Limit Laws for Empirical Optimal Transport

Dissertation

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

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

vorgelegt von
Marcel Klatt
aus Hannover

Göttingen, 2021

Thesis Committee:

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

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

Prof. Dr. Daniel Rudolf
Institute for Computer Science and Mathematics, University of Passau

Members of the Examination Board:

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

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

Further Members of the Examination Board:

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

Prof. Dr. Gerlind Plonka-Hoch
Institute for Numerical and Applied Mathematics, University of Göttingen

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

Date of the Oral Examination: February 9, 2022

¡Vaya camina por arriba ‘el mambo!

Alexander Abreu Manresa,
Havana D’Primera

Preface

Playing with sand is undoubtedly one of the favorite games of little kids, and of course, I was not the exception. Spending hours of my summer holidays with my family at the beach digging holes and piling sand for creating castles was certainly amusing but exhausting. Indeed, I could have struggled less building the castles if I had ever wondered how to transport the sand with minimal effort. In those days, I neither had the idea that behind this question hides the beautiful and surprisingly rich mathematical theory of *optimal transport* (OT), nor that it would become the basis for my thesis almost thirty years later.

The origin of the OT theory dates back to 1781 when the French mathematician and engineer Gaspard Monge investigated how to transport mass from one collection of locations to another in the most economical manner. In his manuscript *Mémoire sur la théorie des déblais et des remblais*, Monge defined the transport cost as the sum over all transport paths of the mass travelled times their respective distance. Monge was primarily interested in finding a solution for optimal transportation but he could not give a satisfactory answer. Despite the work of Appell later in 1887, OT appeared to be sinking into oblivion. In 1942, more than 150 years after Monge's initial contributions, the Soviet mathematician Leonid V. Kantorovich proposed a relaxation of Monge's original OT formulation. Kantorovich's probabilistic perspective on OT, that later became popular as the *Monge-Kantorovich transportation problem*, turned into a new upswing in the mid-twentieth century, primarily due to OT's practical relevance in economics. OT was then investigated within a framework of what today is popular as a standard linear program and indeed paved the way for the theory of general linear programming. Significant contributors to the development were Tolstoí, Hitchcock, Dantzig[1], and Koopman. The latter received the Nobel prize in economics together with Kantorovich in 1975. Since then, an enormous and surprisingly rich OT theory has emerged, not only giving a detailed and precise answer to Monge's initial problem of finding optimal

[1] Vershik (2002) and Cottle et al. (2007) give a detailed historical account with special focus to the contributions made by Kantorovich and Dantzig, respectively.

solutions[2] but pushing OT even further and beyond a purely mathematical context. Indeed, OT has quickly been realized to be related to various topics[3] and is recently gaining increasing popularity in fields ranging from biology, economics, and social science to machine learning, physics and *statistics*. However, despite OT's conceptual appeal, new challenges have arisen with the dawn of the big-data era. Solving OT problems is tremendously time-consuming, even for small-size instances, questioning its versatile use in applications. These difficulties have motivated a new line of research on the computational side of OT, with perhaps the most prominent approach based on *entropy regularization*.

The central theme of the thesis revolves around *statistical* aspects of OT and its entropy regularized surrogates, and aims to contribute to a better understanding of *empirical* OT and their related *empirical* entropy regularized surrogates. This thesis is a compilation of the main results of the three articles Klatt et al. (2020a), Klatt et al. (2020b) and Hundrieser et al. (2021c) found in the Addenda listed in A, B and C, respectively.

Chapter 1 contains a short account on the statistical framework inherent in all three articles. The primary purpose hereby is to give necessary background and to set notation. Further, the basic theory of OT, standard linear programming and the entropy regularization approach is presented. **Chapter 2** summarizes the main results from Klatt et al. (2020a) and Klatt et al. (2020b) together with a detailed discussion and background on related work. **Chapter 3** follows an identical exposition and presents the main results from Hundrieser et al. (2021c) and a thorough discussion including related work.

The three articles focus on *distributional limit laws* for various empirical OT and empirical entropy regularized OT quantities. Distributional limit laws capture the asymptotic fluctuation of empirical estimators around their population quantities after proper standardization. They are essential for statistical data analysis, e.g., for hypothesis testing and deriving asymptotic confidence bands. The presented articles are broadly characterized in two categories: Distributional limit laws for empirical OT and probability measures supported (1) on finite and (2) on more general spaces.

Category (1) involves the articles Klatt et al. (2020a) and Klatt et al. (2020b). In particular, Klatt et al. (2020a) investigates the OT between probability measures with finite support and contains novel contributions on distributional limit laws for empirical OT

[2] The first answer to Monge's problem appeard in Sudakov (1979). Some of the proof arguments were later fixed by Evans and Gangbo (1999), and Ambrosio and Pratelli (2003).

[3] Important contributors in this regard were Zolotarev, Sudakov, Kellerer, Rachev, Rüschendorf, Brenier, Smith, Knott, McCann, Evans, Gangbo, Ambrosio. The list is far from being complete and many more can be found, e.g., in the textbooks by Rachev and Rüschendorf (1998), Villani (2009) and Santambrogio (2015).

plans. Indeed, the results are more general and focus on distributional limit laws for empirical optimal solutions to random linear programs. Klatt et al. (2020b) explores the entropy regularized OT between probability measures with finite support. Apart from entropy regularization, the article considers more general regularization methods. The main result subsumes central limit laws for the empirical entropy regularized OT plan and its corresponding transport cost popularized as *Sinkhorn divergence*. The statistical results are applied to colocalization analysis of protein interaction networks.

Hundrieser et al. (2021c), placed in category (2), investigates OT between probability measures with support in general Polish spaces. The article provides a unifying approach to distributional limit laws for the empirical OT cost. The main statement implies well-known results, e.g., for probability measures with finite support or support on the real line, but also contains novel extensions for the empirical OT cost between probability measures supported on low-dimensional Euclidean or Polish spaces. In particular, the article demonstrates the importance of two fundamental principles that are present throughout this thesis and seamlessly integrate within a statistical context: *Duality* for (in)finite dimensional linear programs and *weak convergence* of underlying empirical processes.

Own Contributions

- Klatt et al. (2020a) (Addendum A) was written jointly with Y. Zemel and A. Munk. Most of the statements were achieved during meetings and discussions in Göttingen and Cambridge. Y. Zemel and I contributed equally, and we received helpful comments and suggestions from A. Munk.

- Klatt et al. (2020b) (Addendum B) is mostly my own contribution. The applications of the statistical statements to colocalization analysis of protein interaction networks is the result from a close collaboration with C. Tameling and A. Munk.

- Hundrieser et al. (2021c) (Addendum C) is the joint work of all authors. My first attempt for a rigorous differentiability analysis of the OT functional was corrected by S. Hundrieser. The central limit theorems for empirical OT costs were achieved during joint meetings. The preparation of the manuscript was distributed into equal parts between S. Hundrieser and me. Authors T. Staudt and A. Munk made careful corrections and suggestions.

Acknowledgments

I owe my deepest gratitude to my supervisor ***Axel Munk*** for introducing me to the topic of optimal transport and the opportunity to be his PhD student. His ideas to approach mathematical problems together with his guidance have strongly influenced my mathematical thinking and my career as a (young) researcher. I would also like to thank ***Dominic Schuhmacher***. His open-door policy made discussions easy and often illuminating. Further, I enjoyed my very first mathematical conference in Valladolid with him in 2018 and especially all the restaurants we both had to visit after a hard working day! I am also indebted to ***Daniel Rudolf*** and his various helpful comments and suggestions.

I am particularly grateful for the assistance given by ***Yoav Zemel***. I have greatly benefited from his excellent mathematical background. Despite his duties as a father of two young children he always found time to answer all my questions. Thanks Yoav for not only being a close collaborator but also a good friend!

My time as a PhD student in Göttingen would have certainly been less successful without the support of the members of the Institute for Mathematical Stochastics. Special thanks in this regard go to the *Optimal Transport Crew* including ***Shayan Hundrieser***, ***Florian Heinemann***, ***Giacomo Nies***, ***Christoph Weitkamp***, ***Thomas Staudt***, ***Gilles Mordant***, ***Max Sommerfeld*** and ***Carla Tameling***. It has been a pleasure working with you all, exchanging ideas and preparing articles together.
I should not forget to mention ***Christian Böhm***. He is the good soul of the institute and always has an ear for my concerns such as a crashing stock market, problems with my car or any technical issues.

Thanks also to the ***Research Training Group 2088*** for the financial support and the manifold opportunities to present my research in yearly retreats, together with the workshops and colloquia that definitely broadened my mathematical horizon.

Writing a thesis is only fun if there is also sufficient support from close friends who at its best share the same hobbies. I would like to thank ***Anne Hobert*** and ***Wladimir Bosten*** for taking this important role during that time. Special thanks go also to all ***Salseros Göttingen*** members.

Cualquier intento a cualquier nivel no habría sido completado satisfactoriamente sin la presencia de ***Laura Margarita Sánchez Galindo*** y su pequeño rayo de sol, ***Teofilo José***. Estoy muy agradecido por todo lo que ambos han hecho por mí. Ustedes son mi roca en medio del mar agitado.

Zu guter Letzt möchte ich meinen Eltern, ***Simone*** und ***Rüdiger***, als auch meinem Bruder, ***Niklas*** und seiner Freundin ***Wiebke*** danken. Besonderer Dank gebührt ebenso meiner Oma, ***Helga***, die durch ihre zahlreichen Zusendungen an Kaffee und anderen nützlichen Dingen meine Studienzeit noch angenehmer gestaltet hat. Ohne die Unterstützung meiner Familie würde es diese Dissertationsschrift vermutlich nicht geben. Ich bin stolz, ein Teil dieser Familie zu sein!

Contents

List of Symbols

$\mathbb{N}$	Set of non-negative integers
$\mathbb{R}_+$	Set of non-negative real numbers
$\mathbb{R}^d$	Euclidean space of dimension d
$\mathbb{1}_M, 0_M$	Vector of ones or zeroes of dimension M, respectively
$\mathbb{I}_M$	Identity matrix of dimension $M \times M$
$\langle \cdot , \cdot \rangle$	Dot (scalar) product in Euclidean space
$\mathcal{X}, \mathcal{Y}$	Polish spaces
$\Vert \cdot \Vert$	(Euclidean) Norm
$\mathcal{P}(\mathcal{X})$	Set of probability measures on the space $\mathcal{X}$
μ, ν	Probability measures on the space $\mathcal{X}$
$T_{\#}\mu$	Push-forward measure of μ by T (p. 2)
$\Delta(\mathcal{X})$	Probability simplex for a finite (countable) space $\mathcal{X}$
r, s	Probability measures with finite (countable) support on $\mathcal{X}$
$\xrightarrow{\mathcal{D}}$	Convergence in distribution (weak convergence)
$\Pi(\mu, \nu)$	Set of transport plans (p. 3)
$\Pi(r, s)$	Transportation simplex (p. 7)
$\mathcal{OT}_c(\mu, \nu)$	Optimal transport cost between μ and ν with cost function c (p. 6)
$\mathcal{W}_p(\mu, \nu)$	p-Wasserstein distance between μ and ν (p. 6)
$\mathcal{OT}_{c,\lambda}(r, s)$	Entropy regularized optimal transport between r and s with cost function c and regularization parameter $\lambda > 0$ (p. 14)
$\mathcal{SD}_{c,\lambda}(r, s)$	Sinkhorn divergence between r and s with cost function c and regularization parameter $\lambda > 0$ (p. 14)
$\mathcal{SL}_{c,\lambda}(r, s)$	Sinkhorn loss between r and s with cost function c and regularization parameter $\lambda > 0$ (p. 14)
$\mathcal{EOT}_{c,\lambda}(\mu, \nu)$	Entropy regularized optimal transport between μ and ν with cost function c and regularization parameter $\lambda > 0$ (p. 15)

CHAPTER 1

Introduction

Extracting relevant information from data while aiming for evidence is a challenging task. Indeed, data is typically random that requires to deal with uncertainty – a major topic of mathematical statistics. Consider the fundamental estimation problem concerned with finding a plausible decision for some unknown quantity θ. Instead of access to θ, only a collection of random variables[1] $X_1, \ldots, X_n$ with values in a sample space is available. Within the framework of a statistical model, these random variables follow a distribution P_θ defined on some measurable space $(\mathcal{X}, \mathcal{A})$ and governed by θ. An estimator or statistic for θ is nothing but a measurable function $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ of the random variables. Classical estimation theory is devoted to studying the quality of estimators such as their consistency and rates of convergence as more random variables enter the estimation procedure. A refined analysis of estimators is provided by a *distributional limit law*: If the number of random variables n and a sequence of positive real values r_n tend to infinity, then[2] $r_n(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} Z$ with limiting random variable Z following a specific distribution, also termed distributional limit law. For sufficiently large sample size n, the fluctuation of the random quantity $r_n(\hat{\theta}_n - \theta)$ is thus approximated by the law of the limiting random variable Z. Often the quantity of interest is $\phi(\theta)$, for some known function ϕ, and the *plug-in* estimator $\phi(\hat{\theta}_n)$ is a reasonable choice. If the estimator $\hat{\theta}_n$ meets a distributional limit law, then a related statement involving the random quantity $\phi(\hat{\theta}_n)$ also holds thanks to the *delta method*. Under suitable differentiability assumptions on ϕ, a distributional limit law of the form

$$r_n\left(\phi(\hat{\theta}_n) - \phi(\theta)\right) \xrightarrow{\mathcal{D}} \phi'_\theta(Z), \tag{1.1}$$

[1] Implicit in the definition of a random variable is its measurability with respect to some probability space. This thesis makes no attempt to restate all the necessary mathematical machinery underlying the definitions. For the formalism please refer to the book by Billingsley (2008).

[2] The notation $\xrightarrow{\mathcal{D}}$ refers to convergence in distribution or weak convergence (Billingsley, 2013).

remains valid as the sample size n tends to infinity. The limiting random variable $\phi'_\theta(Z)$ is equal to the derivative of ϕ at θ evaluated at the weak limit of $r_n(\hat{\theta}_n - \theta)$ (Shapiro, 1991; Dümbgen, 1993).

This thesis establishes distributional limit laws of the form (1.1) within the context of *optimal transport* (OT). The general framework considers a complete, separable metric space $\mathcal{X}$, with $\mathcal{P}(\mathcal{X})$ the set of probability measures on $\mathcal{X}$. A probability measure $\mu \in \mathcal{P}(\mathcal{X})$ plays the role of θ. The corresponding estimator $\hat{\theta}_n$ is replaced by the *empirical measure* $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ based on independently, identically distributed (i.i.d.) random variables $X_1, \ldots, X_n \sim \mu$. For a scaling rate $r_n = \sqrt{n}$, this replacement leads to consider the *empirical process* $\sqrt{n}(\hat{\mu}_n - \mu)$ and its weak convergence in some suitable normed space (van der Vaart and Wellner, 1996). The function ϕ is motivated by OT based quantities and their entropy regularized surrogates. Two particular examples for ϕ are the OT cost between two probability measures and the optimal solution of the entropy regularized OT problem, both of which are defined below.

The primary purpose of the following sections is to briefly introduce OT and entropy regularized OT based quantities more formally. For further reading, please refer to the excellent monographs on OT by Rachev and Rüschendorf (1998), Villani (2009) and Santambrogio (2015) as well as the detailed textbooks on linear programming by Luenberger and Ye (1984) and Bertsimas and Tsitsiklis (1997). The book by Peyré and Cuturi (2019) contains a thorough discussion of entropy regularized OT.

1.1 Optimal Transport (OT)

Initial theoretical approaches for OT date back to Monge in 1781. His primary concerns dealt with optimizing the acquisition of raw material for constructions. Monge's initial formulation might state within the Euclidean space $\mathbb{R}^d$ equipped with Borel σ-algebra $\sigma(\mathbb{R}^d)$ and two probability measures[3] $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. For a measurable map $T\colon \mathbb{R}^d \to \mathbb{R}^d$, the quantity $T_\#\mu$ denotes the push-forward measure of μ by T defined as $(T_\#\mu)(A) = \mu(T^{-1}(A))$ for every $A \in \sigma(\mathbb{R}^d)$. Additionally, the formulation involves a measurable cost function $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ that describes with $c(x, y)$ the cost of transportation from location x to y. Then, OT as considered by Monge comprises the optimization problem

$$\inf_{T_\#\mu=\nu} \int_{\mathbb{R}^d} c\,(x, T(x))\, \mathrm{d}\mu(x). \tag{1.2}$$

[3] Within the framework of Monge's seminal work *Mémoire sur la théorie des déblais et des remblais*, the probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ corresponds to material extracted from the soil (*déblais*). The probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$ models a configuration the material is used for construction (*remblais*).

