Efficient Calculation of Probability Metrics of the f-$\mathcal{E}$-Class

DISSERTATION

zur Erlangung des akademischen Grades
Dr. rer. nat.

eingereicht an der
Mathematisch-Naturwissenschaftlich-Technischen Fakultät
der Universität Augsburg
Juli 2021

von

Christian Drescher

Datum der Einreichung:	22.07.2021
Tag der mündlichen Prüfung:	12.10.2021
Erster Gutachter:	Professor Dr. Ralf Werner
Zweiter Gutachter:	Professor Dr. Vitali Wachtel

Abstract

Throughout this thesis, we introduce the class of f-$\mathcal{E}$-metrics, a parameterizable class of metrics for probability distributions, containing the Prokhorov and Wasserstein-∞ metrics. Starting with the theoretical foundations, we show the similarities and differences between the latter two metrics and explore the two topologies the f-$\mathcal{E}$-class induces. This provides a joint framework for the previously mostly independently considered metrics, highlighting their connections. Figuratively speaking, this is a way of comparing how much mass has to be transported how far to transform one distribution into the other. The f-$\mathcal{E}$-metric is then attained at the balance of the distance and the f-weighted mass. In contrast to the Wasserstein-p metric, which averages all transported mass together with the distances, f-$\mathcal{E}$-metrics are only considering a cutoff point. In Proposition 2.5.1 and Proposition 2.5.2, we provide two generally valid algorithms for the exact computation of the distance of finite support distributions of size $m \geq n$. Obtaining a worst case complexity of $\mathcal{O}\left(mn^2 \log(m)\right)$, the computation of the f-$\mathcal{E}$-class is instantly on par with the well researched Wasserstein-p metric.

We further introduce for the first time quasi-convex metrics, a concept linking metric and ordered spaces. This allows for sorting the supports of our probability distributions in corresponding metric spaces while keeping a strong link with the underlying metric. The theoretical foundation lies within Monge sequences, which we cover throughout this thesis. Combining these, we can significantly improve the general complexity to quasi-linearity for the Prokhorov and Wasserstein-∞ metric. We proof correctness and worst case complexities for all algorithms, setting them on par with the Wasserstein-p metric. In detail, we obtain a general complexity of $\mathcal{O}\left(mn^2 \log(m)\right)$ and refined for the quasi-convex case to a strongly quasi-linear $\mathcal{O}\left(m \log(m)\right)$ for the Wasserstein-∞ metric in Corollary 4.2.18 and a weakly quasi-linear $\mathcal{O}\left(m \max\{\log(m), \frac{1}{acc_X}\}\right)$ depending on the support of the distributions for the Prokhorov metric in Theorem 4.1.12.

We compare the f-$\mathcal{E}$-class with existing metrics to embed it in the current tool

set and show their relationships. We conclude with a numerical analysis of our algorithms to check correctness and complexity based on implementations in MATLAB.

In total, we have newly developed a broad class of probability metrics, containing the well known Prokhorov and Wasserstein-∞ metric, analyzed their theoretical properties and provided a comprehensive set of exact and efficient algorithms for their computation for finitely supported measures.

Zusammenfassung

In dieser Arbeit führen wir die Klasse der f-$\mathcal{E}$-Metriken ein, eine parametrisierbare Klasse von Metriken für Wahrscheinlichkeitsverteilungen, die die Prokhorov und Wasserstein-∞-Metriken enthält. Beginnend mit den theoretischen Grundlagen zeigen wir die Gemeinsamkeiten und Unterschiede zwischen letzteren beiden Metriken und untersuchen die zwei induzierten Topologien, die die f-$\mathcal{E}$-Klasse enthält. Damit wird ein gemeinsamer Rahmen für die bisher meist unabhängig voneinander betrachteten Metriken geschaffen und ihre Zusammenhänge verdeutlicht. Bildlich gesprochen wird hierbei verglichen, wie viel Masse wie weit transportiert werden muss, um eine Verteilung in die andere zu transformieren. Die f-$\mathcal{E}$-Metrik ergibt sich dann aus dem Gleichgewicht zwischen der Entfernung und der mit f-gewichteten Masse. Im Gegensatz zur Wasserstein-p Metrik, die die gesamte transportierte Masse zusammen mit den Entfernungen mittelt, wird bei der f-$\mathcal{E}$-Metrik nur ein Trennpunkt betrachtet. In Proposition 2.5.1 und Proposition 2.5.2 stellen wir zwei allgemeingültige Algorithmen für die exakte Berechnung des Abstands von Wahrscheinlichkeitsverteilungen mit endlichen Trägern der Größen $m \geq n$ vor. Mit einer Worst-Case-Komplexität von $\mathcal{O}\left(mn^2 \log(m)\right)$ ist die Berechnung der f-$\mathcal{E}$-Klasse somit aus dem Stand gleichwertig mit der gut erforschten Wasserstein-p Metrik.

Weiterhin führen wir als Erste quasikonvexe Metriken ein, ein Konzept, das metrische und geordnete Räume miteinander verbindet. Dies erlaubt es uns, die Träger unserer Wahrscheinlichkeitsverteilungen in entsprechenden metrischen Räumen zu ordnen und dabei eine starke Verbindung mit der zugrundeliegenden Metrik zu bewahren. Die theoretische Grundlage hierfür bieten Monge-Folgen, die wir ebenfalls in dieser Arbeit behandeln. Indem wir diese kombinieren, können wir die allgemeine Komplexität deutlich verbessern und erhalten Quasi-Linearität für die Prokhorov und Wasserstein-∞ Metrik . Wir beweisen Korrektheit und Worst-Case-Komplexität für alle Algorithmen und erhalten die selben Komplexitäten wie für die Wasserstein-p Metrik. Im Detail erhalten wir eine allgemeine Komplexität von $\mathcal{O}\left(mn^2 \log(m)\right)$ und

verbessern diese für den quasi-konvexen Fall zu einem stark quasi-linearen $\mathcal{O}\left(m\log(m)\right)$ für die Wasserstein-∞-Metrik in Corollary 4.2.18 und einem schwach quasi-linearen $\mathcal{O}\left(m\max\{\log(m),\frac{1}{acc_X}\}\right)$ in Abhängigkeit vom Träger der Verteilungen für die Prokhorov Metrik in Theorem 4.1.12.
Wir vergleichen die f-$\mathcal{E}$-Klasse mit bestehenden Metriken und zeigen ihre Relation zu diesen. Wir schließen mit einer numerischen Analyse unserer Algorithmen zur Überprüfung der Korrektheit und Komplexität anhand von Implementierungen in MATLAB ab.

Insgesamt haben wir in dieser Arbeit eine breite Klasse von Wahrscheinlichkeitsmetriken neu entwickelt, die die bekannten Prokhorov und Wasserstein-∞ Metriken enthält, ihre theoretischen Eigenschaften analysiert und einen umfassenden Satz von exakten und effizienten Algorithmen für deren Berechnung für Wahrscheinlichkeitsverteilungen mit endlichem Träger vorgestellt.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

Let us start by motivating the use of metrics for probability distributions and why we see the need for the f-$\mathcal{E}$-class.

1.1 History of Metrics in Probability Theory

The usual way of introducing convergence in a classical calculus course is via metrics. It presents a way of mathematically defining convergence as one would understand it figuratively, by quantifying closeness. A metric on the space S is defined as bivariate function $\mathrm{d} : S \times S \to \mathbb{R}^{\geq}$ fulfilling only three simple properties:

- $\mathrm{d}(x, y) = 0 \Leftrightarrow x = y$ for all $x, y \in S$ (identity of indiscernibles)
- $\mathrm{d}(x, y) = \mathrm{d}(y, x)$ for all $x, y \in S$ (symmetry)
- $\mathrm{d}(x, z) \leq \mathrm{d}(x, y) + \mathrm{d}(y, z)$ for all $x, y, z \in S$ (triangle inequality)

These properties equal what one would expect from a distance in everyday life, the "identity of indiscernibles" states that two objects have no distance if and only if they are the same, the "symmetry" keeps the distance the same whether one travels from x to y or the other direction from y to x and the "triangle inequality" ensures detours can't shorten the total way traveled. A sequence of points $x_n \subset S$ is then said to converge to a point $x \in S$, if $\mathrm{d}(x_n, x) \xrightarrow[n\to\infty]{} 0$, i.e. $\forall \varepsilon > 0 \exists N \in \mathbb{N} : \mathrm{d}(x_n, x) < \varepsilon$ for all $n \geq N$. This translates convergence to a finally decreasing to zero distance, as one would expect. Of course, many generalizations and variants exist, including quasimetrics, metametrics and semimetrics, all varying one or more properties. The advantage of working with metrics is their generality – the concept is independent of the actual space S. One intuitively thinks of the real line $\mathbb{R}$ and the euclidean metric $\mathrm{d}(x,y) = |x-y|$ as the standard example of a metric, but many more metric spaces exist, like the space of continuous functions or as in this exposition, the space of probability measures.

By imposing stricter rules on a metric, the typical next step are norms on vector spaces. On the contrary easing up the concept of metrics leads to topologies, which we will look further into. Instead of quantifying a distance to define closeness, a topology directly lists all respective neighborhoods $N \subset S$ of every point $x \in S$, fulfilling certain properties. This states a relaxation of the concept of metrics, as no general comparison of the neighborhoods is possible in contrast to the classic distance in a metric sense.

In probability theory, the classical concepts lie somewhere in between metric and topological spaces: To do so, we consider measures $\mathbb{M}_1, \mathbb{M}_2, \dots$ on a metric space (S, d).
The most common definition of convergence of probability measures is the weak convergence, $\mathbb{M}_n \xrightarrow[n\to\infty]{w} \mathbb{M} :\Longleftrightarrow \mathbb{E}_{\mathbb{M}_n}[f] \xrightarrow[n\to\infty]{} \mathbb{E}_{\mathbb{M}}[f]$ for all bounded and continuous functions $f : S \to \mathbb{R}$. As its name suggests, it is the weakest form of convergence, i.e. other types of convergence imply weak convergence, but not vice versa. We only require the convergence of the expected values, but do not require or quantify a certain speed. This is obviously due to the fact that the convergence speed of $2 \cdot f$ is half the speed of f, i.e. the convergence speed can be slowed down indefinitely. While this problem can be overcome quite easily, e.g. by imposing a constraint on the maximal value $\|f\|_\infty$, it complicates the otherwise straightforward definition. But this also raises concerns regarding the "right" choice of constraint. More common than limiting the maximum is limiting the Lipschitz-norm[1] $\|f\|_{Lip} \leq 1$, which defines the Wasserstein-metric[2] (Gibbs and Su, 2002, Chapter 2, Section

[1]The Lipschitz-norm is defined as $\|f\|_{Lip} := \sup_{x,y\in S} \frac{|f(x)-f(y)|}{\mathrm{d}(x,y)}$
[2]The Wasserstein-metric is defined as $\mathrm{d}_W(\mathbb{M}_1, \mathbb{M}_2) := \sup_{\|f\|_{Lip}\leq 1} \left|\mathbb{E}_{\mathbb{M}_1}[f] - \mathbb{E}_{\mathbb{M}_2}[f]\right|$

Wasserstein). The most prominent occurrence of weak convergence, now for random variables, is the central limit theorem:

Theorem. *Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow[n\to\infty]{w} \mathcal{N}(0, \sigma^2)$$

Application include the approximation of i.i.d. sums by normal distributions to simplify modeling and calculations. This justifies the regularly made assumption of normally distributed errors in statistics.

Stronger types of convergence of sequences of random variables include convergence in probability and almost sure convergence. In contrast to weak convergence, they relate to sequences of real-valued random variables $X_1, X_2, \ldots$ instead of probability distributions $\mathbb{M}_1, \mathbb{M}_2, \ldots$.

Definition. *Let $X_1, X_2, \ldots$ and X be random variables, X_n is said to converge in probability to X, if*

$$\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \varepsilon\right] = 0 \qquad \textit{for all } \varepsilon > 0.$$

The sequence is said to converge almost surely, if

$$\mathbb{P}\left[\lim_{n\to\infty} X_n = X\right] = 1.$$

As both notations define a stricter sense of convergence, they imply weak convergence for the associated distributions. Common occurrences are the law of large numbers and the law of iterated logarithm.

Theorem. *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean 0 and variance 1, then*

$$\limsup_{n\to\infty} \frac{\pm \sum_{i=1}^{n} X_i}{\sqrt{2n \log\log n}} = 1 \quad \textit{almost surely.}$$

In contrast to other classical convergence results stating characteristics of the limit, the law of iterated logarithms also quantifies how the limit is approached. It states how far the sample average will deviate from the mean on a regular basis.

A result lying in between classical convergence results and our desired focus on quantifying distances is the Dvoretzky–Kiefer–Wolfowitz inequality for real valued random variables.

Theorem. *Dvoretzky et al. (1956)*
Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with cumulative distribution function F and denote by F_n the empirical distribution function of the first n observations, i.e. $F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$, then

$$\mathbb{P}\left[\sup_{x\in\mathbb{R}}\left|F_n(x) - F(x)\right| > \varepsilon\right] \leq Ce^{-2n\varepsilon^2} \qquad \text{for all } \varepsilon > 0,$$

for some constant $C > 0$. Massart (1990) later showed $C = 2$.

Essentially, the probability of large deviations in the distribution function declines exponentially with the sample size. This can be used in two ways, one by having a fixed sample size N and obtaining a-posteriori exceedance probabilities for all error sizes ε. Or by a-priori defining maximal exceedance probabilities for certain error levels and calculating a necessary sample size guaranteeing these bounds. This is especially relevant for regulated areas, i.e. risk management in finance or medicines studies.
The deviation measured in the Dvoretzky–Kiefer–Wolfowitz inequality is well known as the Kolmogorov metric

Definition. *Let F, G be cumulative distribution functions, then their Kolmogorov distance is defined as*

$$\mathrm{d}_K(F, G) := \sup_{x\in\mathbb{R}}\left|F(x) - G(x)\right|.$$

It is well known from the Kolmogorov(-Smirnov)-test, a statistical method to check whether a sampled distribution stems from a given distribution (zero-hypothesis) or not. While being restricted to real-valued distributions, its computation is already challenging. As $F(x) - G(x)$ is the difference of two cumulative distribution functions, it possesses in general no helpful properties besides being continuous from right with the left sided limit existing (càdlàg). Even if densities are available, the constant sign changes complicate matters. A setting providing easier means of computation is if at least one distribution is discrete or even has finite support, i.e. $G(x) = \sum_{i=1}^{l} \mathbb{1}_{X_i \leq x}$ for some X_i. If we denote the support as $\Gamma \coloneqq \{X_i \mid i = 1, \dots, l\}$, the calculation of the supremum can be reduced to a finite set

$$\sup_{x \in \mathbb{R}} \left| F(x) - G(x) \right| = \max_{i=1,\dots,l} \max\{\left| F(X_i) - G(X_i) \right|, \left| F(X_i^-) - G(X_i^-) \right|\},$$

with $F(x^-) \coloneqq \lim_{y \uparrow x} F(y)$ denoting the limit from below.

Our interest in probability metrics was sparked by a problem in stochastic optimization, how sensitive the optimal solution is with respect to deviations in the distribution. Consider a classical stochastic optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f_{\mathbb{M}_1}(x, Z) \\ \text{subject to} \quad & x \in C \end{aligned} \tag{$SP_{\mathbb{M}_1}$}$$

for an optimal variable $x \in \mathbb{R}^d$ constrained to a set C and a random parameter Z. Here the objective function $f_{\mathbb{M}_1}$ is parametrized by a probability measure $\mathbb{M}_1$, e.g. $\mathbb{E}_{\mathbb{M}_1}[g(x, Z)]$ or a quantile of the distribution of $g(x, Z)$ under $\mathbb{M}_1$. Let us denote the optimal solution of $SP_{\mathbb{M}_1}$ by $x^*_{\mathbb{M}_1}$. This setting assumes exact knowledge of the distribution $\mathbb{M}_1$, which often holds not true, either by necessary discretizations or simply missing information. So instead of the real problem $SP_{\mathbb{M}_1}$, one might find oneself solving $SP_{\mathbb{M}_2}$ for a slightly different distribution $\mathbb{M}_2$, obtaining a different optimal control $x^*_{\mathbb{M}_2}$. Important questions then include how much the two optimal solutions deviate from each other in relation to the distance between $\mathbb{M}_1$ and $\mathbb{M}_2$,

$$\left\| x^*_{\mathbb{M}_1} - x^*_{\mathbb{M}_2} \right\| \leq h\left(\mathrm{d}\left(\mathbb{M}_1, \mathbb{M}_2 \right) \right)$$

for some $h : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ or how much worse $x^*_{\mathbb{M}_2}$ performs than $x^*_{\mathbb{M}_1}$ in the real problem

$$f_{\mathbb{M}_1}(x^*_{\mathbb{M}_2}, Z) - f_{\mathbb{M}_1}(x^*_{\mathbb{M}_1}, Z) < h\left(\mathrm{d}\left(\mathbb{M}_1, \mathbb{M}_2 \right) \right).$$

That means given $\mathrm{d}(\mathbb{M}_1, \mathbb{M}_2)$, we can bound the maximal deviations of the optimal control and its objective value. In the case of discretization, this can be used to calculate a priori an acceptable discretization level for approximatively solving $SP_{\mathbb{M}_1}$.

1.2 Joining Probability Theory and Optimization

While there certainly is an interest in probability metrics from a probability theory point of view, their computation has only been covered selectively. Probably the most researched metric is the Wasserstein metric due to its link to optimal transport. Figuratively, optimal transport is the problem of matching given supplies and demands of one good while minimizing the occurring costs. Imagine heaps of earth of different sizes $p_1, \ldots, p_m$ at locations $X_1, \ldots, X_m$ and a plan how the earth should be distributed at locations $Y_1, \ldots, Y_n$ with masses $q_1, \ldots, q_n$; and denote the distance between the sites X_i and Y_j by $c_{i,j}$.

The problem dates back to the eighteenth century, when it was introduced by Monge (1781). In its simplest form of equal supply and demand on each site, i.e. $p_1 = \cdots = p_m = q_1 = \cdots = q_n$ with $m = n$, this reduces to an assignment problem. The optimal transportation plan can be expressed as an permutation $\sigma : \{1, \ldots, m\} \to \{1, \ldots, m\}$, linking each point of supply i with exactly one point of demand $\sigma(i)$. As they allow a specifically tailored solving approach and constitute a whole area of research themselves, we do not go into detail here and only mention the most prominent algorithm: assignment problems can be solved by the so called “Hungarian method”, introduced by Kuhn (1955), based on ideas by the Hungarian mathematicians Dénes Kőnig and Jenő Egerváry.

The problem was generalized by Kantorovitch (1958), therefore also referred to as the Kantorovich-Monge problem. As linear programming, it developed as a way to quantify logistical problem during World War II, like the rationing of food and reallocating troops. Allowing arbitrary positive supplies and demands, only requiring equal totals $\sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j$, gives a min cost flow problem by introducing source s and sink t:

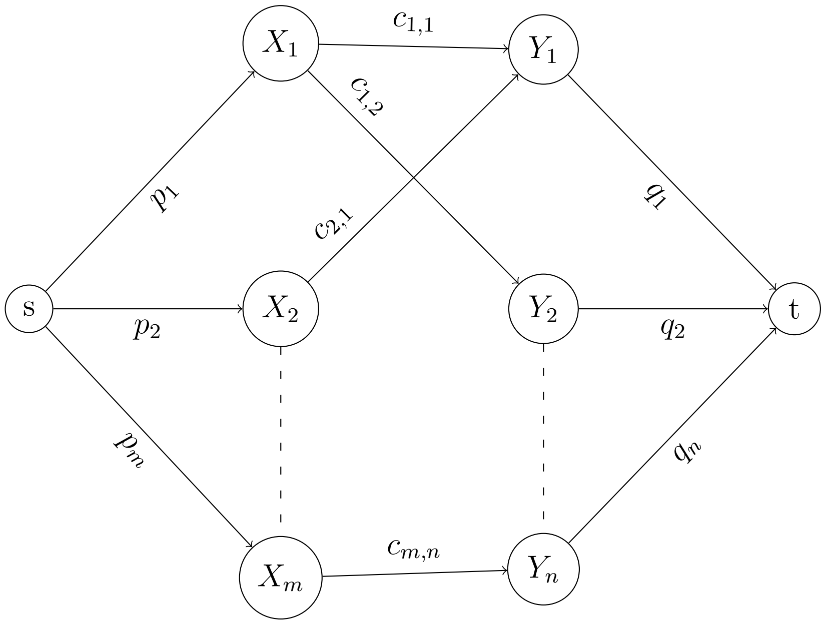


Figure 1.1: Network of a min cost flow formulation of the optimal transport problem with supply nodes $X_1, \ldots, X_m$, demand nodes $Y_1, \ldots, Y_n$. Labels below edges are flow constraints, labels above costs, missing labels are no flow constraints and zero costs.

In recent years, optimal transport became more and more relevant again, especially due to Villani (2009), a book thoroughly dedicated to this topic. Over the years, nearly all important scientific prizes have been awarded to related topics, especially the Nobel prize in economics 1975 to Leonid Kantorovich and Tjalling Koopmans “"for their contributions to the theory of optimum allocation of resources"”, the Henri Poincaré prize 2009 and the Fields medal in 2010 to Cédric Villani. Villani (2009) covers in detail the theory of continuous optimal transport, instead of finite sets of supply and demand, continuous areas are considered. This is achieved by using density functions f and g instead of probability masses p_i and q_j, norming the total mass to 1. While previously transport maps were represented by matrices $m_{i,j}$, now maps $m(x, y)$ are used. Their interpretation of the concepts remains the same, but mathematically, new opportunities arise. The gradient $\nabla m(x, y)$ is of special interest as it has geometric meaning in the field of gradient flows. Throughout this thesis, we focus solely on the discrete setting, but refer to Ambrosio et al. (2005) for a detailed introduction.

This connects to our area of interest by interpreting every outcome X_i/Y_j and their respective probabilities p_i/q_j of a random variables as supply (demand) and its mass. An obvious consequence is a total mass of supply and demand of $\sum_{i=1}^{m} p_i = 1 = \sum_{j=1}^{n} q_j$. The distance between two probability distributions is therefore the cost of transforming one into the other.

So far, we have just superficially used the term "cost" of transportation without specifying how to calculate it. The most intuitive approach is summing up all occurring path costs multiplied by the amount of goods transported on it, $m_{i,j}$ from X_i to Y_j, $\sum_{i,j=1}^{m,n} m_{i,j}c_{i,j}$. This defines the Wasserstein-1 metric, the most prominent probability metric. In probability theoretic terms, this can be interpreted as minimizing the mean cost. By squaring the path costs, $\sum_{i,j=1}^{m,n} m_{i,j}c_{i,j}^2$, we analogously minimize the variance of cost, with its square-root defining the Wasserstein-2 metric. In general, the Wasserstein-p metric for $1 \leq p < \infty$ is defined as $\sqrt[p]{\sum_{i,j=1}^{m,n} m_{i,j}c_{i,j}^p}$. These are well studied problems, both mathematically and numerically, discrete and continuously.

One closely related metric we will cover in this thesis is the Wasserstein-∞ metric, the limit $p \to \infty$ of the Wasserstein-p metric. It follows the classical intuition of p-metrics, increasing the weight of bigger costs and in the limit only taking the most expensive path into account $\max_{i,j} \mathbb{1}_{m_{i,j}>0} \cdot c_{i,j}$. Applications include budgeting rules minimizing maximum costs per path or contractor, or if one considers the transportation time as its cost, minimizing the time it takes to transport all goods.

The Wasserstein metrics stem directly from economic concepts, but a multitude of mathematically interesting cost functions is available. A major problem in real world statistics is the occurrence of outliers and their methodical treatment. How does one decide whether an extreme observation is the result of an unlikely event or simply a measurement error? One metric dealing with this trade-off is the Prokhorov metric Prokhorov (1956). It can roughly be stated as “the smallest value δ such that at most δ goods have to be transported at costs higher than δ”. This eases outliers by ignoring the costs for transporting a small fraction δ of the total mass along high cost paths.

1.3 Literature around the f-$\mathcal{E}$-class

While we are the first to introduce f-$\mathcal{E}$-metrics, two representatives are well known, the Prokhorov and the Wasserstein-∞ metric. As the name suggests, the Prokhorov metric was first introduced by Prokhorov (1956). His motivation was to quantify weak convergence for stochastic processes, but the resulting metric can be used for any metric space. Strassen (1965) built on this to obtain results for the existence of specific joint measures given their marginals by looking at the neighborhood of sets. This can be seen as the first graphical interpretation of the Prokhorov metric. Dudley (1968) was able to further improve this visual interpretation, and linking it to matching problems, a special case of the previously discussed transportation problems. While results up to then were for general probability measures, Schay (1974) restricted the setting to probability measures with finite support. This was the first step towards computing the Prokhorov metric, introducing a linear programming formulation of an occurring subproblem. The dual problem, the focus of García-Palomares and Giné (1977), provides an alternative approach to the results of Schay (1974).

The restriction to probability measures with finite support raised the question how good these approximate continuous ones. As the Prokhorov metric metrizises weak convergence, the only common metric to do so without further restrictions, the convergence of the sampled distribution to the original one is given for increasing sample size. Kersting (1978) addressed the question of convergence speed, providing results for distributions on $\mathbb{R}$ under some regularity constraints.

Literature around the f-$\mathcal{E}$-class

We will build on Garel (1981); Garel and Massé (2009), the first to provide a computation algorithm for the Prokhorov metric for probability measures with finite support. As they were focused on fast approximation via discretization of continuous measures, they were not interested in exactness. However, we will refine one of their steps to develop exact and efficient computation algorithms. As the Prokhorov metric has gained mostly theoretical attention, we are not aware of other computational results. This makes us the first to provide exact algorithms with known worst case complexity.

The Prokhorov metric uses a fixed $1:1$ ratio to interchange costs and outlier mass. We focus on investigating a class of probability metrics which are robust with respect to such outliers, but allowing a free and dynamic choice of the exchange rate. To the best of our knowledge, we are to first to introduce this class of metrics, based on a result by Dudley (1968), which we refer to as the f-$\mathcal{E}$-class. It incorporates the Wasserstein-∞ metric and the Prokhorov metric and embeds them in a consistent framework.

The Wasserstein-∞ metric is usually defined as the limit $p \to \infty$ of the Wasserstein-p metric. A different approach, as discussed by Givens and Shortt (1984), defines it as the minimized maximal covered distance of the transportation plan, allowing for easy visualization. This is in-line with the expected behavior, as an higher exponent weights large deviances more, ultimately only weighting the largest deviance. Champion et al. (2008) characterizes the existence of optimal transport maps realizing the Wasserstein-∞ metric. Similar to the Prokhorov metric, advances for the convergence rate of sampled measures have been made, notably by Kloeckner, B. (2012); Trillos and Slepčev (2015); Liu et al. (2018).
Bobkov and Ledoux (2016) provided an explicit characterization of an optimal transportation plan for probability measures on the real line. We generalize this result for certain ordered, metric spaces, and develop exact and efficient algorithms for probability measures with finite support on such metric spaces.

Our f-$\mathcal{E}$-class provides a uniform framework for the Prokhorov and

Wasserstein-∞ metric, while also introducing a whole new class of metrics along their structure. A first step in this direction has been done by Rachev and Rüschendorf (1992), but only for a limited, parametrized class of metrics.

