases the objective value by $C^p\mathbb{M}_0 > 0$. Hence, $\mu^\star$ is not a (p, C)-barycenter and for any (p, C)-barycenter $\mu^\star$ it holds $\mathbb{M}(\mu^\star) \leq \sum_{i=1}^J \mathbb{M}(\mu^i)$. Denote $\mu^\star = \sum_{k=1}^K a_k \delta_{x_k}$. By the construction of (3.12), each of the x_k for $k = 1, \ldots, K$ is constructed from at least $J/2$ support points of the μ^i. Combined with the previous, rough upper bound this implies $\frac{J}{2} \sum_{k=1}^K a_k \leq \sum_{i=1}^J \mathbb{M}(\mu^i)$ and therefore $\mathbb{M}(\mu^\star) \leq \frac{2}{J} \sum_{i=1}^J \mathbb{M}(\mu^i)$. Hence, the total mass intensity of a (p, C)-barycenter is bounded by double the mean of the individual total masses. Notably, while the Fréchet value is monotone in C, i.e. for $C_1 < C_2$, it holds $F_{p,C_1}(\mu^\star_{C_1}) < F_{p,C_2}(\mu^\star_{C_2})$, the total mass intensity of the (p, C)-barycenter is not necessarily monotone in C. If the measures $\mu^1, \ldots, \mu^J$ have pairwise disjoint support, then for sufficiently small C, the (p, C)-barycenter has mass zero. For increasing C, the sum of the penalty terms starts to dominate the value of the (p, C)-Fréchet functional, thus intuitively for sufficiently large C, the mass of the barycenter should be the minimiser of

$$a \mapsto \sum_{i=1}^J \frac{a + \mathbb{M}(\mu^i)}{2} - \min(a, \mathbb{M}(\mu^i)) = \frac{1}{2} \sum_{i=1}^J |a - \mathbb{M}(\mu^i)|.$$

The minimiser of this function is the median of $\mathbb{M}(\mu^1), \ldots, \mathbb{M}(\mu^J)$. The worst-case threshold for this turns out to be $C > J^{1/p}\mathrm{diam}(\mathcal{Y})$, i.e. if $C > J^{1/p}\mathrm{diam}(\mathcal{Y})$ then there always exists a barycenter $\mu^\star$, such that $\mathbb{M}(\mu^\star) = \mathrm{med}(\mathbb{M}(\mu^1), \ldots, \mathbb{M}(\mu^J))$. Though, for certain examples, this can also already occur for smaller values of C (e.g. in Figure 3.4(b) the $(2, 1.5)$-barycenter of the measures coincides with the 2-Wasserstein barycenter shown there in orange and has mass one as the three underlying measures). For intermediate values of C, however, it is possible for $\mathbb{M}(\mu^\star)$ to exceed the median of

the individual total mass intensities.

The (p, C)-barycenter is also closely connected to barycenters with respect to the TV and the p-Wasserstein distance. If $C > 2^{\frac{1}{p}}\mathrm{diam}(\mathcal{Y})$ and $\mathbb{M}(\mu^1) = \mathbb{M}(\mu^2) = \cdots = \mathbb{M}(\mu^J)$, then any p-Wasserstein barycenter is also a (p, C)-barycenter and vice versa. To consider the relation to TV-barycenters, let $\mathcal{Z} := \bigcup_{i=1}^J \mathcal{X}_i \cup C_{KR}(J, p)$ and define $d'_{\min} := \min_{x\in\mathcal{Z}\setminus C_{KR}(J,p),\, y\in C_{KR}(J,p)} d(x, y)$. If $C \leq d'_{\min}$, then the (p, C)-barycenter $\mu^\star$ is given by

$$\mu^\star = \sum_{x_k\in\cup_{i=1}^J \mathcal{X}_i} \mathrm{med}(a_k^1, \ldots, a_k^J)\delta_{x_k},$$

where $\{a_1^i, \ldots, a_K^i\}$ are the weights of μ^i considered as a measure on $\cup_{i=1}^J \mathcal{X}_i$. In particular, this implies that

$$\mu^\star \in \operatorname*{arg\,min}_{\mu\in\mathcal{M}_+(\mathcal{Y})} \sum_{i=1}^J \mathrm{TV}(\mu^i, \mu).$$

While these relations allow to characterise (p, C)-barycenter for relatively large and small regimes of C, it is in general difficult to obtain more precise characterisation, besides the structure of the centroid set, for intermediate values of C. One exception is the case where all measures are supported on well-separated clusters and C is adapted suitably to the cluster size. Let $\mu^1, \ldots, \mu^J \in \mathcal{M}_+(\mathcal{Y})$ with finite support set $\mathcal{X}_1, \ldots, \mathcal{X}_J$ such that for all $i = 1, \ldots, J$ it holds $\mathcal{X}_i \subset \cup_{r=1}^R B_r$ for some $B_1, \ldots, B_R \subset \mathcal{Y}$ with $\mathrm{diam}(B_r) \leq C$ for all $r = 1, \ldots, R$ and $d(B_k, B_l) > 2^{1/p}C$ for all $k \neq l$. For $r = 1, \ldots, R$, let

$$\mu_r^\star \in \operatorname*{arg\,min}_{\mu\in\mathcal{M}_+(\mathrm{conv}(B_r))} \frac{1}{J}\sum_{i=1}^J \mathrm{KR}^p_{p,C}(\mu, \mu^i{}_{|B_r}), \tag{3.14}$$

where $\mathrm{conv}(B_r)$ is the convex hull of B_r for $r = 1, \ldots, R$. Then, the measure $\sum_{r=1}^R \mu_r^\star$ is a (p, C)-barycenter of $\mu^1, \ldots, \mu^J$. Thus, for measures supported on well-separated clustered, the respective (p, C)-barycenter can be decomposed into R independent, smaller (p, C)-barycenter problem. In particular, (3.14) implies that the (p, C)-barycenter respects the cluster structure within the supports of the measures if the clustered are sufficiently separated and C is adapted according to the cluster size, which is potentially interesting for data analysis.

Lifts to Optimal Transport

The backbone of the proofs of the properties of the (p, C)-KRD and its barycenter is their equivalence to certain balanced OT problems. The fundamental idea is to *augment*

the spaces $\mathcal{X}$ and $\mathcal{Y}$ by a *dummy point* $\mathfrak{d}$. For a fixed $C > 0$, the *augmented spaces* are defined as $\tilde{\mathcal{X}} \coloneqq \mathcal{X} \cup \{\mathfrak{d}\}$ and $\tilde{\mathcal{Y}} \coloneqq \mathcal{Y} \cup \{\mathfrak{d}\}$ and the metric on them is given by

$$\tilde{d}_C^p(y, y') = \begin{cases} d^p(y, y') \wedge C^p, & y, y' \in \mathcal{Y}, \\ \frac{C^p}{2}, & y \in \mathcal{Y},\ y' = \mathfrak{d}, \\ \frac{C^p}{2}, & y = \mathfrak{d},\ y' \in \mathcal{Y}, \\ 0, & y = y' = \mathfrak{d}. \end{cases} \tag{3.15}$$

Here, the metric on $\tilde{\mathcal{X}}$ is assumed to be the restriction of $\tilde{d}_C$ to $\tilde{\mathcal{X}}$. For any $p \geq 1$, let $\tilde{\mathrm{W}}_p$ denote the p-Wasserstein distance with respect to $\tilde{d}_C$ on $\tilde{Y}$. Consider the subset $\mathcal{M}_+^B(\mathcal{Y}) \coloneqq \{\mu \in \mathcal{M}_+(\mathcal{Y}) \mid \mathbb{M}(\mu) \leq B\} \subset \mathcal{M}_+(\mathcal{Y})$ of non-negative measures whose total mass is bounded by B. Setting $\tilde{\mu} \coloneqq \mu + (B - \mathbb{M}(\mu))\delta_{\mathfrak{d}}$, any measure $\mu \in \mathcal{M}_+^B(\mathcal{X})$ defines an *augmented measure* $\tilde{\mu}$ on $\mathcal{X}$ such that $\mathbb{M}(\tilde{\mu}) = B$. Following the arguments in Guittet [2002] it is straightforward to see that for any $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ and their augmented versions $\tilde{\mu}, \tilde{\nu} \in \mathcal{M}_+(\mathcal{X})$ it holds

$$\mathrm{KR}_{C,p}^p(\mu, \nu) = \tilde{W}_p^p(\tilde{\mu}, \tilde{\nu}).$$

Similarly, the (p, C)-barycenter problem can be augmented. For this, let $\tilde{\mathcal{Y}} \coloneqq \mathcal{Y} \cup \{\mathfrak{d}\}$ be endowed with the metric $\tilde{d}_C$ in (3.15) and augment the measures $\mu^1, \dots, \mu^J$ to $\tilde{\mu}^1, \dots, \tilde{\mu}^J$ where $\tilde{\mu}^i = \mu^i + \sum_{j \neq i} \mathbb{M}(\mu^j)\delta_{\mathfrak{d}}$ for $1 \leq i \leq J$. This augmentation is valid, since by an earlier argument, the mass of any (p, C)-barycenter is bounded from above by the sum of all total mass intensities. In particular, $\mathbb{M}(\tilde{\mu}^i) = \sum_{j=1}^J \mathbb{M}(\mu^j)$ and the *augmented p-Fréchet functional* is defined as

$$\tilde{F}_{p,C}(\mu) \coloneqq \frac{1}{J} \sum_{i=1}^J \tilde{W}_p^p(\tilde{\mu}^i, \mu).$$

Any minimiser of $\tilde{F}_{p,C}$ on $\mathcal{M}_+(\tilde{\mathcal{Y}})$ is referred to as as augmented (p, C)-barycenter. Notably, the restriction of an augmented (p, C)-barycenter to $\mathcal{Y}$ yields a (p, C)-barycenter. The augmented formulation can also be used to derive a related multi-marginal OT problem.

LP-Formulation for the (p, C)-Barycenter

Combining the finite support set of the (p, C)-barycenter with the augmented problem formulation it is possible to rewrite the augmented (p, C)-barycenter problem as a linear

program based on the *augmented restricted centroid set* $\tilde{C}_{KR}(J,p,C) = C_{KR}(J,p,C) \cup \{\mathfrak{d}\}$ (recall (3.12) for the definition of $C_{KR}(J,p,C)$) of the augmented measures. This yields the LP

$$\begin{aligned}
\min_{\pi^{(1)},\dots,\pi^{(J)},a} \quad & \frac{1}{J}\sum_{i=1}^{J}\sum_{j=1}^{|\tilde{C}_{KR}(J,p,C)|}\sum_{k=1}^{M_i}\pi_{jk}^{(i)}c_{jk}^{i} \\
\text{s.t.} \quad & \sum_{k=1}^{M_i}\pi_{jk}^{(i)} = a_j, && \forall\, i = 1,\dots,J, \forall\, j = 1,\dots,|\tilde{C}_{KR}(J,p,C)|, \\
& \sum_{j=1}^{|\tilde{C}_{KR}(J,p,C)|}\pi_{jk}^{(i)} = a_k^i, && \forall\, i = 1,\dots,J, \forall k = 1,\dots,M_i, \\
& \pi_{jk}^{(i)} \geq 0, && \forall i = 1,\dots,J, \forall\, j = 1,\dots,|\tilde{C}_{KR}(J,p,C)|, \\
& && \forall k = 1,\dots,M_i,
\end{aligned} \tag{3.16}$$

where for each $1 \leq i \leq J$ $M_i = |\tilde{\mathcal{X}}_i|$ is the cardinality of the support of the augmented measure $\tilde{\mu}^i$. Here, c_{jk}^i denotes the distance between the j-th point of $|\tilde{C}_{KR}(J,p,C)|$ and the k-th point in the support of $\tilde{\mu}^i$, while a^i is the vector of masses corresponding to $\tilde{\mu}^i$.

3.2 Estimation of Unbalanced Optimal Transport

As in the OT setting, in practice the population measures $\mu, \nu, \mu_1, \dots, \mu_J$ are not necessarily available or have to be sampled for randomised but feasible computations. However, while for a probability measure it is straightforward to generate samples, there is no single well-defined approach to sample from a measure of arbitrary intensity.

Statistical Models

In the following, the statistical models discussed in Heinemann et al. [2022a] are presented. It should be stressed, however, that the considered approach is not specific to these models and thus not limited to them. In particular, it provides a template to potentially analyse a range of statistical models in the UOT context.

Multinomial Model

In Chapter 2 it has been seen that randomised algorithms for the computation of p-Wasserstein distance or p-Wasserstein barycenters enable fast approximations of these quantities and non-asymptotic statistical deviation bounds allow to control the expected error of the approximations. Aiming to enable analog tools for the (p,C)-KRD and its

barycenter the *multinomial model* first normalises the measures to define probability measures and then samples in the classical sense. The resulting empirical estimators are rescaled to the original total intensities to keep retain the information on the measures' total mass intensities. Formally, consider independent and identically distributed (i.i.d.) random variables $X_1, \dots, X_N \overset{\text{i.i.d.}}{\sim} \frac{\mu}{\mathbb{M}(\mu)}$, where the total intensity $\mathbb{M}(\mu)$ is assumed to be known. The corresponding unbiased empirical estimator is then defined as

$$\hat{\mu}_N \coloneqq \frac{\mathbb{M}(\mu)}{N} \sum_{x \in \mathcal{X}} |\{k \in \{1, \dots, N\} \mid X_k = x\}| \, \delta_x. \tag{3.17}$$

This approach is best suited to resampling approaches similar to those discussed in Chapter 2. In this context the total mass intensity of the measures is known and the sample size N provides an upper bound on the computational complexity of the UOT problem between the estimators.

Bernoulli Model

The *Bernoulli model* is motivated by tasks in fluorescence cell microscopy, where a fluorescent marker is introduced into the sample to make the objects of interest visible in the proper experimental setup [Kulaitis et al., 2021]. However, this process is limited by the marker's labelling efficiency which determines the probability with which a location is in fact labelled. Formally, consider a point cloud encoded as a measure μ with mass one at each location. Let $B_x \sim \text{Ber}(s_x)$ with a fixed *success probability* $s_x \in [0, 1]$ for each location $x \in \mathcal{X}$. The vector $s_{\mathcal{X}} \coloneqq (s_{x_1}, \dots, s_{x_{|\mathcal{X}|}})$ is referred to as *success vector*. A suitable unbiased estimator for μ is defined by

$$\hat{\mu}_{s_{\mathcal{X}}} \coloneqq \sum_{x \in \mathcal{X}} \frac{B_x}{s_x} \delta_x. \tag{3.18}$$

Poisson Intensity Model

The *Poisson intensity model* is closely related to the Bernoulli model. However, in this model there exists two independent layers of randomness. Motivated by various tasks in photonic imaging [see e.g. Munk et al., 2020] any given point $x \in X$ is observed with a fixed probability of $s \in [0, 1]$, then, if it is observed, a Poisson random variable is observed based on the intensity of the underlying measure at x. In particular, this model is no longer restricted to point clouds. Formally, consider a collection of independent Bernoulli random variables $B_x \sim \text{Ber}(s)$ for each $x \in X$. Additionally, consider an independent Poisson random variable $P_x \sim \text{Poi}(t\mu(x))$ with intensity $t\mu(x)$, where the parameter $t > 0$ usually models the observation time of the experiment. A suitable

unbiased estimator for μ is defined by

$$\hat{\mu}_{t,s} \coloneqq \frac{1}{st} \sum_{x \in \mathcal{X}} B_x P_x \delta_x. \tag{3.19}$$

Sampling Bounds

This section summarises the results on the expected Kantorovich-Rubinstein deviation for the the three statistical models introduced above. Inspired by the approach of Sommerfeld et al. [2019] to control the empirical deviation of the p-Wasserstein distance, the main tool in this section is an approximation of the finite ground space $\mathcal{X}$ by an ultrametric tree, which then allows to use the closed form solution in (3.7).

Tree Approximation of the Kantorovich-Rubinstein Distance

Fix some depth level $L \in \mathbb{N}$. For some $q > 1$ and level $j = 0, \dots, L$ consider the covering set $Q_j \coloneqq \mathcal{N}(\mathcal{X}, q^{-j}\mathrm{diam}(\mathcal{X})) \subset \mathcal{X}$ and let $Q_{L+1} \coloneqq \mathcal{X}$. Any point $x \in Q_j$ is considered as a node at level j of a tree $\mathcal{T}$ and denoted as (x, j) to emphasise its level position. For level $j = 0$ this yields a single element in Q_0 which serves as the root of the tree. For $j = 0, \dots, L$ a node (x, j) at level j is connected to one node $(x', j+1)$ at level $j+1$ if their distance satisfies $d(x, x') \leq q^{-j}\mathrm{diam}(\mathcal{X})$ (ties are broken arbitrarily). The edge weight of the corresponding edge is set equal to $q^{-j}\mathrm{diam}(\mathcal{X})$. Consequently, the height of each node only depends on its assigned level $0 \leq l \leq L+1$ and is defined as $h_{q,L} \colon \{0, \dots, L+1\} \to \mathbb{R}$ by

$$h_{q,L}(l) = \sum_{j=l}^{L} q^{-j}\mathrm{diam}(\mathcal{X}) = \frac{q^{1-l} - q^{-L}}{q-1}\mathrm{diam}(\mathcal{X}). \tag{3.20}$$

By definition the space $\mathcal{X}$ is embedded in level $L+1$ as the leaf nodes of $\mathcal{T}$ with height $h_{q,L}(L+1) = 0$. By a straightforward computation it holds for two points $x, x' \in \mathcal{X}$ considered to be embedded in $\mathcal{T}$ as $(x, L+1)$ and $(x', L+1)$ that $d^p(x, x') \leq d^p_{\mathcal{T}}((x, L+1), (x', L+1))$. The measures μ, ν are embedded into $\mathcal{T}$ as measures μ^L, ν^L supported only on leaf nodes of $\mathcal{T}$ and thus it follows that

$$\mathrm{KR}_{p,C}(\mu, \nu) \leq \mathrm{KR}_{d^p_{\mathcal{T}},C}\left(\mu^L, \nu^L\right),$$

where $\mathrm{KR}_{d^p_{\mathcal{T}},C}$ denotes the (p, C)-KRD on the space $(\mathcal{T}, d_{\mathcal{T}})$. In combination with the closed formula from (3.7) this yields an upper bound on the (p, C)-KRD. Whenever clear from the context the notation is alleviated by writing $\mathsf{v} \in Q_l$ instead of $(\mathsf{v}, l) \in Q_l$.

Having established the framework of the tree-approximation, it is possible able to state the main deviation bounds explicitly. There exist constants $\mathcal{E}^{\text{Mult}}_{p,\mathcal{X},\mu}(C)$, $\mathcal{E}^{\text{Pois}}_{p,\mathcal{X},\mu}(C)$, $\mathcal{E}^{\text{Ber}}_{p,\mathcal{X},\mu}(C)$ such that for any $p \geq 1$ and for any measure μ and its estimator $\hat{\mu}$ derived from either (3.17), (3.18) or (3.19), respectively, it holds

$$\mathbb{E}\left[\text{KR}_{p,C}\left(\hat{\mu},\mu\right)\right] \leq \begin{cases} \mathcal{E}^{\text{Mult}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, N^{-\frac{1}{2p}}, & \text{if } \hat{\mu} = \hat{\mu}_N, \\ \mathcal{E}^{\text{Poi}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, \phi(t,s)^{\frac{1}{p}}, & \text{if } \hat{\mu} = \hat{\mu}_{t,s}, \\ \mathcal{E}^{\text{Ber}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, \psi(s_{\mathcal{X}})^{\frac{1}{p}}, & \text{if } \hat{\mu} = \hat{\mu}_{s_{\mathcal{X}}}, \end{cases}$$

where

$$\phi(t,s) = \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}} \sum_{x\in\mathcal{X}} \sqrt{\mu(x)}\right), & C \leq \min_{x\neq x'} d(x,x') \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s} \sum_{x\in\mathcal{X}} \mu(x)^2\right)^{\frac{1}{2}}, & \text{else,} \end{cases}$$

and

$$\psi(s_{\mathcal{X}}) = \begin{cases} \left(2\sum_{x\in\mathcal{X}}(1-s_x)\right), & C \leq \min_{x\neq x'} d(x,x') \\ \left(\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}\right)^{\frac{1}{2}}, & \text{else.} \end{cases}$$

For

$$A_{q,p,L,\mathcal{X}}(l) := \text{diam}(\mathcal{X})^p 2^{p-1} \left(q^{-Lp}|\mathcal{X}|^{\frac{1}{2}} + \left(\frac{q}{q-1}\right)^p \sum_{j=l}^{L} q^{p-jp}|Q_j|^{\frac{1}{2}} \right),$$

the constant $\mathcal{E}^{\text{Poi}}_{p,\mathcal{X},\mu}(C,q,L)$ is equal to

$$\mathcal{E}^{\text{Poi}}_{p,\mathcal{X},\mu}(C,q,L) = \begin{cases} \left\{ \left(\frac{C^p}{2} - 2^{p-1} \left(\frac{q-q^{-L}}{q-1} \text{diam}(\mathcal{X}) \right)^p \right) + A_{q,p,L,\mathcal{X}}(1) \right\}, \\ \qquad\qquad C \geq 2h_{q,L}(0), \\ A_{q,p,L,\mathcal{X}}(l), \\ \qquad\qquad 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2}, \\ \qquad\qquad C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right), \end{cases}$$

and the constant $\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C,q,L)$ is given by

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C,q,L) = \begin{cases} A_{q,p,L,\mathcal{X}}(1) & C \geq 2h_{q,L}(0), \\ A_{q,p,L,\mathcal{X}}(l), & 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2}\sqrt{\mathbb{M}(\mu)}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)} & C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right). \end{cases}$$

Finally, it holds $\mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C,q,L) = \mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L)$. The bounds in this theorem are reminiscent of the deviation bounds for OT in Chapter 2. However, the key difference between the two is the fact that the bounds in the UOT context are sensitive to the size of C relative to the distances in the tree-approximation of $\mathcal{X}$. Here, three different regimes are distinguished. For sufficiently small C controlling the empirical (p,C)-KRD corresponds to controlling the TV distance between the measures. Hence, the specific properties of the tree structure do not enter into the constants and just the value of C is relevant, since the TV distance is oblivious to the geometric structure of the support sets of the measures. For sufficiently large values of C the constants $\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L), \mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C,q,L)$ differ from $\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C,q,L)$ by an additional summand based on the estimation error in the total mass intensity. This summand does not occur in the multinomial model, since here the total mass intensity is assumed to be known. Therefore, $\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C,q,L)$ essentially recovers the constants in the OT deviation bounds. For intermediate values of C the constants coincide and depend on the specific tree structure of the approximation. As for the balanced setting, it is possible to construct more explicit upper bounds on these constants on Euclidean spaces. For the Wasserstein distance such bounds yield case distinction between the cases $2p < d$, $2p = d$ and $2p > d$ (different behaviour in these specific cases is common in the statistical analysis of UOT, see e.g. Fournier and Guillin [2015] or Weed and Bach [2019]). For the (p,C)-KRD these bounds now need to distinguish these three cases as well as the inherent distinction induced by C. The most critical dependence in this constants is the dependence on $|\mathcal{X}|$. If C is sufficiently large, the upper bounds on the constants roughly scale as $\log_2(|\mathcal{X}|)$ for $p = 2D$ and $|\mathcal{X}|^{-p/d}$ for $p < 2D$. For $C < d_{\min}(\mathcal{X})$ or for $2p > d$, the constants can be bounded independently of $|\mathcal{X}|$. Notably, in the former case the bound still implicitly depends on $\mathcal{X}$ as it contains the sum $\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}$.

While a change in C induces a change in the constants in all three models, the $N^{-1/2}$ rate in the multinomial model does not change when varying C. Notably, this rate is sharp, since for any probability measure μ and sufficiently large C, the (p,C)-KRD between μ and its estimator $\hat{\mu}_N$ coincides with their p-Wasserstein distance. Since by the $N^{-1/2}$

rate is already optimal in this setting [Fournier and Guillin, 2015], it is also sharp of the (p, C)-KRD. For the Poisson model there is a change in behaviour for sufficiently small C. Though, it should be noted that for success probability $s = 1$ the bounds exhibit a $t^{-1/2}$ rate independently of C. Thus, this change in behaviour seems to be driven by the success probability of the Bernoulli experiments and not the Poisson part. This is supported by the fact that the Bernoulli model exhibits a similar behaviour. Fixing $s = 1$ and using the closed form solution of the mean absolute deviation of a Poisson random variable [Ramasubban, 1958] and Stirling's formula it is straightforward to see that the $t^{-1/2}$ rate occurring in this scenario is sharp.

Barycenter Bounds

Following arguments analog to those seen in Chapter 2 allows to extend the deviation bounds on the (p, C)-KRD to (p, C)-barycenters. Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = \operatorname{supp}(\mu^i)$ for $i = 1, \dots, J$. Consider random estimators $\hat{\mu}^1, \dots, \hat{\mu}^J \in \mathcal{M}_+(\mathcal{X})$ derived from either (3.17), (3.18) or (3.19), respectively. Let F_p be the Fréchet functional w.r.t. $\mu^1, \dots, \mu^J$. Let $\mu^\star$ be a (p, C)-barycenter of $\mu^1, \dots, \mu^J$ and let $\hat{\mu}^\star$ be (p, C)-barycenter of $\hat{\mu}^1, \dots, \hat{\mu}^J$. Then, it holds

$$\mathbb{E}\left[|F_{p,C}(\mu^\star) - F_{p,C}(\hat{\mu}^\star)|\right] \leq 2p \min\{\operatorname{diam}(\mathcal{Y}), C\}^{p-1} \begin{cases} \bar{\mathcal{E}}^{\mathrm{Mult}}(C)\, N^{-\frac{1}{2}}, & \text{if } \hat{\mu}^i = \hat{\mu}_N, \\ \bar{\mathcal{E}}^{\mathrm{Poi}}(C)\, \phi(t, s), & \text{if } \hat{\mu}^i = \hat{\mu}^i_{t,s}, \\ \bar{\mathcal{E}}^{\mathrm{Ber}}(C)\, \psi(s_\mathcal{X}), & \text{if } \hat{\mu}^i = \hat{\mu}_{s_\mathcal{X}}, \end{cases}$$

where $\bar{\mathcal{E}}^{\mathrm{Mult}}(C) = \frac{1}{J} \sum_{i=1}^J \mathcal{E}^{\mathrm{Mult}}_{1,\mathcal{X}_i,\mu^i}(C)$ and $\bar{\mathcal{E}}^{\mathrm{Poi}}(C), \bar{\mathcal{E}}^{\mathrm{Ber}}(C)$ are defined analogously.
The statistical control on the distance between the optimal set on a population level and its empirical counterpart requires a more elaborate statement. This again relies on the fact that the problem can be rewritten as a LP (recall (3.16)). Let $\mathbf{B}^\star$ be the set of (p, C)-barycenters of $\mu^1, \dots, \mu^J$ and $\hat{\mathbf{B}}^\star$ the set of (p, C)-barycenters of $\hat{\mu}^1, \dots, \hat{\mu}^J$. Then, for $p \geq 1$ it holds that

$$\mathbb{E}\left[\sup_{\hat{\mu}^\star \in \hat{\mathbf{B}}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} \mathrm{KR}^p_{p,C}(\mu^\star, \hat{\mu}^\star)\right] \leq \frac{p \min\{\operatorname{diam}(\mathcal{Y}), C\}^{p-1}}{V_P} \begin{cases} \bar{\mathcal{E}}^{\mathrm{Mult}}(C)\, N^{-\frac{1}{2}}, & \text{if } \hat{\mu}^i = \hat{\mu}^i_N, \\ \bar{\mathcal{E}}^{\mathrm{Poi}}(C)\, \phi(t, s), & \text{if } \hat{\mu}^i = \hat{\mu}^i_{t,s}, \\ \bar{\mathcal{E}}^{\mathrm{Ber}}(C)\, \psi(s_\mathcal{X}), & \text{if } \hat{\mu}^i = \hat{\mu}^i_{s_\mathcal{X}}, \end{cases}$$

where the constant V_P is strictly positive and given by

$$V_P := V_P(\mu^1, \dots, \mu^J) := (J+1)\mathrm{diam}(\mathcal{X})^{-p} \min_{v \in V \setminus V^\star} \frac{c^T v - f^\star}{d_1(v, \mathcal{M})},$$

where V is the set of feasible vertices from the linear program in (3.16), $V^\star$ is the subset of optimal vertices, c is the cost vector of the program, $f^\star$ is the optimal value, $\mathcal{M}$ is the set of minimisers and $d_1(x, \mathcal{M}) = \inf_{y \in \mathcal{M}} \|x - y\|_1$. As for the bound in (2.1), the difficulty of this estimation problem scales roughly as the average of the J estimation problems for the individual measures. The bound also again depends on the geometry of the constraint polyhedron of the LP formulation of the problem and the well-separateness of its set of minimisers.

It remains to see that the rates in two bounds above are sharp. For this, note that for $J = 1$ and any $p \geq 1, C > 0$ the (p, C)-barycenter of μ^1 is μ^1. Thus, the optimal value of the (p, C)-Fréchet functional is zero and it holds

$$F(\hat{\mu}^\star) - F(\mu^\star) = \mathrm{KR}^p_{p,C}(\mu^1, \hat{\mu}^1).$$

Consequently, it also holds

$$\sup_{\hat{\mu}^\star \in \hat{\mathbf{B}}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} \mathrm{KR}^p_{p,C}(\mu^\star, \hat{\mu}^\star) = \mathrm{KR}^p_{p,C}(\mu^1, \hat{\mu}^1).$$

Thus, the rate for the convergence of the (p, C)-barycenter of the empirical measures, can in general not be faster than the convergence rate of a single estimator. In particular, the rates above are sharp, since the rates for single measures are sharp by the discussion in the previous section.

3.3 Discussion and Related Work

Following the surge in popularity of OT, a range of extensions to UOT trying to circumvent the inherent limitations of OT to measures of equal total mass intensity emerged. In fact, this line of research already goes back to Kantorovich himself and indeed the notion of UOT presented in this chapter is based on the seminal work in Kantorovich and Rubinstein [1958]. This approach was considered again in the theory of Lipschitz spaces [Hanin, 1992]. Following this work, Guittet [2002] extended these results to a more general and modern framework. In particular, this work established the lift from UOT to a balanced OT problem and thus the ability to pose this UOT as a LP.

Closely related to the (p, C)-KRD is the notion of *generalised p-Wasserstein distances or Picolli-Rossi distances* [Piccoli and Rossi, 2016]. For $p \geq 1$ and $a, b \in (0, \infty)$, the generalised p-Wasserstein distance is given by

$$\mathrm{W}_p^{a,b}(\mu, \nu) = \left(\inf_{\substack{\mathbb{M}(\tilde{\mu})=\mathbb{M}(\tilde{\nu}) \\ \tilde{\mu},\tilde{\nu}\in\mathcal{M}_+(\mathcal{Y})}} a^p \mathrm{TV}^p(\mu, \tilde{\mu}) + a^p \mathrm{TV}^p(\nu, \tilde{\nu}) + b^p \mathrm{W}_p^p(\tilde{\mu}, \tilde{\nu}) \right)^{\frac{1}{p}}. \tag{3.21}$$

For $p = 1$ and $b = 1$ it holds

$$\begin{aligned} \mathrm{W}_1^{a,1}(\mu, \nu) &= \min_{\substack{\mathbb{M}(\tilde{\mu})=\mathbb{M}(\tilde{\nu}) \\ \tilde{\mu},\tilde{\nu}\in\mathcal{M}_+(\mathcal{X})}} a\mathrm{TV}(\mu, \tilde{\mu}) + a\mathrm{TV}(\nu, \tilde{\nu}) + \mathrm{W}_1(\tilde{\mu}, \tilde{\nu}) \\ &= \min_{\pi\in\Pi_\leq(\mu,\nu)} a\left(\mathrm{TV}(\mu, \pi_1) + \mathrm{TV}(\nu, \pi_2)\right) + \sum_{x,x'\in\mathcal{X}} d(x, x')\pi(x, x') \\ &= \min_{\pi\in\Pi_\leq(\mu,\nu)} \sum_{x,x'\in\mathcal{X}} d(x, x')\pi(x, x') + a\left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi)\right), \end{aligned}$$

where the second equality follows since the minimum in the first can be restricted [Piccoli and Rossi, 2014] to measures $\tilde{\mu}, \tilde{\nu}$ such that $\tilde{\mu}(x) \leq \mu(x)$ and $\tilde{\nu}(x) \leq \nu(x)$ for all $x \in X$ and the third by a straightforward computation. In particular, it holds $\mathrm{W}_1^{C,1} = \mathrm{KR}_{1,C}$. For $p > 1$ the two distances differ in general. The most notable difference between the two approaches, is the fact for the (p, C)-KRD the weights of the individual measures enter linearly, while for the generalised Wasserstein distance they enter polynomially. There are two key consequences from this. First, the (p, C)-KRD is not geodesic for $p > 1$, while the generalised p-Wasserstein distance is geodesic for any $p \geq 1$[5]. Considering the augmented formulation of the problem, it is clear that the space $\tilde{Y}$ is not geodesic by construction, since the point $\mathfrak{d}$ is isolated. For an example of two measures without a geodesic between each other, consider $\mu = 2\delta_0$ and $\nu = \delta_2$. In particular, it holds $\mathrm{KR}_{2,4}(\mu^1, \mu^2) = \sqrt{12}$. Assume there exists a $(2, 4)$-geodesic $(\nu_t)_{t\in[0,1]}$ between μ^1 and μ^2. Then, it holds $\mathrm{KR}_{2,4}(\mu^1, \nu_{0.5}) = \mathrm{KR}_{2,4}(\mu^2, \nu_{0.5}) = \sqrt{3}$. This implies

$$8|2 - \mathbb{M}(\nu_{0.5})| + 4a^2 = 3 = 8|1 - \mathbb{M}(\nu_{0.5})| + 4(1 - a)^2, \tag{3.22}$$

where a corresponds to the timepoint of the d_2-geodesic between 0 and 2 at which the transported mass of 1 is located. Since $C = 4 > 2$, for any point of the geodesic ν_t it is optimal to transport any mass up to 1 to μ^2 and any mass up to 2 to μ^1. In particular, the mass of the geodesic at any time point must be in $[1, 2]$. The equation in (3.22)

[5]This can be seen from the analog of the Benamou-Brenier formula derived in Piccoli and Rossi [2016].

does not have a solution $a \in \mathbb{R}_+$, hence the assumption is wrong and there is no $\mathrm{KR}_{2,4}$ geodesic between μ^1 and μ^2. Thus, the (p, C)-KRD is not a geodesic distance in general. This is unfortunate for certain geometrical applications. However, the geodesity of the generalised p-Wasserstein distance comes at a cost. A second consequence of the fact that the objective function of this distance is not linear in the weights of its marginals, is that it is not possible to pose the generalised p-Wasserstein distance as a LP for $p > 1$. Thus, while the $\mathrm{W}_p^{a,b}$ distance has favourable geometric properties for studying dynamic/geodesic formulation of UOT, the KRD gains a much clearer structural description of its solutions and the impact of the structure of the measures on its value. Additionally, it can be solved, with minor modification, by any off-the-shelf OT solver or general LP solver, while the generalised p-Wasserstein distance requires more involved approaches.

Another well-known proposal [Caffarelli and McCann, 2010, Figalli, 2010] to define a notion of UOT between two measures μ^1 and μ^2 is the *partial optimal transport (POT)* problem. The idea of POT is to fix a value $M \in [0, \min(\mathbb{M}(\mu^1), \mathbb{M}(\mu^2))]$ in advance and then restrict any OT plan between μ^1 and μ^2 to have total intensity M. Formally, for any $p \geq 1$ it is defined as

$$POT_{p,M}(\mu^1, \mu^2) = \underset{\pi \in \mathcal{M}(X^2), \mathbb{M}(\pi) = M}{\arg\min} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} d(x_k^1, x_l^2)^p \pi_{kl}. \tag{3.23}$$

POT and UOT are closely related. In fact, it is straightforward to see that for every $C > 0$ there exists $M > 0$ such that the solutions of $POT_{p,M}$ and $\mathrm{UOT}_{p,C}$ coincide. Let π^{UOT} be an UOT plan for $\mathrm{KR}_{p,C}(\mu, \nu)$. Let $M_C = \mathbb{M}(\pi_C)$ and let π^{POT} be a POT_{p,M_C} plan between μ and ν. It holds by optimality of π^{UOT} that

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{l=1}^{L} d^p(x_k, y_L) \pi_{kl}^{\mathrm{UOT}} + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi^{\mathrm{UOT}}) \right) \\ \leq &\sum_{k=1}^{K} \sum_{l=1}^{L} d^p(x_k, y_L) \pi_{kl}^{POT} + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi^{POT}) \right). \end{aligned}$$

Since by construction it holds $\mathbb{M}(\pi^{\mathrm{UOT}}) = \mathbb{M}(\pi^{POT})$, the two penalty terms coincide and this implies

$$\sum_{k=1}^{K} \sum_{l=1}^{L} d^p(x_k, y_L) \pi_{kl}^{\mathrm{UOT}} \leq \sum_{k=1}^{K} \sum_{l=1}^{L} d^p(x_k, y_L) \pi_{kl}^{POT}.$$

However, by optimality of π^{POT} for the POT_{p,M_C} problem the converse inequality is also true and it holds

$$\sum_{k=1}^{K}\sum_{l=1}^{L} d^p(x_k, y_L)\pi_{kl}^{POT} = \sum_{k=1}^{K}\sum_{l=1}^{L} d^p(x_k, y_L)\pi_{kl}^{\mathrm{UOT}}.$$

Thus, π^{UOT} is also a POT_{p,M_C} plan and by equality of the penalty terms, π^{POT} is also a solution to $\mathrm{KR}_{p,C}$. Hence, for any $C > 0$ there exists an $M > 0$ such that the solutions to the two problems coincide. To see that the opposite is not true it suffices two consider $\mu = \delta_0$ and $\nu = \delta_1$. For $C < 1$ the UOT plan is given by $\pi = 0$ and for $C > 1$ the UOT plan is equal to $\pi = \delta_{(0,1)}$. For $C = 1$ both plans are optimal. Thus, for all $C > 0$ any UOT plan has either mass 0 or 1. Hence, for e.g. $M = 0.5$, there is no $C > 0$ such that the solutions to UOT and POT coincide. However, inferring which M corresponds to a given C requires solving the $\mathrm{KR}_{p,C}$ problem. Despite their close connection, the two models have quite different viewpoints. For $M = 0$ and $C < d_{\min}$ for disjoint measures μ,ν both approaches yield a plan with mass zero. For $M = \min(\mathbb{M}(\mu), \mathbb{M}(\nu))$ and $C > \mathrm{diam}(\mathcal{Y})$ both approaches intuitively transport as much mass as possible between the two measures. However, for intermediate values of M there is little intuition on the geometry of the POT plan, besides the fact that its total mass intensity is fixed. Hence, in this model it is not possible to incorporate geometric, structural information on the measures into the choice of the parameter. For the $\mathrm{KR}_{p,C}$ on the other side, there is a clear structural connection between the choice of C and the geometry of UOT plan as detailed before. In particular, C governs the largest scale at which transport can occur in OT plans. From a geometric viewpoint, the UOT model is therefore often preferable to the POT model. Though, naturally the POT model has the advantage if the applications provides prior information or constraints on the amount of mass to be transported. For geometrical data analysis however the former setting seems to be more likely to arise. More recently, the rich class of *entropy-optimal transport (EOT)*[6] problems has been considered [Liero et al., 2018]. For a parameter $\lambda > 0$, a cost function c and an entropy function E, it is defined as

$$\mathrm{EOT}_\lambda(\mu^1, \mu^2) = \operatorname*{arg\,min}_{\pi \in \mathcal{M}(\mathcal{X}^2)} \sum_{k=1}^{M_1}\sum_{l=1}^{M_2} c(x_k^1, x_l^2)\pi_{kl} + \lambda\left(E(\mu^1, \pi_1) + E(\mu^2, \pi_2)\right),$$

where π_i denotes the i-th marginal of π. For a comprehensive, general overview of EOT see Liero et al. [2018]. Some specific choices of cost and entropy have received detailed attention. In the context of machine learning this concept has also been considered under

[6] Despite the similarity in names, this is not to be confused with the concept of entropy-regularised OT discussed above.

the name of *robust OT* [Mukherjee et al., 2021]. Setting E to be the *Kullback-Leibler divergence (KLD)*[7]

$$KL(\mu^1, \mu^2) = \sum_{x \in \mathcal{X}} \mu^1(x) \log\left(\frac{\mu^1(x)}{\mu^2(x)}\right),$$

and c as the squared Euclidean distance yields the *Gaussian Hellinger-Kantorovich distance*

$$GHK_\lambda(\mu^1, \mu^2) = \underset{\pi \in \mathcal{M}(\mathcal{X}^2)}{\arg\min} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} d_2^2(x_k^1, x_l^2)\pi_{kl} + \lambda\left(KL(\mu^1, \pi_1) + KL(\mu^2, \pi_2)\right).$$

As for POT it is difficult to infer properties of the optimal value or the optimal coupling from the value of λ. By construction, the GHK distance is increasing in λ, but any further connection is in general unclear.
A different approach which recently gained popularity is obtained by keeping E as the Kullback-Leibler divergence, but setting

$$c(x, y) = -\log(\cos_\sigma^2(d_2(x_k^1, x_l^2)),$$

where $\sigma \in (0, \pi/2]$ and $\cos_\sigma(z) = \cos(\min(z, \sigma))$. This yields the *Hellinger-Kantorovich* [Liero et al., 2018, Chizat et al., 2018a] distance (also known as *Wasserstein-Fisher-Rao distance*) given by

$$HK_{\lambda,\sigma}(\mu^1, \mu^2) = \underset{\pi \in \mathcal{M}(\mathcal{X}^2)}{\arg\min} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} -\log(\cos_\sigma^2(d_2(x_k^1, x_l^2))\pi_{kl} + \lambda\left(KL(\mu^1, \pi_1) + KL(\mu^2, \pi_2)\right).$$

While the effect of λ is similar to the previous case of the GHK distance, the HK distance has an additional parameter σ which is usually referred to as *cut-off locus*. It controls, in a non-trivial way, the maximum distance at which mass is still transported on the corresponding optimal plan. However, in general the relation between σ, λ and the optimal plan and value of the distance remains open.
Setting E to be the total variation distance and $\lambda = C^p$ recovers the (p, C)-KRD. Denote

$$\mathrm{EOT}_{p,\lambda}^{\mathrm{TV}}(\mu^1, \mu^2) = \underset{\pi \in \mathcal{M}(\mathcal{Y}^2)}{\arg\min} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} d^p(x_k^1, x_l^2)\pi_{kl} + \lambda\left(\mathrm{TV}(\mu^1, \pi_1) + \mathrm{TV}(\mu^2, \pi_2)\right). \quad (3.24)$$

The key property that needs to be established is the fact that restricting the minimisation in (3.24) to $\Pi_\leq(\mu^1, \mu^2)$ does change the result. Let π a solution of (3.24) such that w.l.o.g.

[7] The KLD is neither symmetric, nor does it fulfil the triangle inequality. Hence it is not a distance.

$\pi_1 \geq \mu_1$ and denote $\mu^1_> = \pi_1, \mu^2_> = \pi_2$. Define $\mu^1_\leq$ by

$$\mu^1_\leq(x) = \min\{\mu^1(x), \mu^1_>(x)\}$$

and let $\mu^2_\leq$ be the image of $\mu^1_\leq$ under π. It holds

$$\begin{aligned}
\mathrm{TV}(\mu^1, \mu^1_\leq) + \mathrm{TV}(\mu^1_\leq, \mu^1_>) &= \sum_{x\in\mathcal{X}_1} |\mu^1(x) - \mu^1_\leq(x)| + |\mu^1_\leq(x) - \mu^1_>(x)| \\
&= \sum_{x\in\mathcal{X}_1} \mu^1(x) - \mu^1_>(x) \\
&\leq \sum_{x\in\mathcal{X}_1} |\mu^1(x) - \mu^1_>(x)| \\
&= \mathrm{TV}(\mu^1, \mu^1_>),
\end{aligned}$$

where the second equality follows from the construction of $\mu^1_\leq$. Since by construction it holds $\mu^1_\leq \leq \mu^1_>, \mu^2_\leq \leq \mu^2_>$ and $\mathbb{M}(\mu^1_\leq) = \mathbb{M}(\mu^2_\leq), \mathbb{M}(\mu^1_>) = \mathbb{M}(\mu^2_>)$, it follows that

$$\begin{aligned}
\mathrm{TV}(\mu^1_>, \mu^1_\leq) &= \sum_{x\in\mathcal{X}_1} |\mu^1_>(x) - \mu^1_\leq(x)| = \sum_{x\in\mathcal{X}_1} \mu^1_>(x) - \mu^1_\leq(x) \\
&= \mathbb{M}(\mu^1_>) - \mathbb{M}(\mu^1_\leq) = \mathbb{M}(\mu^2_>) - \mathbb{M}(\mu^2_\leq) = \mathrm{TV}(\mu^2_>, \mu^2_\leq).
\end{aligned}$$

From this and the previous calculation it follows

$$\begin{aligned}
\mathrm{TV}(\mu^1, \mu^1_\leq) + \mathrm{TV}(\mu^2, \mu^2_\leq) &\leq \mathrm{TV}(\mu^1, \mu^1_>) - \mathrm{TV}(\mu^1_>, \mu^1_\leq) + \mathrm{TV}(\mu^2, \mu^2_>) + \mathrm{TV}(\mu^2_>, \mu^2_\leq) \\
&= \mathrm{TV}(\mu^1, \mu^1_>) + \mathrm{TV}(\mu^2, \mu^2_>).
\end{aligned}$$

Since $\pi^\leq$ is by construction a restriction of π, it also holds that

$$\sum_{k=1}^{M_1}\sum_{l=1}^{M_2} d^p(x^1_k, x^2_l)\pi^\leq_{kl} \leq \sum_{k=1}^{M_1}\sum_{l=1}^{M_2} d^p(x^1_k, x^2_l)\pi_{kl}.$$

Exchanging the roles of μ^1 and μ^2 yields the converse result for μ^2. Thus, if π is not a subcoupling, then it is possible to construct a subcoupling $\pi^\leq$ which has at least the same objective value. Therefore, it suffices to restrict the optimisation in (3.24) to the set $\Pi_\leq(\mu^1, \mu^2)$. From this it follows

$$\begin{aligned}
\mathrm{EOT}^{\mathrm{TV}}_{p,\lambda}(\mu^1, \mu^2) &= \operatorname*{arg\,min}_{\pi\in\Pi_\leq(\mu^1,\mu^2)} \sum_{k=1}^{M_1}\sum_{l=1}^{M_2} d^p(x^1_k, x^2_l)\pi_{kl} + \lambda\left(\mathrm{TV}(\mu^1, \pi_1) + \mathrm{TV}(\mu^2, \pi_2)\right) \\
&= \operatorname*{arg\,min}_{\pi\in\Pi_\leq(\mu^1,\mu^2)} \sum_{k=1}^{M_1}\sum_{l=1}^{M_2} d^p(x^1_k, x^2_l)\pi_{kl} + \frac{\lambda}{2}\sum_{x\in\mathcal{X}} (\mu^1(x) - \pi_1(x) + \mu^2(x) - \pi_2(x))
\end{aligned}$$

$$= \operatorname*{arg\,min}_{\pi \in \Pi_{\leq}(\mu^1,\mu^2)} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} d^p(x_k^1, x_l^2)\pi_{kl} + \lambda \left(\frac{\mathbb{M}(\mu^1) + \mathbb{M}(\mu^2)}{2} - \mathbb{M}(\pi) \right).$$

Hence, it holds $\mathrm{EOT}_{p,\lambda}(\mu^1, \mu^2) = \mathrm{KR}^p_{p,\lambda^{1/p}}(\mu^1, \mu^2)$ for all $\mu^1, \mu^2, \mathcal{M}_+(\mathcal{Y})$.
One notable feature of the EOT problems in general is that their solutions can be approximated with a modification of entropy regularised OT [Chizat et al., 2018b]. While this is naturally prone to the same computational difficulties emerging from balancing the penalisation between approximation accuracy and numerical stability that entropy regularised OT is faced with for balanced OT, this still enables fast approximations of the corresponding EOT plans.

UOT Barycenter

Any concept of UOT which defines a distance d_{UOT} on $\mathcal{M}_+(\mathcal{Y})$ can also be used to define notion of barycenter of measures $\mu^1, \ldots, \mu^J \in \mathcal{M}_+(\mathcal{Y})$ similar to the (p, C)-barycenter given by

$$\mu^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathcal{Y})} \sum_{i=1}^{J} d_{\mathrm{UOT}}(\mu, \mu^i). \tag{3.25}$$

However, as for the distances themselves, there is limited insight into the geometrical properties of these barycenters. In particular, even some natural properties of the p-Wasserstein barycenter do not apply in general. For instance, for the HK-barycenter, the barycenter of finitely supported measures does not necessarily have finite support. Characterising the support of the HK-barycenter is also difficult even if it is finitely supported. Even for three Dirac measures on $\mathbb{R}^d$ describing the HK-barycenter is highly involved and requires a multitude of case distinctions between the locations of three measures and the masses on these locations. In particular, the information on the supports of the J measures and the model parameter of the HK-barycenter are not sufficient to understand its support structure. For a detailed analysis of this example see Friesecke et al. [2021].
Recalling the previously discussed equivalence between the $POT_{p,M}$ plan and the $\mathrm{UOT}_{p,C}$ plan, it is natural to extend this comparison to the context of barycenters. A straightforward notion of $POT_{p,C}$ barycenter is given by

$$\mu^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathcal{Y})} \sum_{i=1}^{J} POT_{p,M}(\mu, \mu^i). \tag{3.26}$$

However, this problem is ill-posed, since $POT_{p,M}$ is not well-defined for $\mathbb{M}(\mu) \leq M$. Thus, the optimisation has to be restricted to measures of total mass intensity at least M:

$$\mu^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathcal{Y}), \mathbb{M}(\mu) \geq M} \sum_{i=1}^{J} POT_{p,M}(\mu, \mu^i). \tag{3.27}$$

Unfortunately, this definition is problematic for any sort of geometrical data analysis. Since the deviation of mass between the measure μ in (3.26) and the μ^i is not penalised, choosing $\mu = \sum_{i=1}^{J} \mu^i$ yields an objective value of zero for any M and is hence optimal. A more suitable notion of barycenter might be found in the restriction

$$\mu^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathcal{Y}), \mathbb{M}(\mu) = M} \sum_{i=1}^{J} POT_{p,M}(\mu, \mu^i), \tag{3.28}$$

where the previous issue is resolved by fixing the mass of the barycenter to be M. By construction, for each $i = 1, \ldots, J$ the entire mass of $\mu^\star$ is transported to μ^i. In particular, this implies that any support point x of $\mu^\star$ is a p-barycenter (with respect to d) of points $x_1, \ldots, x_J$ with $x_i \in \operatorname{supp}(\mu^i)$ for all $i = 1, \ldots, J$. Consequently, any such POT barycenter is supported on the centroid set C of the p-Wasserstein barycenter in (1.11). In particular, the $POT_{p,M}$-barycenter is significantly less flexible than the (p, C)-barycenter, as it is restricted to the same structure of support set as the p-Wasserstein barycenter. Thus, it is likely to fail capturing the geometry of certain, e.g. clustered, support sets. One approach to rectify this, is to allow different values of M for the individual μ^i. Fix $M > 0$ and $M_1, \ldots, M_J \leq M$ and consider

$$\mu^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathcal{Y}), \mathbb{M}(\mu) = M} \sum_{i=1}^{J} POT^p_{p,M_i}(\mu, \mu^i). \tag{3.29}$$

Any solution of (3.29) is supported on the set

$$\begin{aligned} C^{POT}_{full} = \Big\{ y \in \mathcal{Y} \mid \exists L = 1, \ldots, J,\ \exists (i_1, \ldots, i_L) \subset \{1, \ldots, J\}, \\ x_1, \ldots, x_L :\ x_l \in \operatorname{supp}(\mu^{i_l}) \\ \forall l = 1, \ldots, L :\ y = T^{L,p}(x_1, \ldots, x_L) \Big\}, \end{aligned} \tag{3.30}$$

This set is in particular larger than the centroid set for the (p, C)-barycenter problem, as it allows all values of $L = 1, \ldots, J$ instead of requiring $L \geq J/2$. However, while knowledge on the geometric structure of the measures allows to tune the value of C for the (p, C)-barycenter to certain measures, there is far less intuition on how choosing $M, M_1, \ldots, M_J$ influences the $POT_{p,M}$-barycenter. This formulation of the

POT-barycenter is closely related to the (p, C)-barycenter. Let $\mu^\star_{KR}$ be a (p, C)-barycenter of $\mu_1, \dots, \mu_J$ and let π^i be an UOT plan between $\mu^\star_{KR}$ and μ^i. Set $M = \mathbb{M}(\mu^\star)$ and $M_i = \mathbb{M}(\pi^i)$ for $i = 1, \dots, J$, then it holds

$$\frac{1}{J}\sum_{i=1}^{J} POT_{p,M_i}(\mu^i, \mu^\star_{KR}) = \frac{1}{J}\sum_{i=1}^{J} \mathrm{KR}^p_{p,C}(\mu^i, \mu^\star_{KR}).$$

In particular, by the same argument as for the UOT plans, this implies that for any $C > 0$ there exist $M, M_1, \dots, M_J$ such that any (p, C)-barycenter is also a solution to (3.29). However, again the converse does not hold true. As before it suffices to consider $J = 2$ two measures $\mu_1 = \delta_0$ and $\mu_2 = \delta_1$. For any $C > 0$ the (p, C)-barycenter of μ_0 and μ_1 has either mass 0 or 1. Hence, for any other value of M there exists no (p, C)-barycenter problem which is equivalent to this POT-barycenter problem.

Statistical Modelling for alternative UOT Models

Naturally, the statistical models discussed above can also be applied to the alternative formulations of UOT discussed in the previous section. For the POT model deviation bounds on the expected error $\mathbb{E}\left[POT_M(\mu^1, \hat{\mu}^1)\right]$ can be derived from deviation bounds on (p, C)-KRD. For any M for which there exists a C_M such that the two problems are equivalent, the corresponding deviation bounds for $\mathrm{KR}_{p,C_M}(\mu^1, \hat{\mu}^1)$ apply. If there is no such C_M, then it is still possible to choose the C_{M_0} corresponding to closest value $M_0 > M$ for which the equivalence holds. In this case the POT_M error can still be controlled by the $\mathrm{KR}_{p,C_{M_0}}$ error. A similar argument can be used to control the Picolli-Rossi distance for $p = 1$ due to its equivalence to the (p, C)-KRD. Beyond the models of UOT which have a direct correspondence to a certain variant of the (p, C)-KRD, the statistical deviation of said distances have to be controlled explicitly. In theory, it would be possible to utilise the same approach of approximating the ground space by a tree and then control the distance on this tree. However, while the (p, C)-KRD, similar to the p-Wasserstein distance, enjoys a closed form solution on ultrametric trees, there are currently no analog results on the other UOT functionals considered. If for any UOT formulation a similar formula or at least a sharp upper bound of it on ultrametric trees could be derived, then this could be used to recreate the deviation bounds on the (p, C)-KRD. One bound for which this would not be sufficient though, is the one on the distance between the sets of population level and empirical barycenters. Their proofs make particular use of the structural properties of the (p, C)-barycenter which allow to pose it as the solution to a LP problem. Without an analog characterisation an alternative proof strategy for this bound would be necessary. In particular, this approach

fails if for UOT barycenters which are not guaranteed to have finite support.
One possible avenue to recreate the deviation bounds when there is no knowledge on specific properties of the UOT model on trees, is to make use of a possible dual formulation of the problem. Classical OT omits a strongly related dual formulation (recall LPs and their dual formulation in Chapter 1) and many UOT models, such as the (p, C)-KRD or any EOT model [Liero et al., 2018] yields an equivalent dual problem. In particular, the dual problem in this context can be understood as taking the supremum of an expectation of a function class. This naturally leads to the rich field of empirical process theory (for an overview see Vaart and Wellner [1996] and Wainwright [2019]). Similar arguments have already been used to establish deviation bounds in the usual OT context [Chizat et al., 2020]. These line of reasoning might lead to a general statistical theory for any EOT models. In particular, it could allow for extensions of the results presented in this chapter to a (p, C)-KRD defined for measure without finite support. Pursuing this sort of generalisations and extensions is an interesting direction for further research.
Another interesting question is whether considering the dual formulation of the (p, C)-barycenter problem directly and then using tools from empirical process theory would allow to recover a phenomenon similar to the lower complexity adaptation [Hundrieser et al., 2022]. The constants in empirical deviation bounds in this chapter depend on the average of the constants arising from J individual estimation problem. It seems worthwhile to investigate whether this can be improved to depend on the minimum of these values instead of their mean.

Alternative Statistical Modelling

Besides considering the properties of the statistical models in alternative UOT models, another obvious avenue of extension, is to consider different statistical models to generate the empirical measures. It should be stressed that the tree-approximation does not only allow to derive bounds for the three models previously discussed, but creates a framework which theoretically allows to analyse a range of statistical models. The key feature required to deploy this approach for a given statistical model is control on the mean absolute deviation of the mass placed into a given subtree of the tree approximation. Highly beneficial in this regard is the ability to aggregate the individual random weights within this subtree efficiently. This is for instance the case for the multinomial and Poisson model. There, it is irrelevant which specific points within a subtree are assigned the mass, since the distribution of the sum of the individual masses is easily derived and controlled. As long as the distribution has second moments, the mean absolute deviation can then be controlled by the root of the variance, which is a

good approximation in many scenarios.

Let $u_x \sim \mathcal{P}_{x,\theta}$ for all $x \in \mathcal{X}$ for some parametric family $\mathcal{P}_{x,\theta}$ and some parameter θ. Now, let $\hat{\mu}$ be an unbiased estimator based on the u_x. The key step now is to control $\mathbb{E}|\mu(A) - \hat{\mu}(A)|$ for specific sets $A \subset \mathcal{X}$ by a function of $\phi(\theta)$. If there exists θ_0, such that $\phi(\theta) \to 0$ for $\theta \to \theta_0$, then $\mathbb{E}\left[\mathrm{KR}^p_{p,C}(\mu, \hat{\mu})\right]$ converges at the same rate.

One way to construct such an estimator, similar to the approach used to construct the estimator in the Bernoulli model, is to set

$$\hat{\mu} = \sum_{x \in \mathcal{X}} \frac{u_x}{\mathbb{E}[u_x]} \delta_x.$$

This is an unbiased estimator by construction. Assume that the random variables u_x are independent. Bounding the mean absolute deviation by the variance yields a naive upper bound. It remains to control

$$\mathrm{Var}(\hat{\mu}(A)) = \sum_{x \in A} \frac{\mathrm{Var}(u_x)}{\mathbb{E}[u_x]^2} = \sum_{x \in A} \left(\frac{\mathbb{E}\left[u_x^2\right]}{\mathbb{E}[u_x]^2} - 1 \right). \tag{3.31}$$

Hence, the determining factor in the convergence in this model is the rate at which the ratio of expectation of the square and square of the expectation converges to one. Thus, if $\mathbb{E}\left[u_x^2\right] / \mathbb{E}[u_x]^2 \to 1$ for $\theta \to \theta_0$ for all $x \in \mathcal{X}$ at a rate of $\phi(\theta)$, then $\mathbb{E}\left[\mathrm{KR}^p_{p,C}(\mu, \hat{\mu})\right]$ converges at a rate of $\sqrt{\phi(\theta)}$. It should be stressed though, that this approach does in general not allow for the aggregation of the weights of the empirical measure on the subset A. It is also likely to yield problematic constants in many scenarios, as the pointwise consideration of the elements of A is likely to translate to a $\sqrt{|\mathcal{X}|}$ dependence for the empirical (p, C)-KRD. This is essentially the dependence obtained by controlling the total variation distance between the measures and ignoring the geometry of their supports entirely. Statistical models where it is possible to control the variance (or even better the mean absolute deviation) of $\hat{\mu}(A)$ directly are therefore preferable for this framework.

For a simply example where the ratio in (3.31) does not converge assume that u_x follows an exponential distribution with mean $\mu(x)\theta$ for some parameter $\theta > 0$ for all $x \in \mathcal{X}$. Since for the exponential distribution it holds $\mathrm{Var}(u_x) = \mathbb{E}[u_x]^2$, the ratio is one and does not converge to zero for any sequence of θ. In particular, even though the estimator $\hat{\mu}$ is unbiased it does not converge to μ for any sequence of θ.