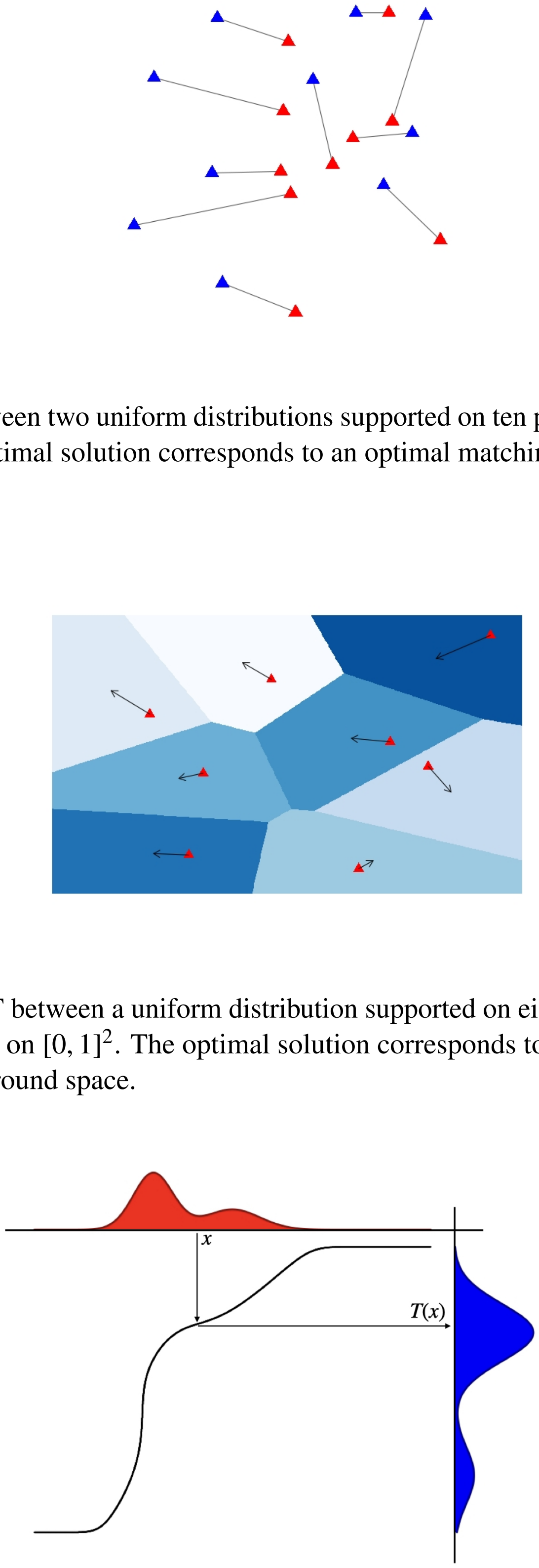
Originally, the l_1 cost $c(x, y) = \|x - y\|_1$ was the first that received Monge's attention. A feasible T in (1.2) defines a transport *map* determining for each particle located at x its destination $T(x)$. Existence statements for an *optimal* transport map T in (1.2) are difficult to prove and generally fail to hold. For instance, if $\mu = \delta_{x_0}$ is equal to a Dirac measure, then $(T_{\#}\mu)(A) = 1 = \nu(A)$ for $A = \{T(x_0)\}$ and a transport map only exists if ν is also a Dirac measure. Nonetheless, for transport from arbitrary ν to $\mu = \delta_{x_0}$, the map $T(x) = x_0$ is feasible for $T_{\#}\nu = \mu$ and also optimal for any cost function. Overall, the *non-linear* constraint $T_{\#}\mu = \nu$ makes the optimization (1.2) challenging, non-symmetric in the measures and often infeasible. A sufficient condition guaranteeing the existence of an OT map T for the cost $c(x, y) = \|x - y\|_1$ involves the probability measure μ to be absolutely continuous with respect to Lebesgue measure (Ambrosio, 2003). A first proof came with Sudakov (1979) more than 200 years after Monge's initial contributions. Some of Sudakov's arguments were later fixed by Evans and Gangbo (1999), and Ambrosio and Pratelli (2003). However, was the Russian mathematician Kantorovich (1942) and his probabilistic perspective on Monge's OT problem (1.2) who made these achievements possible. In modern mathematical terms, Kantorovich's formulation is based on two complete, separable metric spaces $\mathcal{X}$, $\mathcal{Y}$ equipped with respective Borel σ-algebras. For two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, and a measurable cost function $c \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the *Monge-Kantorovich OT* problem is stated as

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \,\mathrm{d}\pi(x, y). \tag{1.3}$$

The set $\Pi(\mu, \nu) := \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid (p_{\mathcal{X}})_{\#}\pi = \mu,\ (p_{\mathcal{Y}})_{\#}\pi = \nu\}$ contains all transport *plans*, where $p_{\mathcal{X}}$ and $p_{\mathcal{Y}}$ are the respective coordinate projections. In the following, OT in this thesis refers to the Monge-Kantorovich formulation (1.3). From a probabilistic perspective, a transport plan $\pi \in \Pi(\mu, \nu)$ is equal to a coupling on the product space $\mathcal{X} \times \mathcal{Y}$ with marginal distributions μ and ν, respectively[4]. Any transport plan π in (1.3) describes for each pair (x, y) the amount of mass transported from location x to destination y. Minimizing over transport plans in (1.3) rather than maps in (1.2) enables mass *splitting* prohibited in Monge's original proposal[5]. Mass splitting designs OT more flexible and allows to consider transportation problems between general probability measures. An illustration is given in Figure 1.1 for discrete-discrete (a), discrete-continuous (semi-discrete) (b) and continuous-continuous OT (c). Apart from the increased generality, OT in (1.3) is symmetric for symmetric cost functions and mathematical analysis is alleviated. For instance, the problem of existence of optimal

[4] The OT in (1.3) might equivalently state as finding a joint law $(X, Y) \sim \pi$ of the random variables $X \sim \mu$ and $Y \sim \nu$ that minimizes the expectation $\mathbb{E}_{\pi}[c(X, Y)]$.

[5] By definition, any transport map T in (1.2) has to transport all mass at x to location $T(x)$.



(a) Discrete OT between two uniform distributions supported on ten points in red and blue, respectively. The optimal solution corresponds to an optimal matching (see Footnote 8 on p. 7 for details).

(b) Semi-discrete OT between a uniform distribution supported on eight points (red) and a uniform distribution on $[0, 1]^2$. The optimal solution corresponds to a Laguerre-Voronoi tessellation of the ground space.

(c) Continuous OT between two absolutely continuous distributions with densities on the real line. The optimal solution corresponds to a map T.

Figure 1.1: **Optimal Transport.** The Monge-Kantorovich OT problem (1.3) with squared Euclidean cost $c(x, y) = \|x - y\|^2$.

solutions is greatly simplified in this framework. First, the set of transport plans $\Pi(\mu, \nu)$ is non-empty as it contains at least the *independence* plan $\mu \otimes \nu$. Second, the existence of a minimizer for (1.3) is a simple consequence of Weierstrass type arguments valid for any non-negative and lower semi-continuous cost function (Villani, 2009, Thm. 4.1). In general, OT in (1.3) is relaxed compared to Monge's formulation (1.2). Any transport map $T\colon \mathcal{X} \to \mathcal{Y}$ induces a transport plan $\pi_T := (Id \times T)_{\#}\mu$ with identical transport cost. Indeed, under certain regularity assumptions on the probability measures and underlying spaces, optimal objective values for (1.2) and (1.3) are equal (Santambrogio, 2015, Sec. 1.5). A particular example is given for squared Euclidean cost $c(x, y) = \|x - y\|^2$ on $\mathbb{R}^d$. If μ is absolutely continuous with respect to Lebesgue measure and μ, ν have finite second moments, then there exists a Lebesgue almost surely unique OT map which is the gradient of a convex function. This statement, popularized as *Brenier's Theorem*, was discovered independently by Brenier (1987), Smith and Knott (1987), and Rüschendorf and Rachev (1990).

Beyond the relaxed OT formulation (1.3), the perhaps most important contribution by Kantorovich was a general *duality* statement. The *dual* OT problem is defined as

$$\sup_{(f,g)\in\Phi_c} \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} g(y)\,\mathrm{d}\nu(y) \tag{1.4}$$

with $\Phi_c := \Big\{(f, g) \in L^1(\mu) \times L^1(\nu) \mid f(x) + g(y) \leq c(x, y)\,\forall (x, y) \in (\mathcal{X} \times \mathcal{Y})\Big\}$. For a non-negative and lower semi-continuous cost function, the maximization problem in (1.4) attains the same optimal value as OT in (1.3) (Villani, 2009, Thm. 5.10). The equality of optimal values is known as *strong duality*[6]. If additionally the cost function is continuous and satisfies $c(x, y) \leq c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y)$ for two functions $(c_{\mathcal{X}}, c_{\mathcal{Y}}) \in L^1(\mu) \times L^1(\nu)$, then maximizers for (1.4) exist (Villani, 2009, Thm. 5.10).

An alternative formulation for (1.4) is obtained employing the *double convexification trick*. If (f, g) is feasible for (1.4), then also (f, f^c) with $f^c(y) := \inf_{x\in\mathcal{X}} c(x, y) - f(x)$ is feasible. The pair (f, f^c) attains an objective value in (1.4) at least as large as the one attained by (f, g). Continuing this iteration with the feasible pair (f^{cc}, f^c) and $f^{cc}(x) := \inf_{y\in\mathcal{Y}} c(x, y) - f^c(y)$ leads to another possible improvement of the objective value for (1.4). Eventually, this process is stationary in the sense that $f^{ccc}(y) = f^c(y)$.

[6] Kantorovich (1942) initially proved strong duality for metric costs $c(x, y) = d(x, y)$. However, the duality statement holds for more general cost functions and even beyond complete, separable metric spaces (Ramachandran and Rüschendorf, 1995).

Consequently, the iteration generates the equivalent dual formulation

$$\sup_{f\in\mathcal{F}_c}\int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x)+\int_{\mathcal{Y}} f^c(y)\,\mathrm{d}\nu(y). \tag{1.5}$$

The supremum is taken over $\mathcal{F}_c$ the set of c-concave functions, i.e., for any $f\in\mathcal{F}_c$ there exists a function $g\colon\mathcal{Y}\to\mathbb{R}\cup\{-\infty\}$ such that $f=g^c$ (Villani, 2009, Thm. 5.10). Any function $f\in\mathcal{F}_c$ attaining the supremum in (1.5) is termed *Kantorovich potential*. As a particular example, for $\mathcal{X}=\mathcal{Y}$ and metric costs $c(x,y)=d(x,y)$, the double convexification trick leads to the *Kantorovich-Rubinstein* dual statement

$$\sup_{f\in\mathrm{Lip}_1(\mathcal{X})}\int_{\mathcal{X}} f(x)\,\mathrm{d}(\mu-\nu)(x), \tag{1.6}$$

where the supremum is taken over Lipschitz functions with Lipschitz modulus bounded by one (Villani, 2009, Part. Case 5.16). Duality goes far beyond the equality of optimal values for (1.3) and (1.4) (resp. (1.5)) and also serves to characterize the support of an OT plan related to the theory of *c-cyclical monotonicity*. Duality is then an important tool for regularity and stability properties of optimal solutions. Most importantly for this thesis, duality enables a suitable framework for a statistical analysis.

1.1.1 OT based Distances

Based on the Monge-Kantorovich OT problem (1.3), the *OT cost*

$$\mathcal{OT}_c(\mu,\nu):=\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi(x,y) \tag{1.7}$$

provides a notion of similarity between two probability measures μ,ν. In particular, suppose that both probability measures $\mu,\nu\in\mathcal{P}(\mathcal{X})$ are supported on the same complete, separable metric space. Then, the smaller the OT cost, the less far, on average, mass has to be transported from μ to ν. From a slightly different perspective, the transformation from μ into ν is less costly and suggests a degree of similarity inherent to both probability measures. A refined mathematical statement involves properties of the underlying cost function. Indeed, for a cost function equal to the metric d on $\mathcal{X}$ and any $p\geq 1$, the OT cost

$$\mathcal{W}_p(\mu,\nu):=\left\{\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{X}} d^p(x,y)\,\mathrm{d}\pi(x,y)\right\}^{1/p} \tag{1.8}$$

defines a finite metric on the space of probability measures

$$\mathcal{P}_p(\mathcal{X}):=\left\{\mu\in\mathcal{P}(\mathcal{X})\,\middle|\,\int_{\mathcal{X}} d^p(x,x_0)\,\mathrm{d}\mu(x)<+\infty\right\},$$

where the point $x_0 \in \mathcal{X}$ is arbitrary (Villani, 2009, Sec. 6). The metric space $\big(\mathcal{P}_p(\mathcal{X}), \mathcal{W}_p\big)$ is commonly referred to as the p-Wasserstein space and $\mathcal{W}_p$ as the p-Wasserstein metric[7]. If $\mathcal{X}$ is a Polish space, then the p-Wasserstein space is also Polish. Convergence of probability measures in the Wasserstein space is equivalent to weak convergence and convergence of their p-th metric moments (Villani, 2009, Ch. 6). Such convergence properties are an attractive tool for statistical purposes (Panaretos and Zemel, 2019). Further, the Wasserstein distance and related OT based similarity measures provide a flexible notion to compare general probability measures while incorporating the underlying ground metric structure. Notably, for two Dirac measures $\mu = \delta_x$ and $\nu = \delta_y$, their p-Wasserstein distance is equal to $W_p(\delta_x, \delta_y) = d(x, y)$. Therefore, these distances are well suited in applications where the importance of a metric is prevalent. Overall, these observations explain the recent advent of OT based similarity measures throughout various disciplines ranging from economics (Galichon, 2016) and finance (Rachev et al., 2011) to machine learning (Peyré and Cuturi, 2019) and biology (Schiebinger et al., 2019; Tameling et al., 2021).

1.1.2 OT on Finite Spaces

If the ground space $\mathcal{X} = \mathcal{Y} = \{x_1, \ldots, x_M\}$ is finite, probability measures are identified as elements of the probability simplex $\Delta(\mathcal{X}) := \big\{r \in \mathbb{R}_+^M \mid \sum_{i=1}^M r(x_i) = 1\big\}$. For a cost function $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and two probability measures $r, s \in \Delta(\mathcal{X})$, the OT in (1.3) takes the particular form[8]

$$\begin{aligned} \min_{\pi \in \mathbb{R}_+^{M\times M}} & \sum_{i,j=1}^{M} c(x_i, x_j)\pi(x_i, x_j) \\ \text{s.t.} & \sum_{j=1}^{M} \pi(x_i, x_j) = r(x_i), \quad \forall\, 1 \le i \le M, \\ & \sum_{i=1}^{M} \pi(x_i, x_j) = s(x_j), \quad \forall\, 1 \le j \le M. \end{aligned} \tag{1.9}$$

[7] The name is dedicated to Vaserstein (1969). However, the metric has been rediscovered several times, e.g., as Mallows distance (Mallows, 1972) in statistics or Earth Mover's distance (Rubner et al., 2000) in computer vision. As Vershik (2002) argues, a proper specification is Kantorovich metric.

[8] According to Birkhoff (1946), if $r \in \Delta(\mathcal{X})$ and $s \in \Delta(\mathcal{Y})$ are two uniform distributions on different finite spaces $\mathcal{X} \neq \mathcal{Y}$ of equal cardinality M, then OT is equivalent finding an optimal matching

$$\frac{1}{M} \min_{\sigma \in \mathcal{S}_M} \sum_{i=1}^{M} c\left(x_i, y_{\sigma(i)}\right),$$

where $\mathcal{S}_M$ denotes the set of all possible permutations of a set with M elements. In particular, the Monge formulation (1.2) and Monge-Kantorovich formulation (1.3) coincide.

Any transport plan π in (1.9) is represented by a $M \times M$ transport matrix with non-negative entries and row and column sums that match the probability measures r and s, respectively. The set of all feasible transport plans π for (1.9) is denoted by $\Pi(r, s)$. The *dual* problem for (1.9) is simply the discrete analogue for (1.4) and reads as

$$\begin{aligned} &\max_{f,g \in \mathbb{R}^M} \sum_{i=1}^{M} f(x_i)r(x_i) + \sum_{j=1}^{M} g(x_j)s(x_j) \\ &\text{s.t.}\, f(x_i) + g(x_j) \leq c(x_i, x_j), \quad \forall\, 1 \leq i, j \leq M. \end{aligned} \tag{1.10}$$

On finite spaces, the optimal objective values for (1.9) and (1.10) are always equal (strong duality) and existence of minimizers and maximizers is guaranteed.