1.4 Contributions and Structure of this Thesis

We start by introducing what we call the f-$\mathcal{E}$-class in Definition 2.1.10, a framework of probability metrics inspired by the Prokhorov-metric in Chapter 2. We are able to separate the class into two disjoint sets in Theorem 2.1.21 and Lemma 2.1.23 by their created topologies and therefore showing the existence of uncountable infinite metrics equivalent to the weak topology. After analyzing the theoretical properties, we provide two exact and efficient algorithms in Proposition 2.5.1 and Proposition 2.5.2 for their computation for finitely-supported probability distributions. To the best of our knowledge, we are the first to do so. While for the Wasserstein-∞ metric special cases like $\mathbb{R}$ have been analyzed, the Prokhorov metric has never been computationally analyzed for exact algorithms. We achieve the same computational complexity as the widely used Wasserstein-p metric. We therefore provide a metric robust to outliers as an viable alternative to the current state of the art.

In Chapter 3 we introduce quasiconvex metrics in Definition 3.3.1, motivated by Norfolk (1991), a generalization of one-dimensional spaces retaining necessary ordering properties allowing fast computation of our metrics. While the concept is a straight forward combination of ordered and metric spaces, they have not been the subject of research until now. This leads to the discussion of Monge sequences, a setting where greedy algorithms are exact algorithms, see Theorem 3.2.3. We show how we can modify our algorithms to take advantage of the ordering of the support and achieve significantly reduced complexities in Theorem 3.4.3.

Chapter 4 is dedicated to the Prokhorov and Wasserstein-∞ metric in detail, covering their theory and providing specifically tailored algorithms. For the Prokhorov metric, we show a significant improvement Corollary 4.1.8 of the approach of Garel and Massé (2009). We then focus on the Prokhorov metric on $\mathbb{R}$ in Subsection 4.1.5 and present an exact, weakly quasi-linear algorithm Theorem 4.1.12 to show the computationally accessibility of the Prokhorov metric. We hope this inspires further research along the Prokhorov metric, which so far has been appealing only from a theoretical point of view, but was not considered to be computationally feasible.

Along the same idea, we analyze the Wasserstein-∞ metric on quasi-convex spaces in Subsection 4.2.5 and present an exact, strongly quasi-linear algorithm Corollary 4.2.18. This generalizes the results of (Bobkov and Ledoux, 2016, Equation (2.3)) from $\mathbb{R}$ with the euclidean metric to general quasi-convex spaces. We summarize our findings in Section 4.3, showing the Prokhorov and Wasserstein-∞ metrics are computationally on par with the Wasserstein-p metric.

We embed the f-$\mathcal{E}$-class into well known probability metrics, by comparing it to the Wasserstein-p and Lévy metrics in Chapter 5.

To show the practicality of our newly introduced algorithms, we conduct a numerical analysis in Chapter 6 of the Prokhorov metric, verifying the theoretical complexities and efficiency of the quasi-convex setting.

We conclude with Chapter 7, summarizing our achievements and providing an outlook to inspire further research.

2 Metrics of the f-$\mathcal{E}$-Class

In this thesis, we cover two metrics for probability measure from a class going back to (Dudley, 1968, Theorem 1).

2.1 Definition and Properties

As these metrics take into account deviations in the probabilities as well as in the support, we need to define the neighborhood of a Borel set. We will follow Huber (1981) throughout this part.

Let Ω be a Polish space (i.e. separable and complete) with a metric d_Ω and Borel-σ-algebra $\mathfrak{A} := \mathcal{B}(\Omega)$. Furthermore, let $\mathcal{M}_1(\Omega, \mathfrak{A})$ be the set of all probability measures on $[\Omega, \mathfrak{A}]$.

Definition 2.1.1
For $\delta \geq 0$, let the closed δ-neighborhood A^δ of a Borel set $A \in \mathfrak{A}$ be given by

$$A^\delta := \left\{ x \in \Omega \mid \mathrm{d}_\Omega(x, A) \leq \delta \right\},$$

where $\mathrm{d}_\Omega(x, A) := \inf_{y \in A} \mathrm{d}_\Omega(x, y)$ is the distance of a point x to the set A.

(Dudley, 1968, Theorem 1) covers the relationship between the behavior of $\mathbb{M}_1[A]$ and $\mathbb{M}_2[A^\delta]$ for two probability measure $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ and an optimal coupling $\mathbb{M}$ of these two. To fix notation, let us denote by $\mathcal{M}_1^{\otimes}(\Omega, \mathfrak{A}) := \mathcal{M}_1(\Omega \times \Omega, \mathfrak{A} \otimes \mathfrak{A})$ the product space and abbreviate

$$\mathbb{M}\left[\mathrm{d}_\Omega(x, y) > \delta\right] := \mathbb{M}\left[\left\{(x, y) \in \Omega \times \Omega \,\middle|\, \mathrm{d}_\Omega(x, y) > \delta\right\}\right]$$

for $\mathbb{M} \in \mathcal{M}_1^{\otimes}(\Omega, \mathfrak{A})$. Further, let us abbreviate the set of probability measures on $\mathcal{M}_1^{\otimes}(\Omega, \mathfrak{A})$ with given marginal measures $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ as $\mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)$, i.e.

$$\mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2) := \{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\Omega, \mathfrak{A}) \mid \pi_1(\mathbb{M}) = \mathbb{M}_1 \text{ and } \pi_2(\mathbb{M}) = \mathbb{M}_2\},$$

with projections $\pi_1(\mathbb{M})[A] = \mathbb{M}[A \times \Omega]$ and $\pi_2(\mathbb{M})[B] = \mathbb{M}[\Omega \times B]$ for all $A, B \in \mathfrak{A}$.

Theorem 2.1.2
(Dudley, 1968, Theorem 1) Let $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ *and* $\delta, \varepsilon \geq 0$*, then the following are equivalent:*

1. $\mathbb{M}_1[A] \leq \mathbb{M}_2[A^\delta] + \varepsilon$ *for all closed* $A \subset \Omega$.
2. *For any* $\gamma > 0$ *there exists a* $\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)$ *with* $\mathbb{M}\left[\, \mathrm{d}_\Omega(x, y) > \delta + \gamma \right] \leq \varepsilon + \gamma$.

If furthermore $[\Omega, \mathrm{d}_\Omega]$ is inner regular, i.e. for all probability measures $\mathbb{M}_1 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ and all Borel sets $A \in \mathfrak{A}$ it holds

$$\mathbb{M}_1[A] = \sup \left\{ \mathbb{M}_1[K] \mid K \subset A, K \text{ compact} \right\},$$

then the same holds for $\gamma = 0$.

Lemma 2.1.3
(Dudley, 1968, p. 1566 ff.) If $[\Omega, \mathrm{d}_\Omega]$ *is a Polish space,* $[\Omega, \mathfrak{A}]$ *is inner regular.*

Theorem 2.1.4
(Dudley, 1968, Theorem 2) Let $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ *and* $\delta, \varepsilon \geq 0$*, then, if* $[\Omega, \mathrm{d}_\Omega]$ *is inner regular, the following are equivalent:*

1. $\mathbb{M}_1[A] \leq \mathbb{M}_2[A^\delta] + \varepsilon$ *for all closed* $A \subset \Omega$.
2. *There exists a* $\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)$ *with* $\mathbb{M}\left[\, \mathrm{d}_\Omega(x, y) > \delta \right] \leq \varepsilon$.

As d_Ω is symmetric, an obvious consequence of Theorem 2.1.2 is the symmetry of the first property:

Corollary 2.1.5
For all $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ *and* $\delta, \varepsilon \geq 0$

$\mathbb{M}_1[A] \leq \mathbb{M}_2[A^\delta]+\varepsilon$ *for all closed* $A \subset \Omega \iff \mathbb{M}_2[A] \leq \mathbb{M}_1[A^\delta]+\varepsilon$ *for all closed* $A \subset \Omega$

holds.

To obtain the smallest ε fulfilling this relationship, we introduce the exceedance function, i.e. how much more mass lies outside the δ neighborhood A^δ.

Definition 2.1.6
Let us define the exceedance of $\mathbb{M}_1$ *and* $\mathbb{M}_2$ *for* $\delta \geq 0$ *as*

$$\mathcal{E}^{\mathbb{M}_1,\mathbb{M}_2}(\delta) := 1 - \sup_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1,\mathbb{M}_2)} \mathbb{M}\left[\,\mathrm{d}_\Omega(x,y) \leq \delta\right] = \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1,\mathbb{M}_2)} \mathbb{M}\left[\,\mathrm{d}_\Omega(x,y) > \delta\right],$$

where we omit the margins if they are clear from the context, i.e. $\mathcal{E}(\delta) := \mathcal{E}^{\mathbb{M}_1,\mathbb{M}_2}(\delta)$.

Again, inner regularity allows for an easier formulation.

Corollary 2.1.7
If $[\Omega,\, \mathrm{d}_\Omega]$ *is inner regular, we have*

$$\sup_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1,\mathbb{M}_2)} \mathbb{M}\left[\,\mathrm{d}_\Omega(x,y) \leq \delta\right] = \max_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1,\mathbb{M}_2)} \mathbb{M}\left[\,\mathrm{d}_\Omega(x,y) \leq \delta\right]$$

for all $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ *and* $\delta \geq 0$*, and especially for our notation*

$$\mathcal{E}(\delta) = \min_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1,\mathbb{M}_2)} \mathbb{M}\left[\,\mathrm{d}_\Omega(x,y) > \delta\right].$$

Observation 2.1.8
The definition of the exceedance $\mathcal{E}$ *is in line with the idea of Theorem 2.1.4 in*

case of inner regularity:

$$\mathbb{M}_1[A] \leq \mathbb{M}_2[A^\delta] + \mathcal{E}(\delta) \text{ for all } \delta > 0 \text{ and } \mathbb{M}_1,\, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A}).$$

Furthermore as $\mathcal{E}$ is the smallest value admitting this, an alternative definition of $\mathcal{E}$ is obviously

$$\mathcal{E}(\delta) = \sup_{\substack{A \subset \Omega \\ A \text{ closed}}} \mathbb{M}_1[A] - \mathbb{M}_2[A^\delta]$$

The motivation behind the exceedance $\mathcal{E}(\delta)$ is the amount of mass that has to be transported further than δ in an optimal transportation plan.

As we will extensively study the exceedance $\mathcal{E}$, let us give an overview of its properties.

Lemma 2.1.9
The exceedance $\mathcal{E}$ satisfies the following:

1. *$\mathcal{E}$ is monotonically decreasing with $\lim_{\delta \to \infty} \mathcal{E}(\delta) = 0$.*
2. *$\mathcal{E}$ is càdlàg, i.e. continuous from right and the limit from the left exists.*
3. *$\mathcal{E}(\delta) \in [0, 1]$ for all $\delta \geq 0$.*
4. *$\mathcal{E}(0) = 0 \iff \mathbb{M}_1 = \mathbb{M}_2$.*

Proof. All of these properties are due the close relationship of $\mathcal{E}$ to the cumulative distribution function. □

This allows us to define a broad class of metrics for probability measures by prescribing a certain relationship between δ and $\mathcal{E}(\delta)$ and searching for the smallest such δ:

Definition 2.1.10
Let $\mathbb{M}_1,\, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ and $f : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ be monotonically increasing, then

$$\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) \coloneqq \inf\left\{\delta \geq 0 \mid f(\delta) \geq \mathcal{E}^{\mathbb{M}_1, \mathbb{M}_2}(\delta)\right\}$$

defines a prototype for distances on $\mathcal{M}_1(\Omega, \mathfrak{A})$, which we will call f-$\mathcal{E}$-prototypes. We say f defines a f-$\mathcal{E}$-metric if $\mathrm{d}_f(\cdot,\cdot)$ fulfills the metric properties.[1]

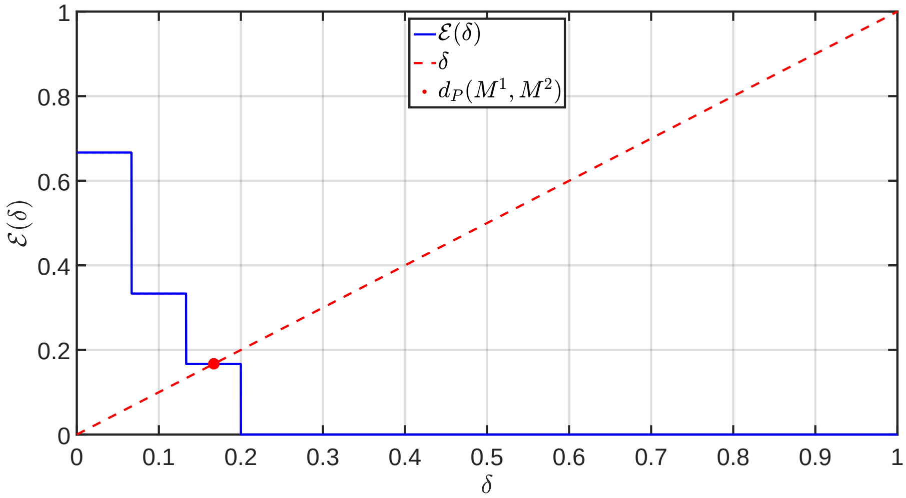


Figure 2.1: Illustration of an f-$\mathcal{E}$-metric for $f(\delta) = \delta$, which we will later identify as the Prokhorov metric d_P, for a small example with $\mathbb{M}_1[\{0\}] = \mathbb{M}_1[\{\frac{1}{5}\}] = \mathbb{M}_1[\{\frac{2}{5}\}] = \mathbb{M}_1[\{\frac{3}{5}\}] = \mathbb{M}_1[\{\frac{4}{5}\}] = \mathbb{M}_1[\{1\}] = \frac{1}{6}$ and $\mathbb{M}_2[\{0\}] = \mathbb{M}_2[\{\frac{1}{3}\}] = \mathbb{M}_2[\{\frac{2}{3}\}] = \mathbb{M}_2[\{1\}] = \frac{1}{4}$.

To assure this defines a metric on $\mathcal{M}_1(\Omega, \mathfrak{A})$, further restrictions for f must hold:

Lemma 2.1.11

1. *If f is càdlàg and $f(0) = 0$ holds, $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) = 0$ if and only if $\mathbb{M}_1 = \mathbb{M}_2$.*
2. *If $f(\delta_0) > 0$ for some $\delta_0 > 0$, $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) < \infty$ for all $\mathbb{M}_1$, $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$.*

[1] Alternatively, one could prescribe a relationship between ε and $g(\varepsilon)$, following our nomenclature called $g(\varepsilon)$-ε-metrics, but we will not pursue this approach.

3. *If f is super-additive, the triangle inequality holds:* $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_3) \leq \mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) + \mathrm{d}_f(\mathbb{M}_2, \mathbb{M}_3)$ *for all* $\mathbb{M}_1$, $\mathbb{M}_2$, $\mathbb{M}_3 \in \mathcal{M}_1(\Omega, \mathfrak{A})$

Proof.

1. "$\Rightarrow$" Let $\mathbb{M}_1 \neq \mathbb{M}_2$, that is $\mathcal{E}(0) > 0$, and as $\mathcal{E}$ and f are càdlàg:

$$\lim_{\delta \downarrow 0} \mathcal{E}(\delta) = \mathcal{E}(0) > 0 = f(0) = \lim_{\delta \downarrow 0} f(\delta),$$

   and therefore $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) > 0$.
   "$\Leftarrow$" For $\mathbb{M}_1 = \mathbb{M}_2$ we have $\mathcal{E}(\delta) = 0$ for all $\delta \geq 0$ and therefore $\mathcal{E}(\delta) \leq f(\delta)$ for all $\delta \geq 0$.

2. We have $\lim_{\delta \to \infty} \mathcal{E}(\delta) = 0 < f(\delta_0) \leq \lim_{\delta \to \infty} f(\delta)$, so there exists a $\delta \geq 0$ such that $f(\delta) \geq \mathcal{E}(\delta)$.

3. Let $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) = \delta_{1,2}$ and $\mathrm{d}_f(\mathbb{M}_2, \mathbb{M}_3) = \delta_{2,3}$, then

$$\mathbb{M}_1[A] \leq \mathbb{M}_2[A^{\delta_{1,2}}] + f(\delta_{1,2}) \leq \mathbb{M}_3[\left(A^{\delta_{1,2}}\right)^{\delta_{2,3}}] + f(\delta_{1,2}) + f(\delta_{2,3}),$$

   so it suffices to show $\left(A^{\delta_{1,2}}\right)^{\delta_{2,3}} \subset A^{\delta_{1,2}+\delta_{2,3}}$ which follows directly from the triangle inequality for d_Ω and $f(\delta_{1,2}) + f(\delta_{2,3}) \leq f(\delta_{1,2} + \delta_{2,3})$, the super-additivity of f. So we know $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_3) \leq \delta_{1,2} + \delta_{2,3}$.

$\square$

Combing all these properties, we obtain sufficient criteria for f to define a f-$\mathcal{E}$-metric on $\mathcal{M}_1(\Omega, \mathfrak{A})$:

Corollary 2.1.12
Let $f : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ *be*

- *monotonically increasing,*
- *càdlàg*[2] *and*
- *super-additive*[3] *with*

[2]While this restriction can be relaxed to continuity in 0 as super-additivity implies the existence of all left-sided limits, càdlàg allows for a more intuitive interpretation.
[3]This implies monotonically increasing, which we list separately for illustration.

- *$f(0) = 0$ and $f(\delta_0) > 0$ for some $\delta_0 > 0$,*

then $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) := \inf\left\{\delta \geq 0 \mid f(\delta) \geq \mathcal{E}^{\mathbb{M}_1,\mathbb{M}_2}(\delta)\right\}$ *defines a metric on* $\mathcal{M}_1(\Omega, \mathfrak{A})$.

This introduces a broad class of new metrics for probability measures, parametrized by the function f. A more restrictive concept has been explored by Rachev and Rüschendorf (1992), looking at linear functions f. Most metrics for probability measures provide an intuitive explanation or visual concept as to why they are defined the way they are. For the Wasserstein-metric it is the average cost of transporting the supply $\mathbb{M}_1$ to fulfill the demand $\mathbb{M}_2$, hence the name Earthmover's distance. The Kullback-Leibler divergence is closely related to the concept of entropy and the Lévy- and Kolmogorov-metric have graphical interpretations within cumulative distribution functions. The motivation for the f-$\mathcal{E}$-metrics is in the sense of the Wasserstein-metric:
Having $\mathrm{d}_f = \delta$ can be interpreted as "No more than $f(\delta)$ mass has to be transported further than δ".

Famous examples of this class are the Wasserstein-∞ metric $\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2)$ with $f(\delta) \equiv 0$ and the Prokhorov metric $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$ with $f(\delta) = \delta$.

Remark 2.1.13
Note $f(\delta) \equiv 0$ does not fulfill Corollary 2.1.12, but still defines a f-$\mathcal{E}$-metric if for all $\mathbb{M}_1, \mathbb{M}_2$, there exists a δ_0 such that $\mathcal{E}^{\mathbb{M}_1,\mathbb{M}_2}(\delta_0) = 0$. If $\sup_{x,y\in\Omega} \mathrm{d}_\Omega(x, y)$ is bounded, this is always fulfilled.

As the main motivation for metrics is the quantification of closeness, their relationship with topologies is of great interest.

Definition 2.1.14
The tuple $\mathcal{T} \subset \mathcal{P}(M)$ is called a topology on the set M, if

1. *$\emptyset \in \mathcal{T}$ and $M \in \mathcal{T}$*

2. $\cup_{i\in\mathcal{I}} A_i \in \mathcal{T}$ *for all* $A_i \in \mathcal{T}$ *for all finite and infinite index sets* $\mathcal{I}$
3. $\cap_{i\in\mathcal{I}} A_i \in \mathcal{T}$ *for all* $A_i \in \mathcal{T}$ *for all finite index sets* $\mathcal{I}$

While metrics provide a quantified distance $\mathrm{d}(\mathbb{P},\cdot)$, topologies only provide the concept of neighborhoods, e.g. open sets containing $\mathbb{P}$. Convergence with respect to a topology is therefore defined as ultimately joining every neighborhood:

Definition 2.1.15
A sequence $\mathbb{P}_n \in \mathcal{M}_1(\Omega,\mathfrak{A})$ *converges to* $\mathbb{P} \in \mathcal{M}_1(\Omega,\mathfrak{A})$ *with respect to the topology* $\mathcal{T}$*, if*

$$\forall A \in \mathcal{T} : \mathbb{P} \in A \quad \exists N \in \mathbb{N} : \quad \mathbb{P}_n \in A \text{ for all } n \geq N$$

and we write $\mathbb{P}_n \xrightarrow[n\to\infty]{\mathcal{T}} \mathbb{P}$.

As every metric space produces a topology by taking all open sets with respect to the metric, one is interested in their relationship.

Definition 2.1.16
We say a metric $\mathrm{d} : M \times M \to \mathbb{R}$ *metrizises a topology* $(M,\mathcal{T})$*, if their notions of convergence are equivalent:*

$$\mathbb{P}_n \xrightarrow[n\to\infty]{\mathcal{T}} \mathbb{P} \quad \Longleftrightarrow \mathrm{d}(\mathbb{P}_n,\mathbb{P}) \xrightarrow[n\to\infty]{} 0 \quad \text{for all } \mathbb{P},\mathbb{P}_n \in M.$$

For probability measures, the most common topology and concept of convergence is the weak topology. We will start by recapping the weak convergence and analyzing its neighborhoods, leading to properties of f necessary and sufficient for d_f to metrizise the weak topology.

Definition 2.1.17
A sequence $\mathbb{P}_n \in \mathcal{M}_1(\Omega,\mathfrak{A})$ *converges weakly to* $\mathbb{P} \in \mathcal{M}_1(\Omega,\mathfrak{A})$*, if*

$$\int_\Omega f(\omega)\,\mathrm{d}\,\mathbb{P}_n \quad \xrightarrow[n\to\infty]{} \quad \int_\Omega f(\omega)\,\mathrm{d}\,\mathbb{P} \quad \text{for all } f : \Omega \to \mathbb{R} \text{ continuous and bounded}$$

and we write $\mathbb{P}_n \xrightarrow[n\to\infty]{w} \mathbb{P}$.

We will cover some of the various equivalent characterizations in the following Lemma.

Lemma 2.1.18

Portmanteau Theorem

The following statements are equivalent for all $\mathbb{P}_n$, $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathfrak{A})$:

1. $\mathbb{P}_n \xrightarrow[n\to\infty]{w} \mathbb{P}$.
2. $\liminf_{n\to\infty} \mathbb{P}_n[G] \geq \mathbb{P}[G]$ *for all open* $G \subset \Omega$.
3. $\limsup_{n\to\infty} \mathbb{P}_n[A] \leq \mathbb{P}[A]$ *for all closed* $A \subset \Omega$.
4. $\lim_{n\to\infty} \mathbb{P}_n[B] = \mathbb{P}[B]$ *for all* $B \subset \Omega$ *with* $\mathbb{P}$*-null boundary* ($\mathbb{P}[int(A)] = \mathbb{P}[A] = \mathbb{P}[cl(A)]$).

A proof can be found in (Huber, 1981, Lemma 2.2).

An easier way to define a topology is through a base:

Definition 2.1.19

We say a collection $\mathcal{B}$ *of subsets of* M *is the base of a topology on* M, *if*

1. $\cup_{B\in\mathcal{B}} B = M$, *i.e.* $\mathcal{B}$ *covers* M *and*
2. *for* $B_1, B_2 \in \mathcal{B}$, *for all* $x \in B_1 \cap B_2$, *there exists a* $B_3 \in \mathcal{B}$ *such that* $x \in B_3 \subseteq B_1 \cap B_2$.

We say $\mathcal{B}$ *is the base of* $\mathcal{T}$, *if*

$$\mathcal{T} = \left\{\cup\mathcal{A} \mid \mathcal{A} \subseteq \mathcal{B}\right\}$$

Directly following from this Lemma, we obtain various bases for neighborhood systems in the weak topology.

Corollary 2.1.20

The following sets form bases for the neighborhood of a probability measure $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathfrak{A})$ with respect to the weak topology for finite sets Φ of bounded continuous functions, $\mathcal{G}$ of open sets, $\mathcal{A}$ of closed sets and $\mathcal{B}$ for sets with $\mathbb{P}$-null boundary and $\varepsilon > 0$:

1. $\left\{\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A}) \middle| \left|\int_\Omega \varphi \,\mathrm{d}\,\mathbb{Q} - \int_\Omega \varphi \,\mathrm{d}\,\mathbb{P}\right| < \varepsilon,\ \forall \varphi \in \Phi\right\}$
2. $\left\{\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A}) \middle| \mathbb{Q}[G] > \mathbb{P}[G] - \varepsilon,\ \forall G \in \mathcal{G}\right\}$
3. $\left\{\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A}) \middle| \mathbb{Q}[A] < \mathbb{P}[A] + \varepsilon,\ \forall A \in \mathcal{A}\right\}$
4. $\left\{\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A}) \middle| \left|\mathbb{Q}[B] - \mathbb{P}[B]\right| < \varepsilon,\ \forall B \in \mathcal{B}\right\}$

Having recapped useful characterizations of the weak topology, we are able to relate it to our f-$\mathcal{E}$-metrics.

Theorem 2.1.21
For all Ω, it holds:

$$\forall \delta > 0 : f(\delta) > 0 \iff \mathrm{d}_f \textit{ metrizises the weak convergence on } \mathcal{M}_1(\Omega, \mathfrak{A}).$$

Note this is a powerful result, as solely characteristics of f influence the convergence, independent of the underlying metric space $(\Omega, \mathrm{d}_\Omega)$.

For the proof, we follow (Huber, 1981, Theorem 2.14) closely:

Proof. "$\Rightarrow$" We start by showing $f(\delta) > 0$ for all $\delta > 0$ is sufficient:
As f is càdlàg with $f(0) = 0$, we get $\forall \varepsilon > 0\, \exists \delta > 0 : 0 < f(\delta) < \varepsilon$.

- Every weak topology neighborhood contains a f-$\mathcal{E}$-neighborhood: Let $\varepsilon > 0$, $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathfrak{A})$ and a closed $A \in \mathfrak{A}$ be given, we show that $\mathcal{U} := \left\{\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A}) \middle| \mathbb{Q}[A] < \mathbb{P}[A] + \varepsilon,\ \forall A \in \mathcal{A}\right\}$ contains a f-$\mathcal{E}$-neighborhood, i.e. $\mathrm{d}_f(\mathbb{P}, \mathbb{Q}) < \delta \Rightarrow \mathbb{Q} \in \mathcal{U}$ for some $\delta > 0$.
  Choose $0 < \delta < \varepsilon$ such that $\mathbb{P}[A^\delta] < \mathbb{P}[A] + \frac{1}{2}\varepsilon$ and $f(\delta) < \frac{1}{2}\varepsilon$, this is possible as A is closed and f continuous in 0. For $\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A})$ with $\mathrm{d}_f(\mathbb{P}, \mathbb{Q})$:

$$\mathbb{Q}[A] \leq \mathbb{P}[A^\delta] + f(\delta) < \mathbb{P}[A] + \varepsilon$$

and therefore $\mathbb{Q} \in \mathcal{U}$.