A more specific approach which is closely related to the multinomial model can be found by sampling without replacement instead of with replacement from μ. In particular, this guarantees the samples $X_1, \ldots, X_N \sim \mu$ to be distinct. For the purpose of randomised computations, the population level measures are known, thus it is reasonable to consider

the estimator

$$\hat{\mu}_N^1 = \frac{\mathbb{M}(\mu^1)}{\sum_{k=1}^N \mu^1(X_k)} \sum_{k=1}^N \mu^1(X_k)\delta_{X_k}. \tag{3.32}$$

One advantage of this model in the context of randomised algorithms compared to the resampling one in (3.17), is the fact that it fully utilises the knowledge on the population measure instead of just their total mass intensity. Additionally, it also controls the computational complexity accurately, as the number of support points of the estimators is always known to be exactly N, while for the resampling approach it can be less than N (in which case a larger sample size would have been possible). As the sample size approaches the support size of the measure, the estimation error vanishes. In particular, for sufficiently large sample size, the subsampling approach is always superior to the resampling one which does not achieve an error of zero for finite sample sizes. However, for smaller sample size, this does not necessarily hold true. In particular, it is possible for the resampling approach to yield a lower error than the subsampling one at small sample sizes and simulations suggest that there are examples (even ones in real data applications) where this occurs. However, outside of simulation studies this effect is hard to quantify. The main reason for this, is the fact that it is difficult to establish deviation bounds for the subsampling model due to the dependency structure within the samples. These dependencies significantly complicate the problem and lead to the analysis of Wielands noncentral hypergeometric distribution for which there is no closed form solution for the mean and variance known. However, due to the importance of randomised methods for large scale OT, it seems highly worthwhile to investigate the relation between these sampling models in more detail. Particularly, the answer to the question in which scenario which approach is preferable is a clear target for further research.

Bibliography

A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. Convergence rates for empirical barycenters in metric spaces: Curvature, convexity and extendable geodesics. *Probability Theory and Related Fields*, 177(1):323–368, 2020.

J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances In Neural Information Processing Systems*, 30, 2017.

J. M. Altschuler and E. Boix-Adsera. Wasserstein barycenters are NP-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2): 389–409, 2016.

J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

D. P. Bertsekas. A distributed algorithm for the assignment problem. *Laboratory for Information and Decision Systems Working Paper, MIT*, 1979.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

L. A. Caffarelli and R. J. McCann. Free boundaries in optimal transport and Monge-Ampere obstacle problems. *Annals of Mathematics*, pages 673–730, 2010.

G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49 (2):1385–1418, 2017.

J. Chevallier. Uniform decomposition of probability measures: Quantization, clustering and rate of convergence. *Journal of Applied Probability*, 55(4):1037–1045, 2018.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018a.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018b.

L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances In Neural Information Processing Systems*, 33:2257–2269, 2020.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances In Neural Information Processing Systems*, 26, 2013.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693. Proceedings of Machine Learning Research, 2014.

G. B. Dantzig. Programming in a linear structure. In *Bulletin of the American Mathematical Society*, volume 54, pages 1074–1074, 1948.

S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: Approximation by empirical measures. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 1183–1203, 2013.

J. J. Dongarra, P. Luszczek, and A. Petitet. The LINPACK benchmark: Past, present and future. *Concurrency and Computation: Practice and Experience*, 15(9):803–820, 2003.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm.

In *International Conference on Machine Learning*, pages 1367–1376. Proceedings of Machine Learning Research, 2018.

K. Fatras, Y. Zine, S. Majewski, R. Flamary, R. Gribonval, and N. Courty. Minibatch optimal transport distances; analysis and applications. *preprint arXiv:2101.01792*, 2021.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. Proceedings of Machine Learning Research, 2019.

A. Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956. doi: 10.4153/CJM-1956-045-5.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

G. Friesecke, D. Matthes, and B. Schmitzer. Barycenters for the Hellinger–Kantorovich distance over $\mathbb{R}^d$. *SIAM Journal on Mathematical Analysis*, 53(1):62–110, 2021.

A. Gavryushkin and A. J. Drummond. The space of ultrametric phylogenetic trees. *Journal of Theoretical Biology*, 403:197–208, 2016.

D. Ge, H. Wang, Z. Xiong, and Y. Ye. Interior-point methods strike back: Solving the Wasserstein barycenter problem. *Advances In Neural Information Processing Systems*, 32, 2019.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. Proceedings of Machine Learning Research, 2018.

S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer, 2007.

K. Guittet. Extended Kantorovich norms: A tool for optimization. Technical report, Technical Report 4402, Institut national de recherche en informatique et en automatique, 2002.

L. G. Hanin. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

F. Heinemann, M. Klatt, and A. Munk. Kantorovich-Rubinstein distance and barycenter for finitely supported measures: Foundations and algorithms. *preprint arXiv:2112.03581*, 2021.

F. Heinemann, M. Klatt, and A. Munk. Kantorovich-rubinstein distance and barycenter for finitely supported measures: Sampling and approximation. *in preparation*, 2022a.

F. Heinemann, A. Munk, and Y. Zemel. Randomized Wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM Journal on Mathematics of Data Science*, 4(1):229–259, 2022b.

F. L. Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics*, 20(1-4):224–230, 1941.

S. Hundrieser, T. Staudt, and A. Munk. Empirical optimal transport between different measures adapts to lower complexity. *preprint arXiv:2202.10434*, 2022.

L. V. Kantorovich. On the translocation of masses. In *Doklady Akademii Nauk SSSR*, volume 37, pages 199–201, 1942.

L. V. Kantorovich and S. Rubinstein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

B. R. Kloeckner. A geometric study of Wasserstein spaces: Ultrametrics. *Mathematika*, 61(1):162–178, 2015.

T. C. Koopmans. Efficient allocation of resources. *Econometrica: Journal of the Econometric Society*, pages 455–465, 1951.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

G. Kulaitis, A. Munk, and F. Werner. What is resolution? A statistical minimax testing perspective on superresolution microscopy. *The Annals of Statistics*, 49(4): 2292–2312, 2021.

T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917, 2017.

T. Le Gouic, Q. Paris, P. Rigollet, and A. J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *Journal of the European Mathematical Society*, 2022.

J. Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

T. Lin, N. Ho, and M. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pages 3982–3991. Proceedings of Machine Learning Research, 2019.

T. Lin, N. Ho, X. Chen, M. Cuturi, and M. Jordan. Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. *Advances In Neural Information Processing Systems*, 33:5368–5380, 2020.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

G. Luise, S. Salzo, M. Pontil, and C. Ciliberto. Sinkhorn barycenters with free support via Frank-Wolfe algorithm. *Advances In Neural Information Processing Systems*, 32, 2019.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Academie Royale des Sciences de Paris*, pages 666–704, 1781.

D. Mukherjee, A. Guha, J. M. Solomon, Y. Sun, and M. Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. Proceedings of Machine Learning Research, 2021.

R. Müller, D. Schuhmacher, and J. Mateu. Metrics and barycenters for point pattern data. *Statistics and Computing*, 30(4):953–972, 2020.

A. Munk, T. Staudt, and F. Werner. Statistical foundations of nanoscale photonic imaging. In *Nanoscale Photonic Imaging*, pages 125–143. Springer, Cham, 2020.

J. Nocedal and S. Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

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

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

B. Piccoli and F. Rossi. On properties of the generalized wasserstein distance. *Archive for Rational Mechanics and Analysis*, 222(3):1339–1365, 2016.

T. Ramasubban. The mean difference and the mean deviation of some discontinuous distributions. *Biometrika*, 45(3-4):549–549, 1958.

B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

M. Sommerfeld, J. Schrieber, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20:105–1, 2019.

K.-T. Sturm. Probability measures on metric spaces of nonpositive curvature, heat kernels and analysis on manifolds, graphs, and metric spaces. *Contemporary Mathematics*, 338:357–390, 2003.

A. N. Tolstoi. Metody nakhozhdeniya naimen'shego summovogo kilometrazha pri planirovanii perevozok v prostranstve. *Planirovanie Perevozok, Sbornik Pervy*, 1: 23–55, 1930.

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

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

L. A. Wolsey and G. L. Nemhauser. *Integer and Combinatorial Optimization*, volume 55. John Wiley & Sons, 1999.

Y. Xie, X. Wang, R. Wang, and H. Zha. A fast proximal point method for computing exact Wasserstein distance. In *Uncertainty in Artificial Intelligence*, pages 433–453. Proceedings of Machine Learning Research, 2020.

Addenda

The following three articles form the basis of this thesis. This short summary provides their respective reference and abstract.

Randomised Wasserstein Barycenter Computation: Resampling with Statistical Guarantees
Florian Heinemann, Axel Munk, and Yoav Zemel.
SIAM Journal on Mathematics of Data Science, 4(1):229–259

Abstract: We propose a hybrid resampling method to approximate finitely supported Wasserstein barycenters on large-scale datasets, which can be combined with any exact solver. Nonasymptotic bounds on the expected error of the objective value as well as the barycenters themselves allow to calibrate computational cost and statistical accuracy. The rate of these upper bounds is shown to be optimal and independent of the underlying dimen- sion, which appears only in the constants. Using a simple modification of the subgradient descent algorithm of Cuturi and Doucet, we showcase the applicability of our method on a myriad of simulated datasets, as well as a real-data example from cell microscopy which are out of reach for state of the art algorithms for computing Wasserstein barycenters.

Kantorovich-Rubinstein distance and barycenter for finitely supported measures: Foundations and Algorithms
Florian Heinemann, Marcel Klatt, and Axel Munk
preprint available arXiv:2112.03581, 2021

Abstract: The purpose of this paper is to provide a systematic discussion of a generalized barycenter based on a variant of unbalanced optimal transport (UOT) that defines a distance between general non-negative, finitely supported measures by allowing for mass creation and destruction modeled by some cost parameter. They are denoted

as Kantorovich-Rubinstein (KR) barycenter and distance. In particular, we detail the influence of the cost parameter to structural properties of the KR barycenter and the KR distance. For the latter we highlight a closed form solution on ultra-metric trees. The support of such KR barycenters of finitely supported measures turns out to be finite in general and its structure to be explicitly specified by the support of the input measures. Additionally, we prove the existence of sparse KR barycenters and discuss potential computational approaches. The performance of the KR barycenter is compared to the OT barycenter on a multitude of synthetic datasets. We also consider barycenters based on the recently introduced Gaussian Hellinger-Kantorovich and Wasserstein-Fisher-Rao distances.

Kantorovich-Rubinstein distance and barycenter for finitely supported measures: A statistical perspective
Florian Heinemann, Marcel Klatt, and Axel Munk
in preparation

Abstract: In this paper we propose and investigate specific statistical models and corresponding sampling schemes for data analysis based on unbalanced optimal transport for finitely supported measures. Specifically, we analyse Kantorovich-Rubinstein (KR) distances with penalty parameter $C > 0$ between measures generated by some underlying statistical model. The main result provides non-asymptotic bounds on the expected error for the empirical KR distance as well as for its barycenters. The impact of the penalty parameter C is studied in detail. Our approach allows for randomised computational schemes for UOT which can be used for fast approximate computations with any exact solver. Using synthetic and real datasets, we empirically analyse the behaviour of the expected errors in simulation studies and illustrate the validity of our theoretical bounds.

CHAPTER A

Randomised Wasserstein Barycenter Computation: Resampling with Statistical Guarantees

Randomised Wasserstein Barycenter Computation: Resampling with Statistical Guarantees

Florian Heinemann* Axel Munk *† Yoav Zemel ‡

May 27, 2021

Abstract

We propose a hybrid resampling method to approximate finitely supported Wasserstein barycenters on large-scale datasets, which can be combined with any exact solver. Nonasymptotic bounds on the expected error of the objective value as well as the barycenters themselves allow to calibrate computational cost and statistical accuracy. The rate of these upper bounds is shown to be optimal and independent of the underlying dimension, which appears only in the constants. Using a simple modification of the subgradient descent algorithm of Cuturi and Doucet, we showcase the applicability of our method on a myriad of simulated datasets, as well as a real-data example from cell microscopy which are out of reach for state of the art algorithms for computing Wasserstein barycenters.

1 Introduction

Recently, optimal transport (OT), and more specifically the Kantorovich (also known as Wasserstein) distance, have achieved renewed interested as they have been recognised as attractive tools in data analysis. Despite its conceptual appeal in many applications (e.g., Rubner et al. [59]; Evans and Matsen [31]; Klatt et al. [44]), optimal transport-based data analysis has been triggered on the one hand by recent computational progress (see e.g., Peyré and Cuturi [57], Schmitzer [62],

*Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen

†Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen

‡Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WB

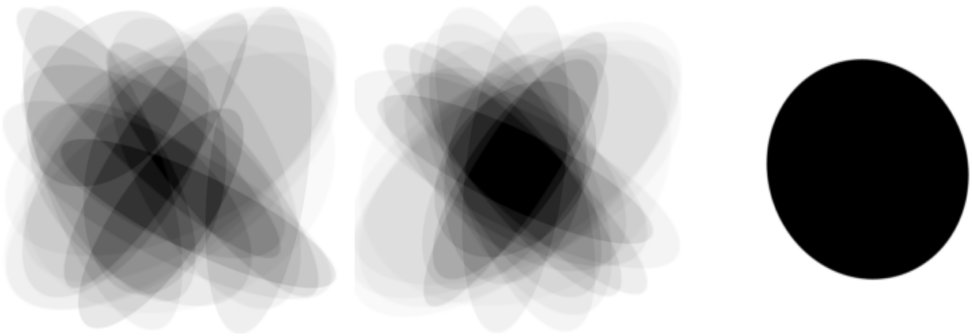


Figure 1: From left to right: Mean, Mean after recentering and Wasserstein barycenter of 20 randomly generated ellipses in $\mathbb{R}^2$.

Solomon et al. [64], Altschuler et al. [4]) and on the other hand by a refined understanding of its statistical properties when estimated from data (e.g., del Barrio et al. [22]; Sommerfeld et al. [65]; Weed and Bach [70]). This also lead to an increasing interest in Fréchet means, or barycenters, with respect to that distance. Since their introduction in the landmark paper of Agueh and Carlier [1], interest in the so-called Wasserstein barycenters sparked. Among the plethora of their potential applications, one can name unsupervised dictionary learning (Schmitz et al. [60]), distributional clustering (Ye et al. [72]), Wasserstein principal component analysis (Seguy and Cuturi [63]), neuroimaging (Gramfort et al. [39]) and computer vision (Rabin et al. [58]; Solomon et al. [64]; Bonneel et al. [13]). An appealing feature of Wasserstein barycenters is their ability to successfully capture geometric properties of complex data objects, thus allowing to define a meaningful notion of average for such objects.
Figure 1 illustrates this by displaying the, numerically computed, Wasserstein barycenter of a collection of ellipses (associated with uniform probability measures on their surface) along with their pixel-wise mean. It is seen that the former is a single "average" ellipse, whilst the latter gives a blurry object that is not representative of a "typical realisation" from this dataset. In fact, in this case the collection of probability measures forms a location-scatter family, and it is known (Álvarez et al. [5]) that the barycenter must again lie in the same family. Hence, the OT-barycenter of these ellipses is again an ellipse.

1.1 The OT-Barycenter Problem

One of the most fundamental questions in statistics and data analysis is inferring the mean of a random quantity on the basis of realisations $x_1, \dots, x_N$ thereof. Whilst this is conceptually straightforward when the data lie in a Euclidean space

or a Hilbert space, many datasets exhibit complex geometries that are far from Euclidean (e.g., Billera et al. [11]; Dryden et al. [25]; Bronstein et al. [17]), catalysing the emergence of the field of non-Euclidean statistics (e.g., Patrangenaru and Ellingson [56]; Huckemann and Eltzner [42]; Dryden and Marron [26]). Utilising the fact that the mean of points $x_1, \dots, x_N \in \mathbb{R}^D$ is characterised as the unique minimiser of $x \mapsto \sum_{i=1}^N \|x_i - x\|^2$, the notion of a barycenter (or Fréchet mean; Fréchet [33]; Huckemann et al. [41]) extends this to the non-Euclidean case by replacing the norm $\|x_i - x\|$ with an arbitrary distance function. Motivated from the preceding paragraphs, our work focusses on OT-barycenters, where the distance defining the barycenter is an optimal transport distance. More specifically, in this paper we are concerned with OT-barycenters of finitely supported probability measures i.e., collections of weighted points. To set up notation, consider a finite metric space $(\mathcal{X}, d)$ and N measures of the form

$$\mu_i = \sum_{k=1}^{M_i} b_k^i \delta_{x_k^i}, \quad x_k^i \in \mathcal{X}, \quad b_k^i \geq 0, \quad \sum_{k=1}^{M_i} b_k^i = 1, \qquad i = 1, \dots, N,$$

where δ_x is a Dirac measure at $x \in \mathcal{X}$. In imaging applications, the $x_k^i = x^{i\prime}$s could be points on a regular grid in $[0,1]^2$ with the weights b_k^i being the greyscale intensity of image k at pixel x^i. The above, more general formulation allows to also treat irregular points clouds that are unrelated to each other. In this setting the p-Wasserstein distance (Kantorovich [43]; Vaserstein [68]) between any two of these measures, μ_i and μ_j, is

$$W_p(\mu_i, \mu_j) = \left(\min_{\pi \in \Pi(\mu_i, \mu_j)} \quad \sum_{k=1}^{M_i} \sum_{l=1}^{M_j} \pi_{kl} d(x_k^i, x_l^j)^p \right)^{1/p},$$

where $p \geq 1$ and the set of **couplings** between μ_i and μ_j contains all the joint distributions on $\mathcal{X}^2$ having μ_i and μ_j as marginal distributions, and is given by

$$\Pi(\mu_i, \mu_j) := \left\{ \pi \in \mathbb{R}^{M_i \times M_j} | \pi \mathbf{1}_{M_i} = b_i, \mathbf{1}_{M_j}^T \pi = b_j^T \right\}, \qquad \mathbf{1}_M = (1, 1, \dots, 1) \in \mathbb{R}^M.$$

In particular, it can be shown that W_p is a distance on the space of probability measures on $\mathcal{X}$ (see e.g., Villani [69, Chapter 6]; or Zolotarev [74] for an early reference). To define barycenters with respect to this distance we need to choose an ambient space $\mathcal{Y}$ containing $\mathcal{X}$, since any reasonable choice of such a barycenter should be allowed to have mass at positions that may differ from the support points of the N data measures. Let $\mathcal{P}(\mathcal{Y})$ be the space of measures on $\mathcal{Y}$. The **p-Fréchet functional** $F : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+$ associated with $\mu_1, \dots, \mu_N$ is defined as

$$F^p(\mu) = \frac{1}{N} \sum_{i=1}^N W_p^p(\mu_i, \mu). \tag{1}$$

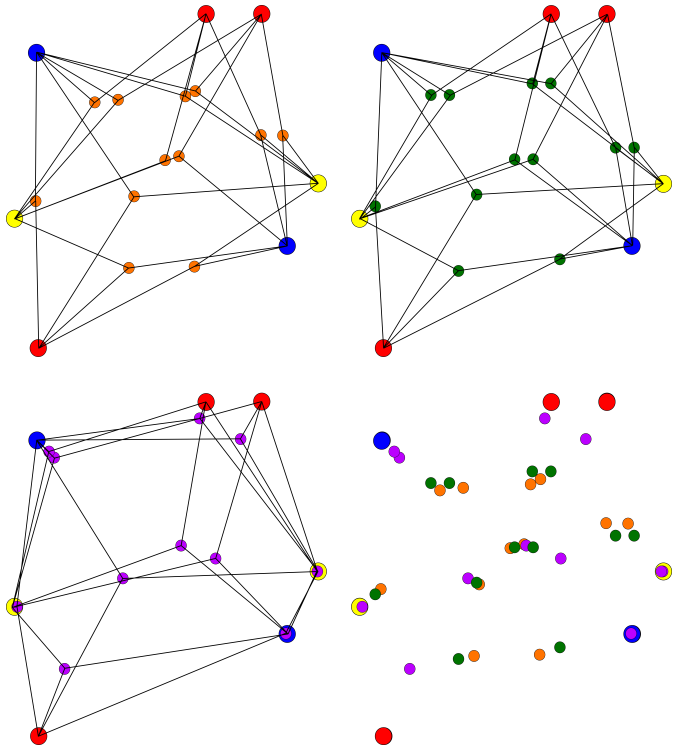


Figure 2: Three different centroid sets for the same set of three measures (red, blue, yellow). The lines connect to the support points of the measures from which a centroid point has been constructed. Here, it holds $|\mathcal{C}| = 12$. **Top left:** Centroids for $p = 3$ in orange. **Top right:** Centroids for $p = 2$ in green. **Bottom left:** Centroids for $p = 1$ in violet. **Bottom right:** All three centroids superimposed onto each other.

We call any minimiser of F^p a **p-Fréchet mean** or **p-Wasserstein barycenter** of $\mu_1, \dots, \mu_N$. When p is omitted, it is assumed to be equal to 2.
The starting point of our paper is an observation made by Le Gouic and Loubes [47, Theorem 8], who showed that when the ambient $\mathcal{Y}$ is geodesic (e.g., $\mathcal{Y} = \mathbb{R}^D$), any p-barycenter of $\mu_1, \dots, \mu_N$ is supported on the p-centroid set

$$\mathcal{C} := \left\{ \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{N} d^p(x_i, y) \middle| x_i \in \operatorname{supp}(\mu_i) \right\} = \left\{ \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{N} d^p(x^i_{k_i}, y) \middle| k_i \in \{1, \dots, M_i\} \right\},$$

where, by a slight abuse of notation, $d(\cdot, \cdot)$ denotes a metric on $\mathcal{Y}$ that extends the original metric d on $\mathcal{X}$ and supp denotes the support of the measure. Thus, even though $\mathcal{Y}$ is usually not a finite space, the minimisation can always be carried out

on measures supported on a finite set $\mathcal{C} \subseteq \mathcal{Y}$. For this reason, defining W_p for finitely supported measures suffices for the purpose of the present paper. In this setting of finitely supported measures barycenters always exist, but they are not necessarily unique.

The choice of the ambient space $\mathcal{Y}$ is not unique. Whilst it can typically assumed to be a Euclidean space $\mathbb{R}^D$, certain applications might warrant setting it to be a curved space of lower dimension $D' < D$ to better represent the given data. This could for instance apply to data on a sphere, where it might be natural to restrict the barycenter to this set as well. The upper bounds we provide in Section 3 depend on the dimension of the chosen ambient space $\mathcal{Y}$. It is usually the case that our bounds improve with decreasing dimension of the ambient space. Therefore, choosing a curved low-dimensional ambient space may not only better capture the structure of the data, but also improve the statistical guarantees on the approximation error.

We would like to stress that the definition and our later algorithms can be adapted in a straightforward way to Fréchet means with non-uniform weights other than $1/N$ in (1). We avoid this generality for brevity and simplicity. When $p = 2$ and d is the Euclidean metric, the argmin is the linear average $\sum x_i/N$ and $\mathcal{C}$ simplifies to

$$\mathcal{C} := \left\{ \frac{1}{N} \sum_{i=1}^{N} x_i \middle| x_i \in \operatorname{supp}(\mu_i) \right\} = \left\{ \frac{1}{N} \sum_{i=1}^{N} x_{k_i}^i \middle| k_i \in \{1, \dots, M_i\} \right\}.$$

As the set $\mathcal{C}$ plays an important role in the following, it is illustrated in Figure 2 for different powers of the Euclidean metric in $\mathbb{R}^2$. In particular, $\mathcal{C}$ is finite with cardinality at most $\prod_{i=1}^{N} M_i$ and the Fréchet functional can be expressed as a sum of sums over at most $|\mathcal{C}| \sum_{i=1}^{N} M_i$ summands. Identifying any candidate barycenter μ with its weights vector $a \in \mathbb{R}^{|\mathcal{C}|}$ on $\mathcal{C}$, this allows to rewrite the minimisation of the p-Fréchet functional as a linear program (LP) which computes optimal transport plans $\pi^{(i)} \in \Pi(\mu, \mu_i)$ between μ and μ_i for $i = 1, \dots, N$, whilst at the same time also minimising over the weights vector a of μ. The LP can then be denoted as

$$
\begin{aligned}
\min_{\pi^{(1)},\dots,\pi^{(N)},a} \quad & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|\mathcal{C}|}\sum_{k=1}^{M_i}\pi_{jk}^{(i)}c_{jk}^{i} \\
\text{subject to} \quad & \sum_{k=1}^{M_i}\pi_{jk}^{(i)} = a_j \quad \forall\, i=1,\dots,N, \quad \forall j=1,\dots,|\mathcal{C}| \\
& \sum_{j=1}^{|\mathcal{C}|}\pi_{jk}^{(i)} = b_k^i \quad \forall\, i=1,\dots,N, \quad \forall k=1,\dots,M_i \\
& \pi_{jk}^{(i)} \geq 0 \quad \forall i=1,\dots,N \quad \forall j=1,\dots,|\mathcal{C}|, \quad \forall k=1,\dots,M_i,
\end{aligned}
\tag{2}
$$

where $c_{jk}^i = d^p(\mathcal{C}_j, x_k^i)$ is the p-th power of the distance between the j-th point of $\mathcal{C}$ and the k-th point in the support of μ_i.
Unfortunately, the size of $\mathcal{C}$ typically grows as $\prod M_i$, rendering this linear program intractable for even moderate values of N and M_i (see Figure 2). Borgwardt and Patterson [15] provide some reformulations and improvements, but ultimately, this approach is currently infeasible for meaningful applications to large-scale data. To illustrate the size of this linear program consider a set of $N = 100$ greyscale images of size $M_i = M = 256 \times 256$. We assume the images to be probability measures supported on an equidistant grid in $[0,1]^2$ equipped with the Euclidean distance. Even if we exploit the fact that for $p = 2$ the set $\mathcal{C}$ is simply an N-times finer grid (see Anderes et al. [6]), this LP still has over 10^{15} variables and over 10^{10} constraints. If the support points would be in more general positions the LP would have over 10^{488} variables and over 10^{483} constraints. To put this into perspective, we note that in November 2020 the highest ranking supercomputer in the TOP500 list (for details on this list see Dongarra et al. [24]) had about 10^{15} bytes of RAM.

1.2 OT-Barycenter Computation

To overcome this computational obstacle and to utilise the impressive geometrical power of the OT-barycenter (recall Figure 1), there has been a great effort in constructing algorithms that yield approximations of the OT-barycenter whilst reducing the computational complexity of the exact LP-formulation by several orders of magnitude. In fact, it is known that already for three measures in two dimensions finding an exact Wasserstein barycenter is NP-hard, in general (Borgwardt and Patterson [14]).
Cuturi and Doucet [21] proposed a subgradient descent method to compute the best approximation of the Wasserstein barycenter, which is supported on a pre-specified support set. If we choose this set to be equal to $\mathcal{C}$, then this algorithm

approximates a true barycenter of the measures. However, as we discussed previously, the set $\mathcal{C}$ is far too large for efficient computations (see Figure 2 for a small scale example). Its size, which controls the number of variables in the linear program (2), can be as large as the support size of the product measure $\mu_1 \otimes \cdots \otimes \mu_N$ of the N data measures. Cuturi and Doucet also introduced an alternating, Lloyd-type procedure, that switches between optimising the weights of the barycenter measure on a specific support set and updating the support set according to the new weights. Since there always exists a barycenter with support size bounded above by $\sum_{i=1}^{N} |M_i| - N + 1$ (Anderes et al. [6]), we can simply choose a support set of this size to approximate the true barycenter. This option to compute an exact barycenter without using the set $\mathcal{C}$ comes at a tangible cost, however. This alternating procedure is highly runtime extensive, since it involves solving the fixed-support barycenter problem in each step. Still, it is vastly superior to the direct LP-approach. However, this alternating procedure yields a non-convex minimisation problem, which is prone to converge to local minima instead of global ones.

The computational burden can be alleviated by means of regularisation, most commonly in the form of an additive entropy penalty, which leads to the so-called Sinkhorn distance (Cuturi [20]). Cuturi and Doucet [21] compute the barycenter with respect to the Sinkhorn distance by means of subgradient descent. Benamou et al. [8] exploit the relation of entropy to the Kullback–Leibler divergence and solve this regularised problem by iterative Bregman projections. These approaches reduce the runtime by orders of magnitude, and approximate a regularised surrogate for the exact Wasserstein barycenter. Unfortunately, (entropic) regularisation is not without its drawbacks, particularly the choice of the regularisation parameter is a delicate issue. Although the regularised solution converges towards the real solution with maximal entropy as the regularisation parameter vanishes ([8]), computations become costly and numerically unstable for small parameters (Altschuler et al. [4]; Dvurechensky et al. [30]; Klatt et al. [44]), which reflects a "no free lunch" scenario. Workarounds like log-stabilisations (Schmitzer [62]) exist, but do not completely address this problem, as with increasingly smaller regularisation there can still be numerical instabilities, and the stabilisation sacrifices a significant portion of the computational gain resulting from the regularisation. Moreover, whilst runtime and memory requirements of the regularised methods scale linearly in the number of measures, the scaling in the support sizes of these measures is still problematic. There have been multiple iterations of improvements on the naive Sinkhorn algorithm (e.g., [4, 30]), but the approximation of Wasserstein barycenters already for relatively small ensembles of medium sized data objects, say 100 images with around 10^5 pixels each, becomes nevertheless infeasible in terms of computation time and required memory.

Recently, Xie et al. [71] have proposed an inexact proximal point method which allows to solve the fixed-support OT-barycenter problem at a similar complexity as the regularised problem, whilst avoiding to introduce a regularisation term. Their simulations also suggest that they achieve sharper images as barycenters compared to the standard entropy-regularised methods. However, their method has an increased memory demand and is still not applicable to large scale data as the scaling in the support size is still identical to the Sinkhorn algorithm. For further recent contributions, see e.g., Tiapkin et al. [67]; Ge et al. [34]; Dvurechenskii et al. [29]; Lin et al. [51]; Li et al. [50].
To summarise the above, the computational aspect of the OT-barycenter problem has attracted significant interest over the last years and an ever growing and improving toolbox of methods is available to tackle this problem. However, the size of problems which can be solved is still rather limited; modern applications e.g. in medical high-resolution imaging, are currently out of reach. Additionally, most methods consider the fixed-support OT-barycenter problem, where all measures are assumed to be supported on the same finite set. These methods cannot exploit sparsity in the support of the measures if it is present. In particular, if the measures have vastly different supports, fixed support methods are likely to be inefficient.

1.3 Our Approach

In this work we are primarily concerned with measures having potentially very different supports (as in the synthetic example in Figure 2). We decrease the problem size by generating random sparse approximations to the data measures. This will have a very limited effect on the runtime of fixed-support methods, but, for instance, the original, alternating subgradient descent by Cuturi and Doucet can fully exploit the advantages of this reduced problem size (see Section 4). In the same spirit any other method which takes advantage of sparse support aligns well with our resampling method. We stress that the proposed resampling method to compute random approximations of Wasserstein barycenters does not rely on any type of regularisation.
Our approach extends work by Sommerfeld et al. [66] on randomised optimal transport computation to the barycenter problem by replacing the original measures by their empirical counterparts, obtained from S independent random samples. Then, the true barycenter μ^* is estimated by the barycenter of the empirical measures $\widehat{\mu}_S^*$. We provide nonasymptotic L_1-type bounds on the objective values of μ^* and $\widehat{\mu}_S^*$, as well as on the distance between the set of empirical barycenters and the true ones. These bounds are optimal in terms of the dependence on the resampling size S. As an example, we will see that for N measures with M support points in

$[0,1]^D$, it holds that

$$\mathbb{E}[|F^2(\mu^*) - F^2(\widehat{\mu}_S^*)|] \leq 4D^{3/2}S^{-\frac{1}{2}} \begin{cases} 2+\sqrt{2} & D=1 \\ 2+\log_2(M) & D=2 \\ (3+\sqrt{2})M^{\frac{1}{2}-\frac{1}{D}} & D\geq 3 \end{cases} \tag{3}$$

where F^2 corresponds to $p=2$ in (1), i.e., the two barycenters are compared with respect to the W_2^2 distance. Two key aspects of our upper bound appear striking to us. First, the bound is uniform in the number of measures N, which has important consequences for randomised computations on a large number of datasets; and second, the convergence rate is $S^{-\frac{1}{2}}$, independently of the dimension D of the ambient space $\mathcal{Y}$ in which the measures reside (see [66] for $N=2$). This is in contrast to the case of absolutely continuous measures, where this rate is typically achieved when $D<2p$, whereas if $D>2p$, convergence holds at the slower rate $S^{-\frac{p}{D}}$ (see e.g., Dudley [27]; Dereich et al. [23]; Lei [49] and references therein), exhibiting a curse of dimensionality. In Figure 3 we showcase two examples of sampling approximations of barycenters in three dimensions, where we obtain visually striking approximations of a barycenter whilst reducing the support size of the measures by an order of magnitude. The bound (3) is particular case of our general results as stated in Theorem 3.1 below. It is important to stress that the upper bound depends on the intrinsic dimension of $\mathcal{X}$ and not on that of the ambient space $\mathcal{Y}$. If, for example, $\mathcal{X}$ can be embedded in a set that is a Lipschitz image of a lower dimensional cube $[0,1]^{D'}$ with $D'<D$, then (3) still holds with D replaced by D'. This is analogous to a similar well-known phenomenon in manifold learning (e.g., Genovese et al. [35]).

To the best of our knowledge, the best complexity bounds on the computation of the optimal transport for two measures are of order $\tilde{\mathcal{O}}(M^{\frac{5}{2}})$ (Lee and Sidford [48]), i.e., this resampling scheme reduces the runtime by a factor of order $\tilde{\mathcal{O}}(\left(\frac{M}{S}\right)^{\frac{5}{2}})$. For instance, by resampling 15% of the data points, the runtime is reduced by a factor of more than 100. For $N=2$ Sommerfeld et al. [66] report empirical relative errors of around 5% for $S=2000$ on 128×128 images, which translates to $S\approx 0.12M$. These results are striking and provided the motivation to extend this method to the barycenter problem for general N, which suffers even more than the optimal transport problem from prohibitive computational cost.

Furthermore, in the case $D=2$, the upper bound depends only logarithmically on the support size M. Thus, in the (ever important) case of two dimensional images, we have only logarithmic dependence on the resolution of the image. This suggests a good performance of our sampling method on high-resolution image-datasets. We explore this empirically in Section 4.

Finally, to perform our simulations, we leverage some empirical observations on the barycenters of uniform, finitely supported measures to modify an existing it-

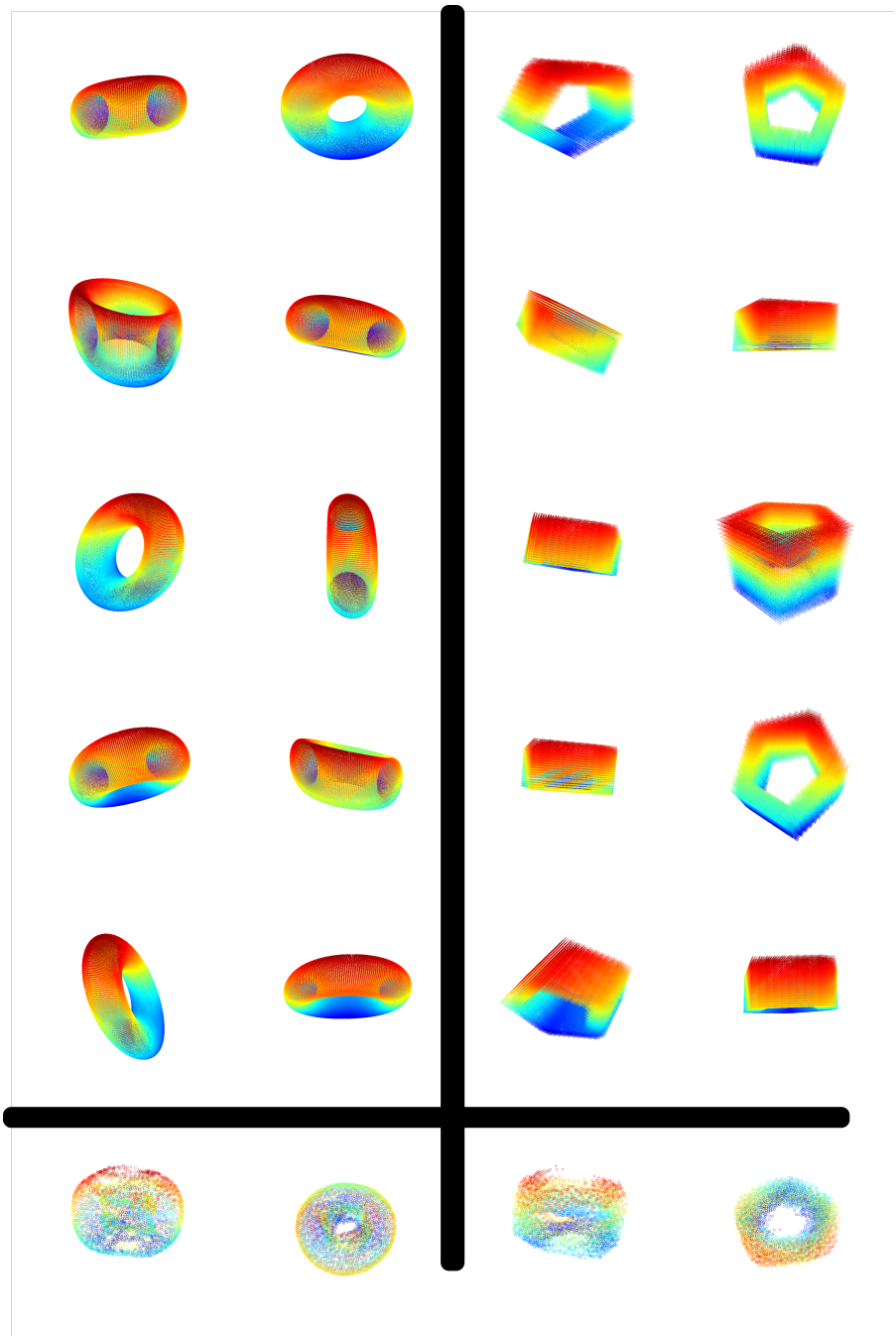


Figure 3: **Upper Left:** Projections of $N = 10$ torsos in $\mathbb{R}^3$, discretised on about $M = 62500$ points. **Lower Left:** Projected stochastic barycenter approximation with $S = 4000$ from the same angle (left) and from an adjusted angle (right). **Upper Right:** Projections of $N = 10$ pentagonal prisms in $\mathbb{R}^3$, discretised on about $M = 85000$ points. **Lower Right:** Projected stochastic barycenter approximation with $S = 4000$ from the same angle (left) and from an adjusted angle (right).

erative method for our needs in order to obtain a stochastic method that excels at solving the barycenter on large, sparsely supported datasets, which we in the following refer to as *Stochastic-Uniform-Approximation (SUA)–method*. We explore the quality of our bounds empirically in large-scale simulation studies and test the visual performance of our method on artificial and real data.

Population Barycenters. Let us stress that in this paper, the collection of measures $\mu_1, \dots, \mu_N$ is viewed as fixed. A different but related problem is that of estimating a population barycenter when the measures μ_i are realisations of a random probability measure $\boldsymbol{\mu}$ in $\mathcal{P}(\mathcal{Y})$ (i.e., the distribution of $\boldsymbol{\mu}$ is an element of $\mathcal{P}(\mathcal{P}(\mathcal{Y}))$); see e.g., Pass [55]; Bigot and Klein [10]. The minimiser of (1) is then an empirical barycenter, whose asymptotic properties (as $N \to \infty$) are studied in the form of consistency (Le Gouic and Loubes [47]), rates of convergence (Ahidar-Coutrix et al. [3]; Le Gouic et al. [37]) and, in very specific cases, central limit theorems (Panaretos and Zemel [54]; Agueh and Carlier [2]; Kroshnin et al. [46]). From a computational perspective, (stochastic) gradient descent-type methods have been shown to converge in some situations (Zemel and Panaretos [73]; Backhoff-Varaguas et al. [7]; Chewi et al. [19]). Most of these papers focus on absolutely continuous measures that are fully observed; closer to our context is the recent work of Dvinskikh [28], where the measures are discrete and the regularised and unregularised settings are both discussed. In summary, the stochasticity in the two problems is of a different nature: In the population barycenter context, it arises in $\mathcal{P}(\mathcal{Y})$, i.e., at the level of the measures; whereas in our setting it takes place on the space $\mathcal{Y}$ itself, i.e., at the level of each individual measure μ_i via the resampling.

Deterministic Approaches. Instead of generating *random* smaller-sized problem, one could approximate the data measures μ_i by some *deterministic* approximations supported on a small number of points. One immediate advantage of our random approach is that our error bounds readily apply in case the data measures are random realisations obtained from some experiment (also compare the previous paragraph on population barycenters). Our results in Section 3 then allow to control the error between the barycenter of the observed data and the barycenter corresponding to the underlying mechanism that generated the data measures.
Though, even if the μ_i's are considered to be fixed, the sampling approach has significant computational advantages, since generating the empirical measures μ_i^S can be done in negligible time. In contrast, a deterministic approximation inevitably involves some sort of optimisation problem. The related problem of finding the *quantiser*, i.e., the best S-supported measure approximating (in W_p) a given measure μ, which is extremely difficult even in one dimension (Graf & Luschgy [38]), provides a serious challenge that to some extent inspired the study of the rate

of convergence of the empirical measure (Dereich et al. [23]). Furthermore, the quantisers will typically not be uniform on their S support points, so computing the barycenter of the quantisers is more complicated than computing the barycenter of empirical measures, for which we can deploy a faster approach due to their uniform weights (see Section 4).

One may also consider the uniform quantisers $\mu^S_{i,unif}$, the best approximation that is uniform on S points (Chevallier [18]). These deterministic approximations converge, in the worst case, at rate $\log S/S$ (faster than $1/\sqrt{S}$ of μ^S_i), and, being uniform, allow for quick computation of the approximate barycenter (see Section 4). Notably, if we replace the empirical measures by uniform quantisers and invoke [18, Theorem 3.3], then we can replicate the proof of Theorem 3.1 and also obtain a worst case rate of $\log S/S$ for the error of the barycenter of the uniform quantisers in the Fréchet functional.

Whilst these stronger bounds are theoretically appealing, there are two significant drawbacks arising from this approach. Firstly, there is little theoretical control on the support of $\mu^S_{i,unif}$; as a consequence, the barycenter of the N uniform quantisers will usually not be supported on the centroid set $\mathcal{C}$, in which any true barycenter must lie. Secondly and more importantly, like quantisers with unrestricted weights, finding $\mu^S_{i,unif}$ or a reasonable surrogate thereof is computationally intractable unless M is small.

Reduction of the support size is not unlike the notion of multi-scale methods for optimal transport (see for instance Gerber and Maggioni [36], Merigot [52], Oberman and Yuanlong [53]). These start by computing a coarse (i.e., with small support size) approximation of the measures and solving the optimal transport between the coarse versions. One then uses the resulting solution as a good initial point to the optimal transport between finer approximations, and the procedure is iterated until the full-scale problem is solved. Thus, these methods speed up the computations substantially by finding a good initial point. Unfortunately, they will ultimately fail in large-scale problems for which even having a good initial point is insufficient. One may argue that in such circumstances one should stop the multi-scale approach on a smaller-scale version of the problem (for instance by merging adjacent grid points). This will inevitably cause a loss of information by blurring the data. In contrast, whilst our sampling approach also reduces the size of the problem, we still have a chance to observe any small feature of the data and if we take a sufficient amount of repeats, we will do that without ever having to move to the full-scale problems.

A potentially more suitable comparison might be given by the "shielding-neighbourhood" approach of Schmitzer [61]. It allows to solve dense, large-scale problems by solving a sequence of smaller, sparse problems instead. In particular, one can obtain an exact optimal solution of the full-scale problem without the need to solve it

directly. Shielding performs particularly well when the measures have regular supports, such as in the case of images. A challenge in applying this method is that without strong a-priori assumptions, it is not clear what size exactly the small problems should have, and their construction could be costly, potentially hindering the computational benefits. In problems having a less regular structure, this method is less efficient. One might then be willing to compute an inexact, approximate, solution, but it is difficult to tune the size of the smaller problems to match the desired computational effort. In contrast, our sampling approach allow for exact control of the size of the reduced problems, which enables precise control of the trade-off between desired computational effort and needed accuracy in the results.

1.4 Outline

The paper is stuctured as follows. Section 2 introduces the randomised optimal transport computation and gives a refined version of the results of [66] for $N = 2$. In Section 3 we extend this to the case $N > 2$ and show theoretical guarantees on the quality of the stochastic approximation. Finally, Section 4 presents numerical results on simulated and real data, and gives more details on the SUA-algorithm. Some proofs are omitted from the main text, and are developed in an Appendix. An implementation of SUA and a selection of the aforementioned algorithms is available as part of the CRAN R package `WSGeometry`.

2 Randomised Optimal Transport

Our approach will be based on reducing the size of the problem by choosing random approximations to the data measures. To this end, we draw an independent sample $X_1, \dots, X_S$ from each μ_i and use the *empirical measure*

$$\mu_i^S(x) := \frac{\#\{k|X_k = x\}}{S}, \quad x \in \mathcal{X}$$

as a proxy for μ_i. We begin with a discussion for $N = 1$ and provide a refined version of Theorem 1 in [66], which we require in the following.

Theorem 2.1. *Let μ be a measure on a finite space $\mathcal{X} = \{x_1, \dots, x_M\}$ endowed with a metric d, and let μ^S be the corresponding empirical measure obtained from a sample of size S from μ. Then*

$$\mathbb{E}[W_p^p(\mu^S, \mu)] \leq \frac{\operatorname{diam}(\mathcal{X})^p \mathcal{E}}{\sqrt{S}},$$

where the constant $\mathcal{E} := \mathcal{E}(\mathcal{X}, p)$ *is given by*

$$\mathcal{E} := 2^{p-1} \inf_{q>1, l_{\max}\in\mathbb{N}_0} q^p \left[q^{-(l_{\max}+1)p} M^{\frac{1}{2}} + \left(\frac{q}{q-1}\right)^p \sum_{l=1}^{l_{\max}} q^{-lp} \sqrt{\mathcal{N}(\mathcal{X}, q^{-l}\mathrm{diam}(\mathcal{X}))} \right].$$

Here $\mathcal{N}(\mathcal{X}, \delta)$ *denotes the* δ*-covering number of* $\mathcal{X}$*, and* $\mathrm{diam}(\mathcal{X}) = \sup_{x,y\in\mathcal{X}} d(x, y)$*. Moreover, if* $p = 1$ *the factor* $\left(\frac{q}{q-1}\right)^p$ *can be removed. When* $(\mathcal{X}, d) \subset (\mathbb{R}^D, \|\cdot\|_2)$*, we have for all integers* $q \geq 2$ *that*

$$\mathbb{E}\left[W_p^p(\mu, \mu^S)\right] \leq S^{-\frac{1}{2}} D^{\frac{p}{2}} 2^{p-1} \mathrm{diam}(\mathcal{X})^p q^p \begin{cases} \left(\frac{q}{q-1}\right)^p \frac{q^{p'}}{1-q^{p'}} & \text{if } p' < 0 \\ 1 + \left(\frac{q}{q-1}\right)^p D^{-1} \log_q M & \text{if } p' = 0 \\ M^{\frac{1}{2}-\frac{p}{D}} + \left(\frac{q}{q-1}\right)^p q^{p'} \frac{M^{\frac{1}{2}-\frac{p}{D}}}{q^{p'}-1} & \text{if } p' > 0, \end{cases}$$

where $p' = D/2 - p$ *and the factor* $\left(\frac{q}{q-1}\right)^p$ *can be omitted if* $p = 1$*. In particular, using* $q = 2$ *and* $p = 1$ *gives*

$$\mathbb{E}\left[W_1(\mu, \mu^S)\right] \leq S^{-\frac{1}{2}} 2 D^{\frac{1}{2}} \mathrm{diam}(\mathcal{X}) \begin{cases} 1 + \sqrt{2} & \text{if } D = 1 \\ 1 + 2^{-1} \log_2 M & \text{if } D = 2 \\ M^{\frac{1}{2}-\frac{1}{D}} + \frac{2^{D/2-1} M^{\frac{1}{2}-\frac{1}{D}}}{2^{D/2-1}-1} & \text{if } D > 2. \end{cases}$$

Theorem 2.1 improves upon [66, Theorem 1] by running the sum over l from 1 instead of zero and allowing a better prefactor than q^{2p}. The proof, however, follows similar lines as that of Theorem 1 in [66] and is based on constructing an ultrametric tree on $\mathcal{X}$ (see Kloeckner [45]) that dominates the original metric d (see also Boissard and Le Gouic [12]). The differences from [66] are given in Appendix A. In our proofs in Section 3, we shall apply the bound in Theorem 2.1 with $p = 1$, in which case the improvement over [66] is a factor of at least 2.

3 Empirical Wasserstein Barycenter

In this section we present our main results. The stochastic approximation in Theorem 2.1 naturally extends from the optimal transport problem to the barycenter problem by computing the barycenter of empirical versions of the N data measures (see Algorithm 1). We point out that at this point step 8 in Algorithm 1 could be performed with any solver for the barycenter problem (see Section 4); no specific method is needed at this step in order for Algorithm 1 to work. However, we will later utilise a specialised version of the iterative method introduced by Cuturi and Doucet [21], which we modify to obtain significantly better performance for

empirical measures (see Section 4). As in the classical optimal transport setting, we can average our results over multiple runs to reduce variability, leading to R empirical barycenters $\bar{\mu}_1, \dots, \bar{\mu}_R$. As a final estimator we take the linear average of those empirical barycenters. This is computationally preferable to using $\bar{\mu}_{r^*} = \operatorname{argmin}_{r \le R} F(\bar{\mu}_r)$, since evaluating the Fréchet functional amounts to solving N large-scale optimal transport problems, and is thus computationally costly. In fact, we expect the linear mean to have good performance since convexity of the Wasserstein distance extends to the Fréchet functional:

$$F^p\left(\frac{1}{R}\sum_{r=1}^{R}\bar{\mu}_r\right) \le \frac{1}{R}\sum_{r=1}^{R}F^p(\bar{\mu}_r), \quad p \ge 1. \tag{4}$$

For a more detailed discussion on the choice of R as well as the estimator obtained from R repeats, we refer to Subsection 4.2.1.

We now provide the theoretical justification for our resampling method, by giving nonasymptotic bounds on the expected error of the empirical barycenter. These exhibit convergence rate of $S^{-\frac{1}{2}}$ independently of the dimension. It will be assumed henceforth that $(\mathcal{X}, d)$ is a subspace of a geodesic space $\mathcal{Y}$, so that the set $\mathcal{C}$ is well-defined. Since any normed vector space is a geodesic space, our results are valid, in particular, when $\mathcal{X}$ can be embedded in a Euclidean space of arbitrary dimension. The minimum in the Fréchet functional is taken over all measures in $\mathcal{P}(\mathcal{Y})$, but can also be equivalently taken only on $\mathcal{P}(\mathcal{C})$. In view of the above inequality (4), it suffices to consider the case $R = 1$ in the following for simplicity (see Remark 3.5 for details).

Algorithm 1 Sampling approximation of the Wasserstein barycenter

Data Measures: $\mu_1, \dots, \mu_N$, sample size S, repeats R
for $r = 1, \dots, R$ **do**
  for $i = 1, \dots, N$ **do**
    Draw $X_1^{(i)}, \dots, X_S^{(i)} \sim \mu_i$
    $\mu_i^S = \frac{1}{S}\sum_{k=1}^{S}\delta_{X_k^{(i)}}$
  end for
  Solve $\bar{\mu}_r \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathcal{C})} \frac{1}{N}\sum_{i=1}^{N} W_p^p(\mu, \mu_i^S)$
end for
Set $\widehat{\mu}_S^* = \frac{1}{R}\sum_{r=1}^{R}\bar{\mu}_r$ **return** Approximation of the empirical Wasserstein barycenter $\widehat{\mu}_S^*$

Theorem 3.1. *Let $\mu_1, \ldots, \mu_N$ be probability measures on $\mathcal{X}$ and let $\mu_1^{S_1}, \ldots, \mu_N^{S_N}$ be the corresponding empirical measures based on $S_i \in \mathbb{N}$ independent samples. Let F^p and $\widehat{F}_S^p$, $p \geq 1$, denote the respective p-Fréchet functionals given by*

$$F^p(\mu) := \frac{1}{N} \sum_{i=1}^{N} W_p^p(\mu_i, \mu), \quad \widehat{F}_S^p(\mu) = \frac{1}{N} \sum_{i=1}^{N} W_p^p(\mu_i^{S_i}, \mu)$$

and let μ^* and $\widehat{\mu}_S^*$ be their respective minimisers. Then

$$\mathbb{E}[|F^p(\mu^*) - F^p(\widehat{\mu}_S^*)|] \leq \frac{2p \ \mathrm{diam}(\mathcal{X})^p}{N} \sum_{i=1}^{N} \frac{\mathcal{E}(supp(\mu_i), 1)}{\sqrt{S_i}},$$

for $\mathcal{E}$ as in Theorem 2.1. In particular, when $p = 2$ and $S_i = S$ for all $i = 1, \ldots, N$, we have

$$\mathbb{E}[|F(\mu^*) - F(\widehat{\mu}_S^*)|] \leq \frac{4 \ \mathrm{diam}(\mathcal{X})^2}{N\sqrt{S}} \sum_{i=1}^{N} \mathcal{E}(supp(\mu_i), 1).$$

Proof. In [65] it is shown that for arbitrary measures $\mu, \mu', \nu \in \mathcal{P}(\mathcal{X})$ and $p \geq 1$, it holds that

$$|W_p^p(\mu, \nu) - W_p^p(\mu', \nu)| \leq W_1(\mu, \mu') \ p \ \mathrm{diam}(\mathcal{X})^{p-1}.$$

Therefore, for any $\mu \in \mathcal{P}(\mathcal{S})$,

$$\begin{aligned}
\mathbb{E}\left[|F^p(\mu) - \widehat{F}_S^p(\mu)|\right] &\leq \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[|W_p^p(\mu_i^S, \mu) - W_p^p(\mu_i, \mu)|\right] \\
&\leq \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[W_1(\mu_i, \mu_i^S)\right] \mathrm{diam}(\mathcal{X})^{p-1} p \\
&\leq \frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{E}(\mathcal{X}_i, 1) \mathrm{diam}(\mathcal{X})^p p}{\sqrt{S_i}},
\end{aligned}$$

where $\mathcal{X}_i := \mathrm{supp}(\mu_i)$. Since μ^* and $\widehat{\mu}_S^*$ are minimisers of their respective Fréchet functionals, deduce that

$$\begin{aligned}
\mathbb{E}[|F^p(\widehat{\mu}_S^*) - F^p(\mu^*)|] &= \mathbb{E}[F^p(\widehat{\mu}_S^*) - F^p(\mu^*)] \\
&\leq \mathbb{E}[F^p(\widehat{\mu}_S^*) - \widehat{F}_S^p(\mu^*)] + \frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{E}(\mathcal{X}_i, 1) \mathrm{diam}(\mathcal{X})^p p}{\sqrt{S_i}} \\
&\leq \mathbb{E}[F^p(\widehat{\mu}_S^*) - \widehat{F}_S^p(\widehat{\mu}_S^*)] + \frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{E}(\mathcal{X}_i, 1) \mathrm{diam}(\mathcal{X})^p p}{\sqrt{S_i}} \\
&\leq \frac{2}{N} \sum_{i=1}^{N} \frac{\mathcal{E}(\mathcal{X}_i, 1) \mathrm{diam}(\mathcal{X})^p p}{\sqrt{S_i}}.
\end{aligned}$$

□

Using the bound on $\mathcal{E}$ from Theorem 2.1 on the space $\mathcal{X} = [0,1]^D$ with $p = 2$, we obtain

$$\mathbb{E}[|F^2(\mu^*) - F^2(\widehat{\mu}_S^*)|] \leq S^{-\frac{1}{2}} \begin{cases} 8\sqrt{2}(2 + \log_2(M)) & D = 2 \\ 183.5M^{\frac{1}{6}} & D = 3. \end{cases}$$

This result gives the rate of approximation in terms of the objective value. Approximating the optimisers is a more delicate matter, which can be addressed using the linear programming structure of the problem. Since the barycenters are not necessarily unique, the best we can hope for is to approximate one of them.

Theorem 3.2. *Let $\mu_1, \ldots, \mu_N$ be probability measures on $\mathcal{X}$ and let $\mu_1^S, \ldots, \mu_N^S$ be empirical measures obtained from S i.i.d. samples. Let $\mathbf{B}^*$ be the set of barycenters of the μ_i and $\mathbf{B}_S^*$ the set of barycenters of the μ_i^S. Then for $p \geq 1$*

$$\mathbb{E}\left[\sup_{\widehat{\mu}_S^* \in \mathbf{B}_S^*} \inf_{\mu^* \in \mathbf{B}^*} W_p^p(\mu^*, \widehat{\mu}_S^*)\right] \leq \frac{p\bar{\mathcal{E}}\ \mathrm{diam}(\mathcal{X})^p}{C_P} S^{-\frac{1}{2}},$$

where $\bar{\mathcal{E}} = \sum_{i=1}^N \mathcal{E}(supp(\mu_i), 1)/N$ and C_P is a strictly positive constant given by

$$C_P := C_P(\mu_1, \ldots, \mu_N) := (N+1)\mathrm{diam}(\mathcal{X})^{-p} \min_{v \in V \setminus V^*} \frac{c^T v - f^*}{d_1(v, \mathcal{M})},$$

where V is the vertex set of the linear programming formulation of the barycenter problem (2), V^* is the subset of optimal vertices, c is the cost vector of the program, f^* is the optimal value, $\mathcal{M}$ is the set of minimisers of the problem (2), and $d_1(v, \mathcal{M}) = \inf_{x \in M} \|v - x\|_1$.

Remark 3.3. *Using the convexity of P and the linearity of the objective function, one can show that the minimum in C_P is attained at a vertex $v \in V \setminus V^*$ which is adjacent to some $v^* \in V^*$.*

Remark 3.4. *A stronger version of Theorem 3.2 can be shown, if one uses total variation instead of the p-Wasserstein distance. This can be achieved by skipping the last inequality of the proof.*

Remark 3.5. *Theorem 3.2 can be generalised, using (4), to a general R, in which case the statement becomes slightly more complicated, however. For $r = 1, \ldots, R$ let $\boldsymbol{B}_{S,r}^*$ be the set of barycenters of the N empirical measures from the r-th repeat. The set $\boldsymbol{B}_S^*$ needs to be replaced by the set*

$$\left\{ \frac{1}{R} \sum_{r=1}^R \mu_r \;\middle|\; \mu_r \in \boldsymbol{B}_{S,r}^* \right\}$$

of measures that can be obtained as a linear average of barycenters from the R repeats. The upper bound remains the same.

Remark 3.6. *The rate $S^{-\frac{1}{2}}$ is optimal. This can already be seen in the case $N = 1$, where the Fréchet functional is simply the p-th power of the p-Wasserstein distance between a measure μ and its empirical version μ^S. For a finitely supported and nondegenerate measure μ, this has a lower bound scaling as $S^{-\frac{1}{2}}$ (e.g., Fournier and Guillin [32]), implying optimality of our rate. For a concrete example let $\mu = \frac{1}{2}(\delta_0+\delta_1)$. By construction we have for some random variable $K \sim Bin(S, 1/2)$ that*

$$\mathbb{E}\left[W_p^p(\mu, \mu^S)\right] = \mathbb{E}\left|\frac{1}{2} - \frac{K}{S}\right| = S^{-1}|\mathbb{E}[K] - K| \geq S^{-1}2^{-\frac{1}{2}}\sqrt{S/4} = \sqrt{2}S^{-\frac{1}{2}}/4,$$

where the inequality follows from the properties of the mean absolute deviation of the binomial distribution (Berend and Kontorovich [9]).

The proof of Theorem 3.2 requires two auxiliary, geometric results that may be of independent interest. To streamline the presentation, their proofs are given in Appendix A.
For a nonempty $P \subseteq \mathbb{R}^L$ and $x \in \mathbb{R}^L$, define $d_1(x, P) = \inf_{z\in P} \|x - z\|_1$. When P is closed, the infimum is attained, since it can be taken on the compact set $P \cap \{y : \|y - x\| \leq d_1(x, P) + 1\}$.

Lemma 3.7. *Let $P \subseteq \mathbb{R}^L$ be a nonempty polyhedron and let $x, y \in \mathbb{R}^L$. Then the function*

$$g(t) = d_1(x + ty, P) = \min_{z\in P}\|x + ty - z\|_1, \qquad t \in \mathbb{R}$$

is convex, Lipschitz and piecewise affine.

An illustration of Lemma 3.7 in $\mathbb{R}^2$ is given in Figure 4.

Lemma 3.8. *Let F^p be the Fréchet functional corresponding to $\mu_1, \ldots, \mu_N \in \mathcal{P}(\mathcal{X})$. Then for any $\mu \in \mathcal{P}(\mathcal{C})$ there exists a $\mu^* \in \arg\min_\mu F^p(\mu)$ such that*

$$F^p(\mu) - F^p(\mu^*) \geq 2C_P W_p^p(\mu, \mu^*),$$

where C_P is the constant from Theorem 3.2.