1.2 Linear Programming

The general theory of linear programming was motivated by findings for the discrete formulation of OT in (1.9). Early contributions were given by Tolstoí (1930), Hitchcock (1941), Kantorovich (1942), Dantzig (1949) and Koopmans (1951). For a finite space $\mathcal{X} = \{x_1, \ldots, x_M\}$ and two probability measures $r, s \in \Delta(\mathcal{X})$ stacking the columns of a transport matrix $\pi \in \Pi(r, s)$ on top of each other defines a vector of length M^2. Further, for right-hand side parameter $b = (r, s)^T$ and $\lambda = (f, g)^T$, OT in (1.9) and its dual in (1.10) are restated as the pair of primal (P) and dual (D) linear programs

$$\begin{aligned} &\min_{\pi \in \mathbb{R}^{M^2}} c^T \pi \\ &\text{s.t.}\, A\pi = b, \pi \geq 0, \end{aligned} \tag{P}$$

$$\begin{aligned} &\max_{\lambda \in \mathbb{R}^{2M}} b^T \lambda \\ &\text{s.t.}\, A^T \lambda \leq c, \end{aligned} \tag{D}$$

respectively, with constraint matrix

$$A = \begin{bmatrix} \mathbb{1}_M^T \otimes \mathbb{I}_M \\ \mathbb{I}_M \otimes \mathbb{1}_M^T \end{bmatrix} \in \mathbb{R}^{2M \times M^2}. \tag{1.11}$$

In general, linear programming is concerned with optimization problems that exhibit a linear objective and whose constraints are given by linear (in)equalities. The primal (P) is a prototypical example of a *standard* linear program where the constraint matrix A is assumed to be of full rank[9]. The primal feasible set $\mathcal{M} := \{\pi \mid A\pi = b, \pi \geq 0\}$ and dual feasible set $\mathcal{N} := \left\{\lambda \mid A^T\lambda \leq c\right\}$ both define a *polyhedron*, respectively, that are assumed

[9] The constraint matrix A in (1.11) underlying OT is not of full rank. Deletion of a row of A and the corresponding entry in b leaves the polyhedron $\mathcal{M} = \{\pi \mid A\pi = b, \pi \geq 0\}$ invariant and in particular does not change the OT problem under considerations. For notational simplicity the discussion is based on the reduced matrix $A \in \mathbb{R}^{(2M-1)\times M^2}$ of full rank.

to be non-empty[10]. For a general polyhedron P an element $\lambda \in P$ is an extreme point if there do not exist two elements $\tau, \eta \in P$ both different to λ and a scalar $t \in (0,1)$ such that $\lambda = t\tau + (1-t)\eta$. Owing to the full rank assumption on A, the optimal solution for (P) (resp. (D)) is attained at an extreme point in $\mathcal{M}$ (resp. $\mathcal{N}$). Fundamental to linear programming is the identification of extreme points with *bases*. A basis $I \subset \{1, \ldots, M^2\}$ is an index set of cardinality $2M-1$ of the columns of A such that the *basis matrix* $A_I \in \mathbb{R}^{(2M-1)\times|I|}$, formed by the columns from A indexed by I, is one-to-one[11]. By full rank assumption on A there always exists a basis I that induces a *primal* and *dual basic solution*

$$\pi(I,b) := \mathrm{Aug}_I\left[(A_I)^{-1}b\right] \in \mathbb{R}^{M^2}, \quad \lambda(I) := (A_I)^{-T}c_I \in \mathbb{R}^{2M-1}, \tag{1.12}$$

respectively. In order to match dimensions, a solution for (P) has dimension M^2 instead of $2M-1$, the linear operator $\mathrm{Aug}_I \colon \mathbb{R}^{2M-1} \to \mathbb{R}^{M^2}$ augments zeroes in the coordinates that are not in I. A basis I determines the indices of possibly non-zero elements (basic variables) in a primal basic solution $\pi(I,b)$, and the binding constraints[12] for a dual basic solution $\lambda(I)$. All feasible basic solutions induced by a basis as in (1.12) are precisely the extreme points of the polyhedra $\mathcal{M}$ and $\mathcal{N}$, respectively. Notably, while the dual basic solution $\lambda(I)$ might be feasible (optimal), this might fail for the primal basic solution $\pi(I,b)$ and vice versa. Although each basis I uniquely induces a basic solution, several bases might lead to the same basic solution. The latter turns out to be a consequence of *degeneracy* for linear programming. A primal feasible basic solution $\pi(I,b)$ is degenerate if less than $2M-1$ of its coordinates are non-zero. A dual feasible basic solution $\lambda(I)$ is degenerate if more than $2M-1$ of the M^2 inequalities in $A^T\lambda(I) \leq c$ are binding. Figure 1.2 illustrates all of the mentioned statements. As the dual (D) is constrained to finitely many linear inequality, the cardinality of dual feasible bases is finite. For fixed right-hand side b in (P), the collection of dual feasible bases $\{I_1, \ldots, I_N\}$ can be partitioned according to their primal feasibility into

$$\pi(I_k,b) \text{ is feasible}, 1 \leq k \leq K; \qquad \pi(I_k,b) \text{ is infeasible}, \ K < k \leq N.$$

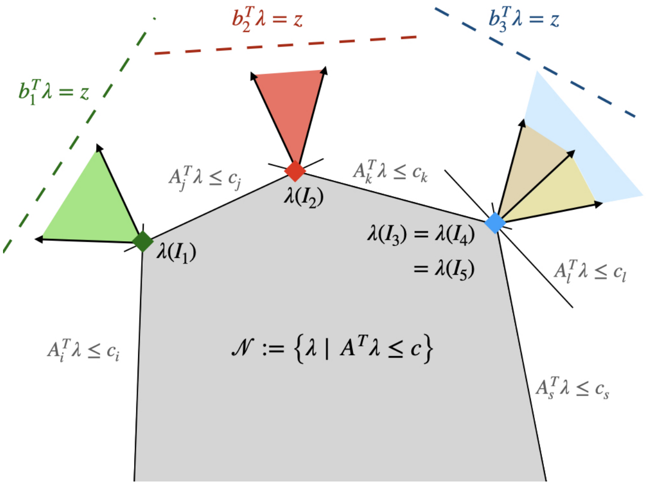
By strong duality, each basis I_k with $1 \leq k \leq K$ induces optimal primal and dual basic solutions in (1.12).

A nonempty, bounded polyhedron can be represented as the convex hull of its extreme

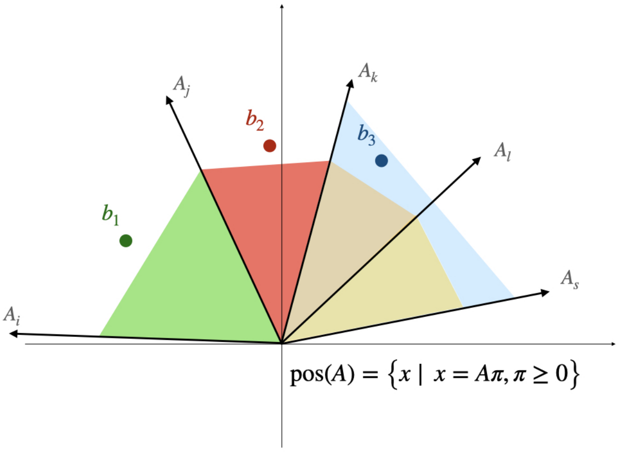
[10] If $\mathcal{M}$ and $\mathcal{N}$ are non-empty, then both (P) and (D) attain an optimal solution and by strong duality their corresponding optimal objective values are equal.

[11] For OT and measures $r, s \in \mathcal{P}(\mathcal{X})$ with $\mathcal{X} = \{x_1, \ldots, x_M\}$, a basis corresponds to a particular spanning tree that is a subset of the complete bipartite graph on $\{1, \ldots, M\} \times \{1, \ldots, M\}$.

[12] A constraint $A_j^T\lambda \leq c_j$ is said to be binding for λ_0 if it holds with equality $A_j^T\lambda_0 = c_j$.



(a) The dual feasible set $\mathcal{N}$ defines a polyhedron with extreme points (green, red and blue diamond shaped). The objective is to maximize the value of $z = b_t^T\lambda$ for $t = 1, 2, 3$. Each extreme point is identified by a basis, e.g., $\lambda(I_1)$ is uniquely identified by $I_1 = \{i, j\}$ and analogously for $\lambda(I_2)$ with $I_2 = \{j, k\}$. The blue extreme point is degenerate and identified by three bases $I_3 = \{k, l\}$, $I_4 = \{k, s\}$ and $I_5 = \{l, s\}$.



(b) The cone pos(A) spanned by all positive linear combinations of the columns of A. The cone is partitioned into sub-cones induced by the polar cones to feasible directions at dual extreme points. These sub-cones are particular choices of a basis index set, e.g., $I_2 = \{j, k\}$ corresponds to the red sub-cone spanned by column A_j and A_k. The right hand side parameter b_2 is contained in that sub-cone and hence I_2 identifies the feasible basic solution $\pi(I_2, b_2)$.

Figure 1.2: **The Geometry of Linear Programs.** For a right-hand side parameter b_1, the basis $I_1 = \{i, j\}$ is primal and dual feasible since $\lambda(I_1)$ is dual feasible (a) and b_1 falls into the sub-cone spanned by $\{A_i, A_j\}$ in (b). Hence, I_1 defines primal and dual optimal basic solutions. Analogous statements hold for b_2 and basis $I_2 = \{j, k\}$. For parameter b_3, the bases $I_3 = \{k, l\}$, $I_4 = \{k, s\}$ define primal and dual optimal basic solutions (a), whereas $I_5 = \{l, s\}$ corresponds to a dual feasible but primal infeasible basis (b).

points. If $\mathcal{M}$ is bounded, then the polyhedron of optimal solutions for (P) is also bounded and equal to the convex hull of the optimal basic solutions[13] $\{\pi(I_1, b), \dots, \pi(I_K, b)\}$ induced by $\{I_1, \dots, I_K\}$.

The correspondence between primal and dual basic solutions is at the heart of linear programming algorithms. The popular *simplex* algorithm by Dantzig (1949) is designed to move from one primal feasible basic solution to another, whilst improving the objective value and checking if the corresponding basis induces a dual feasible basic solution. In the worst case, the simplex algorithm takes an exponential number of iterations (Klee and Minty, 1972). However, the simplex algorithm is observed to have good practical performance and Spielman and Teng (2004) proved polynomial smoothed complexity for its average run-time if input data is randomly perturbed.

Apart from the simplex algorithm solving general linear programs, the particular linear programming structure for OT allows employing a zoo[14] of specialized solvers to compute optimal solutions. Among those are the Hungarian method (Kuhn, 1955), network flow solvers (Ford and Fulkerson, 1956), the transportation simplex (Luenberger and Ye, 1984) or the Auction algorithm (Bertsekas and Castanon, 1989). Typically, the computational complexity of these algorithms scales cubically in the support size of the probability measures (Burkard and Cela, 1999) leading to a severe limitation of OT for real-world applications.

1.3 Entropy Regularized Optimal Transport

With the recent dawn of the big-data era, solving OT for real-world instances is challenging as the high computational complexity hinders finding optimal solutions within reasonable time. Pursuing more attractive computational approaches has encouraged the development of various OT surrogates. Particular variants include the *sliced* (Bonnotte, 2013), *tree-sliced* (Le et al., 2019), *minibatch* (Fatras et al., 2021) or *sampled* OT (Sommerfeld et al., 2019).

The most prominent surrogate, popularized by Cuturi (2013), is based on entropy regularization. On a finite space $\mathcal{X} = \{x_1, \dots, x_M\}$ and measures $r, s \in \Delta(\mathcal{X})$, the negative entropy of a transport plan $\pi \in \Pi(r, s)$ is defined as

$$E(\pi) := \sum_{i,j=1}^{M} \pi(x_i, x_j)\Big(\log\big(\pi(x_i, x_j)\big) - 1\Big) + 1, \tag{1.13}$$

[13] This holds for any linear program in standard form and is trivially satisfied for OT.

[14] The development of efficient algorithms solving OT problems is dedicated an entire research field and more details are found in Peyré and Cuturi (2019).

with the usual convention $0\log(0) = 0$. The negative entropy is strictly convex and minimized among all transport plans in $\Pi(r, s)$ by the independence plan $r \otimes s = (r(x_i)s(x_j))_{i,j}^M$ (Kovačević et al., 2015). For some positive regularization parameter $\lambda > 0$ and a cost function $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the *entropy regularized* OT[15] is defined as

$$\min_{\pi \in \Pi(r,s)} \sum_{i,j=1}^{M} c(x_i, x_j)\pi(x_i, x_j) + \lambda E(\pi). \tag{1.14}$$

The set of transport plans $\Pi(r, s) \subset \mathbb{R}^{M\times M}$ is compact, and the objective function is continuous and strictly convex. Hence, there exists a unique optimal solution π^λ for (1.14). For a regularization parameter tending to zero, the optimal solution π^λ converges to the optimal solution for OT in (1.9) with maximal negative entropy. For large λ, the optimal solution tends to the independence plan $r \otimes s$. Figure 1.3 illustrates the effect of entropy regularization.

Entropy regularization promotes non-sparse OT plans with optimal solutions attained in the interior of the transportation simplex $\Pi(r, s)$, in strong contrast to non-regularized OT plans[16]. Indeed, the first order condition of the Lagrangian for (1.14) implies that the unique optimal solution is of the form

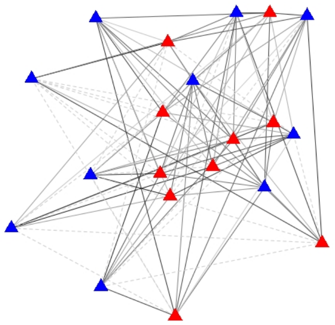
$$\pi^\lambda(x_i, x_j) = \exp\left(\frac{f_\lambda(x_i) + g_\lambda(x_j) - c(x_i, x_j)}{\lambda}\right), \quad \forall\, 1 \leq i, j \leq M \tag{1.15}$$

for two unknown functions (vectors) $f_\lambda, g_\lambda \in \mathbb{R}^M$. Together with the marginal requirement $\pi^\lambda \in \Pi(r, s)$, this leads to the *Sinkhorn* algorithm whose name is dedicated to Sinkhorn (1964) who first proved the convergence. The algorithm iteratively scales the kernel matrix $K := \left(-\frac{c(x_i,x_j)}{\lambda}\right)_{i,j}$ such that in each iteration either the row or column sum constraint is satisfied until convergence. Each iteration is of quadratic complexity and roughly requires $\lfloor \lambda^{-2} \rfloor$ iterations for reasonable approximation guarantees (Altschuler et al., 2017). In total, the computational complexity demonstrates the superior performance of the Sinkhorn algorithm compared to the cubic complexity inherent to non-regularized OT solvers and advocates the use of the entropy regularization for large scale applications.

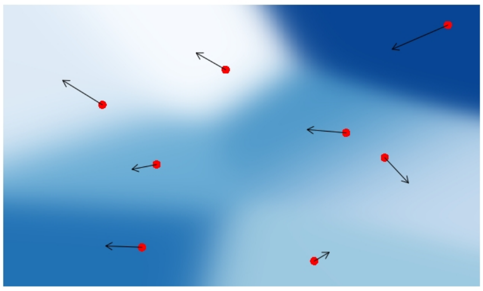
Further, as a strictly convex optimization problem with linear equality constraints, the

[15] In this thesis, the entropy regularized OT on finite spaces is of particular importance. There exists a formulation for general Polish spaces. For details please refer to (1.21) and Peyré and Cuturi (2019).

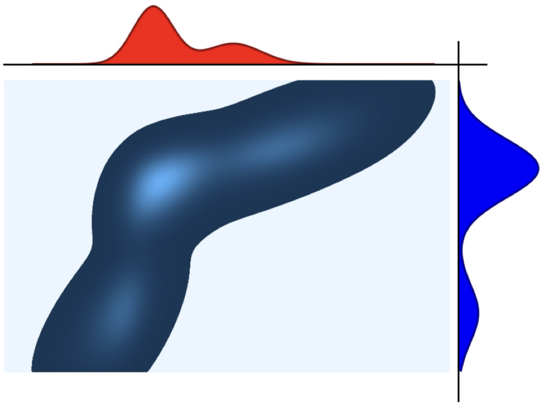
[16] There always exists an OT plan for (1.9) with at most $2M - 1$ positive entries. A simple argument employs the discussion on bases outlined in Section 1.2 and the fact that there exists an OT plan induced by a basis I with cardinality $|I| = 2M - 1$.



(a) Discrete entropy regularized OT between two uniform distributions supported on ten points in red and blue, respectively. Thicker connections correspond to more mass transportation. In contrast to Figure 1.1a, the optimal solution is not equal to a matching and promotes mass transportation from every red to each blue point.



(b) Semi-discrete entropy regularized OT between a uniform distribution supported on eight points (red) and a uniform distribution on $[0, 1]^2$. Compared to Figure 1.1b, the optimal solution is a blurry Laguerre-Voronoi tessellation of the ground space.



(c) Continuous entropy OT between two absolutely continuous distributions with densities on the real line. Different to Figure 1.1c, the optimal solution is far from being a map and smears the transportation over the product space.