- Every f-$\mathcal{E}$-neighborhood contains a weak topology-neighborhood: Let $\delta > 0$ and $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathfrak{A})$ be given, we show there exists $\varepsilon > 0$ such that $\left\{\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A}) \middle\| \mathbb{Q}[B] - \mathbb{P}[B]\right| < \varepsilon, \forall B \in \mathcal{B}\Big\} \subset \left\{\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A}) \middle| \operatorname{d}_f(\mathbb{P}, \mathbb{Q}) < \delta\right\}$ for some $\mathcal{B}$.
  Choose $\varepsilon > 0$ such that $\varepsilon < \frac{1}{2}\delta$ and $\varepsilon < \frac{1}{2}f(\delta)$. As $(\Omega, \operatorname{d}_\Omega)$ is separable, there exists a finite union $\mathcal{B}'$ of $B \in \mathfrak{A}$ with

$$\operatorname{diam}(B) < \varepsilon \quad \text{and} \quad \mathbb{P}\left[\left(\bigcup_{B \in \mathcal{B}'} B\right)^c\right] < \varepsilon$$

  and we can choose the B to be disjoint and have $\mathbb{P}$-null boundary. Let $\mathbb{Q} \in \mathcal{U} := \left\{\mathbb{Q} \in \mathcal{M}_1(\Omega, \mathfrak{A}) \middle\| \mathbb{Q}[B] - \mathbb{P}[B]\right| < \varepsilon \text{ for all } B \in \mathcal{B}\Big\}$ with $\mathcal{B} := \left[\{\bigcup_{B \in \mathcal{I}} B \middle| \mathcal{I} \subset \mathcal{B}'\right\}$ and $A \in \mathfrak{A}$ arbitrary. Define

$$C := \bigcup_{\substack{B \in \mathcal{B} \\ B \cap A \neq \emptyset}} B$$

  as the $\mathcal{B}$-approximation of A. Then

$$A \subset C \cup \left(\bigcup_{B \in \mathcal{B}'} B\right)^c \quad \text{and} \quad C \subset A^\varepsilon$$

  and hence

$$\mathbb{P}[A] < \mathbb{P}[C] + \varepsilon < \mathbb{Q}[C] + \varepsilon + \varepsilon < \mathbb{Q}[A^\varepsilon] + f(\delta) \leq \mathbb{Q}[A^\delta] + f(\delta).$$

  This shows $d_f(\mathbb{Q}, \mathbb{P}) < \delta$ and concludes the sufficiency.

"$\Leftarrow$" Assume there exists a $\theta > 0$ with $f(\theta) = 0$ and consider the probability measures $\mathbb{P}[\{0\}] = 1$ and $\mathbb{P}_n[\{0\}] = 1 - \frac{1}{n}$, $\mathbb{P}_n[\{\theta\}] = \frac{1}{n}$. Clearly,

$$\mathbb{P}[\{0\}] > \mathbb{P}_n\left[\{0\}^\delta\right] + f(\delta) \text{ for all } \delta < \theta$$

and therefore $\operatorname{d}_f(\mathbb{P}, \mathbb{P}_n) \geq \theta$, but $\mathbb{P}_n \xrightarrow[n\to\infty]{w} \mathbb{P}$ holds.

□

Remark 2.1.22
Theorem 2.1.21 implies the topological equivalence of all f-$\mathcal{E}$-metrics with $f(\delta) > 0$ for all $\delta > 0$. The most prominent representative of this class is the Prokhorov metric.

In the same spirit, all f-$\mathcal{E}$-metrics with $f(\delta) = 0$ for some $\delta > 0$ induce the same topology. The most prominent representative of this class is the Wasserstein-∞ metric.

Lemma 2.1.23
Let f and g define f-$\mathcal{E}$-metrics with $f(\delta_f) = 0$ for some $\delta_f > 0$ and $g(\delta_g) = 0$ for some $\delta_g > 0$ and $\mathbb{P}_n$, $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathfrak{A})$, then

$$\mathrm{d}_f(\mathbb{P}_n, \mathbb{P}) \xrightarrow[n\to\infty]{} 0 \quad \Longleftrightarrow \quad \mathrm{d}_g(\mathbb{P}_n, \mathbb{P}) \xrightarrow[n\to\infty]{} 0.$$

Proof. Let $\mathbb{P}_n$, $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathfrak{A})$ with $\mathrm{d}_f(\mathbb{P}_n, \mathbb{P}) \xrightarrow[n\to\infty]{} 0$, i.e.

$$\forall \delta > 0\, \exists N \in \mathbb{N}: \ \mathrm{d}_f(\mathbb{P}_n, \mathbb{P}) < \delta \quad \forall n \geq N,$$

that is

$$\mathbb{P}[A] \leq \mathbb{P}_n[A^\delta] + f(\delta) \quad \forall A \in \mathfrak{A},\, n \geq N$$

and especially for $\delta \leq \min\{\delta_f, \delta_g\}$

$$\mathbb{P}[A] \leq \mathbb{P}_n[A^\delta] + f(\delta) = \mathbb{P}_n[A^\delta] = \mathbb{P}_n[A^\delta] + g(\delta) \quad \forall A \in \mathfrak{A},\, n \geq N$$

and therefore

$$\forall \delta > 0\, \exists N \in \mathbb{N}: \ \mathrm{d}_g(\mathbb{P}_n, \mathbb{P}) < \delta \quad \forall n \geq N.$$

The other direction works the same way with f and g interchanged. □

The class of f-$\mathcal{E}$-metrics therefore only induces two different topologies, with one being stricter than the other:

Lemma 2.1.24
Let f and g define f-$\mathcal{E}$-metrics with $f(\delta_f) > 0$ for all $\delta_f > 0$ and $g(\delta_g) = 0$ for some $\delta_g > 0$ and $\mathbb{P}_n$, $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathfrak{A})$, then

$$\mathrm{d}_g(\mathbb{P}_n, \mathbb{P}) \xrightarrow[n\to\infty]{} 0 \quad \Longrightarrow \quad \mathrm{d}_f(\mathbb{P}_n, \mathbb{P}) \xrightarrow[n\to\infty]{} 0.$$

Proof. Let $\mathbb{P}_n$, $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathfrak{A})$ with $\mathrm{d}_g(\mathbb{P}_n, \mathbb{P}) \xrightarrow[n\to\infty]{} 0$, i.e.

$$\forall \delta > 0 \, \exists N \in \mathbb{N} : \ \mathrm{d}_g(\mathbb{P}_n, \mathbb{P}) < \delta \quad \forall n \geq N,$$

that is

$$\mathbb{P}[A] \leq \mathbb{P}_n[A^\delta] + g(\delta) \quad \forall A \in \mathfrak{A}, \, n \geq N$$

and especially for $\delta \leq \delta_g$

$$\mathbb{P}[A] \leq \mathbb{P}_n[A^\delta] + g(\delta) = \mathbb{P}_n[A^\delta] \leq \mathbb{P}_n[A^\delta] + f(\delta) \quad \forall A \in \mathfrak{A}, \, n \geq N$$

and therefore

$$\forall \delta > 0 \, \exists N \in \mathbb{N} : \ \mathrm{d}_f(\mathbb{P}_n, \mathbb{P}) < \delta \quad \forall n \geq N.$$

□

2.2 Relationship to Metrics for Random Variables

As there is an inherent connection between probability measures and random variables, there is also a close relationship between respective metrics.
Let $(\Omega', \mathfrak{A}', \mathbb{W})$ be a rich enough probability space (see (Rachev et al., 2013, Theorem 2.7.1)) and $X, Y : (\Omega', \mathfrak{A}') \to (\Omega, \mathfrak{A})$ random variables, i.e. measurable function, and let $L^0(\Omega, \mathfrak{A})$ be the set of all random variables. We denote by $\mathbb{M}_1$ the pushforward measure of X, i.e. $\mathbb{M}_1(A) \coloneqq \mathbb{W} \circ X^{-1}(A)$ for all $A \in \mathfrak{A}$, and by $\mathbb{M}_2$ the pushforward measure of Y.

Definition 2.2.1
Let X, $Y \in L^0(\Omega, \mathfrak{A})$ and $f : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ be monotonically increasing, we call

$$\mathrm{d}_f^{\mathbb{W}}(X, Y) \coloneqq \inf \left\{ \varepsilon > 0 \middle| \mathbb{W}[\, \mathrm{d}_\Omega(X, Y) > \varepsilon] \leq f(\varepsilon) \right\}$$

the f-$\mathcal{E}$-metric for random variables.

Instead of optimizing the joint measure, we can optimize over random variables as well:

Corollary 2.2.2
Given two probability measures $\mathbb{M}_1$, $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$, *we have*

$$\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) = \inf_{\substack{X \sim \mathbb{M}_1 \\ Y \sim \mathbb{M}_2}} \mathrm{d}_f^{\mathbb{W}}(X, Y)$$

for random variables X, $Y : (\Omega', \mathfrak{A}') \to (\Omega, \mathfrak{A})$.

Proof. We prove this by constructing a 1:1 relationship between a probability measure $\mathbb{M}$ on $(\Omega \times \Omega, \mathfrak{A} \times \mathfrak{A})$ and a pair of random variables X, $Y : (\Omega', \mathfrak{A}') \to (\Omega, \mathfrak{A})$:

- Given jointly X and Y, we define $\mathbb{W}[A \times B] := \mathbb{M}\left[\{\omega \mid X(\omega) \in A, Y(\omega) \in B\}\right]$
- Given $\mathbb{W} \in L^0(\Omega \times \Omega, \mathfrak{A} \times \mathfrak{A})$, we obtain jointly X and Y as in (Dudley, 1968, Theorem 1).

$\square$

We cover this to simplify notation and proofs later on. This is especially relevant for transformations of measures, e.g. shifting and scaling of the support, where $\mathbb{M}_1 + a$ and $\alpha\mathbb{M}_1$ are less intuitive than $X + a$ and αX.

2.3 Probability Measures with Finite Support

As the focus of this thesis lies on the efficient and exact computation of f-$\mathcal{E}$-metrics, we will focus on probability measures with finite support. Let $\mathcal{X} := \operatorname{supp}(\mathbb{M}_1) = \{X_1, \ldots, X_m\}$ with $\mathbb{M}_1[\{X_i\}] = p_i$ for $i = 1, \ldots, m$ and $\mathcal{Y} := \operatorname{supp}(\mathbb{M}_2) = \{Y_1, \ldots, Y_n\}$ with $\mathbb{M}_2[\{Y_j\}] = q_j$ for $j = 1, \ldots, n$ denote the support and probabilities of two finite-support probability measures $\mathbb{M}_1$ and $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$. We will assume $m \geq n$ for the remainder of this paper without loss of generality, as well as $p_i, q_j > 0$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

This setup has direct consequences for the exceedance $\mathcal{E}$. While some of these have previously been mentioned, we repeat them to create a single Proposition collecting all properties relevant to the computation of f-$\mathcal{E}$-metrics.

Proposition 2.3.1
In case of finite-support measures $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$, the function $\mathcal{E} : \mathbb{R}^{\geq} \to [0,1]$ has the following properties:

i) $\mathcal{E}$ *is monotonically decreasing,*

ii) $\mathcal{E}$ *is piecewise constant with at most $m \cdot n$ jumps,*

iii) $\mathcal{E}$ *is càdlàg (continue à droite, limite à gauche), and*

iv) $\mathcal{E}(\delta) = 0$ *for all* $\delta \geq \mathrm{d}_{\max} := \max\limits_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \mathrm{d}_{\Omega}(X_i, Y_j)$.

Lemma 2.3.2
In the finite support setup, $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) < \infty$ *holds even if* $f \equiv 0$.

Proof. As $\mathcal{E}(\delta) = 0$ and $f(\delta) \geq 0$ holds for $\delta \geq \mathrm{d}_{\max}$, we have $\mathcal{E}(\delta) \leq f(\delta)$ and therefore $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) \leq \mathrm{d}_{\max}$. □

So following Corollary 2.1.12, we obtain a relaxed set of requirements in the finite support setting:

Corollary 2.3.3
If $|\Omega| < \infty$ *and* $f : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ *is*

- *monotonically increasing,*
- *càdlàg and*
- *super-additive with*
- $f(0) = 0$,

then $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) := \inf\left\{\delta \geq 0 \mid f(\delta) \geq \mathcal{E}^{\mathbb{M}_1, \mathbb{M}_2}(\delta)\right\}$ *defines a* f-$\mathcal{E}$*-metric on* $\mathcal{M}_1(\Omega, \mathfrak{A})$.

Remark 2.3.4
Illustrative speaking, we need to ensure $f(\delta) \geq \mathcal{E}(\delta)$ for some $\delta \geq 0$. This was previously achieved by requiring $f(\delta_0) > 0$ for some δ_0, as we have $\mathcal{E}(\delta) \xrightarrow{\delta \to \infty} 0$. Given the finite support setting, we obtain $\mathcal{E}(\delta_1) = 0$ for some $\delta_1 \geq 0$, fulfilling the same inequality.

2.4 Computation of f-$\mathcal{E}$-Metrics for Probability Measures with Finite Support

This sets the stage for the main part of this thesis, the efficient computation of f-$\mathcal{E}$-metrics in a finite support setting.
As a reminder, we want to solve

$$\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) := \inf \left\{ \delta \geq 0 \mid f(\delta) \geq \mathcal{E}^{\mathbb{M}_1, \mathbb{M}_2}(\delta) \right\}.$$

To do this in an efficient way, we will split the computation in two disjoint problems:

- So far, we have no way of calculating $\mathcal{E}(\delta)$. As it will be necessary to continually evaluate the exceedance, we want to find an efficient way to do so. We will refer to this as "solving the subproblem for δ"
- As we will see, the calculation of $\mathcal{E}$ will always be computationally expensive. We therefore need to find an efficient search strategy to reduce the number of evaluations.

Let us start with the evaluation of the exceedance $\mathcal{E}$: In our specific setup, as the set $\mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)$ only consists of finite-support measures with support $\mathcal{X} \times \mathcal{Y} = \{X_1, \ldots, X_m\} \times \{Y_1, \ldots, Y_n\}$, one can directly build upon ideas of Schay (1974) and García-Palomares and Giné (1977) and obtain

$$\begin{aligned}
\mathcal{E}(\delta) = \; 1 - \max_{\mu \in \mathbb{R}_{\geq 0}^{m \times n}} \quad & \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} \mathbb{1}_{\mathrm{d}_\Omega(X_i, Y_j) \leq \delta} \\
\text{s.t.} \quad & \sum_{j=1}^{n} \mu_{ij} = p_i \qquad \text{for } i = 1, \ldots, m \qquad \text{(TP)} \\
& \sum_{i=1}^{m} \mu_{ij} = q_j \qquad \text{for } j = 1, \ldots, n.
\end{aligned}$$

Schay (1974) treated this as a standard linear program without further exploring its structure. Shortly after, García-Palomares and Giné (1977) realized the transportation problem structure, allowing for a more efficient treatment. Garel (1981) recognized that due to the special structure of the objective function (TP) can also be interpreted as an ordinary max flow problem (instead of a more general min cost flow): Consider a network with nodes $X_i,\ i = 1, \dots, m$ and $Y_j,\ j = 1, \dots, n$; the X_i are linked with the source s by edges with capacities p_i and the Y_j are linked with the sink t by edges with capacities q_j. Furthermore, there are edges $X_i \to Y_j$ with capacity 1 if and only if $\mathrm{d}_\Omega(X_i, Y_j) \leq \delta$. Consequently, in Garel (1981) and Garel and Massé (2009) it is suggested to solve (TP) with the Ford-Fulkerson algorithm[4]. Taking into account the progress which has been made in recent years for maximum flow algorithms, the Ford-Fulkerson algorithm should of course be replaced by some more efficient variant. However, due to a special problem structure, the above discussion becomes merely irrelevant: It can easily be observed that the max flow problem actually represents a bipartite max flow problem – a fact that has remained unnoticed until today. Due to this observation, it is most beneficial to solve the bipartite max flow problem for example with Karzanov's algorithm, with a complexity analysis carries out by Gusfield et al. (1987).

Theorem 2.4.1
(Gusfield et al., 1987, Theorem 2.2) The bipartite max flow problem with $m \geq n$ nodes can be exactly solved in $\mathcal{O}(mn^2)$.

We can use this to our advantage and obtain a complexity of $\mathcal{C}(\mathcal{E}) = \mathcal{O}(mn^2)$ per evaluation of the exceedance without further assumptions. However, if the support sizes m and n align smoothly, i.e. $m = l \cdot n$ for a $l \in \mathbb{N}$, together with uniform probabilities $p_1 = \cdots = p_m = \frac{1}{m}$ and $q_1 = \cdots = q_n = \frac{1}{n}$, the complexity reduces further.

[4] For practical purposes, they suggested to rely on the Network Simplex algorithm which is rather efficient in practical settings, see e.g. Kovács (2015). From a theoretical point of view, however, the complexity is not competitive, see e.g. Orlin (1997).

Theorem 2.4.2

(Lee and Sidford, 2014, Theorem 7) The bipartite max flow problem with $m \geq n$ nodes can be exactly solved in $\tilde{\mathcal{O}}(m^{1.5} n \log^2(U))$ with capacity ratio U.

We omit the discussion of the capacity ratio here and refer to the detailed work of Lee and Sidford (2013). In the case of uniform probabilities with $m = n$, we obtain $\mathcal{C}(\mathcal{E}) = \mathcal{O}(m^{2.5})$ with $U = 1$.

Since we have established a quite efficient worst case complexity for the subproblem (TP) by exploiting the bipartite structure, we now turn to the question how to find the optimal δ. Not surprisingly, due to finite number of jumps of the exceedance $\mathcal{E}$, it is sufficient to solve a finite number of subproblems to determine $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)$. This was first noted by (Garel and Massé, 2009, p. 78) for the Prokhorov metric, which coincides with the $\delta - \delta$-metric (i.e. $f(\delta) = \delta$) in our setup.

Theorem 2.4.3

Given the set

$$\mathit{\Delta}(\mathbb{M}_1, \mathbb{M}_2) \coloneqq \{\, \mathrm{d}_\Omega(X_i, Y_j) \mid i = 1, \dots, m,\ j = 1, \dots, n\} \cup \{0\}$$

the f-$\mathcal{E}$-metric $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)$ can be computed as

$$\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) = \min\left(\delta_{min}, f^{-1}(\mathcal{E}(\delta_{max}))\right)$$

where

$$\delta_{min} \coloneqq \min\{\delta \in \mathit{\Delta}(\mathbb{M}_1, \mathbb{M}_2) \mid \mathcal{E}(\delta) \leq f(\delta\})$$

$$\delta_{max} \coloneqq \max\{\delta \in \mathit{\Delta}(\mathbb{M}_1, \mathbb{M}_2) \mid \mathcal{E}(\delta) > f(\delta)\},$$

where we set $\delta_{max} \coloneqq 0$ if $\mathcal{E}(\delta) \leq f(\delta)$ for all $\delta \in \mathit{\Delta}(\mathbb{M}_1, \mathbb{M}_2)$.

Remark 2.4.4

While a closed form solution to $f^{-1}(\mathcal{E}(\delta_{max}))$ might not always be available, we will pretend to do so:

- $\delta_{min} \leq f^{-1}(\mathcal{E}(\delta_{max}))$ *is equivalent to* $f(\delta_{min}) \leq \mathcal{E}(\delta_{max})$ *and can easily be checked*
- $f^{-1}(\mathcal{E}(\delta_{max})) > \delta_{min}$ *is therefore the only problematic case. But this can easily be overcome by a simple bisection as* f *is monotonically with* $f(0) \leq \mathcal{E}(\delta_{max})$ *and* $f(\delta_{min}) > \mathcal{E}(\delta_{max})$

As f *might not be continuous,* $f^{-1}(\varepsilon)$ *refers to* $\min\left\{\delta \middle| f(\delta) \geq \varepsilon\right\}$. *Also regarding the complexity, this additional step is negligible as regular evaluations of* f *are necessary throughout our algorithms and independent of* m *and* n*, so especially* $f \in \mathcal{O}(1)$.

Remark 2.4.5
For $\delta_{max} = 0$ *we have* $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) = 0$. *We will therefore assume* $\delta_{max} > 0$ *from now on.*

Thanks to the above result, the computation of $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)$ reduces to finding δ_{min} and δ_{max} in the set of candidates $\Delta(\mathbb{M}_1, \mathbb{M}_2)$. This can be achieved by calculating $\mathcal{E}(\delta)$ for all $\delta \in \Delta(\mathbb{M}_1, \mathbb{M}_2)$, which coincides with a bipartite max flow problem as explained above. For illustrative purposes, let us visualize this method on a small example with $f(\delta) := \delta$, which we will refer to later on for further explanations.

Example 2.4.6
Let $\mathcal{X} := \{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$ *and* $\mathcal{Y} := \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ *with* $p_1 = \cdots = p_6 = \frac{1}{6}$ *and* $q_1 = \cdots = q_4 = \frac{1}{4}$. *This gives us*

$$\Delta(\mathbb{M}_1, \mathbb{M}_2) = \{0, \frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{4}{15}, \frac{5}{15}, \frac{6}{15}, \frac{7}{15}, \frac{9}{15}, \frac{10}{15}, \frac{12}{15}, 1\}$$

and $\mathcal{E}(0) = 1 - \frac{2}{6}$, $\mathcal{E}(\frac{1}{15}) = 1 - \frac{4}{6}$, $\mathcal{E}(\frac{2}{15}) = 1 - \frac{5}{6}$ *and* $\mathcal{E}(\frac{3}{15}) = 0$*, which is visualized in Figure 2.2. Therefore, we obtain* $\delta_{min} = \frac{3}{15}$ *and* $\delta_{max} = \frac{2}{15}$ *and finally* $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) = \min\left(\frac{3}{15}, \mathcal{E}(\frac{2}{15})\right) = \min\left(\frac{3}{15}, 1 - \frac{5}{6}\right) = \frac{1}{6}$.

While this approach always leads to an exact computation of the f-$\mathcal{E}$-metric,

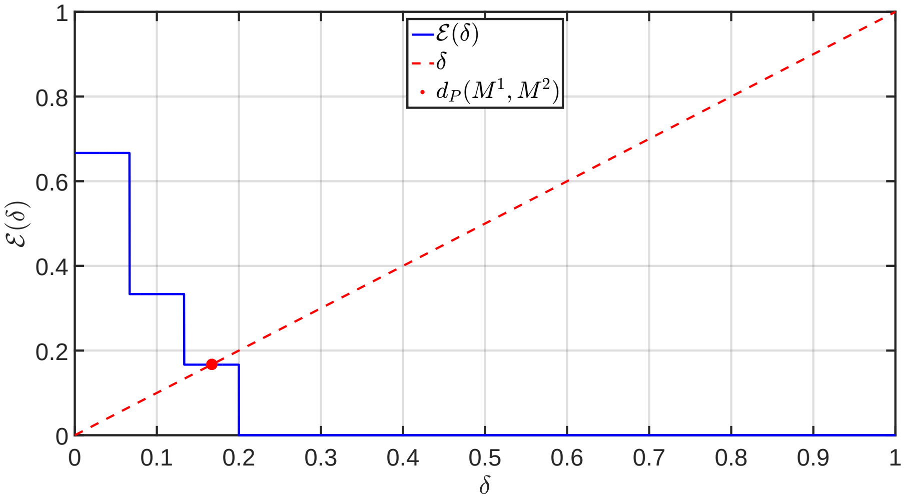


Figure 2.2: Illustration of $\mathcal{E}$ for a small example with $\mathcal{X} = \{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$ and $\mathcal{Y} = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ with $p_1 = \cdots = p_6 = \frac{1}{6}$ and $q_1 = \cdots = q_4 = \frac{1}{4}$.

it is clearly inefficient for large m and n as it requires up to mn solutions of bipartite max flow problems. In total, this results in an aggregated complexity of $\mathcal{O}(m^2 n^3)$.

2.5 Efficient Search Strategies to obtain δ_{max} and δ_{min}

In this section, we introduce two novel and significantly more efficient approaches to obtain δ_{min} and δ_{max}. Both approaches are based on bisection instead of fixpoint iterations. For this purpose, let us use the notation

$$\varDelta(\mathbb{M}_1, \mathbb{M}_2) =: \{\delta_1, \ldots, \delta_L\}$$

with $\delta_1 < \cdots < \delta_L$ sorted and obtain $\big(\mathcal{E}(\delta_k)\big)_k$ to be monotonically decreasing with $k = 1, \ldots, L \leq mn$ as $\mathcal{E}$ is monotonically decreasing by Proposition 2.3.1. Valid starting bounds are ensured as

- $\mathcal{E}(\delta_L) = 0 \leq f(\delta_L)$ and
- $\mathcal{E}(\delta_1) = \mathcal{E}(0) \geq 0 = f(0) = f(\delta_1)$ hold.

2.5.1 Bisecting $\Delta(\mathbb{M}_1, \mathbb{M}_2)$

Starting with $\Delta(\mathbb{M}_1, \mathbb{M}_2)$, we propose to successively half the size of the set of possible candidates via bisection. By the monotonicity of $\left(\mathcal{E}(\delta_k)\right)_{k\in\Delta}$ we can compute $\mathcal{E}(\delta_k)$ for $k \coloneqq \left\lfloor \frac{1}{2}(l+u) \right\rfloor$, starting with $l = 1$ and $u = L$. This reduces the set of candidates to $(\delta_l, \ldots, \delta_k)$ in the case of $\mathcal{E}(\delta_k) \leq \delta_k$ or $(\delta_k, \ldots, \delta_u)$ in the case of $\mathcal{E}(\delta_k) > \delta_k$. This process effectively reduces the number of possible candidates by a factor of two and can be repeated until $\delta_u = \delta_{min}$ and $\delta_l = \delta_{max}$ provide the bounds of the bisection with $u = l + 1$.

Algorithm 1 Δ-Bisection

```
Compute and sort Δ(M_1, M_2) such that δ_1 < ··· < δ_L
Set l := 1,  u := L
while u − l > 1 do
  Set k := ⌊½(l + u)⌋
  Compute E(δ_k)
  if E(δ_k) < f(δ_k) then
    Set u := k
  else if E(δ_k) > f(δ_k) then
    Set l := k
  else
    return  d_f(M_1, M_2) := δ
  end if
end while
return  d_f(M_1, M_2) := min(δ_u, f^{-1}(E(δ_l)))
```

Proposition 2.5.1

Δ-Bisection: strongly polynomial complexity & exactness

For the Δ-Bisection Algorithm 1 holds:

(i) it can be implemented in $\mathcal{O}\left(\log(mn)\max(mn, \mathcal{C}(\mathcal{E}))\right) = \mathcal{O}(mn^2\log(m))$ time, and

(ii) it computes $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)$ *exactly.*

Proof. (i) Computing and sorting $\delta_1, \ldots, \delta_L$ can be done in $\mathcal{O}\big(mn\log(mn)\big)$. The sorting is followed by at most $\lceil \log_2(L) \rceil$ iterations of the while-loop as $(u-l)$ is halved in each iteration and the while-condition is $(u-l) > 1$.
The complexity $\mathcal{C}(\mathcal{E})$ of the subproblem is $\mathcal{O}(mn^2)$.
(ii) By (i) the algorithm terminates. In each step, $\mathcal{E}(\delta_l) \geq f(\delta_l)$ and $\mathcal{E}(\delta_u) \leq f(\delta_u)$ is ensured by selecting the lower or upper half of $(\delta_l, \ldots, \delta_k, \ldots, \delta_u)$. Thus we always have $\delta_l \leq \delta_{max}$ and $\delta_u \geq \delta_{min}$. At the end of the bisection, $u - l \leq 1$ guarantees $\delta_u = \delta_{min}$ and $\delta_l = \delta_{max}$, which is sufficient by Theorem 2.4.3. □

2.5.2 Bisecting $[0, D]$

It is also possible to execute the bisection over the interval $[0, D]$ for a fitting D and obtain an exact value for $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)$. As we need $[0, D]$ to contain δ_{min} and δ_{max} for exactness, $f(D) \geq \mathcal{E}(D)$ has to hold. For this approach one has to determine the necessary accuracy

$$acc_\Delta := \min\left\{\left|\delta - \delta'\right| \mid \delta,\ \delta' \in \Delta(\mathbb{M}_1, \mathbb{M}_2);\ \ \delta \neq \delta'\right\},$$

which can be calculated in $\mathcal{O}(mn)$ given $\Delta(\mathbb{M}_1, \mathbb{M}_2)$ is sorted, i.e. in aggregation in $\mathcal{O}(mn\log(m))$ time. The bisection can then be stopped whenever the interval becomes smaller than acc_Δ. Subsequently, one identifies δ_{min} and δ_{max} to obtain $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)$. This results in a weakly polynomial version of the algorithm with $\lceil \log_2(\frac{D}{acc_\Delta}) \rceil$ while-iterations.