Sketch of proof of Lemma 3.8. Let $P \subset \mathbb{R}^L$ be the feasible polytope corresponding to (2) and let f^* be the optimal objective value of this linear program. To each $\mu \in \mathcal{P}(\mathcal{C})$ corresponds a $\pi \in P$ such that $c^T\pi = F(\mu)$. Fix an element $\pi^* \in \arg\min d_1(\pi, \mathcal{M})$, from which we can construct a minimiser μ^* of F. It holds that

$$F^p(\mu) - F^p(\mu^*) = c^T\pi - c^T\pi^* = \|\pi - \pi^*\|_1 \frac{c^T\pi - f^*}{\|\pi - \pi^*\|_1} \geq \|\pi - \pi^*\|_1\psi(\pi),$$

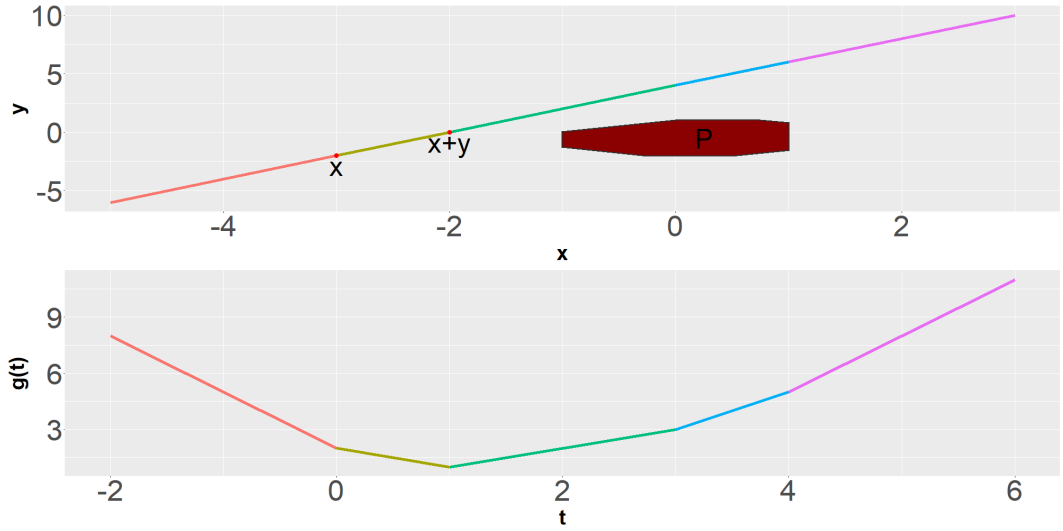


Figure 4: **Top:** the polytope P and the vectors $x = (-3, -2)$ and $y = (1, 2)$. **Bottom:** the function $g(t)$. The different colours correspond to segments on which g is affine. On the segment $t \in (1, 3)$ the minimiser in $d_1(x + ty, P)$ is not unique.

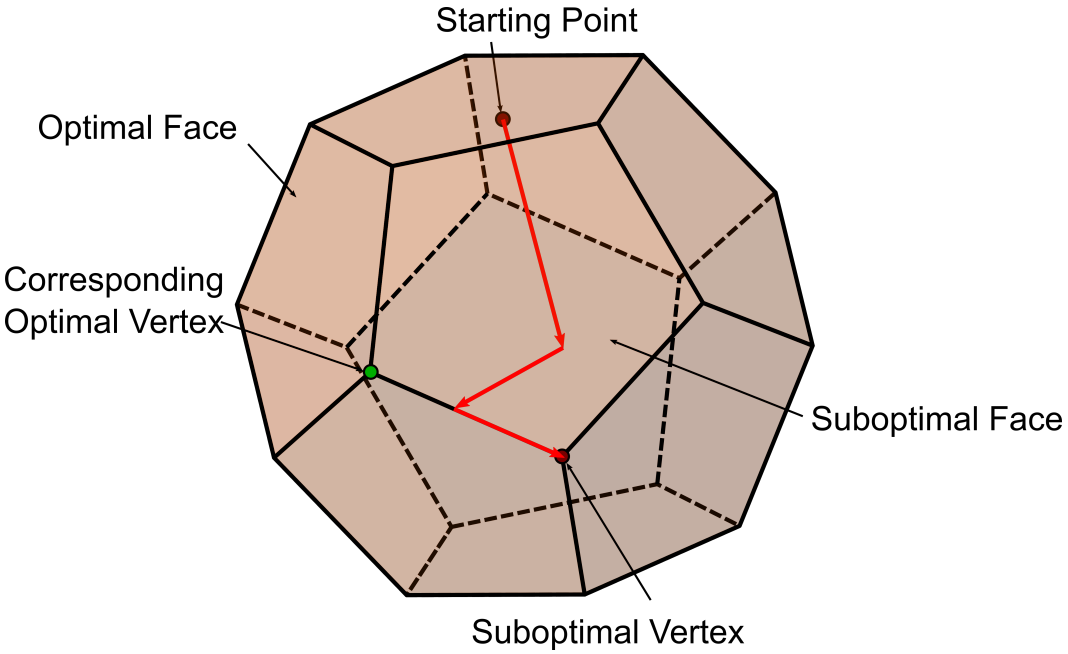


Figure 5: Sketch of a path of on which ψ is nonincreasing. We start at a point in the interior of P, then move to a face/subface until we hit a vertex.

where

$$\psi(\pi) = \frac{c^T\pi - f^*}{d_1(\pi, \mathcal{M})}, \qquad \pi \in P \setminus \mathcal{M}.$$

One can now minimise ψ over all elements of $\mathcal{P}$ and show that the minimum is positive and attained at a vertex of the polytope. The main idea of the proof is that for each point $\pi \in \mathcal{P}$ which is not a vertex, one can find a direction in $\mathcal{P}$ in which one can move without increasing the value of ψ. We can then move into this direction until we hit a face/subface of $\mathcal{P}$, at this point we can then construct a new direction along which we can move unless we have hit a vertex. A sketch of this in three dimensions can be seen in Figure 5. The value of ψ can be controlled using Lemma 3.7, as it is a ratio of an affine and a piecewise affine convex function. Finally, we can use the structure of the linear program to bound $\|\pi - \pi^*\|_1$ by the Wasserstein distance between the corresponding measures μ and μ^*. □

Proof of Theorem 3.2. Invoking Lemma 3.8 with $\mu \in \mathbf{B}_S^*$ and applying Theorem 3.1 yields

$$\frac{2p\bar{\mathcal{E}}\mathrm{diam}(\mathcal{X})^p}{\sqrt{S}} \geq \mathbb{E}\left[F^p(\widehat{\mu}_S^*) - F^p(\mu^*)\right] \geq \mathbb{E}\left[2C_P \sup_{\widehat{\mu}_S^* \in \mathbf{B}_S^*} \inf_{\mu^* \in \mathbf{B}^*} W_p^p(\widehat{\mu}_S^*, \mu^*)\right].$$

Thus

$$\frac{p\bar{\mathcal{E}}\ \mathrm{diam}(\mathcal{X})^p}{C_P} S^{-\frac{1}{2}} \geq \mathbb{E}\left[\sup_{\widehat{\mu}_S^* \in \mathbf{B}_S^*} \inf_{\mu^* \in \mathbf{B}^*} W_p^p(\mu^*, \widehat{\mu}_S^*)\right].$$

□

The constants C_P and $\bar{\mathcal{E}}$ are scale invariant, i.e., they remain the same if the metric d is multiplied by a positive constant. Finding a more explicit lower bound for C_P in specific cases (such as measures supported on a regular grid) is an important task for further research.

Remark 3.9. *If the number of measures N increases whilst S is kept fixed, then the total number of observations NS increases linearly in S. From a sample complexity perspective, one might be interested in fixing the total sample size $L = \sum_{i=1}^N S_i$. In view of Theorem 3.1, consistency of the Fréchet value is achieved if all the S_i's diverge to infinity. In the "uniform" case, where $S_i = L/N$, this translates to requiring $N/L \to 0$.*

4 Computational Results

4.1 The SUA-Algorithm

Over the past years many algorithms that allow to compute approximated Wasserstein barycenters have emerged. Most of them focus on either the regularised problem, or on computing the barycenter on a fixed, pre-specified support set. For reasons discussed in the introduction, we aim to solve the exact, unregularised problem. In the context of stochastic sampling, methods that consider a common, fixed support are inappropriate, as the resampling inevitably leads to empirical measures with different supports. Thus, we employ a modification of the iterative, alternating algorithm proposed by Cuturi and Doucet [21]. Alternating between position and weight updates for general measures creates a massive computational burden. However, there is an interesting empirical observation that renders the algorithm of [21] particularly appealing in the context of stochastic sampling.

Suppose that the measures $\mu_1, \dots, \mu_N$ are all uniform on the same number of points. In the notation of Section 1.1, we assume that $M_1 = M_2 \cdots = M_N = M$ and $b_k^i = 1/M$ for all $i = 1, \dots, N$ and all $k = 1, \dots, M$. When $N = 2$, then we can use the Birkhoff-von Neumann theorem to show that the barycenter of these two measures is also a uniform measure on M points, regardless of the cost function. If $N \geq 3$ (and $M \geq 3$), then the feasible polyhedron of the barycenter linear program (2) has vertices that are not uniform. Consequently, it is not clear whether the minimiser is also uniform on M points (see also Lin et al. [51]). However, we have empirically observed that for the squared Euclidean cost, there is a barycenter that is uniform on M points. This has been tested on an overwhelming amount of simulations in different scenarios and, with no exception, the barycenter was always uniform. We do not have a formal proof of this uniformity conjecture, which poses an interesting question for further research. However, in view of the empirical evidence we shall assume in the following that it is practically valid, and explore its computational consequences.

Our observations suggest two immediate improvements to the alternating procedure. Firstly, we no longer need to perform weight updates: they can be chosen uniform. Secondly, we can reduce the number of support points of the candidate barycenter from $MN - N + 1$ to M. This would have probably been done in practice anyway to reduce computational strain, as, for instance, using a fixed support approximation of the barycenter essentially implies choosing the support size of the barycenter to be M and not performing positions updates. Our per-iteration computational cost is therefore comparable to that of a fixed support barycenter. However, based on this conjecture this support size is actually optimal. Finally, we replace the one-step Newton-type update of [21] by a proper subgradient descent on the positions of the candidate measure. The procedure can

be seen in Algorithm 2. Note that this problem is non-convex, so, in order to avoid local minima, we give the algorithm a "warmstart" by performing a fixed number of stochastic subgradient descent steps to obtain an initial point for the genuine (i.e., non-stochastic) subgradient descent. The left panel of Figure 6 shows the advantages of this small modification, which comes at little computational cost, but provides a noticeable improvement. Here, performance is measured in terms of the (empirical) relative error in the Fréchet functional, namely

$$\frac{F(\widehat{\mu}_S^*) - F(\mu^*)}{F(\mu^*)}.$$

Without the warmstart our runs provided a median relative error of 5.2%, which reduced to 0.8% with a warmstart. Over half of the runs with warmstart yielded a relative error under 1%, which is a good indication that we can obtain good results using very few restarts. However, even without the warmstart, we typically obtain relative errors in the single-digit percent points, which is still tolerable in many applications.

Whilst the measures $\mu_1, \ldots, \mu_N$ are typically not uniform, the uniform approximation can be applied to their empirical versions $\mu_1^S, \ldots, \mu_N^S$; the algorithm is still applicable if some points are duplicate by taking some of the rows of some Y^i to be the same. Thus, the SUA-method reduces the problem size by replacing M_i with S, whilst at the same time enforcing a setting in which a much faster method can be employed. This two-fold gain pushes down the computation time by orders of magnitude, and provides the basis for our simulation results in the following section.

4.2 Simulations

In this subsection we present empirical results on the decay of the approximation error as the resample size S grows, for some classes of measures on $\mathbb{R}^2$. As we discuss the impact of the number of repeats R separately in the following subsection, we set $R = 1$ whilst analysing the behaviour with respect to S.

The right panel of Figure 6 shows the simulated, expected error as a function of S, for the datasets showcased in Figure 9. Noticeably, performance greatly varies between different types of datasets. For simple geometric structures such as ellipses, crescent shapes, or Gaussian samples, we obtain the best results; for S of the order of 10% of the original data size M we obtain expected approximation errors of under 5%. For more complicated structures, such as the nested ellipses and the Cauchy density on a grid, we still obtain decent empirical errors. For $S \approx 0.1M$ we observe an error of under 20% in the Fréchet functional, whilst the runtime is improved by a factor of about 300. On the unstructured data, such as

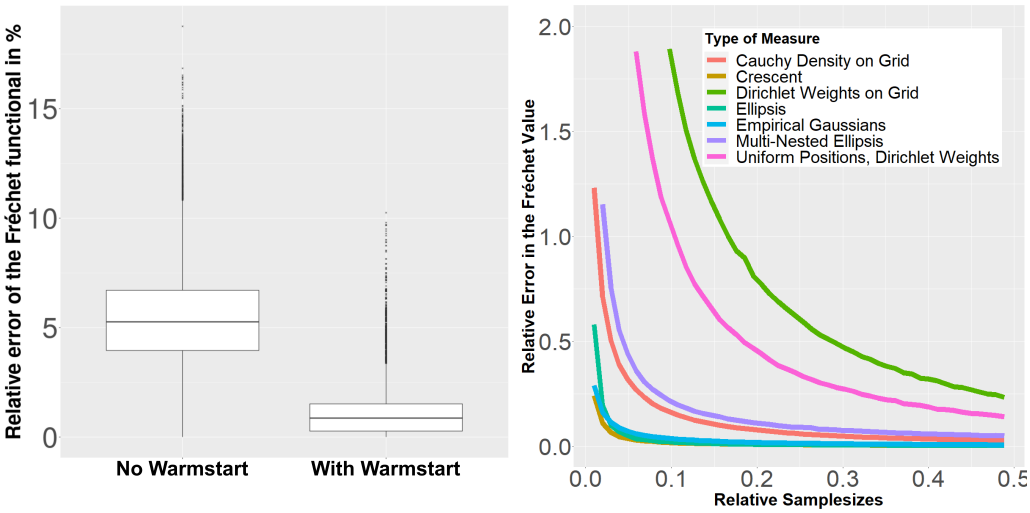


Figure 6: **Left:** Relative Fréchet error from 50000 runs, on $N = 4$ measures with $M = 20$ uniformly random positions with weights $1/M$, with and without warmstart, respectively. **Right:** Simulated expected error of the Fréchet functional of $N = 20$ measures with $M = 1024$ points in $\mathbb{R}^2$ based on 100 repeated runs for each value of S. Example measures from the corresponding dataset and example barycenters for some values of S can be seen in Figure 8 and Figure 9, respectively.

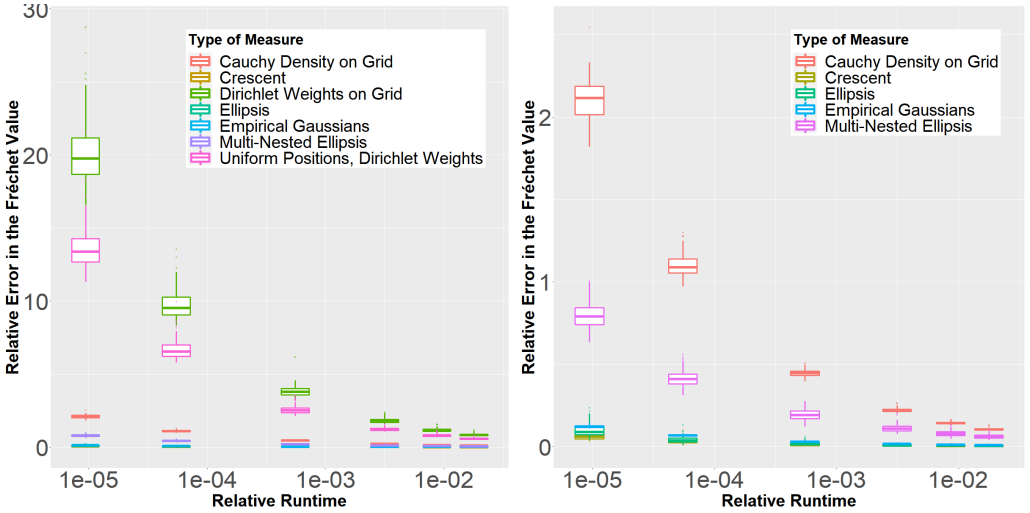


Figure 7: Simulated expected error of the Fréchet functional of $N = 10$ measures with $M = 4096$ points in $\mathbb{R}^2$ based on 100 repeated runs for each value of S compared to their relative runtime compared to the original dataset. On the right panel the two outlier distributions have been removed.

Algorithm 2 Stochastic Uniform Approximation (SUA)–algorithm for the Wasserstein barycenter

Data Measures: $\mu_1, \dots, \mu_N$, sample size S, repeats R, initial position matrix $X^0 \in \mathbb{R}^{S \times d}$, sequence of stepsizes $(\alpha_n)_{n \in \mathbb{N}}$, $\alpha_n > 0 \ \forall \ n \in \mathbb{N}$.
$\Pi = \{T \in \mathbb{R}^{S \times S} \quad | \quad T\mathbf{1}_S = \mathbf{1}_S^T S^{-1} = T^T \mathbf{1}_S^T\}$
for $r = 1, \dots, R$ **do**
    for $i = 1, \dots, N$ **do**
        Draw $X_1^{(i)}, \dots, X_S^{(i)} \sim \mu_i$
        $\mu_i^S = \frac{1}{S} \sum_{k=1}^{S} \delta_{X_k^{(i)}}$
    end for
    $Y^i = \text{supp}(\mu_i^S)$
    while not converged **do**
        for $i = 1 \dots, N$ **do**
            $T_i = \underset{T \in \Pi}{\arg\min} \sum_{k=1}^{S} \sum_{j=1}^{S} T_{kj} \|X_k - Y_j^i\|_2^2$
            $V_i = 2(X - T_i Y^{(i)})$
        end for
        $X^{n+1} = X^n - \frac{\alpha_n}{N} \sum_{i=1}^{N} V_i$
        $n = n + 1$
    end while
    $\bar{\mu}_r = \frac{1}{S} \sum_{k=1}^{S} \delta_{X_k^n}$
end for
Set $\widehat{\mu}_S^* = \frac{1}{R} \sum_{r=1}^{R} \bar{\mu}_r$ **return** Approximation of the empirical Wasserstein barycenter $\widehat{\mu}_S^*$

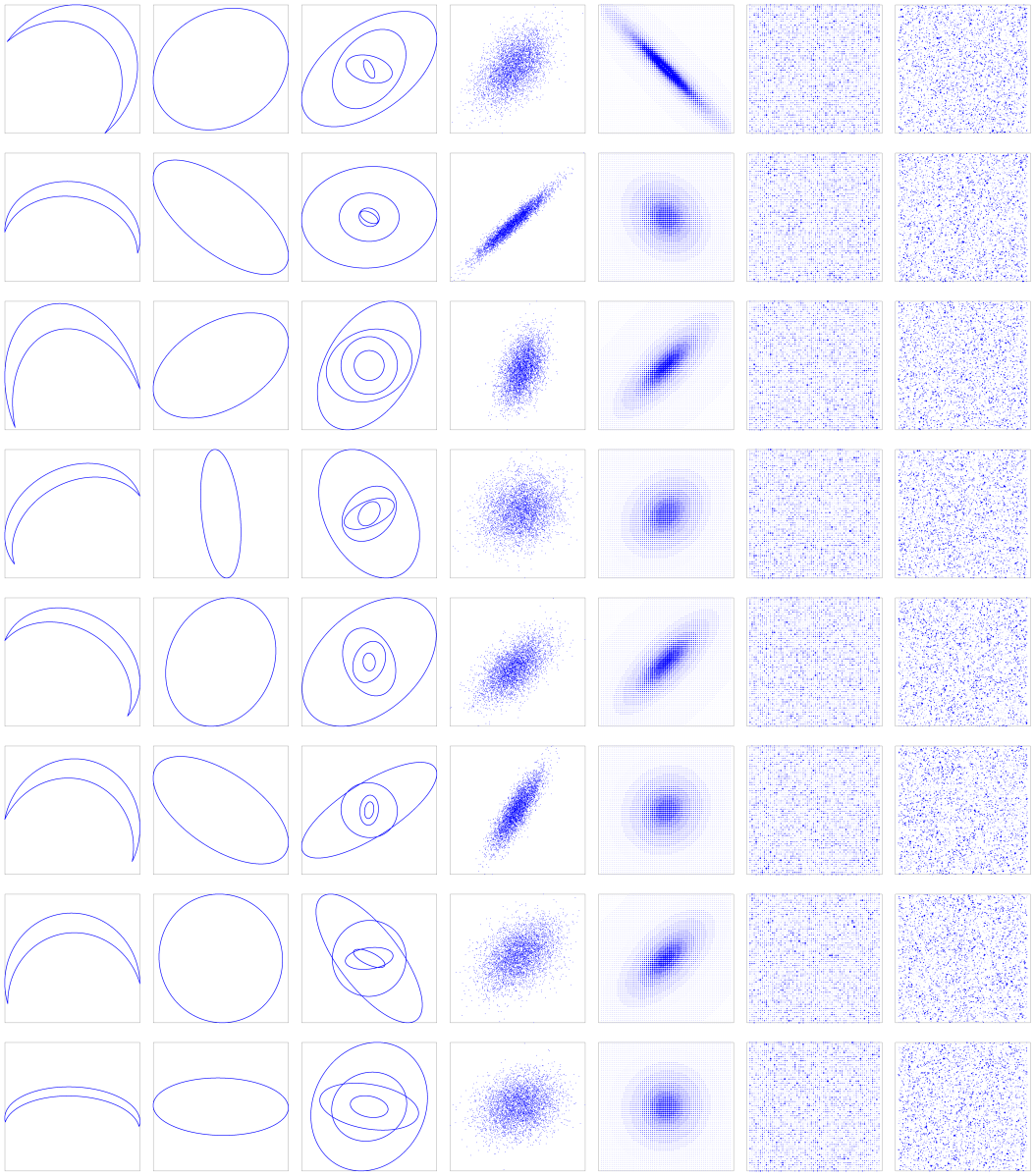


Figure 8: 1.Column: Discretised crescents with uniform weights, 2.Column: Discretised ellipses with uniform weights, 3.Column: Discretised nested ellipses with uniform weights, 4.Column: Gaussian positions with uniform weights, 5.Column: Discretised Cauchy density on a grid, 6.Column: Dirichlet weights on a grid, 7.Column: Dirichlet weights on uniform positions. In each column the measures above the line are excerpts from the used dataset of $N = 100$ measures with $M = 4096 = 64^2$ points.

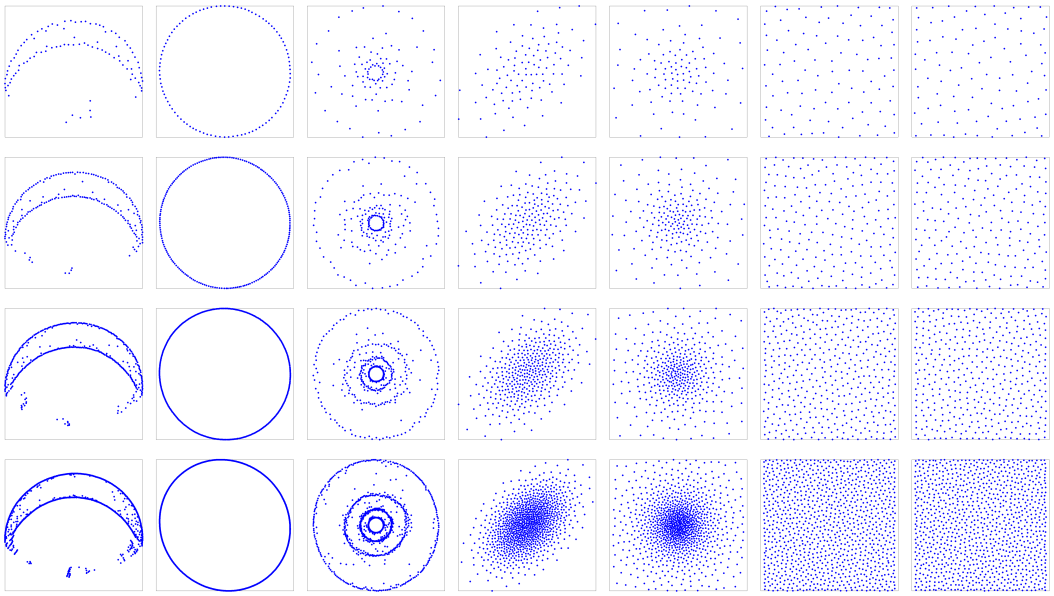


Figure 9: The randomised approximations of the barycenters of the data in Figure 8 obtained from the SUA-algorithm. The first row shows empirical barycenters based on $S = 100$ samples, the second for $S = 200$, third for $S = 400$ and the fourth for $S = 1000$. The computation of one barycenter in the last row took about 40 minutes on a single core of an i7-7700k.

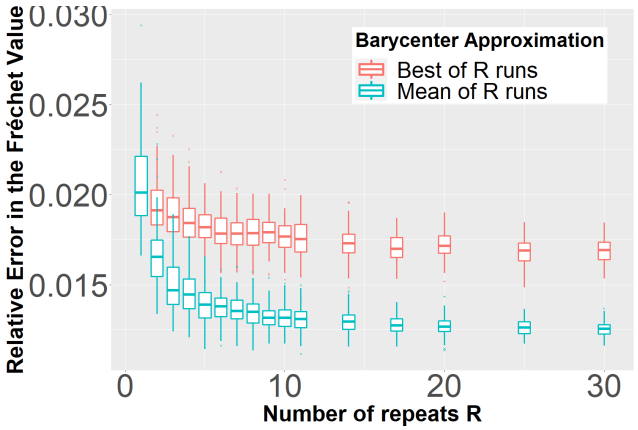


Figure 10: The dependence of the empirical barycenter approximation on the number of repeats R. Here, the data measures are $N = 20$ multi-nested ellipses supported on $M = 100$ support points each. The sample size S was chosen to be 33. For each R, 100 runs have been performed.

Dirichlet weights on a grid/uniform support points, we observe the worst results. For 10% support size we still observe errors upwards of 100%. As barycenters are mostly used in settings where the underlying measures have some form of geometric structure, these results indicate good potential performance of our stochastic barycenter approximation. We point out that the variance of the observed Fréchet values was generally quite small, as can be seen in Figure 7, and decreases with increasing sample size. This suggests that one might be best advised to increase the number of samples S instead of increasing the number of repeats R in Algorithm 1. In fact, this coincides with the observation by Sommerfeld et al. [66] for the empirical Wasserstein distance itself as well as our own findings in the next subsection. It is also noteworthy that the curves at the right panel of Figure 6 suggest that doubling the sample size approximatively halves the approximation error, indicating linear decay in S instead of the rate $S^{-\frac{1}{2}}$. This suggests that, whilst asymptotically we cannot perform better than this rate, we can achieve even better results for small sample sizes.
Since the convergence rate in Theorem 3.1 is independent of the dimension of the ambient space of the data, it is natural to extend our simulations from the plane to higher dimensions. In Figure 3 we provide some examples of datasets in $\mathbb{R}^3$. The approximation indeed performs well, allowing to compute rather good approximate barycenters when S is less than 5% of the original size M. Whilst we were no longer able to approximate the barycenter of these measures directly due to memory constraints, extrapolating from smaller datasets suggests that the corresponding computation would take well over a week to complete. On the contrary, our stochastic approximation took around 90 minutes on a single core of an i7-7700k and used about 800MB of RAM with our implementation.

4.2.1 The effect of the number of repeats R

Apart from the choice of the sample size S, a discussion on the number of repeats R is in order. Two questions arise in this regard: firstly, how to choose R; and secondly, how to combine the R empirical barycenter into a single estimator, when $R > 1$.
We address the second matter first. A natural choice is to take the empirical barycenter that achieves the smallest value of the Fréchet functional across the R options. This approach, however, is computationally demanding and might not even be feasible at all for large problems, since it requires solving optimal transport problems between the barycenter and the data measures. For this reason, we advocate taking the linear average of the R empirical barycenters, a simple procedure that does not involve optimisation. In view of (4), it is guaranteed to have objective value that is better than the average value of the R empirical barycenters. Moreover, our simulations suggest that in fact the linear mean performs better

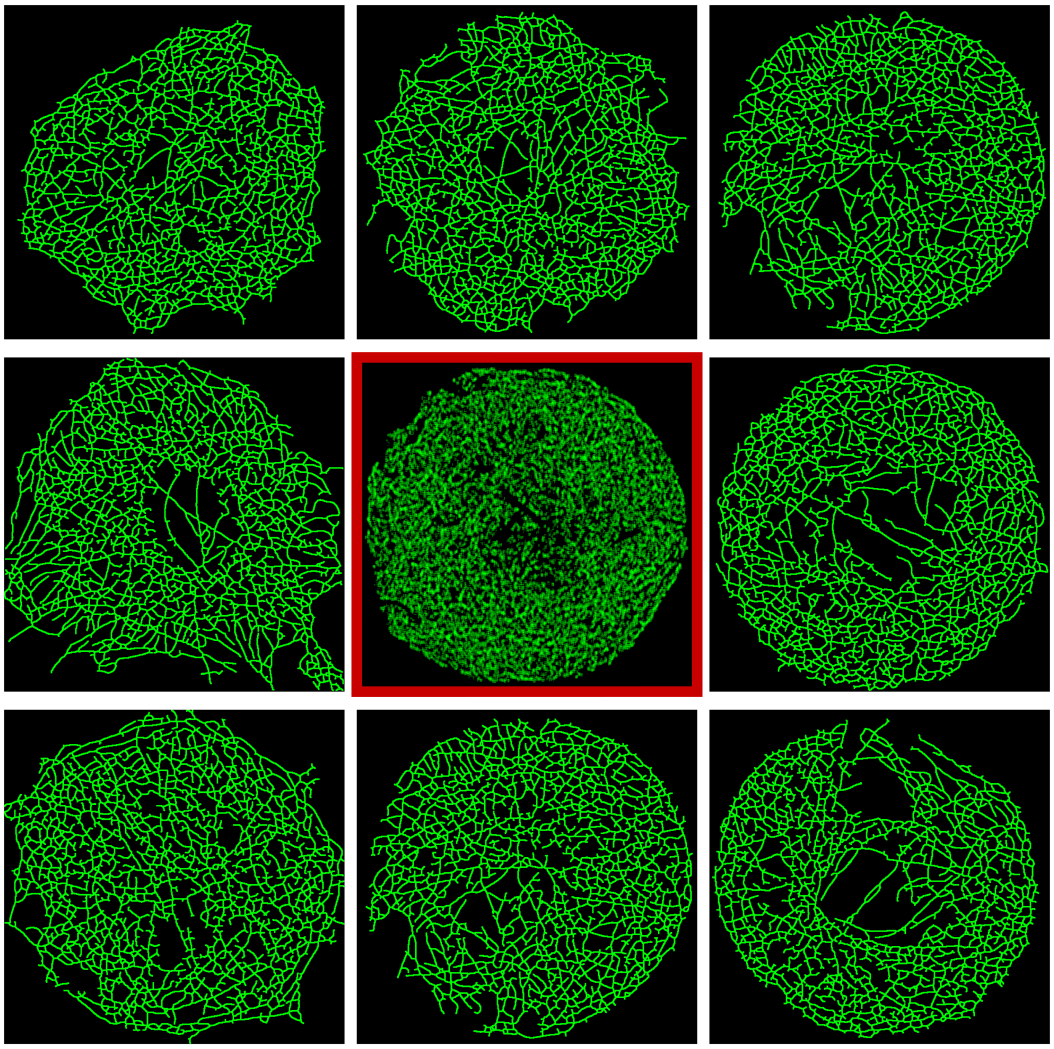


Figure 11: The eight outer confocal microscopy images show fluorescently labelled microtubules in mouse fibroblast cells, which had their graph structure extracted (by use of CytoSeg2.0 [16]). The images have a resolution of 642×642. The center image (framed in red) shows their barycenter approximation obtained from the SUA-algorithm with $S = 20,000$ samples, which has been mapped back onto a 642×642 grid to produce an image.

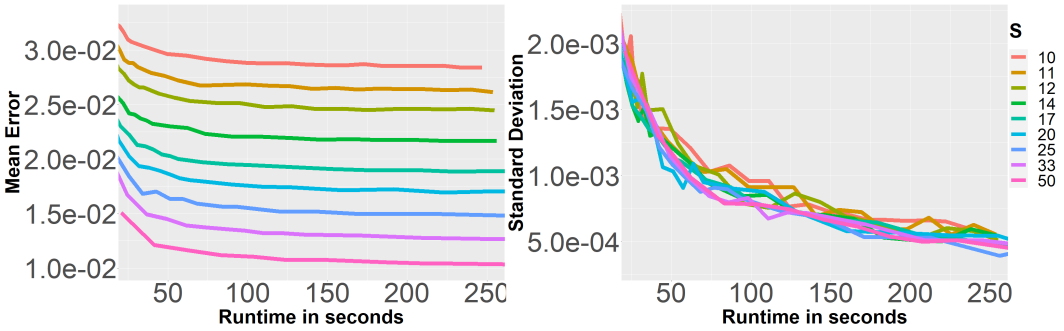


Figure 12: The dependence of the empirical barycenter approximation on the number of repeats R. Here, the data measures are $N = 20$ multi-nested ellipses supported on $M = 100$ support points each. For each pair of R and S, 100 runs have been performed.

than the best of the R runs for essentially any R and S. An example of this can be seen in Figure 11. Intuitively, this superior performance can be attributed to the fact that usually $S << M$, thus the linear mean has the advantage of having a larger support size, which is closer to the support size of the original barycenter. Having discussed the choice of the estimator, we can now proceed to analyse the properties of this estimator with respect to R empirically. To this end, we computed 100 barycenters for a range of combinations of R and S, and estimated the mean and the standard deviation of the estimation error. Naturally, increasing R improves performance, in that both the mean and the variance of the approximation error decrease. But this observation does not account for the extra computational effort resulting in increasing R. For a fixed runtime, increasing R necessarily amounts to decreasing S. Figure 12 shows the empirically observed mean and standard deviation of the error as a function of the runtime, for different choices of S. Noticeably, for a fixed runtime the standard deviation of the error is nearly identical for all considered S, whilst the mean error clearly decreases. It should be stressed that the fact that the lines at the left plot of Figure 12 never intersect implies that, in this example, choosing $R = 1$ is best for any fixed runtime considered in the simulation. In other words, any extra available computational effort would be best spent by increasing S, and not R. We also point out that if we consider both plots in Figure 12, we see that, due to the bias-variance formula, it is also optimal in terms of the mean squared error to increases S instead of R. This is, of course, only one example, but we believe that it provides a rather compelling argument for choosing R small. For this reason, we use $R = 1$ in the following real data example.

4.3 A Cell Microscopy Example

An important line of research in biophysics concerns the understanding of the structure of filament networks and cytoskeletons of proteins and their behaviour over time when placed under certain conditions (for an overview see e.g., Huber et al. [40]). One can envisage using barycenters here in a number of ways. One option is to consider the barycenter of a single filament at different time points, thus "averaging out" variability across time. Another possibility is to compute the barycenter of a number of different filaments at the same time point, thus averaging over the variability within certain types of proteins. In that setting one could generate genetically identical cells from each class of protein. Although cells at each class are genetic twins, some variability may still arise from mutations. This variability can be averaged out by taking a barycenter from each protein class, and by doing this at each time point, one can study the evolution of a typical filament (represented by the barycenters) across time.
A modern, high-resolution image usually has hundreds of thousands of pixels, and high quality images can easily extend this well into the millions. In our example, we consider $N = 8$ images with resolution of $M = 642 \times 642 > 400,000$, which can be seen in Figure 10. Computing the exact barycenter from these 8 images is intractable. One could, of course, decrease the resolution in order to obtain a problem of tractable size. Doing so, however, would result in a substantial information loss, comparable to a decrease in the microscope's resolution, as it would blur the fine structure of the filaments. Instead, we compute an approximate barycenter using our method with $S = 20,000$ (Figure 10, with red frame in the center). Upon visual inspection this approximate barycenter seems to represent the 8 filaments well.

5 Conclusion and Outlook

Whilst the Wasserstein barycenter shows great potential in geometric data analysis, its tremendous computational cost is a severe hurdle for the OT-barycenter to become a widespread tool in analysing complex data. Even using modern solvers (potentially with regularisation), one is quickly faced with massive memory and runtime demands which render efficient computations on personal computers impossible for realistic problem sizes. This is the case for e.g. biological imaging, where there is great interest in obtaining sharp high-resolution images of the studied objects, as, for instance, filaments or other protein structures. Since the number of discretisation points required to achieve a fixed quality grows exponentially with the underlying dimension, the need to alleviate the computation burden is even more paramount in three or higher dimensions in comparison to $\mathbb{R}^2$. Our

SUA-algorithm might be able to bridge the gap between modern applications and the capabilities of state-of-the-art solvers, and may thus open the door to new and interesting data analysis in the natural sciences, where three-dimensional datasets are not uncommon. This is backed by the fact that our convergence rate is independent of the dimension combined with the strong visual performance of our resampled OT-barycenters in three dimensions with relatively low runtime.
Another attractive feature of our method is the possibility to adjust runtime and memory demand depending on the required accuracy of the result, which on the one hand allows to perform quick, rough analysis on personal computers, whilst on the other hand still enables more precise computations on specialised high-performance computing clusters.
Finally, our simulations suggest even better performance than guaranteed by our theoretical bounds in many settings. Whilst we have shown that our rates cannot be improved in general, it is still possible, and in fact suggested by simulations, that for certain classes of measures we can obtain better bounds on the expected error if S is a sufficiently small fraction of M (which is the relevant regime in practice). Providing such improved bounds would render our approach even more appealing, and is another clear avenue for further research.

Acknowledgments

We wish to thank an associate editor and two reviewers for their constructive feedback. We are grateful to Sarah Köster and Julia Börke for providing us with the data used in Section 4.3. F. Heinemann gratefully acknowledges support from the DFG Research Training Group 2088 *Discovering structure in complex data: Statistics meets Optimization and Inverse Problems*. A. Munk gratefully acknowledges support from the DFG CRC 1456 *Mathematics of the Experiment* and the Cluster of Excellence 2067 MBExC *Multiscale bioimaging–from molecular machines to networks of excitable cells*. Y. Zemel was supported in part by Swiss National Science Foundation Grant 178220, and in part by a U.K. Engineering and Physical Sciences Research Council programme grant.

A Auxiliary Proofs

Proof of Theorem 2.1. The proof is very similar to that of Sommerfeld et al. [66]; we only give the difference here. The key idea is to improve the bound on the height function h of [66]. Using the same notation as [66], note that for $1 \leq l \leq l_{\max}$, we

have

$$\frac{h^p(par(x)) - h^p(x)}{\mathrm{diam}(\mathcal{X})^p} = q^{-lp}\left(\left(q + \sum_{j=0}^{l_{\max}-l} q^{-j}\right)^p - \left(\sum_{j=0}^{l_{\max}-l} q^{-j}\right)^p\right) \leq q^{p-lp}\left(\frac{q}{q-1}\right)^p.$$

If x is a leaf, then $h(x) = 0$ and the difference is $q^{-pl_{\max}}$ and if x is a root, then $par(x) = x$ and the difference vanishes. Thus, we can start the sum in the definition of $\mathcal{E}$ at 1 instead of 0. Also, for $p = 1$, we have $q^l(h(par(x)) - h(x)) = q \cdot \mathrm{diam}(\mathcal{X})$, where the fractional factor has vanished. Plugging this into the upper bound for the tree metric yields

$$(W_p^{\mathcal{T}}(r,s))^p \leq 2^{p-1}\mathrm{diam}(\mathcal{X})^p \left[\sum_{l=1}^{l_{\max}} \left(\frac{q}{q-1}\right)^p q^p q^{-lp} \sum_{x \in \tilde{Q}_l} |(S_T r)_x - (S_T s)_x| \right.$$
$$\left. + q^{-l_{\max} p} \sum_{x \in \tilde{Q}_{l_{\max}+1}} |(S_T r)_x - (S_T s)_x| \right].$$

Finally, taking expectations we obtain

$$\mathbb{E}\left[W_p^p(\mu, \mu^S)\right] \leq S^{-\frac{1}{2}} 2^{p-1} \mathrm{diam}(\mathcal{X})^p q^p \left[\sum_{l=1}^{l_{\max}} \left(\frac{q}{q-1}\right)^p q^{-lp} \sqrt{\mathcal{N}(\mathcal{X}, q^{-l}\mathrm{diam}(\mathcal{X}))} \right.$$
$$\left. + q^{-(l_{\max}+1)p} M^{\frac{1}{2}} \right].$$

For the specific case $p = 1$, we have a better bound:

$$\mathbb{E}\left[W_1(\mu, \mu^S)\right] \leq S^{-\frac{1}{2}} \mathrm{diam}(\mathcal{X}) q \left[\sum_{l=1}^{l_{\max}} q^{-l} \sqrt{\mathcal{N}(\mathcal{X}, q^{-l}\mathrm{diam}(\mathcal{X}))} + q^{-(l_{\max}+1)} M^{\frac{1}{2}}\right].$$

Imitating the proof of [66, Theorem 3] with this improved bound yields the result for $\mathcal{X} \subset (\mathbb{R}^D, \|\cdot\|_2)$. The details are a straightforward adaptation of the work in [66, Theorem 3], and are therefore omitted. □

Proof of Lemma 3.7. Convexity follows from convexity of P, and we also have $|g(t) - g(s)| \leq |t - s|\|y\|_1$ by the triangle inequality. To prove piecewise affineness it suffices to show that there exists $t_0 > 0$ such that g is affine on $[0, t_0]$. One can then replace x by $x + t_0 y$ and repeat the argument. Replacing y by $-y$ yields the result for negative values of t.
The minimum in $d_1(x + ty, P)$ is attained by a compactness argument. Let $z_t \in P$ be such that $g(t) = \|x + ty - z_t\|_1$.

Claim: It is possible to choose z_t and z_0 in such a way that $z_t \to z_0$ as $t \to 0$.
We will show the stronger statement that as $t \searrow 0$

$$\sup_{z_t \in \arg\min d_1(x+ty,P)} \inf_{z_0 \in \arg\min d_1(x,P)} \|z_t - z_0\|_1 \to 0.$$

By continuity for $t > 0$ there exist $z_t \in \arg\min\ d_1(x+ty, P)$ and $x_t \in \arg\min\ d_1(x, P)$ which attain the supremum and the infimum. Assume that the converse of the claim is true. Then for some sequence $t_k \to 0$, we can choose $z_{t_k} \in \arg\min d_1(x + t_k y, P)$ such that for all k and all $z \in \arg\min d_1(x, P)$ we have $\|z_{t_k} - z\|_1 > \epsilon > 0$. By compactness of P there is a further subsequence k_l such that $z_{t_{k_l}} \to z$. Since

$$\|z - x\|_1 = \lim_{l\to\infty} \|z_{t_{k_l}} - x - t_{k_l} y\|_1 = \lim_{l\to\infty} d_1(x + t_{k_l} y, P) = d_1(x, P)$$

we have $z \in \arg\min d_1(x, P)$ and therefore

$$0 < \epsilon \leq \liminf_{l\to\infty} \|z_{t_{k_l}} - z\|_1 = 0,$$

a contradiction. This proves the claim.
Since P is a polyhedron, for all $z_0 \in P$ there exists $r_0(z_0) > 0$ such that for all $v \in \mathbb{R}^L$ with $\|v\|_1 \leq r_0$,

$$z_0 + v \in P \quad \implies z_0 + 2v \in P.$$

By the previous claim, there exist $z_0 \in \arg\min d_1(x, P)$, $z_t \in \arg\min d_1(x+ty, P)$, and $T_0 > 0$ such that for all $t \in [0, T_0]$,

$$\|x+ty-z_t\|_1 = d_1(x+tv, P), \quad \|x-z_0\|_1 = d_1(x, P), \qquad \|z_t-z_0\|_1 \leq r_0 = r_0(z_0).$$

Therefore $z_0 + 2(z_t - z_0) \in P$ for all $t \in [0, T_0]$ and

$$d_1(x, P)+d_1(x+2ty, P) \leq \|x-z_0\|_1+\|x+2ty-z_0-2(z_t-z_0)\|_1 = \|x^0\|_1+\|x^0+2u^t\|_1,$$

where $x^0 = x - z_0$ and $u^t = ty - (z_t - z_0) \to 0$. Define

$$I_+ = \{1 \leq i \leq L : x_i^0 > 0\}, \quad I_- = \{1 \leq i \leq L : x_i^0 < 0\}, \quad I_0 = \{1 \leq i \leq L : x_i^0 = 0\}.$$

For t sufficiently small (and such that $t \leq T_0$) we have $2\|u^t\|_1 \leq \min_{i\notin I_0} |x_i^0|$ (where we define the minimum to be infinite if $x = z_0$) and therefore

$$\begin{aligned} d_1(x, P) + d_1(x + 2ty, P) &\leq \sum_{i\in I_+} (x_i^0 + x_i^0 + 2u_i^t) + \sum_{i\in I_-} (-x_i^0 - x_i^0 - 2u_i^t) + \sum_{i\in I_0} 2|u_i^t| \\ &= 2\sum_{i\in I_+} (x_i^0 + u_i^t) + 2\sum_{i\in I_-} (-x_i^0 - u_i^t) + 2\sum_{i\in I_0} |u_i^t| \\ &= 2d_1(x + ty, P). \end{aligned}$$

Thus, there exists $0 < t_0 \leq T_0$ such that for all $t \in [0, t_0]$,

$$g(0) + g(2t) \leq 2g(t).$$

Since g is convex and finite on $\mathbb{R}$, this implies that g is affine on $[0, t_0]$. As stated at the beginning of the proof, this shows that g is piecewise affine on the real line. □

Proof of Lemma 3.8. If μ is optimal, then we can choose $\mu^* = \mu$ and there is nothing to prove. Hence we may assume that μ is not optimal. Let $P \subset \mathbb{R}^L$ be the feasible polytope corresponding to (2). To each $\mu \in \mathcal{P}(\mathcal{C})$ corresponds a $\pi \in P$ such that $c^T\pi = F(\mu)$. Fix an element $\pi^* \in \arg\min d_1(\pi, \mathcal{M})$, from which we can construct a minimiser μ^* of F. It holds that

$$F^p(\mu) - F^p(\mu^*) = c^T\pi - c^T\pi^* = \|\pi - \pi^*\|_1 \frac{c^T\pi - f^*}{\|\pi - \pi^*\|_1} \geq \|\pi - \pi^*\|_1 \psi(\pi),$$

where

$$\psi(\pi) = \frac{c^T\pi - f^*}{d_1(\pi, \mathcal{M})}, \qquad \pi \in P \setminus \mathcal{M}.$$

Next, we aim to show that the infimum of ψ is attained at a vertex of P. If π is not a vertex of P, then there exists a vector $v \in \mathbb{R}^L \setminus \{0\}$ such that $\pi + tv \in P$ for all $t \in [-1, 1]$. Let

$$\phi(t) := \psi(\pi + tv) = \frac{c^T\pi - f^* + tc^Tv}{d_1(\pi + tv, \mathcal{M})} := \frac{at + b}{g(t)}, \qquad b = c^T\pi - f^*,$$

where the nominator is affine by definition and the denominator, $g(t)$, is piecewise affine by Lemma 3.7. Since g is continuous and convex, we have $g(t) = \alpha_+ t + \beta$ for small $t \geq 0$ and $g(t) = \alpha_- t + \beta$ for small $t \leq 0$, with $\alpha_- \leq \alpha_+$ and $\beta > 0$. Taking the derivative we obtain

$$\phi'_\pm(t) = \frac{a(\alpha_\pm t + \beta) - \alpha_\pm(at + b)}{(\alpha_\pm t + \beta)^2} = \frac{a\beta - b\alpha_\pm}{(\alpha_\pm t + \beta)^2},$$

where ϕ'_+ and ϕ'_- denote the right and left derivatives. We now distinguish three cases. If $\phi'_+(0) < 0$, then $\psi(\pi + tv) < \psi(\pi)$ for $t > 0$ small enough. If $\phi'_-(0) > 0$, then $\psi(\pi + tv) > \psi(\pi)$ for $t < 0$ small enough. In both cases π is not a minimiser. Since $\phi'_+(0) \leq \phi'_-(0)$, the only other possibility is that both derivatives vanish. Replacing v by $-v$ if necessary, we may assume that $a \geq 0$. We also have $b > 0$ and $\beta > 0$, since $\pi \in P \setminus \mathcal{M}$ by assumptions. Thus it holds $\alpha_\pm = \alpha = a\beta/b \geq 0$. We can now move in direction v (i.e., increasing t) until one of two things happens. Either we can no longer move in direction v without leaving P, or the denominator

of ϕ has changed since we moved into a different affine segment of g (t cannot go to infinity since $P \subseteq [0,1]^L$ is bounded by definition, and we cannot reach $\mathcal{M}$ because $a \geq 0$.). In the first case, we have reached a face of P (if π is in the interior of P) or a strict subface of one of the faces of P containing π (if π is on the boundary of P) and can now search a new direction v in this (sub)face. Since P is a polyhedron, this can only happen finitely many times until we reach a vertex of P. In the second case we reach a point $t_0 > 0$ at which the right derivative of g is strictly larger than α (by convexity of g). Thus $g(t_0) \geq g(0) > 0$, the right derivative of ϕ becomes negative, and for ϵ small

$$\psi(\pi + (\epsilon + t_0)v) = \phi(\epsilon + t_0) < \phi(t_0) = \phi(0) = \psi(\pi),$$

and consequently π is not a minimiser.
All in all, for any $\pi \in P \setminus \mathcal{M}$, either π does not minimise ψ, or there exists a vertex v of P that is not in $\mathcal{M}$ and such that $\psi(v) \leq \psi(\pi)$. Hence

$$\inf_{\pi \in P \setminus \mathcal{M}} \psi(\pi) = \min_{\pi \in V \setminus V^*} \psi(\pi) = \min_{v \in V \setminus V^*} \frac{c^T v - f^*}{d_1(v, M_f)} =: \widetilde{C}_P,$$

and therefore

$$F^p(\mu) - F^p(\mu^*) = c^T(\pi - \pi^*) \geq \psi(\pi)\|\pi - \pi^*\|_1 \geq \widetilde{C}_P \|\pi - \pi^*\|_1.$$

To relate $\|\pi - \pi^*\|_1$ to the Wasserstein distance recall that π encodes a measure μ as well as the optimal transport plan T^i between μ and each measure μ_i. Hence

$$\begin{aligned}\|\pi - \pi^*\|_1 &\geq \sum_{i=1}^{N} \|T^{(i,1)} - T^{(i,2)}\|_1 + \sum_{x \in \mathcal{C}} |\mu(x) - \mu^*(x)| \\ &\geq N \sum_{x \in \mathcal{C}} |\mu(x) - \mu^*(x)| + \sum_{x \in \mathcal{C}} |\mu(x) - \mu^*(x)| = (N+1) \sum_{x \in \mathcal{C}} |\mu(x) - \mu^*(x)|,\end{aligned}$$

where $T^{(i,1)}$ is an optimal transport plan between μ_i and μ and $T^{(i,2)}$ is an optimal plan between μ_i and μ^*. Thus,

$$\begin{aligned}F^p(\mu) - F^p(\mu^*) &\geq \widetilde{C}_P (N+1) \sum_{x \in \mathcal{C}} |\mu(x) - \mu^*(x)| \\ &= 2(N+1)\widetilde{C}_P \ \mathrm{TV}(\mu, \mu^*) \geq 2 C_P W_p^p(\mu, \mu^*),\end{aligned}$$

with $C_P = (N+1)\widetilde{C}_P \mathrm{diam}(\mathcal{X})^{-p}$. Finally, $\widetilde{C}_P > 0$ because it is a minimum of finitely many positive numbers. Positivity of C_P follows. □

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal of Mathematical Analysis*, 43:904–924, 2011.

[2] Martial Agueh and Guillaume Carlier. Vers un théorème de la limite centrale dans l'espace de Wasserstein? *Comptes Rendus Mathématique*, 355(7):812–818, 2017.

[3] Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probability Theory and Related Fields*, 177(1):323–368, 2020.

[4] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 1964–1974, Red Hook, NY, 2017. Curran.

[5] Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[6] Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84:389–409, 2015.

[7] Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. Bayesian learning with Wasserstein barycenters. *arXiv:1805.10833*, 2018.

[8] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[9] Daniel Berend and Aryeh Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics and Probability Letters*, 83(4):1254–1259, 2013.

[10] Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

[11] Louis J Billera, Susan P Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

[12] Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'IHP Probabilités et statistiques*, 50(2):539–563, 2014.

[13] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71–1, 2016.

[14] Steffen Borgwardt and Stephan Patterson. On the computational complexity of finding a sparse Wasserstein barycenter. *arXiv:1910.07568*, 2019.

[15] Steffen Borgwardt and Stephan Patterson. Improved linear programs for discrete barycenters. *Informs Journal on Optimization*, 2(1):14–33, 2020.

[16] David Breuer, Jacqueline Nowak, Alexander Ivakov, Marc Somssich, Staffan Persson, and Zoran Nikoloski. System-wide organization of actin cytoskeleton determines organelle transport in hypocotyl plant cells. *Proceedings of the National Academy of Sciences*, 114(28):E5741–E5749, 2017.

[17] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[18] Julien Chevallier. Uniform decomposition of probability measures: quantization, clustering and rate of convergence. *Journal of Applied Probability*, 55(4):1037–1045, 2018.

[19] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures–Wasserstein barycenters. *arXiv:2001.01700*, 2020.

[20] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 2292–2300. Curran, Red Hook, NY, 2013.

[21] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, pages 685–693. PMLR, Beijing, 2014.

[22] Eustasio del Barrio, Evarist Giné, and Carlos Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, pages 1009–1071, 1999.

[23] Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: Approximation by empirical measures. *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques*, 49(4):1183–1203, 2013.

[24] Jack J Dongarra, Piotr Luszczek, and Antoine Petitet. The linpack benchmark: past, present and future. *Concurrency and Computation: practice and experience*, 15(9):803–820, 2003.

[25] Ian L Dryden, Alexey Koloydenko, and Diwei Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123, 2009.

[26] Ian L Dryden and James S Marron. *Object Oriented Data Analysis*. forthcoming.

[27] Richard Mansfield Dudley. The speed of mean Glivenko–Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[28] Darina Dvinskikh. SA vs SAA for population Wasserstein barycenter calculation. *arXiv:2001.07697*, 2020.

[29] Pavel Dvurechenskii, Darina Dvinskikh, Alexander Gasnikov, Cesar Uribe, and Angelia Nedich. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 31:10760–10770, 2018.

[30] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden*, pages 1366–1375, 2018.

[31] Steven N Evans and Frederick A Matsen. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

[32] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[33] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, 10(4):215–310, 1948.

[34] Dongdong Ge, Haoyue Wang, Zikai Xiong, and Yinyu Ye. Interior-point methods strike back: solving the Wasserstein barycenter problem. In *Advances in Neural Information Processing Systems*, pages 6894–6905, 2019.

[35] Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13(43):1263–1291, 2012.

[36] Samuel Gerber and Mauro Maggioni. Multiscale strategies for computing optimal transport. *Journal of Machine Learning Research*, 18:1–32, 2017.

[37] Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *arXiv:1908.00828*, 2019.

[38] Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*. Springer, Berlin, 2007.

[39] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

[40] Florian Huber, Adeline Boire, Magdalena Preciado López, and Gijsje H Koenderink. Cytoskeletal crosstalk: when three different personalities team up. *Current Opinion in Cell Biology*, 32:39–47, 2015.

[41] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, pages 1–58, 2010.

[42] Stephan F Huckemann and Benjamin Eltzner. Data analysis on nonstandard spaces. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(3):e1526, 2021.

[43] Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[44] Marcel Klatt, Carla Tameling, and Axel Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

[45] Benoît R Kloeckner. A geometric study of Wasserstein spaces: ultrametrics. *Mathematika*, 61(1):162–178, 2015.

[46] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Statistical inference for Bures–Wasserstein barycenters. *Annals of Applied Probability*, in press.

[47] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

[48] Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in $O(\text{vrank})$ iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 424–433. IEEE, 2014.

[49] Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

[50] Lingxiao Li, Aude Genevay, Mikhail Yurochkin, and Justin M Solomon. Continuous regularized Wsserstein barycenters. *Advances in Neural Information Processing Systems*, 33, 2020.

[51] Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, and Michael I Jordan. Computational hardness and fast algorithm for fixed-support Wasserstein barycenter. *NEURIPS, to appear*, 2020.

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

[53] Adam M Oberman and Yuanlong Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

[54] Victor M Panaretos and Yoav Zemel. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.

[55] Brendan Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264(4):947–963, 2013.

[56] Victor Patrangenaru and Leif Ellingson. *Nonparametric Statistics on Manifolds and their Applications to Object Data Analysis*. CRC Press, 2015.

[57] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[58] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[59] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[60] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[61] Bernhard Schmitzer. A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, 2016.

[62] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

[63] Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pages 3312–3320, 2015.

[64] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

[65] Max Sommerfeld and Axel Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

[66] Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23, 2019.

[67] Daniil Tiapkin, Alexander Gasnikov, and Pavel Dvurechensky. Stochastic saddle-point optimization for Wasserstein barycenters. *arXiv:2006.06763*, 2020.

[68] Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

[69] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, Berlin, 2008.

[70] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[71] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact Wasserstein distance. In *Uncertainty in Artificial Intelligence*, pages 433–453. PMLR, 2020.

[72] Jianbo Ye, Panruo Wu, James Z Wang, and Jia Li. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.

[73] Yoav Zemel and Victor M Panaretos. Fréchet means and Procrustes analysis in Wsserstein space. *Bernoulli*, 25(2):932–976, 2019.

[74] Vladimir Mikhailovich Zolotarev. Metric distances in spaces of random variables and their distributions. *Mathematics of the USSR-Sbornik*, 30(3):373–402, 1976.

CHAPTER B

Kantorovich-Rubinstein distance and barycenter for finitely supported measures: Foundations and Algorithms

Kantorovich-Rubinstein distance and barycenter for finitely supported measures: Foundations and Algorithms

Florian Heinemann * Marcel Klatt * Axel Munk *†‡

July 19, 2022

Abstract

The purpose of this paper is to provide a systematic discussion of a generalized barycenter based on a variant of unbalanced optimal transport (UOT) that defines a distance between general non-negative, finitely supported measures by allowing for mass creation and destruction modeled by some cost parameter. They are denoted as Kantorovich-Rubinstein (KR) barycenter and distance. In particular, we detail the influence of the cost parameter to structural properties of the KR barycenter and the KR distance. For the latter we highlight a closed form solution on ultra-metric trees. The support of such KR barycenters of finitely supported measures turns out to be finite in general and its structure to be explicitly specified by the support of the input measures. Additionally, we prove the existence of sparse KR barycenters and discuss potential computational approaches. The performance of the KR barycenter is compared to the OT barycenter on a multitude of synthetic datasets. We also consider barycenters based on the recently introduced Gaussian Hellinger-Kantorovich and Wasserstein-Fisher-Rao distances.

1 Introduction

Over the past decade, optimal transport (OT) based concepts for data analysis [for a thorough treatment of the mathematical foundations of optimal transport see e.g. Rachev and Rüschendorf, 1998, Villani, 2008, Santambrogio, 2015] have seen increasing popularity. This is mainly due to the fact that OT based methods respect important features of the data's geometric structure. Furthermore, noteworthy advances have been achieved in various areas, such as optimisation [Bertsimas and Tsitsiklis, 1997, Wolsey and Nemhauser, 1999, Grötschel et al., 2012], machine learning [Frogner et al., 2015, Peyré et al., 2019, Xie et al., 2020], computer vision [Gangbo and McCann, 2000, Su et al., 2015, Solomon et al., 2015] and statistical inference [Sommerfeld and Munk, 2018, Panaretos and Zemel, 2020, Hallin et al., 2021], among others. This methodological and computational progress recently also paved the way to novel areas of applications including genetics [Evans and Matsen, 2012, Schiebinger et al., 2019] and cell biology [Gellert et al., 2019, Klatt et al., 2020, Tameling et al., 2021, Wang and Yuan, 2021], to cite but a few. Of particular importance from a data analysis point of view are extensions to compare more than two measures, a prominent proposal being the Fréchet mean [Fréchet, 1948], in the present context known as *Wasserstein barycenter* [Agueh and Carlier, 2011]. Wasserstein barycenters allow for a notion of average on the space of probability measures, which is well-adapted

*Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen
†Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen
‡University Medical Center Göttingen, Cluster of Excellence 2067 Multiscale Bioimaging - From molecular machines to networks of excitable cells

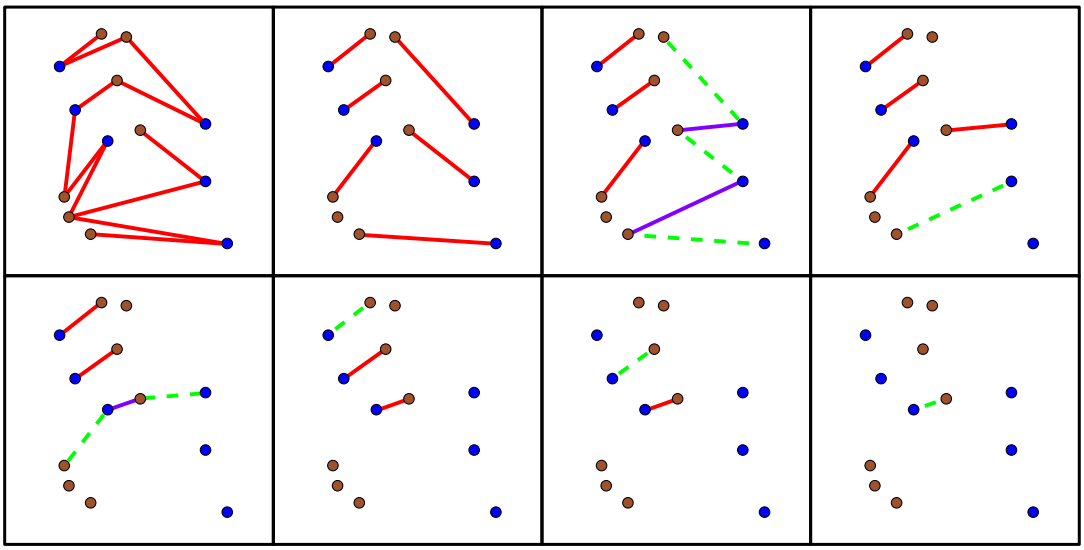


Figure 1: (Unbalanced) OT between two measures (support in blue and brown, respectively) with weights equal to one at each support point. **Top-Left:** OT plan (red) between normalised versions of the two measures. **Rest:** UOT plans (red/purple) between non-normalised measures. From top-left to bottom-right C is decreasing. The edges, which have been removed most recently due to the reduction of C, are shown in green. Edges which have been added to the UOT graph due to the most recent reduction of C are marked in purple.

to the geometry of the data [Álvarez-Esteban et al., 2016, Anderes et al., 2016]. With recent progress on their computation [Cuturi and Doucet, 2014, Carlier et al., 2015, Bonneel et al., 2015, Kroshnin et al., 2019, Ge et al., 2019, Heinemann et al., 2022] they establish themselves even further as a promising tool in many fields of data analysis, such as texture mixing [Rabin et al., 2011], distributional clustering [Ye et al., 2017], histogram regression [Bonneel et al., 2016], domain adaptation [Montesuma and Mboula, 2021] and unsupervised learning [Schmitz et al., 2018], among others.
However, a well known drawback of the Wasserstein distance and its barycenters in various applications is their limitation to measures with equal total mass. In fact, in many real world instances the difference in total mass intensity is of crucial importance. Employing vanilla Wasserstein based tools on general positive measures necessitates the usage of a normalisation procedure to enforce mass equality between the measures. This approach is, by design, oblivious to the mass differences between the original measures and can limit its use in applications. Exemplary, we mention that normalisation destroys stoichiometric features in the analysis of protein interaction and pathways as pointed out in Tameling et al. [2021]. Overall, this might lead to incorrect conclusions on specific applications. An illustrative example is given in Figure 1.