Figure 1.3: **Entropy Optimal Transport.** The Entropy regularized OT problem (1.14) with squared Euclidean cost $c(x, y) = \|x - y\|^2$.

entropy regularized OT in (1.14) admits the dual formulation

$$\max_{f,g\in\mathbb{R}^M} \sum_{i=1}^{M} f(x_i)r(x_i) + \sum_{j=1}^{M} g(x_j)s(x_j) - \lambda \sum_{i,j=1}^{M} \left(\exp\left(\frac{f(x_i)+g(x_j)-c(x_i,x_j)}{\lambda}\right) - 1\right). \tag{1.16}$$

For the unconstrained maximization problem (1.16) exists a unique (up to constant shifts) pair of dual optimal solutions (f_λ, g_λ) and strong duality holds, i.e., the optimal objective values for (1.14) and (1.16) are equal. In fact, the optimal pair (f_λ, g_λ) fulfills (1.15) and the dual solutions are recovered from the Sinkhorn algorithm[17].

1.3.1 Entropy Regularized OT based Divergences

Entropy regularized OT defines an alternative notion to measure the similarity between probability measures $r, s \in \Delta(\mathcal{X})$ with $\mathcal{X} = \{x_1, \ldots, x_M\}$. The first approach is to employ the optimal value for (1.14) and define

$$\mathcal{OT}_{c,\lambda}(r,s) := \min_{\pi\in\Pi(r,s)} \sum_{i,j=1}^{M} c(x_i,x_j)\pi(x_i,x_j) + \lambda E(\pi). \tag{1.17}$$

Similarly, the *Sinkhorn divergence* is based on evaluation of the transport cost for the optimal solution π^λ for (1.14), i.e.,

$$\mathcal{SD}_{c,\lambda}(r,s) := \sum_{i,j=1}^{M} c(x_i,x_j)\pi^\lambda(x_i,x_j). \tag{1.18}$$

Even if the cost function is induced from an underlying metric $c(x_i,x_j) = d(x_i,x_j)$, neither the optimal value $\mathcal{OT}_{c,\lambda}(r,s)$ nor the Sinkhorn divergence $\mathcal{SD}_{c,\lambda}(r,s)$ define a true notion of a distance on the set of probability measures $\Delta(\mathcal{X})$. In particular, for $r = s$ both quantities are not equal to zero. The recently proposed Sinkhorn *loss* as introduced by Genevay et al. (2018),

$$\mathcal{SL}_{c,\lambda}(r,s) := \mathcal{OT}_{c,\lambda}(r,s) - \frac{1}{2}\left(\mathcal{OT}_{c,\lambda}(r,r) + \mathcal{OT}_{c,\lambda}(s,s)\right), \tag{1.19}$$

accounts for the identity of indiscernible. Still, even for underlying metric costs, the Sinkhorn loss does not fulfill the triangle inequality. However, the definitions (1.17), (1.18) and (1.19) serve as a notion of *divergence* between probability measures and converge to the OT cost in (1.7) as the regularization parameter tends to zero. Moreover,

[17] Optimization of the dual problem (1.16) might be carried out by a block coordinate ascent keeping one variable fixed while optimizing over the other and vice versa. These iterations are equivalent to the Sinkhorn algorithm.

the computational accessibility together with differentiability properties shape these divergences as noteworthy surrogates. Applications range from incorporation in variational OT problems such as OT barycenters (Cuturi and Doucet, 2014) to defining an alternative loss function in machine learning (Genevay et al., 2018) and many more are found in Peyré and Cuturi (2019).

The entropy regularized OT might be defined between probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ supported on general Polish spaces $\mathcal{X}$. To this end, the entropy $E(\pi)$ in (1.13) is replaced by the *Kullback-Leibler divergence* (also known as mutual information or relative entropy)[18]

$$\mathrm{KL}(\mu \,\|\, \nu) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\mu(x) & \text{if } \mu \ll \nu, \\ +\infty & \text{else.} \end{cases} \tag{1.20}$$

For a cost function $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and regularization parameter $\lambda > 0$, the entropy regularized OT is then defined as the strictly convex optimization problem

$$\mathcal{EOT}_{c,\lambda}(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y)\, \mathrm{d}\pi(x, y) + \lambda \mathrm{KL}(\pi \,\|\, \mu \otimes \nu). \tag{1.21}$$

Under regularity assumption on the cost function and the probability measures, the entropy regularized OT (1.21) admits the dual formulation (Chizat et al., 2018)

$$\begin{aligned} \mathcal{EOT}_{c,\lambda}(\mu, \nu) = \sup_{\substack{f \in L_1(\mu) \\ g \in L_1(\nu)}} & \int_{\mathcal{X}} f(x)\, \mathrm{d}\mu(x) + \int_{\mathcal{X}} g(y)\, \mathrm{d}\nu(y) \\ & -\lambda \int_{\mathcal{X}\times\mathcal{X}} \exp\left(\frac{f(x) + g(y) - c(x, y)}{\lambda}\right) \mathrm{d}\mu(x)\, \mathrm{d}\nu(y). \end{aligned} \tag{1.22}$$

Analogously as to the finite setting in (1.14), any pair of dual optimal solutions for (1.22) is denoted as (f_λ, g_λ).

[18] Already for countable spaces $\mathcal{X} = \{x_1, x_2, \dots\}$ and probability measures $r, s \in \mathcal{P}(\mathcal{X})$, for any coupling $\pi \in \Pi(r, s)$ its entropy $E(\pi)$ is finite only if r and s have a finite entropy (Cover, 1999). In particular, the optimization problem (1.14), if extended to countable spaces, is only finite on a particular subset of $\mathcal{P}(\mathcal{X})$. Contrary, $\mathrm{KL}(\pi \,\|\, r \otimes s)$ is finite at least for $\pi = r \otimes s \in \Pi(r, s)$.

CHAPTER 2

Limit Laws for Empirical (Regularized) OT on Finite Spaces

This chapter deals with distributional limit laws for empirical OT and empirical entropy regularized OT quantities defined on finite spaces. The main results presented in Section 2.1 subsume the two articles Klatt et al. (2020a) and Klatt et al. (2020b). To this end, let $\mathcal{X} = \{x_1, \dots, x_M\}$ be a finite discrete space equipped with a metric $d\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. There exist two probability measures $r, s \in \Delta(\mathcal{X})$ of full support (possibly $r = s$) but only s is completely known[1]. The probability measure r is unknown and needs to be estimated. Concerning the estimation of r, it is assumed that there is access to a collection of independent and identically distributed (i.i.d.) $\mathcal{X}$-valued random variables $X_1, \dots, X_n \sim r$. The standard estimator for r is the empirical measure $\hat{r}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$. By the strong law of large numbers and increasing sample size n, the estimator $\hat{r}_n$ converges to r almost surely. A refined statement follows from the multivariate central limit theorem

$$\sqrt{n}\,(\hat{r}_n - r) \xrightarrow{\mathcal{D}} \mathbb{G}_r, \tag{2.1}$$

as the sample size n tends to infinity. The limiting random variable $\mathbb{G}_r$ follows a M-dimensional centered Gaussian distribution $\mathcal{N}_M\,(0, \Sigma(r))$ with covariance matrix

$$\Sigma(r) = \begin{cases} -r(x)r(x') & \text{if } x \neq x', \\ r(x)(1 - r(x)) & \text{else.} \end{cases} \tag{2.2}$$

Within this prescribed setup, the random quantities of interest in this chapter are plug-in estimators for OT and entropy regularized OT based quantities as defined in the

[1] To streamline the presentation of the two articles Klatt et al. (2020a,b), the main results are tailored to the one-sample case, where $r \in \Delta(\mathcal{X})$ is unknown while $s \in \Delta(\mathcal{X})$ is assumed to be known. Both articles also contain statements for the two-sample case in which both probability measures are unknown and hence need to be estimated.

Introduction:

- The empirical OT plans (Section 1.1.2):

$$\hat{\pi}_n \in \operatorname*{arg\,min}_{\pi\in\Pi(\hat{r}_n,s)} \sum_{i,j=1}^{M} d(x_i,x_j)\pi(x_i,x_j) \tag{2.3}$$

- The empirical entropy regularized OT plan (Section 1.3):

$$\hat{\pi}_n^\lambda = \operatorname*{arg\,min}_{\pi\in\Pi(\hat{r}_n,s)} \sum_{i,j=1}^{M} d(x_i,x_j)\pi(x_i,x_j) + \lambda E(\pi) \tag{2.4}$$

- The empirical objective value for entropy regularized OT in (1.17):

$$\mathcal{OT}_\lambda(\hat{r}_n,s) = \min_{\pi\in\Pi(\hat{r}_n,s)} \sum_{i,j=1}^{M} d(x_i,x_j)\pi(x_i,x_j) + \lambda E(\pi) \tag{2.5}$$

- The empirical Sinkhorn divergence in (1.18):

$$\mathcal{SD}_\lambda(\hat{r}_n,s) := \sum_{i,j=1}^{M} d(x_i,x_j)\hat{\pi}_n^\lambda(x_i,x_j) \tag{2.6}$$

- The empirical Sinkhorn loss in (1.19):

$$\mathcal{SL}_\lambda(\hat{r}_n,s) := \mathcal{OT}_\lambda(\hat{r}_n,s) - \frac{1}{2}\left(\mathcal{OT}_\lambda(\hat{r}_n,\hat{r}_n) + \mathcal{OT}_\lambda(s,s)\right) \tag{2.7}$$

The primary purpose of the two articles Klatt et al. (2020a,b) is to quantify the asymptotic fluctuation of the empirical plug-in estimators (2.3) – (2.7) around their respective population quantities (after proper scaling) by a central limit theorem.

Some remarks concerning the presentation of the main results are in order: First, the statistical results to follow remain valid (upto minor changes) if the metric d inherent in (2.3) – (2.7) is replaced by a general cost function $c\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}_+$. Second, the article Klatt et al. (2020a) contains more general results concerning distributional limit laws for empirical solutions to linear programs in standard form. OT is nothing but a special case of this theory (cf. Section 1.2). Third, the article Klatt et al. (2020b) considers various regularization methods for OT. The statistical theory for entropy regularization of OT is again a particular case. Lastly, and most importantly, some results require the notion of a *reduction*. In particular, any probability measure $r \in \Delta(\mathcal{X})$ is identified by a vector of dimension M. The associated probability measure $r_\dagger$ is equal to r except with the last coordinate removed, hence of dimension $M-1$. The reduced perspective $r_\dagger$ serves two important purposes: On the one hand, it removes the *rank deficiency* inherent in the definition of OT couplings that now might equivalently be defined as M^2 dimensional vectors

$$\Pi(r_\dagger, s) := \left\{ \pi \in \mathbb{R}^{M^2} \,\middle|\, A_\dagger \pi = \begin{bmatrix} r_\dagger \\ s \end{bmatrix}, \pi \geq 0 \right\} \tag{2.8}$$

with constraint matrix $A_\dagger$ of full rank (see Section 1.2 and Footnote 9 on p. 8). On the other hand, the related central limit theorem in (2.1) for $\sqrt{n}\left(\hat{r}_{\dagger n} - r_\dagger\right) \xrightarrow{\mathcal{D}} \mathbb{G}_{r_\dagger} \sim \mathcal{N}_{M-1}(0, \Sigma(r_\dagger))$ leads to an asymptotic centered Gaussian distribution that is *absolutely continuous* with respect to Lebesgue measure. Both observations alleviate the underlying proof techniques employed in the two articles.

2.1 Main Results

Klatt et al. (2020a): The exposition of the main results starts with the article Klatt et al. (2020a) and distributional limit laws for empirical OT plans (2.3). For fixed probability measures $r, s \in \Delta(\mathcal{X})$, the set of optimal solutions is denoted by

$$\Pi^\star(r, s) := \left\{ \pi \in \Pi(r, s) \,\middle|\, \pi \in \underset{\pi \in \Pi(r,s)}{\arg\min} \sum_{i,j=1}^{M} d(x_i, x_j)\pi(x_i, x_j) \right\}. \tag{2.9}$$

Suppose that $I_1, \ldots, I_K$ are all dual feasible bases that induce primal feasible and hence optimal basic solutions (see Section 1.2 for details). For each $1 \leq k \leq K$, denote by $\mathrm{DP}_k := \{i \in I_k \mid \pi_i(I_k, [r_\dagger, s]) = 0\}$ the collection of those indices in basis I_k for which the particular basic variables are equal to zero. For each basis I_k there exists a closed and convex *cone* of feasible perturbations[2]

$$H_k := \left\{ v \in \mathbb{R}^{M-1} \,\middle|\, \pi_i(I_k, [v, 0_M]) \geq 0 \,\forall\, i \in \mathrm{DP}_k \right\}. \tag{2.10}$$

More precisely, to each $v \in H_k$ there exists a scalar $t > 0$ such that the basis I_k remains primal and dual optimal for the standard linear program (P) with parameter $b = (r_\dagger, s)^T$ replaced by $\tilde{b} = (r_\dagger, s)^T + t\,(v, 0_M)^T$.

The main result in Klatt et al. (2020a) states the following distributional limit law for the empirical OT plans in (2.3): There exists a random sequence $(\pi_n^\star)_{n\in\mathbb{N}} \subset \Pi^\star(r, s)$ of

[2] From a linear programming perspective the set DP_k encodes degeneracy: The optimal primal basic solution $\pi(I_k, [r_\dagger, s])$ is degenerate if and only if the set $\mathrm{DP}_k \neq \emptyset$. If $\pi(I_k, [r_\dagger, s])$ is non-degenerate, then $\mathrm{DP}_k = \emptyset$ and the corresponding cone H_k in (2.10) is equal to $\mathbb{R}^{M-1}$.

optimal OT plans such that if the sample size n tends to infinity, then

$$\boxed{\sqrt{n}\,(\hat{\pi}_n - \pi_n^\star) \xrightarrow{\mathcal{D}} M\left(\mathbb{G}_{r_\dagger}\right) := \sum_{\mathcal{K}} \mathbb{1}_{\left\{\mathbb{G}_{r_\dagger}\in\left\{H_{\mathcal{K}}\setminus\cup_{k\notin\mathcal{K}}H_k\right\}\right\}}\, \alpha^{\mathcal{K}} \otimes \pi\left(I_{\mathcal{K}}, [\mathbb{G}_{r_\dagger}, 0_M]\right).} \tag{2.11}$$

The sum in the limiting random variable $M\left(\mathbb{G}_{r_\dagger}\right)$ runs over non-empty subsets $\mathcal{K}$ of $\{1,\dots,K\}$, with $H_{\mathcal{K}} := \cap_{k\in\mathcal{K}}H_k$ and $\alpha^{\mathcal{K}}$ is a random vector in the (essentially $|\mathcal{K}|$-dimensional)[3] unit simplex $\Delta_{\mathcal{K}} := \left\{\alpha \in \mathbb{R}_+^N \,\middle|\, \|\alpha\|_1 = 1,\, \alpha_k = 0\,\forall\, k \notin \mathcal{K}\right\}$.

Based on the general distributional limit theorem in (2.11) and under certain *non-degeneracy* assumptions, various simplifications of the limiting random variable follow. A particular setting arises if the set of all dual basic solutions is non-degenerate. Within the OT context, dual non-degeneracy is satisfied if and only if[4] for the dual optimal basic solutions $\lambda_{\dagger 1},\dots,\lambda_{\dagger K}$ (recall the notation $\lambda_\dagger = (f_\dagger, g)^T$ for f, g dual optimal solutions for (D)) it holds that

$$\lambda(I_j)_\dagger \neq \lambda(I_k)_\dagger,\; 1 \leq j < k \leq K. \tag{2.12}$$

Under assumption (2.12) the primal optimal basic solution is unique $\Pi^\star(r,s) = \{\pi^\star\}$. Together with the absolute continuity of the limit law $\mathbb{G}_{r_\dagger}$ for the reduced multinomial process $\sqrt{n}(\hat{r}_{\dagger n} - r_\dagger)$, the corresponding distributional limit theorem in (2.11) simplifies to

$$\boxed{\sqrt{n}\,(\hat{\pi}_n - \pi^\star) \xrightarrow{\mathcal{D}} \sum_{k=1}^{K} \mathbb{1}_{\left\{\mathbb{G}_{r_\dagger}\in H_k\right\}}\, \pi\left(I_k, [\mathbb{G}_{r_\dagger}, 0_M]\right).} \tag{2.13}$$

A sufficient condition for dual non-degeneracy in (2.12) and therefore verification of the statement in (2.13) is given in terms of the underlying metric: If for any $l \geq 2$ and any family of points $\left\{(x_{i_k}, x_{j_k})\right\}_{1\leq k\leq l} \subset \mathcal{X}\times\mathcal{X}$ with all x_{i_k} and all x_{j_k} pairwise different it holds that

$$\sum_{k=1}^{l} d\left(x_{i_k}, x_{j_k}\right) \neq \sum_{k=1}^{l} d\left(x_{i_k}, x_{j_{k-1}}\right), \quad x_{j_0} := x_{j_l}, \tag{2.14}$$

then every dual feasible basic solution is non-degenerate and (2.12) is satisfied. Condition (2.14) fails to hold if the underlying ground space $\mathcal{X}$ is sufficiently symmetric. A prototypical example is a regular grid with underlying grid metric for which (2.14) is usually never satisfied[5].