Proposition 2.5.2
$[0, D]$*-Bisection: weakly polynomial complexity & exactness*
For the $[0, D]$*-Bisection Algorithm 2 with* $D \geq \delta_{\max}$ *holds:*

(i) it can be implemented in $\mathcal{O}\Big(\max(mn\log(mn), \log(\frac{D}{acc_\Delta})\mathcal{C}(\mathcal{E}))\Big)$*, and*

(ii) it computes $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)$ *exactly.*

Proof. (i) Computing and sorting $\delta_1, \ldots, \delta_L$ can be done in $\mathcal{O}\big(mn\log(mn)\big)$.

This is followed by at most $\lceil \log_2(\frac{D}{acc_\Delta}) \rceil$ iterations of the while-loop as $(\delta_u - \delta_l)$ is halved in each iteration and the while-condition is $(\delta_u - \delta_l) > acc_\Delta$. After the loop, $\bar{\delta} \in \Delta(\mathbb{M}_1, \mathbb{M}_2) \cap [\delta_l, \delta_u]$ can be found in $\mathcal{O}(mn)$ if it exists.
(ii) By (i) the algorithm terminates. In each step, $\mathcal{E}(\delta_l) \geq f(\delta_l)$ and $\mathcal{E}(\delta_u) \leq f(\delta_u)$ is ensured by selecting the lower or upper half of $[\delta_l, \delta_u]$. After the loop and by the choice of acc_Δ, we have $\left|\Delta(\mathbb{M}_1, \mathbb{M}_2) \cap [\delta_l, \delta_u]\right| \leq 1$ which guarantees a unique $\bar{\delta}$ if the set is not empty. An empty set implies $\mathcal{E}(\delta_l) = \mathcal{E}(\delta_u)$ which gives us $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) = f^{-1}(\mathcal{E}(\delta_u))$. If this is not the case, δ_{min} has to be determined which is equal to the last iteration of the Δ-Bisection. Note that δ_{max} is not required, as we have $\mathcal{E}(\delta_l) = \mathcal{E}(\delta_{max})$. As before, this is sufficient by Theorem 2.4.3. □

To allow the practical computation of $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)$, we present valid choices of D.

Lemma 2.5.3
The choices

1. $D := \delta_L$ *and*
2. $D := \lceil \frac{1}{\lambda} \rceil \delta_>$ *for all* $\delta_> > 0$ *with* $f(\delta_>) = \lambda > 0$

ensure valid starting bounds for the $[0, D]$*-Bisection.*

Proof.

1. We have $\mathcal{E}(\delta_L) = 0 \leq f(\delta_L)$ and therefore $\delta_{min} \leq \delta_L$.
2. Setting $D := \lceil \frac{1}{\lambda} \rceil \delta_>$ ensures $f(D) \geq 1$ by the super-additivity of f and, as $\mathcal{E}$ is bounded by 1, $f(D) \geq 1 \geq \mathcal{E}(D)$.

□

Corollary 2.5.4
As $D := \lceil \frac{1}{\lambda} \rceil \delta_>$ *is independent of* $\mathbb{M}_1$ *and* $\mathbb{M}_2$ *and only depends on* f*, it can be considered constant for our complexity analysis, the* $[0, D]$*-Bisection can therefore be implemented in* $\mathcal{O}\left(\max(mn \log(mn), \log(\frac{1}{acc_\Delta})\mathcal{C}(\mathcal{E}))\right)$.

However from a computational point of view, the choice of D will certainly impact the runtime. It can be easily seen that the $[0, D]$-Bisection might require more iterations than the Δ-Bisection. This is due to $\frac{\delta_L}{acc_\Delta} \geq \left|\Delta(\mathbb{M}_1, \mathbb{M}_2)\right| - 1 = L-1$, with equality only for the optimal equidistant spacing of the δ_i in $[0, \delta_L]$. This can be overcome by updating the stopping criterion $\delta_u - \delta_l \leq acc_\Delta$ to $\delta_u - \delta_l \leq acc_{\Delta_{l,u}}$ with $acc_{\Delta_{l,u}} := \min\left\{|\delta - \delta'| > 0 \mid \delta,\ \delta' \in \Delta(\mathbb{M}_1, \mathbb{M}_2) \cap [\delta_l, \delta_u]\right\}$ in each step, which still gives a unique δ after the loop as before. The cost of updating $acc_{\Delta_{l,u}}$ is $\mathcal{O}(mn)$ and does therefore not change the complexity per iteration which is dominated by $\mathcal{C}(\mathcal{E})$.

Remark 2.5.5
The $[0, D]$-Bisection can also be modified to obtain an approximation of the f-$\mathcal{E}$-metric, say up to acc precision. Changing the stopping criterion to $\delta_u - \delta_l < acc$ gives $\left|\min(\delta_u, f^{-1}(\mathcal{E}(\delta_l))) - \mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2)\right| < acc$. As the defining property of the f-$\mathcal{E}$-metric stems solely from the x-axis, this is independent of f and f^{-1}.

Summarising, we obtain the following complexities for the computation of the f-$\mathcal{E}$-metric in the general case:

Straightforward	Δ-Bisection	$[0, D]$-Bisection (exact) $[0, D]$-Bisection (exact up to acc)
$\mathcal{O}(mn \cdot \mathcal{C}(\mathcal{E}))$	$\mathcal{O}\left(\log(mn) \max(mn, \mathcal{C}(\mathcal{E}))\right)$	$\mathcal{O}\left(\max(mn\log(mn), \log(\frac{1}{acc_\Delta})\mathcal{C}(\mathcal{E}))\right)$ $\mathcal{O}\left(\max(mn\log(mn), \log(\frac{1}{acc})\mathcal{C}(\mathcal{E}))\right)$

Table 2.1: Complexities of the computation of the f-$\mathcal{E}$-metric for general metrics

Algorithm 2 $[0, D]$-Bisection

Compute $\mathit{\Delta}(\mathbb{M}_1, \mathbb{M}_2)$
Set $\delta_l := 0, \delta_u := D$
while $\delta_u - \delta_l > acc_{\mathit{\Delta}}$ **do**
  Set $\delta := \frac{1}{2}(\delta_l + \delta_u)$
  Compute $\mathcal{E}(\delta)$
  if $\mathcal{E}(\delta) < f(\delta)$ **then**
    Set $\delta_u := \delta$
  else if $\mathcal{E}(\delta) > f(\delta)$ **then**
    Set $\delta_l := \delta$
  else
    return $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) := \delta$
  end if
end while
if $\mathit{\Delta}(\mathbb{M}_1, \mathbb{M}_2) \cap [\delta_l, \delta_u] = \emptyset$ **then**
  return $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) := f^{-1}(\mathcal{E}(\delta_l))$
else
  $\mathit{\Delta}(\mathbb{M}_1, \mathbb{M}_2) \cap [\delta_l, \delta_u] = \{\bar{\delta}\}$
  if $\mathcal{E}(\bar{\delta}) < f(\bar{\delta})$ **then**
    $\delta_{min} := \bar{\delta}$
    return $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) := \min\{\delta_{min}, f^{-1}(\mathcal{E}(\delta_l))\}$
  else
    $\delta_{min} := \min\{\delta' \in \mathit{\Delta}(\mathbb{M}_1, \mathbb{M}_2) \mid \delta' \geq \bar{\delta}\}$
    return $\mathrm{d}_f(\mathbb{M}_1, \mathbb{M}_2) := \min\{\delta_{min}, f^{-1}(\mathcal{E}(\delta_l))\}$
  end if
end if

3 Quasi-Convex Metrics

In this chapter, we will discuss settings $[\Omega, \mathrm{d}_\Omega]$ allowing faster computation of $\mathcal{E}$ and therefore of f-$\mathcal{E}$-metrics. The most common space allowing efficient computation will be $\mathbb{R}$ with the classical Euclidean distance $|x-y|$, but we will embed this in a more general context we call quasi-convex metrics. Figuratively, quasi-convex metrics relate an ordering of the points to their distances and require them to "agree" in a certain sense. If the space is accordingly ordered, greedy algorithms can exactly compute the exceedance $\mathcal{E}$, allowing us to drop the bipartite max flow. This is based on Drescher et al. (2018) and Drescher and Werner (2019).

3.1 Treating $\mathcal{E}$ as a Transportation Problem

For the sake of completeness, let us repeat (TP) here

$$
\begin{aligned}
\mathcal{E}(\delta) = \; 1 - \max_{\mu \in \mathbb{R}_{\geq 0}^{m \times n}} \quad & \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} \mathbb{1}_{\mathrm{d}_\Omega(x_i, y_j) \leq \delta} \\
s.t. \quad & \sum_{j=1}^{n} \mu_{ij} = p_i \qquad \text{for } i = 1, \dots, m \\
& \sum_{i=1}^{m} \mu_{ij} = q_j \qquad \text{for } j = 1, \dots, n.
\end{aligned}
\tag{TP}
$$

In Section 2.4, we recommended to treat this as a bipartite max flow problem to obtain a complexity of $\mathcal{C}(\mathcal{E}) = \mathcal{O}(mn^2)$ if no further structure is available, holding for every choice of $[\Omega, \mathrm{d}_\Omega]$. From now on, we will consider (TP) as an arbitrary transportation problem:

$$\begin{aligned}
\mathcal{E}(\delta) = 1 + \min_{\mu \in \mathbb{R}^{m\times n}_{\geq 0}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} \mu_{ij} c^{\delta}_{ij} & & \\
s.t. \quad & \sum_{j=1}^{n} \mu_{ij} = p_i & \text{for } i = 1,\dots,m & \qquad (TP^C)\\
& \sum_{i=1}^{m} \mu_{ij} = q_j & \text{for } j = 1,\dots,n &
\end{aligned}$$

with cost-matrix $C^\delta \coloneqq (c^\delta_{ij})_{ij}$ defined as

$$c^\delta_{ij} \coloneqq \begin{cases} -1 & \text{for } \mathrm{d}_\Omega(X_i, Y_j) \leq \delta, \\ 0 & \text{otherwise.} \end{cases}$$

We introduce a total ordering $\leq$ of Ω, and write $[\Omega, \leq, \mathrm{d}_\Omega]$ for the ordered metric space $[\Omega, \mathrm{d}_\Omega]$. Let us assume $X_1 < \cdots < X_m$ and $Y_1 < \cdots < Y_n$ to be sorted and distinct, i.e. $X_{i_1} \neq X_{i_2}$ for $i_1 \neq i_2$ and $Y_{j_1} \neq Y_{j_2}$ for $j_1 \neq j_2$. This can always be achieved with a complexity of at most $\mathcal{O}(m \log(m))$. In the case of identical points, we add their probabilities to preserve the distribution. We follow the concept of Monge sequences in the sense of Hoffman (1963) and Alon et al. (1989), who also show the optimality of greedy-type methods.

3.2 Monge Sequences

We call a bijective mapping $\pi : I \coloneqq \{(i,j) \mid i = 1,\dots,m \text{ and } j = 1,\dots,n\} \longrightarrow \{1,2,\dots,mn\}$ (a numerical representation of) a sorting of the indices I and write $(i,j) <_\pi (r,s) \iff \pi((i,j)) < \pi((r,s))$. Furthermore we define $(i_k, j_k) \coloneqq \pi^{-1}(k)$ to obtain indices from their rank.

Definition 3.2.1
A sorting π of the indices I is called a Monge sequence *for C^δ, if and only if, for every $1 \leq i,r \leq m, 1 \leq j,s \leq n$ the following condition is satisfied:*

If (i,j) precedes both (i,s) and (r,j) with respect to π, then
$$c^\delta_{ij} + c^\delta_{rs} \leq c^\delta_{is} + c^\delta_{rj}.$$

We are now ready to formulate the greedy-type algorithm, i.e. a north-west-corner-rule after sorting, to obtain a feasible transportation plan.

Algorithm 3 Greedy transportation plan (for given sorting π)

Set $\mu_{i,j} := 0$ for all $i = 1, \ldots, m,\ j = 1, \ldots, n$
for $k = 1, \ldots, mn$ **do**
  $\mu_{i_k j_k} := \min(p_{i_k}, q_{j_k})$
  $p_{i_k} := p_{i_k} - \mu_{i_k, j_k}$
  $q_{j_k} := q_{j_k} - \mu_{i_k, j_k}$
end for
return μ

We refer to its output ${}^{\pi}\mu^{g}$ as the greedy solution with respect to the sorting π. It is well known that the greedy transportation plan algorithm provides a feasible solution for (TP), in fact even a basis solution.

Proposition 3.2.2
At most $m + n - 1$ entries of the greedy transportation plan ${}^{\pi}\mu^{g}$ are non-zero and ${}^{\pi}\mu^{g}$ is a feasible transportation plan.

Proof. In the beginning, the transportation plan μ is initialised as zero. After each positive assignment to μ_{i_k, j_k}, p_{i_k} or q_{j_k} will be reduced to zero. Thus $p_i = q_j = 0$ holds for all $i = 1, \ldots, m,\ j = 1, \ldots, n$ after at most $m+n$ positive assignments.
After $m + n - 1$ positive assignments, at most one p_i or q_j can be non-zero. Assuming $p_i > 0$ or $q_j > 0$ leads to a contradiction to $\sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j$, which holds in the beginning and, as both sides are decreased simultaneously, in every step. Thus at most $m + n - 1$ entries in μ are non-zero.

□

If we could choose the sorting π in such a way that the corresponding cost matrix has the Monge property (after sorting), we would obtain optimality of the greedy transportation plan as shown by Hoffman (1963).

Theorem 3.2.3
If π is an Monge sequence for C^δ, the greedy solution ${}^\pi\mu^g$ is optimal for (TP).

The existence of a Monge sequence for a given instance of (TP) can be checked – along with its construction, if existent – with the algorithm of (Alon et al., 1989, Theorem 5.4), which has complexity $\mathcal{O}(mn^2 \log(m))$. However, solving (TP) directly with a bipartite max flow algorithm has complexity $\mathcal{O}(mn^2)$. Therefore, other and more efficient ways of determining a Monge sequence in our context are asked for. For this purpose, let us subsequently consider quasi-convex metrics, as in this case Monge sequences can be easily obtained in linear time.

3.3 Quasi-Convex Metrics

In the following, we introduce and analyze quasi-convex metrics, as these will then allow for an easy construction of Monge sequences. Although quasi-convex metrics might seem to be a rather special construction, let us emphasize that at least the most important space for applications, i.e. $(\mathbb{R}, \leq, |\cdot|)$, represents a totally ordered space with a quasi-convex metric.

Definition 3.3.1
Motivated by Norfolk (1991), for a totally ordered space $[\Omega, \leq]$ let us call a metric $\mathrm{d}_\Omega(\cdot,\cdot) : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$ quasi-convex*, if and only if for all $x^l \leq x^u$:*

$$\forall x^m \in [x^l, x^u] : \quad \max\{\mathrm{d}_\Omega(x^l, x^m), \mathrm{d}_\Omega(x^m, x^u)\} \leq \mathrm{d}_\Omega(x^l, x^u),$$

where $[x^l, x^u] \coloneqq \{x \in \Omega \mid x^l \leq x \leq x^u\}$ denotes the corresponding order interval.

Instead of referring to midpoints inside the interval, it is also possible to consider extreme points only:

Definition 3.3.2

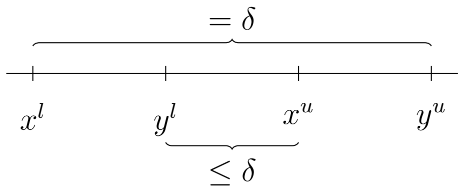
We say a metric $\mathrm{d}_\Omega(\cdot,\cdot) : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$ *has the* crossing property*, if and only if for all* $x^l < x^u$ *and* $y^l < y^u$ *it holds:*

$$\begin{aligned} \mathrm{d}_\Omega(x^l, y^l) &\leq \max\{\, \mathrm{d}_\Omega(x^l, y^u),\, \mathrm{d}_\Omega(x^u, y^l)\}, \quad \textit{and} \\ \mathrm{d}_\Omega(x^u, y^u) &\leq \max\{\, \mathrm{d}_\Omega(x^l, y^u),\, \mathrm{d}_\Omega(x^u, y^l)\}. \end{aligned}$$

While quasi-convexity addresses the inner points of an interval, the crossing property describes the relationship between two different intervals. Intuitively, these two concepts seem to be quite unrelated, however, the following proposition shows that these properties are indeed equivalent.

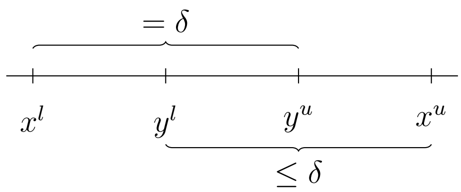
Proposition 3.3.3
A metric in a totally ordered space $[\Omega, \leq]$ *is quasi-convex if and only if it has the crossing property.*

Proof. "⇒" Let $x^l < x^u$, $y^l < y^u$ and w.l.o.g. $\mathrm{d}_\Omega(x^l, y^u) = \delta$ and $\mathrm{d}_\Omega(x^u, y^l) \leq \delta$ for some δ. We have to show $\mathrm{d}_\Omega(x^l, y^l) \leq \delta$ and $\mathrm{d}_\Omega(x^u, y^u) \leq \delta$. Let us present the two major cases exemplarily, the remaining cases easily follow by the same line of arguments:
a) For $x^u, y^l \in [x^l, y^u]$ we have $\mathrm{d}_\Omega(x^u, y^u) \leq \mathrm{d}_\Omega(x^l, y^u) \leq \delta$ and $\mathrm{d}_\Omega(x^l, y^l) \leq \mathrm{d}_\Omega(x^l, y^u) \leq \delta$ by quasi-convexity.



Case a): one interval is contained in the other

b) If this is not the case, we have two overlapping intervals, e.g. for $x^l \leq y^l < y^u \leq x^u$:



Case b): overlapping intervals

Here we have $y^u \in [y^l, x^u]$ and $y^l \in [x^l, y^u]$ and obtain the crossing property again by quasi-convexity.
"$\Leftarrow$" Follows from the case where $x^u = y^l$ for $x^l < y^u$ or $y^u = x^l$ otherwise. $\square$

As we have already indicated above, the most important case of distributions on the real line belongs to the class of quasi-convex metrics, if the usual distance is chosen.

Proposition 3.3.4
The Euclidean metric $\mathrm{d}_\Omega(x, y) = |x - y|$ *on* $(\mathbb{R}, \leq)$ *is quasi-convex.*

Proof. Let $x^l < x^m < x^u$ and $\mathrm{d}_\Omega(x^l, x^u) \leq \delta$ for some $\delta \geq 0$.
This gives $\mathrm{d}_\Omega(x^l, x^m) = x^m - x^l \leq x^u - x^l = \mathrm{d}_\Omega(x^l, x^u) \leq \delta$ and $\mathrm{d}_\Omega(x^m, x^u) = x^u - x^m \leq x^u - x^l = \mathrm{d}_\Omega(x^l, x^u) \leq \delta$. $\square$

Remark 3.3.5
Proposition 3.3.4 can easily be generalized to monotonically increasing transformations of the Euclidean metric.

Let us point out that it is important that the usual metric is chosen on $\mathbb{R}$, as other choices might not lead to a quasi-convex metric:

Example 3.3.6

On $\mathbb{R}$, consider the metric

$$\mathrm{d}_{\mathbb{R}}(x,y) = \begin{cases} 0 & \textit{if } x = y \\ 1 & \textit{if } 0 < |x-y| \leq 2 \\ \frac{1}{2} + \frac{1}{|x-y|} & \textit{if } 2 < |x-y| \end{cases}$$

with $x^l = 0, x^m = 2$ and $x^u = 4$, we see $\frac{3}{4} = \mathrm{d}_{\mathbb{R}}(x_l, x_u) < \mathrm{d}_{\mathbb{R}}(x_l, x_m) = \mathrm{d}_{\mathbb{R}}(x_m, x_u) = 1$, which contradicts quasi-convexity.

To answer the question whether it is possible to extend the above proposition to higher dimension, let us provide the following negative answer. For this purpose, let us give a metric on $\mathbb{R}^2$ which is not quasi-convex:

Example 3.3.7
Consider $\mathbb{R}^2$ with the lexicographic ordering and the Euclidean norm for the points $x^l = (0,0), x^m = (1,2)$ and $x^u = (2,0)$. We have $x^l < x^m < x^u$, but $||x^u - x^l|| = 2 < ||x^m - x^l|| = ||x^u - x^m|| = \sqrt{5}$.

Let us now consider the main benefit when working in a totally ordered space with quasi-convex metric. Indeed, the crossing property has direct consequences for the (negative) cost matrix of (TP):

Corollary 3.3.8
If the metric d_{Ω} has the crossing property, the corresponding cost matrix C^{δ} for (TP) satisfies

$$c^{\delta}_{i^l j^u} = -1 \textit{ and } c^{\delta}_{i^u j^l} = -1 \quad \Longrightarrow \quad c^{\delta}_{i^l j^l} = -1 \textit{ and } c^{\delta}_{i^u j^u} = -1$$

for all $i^l < i^u$ and $j^l < j^u$.

Remark 3.3.9
Considering the network representation of the corresponding max flow problem,

the motivation for the naming of the crossing property becomes transparent: for ordered vertices X_i and Y_j with crossing edges $X_{i^l} - Y_{j^u}$ and $X_{i^u} - Y_{j^l}$, the direct edges $X_{i^l} - Y_{j^l}$ and $X_{i^u} - Y_{j^u}$ have to exist as well. Removing the source and the sink of the network hence gives a strongly ordered graph in the sense of (Spinrad et al., 1986, Definition 3).

3.4 Constructing Monge Sequences in the Case of Quasi-convex Metrics

Thanks to the above preparations, we can now easily construct a Monge sequence π^δ for C^δ. The main idea of the Algorithm 4 is to split the indices I into two parts which we will call $I^\delta_{-1} \coloneqq \{(i,j) \mid \mathrm{d}_\Omega(X_i, Y_j) \leq \delta\}$ and $I^\delta_0 \coloneqq \{(i,j) \mid \mathrm{d}_\Omega(X_i, Y_j) > \delta\}$ depending on the associated costs. These are then sorted with respect to the rows of C^δ. [1]

Lemma 3.4.1
If d_Ω is quasi-convex, π^δ obtained from Algorithm 4 is a Monge sequence for C^δ. The complexity of Algorithm 4 is $\mathcal{O}(mn)$.

Proof. Let $1 \leq i, r \leq m, 1 \leq j, s \leq n$ and (i,j) precede both (i,s) and (r,j) in π^δ. We will show $c^\delta_{ij} + c^\delta_{rs} \leq c^\delta_{is} + c^\delta_{rj}$.
In the case of $c^\delta_{ij} = 0$:
By definition (i,j) is in I^δ_0 and so are (i,s) and (r,j) because they are preceeded by (i,j) in π^δ. Thereby we have $c^\delta_{ij} = c^\delta_{is} = c^\delta_{rj} = 0$ and since c^δ_{rs} is either 0 or -1 the assumption holds.
In the case $c^\delta_{ij} = -1$:
If $c^\delta_{is} + c^\delta_{rj} \geq -1$ the assumption follows directly. If $c^\delta_{is} + c^\delta_{rj} = -2$ we have $c^\delta_{rs} = -1$ from Corollary 3.3.8.
The complexity $\mathcal{O}(mn)$ is due to the fact that each tuple (i,j) has to be

[1] I would like to especially thank Jonas Schwinn for his contributions to this part by guiding me through the theory of Monge sequences and how they relate to transportation problems, which he analyzed in his thesis Schwinn (2019).

Algorithm 4 Monge sequence for quasi-convex d_Ω

Set $\pi^\delta_{-1}(i,j) := 0$, $\pi^\delta_0(i,j) := 0$, $\pi^\delta(i,j) := 0$ for all $i = 1,\dots,m, j = 1,\dots,n$

```
Set k := 1, l := 0
for i = 1,...,m do
  for j = 1,...,n do
    if d_Ω(X_i, Y_j) ≤ δ then
      Set π^δ_{-1}(i,j) := k
      Set k := k + 1
    else
      Set π^δ_0(i,j) := l
      Set l := l + 1
    end if
  end for
end for
for i = 1,...,m, j = 1,...,n do
  if π^δ_{-1}(i,j) ≠ 0 then
    Set π^δ(i,j) = π^δ_{-1}(i,j)
  else
    Set π^δ(i,j) = π^δ_0(i,j) + k
  end if
end for
return π^δ
```

considered and that the effort for each tuple is constant.

□

Remark 3.4.2
In direct consequence of Lemma 3.4.1, this reduces the complexity $\mathcal{C}(\mathcal{E})$ from $\mathcal{O}(mn^2)$ to $\mathcal{O}(mn)$ in case of a quasi-convex metric. Table 2.1 remains valid with this reduced complexity:

Straightforward	$\varDelta$-Bisection	[0, D]-Bisection (exact) [0, D]-Bisection (exact up to acc)
$\mathcal{O}(m^2n^2)$	$\mathcal{O}\left(\log(mn)mn\right)$	$\mathcal{O}\left(mn\log(\max(mn, \frac{1}{acc_\varDelta}))\right)$ $\mathcal{O}\left(mn\log(\max(mn, \frac{1}{acc}))\right)$

Table 3.1: Complexities of the computation of the f-$\mathcal{E}$-metric for quasi-convex metrics

As we have seen in Proposition 3.2.2, the greedy solution ${}^{\pi}\mu^g$ only needs at most $m+n-1$ elements of the Monge sequence. If it is possible to directly obtain these elements, then the complexity will be further reduced from $\mathcal{O}(mn)$ as in Lemma 3.4.1 to a linear time complexity of $\mathcal{O}(m)$. For this purpose, let us subsequently introduce the reduced Monge sequence & transportation plan algorithm to avoid the calculation of the full Monge sequence, i.e. to avoid the full sorting. Instead, the reduced Monge sequence algorithm directly executes a greedy algorithm on a reduced sequence to achieve this further reduction in complexity.

The algorithm consists of two major phases: In the first phase, it traverses the sorted sets $\mathcal{X}$ and $\mathcal{Y}$ to assign flow to indices (i, j) satisfying $\mathrm{d}_\Omega(X_i, Y_j) \leq \delta$ to obtain $\mathcal{E}(\delta)$. As they are sorted to a Monge sequence, we obtain an optimal flow. In the optional second phase, the transportation plan is extended to a feasible and optimal solution of the transportation problem.

Note that Algorithm 6 is only necessary if one is interested in an optimal transportation plan μ realizing the optimal value $\mathcal{E}(\delta)$ which is determined in Algorithm 5.

Theorem 3.4.3

If d_Ω is quasi-convex on a totally ordered space, then the reduced Monge algorithm 5 computes the greedy solution with respect to the Monge sequence π^δ for (TP). In this case, the greedy solution is optimal and the complexity of the subproblem $\mathcal{C}(\mathcal{E})$ reduces from $\mathcal{O}(mn^2)$ to $\mathcal{O}(m)$.