1.1 Prior Work

The limitation of OT based concepts dealing only with measures of equal total mass has opened a wealth of approaches to account for more general measures. As an early proposal of this idea, the *partial OT formulation* [Caffarelli and McCann, 2010, Figalli, 2010] suggests to fix the total mass of the OT plan in advance, while relaxing the marginal constraints. Comparably more recent are *entropy transport formulations*[1]. This general

[1]Critically, this is not be confused with entropy *regularized* optimal transport, which is a popular computational approach adding an entropy penalty term to the OT problem to allow for efficient, approximate

framework removes the marginal constraints and instead uses a divergence functional to measure the deviation between the transport marginals and the input measures. The entropy transport framework encompasses the *Hellinger-Kantorovich* distance [Liero et al., 2018, Chizat et al., 2018b], also known as *Wasserstein-Fisher-Rao* distance [Chizat et al., 2018a] and the *Gaussian Hellinger-Kantorovich* distance [Liero et al., 2018]. Inherent to all of these models is their dependency on parameters whose exact influence on the models' properties is generally not well understood. An alternative idea is based on extending the well-studied dynamic formulation of OT [Benamou and Brenier, 2000] to measures with different total masses. With a focus on its geodesic properties, this approach has been studied in several works [Chizat et al., 2018a,c, Gangbo et al., 2019].
In this paper, we rely on a simple and intuitive idea based on the seminal work of Kantorovich and Rubinstein [1958]. This accounts for mass construction and deletion at a cost modeled by some prespecified parameter [for details see also Hanin, 1992, Guittet, 2002]. It leads to the *Kantorovich-Rubinstein distance (KRD)* which curiously has been revisited several times under different names by various authors. For $p = 1$, it has been referred to as Earth Mover's Distance [Pele and Werman, 2008], and generalized Wasserstein distance [Piccoli and Rossi, 2014], while for general $p \geq 1$ common terminology includes Kantorovich distance [Gramfort et al., 2015], generalized KRD [Sato et al., 2020], transport-transform metric [Müller et al., 2020] and robust optimal transport distance [Mukherjee et al., 2021].

1.2 Contributions

In this work, we define barycenters with respect to the KRD and investigate their fundamental properties from a data analysis point of view. This extends the popular notion of Wasserstein barycenters to unbalanced barycenters (UBCs), i.e., barycenters of measures of different total masses. Similary, UBCs have been considered explicitly for the Hellinger-Kantorovich distance [Chung and Phung, 2020, Friesecke et al., 2021] and for the partial OT distance for absolutely continuous measures [Kitagawa and Pass, 2015]. Notably, the well-known approach of matrix scaling algorithms has been shown to provide a general framework to approximate any UBC based on entropy optimal transport [Chizat et al., 2018b] of finitely supported measures. Closely related to our approach is the work by Müller et al. [2020] approximating the KR barycenter in the special case of point patterns.
The KR distance: Let $(\mathcal{X}, d)$ be a finite metric space, where $\mathcal{X} = \{x_1, \dots, x_N\}$ and

$$\mathcal{M}_+(\mathcal{X}) \coloneqq \left\{ \mu \in \mathbb{R}^{|\mathcal{X}|} \mid \mu(x) \geq 0 \, \forall x \in \mathcal{X} \right\}$$

is the set of non-negative measures[2] on $\mathcal{X}$. For a measure $\mu \in \mathcal{M}_+(\mathcal{X})$ its total mass is defined as $\mathbb{M}(\mu) \coloneqq \sum_{x \in \mathcal{X}} \mu(x)$ and the subset of non-negative measures with total mass equal to one is the set of probability measures $\mathcal{P}(\mathcal{X})$. If $\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X})$ is a measure on the product space $\mathcal{X} \times \mathcal{X}$ its marginals are defined as $\pi(x, \mathcal{X}) \coloneqq \sum_{x' \in \mathcal{X}} \pi(x, x')$ and $\pi(\mathcal{X}, x') \coloneqq \sum_{x \in \mathcal{X}} \pi(x, x')$, respectively. For two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ we define the set of *non-negative sub-couplings* as

$$\begin{aligned} \Pi_{\leq}(\mu, \nu) \coloneqq \{ \pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X}) \mid & \pi(x, \mathcal{X}) \leq \mu(x), \\ & \pi(\mathcal{X}, x') \leq \nu(x') \, \forall \, x, x' \in \mathcal{X} \}. \end{aligned} \tag{1}$$

computations [Cuturi, 2013, Benamou et al., 2015, Carlier et al., 2017]

[2] A non-negative measure on a finite space $\mathcal{X}$ is uniquely characterized by the values it assigns to each singleton $\{x\}$. To ease notation we write $\mu(x)$ instead of $\mu(\{x\})$. The corresponding σ-field is always to be understood as the powerset of $\mathcal{X}$.

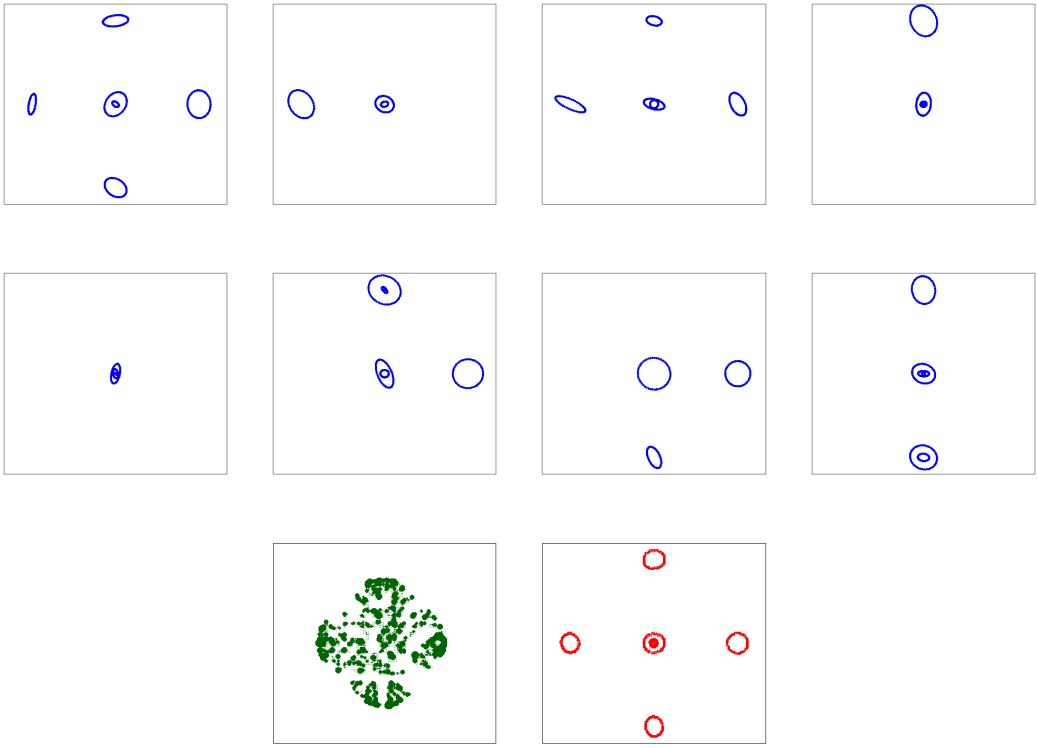


Figure 2: **Upper two rows:** An excerpt of eight instances of a dataset of $N = 100$ nested ellipses at up to 5 different clusters in $[0, 1]^2$. The number of ellipses in each cluster follows a Poisson distribution. For the cluster in the center the intensity is 2 and for the four outer clusters the intensity is 1. Each ellipse is discretized into 50 points with mass 1 at each location. **Bottom-Left:** The Wasserstein barycenter of the normalized versions of these measures. **Bottom-Right:** The $(2, 0.2)$-barycenter of these measures. The $(2, C)$-barycenter for different values of C can be seen in Figure 8.

Similarly, we denote the set of *couplings* between μ and ν as $\Pi_=(\mu, \nu)$, where the inequality constraints in (1) are replaced by equalities. For $p \geq 1$ and a parameter $C > 0$, *unbalanced optimal transport* (UOT) between two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ is defined as

$$\begin{aligned} \mathrm{UOT}_{p,C}(\mu, \nu) \coloneqq \min_{\pi \in \Pi_\leq(\mu,\nu)} & \sum_{x,x' \in \mathcal{X}} d^p(x, x')\pi(x, x') \\ & + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi) \right). \end{aligned} \tag{2}$$

Notably, $\mathrm{UOT}_{p,C}(\mu, \nu)$ is finite for all measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ with possibly different total masses and a solution of (2) always exists. Here, the parameter C penalizes deviation of mass from the marginals of π with respect to the input measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$. In particular and unlike the (balanced) OT problem

$$\mathrm{OT}_p(\mu, \nu) \coloneqq \min_{\Pi_=(\mu,\nu)} \sum_{x,x' \in \mathcal{X}} d^p(x, x')\pi(x, x')$$

defined only for measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ with equal total mass $\mathbb{M}(\mu) = \mathbb{M}(\nu)$, UOT in (2) relaxes the marginal constraint and allows optimal solutions to have more flexible

marginals. Based upon UOT we define the p-th order *Kantorovich-Rubinstein distance* between two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ as

$$\mathrm{KR}_{p,C}(\mu,\nu) \coloneqq (\mathrm{UOT}_{p,C}(\mu,\nu))^{1/p}. \tag{3}$$

For any $p \geq 1$, it defines a distance on the space of non-negative measures $\mathcal{M}_+(\mathcal{X})$ and it is an extension of the well-known *p-Wasserstein distance* $W_p(\mu,\nu) \coloneqq (\mathrm{OT}_p(\mu,\nu))^{1/p}$ defined only for measures of equal total mass. Indeed, the KRD is shown to interpolate in-between *OT on small scales* and *point-wise comparisons on large scales* (Theorem 2.2) relative to the parameter C. This allows for an intuitive interpretation of the KRD. More precisely, in Lemma 2.1, we detail a clear geometrical connection between the value of C and the structure of the UOT. In particular, this contrasts the closely related partial OT problem [Figalli, 2010] mentioned above. Employing Lagrange multipliers one can see that for any choice of C, there exists a fixed mass m of the partial OT problem, such that these two problems are equivalent. However, finding this value of m requires to solve the UOT problem. We stress that the influence of m on the resulting transport is in general hard to determine, while the impact of C is intuitively clear. Thus, this perspective seems better suited to many applications. For the specific case of measures supported on ultra-metric trees (Section 2.1.1) we prove (Theorem 2.3) an analogue of the well-known closed formula for the p-Wasserstein distance [Kloeckner, 2015]. Additionally, the computation of the KRD is known to be equivalent to solving a related balanced OT problem [Guittet, 2002], allowing to apply any state-of-the-art solver with minimal modifications to compute the KRD and plan.

The KR barycenter: The KRD also lends itself to define a notion of a barycenter for a collection of measures as a generalization of the p-Wasserstein barycenter defined for probability measures $\mu_1, \ldots, \mu_J \in \mathcal{P}(\mathcal{X})$ as

$$\tilde{\mu} \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathcal{Y})} \frac{1}{J} \sum_{i=1}^{J} W_p^p(\mu, \mu_i). \tag{4}$$

Here, $(\mathcal{X}, d)$ is assumed to be embedded in some ambient space $(\mathcal{Y}, d)$, e.g., an Euclidean space with $\mathcal{X} \subset \mathcal{Y}$. The distance d on $\mathcal{X}$ is understood to be the distance on $\mathcal{Y}$ restricted to $\mathcal{X}$. For $\mu_1, \ldots, \mu_J \in \mathcal{M}_+(\mathcal{X})$, any measure

$$\mu^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu) \coloneqq \frac{1}{J} \sum_{i=1}^{J} \mathrm{KR}_{p,C}^p(\mu_i, \mu) \tag{5}$$

is said to be a (p,C)-*Kantorovich-Rubinstein barycenter* or (p,C)-*barycenter* for short[3]. We refer to the objective functional $F_{p,C}$ as (unbalanced) (p,C)-*Fréchet functional*. Notably, (p,C)-barycenters' support is not restricted to the finite space $\mathcal{X}$ which raises fundamental questions on its structural properties. In the following, we establish that there exists a finite set containing the support of any (p,C)-barycenter (Section 2.2). Indeed, this set can be explicitly constructed from the support of the individual μ_i's, but its size grows exponentially in the number of individual measures. However, we prove that there always exists a *sparse* (p,C)-barycenter whose support size is at most linear in the number of measures (Theorem 2.5). We note that these properties are analogs of well-known properties of Wasserstein barycenters [Anderes et al., 2016], that we re-establish for the unbalanced setting.

[3]For the sake of readability, the weights in this definition are fixed to $1/J$, though it is easy to adapt all instances of their occurrence in this work to arbitrary positive weights $\lambda_1, \ldots, \lambda_J$, summing to 1.

Comparably, employing more general entropy transport distances, we are not aware of any similar structural description of their barycenters in terms of the input measures and the parameter. Notably, the entropy optimal transport barycenter of dirac measures is not necessarily finitely supported itself [for an example see Friesecke et al., 2021]. In contrast, our explicit structural description of the support of KR barycenters provides an immediate understanding of its properties for a given choice of C. This clear link between C and the (p, C)-barycenter also allows to incorporate previous knowledge of the measures or the ground space into the choice C. The (p, C)-barycenter can be tuned to be more flexible and provide superior performance compared to its p-Wasserstein counterpart by avoiding to normalise each measure. An illustrative example is included in Figure 2, where the (p, C)-barycenter detects all clusters correctly, while the Wasserstein barycenter does not provide any structural information on the underlying measures. This showcases potentially superior robustness and flexibility of the (p, C)-barycenter compared to the Wasserstein barycenter. We study this comparison in more detail on multiple synthetic data sets in Section 4. Here, the computational results[4] are based on the fact that, due to our structural analysis of the support of the (p, C)-barycenter, it is straightforward to modify any given state-of-the-art solver for the Wasserstein barycenter problem to solve the (p, C)-barycenter problem (Section 4.1).

2 Kantorovich-Rubinstein Distance and $(\mathbf{p}, \mathbf{C})$-Barycenter

In this section, we provide some theoretical analysis of the structural properties inherent in the UOT in (2) and as a consequence to the KRD in (3). We also focus on the variational formulation defining the (p, C)-barycenter in (5).

2.1 KR Distance

In this subsection, we focus on structural properties of minimizers for UOT in (2) and their consequences for the KRD. Notably, one can equivalently restate the penalization of total mass in (2) as

$$\begin{aligned} & C\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2}-\mathbb{M}(\pi)\right) \\ =& \frac{C}{2}\left(\sum_{x\in\mathcal{X}}(\mu(x)-\pi(x,\mathcal{X}))+\sum_{x'\in\mathcal{X}}\left(\nu(x')-\pi(\mathcal{X},x')\right)\right). \end{aligned} \tag{6}$$

While in (2) the parameter $C > 0$ controls the deviation of the total mass of π, the alternative representation (6) demonstrates its marginal characterization. Indeed, the parameter C specifies the maximal distance (scale) for which transportation is cheaper than creation or destruction of mass. More precisely, each optimal solution π_C for (2) induces a directed transportation graph $G(\pi_C)$ between the support points of μ (source points) and the support points of ν (sink points). By definition, the graph $G(\pi_C)$ contains a directed edge (x, x') if and only if $\pi_C(x, x') > 0$. For a directed path $P = (x_{i_1}, \ldots, x_{i_k})$ in $G(\pi_C)$ its path length is defined as $\mathcal{L}(P) = \sum_{j=1}^{k-1} d^p(x_{i_j}, x_{i_{j-1}})$. The parameter $C > 0$ determines the maximal path length for any path in $G(\pi_C)$ as the following statement demonstrates.

[4] An implementation can be found in the R package *WSGeometry* at https://github.com/F-Heinemann/WSGeometry/.

Lemma 2.1. *For $p \geq 1$, parameter $C > 0$ and measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ consider the UOT (2) with an optimal solution π_C. The length of any directed path P from the corresponding transport graph $G(\pi_C)$ is bounded by*

$$\mathcal{L}(P) \leq C^p.$$

In particular, if $d(x, x') > C$ then for any optimal solution of 2 it holds $\pi_C(x, x') = 0$.

A proof is included in Appendix A.2. Lemma 2.1 shows that the underlying transportation graph has maximal path length C^p which limits the interaction between source and sink points. It will be of crucial importance for closed formulas on ultra-metric trees in the following subsection. As an immediate consequence we obtain some important statements on the KRD in (3) along with its metric property.

Theorem 2.2. *For any $p \geq 1$ and parameter $C > 0$ the following statements hold:*

(i) The p-th order KRD in (3) defines a metric on the space of non-negative measures $\mathcal{M}_+(\mathcal{X})$.

(ii) If $C \leq \min_{x \neq x'} d(x, x')$, then it holds that

$$KR^p_{p,C}(\mu, \nu) = \frac{C^p}{2} TV(\mu, \nu),$$

where $TV(\mu, \nu) := 1/2 \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$ is the total variation distance. The same equality holds for all $C > 0$ if $\mu(x) \geq \nu(x)$ for all $x \in \mathcal{X}$ or if $\mu(x) \leq \nu(x)$ for all $x \in \mathcal{X}$.

(iii) If $C \geq \max_{x,x'} d(x, x')$ and $\mathbb{M}(\mu) = \mathbb{M}(\nu)$, then it holds that

$$KR^p_{p,C}(\mu, \nu) = W^p_p(\mu, \nu).$$

(iv) If $C_1 \leq C_2$, then it holds

$$KR^p_{p,C_1}(\mu, \nu) \leq KR^p_{p,C_2}(\mu, \nu).$$

We stress that the metric property of the KRD in Theorem 2.2 (i) has already been established in specific instances, e.g., for $p = 1$ [Piccoli and Rossi, 2014]. Our proof follows that of Theorem 2 in Müller et al. [2020] for uniform measures on point patterns with minor modifications.
Theorem 2.2 demonstrates how two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ are compared with respect to KRD. Depending on the parameter $C > 0$ the optimal value interpolates between p-th order Wasserstein distance on small scales and total variation on larger scales with respect to C. Equivalently, these properties can be shown by considerations of the *dual* program for UOT in (2) given by

$$\mathrm{UOT}_{p,C}(\mu, \nu) = \max_{\substack{f,g\colon \mathcal{X} \to \mathbb{R} \\ f \leq C^p/2,\, g \leq C^p/2}} \sum_{x \in \mathcal{X}} f(x)\mu(x) + \sum_{x' \in \mathcal{X}} g(x')\nu(x') \qquad (\mathrm{DUOT}_{p,C})$$

$$\text{s.t. } f(x) + g(x') \leq d^p(x, x'),\ \forall x, x' \in \mathcal{X},$$

where the equality holds due to *strong duality*. For $p = 1$ this can be further specified to

$$\mathrm{UOT}_{1,C}(\mu, \nu) = \max_{\substack{f\colon \mathcal{X} \to \mathbb{R} \\ f\ 1-\text{Lipschitz} \\ \|f\|_\infty \leq C/2}} \sum_{x \in \mathcal{X}} f(x)(\mu(x) - \nu(x))$$

which reveals its relation to the *flat metric* [Bogachev, 2007] as observed in Lellmann et al. [2014], Schmitzer and Wirth [2019]. As in general $\mathbb{M}(\mu) \neq \mathbb{M}(\nu)$, the bound $f, g \leq C^p/2$ on dual feasible solutions f, g is necessary for the dual to be finite. However, if the measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ have equal total mass $\mathbb{M}(\mu) = \mathbb{M}(\nu)$ and $C \geq \max_{x,x'} d(x, x')$, then the bound on dual feasible solutions is redundant and we obtain the dual of the usual OT problem

$$\mathrm{OT}_p(\mu, \nu) = \max_{f,g\colon \mathcal{X}\to\mathbb{R}} \sum_{x\in\mathcal{X}} f(x)\mu(x) + \sum_{x'\in\mathcal{X}} g(x')\nu(x') \qquad (\mathrm{DOT}_p)$$
$$\text{s.t. } f(x) + g(x') \leq d^p(x, x'),\ \forall x, x' \in \mathcal{X}.$$

2.1.1 KR Distance on Ultrametric Trees

For OT, the approximations of the underlying distance by a tree metric are common tools for theoretical and practical purposes. The former is usually employed for rates of convergence for the expectation of empirical OT costs [Sommerfeld et al., 2019] while in the latter tree approximations serve to reduce the computational complexity inherent in OT [Le et al., 2019]. OT on ultramatric trees is also applied for the analysis of phylogenetic trees [Gavryushkin and Drummond, 2016]. For an efficient computational implementation of UOT on tree metrics we refer to Sato et al. [2020]. Notably, while OT with tree metric costs has a closed form solution, this fails to hold for its UOT counterpart. An exception is given in terms of ultrametric trees for which not only OT [Kloeckner, 2015] but also UOT admits a closed form solution, which we establish in this subsection.

To this end, consider a tree $\mathcal{T}$ with nodes V, edges E attached with (non-negative) weights $w(e)$ for $e \in E$ and a designated root r. Two nodes $\mathsf{v}, \mathsf{w} \in V$ are connected by a unique path denoted $\mathcal{P}(\mathsf{v}, \mathsf{w})$ either represented by a sequence of nodes or as a sequence of edges. The distance $d_{\mathcal{T}}(\mathsf{v}, \mathsf{w})$ is equal to the sum of the weights of those edges contained in $\mathcal{P}(\mathsf{v}, \mathsf{w})$. A *leaf* of $\mathcal{T}$ is any node such that its degree (number of edges attached to the node) is equal to one and the set of all leaf nodes is denoted as $L \subset V$. A node $\mathsf{v}^\star$ is termed *parent* of node v denoted by $\mathrm{par}(\mathsf{v}) = \mathsf{v}^\star$ if both are connected by a single edge but $\mathsf{v}^\star$ is closer to the root than v. The parent of the root node is set to $\mathrm{par}(\mathsf{r}) = \mathsf{r}$. For a node v its *children* are the elements of the set $\mathcal{C}(\mathsf{v}) = \{\mathsf{w} \in V \mid \mathsf{v} \in \mathcal{P}(\mathsf{w}, \mathsf{r})\}$. Notice that with this definition v is a child of itself (Figure 3 (a) for an illustration).

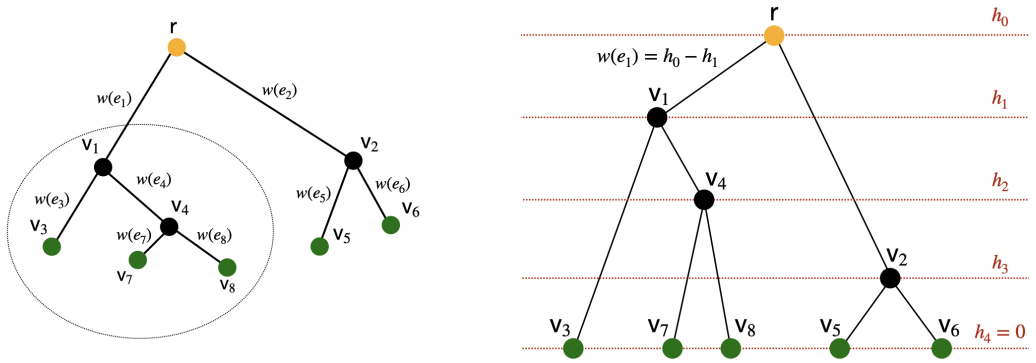
A tree $\mathcal{T}$ is termed *ultrametric tree* if all its leaf nodes are at the same distance to the root. Equivalently, there exists a *height function* $h\colon V \to \mathbb{R}_+$ that is monotonically decreasing meaning that $h(\mathrm{par}(\mathsf{v})) \geq h(\mathsf{v})$ and such that $h(\mathsf{v}) = 0$ for $\mathsf{v} \in L$. The distance is set to $d_{\mathcal{T}}(\mathsf{v}, \mathrm{par}(\mathsf{v})) = |h(\mathsf{v}) - h(\mathrm{par}(\mathsf{v}))|$ and extended on the full tree (Figure 3 (b) for an illustration).

Consider an ultrametric tree $\mathcal{T}$ with height function h and measures μ^L, ν^L supported on the leaf nodes $L \subset V$. We prove that the p-th order KRD admits a *closed formula* for such a setting. Intuitively, the parameter C restricts transportation of mass up to a certain threshold allowing to decompose $\mathcal{T}$ into subtrees. Mass transportation is restricted solely within each subtree whereas mass abundance or deficiency is penalized with parameter C for each particular subtree (Figure 4 for an illustration). We define the set

$$\mathcal{R}(C) \coloneqq \left\{\mathsf{v} \in V \mid h(\mathsf{v}) \leq \frac{C}{2} < h(\mathrm{par}(\mathsf{v}))\right\} \qquad (7)$$

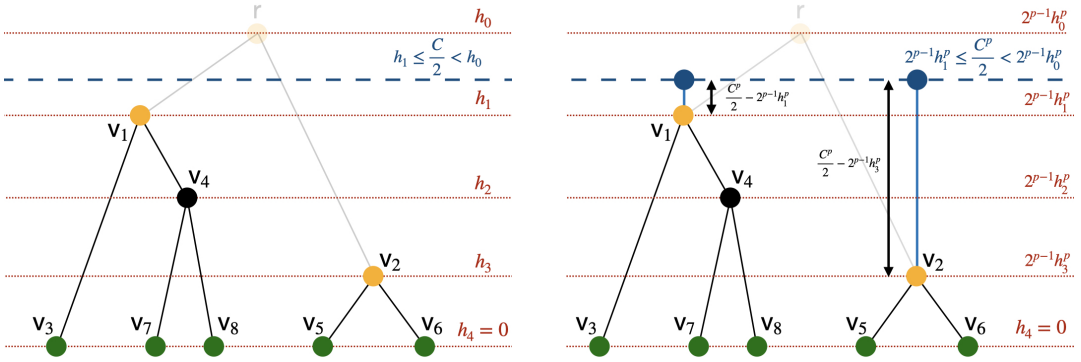
with the convention that $\mathcal{R}(C) = \{\mathsf{r}\}$ if $C/2 \geq h(\mathsf{r})$ and for a node $\mathsf{v} \in V$ set

$$\mu^L(\mathcal{C}(\mathsf{v})) \coloneqq \sum_{\mathsf{w}\in\mathcal{C}(v)\cap L} \mu^L(w).$$



(a) (b)

Figure 3: **General Tree Structures: (a)** A tree graph $\mathcal{T}$ with root r (orange), internal nodes (black) and leaf nodes L (green). By definition $\mathrm{par}(\mathsf{v}_5) = \mathrm{par}(\mathsf{v}_6) = \mathsf{v}_2$ and the children of v_1 are equal $\mathcal{C}(\mathsf{v}_1) = \{\mathsf{v}_1, \mathsf{v}_3, \mathsf{v}_4, \mathsf{v}_7, \mathsf{v}_8\}$. The distance from each leaf node to the root may vary. **(b)** An ultrametric tree $\mathcal{T}$ with height function h (red) such that $0 = h_4 3 < h_3 < h_2 < h_1 < h_0$. Edge weights are defined by the the difference of consecutive height values, e.g. $w(e_1) = h_0 - h_1$. Each leaf node (green) is at the same distance to the root r (orange).



(a) (b)

Figure 4: **Closed formula for the KRD on ultrametric trees: (a)** Depending on the regularization $C > 0$ and the underlying height function h the ultrametric tree $\mathcal{T}$ introduced in Figure 3 (b) is decomposed into two subtrees. Each node in the set $\mathcal{R}(C) = \{\mathsf{v}_1, \mathsf{v}_2\}$ (orange) serves as a new root and corresponding subtrees $\mathcal{T}(\mathsf{v}_1) \coloneqq \mathcal{C}(\mathsf{v}_1)$ and $\mathcal{T}(\mathsf{v}_2) \coloneqq \mathcal{C}(\mathsf{v}_2)$ are equal their respective set of children with corresponding edges. **(b)** The p-th height transformation $\mathcal{T}_p(\mathsf{v}_1)$ and $\mathcal{T}_p(\mathsf{v}_2)$ of the induced subtrees $\mathcal{T}(\mathsf{v}_1)$ and $\mathcal{T}(\mathsf{v}_2)$, respectively. Each subtree is extended by a new root (blue) with an edge (lightblue) whose distance is equal the difference of regularization $C^p/2$ and the p-th height transformed value of the former root.

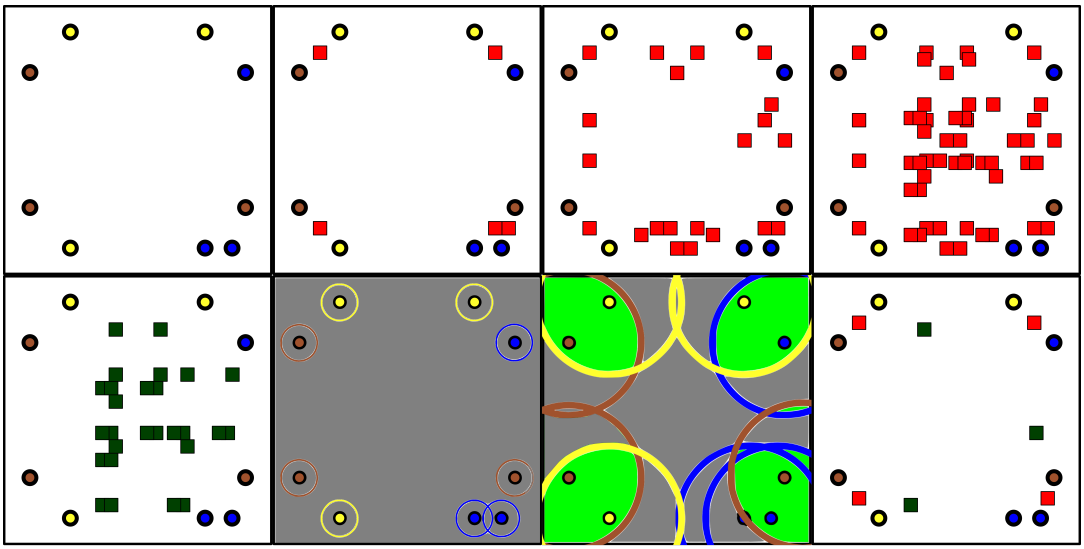


Figure 5: Centroid sets and barycenters: The support points of three ($J = 3$) measures (yellow, brown and blue dots) with unit mass at each position. **Top:** Different centroid sets $\mathcal{C}_{KR}(3, 2, C)$ (red squares) with increasing value of C from left to right. **Bottom-Left:** The centroid set $\mathcal{C}^W(3, 2)$ (dark green squares) corresponding to the 2-Wasserstein barycenter. **Bottom-Center:** Circles corresponding to C^p (for two different choices of C) balls around the support points. The grey colouring indicates that there is no overlap of at least two circles in this area and thus no $(2, C)$-barycenter can have mass in this area. Conversely, the green colouring indicates overlap and thus the potential support area of the barycenter. **Bottom-Right:** The $(2, C)$-barycenter (red squares) for a specific choice of C and the 2-Wasserstein barycenter (dark green squares).

Theorem 2.3 (KR on ultrametric trees)**.** *Consider an ultrametric tree $\mathcal{T}$ with leaf nodes L and height function $h\colon V \to \mathbb{R}_+$ inducing the tree metric $d_{\mathcal{T}}$. For any $p \geq 1$ and two measures $\mu^L, \nu^L \in \mathcal{M}_+(L)$ supported on the leaf nodes of $\mathcal{T}$ it holds that*

$$KR^p_{p,C}\left(\mu^L, \nu^L\right) =$$
$$\sum_{v \in \mathcal{R}(C)} \left(2^{p-1} \sum_{w \in \mathcal{C}(v) \setminus \{v\}} \left(\left(h(par(w))^p - h(w)^p\right) \left|\mu^L(\mathcal{C}(w)) - \nu^L(\mathcal{C}(w))\right| \right) \right.$$
$$\left. + \left(\frac{C^p}{2} - 2^{p-1} h(v)^p \right) \left|\mu^L(\mathcal{C}(v)) - \nu^L(\mathcal{C}(v))\right| \right).$$

The closed formula in Theorem 2.3 decomposes the underlying UOT into two tasks. While summing over subtrees carried out by the outer sum, the inner sum consists of two terms. The first considers OT within each subtree whereas the second accounts for mass deviation on that particular subtree.
The proof of this formula is given in Appendix A.2.1.

2.2 $(\mathbf{p}, \mathbf{C})$-Barycenters

In the finite setting considered in this work a (p, C)-barycenter as defined in (5) always exists, but is not necessarily unique. Moreover, the location and structure of the support of the (p, C)-barycenter are not fixed and hence unknown. For the Wasserstein barycenter

there exists a finitely supported, sparse barycenter in this context [Anderes et al., 2016, Le Gouic and Loubes, 2017]. We establish analog properties of the (p, C)-barycenter.

Definition 2.4. *Let $(\mathcal{Y}, d)$ be a metric space, $p \geq 1$ and $J \in \mathbb{N}$. A Borel barycenter application $T^{J,p}$ associates to any points $(y_1, \dots, y_J) \in \mathcal{Y}^J$ a minimum $y^\star \in \mathcal{Y}$ of $\sum_{i=1}^J d^p(y_i, y)$, i.e.,*

$$T^{J,p}(y_1, \dots, y_J) \in \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{J} d^p(y_i, y).$$

A Borel barycenter application is in general not a function since the minimum does not need to be unique. In particular, $y = T^{J,p}(y_1, \dots, y_J)$ only means that y is one of the minima of the average distance function. As the measures $\mu_1, \dots, \mu_J$ are defined on $\mathcal{X}$ we usually restrict the Borel barycenter application to inputs from the space $\mathcal{X} \subset \mathcal{Y}$. We define the *full centroid set* of the measures $\mu_1, \dots, \mu_J \in \mathcal{M}_+(\mathcal{X})$ as

$$\begin{aligned} \mathcal{C}_{KR}(J,p) = \Big\{ y \in \mathcal{Y} \mid \exists L \geq \lceil J/2 \rceil,\ \exists (i_1, \dots, i_L) \subset \{1, \dots, J\}, \\ x_1, \dots, x_L :\ x_l \in \operatorname{supp}(\mu_{i_l}) \\ \forall l = 1, \dots, L :\ y = T^{L,p}(x_1, \dots, x_L) \Big\}, \end{aligned} \tag{8}$$

and the *restricted centroid set*

$$\begin{aligned} \mathcal{C}_{KR}(J,p,C) = \Big\{ y = T^{L,p}(x_1, \dots, x_L) \in \mathcal{C}_{KR}(J,p) \mid \forall 1 \leq l \leq L : \\ d^p(x_l, y) \leq C^p;\ \sum_{i=1}^{L} d^p(x_l, y) \leq \frac{C^p(2L - J)}{2} \Big\}. \end{aligned} \tag{9}$$

We stress that for each L-tupel $(x_1, \dots, x_L)$ one fixed representative of $T^{L,p}(x_1, \dots, x_L)$ is chosen for the construction of the centroid set $C_{KR}(J, p, C)$. To streamline the presentation any statement concerning $\mathcal{C}_{KR}(J, p, C)$ in the following theorem is to be understood in the sense that there exists a choice of $\mathcal{C}_{KR}(J, p, C)$ such that the statement holds true.

Theorem 2.5. *Let $\mu_1, \dots, \mu_J \in \mathcal{M}_+(\mathcal{X})$ be a collection of non-negative measures on the finite discrete space $\mathcal{X} \subset \mathcal{Y}$. For any $C > 0$ it holds that*

(i)

$$\inf_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu) = \inf_{\substack{\mu \in \mathcal{M}_+(\mathcal{Y}) \\ supp(\mu) \subseteq \mathcal{C}_{KR}(J,p,C)}} F_{p,C}(\mu).$$

Moreover, any (p, C)-barycenter $\mu^\star$ satisfies $supp(\mu^\star) \subseteq \mathcal{C}_{KR}(J, p, C)$ and its total mass is bounded by

$$0 \leq \mathbb{M}(\mu^\star) \leq \frac{2}{J} \sum_{i=1}^{J} \mathbb{M}(\mu_i).$$

(ii) For any (p, C)-barycenter μ^* and any point $y \in supp(\mu^*)$, there exist UOT plans π_i between μ^* and μ_i for $i = 1, \dots, J$, respectively, such that if $\pi_i(y, x) > 0$, then there exists $L \geq \lceil J/2 \rceil$, $x_l \in supp(\mu_{i_l})$ for $l = 2, \dots, L$, $(i_2, \dots, i_L) \subset \{1, \dots, J\}$ and $i_l \neq i$ for $l = 2, \dots, L$ with $y = T^{L,p}(x, x_{i_2}, \dots, x_{i_L})$, $\pi_j(y, x_j) > 0$ if $j \in \{i_2, \dots, i_L\}$. Additionally, if for any $(x_1, \dots, x_L) \in \mathcal{Y}^L$ it holds that

$$T^{L,p}(x_1, \dots, x_L) = T^{L,p}(y_1, x_2, \dots, x_L) \Leftrightarrow x_1 = y_1, \tag{10}$$

then $\pi_i(y, x) \in \{0, \mu^*(y)\}$ for $i = 1, \dots, J$.

(iii) *If* $M_i := |supp(\mu_i)|$ *for* $1 \leq i \leq J$ *then there exists a* (p, C)*-barycenter* $\mu^\star$ *such that*

$$|supp(\mu^\star)| \leq \min\left\{|\mathcal{C}_{KR}(J, p, C)|, \sum_{i=1}^{J} M_i\right\}.$$

(iv) *If* $C_1 \leq C_2$*, then it holds*

$$\inf_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C_1}(\mu) \leq \inf_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C_2}(\mu).$$

(v) *Furthermore, set* $\mathcal{Z} := \bigcup_{i=1}^{J} supp(\mu_i) \cup \mathcal{C}_{KR}(J, p)$ *and define*

$$d'_{\min} := \min_{x \in \mathcal{Z} \setminus \mathcal{C}_{KR}(J,p),\ y \in \mathcal{C}_{KR}(J,p)} d(x, y).$$

If $C \leq d'_{\min}$*, then the* (p, C)*-barycenter* $\mu^\star$ *is given by*

$$\mu^\star = \sum_{x \in \mathcal{X}} med(\mu_1(x), \ldots, \mu_J(x))\delta_x.$$

(vi) *Let* $C > J^{1/p} diam(\mathcal{Z})$ *and let* $\mu_1, \ldots, \mu_J$ *be ordered such that* $\mathbb{M}(\mu_i) \leq \mathbb{M}(\mu_j)$ *for* $i \leq j$*. Suppose that* J *is odd or there there exists no point* $y \in \mathcal{Y}$ *contained in at least* $J/2$ *different support sets. Then, for any* (p, C)*-barycenter* $\mu^\star$ *it holds that* $\mathbb{M}(\mu^\star) = \mathbb{M}\left(\mu_{\lceil J/2 \rceil}\right)$*. Else, there exists at least one* (p, C)*-barycenter with this total mass.*

The proof is based on the fact that finding a (p, C)-barycenter can be proven to be equivalent to solving a *multi-marginal optimal transport problem* (Section 3.2). Statement (i) provides insights into the structure of the support of any (p, C)-barycenter and its dependency with respect to the magnitude of C. The definition of $\mathcal{C}_{KR}(J, p, C)$ can be understood as a joint restriction on $\sum_{i=1}^{L} d^p(x_i, y)$ combined with an individual restriction on each $d^p(x_i, y)$ of the original centroid points of $\mathcal{C}_{KR}(J, p)$. The joint restriction ensures that simply deleting any mass at a given centroid point (and thus reducing the total mass of the measure) does not improve the objective value. This is a minimal feasibility assumption on the considered centroid point, as otherwise no measure containing this point can be optimal. The second restriction concerns each point individually. If a point x_i has a distance larger than C^p from a point y, then, by Lemma 2.1, there is no transport between y and x_i. Thus, centroids which have have a larger distance to one of the points $x_1, \ldots, x_L$ they are constructed from can not be in the support of any (p, C)-barycenter. This also gives rise to some helpful intuition for the support structure of any (p, C)-barycenter. Considering all C^p-neighbourhoods around any of the support points of the μ_i, then a (p, C)-barycenter can only have support in regions where at least balls from $\lceil J/2 \rceil$ different measures intersect. A visual representation of this is given in the center of bottom row of Figure 5. By definition, the sets $\mathcal{C}_{KR}(J, p, C)$ are equipped with a natural ordering in the sense that if $C_1 \leq C_2$ then $\mathcal{C}_{KR}(J, p, C_1) \subseteq \mathcal{C}_{KR}(J, p, C_2)$. Moreover, if C is large enough then $\mathcal{C}_{KR}(J, p, C) = \mathcal{C}_{KR}(J, p)$. We illustrate these sets in the top row Figure 5. We observe that the cardinality of the restricted centroid set in (9) decreases with decreasing C. In the extremes for large C the restricted centroid sets coincides with the full centroid sets in (8) that is independent of C. For small C, if there is no point which is contained in the support of at least $J/2$ measures, the restricted centroid set is empty. For an illustration we refer to the top row of Figure 5.
Property (ii) is an analogue to a well-known characterization [Anderes et al., 2016] of the p-Wasserstein barycenter on $\mathbb{R}^d$ with Euclidean distance d_2, where the transport from the

barycenter to the underlying measures is characterized by a transport map. The corresponding statement for the (p, C)-barycenter holds true as well in this context. Indeed, on $(\mathbb{R}^d, d_2)$ condition (10), which can be understood as an injectivity-type assumption on the barycentric application, is satisfied due to the fact that $T^L(x_1, \dots, x_L) = \frac{1}{L}\sum_{l=1}^{L} x_l$. However, for $(\mathbb{R}^d, d_1)$ this assertion does not hold. Consider $x_1 < x_2 < x_3 < x_4 \in \mathbb{R}$ and measures $\mu_1 = \delta_{x_1} + \delta_{x_2}, \mu_2 = \delta_{x_3} + \delta_{x_4}$, then any measure of the form $\mu^* = 2\delta_y$ for any $y \in [x_2, x_3]$ is a (p, C)-barycenter for $C > 2|x_1 - x_4|$. Thus, there only exist mass-splitting UOT plans between μ^* and μ_1, μ_2 and the transport is not characterized by a transport map. On more general spaces such as a tree $\mathcal{T}$ rooted at r, three leaves x_1, x_2, x_3 and positive edge weights $e_1, \dots, e_3 \in (0, 1)$ the barycenter on $\mathcal{T}$ of any two leafs $x_i \neq x_j$, is the root r. In particular, in this example, or in fact in any tree $\mathcal{T} = (V, E)$ which has a vertex y with degree of at least three[5] condition (10) fails. The unique $(2, 2)$-barycenter of two measures $\mu_1 = \delta_{x_1} + \delta_{x_2}$ and $\mu_2 = \delta_{x_2} + \delta_{x_3}$ is given by $\mu^* = 2\delta_r$. Thus, there are again only mass-splitting UOT plans between μ^* and μ_1 and μ_2. However, for the unit circle $\mathcal{S}^1$ equipped with its natural arc-length distance property (10) does hold. Assume $a_0 = T^L(x_1, \dots, x_L) = T^{L,p}(y_1, \dots, x_L)$, $a_1 = T^{L-1,p}(x_2, \dots, x_L)$ and for each $x \in \mathcal{S}_1$ denote $H_r(x)$ and $H_l(x)$ as the halfcircle right and left of x, respectively. It is straightforward to see by contraposition that if it holds $a_1 \in H_r(a_0)$, then this implies $x_1, y_1 \in H_l(a_1)$ and $x_1, y_1 \in H_l(a_0)$. However, it also holds $d(x_1, a_0) = d(y_1, a_0)$, and thus $\langle x_1 - y_1, a_0 \rangle = 0$. In particular, this implies that either $x_1 \in H_l(a_0)$ and $y_1 \in H_r(a_0)$ or vice versa and hence $x_1 = y_1$. The case $a_1 \in H_l(a_0)$ is analog and the case $a_0 = a_1$ clear.

Property (iii) guarantees the existence of *sparse* (p, C)-barycenters. For large C the size $\mathcal{C}_{KR}(J, p, C)$ scales as $\prod_{i=1}^{J} M_i$, growing essentially exponentially in J. However, here we see that there always exists a (p, C)-barycenter supported on a sparse subset of $\mathcal{C}_{KR}(J, p, C)$ which has cardinality growing only linearly in J. Part (iv) simply extends the montonicity of the (p, C)-KRD to the (p, C)-Fréchet functional. Statement (v) yields a critical point after which decreasing C does no longer change the resulting (p, C)-barycenter and provides a closed form characterisation of the (p, C)-barycenter in this context. Finally, statement (vi) enables control on the total mass of the (p, C)-barycenter for large values of C. In particular, since the total mass is close to the median of the total masses of the μ_i, we point out that the total mass of the (p, C)-barycenter in this setting is robust against outliers. A small amount of measures with unreasonably high mass has no impact on the total mass of the (p, C)-barycenter.

Naturally, we compare the (p, C)-barycenter to its popular Wasserstein analogue in (4). As proven in Le Gouic and Loubes [2017] [and initially for $p = 2$ for $\mathbb{R}^d$ by Anderes et al., 2016] the support of any p-Wasserstein barycenter is contained in

$$\mathcal{C}_W(J, p) = \left\{ y \in \mathcal{Y} \mid y = T^{J,p}(x_1, \dots, x_J), x_i \in \operatorname{supp}(\mu_i) \right\}. \tag{11}$$

Compared to the p-Wasserstein barycenter of the probability measures $\mu_1, \dots, \mu_J$ the restricted centroid set $\mathcal{C}_{KR}(J, p, C)$ allows more flexibility for specific cases and can provide a more reasonable representation of the data. We illustrate this in Figure 5 (bottom-left/right) where the $(2, C)$-barycenter clearly represents all clusters while the 2-Wasserstein barycenter fails to capture them. Nevertheless, if C is large enough and all measures have equal total mass both barycenters coincide.

Corollary 2.6. *If $C > 2^{\frac{1}{p}} diam(\mathcal{Z})$ and $\mathbb{M}(\mu_1) = \mathbb{M}(\mu_2) = \dots = \mathbb{M}(\mu_J)$, then any p-Wasserstein barycenter is also a (p, C)-barycenter and vice versa.*

While this shows that the (p, C)-barycenter is a strict generalisation of the usual p-Wasserstein barycenter as the solutions coincide for large C, for smaller values of C there

[5] The degree of a vertex in a graph is the number of vertices which are adjacent to it.

can be significant differences. One such striking difference between the p-Wasserstein barycenter and the (p, C)-barycenter comes in the form of a localization property. Let $B_1, \dots, B_R \subset \mathcal{Y}$ such that $\operatorname{supp}(\mu_i) \subset \cup_{r=1}^R B_r$ with $\operatorname{diam}(B_r) \leq C$ for all $r = 1, \dots, R$ and $d(B_k, B_l) > 2^{1/p} C$ for all $k \neq l$. Here, the $(p, C)-$barycenter tends to place mass between the clusters $B_1, \dots, B_R$. However, a (p, C)-barycenter is obtained by combining R barycenters of the measures restricted to the $B_1, \dots, B_R$, respectively.

Lemma 2.7. *Let $\mu_1, \dots, \mu_J \in \mathcal{M}_+(\mathcal{X})$ such that for all $i = 1, \dots, J$ it holds $\operatorname{supp}(\mu_i) \subset \cup_{r=1}^R B_r$ for some $B_1, \dots, B_R \subset \mathcal{Y}$ with $\operatorname{diam}(B_r) \leq C$ for all $r = 1, \dots, R$ and $d(B_k, B_l) > 2^{1/p} C$ for all $k \neq l$. For $r = 1, \dots, R$, let*

$$\mu_r^* \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathit{conv}(B_r))} \frac{1}{J} \sum_{i=1}^{J} KR^p_{p,C}(\mu, \mu_{i|B_r}),$$

where $\mathit{conv}(B_r)$ is the convex hull of B_r for $r = 1, \dots, R$. Then, the measure $\sum_{r=1}^{R} \mu_r^*$ is a (p, C)-barycenter of $\mu_1, \dots, \mu_J$.

In particular, Lemma 2.7 implies that the (p, C)-barycenter respects the cluster structure within the supports of the measures if the clustered are sufficiently separated and C is adapted according to the cluster size. Examples of this setting can be seen in Figure 2 and Figure 5.

3 A Lift to Optimal Transport, Wasserstein Barycenters and Multi-Marginal Optimal Transport

In this section, we provide the necessary tools and framework to establish our results in the previous section. Following the ideas of Guittet [2002] we state UOT in (2) as an equivalent balanced OT problem. We extend this idea to the (p, C)-barycenter, showing it to be equivalent to a specific Wasserstein barycenter problem as well as a balanced multi-marginal optimal transport problem.

3.1 A Lift to Optimal Transport

We fix a parameter $C > 0$, introduce an additional dummy point $\mathfrak{d}$ and define the augmented space $\tilde{\mathcal{X}} := \mathcal{X} \cup \{\mathfrak{d}\}$ with metric cost

$$\tilde{d}^p_C(x, x') = \begin{cases} d^p(x, x') \wedge C^p, & x, x' \in \mathcal{X}, \\ \frac{C^p}{2}, & x \in \mathcal{X},\ x' = \mathfrak{d}, \\ \frac{C^p}{2}, & x = \mathfrak{d},\ x' \in \mathcal{X}, \\ 0, & x = x' = \mathfrak{d}. \end{cases} \tag{12}$$

Notably, $\tilde{d}_C \colon \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \mathbb{R}_+$ defines a metric on $\tilde{\mathcal{X}}$ [Müller et al., 2020, Lemma A1]. Consider the subset $\mathcal{M}_+^B(\mathcal{X}) := \{\mu \in \mathcal{M}_+(\mathcal{X}) \mid \mathbb{M}(\mu) \leq B\} \subset \mathcal{M}_+(\mathcal{X})$ of non-negative measures whose total mass is bounded by B. Setting $\tilde{\mu} := \mu + (B - \mathbb{M}(\mu))\delta_{\mathfrak{d}}$, any measure $\mu \in \mathcal{M}_+^B(\mathcal{X})$ defines an *augmented measure* $\tilde{\mu}$ on $\tilde{\mathcal{X}}$ such that $\mathbb{M}(\tilde{\mu}) = B$. Hence, for two measures $\mu, \nu \in \mathcal{M}_+^B(\mathcal{X})$ we can define the OT problem on $\tilde{\mathcal{X}}$ between their augmented measures $\tilde{\mathrm{OT}}_{\tilde{d}^p_C}(\tilde{\mu}, \tilde{\nu})$. In fact, it holds that

$$\mathrm{UOT}_{p,C}(\mu, \nu) = \mathrm{UOT}_{d^p \wedge C^p, C}(\mu, \nu) = \tilde{\mathrm{OT}}_{\tilde{d}^p_C}(\tilde{\mu}, \tilde{\nu}),$$

where the first equality follows by Lemma 2.1 as for any optimal solution π_C it holds $\pi_C(x, x') = 0$ if $d^p(x, x') > C^p$ and the second follows by [Guittet, 2002, Lemma 3.1]. The same equalities remain valid replacing B by an arbitrarily large constant as summarized by the following lemma.

Lemma 3.1. *Consider $\mu, \nu \in \mathcal{M}_+^B(\mathcal{X})$ with extended versions $\tilde{\mu}, \tilde{\nu}$. Then for any $a > 0$ it holds that*

$$\tilde{OT}_{\tilde{d}_C^p}(\tilde{\mu}, \tilde{\nu}) = \tilde{OT}_{\tilde{d}_C^p}(\tilde{\mu} + a\delta_{\mathfrak{d}}, \tilde{\nu} + a\delta_{\mathfrak{d}}).$$

Proof. For $p = 1$, the result is trivial since by duality $\tilde{\mathrm{OT}}_{\tilde{d}_C}(\tilde{\mu}, \tilde{\nu})$ only depends on the difference of the measures. For $p > 1$ we invoke $\tilde{d}_C$-cyclical monotonicity [Villani, 2008, Thm. 5.10] of any OT plan π and use the property that $\tilde{d}_C^p(x, \mathfrak{d}) = C^p/2$. This yields that $(\mathfrak{d}, \mathfrak{d}) \in \mathrm{supp}(\pi)$ which leads to the desired conclusion. □

3.2 A Lift to Wasserstein Barycenters

We can also lift the optimization problem defining a (p, C)-barycenter to an equivalent p-Wasserstein barycenter formulation (4). Augmentation of the underlying measures, however, is not straightforward as the total mass of the (p, C)-barycenter is unknown. A first crude upper bound on its total mass leads to a feasible approach.

Lemma 3.2. *Consider $\mu_1, \dots, \mu_J \in \mathcal{M}_+(\mathcal{X})$ and let $F_{p,C}$ be their associated unbalanced Fréchet functional. Then it holds that*

$$\operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu) = \operatorname*{arg\,min}_{\substack{\mu \in \mathcal{M}_+(\mathcal{Y}) \\ \mathbb{M}(\mu) \le \sum_{i=1}^J \mathbb{M}(\mu_i)}} F_{p,C}(\mu).$$

More precisely, any (p, C)-barycenter $\mu^\star$ of $\mu_1, \dots, \mu_J$ satisfies $\mathbb{M}(\mu^\star) \le \sum_{i=1}^J \mathbb{M}(\mu_i)$.

Proof. Assume first that there exists a measure $\mu \in \mathcal{M}_+(\mathcal{Y})$ such that $\mu = \nu_1 + \nu_2$ where no transport between ν_2 and any μ_i occurs in the optimal solution of $\mathrm{UOT}_{p,C}(\mu, \mu_i)$ for $1 \le i \le J$ and it holds $\mathbb{M}(\nu_2) > 0$. Thus it holds

$$F_{p,C}(\mu) = F_{p,C}(\nu_1 + \nu_2) = F_{p,C}(\nu_1) + (C^p/2)\mathbb{M}(\nu_2) > F_{p,C}(\nu_1)$$

and we improve the objective value of μ by removing ν_2. Hence, let $\mu \in \mathcal{M}_+(\mathcal{Y})$ be any measure such that $\nu_2 \equiv 0$. Consider π_i the optimal solution for $\mathrm{UOT}_{p,C}(\mu, \mu_i)$ for each $1 \le i \le J$. Decompose the measure $\mu = \sum_{i=1}^J \tau_i$, where τ_i is the mass of μ transported to μ_i according to π_i and which is not yet included in any τ_j for $j < i$. Clearly, $\mathbb{M}(\mu) = \sum_{i=1}^J \mathbb{M}(\tau_i) \le \sum_{i=1}^J \mathbb{M}(\mu_i)$ and we conclude that

$$\min_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu) = \min_{\substack{\mu \in \mathcal{M}_+(\mathcal{Y}) \\ \mathbb{M}(\mu) \le \sum_{i=1}^J \mathbb{M}(\mu_i)}} F_{p,C}(\mu).$$

By our first considerations the claim follows. □

Given the upper bound on the total mass of any (p, C)-barycenter at our disposal we can formulate a lift of the (p, C)-barycenter problem to a related p-Wasserstein barycenter problem. For this, let $\tilde{\mathcal{Y}} := \mathcal{Y} \cup \{\mathfrak{d}\}$ endowed with the metric $\tilde{d}_C$ in (12) (replace $\mathcal{X}$ by $\mathcal{Y}$ and recall that $\mathcal{X} \subset \mathcal{Y}$) and augment the measures $\mu_1, \dots, \mu_J$ to $\tilde{\mu}_1, \dots, \tilde{\mu}_J$ where $\tilde{\mu}_i = \mu_i + \sum_{j \ne i} \mathbb{M}(\mu_j)\delta_{\mathfrak{d}}$ for $1 \le i \le J$. In particular, $\mathbb{M}(\tilde{\mu}_i) = \sum_{j=1}^J \mathbb{M}(\mu_j)$ and we can define the *augmented p-Fréchet functional*

$$\tilde{F}_{p,C}(\mu) := \frac{1}{J} \sum_{i=1}^J \tilde{\mathrm{OT}}_{\tilde{d}_C^p}^p(\tilde{\mu}_i, \mu),$$

where by definition $\tilde{F}_{p,C}$ is restricted to measures μ with mass $\mathbb{M}(\mu) = \sum_{i=1}^J \mathbb{M}(\mu_i)$.

Lemma 3.3. *For $1 \leq i \leq J$ consider measures $\mu_i \in \mathcal{M}_+(\mathcal{X})$ and their augmented versions $\tilde{\mu}_i \coloneqq \mu_i + \sum_{j \neq i} \mathbb{M}(\mu_j)\delta_{\mathfrak{d}}$, respectively. Then it holds that*

$$F_{p,C}(\mu) = \tilde{F}_{p,C}\left(\mu + \left(\sum_{i=1}^{J} \mathbb{M}(\mu_i) - \mathbb{M}(\mu)\right)\delta_{\mathfrak{d}}\right)$$

for all $\mu \in \mathcal{M}_+(\mathcal{Y})$ such that $\mathbb{M}(\mu) \leq \sum_{i=1}^{J} \mathbb{M}(\mu_i)$ and in particular

$$\min_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu) = \min_{\substack{\mu \in \mathcal{M}_+(\tilde{\mathcal{Y}}) \\ \mathbb{M}(\mu) = \sum_{i=1}^{J} \mathbb{M}(\mu_i)}} \tilde{F}_{p,C}(\mu).$$

The proof of this Lemma is given in Appendix A.1.

Remark 3.4 (Optimal (p, C)-barycenters)**.** *Lemma 3.3 states that the optimal objective value for the (p, C)-barycenter is equal the related p-Wasserstein barycenter problem on the augmented space. In particular, the proof also reveals that if $\tilde{\mu}^\star$ is a p-Wasserstein barycenter for the augmented measures $\tilde{\mu}_1, \dots, \tilde{\mu}_J$ then $\mu^\star \coloneqq \tilde{\mu}^\star - \tilde{\mu}^\star(\mathfrak{d})\delta_{\mathfrak{d}}$ is a (p, C)-barycenter for the measures $\mu_1, \dots, \mu_J$. Vice versa, if $\mu^\star$ is a (p, C)-barycenter for the measures $\mu_1, \dots, \mu_J$ then $\tilde{\mu}^\star \coloneqq \mu^\star + \left(\sum_{i=1}^{J} \mathbb{M}(\mu_i) - \mathbb{M}(\mu^\star)\right)\delta_{\mathfrak{d}}$ is a p-Wasserstein barycenter for the augmented measures $\tilde{\mu}_1, \dots, \tilde{\mu}_J$.*

3.3 A Lift to Multi-Marginal Optimal Transport

On the augmented space $\tilde{\mathcal{Y}} \coloneqq \mathcal{Y} \cup \{\mathfrak{d}\}$ equipped with metric $\tilde{d}_C$ in (12), we define for $p \geq 1$ and $J \in \mathbb{N}$ a Borel barycenter application $\tilde{T}_C^{J,p} \colon \tilde{\mathcal{Y}}^J \to \tilde{\mathcal{Y}}$ that takes as input $(y_1, \dots, y_J) \in \tilde{\mathcal{Y}}$ and outputs any minimizer $y \in \tilde{\mathcal{Y}}$ of the function

$$f(y) = \sum_{i=1}^{J} \tilde{d}_C^p(y_i, y).$$

Of particular interest to us is the barycentric application restricted to inputs from $\tilde{\mathcal{X}}$. However, we collect some of its key properties for general input $(y_1, \dots, y_J) \in \tilde{\mathcal{Y}}^J$. For this, we define the index set

$$\mathcal{B}(y_1, \dots, y_J) \coloneqq \{i \mid y_i = \mathfrak{d},\, 1 \leq i \leq J\}.$$

If clear from the context, then the dependence on $y_1, \dots, y_J$ is suppressed and the set is simply denoted as $\mathcal{B}$.