If additionally to dual non-degeneracy (2.12), also the unique primal optimal basic

[3] The set of all dual feasible bases is assumed to be of cardinality N with $K \leq N$ (see Section 1.2).

[4] For general linear programs, the characterization for non-degeneracy in terms of condition (2.12) requires the set of optimal primal solutions to be non-empty and bounded. For OT this is trivially satisfied.

[5] If $\mathcal{X} \subset \mathbb{R}$ is a finite subset on the real line, then the non-degeneracy condition (2.14) is satisfied for a cost function $c(x,y) = h(|x-y|)$ with h strictly convex or strictly concave and $h(0) = 0$. It follows that (2.12) holds for $c(x,y) = |x-y|^p$ and $p \in (0,\infty)$, $p \neq 1$.

solution is non-degenerate, then primal and dual optimal solutions are unique and there exists a unique basis I_1 that induces them. The distributional limit theorem for the empirical OT plan then takes the simple form

$$\boxed{\sqrt{n}\,(\hat{\pi}_n - \pi^\star) \xrightarrow{\mathcal{D}} \pi\left(I_1, [\mathbb{G}_{r_\dagger}, 0_M]\right).} \tag{2.15}$$

The limiting random variable in (2.15) is equal to the linear transformation $\left(A_{\dagger I_1}\right)^{-1}$ with unique basis I_1 of the augmented limiting random variable $[\mathbb{G}_{r_\dagger}, 0_M]$. The non-negative entries jointly follow a centered Gaussian distribution with covariance matrix equal to $(A_\dagger)_{I_1}^{-1}\,\Sigma(r_\dagger)\,(A_\dagger)_{I_1}^{-T}$. A sufficient condition for primal non-degeneracy is given in terms of the probability measures r and s: If all proper subsets $A, B \subset \mathcal{X}$ with $A \neq B$ are assigned different total mass

$$\sum_{x\in A} r(x) \neq \sum_{x\in B} s(x), \tag{2.16}$$

then every primal feasible basic solution is non-degenerate (Klee and Witzgall, 1968). Overall, distributional limit theorems for empirical OT plans on finite spaces are driven by the geometry of the boundary of the underlying transportation simplex $\Pi(r, s)$ and the amount of degeneracy present in the primal (P) and dual (D) linear programs. The result (2.11) holds generally and reflects the complex geometric and combinatorial nature for degenerate linear programs. If dual optimal basic solutions are non-degenerate, then the simplified statement (2.13) follows. Both results (2.11) and (2.13) hold for the *alternative* $r \neq s$ as well as for the *null* $r = s$. These two different regimes are reflected in the number of cones (2.10) and summands defining the limiting random variable. The null $r = s$ is the most challenging case since the primal optimal basic solution is highly degenerate. More precisely, the transport plan $\pi \in \mathbb{R}^{M\times M}$ is supported on the diagonal with only M instead of $2M - 1$ non-zero entries. For an illustration of the limit law in such a case, please refer to Klatt et al. (2020b, Example 6.3).

In case all primal and dual optimal basic solutions are non-degenerate, the central limit theorem (2.15) is valid. The corresponding non-negative entries jointly follow a Gaussian law. Sufficient conditions for non-degeneracy are given in (2.14) for the dual and in (2.16) for the primal and suggest certain non-regularity either in the probability measures or in the underlying cost function. Under small perturbations of the metric cost and the probability measures $r, s \in \Delta(\mathcal{X})$, the conditions (2.14) and (2.16) are usually satisfied (see Sommerfeld (2017); Klatt et al. (2020b) for details). Nonetheless, degeneracy in linear programs is generic rather than the exception for most practical situations (Bertsimas and Tsitsiklis, 1997).

Klatt et al. (2020b): The article Klatt et al. (2020b) focuses on central limit theorems for empirical entropy regularized OT quantities. The entropy regularized OT (1.14) is a strictly convex optimization problem with a unique optimal solution π^λ attained in the interior of the transportation simplex $\Pi(r, s)$. An application of standard methods from sensitivity and stability analysis in *non-linear* programming (Fiacco, 1983) proves the solution $\pi^\lambda = \varphi_\lambda(r_\dagger, s)$ to be implicitly parameterized by some smooth function φ defined on $\Delta(\mathcal{X})_\dagger \times \Delta(\mathcal{X})$. The derivative of φ might be computed explicitly

$$\nabla_{(r_\dagger, s)}\varphi_\lambda = D\left(\pi^\lambda\right) A_\dagger \left[A_\dagger\, D\left(\pi^\lambda\right) A_\dagger^T\right]^{-1} \in \mathbb{R}^{M^2\times(2N-1)} \tag{2.17}$$

with constraint matrix $A_\dagger$ in (2.8) and $D(\pi^\lambda) \in \mathbb{R}^{M^2\times M^2}$ a diagonal matrix with diagonal[6] equal to π^λ. Central limit theorems then follow as a consequence of the delta method: For two probability measures $r, s \in \Delta(\mathcal{X})$, the empirical entropy regularized transport plan $\hat{\pi}_n^\lambda$ in (2.4), estimated from the empirical measure $\hat{r}_n$, meets the central limit theorem

$$\boxed{\sqrt{n}\left(\hat{\pi}_n^\lambda - \pi^\lambda\right) \xrightarrow{\mathcal{D}} Z_{\pi^\lambda} \sim \mathcal{N}_{M^2}\left(0, \Sigma_\lambda(r \mid s)\right),} \tag{2.18}$$

as the sample size n tends to infinity. The limiting random variable Z_{π^λ} follows a M^2-dimensional centered Gaussian distribution. The covariance matrix is equal to $\Sigma_\lambda(r \mid s) = [\nabla_{r_\dagger}\varphi_\lambda]\, \Sigma(r_\dagger)\, [\nabla_{r_\dagger}\varphi_\lambda]^T$ and specifically depends on the unique solution π^λ. The central limit theorem (2.18) is valid and non-degenerate for the *null* $r = s$ as well as for the *alternative* $r \neq s$.

The general asymptotic analysis for the empirical entropy regularized OT plan in (2.18) enables a straightforward investigation of central limit theorems for the corresponding empirical entropy divergences in (2.5), (2.6) and (2.7). In particular, all divergences are smooth functions of the empirical entropy regularized OT plan. Their asymptotic statements follow by another application of the delta method. The scaling rate $\sqrt{n}$ in (2.18) is passed to the corresponding central limit theorems and the limiting random variables always follow a certain centered Gaussian law.

The empirical Sinkhorn divergence $\mathcal{SD}_\lambda(\hat{r}_n, s)$ in (2.6) is a linear function of the empirical entropy regularized OT plan and hence the simplest random object in this regard. The corresponding central limit theorem states that if the sample size n tends to infinity,

[6] The article Klatt et al. (2020b) considers the entropy plan π^λ as a vector of dimension M^2 rather than a matrix of dimension $M \times M$. Moreover, the main results in the article consider general *proper* regularization functions f replacing the entropy E in (1.14). Then, the diagonal matrix $D\left(\pi^\lambda\right)$ in (2.17) is replaced by the inverse of the Hessian $(\nabla_\pi^2 f)^{-1}$ for the regularization function f. For entropy E, the inverse of its Hessian is exactly of described diagonal form $D(\pi^\lambda)$.

then

$$\boxed{\sqrt{n}\left(\mathcal{SD}_\lambda(\hat{r}_n, s) - \mathcal{SD}_\lambda(r, s)\right) \xrightarrow{\mathcal{D}} \langle Z_{\pi^\lambda}, d\rangle \sim \mathcal{N}\left(0, \sigma_\lambda^2(r \mid s)\right)} \tag{2.19}$$

with d the vectorized version of the underlying metric on $\mathcal{X}$. The limiting random variable $\langle Z_{\pi^\lambda}, d\rangle$ follows a centered Gaussian distribution with variance equal to $\sigma_\lambda^2(r \mid s) = d^T \Sigma_\lambda(r \mid s) d^T$. The central limit law (2.19) holds for the null $r = s$ and the alternative $r \neq s$. Notably, even for $r = s$ the central limit theorem for the empirical quantity $\mathcal{SD}_\lambda(\hat{r}_n, r)$ is not centered around zero as the optimal objective value $\mathcal{SD}_\lambda(r, r)$ is strictly positive. A similar asymptotic behavior is obtained for the empirical objective value $\mathcal{OT}_\lambda(\hat{r}_n, s)$ in (2.5). If the sample size n tends to infinity, then

$$\boxed{\sqrt{n}\left(\mathcal{OT}_\lambda(\hat{r}_n, s) - \mathcal{OT}_\lambda(r, s)\right) \xrightarrow{\mathcal{D}} \left\langle \mathbb{G}_r, f_\lambda^{(r,s)}\right\rangle \sim \mathcal{N}\left(0, \tilde{\sigma}_\lambda^2(r \mid s)\right).} \tag{2.20}$$

The limiting random variable $\left\langle \mathbb{G}_r, f_\lambda^{(r,s)}\right\rangle$ is equal to the scalar product between the limiting random variable $\mathbb{G}_r$ from (2.1) and the optimal solution $f_\lambda^{(r,s)}$ for the dual of the entropy regularized OT (1.16). The asymptotic distribution is equal to a centered Gaussian distribution with variance equal to $\tilde{\sigma}_\lambda^2(r \mid s) = \mathrm{Var}_{X\sim r}[f_\lambda^{(r,s)}(X)]$. The central limit law (2.20) holds for the null $r = s$ and the alternative $r \neq s$ and degenerates to a Dirac at zero if and only if $f_\lambda^{(r,s)}$ is constant[7].

Analogous central limit laws for the empirical Sinkhorn loss (2.7) state for sample size n tending to infinity that

$$\boxed{\sqrt{n}\left(\mathcal{SL}_\lambda(\hat{r}_n, s) - \mathcal{SL}_\lambda(r, s)\right) \xrightarrow{\mathcal{D}} \left\langle \mathbb{G}_r, f_\lambda^{(r,s)} - \frac{1}{2}\left(f_\lambda^{(r,r)} + g_\lambda^{(r,r)}\right)\right\rangle.} \tag{2.21}$$

Compared to (2.20), the additional term $\frac{1}{2}(f_\lambda^{(r,r)} + g_\lambda^{(r,r)})$ in the limiting random variable accounts for the randomness in $\mathcal{SL}_\lambda(\hat{r}_n, \hat{r}_n)$ inherent in the definition of the empirical Sinkhorn loss (2.7). The limiting random variable follows a centered Gaussian distribution. The variance of the limit distribution is equal to $\mathrm{Var}_{X\sim r}[f_\lambda^{(r,s)}(X) - \frac{1}{2}(f_\lambda^{(r,r)}(X) - g_\lambda^{(r,r)}(X))]$. Herein, $f_\lambda^{(r,s)}$ is the dual optimal solution for $\mathcal{OT}_\lambda(r, s)$ and $(f_\lambda^{(r,r)}, g_\lambda^{(r,r)})$ is the pair of dual optimal solutions for $\mathcal{OT}_\lambda(r, r)$. Under the null $r = s$, the central limit law in (2.21) degenerates $\sqrt{n}\,\mathcal{SL}_\lambda(\hat{r}_n, r) \xrightarrow{\mathcal{D}} \delta_0$ to a Dirac at zero and please refer to the following section and (2.32) for a thorough discussion.

Overall, central limit theorems for empirical entropy regularized OT quantities defined on finite spaces reveal limiting random variables following a centered Gaussian limit law. In particular, these asymptotic results demonstrate a substantially different

[7] The variance $\mathrm{Var}_{X\sim r}[f_\lambda^{(r,s)}(X)]$ is equal to zero if and only if $f_\lambda^{(r,s)}(X)$ is constant. This is verified by considerations of the covariance matrix $\Sigma(r)$ in (2.2) with kernel $\ker(\Sigma(r)) = \{v \in \mathbb{R}^M \mid v = a\mathbb{1}_M,\ a \in \mathbb{R}\}$.

statistical behavior when compared to their non-regularized OT counterparts. At the heart of this phenomenon is the entropy function that renders the optimization problem (1.14) strictly convex and sufficiently smooth in contrast to the linear formulation of OT in (1.9). Further, entropy regularization modifies the unique optimal solution π^λ non-degenerate in the sense that π^λ belongs to the interior of the transportation simplex $\Pi(r, s)$ with all its coordinates strictly positive. From a statistical perspective and the fact that the underlying multinomial process of empirical frequencies asymptotically follows a Gaussian law (2.1), the smoothness translates via the delta method into Gaussian limit laws. All central limit laws for empirical entropy OT quantities are usually non-degenerate for both the null $r = s$ and the alternative $r \neq s$, with the notable exception of the empirical Sinkhorn loss under the null.

2.2 Discussion and Related Work

Klatt et al. (2020a): The asymptotic analysis of empirical OT plans is strongly related to an analysis of linear programs with random parameters. Indeed, the latter is a topic covered by a large body of literature. Shortly after Dantzig (1949) proposed the simplex algorithm, the need to deal with random input data was inevitable. For a linear program in standard form

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} c^T x \\ &\text{s.t.}\, Ax = b,\ x \geq 0, \end{aligned} \tag{2.22}$$

the parameters $(b, c) \in \mathbb{R}^m \times \mathbb{R}^n$ usually model practical needs such as budget, prices or capacities[8] that are often not available exactly but estimated. Early contributions dealing with uncertain parameters are the work by Dantzig (1955), Beale (1955) Ferguson and Dantzig (1956) concerned with feasible computational approaches.

The *distribution problem* (Wets, 1980) is among the first that concerns statistical questions relevant for this thesis. Herein, the standard linear program (2.22) is considered as the function $b \mapsto O(b)$ that maps the parameter b to its corresponding objective value $O(b)$ equal to the optimal value for (2.22). If b is a random variable following some known distribution μ, then the distribution problem seeks to characterize the push-forward measure $O_{\#}\mu$. Various contributions rely on so-called *decision regions* from sensitivity analysis that is detailled below. In particular, a careful integration over the collection of all these regions leads to a characterization of the cumulative distribution function for the push-forward measure $O_{\#}\mu$ (see Ewbank et al. (1974) and references therein).

[8] For OT, the parameter c models the underlying transport cost, whereas the vector $b = (r, s)^T$ contains the two probability distributions $r, s \in \Delta(\mathcal{X})$.

Moreover, modeling uncertainty in linear programs has opened a wealth of approaches subsumed in the theory of stochastic programming. Typical models deviate from the standard linear program (2.22) with random parameters, but include *chance constrained*, *two-* or *multiple-stage* linear programs (Shapiro et al., 2021). For instance, a simple two stage linear recourse model considers the problem

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} c^T x + \mathbb{E}_\mu \left[Q(x, \xi) \right] \\ &\text{s.t.}\, Ax = b,\ x \geq 0, \end{aligned} \tag{2.23}$$

where $x \mapsto \mathbb{E}_\mu \left[Q(x, \xi) \right]$ denotes the recourse function defined as

$$Q(x, \xi) = \inf_z \left\{ q(\xi)^T z \,\middle|\, W(\xi) z = d(\xi) + T(\xi) x,\ z \geq 0 \right\}.$$

Herein, $\xi \sim \mu$ is a random vector with components $(q(\xi), W(\xi), d(\xi), T(\xi))$ that are componentwise either random matrices or vectors, respectively. The primary purpose of such a model is to find an optimal decision x in a first stage such that the decision minimizes the loss (maximizes the profit) $\mathbb{E}_\mu \left[Q(x, \xi) \right]$ depending on a specific realization of the random variable ξ in a second stage. A classical example is the *news vendor problem* (Birge and Louveaux, 2011). A statistical analysis of these models, among those also central limit theorems for optimal objective values, is the subject of various articles and books (see Ruszczyński and Shapiro (2003) and Eichhorn and Römisch (2007)[9] and references therein).

Further, the asymptotic analysis for empirical OT plans and the corresponding distributional limit theorem in (2.11) are closely connected to the theory of (deterministic) parametric programming for standard linear programs. Early pioneers[10] developing the theory for parametric programming considered standard linear programs of the form

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} (c + \eta c')^T x \\ &\text{s.t.}\, Ax = (b + \delta b'),\ x \geq 0, \end{aligned} \tag{2.24}$$

for specified perturbations $c' \in \mathbb{R}^n$ and $b' \in \mathbb{R}^m$ and scalars $\eta, \delta \in \mathbb{R}$. For instance, the work by Saaty and Gass (1954) deals with the case $\delta = 0$ and $\eta \in \mathbb{R}$ in (2.24). In the context of the bases-driven approach in Klatt et al. (2020a), their main results state that

[9] Eichhorn and Römisch (2007) prove central limit theorems for the empirical optimal objective value of a two-stage stochastic mixed-integer linear programs if the probability measure μ is estimated by its empirical measure $\hat{\mu}_n$. Their proof technique shares much similarity to the approach taken by Hundrieser et al. (2021c) discussed in Chapter 3.