Proof. In the first part the algorithm considers all indices (i,j) such that $\mathrm{d}_\Omega(X_i,Y_j) \leq \delta$ in row-wise order with respect to C^δ. Therefore the indices are treated in the order of π^δ_{-1}, the first part of the Monge sequence π^δ, with the exception that indices (r,s) where $p_r = 0$ or $q_s = 0$ are skipped. This reduces the complexity to $\mathcal{O}(m)$ instead of $\mathcal{O}(mn)$. When all indices in π^δ_{-1} have been considered, the algorithm optionally switches to Algorithm 6 and likewise considers indices (i,j) such that $\mathrm{d}_\Omega(X_i,Y_j) > \delta$ in the order of π^δ_0. Since the algorithm imposes the maximal value on μ_{ij} whenever indices (i,j) are considered, it computes the greedy solution with respect to π^δ.

The second claim on the optimality of the greedy transportation plan follows directly from Theorem 3.2.3.

Concerning the statement on the reduced complexity, the argumentation is as follows: In each iteration of Algorithm 5 at least one of the indices i or j will be increased by 1. This can either happen by reducing the corresponding probability to 0 or by pushing the point to the queue. Therefore, we have at most $m+n-1$ iterations of the while-loop. Analogously, in each iteration of the optional Algorithm 6 at least one element in $\mathcal{X}^R$ or $\mathcal{Y}^R$ is removed and thereby Algorithm 6 also has at most $m+n-1$ iterations.

□

Remark 3.4.4
While i or j are increased by 1 in each iteration, it can happen that both i and j are increased simultaneously. This will occur for $p_i = q_j$ and decrease the necessary number of iterations and therefore computation time. For example, taking $m = n$ and $p_i = q_j = \frac{1}{m}$ for all i,j halves the number of necessary iterations.

Naturally, the question arises, whether one can still find a Monge-sequence for the cost-matrix of TP^C in $\mathbb{R}^d$ for $d \geq 2$. However, this is not even possible for the Euclidean norm. Consider the counterexample in Figure 3.1 for $d = 2$ and $m = n = 3$ where an edge (X_i, Y_j) is equivalent to $\mathrm{d}_\Omega(X_i,Y_j) \leq \delta$. The

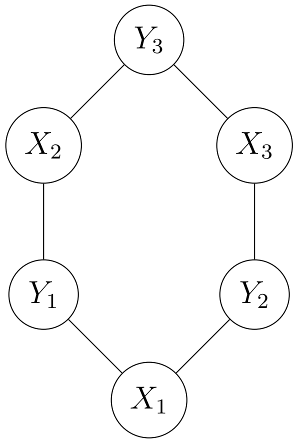


Figure 3.1: Counterexample for the existence of Monge-sequences in the case of $d = 2$.

corresponding cost-matrix is then given by

$$C^\delta = \begin{pmatrix} -1 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & -1 \end{pmatrix}$$

where the non-existence of Monge-sequences can be proved for instance by the algorithm presented in Alon et al. (1989).
Observe though that this does not necessarily mean the greedy-algorithm will not succeed on this instance, but there are choices of the p_i and q_i such that the solution of the greedy-algorithm will be suboptimal. For the exact Δ-Bisection, it has to be noted that now the sorting of the distances dominates the complexity. For the $[0, D]$-Bisection, the distances Δ do not have to be sorted, in fact they can remain unknown. However, while we obtain a strongly polynomial complexity estimate for the Δ-Bisection, we could only obtain a weakly polynomial complexity for the $[0, D]$-Bisection. It has to be noted that in contrast to the previous section, the $[0, D]$-Bisection does not yield the exact solution any more, but is only exact up to an accuracy of $\log(\frac{1}{acc})$.

Summarising, we obtain the following complexities based on the reduced Monge sequence & transportation plan algorithm:

Δ-Bisection (exact)	$[0, D]$-Bisection (exact up to acc)
$\mathcal{O}(mn\log(mn))$	$\mathcal{O}\left(m\max(\log(m), \log(\frac{1}{acc}))\right)$

Table 3.2: Complexities of the computation of the f-$\mathcal{E}$-metric for quasi-convex metrics on $(\Omega, \leq)$

This wraps up the computation of f-$\mathcal{E}$-metrics for probability measures with finite support for general functions f. In the following chapters, we will investigate the two most prominent members of this class, the Prokhorov- and the Wasserstein-∞ metric. We will investigate how we can use specific properties of f to further improve computational efficiency.

Algorithm 5 Reduced Monge sequence & transportation plan

Set $\varepsilon := 1$ and $\mu_{ij} := 0$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$

Set $\mathcal{X}^R := \emptyset$ and $\mathcal{Y}^R := \emptyset$

Set $i := 1$, $j := 1$

while $i \le m$ and $j \le n$ **do**

  if $\mathrm{d}_\Omega(X_i, Y_j) \le \delta$ **then**

    $\mu_{ij} := \min\big(p_i, q_j\big)$

    $p_i := p_i - \mu_{ij}$

    $q_j := q_j - \mu_{ij}$

    $\varepsilon := \varepsilon - \mu_{ij}$

    if $p_i = 0$ **then**

      $i := i + 1$

    end if

    if $q_j = 0$ **then**

      $j := j + 1$

    end if

  else

    if $X_i < Y_j$ **then**

      $\mathcal{X}^R := \mathcal{X}^R \cup \{i\}$

      $i := i + 1$

    else

      $\mathcal{Y}^R := \mathcal{Y}^R \cup \{j\}$

      $j := j + 1$

    end if

  end if

end while

while $i \le m$ **do**

  $\mathcal{X}^R := \mathcal{X}^R \cup \{i\}$

  $i := i + 1$

end while

while $j \le n$ **do**

  $\mathcal{Y}^R := \mathcal{Y}^R \cup \{j\}$

  $j := j + 1$

end while

optional: call transportation plan($\mathcal{X}^R, p, \mathcal{Y}^R, q, \mu$) to obtain feasible μ

return $\mathcal{E}(\delta) := \varepsilon$

Algorithm 6 Transportation plan

while $\mathcal{X}^R \neq \emptyset$ and $\mathcal{Y}^R \neq \emptyset$ **do**
  Denote by i the first node in $\mathcal{X}^R$ and by j the first node in $\mathcal{Y}^R$.
  $\mu_{ij} := \min\left(p_i, q_j\right)$
  $p_i := p_i - \mu_{ij}$
  $q_j := q_j - \mu_{ij}$
  if $p_i = 0$ **then**
    $\mathcal{X}^R := \mathcal{X}^R \setminus \{i\}$
  end if
  if $q_j = 0$ **then**
    $\mathcal{Y}^R := \mathcal{Y}^R \setminus \{j\}$
  end if
end while
return μ

4 Well-known f-$\mathcal{E}$-metrics

While the class of f-$\mathcal{E}$-metrics is very broad, we have seen its topological split in only two distinct subclasses. We will investigate one representative per subclass

- Prokhorov metric – with $f(\delta) = \delta$, it is metrizising the weak topology
- Wasserstein-∞ – with $f(\delta) = 0$, it is inducing the stricter topology

Fixing f allows us to take advantage of its specific form to obtain more efficient, but also more specific, computation algorithms.

4.1 The Prokhorov Metric

This chapter is based on the results of Drescher et al. (2018).

4.1.1 Definition and Elementary Properties

The usual definition of the Prokhorov goes back to Prokhorov (1956):

Definition 4.1.1
For two probability measures $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$*, their Prokhorov distance is defined as*

$$\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) \coloneqq \inf \left\{ \delta > 0 \mid \mathbb{M}_1[A] \le \mathbb{M}_2[A^\delta] + \delta \textit{ for all } A \in \mathfrak{A} \right\}.$$

This fits within our f-$\mathcal{E}$-metrics concept for the simple choice $f(\delta) = \delta$.

At this point, let us introduce the Ky-Fan metric for random variables. As a short reminder of our notation, random variables map from a rich enough probability space $(\Omega', \mathfrak{A}', \mathbb{W})$ into our metric space $(\Omega, \mathfrak{A})$.

Definition 4.1.2
The Ky-Fan metric of two random variables X, $Y : (\Omega', \mathfrak{A}') \to (\Omega, \mathfrak{A})$ *is defined as*

$$\mathrm{d}_{KF}^{\mathbb{W}}(X, Y) \coloneqq \inf \left\{ \varepsilon > 0 \middle| \mathbb{W}[\mathrm{d}_{\Omega}(X, Y) > \varepsilon] \leq \varepsilon \right\},$$

i.e. our f*-*$\mathcal{E}$*-metric for random variables with* $f \equiv 0$.

Following Corollary 2.2.2, we see the relationship between the Prokhorov metric and the Ky-Fan metric.

Lemma 4.1.3
Given two probability measures $\mathbb{M}_1$, $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$*, we have*

$$\mathrm{d}_{P}(\mathbb{M}_1, \mathbb{M}_2) = \inf_{\substack{X \sim \mathbb{M}_1 \\ Y \sim \mathbb{M}_2}} \mathrm{d}_{KF}^{\mathbb{W}}(X, Y)$$

for random variables X, $Y : (\Omega', \mathfrak{A}') \to (\Omega, \mathfrak{A})$.

Let us emphasize the well-known fact that the Prokhorov metric metrizises the weak topology on $\mathcal{M}_1(\Omega, \mathfrak{A})$, see e.g. (Huber, 1981, Theorem 3.8), without any further prerequisites and thus constitutes one of the most important probability metrics.

Lemma 4.1.4
The Prokhorov metric metrizises the weak topology on $\mathcal{M}_1(\Omega, \mathfrak{A})$ *for metric spaces* $(\Omega, \mathrm{d}_{\Omega})$.

Proof. We obviously have $f(\delta) = \delta > 0$ for all $\delta > 0$, so Theorem 2.1.21 yields the statement. □

As we only cover the computation for probability measures with finite support, we need to approximate general distributions. As e.g. demonstrated in Graf and Luschgy (2009), arbitrary measures $\mathbb{P}$ can be approximated sufficiently well by finite-support measures $\mathbb{P}^m$. This can for instance be achieved via optimal quantizers, quantiles, but also by straightforward random sampling. Luckily, the joint approximation error can be bounded by the sum of the separate approximation errors:

Lemma 4.1.5
Let (M, d) *be a metric space and* $A, \tilde{A}, B, \tilde{B} \in M$. *Then*

$$\left| \mathrm{d}(A,B) - \mathrm{d}(\tilde{A},\tilde{B}) \right| \leq \mathrm{d}(A,\tilde{A}) + \mathrm{d}(B,\tilde{B}).$$

Proof. Without loss of generality assume $\mathrm{d}(A,B) \geq \mathrm{d}(\tilde{A},\tilde{B})$:

$$\mathrm{d}(A,B) - \mathrm{d}(\tilde{A},\tilde{B}) = \mathrm{d}(A,B) - \mathrm{d}(A,\tilde{A}) + \underline{\mathrm{d}(A,\tilde{A}) + \mathrm{d}(B,\tilde{B})} - \mathrm{d}(B,\tilde{B}) - \mathrm{d}(\tilde{A},\tilde{B}),$$

it remains to be shown:

$$\mathrm{d}(A,B) - \mathrm{d}(A,\tilde{A}) - \mathrm{d}(B,\tilde{B}) - \mathrm{d}(\tilde{A},\tilde{B}) \leq 0.$$

This is a direct consequence of the triangle inequality of d

$$\mathrm{d}(A,B) \leq \mathrm{d}(A,\tilde{A}) + \mathrm{d}(\tilde{A},\tilde{B}) + \mathrm{d}(B,\tilde{B}).$$

□

All these approaches lead to weak convergence by increasing m and n to obtain finer discretizations and therefore

$$\left| \mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) - \mathrm{d}_P(\mathbb{M}_1^m, \mathbb{M}_2^n) \right| \xrightarrow[n,m\to\infty]{} 0.$$

A convergence rate of $m^{-\frac{1}{d}}$ in a d-dimensional space Ω for absolutely continuous measures with compact support is shown in (Graf and Luschgy, 2009, Theorem 4.3), statements for further distribution classes can also be found there. (Kersting, 1978, Theorem 1) also discusses convergence rates, under very specific properties of the cumulative distribution function.

4.1.2 The Prokhorov Metric on $\mathbb{R}^d$

While our setup works for general metrics spaces $[\Omega, \mathrm{d}_\Omega]$, the most prevalent case will certainly be $[\mathbb{R}^d, \|\cdot\|]$. We dedicate this subsection to $\mathbb{R}^d$ to cover specific properties of the Prokhorov metric.
To do so, let us fix some notation regarding the transformation of probability measures:

- $\mathbb{P}\alpha$ for $\alpha > 0$ is the scaling of the support, i.e. $\mathbb{P}\alpha[B] := \mathbb{P}[\frac{1}{\alpha}B]$ for all $B \in \mathcal{B}(\mathbb{R}^d)$,
- $\mathbb{P} + a$ for $a \in \mathbb{R}^d$ is the shifting of the support, i.e. $\mathbb{P} + a[B] := \mathbb{P}[B - a]$ for all $B \in \mathcal{B}(\mathbb{R}^d)$ and
- $\mathbb{P}\varphi$ for a mapping $\varphi : \mathbb{R}^d \to \mathbb{R}^k$ is the transformation of the support, i.e. $\mathbb{P}\varphi[B] := \mathbb{P}[\varphi^{-1}(B)]$ for all $B \in \mathcal{B}(\mathbb{R}^k)$

with $\alpha B = \{\alpha b \mid b \in B\}$ etc. This notation is inspired by corresponding random variables as we will see in the following proofs.

These transformations have direct consequences for the Prokhorov metric.

Lemma 4.1.6
Let $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. *Then*

1. $\alpha \,\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) \leq \mathrm{d}_P(\mathbb{M}_1\alpha, \mathbb{M}_2\alpha) \leq \mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$ *for all* $\alpha \in (0, 1)$,
2. $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) \leq \mathrm{d}_P(\mathbb{M}_1\beta, \mathbb{M}_2\beta) \leq \beta \,\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$ *for all* $\beta > 1$,
3. $\mathrm{d}_P(\mathbb{M}_1 + a, \mathbb{M}_2 + a) = \mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$ *for all* $a \in \mathbb{R}^d$ *and*
4. *for linear projections* $\pi : \mathbb{R}^d \to \mathbb{R}^k$, $\mathrm{d}_P(\mathbb{M}_1\pi, \mathbb{M}_2\pi) \leq \mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$.

Proof. We prove all properties for the Ky-Fan metric $\mathrm{d}^{\mathbb{W}}_{KF}(X, Y)$ for two random variables $X \sim \mathbb{M}_1$ and $Y \sim \mathbb{M}_2$, so the Lemma follows by taking the infimum as in Corollary 2.2.2.

1. While there exists a proof by Startek (2010), we provide a significantly simplified proof here:

Let $\alpha \in (0,1)$, then

$$\begin{aligned}
\mathrm{d}_{KF}^{\mathbb{W}}(\alpha X, \alpha Y) &= \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|\alpha X - \alpha Y\| > \varepsilon] \leq \varepsilon\} \\
&\overset{\text{pos. hom.}}{=} \inf\{\varepsilon > 0 \mid \mathbb{W}[\underbrace{\alpha\, \|X - Y\|}_{<\|X-Y\|} > \varepsilon] \leq \varepsilon\} \\
&\leq \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|X - Y\| > \varepsilon] \leq \varepsilon\} \\
&= \mathrm{d}_{KF}^{\mathbb{W}}(X, Y)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{d}_{KF}^{\mathbb{W}}(\alpha X, \alpha Y) &= \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|\alpha X - \alpha Y\| > \varepsilon] \leq \varepsilon\} \\
&\overset{\text{pos. hom.}}{=} \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|X - Y\| > \frac{\varepsilon}{\alpha}] \leq \varepsilon\} \\
&= \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|X - Y\| > \varepsilon] \leq \alpha\varepsilon\} \\
&\geq \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|X - Y\| > \alpha\varepsilon] \leq \alpha\varepsilon\} \\
&= \alpha\, \mathrm{d}_{KF}^{\mathbb{W}}(X, Y).
\end{aligned}$$

2. The argument for $\beta > 1$ follows the idea of $\alpha < 1$.
3. Let $a \in \mathbb{R}^d$, we obviously have $\big\|(X+a) - (Y+a)\big\| = \|X - Y\|$ and therefore $\mathrm{d}_{KF}^{\mathbb{W}}(X+a, Y+a) = \mathrm{d}_{KF}^{\mathbb{W}}(X, Y)$.
4.

$$\begin{aligned}
\mathrm{d}_{KF}^{\mathbb{W}}(\pi X, \pi Y) &= \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|\pi X - \pi Y\|_k > \varepsilon] \leq \varepsilon\} \\
&= \inf\{\varepsilon > 0 \mid \mathbb{W}[\underbrace{\|X - Y\|}_{\leq \|X-Y\|_d} > \varepsilon] \leq \varepsilon\} \\
&\leq \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|X - Y\|_d > \varepsilon] \leq \varepsilon\} \\
&= \mathrm{d}_{KF}^{\mathbb{W}}(X, Y)
\end{aligned}$$

□

Remark 4.1.7

These properties are not unique to $\mathbb{R}^d$ but can easily be transferred to other metric spaces and other metrics of th f-$\mathcal{E}$-class: 1. *and* 2. *hold if d_Ω is positive* *homogeneous and f is linear,* 3. *holds as long as* d_Ω *is translation invariant and* 4. *holds for all contractions* f.

Especially the fourth property is helpful to obtain lower bounds for the Prokhorov metric. In Subsection 4.1.5, we show the existence of an highly efficient algorithm for probability measures on $\mathbb{R}$, allowing us to obtain fast bounds in $\mathbb{R}^d$.

4.1.3 Fixpoint Iteration

While we are the first to provide an exact computation algorithm for f-$\mathcal{E}$-metrics, previous approximation ideas exist. Specifically for the Prokhorov metric, Garel (1981) and Garel and Massé (2009) propose a fixpoint iteration, as $\mathcal{E}(\delta) = \delta$ implies $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) = \delta$. But as we will see in a following example, it is not necessary, nor always successful.

In the example in Figure 4.1 (as discussed in Figure 2.2), it holds that $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) = \frac{1}{6} = \delta = \mathcal{E}(\frac{1}{6})$, which represents a fixpoint of the function $\mathcal{E}$. This observation immediately leads to the idea of using a fixpoint iteration. Consequently, this was already suggested in the numerical study (Garel and Massé, 2009, Section 4). There, they employ the fixpoint strategy $\delta_{k+1} \coloneqq \mathcal{E}(\delta_k)$ to obtain the next iterate for δ. They note that this approach often leads to convergence to $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$. However, two major issues seem to have been unnoticed by Garel and Massé (2009):

- First of all, as the function $\mathcal{E}$ is not continuous, the existence of a fixpoint is not always guaranteed. For example, this would be the case if the red line intersects with a vertical segment of the blue line in Figure 2.2.
- Second, even if a fixpoint exists, the fixpoint iteration does not necessarily converge to a fixpoint.

For example, as can be seen in Figure 4.2, any starting point other than the true Prokhorov distance leads to a failure of the fixpoint iteration. Furthermore, we sketch a strategy for building a counterexample for arbitrary m, n. Take

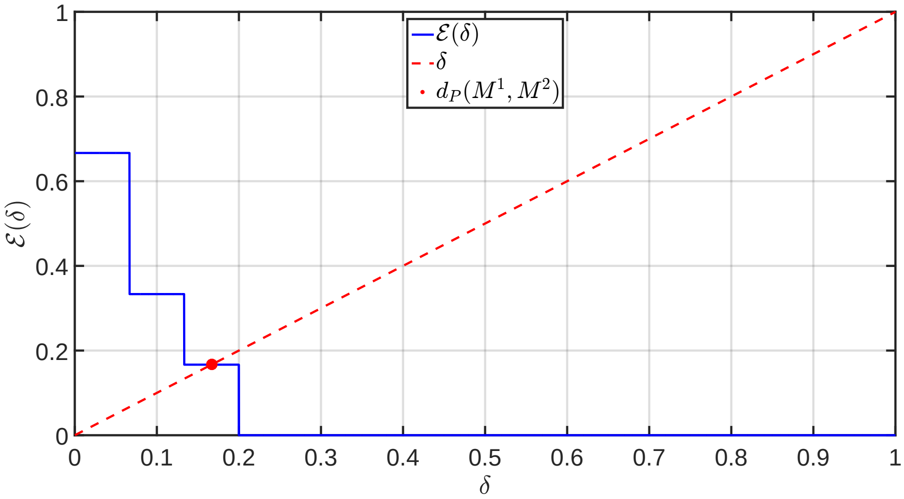


Figure 4.1: Illustration of $\mathcal{E}$ for a small example with $\mathcal{X} = \{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$ and $\mathcal{Y} = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ with $p_1 = \cdots = p_6 = \frac{1}{6}$ and $q_1 = \cdots = q_4 = \frac{1}{4}$.

the X_i i.i.d. with mean 0 and the Y_j i.i.d. with mean 0.5, both with a very small variance. This forces the exceedance function to be flat outside of a small interval around 0.5 and very steep within. Starting the fixpoint iteration outside the interval obviously fails within the first step. Therefore, fixpoint iteration has to be discarded as an reliable approach for the computation of the Prokhorov metric.

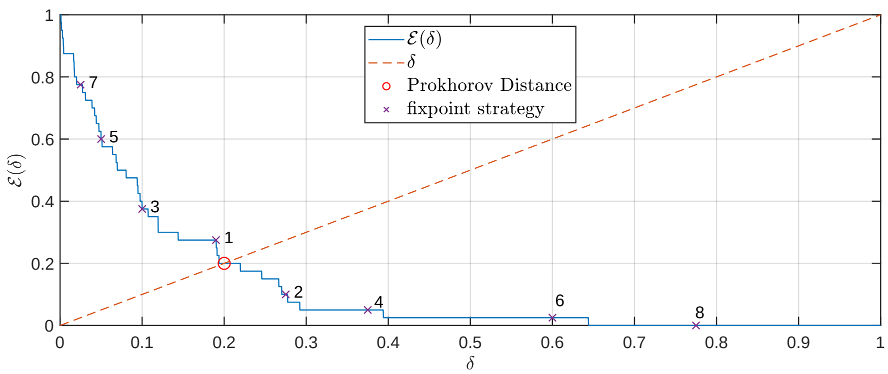


Figure 4.2: Illustration of an example where the fixpoint iteration of $\mathcal{E}$ fails, although a fixpoint exists.

4.1.4 Applying the Δ-Bisection to the Prokhorov Metric

As the Prokhorov metric is a f-$\mathcal{E}$-metric, it can be computed by the bisections we presented in Subsection 2.5.1 and Subsection 2.5.2.

Corollary 4.1.8
The Prokhorov metric for probability measures with finite support can be exactly computed in $\mathcal{O}(mn^2 \log(m))$ with the Δ-Bisection.

This is a significant improvement over the previous best complexity of $\mathcal{O}(m^2n^3)$ due to Garel and Massé (2009) .

Due to the nice behavior of f, it is also simple to obtain a valid upper bound D for the $[0, D]$-Bisection.

Lemma 4.1.9
Letting $D := 1$ is a valid starting point for the $[0, D]$ bisection, i.e. $\delta_{min} \leq 1$ and $\delta_{max} \leq 1$.

Proof. We follow Lemma 2.5.3 and obtain $D = n\frac{1}{n}$ for all $n \in \mathbb{N}$ as $f(\frac{1}{n}) = \frac{1}{n}$ guarantees $f(1) = 1 \geq \mathcal{E}(1)$. □

4.1.5 Computation of d_P for the Euclidean Distance on $\mathbb{R}$ in Quasi-Linear Time

Conducting the bisection over the set of distances $\Delta(\mathbb{M}_1, \mathbb{M}_2)$ as described in Section 2.5.1 requires the computation and sorting of up to mn distances in $\Delta(\mathbb{M}_1, \mathbb{M}_2)$. This still dominates the computation time of the reduced Monge algorithm and results in a strongly polynomial complexity of $\mathcal{O}\left(mn \log(mn)\right)$, which is already a significant improvement over the previous complexity. However, as we can see in Table 3.2, there is a mismatch between the exact Δ-Bisection and the inexact $[0, 1]$-Bisection, which we did not observe in the general setting. Previously, we were able to use the $[0,1]$-Bisection as an exact and an inexact algorithm, solely depending on the selection of acc. This is no longer possible as we need $acc \leq acc_{\Delta}$ for exactness, which we can not determine without knowing $\Delta(\mathbb{M}_1, \mathbb{M}_2)$ – which has at least complexity $\mathcal{O}(mn)$ and thus has larger complexity than the inexact $[0,1]$-Bisection. After Proposition 2.5.2, we showed the sufficiency of an alternative stopping criterion leading to exactness by updating acc_{Δ} to $acc_{\Delta_{l,u}}$, the required accuracy for the remaining candidates. On $(\mathbb{R}, \leq, |\cdot|)$, we can use this to our advantage by starting with the $[0,1]$-Bisection until we are able to efficiently calculate a similar stopping criterion leading to the exact Prokhorov distance. Subsequently we combine these results to obtain a weakly quasi-linear algorithm for the exact computation of the Prokhorov metric on $(\mathbb{R}, \leq, |\cdot|)$.

To avoid the the still dominating sorting of all distances, we propose a different approach consisting of a combination of both previous bisections. Given the minimal distance acc_X of two elements in X, we first bisect the interval $[0,1]$ until the difference between the maximal possible Prokhorov distance δ_u and the minimal possible Prokhorov distance δ_l is smaller than acc_X. At this point it is possible to efficiently compute a set $(\delta_l = \delta_1, \delta_2 \ldots, \delta_{R-1}, \delta_R = \delta_u)$ of remaining candidates for δ_{min} and δ_{max} with $R \leq 2n+2$. Then, these candidates are found by matching each element Y_j with at most two elements in X, as can be seen in Figure 4.3. For an efficient computation of this list, we heavily exploit properties of $(\mathbb{R}, \leq, |\cdot|)$ beyond those used in Subsection 4.2.5. We then continue with Algorithm 1 for the remaining candidates.

We are now ready to present the final version of the efficient algorithm for $(\mathbb{R}, \leq, |\cdot|)$:

Proposition 4.1.10
The quasi-linear Prokhorov algorithm can be implemented in $\mathcal{O}\left(m \cdot \max(\log(m), \log(\frac{1}{acc_X}))\right)$ *on* $(\mathbb{R}, \leq, |\cdot|)$.

Proof. Initialization: Sorting $\mathcal{X}$ and $\mathcal{Y}$ can be done in $\mathcal{O}\left(m\log(m)\right)$ and

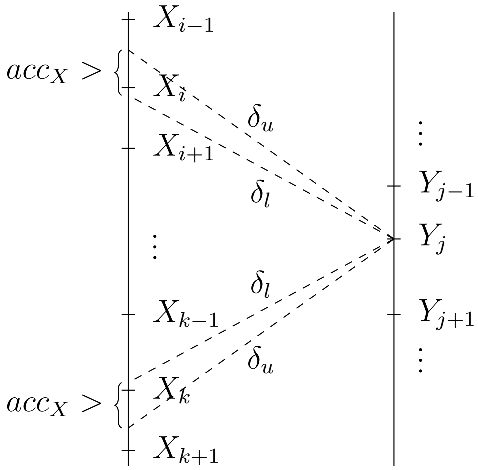


Figure 4.3: By definition of acc_X, for any elements X_s and X_l, $s \neq l$, it holds $\mathrm{d}_\Omega(X_s, X_l) \geq acc_X$. This implies, for $|\delta_u - \delta_l| < acc_X$, we find at most two elements X_i and X_k for each element Y_j such that $X_i \in [Y_j - \delta_u, Y_j - \delta_l]$ and $X_k \in [Y_j + \delta_l, Y_j + \delta_u]$.