Lemma 3.5. *Fix some parameter $C > 0$ and consider the space $\tilde{\mathcal{Y}}$ with metric $\tilde{d}_C$ as defined in (12). For points $(y_1, \dots, y_J) \in \tilde{\mathcal{Y}}^J$ it holds that*

(i) $\tilde{T}_C^{J,p}(y_1, \dots, y_J) = \mathfrak{d}$ if and only if $\sum_{i \notin \mathcal{B}} \tilde{d}_C^p(y_i, y) \geq (J - 2|\mathcal{B}|)C^p/2$ for any $y \in \tilde{\mathcal{Y}}$. In particular, if strict inequality holds then $\tilde{T}_C^{J,p}(y_1, \dots, y_J) = \mathfrak{d}$ is unique.

(ii) If $2|\mathcal{B}| \geq J$ then it holds $\tilde{T}^{J,p}(y_1, \dots, y_J) = \mathfrak{d}$ with uniqueness if $2|\mathcal{B}| > J$.

(iii) If $\tilde{T}^{J,p}(y_1, \dots, y_J) \neq \mathfrak{d}$ then it holds

$$\tilde{T}_C^{J,p}(y_1, \dots, y_J) = \arg\min_{y \in \mathcal{Y}} \sum_{i \notin \mathcal{B}} \tilde{d}_C^p(y_i, y).$$

(iv) If $C > 2^{\frac{1}{p}} diam(\mathcal{Y})$, then for any points $y_1, \dots, y_J \in \mathcal{Y}$ with $|\mathcal{B}| = 0$ it holds that $\tilde{T}_C^{J,p}(y_1, \dots, y_J) = T^{J,p}(y_1, \dots, y_J)$ where the latter one is defined with respect to the usual metric d^p on $\mathcal{Y}$.

A proof of this result is provided in Appendix A.1. Lemma 3.5 allows to characterize the centroid sets of the augmented measures $\tilde{\mu}_1, \dots, \tilde{\mu}_J$ defined as

$$\tilde{\mathcal{C}}_{KR}(J,p,C) := \Big\{ y \in \tilde{\mathcal{Y}} \mid y = \tilde{T}_C^{J,p}(x_1, \dots, x_J), x_i \in \mathrm{supp}(\tilde{\mu}_i); \\ d^p(y, x_i) \leq C^p \ \forall \ x_i \neq \mathfrak{d} \Big\}. \tag{13}$$

Remark 3.6. *We point out that computing $\tilde{T}_C^{J,p}$ is in general a difficult optimisation problem. While for squared euclidean distance, computing the barycentric application simply amounts to taking the mean of the x_i, even on the non-augmented space, there are no closed form solutions available for most choices of distances and values of p. This problem is exacerbated by the truncation of the distance $\tilde{d}$ at C^p [as also pointed out in Müller et al., 2020], since it implies that disregarding a certain subset of points and just computing the barycenter with respect to the remaining x_i might in fact be optimal. However, initially it is not clear which x_i to choose, turning this into a difficult combinatorial problem.*

Recall that for any measure μ its support is contained in $\mathcal{X}$ a subset of $\mathcal{Y}$. The augmented measure $\tilde{\mu}$ is extended by an additional support point at $\{\mathfrak{d}\}$. In particular, while the centroid set is a subset of $\tilde{\mathcal{Y}}$ it only depends on the support of the measures $\tilde{\mu}_i$ contained in $\tilde{\mathcal{X}} := \mathcal{X} \cup \{\mathfrak{d}\}$.

Corollary 3.7. *For the centroid sets of the augmented measures $\tilde{\mu}_i \in \mathcal{M}_+(\tilde{\mathcal{X}})$ with $1 \leq i \leq J$ it holds*

$$\tilde{\mathcal{C}}_{KR}(J,p,C) \subset \mathcal{C}_{KR}(J,p,C) \cup \{\mathfrak{d}\} \subset \mathcal{C}_{KR}(J,p) \cup \{\mathfrak{d}\}.$$

Proof. The first inclusion follows by statements (i) and (iii) in Lemma 3.5 and the observation that $|\mathcal{B}| = J - L$. The second by applying $\mathcal{C}_{KR}(J,p,C) \subset \mathcal{C}_{KR}(J,p)$. □

Remark 3.8. *One could define $\mathcal{C}_{KR}(J,p,C)$ in terms of $\tilde{d}_C$ instead of d to obtain equality in the first inclusion. Replacing $T^{L,p}$ by $\tilde{T}_C^{L,p}$ in the definition of the centroid set would not alter any of the related proofs and yield slightly sharper control on the support of (p,C)-barycenter. However, as we consider the given definition to be more intuitive, we omit this improvement in the statement of the theorem.*

Let $\Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_J)$ be the set of measures on $\tilde{\mathcal{Y}} \times \ldots \times \tilde{\mathcal{Y}}$ whose i-th marginal is equal to $\tilde{\mu}_i$ for all $1 \leq i \leq J$. We refer to the elements of this set as *multi-couplings* of $\tilde{\mu}_1, \dots, \tilde{\mu}_J$. For $p \geq 1$ define the *augmented multi-marginal transport problem* as

$$\min_{\pi \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_J)} \int_{\tilde{\mathcal{Y}}^J} c_{p,C}(y_1, \dots, y_J) \pi(dy_1, \dots, dy_J), \tag{14}$$

where

$$c_{p,C}(y_1, \dots, y_J) := \frac{1}{J} \sum_{i=1}^{J} \tilde{d}_C^p \left(y_i, \tilde{T}_C^{J,p}(y_1, \dots, y_J) \right).$$

The relation between the augmented multi-marginal transport formulation (14) and the (p,C)-barycenter is as follows.

Proposition 3.9. *Let $\mu_1, \dots, \mu_J \in \mathcal{M}_+(\mathcal{X})$ and $\tilde{\mu}_1, \dots, \tilde{\mu}_J$ be their augmented counterparts. If $\pi \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_J)$ is a solution to the augmented multi-marginal problem (14), then the measure $\mu^\star := (\tilde{T}_C^{J,p} \# \pi)_{|\mathcal{Y}} \in \mathcal{M}_+(\mathcal{Y})$ is a (p, C)-barycenter of the measures $\mu_1, \dots, \mu_J$, where $\tilde{T}_C^{J,p} \# \pi$ denotes the pushforward of π under $\tilde{T}_C^{J,p}$. Moreover, for every (p, C)-barycenter $\mu^\star$, there exists a solution π to the augmented multi-marginal transport problem, such that*

$$\mu^\star + \left(\sum_{i=1}^{J} \mathbb{M}(\mu_i) - \mathbb{M}(\mu^\star) \right) \delta_{\mathfrak{d}} = \tilde{T}_C^{J,p} \# \pi.$$

In particular, it holds that

$$\min_{\pi \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_J)} \int_{\tilde{\mathcal{Y}}^J} c_{p,C}(y_1, \dots, y_J) \pi(dy_1, \dots, dy_J) = \inf_{\mu \in \mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu).$$

The proof follows straightforwardly along the lines of related statements for the multi-marginal optimal transport problem (Le Gouic and Loubes, 2017, Theorem 8; Masarotto et al., 2019, Lemma 8 or Panaretos and Zemel, 2020, Proposition 3.1.2). This correspondence between the (p, C)-barycenter problem and a balanced multi-marginal optimal transport serves as one of the key components in the proof of Theorem 2.5.

4 Computational Issues and Numerical Experiments

We present approaches to compute the (p, C)-barycenter problem by solving related OT problems. Based on this, we investigate the performance of the Wasserstein and (p, C)-barycenters on multiple synthetic datasets. For reference, we also report on results for two related concepts of unbalanced barycenters (UBCs), namely the Gaussian-Hellinger-Kantorovich and Wasserstein-Fisher-Rao barycenter.

4.1 Algorithms

Theorem 2.5 and Proposition 3.9 both allow to pose the augmented problem (recall Section 3) as a linear program and using Lemma 3.3 one can obtain a solution to the original problem by solving the augmented one. Using any linear program solver this enables the direct computation of an exact solution of this problem. However, the number of variables in this approach scales as the size of $\mathcal{C}_{KR}(J, p, C)$ and hence it turns out to be infeasible already for relatively small instance sizes. To compute (p, C)-barycenters at larger scales we revisit iterative methods to solve the (balanced) Wasserstein barycenter problem and give instructions how to use modifications of them to compute (p, C)-barycenters. In particular, we detail a multi-scale method which solves successive fixed-support (p, C)-barycenter LPs on increasingly refined support sets. This provides a meta-framework to adjust state-of-the-art solvers for the Wasserstein barycenter for (p, C)-barycenter computations.
To construct the augmented problem we add the dummy point $\mathfrak{d}$ to the support of the μ_i's, while setting its distance to all other locations to be $C^p/2$. Note, that by Lemma 2.1 and Lemma 3.1 the truncation of $\tilde{d}$ at C^p can be omitted if $\mathbb{M}(\tilde{\mu}_i) > 3 \max_{i=1,\dots,J} \mathbb{M}(\mu_i)$. If this is not the case, we can enforce it by adding additional mass at $\mathfrak{d}$ in all augmented measures without changing the optimal value.

4.1.1 LP-Formulation for the (p, C)-Barycenter

Using property (i) from Theorem 2.5, we can rewrite the augmented (p, C)-barycenter problem as a linear program similarly to the usual p-Wasserstein barycenter problem (4). However, compared to the latter one, we replace the standard centroid set $\mathcal{C}_W(J, p)$ from (11), by the centroid set $\tilde{\mathcal{C}}_{KR}(J, p, C)$ of the augmented measures from (13). This yields

$$
\begin{aligned}
\min_{\pi^{(1)},\dots,\pi^{(J)},a} \quad & \frac{1}{J}\sum_{i=1}^{J}\sum_{j=1}^{|\tilde{\mathcal{C}}_{KR}(J,p,C)|}\sum_{k=1}^{M_i}\pi_{jk}^{(i)}c_{jk}^{i} \\
\text{s.t.} \quad & \sum_{k=1}^{M_i}\pi_{jk}^{(i)} = a_j, && \forall\ i=1,\dots,J, \forall j=1,\dots,|\tilde{\mathcal{C}}_{KR}(J,p,C)|, \\
& \sum_{j=1}^{|\tilde{\mathcal{C}}_{KR}(J,p,C)|}\pi_{jk}^{(i)} = b_k^i, && \forall\ i=1,\dots,J, \forall k=1,\dots,M_i, \\
& \pi_{jk}^{(i)} \geq 0 && \forall i=1,\dots,J, \forall j=1,\dots,|\tilde{\mathcal{C}}_{KR}(J,p,C)|, \\
& && \forall k=1,\dots,M_i,
\end{aligned}
$$

where $M_i = |\tilde{\mathcal{X}}_i|$ is the cardinality of the support of the augmented measure $\tilde{\mu}_i$. Here, c_{jk}^i denotes the distance between the j-th point of $|\tilde{\mathcal{C}}_{KR}(J, p, C)|$ and the k-th point in the support of $\tilde{mu}_i$, while b^i is the vector of masses corresponding to $\tilde{\mu}_i$. For practical purposes it may be advantageous to solve the multi-marginal problem instead of the (p, C)-barycenter problem. This changes the number of variables from $|\tilde{\mathcal{C}}_{KR}(J, p, C)|(1 + \sum_{i=1}^{J} M_i)$ to $\prod_{i=1}^{J} M_i$ and the number of constraints from $J|\tilde{\mathcal{C}}_{KR}(J, p, C)| + \sum_{i=1}^{J} M_i$ to $\sum_{i=1}^{J} M_i$. Depending on the value of C, and hence the cardinality of $\tilde{\mathcal{C}}_{KR}(J, p, C)$, it is possible to pick the problem with the smaller complexity.

While this formulation is appealing for proving theoretical statements as provided in Theorem 2.5, it quickly becomes computationally infeasible even for small scale problems as the number of variables in the LP grows potentially as $\prod M_i$. However, it still enables exact computations of (p, C)-barycenters for small scale examples, which is currently impossible for general UBCs. Though, while there has been some recent advancement for the 2-Wasserstein barycenter in special cases [Altschuler and Boix-Adsera, 2021] these LP-based algorithms ultimately do not scale to large instance sizes.

4.2 Iterative Algorithms and the Multi-Scale Approach

For the Wasserstein barycenter, iterative methods computing approximate barycenters, with a per iterations complexity only linear in the number of measures, enjoy great popularity. Most well known is the *fixed-support Wasserstein barycenter* [Ge et al., 2019, Lin et al., 2020, Xie et al., 2020] approach, aiming to find the best approximation of the barycenter on a pre-specified support set, for which a variety of methods is available. We utilise this fixed-support approach for the augmented (p, C)-barycenter problem by adding the dummy point $\mathfrak{d}$ to the given support and constructing the cost as described above. This yields a meta-framework which allows to employ fixed-support Wasserstein barycenter algorithms for fixed-support (p, C)-barycenter computation. One can also modify more general *free support* methods [Cuturi and Doucet, 2014, Ge et al., 2019, Luise et al., 2019], which usually alternate between updating the support set of the barycenter and its weights on this set, to provide approximate (p, C)-barycenters. However, the necessary position updates usually explicitly or implicitly rely on being able to compute the barycentric application $\tilde{T}^{J,p}$ efficiently. Recalling Remark 3.6, this is in general not tractable for the

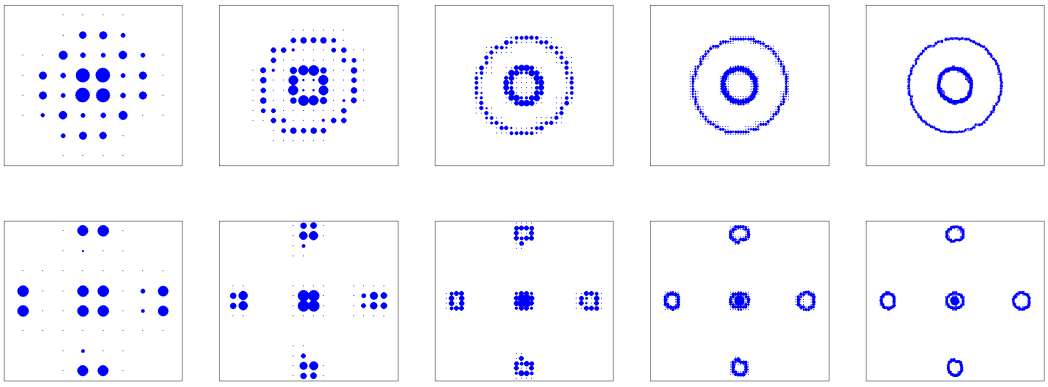


Figure 6: An illustration of the multi-scale approach on two different datasets. The fixed-support solutions are shown on grids of the sizes 8×8, 16×16, 32×32, 64×64 and 128×128 increasing from left to right. The corresponding run-times on a single core of an Intel Core $i7$ $12700K$ in the first/second row were 2.5/5 seconds, 14/16 seconds, 145/42 seconds, 13/3 minutes and 143/22 minutes. **Top:** The dataset of nested ellipses from Figure 7. **Bottom:** The dataset of ellipses with clustered support structure from Figure 2.

augmented problem, which severely hinders the use of these approaches. Thus, it is tempting to avoid these issues by approximating $\mathcal{Y}$ with a large finite space, i.e., by taking a grid of high-resolution, and solving the fixed support (p, C)-barycenter problem on this set. However, solving the fixed-support problem on this large space requires significant computational effort. We advocate an alternative by adapting the ideas of multi-scale methods for the Wasserstein distance/barycenter [Mérigot, 2011, Gerber and Maggioni, 2017, Schmitzer, 2019] to the (p, C)-barycenter setting. The idea of this approach is to start with a coarse version of the problem and then successively solve refined problems, while using the knowledge of the coarse solution to reduce the complexity of the finer ones. Thus, we initialise the support set of the barycenter as a fixed grid of size $K_1 \times \cdots \times K_d$ in $\mathbb{R}^d$. In the j-th step of the algorithm, after solving the fixed-support problem, we remove the grid points which have zero mass and replace the remaining ones with its 2^d closest points in a refined version of the original grid of size $2^j K_1 \times \cdots \times 2^j K_d$. This can be understood as solving the fixed-support problem on successively finer grids, while incorporating information provided by having already solved a coarser solution of the problem. We terminate the method once a pre-specified resolution has been reached. This allows to obtain fixed-support approximation of the (p, C)-barycenter on fine grids without having to optimise over the full support set.

We point out that this approach, while inspired by multi-scale approaches is more closely related to the formerly mentioned free-support methods. As such it does in general not yield a globally optimal fixed-support (p, C)-barycenter at the finest resolution. Instead it converges to a local minimum of the unbalanced Fréchet functional depending on the resolution of the initial grid. This is a common problem among alternating procedures for the free-support barycenter problem and can be attributed to the fact that the Fréchet functional is non-convex in the support locations of the measures. However, we stress that with this approach we observe reasonable approximations of the (p, C)-barycenter while avoiding the inherent problems of generalising usual position update procedures discussed above. In particular, we do not have to solve the $\tilde{T}_C^{J,p}$ barycenter problem at any point. Additionally, we note that the initial grid size should be chosen at least fine enough that

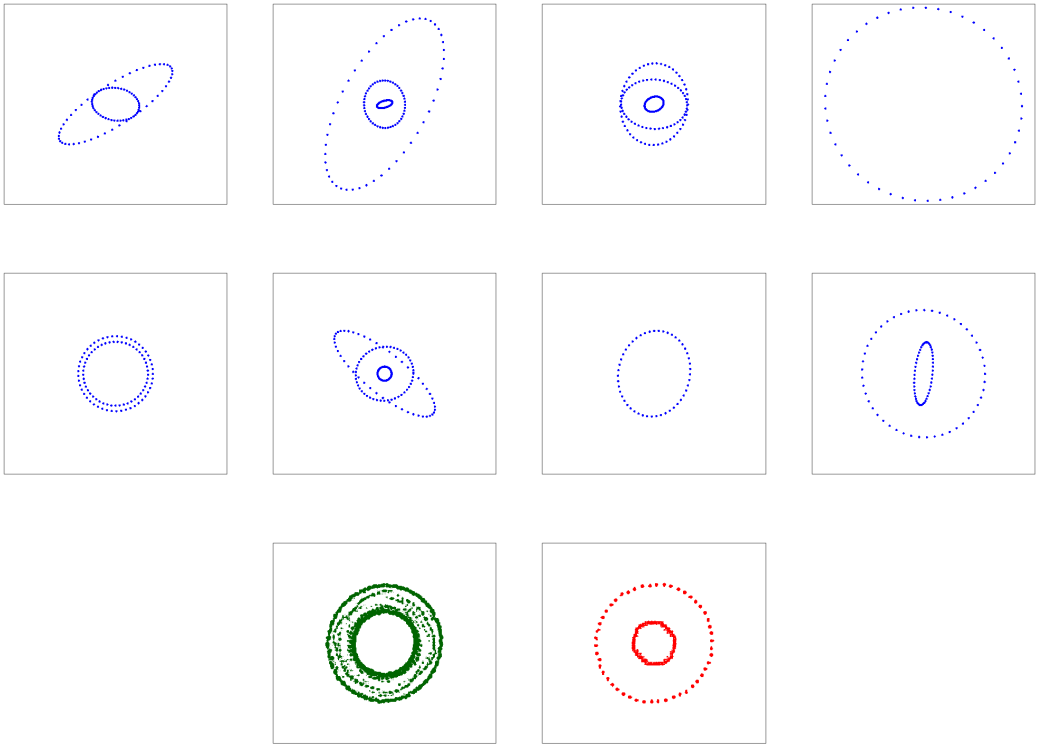


Figure 7: An excerpt of a dataset of $N = 100$ discretized ellipses. Each measure contains between 1 and 3 ellipses with equal probability. Each ellipse consists of 50 points with mass 1 in $[0, 1]^2$. **Left:** In darkgreen the 2-Wasserstein barycenter, where all measures are normalized to be probability measures. **Right:** In red, the $(2, 1.5)$-barycenter.

the distance between two adjacent grid points is smaller than C. Otherwise it is possible that support points lying between two grid points, having distance larger C to both, are not accounted for. For a visual illustration of the algorithm we refer to Figure 6.

4.3 Synthetic Data Simulations

We test the performance of the (p, C)-barycenter as a data analytic tool compared to the usual p-Wasserstein barycenter on a multitude of datasets. We base our computations on the MAAIPM method [Ge et al., 2019], which allows for high-precision approximations of barycenters up to moderate data sizes. The algorithm has been deployed to solve the fixed-support (p, C)-barycenter problems arising in the multi-scale method detailed above. Implementations of our used method and some alternatives can be found as part of the R-package *WSGeometry* (https://github.com/F-Heinemann/WSGeometry/).

Mismatched Shapes

This first set of examples mainly serves as starting point to illustrate improved performance of the (p, C)-barycenter compared to the p-Wasserstein barycenter. A prototypical benchmark for the p-Wasserstein barycenter are two nested ellipses as popularized in Cuturi and Doucet [2014]. For our example of nested ellipses, we assume that the support of each measure consists of nested ellipses, but the number of ellipses varies between the individual underlying measures. Specifically, we assume that for each μ_i the number of

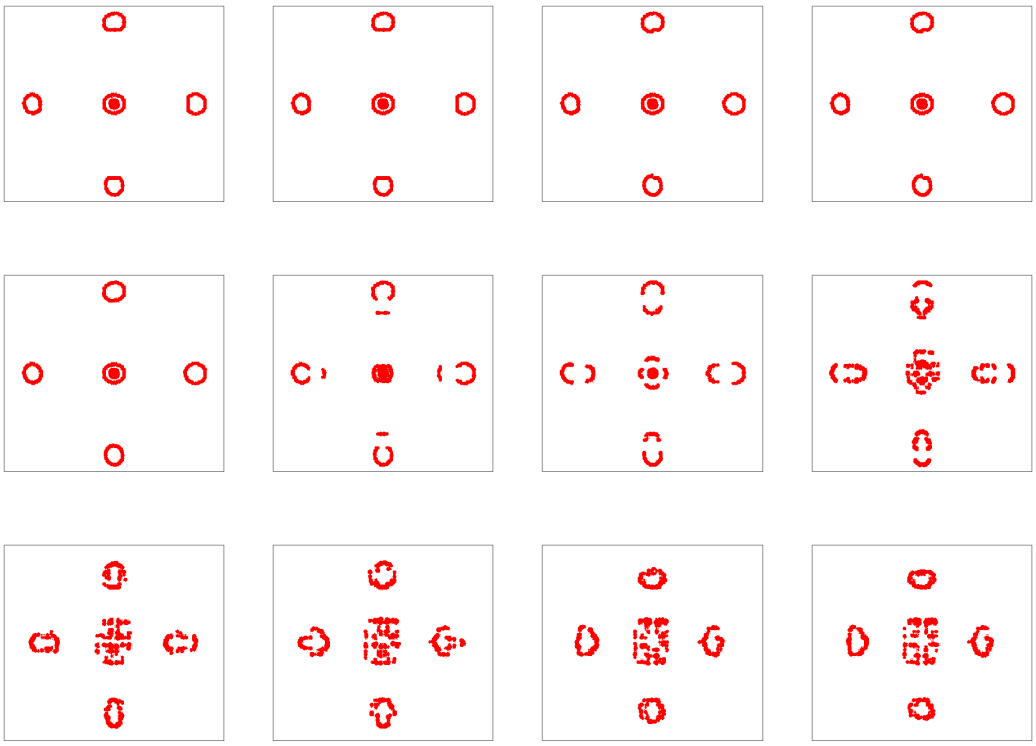


Figure 8: The $(2, C)$-barycenters for the measures in Figure 2 for different values of C. From top-left to bottom-right the values of C are equal to $0.1, 0.15, 0.2, 0.25, 0.275, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6$, respectively.

ellipses is uniformly random in $\{1, 2, 3\}$ and that each ellipse is discretised onto M support points with unit mass, respectively. This can be seen in Figure 7. We observe that while the p-Wasserstein barycenter recovers the elliptic shape of the underlying measures, it fails to produce distinct ellipses and instead produces something akin to a ring. In contrast, the (p, C)-barycenter yields two distinct ellipses, which coincides with the expected number of ellipses in one of the measures. This aligns well with intuition that the (p, C)-barycenter will simply disregard any additional structures which are not present in a sufficient amount of underlying measures. In contrast, the p-Wasserstein barycenter does not allow for this flexibility which enforces additional support points.

Local Scale Cluster Detection

Recall the setting of Figure 2. In the following class of examples, we are interested in datasets which possesses a natural cluster structure. Let $B_1, \dots, B_R \subset \mathbb{R}^D$ be convex, disjoint sets and assume that $\operatorname{supp}(\mu_i) \subset \cup_{r=1}^{R} B_r$ for all $i = 1, \dots, J$. If the diameter of all B_r is bounded from above by C and that the distance between each two B_r, B_s is at least $2^{1/p}C$, then Lemma 2.7 guarentees that the (p, C)-barycenter detects all of the R clusters in which at least $J/2$ measures have positive mass. In particular, by Theorem 2.5 (v) the (p, C)-barycenter will have mass in all of those clusters. Intuitively, this setting is reasonable if, for instance, it is already known that any interactions between support points of different measures are limited to scales below a certain threshold, which should then be chosen as C. The lower bound on the inter-cluster distance ensures that any pair of two clusters is well-separated, ensuring that it is always possible to distinguish between

two different clusters, as they can not be arbitrarily close to each other.
In Figure 2 the p-Wasserstein barycenter completely fails to capture the geometric data structure. Most of its mass is between the clusters and the outer clusters have nearly no mass. Moreover, the elliptic structure within each cluster is clearly not captured. In contrast, the (p, C)-barycenter not only captures all clusters, it also distinguishes between the difference in intensity (expected number of ellipses) in the clusters, matching the theoretical guarantees of Lemma 2.7. We stress that for this example the choice of C is of particular importance. If we choose C too large, the (p, C)-barycenter will fail to recover the data's support structure (for an illustration of the (p, C)-barycenter in this example over a range of values of C see Figure 8). Consequently, it is crucial to choose C appropriately. In this example, the barycenter appears to be stable and detect all clusters for $C \in [0.1, 0.275]$. Notably, if the locations of the clusters are already known, this setting also allows for parallel computations of the (p, C)-barycenter, where the problems are solved separately on each cluster and recombined at the end (Lemma 2.7).

Randomly distorted Measures

In a statistical context it is important to investigate the stability of the (p, C)-barycenter under random distortions. We fix a reference measure μ_0 on $\mathbb{R}^d$ and generate a set of measures by random modifications of μ_0. We then attempt to recover μ_0 by computing the p-Wasserstein and (p, C)-barycenter of these measures, respectively.
In the following, let $B(p)$ denote a Bernoulli random variable with mean p, $Poi(\lambda)$ a Poisson distribution with mean λ and $U[a, b]$ a uniform distribution on $[a, b]$. We generate $\mu_1, \dots, \mu_J$ as follows:
For $i = 1, \dots, J$ initialise $\mu_i = \mu_0$, then succesively modify μ_i based on the four following steps.

(i) **Point Deletion:** Fix $p_{del} \in [0, 1]$ and $\lambda_{del} \in \mathbb{R}_+$. We draw a $\text{Ber}(p_{del})$ random variable. If it takes the value 1, then we draw $D \sim Poi(\lambda_{del})$ and select $\min(D, |\text{supp}(\mu_0)|)$ points in the support of μ_0 uniformly by drawing without replacement. These points (and their mass) are not contained in μ_i, since they have been deleted.

(ii) **Point Addition:** We fix parameters $p_{add} \in [0, 1], \lambda_{add} \in \mathbb{R}_+, m_{add} \in \mathbb{R}^2, \sigma_{add} \in \mathbb{R}^{2\times 2}, u_0, u_1 \in \mathbb{R}$. Draw a $\text{Ber}(p_{add})$ random variable. If it takes the value 1, draw a $\text{Poi}(\lambda_{add})$ random variable α. Then, generate α random variables following a normal distribution with mean m_{add} and covariance matrix σ_{add}. Add these support points to μ_i, where the weight of each of these points is determined by independent $U[u_0, u_1]$ random variables.

(iii) **Position Change:** Fix parameters $a_1, a_2, b_1, b_2 \in \mathbb{R}$ with $a_1 \leq b_1$ and $a_2 \leq b_2$. For each x_0 in the support of μ_i, we draw a $U([a_1, b_1] \times [a_2, b_2])$ random variable and shift the position of x_0 by it.

(iv) **Weight Change:** Fix parameters $l, u \in \mathbb{R}$ with $l \leq u$. For each support point x_0 of μ_0 with weight w_0, we draw a $U[l, u]$ random variable U and change the weight of x_0 in μ_i to be $w_0 + U$.

An example of this setting can be seen in Figure 9. Comparing the two barycenters displayed there to the original measure reveals that, while the rough shape of the 2-Wasserstein barycenter is correct, its mass is spread out over a larger area and it has a significantly larger number of support points. Since all measures have been normalised, we have also lost all information on the mass of μ_0. Contrary to that, the (p, C)-barycenter retrieves the original measures recovering the location and number of the of support points

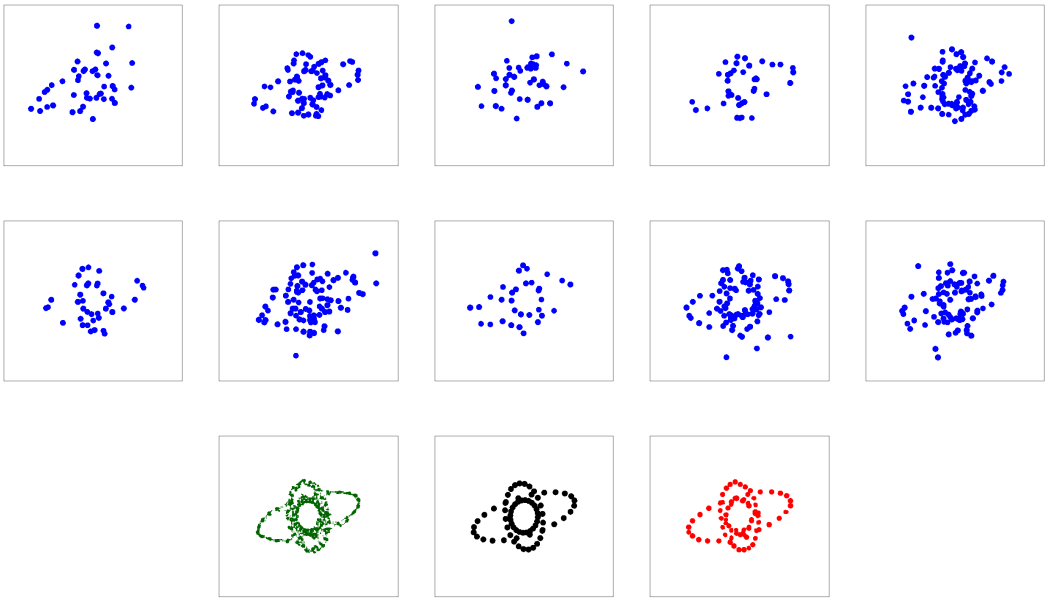


Figure 9: An excerpt from a dataset of $N = 100$ noisy nested ellipses supported in $[0,1]^2$. The parameters are $p_{del} = 1/3$, $\lambda_{del} = 75$, $p_{add} = 1/3$, $\lambda_{add} = 25$, $m_{add} = (0.5, 0.5)^T$, $\sigma_{add} = 0.15I_2$, $u_0 = 0.9$, $u_1 = 1.1$, $a_1 = a_2 = -0.025$, $b_1 = b_2 = 0.025$, $l = u = 0.1$. **Left:** The 2-Wasserstein barycenter (dark green). **Center:** The original measure μ_0 (black). **Right:** The $(2, 1.5)$-barycenter (red).

closely. Additionally, it also has a mass which only deviates from the original mass by about 0.23%. If one is only interested in recovering the general shape of the data, both approaches provide comparable performance. However, if the measures total mass and more detailed support structure are of importance the (p, C)-barycenter appears to be preferable.

Total Mass Intensity

While the p-Wasserstein barycenter of J probability measures has mass one, the mass of the (p, C)-barycenter depends on C as well as the geometry of the measures $\mu_1, \dots, \mu_J \in \mathcal{M}_+(\mathcal{X})$. Exact values for the mass of a (p, C)-barycenter without detailed computations, are only available in the limiting scenarios where C is extremely small or large relative to the other distances in $\mathcal{X}$. For the former, we know by Theorem 2.5 (v) that the barycenter has mass zero for disjoint measures and for the latter, Theorem 2.5 (vi) yields that there exists a (p, C)-barycenter with total mass intensity equal to the median of $\mathbb{M}(\mu_1), \dots, \mathbb{M}(\mu_J)$. For intermediate values of C, Theorem 2.5 (i) yields the upper bound by $2J^{-1}\sum_{i=1}^{J}\mathbb{M}(\mu_i)$. To highlight some possible behaviours of the total mass intensity of (p, C)-barycenter we consider three specific examples in Figure 10. We note that in all three cases at about $C = 0.6$ the mass of the barycenters is at the median of their respective $\mu_1, \dots, \mu_J$ and does no longer change with increasing C. This is significantly smaller than the requirement in Theorem 2.5 (vi), which underlines the fact that while in the worst case, this lower bound is sharp, in many examples the total mass of the (p, C)-barycenter stabilises significantly earlier. Moreover, none of the three curves is monotone. Instead the total mass of the barycenter is increasing up to a certain point, after which it decreases until it reaches the median of the masses. This makes intuitive sense, as the measures are disjoint, thus for small C the barycenter is empty and starts

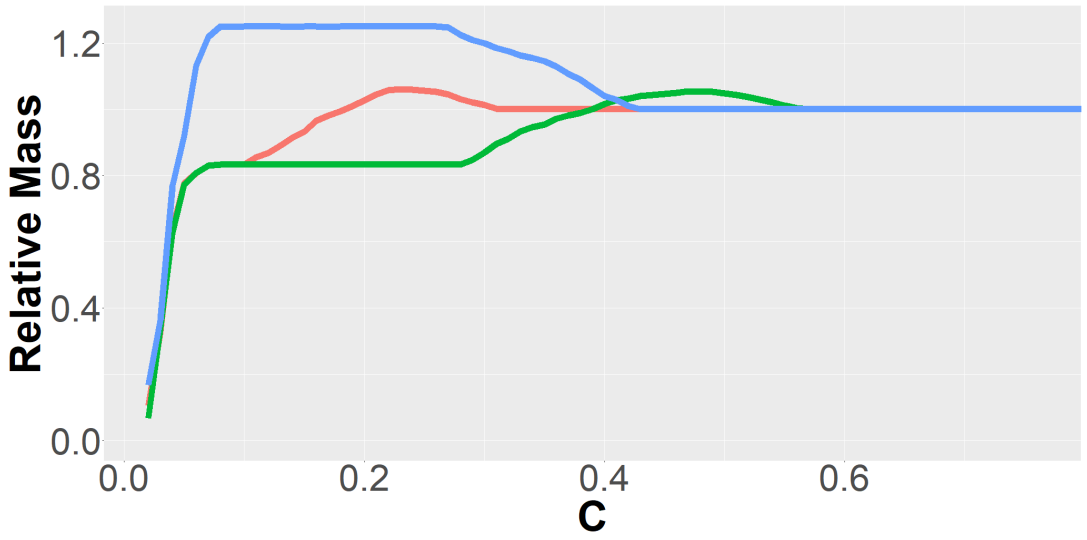


Figure 10: The mass of a (p, C)-barycenter for three sets of measures relative to the median of the total mass intensities of these measures. The green line corresponds to $J = 25$ measures from the same class as considered in Figure 2. The red line corresponds to the same measures where the four outer clusters have been moved closer to the central one, such that their distance has been halved. The blue line corresponds to $J = 5$ measures with the same cluster structure as in Figure 2, where the total number of ellipses in all clusters is fixed to be equal to four for all J measures.

to grow in mass quickly as the points within the clusters can be matched. In particular, the differences in intensity between clusters might lead to a total mass over the median $\mathbb{M}(\mu_1), \ldots, \mathbb{M}(\mu_J)$, as by Lemma 2.7 the total mass intensity of the (p, C)-barycenter is $\sum_{r=1}^{R} med(\mathbb{M}(\mu_{1_{|B_r}}), \ldots, \mathbb{M}(\mu_{J_{|B_r}})$, where $B_1, \ldots, B_5$ denote the respective cluster locations. For larger C these clusters start to merge and support points between the clusters reduce the total mass. In particular, these points can be seen clearly in the plot. Up until about $C = 0.1$, which is the cluster size, the mass of the barycenters rises sharply, before stabilising until the intercluster distance is reached. This is about 0.3 for the green and blue lines and about 0.15 for the red line (since the measures in this example are generated by halving the intercluster distance from the green one). This behaviour highlights the sensitivity of the mass of the (p, C)-barycenter to the geometry of the measures. It is therefore impossible to infer the total mass of the (p, C)-barycenter from the magnitude of C alone without accounting for the specific measures. However, analysing the structural properties of the support sets of the measures might provide a good indication at what values of C changes in drastic behaviour of the total mass are to be expected.

4.4 Comparison with Related Unbalanced Barycenter Concepts

We compare the (p, C)-barycenter with two alternative UBC approaches.

The Gaussian-Hellinger-Kantorovich Barycenter: This example falls in the general framework of optimal entropy transport problems. Measuring deviation between a feasible solution and the input marginals is carried out via the *Kullback-Leibler divergence* defined

for $\mu \ll \nu$ [6] as

$$KL(\mu, \nu) = \sum_{x \in X} \mu(x) \log \left(\frac{\mu(x)}{\nu(x)} \right).$$

If $\mu \not\ll \nu$ the value of KL is set to be $+\infty$. For a parameter $\lambda > 0$, the *Gaussian-Hellinger-Kantorovich Distance* [Liero et al., 2018] is defined as

$$GHK_\lambda(\mu, \nu) = \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X})} \sum_{x, x' \in \mathcal{X}} d^2(x, x')\pi(x, x') + \lambda KL(\mu, \pi_1) + \lambda KL(\nu, \pi_2),$$

where π_1 and π_2 denote the respective marginals of π. The GHK_λ barycenter is defined as

$$\arg\min_{\mu \in \mathcal{M}_+(\mathcal{Y})} \sum_{\mathrm{i}=1}^{\mathrm{N}} GHK_\lambda(\mu_i, \mu).$$

The Hellinger-Kantorovich Barycenter: The Hellinger-Kantorovich distance, also known as Wasserstein-Fisher-Rao distance [Liero et al., 2018, Chizat et al., 2018a], is closely related to the Gaussian-Hellinger-Kantorovich distance. For fixed parameter $\sigma \in (0, \pi/2]$, referred to as the *cut-locus*, it is defined as

$$\begin{aligned} HK_\sigma(\mu, \nu) = \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X})} & \sum_{x, x' \in \mathcal{X}} (-\log(\cos^2_\sigma(d(x, y)))\pi(x, x') \\ & + KL(\mu, \pi_1) + KL(\nu, \pi_2), \end{aligned}$$

where $cos_\sigma : z \mapsto \cos(\min(z, \sigma))$. For a fixed cut-off locus σ, the HK_σ barycenter is defined as

$$\arg\min_{\mu \in \mathcal{M}_+(\mathcal{Y})} \sum_{\mathrm{i}=1}^{\mathrm{N}} HK_\sigma(\mu_i, \mu).$$

Comparing the barycenters: As the resulting barycenters vary significantly in all three cases, depending on the parameters C, λ, σ, we compare their behaviour upon change of parameter. As a simple example, we consider four measures supported on subsets of a grid on $[0, 1]^2$, displayed in Figure 11. To ensure fair comparison, we deploy the same method based on the general scaling method [Chizat et al., 2018b] to approximate the UBC in all three cases. However, we point out that this implies disregarding the ambient space and instead taking the minimum over all positive measures supported on a prespecified grid in $[0, 1]^2$.

For high parameter values all three approaches yield similar results. This is, of course, to be expected, since these distances interpolate between p-Wasserstein distance and total variation/Kullback-Leibler distance and large parameters correspond to a setting being close to the Wasserstein distance. The KR barycenter has mass zero for small choice of C by Theorem 2.5 (iv), since the four measures have disjoint support. After reaching a threshold of $C \approx 0.1$, the mass in the $(2, C)$-barycenter starts to increase as mass is added in the center of the unit square until at $C \approx 0.3$ the mass of an individual data measure is reached.

For small λ the GHK_λ barycenter has small mass and its support is close to that of a linear mean of the four measures, though the total mass intensity is significantly lower than for the original measures. With increasing λ the mass starts to increase and to smear into the middle of the unit square, until a large square, encompassing all four data supports, is formed. After this point increasing λ causes the square to contract while its mass

[6] A measure $\mu \in \mathcal{M}_+(\mathcal{X})$ is said to be *absolutely continuous* (denoted $\mu \ll \nu$) with respect to another measure $\nu \in \mathcal{M}_+(\mathcal{X})$ if $\nu(A) = 0$ implies $\mu(A) = 0$ for any measurable set A.

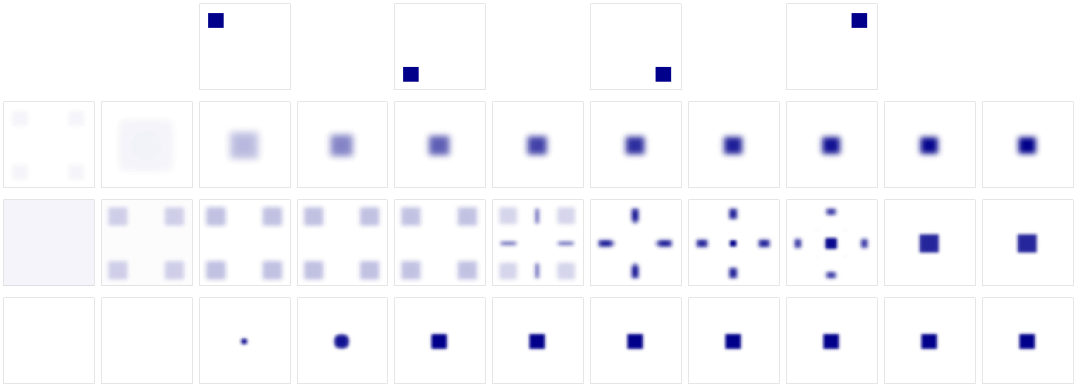


Figure 11: Comparison of the three unbalanced barycenters when varying their parameter. All measures are supported on an equidistant 64×64 grid in $[0,1]^2$. **First row:** The four underlying measures. **Second row:** The Gaussian-Hellinger-Kantorovich barycenter for $\lambda = 0.01, 0.15, \ldots, 1.83, 1.97$. **Third row:** The Hellinger-Kantorovich barycenter for $\sigma = 0.01, 0.08, \ldots, 0.64, 0.71$. **Fourth row:** The KR barycenter for $C = 0.01, 0.08, \ldots, 0.64, 0.71$.

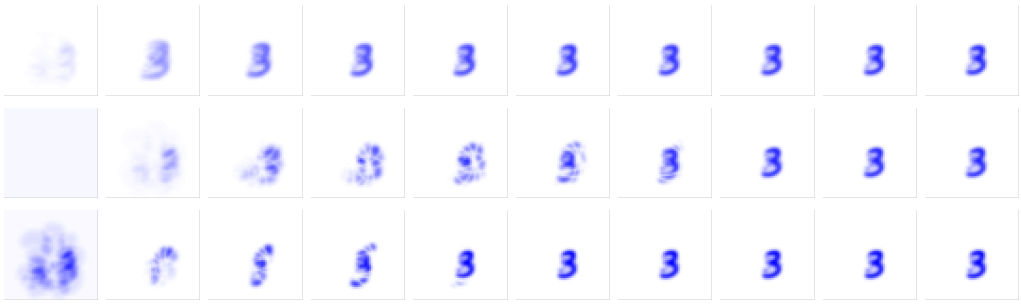
Figure 12: Images displaying the underlying measures used for barycenter computation in Figure 13. Each row corresponds to a dataset of ten elements of the classical MNIST dataset which have been randomly rescaled and shifted within a 50×50 grid in $[0,1]^2$. Their total mass intensities have not been normalised.

increases. Finally, we approach a single square at roughly the same size as the squares in the underlying measures for large λ.

The HK_σ barycenter is close to a linear mean of the four measures for small cut-off. Increasing σ initially reduces the mass at each of the square locations. At a threshold of $\sigma \approx 0.34$, we observe a change, where part of the mass is moved vertically or horizontally to the mid points between the squares in a rectangular shape. Until $\sigma \approx 0.43$ all mass is shifted to these "middle-rectangles", at which point a second shift occurs, where the mass from these rectangles starts to move towards a square in the center. At $\sigma \approx 0.6$, all mass has been shifted towards a square in the center and there is no further change in the HK barycenter, when increasing σ.

Additionally, we consider Figure 13, where the three unbalanced barycenter models are compared on three exemplary classes based on the MNIST dataset. Here, the original 28×28 images have been rescaled to sizes between 14×14 and 42×42 and embedded in a random subgrid of a 50×50 image. In this setting, there is a notable distinction between the GHK barycenter and the KR and HK barycenters. While for the former, the overall shape is recovered even for small parameter values, the latter two barycenters produce unstructured results for small parameters. The GHK distance is not constructed

(a)



(b)

(c)

Figure 13: Comparison of the three unbalanced barycenters when varying their parameter. The set of underlying measures for (a) is the first row of Figure 12. For (b) it is the second for (c) the third. For each class of examples the three different UOT barycenter models are considered in different rows: **First row:** The Gaussian-Hellinger-Kantorovich barycenter for $\lambda = 0.01, 0.12, 0.23, \ldots, 1$. **Second row:** The Hellinger-Kantorovich barycenter for $\sigma = 0.01, 0.08, \ldots, 0.64$. **Third row:** The KR barycenter for $C = 0.1, 0.2, \ldots, 1$.

to have a maximal transport distance comparable to the impact of C or σ in the other two cases, which allows to transport across larger distance and recover the correct shape for smaller values of λ. However, the mass of the GHK barycenter is significantly smaller than that of the original measures for small values of λ and only increases to the correct magnitude for larger penalty values. The HK and KR barycenters consist of fragments of the final shape which move towards a joint location for increasing parameters. For large penalties all three models are nearly identical and display the corresponding number correctly. This makes sense, as in this setting the minimisation in any individual term of the (p, C)-Fréchet functional is driven by minimising an OT term. We point out that for the (p, C)-barycenter this regime is guaranteed to be reached by choosing C larger than the diameter of the space, while for the other two models the suitable parameter choice for this example is ambiguous without actually computing the result for specific values.

Overall, for large parameter values all considered UBCs perform similarly. In small parameter regimes we observe significant differences. This difference in behavior is to be expected as the dependence of the UOT models on their parameters varies significantly. One key advantage of the KR barycenter is that its connection between the choice of C and the properties of the resulting barycenter is immediate and intuitive. While the cut-off locus σ for the HK barycenter fulfils a similar role, imposing control at the maximum scale at which transport does occur, the consequences of changing σ from one value to another are far less immediate due to the involved structure of the cost functional in this setting. Similarly to the KR barycenter, it is worth noticing that the HK barycenter does allow for mass at locations given by centroids of support points of $L < N$ measures. Though, while for the KRD a feature of the underlying measures is only contained in the barycenter if it is present in more than $L = N/2$ measures, the HK barycenter also allows for mass at locations constructed from less support points. Thus, the HK barycenter is prone to being more susceptible to errors due to noise within the data. Compared to the other two choices, the parameter λ of the GHK barycenter does appear to have less interpretation, with the only clear connection being that increasing λ increases the mass of the GHK barycenter. There does also not appear to be any well-founded method how to approach the choice of λ for a given dataset.

Acknowledgments

F. Heinemann and M. Klatt gratefully acknowledge support from the DFG Research Training Group 2088 *Discovering structure in complex data: Statistics meets optimization and inverse problems.* A. Munk gratefully acknowledges support from the DFG CRC 1456 *Mathematics of the Experiment A04, C06* and the Cluster of Excellence 2067 MBExC *Multiscale bioimaging–from molecular machines to networks of excitable cells.* We kindly thank two anonymous referees for their helpful comments and suggestions.

References

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

J. M. Altschuler and E. Boix-Adsera. Wasserstein barycenters can be computed in polynomial time in fixed dimension. *Journal of Machine Learning Research*, 22:44–1, 2021.

P. C. Álvarez-Esteban, E. Del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2):389–409, 2016.

J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

V. I. Bogachev. *Measure Theory*, volume 1. Springer Science & Business Media, 2007.

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

N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71–1, 2016.

L. A. Caffarelli and R. J. McCann. Free boundaries in optimal transport and Monge-Ampere obstacle problems. *Annals of Mathematics*, pages 673–730, 2010.

G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49 (6):1621–1642, 2015.

G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2): 1385–1418, 2017.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018a.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018b.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018c.

N.-P. Chung and M.-N. Phung. Barycenters in the Hellinger–Kantorovich space. *Applied Mathematics & Optimization*, pages 1–30, 2020.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693. PMLR, 2014.

S. N. Evans and F. A. Matsen. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

A. Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310, 1948.

G. Friesecke, D. Matthes, and B. Schmitzer. Barycenters for the Hellinger–Kantorovich distance over $\mathbb{R}^d$. *SIAM Journal on Mathematical Analysis*, 53(1):62–110, 2021.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems*, 28:2053–2061, 2015.

W. Gangbo and R. J. McCann. Shape recognition via Wasserstein distance. *Quarterly of Applied Mathematics*, pages 705–737, 2000.

W. Gangbo, W. Li, S. Osher, and M. Puthawala. Unnormalized optimal transport. *Journal of Computational Physics*, 399:108940, 2019.

A. Gavryushkin and A. J. Drummond. The space of ultrametric phylogenetic trees. *Journal of theoretical biology*, 403:197–208, 2016.

D. Ge, H. Wang, Z. Xiong, and Y. Ye. Interior-point methods strike back: Solving the wasserstein barycenter problem. *Advances in Neural Information Processing Systems*, 32, 2019.

M. Gellert, M. F. Hossain, F. J. F. Berens, L. W. Bruhn, C. Urbainsky, V. Liebscher, and C. H. Lillig. Substrate specificity of thioredoxins and glutaredoxins–towards a functional classification. *Heliyon*, 5(12):e02943, 2019.

S. Gerber and M. Maggioni. Multiscale strategies for computing optimal transport. *Journal of Machine Learning Research*, 18:1–32, 2017.

A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2. Springer Science & Business Media, 2012.

K. Guittet. Extended Kantorovich norms: A tool for optimization. Technical report, Technical Report 4402, INRIA, 2002.

M. Hallin, G. Mordant, and J. Segers. Multivariate goodness-of-fit tests based on wasserstein distance. *Electronic Journal of Statistics*, 15(1):1328–1371, 2021.

L. G. Hanin. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

F. Heinemann, A. Munk, and Y. Zemel. Randomized wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM Journal on Mathematics of Data Science*, 4(1):229–259, 2022.

L. V. Kantorovich and S. Rubinstein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

J. Kitagawa and B. Pass. The multi-marginal optimal partial transport problem. In *Forum of Mathematics, Sigma*, volume 3. Cambridge University Press, 2015.

M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

B. R. Kloeckner. A geometric study of Wasserstein spaces: Ultrametrics. *Mathematika*, 61(1):162–178, 2015.

A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe. On the complexity of approximating Wasserstein barycenters. In *International Conference on Machine Learning*, pages 3530–3540. PMLR, 2019.

T. Le, M. Yamada, K. Fukumizu, and M. Cuturi. Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 12304–12315, 2019.

T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917, 2017.

J. Lellmann, D. A. Lorenz, C. Schonlieb, and T. Valkonen. Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

T. Lin, N. Ho, X. Chen, M. Cuturi, and M. Jordan. Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. *Advances in Neural Information Processing Systems*, 33, 2020.

D. G. Luenberger, Y. Ye, et al. *Linear and Nonlinear Programming*, volume 2. Springer, 1984.

G. Luise, S. Salzo, M. Pontil, and C. Ciliberto. Sinkhorn barycenters with free support via frank-wolfe algorithm. *Advances in neural information processing systems*, 32, 2019.

V. Masarotto, V. M. Panaretos, and Y. Zemel. Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *Sankhya A*, 81(1):172–213, 2019.

Q. Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.

E. F. Montesuma and F. M. N. Mboula. Wasserstein barycenter for multi-source domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16785–16793, 2021.

D. Mukherjee, A. Guha, J. M. Solomon, Y. Sun, and M. Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. PMLR, 2021.

R. Müller, D. Schuhmacher, and J. Mateu. Metrics and barycenters for point pattern data. *Statistics and Computing*, pages 1–20, 2020.

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

O. Pele and M. Werman. A linear time histogram metric for improved SIFT matching. In *European Conference on Computer Vision*, pages 495–508. Springer, 2008.

G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 55. Springer, 2015.

R. Sato, M. Yamada, and H. Kashima. Fast unbalanced optimal transport on a tree. *Advances in neural information processing systems*, 33:19039–19051, 2020.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

B. Schmitzer and B. Wirth. A framework for Wasserstein-1-type metrics. *Journal of Convex Analysis*, 26(2):353–396, 2019.

J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society Series B*, 80(1):219–238, 2018.

M. Sommerfeld, J. Schrieber, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20 (105):1–23, 2019.

Z. Su, W. Zeng, Y. Wang, Z.-L. Lu, and X. Gu. Shape classification using Wasserstein distance for brain morphometry analysis. In *International Conference on Information Processing in Medical Imaging*, pages 411–423. Springer, 2015.

C. Tameling, S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk. Colocalization for super-resolution microscopy via optimal transport. *Nature Computational Science*, 1 (3):199–211, 2021.

C. Villani. *Topics in Optimal Transportation*. Number 58. American Mathematical Soc., 2003.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

S. Wang and M. Yuan. Revisiting colocalization via optimal transport. *Nature Computational Science*, 1(3):177–178, 2021.

L. A. Wolsey and G. L. Nemhauser. *Integer and Combinatorial Optimization*, volume 55. John Wiley & Sons, 1999.

Y. Xie, X. Wang, R. Wang, and H. Zha. A fast proximal point method for computing exact Wasserstein distance. In *Uncertainty in Artificial Intelligence*, pages 433–453. PMLR, 2020.

J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.

A Proofs

A.1 Proofs of Section 3

Proof of Lemma 3.3. Let $\mu \in \mathcal{M}_+(\mathcal{Y})$ be such that $\mathbb{M}(\mu) \leq \sum_{i=1}^J \mathbb{M}(\mu_i)$. Then

$$\begin{aligned} F_{p,C}(\mu) &\overset{(i)}{=} \frac{1}{J}\sum_{i=1}^J \tilde{\mathrm{OT}}^p_{\tilde{d}^p_C}\left(\mu + \mathbb{M}(\mu_i)\delta_{\mathfrak{d}}, \mu_i + \mathbb{M}(\mu)\delta_{\mathfrak{d}}\right) \\ &\overset{(ii)}{=} \frac{1}{J}\sum_{i=1}^J \tilde{\mathrm{OT}}^p_{\tilde{d}^p_C}\left(\mu + \left(\sum_{i=1}^J \mathbb{M}(\mu_i) - \mathbb{M}(\mu)\right)\delta_{\mathfrak{d}}, \tilde{\mu}_i\right) \\ &= \tilde{F}_{p,C}\left(\mu + \left(\sum_{i=1}^J \mathbb{M}(\mu_i) - \mathbb{M}(\mu)\right)\delta_{\mathfrak{d}}\right), \end{aligned}$$

where (i) follows from the lift to an OT problem (Section 3.1) and (ii) follows from Lemma 3.1 by adding mass $\sum_{j\neq i}\mathbb{M}(\mu_j) - \mathbb{M}(\mu)$ at $\mathfrak{d}$. We then have that

$$\begin{aligned} \min_{\substack{\mu\in\mathcal{M}_+(\mathcal{Y}) \\ \mathbb{M}(\mu)\leq\sum_{i=1}^J \mathbb{M}(\mu_i)}} F_{p,C}(\mu) &= \min_{\substack{\mu\in\mathcal{M}_+(\mathcal{Y}) \\ \mathbb{M}(\mu)\leq\sum_{i=1}^J \mathbb{M}(\mu_i)}} \tilde{F}_{p,C}\left(\mu + \left(\sum_{i=1}^J \mathbb{M}(\mu_i) - \mathbb{M}(\mu)\right)\delta_{\mathfrak{d}}\right) \\ &\geq \min_{\substack{\mu\in\mathcal{M}_+(\tilde{\mathcal{Y}}) \\ \mathbb{M}(\mu)=\sum_{i=1}^J \mathbb{M}(\mu_i)}} \tilde{F}_{p,C}(\mu) \end{aligned}$$

and

$$\min_{\substack{\mu\in\mathcal{M}_+(\tilde{\mathcal{Y}}) \\ \mathbb{M}(\mu)=\sum_{i=1}^J \mathbb{M}(\mu_i)}} \tilde{F}_{p,C}(\mu) = \min_{\substack{\mu\in\mathcal{M}_+(\tilde{\mathcal{Y}}) \\ \mathbb{M}(\mu)=\sum_{i=1}^J \mathbb{M}(\mu_i)}} F_{p,C}(\mu_{|\mathcal{Y}}) \geq \min_{\substack{\mu\in\mathcal{M}_+(\mathcal{Y}) \\ \mathbb{M}(\mu)\leq\sum_{i=1}^J \mathbb{M}(\mu_i)}} F_{p,C}(\mu).$$

Combining both inequalities and using Lemma 3.2 then finishes the proof. □

Proof of Lemma 3.5. **(i)** By definition, the objective value for $\tilde{T}_C^{J,p}(y_1, \ldots, y_J)$ at $\mathfrak{d}$ is equal to $(J - |\mathcal{B}|)C^p/2$. Thus, $\tilde{T}_C^{J,p}$ outputs $\mathfrak{d}$ if and only if for any $y \in \mathcal{Y}$ it holds

$$\sum_{i=1}^J \tilde{d}^p_C(y_i, y) \geq (J - |\mathcal{B}|)C^p/2$$

which is equivalent to

$$\sum_{i\notin\mathcal{B}} \tilde{d}_C^p(y_i, y) \geq (J - 2|\mathcal{B}|)C^p/2.$$

In particular, if all inequalities are strict $\mathfrak{d}$ is the unique output for $\tilde{T}_C^{J,p}(y_1, \dots, y_J)$. Statement **(ii)** is a direct consequence of (i). For statement **(iii)** we again use that by definition $\tilde{d}_C^p(y, \mathfrak{d}) = C^p/2$ for any $y \in \mathcal{Y}$ and hence

$$\min_{y\in\mathcal{Y}} \sum_{i=1}^{J} \tilde{d}_C^p(y_i, y) = |\mathcal{B}|\frac{C^p}{2} + \min_{y\in\mathcal{Y}} \sum_{i\notin\mathcal{B}} \tilde{d}_C^p(y_i, y).$$

Proving **(iv)**, let $C > 2^{1/p}\mathrm{diam}(\mathcal{Y})$, pick points $y_1, \dots, y_J \in \mathcal{Y}$ and observe that for any $y \in \mathcal{Y}$ it holds that

$$\sum_{i=1}^{J} \tilde{d}_C^p(y_i, \mathfrak{d}) = J\frac{C^p}{2} > J\mathrm{diam}(\mathcal{Y})^p \geq \sum_{i=1}^{J} d^p(y_i, y).$$

Thus, $\tilde{T}_C^{J,p}(x_1, \dots, x_J) \neq \mathfrak{d}$ and since $|\mathcal{B}| = 0$, the claim follows from (iii). □

A.2 Proofs of Section 2

Proof for Lemma 2.1. Suppose that π_C is optimal but its induced graph $G(\pi_C)$ contains a path $P = (x_{i_1}, \dots, x_{i_k})$ such that $\mathcal{L}(P) > C^p$. By definition of $G(\pi_C)$ it holds that $\pi_C(x_{i_j}, x_{i_{j+1}}) > 0$ for all $1 \leq j \leq k-1$. We define a new transport plan with augmented transport along the path P. For this, define $\epsilon := \min_{1\leq j\leq k-1} \pi_C(x_{i_j}, x_{i_{j+1}})$ and construct the new plan

$$\tilde{\pi}_C(x, x') = \begin{cases} \pi_C(x, x') - \epsilon, & \text{if } \exists\, 1 \leq j \leq k-1,\ x = x_{i_j},\ x' = x_{i_{j+1}} \\ \pi_C(x, x'), & \text{else.} \end{cases}$$

Compared to π_C the transportation cost for $\tilde{\pi_C}$ is reduced by $\epsilon\mathcal{L}(P)$ while the marginal deviation is increased by ϵC^p. In particular, it holds that

$$\begin{aligned} &\sum_{x,x'} d^p(x, x')\tilde{\pi}_C(x, x') + \frac{C^p}{2}\left(\sum_x \mu(x) - \tilde{\pi}_C(x, \mathcal{X}) + \sum_{x'} \nu(x') - \tilde{\pi}_C(\mathcal{X}, x')\right) \\ =&\sum_{x,x'} d^p(x, x')\pi_C(x, x') + \frac{C^p}{2}\left(\sum_x \mu(x) - \pi_C(x, \mathcal{X}) + \sum_{x'} \nu(x') - \pi_C(\mathcal{X}, x')\right) \\ &+ \epsilon\left(C^p - \mathcal{L}(P)\right). \end{aligned}$$

As $\epsilon > 0$ and $\mathcal{L}(P) > C^p$ this contradicts the optimality for π_C. Consequently, any path P in the induced graph $G(\pi_C)$ necessarily has path length at most C^p. If $d(x, x') > C$ this implies that $d^p(x, x') > C^p$ and hence by the statement on induced graphs that $\pi_C(x, x') = 0$. □

Proof for Theorem 2.2. We first establish the metric properties **(i)**. It is straightforward to show $\mathrm{KR}_{p,C}(\mu, \nu) = 0$ if and only if $\mu = \nu$ and that $\mathrm{KR}_{p,C}$ is symmetric. For the triangle inequality let $\mu, \nu, \tau \in \mathcal{M}_+(\mathcal{X})$ and choose $B \geq \max\{\mathbb{M}(\mu), \mathbb{M}(\nu), \mathbb{M}(\tau)\}$. Then by augmenting the measures accordingly (Section 3.1) we find that

$$\begin{aligned} \mathrm{KR}_{p,C}(\mu, \nu) &= \left(\tilde{\mathrm{OT}}_{\tilde{d}_C^p}(\tilde{\mu}, \tilde{\nu})\right)^{1/p} \\ &\leq \left(\tilde{\mathrm{OT}}_{\tilde{d}_C^p}(\tilde{\mu}, \tilde{\tau})\right)^{1/p} + \left(\tilde{\mathrm{OT}}_{\tilde{d}_C^p}(\tilde{\tau}, \tilde{\nu})\right)^{1/p} = \mathrm{KR}_{p,C}(\mu, \tau) + \mathrm{KR}_{p,C}(\tau, \nu) \end{aligned}$$

where the inequality follows by the triangle inequality for the Wasserstein distance [Villani, 2003, Theorem 7.3]. Statement **(ii)** follows from Lemma 2.1 by noting that there exists at least one optimal solution π_C equal to zero except on the diagonal for which $\pi_C(x,x) = \mu(x) \wedge \nu(x)$. Plugging into the objective of (2) yields the claim. Additionally, suppose that w.l.o.g. $\mu(x) \geq \nu(x)$ for all $x \in \mathcal{X}$. Then independent to the choice of $C > 0$ and $p \geq 1$ the unique optimal solution is to remain all shared mass at its common place and to delete surplus material which is exactly the solution $\pi_C(x,x) = \mu(x) \wedge \nu(x)$ described before. Statement **(iii)** follows by noting that for $C \geq \max_{x,x'} d(x,x')$ the dual formulation in ($\mathrm{DUOT}_{p,C}$) and in (DOT_p) coincide.
Finally, for statement **(iv)** we note that by construction it holds $\tilde{d}^p_{C_1}(x,y) \leq \tilde{d}^p_{C_2}(x,y)$ for all $x, y \in \tilde{\mathcal{Y}}$. Hence, for any coupling π of the augmented measures $\tilde{\mu}, \tilde{\nu}$ it holds

$$\sum_{x,x' \in \tilde{\mathcal{Y}}} \tilde{d}^p_{C_1}(x,x')\pi(x,x') \leq \sum_{x,x' \in \tilde{\mathcal{Y}}} \tilde{d}^p_{C_2}(x,x')\pi(x,x').$$

Taking the minimum over all couplings of $\tilde{\mu}$ and $\tilde{\nu}$ on both sides completes the proof. □

A.2.1 Proof for Theorem 2.3

Using the lift to the OT problem, we can now start to prove the closed formula on ultrametric trees. For this, consider an ultrametric tree $\mathcal{T}$ with height function $h\colon V \to \mathbb{R}_+$ and define its *p-height transformed tree* denoted $\mathcal{T}_p := \mathcal{T}$ as the same tree but with height function $h_p(\mathsf{v}) = 2^{p-1}h(\mathsf{v})^p$. An illustration is given in Figure 4. Notice that by monotonicity $\mathcal{T}_p$ is again an ultrametric tree.