[10] For a detailed historical account on parametric programming and sensitivity analysis please refer to the book by Gal and Greenberg (2012).

there exists an interval (α_1, α_2) on the real line that is partitioned into *critical regions* (intervals) Ω_k for $k = 1, \dots, K$ such that for each $\eta \in \Omega_k$ the basis I_k remains primal and dual optimal. Analogous statements for $\eta = 0$ and $\delta \in \mathbb{R}$ are contained in the article by Manne (1953).

However, the nature of the asymptotic analysis for empirical OT plans requires a more elaborate analysis based on random perturbations. Sensitivity analysis for linear programs broadens the theory to more general perturbations (Bonnans and Shapiro, 2013). For finite dimensional linear programs such a sensitivity analysis may be analysed from an intuitive geometric perspective related to the *basis decomposition theorem* by Walkup and Wets (1969): If the standard linear program (2.22) is bounded, then there exists a decomposition of the cone $\text{pos}(A) = \{y \mid y = Ax,\ x \geq 0\} \subset \mathbb{R}^m$ into a finite closed polyhedral complex[11] $\mathcal{C}$ whose elements are closed convex cones with vertex at the origin and an one-to-one correspondence between the one-dimensional cones of $\mathcal{C}$ and selected columns of the constraint matrix A, such that

(a) the closed m-dimensional elements of $\mathcal{C}$ cover $\text{pos}(A)$,

(b) the m columns of A associated with the edges of a closed m-dimensional element C of $\mathcal{C}$ constitute an optimal basis for all b in C.

As correctly noted by Walkup and Wets (1969), the polyhedral complex is not unique. Non-uniqueness is related to degeneracy of the dual program for (2.22), i.e.,

$$\begin{aligned} &\max_{\lambda \in \mathbb{R}^{2M}} b^T \lambda \\ &\text{s.t.}\, A^T \lambda \leq c. \end{aligned} \tag{2.25}$$

Figure 2.1a illustrates the basis decomposition theorem. Rather than working with the elements of a particular complex $\mathcal{C}$, the asymptotic analysis for the empirical OT plan in Klatt et al. (2020a) employs the cones H_k in (2.10) of feasible directions. To each cone H_k is attached a primal and dual feasible basis I_k. In terms of the basis decomposition theorem, the basis I_k corresponds to a closed convex cone (*decision region*) C_k spanned by the columns of A_{I_k}. The cone C_k is an element of a particular polyhedral complex $\mathcal{C}$ and H_k is equal to all directions $v \in \mathbb{R}^m$ for which there exists a $t > 0$ such that $b + tv$ remains in C_k. In other words, for each $v \in H_k$ the basis I_k remains primal and dual optimal for (2.22) with parameter b replaced by $b + tv$. Under degeneracy in linear programs, there exist several polyhedral complexes and consequently the cones of feasible directions might be included in each other. Figure 2.1b illustrates these cases.

[11] A finite closed polyhedral complex $\mathcal{C}$ is a finite collection of closed convex cones such that: (a) If $C \in \mathcal{C}$, then every closed face of C is an element in $\mathcal{C}$. (b) If C_1, C_2 are distinct cones in $\mathcal{C}$, then either they are disjoint or they intersect in a common face.

Depending on the initial location of b in the cone $\mathrm{pos}(A)$, a small (random) perturbation either leaves a basis optimal or induces a change from one basis to another. Under degeneracy such a change is usually non-unique that complicates the sensitivity analysis (see also Gal (1986); Jansen et al. (1997); Koltai and Terlaky (2000) for details)[12]. The limiting random variable in (2.11) is designed to take all possible bases changes into account. Under primal or dual non-degeneracy the bases changes are less complicated and consequently lead to distributional limit theorems with a simplified limiting random variable in (2.13) and (2.15), respectively.

In modern mathematical formulations, the proof technique underlying the main result (2.11) employs the fundamental observation that the convex polyhedron $\mathrm{pos}(A)$ allows for a *cone triangulation* depending on the corresponding dual linear program (De Loera et al., 2010)[13]. The triangulation of $\mathrm{pos}(A)$ is not only crucial for distributional limit theorems but has also recently been applied to describe the optimal solution minimizing

$$\mathcal{OT}_d(r, \mathcal{M}) = \min_{s\in\mathcal{M}} \min_{\pi\in\Pi(r,s)} \sum_{i,j=1}^{M} d(x_i, x_j)\pi(x_i, x_j), \tag{2.26}$$

the OT cost between a fixed probability measure $r \in \Delta(\mathcal{X})$ and an algebraic variety $\mathcal{M}$ inside the probability simplex $\Delta(\mathcal{X})$ (Sturmfels and Venturello, 2020). When restricted to a particular cone C of a specified complex $\mathcal{C}$, the inner minimization problem in (2.26) is a simple linear function. Hence, finding an optimal solution $s \in \mathcal{M}$ for (2.26) amounts to minimize a piecewise linear function that simplifies exact computations.

Closely related to the asymptotic analysis for empirical OT plans is the work by Sommerfeld and Munk (2018) proving limit laws for the empirical $\mathcal{OT}_d(\hat{r}_n, s)$ cost. The distributional limit theorem in (2.11) recovers their main result[14]. For $I_1, \dots, I_K$ optimal primal and dual bases for the standard linear program $\mathcal{OT}_d(r, s)$, it holds, for sample size n tending to infinity, that

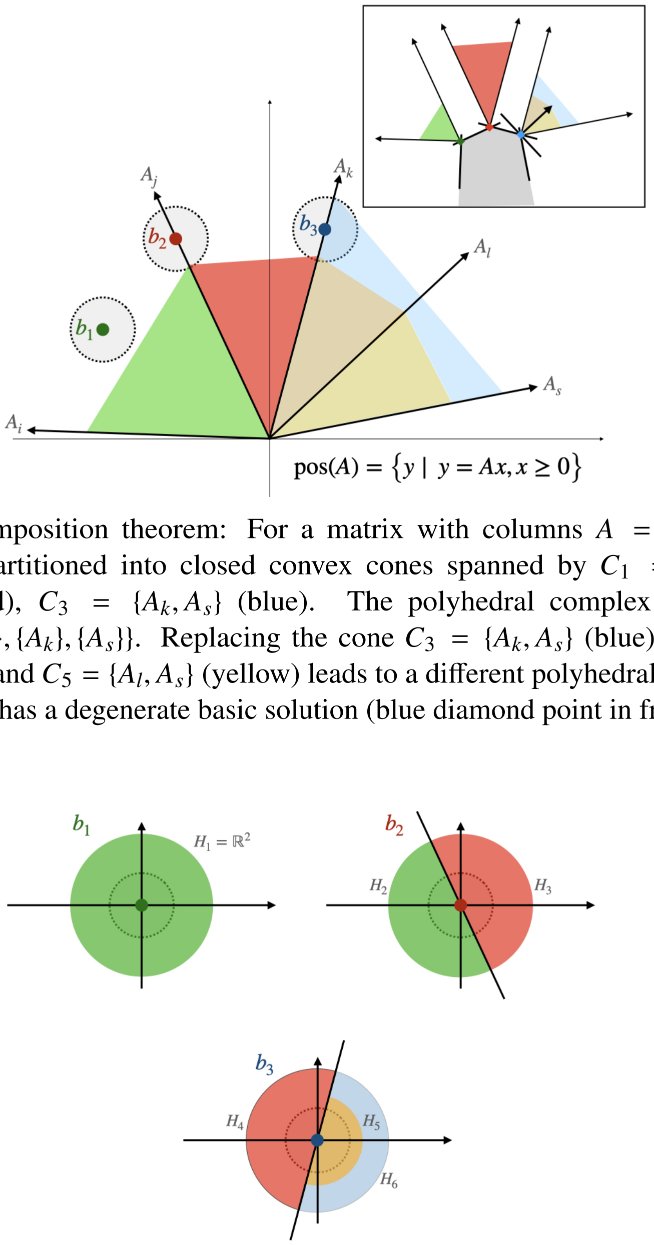
$$\sqrt{n}\left(\mathcal{OT}_d(\hat{r}_n, s) - \mathcal{OT}_d(r, s)\right) \xrightarrow{\mathcal{D}} \max_{1\le j\le K} \left\langle f_\dagger(j), \mathbb{G}_{r_\dagger} \right\rangle \tag{2.27}$$

with $f_\dagger(j)$ the dual optimal basic solution induced by basis I_j with $1 \le j \le K$. Indeed, a simple application of the continuous mapping theorem with linear function $\langle d, \cdot\rangle$

[12] In terms of the primal and dual optimal bases $I_1, \dots, I_K$, primal degeneracy leads to a discussion on *degeneracy graphs* (Gal, 1985; Greenberg, 1986).

[13] In terms of De Loera et al. (2010, Sec. 1.2) and with respect to Figure 2.1a, the collection of sub-cones $\{C_1, C_2, C_3\}$ is a *cone triangulation* of $\mathrm{pos}(A)$. Each individual sub-cone is *simplicial* since the corresponding matrix is of full rank, e.g., for $C_1 = \{A_j, A_k\}$ with matrix $[A_j, A_k]$.

[14] The central limit theorem for empirical $\mathcal{OT}_d(\hat{r}_n, s)$ cost is also of particular interest in Chapter 3 and please refer to Section 3.2 for a detailed discussion.



(a) The basis decomposition theorem: For a matrix with columns $A = [A_i, A_j, A_k, A_l, A_s]$ the set $\mathrm{pos}(A)$ is partitioned into closed convex cones spanned by $C_1 = \{A_i, A_j\}$ (green), $C_2 = \{A_j, A_k\}$ (red), $C_3 = \{A_k, A_s\}$ (blue). The polyhedral complex is equal to $C = \{C_1, C_2, C_3, \{A_i\}, \{A_j\}, \{A_k\}, \{A_s\}\}$. Replacing the cone $C_3 = \{A_k, A_s\}$ (blue) with the two sub-cones $C_4 = \{A_k, A_l\}$ and $C_5 = \{A_l, A_s\}$ (yellow) leads to a different polyhedral complex. Notably, the dual polyhedron has a degenerate basic solution (blue diamond point in framed figure).

(b) The cones H_k of feasible directions defined in (2.10): The vector $b_1 + tv$ for any perturbation v and small enough $t > 0$ remains in the green cone in (a) and consequently $H_1 = \mathbb{R}^2$. The set of feasible directions for b_2 is divided into two regions depending on the green and red cone in (a). This leads to the two cones H_2 and H_3. Owing to degeneracy, the situation for parameter b_3 is different. While for certain directions v the parameter $b_3 + tv$ belongs to the red cone in (a), there exist directions $\tilde{v}$ such that $b_3 + t\tilde{v}$ belongs to two cones (blue and yellow).

Figure 2.1: **The Sensitivity of Linear Programs.** The right hand side parameter b_i for $i = 1, 2, 3$ belongs to the cone $\mathrm{pos}(A) = \{y \mid y = Ax,\ x \geq 0\}$ in (a) that is partitioned according to the basis decomposition theorem. Depending on the location of b_i, its set of feasible perturbations is partitioned by cones H_k in (b). For each direction $v \in H_k$, there exists a $t > 0$ such that the basis I_k remains optimal for the linear program (2.22) with right hand side $b_i + tv$.

to the main results (2.11) leads to the limit law in (2.27). A careful rewriting of the corresponding limiting random variable employs duality. To each indicator function in (2.11) is attached a dual feasible basic solution $f_\dagger(j)$ induced by bases I_j with $1 \leq j \leq K$ and it holds that

$$\langle d, M(\mathbb{G}_{r_\dagger})\rangle = \sum_{\mathcal{K}} \mathbb{1}_{\left\{\mathbb{G}_{r_\dagger} \in \left\{H_{\mathcal{K}} \setminus \cup_{k \notin \mathcal{K}} H_k\right\}\right\}} \left\langle f_\dagger(k), \mathbb{G}_{r_\dagger} \right\rangle. \tag{2.28}$$

If $\mathbb{G}_{r_\dagger} \in \{H_{\mathcal{K}} \setminus \cup_{k \notin \mathcal{K}} H_k\}$, then any dual feasible basic solution $f_\dagger(k)$ for $k \in \mathcal{K}$ maximizes the random linear objective $\langle \cdot, \mathbb{G}_{r_\dagger}\rangle$ among all dual optimal basic solutions $f_\dagger(j)$ with $1 \leq j \leq K$. In particular, the limiting random variable in (2.27) is equal to the right hand side in (2.28)[15]. Similarly, the main result (2.11) allows for statements on the Hausdorff distance

$$d_H(\Pi^\star(r, s), \Pi^\star(\hat{r}_n, s)) = \max_{\hat{\pi} \in \Pi^\star(\hat{r}_n, s)} \min_{\pi \in \Pi^\star(r,s)} \|\hat{\pi} - \pi\| \vee \max_{\pi \in \Pi^\star(r,s)} \min_{\hat{\pi} \in \Pi^\star(\hat{r}_n, s)} \|\hat{\pi} - \pi\| \tag{2.29}$$

between the optimality set $\Pi^\star(r, s)$ and its empirical counterpart $\Pi^\star(\hat{r}_n, s)$. By the strong law of large numbers $\|\hat{r}_n - r\| = O_{\mathbb{P}}(n^{-1/2})$ and from the method employed in Klatt et al. (2020b) follows $d_H(\Pi^\star(r, s), \Pi^\star(\hat{r}_n, s)) = O_{\mathbb{P}}(n^{-1/2})$.

Apart from the statistical results on finite spaces, recent attention is given to the estimation of OT maps

$$T_0 \in \operatorname*{arg\,min}_{T\#\mu = \nu} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \, \mathrm{d}\mu(x) \tag{2.30}$$

between more general probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with respect to squared Euclidean cost (see Manole et al. (2021) and references therein). In contrast to the finite dimensional linear program approach by Klatt et al. (2020a), at the heart of the statistical analysis for suitable OT map estimators on Euclidean spaces is *Brenier's Theorem*[16]. A reasonable estimator $\hat{T}$ for T_0 is obtained by solving (2.30) with μ and ν replaced by suitable estimators derived from data. The work by Deb et al. (2021) analyses estimators $\hat{T}_{n,m}$ based on a barycentric projection of the empirical OT plan between two empirical probability measures $\hat{\mu}_n$ and $\hat{\nu}_m$. Different to the rate $n^{-1/2}$ for finite spaces as deduced from Klatt et al. (2020a), the convergence rates in Deb et al.

[15] The arguments leading to the distributional limit law (2.27) for the empirical OT cost hold for more general linear programs in standard form. A direct approach via Hadamard directional differentiability of the optimal objective value function in conjunction with the delta method is included in the book by Rubinstein and Shapiro (1993, Thm. 6.4.2).

[16] If μ is absolutely continuous with respect to Lebesgue measure and μ, ν have finite second moments, then there exists a Lebesgue almost surely unique OT map which is the gradient of a convex function (cf. Section 1.1 in the Introduction).

(2021) depend on the underlying dimension. If $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ have compact support and densities bounded away from zero and infinity, then for some constant $C > 0$ it holds that

$$\mathbb{E}\left[\int_{\mathbb{R}^d} \|\hat{T}_{n,m}(x) - T_0(x)\|^2 \,\mathrm{d}\hat{\mu}_n(x)\right] \leq C \begin{cases} n^{-1/2} + m^{-1/2} & \text{if } d = 2, 3, \\ n^{-1/2}\log(1+n) + m^{-1/2}\log(1+m) & \text{if } d = 4, \\ n^{-2/d} + m^{-2/d} & \text{if } d \geq 5. \end{cases} \tag{2.31}$$

The rate of convergence is proven to be minimax under certain regularity conditions (Hütter and Rigollet, 2021). Besides Brenier's Theorem, the proof technique for (2.31) relies on stability estimates that lead to a supremum of an integral involving the empirical measure. The supremum might then be bounded by empirical process techniques and Dudley's entropy integral. For sample size $n = m$, the barycentric projection $\hat{T}_{n,n}$ is equal to the empirical OT plan (see footnote 8 on p. 7). Then, linear smoothers and least square estimators are employed to define an empirical OT map on the whole Euclidean space. The rate of convergence for such estimators is analyzed by Manole et al. (2021) with qualitatively similar results as in (2.31). Further, the OT map in (2.30) between two Gaussian measures is linear and only depends on the underlying means and covariance matrices of the Gaussian measures. If $\mu = \mathcal{N}(\theta_\mu, \Sigma_\mu)$ and $\nu = \mathcal{N}(\theta_\nu, \Sigma_\nu)$ are Gaussian distributions on $\mathbb{R}^d$, then the OT map transporting μ optimally to ν with respect to squared Euclidean cost is equal to $T(x) = \theta_\nu + A(x - \theta_\mu)$ with matrix $A = \Sigma_\mu^{-1/2}(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})\Sigma_\mu^{-1/2}$ (cf. formula (3.37)). This leads to estimators of linear OT maps in the Gaussian case with dimension-free parametric convergence rates $n^{-1/2}$ (Flamary et al., 2019).