$\mathcal{O}\big(n\log(n)\big)$ respectively. Afterwards acc_X can be determined in $\mathcal{O}(m)$ steps by traversing the sorted elements in $\mathcal{X}$, as we have

$$acc_X = \min_{1\leq i\leq m-1} |X_i - X_{i+1}| = \min_{1\leq i<j\leq m} \big|X_i - X_j\big|$$

by the quasi-convexity of $|\cdot|$.

$[0,1]$-*Bisection*: After $\lceil \log_2(\frac{1}{acc_X})\rceil$ loops of the first bisection, we have $|\delta_u - \delta_l| < acc_X$ as the bisection halves $|\delta_u - \delta_l|$ in each step. In each iteration, a subproblem is solved in $\mathcal{O}(m)$ with the Reduced Monge Algorithm.

Candidate list: Now we can traverse the list $Y_1, \ldots, Y_n$ and find for each Y_j at most two elements X_i and X_k such that $X_i \in [Y_j - \delta_u, Y_j - \delta_l]$ and $X_k \in [Y_j + \delta_l, Y_j + \delta_u]$ as in Figure 4.3 and compute their distances $\mathrm{d}_\Omega(X_i, Y_j)$ and $\mathrm{d}_\Omega(X_k, Y_j)$. Since $\mathcal{X}$ is sorted and $|\cdot|$ is quasi-convex, this can be done via bisection in $\mathcal{O}\big(n\log(m)\big)$ as it is done illustratively in Algorithm 8 for $[Y_j - \delta_u,\ Y_j - \delta_l]$.

We can not have another X_h in $[Y_j + \delta_l,\ Y_j + \delta_u]$ as then w.l.o.g. $Y_j < X_k < X_h$ gives $|X_h - X_k| = \big|X_h - Y_j\big| - \big|X_k - Y_j\big| \leq \delta_u - \delta_l < acc_X$, which contradicts the definition of acc_X. The same argument holds for $[Y_j - \delta_u,\ Y_j - \delta_l]$. Note

that this argument is unique to $(\mathbb{R}, \leq, |\cdot|)$ and does not hold for general ordered spaces with quasi-convex metrics.
Δ^R-Bisection: The algorithm is completed by a final bisection in at most $\lceil \log_2(R) \rceil \leq \lceil \log_2(2n+2) \rceil$ steps over the set $\delta_1, \ldots, \delta_R$, where again a subproblem is solved in each loop.
Thereby, summing up the three parts of the algorithm and taking $m \geq n$ into account, one obtains an overall complexity of $\mathcal{O}\left(m \max(\log(m), \log(\frac{1}{acc_X}))\right)$ for the quasi-linear Prokhorov computation on $(\mathbb{R}, \leq, |\cdot|)$. $\square$

Note that one likewise obtains the complexity $\mathcal{O}\left(m \max(\log(m), \log(\frac{1}{acc_Y}))\right)$ by exchanging the roles of $\mathcal{X}$ and $\mathcal{Y}$ in the algorithm.

Proposition 4.1.11
The quasi-linear Prokhorov Algorithm computes $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$ *on* $(\mathbb{R}, \leq, |\cdot|)$ *exactly.*

Proof. After the first bisection, the sorted remaining candidates $\delta_1 < \cdots < \delta_R$ satisfy $\mathcal{E}(\delta_1) > \delta_1$ and $\mathcal{E}(\delta_R) \leq \delta_R$ by Proposition 2.5.2. Furthermore, we have $[\delta_l, \delta_u] \cap \Delta(\mathbb{M}_1, \mathbb{M}_2) = \{\delta_2, \ldots, \delta_{R-1}\}$: Assume a $\delta' \in [\delta_l, \delta_u] \cap \Delta(\mathbb{M}_1, \mathbb{M}_2)$ is not found between the two bisections. Then there exist i, j such that $\mathrm{d}_\Omega(X_i, Y_j) = \delta'$ and therefore $X_i \in [Y_j - \delta_u, Y_j - \delta_l]$ or $X_i \in [Y_j + \delta_l, Y_j + \delta_u]$ most hold. As we previously have checked all these intervals, δ' was already found.
As in Proposition 2.5.2, we have $\mathcal{E}(\delta_1) \geq \mathcal{E}(\delta_{max})$, but δ_{min} can not be obtained as before. If we have $\delta_{min} \in \Delta^R(\delta_l, \delta_u)$, it will be found by the Δ-Bisection. Assume $\delta_{min} \notin \Delta^R(\delta_l, \delta_u)$, that is $\delta_u = \delta_R < \delta_{min}$. As we already know $\mathcal{E}(\delta_R) \leq \delta_R$, we can conclude $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) \leq \delta_R < \delta_{min}$ and therefore $\mathcal{E}(\delta_{max}) < \delta_{min}$. This shows us that $\delta_{min} \in \Delta^R(\delta_l, \delta_u)$ if and only if δ_{min} influences $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$. We can therefore start Algorithm 1, the Δ-Bisection, with the reduced set $\Delta^R(\delta_l, \delta_u)$. Thus the statement reduces to Proposition 2.5.1. $\square$

Combing the exactness and complexity results, we obtain one of our central results:

Theorem 4.1.12
The quasi-linear Prokhorov Algorithm computes $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$ *on* $(\mathbb{R}, \leq, |\cdot|)$ *exactly in weakly quasi-linear time.*

Proof. The exactness has been shown in Proposition 4.1.11, the weak quasi-linearity in Proposition 4.1.10. □

4.2 The Wasserstein-∞ Metric

This chapter is based on the results of Drescher and Werner (2019).

4.2.1 Introduction

To the best of our knowledge, there is currently no general method available for an exact computation of the Wasserstein-∞ metric; not even for the case of measures with finite support. At the time being, the only two computationally tractable cases known to us are:

- The computation of the Wasserstein-∞ metric on the real line – which leads to a one-dimensional global minimization problem, see Bobkov and Ledoux (2016), Equation (2.3). In case of measures with finite support of size m and n ($m \geq n$), this quite obviously reduces to a $\mathcal{O}(m \log(m))$ algorithm.
- The computation of the Wasserstein-∞ metric in $\mathbb{R}^k$, restricted to the special case of measures with equal mass $1/m$ on each of m distinct points and equal sized supports (i.e. $m = n$), see (Efrat et al., 2001, Table 1). There, exploiting the geometry of $\mathbb{R}^k$ and using an reduction to a matching problem, an $\mathcal{O}(m^{3/2} \log(m))$ algorithm is given for the Wasserstein-∞ metric in $\mathbb{R}^2$ for arbitrary l_p-distance, while for $\mathbb{R}^k$, $k \geq 2$, a complexity of $\mathcal{O}(m^{3/2} \log^{k-1}(m))$ only holds for the l_∞-distance.

In the following, we are providing an $\mathcal{O}(mn^2 \log(m))$ algorithm for the general case of measures with finite support of size m and n, $m \geq n$, in arbitrary metric spaces. We would like to emphasize that the general $\mathcal{O}(mn^2 \log(m))$ algorithm is the first exact algorithm for the Wasserstein-∞ metric in the general setup. This is especially important as this can for example be used to approximate arbitrary measures with compact support by measures of finite support, cf. (Kloeckner, B., 2012, Lemma 3.5), which in turn allows to approximate the Wasserstein-∞ distance for measures with compact support in arbitrary metric spaces. Further, we improve the general algorithm to $\mathcal{O}(m \log(m))$ for totally ordered spaces with a quasi-convex distance. This especially includes the case of the real line and thus matches above mentioned complexity. While not improving over the existing algorithm for $\mathbb{R}$, it still introduces a new feature to the analysis, as it allows to consider different (quasi-convex) metrics than the usual absolute value. Further, it offers a different, yet novel view on the situation in $\mathbb{R}$.

To motivate our approach, note that while the Wasserstein-∞ metric is usually defined in terms of the (monotonous limit of the) Wasserstein-p metric, it is surprisingly closely related to the Prokhorov metric. We will exploit this connection in the following for the analysis of the subsequent algorithms.

4.2.2 The Wasserstein-∞ Metric – Definition and Elementary Properties

We do not follow the usual motivation of the Wasserstein-∞ metric to stay close to the idea of f-$\mathcal{E}$-metrics. A thorough discussion of the usual approach Definition 5.1.6 via the Wasserstein-p metric can be found in Section 5.1. The following theorem therefore provides a different representation of the Wasserstein-∞ metric, which already dates back to Givens and Shortt (1984).

Theorem 4.2.1
(Bobkov and Ledoux, 2016, Theorem 2.8) The Wasserstein-∞ metric d_{W_∞} of $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ can also be defined as

$$\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) = \inf \left\{ \delta > 0 \mid \mathbb{M}_1[A] \leq \mathbb{M}_2[A^\delta] \textit{ for all } A \in \mathfrak{A} \right\}.$$

Remark 4.2.2

Technically, W_∞ is only a metric if it is restricted to the set of measures $\left\{\tilde{\mathbb{M}} \mid \int_\Omega \mathrm{d}_\Omega(x,y)\,\mathrm{d}\tilde{\mathbb{M}}[x] < \infty\right\}$ for some $y \in \Omega$. Allowing all measures, a distance of ∞ might occur. We will implicitly assume a finite distance for all measures when referring to the Wasserstein-∞ metric, i.e. restricting $\mathcal{M}_1(\Omega,\mathfrak{A})$.

This definition lets us put the Wasserstein-∞ metric within our f-$\mathcal{E}$-metric concept.

Corollary 4.2.3

The Wasserstein-∞ metric is a f-$\mathcal{E}$-metric with $f(\delta) = 0$ for all $\delta \geq 0$.

From this representation, some immediate and well-known bounds for the Wasserstein-∞ metric can be obtained rather easily. Let us mention that although the second statement seems to be common knowledge, we could not find a proof for it; therefore we have opted to provide a short proof for this inequality.

Proposition 4.2.4

For all $\mathbb{M}_1, \mathbb{M}_2$ in $\mathcal{M}_1(\Omega,\mathfrak{A})$ we have the following bounds:

i) $\mathrm{d}_{W_\infty}(\mathbb{M}_1,\mathbb{M}_2) \geq \mathrm{d}_P(\mathbb{M}_1,\mathbb{M}_2)$ *for the Prokhorov metric and*

ii) $\mathrm{d}_{W_\infty}(\mathbb{M}_1,\mathbb{M}_2) \geq \mathrm{d}_H(\mathrm{supp}(\mathbb{M}_1),\mathrm{supp}(\mathbb{M}_2))$ *for the Hausdorff distance of the supports.*

Proof. The first statement follows directly from the the previous representation of the Wasserstein-∞ metric in Theorem 4.2.1, when compared to the definition of the Prokhorov metric: instead of requiring $\mathbb{M}_1[A] \leq \mathbb{M}_2[A^\delta]$ for the Wasserstein-∞ metric, the Prokhorov metric only requires $\mathbb{M}_1[A] \leq \mathbb{M}_2[A^\delta] + \delta$ for all $A \in \mathfrak{A}$, which obviously represents a relaxation. For the second statement, let $\mathcal{X} := \mathrm{supp}(\mathbb{M}_1)$ and $\mathcal{Y} := \mathrm{supp}(\mathbb{M}_2)$. By definition, $d_H(\mathcal{X},\mathcal{Y}) = \max(e(\mathcal{X},\mathcal{Y}), e(\mathcal{Y},\mathcal{X}))$ for $e(\mathcal{X},\mathcal{Y}) = \sup_{x\in\mathcal{X}} \mathrm{d}_\Omega(x,\mathcal{Y})$ with $\mathrm{d}_\Omega(x,\mathcal{Y}) := \inf_{y\in\mathcal{Y}} d(x,y)$. W.l.o.g. we assume $e(\mathcal{X},\mathcal{Y}) \geq e(\mathcal{Y},\mathcal{X})$. Fix $\varepsilon > 0$ arbitrary and choose some $x^* \in \mathcal{X}$ such that $\mathrm{d}_\Omega(x^*, \mathcal{Y}) \geq \sup_{x\in\mathcal{X}} \mathrm{d}_\Omega(x, \mathcal{Y}) - \varepsilon/3$. As $x^* \in \mathcal{X} = \mathrm{supp}(\mathbb{M}_1)$, $\mathbb{M}_1\left[\{x^*\}^{\varepsilon/3}\right] > 0$ has to hold. We therefore have that for $\delta := \mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) + \varepsilon/3$, it holds by the representation of the Wasserstein-∞ metric in Theorem 4.2.1 that $\mathbb{M}_2\left[[\{x^*\}^{\varepsilon/3}]^\delta\right] > 0$. Since $\mathcal{Y} = \mathrm{supp}(\mathbb{M}_2)$, this implies that $[\{x^*\}^{\varepsilon/3}]^\delta \cap \mathcal{Y} \neq \emptyset$. As $[\{x^*\}^{\varepsilon/3}]^\delta \subset \{x^*\}^{\varepsilon/3+\delta}$, there is a $y^* \in \mathcal{Y}$ with $\mathrm{d}_\Omega(x^*, y^*) \leq \varepsilon/3 + \delta = \mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) + 2\varepsilon/3$. Thus, $\mathrm{d}_\Omega(x^*, \mathcal{Y}) \leq \mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) + 2\varepsilon/3$, which, by the choice of x^* with $d_H(\mathcal{X}, \mathcal{Y}) = e(\mathcal{X}, \mathcal{Y}) = \sup_{x\in\mathcal{X}} \mathrm{d}_\Omega(x, \mathcal{Y}) \leq \mathrm{d}_\Omega(x^*, \mathcal{Y}) + \varepsilon/3$ leads to $d_H(\mathcal{X}, \mathcal{Y}) \leq \mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) + \varepsilon$. As $\varepsilon > 0$ was arbitrary, the third statement follows.

□

Proposition 4.2.5
In case of finite-support measures $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$, we have the following bounds[1]:

i) $\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) \geq \max_{i=1,\dots,m} \min_{j=1,\dots,n} \mathrm{d}_\Omega(X_i, Y_j)$

ii) $\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) \geq \max_{j=1,\dots,n} \min_{i=1,\dots,m} \mathrm{d}_\Omega(X_i, Y_j)$

iii) $\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) \leq \mathrm{d}_{\max}$

Proof. The first two properties follow from Proposition 4.2.4 ii) and for the third, consider $\mathcal{E}(\mathrm{d}_{\max}) = 0$. □

As for the Prokhorov metric, let us relate the Wasserstein-∞ metric to its counterpart for random variables. As a short reminder of our notation, random variables map from a rich enough probability space $(\Omega', \mathfrak{A}', \mathbb{W})$ into our metric space $(\Omega, \mathfrak{A})$. Let us introduce the L^∞-metric for random variables, inspired by the common L^∞-norm.

Definition 4.2.6

[1]Let us remark that these bounds can be exploited to reduce run time (but not complexity) of the general algorithm. However, as they cannot be used efficiently for the specific algorithm, we have opted to not mention these bounds further in the exposition.

We define the L^∞-metric of two random variables X, $Y : (\Omega', \mathfrak{A}') \to (\Omega, \mathfrak{A})$ as

$$\mathrm{d}_\Omega^\infty(X, Y) := \inf \left\{ \varepsilon > 0 \middle| \mathbb{W}[\,\mathrm{d}_\Omega(X, Y) > \varepsilon] \leq 0 \right\},$$

i.e. our f-$\mathcal{E}$-metric for random variables with $f \equiv 0$.

Remark 4.2.7
While this way of defining the L^∞-metric is motivated by our concept of f-$\mathcal{E}$-metrics, an equivalent but more accessible definition is available:

$$\mathrm{d}_\Omega^\infty(X, Y) = \operatorname{ess\,sup}_{\mathbb{W}} \mathrm{d}_\Omega(X, Y).$$

The metric properties are now obvious consequences of the metric properties of d_Ω.

Remark 4.2.8
This is motivated by the L^∞-norm for random variables on normed spaces, i.e. if $\mathrm{d}_\Omega(x, y) := \|x - y\|$ for a norm $\|\cdot\|$ on Ω, we have

$$\|X - Y\|_\infty = \operatorname{ess\,sup}_{\mathbb{W}} \|X - Y\| = \operatorname{ess\,sup}_{\mathbb{W}} \mathrm{d}_\Omega(X, Y) = \inf \left\{ \varepsilon > 0 \middle| \mathbb{W}[\,\mathrm{d}_\Omega(X, Y) > \varepsilon] \leq 0 \right\}.$$

Following Corollary 2.2.2, we see the relationship between the Wasserstein-∞ metric and the L^∞-metric.

Lemma 4.2.9
Given two probability measures $\mathbb{M}_1$, $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$, we have

$$\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) = \inf_{\substack{X \sim \mathbb{M}_1 \\ Y \sim \mathbb{M}_2}} \mathrm{d}_\Omega^\infty(X, Y)$$

for random variables X, $Y : (\Omega', \mathfrak{A}') \to (\Omega, \mathfrak{A})$.

4.2.3 The Wasserstein-∞ Metric in $\mathbb{R}^d$

This part follows Subsection 4.1.2 very closely to provide similar relationships for the Wasserstein-∞ metric on $\mathbb{R}^d$. We keep the notation from before, $\mathbb{P}\alpha$ is the scaled probability measure, $\mathbb{P}+a$ the shifted and $\mathbb{P}\varphi$ the transformed one. These transformations have direct consequences for the Wasserstein-∞ metric.

Lemma 4.2.10
Let $\mathbb{M}_1,\ \mathbb{M}_2 \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, *then*

1. $\mathrm{d}_{W_\infty}(\mathbb{M}_1\alpha, \mathbb{M}_2\alpha) = \alpha\, \mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2)$ *for all* $\alpha > 0$,
2. $\mathrm{d}_{W_\infty}(\mathbb{M}_1 + a, \mathbb{M}_2 + a) = \mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2)$ *for all* $a \in \mathbb{R}^d$ *and*
3. *for linear projections* $\pi : \mathbb{R}^d \to \mathbb{R}^k$, $\mathrm{d}_{W_\infty}(\mathbb{M}_1\pi, \mathbb{M}_2\pi) \leq \mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2)$.

Proof. We prove all properties for the L^∞-metric $\mathrm{d}_\Omega^\infty(X, Y)$ for two random variables $X \sim \mathbb{M}_1$ and $Y \sim \mathbb{M}_2$, so the lemma follows by taking the infimum as in Corollary 2.2.2.

1. Let $\alpha > 0$, then

$$\begin{aligned}
\mathrm{d}_\Omega^\infty(\alpha X, \alpha Y) &= \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|\alpha X - \alpha Y\| > \varepsilon] \leq 0\} \\
&\overset{\text{pos. hom.}}{=} \inf\{\varepsilon > 0 \mid \mathbb{W}[\alpha\ \|X - Y\| > \varepsilon] \leq 0\} \\
&= \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|X - Y\| > \frac{\varepsilon}{\alpha}] \leq 0\} \\
&= \alpha\, \mathrm{d}_\Omega^\infty(X, Y)
\end{aligned}$$

2. Let $a \in \mathbb{R}^d$, we obviously have $\big\|(X+a) - (Y+a)\big\| = \|X - Y\|$ and therefore $\mathrm{d}_\Omega^\infty(X + a, Y + a) = \mathrm{d}_\Omega^\infty(X, Y)$.

3.

$$\begin{aligned}
\mathrm{d}_\Omega^\infty(\pi X, \pi Y) &= \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|\pi X - \pi Y\|_k > \varepsilon] \leq 0\} \\
&= \inf\{\varepsilon > 0 \mid \mathbb{W}[\underbrace{\|X - Y\|}_{\leq \|X-Y\|_d} > \varepsilon] \leq \varepsilon\} \\
&\leq \inf\{\varepsilon > 0 \mid \mathbb{W}[\,\|X - Y\|_d > \varepsilon] \leq 0\} \\
&= \mathrm{d}_\Omega^\infty(X, Y)
\end{aligned}$$

$\square$

Remark 4.2.11
Again, these properties are not unique to $\mathbb{R}^d$ but can easily be transferred to other metric spaces: 1. *holds if d_Ω is positive homogeneous and f is linear,* 2. *holds as long as d_Ω is translation invariant and* 3. *holds for all contractions f.*

Especially the third property is helpful to obtain lower bounds for the Wasserstein-∞ metric. In Subsection 4.2.5, we show the existence of an highly efficient algorithm for probability measures on $\mathbb{R}$, allowing us to obtain fast bounds in $\mathbb{R}^d$.

4.2.4 Applying the $\varDelta$-Bisection to the Wasserstein-∞ Metric

As the Wasserstein-∞ metric is a f-$\mathcal{E}$-metric, it can be computed by the bisections we presented in Subsection 2.5.1 and Subsection 2.5.2.

Corollary 4.2.12
The Wasserstein-∞ metric for probability measures with finite support can be exactly computed in $\mathcal{O}(mn^2 \log(m))$ with the $\varDelta$-Bisection.

We are the first to provide such an algorithm without further assumptions. Previous approaches covered only special cases like $\Omega = \mathbb{R}$ (Bobkov and Ledoux (2016)), or $p_i = q_j$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$((Efrat et al., 2001, Table 1).

As before, we could also bisect the interval $[0, \mathrm{d}_{\max}]$, but the bisection of $\varDelta$ is computationally more efficient.

4.2.5 A Quasi-linear Time Algorithm for the Quasi-convex Setting

While the evaluation of the exceedance in a general setting reduces to a classical bipartite max flow problem, it is possible to exploit the given problem structure for a significantly more efficient approach in important special cases. For this purposes, let us from now on assume that $(\Omega, \leq)$ represents a totally ordered space. Let us further assume $X_1 < \cdots < X_m$ and $Y_1 < \cdots < Y_n$ to be sorted (with respect to the total ordering). Let us further recall a well-known representation of the Wasserstein-p metric on $\mathbb{R}$, see e.g. (Major, 1978, Theorem 8.1):

Theorem 4.2.13

Consider $(\Omega, \mathfrak{A}) = (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ *with the Euclidean distance and let* $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$. *For* $f : \mathbb{R} \to \mathbb{R}$ *convex (e.g.* $f(z) = |z|^p$ *for some* $p \geq 1$*), the following equality holds:*

$$\inf_{\substack{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\Omega, \mathcal{B}(\Omega)) \\ \pi_1(\mathbb{M}) = \mathbb{M}_1 \\ \pi_2(\mathbb{M}) = \mathbb{M}_2}} \int_{\mathbb{R}\times\mathbb{R}} f(x-y) \,\mathrm{d}\,\mathbb{M}(x,y) = \int_0^1 f\left(\mathbb{M}_1^{-1}(u) - \mathbb{M}_2^{-1}(u)\right) \,\mathrm{d}\,u.$$

Here, $\mathbb{M}_1^{-1}(u) := \inf \left\{ x \in \mathbb{R} \mid \mathbb{M}_1 \left[\left\{ z \in \mathbb{R} \mid z \leq x \right\} \right] \geq u \right\}$ *denotes the inverse of the cumulative distribution function.*

This explicitly and uniquely characterizes the optimal joint measure $\mathbb{M}^*$ via the distribution functions as $\mathbb{M}^* \left[(-\infty, x], (-\infty, y] \right] := \min\left(\mathbb{M}_1[(-\infty, x]], \mathbb{M}_2[(-\infty, y]] \right)$. The corresponding copula is also known as the comonotonicity copula. Quite importantly, note that the assumption of convexity can not be loosened to quasi-convexity: For $X_1 = 0$, $X_2 = Y_1 = 1$ and $Y_2 = 2$ with $p_1 = p_2 = q_1 = q_2 = \frac{1}{2}$ and $f(x) = \sqrt{x}$, the right hand side equals 1, while the minimum is $\frac{\sqrt{2}}{2}$.

Using $f(z) = |z|^p$, taking the p-th root on both sides and letting $p \to \infty$ indicates[2] that the analogous statement also holds for $\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2)$. Instead

[2] A rigorous proof actually needs that the objective functions parametrized by p converge

of relying on such a limit argument, we instead resort to the main idea of the proof of (Major, 1978, Theorem 8.1) and transfer it to our setting of totally ordered spaces with a quasi-convex metric to obtain the following even more general result:

Theorem 4.2.14
Let Ω be a totally ordered space with a quasi-convex metric d_{Ω} and let $\mathbb{M}_1$, $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ be measures with finite support. Then it holds that

$$\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) = \sup_{u\in[0,1]} \mathrm{d}_{\Omega}\left(\mathbb{M}_1^{-1}(u), \mathbb{M}_2^{-1}(u)\right),$$

where $\mathbb{M}_1^{-1}(u) := \min\left\{x \in \mathcal{X} \mid \mathbb{M}_1\left[\left\{z \in \mathcal{X} \mid z \leq x\right\}\right] \geq u\right\}$ *denotes (a generalized variant of) the inverse of the cumulative distribution function.*

Theorem 4.2.14 matches (Bobkov and Ledoux, 2016, Equation (2.3)), covering the real line with the Euclidean distance, for measures with finite support in our generalized framework. As mentioned above, for the proof of Theorem 4.2.14 we resort to the proof of (Major, 1978, Theorem 8.1), adjusted to our framework.

Proof. Thanks to Givens and Shortt (1984), Equation (2), we have

$$\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) = \inf_{\substack{\mathbb{M}\in\mathcal{M}_1^{\otimes}(\Omega,\mathcal{B}(\Omega))\\ \pi_1(\mathbb{M})=\mathbb{M}_1\\ \pi_2(\mathbb{M})=\mathbb{M}_2}} \operatorname*{ess\,sup}_{(x,y)\in\mathcal{X}\times\mathcal{Y}}{}_{\mathbb{M}}\ \mathrm{d}_{\Omega}(x, y).$$

As in Theorem 2.1.4, the infimum is attained by some (discrete) joint measure $\mathbb{M}^*$. To shorten notation, let us introduce lowercase letters for the point masses $m^*(x, y) := \mathbb{M}^*[\{(x, y)\}]$ We will argue in the following that we can assume a property of the support of the optimal measure $\mathbb{M}^*$ which is in gist similar to the crossing property for a metric:

($\boldsymbol{\Delta}$) For all $x^l < x^u \in \mathcal{X}$ and for all $y^l < y^u \in \mathcal{Y}$: $\min\left\{m^*(x^l, y^u), m^*(x^u, y^l)\right\} = 0$

in an uniform manner.

For this purpose, assume that there exist $x^l < x^u \in \mathcal{X}$ and $y^l < y^u \in \mathcal{Y}$ such that property ($\boldsymbol{\Delta}$) is violated, i.e. such that

$$m \coloneqq \min\Big\{m^*(x^l, y^u), m^*(x^u, y^l)\Big\} > 0.$$

Then, a related measure $\widetilde{\mathbb{M}^*}$ can be defined as follows, see also Figure 4.4 for an illustration:

$$\begin{aligned}
\widetilde{m^*}(x^l, y^l) &\coloneqq m^*(x^l, y^l) + m,\\
\widetilde{m^*}(x^u, y^u) &\coloneqq m^*(x^u, y^u) + m,\\
\widetilde{m^*}(x^u, y^l) &\coloneqq m^*(x^u, y^l) - m,\\
\widetilde{m^*}(x^l, y^u) &\coloneqq m^*(x^l, y^u) - m,\\
\widetilde{m^*}(x, y) &\coloneqq m^*(x, y) \text{ otherwise.}
\end{aligned}$$

The marginal distributions of $\widetilde{\mathbb{M}^*}$ obviously remain unchanged, hence $\widetilde{\mathbb{M}^*} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)$ holds.