Lemma A.1. *Let $\mathcal{T}$ be an ultrametric tree with height function $h\colon V \to \mathbb{R}_+$ and consider its p-height transformed tree $\mathcal{T}_p$. Then it holds that*

$$d^p_{\mathcal{T}}(\mathsf{v},\mathsf{w}) = d_{\mathcal{T}_p}(\mathsf{v},\mathsf{w})$$

for all leaf nodes $\mathsf{v},\mathsf{w} \in L \subset V$.

Proof. Let $\mathsf{v},\mathsf{w} \in L$ be two leaf nodes in the ultrametric tree $\mathcal{T}$ with height function h and let a be their common ancestor[7]. Since paths between any two vertices are unique and all leaf nodes have the same distance to the root, it holds that

$$\begin{aligned} d_{\mathcal{T}}(\mathsf{v},\mathsf{a}) - d_{\mathcal{T}}(\mathsf{w},\mathsf{a}) &= d_{\mathcal{T}}(\mathsf{v},\mathsf{a}) + d_{\mathcal{T}}(\mathsf{a},\mathsf{r}) - d_{\mathcal{T}}(\mathsf{a},\mathsf{r}) - d_{\mathcal{T}}(\mathsf{w},\mathsf{a}) \\ &= d_{\mathcal{T}}(\mathsf{v},\mathsf{r}) - d_{\mathcal{T}}(\mathsf{w},\mathsf{r}) = 0. \end{aligned}$$

Hence,

$$(d_{\mathcal{T}}(\mathsf{v},\mathsf{w}))^p = (d_{\mathcal{T}}(\mathsf{v},\mathsf{a}) + d_{\mathcal{T}}(\mathsf{w},\mathsf{a}))^p = 2^p(h(\mathsf{a}) - h(\mathsf{v}))^p = 2^p h(\mathsf{a})^p,$$

where we use that $h(\mathsf{v}) = 0$. Repeating the argument for the ultrametric tree $\mathcal{T}_p$ we conclude that $d_{\mathcal{T}_p}(\mathsf{v},\mathsf{w}) = 2d_{\mathcal{T}_p}(\mathsf{v},\mathsf{a}) = 2^p h(\mathsf{a})^p$. □

Equipped with this result we are now able to prove the closed formula from Theorem 2.3.

Proof for Theorem 2.3. Let $\mathrm{KR}^p_{p,C}(\mu,\nu) = \mathrm{UOT}_{p,C}(\mu,\nu)$ refer to UOT w.r.t. the distance on $\mathcal{T}$, which only depends on the distance between individual leaf nodes. Considering the p-th height transformed tree $\mathcal{T}_p$ and applying Lemma A.1 we conclude that

[7] If $\mathsf{v},\mathsf{w} \in L$ are leaf nodes their common ancestor is defined as the node included in the path from v to w closest to the root.

$$\mathrm{KR}^p_{p,C}(\mu^L,\nu^L) = \min_{\pi\in\mathcal{M}_+(L\times L)} \sum_{\mathsf{v},\mathsf{v}'\in L} d_{\mathcal{T}_p}(\mathsf{v},\mathsf{v}')\pi(\mathsf{v},\mathsf{v}') + \frac{C^p}{2}\left(\sum_{\mathsf{v}\in L}\left(\mu^L(\mathsf{v})-\pi(\mathsf{v},L)\right)+\sum_{\mathsf{v}'\in L}\left(\nu^L(\mathsf{v}')-\pi(L,\mathsf{v}')\right)\right)$$

$$\text{s.t. } \pi(\mathsf{v},L)\le \mu^L(\mathsf{v}), \quad \forall\,\mathsf{v}\in L,$$

$$\pi(L,\mathsf{v}')\le \nu^L(\mathsf{v}'), \quad \forall\,\mathsf{v}'\in L.$$

The linear optimization problem can be decomposed on several subtrees. For this recall that by Lemma 2.1 (i) there exists an optimal solution such that mass transportation is only considered on metric scales between two leaf nodes $\mathsf{v},\mathsf{v}'\in L$ such that $d_{\mathcal{T}_p}(\mathsf{v},\mathsf{v}')\le C^p$. If v_0 is the common ancestor of v,v' then by the ultrametric tree properties of $\mathcal{T}$ (see also the proof of Lemma A.1) the inequality $d_{\mathcal{T}_p}(\mathsf{v},\mathsf{v}')\le C^p$ is equivalent to the height function $h(\mathsf{v}_0)\le \frac{C}{2}$. Consider the set $\mathcal{R}(C)$ in (7) and for each $\mathsf{v}\in\mathcal{R}(C)$ define subtrees $\mathcal{C}(\mathsf{v})$ consisting of the children of v and the subset of corresponding edges. By construction if $\mathsf{v}_i,\mathsf{v}_j\in\mathcal{R}(C)$ with $\mathsf{v}_i\neq\mathsf{v}_j$ then the subtrees are disjoint $\mathcal{C}(\mathsf{v}_i)\cap\mathcal{C}(\mathsf{v}_j)=\emptyset$ (Figure 4 (a) for an illustration). In particular, the linear optimization problem $\mathrm{KR}^p_{p,C}(\mu^L,\nu^L)$ is decomposed on each individual subtree $\mathcal{C}(\mathsf{v})$ for each $\mathsf{v}\in\mathcal{R}(C)$. The distance on individual subtrees is set to be the p-th height transformed tree distance $d_{\mathcal{T}_p}$ which exactly captures the pairwise p-th power distance between leaf nodes belonging to the same subtree (Lemma A.1). For an element $\mathsf{v}\in\mathcal{R}(C)$ consider its subtree $\mathcal{C}(\mathsf{v})$ with distance $d_{\mathcal{T}_p}$. By definition the maximal distance between its leaf nodes is bounded by $C^p/2$. We augment the subtree $\mathcal{C}(\mathsf{v})$ with a dummy node $\tilde{\mathsf{v}}$ and introduce an edge $e=(\mathsf{v},\tilde{\mathsf{v}})$ with edge weight $\frac{C^p}{2}-2^{p-1}h(\mathsf{v})^p$ (Figure 4 (b) for an illustration). Denote the augmented tree by $\tilde{\mathcal{C}}(\mathsf{v})$. Considering the measures μ^L,ν^L restricted to $\mathcal{C}(\mathsf{v})$ we augment μ^L adding mass $\left(\mu^L(\mathcal{C}(\mathsf{v}))-\nu^L(\mathcal{C}(\mathsf{v}))\right)_+$ at $\tilde{\mathsf{v}}$ and vice versa augment ν^L adding mass $\left(\nu^L(\mathcal{C}(\mathsf{v}))-\mu^L(\mathcal{C}(\mathsf{v}))\right)_+$ at $\tilde{\mathsf{v}}$. This construction defines an equivalent OT problem on $\tilde{\mathcal{C}}(\mathsf{v})$ [Guittet, 2002]. Hence, applying the closed formula for OT on general metric trees [Evans and Matsen, 2012, p.575] yields

$$2^{p-1}\sum_{\mathsf{w}\in\mathcal{C}(\mathsf{v})\setminus\{\mathsf{v}\}}\left((h(par(\mathsf{w}))^p-h(\mathsf{w})^p)\left|\mu^L(\mathcal{C}(\mathsf{w}))-\nu^L(\mathcal{C}(\mathsf{w}))\right|\right) + \left(\frac{C^p}{2}-h(\mathsf{v})\right)\left|\mu^L(\mathcal{C}(\mathsf{v}))-\nu^L(\mathcal{C}(\mathsf{v}))\right|.$$

Summing over all subtrees indexed by the set $\mathcal{R}(C)$ finishes the proof. □

A.2.2 Proofs for the Barycenter

Proof for Theorem 2.5. **(ii)** Let μ be a (p,C)-barycenter and $\tilde{\mu}$ its augmented counterpart. Then, by Proposition 3.9 there exists an optimal multi-coupling π, such that it holds $\mu=\tilde{\mu}_{|\mathcal{Y}}=(\tilde{T}_C^{J,p}\#\pi)_{|\mathcal{Y}}$. Hence, for each $y\in\mathrm{supp}(\tilde{\mu})$ there exists $K_y\ge 1$ and K_y J-tupels $(x^1_{y,1},\dots,x^J_{y,1}),\dots,(x^1_{y,K_y},\dots,x^J_{y,K_y})$ such that for $k=1,\dots,K_y$ it holds

$$y=\tilde{T}_C^{J,p}(x^1_{y,k},\dots,x^J_{y,k})$$

and $\mu(y)=\sum_{k=1}^{K_y}a^y_k$, where $a^y_k=\pi(x^1_{y,k},\dots,x^J_{y,k})$. For $i=1,\dots,J$ define $\tilde{\pi}_i\in\mathcal{M}_+(\mathcal{Y}\times\mathcal{Y})$ by $\tilde{\pi}_i(y,x^i_{y,k})=a^y_k$ for all $y\in\mathrm{supp}(\tilde{\mu})$, $i=1,\dots,J$ and $k=1,\dots,K_y$. Set $\tilde{\pi}_i$ to be zero

everywhere else for $i = 1, \dots, J$. By construction, $\tilde{\pi}_i$ defines an OT plan between $\tilde{\mu}$ and $\tilde{\mu}_i$ for $i = 1, \dots, J$. It holds

$$\begin{aligned}\frac{1}{J}\sum_{i=1}^{J} \tilde{OT}_{\tilde{d}_C^p}(\tilde{\mu}, \tilde{\mu}_i) &= \sum_{x \in \mathrm{supp}(\pi)} c_{p,C}(x)\pi(x) \\ &= \sum_{x \in \mathrm{supp}(\pi)} \frac{1}{J}\sum_{i=1}^{J} \tilde{d}_C^p\left(x_i, \tilde{T}_C^{J,p}(x)\right)\pi(x) \\ &= \frac{1}{J}\sum_{y \in \mathrm{supp}(\tilde{\mu})}\sum_{i=1}^{J}\sum_{k=1}^{K_y} \tilde{d}_C^p(x_{y,k}^i, y)a_k^y \\ &= \frac{1}{J}\sum_{i=1}^{J}\sum_{y \in \mathrm{supp}(\tilde{\mu})}\sum_{k=1}^{K_y} \tilde{d}_C^p(x_{y,k}^i, y)\tilde{\pi}^i(y, x_{y,k}^i),\end{aligned}$$

where the first equality follows from Proposition 3.9 and the third and fourth by construction. Since $\tilde{\pi}_i$ is an OT plan between $\tilde{\mu}$ and $\tilde{\mu}_i$ it holds for all $i = 1, \dots, J$ that

$$\sum_{y \in \mathrm{supp}(\tilde{\mu})}\sum_{k=1}^{K_y} \tilde{d}_C^p(x_{y,k}^i, y)\tilde{\pi}_i(y, x_{y,k}^i) \geq \tilde{OT}_{p,C}^p(\tilde{\mu}, \tilde{\mu}_i).$$

Thus, it follows together with the previous equations that

$$\tilde{OT}_{p,C}^p(\tilde{\mu}, \tilde{\mu}_i) = \sum_{y \in \mathrm{supp}(\tilde{\mu})}\sum_{k=1}^{K_y} \tilde{d}_C^p(x_{y,k}^i, y)\tilde{\pi}_i(y, x_{y,k}^i),$$

i.e. $\tilde{\pi}_i$ is optimal. Lemma 3.5 now yields the first part of the statement.
For the second part assume that for any $(x_1, \dots, x_L) \in \mathcal{Y}^L$ it holds that $T^{L,p}(x_1, \dots, x_L) = T^{L,p}(y_1, x_2, \dots, x_L)$ is equivalent to $x_1 = y_1$. Let $y \in \mathrm{supp}(\tilde{\mu})$ and consider OT plans $\tilde{\pi}^1, \dots, \tilde{\pi}^J$ between $\tilde{\mu}$ and $\tilde{\mu}_i$, respectively. For $i = 1, \dots, J$ consider x^i such that $\tilde{\pi}_i(y, x^i) = a_i > 0$. Assume that it holds $y \neq \tilde{T}_C^{J,p}(x_1, \dots, x_J)$. Denote the minimum of the a_i as $a_0 = \min_{i=1,\dots,J} a_i$. By construction, it follows that

$$\tilde{F}_{p,C}(\tilde{\mu} - a_0\delta_y + a_0\delta_{\tilde{T}_C^{J,p}(x_1,\dots,x_J)}) < F_{p,C}(\tilde{\mu}),$$

which is a contradiction to $\tilde{\mu}$ being a barycenter of $\tilde{\mu}_1, \dots, \tilde{\mu}_J$. Thus, it holds $y = \tilde{T}_C^{J,p}(x_1, \dots, x_J)$. Now, assume w.l.o.g. there exists $x_1, z_1 \in \mathcal{Y}$, such that it holds $\pi^1(y, x_1) > 0$ and $\pi^1(y, z_1) > 0$. However, by the previous argument this implies

$$\tilde{T}_C^{J,p}(x_1, \dots, x_J) = y = \tilde{T}_C^{J,p}(z_1, \dots, x_J).$$

By assumption this is equivalent to $x_1 = z_1$, thus it holds for all $x, y \in \mathcal{Y}$ and $i = 1, \dots, J$ that $\pi^i(y, x) \in \{0, \mu(y)\}$.
(i) By Proposition 3.9 the objective value of the balanced multi-marginal and (p, C)-barycenter problem coincide and a (p, C)-barycenter is obtained as the push-forward of an optimal balanced multi-coupling under the map $\tilde{T}_C^{J,p}$ restricted to $\mathcal{Y}$. By construction and Corollary 3.7 any such measure is supported in $\mathcal{C}_{\mathrm{KR}}(J, p, C)$. Thus, there always exists a (p, C)-barycenter whose support is restricted to $\mathcal{C}_{\mathrm{KR}}(J, p, C)$ and the minimum over Y and $\mathcal{C}_{\mathrm{KR}}(J, p, C)$ coincide.
The second part is similar and we let $\tilde{\mu}$ be any p-Wasserstein barycenter. Then by Proposition 3.9, there exists a multi-coupling of $\tilde{\mu}_1, \dots, \tilde{\mu}_J$, such that $\tilde{\mu} = \tilde{T}_C^{J,p}\#\tilde{\pi}$. Since any such

push-forward measure can only have support in $\mathcal{C}_{\mathrm{KR}}(J,p,C)\cup\{\mathfrak{d}\}$, it holds for $\mu=\tilde{\mu}_{|\mathcal{Y}}$ that $\mathrm{supp}(\mu)\subset\mathcal{C}_{\mathrm{KR}}(J,p,C)$. It remains to show the upper bound on the total mass. By the equivalence to the multi-marginal problem and by Lemma 3.5 (ii) any (p,C)-barycenter μ cannot have mass on a point which is constructed from a set of points $(x_1,\dots,x_J)$ for which $2|\mathcal{B}(x_1,\dots,x_J)|\geq J$. Additionally, by part (ii) we know that there exists UOT plans, such that the mass of each (p,C)-barycenter support point is fully transported to points it is constructed from. Let $(a_1,\dots,a_K)$ be the weight vector of the support points of the (p,C)-barycenter, then it holds that

$$\sum_{i=1}^{J}\mathbb{M}(\mu_i)-\lceil J/2\rceil\sum_{k=1}^{K}a_k\geq 0,$$

since by the previous argument and Lemma 3.5, any (p,C)-barycenter support point x_k reduces the maximum available mass by at least $\lceil J/2\rceil a_k$ and by Lemma 3.2, the total mass of the (p,C)-barycenter is bounded by the sum of the total masses of the μ_i. Therefore it holds that

$$\mathbb{M}(\mu)=\sum_{k=1}^{K}a_k\leq\lceil J/2\rceil^{-1}\sum_{i=1}^{J}\mathbb{M}(\mu_i)\leq\frac{2}{J}\sum_{i=1}^{J}\mathbb{M}(\mu_i).$$

(iii) The multi-marginal problem between $\tilde{\mu}_1,\dots,\tilde{\mu}_J$ is a balanced problem, thus we can pose this as a linear program with a total of $\prod_{i=1}^{J}M_i$ variables and $\sum_{i=1}^{N}M_i+J$ constraints. As all measures have the same total mass, we can drop one arbitrary marginal constraint for each measure besides the first. Thus, the rank of the constraint matrix in the corresponding constraint is bounded by $\sum_{i=1}^{N}M_i+1$. Hence, each basic feasible solution of the linear program has at most $\sum_{i=1}^{N}M_i+1$ non-zero entries (see [Luenberger et al., 1984] for details). Let π be such a solution. By Proposition 3.9 the measure $\tilde{\mu}=\tilde{T}_C^{J,p}\#\pi$ is a p-Wasserstein barycenter and by construction it has at most $\sum_{i=1}^{N}M_i+1$ support points. Due to the upper bound on the total mass of the (p,C)-barycenter in property (i), we can guarantee that there is non-zero mass at $\mathfrak{d}$ for $J>2$, hence in this case, restricting the measure to $\mathcal{Y}$ reduces the support size by one. For $J=2$, we note that the multi-marginal problem is just the augmented UOT problem. By construction we either have a point x in the support of one of the two measures, such that there is transport between x and $\mathfrak{d}$ or both measures have equal mass at $\mathfrak{d}$ and it is optimal to leave this mass in place. In the first case, we have mass at $\tilde{T}_C^{J,p}(x,\mathfrak{d})=\mathfrak{d}$, thus the support size can be reduced by one and in the second the problem is equivalent to the OT problem and thus the barycenter has at most M_1+M_2-1 support points. Finally, by property (i) the support of any (p,C)-barycenter is contained in $\mathcal{C}_{KR}(J,p,C)$, thus the cardinality of this set also provides a trivial upper bound on the support size of any (p,C)-barycenter. Taking the minimum over both quantities, we conclude

$$|\mathrm{supp}(\mu)|\leq\min\left\{|\mathcal{C}_{KR}(J,p,C)|,\sum_{i=1}^{J}M_i\right\}.$$

(iv) For any $\mu\in\mathcal{M}_+(\mathcal{Y})$, it holds

$$F^p_{p,C_1}(\mu)=\frac{1}{J}\sum_{i=1}^{J}\mathrm{KR}^p_{p,C_1}(\mu,\mu_i)\leq\frac{1}{J}\sum_{i=1}^{J}\mathrm{KR}^p_{p,C_2}(\mu,\mu_i)=F^p_{p,C_2}(\mu),$$

where the inequality follows from Theorem 2.2 (iv). Taking the infimum over all measures in $\mathcal{M}_+(\mathcal{Y})$ on both sides completes the proof.

(v) Let $C \leq d'_{\min}$, then by Theorem 2.2 (ii) it holds

$$\begin{aligned}\operatorname*{arg\,min}_{\mu\in\mathcal{M}_+(\mathcal{Y})} F_{p,C}(\mu) &= \operatorname*{arg\,min}_{\mu\in\mathcal{M}_+(\mathcal{Y})} \frac{C^p}{2J}\sum_{i=1}^{J} TV(\mu,\mu_i)\\ &= \operatorname*{arg\,min}_{a\in\mathbb{R}_+^K} \frac{C^p}{2J}\sum_{i=1}^{J}\sum_{k=1}^{K}|a_k - a_k^i|\\ &= \operatorname*{arg\,min}_{a\in\mathbb{R}_+^K} \frac{C^p}{2J}\sum_{k=1}^{K}\sum_{i=1}^{J}|a_k - a_k^i|,\end{aligned}$$

where the change in the arg min in the second line follows from the fact that the total variation can only increase if we place mass outside of the support of the measures. Thus it suffices to consider measures supported on the union of the supports. Now, we note that the K summands are independent to each other, thus we can minimise them separately. Hence, for the k-th entry of a it holds that

$$a_k \in \operatorname*{arg\,min}_{a\in\mathbb{R}_+}\sum_{\mathrm{i}=1}^{\mathrm{N}}|a_k - a_k^i| = med(a_k^1,\dots,a_k^J)$$

which yields the claim.

(vi) Let $\mathbb{M}_i = \mathbb{M}(\mu_i)$ for $i = 1,\dots,J$ and set $\mathbb{M}_0 = 0$. Assume that J is odd. Let μ be a (p,C)-barycenter of $\mu_1,\dots,\mu_J$ with $\mu(\mathcal{Y}) \in [\mathbb{M}_{k-1},\mathbb{M}_k]$. In particular, μ fulfills the non-mass-splitting property in (ii). Let $a \in (0, \mathbb{M}_k - \mathbb{M}(\mu)]$ and $\tilde{\mu}$ the augmented measure for μ. By construction, we can find support points $x_k,\dots,x_J \neq \mathfrak{d}$ of the augmented measures $\tilde{\mu}_k,\dots,\tilde{\mu}_J$ from which w.l.o.g. mass a is transported to $\mathfrak{d}$ in μ. If one of the points has mass smaller a, we can just replace a with the minimum of the masses of the points and repeat the argument until we have considered a total mass of a. Set $x_0 = \tilde{T}_C^{J,p}(\mathfrak{d},\dots,\mathfrak{d},x_k,\dots,x_J)$ and notice that if $x_0 = \mathfrak{d}$, we do not change the objective function in the augmented problem (Lemma 3.1) by adding this point which means w.l.o.g. $x_0 \neq \mathfrak{d}$. In this case, we have

$$x_0 = \operatorname*{arg\,min}_{x\in\mathcal{Y}}\sum_{i=k}^{J}\tilde{d}_C^p(x_i,x).$$

Now, the objective cost of not having mass a at x_0 is $aC^p(J-k)/2$, while the cost of adding $a\delta_{x_0}$ to μ is equal to $a(kC^p/2 + \sum_{i=k}^{J}\tilde{d}_C^p(x_i,x_0))$. Hence, adding the point improves the value of the Fréchet functional, if

$$\sum_{i=k}^{J}\tilde{d}_C^p(x_i,x_0) \leq C^p(J-2k)/2.$$

For $2k > J$, the right hand side will always be negative, so we can not improve. Thus, we assume $2k < J$. By assumption it holds $C \geq J^{\frac{1}{p}}\mathrm{diam}(\mathcal{Z})$. Hence,

$$\begin{aligned}&\frac{C^p}{2} \geq \frac{J}{2}diam(\mathcal{Z})^p\\ \Leftrightarrow\quad &\frac{C^p}{2\mathrm{diam}(\mathcal{Z})^p} \geq \frac{J}{2} \geq \frac{J-k}{J-2k}\\ \Leftrightarrow\quad &C^p(J-2k)/2 \geq \mathrm{diam}(\mathcal{Z})^p(J-k) \geq \sum_{i=k}^{J}\tilde{d}_C^p(x_i,x_0).\end{aligned}$$

Therefore, for $2k < J$ the objective value of μ can always be improved by increasing its mass by a, as long as $k < \lceil J/2 \rceil$. Thus, since μ is a barycenter it holds $\mathbb{M}(\mu) \geq \mathbb{M}\mu_{\lceil J/2 \rceil}$. An analog, converse argument yields that if $k > J/2$, we can always improve the objective value of μ, since removing and then re-adding any mass to μ increases the objective value by the previous argument. Hence, it holds $\mathbb{M}(\mu) = \mathbb{M}\mu_{\lceil J/2 \rceil}$.
Now, assume J is even. For $2k \neq J$ nothing in the previous argument changes. However, for $2k = J$ (note that this can only hold now that J is even), the right hand side is zero, however, if all the x_i for $i = k, \ldots, J$, are identical to x_0 (in particular, there exists a point contained in the support of at least half of the measures), then the left hand side will also be zero. In this case, the presence of this point does not change the objective value and there are (p, C)-barycenters of different total masses. However, we can still always choose to not place mass in such cases, to obtain a (p, C)-barycenter of the desired total mass. □

Proof for Lemma 2.7. It suffices to show that there is no centroid point, which is constructed from points from two or more different sets B_r. Assume there is a point $y_0 \in \mathcal{C}_{KR}(J, p, C)$, such that y_0 is constructed, among others, from $x_1 \in B_r$ and $x_2 \in B_s$ for $r \neq s$. We distinguish two cases. Assume $y_0 \in B_r$, then it holds $d^p(x_1, y_0) > 2^{p-1}C^p \geq C^p$ and y_0 would not be in the restricted centroid set. The analogue argument holds for $y_0 \in B_s$. Now, assume y_0 is neither in B_r nor B_S. Since $d(B_r, B_S) > 2^{1/p}C$, it holds either $d^p(B_r, y_0) > C^p$ or $d^p(B_s, y_0) > C^p$. Thus, we obtain another contradiction to $y_0 \in \mathcal{C}_{KR}(J, p, C)$. Hence, $\mathcal{C}_{KR}(J, p, C)$ only contains centroids constructed from points within one B_r and by convexity of the B_r, any centroid point constructed from points within B_r is again in B_r. Theorem 2.5 (ii) yields that there will always be an optimal solution which only transports within each B_r, thus the R problems are in fact independent and we can separate them without changing the objective value. □

CHAPTER C

$(\mathbf{p}, \mathbf{C})$-Kantorovich-Rubinstein distance and barycenter for finitely supported measures: A statistical perspective

$(\mathbf{p}, \mathbf{C})$-Kantorovich-Rubinstein distance and barycenter for finitely supported measures: A statistical perspective

Florian Heinemann* Marcel Klatt* Axel Munk*†‡

July 19, 2022

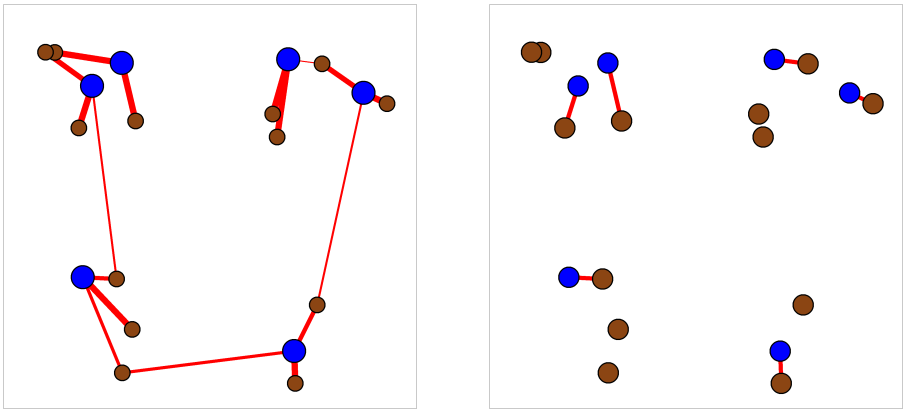
Abstract

In this paper we propose and investigate specific statistical models and corresponding sampling schemes for data analysis based on unbalanced optimal transport for finitely supported measures. Specifically, we analyse Kantorovich-Rubinstein (KR) distances with penalty parameter $C > 0$ between measures generated by some underlying statistical model. The main result provides non-asymptotic bounds on the expected error for the empirical KR distance as well as for its barycenters. The impact of the penalty parameter C is studied in detail. Our approach allows for randomised computational schemes for UOT which can be used for fast approximate computations with any exact solver. Using synthetic and real datasets, we empirically analyse the behaviour of the expected errors in simulation studies and illustrate the validity of our theoretical bounds.

1 Introduction

Optimal transport (OT) [for a detailed mathematical discussion see e.g. Villani, 2008, Santambrogio, 2015] has been a focus of attention in various research fields in recent years. Its powerful geometric features promoted by improved computational tools [Chizat et al., 2018a, Peyré and Cuturi, 2019, Guo et al., 2020] have turned OT into a promising new tool for modern data analysis with applications in machine learning [Frogner et al., 2015, Arjovsky et al., 2017, Schmitz et al., 2018, Yang et al., 2018], computer vision [Baumgartner et al., 2018, Yang et al., 2021], genetics [Evans and Matsen, 2012, Schiebinger et al., 2019], cell biology [Gellert et al., 2019, Klatt et al., 2020, Tameling et al., 2021] and image processing [Pitié et al., 2007, Rabin and Papadakis, 2015, Tartavel et al., 2016].
However, the wide range of OT applications also surfaced some limitations of classical OT. In particular, the assumption of equal total mass intensity of the measures is often inappropriate. A straightforward strategy to overcome this issue in settings with different total mass intensities is to normalise the measures' total intensities. However, this preprocessing step has an immediate impact on the corresponding transport plan. For example, when matching point clouds of different sizes the resulting plan distributes mass among several points, whereas often it is desired to match points one-to-one which is favourable in many applications (see Figure 1). Attempts to circumvent this issue have led to a range of *unbalanced optimal transport* (UOT) proposals [Figalli, 2010, Liero et al., 2018, Chizat et al., 2018b, Balaji et al., 2020, Mukherjee et al., 2021, Heinemann et al., 2021]. These formulations extend optimal transport concepts to general positive measures by either fixing the total amount of mass to be transported in advance or by penalising marginal

*Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen
†Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen
‡University Medical Center Göttingen, Cluster of Excellence 2067 Multiscale Bioimaging - From molecular machines to networks of excitable cells



(a) (b)

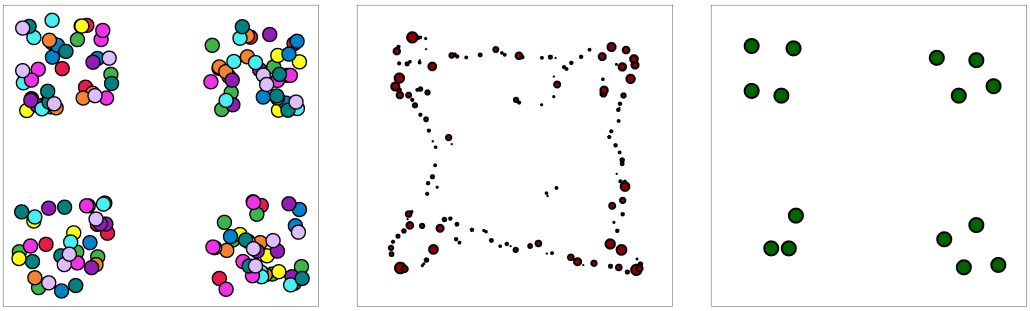
Figure 1: Transport between two measures (blue and brown) with their support points located in $[0,1]^2$. The respective transport plans between them are displayed by red lines where the thickness of a line is proportional to the transported mass. **(a)** The measures have been normalised to probability measures (the blue points have mass $1/6$ and the brown points with mass $1/13$). **(b)** The UOT plan for the $(2,2)$-KRD between the two unnormalised measures (all points have mass 1).

constraints by replacing the hard marginal constraints inherent in OT. These approaches also give rise to an associated barycenter, generalising the popular notion of *OT barycenters* [Agueh and Carlier, 2011] to measures of unequal total intensity.
A first attempt to cope with unbalanced mass already goes back to Kantorovich and Rubinstein [1958]. This was build on by following work of Hanin [1992] and Guittet [2002]. Recently, these ideas have been further studied and the structural properties of the resulting (p, C)-Kantorovich-Rubinstein distance (KRD), in (2), and its barycenter have been considered [Heinemann et al., 2021]. A comparison between the (p, C)-barycenter and the p-Wasserstein barycenter in a simple example can be seen in Figure 2. From a data analysis point of view we find it particularly appealing that for the (p, C)-KRD there is a clear geometrical connection between its penalty C and the structural properties of the UOT plans and (p, C)-barycenter. In particular, C controls the largest scale at which mass transport is possible in an OT plan. Furthermore, each support point of any (p, C)-barycenter is contained in a finite set characterised by the value of C. This intuitive understanding of the (p, C)-KRD allows to easily design it to respect different structural properties of the data and thus makes it a prime candidate for statistical tasks in data analysis.
Though, due to the unbalanced nature of the problem, the task of sampling from the measures requires alternative sampling schemes and different statistical modeling. In this work, we treat three specific models motivated by applications in randomised algorithms and microscopy tasks. These statistical models also serve as meaningful data models for the alternative UOT concepts mentioned above. Notably, our approach also provides a framework which potentially allows to treat alternative models.

1.1 Kantorovich-Rubinstein Distance

Let $(\mathcal{X}, d)$ be a finite metric space with cardinality M and denote by

$$\mathcal{M}_+(\mathcal{X}) \coloneqq \left\{ \mu \in \mathbb{R}^{|\mathcal{X}|} \mid \mu(x) \geq 0 \ \forall x \in \mathcal{X} \right\}$$



(a) (b) (c)

Figure 2: **(a)** $J = 10$ measures (each colour corresponds to a different measure) superimposed on top of each other with uniform mass on their support points in $[0,1]^2$. **(b)** The OT barycenter (for squared Euclidean cost) of the normalised measures. **(c)** The $(2, 0.3)$-barycenter of the unnormalised measures (see (3) for a rigorous definition).

the set of non-negative measures[1] on $\mathcal{X}$. For a measure $\mu \in \mathcal{M}_+(\mathcal{X})$ its total mass is defined as $\mathbb{M}(\mu) \coloneqq \sum_{x\in\mathcal{X}} \mu(x)$ and the subset $\mathcal{P}(\mathcal{X}) \subset \mathcal{M}_+(\mathcal{X})$ of measures with total mass one is the set of probability measures. If $\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X})$ is a measure on the product space $\mathcal{X} \times \mathcal{X}$ its marginals are defined as $\pi(x, \mathcal{X}) \coloneqq \sum_{x'} \pi(x, x')$ and $\pi(\mathcal{X}, x') \coloneqq \sum_{x'\in\mathcal{X}} \pi(x, x')$, respectively. For two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ define the set of *non-negative sub-couplings* as

$$\Pi_{\leq}(\mu, \nu) \coloneqq \{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X}) \,|\, \pi(x, \mathcal{X}) \leq \mu(x),\ \pi(\mathcal{X}, x') \leq \nu(x')\, \forall\, x, x' \in \mathcal{X}\}. \tag{1}$$

For $p \geq 1$ and a parameter $C > 0$, the $(p, C)-$*Kantorovich-Rubinstein distance* (KRD) between two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ is defined as

$$\begin{aligned} \mathrm{KR}_{p,C}(\mu, \nu) \coloneqq \Bigg(& \min_{\pi\in\Pi_{\leq}(\mu,\nu)} \sum_{x,x'\in\mathcal{X}} d^p(x, x')\pi(x, x') \\ & + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi) \right) \Bigg)^{\frac{1}{p}}. \end{aligned} \tag{2}$$

For any $p \geq 1$, it defines a distance on the space of non-negative measures $\mathcal{M}_+(\mathcal{X})$ and it naturally extends the well-known *p-th order OT distance* defined for measures of equal total mass [Heinemann et al., 2021].

1.2 Kantorovich-Rubinstein Barycenters

The (p, C)-KRD also allows to define a notion of a barycenter for a collection of measures with (potentially) different total masses. Assume $(\mathcal{X}, d)$ to be embedded in some connected ambient space[2] $(\mathcal{Y}, d)$, e.g., an Euclidean space. Define the (unbalanced) (p, C)-*Fréchet functional*

$$F_{p,C}(\mu) = \frac{1}{J} \sum_{i=1}^{J} KR^p_{p,C}(\mu^i, \mu). \tag{3}$$

[1] A non-negative measure on a finite space $\mathcal{X}$ is uniquely characterized by the values it assigns to each singleton $\{x\}$. To ease notation we write $\mu(x)$ instead of $\mu(\{x\})$. The corresponding σ-field is always to be understood as the powerset of $\mathcal{X}$.

[2] We assume the metric on $\mathcal{X} \subset \mathcal{Y}$ to be the metric of $\mathcal{Y}$ restricted to $\mathcal{X}$.

Any minimiser of this functional in $\mathcal{M}_+(\mathcal{Y})$ is said to be a (p, C)-*Kantorovich-Rubinstein barycenter* or (p, C)-*barycenter* for short of $\mu_1, \dots, \mu_J$[3]. The objective functional $F_{p,C}$ is referred as (unbalanced) (p, C)-*Fréchet functional*. Let $T^{L,p}$ is the so called *Borel barycenter application*[4]

$$T^{L,p}(x_1, \dots, x_L) \in \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{L} d^p(x_i, y).$$

Define the *full centroid set* of the measures

$$\begin{aligned} \mathcal{C}_{KR}(J, p) = \Big\{ y \in \mathcal{Y} \mid \exists L \geq \lceil J/2 \rceil,\ \exists (i_1, \dots, i_L) \subset \{1, \dots, J\}, \\ x_1, \dots, x_L :\ x_l \in \operatorname{supp}(\mu_{i_l}) \\ \forall l = 1, \dots, L :\ y = T^{L,p}(x_1, \dots, x_L) \Big\}, \end{aligned} \tag{4}$$

and based on it the *restricted centroid set*

$$\begin{aligned} \mathcal{C}_{KR}(J, p, C) = \Big\{ y = T^{L,p}(x_1, \dots, x_L) \in \mathcal{C}_{KR}(J, p) \mid \forall 1 \leq l \leq L : \\ d^p(x_l, y) \leq C^p;\ \sum_{i=1}^{L} d^p(x_l, y) \leq \frac{C^p(2L - J)}{2} \Big\}. \end{aligned} \tag{5}$$

According to Heinemann et al. [2021], any (p, C)-barycenter is finitely supported and its support is included in the restricted centroid set $\mathcal{C}_{KR}(J, p, C)$.

1.3 Statistical Models and Contributions

In practice, one does not necessarily have access to the population measures $\mu, \nu, \mu^1, \dots, \mu^J$, respectively. Instead these measures have to be estimated from data. For probability measures the most common statistical model assumes access to (i.i.d.) data $X_1, \dots, X_N \sim \mu$. A reasonable estimator is the empirical measure $\hat{\mu}_N = (1/N) \sum_{k=1}^{N} \delta_{X_k}$, where δ_X denotes the (random) point measure at location X. However, sampling from measures with arbitrary total mass is not immediate and requires the need for alternative statistical modelling. We propose three approaches motivated by different applications.

Multinomial Model

For the *multinomial model* the measures are normalised to define probability measures such that sampling as described above is possible. The resulting empirical estimators are rescaled to the original total intensities. More precisely, consider i.i.d. random variables $X_1, \dots, X_N \overset{\text{i.i.d.}}{\sim} \frac{\mu}{\mathbb{M}(\mu)}$, where the total intensity $\mathbb{M}(\mu)$ is assumed to be known. The corresponding unbiased empirical estimator is then defined as

$$\hat{\mu}_N \coloneqq \frac{\mathbb{M}(\mu)}{N} \sum_{x \in \mathcal{X}} |\{k \in \{1, \dots, N\} \mid X_k = x\}|\ \delta_x. \tag{6}$$

This approach can be understood as extension of the classical sampling approach to measures of arbitrary total intensities. The key motivation for this model is resampling for randomised computations of UOT. In real world data analysis it is common to encounter

[3] For the sake of readability, the weights in this definition are fixed to $1/J$. Adaptation of all results to arbitrary positive weights $\lambda_1, \dots, \lambda_J$ summing to one is straightforward.

[4] There are scenarios where multiple sets fulfil the definition of the restricted centroid set, since there might be multiple points that minimise the barycentric application. In this case, a fixed representative is chosen and there still exists a choice of centroid set which contains the support of the (p, C)-barycenter.

data (e.g. high-resolution images) which are out of reach for current state of the art solvers for OT. One idea in this scenario is to replace each measure by its empirical version and then use these measures as surrogates. Statistical deviation bounds now allow to balance computational complexity and accuracy of approximation in terms of the sample size N. This idea has been considered in detail for the p-Wasserstein distance [Sommerfeld et al., 2019] and the p-Wasserstein barycenter [Heinemann et al., 2022]. In the context of this model we extend their results to measures of unequal total intensities.

Bernoulli Model

For the *Bernoulli model* we consider measures μ with $\mu(x) = 1$ for all $x \in \operatorname{supp}(\mu)$. Thus, the measure μ represents a point cloud in the ambient space $\mathcal{Y}$. We then assume to observe independent Bernoulli random variables $B_x \sim \operatorname{Ber}(s_x)$ with a fixed *success probability* $s_x \in [0,1]$ for each location $x \in \mathcal{X}$. We denote $s_{\mathcal{X}} := (s_{x_1}, \dots, s_{x_M})$ and refer to $s_{\mathcal{X}}$ as *success vector*. A suitable unbiased estimator for μ is defined by

$$\hat{\mu}_{s_{\mathcal{X}}} := \sum_{x \in \mathcal{X}} \frac{B_x}{s_x} \delta_x. \tag{7}$$

The corresponding Bernoulli field $(B_x)_{x \in \mathcal{X}}$ arises e.g. in fluorescence cell microscopy where a fluorescent marker is excited with a laser beam and emitted photons indicate the position of the objects of interest in the proper experimental setup [Kulaitis et al., 2021]. However, the marker has a limited *labeling efficiency* $s_x \in (0,1]$ at each location $x \in \mathcal{X}$ and we only observe a location which has been labelled by the marker and emits photons.

Poisson Intensity Model

For the *Poisson intensity model* we fix a parameter $t > 0$ and a success probability $s \in [0,1]$. Consider a collection of $|\mathcal{X}|$ independent Poisson random variables $P_x \sim \operatorname{Poi}(t\mu(x))$ with intensity $t\mu(x)$ at each location $x \in \mathcal{X}$ and independent from that Bernoulli random variables $B_x \sim \operatorname{Ber}(s)$ for each $x \in X$. A suitable unbiased estimator for μ is defined by

$$\hat{\mu}_{t,s} := \frac{1}{st} \sum_{x \in \mathcal{X}} B_x P_x \delta_x. \tag{8}$$

This model is closely related to the Bernoulli model. Notably, in this model the success probability is assumed to be homogeneous (as opposed to the inhomogeneous probabilities in the Bernoulli model) and the values of the population measures at each support point are not necessarily equal to one. Hence, we have two independent layers of randomness in the construction of this empirical measure. First, we observe a location with a certain probability s, then we observe random mass driven by a Poisson distribution based on the mass of μ and the value of t.
This model is motivated by various tasks in photonic imaging [see e.g. Munk et al., 2020], for example, fluorescence microscopy, X-ray imaging and positron emission tomography (PET). The finite space $\mathcal{X}$ represents the center of bins of a detection interface used to measure the emitted photons. The value $\mu(x)$ corresponds to the integrated underlying photon intensity over its respective bin. This intensity is proportional to an external source, such as a laser intensity in fluorescence microscopy and modelled by the parameter $t > 0$. The Bernoulli random variable B_x models the possibility that during the observation time in the bin of x no photon is recorded. This might be due to various effects that cause thinning, such as limited labelling efficiency, dead time of cameras or a loss of photons due to sparse detector tubes. The value of P_x corresponds to the number of photons which have been measured at the bin of x. Note, that besides B_x, there might also

be additional effects present which do not disable the whole bin, but just prevent a single photon from being measured. All this is incorporated in the probability $s' \in (0,1]$ that a single photon at any bin and any point in time can not be measured. In this case, the model can be shown to be equivalent to a Poisson model with parameter $ts' > 0$ instead of t. Hence, this kind of thinning corresponds to a reparametrisation of the original model and is thus a special case of this general Poisson intensity model.

1.3.1 Sampling Bounds

Let $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ and let $\hat{\mu}, \hat{\nu}$ be empirical versions of these measures generated with one of the aforementioned sampling mechanisms. We are interested in tight upper bounds for the quantities

$$(i)\ \mathbb{E}\left[KR_{p,C}(\mu,\hat{\mu})\right], \quad (ii)\ \mathbb{E}\left[\left|KR_{p,C}(\hat{\mu},\hat{\nu}) - KR_{p,C}(\mu,\nu)\right|\right].$$

In particular, we show that there exist constants $\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C)$, $\mathcal{E}^{\mathrm{Pois}}_{p,\mathcal{X},\mu}(C)$, $\mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C)$ such that for any $p \geq 1$ and for any measure μ and its estimator $\hat{\mu}$ in one of the three statistical models it holds

$$\mathbb{E}\left[\mathrm{KR}_{p,C}\left(\hat{\mu},\mu\right)\right] \leq \begin{cases} \mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, N^{-\frac{1}{2p}}, & \text{if } \hat{\mu} = \hat{\mu}_N, \\ \mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, \phi(t,s)^{\frac{1}{p}}, & \text{if } \hat{\mu} = \hat{\mu}_{t,s}, \\ \mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, \psi(s_{\mathcal{X}})^{\frac{1}{p}}, & \text{if } \hat{\mu} = \hat{\mu}_{s_{\mathcal{X}}}, \end{cases}$$

where

$$\phi(t,s) = \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}} \sum_{x\in\mathcal{X}} \sqrt{\mu(x)}\right), & C \leq \min_{x\neq x'} d(x,x') \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s} \sum_{x\in\mathcal{X}} \mu(x)^2\right)^{\frac{1}{2}}, & \text{else,} \end{cases}$$

and

$$\psi(s_{\mathcal{X}}) = \begin{cases} \left(2\sum_{x\in\mathcal{X}}(1-s_x)\right), & C \leq \min_{x\neq x'} d(x,x') \\ \left(\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}\right)^{\frac{1}{2}}, & \text{else.} \end{cases}$$

Notably, for $N \to \infty$, $t \to \infty$, $s \to 1$ and $s_{\mathcal{X}} \to \mathbf{1}_M$, these upper bounds vanish. This extends recent progress made on estimating OT distances in the finitely supported setting to unbalanced measures under the (p,C)-KRD. Inspired by partition strategies used to control the empirical deviation of the OT distance between probability measures [Dereich et al., 2013, Weed and Bach, 2019], the proof is based on a tree approximation of the underlying space $\mathcal{X}$ also used to control the empirical deviation of the OT distance between finitely supported measures [Sommerfeld et al., 2019]. Using a recently established closed form solution of the (p,C)-KRD on ultra-metric trees [Heinemann et al., 2021] we then obtain an upper bound on the expected error in (i) and (ii). To further extend this control to the context of the (p,C)-Fréchet functional and (p,C)-barycenter (3), we utilise certain structural properties of the (p,C)-barycenter. Let μ^* be any (p,C)-barycenter of the population measures $\mu^1,\dots,\mu^J$ and let $\hat{\mu}^*$ be any barycenter of the empirical measures $\hat{\mu}^1,\dots,\hat{\mu}^J$. From (i), we can derive a bound on the error in the Fréchet functional and in conjugation with theory from linear programming, we then extend this control to a more refined statement on the (p,C)-barycenters (see Theorem 3.2). The caveat is the fact that

neither μ^* nor $\hat{\mu}^*$ is necessarily unique. Thus, we need to control the error in terms of the respective optimal set $\mathbf{B}^*$ and its empirical counterpart $\hat{\mathbf{B}}^*$. For this, we consider the quantities

$$(iii)\ \mathbb{E}\left[F_{p,C}(\hat{\mu}^*) - F_{p,C}(\mu^*)\right], \quad (iv)\ \mathbb{E}\left[\sup_{\hat{\mu}^\star \in \hat{\mathbf{B}}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} KR^p_{p,C}(\mu^\star, \hat{\mu}^\star)\right].$$

The term (iv) is the expected deviation of the empirical barycenter which has the largest (p, C)-KRD to the set of barycenters of the population measures. Bounds on $(i) - (iv)$ do not only allow for statistical error control when the population measures are estimated from data, but they also enable randomised computations. If the size of the population measures is computationally infeasible, then empirical versions of these measures can be used as a proxy for the population level distances and barycenter. The bounds $(i) - (iv)$ allow to tune the problem size against the quality of the approximation. While theoretically all three models allow this approximation approach, the most suitable candidate is clearly the multinomial model. In this resampling scenario, the assumption of known total intensities is natural, as the population measures are known, but computationally infeasible, and the sample size provides a strict upper bound on the computational complexity of a given approximation. The latter does not hold true for the other two models, where the support size of the estimated measures can not be controlled as clearly. One alternative approach is to make use of a subsampling[5] method instead of a resampling one, i.e. to replace the i.i.d. samples $X_1, \ldots, X_N$ from μ by ones drawn without replacement. A natural choice of estimator in this scenario is

$$\hat{\mu}_N = \frac{1}{\sum_{i=1}^N \mu(X_i)} \sum_{i=1}^N \mu(X_i)\delta_{X_i}, \tag{9}$$

where the mass at each drawn location $x \in \mathcal{X}$ is proportional to the mass of the population measure at x and the total mass intensity is rescaled to the known, true total intensity. This estimator is, due to the sampling without replacement, guaranteed to have N support points, which yields close control on the required runtime for a given approximation. This approach has become popular within the machine learning community where it is referred to as mini-batch OT [Fatras et al., 2021].
We study the convergence properties of $(i) - (iv)$ for the three described models in terms of non-asymptotic deviation bounds and in extended simulation studies on a wide range of synthetic datasets. Finally, we test the performance of the resampling approach to approximate the (p, C)-KRD on a real dataset from confocal nanoscopy and compare its empirical performance to that of the subsampling approach described above[6].

2 Sampling Bounds for Kantorovich-Rubinstein Distances

In this section we replace the population measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ by their respective estimators. Exemplary, we investigate the Poisson model in detail. We provide theoretical guarantees for the accuracy of the approximation in terms of their *expected Kantorovich-Rubinstein deviation*. Results for the multinomial and Bernoulli model follow along the

[5] Here, subsampling refers to sampling without replacement, while resampling refers to sampling with replacement.

[6] We do not provide any deviation bounds for the subsampling approach, but just simulation results on a real data set. The dependency structure of the samples in this setting yields quite involved combinatorical problems, that make it complicated to derive deviation bounds for this model.

same reasoning. Corresponding deviation bounds and the proofs are provided in Appendix A and Appendix B.
The proof of the main result and the explicit construction of its constant relies on a tree approximation of the space $\mathcal{X}$. This approximation requires some *depth level* $L \in \mathbb{N}$ and *resolution* $q > 1$, based on which we construct minimal $q^{-j}\mathrm{diam}(\mathcal{X})$-coverings[7] on $\mathcal{X}$ to obtain our bounds. An illustration of this approximation is given in Figure 3.

2.1 Tree Approximation for the Kantorovich-Rubinstein Distance

Let $\mathcal{T} = (V, E)$ be a rooted, ultrametric tree with height function $h : V \longrightarrow \mathbb{R}_+$ and root r. For two nodes $\mathsf{u}, \mathsf{v} \in V$, denote the unique path between u and v in $\mathcal{T}$ by $\mathcal{P}(\mathsf{u}, \mathsf{v})$. For a node $\mathsf{v} \in V$ its *children* are the elements of the set $\mathcal{C}(\mathsf{v}) = \{\mathsf{w} \in V \mid \mathsf{v} \in \mathcal{P}(\mathsf{w}, \mathsf{r})\}$. The *parent* $par(\mathsf{v})$ of a node v is the unique node with $(par(\mathsf{v}), \mathsf{v}) \in E$ and $h(\mathsf{v}) < h(par(\mathsf{v}))$. For any $C > 0$, define the set

$$\mathcal{R}(C) \coloneqq \{\mathsf{v} \in V \mid h(\mathsf{v}) \leq C/2 < h(\mathrm{par}(\mathsf{v}))\} \tag{10}$$

with the convention that $\mathcal{R}(C) = \{\mathsf{r}\}$ if $\frac{C}{2} \geq h(\mathsf{r})$. The goal is to control the (p, C)-KRD on the finite metric space $(\mathcal{X}, d)$ by bounding it from above by a dominating distance $d_{\mathcal{T}}$ induced[8] from a tree $\mathcal{T}$ with the elements of $\mathcal{X}$ as vertices and a height function h such that $d(x, x') \leq d_{\mathcal{T}}(x, x')$. In this case and by the definition of the Kantorovich-Rubinstein distance it holds for all measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ that

$$\mathrm{KR}_{p,C}(\mu, \nu) \leq \mathrm{KR}_{d_{\mathcal{T}}^p,C}(\mu, \nu), \tag{11}$$

where $\mathrm{KR}_{d_{\mathcal{T}}^p,C}(\mu, \nu)$ denotes the (p, C)-KRD w.r.t. the ground space $(\mathcal{X}, d_{\mathcal{T}})$. Moreover, if $\mathcal{T}$ is an ultrametric tree with leaf nodes L and height function $h\colon V \to \mathbb{R}_+$ inducing[8] the tree metric $d_{\mathcal{T}}$ and the two measures $\mu^L, \nu^L \in \mathcal{M}_+(L)$ supported on the leaf nodes of $\mathcal{T}$, then it holds [Heinemann et al., 2021] that

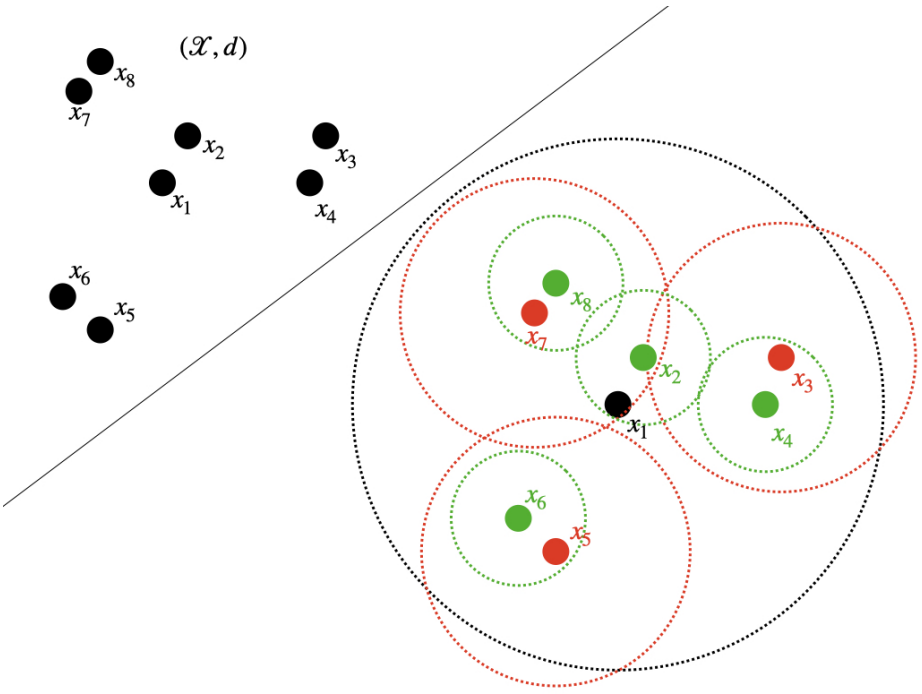
$$\begin{aligned}
&\mathrm{KR}^p_{d_{\mathcal{T}}^p,C}\left(\mu^L, \nu^L\right) = \\
&\sum_{\mathsf{v} \in \mathcal{R}(C)} \Bigg(2^{p-1} \sum_{\mathsf{w} \in \mathcal{C}(\mathsf{v}) \setminus \{\mathsf{v}\}} \Big(\left(h(par(\mathsf{w}))^p - h(\mathsf{w})^p\right) \left|\mu^L(\mathcal{C}(\mathsf{w})) - \nu^L(\mathcal{C}(\mathsf{w}))\right| \Big) \\
&+ \left(\frac{C^p}{2} - 2^{p-1} h(\mathsf{v})^p\right) \left|\mu^L(\mathcal{C}(\mathsf{v})) - \nu^L(\mathcal{C}(\mathsf{v}))\right| \Bigg).
\end{aligned} \tag{12}$$

The construction of $\mathcal{T}$, such that (11) holds, is as follows. Fix some depth level $L \in \mathbb{N}$. For some $q > 1$ and level $j = 0, \ldots, L$ define the covering set $Q_j \coloneqq \mathcal{N}(\mathcal{X}, q^{-j}\mathrm{diam}(\mathcal{X})) \subset \mathcal{X}$ and let $Q_{L+1} \coloneqq \mathcal{X}$. Any point $x \in Q_j$ is considered as a node at level j of a tree $\mathcal{T}$ and denoted as (x, j) to emphasise its level position. For level $j = 0$ this yields a single element in Q_0 which serves as the root of the tree. For $j = 0, \ldots, L$ a node (x, j) at level j is connected to one node $(x', j+1)$ at level $j+1$ if their distance satisfies $d(x, x') \leq q^{-j}\mathrm{diam}(\mathcal{X})$ (ties are broken arbitrarily). The edge weight of the corresponding edge is set equal to $q^{-j}\mathrm{diam}(\mathcal{X})$. Consequently, the height of each node only depends on its assigned level $0 \leq l \leq L+1$ and is defined as $h_{q,L}\colon \{0, \ldots, L+1\} \to \mathbb{R}$ by

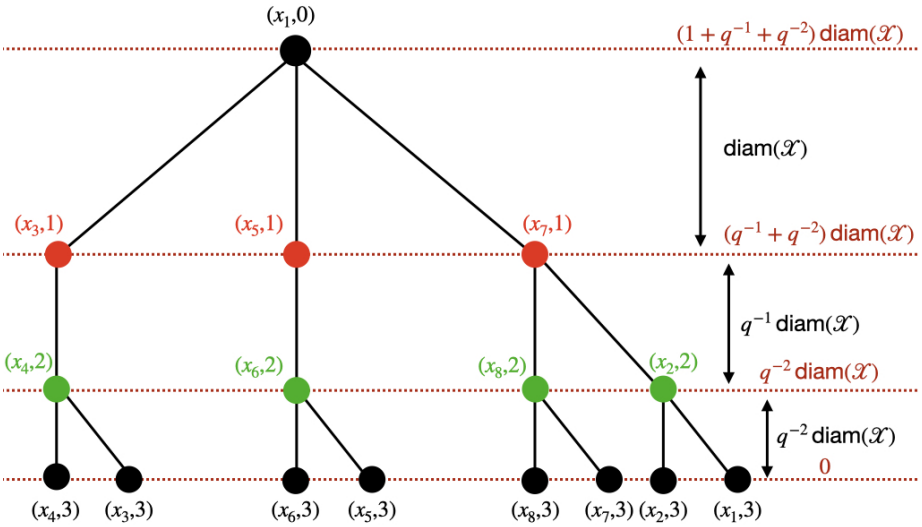
$$h_{q,L}(l) = \sum_{j=l}^{L} q^{-j}\mathrm{diam}(\mathcal{X}) = \frac{q^{1-l} - q^{-L}}{q-1}\mathrm{diam}(\mathcal{X}). \tag{13}$$

[7] For a metric space $(\mathcal{X}, d)$ an ϵ-cover is a set of points $\{x_1, \ldots, x_m\} \subset \mathcal{X}$ such that for each $x \in \mathcal{X}$, there exists some $1 \leq i \leq m$ such that $d(x, x_i) \leq \epsilon$. The smallest such set is denoted as $\mathcal{N}(\mathcal{X}, d, \epsilon)$.