Klatt et al. (2020b): Compared to empirical non-regularized OT, the statistical analysis for empirical *entropy regularized* OT is still in its infancy. Parallel and indepedently to Klatt et al. (2020b), contributions by Bigot et al. (2019) prove central limit theorems for empirical entropy regularized OT quantities defined between probability measures $r, s \in \Delta(\mathcal{X})$ supported on a finite space $\mathcal{X} = \{x_1, \ldots, x_M\}$. Their purpose is to construct test statistics for measuring differences in multivariate probability distributions based on the empirical entropy regularized objective value $\mathcal{OT}_\lambda(\hat{r}_n, s)$ in (2.5) and the empirical Sinkhorn loss $\mathcal{SL}_\lambda(\hat{r}_n, s)$ in (2.7). This approach is motivated from the respective superior computational complexity when compared to non-regularized OT. The main theoretical results in Bigot et al. (2019) state central limit theorems identical to the statements in (2.20) and (2.21). Their proof technique is similar to Klatt et al. (2020b) and relies on differentiability of the entropy regularized quantities, considered as functions defined on the probability simplex $\Delta(\mathcal{X})$. Together with an application of the delta method, distributional limit results follow. Notably, the differentiability analysis of the entropy regularized OT plan in Klatt et al. (2020b) is more general and implies differentiability of the corresponding optimal value and the Sinkhorn loss (see also Luise et al. (2018) for analogous differentiability statements).

The law of the limiting random variable in (2.20) degenerates to a Dirac at zero if and only if the dual solution $f_\lambda^{(r,s)}$ is constant. Degeneracy is a rather exceptional case as is seen by considerations of the correspondence between primal and dual entropy regularized optimizers in (1.15). From an algorithmic point of view, a constant $f_\lambda^{(r,s)}$ implies that a right scaling of the kernel matrix $K := \left(-\frac{d(x_i,x_j)}{\lambda}\right)_{i,j}$ by a diagonal matrix D suffices, in order to ensure that their product KD fulfills (up to a constant) the marginal constraints imposed by r and s. In particular, the Sinkhorn algorithm finds an exact optimal solution after only one iteration (see Section 1.3). Similarly, the central limit theorem for the Sinkhorn loss under the alternative $r \neq s$ is non-degenerate.

An interesting case arises for the Sinkhorn loss under the null $r = s$. For this case, $\mathcal{SL}_\lambda(r,r) = 0$ and the gradient $\nabla_r \mathcal{SL}_\lambda(r,r)$ vanishes. This suggests the central limit law for the empirical quantity $\mathcal{SL}_\lambda(\hat{r}_n, r)$ to be of second order. Indeed, Bigot et al. (2019, Thm. 2.8) prove that for increasing sample size n, the empirical Sinkhorn loss asymptotically follows a mixture of chi-squared distributed random variables

$$n\,\mathcal{SL}_\lambda(\hat{r}_n, r) \xrightarrow{\mathcal{D}} \frac{1}{2}\sum_{i=1}^{M} \rho_i \chi_i^2(1) \tag{2.32}$$

with $\chi_1^2(1), \dots, \chi_M^2(1)$ i.i.d. chi-squared random variables of degree one and $\rho_1, \dots, \rho_M$ non-negative eigenvalues depending on the Hessian of the Sinkhorn loss.

Further, the entropy regularized OT quantities (1.17), (1.18) and (1.19) converge with decreasing regularization parameter $\lambda > 0$ to their respective non-regularized OT cost (Weed, 2018; Genevay et al., 2019). From a statistical perspective, the central limit theorems (2.19), (2.20) and (2.21) with decreasing regularization parameter $\lambda(n)$ depending on the sample size n retrieve the central limit theorem in (2.27) for the empirical OT cost. As a prototypical example consider the empirical Sinkhorn divergence $\mathcal{SD}_\lambda(\hat{r}_n, s)$ and the corresponding central limit theorem in (2.19). For a regularization parameter $\lambda(n) = o(n^{-1/2})$ and sample size n tending to infinity, it holds that

$$\sqrt{n}\left(\mathcal{SD}_{\lambda(n)}(\hat{r}_n, s) - \mathcal{SD}_{\lambda(n)}(r, s)\right) \xrightarrow{\mathcal{D}} \max_{1\le j\le K} \langle f(j), \mathbb{G}_r\rangle \tag{2.33}$$

with $f(j)$ for $j = 1, \dots, K$ the dual optimal basic solutions for the dual linear program of $\mathcal{OT}_d(r,s)$. Statement (2.33) is valid for the null $r = s$ as well as for the alternative

$r \neq s$ (Klatt et al., 2020b, Sec. 3.2). An analogous result for the Sinkhorn loss under the alternative is included in Bigot et al. (2019, Thm. 2.11).

The work by Hundrieser et al. (2021a) extends the statistical analysis for empirical entropy regularized OT quantities to countable spaces $\mathcal{X} = \{x_1, x_2, \dots\}$. Indeed, the central limit theorems for empirical entropy regularized OT quantitites remain valid for probability measures with countable support. However and in contrast to finite spaces, the weak convergence (2.1) for the multinomial process $\sqrt{n}(\hat{r}_n - r)$ towards a centered Gaussian process $\mathbb{G}_r$ in $l^1(\mathcal{X})$ requires the *Borisov-Dudley-Durst*[17] condition

$$\sum_{x \in \mathcal{X}} \sqrt{r(x)} < +\infty. \tag{2.34}$$

If restricted to a bounded cost function $\sup_{x,x'} |c(x,x')| < +\infty$ and probability measures $r, s \in \Delta(\mathcal{X})$ such that r satisfies condition (2.34), then the empirical optimal value $\mathcal{EOT}_{c,\lambda}(\hat{r}_n, s)$ in (1.21) (after proper centering and standardization) weakly converges

$$\sqrt{n}\left(\mathcal{EOT}_{c,\lambda}(\hat{r}_n, s) - \mathcal{EOT}_{c,\lambda}(r, s)\right) \xrightarrow{\mathcal{D}} \left\langle \mathbb{G}_r, f_\lambda^{(r,s)} \right\rangle, \tag{2.35}$$

as the sample size n tends to infinity. The limiting random variable in (2.35) follows a centered Gaussian distribution with variance equal to $\mathrm{Var}_{X\sim r}[f_\lambda^{(r,s)}(X)]$. Herein, $f_\lambda^{(r,s)}$ denotes the dual solution for (1.22). Analogously to Klatt et al. (2020b), the main result in Hundrieser et al. (2021a) states a weak limit for the empirical entropy regularized OT plan $\pi^\lambda(\hat{r}_n, s)$ for (1.21). If r satisfies condition (2.34), then, for sample size n tending to infinity, it holds that

$$\sqrt{n}\left(\pi^\lambda(\hat{r}_n, s) - \pi^\lambda(r, s)\right) \xrightarrow{\mathcal{D}} \mathcal{D}^H_{|(r,s)}\pi^\lambda\left(\mathbb{G}_r, 0\right). \tag{2.36}$$

The limiting random variable in (2.36) is equal to a suitable derivative of the entropy regularized OT plan considered as a function over the probability simplex $\Delta(\mathcal{X})$ evaluated at the weak limit $\mathbb{G}_r$ for the multinomial empirical process. The derivative is linear and hence the limiting random variable follows a centered Gaussian process inline with results by Klatt et al. (2020b) for finite spaces. Extensions for the central limit theorems in (2.35) and (2.36) to unbounded costs require a careful modification to a *weighted* version of the Borisov-Dudley-Durst condition (2.34). The weights are linked to an exponential penalty term appearing in the dual formulation in (1.22) and suitable functions lower and upper bounding the cost function.

[17] The Borisov-Dudley-Durst condition (2.34) naturally appears in empirical process theory. It states that for any probability measure $r \in \Delta(\mathbb{N})$, the power set $2^{\mathbb{N}}$ is r-Donsker if and only if condition (2.34) holds (Dudley, 2014, Thm. 7.3.1). More details are included in Section 3.2.

The results for empirical entropy regularized OT with probability measures defined on finite and countable spaces leave open statements for more general spaces. However, the literature on a related asymptotic analysis is still scarce and limited to particular cases. For squared Euclidean cost $c(x, y) = \frac{1}{2}||x - y||^2$ and *sub-Gaussian* probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, Mena and Niles-Weed (2019) analyze statistical properties for the empirical entropy regularized value $\mathcal{EOT}_{||\cdot||^2,\lambda}(\hat{\mu}_n, \nu)$ in (1.21). They prove the upper bound

$$\mathbb{E}_\mu \left[|\mathcal{EOT}_{||\cdot||^2,\lambda}(\hat{\mu}_n, \nu) - \mathcal{EOT}_{||\cdot||||^2,\lambda}(\mu, \nu)| \right] \leq K_{d,\lambda,\sigma} n^{-1/2}$$

with notably parametric rate $n^{-1/2}$ and constant $K_{d,\lambda,\sigma}$ depending on the underlying dimension d, regularization parameter λ and sub-Gaussian parameter σ (see Genevay et al. (2019) for compactly supported measures)[18]. Mena and Niles-Weed (2019) also provide a central limit theorem. For sample size n increasing to infinity, it holds that

$$\sqrt{n} \left(\mathcal{EOT}_{||\cdot||^2,\lambda}(\hat{\mu}_n, \nu) - \mathbb{E}_\mu \left[\mathcal{EOT}_{||\cdot||^2,\lambda}(\hat{\mu}_n, \nu) \right] \right) \xrightarrow{\mathcal{D}} Z \sim \mathcal{N} \left(0, \mathrm{Var}_{X \sim \mu} \left[f_\lambda^{(\mu,\nu)}(X) \right] \right) \tag{2.37}$$

and $f_\lambda^{(\mu,\nu)}$ the dual solution for (1.22). The central limit theorem (2.37) is the generalization of statement (2.20) when restricted to finite spaces. A notable exception, however, is the centering around $\mathbb{E}_\mu \left[\mathcal{EOT}_{||\cdot||^2,\lambda}(\hat{\mu}_n, \nu) \right]$ in statement (2.37) rather than the population quantity $\mathcal{EOT}_{||\cdot||^2,\lambda}(\mu, \nu)$. The centering is a consequence of the proof technique for (2.37) inspired by del Barrio and Loubes (2019) and based on the Efron-Stein variance inequality[19]. In fact, apart from finite and countable spaces, it remains open if the random sequence $\sqrt{n} \left(\mathcal{EOT}_{||\cdot||^2,\lambda}(\hat{\mu}_n, \nu) - \mathcal{EOT}_{||\cdot||^2,\lambda}(\mu, \nu) \right)$ weakly converges to a limiting random variable with non-degenerate limit law.

Different to the plug-in approach based on the empirical measure, Bercu and Bigot (2021) propose an estimator $\widehat{\mathcal{EOT}}_{c,\lambda,n}$ suitable for the semi-discrete setting with $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $s \in \Delta(\mathcal{X})$ for a finite space $\mathcal{X} = \{x_1, \ldots, x_M\}$ in (1.21). The estimator is based on a stochastic Robbins-Monro algorithm with appropriate chosen algorithmic parameters. Among a general stochastic analysis of the algorithm designed for entropy regularized OT, Bercu and Bigot (2021) prove a distributional limit law. If the cost function c fulfills

$$\int_{\mathbb{R}^d} c^4(x, y) \, \mathrm{d}\mu(x) < +\infty \quad \forall y \in \mathbb{R}^d, \tag{2.38}$$

then, for n tending to infinity, the random sequence $\sqrt{n} \left(\widehat{\mathcal{EOT}}_{c,\lambda,n} - \mathcal{EOT}_{c,\lambda}(\mu, s) \right)$

[18] This stands in stark contrast to convergence rates for *non-regularized* OT and please refer to Section 3.2 and the upper bounds in (3.16) for further details.

[19] See also the discussion on p. 47 and the central limit theorem in (3.34) for further details.

weakly converges towards a limiting random variable. The limit law is a centered Gaussian distribution with an asymptotic variance equal to $\mathrm{Var}_{X\sim\mu}\left[f_\lambda^{(\mu,s)}(X)\right]$ with $f_\lambda^{(\mu,s)}$ the semi-discrete dual solution for (1.22) and hence identical to (2.37) if restricted to the semi-discrete setting.

CHAPTER 3

A Unifying Approach to Limit Laws for Empirical OT

This chapter is based on the article Hundrieser et al. (2021c) presenting a unifying approach to distributional limit laws for the empirical OT cost. The general setting covers two Polish spaces $\mathcal{X}, \mathcal{Y}$ and a non-negative cost function that fulfills the following assumption.

$$\text{The cost } c\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+ \text{ is continuous with } \|c\|_\infty := \sup_{x,y}|c(x,y)| < +\infty. \tag{\textbf{C1}}$$

For any two probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and a cost function satisfying assumption (**C1**), the $\mathcal{OT}_c(\mu,\nu)$ cost is finite and enjoys Kantorovich duality

$$\mathcal{OT}_c(\mu,\nu) = \sup_{f\in\mathcal{F}_c} \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} f^c(y)\,\mathrm{d}\nu(y). \tag{3.1}$$

The function $f^c(y) = \inf_{x\in\mathcal{X}} f(x) - c(x,y)$ is the c-transform of f and the supremum is taken over uniformly bounded c-concave functions on $\mathcal{X}$ (cf. Section 1.1) defined as

$$\mathcal{F}_c := \left\{ f\colon \mathcal{X}\to\mathbb{R} \,\middle|\, \exists\, g\colon \mathcal{Y}\to\mathbb{R},\ -\|c\|_\infty \le g \le 0,\ f(x) = \inf_{y\in\mathcal{Y}} c(x,y) - g(y) \right\}. \tag{3.2}$$

There exists a Kantorovich potential $f \in \mathcal{F}_c$ (optimizer) attaining the supremum in the dual formulation (3.1) and the set of all Kantorovich potentials is denoted as

$$S_c(\mu,\nu) := \left\{ f\in\mathcal{F}_c \,\middle|\, \mathcal{OT}_c(\mu,\nu) = \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} f^c(y)\,\mathrm{d}\nu(y) \right\}. \tag{3.3}$$

The Kantorovich duality (3.1) sets the $\mathcal{OT}_c(\mu,\nu)$ cost within a functional framework suitable for a statistical analysis. Denote by $l^\infty(\mathcal{F}_c)$ the normed space of bounded functionals from $\mathcal{F}_c$ to the reals equipped with the uniform norm $\|\Psi\|_{\mathcal{F}_c} = \sup_{f\in\mathcal{F}_c} |\Psi(f)|$

and analogously for $l^\infty(\mathcal{F}_c^c)$ with $\mathcal{F}_c^c := \{f^c \mid f \in \mathcal{F}_c\}$. A probability measure $\mu \in \mathcal{P}(\mathcal{X})$ (resp. $\nu \in \mathcal{P}(\mathcal{Y})$) defines an element in $l^\infty(\mathcal{F}_c)$ (resp. $l^\infty(\mathcal{F}_c^c)$) by integration $\mu(f) := \int_\mathcal{X} f(x)\,\mathrm{d}\mu(x)$. This perspective restates the $\mathcal{OT}_c$ cost in (3.1) as the functional $\mathcal{OT}_c \colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \subseteq l^\infty(\mathcal{F}_c) \times l^\infty(\mathcal{F}_c^c) \to \mathbb{R}_+$ with

$$(\mu, \nu) \mapsto \sup_{f \in \mathcal{F}_c} \mu(f) + \nu(f^c). \tag{3.4}$$

The functional (3.4) is *Hadamard directionally differentiable* at $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$. The proof technique requires either *one* of the following assumptions on the cost function and the underlying spaces[1].

$$\mathcal{X} \text{ compact; } \{c(\cdot, y) \mid y \in \mathcal{Y}\} \text{ equicontinuous on } \mathcal{X}. \tag{\textbf{C2a}}$$

$$\mathcal{X}, \mathcal{Y} \text{ locally compact; } \{c(x, \cdot) \mid x \in \mathcal{X}\}, \{c(\cdot, y) \mid y \in \mathcal{Y}\} \text{ equicontinuous.} \tag{\textbf{C2b}}$$

Under assumption (**C1**) combined with either (**C2a**) or (**C2b**), the Hadamard directional derivative for (3.4) is defined along directions (Δ_μ, Δ_ν) belonging to the product of the *Bouligand cones*

$$\overline{\left\{\frac{\mu' - \mu}{t} \,\middle|\, t > 0,\ \mu' \in \mathcal{P}(\mathrm{supp}(\mu))\right\}} \times \overline{\left\{\frac{\nu' - \nu}{t} \,\middle|\, t > 0,\ \nu' \in \mathcal{P}(\mathrm{supp}(\nu))\right\}},$$

where the closure is defined in $l^\infty(\mathcal{F}_c)$ and $l^\infty(\mathcal{F}_c^c)$, respectively. The derivative turns out to be equal to

$$(\Delta_\mu, \Delta_\nu) \mapsto \sup_{f \in S_c(\mu,\nu)} \Delta_\mu(f) + \Delta_\nu(f^c). \tag{3.5}$$

The Hadamard directional differentiability together with the weak convergence of an underlying empirical process is at the heart of the main statistical results. Employing the functional delta method (Römisch, 2004) then implies a distributional limit law for the empirical OT cost after centering around the population quantity and proper scaling.