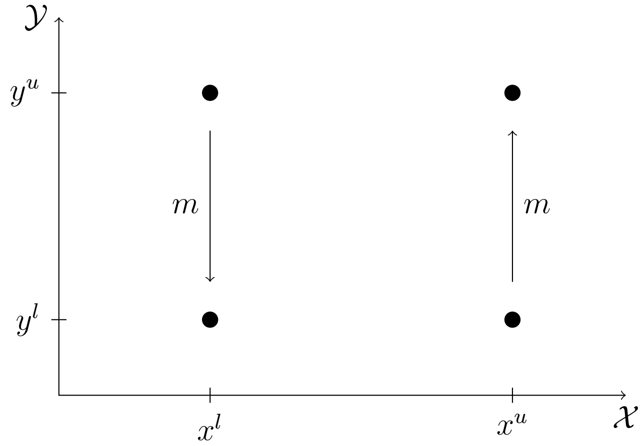


Figure 4.4: Illustration of mass transfer ”towards the diagonal”.

As $\widetilde{\mathbb{M}^*}$ can only differ from $\mathbb{M}^*$ on the set $D \coloneqq \big\{(x^l, y^l), (x^l, y^u), (x^u, y^l), (x^u, y^u)\big\}$, it suffices to show

$$\operatorname*{ess\,sup}_{\rho \in D}{}_{\widetilde{\mathbb{M}^*}}\, \rho \le \operatorname*{ess\,sup}_{\rho \in D}{}_{\mathbb{M}^*}\, \rho.$$

The crossing property of d_Ω yields

$$\max\{\mathrm{d}_\Omega(x^l,y^l),\,\mathrm{d}_\Omega(x^u,y^u)\} \le \max\{\mathrm{d}_\Omega(x^l,y^u),\,\mathrm{d}_\Omega(x^u,y^l)\}.$$

By changing from $\mathbb{M}^*$ to $\widetilde{\mathbb{M}^*}$, mass is always shifted "towards the diagonal", i.e. away from the two points (x^l,y^u) and (x^u,y^l). Obviously, this can only decrease, but never increase the essential supremum over D. In summary,

$$\operatorname*{ess\,sup}_{(x,y)\in\mathcal{X}\times\mathcal{Y}}{}_{\widetilde{\mathbb{M}^*}}\ \mathrm{d}_\Omega(x,y) \le \operatorname*{ess\,sup}_{(x,y)\in\mathcal{X}\times\mathcal{Y}}{}_{\mathbb{M}^*}\ \mathrm{d}_\Omega(x,y),$$

and, by repeating this step at most a finite number of times, the existence of a solution with property ($\boldsymbol{\Delta}$) is obtained as claimed in the beginning.

We further note that property ($\boldsymbol{\Delta}$), together with the marginal distributions, uniquely defines a joint distribution, namely the comonotonic coupling (this can for example easily be seen by a north-west-corner-rule like argument). Since the comonotonic coupling is the only joint measure satisfying ($\boldsymbol{\Delta}$), it must already be optimal.

□

Remark 4.2.15

We have stated Theorem 4.2.14 only for the case of measures with finite support. Of course, the statement can be generalised to arbitrary measures with compact support by approximating these measures with a suitable quantization in the gist of (Kloeckner, B., 2012, Lemma 3.5). The difficulty of approximating $\mathbb{M}_1$ *lies in covering* $\operatorname{supp}(\mathbb{M}_1)$*, not the selection of optimal probabilities. This is closely linked to centroidal Voronoi tessellations. If the measure* $\mathbb{M}_1$ *is unknown and has to be approximated by sampling, further assumptions are necessary, i.e. if* $\operatorname{supp}(\mathbb{M}_1)$ *is not connected, the sample measure will almost surely not converge with respect to the Wasserstein-∞ metric. Under further density assumptions, (Trillos and Slepčev, 2015, Theorem 1.1) and (Liu et al., 2018, Theorem 1.1) provide explicit convergence rates in* m*. As a rigorous proof nevertheless requires some more technical steps and the result is not of use in our discrete setup, we have preferred to focus on the discrete setup presented here.*

Remark 4.2.16

This result is closely linked to (Jylhä, 2015, Theorem 3.15), where a sufficient optimality criterion for distances of the form $\mathrm{d}_{\mathbb{R}}(x, y) = g(x - y)$ *for quasi-convex* $g : \mathbb{R} \to \mathbb{R}$ *for the real line is shown. While the property* $(\boldsymbol{\Delta})$*, called (IM) in Jylhä (2015), is sufficient in totally ordered spaces, it is no more in higher dimensions. However, it can be lifted to (ICM), a more complex variation already discussed in (Champion et al., 2008, Definition 3.1), which is harder to verify.*

Remark 4.2.17

By close inspection of the proof of Theorem 4.2.14, we see that the statement can be generalized to

$$\inf_{\substack{\mathbb{M}\in\mathcal{M}_1^{\otimes}(\Omega,\mathcal{B}(\Omega))\\ \pi_1(\mathbb{M})=\mathbb{M}_1\\ \pi_2(\mathbb{M})=\mathbb{M}_2}} \operatorname*{ess\,sup}_{(x,y)\in\mathcal{X}\times\mathcal{Y}}{}_{\mathbb{M}} G(\mathrm{d}_{\Omega}(x, y)) = \sup_{u\in[0,1]} G\Big(\mathrm{d}_{\Omega}\left(\mathbb{M}_1^{-1}(u), \mathbb{M}_2^{-1}(u)\right)\Big).$$

where $G : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *is some arbitrary monotonically increasing function. We especially note that the loss of the triangle inequality after transformation is of no harm to the efficient computation.*

Corollary 4.2.18

Efficient Wasserstein-∞ computation

In a totally ordered space with quasi-convex metric, $\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2)$ *can be computed exactly in* $\mathcal{O}\left(m \log(m)\right)$ *time for measures with finite support.*

Proof. Thanks to the analytical characterization of the optimal coupling in Theorem 4.2.14, it now only remains to calculate $\max_{u\in[0,1]} \mathrm{d}_{\Omega}\left(\mathbb{M}_1^{-1}(u), \mathbb{M}_2^{-1}(u)\right)$ efficiently. This can for example be achieved by jointly sorting $\mathcal{X} \cup \mathcal{Y}$ to obtain all potential points of discontinuity of $\mathrm{d}_{\Omega}\left(\mathbb{M}_1^{-1}(u), \mathbb{M}_2^{-1}(u)\right)$. The maximum is of course attained in one of these points. As the sorting can be carried out in $\mathcal{O}(m \log(m))$ and the maximum can then be found in $\mathcal{O}(m)$, the quasi-linear complexity follows.

□

While the complexity obtained in Corollary 4.2.18 does not improve over the existing complexity for $(\mathbb{R}, \leq, |\cdot|)$, our approach covers several additional settings which were not covered before.

Example 4.2.19
Consider for example the Huber-like distance

$$\mathrm{d}_{\Omega}(x,y) := \begin{cases} |x-y|^2 & \text{if } 0 \leq |x-y| \leq 1 \\ |x-y| & \text{if } |x-y| > 1, \end{cases}$$

which is differentiable around 0 in contrast to the usual absolute value. This distance is now covered by Theorem 4.2.14 together with the preceding remark.

Example 4.2.20
The most popular application is probably the metric $\mathrm{d}_c(x,y) := \frac{|x-y|}{1+|x-y|}$ *on* $\mathbb{R}$*, as this metric compactifies* $\mathbb{R}$*. We especially note that* d_c *represents again a quasi-convex metric, although the same argumentation as in the previous example would have been sufficient to guarantee fast computations for the Wasserstein-*∞ *metric based on* d_c*.*

4.3 Algorithmic Conclusion

We were able to provide exact algorithms for the computation of the Prokhorov and the Wasserstein-∞ metric, both with a complexity of $\mathcal{O}(mn^2 \log(m))$. To the best of our knowledge, we are the first to provide an exact algorithm for the Wasserstein-∞ for measures with finite support. For the Prokhorov metric, we were able to significantly improve upon the previously known complexity of $\mathcal{O}(m^2 n^3)$.

In the case of a quasi-convex metrics, we generalize the Wasserstein-∞ metric characterization of Jylhä (2015) from $\mathbb{R}$ to arbitrary totally ordered spaces. Our results include an explicit representation of the joint measure and a quasi-linear $\mathcal{O}(m\log(m))$ algorithm. For the Prokhorov metric, we achieved a weakly quasi-linear complexity of $\mathcal{O}\left(m \cdot \max(\log(m), \log(\frac{1}{acc_X}))\right)$.

This sets them computationally on par with the widely used Wasserstein-p metrics.

	d_P	d_{W_∞}	W_p
$(\Omega, \mathrm{d}_\Omega)$	$\mathcal{O}(mn^2\log(m))$	$\mathcal{O}(mn^2\log(m))$	$\mathcal{O}(mn^2\log(m))^3$
$(\mathbb{R}, \lvert\cdot\rvert)$	$\mathcal{O}\left(m \cdot \max(\log(m), \log(\frac{1}{acc_X}))\right)$	$\mathcal{O}(m\log(m))$	$\mathcal{O}(m\log(m))$

Table 4.1: Comparing the complexities of the Prokhorov, Wasserstein-∞ and Wasserstein-p metric

[3]This only holds for $m = k \cdot n$ for a $k \in \mathbb{N}$, see Kleinschmidt and Schannath (1995).

Algorithm 7 quasi-linear computation of the Prokhorov metric on $(\mathbb{R}, \leq, |\cdot|)$

Sort $\mathcal{X}$ and $\mathcal{Y}$ such that $X_1 < \cdots < X_m$ and $Y_1 < \cdots < Y_n$
Compute $acc_X = \min_{i=1,\ldots,m-1} |X_i - X_{i+1}|$
Set $\delta_l := 0, \delta_u := 1$
while $|\delta_u - \delta_l| \geq acc_X$ **do**
  $\delta := \frac{\delta_u + \delta_l}{2}$
  Compute $\mathcal{E}(\delta)$ with Reduced Monge Algorithm
  if $\mathcal{E}(\delta) \leq \delta$ **then**
    $\delta_u := \delta$
  else
    $\delta_l := \delta$
  end if
end while

Bisection of $[0, 1]$

Set $\varDelta^R(\delta_l, \delta_u) := (\varDelta(\mathbb{M}_1, \mathbb{M}_2) \cap [\delta_l, \delta_u]) \cup \{\delta_l\} \cup \{\delta_u\}$
Compute list of candidates $\delta_l = \delta_1, \delta_2 \ldots, \delta_{R-1}, \delta_R = \delta_u$ in $\varDelta^R(\delta_l, \delta_u)$
Sort $\delta_1 < \cdots < \delta_R$
Set $l := 1,\ u := R$
while $u - l > 1$ **do**
  $k := \left\lfloor \frac{1}{2}(l + u) \right\rfloor$
  Compute $\mathcal{E}(\delta_k)$ with Reduced Monge Algorithm
  if $\mathcal{E}(\delta_k) \leq \delta_k$ **then**
    $u := k$
  else
    $l := k$
  end if
end while

Bisection of $\varDelta^R(\delta_l, \delta_u)$

return $\mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2) := \min\left(\delta_u, \mathcal{E}(\delta_l)\right)$

Algorithm 8 Finding the remaining X_h in $[Y_j - \delta_u,\ Y_j - \delta_l]$

```
Set i := 1, k := m
while k − i > 1 do
  h := ⌊½(i + k)⌋
  if X_i > Y_j − δ_l or X_k < Y_j − δ_u then
    return ∅
  else if X_h > Y_j − δ_l then
    Set k := h
  else if X_h < Y_j − δ_u then
    Set i := h
  else
    return {X_h}
  end if
end while
if Y_j − δ_u ≤ X_i ≤ Y_j − δ_l then
  return {X_i}
else if Y_j − δ_u ≤ X_k ≤ Y_j − δ_l then
  return {X_k}
else
  return ∅
end if
```

5 Relationship to other metrics

As there is a multitude of probability metrics available, we dedicate this chapter to the relationships between our f-$\mathcal{E}$ metrics and commonly used metrics. For a short and comprehensive overview of probability metrics and their relations, let us refer to Gibbs and Su (2002), especially Figure 1.

5.1 Wasserstein-p metric

The Wasserstein metric goes back to the Monge–Kantorovich transportation problem, see Kantorovitch (1958), where goods have to be transported in a cost efficient manner from points of supply to match demand. For an overview, let us refer to Villani (2009) and the references therein.

Definition 5.1.1
(Villani, 2009, Definition 6.1) The Wasserstein-p metric W_p *for* $1 \leq p < \infty$ *between two probability measures* $\mathbb{M}_1$, $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ *is defined as*

$$W_p(\mathbb{M}_1, \mathbb{M}_2) \coloneqq \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \left(\int_\Omega \mathrm{d}_\Omega(x,y)^p \, \mathrm{d}\,\mathbb{M}[(x,y)] \right)^{1/p}$$

We call each $\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)$ a *transport plan* or a *coupling* of $\mathbb{M}_1$ and $\mathbb{M}_2$.

Proposition 5.1.2
(Villani, 2009, p. 106) The Wasserstein-p metric defines a metric on $\mathcal{M}_1(\Omega, \mathfrak{A})$ *for each* $1 \leq p < \infty$*.*

Remark 5.1.3

Technically, W_p is only a metric if it is restricted to the set of measures $\left\{\tilde{\mathbb{M}} \mid \int_\Omega \mathrm{d}_\Omega(x,y)\,\mathrm{d}\tilde{\mathbb{M}}[x] < \infty\right\}$ for some $y \in \Omega$. Allowing all measures, a distance of ∞ might occur.

Remark 5.1.4

Let us emphasize the well-known fact that the Wasserstein metric metrizises the weak topology on $\mathcal{M}_1(\Omega, \mathfrak{A})$, if the diameter of Ω is bounded, i.e. $\sup\limits_{x,y\in\Omega} \mathrm{d}_\Omega(x,y)^p < \infty$, see Villani (2009), Theorem 6.9.

Similar to the L^p metrics, the Wasserstein-p metric is monotone with respect to p.

Proposition 5.1.5

(Villani, 2009, Remark 6.6) For $\mathbb{M}_1$, $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ and $1 \leq p \leq q < \infty$, $W_p(\mathbb{M}_1, \mathbb{M}_2) \leq W_q(\mathbb{M}_1, \mathbb{M}_2)$ holds.

As the name Wasserstein-∞ metric suggests, it is usually defined as the limit $p \to \infty$ of the Wasserstein-p metric.

Definition 5.1.6

(Bobkov and Ledoux, 2016, Sec 2.1) The Wasserstein-∞ metric W_∞ of $\mathbb{M}_1$, $\mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ can also be defined as the limit $p \to \infty$ of $W_p(\mathbb{M}_1, \mathbb{M}_2)$:

$$\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) := \lim_{p\to\infty} W_p(\mathbb{M}_1, \mathbb{M}_2) = \sup_{p\geq 1} W_p(\mathbb{M}_1, \mathbb{M}_2)$$

The equivalence of the two provided definitions is shown in Theorem 4.2.1.

Although Definition 5.1.6 represents the most common definition of the Wasserstein-∞ metric, it is not the most convenient representation to obtain an exact algorithm for its computation.

This remains true even in the case of measures with finite support (where the individual Wasserstein-p metric can be computed exactly) due to the fact that still the computation of a limit is required.

Corollary 5.1.7
For all $\mathbb{M}_1, \mathbb{M}_2$ *in* $\mathcal{M}_1(\Omega, \mathfrak{A})$ *we have* $\mathrm{d}_{W_\infty}(\mathbb{M}_1, \mathbb{M}_2) \geq W_p(\mathbb{M}_1, \mathbb{M}_2)$ *for all* $p \geq 1$.

Proof. This is an immediate consequence of the monotonicity of W_p. □

One a more conceptual level, the Wasserstein-p metric takes all occurring distances of Δ into account, where as f-$\mathcal{E}$ metrics only look at one specific distance. This is in-line with the classical jump from L^p metrics to the L^∞ metric and therefore most obvious with the Wasserstein-∞ metric. However, this also allows us to relate the Wasserstein-p to the exceedance $\mathcal{E}$:

Theorem 5.1.8
Let $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\Omega, \mathfrak{A})$ *and* $p \geq 1$, *then*

$$W_p(\mathbb{M}_1, \mathbb{M}_2) = \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \left(\int_0^\infty p \mathcal{E}^{\mathbb{M}}(t) s^{p-1} \,\mathrm{d}\, s \right)^{1/p},$$

where we use the notation $\mathcal{E}^{\mathbb{M}}(\delta) \coloneqq \mathbb{M}\left[\, \mathrm{d}_\Omega(x, y) > \delta \right]$.

Proof. We have

$$\begin{aligned}
W_p(\mathbb{M}_1, \mathbb{M}_2)^p &= \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \int_\Omega \mathrm{d}_\Omega(x, y)^p \,\mathrm{d}\, \mathbb{M}[(x, y)] \\
&= \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \mathbb{E}_{\mathbb{M}} \left[\, \mathrm{d}_\Omega(x, y)^p \right] \\
&= \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \int_0^\infty \left(1 - F_{\mathrm{d}_\Omega(x,y)^p}(t)\right) \,\mathrm{d}\, t \\
&= \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \int_0^\infty \left(1 - F_{\mathrm{d}_\Omega(x,y)}(t^{\frac{1}{p}})\right) \,\mathrm{d}\, t \\
&\overset{t=s^p}{=} \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \int_0^\infty \left(1 - \underbrace{F_{\mathrm{d}_\Omega(x,y)}(s)}_{=\mathbb{M}[\,\mathrm{d}_\Omega(x,y) \leq s]}\right) p\, s^{p-1} \,\mathrm{d}\, s
\end{aligned}$$

$$= \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \int_0^\infty p\mathcal{E}^{\mathbb{M}}(s)^p s^{p-1} \,\mathrm{d}\,(s)$$

$\square$

Most notably, we obtain for $p = 1$:

Corollary 5.1.9
Let $\mathbb{M}_1, \mathbb{M}_2$ in $\mathcal{M}_1(\Omega, \mathfrak{A})$, then

$$W_1(\mathbb{M}_1, \mathbb{M}_2) = \inf_{\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)} \int_0^\infty \mathcal{E}^{\mathbb{M}}(t) \,\mathrm{d}\,(t).$$

Remark 5.1.10
Note the difference between $\mathcal{E}^{\mathbb{M}}(\cdot)$ and $\mathcal{E}(\cdot)$, where the first fixes the joint measure $\mathbb{M}$ and then varies the argument δ in contrast to second, which optimizes the joint measure $\mathbb{M}$ for each δ individually. Therefore, the inequality $\mathcal{E}(\delta) \leq \mathcal{E}^{\mathbb{M}}(\delta)$ holds for all δ and $\mathbb{M} \in \mathcal{M}_1^{\otimes}(\mathbb{M}_1, \mathbb{M}_2)$.

Corollary 5.1.11
Let $\mathbb{M}_1, \mathbb{M}_2$ in $\mathcal{M}_1(\Omega, \mathfrak{A})$ and $p \geq 1$, then

$$W_p(\mathbb{M}_1, \mathbb{M}_2) \geq \left(\int_0^\infty p\,\mathcal{E}(s) s^{p-1} \,\mathrm{d}\,s \right)^{1/p}.$$

This can be used to construct a computational advantage. Given $W_p(\mathbb{M}_1, \mathbb{M}_2)$, an upper bound for the integral, we can invert the idea of upper Darboux integrals to bound $\mathcal{E}(\cdot)$. Assume having calculated $\mathcal{E}(\delta)$ for $\delta \in [0, \delta_{low}] \cup [\delta_{high}, \delta_{max}]$. We can now estimate an upper bound for the integral $\left(\int_{\delta_{low}}^{\delta_{high}} p\,\mathcal{E}(s) s^{p-1} \,\mathrm{d}\,s \right)^{1/p}$ and therefore also limit $\mathcal{E}(\delta)$ for $\delta \in (\delta_{low}, \delta_{high})$

5.2 Lévy Metric

For $\Omega = \mathbb{R}$, the Lévy metric constitutes a simplified version of the Prokhorov metric.

Definition 5.2.1
(Huber, 1981, Def 2.7) Let F, G be to cumulative distribution functions on $\Omega = \mathbb{R}$, then their Lévy metric is defined as

$$\mathrm{d}_L(F,G) := \inf\left\{\varepsilon > 0 \mid G(x-\varepsilon) - \varepsilon \le F(x) \le G(x+\varepsilon) + \varepsilon \textit{ for all } x \in \mathbb{R}\right\}.$$

Lemma 5.2.2
(Huber, 1981, Lemma 2.8) The Lévy metric defines a metric on $\mathcal{M}_1(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

As the Prokhorov metric, the Lévy metric metrizises the weak topology.

Theorem 5.2.3
(Huber, 1981, Theorem 2.9) The Lévy metric metrizises the weak topology.

An immediate consequence is the topological equivalence of both metrics.

Corollary 5.2.4
The Prokhorov metric and the Lévy metric are topologically equivalent.

However, we are unable to bound the Prokhorov metric terms of the Lévy metric and obtain only an one-way inequality.

Lemma 5.2.5
(Huber, 1981, Equation 2.24) For all $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{M}_1(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and their respective cumulative distribution functions $F_{\mathbb{M}_1}$ and $F_{\mathbb{M}_2}$,

$$\mathrm{d}_L(F_{\mathbb{M}_1}, F_{\mathbb{M}_2}) \le \mathrm{d}_P(\mathbb{M}_1, \mathbb{M}_2)$$

holds.

Proof. Clearly, the set of all Borel sets considered by the Prokhorov metric includes alls sets of the form $(-\infty, x)$ for all $x \in \mathbb{R}$ and as $\mathbb{M}_1[(-\infty, x)] = F_{\mathbb{M}_1}(x)$, the lemma follows. $\square$

6 Numerical Analysis

This chapter is dedicated to multiple numerical analyses, including practical running times, verification of worst case complexities and efficiency of lower bounds. All tests were carried out on a standard laptop (processor: Intel Core i7-5600U, 2.60 GHz, RAM: 16GB). The implementations were done in MATLAB. We used the network-simplex of IBMs ILOG CPLEX 12.6.3 for MATLAB R2015b to solve the max flow problems when necessary. We do not present the times of the $[0, 1]$-Bisection, as they are very similar to the $\varDelta$-Bisection.

6.1 Verifying our Theoretical Complexities

In this section, we report a numerical analysis to validate the complexities of Algorithm 1 ($\varDelta$-Bisection) and Algorithm 7 (Quasi-linear Prokhorov) for the euclidean metric on $\Omega = \mathbb{R}$. The input for both algorithms was given in the form of unsorted $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ as well as $p_1 = \cdots = p_m = \frac{1}{m}$ and $q_1 = \cdots = q_n = \frac{1}{n}$ as one would obtain them from random observations. The computation times therefore include the sorting of the support (Quasi-linear Prokhorov) and the calculation and sorting of the distances ($\varDelta$-Bisection).
We selected the following three problem classes

- m-discretization of $\mathcal{N}(0, 1)$ versus n-discretization of $\mathcal{N}(0, 1)$
- m-discretization of $\mathcal{U}(0, 1)$ versus n-discretization of $\mathcal{U}(0, 1)$
- m-discretization of $\mathcal{N}(0, 1)$ versus n-discretization of standard Cauchy

and combined their computation times. Each entry of the tables consists of the mean time over all problem classes and its standard error. We omit instances where the X and Y are drawn from multivariate distributions, as the Δ-Bisection is independent of the dimension. As mentioned in Remark 3.4.4, the instances with similar m and n have decreased computation time. To verify the complexity, we have therefore increased n by one, which we omit in the tables.

6.1.1 Quasi-Linear Prokhorov

The theoretical complexity of the Quasi-linear Prokhorov algorithm for the Euclidean metric is $\mathcal{O}\left(m \cdot \max(\log(m), \log(\frac{1}{acc_X}))\right)$. By $\mathrm{steps}(m, n)$, we denote the average number of steps of the first bisection, the reduction of candidates, to take $\log(\frac{1}{acc_X})$ into account. Table 6.1 lists the average computation times $T^Q(m, n)$ and its standard error in seconds.

In Table 6.2, we divide $T^Q(m, n)$ by $T^Q_{reg}(m, n) := (m + n) \cdot (\log_2(m + n) + \mathrm{steps}(m, n))$ to obtain a nearly constant ratio. This verifies the theoretical complexity as shown in Proposition 4.1.10. In our test cases, the $\mathrm{steps}(m, n)$ term dominates the $\log_2(m+n)$ term. For example, for the $\mathcal{U}[0, 1]$ distribution, acc_X is always less or equal to $\frac{1}{m}$ and likely smaller than $\frac{1}{m+n}$.

In Table 6.3, we list the computation times for big "$\mathcal{N}(0, 1)$ versus $\mathcal{N}(0, 1)$" problems, averaged over 200 instances per discretization level.

As a rough guideline, we are able to solve instances of size $m = n = 40.000$ in one second, and of size $m = n = 3.000.000$ in one minute.

6.1.2 Δ-Bisection

The theoretical complexity of the Δ-Bisection for the Euclidean metric is $\mathcal{O}\left(mn^2 \log(mn)\right)$. Table 6.4 lists the average computation times $T^\Delta(m, n)$ and its standard error in seconds.

In Table 6.5, we divide $T^\Delta(m, n)$ by $T^\Delta_{NWS}(m, n) := (m + n)mn \log_2(m +$

m vs. n	1 000	2 000	5 000	10 000	20 000	30 000	40 000	50 000
1 000	0.02	0.03	0.05	0.10	0.20	0.30	0.40	0.51
	0.0001	0.0002	0.0003	0.0005	0.0010	0.0015	0.0021	0.0027
2 000	0.03	0.03	0.06	0.12	0.22	0.32	0.43	0.55
	0.0002	0.0002	0.0004	0.0006	0.0011	0.0017	0.0022	0.0028
5 000	0.05	0.07	0.10	0.15	0.26	0.37	0.50	0.61
	0.0003	0.0004	0.0006	0.0009	0.0015	0.0020	0.0029	0.0032
10 000	0.10	0.12	0.15	0.21	0.33	0.44	0.57	0.70
	0.0006	0.0007	0.0010	0.0014	0.0019	0.0025	0.0034	0.0038
20 000	0.20	0.22	0.26	0.33	0.45	0.59	0.72	0.85
	0.0011	0.0012	0.0015	0.0020	0.0028	0.0035	0.0044	0.0047
30 000	0.30	0.32	0.37	0.44	0.59	0.72	0.86	0.99
	0.0016	0.0018	0.0021	0.0026	0.0036	0.0047	0.0050	0.0056
40 000	0.40	0.43	0.50	0.58	0.72	0.86	0.98	1.12
	0.0022	0.0023	0.0029	0.0034	0.0045	0.0051	0.0063	0.0062
50 000	0.52	0.55	0.62	0.70	0.85	0.99	1.13	1.26
	0.0028	0.0030	0.0034	0.0040	0.0050	0.0059	0.0068	0.0080

Table 6.1: Average computation time $T^Q(m,n)$ and standard error of the Quasi-linear Prokhorov algorithm in seconds

$n)\log_2(mn)$, the worst case complexity of the network simplex, and observe a better average complexity than expected in worst case. In Table 6.6, we divide $T^{\Delta}(m,n)$ by $T^{\Delta}_{BMF}(m,n) \coloneqq (m \vee n)(m \wedge n)^2 \log_2(mn)$, the bipartite max flow worst case complexity, and observe a different behaviour. For m and n of similar order, Cplex is faster than the bipartite max flow, but for $n \ll m$, the bipartite max flow is superior as it was observed in Gusfield et al. (1987). Regressing $T^{\Delta}(m,n)$ against polynomials of m and n suggests a complexity of $T^{\Delta}_{reg}(m,n) \coloneqq (mn)^{1.13}\log_2(mn)$. In contrast to the Reduced Monge Algorithm, the complexity of the Δ-Bisection differs from its worst case complexity. We suppose this is due to the fact that the complexity $\mathcal{C}(\mathcal{E})$ varies with δ. The amount of edges in the network is determined by δ and is only bounded by mn

m vs. n	1 000	2 000	5 000	10 000	20 000	30 000	40 000	50 000
1 000	1.13	1.12	1.03	0.97	0.91	0.87	0.86	0.86
2 000	1.12	1.11	1.07	1.02	0.95	0.92	0.91	0.91
5 000	1.04	1.08	1.07	1.05	1.01	0.98	0.99	0.96
10 000	0.97	1.02	1.06	1.05	1.03	1.01	1.02	1.01
20 000	0.91	0.96	1.01	1.04	1.03	1.05	1.04	1.04
30 000	0.88	0.92	0.98	1.01	1.05	1.04	1.05	1.04
40 000	0.86	0.90	0.99	1.02	1.05	1.05	1.04	1.04
50 000	0.86	0.91	0.97	1.01	1.04	1.05	1.05	1.04

Table 6.2: Ratio $T^Q(m,n)\,/\,T^Q_{reg}(m,n)$

$m = n$	100 000	500 000	1 000 000	2 000 000	3 500 000	5 000 000
avg. time in s	2.90	16.47	17.75	37.23	67.28	98.59
standard error	0.008	0.026	0.030	0.059	0.092	0.138

Table 6.3: Average computation time $T^Q(m,n)$ and standard error of the Quasi-linear Prokhorov algorithm in seconds for large instances

from above, but most likely smaller.