[8] For two vertices of the tree $\mathcal{T}$, we define their distance $d_{\mathcal{T}}$ as the sum of the weights of the edges included in the unique path between the two vertices. Here, the weight of an edge joining two vertices v and $par(v)$ is given by $h(par(v)) - h(v)$.



(a)



(b)

Figure 3: **Ground metric approximation by an ultrametric tree distance: (a)** A finite metric space $(\mathcal{X}, d)$ and its covering sets Q_0 (black), Q_1 (red) and Q_2 (green) for $L = 2$. **(b)** Based on the covering sets from (a) an ultrametric tree is constructed. The metric space $\mathcal{X}$ is embedded in level $L + 1 = 3$ and equal to all leaf nodes of that tree.

By definition the space $\mathcal{X}$ is embedded in level $L+1$ as the leaf nodes of $\mathcal{T}$ with height $h_{q,L}(L+1) = 0$. By a straightforward computation it holds for two points $x, x' \in \mathcal{X}$ considered as embedded in $\mathcal{T}$ as $(x, L+1)$ and $(x', L+1)$ that

$$d^p(x, x') \leq d^p_{\mathcal{T}}((x, L+1), (x', L+1)). \tag{14}$$

The measures μ, ν are embedded into $\mathcal{T}$ as measures μ^L, ν^L supported only on leaf nodes of $\mathcal{T}$ and thus it follows from (14) that

$$\mathrm{KR}_{p,C}(\mu, \nu) \leq \mathrm{KR}_{d^p_{\mathcal{T}},C}\left(\mu^L, \nu^L\right).$$

In combination with the closed formula from (12) this yields an upper bound on the (p, C)-KRD. Whenever clear from the context the notation is alleviated by writing $\mathsf{v} \in Q_l$ instead of $(\mathsf{v}, l) \in Q_l$.

Lemma 2.1. *Let $(\mathcal{X}, d)$ be a finite metric space and let $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ with total mass $\mathbb{M}(\mu)$ and $\mathbb{M}(\nu)$, respectively. Let $p \geq 1$ and $C > 0$. Then for any resolution $q > 1$, depth $L \in \mathbb{N}$ and height function (13) with*

$$h_{q,L}(k) = \frac{q^{1-k} - q^{-L}}{q-1} diam(\mathcal{X})$$

it holds that

$$KR^p_{p,C}(\mu, \nu) \leq \begin{cases} \left(\frac{C^p}{2} - 2^{p-1} h_{q,L}(0)^p\right) |\mathbb{M}(\mu) - \mathbb{M}(\nu)| \\ +B_{q,p,L,\mathcal{X}}(1), \\ \qquad\qquad \text{if } C \geq 2h_{q,L}(0), \\ B_{q,p,L,\mathcal{X}}(l), \\ \qquad\qquad \text{if } 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2} TV(\mu, \nu), \\ \qquad\qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x \neq x'} d(x, x')\right), \end{cases}$$

where $TV(\mu, \nu) = \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$ is the total variation distance *and*

$$B_{q,p,L,\mathcal{X}}(l) = 2^{p-1} \sum_{j=l}^{L+1} \sum_{x \in Q_j} \left(h_{q,L}(j-1)^p - h_{q,L}(j)^p\right) \left|\mu^L(\mathcal{C}(x)) - \nu^L(\mathcal{C}(x))\right|.$$

2.2 (p, C)-Kantorovich-Rubinstein Deviation Bound

Given Lemma 2.1 based on the tree-approximation, we are able to state our main result explicitly.

Theorem 2.2. *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$. Let $\hat{\mu}_{t,s}$ be the estimator from (8). Then, for any $p \geq 1$, resolution $q > 1$ and depth $L \in \mathbb{N}$ it holds that*

$$\mathbb{E}\left[KR_{p,C}\left(\hat{\mu}_{t,s}, \mu\right)\right] \leq \mathcal{E}^{Poi}_{p,\mathcal{X},\mu}(C, q, L)^{1/p} \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}\right)^{\frac{1}{p}}, \\ \qquad\qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x \neq x'} d(x, x')\right) \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s} \sum_{x \in \mathcal{X}} \mu(x)^2\right)^{\frac{1}{2p}}. \\ \qquad\qquad \text{else.} \end{cases}$$

For

$$A_{q,p,L,\mathcal{X}}(l) := diam(\mathcal{X})^p 2^{p-1}\left(q^{-Lp}|\mathcal{X}|^{\frac{1}{2}} + \left(\frac{q}{q-1}\right)^p \sum_{j=l}^{L} q^{p-jp}|Q_j|^{\frac{1}{2}}\right),$$

the constant is equal to

$$\mathcal{E}^{Poi}_{p,\mathcal{X},\mu}(C,q,L) = \begin{cases} \left(\frac{C^p}{2} - 2^{p-1}\left(\frac{q-q^{-L}}{q-1} diam(\mathcal{X})\right)^p\right) + A_{q,p,L,\mathcal{X}}(1), \\ \qquad\qquad \text{if } C \geq 2h_{q,L}(0), \\ A_{q,p,L,\mathcal{X}}(l), \\ \qquad\qquad \text{if } 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2}, \\ \qquad\qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right), \end{cases}$$

Furthermore, for $p = 1$ the factor $\frac{q}{(q-1)}$ in $A_{q,1,L,\mathcal{X}}(l)$ can be removed for all $l = 1,\dots,L$.

The constant $\mathcal{E}^{\text{Poi}}_{p,\mathcal{X},\mu}(C,q,L)$ is reminiscent of the constants for similar deviation bounds for optimal transport between finitely supported measures [Sommerfeld et al., 2019]. However, in the case of UOT one finds an interesting case distinction into roughly three cases depending on the relation between the penalty parameter C of the (p,C)-KRD and the resolution q and depth L of the tree approximation. The different constants arise from the fact that C controls the maximal range at which transport occurs in an UOT plan. In particular, if $d(x,x') > C^p$, then for any UOT plan π it holds $\pi(x,x') = 0$. If C is sufficiently large, i.e. larger than the diameter of $\mathcal{X}$, then $\mathcal{E}^{\text{Poi}}_{p,\mathcal{X},\mu}(C,q,L)$ coincides with the deviation bounds for usual optimal transport, however, there is an additional summand arising from the necessary estimation of the true total mass $\mathbb{M}(\mu)$. For sufficiently small C, e.g. $C < \min_{x\neq x'} d(x,x')$, the (p,C)-KRD is proportional to the TV distance, hence we obtain a constant which is oblivious to the geometry of the ground space (see Lemma 2.5). For an intermediate value of C, the UOT problem on the ultra-metric tree $\mathcal{T}$ decomposes into smaller problems on subtrees of $\mathcal{T}$ (for details see the proof of (12) in Heinemann et al. [2021]) depending on C. The expected (p,C)-KRD error then depends on the size of these subtrees and the mass estimation error inherent to the total mass on these subtrees.

Remark 2.3. *Since the deviation bound holds for any resolution $q > 1$ and depth $L \in \mathbb{N}$ one can optimise and equivalently state upper bounds in terms of the infimum over those parameters. When the dependence on q or L is omitted, it is assumed that the infimum over those parameters has been taken, i.e.*

$$\mathcal{E}^{Poi}_{p,\mathcal{X},\mu}(C) = \inf_{L\in\mathbb{N},q>1} \mathcal{E}^{Poi}_{p,\mathcal{X},\mu}(C,q,L).$$

From the reverse triangle inequality we immediately obtain the following corollary from Theorem 2.2.

Corollary 2.4. *Let $(\mathcal{X},d)$ be a finite metric space and $\mu,\nu \in \mathcal{M}_+(\mathcal{X})$. Let $\hat{\mu}_{t,s},\hat{\nu}_{t,s}$ be the estimator from (8) for each of these measures, respectively. Then, for any $p \geq 1$, resolution $q > 1$ and depth $L \in \mathbb{N}$ it holds that*

$$\mathbb{E}\left[\left|KR_{p,C}\left(\hat{\mu}_{t,s},\hat{\nu}_{t,s}\right) - KR_{p,C}\left(\mu,\nu\right)\right|\right] \leq$$
$$2\mathcal{E}^{Poi}_{p,\mathcal{X},\mu}(C,q,L)^{1/p} \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}\right)^{\frac{1}{2p}}, & \text{if } C < \min_{x,x'\in\mathcal{X}} d(x,x'), \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}\mu(x)^2\right)^{\frac{1}{2p}}, & \text{else,} \end{cases}$$

where $\mathcal{E}^{Poi}_{p,\mathcal{X},\mu}(C,q,L)$ *is defined as in Theorem 2.2.*

2.3 Rate Optimality

In the following, we provide some intuition and discussion of the deviation bound provided in Theorem 2.2. The term $\phi(s,t)$ must necessarily contain a sum of a term depending on s and a term depending on t. In particular, neither choosing $s = 1$ nor letting t go to infinity for $s < 1$, would yield a zero error. For any fixed $t > 0$ and $s = 1$ the expected (p, C)-KRD error is clearly non-zero as the mass of the measures at each location is in general not estimated correctly. Similarly, for any fixed $s < 1$ letting $t \to \infty$ can not yield an expected (p, C)-KRD error of zero as on average $(1-s)|\mathcal{X}| > 0$ support points of μ are not observed. However, for $s = 1$ the error vanishes for $t \to \infty$, as we observe all support points of μ and then the strong law of large numbers guarantees the convergence of the weights at each location. It remains to verify whether the rate in t is optimal. For this, fix $s = 1$ and observe that

$$\min\{C, \min_{x\neq x'} d(x,x')\}^p TV(\mu,\nu) \leq KR^p_{p,C}(\mu,\nu) \leq \min\{C, \operatorname{diam}(\mathcal{X})\}^p TV(\mu,\nu). \tag{15}$$

Hence, it suffices to show that $t^{-\frac{1}{2}}$ is the optimal rate for the convergence in total variation distance. It holds

$$\mathbb{E}\left[TV(\mu,\hat{\mu}_{t,1})\right] = \sum_{x\in\mathcal{X}} \mathbb{E}\left[|\mu(x) - \hat{\mu}_{t,1}(x)|\right] = \sum_{x\in\mathcal{X}} t^{-1}\mathbb{E}\left[|\mathbb{E}[P_x] - P_x|\right],$$

where $P_x \sim Poi(\mu(x)t)$. Using the closed form solution of the mean absolute deviation of a Poisson random variable [Ramasubban, 1958] and Stirling's formula, we obtain the asymptotic equivalence (in the sense that their ratio converges to one as $t \to \infty$)

$$\mathbb{E}\left[TV(\mu,\hat{\mu}_{t,1})\right] \approx \sum_{x\in\mathcal{X}} (t\mu(x))^{t\mu(x)-\lfloor t\mu(x)\rfloor} t^{-1/2}\sqrt{\mu(x)}\sqrt{2/\pi}.$$

For $t > \inf_{x\in\operatorname{supp}(\mu)} \mu(x)$, this is bounded from below by

$$\left(\sqrt{2/\pi}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}\right) t^{-1/2}.$$

All combined, the expectation $\mathbb{E}\left[KR^p_{p,C}(\mu,\hat{\mu}_{t,1})\right]$ has a lower bound which is asymptotically equivalent to $t^{-\frac{1}{2}}$ and hence the rate in t of Theorem 2.2 is sharp.

2.4 Explicit Bounds for Euclidean Spaces

While the constants in the previous theorem are valid on arbitrary metric spaces, more explicit bounds can be derived for many practical applications. On Euclidean spaces we can follow along the lines of the proofs in Sommerfeld et al. [2019] (see their appendix for details) and obtain more explicit upper bounds on these constants.
Thus, assume that $\mathcal{X} \subset \mathbb{R}^D$ and that d is the Euclidean distance d_2. We fix $q = 2$, since it minimises the constants from Theorem 2.2 for all choices of q on the interval $[2,\infty)$. Let $d_\infty(x,x') = \max_{d=1,\dots,D}|x_d - x'_d|$ be the uniform distance and $\operatorname{diam}_\infty(\mathcal{X})$ be the diameter of $\mathcal{X}$. Using the fact that

$$d_2(x,x') \leq \sqrt{D} d_\infty(x,x'),$$

we bound the (p, C)-KRD with respect to the Euclidean distance by the (p, C)-KRD with respect to the uniform distance at a price of a factor of $\sqrt{D}$. In particular, we also apply this to the definition of the height function (13) given for $q = 2$ and any $l \in \{0, \dots, L\}$ by

$$h_L(l) = \left(2^{1-l} - 2^{-L}\right) \operatorname{diam}_\infty(\mathcal{X}).$$

Within this framework, we can compute explicit upper bounds on the constants in Theorem 2.2. For $D < 2p$ and $L \to \infty$, it holds

$$\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C) \le D^{p/2} \begin{cases} \frac{C^p}{2\sqrt{D}} - 2^{2p-1}\operatorname{diam}^p_\infty(\mathcal{X}) + \operatorname{diam}^p_\infty(\mathcal{X}) 2^{3p-1} \frac{2^{D/2-p}}{1-2^{D/2-p}}, \\ \qquad \text{if } C \ge 2h_L(0), \\ \operatorname{diam}^p_\infty(\mathcal{X}) 2^{3p-1} \frac{2^{l(D/2-p)}}{1-2^{D/2-p}}, \\ \qquad \text{if } 2h_L(l) \le C < 2h_L(l-1), \\ \frac{C^p}{2}, \\ \qquad \text{if } C \le \left(2h_L(L) \wedge \min_{x \neq x'} d_\infty(x, x')\right). \end{cases}$$

For $D = 2p$ and $L = \lfloor \frac{1}{D} \log_2(|\mathcal{X}|) \rfloor$, it holds

$$\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C) \le D^{p/2} \begin{cases} \frac{C^p}{2\sqrt{D}} - 2^{p-1}\left(2 - |\mathcal{X}|^{-\frac{1}{D}}\right)^p \operatorname{diam}^p_\infty(\mathcal{X}) \\ + \operatorname{diam}^p_\infty(\mathcal{X}) 2^{3p-1}(2^{-2p} + D^{-1}\log_2(|\mathcal{X}|)), \\ \qquad \text{if } C \ge 2h_L(0), \\ \operatorname{diam}^p_\infty(\mathcal{X}) 2^{3p-1}\left(2^{-2p} + D^{-1}\log_2(|\mathcal{X}|) - l\right), \\ \qquad \text{if } 2h_L(l) \le C < 2h_L(l-1), \\ \frac{C^p}{2}, \\ \qquad \text{if } C \le \left(2h_L(L) \wedge \min_{x \neq x'} d_\infty(x, x')\right). \end{cases}$$

For $D > 2p$ and $L = \lfloor \frac{1}{D} \log_2(|\mathcal{X}|) \rfloor$, it holds

$$\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C) \le D^{p/2} \begin{cases} \frac{C^p}{2\sqrt{D}} - 2^{p-1}\left(2 - |\mathcal{X}|^{-\frac{1}{D}}\right)^p \operatorname{diam}^p_\infty(\mathcal{X}) \\ + \operatorname{diam}^p_\infty(\mathcal{X}) 2^{p-1} |\mathcal{X}|^{\frac{1}{2}-\frac{p}{D}} \left(1 + \frac{2^{p+D/2}}{2^{D/2-p}-1}\right), \\ \qquad \text{if } C \ge 2h_L(0), \\ \operatorname{diam}^p_\infty(\mathcal{X}) \Bigg(|\mathcal{X}|^{\frac{1}{2}-\frac{p}{D}} \\ + \frac{2^{p+D/2}}{2^{D/2-p}-1} \left(|\mathcal{X}|^{\frac{1}{2}-\frac{p}{D}} - 2^{(D/2-p)(l-1)}\right) \Bigg), \\ \qquad \text{if } 2h_L(l) \le C < 2h_L(l-1), \\ \frac{C^p}{2}, \\ \qquad \text{if } C \le \left(2h_L(L) \wedge \min_{x \neq x'} d_\infty(x, x')\right). \end{cases}$$

These bounds distinguish cases, based on the value of the penalty C, as well as the dimension D. The most critical part of these these bounds is the impact of the cardinality of $\mathcal{X}$. If $D < 2p$, then there is no dependence on $|\mathcal{X}|$ and the convergence of approximation error of the empirical measure is independent of its support size. If $D = 2p$, then $|\mathcal{X}|$ enters through a logarithmic term. If $D > 2p$, then the dependence becomes polynomial in $|\mathcal{X}|$. Though, we stress that this polynomial scaling is always superior to the scaling of $|\mathcal{X}|^{\frac{1}{2}}$ which one would obtain from a naive total variation bound (recall (15)). These phase transitions for the dependence on the support size match those for empirical optimal transport [Sommerfeld et al., 2019]. A novelty for the UOT setting is the fact that additionally we between different scales of C. This is explained by the previously discussed control of C on the maximal distance at which transport occurs in an optimal plan. The height function is again used to specify the scale induced by a particular choice of the parameter C. Notably, the dependence on $|\mathcal{X}|$ does not change on most scales of C. There is an exception, however, for sufficiently small values of C, where the (p, C)-KRD is equal to a scaled total variation distance. Thus, these bounds are completely oblivious to the geometry of $\mathcal{X}$ in $\mathcal{Y}$, though they scale as $|\mathcal{X}|^{\frac{1}{2}}$. As a final observation, we note that for $C > 2h_L(0)$, these bounds essentially recover analog bounds for empirical optimal transport. However, for the (p, C)-KRD the bounds include an additional summand based on the estimation error for the measure's total mass intensity.

2.5 Proofs

As a preparatory step to prove Theorem 2.2, we treat the significantly simpler case of an empirical deviation bound with respect to the total variation distance.

Lemma 2.5 (Total Variation Bound)**.** *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with total mass $\mathbb{M}(\mu)$. Let $\hat{\mu}_{t,s}$ be the estimator from (8). Then, for any $p \geq 1$ it holds that*

$$\mathbb{E}\left[KR^p_{p,C}\left(\hat{\mu}_{t,s}, \mu\right)\right] \leq \frac{C^p}{2}\left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}\right).$$

Proof. Let $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$. By retaining all common mass between μ and ν at place and delete (resp. create) excess mass (resp. deficient mass) we obtain a feasible solution for (2) with objective value in terms of a total variation distance between μ and ν. Thus, it holds

$$\mathrm{KR}^p_{p,C}\left(\mu, \nu\right) \leq \frac{C^p}{2}\mathrm{TV}\left(\mu, \nu\right).$$

In particular, this holds for $\nu = \hat{\mu}_{t,s}$. Taking expectations yields

$$\begin{aligned}
\mathbb{E}\left[\mathrm{TV}\left(\hat{\mu}_{t,s}, \mu\right)\right] &= \frac{1}{st}\sum_{x\in\mathcal{X}}\mathbb{E}\left[|P_x B_x - st\mu(x)|\right] \\
&= \frac{1}{st}\sum_{x\in\mathcal{X}} s\mathbb{E}\left[|P_x - st\mu(x)|\right] + (1-s)st\mu(x) \\
&\leq \frac{1}{st}\sum_{x\in\mathcal{X}} s(1-s)\mathbb{E}\left[P_x\right] + s^2\mathbb{E}\left[|P_x - t\mu(x)|\right] + (1-s)st\mu(x) \\
&\leq \frac{1}{st}\sum_{x\in\mathcal{X}} 2s(1-s)t\mu_x + s^2\sqrt{t}\sqrt{\mu(x)} \\
&= 2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}.
\end{aligned}$$

□

With Lemma 2.1 and Lemma 2.5 at our disposal we are able to prove Theorem 2.2.

Proof of Theorem 2.2. Let $\hat{\mu}_{t,s}$ be the estimator from (8). We fix $p = 1$ and detail the case $p > 1$ at the end of the proof. Suppose first that $C \leq \min_{x \neq x'} d(x, x')$. According to [Heinemann et al., 2021, Theorem 2.2 (ii)] it holds that

$$\mathbb{E}\left[\mathrm{KR}_{1,C}(\hat{\mu}_{t,s}, \mu)\right] = \frac{C}{2}\mathbb{E}\left[\sum_{x \in \mathcal{X}} |\hat{\mu}_{t,s} - \mu(x)|\right] = \frac{C}{2}\mathbb{E}\left[\mathrm{TV}(\hat{\mu}_{t,s}, \mu)\right].$$

This yields the total variation bounds (see Lemma 2.5). Next, consider the tree approximation as outlined in Section 2.1 and construct an ultrametric tree $\mathcal{T}$ such that $\mathrm{KR}_{1,C}(\hat{\mu}_{t,s}, \mu) \leq \mathrm{KR}_{d_{\mathcal{T}},C}\left(\hat{\mu}^L_{t,s}, \mu^L\right)$. Applying Lemma 2.1 for $p = 1$ where by definition the difference of height function is equal to

$$h_{q,L}(j-1) - h_{q,L}(j) = \frac{\mathrm{diam}(\mathcal{X})}{q-1}\left(q^{2-j} - q^{1-j}\right) = \mathrm{diam}(\mathcal{X}) q^{1-j}$$

and yields the upper bound

$$\mathbb{E}\left[\mathrm{KR}_{1,C}(\hat{\mu}_{t,s}, \mu)\right] \leq \mathbb{E}\left[\mathrm{KR}_{d_{\mathcal{T}},C}\left(\hat{\mu}^L_{t,s}, \mu^L\right)\right]$$

$$= \begin{cases} \left(\frac{C}{2} - h_{q,L}(0)\right)\mathbb{E}\left[|\mathbb{M}(\hat{\mu}_{t,s}) - \mathbb{M}(\mu)|\right] \\ +\mathrm{diam}(\mathcal{X}) \sum_{j=1}^{L+1} q^{1-j} \sum_{x \in Q_j} \mathbb{E}\left[\left|\hat{\mu}^L_{t,s}(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right], & C \geq 2h_{q,L}(0) \\ \mathrm{diam}(\mathcal{X}) \sum_{j=l}^{L+1} q^{1-j} \sum_{x' \in Q_j} \mathbb{E}\left[\left|\hat{\mu}^L_{t,s}(\mathcal{C}(x')) - \mu^L(\mathcal{C}(x'))\right|\right], & 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2}\mathbb{E}\left[\mathrm{TV}(\hat{\mu}_{t,s}, \mu)\right], & C \leq \left(2h_{q,L}(L) \vee \min_{x \neq x'} d(x, x')\right). \end{cases}$$

For the estimator from (8) with $B_x \sim \mathrm{Ber}(s)$ and $P_x \sim \mathrm{Poi}(t\mu(x))$ for all $x \in \mathcal{X}$, it holds

$$\begin{aligned} &\sum_{x \in Q_l} \mathbb{E}\left[\left|\hat{\mu}^L_{t,s}(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right] \\ &= \sum_{x \in Q_l} \frac{1}{st}\mathbb{E}\left[\left|\sum_{y \in \mathcal{C}(x)} P_y B_y - st \sum_{y \in \mathcal{C}(x)} \mu(y)\right|\right] \\ &\leq \sum_{x \in Q_l} \frac{1}{st}\sqrt{\mathrm{Var}\left(\sum_{y \in \mathcal{C}(x)} P_y B_y\right)} = \sum_{x \in Q_l} \frac{1}{st}\sqrt{\sum_{y \in \mathcal{C}(x)} \mathrm{Var}(P_y B_y)} \\ &= \sum_{x \in Q_l} \frac{1}{st}\sqrt{\sum_{y \in \mathcal{C}(x)} s(1-s)t\mu(y) + s(1-s)\mu(y)^2 + t\mu(y)s^2} \\ &= \sum_{x \in Q_l} \sqrt{\frac{1-s}{st}\sum_{y \in \mathcal{C}(x)} \mu(y) + \frac{1-s}{s}\sum_{y \in \mathcal{C}(x)} \mu(y)^2 + \frac{1}{t}\sum_{y \in \mathcal{C}(x)} \mu(y)} \end{aligned}$$

$$= \sum_{x\in Q_l} \sqrt{\frac{1}{st}\mu\left(\mathcal{C}(x)\right) + \frac{1-s}{s}\sum_{y\in\mathcal{C}(x)}\mu(y)^2}$$
$$\leq \sqrt{|Q_l|}\sqrt{\frac{1}{st}\sum_{x\in Q_l}\mu(\mathcal{C}(x)) + \frac{1-s}{s}\sum_{x\in Q_l}\sum_{y\in\mathcal{C}(x)}\mu(y)^2}$$
$$= \sqrt{|Q_l|}\sqrt{\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}\mu(x)^2}$$

Following an analogous computation one bounds the estimation error for the total mass intensity as

$$\mathbb{E}\left[|\mathbb{M}(\hat{\mu}_{t,s}) - \mathbb{M}(\mu)|\right] \leq \sqrt{\mathrm{Var}(\mathbb{M}(\hat{\mu}_{t,s}))} \leq \sqrt{\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}\mu(x)^2}.$$

Applying both of these bounds to the previous upper bound on the (p, C)-KRD in Lemma 2.1 yields the claim.
For $p > 1$, we first observe again that if $C \leq \min_{x\neq x'} d(x, x')$ then according to Heinemann et al. [2021] it holds that

$$\mathbb{E}\left[\mathrm{KR}^p_{p,C}(\hat{\mu}_{t,s}, \mu)\right] = \frac{C^p}{2}\mathbb{E}\left[\mathrm{TV}(\hat{\mu}_{t,s}, \mu)\right]$$

which yields the total variation bounds. For more general C, we simple repeat the previous calculations with the upper bounds on the difference of height function $h_{q,L}(j-1)^p - h_{q,L}(j)^p \leq \mathrm{diam}(\mathcal{X})^p \left(\frac{q}{q-1}\right)^p q^{p-jp}$. Since $h_{q,L}(L+1) = 0$ we also have $h_{q,L}(L)^p - h_{q,L}(L+1)^p = \mathrm{diam}(\mathcal{X})^p q^{-Lp}$. The expectations are bounded identically as before. Finally, using Jensen's inequality to bound

$$\mathbb{E}\left[KR_{p,C}(\mu, \hat{\mu}_{t,s})\right] \leq \left(\mathbb{E}\left[KR^p_{p,C}(\mu, \hat{\mu}_{t,s})\right]\right)^{\frac{1}{p}}$$

finishes the proof. □

Remark 2.6. *Omitting Jensen's inequality in the last step of the proof of Theorem 2.2 implies the slightly stronger result*

$$\mathbb{E}\left[KR^p_{p,C}\left(\hat{\mu}_{t,s}, \mu\right)\right] \leq \mathcal{E}^{Poi}_{p,\mathcal{X},\mu}(C, q, L) \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}\right), \\ \qquad\qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x, x')\right), \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}\mu(x)^2\right)^{\frac{1}{2}}, \\ \qquad\qquad \text{else.} \end{cases}$$

3 Empirical Kantorovich-Rubinstein Barycenters

Consider measures $\mu^1, \ldots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ which we replace by $\hat{\mu}^1_{t_1,s_1}, \ldots, \hat{\mu}^J_{t_J,s_J} \in \mathcal{M}_+(\mathcal{X})$ as defined in (8). We again focus on the Poisson model and treat the remaining two models in the appendix. The previous upper bound on the Kantorovich-Rubinstein distance in Theorem 2.2 between a measure and its empirical version is used to achieve control on

the mean absolute deviation of (p, C)-barycenters in terms of their p-Fréchet functional $F_{p,C}(\mu) = \frac{1}{J}\sum_{i=1}^{J} \mathrm{KR}^p_{p,C}(\mu^i, \mu)$ from (3). We denote

$$\mu^\star \in \operatorname*{arg\,min}_{\mu\in\mathcal{M}_+(\mathcal{Y})} \frac{1}{J}\sum_{i=1}^{J} \mathrm{KR}^p_{p,C}(\mu^i, \mu), \quad \hat{\mu}^\star \in \operatorname*{arg\,min}_{\mu\in\mathcal{M}_+(\mathcal{Y})} \frac{1}{J}\sum_{i=1}^{J} \mathrm{KR}^p_{p,C}(\hat{\mu}^i_{t_i,s_i}, \mu)$$

and measure the accuracy of approximation of $\mu^\star$ by $\hat{\mu}^\star$ in terms of their mean absolute p-Fréchet deviation.

Theorem 3.1. *Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = supp(\mu^i)$ for $i = 1, \dots, J$. Consider random estimators $\hat{\mu}^1_{t_1,s_1}, \dots, \hat{\mu}^J_{t_J,s_J} \in \mathcal{M}_+(\mathcal{X})$ derived from (8). Then,*

$$\mathbb{E}\left[|F_{p,C}(\mu^\star) - F_{p,C}(\hat{\mu}^\star)|\right] \le \frac{2p\min\{\mathrm{diam}(\mathcal{Y}), C\}^{p-1}}{J}\sum_{i=1}^{J}\mathcal{E}^{Poi}_{1,\mathcal{X}_i,\mu^i}(C)\phi(t_i, s_i),$$

where ϕ is given by

$$\phi(t,s) = \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}\right), & C \le \min_{x\neq x'} d(x,x') \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}\mu(x)^2\right)^{\frac{1}{2}}, & \text{else.} \end{cases}$$

A more elaborate statement gives control over the set of empirical (p, C)-barycenters itself. This involves a related linear program that is presented in detail in Appendix C.

Theorem 3.2. *Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = supp(\mu^i)$ for $i = 1, \dots, J$. Consider random estimators $\hat{\mu}^1_{t_1,s_1}, \dots, \hat{\mu}^J_{t_J,s_J} \in \mathcal{M}_+(\mathcal{X})$ derived from (8). Let $\mathbf{B}^\star$ be the set of (p, C)-barycenters of $\mu^1, \dots, \mu^J$ and $\hat{\mathbf{B}}^\star$ the set of (p, C)-barycenters of $\hat{\mu}^1_{t_1,s_1}, \dots, \hat{\mu}^J_{t_J,s_J}$. Then, for $p \ge 1$ it holds that*

$$\mathbb{E}\left[\sup_{\hat{\mu}^\star\in\hat{\mathbf{B}}^\star}\inf_{\mu^\star\in\mathbf{B}^\star} KR^p_{p,C}(\mu^\star, \hat{\mu}^\star)\right] \le \frac{p\min\{\mathrm{diam}(\mathcal{Y}), C\}^{p-1}}{V_P J}\sum_{i=1}^{J}\mathcal{E}^{Poi}_{1,\mathcal{X}_i,\mu^i}(C)\phi(t_i, s_i),$$

where ϕ is defined as in Theorem 3.1. The constant V_P is strictly positive and given by

$$V_P \coloneqq V_P(\mu^1, \dots, \mu^J) := (J+1)\mathrm{diam}(\mathcal{X})^{-p}\min_{v\in V\setminus V^\star}\frac{c^T v - f^*}{d_1(v, \mathcal{M})},$$

where V is the set of feasible vertices from the linear program in Appendix C, $V^\star$ is the subset of optimal vertices, c is the cost vector of the program, $f^\star$ is the optimal value, $\mathcal{M}$ is the set of minimisers of the linear program (19) and $d_1(x, \mathcal{M}) = \inf_{y\in\mathcal{M}}\|x - y\|_1$.

The proofs of Theorem 3.1 and Theorem 3.2 are deferred to Appendix D.

Remark 3.3. *For $J = 1$ and any $p \ge 1, C > 0$ the (p, C)-barycenter of μ^1 is just μ^1. Thus, the optimal value of the Fréchet functional is zero and it holds*

$$F(\hat{\mu}^*) - F(\mu^*) = KR^p_{p,C}(\mu^1, \hat{\mu}^1_{t,s}).$$

Consequently, it also holds

$$\sup_{\hat{\mu}^\star\in\hat{\mathbf{B}}^\star}\inf_{\mu^\star\in\mathbf{B}^\star} KR^p_{p,C}(\mu^*, \hat{\mu}^*) = KR^p_{p,C}(\mu^1, \hat{\mu}^1_{t,s}).$$

Thus, the rate for the convergence of the (p, C)-barycenter of the empirical measures, can in general not be faster than the convergence rate of a single estimator. In particular, the rates in t in Theorem 3.1 and Theorem 3.2 are sharp.

4 Simulations

In this section we investigate empirically the decay in the expected error for the Poisson model for measures within $\mathcal{X} \subset [0,1]^2$. For the (p,C)-KRD we consider two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ and the *relative* (p,C)-*KRD error*[9]

$$\mathbb{E}\left[\left|\frac{KR_{p,C}(\hat{\mu}_{t,s}, \hat{\nu}_{t,s}) - KR_{p,C}(\mu,\nu)}{KR_{p,C}(\mu,\nu)}\right|\right]. \tag{16}$$

For the setting of barycenters we consider the *relative* (p,C)-*Fréchet error*[9]

$$\mathbb{E}\left[\frac{F_{p,C}(\hat{\mu}^*) - F_{p,C}(\mu^*)}{F_{p,C}(\mu^*)}\right]. \tag{17}$$

In both cases, the relative error allows for easier comparisons between models than the absolute error. Additionally, for the (p,C)-barycenter (17) is readily available from simulations, while numerically considering the quantity in Theorem 3.2 is difficult, as it requires all optimal solutions instead of a single one.

4.1 Synthetic Datasets

We consider eight types of measures for our simulations. Let us fix some notation. Let $J \in \mathbb{N}$ be the number of measures generated. Let $U[0,1]^2$ denote the uniform distribution and let $\mathrm{Poi}(\lambda)$ denote a Poisson distribution with intensity λ. In all eight settings considered below, the measures are of the form

$$\mu^i = \sum_{k=1}^{K_i} w_k^i \delta_{l_k^i}$$

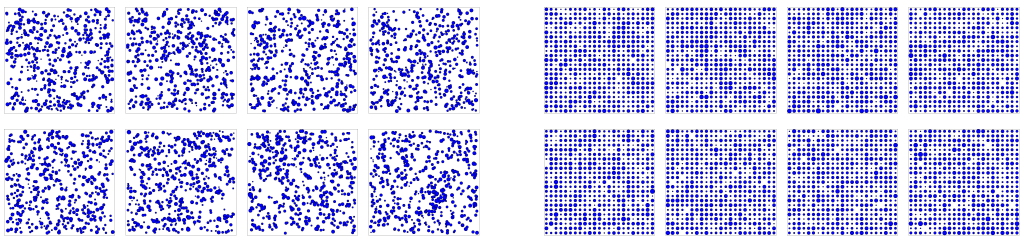
for some weights w_k^i, locations l_k^i and $K_i \in \mathbb{N}$. If $K_i = K_j$ for all $i,j = 1,\dots,J$, then we omit the index and denote the number of points by K. Note, that all measures have been constructed to have their support included in $[0,1]^2$.

Poisson Intensities on Uniform Positions (PI), see Figure 4 (a)

Let $w_1^i, \dots, w_K^i \sim \mathrm{Poi}(\lambda)$ for some intensity $\lambda > 0$ and $l_1^i, \dots, l_K^i \sim U[0,1]^2$ for $1 \le i \le J$.

Poisson Intensities on a Grid (PIG), see Figure 4 (b)

Set $K = M^2$ for $M \in \mathbb{N}$ and let $w_1^i, \dots, w_{M^2}^i \sim \mathrm{Poi}(\lambda)$ and $l_1^i, \dots, l_{M^2}^i$ be the location of an equidistant $M \times M$ grid in $[0,1]^2$ for $1 \le i \le J$.



(a) (b)

Figure 4: **(a)** An example of $J = 8$ measures from the *PI* dataset with $M = 500$ and $\lambda = 5$. **(b)** An example of $J = 8$ measures from the *PIG* dataset with $M = 23^2$ and $\lambda = 5$.

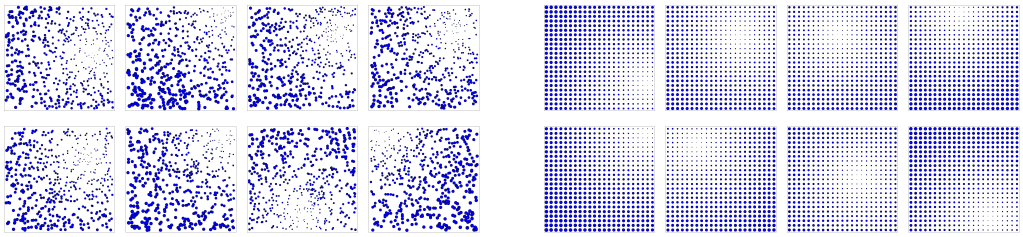
[9] We define $0/0 := 0$.

Norm-based Intensities on Uniform Positions (NI), see Figure 5 (a)

Fix J locations $l_0^1, \dots, l_0^J \in [0,1]^2$. Let $l_1^i, \dots, l_K^i \sim U[0,1]^2$ and let $w_k^i = \|l_k^i - l_0^i\|_2$ for $1 \leq i \leq J$.

Norm-based Intensities on a Grid (NIG), see Figure 5 (b)

Let $K = M^2$ for $M \in \mathbb{N}$. Fix J locations $l_0^1, \dots, l_0^J \in [0,1]^2$ and let $l_1^i, \dots, l_{M^2}^i$ be the location of an equidistant $M \times M$ grid in $[0,1]^2$ for each $1 \leq i \leq J$. Set $w_k^i = \|l_k^i - l_0^i\|_2$ for $1 \leq i \leq J$.



(a) (b)

Figure 5: **(a)** An example of $J = 8$ measures from the *NI* dataset with $M = 500$. **(b)** An example of $J = 8$ measures from the *NIG* dataset with $M = 23^2$.

Nested Ellipses (NE), see Figure 6 (a)

Let $G_1, \dots, G_J \sim U\{1,2,3,4,5\}$ and let $K_i = MG_i$ for $M \in \mathbb{N}$. Set all w_k^i equal to 1 for each $1 \leq k \leq K_i$ for $i = 1, \dots, J$. Let $t_1, \dots, t_M$ be a discretization of $[0, 2\pi]$. Let $U_1^i, \dots, U_K^i, V_1^i, \dots, V_K^i \sim U[0.2, 1]$. For $1 \leq i \leq J$, set

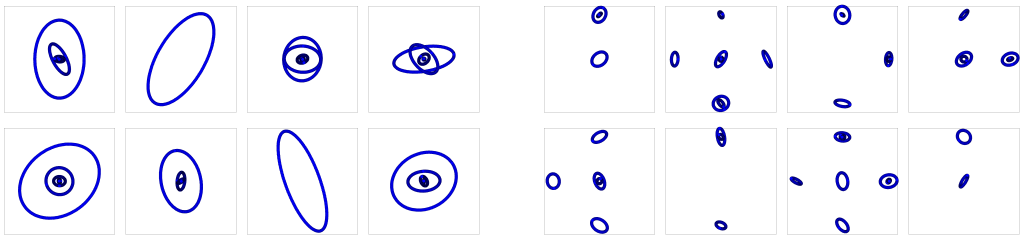
$$l^i_{M(j_i-1)+k} = 0.5(1 + 3^{-j}(U_{M(j_i-1)+k} \sin(t_k), V_{M(j_i-1)+k} \cos(t_k))^T), \quad j_i = 0, \dots, G_i.$$

Clustered Nested Ellipses (NEC), see Figure 6 (b)

Let $G_1^c, \dots, G_J^c \sim \text{Poi}(\lambda_c)$ for $c = 1, \dots, 5$. Let $\lambda_3 = 2$ and set $\lambda_c = 1$ else. Let $K_i = M \sum_{c=1}^5 G_i^c$. Set w_k^i equal to 1 for $1 \leq k \leq K_i$ for $i = 1, \dots, J$. Let $t_1, \dots, t_M$ be a discretization of $[0, 2\pi]$. Let $U_1^i, \dots, U_K^i, V_1^i, \dots, V_K^i \sim U[0.2, 1]$. Let $\alpha = (2, 12, 12, 22, 12)^T$ and $\beta = (12, 2, 12, 12, 22)^T$. Set for $c = 1, \dots, 5$ and $j_i = 0, \dots, G_i^c$

$$l^i_{M(\sum_{r=1}^{c-1} G_i^r)+M(j_i-1)+k} = \frac{1}{24}((3^{-j} U_{M(j_i-1)+k} \sin(t_k) + \alpha_c, 3^{-j} V_{M(j_i-1)+k} \cos(t_k) + \beta_c))^T,$$

where we use the convention that a sum is zero if its last index is smaller than its first one.



(a) (b)

Figure 6: **(a)** An example of $J = 8$ measures from the *NE* dataset with $M = 200$. **(b)** An example of $J = 8$ measures from the *NEC* dataset with $M = 95$.

Spirals of varying Length (SPI), Figure 7 (a)

Let $a_i \sim U[2, 4]$ and $b_i \sim U[3, 6]$ for $i = 1, \dots, J$. Let $K_i = \lceil b_i M \rceil$ and let $t_1, \dots, t_K$ be a discretization of $[0, b\pi]$. Set $w_k^i = 1$ for $k = 1, \dots, K_i$ and $i = 1, \dots, J$. Set

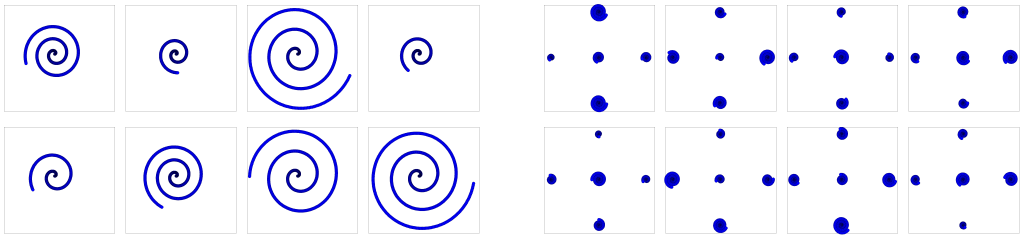
$$l_k^i = a_i((t_k \sin(t_k) + 64)/140, (t_k \cos(t_k) + 70)/130)^T.$$

Clustered Spirals (SPIC), see Figure 7 (b)

Let $a_i^c \sim U[2, 4]$ and $b_i^c \sim U[3, 6]$ for $i = 1, \dots, J$, $c = 1, \dots, 5$. Let $K_i^c = \lceil b_i^c M \rceil$ and let $t_1, \dots, t_K$ be a discretization of $[0, b\pi]$. Set $w_k^i = 1$ for $k = 1, \dots, K_i$ and $i = 1, \dots, J$ and let $\alpha = (0, 3, 3, 6, 3)^T$ and $\beta = (3, 0, 3, 3, 6)^T$. Set

$$l^i_{\sum_{r=1}^{c-1} K_i^r + k} = (1/7)(((a_i^c t_k \sin(t_k) + 64)/140) + \alpha_c, (a_i^c t_k \cos(t_k) + 70)/130 + \beta_c)^T,$$

where we again use the convention that a sum is zero if its last index is smaller than its first one.



(a) (b)

Figure 7: **(a)** An example of $J = 8$ measures from the *SPI* dataset with $M = 110$. **(b)** An example of $J = 8$ measures from the *SPIC* dataset with $M = 22$.

4.2 Simulation Results for the $(2, C)$-Kantorovich-Rubinstein Distance

In the following, we discuss the results from our simulation studies for the Poisson model for the $(2, C)$-KRD between two measures within one of the eight classes of measures introduced above. Note, that for brevity the plots for some of the classes have been omitted to Appendix E and exemplary the classes PI, NE and NEC are discussed. The remaining plots in the Figures 19, 20, 21, 22 and 23 are found in Appendix E.
For the error of the NE class in Figure 8 it can be seen that the error is decreasing in s and t, but increasing in C. Both of these behaviours are in line with the bound in Theorem 2.2. The decrease of the error for increasing s and t is immediately clear from

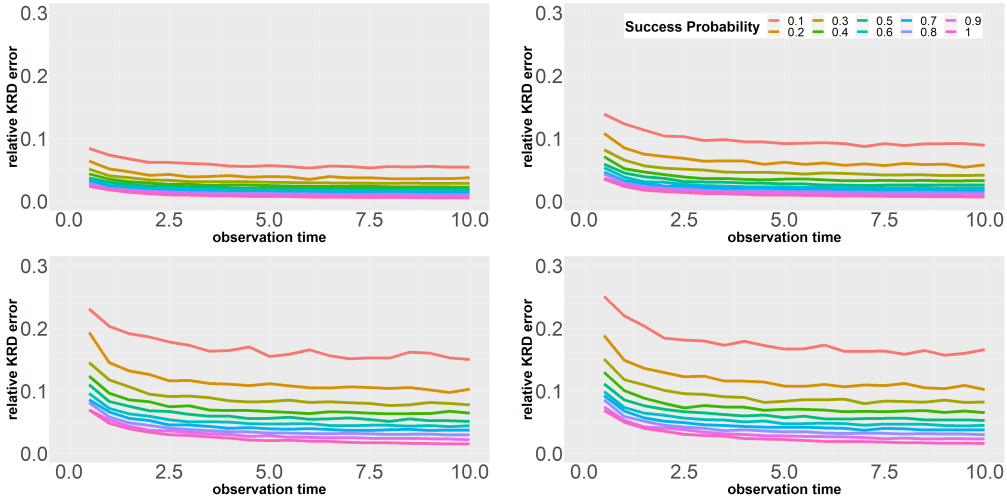


Figure 8: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the NE class and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 100$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

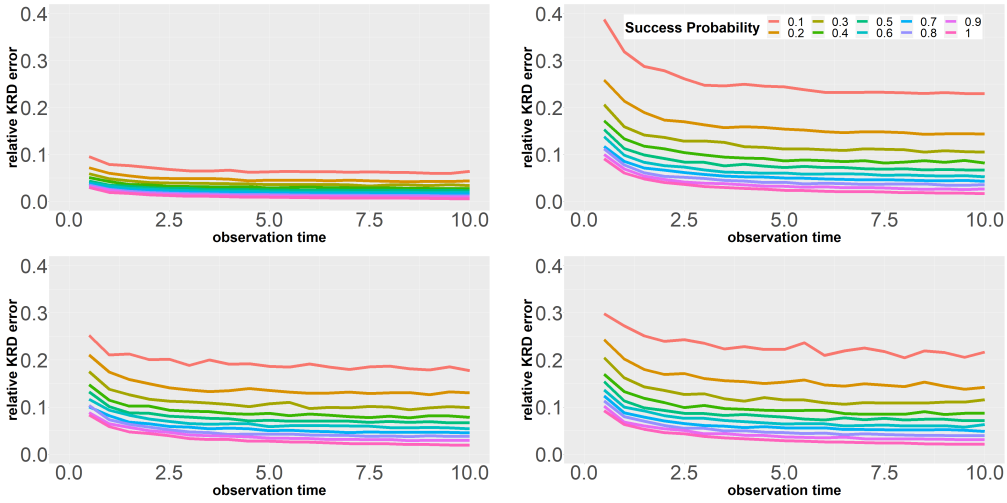


Figure 9: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the NEC class and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 75$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

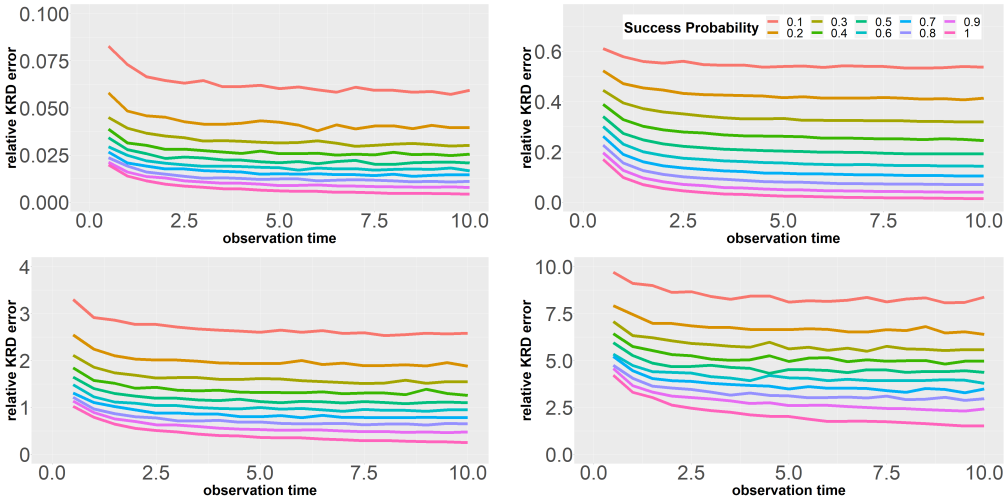


Figure 10: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the PI class and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 450$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

our theoretical results. The increase of error for increasing C is based on the fact that in the Poisson model the population total intensities of μ and ν are unknown and have to be estimated from the data. The (p, C)-KRD penalises mass deviation with a factor scaling with C, so naturally for increasing C, the errors in the estimation of the true difference of masses yields an increase in the expected relative (p, C)-KRD error. Notably, while the decrease in s and t is similar for the error of the NEC class in Figure 9, the error is no longer increasing in C. Instead the errors increase from $C = 0.01$ to $C = 0.1$, but then decrease going to $C = 1$. Afterwards they increase again at $C = 10$. This difference in behaviour is explained by the cluster structure of the measure in NEC. There is still the general trend of increasing error for increasing C, as present in the NE class, but now there is an additional change in behaviour based on the fact if transport occurs within clusters or between clusters. From $C = 0.1$ to $C = 1$, we pass the size of the clusters and the distance between the clusters. Thus, $C = 1$ is the first value in our simulation for which inter-cluster transports can occur. This causes a decrease in error, as the impact of the estimation of the total mass intensity of a measure within one cluster is decreased. After this point the usual increase in error for increasing C due to the estimation of the total mass intensity occurs again. For the (p, C)-KRD error in the PI class in Figure 10 this effect is particularly strong. The error increases on average about two orders of magnitude from $C = 0.01$ to $C = 10$. This is explained by the fact that the total mass intensity in this class is significantly larger than for the classes NE and NEC, where each location in the support of the measures has mass one. This also causes an increase of the variance of the mass of the empirical measures at each location, which causes a faster increase of error for increasing C at all scales of C.

4.3 Simulation Results for the $(2, C)$-Barycenter

It remains to discuss the results from our simulation studies for the Poisson model for the $(2, C)$-barycenter between sets of measures within one of the eight classes of measures introduced above. We again exemplary discuss the classes NE, NEC and PI. The remaining plots in the Figures 24, 25, 26, 27 and 28 are found in Appendix E. We also restrict our analysis to the values of $C = 0.1, 1, 10$, since for $C = 0.01$ the (p, C)-KRD is close to the TV

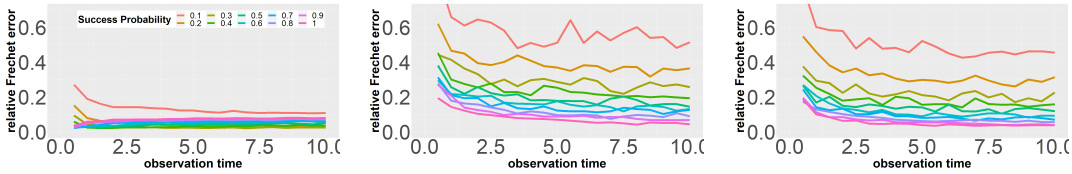


Figure 11: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the NE class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 100 independent runs. Set $M = 100$. From left to right we have $C = 0.1, 1, 10$, respectively.

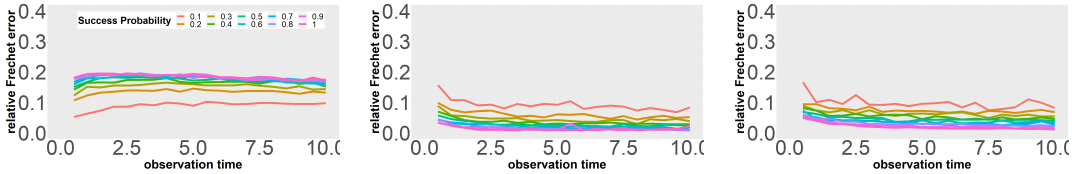


Figure 12: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the NEC class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 100 independent runs. Set $M = 75$. From left to right we have $C = 0.1, 1, 10$, respectively.

distance. In particular, for all classes except PIG and NIG, where all measures share the same support grid, the $(2, C)$-barycenter will be close or identical to the zero measure, since the measures in the other classes are almost surely disjoint. Additionally, if the barycenter of the population measures is the zero measure, any empirical barycenter has mass zero as well. Thus, there is little merit in simulating the barycenters in these cases. For the classes NIG and PIG the barycenters are essentially TV-barycenters for small C which removes any geometrically interesting features from the barycenter. Finally, for extremely small values of C the (p, C)-barycenter computations tend to become numerically unstable due to either involving values close to machine accuracy and UOT plans for this values of C often being close to the zero measure. Hence, empirical simulations of the expected relative Fréchet error would also be less reliable in this regime of values for C. In summary, empirical analysis of the properties of the (p, C)-barycenter for values of C which are several orders of magnitude smaller than the diameter of $\mathcal{Y}$ is inadvisable.

For the relative Fréchet errors we observe significant changes in behaviour compared to the relative (p, C)-KRD before. Considering the error for the NE class in Figure 11, we note that for $C = 0.1$ the behaviour in s and t is different than for the (p, C)-KRD. Namely, for fixed s, the error is in general not strictly decreasing in t and vice versa for fixed t, the error is not always strictly decreasing in s. This is an interesting effect arising for small values of the product st. A point $y \in \mathcal{Y}$ can only be a support point of a (p, C)-barycenter if it is in the intersection of at least $J/2$ balls of size $C^p/2$ around support points of different measures (compare the construction of the centroid set in (5)). Now, for small s and t many support points of the population barycenter are not included in the support of the empirical one, since centroid set of the empirical measures is significantly smaller than the population level one. In particular, this can create situations where an increase in s or t on average adds support points to empirical barycenter, which cause the relative error to increase, since placing mass zero at this location, for small C, is actually better than placing a potentially larger mass (since we assumed $(ts)^{-1}$ to be relatively

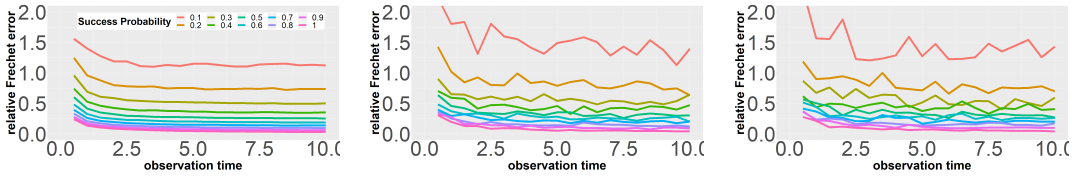


Figure 13: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the PI class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 100 independent runs. The parameter K is set to 450. From left to right we have $C = 0.1, 1, 10$, respectively.

small) at this location. Thus, while asymptotically, the rate in Theorem 3.1 is optimal, for certain, sufficiently small, values of s, t and C, the behaviour of the relative Fréchet error might be counter-intuitive. For $C = 0.1$ and $C = 1$, the errors behave quite similarly to the (p, C)-KRD setting, though there is essentially no increase in error going from $C = 1$ to $C = 10$. This is explained by two points. First, the location of the (p, C)-barycenter tends to be more centered within the support of the measures (all measures are support on subsets of the unit square), so little transport between the barycenter and the μ^i occurs at a distance larger than one. Second, the key factor for the increasing error for increasing C in the (p, C)-KRD case is the estimation error for the total mass intensities. However, for sufficiently large C the mass of the (p, C)-barycenter is the median of the total masses of the μ^i. Since, this quantity is significantly more stable under estimation than the indivual total mass intensities, it is to be expected that the mass estimation has little effect on the relative Fréchet error. For the error of the NEC class in Figure 12 the results are similar to the NE case. We observe similar effects on the dependence of s and t for $C = 0.1$ and for $C = 1$ and $C = 10$, the errors look extremely similar. One notable distinction is the fact the from $C = 0.1$ to $C = 1$ the errors decrease on average. As before for the NEC class in the (p, C)-KRD setting, this can be explained by its cluster structure and $C = 1$ being the first value for which inter-cluster transport becomes possible in an UOT plan. This is therefore also the first value of C which allows the (p, C)-barycenter to have mass between clusters. Finally, for the error of the PI class in Figure 13 the value of C only has a minimal effect on the resulting errors. Notably and contrary to the two prior classes, we do not encounter any additional effects for $C = 0.1$. This is explained by the in general higher mass intensities of the measures in the PI class, which make the previously described effects due to low values of s and t less likely. Additionally, these measures do not possess any geometrical structures in their support, which could impact the behaviour on different scales. There is again little increase in error for increasing C, which is in stark contrast to the PI class in the (p, C)-KRD setting, where the error increased by multiple orders of magnitude. This is another strong indicator, that the Fréchet error is significantly more stable under C, due to the stability of total mass intensity of the empirical barycenter opposed to the total mass intensity of the individual measures.

4.4 Real Data Example

In Figure 14 we consider the $(2, 0.1)$-KRD between images which are an excerpt from a real dataset from confocal nanoscopy of adult human dermal fibroblast cells (for the full dataset see Tameling et al. [2021]). The images in the Figures 14(a),(b) and Figures 14(c),(d) are visually similar, as they correspond to measurements taken based on two different markers (one at the inner mitochondrial membrane and one at the outer) in the same cells. The $(2, 0.1)$-KRD captures this fact in the sense, that the pairwise distance between the

measures are smallest for these pairs of images. Utilising UOT on this type of datasets is a potential way of quantifying dissimilarity between the respective measures and extending OT based dissimilarity analysis to measures of unequal total intensity. However, for high-resolution images numerical computations become intractable which requires the use of surrogates such as estimators of the measures for randomised computations. The 300×300 images here are specifically chosen such that the true distances can still be computed which allows to compare the expected error of the empirical (p, C)-KRD for given sample sizes on this data set. We compare the results obtained from the resampling approach (i.e. the estimator from (6)) considered in the multinomial model to the subsampling approach (i.e. the estimator from (9)) obtained by sampling without replacement from the measures instead. In these simulations the maximum sample size is about $1/5$ of the support sizes. This corresponds to a runtime of about 2.5% of the original problem size. While it is clear by construction that for sufficiently large sample sizes, subsampling yields a smaller error than the resampling (as the error approaches zero if the sample size approaches the support size), for smaller sample sizes the resampling can have a better performance. It yields a relative error below 5% at less than 10% of the original support size in all considered instances. This approximation can be achieved in around 0.5% of the original runtime. The subsampling approach does not reach this level of accuracy for the considered sample sizes. Thus, these simulations suggest that randomised computations based on the multinomial model allow for high accuracy approximations of the (p, C)-KRD in real data applications at a significantly lower computational cost than the original problem and that for small sample sizes there are scenarios where the resampling approach yields significantly better performance than the subsampling one.

Acknowledgments

F. Heinemann and M. Klatt gratefully acknowledge support from the DFG Research Training Group 2088 *Discovering structure in complex data: Statistics meets optimization and inverse problems.* A. Munk gratefully acknowledges support from the DFG CRC 1456 *Mathematics of the Experiment A04, C06* and the DFG Cluster of Excellence 2067 MBExC *Multiscale bioimaging–from molecular machines to networks of excitable cells.*

References

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

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. Proceedings of Machine Learning Research, 2017.

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu. Visual feature attribution using Wasserstein GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018.

D. Berend and A. Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013.

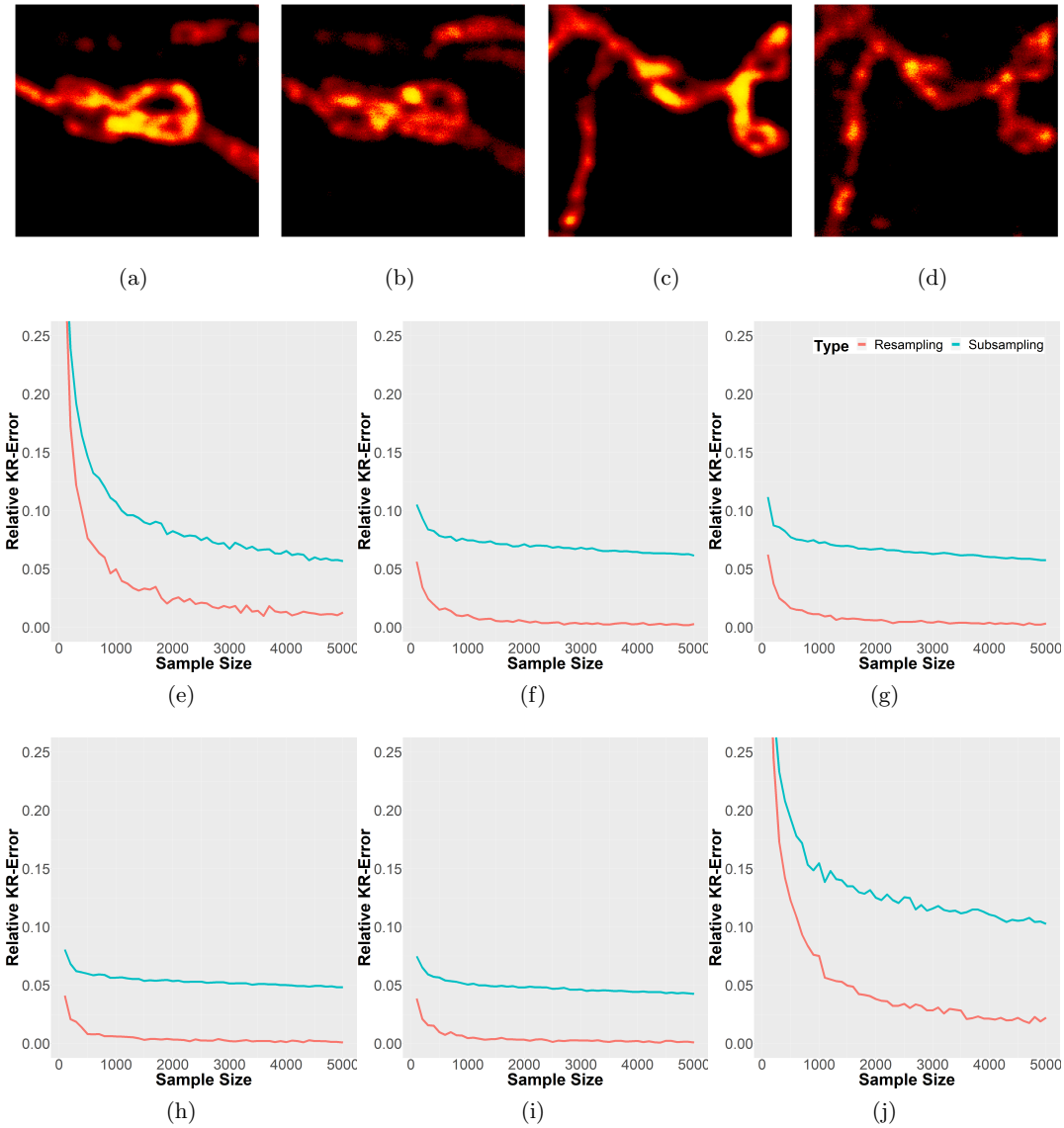


Figure 14: **(a)-(d):** Excerpts of size 300×300 from the confocal nanoscopy data of adult Human Dermal Fibroblasts in Tameling et al. [2021]. The images have on average about 25000 non-zero pixels. (a) and (c) have been labelled at MIC60 (a mitochondrial inner membrane complex); (b) and (d) have been labelled at TOM20 (translocase of the outer mitochondrial membrane). **(e)-(j):** The relative error for the empirical $(2, 0.1)$-KRD (obtained from resampling and subsampling) between the four filament structures in (a)-(d) considered as measures in $[0, 1]^2$. **(e)** Between (a) and (b). **(f)** Between (a) and (c). **(g)** Between (a) and (d). **(h)** Between (b) and (c). **(i)** Between (b) and (d). **(j)** Between (c) and (d).

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018a.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018b.

S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: Approximation by empirical measures. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 1183–1203, 2013.

S. N. Evans and F. A. Matsen. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

K. Fatras, Y. Zine, S. Majewski, R. Flamary, R. Gribonval, and N. Courty. Minibatch optimal transport distances; analysis and applications. *preprint arXiv:2101.01792*, 2021.

A. Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems*, 28:2053–2061, 2015.

M. Gellert, M. F. Hossain, F. J. F. Berens, L. W. Bruhn, C. Urbainsky, V. Liebscher, and C. H. Lillig. Substrate specificity of thioredoxins and glutaredoxins – towards a functional classification. *Heliyon*, 5(12):e02943, 2019.

K. Guittet. Extended Kantorovich norms: a tool for optimization. Technical report, Technical Report 4402, INRIA, 2002.

W. Guo, N. Ho, and M. Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pages 2088–2097. Proceedings of Machine Learning Research, 2020.

L. G. Hanin. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

F. Heinemann, M. Klatt, and A. Munk. Kantorovich-Rubinstein distance and barycenter for finitely supported measures: Foundations and algorithms. *preprint arXiv:2112.03581*, 2021.

F. Heinemann, A. Munk, and Y. Zemel. Randomized Wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM Journal on Mathematics of Data Science*, 4(1):229–259, 2022.

L. V. Kantorovich and S. Rubinstein. On a space of totally additive functions. *Vestnik of the Saint Petersburg University: Mathematics*, 13(7):52–59, 1958.

M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

G. Kulaitis, A. Munk, and F. Werner. What is resolution? A statistical minimax testing perspective on superresolution microscopy. *The Annals of Statistics*, 49(4):2292–2312, 2021.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

D. Mukherjee, A. Guha, J. M. Solomon, Y. Sun, and M. Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. Proceedings of Machine Learning Research, 2021.

A. Munk, T. Staudt, and F. Werner. Statistical foundations of nanoscale photonic imaging. In *Nanoscale Photonic Imaging*, pages 125–143. Springer, Cham, 2020.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

F. Pitié, A. C. Kokaram, and R. Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1-2):123–137, 2007.

J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *International conference on scale space and variational methods in computer vision*, pages 256–269. Springer, 2015.

T. Ramasubban. The mean difference and the mean deviation of some discontinuous distributions. *Biometrika*, 45(3-4):549–549, 1958.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 55 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer, 2015.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society Series B*, 80(1):219–238, 2018.

M. Sommerfeld, J. Schrieber, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20 (105):1–23, 2019.

C. Tameling, S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk. Colocalization for super-resolution microscopy via optimal transport. *Nature Computational Science*, 1 (3):199–211, 2021.

G. Tartavel, G. Peyré, and Y. Gousseau. Wasserstein loss for image synthesis and restoration. *SIAM Journal on Imaging Sciences*, 9(4):1726–1755, 2016.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. Low-dose ct image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE transactions on Medical Imaging*, 37 (6):1348–1357, 2018.

X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian. Rethinking rotated object detection with Gaussian Wasserstein distance loss. In *International Conference on Machine Learning*, pages 11830–11841. Proceedings of Machine Learning Research, 2021.

A Bounds for the Multinomial Model

In this section we provide results analog to Theorem 2.2, Theorem 3.1 and Theorem 3.2 for the estimator in the multinomial model in (6). The proofs only differ in the way the respective expectations are bounded, so whenever suitable, we only provide these differences in the proofs.

Lemma A.1 (Total Variation Bound)**.** *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with total mass $\mathbb{M}(\mu)$. Let $\hat{\mu}_N$ be the estimator from (6). Then, for any $p \geq 1$ it holds that*

$$\mathbb{E}\left[KR^p_{p,C}\left(\hat{\mu}_N, \mu\right)\right] \leq \left(\frac{C^p}{2}\sqrt{\mathbb{M}(\mu)}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}\right) N^{-\frac{1}{2}}.$$

Proof. This proof is identical to the proof of Lemma 2.5 except for the bound on the expectation. For this, we note

$$\begin{aligned}\mathbb{E}\left[\mathrm{TV}\left(\hat{\mu}_N, \mu\right)\right] &= \sum_{x\in\mathcal{X}} \mathbb{E}\left[\left|\,\hat{\mu}_N(x) - \mu(x)\,\right|\right] \\ &= \frac{\mathbb{M}(\mu)}{N}\sum_{x\in\mathcal{X}} \mathbb{E}\left[\left|\sum_{i=1}^N \mathbb{1}\{X_i = x\} - N\frac{\mu(x)}{\mathbb{M}(\mu)}\right|\right] \\ &\leq \frac{\mathbb{M}(\mu)}{N}\sum_{x\in\mathcal{X}}\sqrt{N\frac{\mu(x)}{\mathbb{M}(\mu)}\left(1-\frac{\mu(x)}{\mathbb{M}(\mu)}\right)} \leq N^{-\frac{1}{2}}\sqrt{\mathbb{M}(\mu)}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)},\end{aligned}$$

where the inequality follows from the fact the $X_i \sim Ber\left(\frac{\mu(x)}{\mathbb{M}(\mu)}\right)$ for $i = 1, \ldots, N$. □

Theorem A.2. *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with total mass $\mathbb{M}(\mu)$. Let $\hat{\mu}_N$ be the estimator from (6). Then, for any $p \geq 1$, resolution $q > 1$ and depth $L \in \mathbb{N}$ it holds that*

$$\mathbb{E}\left[KR_{p,C}\left(\hat{\mu}_N, \mu\right)\right] \leq \mathcal{E}^{Mult}_{p,\mathcal{X},\mu}(C)^{1/p} N^{-\frac{1}{2p}}.$$

For

$$A_{q,p,L,\mathcal{X}}(l) := diam(\mathcal{X})^p 2^{p-1}\left(q^{-Lp}|\mathcal{X}|^{\frac{1}{2}} + \left(\frac{q}{q-1}\right)^p \sum_{j=l}^{L} q^{p-jp}|Q_j|^{\frac{1}{2}}\right),$$

the constant is equal to

$$\mathcal{E}^{Mult}_{p,\mathcal{X},\mu}(C,q,L) = \begin{cases} A_{q,p,L,\mathcal{X}}(1), & \text{if } C \geq 2h_{q,L}(0), \\ A_{q,p,L,\mathcal{X}}(l), & \text{if } 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2}\sqrt{\mathbb{M}(\mu)}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}, & \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right). \end{cases}$$

Furthermore, for $p = 1$ the factor $\frac{q}{(q-1)}$ in $A_{q,1,L,\mathcal{X}}(a,b,l)$ can be removed. Denote

$$\mathcal{E}^{Mult}_{p,\mathcal{X},\mu}(C) := \inf_{L\in\mathbb{N},q>1} \mathcal{E}^{Mult}_{p,\mathcal{X},\mu}(C,q,L).$$

Proof. The proof of this result only differs from the proof of Theorem 2.2 by the upper bounds on the relevant expectations. By definition $\mathbb{E}\left[|\mathbb{M}(\hat{\mu}_N) - \mathbb{M}(\mu)|\right] = 0$. Furthermore, scaling the expectation by total mass

$$\mathbb{E}\left[\left|\hat{\mu}^L_N(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right] = \mathbb{M}(\mu)\mathbb{E}\left[\left|\frac{\hat{\mu}^L_N(\mathcal{C}(x))}{\mathbb{M}(\mu)} - \frac{\mu^L(\mathcal{C}(x))}{\mathbb{M}(\mu)}\right|\right],$$

we notice that $\frac{\hat{\mu}^L_N(\mathcal{C}(x))}{\mathbb{M}(\mu)} \overset{D}{=} \frac{1}{N}\sum_{i=1}^N X_i(x)$, where $X_1(x),\dots,X_N(x) \overset{\text{i.i.d.}}{\sim} \text{Ber}\,(a(x))$ with $a(x) := \frac{\mu^L(\mathcal{C}(x))}{\mathbb{M}(\mu)}$. Consequently, it holds that

$$\begin{aligned}\sum_{x\in Q_l} \mathbb{E}\left[\left|\hat{\mu}^L_N(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right] &= \mathbb{M}(\mu)\sum_{x\in Q_l}\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N X_i(x) - a(x)\right|\right] \\ &\leq \mathbb{M}(\mu)\sum_{x\in Q_l}\sqrt{\frac{a(x)(1-a(x))}{N}} \\ &\leq \mathbb{M}(\mu)\sqrt{\frac{|Q_l|}{N}}.\end{aligned}$$

□

Notably, compared to $\mathcal{E}^{\text{Poi}}_{p,\mathcal{X},\mu}(C,q,L)$, the constant $\mathcal{E}^{\text{Mult}}_{p,\mathcal{X},\mu}(C,q,L)$ misses an additional summand for large C. This summand corresponds to the estimation error of the total mass intensity of $\hat{\mu}_N$ which is zero by assumptions of the model.

Remark A.3. *If $C > \text{diam}(\mathcal{X})$ and $\mathbb{M}(\mu) = \mathbb{M}(\nu)$ UOT between μ and ν is equal to OT between these two measures. In particular, for $C > 2_{q,L(0)}$ we recover the respective deviation bounds for empirical optimal transport in Sommerfeld et al. [2019]. Since in the multinomial model for all $N \in \mathbb{N}$ it holds $\mathbb{M}(\hat{\mu}_N) = \mathbb{M}(\mu)$, this implies that for $C > \text{diam}(\mathcal{X})$ the (p,C)-KRD error is equal to the OT error. Since for empirical OT the parametric $N^{-\frac{1}{2}}$ rate is already known to be optimal [Fournier and Guillin, 2015], our rate in N is sharp.*

Theorem A.4. *Let $\mu^1,\dots,\mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = supp(\mu^i)$ for $i = 1,\dots,J$. Consider random estimators $\hat{\mu}^1_{N_i},\dots,\hat{\mu}^J_{N_J} \in \mathcal{M}_+(\mathcal{X})$ derived from (6) and based on sample size $N_1,\dots,N_J$, respectively. Then it holds for any barycenter μ^* of the population measures and any barycenter $\hat{\mu}^*$ of the estimators,*

$$\mathbb{E}\left[|F_{p,C}(\mu^\star) - F_{p,C}(\hat{\mu}^\star)|\right] \leq \frac{2p\min\{\text{diam}(\mathcal{Y}),C\}^{p-1}}{J}\sum_{i=1}^J \mathcal{E}^{Mult}_{1,\mathcal{X}_i,\mu^i}(C)N_i^{-\frac{1}{2}}.$$

Theorem A.5. *Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = supp(\mu^i)$ for $i = 1, \dots, J$. Consider random estimators $\hat{\mu}^1_{N_i}, \dots, \hat{\mu}^J_{N_J} \in \mathcal{M}_+(\mathcal{X})$ derived from (6) and based on sample size $N_1, \dots, N_J$, respectively. Let $\mathbf{B}^\star$ be the set of (p, C)-barycenters of $\mu^1, \dots, \mu^J$ and $\hat{\mathbf{B}}^\star$ the set of (p, C)-barycenters of $\hat{\mu}^1_{N_i}, \dots, \hat{\mu}^J_{N_J}$. Then, for $p \geq 1$ it holds that*

$$\mathbb{E}\left[\sup_{\hat{\mu}^\star \in \hat{\mathbf{B}}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} KR^p_{p,C}(\mu^\star, \hat{\mu}^\star)\right] \leq \frac{p \min\{\mathrm{diam}(\mathcal{Y}), C\}^{p-1}}{V_P J} \sum_{i=1}^{J} \mathcal{E}^{Mult}_{1,\mathcal{X}_i,\mu^i}(C) N_i^{-\frac{1}{2}},$$

where the constant V_P is defined as in Theorem 3.2.

The proofs of Theorem A.4 and Theorem A.5 are deferred to Appendix D.

Remark A.6. *By the same argument as for the Poisson model, the approximation rate of the sampling estimator for the barycenter never decreases faster to zero than for a single measure. Thus, the $N^{-\frac{1}{2}}$ rate is sharp.*

A.1 Explicit

Following the arguments in Section 2.4, we can also provide upper bounds on d-dimensional Euclidean spaces for the constant in Theorem A.2.
For $D < 2p$ and $L \to \infty$, it holds

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C) \leq D^{p/2} \begin{cases} \mathrm{diam}^p_\infty(\mathcal{X}) 2^{3p-1} \frac{2^{D/2-p}}{1-2^{D/2-p}}, & \text{if } C \geq 2h_L(0), \\ \mathrm{diam}^p_\infty(\mathcal{X}) 2^{3p-1} \frac{2^{l(D/2-p)}}{1-2^{D/2-p}}, & \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2} \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}, & \text{if } C \leq \left(2h_L(L) \wedge \min_{x \neq x'} d_\infty(x, x')\right). \end{cases}$$

For $D = 2p$ and $L = \lfloor \frac{1}{D} \log_2(|\mathcal{X}|) \rfloor$, it holds

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C) \leq D^{p/2} \begin{cases} \mathrm{diam}^p_\infty(\mathcal{X}) 2^{3p-1} (2^{-2p} + D^{-1} \log_2(|\mathcal{X}|)), \\ \qquad \text{if } C \geq 2h_L(0), \\ \mathrm{diam}^p_\infty(\mathcal{X}) 2^{3p-1} \left(2^{-2p} + D^{-1} \log_2(|\mathcal{X}|) - l\right), \\ \qquad \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2} \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}, \\ \qquad \text{if } C \leq \left(2h_L(L) \wedge \min_{x \neq x'} d_\infty(x, x')\right). \end{cases}$$

For $D > 2p$ and $L = \lfloor \frac{1}{D} \log_2(|\mathcal{X}|) \rfloor$, it holds

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C) \leq D^{p/2} \begin{cases} \mathrm{diam}^p_\infty(\mathcal{X}) 2^{p-1} |\mathcal{X}|^{\frac{1}{2}-\frac{p}{D}} \left(1 + \frac{2^{p+D/2}}{2^{D/2-p}-1}\right), \\ \qquad \text{if } C \geq 2h_L(0), \\ \mathrm{diam}^p_\infty(\mathcal{X}) \left(|\mathcal{X}|^{\frac{1}{2}-\frac{p}{D}} + \frac{2^{p+D/2}}{2^{D/2-p}-1} \left(|\mathcal{X}|^{\frac{1}{2}-\frac{p}{D}} - 2^{(D/2-p)(l-1)}\right)\right), \\ \qquad \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2} \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}, \\ \qquad \text{if } C \leq \left(2h_L(L) \wedge \min_{x \neq x'} d_\infty(x, x')\right). \end{cases}$$

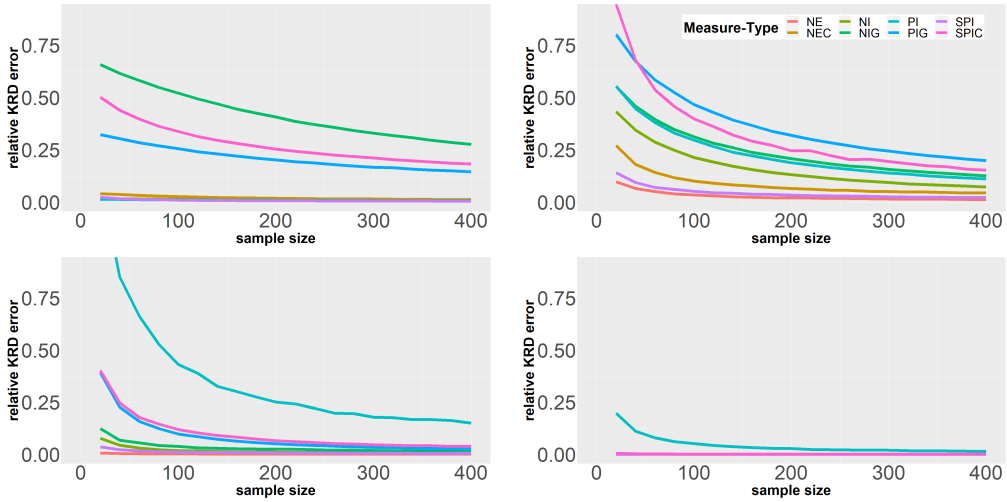


Figure 15: Expected relative $(2, C)$-KRD error for two measures in the multinomial model for the eight classes in Section 4. For each sampling size N the expectation is estimated from 1000 independent runs. For each class the parameters are set, such that the measures have on average 300 support points. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

We stress that while these constants do not include the additional term for the estimation of the total mass intensity, their dependency on $|\mathcal{X}|$ is identical to that of the upper bounds on $\mathcal{E}^{\text{Pois}}_{p,\mathcal{X},\mu}(C)$. In particular, the phase transitions still occur depending on whether D is larger than $2p$, smaller than $2p$ or equal to it.

A.2 Simulations

We repeat the simulations from Section 4 for the multinomial model. For the (p, C)-KRD the results (in Figure 15) slightly differ from the results in the Poisson model. Notably, for increasing C, the error is decreasing. This is explained by the fact that in the multinomial scheme, we do not have to estimate the total intensities of the measures and it is precisely this estimation error that drives the error for increasing C in the Poisson model. Similarly to the Poisson scheme, we observe a decrease in error for the measure classes with clustered support structures when C surpasses the distance between two individual clusters.
For the (p, C)-barycenters under the multinomial sampling model (in Figure 16) there is an initial increase in error for small sample sizes. Specifically, this occurs for $C = 0.1$ and the NEC and SPIC classes. This value of C is below the cluster size. This effect is most likely for these measure classes. For increasing C there is a significant reduction in estimation error. In particular, for some classes the error reduces by two orders of magnitude going from $C = 0.1$ to $C = 10$. Since the total mass intensities of the individual measures do not need to be estimated in this sampling model, we already observe an decrease in error for increasing C for the (p, C)-KRD and naturally there is a similar effect for the Fréchet functional.

B Bounds for the Bernoulli Model

In this section we provide results analog to Theorem 2.2, Theorem 3.1 and Theorem 3.2 for the estimator in the multinomial model in (7). The respective proofs only differ in the way the respective expectation are bounded, so whenever suitable, we only provide these

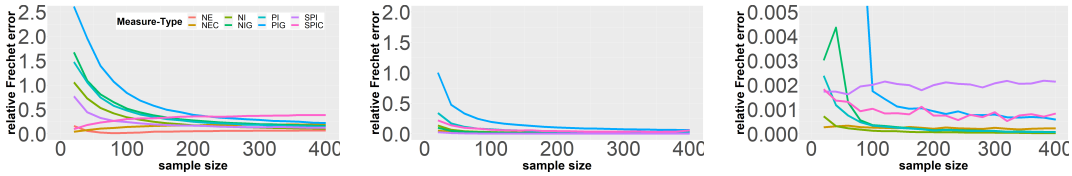


Figure 16: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the PI class for the multinomial model with different sample sizes N. For each sample size the expectation is estimated from 100 independent runs. For each class the parameters are set, such that the measures have on average 300 support points. From left to right we have $C = 0.1, 1, 10$, respectively.

differences in the proofs.

Lemma B.1 (Total Variation Bound)**.** *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with $\mu(x) \in \{0, 1\}$ for $x \in \mathcal{X}$. Let $\hat{\mu}_{s_\mathcal{X}}$ be the measure in (7). Then, for any $p \geq 1$ it holds that*

$$\mathbb{E}\left[KR^p_{p,C}\left(\hat{\mu}_{s_\mathcal{X}}, \mu\right)\right] \leq C^p \sum_{x \in \mathcal{X}} (1 - s_x)\,.$$

Proof. This proof is identical to the proof of Lemma 2.5 except for the bound on the expectation. For this, note that

$$\mathbb{E}\left[\mathrm{TV}\left(\hat{\mu}_{s_\mathcal{X}}, \mu\right)\right] = \sum_{x \in \mathcal{X}} \mathbb{E}\left[\left|\frac{1}{s_x} B_x - 1\right|\right] = \sum_{x \in \mathcal{X}} (1 - s_x) + s_x\left(\frac{1}{s_x} - 1\right) = 2 \sum_{x \in \mathcal{X}} (1 - s_x),$$

with $B_x \sim Ber(s_x)$ for $s_x \in [0, 1]$ for all $x \in \mathcal{X}$. □

Theorem B.2. *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with $\mu(x) \in \{0, 1\}$ for $x \in \mathcal{X}$. Let $\hat{\mu}_{s_\mathcal{X}}$ be the measure in (7). Then, for any $p \geq 1$, resolution $q > 1$ and depth $L \in \mathbb{N}$ it holds that*

$$\mathbb{E}\left[KR_{p,C}\left(\hat{\mu}_{s_\mathcal{X}}, \mu\right)\right] \leq \mathcal{E}^{Ber}_{p,\mathcal{X},\mu}(C, q, L)^{1/p} \begin{cases} \left(2\sum_{x \in \mathcal{X}}(1 - s_x)\right)^{\frac{1}{p}}, \\ \qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x \neq x'} d(x, x')\right) \\ \left(\sum_{x \in \mathcal{X}} \frac{1 - s_x}{s_x}\right)^{\frac{1}{2p}}, \\ \qquad \text{else.} \end{cases}$$

The constant $\mathcal{E}^{Ber}_{p,\mathcal{X},\mu}(C, q, L)$ is equal to $\mathcal{E}^{Poi}_{p,\mathcal{X},\mu}(C, q, L)$ for all $C > 0$, $q > 1$ and $L \in \mathbb{N}$. We denote

$$\mathcal{E}^{Ber}_{p,\mathcal{X},\mu}(C) := \inf_{L \in \mathbb{N}, q > 1} \mathcal{E}^{Ber}_{p,\mathcal{X},\mu}(C, q, L).$$

Proof. The proof of this result only differs from the proof of Theorem 2.2 by the upper bounds on the relevant expectations. Recall the estimator $\hat{\mu}_{s_\mathcal{X}}$ from (7) and let $B_x \sim$

$\mathrm{Ber}(s_x)$ for $s_x \in [0,1]$ for all $x \in \mathcal{X}$. It holds that

$$\begin{aligned}\sum_{x\in Q_l} \mathbb{E}\left[\left|\hat{\mu}^L_{s_\mathcal{X}}(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right] &= \sum_{x\in Q_l} \mathbb{E}\left[\left|\sum_{y\in\mathcal{C}(x)} \frac{B_y}{s_y} - \sum_{y\in\mathcal{C}(x)} s_y\right|\right] \\ &\leq \sum_{x\in Q_l} \sqrt{\mathrm{Var}\left(\sum_{y\in\mathcal{C}(x)} \frac{B_y}{s_y}\right)} = \sum_{x\in Q_l} \sqrt{\sum_{y\in\mathcal{C}(x)} s_y^{-2}\mathrm{Var}(B_y)} \\ &= \sum_{x\in Q_l} \sqrt{\sum_{y\in\mathcal{C}(x)} \frac{1-s_y}{s_y}} \leq \sqrt{|Q_l|}\sqrt{\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}}.\end{aligned}$$

The total mass can be bounded analogously as

$$\mathbb{E}\left[|\hat{\mu}^L_{s_\mathcal{X}}(\mathcal{X}) - \mu^L(\mathcal{X})|\right] \leq \sqrt{\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}}.$$

□

Since the constants for the deviation bounds for this model coincide with those for the Poisson model we refer to the previous discussion on their properties.

Remark B.3. *Consider $s_\mathcal{X}$ such that $s_x = s$ for some $s \in [0,1]$ and all $x \in \mathcal{X}$. Note, that for sufficiently small C the upper bound is an equality, since the (p,C)-KRD in this setting is proportional to the TV distance and that distance has a closed form solution here. For larger C, the expectation in the proof of Theorem B.2 amounts to bounding the mean absolute deviation of a binomial distribution. This has a closed form solution which scales as the standard deviation of the respective binomial for s not too close to 0 or 1 [Berend and Kontorovich, 2013]. Hence, in this context the upper bound on the mean absolute deviation in the proof is sharp. So based on the presented approach for the deviation bounds, the upper bound is non-improvable.*

Theorem B.4. *Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = supp(\mu^i)$ for $i = 1, \dots, J$. Consider (random) estimators $\hat{\mu}^1_{s_{\mathcal{X}_1}}, \dots, \hat{\mu}^J_{s_{\mathcal{X}_J}} \in \mathcal{M}_+(\mathcal{X})$ derived from (7). Then,*

$$\mathbb{E}\left[|F_{p,C}(\mu^\star) - F_{p,C}(\hat{\mu}^\star)|\right] \leq \frac{2p\min\{\mathrm{diam}(\mathcal{Y}), C\}^{p-1}}{J} \sum_{i=1}^J \mathcal{E}^{Ber}_{1,\mathcal{X}_i,\mu^i}(C)\psi(s_{\mathcal{X}_i}),$$

where ψ is given by

$$\psi(s_\mathcal{X}) = \begin{cases} (2\sum_{x\in\mathcal{X}}(1-s_x), & \text{if } C \leq \min_{x\neq x'} d(x,x') \\ \left(\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}\right)^{\frac{1}{2}}, & \text{else.} \end{cases}$$

Theorem B.5. *Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = supp(\mu^i)$ for $i = 1, \dots, J$. Consider (random) estimators $\hat{\mu}^1_{s_{\mathcal{X}_1}}, \dots, \hat{\mu}^J_{s_{\mathcal{X}_J}} \in \mathcal{M}_+(\mathcal{X})$ derived from (7). Let $\mathbf{B}^\star$ be the set of (p,C)-barycenters of $\mu^1, \dots, \mu^J$ and $\hat{\mathbf{B}}^\star$ the set of (p,C)-barycenters of $\hat{\mu}^1_{s_{\mathcal{X}_1}}, \dots, \hat{\mu}^J_{s_{\mathcal{X}_J}}$. Then, for $p \geq 1$ it holds that*

$$\mathbb{E}\left[\sup_{\hat{\mu}^\star\in\hat{\mathbf{B}}^\star} \inf_{\mu^\star\in\mathbf{B}^\star} KR^p_{p,C}(\mu^\star, \hat{\mu}^\star)\right] \leq \frac{p\min\{\mathrm{diam}(\mathcal{Y}), C\}^{p-1}}{V_P J} \sum_{i=1}^J \mathcal{E}^{Ber}_{1,\mathcal{X}_i,\mu^i}(C)\psi(s_{\mathcal{X}_i}),$$

where ψ is defined as in Theorem B.4 and V_P is defined as in Theorem 3.1.

The proofs of Theorem B.4 and Theorem B.5 are deferred to Appendix D.

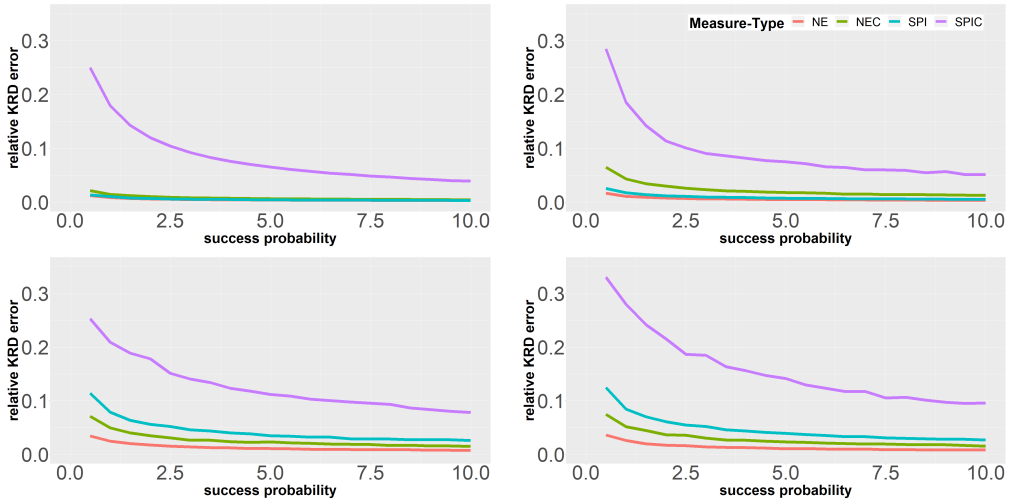


Figure 17: Expected relative $(2, C)$-KRD error for two measures in the Bernoulli Model for the measure classes from Section 4. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. The parameters are chosen such that the measures have on average 300 support points. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

B.1 Simulations

To construct a reasonable framework for the simulations, we fix $s^0 \in \mathbb{R}_+$ and assume that

$$s_x = \frac{s_0}{\|x - (0.5, 0.5)^T\|_2 + s_0}.$$

Intuitively, the success probability at a given point x is larger, if x is closer to the center of $[0, 1]^2$ and smaller if it is further away from the center. However, it still holds that for $s^0 \to \infty$ the success probability at each location converges to one. For the simulations, we now consider the error as a function of s^0. Note that in this simulation study only the classes of measures with mass one at each support point are considered in accordance with the Bernoulli model in (7). One notable observation for the empircal (p, C)-KRD (in Figure 17) is that the error of the SPIC class is significantly higher than for the NEC class, even though they share the same cluster locations. This can be explained by the fact that, by construction, the measures in the NEC class have a higher proportion of their mass in their central clusters, which is close to $(0.5, 0.5)^T$ and thus has a high probability of being observed. This effect also carries over to the (p, C)-barycenter (in Figure 18). In general, for the (p, C)-KRD the error in this model is increasing in C (which is again explained by the estimation error for the true total mass intensity). However, the effect is less pronounced than in the Poisson model. For the clustered data types a small decrease of error for increasing C over the cluster size can again be noted. Though, also this effect is less significant than in the other models. For the (p, C)-barycenter a decrease in error in C can be observed which is consistent with the previous results for the Poisson model and again explained by the increased stability of the total mass intensity of the barycenter compared to the individual ones.

C Lifts to Balanced Optimal Transport Problems

A key tool in establishing properties of the (p, C)-KRD and the (p, C)-barycenter is the lift of these problems to the space of probability measures by augmenting the space $\mathcal{X}$

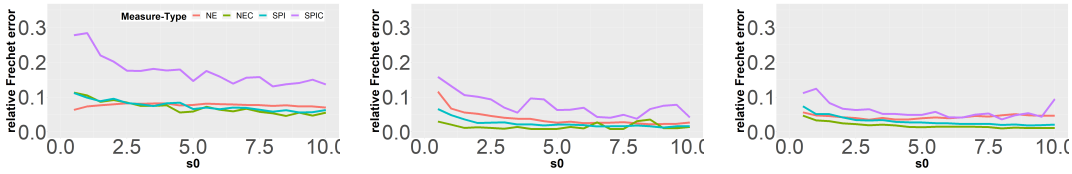


Figure 18: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the PI class for the Bernoulli model with different success vectors $s_{\mathcal{X}}$. For each sample size the expectation is estimated from 100 independent runs. The parameters are set such that the measures in all classes have on average 300 support points. From left to right we have $C = 0.1, 1, 10$, respectively.

with a dummy point having a fixed distance to all points in $\mathcal{X}$. For a fixed parameter $C > 0$, consider a dummy point $\mathfrak{d}$ and define the augmented space $\tilde{\mathcal{X}} := \mathcal{X} \cup \{\mathfrak{d}\}$ with metric cost

$$\tilde{d}_C^p(x, x') = \begin{cases} d^p(x, x') \wedge C^p, & \text{if } x, x' \in \mathcal{X}, \\ \frac{C^p}{2}, & \text{if } x \in \mathcal{X},\ x' = \mathfrak{d}, \\ \frac{C^p}{2}, & \text{if } x = \mathfrak{d},\ x' \in \mathcal{X}, \\ 0, & \text{if } x = x' = \mathfrak{d}. \end{cases} \tag{18}$$

Consider the subset $\mathcal{M}_+^B(\mathcal{X}) := \{\mu \in \mathcal{M}_+(\mathcal{X}) \mid \mathbb{M}(\mu) \leq B\} \subset \mathcal{M}_+(\mathcal{X})$ of non-negative measures whose total mass is bounded by B. Setting $\tilde{\mu} := \mu + (B - \mathbb{M}(\mu))\delta_{\mathfrak{d}}$, any measure $\mu \in \mathcal{M}_+^B(\mathcal{X})$ defines an *augmented measure* $\tilde{\mu}$ on $\tilde{\mathcal{X}}$ such that $\mathbb{M}(\tilde{\mu}) = B$. For any $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ and their augmented versions $\tilde{\mu}, \tilde{\nu} \in \mathcal{M}_+(\tilde{\mathcal{X}})$ it holds

$$KR_{C,p}^p(\mu, \nu) = \tilde{OT}_p^p(\tilde{\mu}, \tilde{\nu}).$$

Here, $\tilde{OT}_p^p$ denotes the p-OT distance defined for measures μ, ν on $(\tilde{\mathcal{X}}, \tilde{d})$ with $\mathbb{M}(\mu) = \mathbb{M}(\nu)$ as

$$\tilde{OT}_{p,C}^p(\mu, \nu) := \min_{\pi \in \Pi_=(\mu,\nu)} \sum_{x,x' \in \mathcal{X}} d^p(x, x')\pi(x, x'),$$

where the set of couplings $\Pi_=(\mu, \nu)$ is the set $\Pi_\leq(\mu, \nu)$ with inequalities replaced by equalities. Similarly, the (p, C)-barycenter problem can be augmented. For this, let $\tilde{\mathcal{Y}} := \mathcal{Y} \cup \{\mathfrak{d}\}$ endowed with the metric $\tilde{d}_C$ in (18) (replace $\mathcal{X}$ by $\mathcal{Y}$ and recall that $\mathcal{X} \subset \mathcal{Y}$) and augment the measures $\mu^1, \ldots, \mu^J$ to $\tilde{\mu}^1, \ldots, \tilde{\mu}^J$ where $\tilde{\mu}_i = \mu^i + \sum_{j \neq i} \mathbb{M}(\mu^i)\delta_{\mathfrak{d}}$ for $1 \leq i \leq J$. In particular, it holds $\mathbb{M}(\tilde{\mu}_i) = \sum_{i=1}^J \mathbb{M}(\mu^i)$ and the *augmented p-Fréchet functional* is defined as

$$\tilde{F}_{p,C}(\mu) := \frac{1}{J} \sum_{i=1}^J \tilde{\mathrm{OT}}_p^p(\tilde{\mu}_i, \mu).$$

Any minimiser of $\tilde{F}_{p,C}$ is referred to as augmented (p, C)-barycenter.

LP-Formulation for the $(\mathbf{p}, \mathbf{C})$-Barycenter

According to Heinemann et al. [2021], the augmented (p, C)-barycenter problem can be rewritten as a linear program based on the centroid set $\tilde{\mathcal{C}}_{KR}(J, p, C) = \mathcal{C}_{KR}(J, p, C) \cup \{\mathfrak{d}\}$

(recall (5) for the definition of $\mathcal{C}_{KR}(J,p,C)$) of the augmented measures. This yields

$$\begin{aligned}
\min_{\pi^{(1)},\dots,\pi^{(J)},a} \quad & \frac{1}{J}\sum_{i=1}^{J}\sum_{j=1}^{|\tilde{\mathcal{C}}_{KR}(J,p,C)|}\sum_{k=1}^{M_i}\pi_{jk}^{(i)}c_{jk}^{i} \\
\text{s.t.} \quad & \sum_{k=1}^{M_i}\pi_{jk}^{(i)} = a_j, && \forall\, i=1,\dots,J, \forall j=1,\dots,|\tilde{\mathcal{C}}_{KR}(J,p,C)|, \\
& \sum_{j=1}^{|\tilde{\mathcal{C}}_{KR}(J,p,C)|}\pi_{jk}^{(i)} = b_k^i, && \forall\, i=1,\dots,J, \forall k=1,\dots,M_i, \\
& \pi_{jk}^{(i)} \geq 0, && \forall i=1,\dots,J, \forall j=1,\dots,|\tilde{\mathcal{C}}_{KR}(J,p,C)|, \\
& && \forall k=1,\dots,M_i,
\end{aligned} \tag{19}$$

where $M_i = |\tilde{\mathcal{X}}_i|$ for each $1 \leq i \leq J$ is the cardinality of the support of the augmented measure $\tilde{\mu}_i$. Here, c_{jk}^i denotes the distance between the j-th point of $|\tilde{\mathcal{C}}_{KR}(J,p,C)|$ and the k-th point in the support of $\tilde{\mu}_i$, while b^i is the vector of masses corresponding to $\tilde{\mu}_i$.

D Omitted Proofs

D.1 Proofs for the Empirical (p, C)-Barycenters

Proof of Theorem 3.1, Theorem A.4 and Theorem B.4. Let $\hat{\mu}^1,\dots,\hat{\mu}^J$ be any of the three estimators from (6), (7) or (8) respectively. Further, for each $1 \leq i \leq J$ let $\mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)$ be the corresponding constant in Theorem 2.2, Theorem A.2 or Theorem B.2 for μ^i, respectively, and let θ_i denote the respective sampling parameter dependences $N_i^{-1/2}$, $\phi(t_i,s_i)$ or $\psi(s_{\mathcal{X}_i})$, respectively. Due to the construction of the lifted problem, it holds for any $\mu^1,\mu^2,\mu^3 \in \mathcal{M}_+(\mathcal{Y})$ that

$$\begin{aligned}
|KR_{p,C}^p(\mu^1,\mu^3) - KR_{p,C}^p(\mu^2,\mu^3)| &= |\tilde{OT}_p^p(\tilde{\mu}^1,\tilde{\mu}^3) - \tilde{OT}_p^p(\tilde{\mu}^2,\tilde{\mu}^3)| \\
&\leq \operatorname{diam}(\tilde{\mathcal{Y}})^{p-1} p\, \tilde{OT}_1(\tilde{\mu}^1,\tilde{\mu}^2) \\
&= \min\{\operatorname{diam}(\mathcal{Y}), C\}^{p-1} p\, KR_{1,C}(\mu^1,\mu^2),
\end{aligned}$$

where the inequality follows from Sommerfeld and Munk [2018]. Taking expectation and applying the previous display together with Theorem 2.2 yields

$$\begin{aligned}
\mathbb{E}\left[|F_{p,C}(\mu) - \hat{F}_{p,C}(\mu)|\right] &\leq \frac{1}{J}\sum_{i=1}^{J}\mathbb{E}\left[|KR_{p,C}^p(\mu^i,\mu) - KR_{p,C}^p(\hat{\mu}^i,\mu)|\right] \\
&\leq p\, \min\{\operatorname{diam}(\mathcal{Y}), C\}^{p-1}\frac{1}{J}\sum_{i=1}^{J}\mathbb{E}\left[KR_{1,C}(\mu^i,\hat{\mu}^i)\right] \\
&\leq p \min\{\operatorname{diam}(\mathcal{Y}), C\}^{p-1}\frac{1}{J}\sum_{i=1}^{J}\mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\theta_i.
\end{aligned}$$

Let $\mu^\star$ and $\hat{\mu}^\star$ be minimizers of their respective p-Fréchet functional $F_{p,C}$ and $\hat{F}_{p,C}$. Then, it follows that

$$\begin{aligned}
&\mathbb{E}[|F_{p,C}(\hat{\mu}^\star) - F_{p,C}(\mu^\star)|] \\
&= \mathbb{E}\left[F_{p,C}(\hat{\mu}^\star) - \hat{F}_{p,C}(\mu^\star) + \hat{F}_{p,C}(\mu^\star) - F_{p,C}(\mu^\star)\right] \\
&\leq \mathbb{E}\left[F_{p,C}(\hat{\mu}^\star) - \hat{F}_{p,C}(\mu^\star)\right] + \mathbb{E}\left[\hat{F}_{p,C}(\mu^\star) - F_{p,C}(\mu^\star)\right] \\
&\leq \mathbb{E}\left[F_{p,C}(\hat{\mu}^\star) - \hat{F}_{p,C}(\mu^\star)\right] + p\ \min\{\mathrm{diam}(\mathcal{Y}), C\}^{p-1} \frac{1}{J}\sum_{i=1}^{J} \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\theta_i \\
&\leq \mathbb{E}\left[F_{p,C}(\hat{\mu}^\star) - \hat{F}_{p,C}(\hat{\mu}^\star)\right] + p\ \min\{\mathrm{diam}(\mathcal{Y}), C\}^{p-1} \frac{1}{J}\sum_{i=1}^{J} \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\theta_i \\
&\leq p\ \min\{\mathrm{diam}(\mathcal{Y}), C\}^{p-1} \frac{2}{J}\sum_{i=1}^{J} \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\theta_i,
\end{aligned}$$

where the fourth inequality follows from $\hat{\mu}^*$ being a minimiser of $\hat{F}_{p,C}$. □

Proof of Theorem 3.2, Theorem A.5 and Theorem B.5. Let $\hat{\mu}^1, \dots, \hat{\mu}^J$, $\mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)$ and θ_i for all $i = 1, \dots, J$ as in the previous proof. Let $\mathbf{B}$ be the set of (p, C)-barycenters of the measures $\mu^1, \dots, \mu^J$ and define $\tilde{\mathbf{B}}$ as the set of OT_p-barycenters of the augmented measures $\tilde{\mu}^1, \dots, \tilde{\mu}^J$. Similar, we denote $\hat{\mathbf{B}}$ the set of (p, C)-barycenters of the estimated measures $\hat{\mu}^1, \dots, \hat{\mu}^J$ and let $\hat{\tilde{\mathbf{B}}}$ be the set of p-barycenters of their augmented versions. Define the lift of a measure $\mu \in \mathcal{M}_+(\mathcal{Y})$ to a measure $\tilde{\mu} \in \mathcal{M}(\tilde{\mathcal{Y}})$ by

$$\phi_{\mu^1,\dots,\mu^J}(\mu) = \mu + \left(\sum_{i=1}^{J} \mathbb{M}(\mu^i) - \mathbb{M}(\mu)\right)\delta_{\mathfrak{d}}.$$

If $\mu \in \mathbf{B}$ then it follows by Lemma 3.3 in Heinemann et al. [2021] that $\phi_{\mu^1,\dots,\mu^J}(\mu) \in \tilde{\mathbf{B}}$. Conversely, for any $\tilde{\mu} \in \tilde{\mathbf{B}}$ it holds that $\phi^{-1}_{\mu^1,\dots,\mu^J}(\tilde{\mu}) \in \mathbf{B}$. We denote by $\phi(\mathbf{B}) := \{\phi_{\mu^1,\dots,\mu^J}(\mu) | \mu \in \mathbf{B}\}$ and analogously $\phi^{-1}\left(\tilde{\mathbf{B}}\right) := \{\phi^{-1}_{\mu^1,\dots,\mu^J}(\tilde{\mu}) | \tilde{\mu} \in \tilde{\mathbf{B}}\}$. With this we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{\hat{\mu}\in\hat{\mathbf{B}}} \inf_{\mu\in\mathbf{B}} \mathrm{KR}^p_{p,C}(\mu, \hat{\mu})\right] &= \mathbb{E}\left[\sup_{\hat{\mu}\in\phi^{-1}\left(\hat{\tilde{\mathbf{B}}}\right)} \inf_{\mu\in\phi^{-1}\left(\tilde{\mathbf{B}}\right)} \mathrm{KR}^p_{p,C}(\mu, \hat{\mu})\right] \\
&= \mathbb{E}\left[\sup_{\hat{\mu}\in\phi^{-1}\left(\hat{\tilde{\mathbf{B}}}\right)} \inf_{\mu\in\phi^{-1}\left(\tilde{\mathbf{B}}\right)} \tilde{OT}^p_p(\phi(\mu), \phi(\hat{\mu}))\right] \\
&= \mathbb{E}\left[\sup_{\hat{\mu}\in\hat{\tilde{\mathbf{B}}}} \inf_{\mu\in\tilde{\mathbf{B}}} \tilde{OT}^p_p(\tilde{\mu}, \hat{\tilde{\mu}})\right].
\end{aligned} \tag{20}$$

We continue by recalling a slightly adapted version of Lemma 3.8 in Heinemann et al. [2022]. Since we only apply this Lemma to the augmented, balanced OT problem, the proof remains unchanged and is therefore omitted.

Lemma D.1. *Let $\tilde{F}_{p,C}$ be the augmented Fréchet functional corresponding to $\tilde{\mu}^1, \dots, \tilde{\mu}_N \in \mathcal{M}_+(\tilde{\mathcal{Y}})$. Then, for any $\tilde{\mu} \in M_+(\mathcal{C}_{KR}(J, p, C) \cup \{\mathfrak{d}\})$ with $\mathbb{M}(\tilde{\mu}) = \sum_{i=1}^J \mathbb{M}(\mu^i)$ there exists a $\tilde{\mu}^* \in \arg\min_{\tilde{\nu}} \tilde{F}_{p,C}(\tilde{\nu})$ such that*

$$\tilde{F}_{p,C}(\tilde{\mu}) - \tilde{F}_{p,C}(\tilde{\mu}^*) \geq 2V_P \tilde{OT}^p_p(\tilde{\mu}, \tilde{\mu}^*),$$

where V_P is the constant from Theorem 3.2.

Invoking Lemma D.1 with $\hat{\mu} \in \hat{\tilde{\mathbf{B}}}$ and applying Theorem 3.1 yields

$$\begin{aligned}\frac{2p}{J}\sum_{i=1}^{J}\mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\min\{\mathrm{diam}(\mathcal{X}),C\}^{p-1}\theta &\geq \mathbb{E}\left[F_{p,C}(\hat{\tilde{\mu}})-F_{p,C}(\tilde{\mu})\right]\\ &\geq \mathbb{E}\left[2V_p\sup_{\hat{\tilde{\mu}}\in\hat{\tilde{\mathbf{B}}}}\inf_{\tilde{\mu}\in\tilde{\mathbf{B}}}\tilde{OT}_p^p\left(\tilde{\mu},\hat{\tilde{\mu}}\right)\right]\\ &= \mathbb{E}\left[2V_p\sup_{\hat{\mu}\in\hat{\mathbf{B}}}\inf_{\mu\in\mathbf{B}}\mathrm{KR}_{p,C}^p\left(\mu,\hat{\mu}\right)\right],\end{aligned}$$

where the equality follows from (20) and hence

$$\frac{\frac{p}{J}\sum_{i=1}^{J}\mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\ \min\{\mathrm{diam}(\mathcal{X}),C\}^{p-1}}{V_P}\theta \geq \mathbb{E}\left[\sup_{\hat{\mu}\in\hat{\mathbf{B}}}\inf_{\mu\in\mathbf{B}}\mathrm{KR}_{p,C}^p\left(\mu,\hat{\mu}\right)\right].$$

□

E Additional Figures for the Poisson Sampling

E.1 Distance

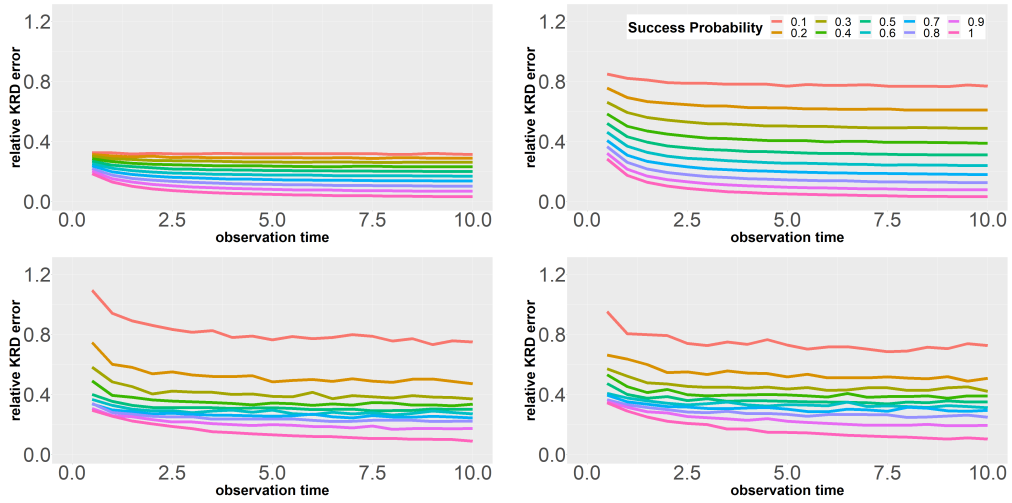


Figure 19: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the PIG class and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 22$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

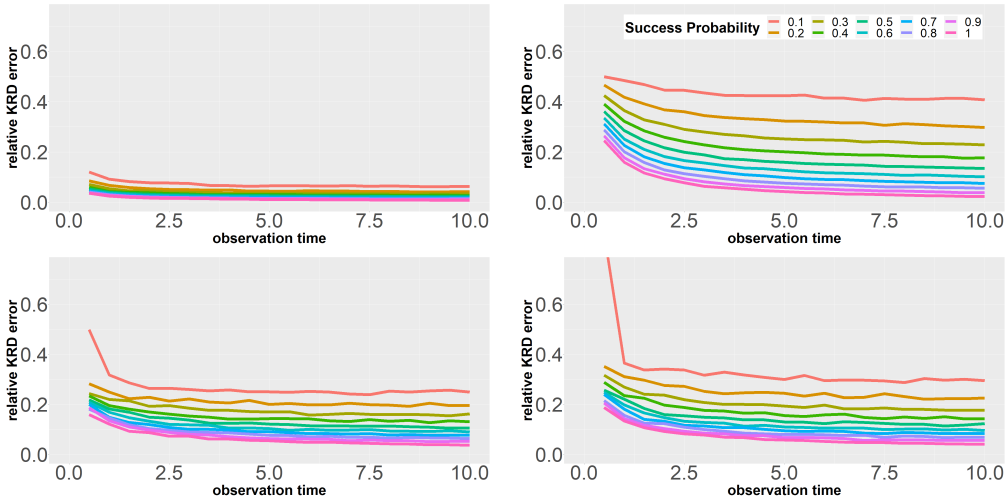


Figure 20: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the NI class and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 300$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

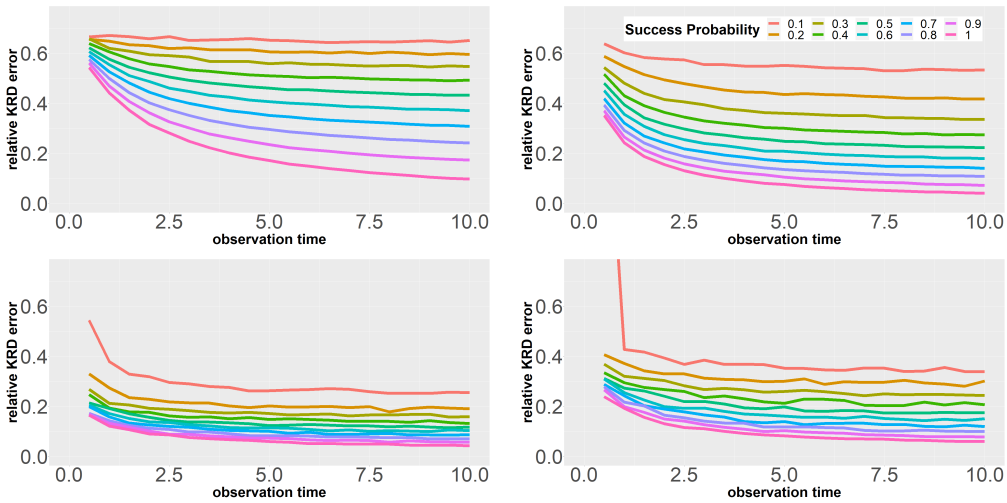


Figure 21: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the NIG class and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 17$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

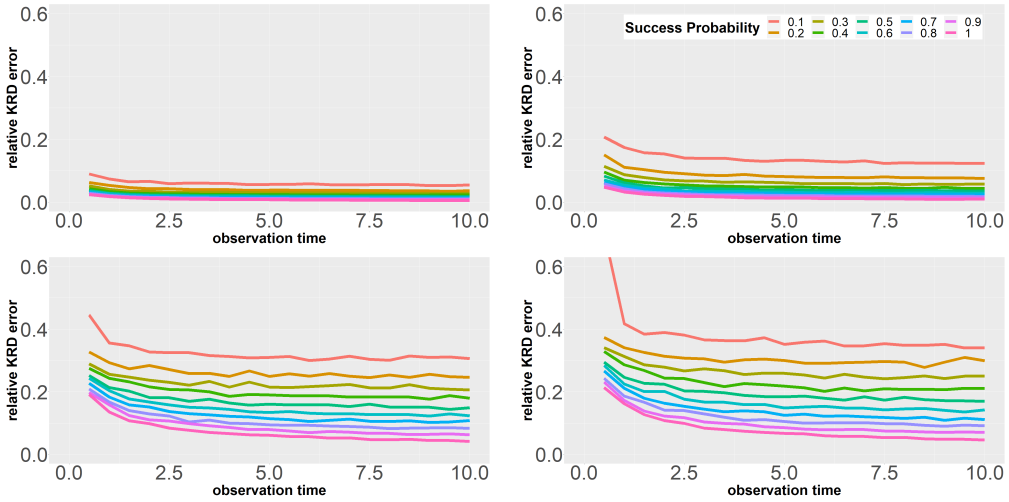


Figure 22: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the SPI class and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 65$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

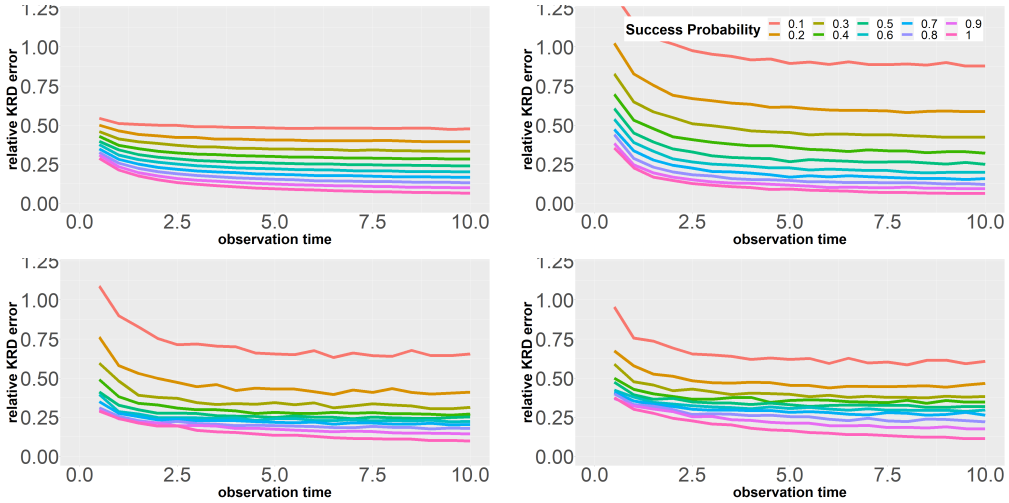


Figure 23: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the SPIC class and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 12$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

E.2 Barycenter

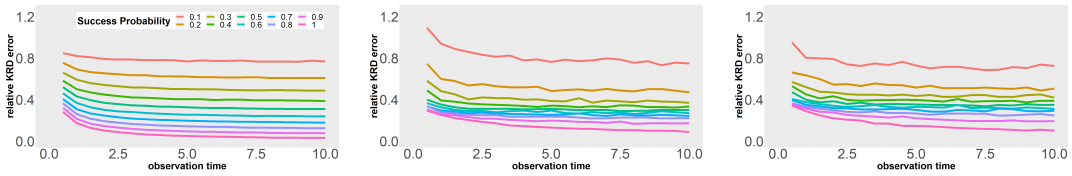


Figure 24: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the PIG class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 22$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

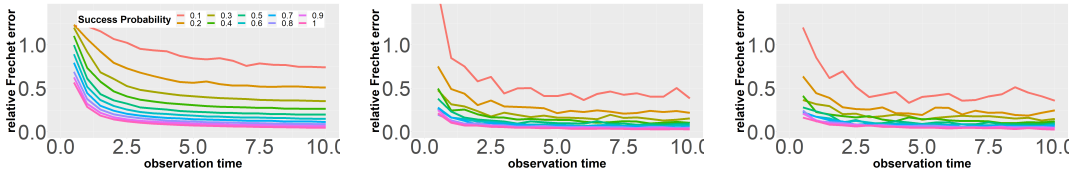


Figure 25: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the NI class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 300$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

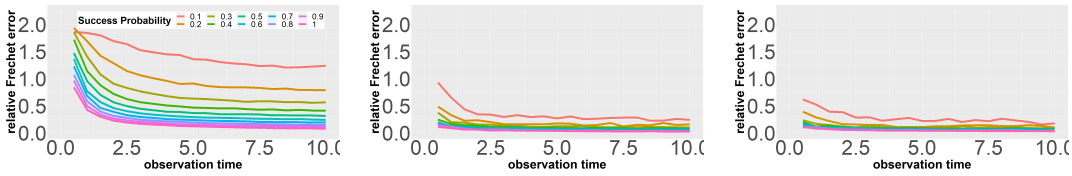


Figure 26: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the NIG class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 17$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

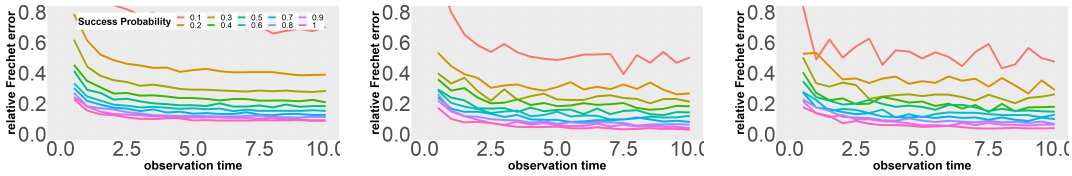


Figure 27: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the SPI class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 65$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

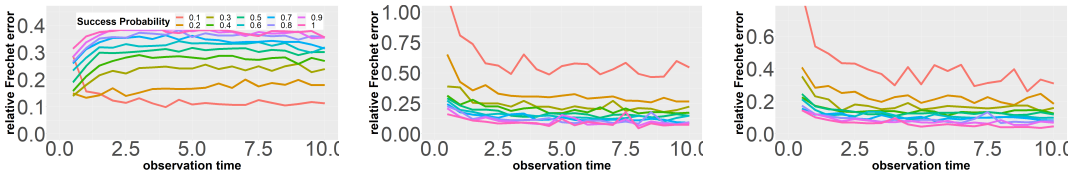


Figure 28: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the SPIC class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. Set $M = 12$. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.