In total, we are able to solve instances of size $m = n = 4.000$ in roughly a minute.

6.1.3 Comparison of the Algorithms

As expected, the Quasi-linear Prokhorov algorithm is far superior to the Δ-Bisection regarding its running time. To be precise, a 1 000 vs. 1 000 evaluation with the Δ-Bisection takes on average over twice as long as a 50 000 vs. 50 000 evaluation with the Quasi-linear Prokhorov algorithm. Additionally, the Δ-Bisection is more demanding in terms of memory, $\mathcal{O}(mn)$ vs. $\mathcal{O}(m)$ respectively. Thus, due to memory limitations, the biggest instances we could solve were $m = n = 10\,000$ in about 10 minutes for the Δ-Bisection

m vs. n	1 000	1 500	2 000	3 000	4 000
1 000	2.69	4.27	5.88	9.56	13.65
	0.022	0.038	0.061	0.107	0.159
1 500	4.29	6.72	9.51	15.39	22
	0.038	0.064	0.100	0.177	0.268
2 000	6.02	9.52	13.29	21.72	30.56
	0.055	0.099	0.147	0.265	0.395
3 000	9.69	15.46	21.83	35.06	49.37
	0.098	0.175	0.259	0.462	0.680
4 000	13.91	22.18	30.78	49.64	69.46
	0.149	0.261	0.386	0.664	1.010

Table 6.4: Average computation time $T^{\Delta}(m,n)$ and standard error of the Δ-Bisection in seconds based on the Network simplex for $\mathcal{E}(\cdot)$

and $m = n = 100\,000\,000$ in about 40 minutes for the Quasi-linear Prokhorov algorithm.

6.2 Obtaining Fast Lower Bounds

As we have a significant speed up and, more importantly, maximum problem size for one-dimensional spaces, we can project from a multidimensional space to one-dimensional subspaces and obtain a lower bound. While this bound is not exact enough to quantify distances or convergence speeds, it can be used to reject convergence. This is based on the results Lemma 4.1.6 and Lemma 4.2.10, stating a monotonically decreasing behavior with respect to linear projections.

To show the speed up, we take a sequence of two-dimensional normal distri-

m vs. n	1 000	1 500	2 000	3 000	4 000
1 000	1.98	1.57	1.30	0.99	0.81
1 500	1.58	1.31	1.15	0.91	0.76
2 000	1.33	1.15	1.01	0.84	0.71
3 000	1.00	0.91	0.84	0.72	0.63
4 000	0.83	0.77	0.71	0.63	0.56

Table 6.5: Ratio $T^{\Delta}(m,n)\,/\,T^{\Delta}_{NWS}(m,n)$

m vs. n	1 000	1 500	2 000	3 000	4 000
1 000	1.28	1.31	1.33	1.40	1.47
1 500	1.32	0.89	0.93	0.97	1.02
2 000	1.36	0.93	0.72	0.76	0.79
3 000	1.42	0.98	0.76	0.53	0.55
4 000	1.50	1.03	0.79	0.55	0.43

Table 6.6: Ratio $T^{\Delta}(m,n)\,/\,T^{\Delta}_{BMF}(m,n)$

butions F_k with mean $(0,0)$, and covariance

$$\sigma_k := \begin{pmatrix} 0.1 & 0.05 + 0.05 \cdot \sqrt[10]{\frac{1}{k}} \\ 0.05 + 0.05 \cdot \sqrt[10]{\frac{1}{k}} & 0.1 \end{pmatrix}$$

which we compare with the two-dimensional normal distribution F, also with mean $(0,0)$ but covariance

$$\sigma_i := \begin{pmatrix} 0.1 & 0.03 \\ 0.03 & 0.1 \end{pmatrix}.$$

As $\sigma_k \nrightarrow \sigma$, we expect $\mathrm{d}_P(F_k, F) \nrightarrow 0$ and also $\hat{\mathrm{d}}_P(F_k, F) \nrightarrow 0$ for the lower bounds obtained from projection. For each sample point X_j, we obtain its one-dimensional estimator by projecting it on a one-dimensional linear subspace containing $(0,0)$. To increase the accuracy of our bound, we compute 4 projections (parametrized by the angle between the x-axis and the subspace, 0°, 45°, 90° and 135°) and use the maximal obtained distance per distribution as lower bound. We use sample sizes $m = n = 2000$ and report the exact and estimated distances as well as their respective computation times in Figure 6.1.

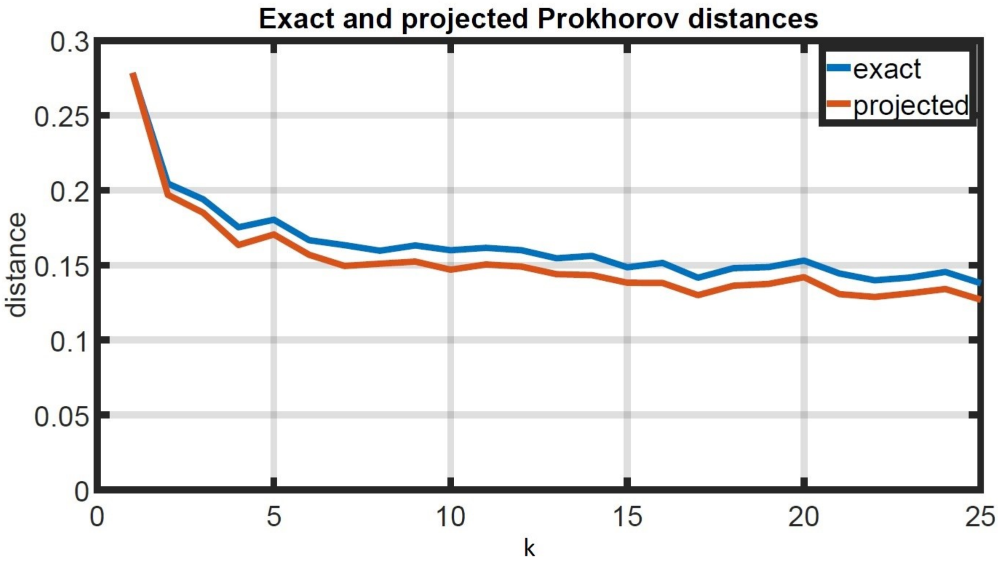


Figure 6.1: Comparison of exact Prokhorov distance with a total computation time of 28.5 minutes and lower bound obtained from one-dimensional projections with a total computation time of 2.5 seconds.

It can be clearly seen from the projections that no convergence happens and the projection is indeed a lower bound. Moreover, while the exact computation took roughly 28.5 minutes for all 25 distances, the projections were obtained in less than 2.5 seconds. This is a time gain of factor > 675, a significant advantage.

Besides faster computation the projection method can also be used to deal with sample sizes > 10000, the memory limit we explored in Subsection 6.1.3.

6.3 Verifying the Convergence Rate of Sampled Distributions

In this part we numerically verify the theoretical convergence rates of Graf and Luschgy (2009) and Kersting (1978) of $N^{-\frac{1}{d}}$ and see better rates in practice. Especially for the uniform distribution Figure 6.2 on $[0, 1]$, we see a polynomial convergence instead of the linear worst case.

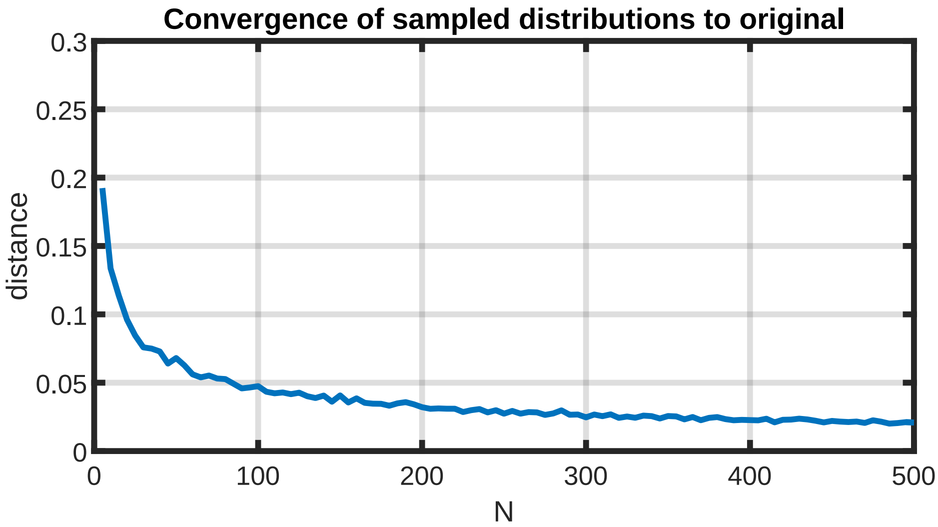


Figure 6.2: Distance of a sampled distribution with N samples to its original one, a $[0, 1]$ uniform distribution, averaged over 50 draws per N.

6.4 Verifying Weak Convergence

In Figure 6.3, we randomly discretize normal distributions with varying mean $\mu \in [-0.5, 0.5]$ and $\sigma \in [0.5, 1.5]$ for $m = n = 25\,000$ and plot their approximate distance $\mathrm{d}_P(\mathcal{N}(0, 1), \mathcal{N}(\mu, \sigma))$. We see the expected convergence for $\mu \to 0$ and $\sigma \to 1$ as the Prokhorov distance decreases.

In Figure 6.4 we keep $\sigma = 1$ fixed and vary only $\mu \in [-0.1, 0.1]$ and in Figure 6.5 we fix $\mu = 1$ and vary $\sigma \in [0.6, 1.4]$. In this cases, the discretizations

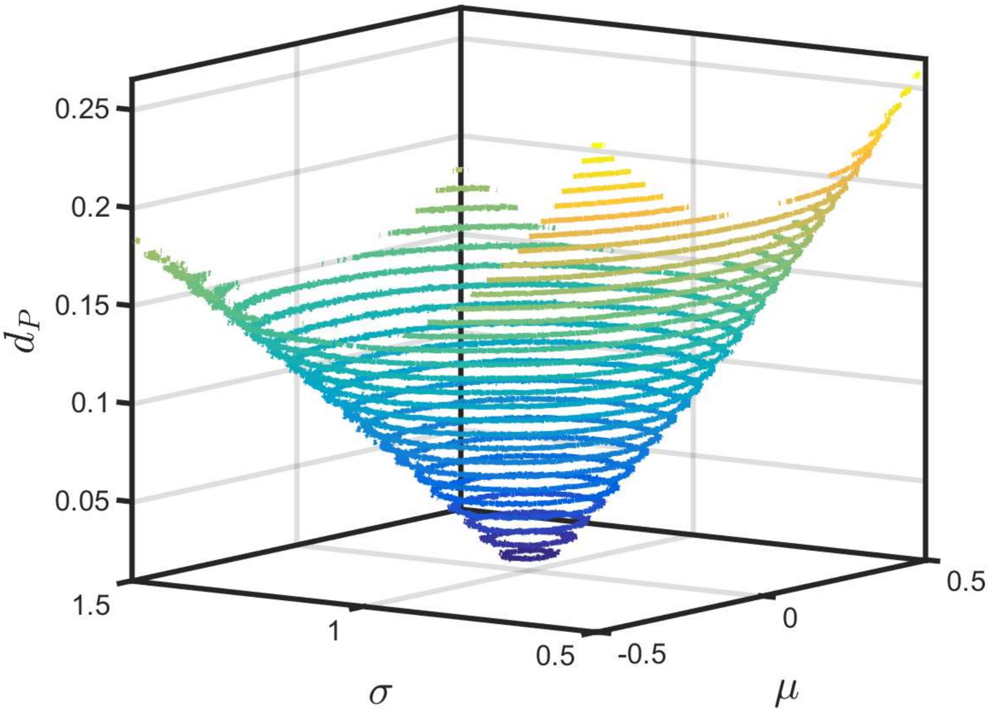


Figure 6.3: Prokhorov distance of $\mathcal{N}(0,1)$ to $\mathcal{N}(\mu,\sigma)$

were obtained from the quantiles to smooth the graph. As above, we see the convergence of the distributions for $\mu \to 0$ and $\sigma \to 0$ respectively.

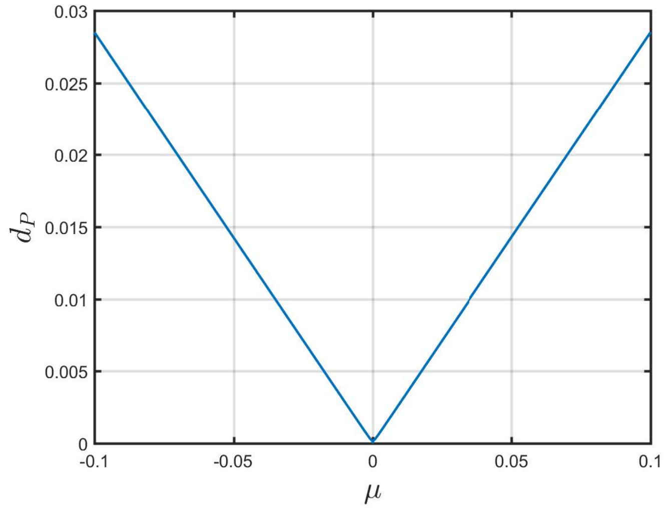


Figure 6.4: Prokhorov distance of $\mathcal{N}(0,1)$ to $\mathcal{N}(\mu,1)$

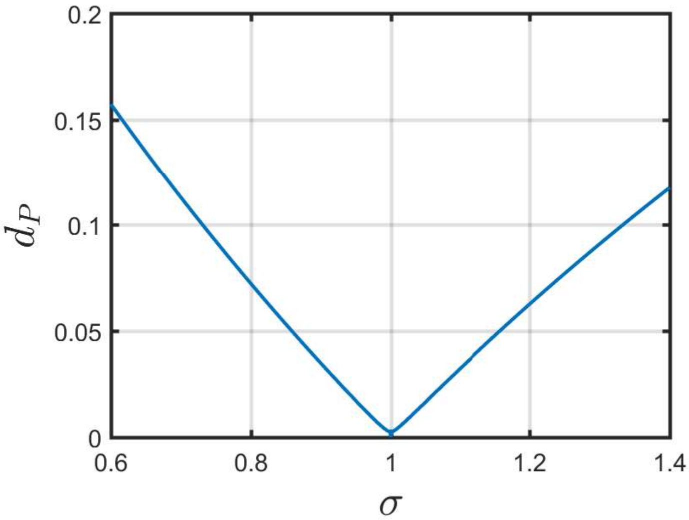


Figure 6.5: Prokhorov distance of $\mathcal{N}(0,1)$ to $\mathcal{N}(0,\sigma)$

7 Outlook

Achievements

Throughout this thesis, we have newly introduced the f-$\mathcal{E}$-class of probability metrics. This is an extensive set of metrics, and provides a joint theoretical and computational framework for the Prokhorov and Wasserstein-∞ metrics. We showed the split in two separate topologies gives a straightforward classification of the f-$\mathcal{E}$-class, based solely on the behavior of the f function around 0. While the Prokhorov and the Wasserstein-∞ metric are the most well known representatives of each class, a new broad class of metrics is now accessible for further research and applications.
We focused on exact and efficient computation of these metrics for finite-support distributions, we were able to show correctness and worst case complexities for our algorithms. For the Prokhorov metric, we are in general the first do so, and for the Wasserstein-∞ metric, algorithms were only available for special settings, i.e. on $\mathbb{R}$. Most notably, we achieved the same complexities as the Wasserstein-p metric in general settings, the most widely used and researched probability metric by far. This broadens the set of computationally accessible metrics One shortcoming of the Wasserstein-p metric, metrizising the weak topology only on specific underlying metric spaces, can be overcome by selecting a metric of the Prokhorov class of f-$\mathcal{E}$-metrics, i.e. $f(\delta) > 0$ for all $\delta > 0$.

As a next step, we introduced the concept of quasi-convex metric, a framework combining ordered and metric spaces. This includes the most common case of

$\mathbb{R}$ with the usual order and the Euclidean metric, but generalizes it. For this, we were able to show the existence of Monge sequences for the evaluation of the exceedance, a setting where greedy flow algorithms are exact, and therefore highly efficient.

We further refined this by explicitly looking at $\mathbb{R}$ with the usual order and the Euclidean metric for the Prokhorov metric, and provided an exact weakly quasi-linear algorithm. This is by far the most detailed analysis of the computation of the Prokhorov metric to date to the best of our knowledge. Similar steps were carried out for the Wasserstein-∞ metric to obtain an exact strongly quasi-linear algorithm for general quasi-convex settings.

In total, we have developed a broad class of probability metrics, analyzed their theoretical properties and provide a comprehensive set of exact and efficient algorithms for finitely supported measures.

Applications

As the constraints of the f-$\mathcal{E}$-class are the classical transportation constraints, our metrics provide a broader class of objective functions.
A Prokhorov metric of 0.05 can colloquially be interpreted as "At most 5% of the mass has higher transportation cost than 0.05". In Applications, this could be used as a measure for reliability, i.e. "At least 95% have low costs". The extreme case, the Wasserstein-∞-metric, says all mass has to be transported for cost no higher than its value, i.e. the single most expensive part drives the value. Interesting settings for this objective include time optimization, i.e. after which time will all goods be transported.

In Computer Vision, the classical approach is via various types of neural nets, usually convolutional neural nets. These are powerful to extract single objects of interest out of the picture, i.e. for cars “Is there a pedestrian on the road?”. While this constitutes a prevalent use case within computer vision, it can not be used to check “optical” similarity of two pictures, i.e. do they look similar to the human eye. This can be answered by transportation metrics, and therefore also by our f-$\mathcal{E}$-class. Potential use cases lie within copyright infringements, however these metrics are not stable with respect to simple transformation like trimming and mirroring. Image comparison also constitutes one of the common benchmarks for transportation algorithms Schrieber et al. (2017)

A typical statistical test is whether an empirical distribution G stems from a certain distribution F or not. This can be done by taking sampled distributions $F_1, F_2, \dots$ of the same sample size n as G of F and calculating their distances $\mathrm{d}_i := \mathrm{d}(F, F_i)$ by a metric of ones choice. Sorting the d_i constructs confidence intervals of distances, generating a cut-off value C solely based on the confidence level. For $\mathrm{d}(F, G) > C$ we reject the hypothesis “G is sampled from F”, otherwise we accept it. The advantage over other tests lies in the topological properties of the f-$\mathcal{E}$-class, i.e. robustness against outliers for the Prokhorov metric or higher sensitivity for the Wasserstein-∞ metric.
In general, metrics of the f-$\mathcal{E}$-class can be used wherever other probability metrics have been used previously and broaden the tool set by allowing more flexibility.

Open Problems

While we were fully able to match the complexities of exact algorithms for the Wasserstein-p metric, $\mathcal{O}(mn^2 \log(m))$, in the general setting with the Prokhorov and Wasserstein-∞ metric, we were only able to match the strongly quasi-linear complexity $\mathcal{O}(m \log(m))$ with the Wasserstein-∞ metric. This leaves a gap for the Prokhorov metric, as our algorithm is weakly quasi-linear, $\mathcal{O}\left(m \cdot \max(\log(m), \log(\frac{1}{acc_X}))\right)$. Practically this has no big impact, as $\frac{1}{acc_X}$ usually does not exceed m by relevant magnitude. But it remains open whether this gap can be closed or is inherent to its structure.

As our algorithms work solely for measures with finite support, approximation results for continuous distributions are needed to ensure reliable results for arbitrary distributions. For an overview, independent of the choice of metric, we refer to Graf and Luschgy (2000), introducing the problem and covering general geometric observations. Results for the Prokhorov metric can be found in Graf and Luschgy (2009), with results for specific settings dating back to Kersting (1978) and Massart (1988). Roughly speaking, the approximation error declines behaves like $\left(\frac{1}{n}\right)^{\frac{1}{d}}$ in $\mathbb{R}^d$ with support of size n. These convergence questions have obviously only been tackled for the Prokhorov and Wasserstein-∞ metric so far, but not in general for our newly introduced class of f-$\mathcal{E}$. We assume similar convergence rates hold true, based on regularity properties of f and f^{-1}, most likely its Lipschitz constant. We have not covered this topic to focus on the algorithmic aspects, hopefully sparking interest in f-$\mathcal{E}$ metrics for further research.

Bibliography

Alon, N., Cosares, S., Hochbaum, D. S., and Shamir, R. (1989). An algorithm for the detection and construction of Monge sequences. *Linear Algebra and its Applications*, 114–115:pp. 669–680.

Ambrosio, L., Gigli, N., and Savare, G. (2005). Gradient flows in metric spaces and in the space of probability measures.

Bobkov, S. and Ledoux, M. (2016). One-dimensional empirical measures, order statistics, and Kantorovich transport distances. `http://www-users.math.umn.edu/~bobko001/preprints/2016_BL_Order.statistics_Revised.version.pdf`, accessed on 2019/02/22.

Champion, T., De Pascale, L., and Juutinen, P. (2008). The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM J. Math. Analysis*, 40:pp. 1–20.

Drescher, C., Schwinn, J., and Werner, R. (2018). Efficient computation of the Prokhorov metric for finite-support distributions. *Preprint.*

Drescher, C. and Werner, R. (2019). Efficient computation of the Wasserstein-∞ metric for measures with finite support. *Preprint.*

Dudley, R. M. (1968). Distances of probability measures and random variables. *Ann. Math. Statist.*, 39(5):pp. 1563–1572.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):pp. 642–669.

Efrat, A., Itai, A., and Katz, M. J. (2001). Geometry helps in bottleneck matching and related problems. *Algorithmica*, 31(1):pp. 1–28.

García-Palomares, U. and Giné, E. (1977). On the linear programming approach to the optimality property of Prokhorov's distance. *Journal of Mathematical Analysis and Applications*, 60(2):pp. 594–600.

Garel, B. (1981). Calcul pratique de la distance de Prokhorov. *Statistique et analyse des données*, 6(3):pp. 35–46.

Garel, B. and Massé, J.-C. (2009). Calculation of the Prokhorov distance by optimal quantization and maximum flow. *Advances in Statistical Analysis*, 93:pp. 73–88.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70:pp. 419–435.

Givens, C. and Shortt, R. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Math. J.*, 31(2):pp. 231–240.

Graf, S. and Luschgy, H. (2000). *Foundations of Quantizations for Probability Distributions*. Springer, Heidelberg.

Graf, S. and Luschgy, H. (2009). Quantization for probability measures in the Prokhorov metric. *Theory of Probability and Its Applications*, 53(2):pp. 216–241.

Gusfield, D., Martel, C. U., and Fernández-Baca, D. (1987). Fast algorithms for bipartite network flow. *SIAM Journal on Computing*, 16(2):pp. 237–251.

Hoffman, A. J. (1963). On simple transportation problems. *Convexity, Proceedings of Symposia in Pure Mathematics*, 7:pp. 317–327.

Huber, P. J. (1981). *Robust statistics*. John Wiley and Sons, New York.

Jylhä, H. (2015). The L^∞ optimal transport: infinite cyclical monotonicity and the existence of optimal transport maps. *Calculus of Variations and Partial Differential Equations*, 52(1):303–326.

Kantorovitch, L. (1958). On the translocation of masses. *Management Science*, 5(1):pp. 1–4.

Kersting, G. D. (1978). Die Geschwindigkeit der Glivenko-Cantelli-Konvergenz gemessen in der Prohorov-Metrik. *Mathematische Zeitschrift*, 163:65–102.

Kleinschmidt, P. and Schannath, H. (1995). A strongly polynomial algorithm for the transportation problem. *Mathematical Programming*, 68(1):pp. 1–13.

Kloeckner, B. (2012). Approximation by finitely supported measures. *ESAIM: COCV*, 18(2):pp. 343–359.

Kovács, P. (2015). Minimum-cost flow algorithms: an experimental evaluation. *Optimization Methods and Software*, 30(1):94–127.

Kuhn, H. W. (1955). The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Lee, Y. T. and Sidford, A. (2013). Following the path of least resistance: An $\tilde{O}(m\sqrt{n})$ algorithm for the minimum cost flow problem. *CoRR*, abs/1312.6713.

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

Liu, A., Liu, J.-G., and Lu, Y. (2018). On the rate of convergence of empirical measure in $\infty-$Wasserstein distance for unbounded density function. *arXiv e-prints*, page arXiv:1807.08365.

Major, P. (1978). On the invariance principle for sums of independent identically distributed random variables. *Journal of Multivariate Analysis*, 8(4):487 – 517.

Massart, P. (1988). About the Prohorov distance between the uniform distribution over the unit cube in $\mathbb{R}^d$ and its empirical measure. *Probability Theory and Related Fields*, 79(3):pp. 431–450.

Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269 – 1283.

Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais.* De l'Imprimerie Royale.

Norfolk, T. (1991). When does a metric generate convex balls? `http://www.math.uakron.edu/~norfolk/convex.ps`, accessed on 2017/07/07.

Orlin, J. B. (1997). A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129.

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability and Its Applications*, 1(2):pp. 157–214.

Rachev, S. T., Klebanov, L., Stoyanov, S., and Fabozzi, F. (2013). *The Methods of Distances in the Theory of Probability and Statistics*. Springer, New York.

Rachev, S. T. and Rüschendorf, L. and, S. A. (1992). Uniformities for the convergence in law and in probability. *Journal of Theoretical Probability*, 5:33 – 44.

Schay, G. (1974). Nearest random variables with given distributions. *Ann. Probab.*, 2(1):pp. 163–166.

Schrieber, J., Schuhmacher, D., and Gottschlich, C. (2017). Dotmark – a benchmark for discrete optimal transport. *IEEE Access*, 5:271—282.

Schwinn, J. (2019). *Novel applications of Column Generation in large-scale linear programming.* doctoralthesis, Universität Augsburg.

Spinrad, J., Brandstädt, A., and Stewart, L. (1986). Bipartite permutation graphs. *Discrete Applied Mathematics*, 18(3):pp. 279–292.

Startek, M. (2010). The Ky-Fan metric and the change of scale. *Journal of Pure and Applied Mathematics*, 59(4):pp. 375–379.

Strassen, V. (1965). The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):pp. 423–439.

Trillos, G. N. and Slepčev, D. (2015). On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383.

Villani, C. (2009). *Optimal Transport*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag Berlin Heidelberg.