3.1 Main Results

Consider an i.i.d. sequence of random variables $X_1, \ldots, X_n \sim \mu$ and its associated empirical measure $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. The *empirical process* $\sqrt{n}(\hat{\mu} - \mu)$ defines a random element in the normed space $l^\infty(\mathcal{F}_c)$. A function class $\mathcal{F}_c$ is termed *μ-Donsker* if the

[1] Assumptions (**C2a**) and (**C2b**) allow for an application of a suitable version of the *Arzelá-Ascoli* theorem to the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ to conclude for necessary compactness properties. This simplifies the derivative to (3.5) for the OT functional in (3.4).

empirical process weakly converges

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathbb{G}_\mu, \tag{3.6}$$

as n tends to infinity, towards a tight random element $\mathbb{G}_\mu$ in $l^\infty(\mathcal{F}_c)$. By considerations of the marginal distributions, the random element $\mathbb{G}_\mu$ turns out to be a mean-zero Gaussian process indexed over $\mathcal{F}_c$ with covariance for $f_1, f_2 \in \mathcal{F}_c$ depending on μ and equal to

$$\mathbb{E}_\mu\left[\mathbb{G}_\mu(f_1)\mathbb{G}_\mu(f_2)\right] = \mu(f_1 f_2) - \mu(f_1)\mu(f_2). \tag{3.7}$$

The main statement[2] of the article yields that if assumption (**C1**) and either (**C2a**) or (**C2b**) hold and the function class $\mathcal{F}_c$ is μ-Donsker, then a distributional limit law for the empirical OT cost follows

$$\boxed{\sqrt{n}\left(\mathcal{OT}_c(\hat{\mu}_n, \nu) - \mathcal{OT}_c(\mu, \nu)\right) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \mathbb{G}_\mu(f).} \tag{3.8}$$

The limiting random variable is equal to the supremum over the set of Kantorovich potentials $S_c(\mu, \nu)$ in (3.3) of the mean-zero Gaussian process $\mathbb{G}_\mu$ in $l^\infty(\mathcal{F}_c)$. Statement (3.8) is a direct consequence of the functional delta method, together with the Hadamard directional derivative (3.5) and the weak convergence of the empirical process (3.6). Under further assumptions, the limiting random variable in (3.8) simplifies. Indeed, if there exists a *unique*[3] Kantorovich potential f, then

$$\boxed{\sqrt{n}\left(\mathcal{OT}_c(\hat{\mu}_n, \nu) - \mathcal{OT}_c(\mu, \nu)\right) \xrightarrow{\mathcal{D}} \mathbb{G}_\mu(f).} \tag{3.9}$$

The limiting random variable $\mathbb{G}_\mu(f)$ follows a centered Gaussian distribution equal to $\mathcal{N}\left(0, \mathrm{Var}_{X\sim\mu}[f(X)]\right)$. Provided that f is μ-almost surely non-constant, a standardization of the left hand side in (3.9) with the weakly consistent estimator $\mathrm{Var}_{X\sim\hat{\mu}_n}[f_n(X)]$ with $f_n \in S_c(\hat{\mu}_n, \nu)$ for the asymptotic variance might be employed. This leads by Slutzky's Lemma to a *pivotal* limiting random variable following a *standard* Gaussian distribution. A *degenerate* case arises if the unique Kantorovich potential is additionally *trivial*[4]. Then, the limiting random variable $\mathbb{G}_\mu(f) \sim \delta_0$ in (3.9) follows a Dirac distribution at zero. The article contains a thorough analysis on triviality for Kantorovich potentials.

[2] The presentation of the main results is tailored to the one-sample case concerned with sampling out of $\mu \in \mathcal{P}(\mathcal{X})$. The probabiltiy measure $\nu \in \mathcal{P}(\mathcal{Y})$ is throughout assumed to be fixed. The article Hundrieser et al. (2021c) also focuses on sampling out of ν or the two-sample case.

[3] *Uniqueness* means that for all $f_1, f_2 \in S_c(\mu, \nu)$ their difference $f_1 - f_2$ is constant μ-almost surely. The article discusses various results concerning uniqueness statements for Kantorovich potentials.

[4] A Kantorovich potential $f \in S_c(\mu, \nu)$ is *trivial* if f is constant μ-almost surely.

The main statement on triviality depends on the underlying topology of the measures' support. In particular, if the cost function fulfills (**C1**) and either (**C2a**) or (**C2b**) with $\mathcal{X} = \mathcal{Y}$, then trivial Kantorovich potentials exist if and only if there exists a transport plan $\pi \in \Pi(\mu, \nu)$ such that all $(x, y) \in \mathrm{supp}(\pi)$ satisfy

$$c(x, y) = \inf_{x' \in \mathrm{supp}(\mu)} c(x', y). \tag{3.10}$$

A prototypical example for which (3.10) is satisfied appears for $\mu = \nu$ and cost $c(x, x) = 0$ for all $x \in \mathcal{X}$. Notably, even if (3.10) holds, the set of Kantorovich potentials is not necessarily a singleton. Indeed, if μ has disconnected support, e.g., the support is equal to the union of two *cost separated*[5] subsets, then the set of all Kantorovich potentials $S_c(\mu, \mu)$ is not a singleton. Consequently, the limit law for the random sequence $\sqrt{n}\,\mathcal{OT}_c(\hat{\mu}_n, \mu)$ in (3.8) is non-degenerate in such cases.

The perhaps simplest case leading to concrete results of the general statement (3.8) deals with $\mathcal{X} = \mathcal{Y} = \{x_1, x_2, \ldots\}$ countable discrete spaces and probability measures $r, s \in \Delta(\mathcal{X})$. Owing to the discrete topology, any bounded and non-negative cost $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ fulfills assumption (**C1**) and (**C2b**). According to van der Vaart (2000, Thm. 2.10.24) the function class $\mathcal{F}_c$ is r-Donsker if

$$\sum_{x \in \mathcal{X}} \sqrt{r(x)} < +\infty. \tag{3.11}$$

In particular, for any non-negative, bounded cost function and probability measures $r, s \in \mathcal{P}(\mathcal{X})$ such that r fulfills the summability condition (3.11), the central limit theorem (3.8) is valid.

For Euclidean spaces $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, the necessary assumptions (**C1**) and either (**C2a**) or (**C2b**) together with $\mathcal{F}_c$ being μ-Donsker translate into an interplay between sufficient regularity of the cost function and the size of the underlying dimension d. On the real line ($d = 1$) and for a bounded and (α, L)-Hölder continuous cost function $c\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ with $\alpha \in (1/2, 1]$ and some $L > 0$, i.e.,

$$|c(x, y) - c(x', y')| \leq L\left(|x - x'|^{\alpha} + |y - y'|^{\alpha}\right), \quad \forall x, x', y, y' \in \mathbb{R}, \tag{3.12}$$

the function class $\mathcal{F}_c$ is a uniformly bounded subset of real-valued (α, L)-Hölder func-

[5] Two subsets $A, B \subset \mathcal{X}$ are cost separated if $\inf_{x \in A, y \in B} c(x, y) > 0$.

tions. A sufficient condition[6] for $\mathcal{F}_c$ to be μ-Donsker is that

$$\sum_{k\in\mathbb{Z}} \sqrt{\mathbb{P}\left(X \in [k, k+1)\right)} < +\infty \tag{3.13}$$

with a random variable $X \sim \mu$. In particular, in the univariate setting with a bounded cost function that fulfills assumption (3.12) and a probability measure μ such that the tail condition (3.13) holds, the distributional limit law (3.8) for the empirical OT cost is valid.

Distributional limit results extend to dimensions $d = 2, 3$. In these higher-dimensional regimes, more regularity for the cost function is required. If the cost function is bounded and L-Lipschitz and there exists some $\Lambda > 0$ such that for all $k \in \mathbb{Z}^d$ it holds that

$$\begin{aligned} &\exists\, x_k \in [k, k+1) \text{ s.t. } c(\cdot, y) - \Lambda\|\cdot - x_k\|_2^2 \text{ is concave on } [k, k+1)\,\forall\, y \in \mathbb{R}^d, \\ &\exists\, y_k \in [k, k+1) \text{ s.t. } c(x, \cdot) - \Lambda\|\cdot - y_k\|_2^2 \text{ is concave on } [k, k+1)\,\forall\, x \in \mathbb{R}^d, \end{aligned} \tag{3.14}$$

then $\mathcal{F}_c$ is equal to a subset of semi-concave and L-Lipschitz functions. Such a function class is proven to be μ-Donsker if

$$\sum_{k\in\mathbb{Z}^d} \sqrt{\mathbb{P}\left(X \in [k, k+1)\right)} < +\infty \tag{3.15}$$

with a d-dimensional random vector $X \sim \mu$. In particular, on Euclidean spaces $\mathbb{R}^d$ with $d = 2, 3$, for a bounded and L-Lipschitz cost function that fulfills assumption (3.14) and a probability measure μ satisfying the tail condition (3.15), the distributional limit law (3.8) for the empirical OT cost is valid.

Beyond Euclidean spaces with dimension $d \leq 3$, the article contains distributional limit results for the empirical OT cost as long as at least one probability measure μ or ν has a discrete or low-dimensional Euclidean support ($d \leq 3$). This encompasses distributional limit laws for the empirical semi-discrete OT cost. More generally, it leads to statements for the empirical OT cost between a probability measure supported on an Euclidean and a probability measure supported on a Polish space. In fact, these asymmetric settings in the probability measures allow for concrete distributional limit results thanks to the *lower complexity adaptation* – a recent observation that explains the statistical complexity of the empirical OT cost to be driven by the less complex probability measure.

[6] For $\alpha = 1$, the function class $\mathcal{F}_c$ is a subset of bounded Lipschitz functions. In this setting, the condition (3.13) is known to be necessary and sufficient for $\mathcal{F}_c$ to be μ-Donsker (Giné and Zinn, 1986).

3.2 Discussion and Related Work

The analysis of the statistical performance of empirical OT plug-in estimators is a well-established research area of *statistical optimal transport*. The most studied setting is certainly the $\mathcal{OT}_{\|\cdot\|^p}(\mu,\nu)$ cost, popularized as Wasserstein distance (cf. Section 1.1.1), and defined for probability measures μ,ν with support in a general separable Banach spaces $\mathcal{X}$ with norm $\|\cdot\|$ and cost parameter $1 \leq p < \infty$. If $\mu,\nu \in \mathcal{P}(\mathcal{X})$ have finite p-th moments, then $\mathcal{OT}_{\|\cdot\|^p}(\hat{\mu}_n,\nu)$ converges almost surely to its population quantity $\mathcal{OT}_{\|\cdot\|^p}(\mu,\nu)$ (Varadarajan, 1958; Bickel and Freedman, 1981). An analysis initiated by Dudley (1969), followed up by more refined results on Euclidean spaces $\mathbb{R}^d$ from Ajtai et al. (1984); Talagrand (1992); Dobrić and Yukich (1995), concerns the speed of convergence. It is nowadays well understood that the rate of convergence for empirical OT plug-in estimators depends on various factors, such as the underlying dimension of the ground space, the smoothness of the cost function and the regularity properties of the probability measures (see Weed and Bach (2019); Niles-Weed and Rigollet (2019); Manole and Niles-Weed (2021) and references therein). In particular, for a probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ with bounded support[7], Fournier and Guillin (2015) prove tight upper bounds of the form

$$\mathbb{E}_\mu\left[\mathcal{OT}_{\|\cdot\|^p}(\hat{\mu}_n,\mu)\right] \leq C_d \begin{cases} n^{-1/2} & \text{if } d < 2p, \\ n^{-1/2}\log(n) & \text{if } d = 2p, \\ n^{-p/d} & \text{if } d > 2p, \end{cases} \tag{3.16}$$

for C_d a constant depending on the dimension. Compared to the long-standing analysis on upper bounds on the expectation, distributional limit laws for the empirical OT cost have only recently gained increasing attention. Detailed analysis is usually carried out within more concrete settings, such as on countable discrete space, the real line $\mathbb{R}$ or high-dimensional Euclidean spaces $\mathbb{R}^d$ with squared Euclidean cost.

Countable and semi-discrete case: A remarkably complete, although not chronologically first analysis is given for a countable discrete metric space $\mathcal{X} = \{x_1, x_2, \ldots\}$ and metric cost $c(x,y) = d^p(x,y)$ for some $1 \leq p < \infty$ (Sommerfeld and Munk, 2018; Tameling et al., 2019). For a probability measure $r \in \Delta(\mathcal{X})$ and some arbitrary $x_0 \in \mathcal{X}$

[7] For a probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ with unbounded support, the upper bounds (3.16) remain valid provided μ has sufficiently many finite moments $\int_{\mathbb{R}^d} \|x\|^q \, \mathrm{d}\mu(x) < +\infty$ and $q > 2p$ if $d \leq 2p$, or for $q > dp/(d-p)$ if $d > 2p$.

such that

$$\sum_{x \in \mathcal{X}} d^p(x, x_0) \sqrt{r(x)} < +\infty, \tag{3.17}$$

a distributional limit law for the empirical OT cost of the form

$$\sqrt{n}\, \mathcal{OT}_{d^p}(\hat{r}_n, r) \xrightarrow{\mathcal{D}} \sup_{f \in \mathcal{S}_{d^p}(r,r)} \langle \mathbb{G}_r, f \rangle \tag{3.18}$$

holds. The limiting random variable depends on a centered infinite dimensional Gaussian random vector $\mathbb{G}_r$ defined on a weighted $l^1(\mathcal{X})$-space and the set of Kantorovich potentials $\mathcal{S}_{d^p}(r, r)$ for the dual of $\mathcal{OT}_{d^p}(r, r)$. The cost weighted summability condition on the probability measure r in (3.17) guarantees weak convergence of the empirical process $\sqrt{n}(\hat{r}_n - r)$ towards $\mathbb{G}_r$ in some weighted $l^1(\mathcal{X})$-space as the sample size n tends to infinity. Condition (3.17) is a weighted version of (3.11), the latter being the celebrated *Borisov-Dudley-Durst* condition (see footnote 17 on p. 32). Suitable weights in (3.11) do not appear as assumption (**C1**) requires bounded cost functions. The distribution of the limiting random variable in (3.18) might degenerate. A prototypical example is given by a probability measure $r \in \mathcal{P}(\mathbb{Q} \cap [0, 1])$ and squared Euclidean cost $c(x, y) = |x - y|^2$. Then, the set of Kantorovich potentials $\mathcal{S}_{|\cdot|^2}(r, r)$ is a singleton (up to a constant shift) (Santambrogio, 2015, Prop. 7.18) and its unique element is trivial (cf. (3.10) and surrounding discussion). Consequently, the limiting random variable $\langle \mathbb{G}_r, f \rangle \sim \delta_0$ follows a Dirac distribution at zero. In contrast, on finite spaces $\mathcal{X} = \{x_1, \dots, x_M\}$, as considered by Sommerfeld and Munk (2018), the limit law (3.18) is always non-degenerate as the set $\mathcal{S}_{d^p}(r, r)$ is never equal to a singleton. Notably, condition (3.17) is vacuous in the finite case and weak convergence of the underlying empirical process simply follows from the standard multivariate central limit theorem for empirical multinomial frequencies[8]. A qualitatively different situation arises for two different probability measures $r \neq s$. If the probability measure r fulfills the cost weighted summability condition (3.17), then the main result in Tameling et al. (2019) states that

$$\sqrt{n}\, \left(\mathcal{OT}_{d^p}(\hat{r}_n, s) - \mathcal{OT}_{d^p}(r, s)\right) \xrightarrow{\mathcal{D}} \sup_{f \in \mathcal{S}_{d^p}(r,s)} \langle \mathbb{G}_r, f \rangle\,. \tag{3.19}$$

The limit law is typically non-degenerate as trivial Kantorovich potentials usually do not exist in this case. On finite spaces, linear programming provides sufficient conditions for which the set of dual solutions $\mathcal{S}_{d^p}(r, s)$ is a singleton (Klee and Witzgall, 1968)[9]. Then, the limiting random variable in (3.19) follows a centered Gaussian distribution

[8] In the finite setting, the limiting random variable $\mathbb{G}_r$ follows the M-dimensional centered Gaussian distribution $\mathcal{N}(0, \Sigma(r))$ (cf. (2.1) in Chapter 2).

[9] A sufficient condition for dual uniqueness and hence Gaussian limiting random variables is (2.16).

with variance $\mathrm{Var}_{X\sim r}[f(X)]$ and f the associated unique Kantorovich potential for the dual formulation of $\mathcal{OT}_{d^p}(r,s)$.
The work by del Barrio et al. (2021a) extends the previous results to the semi-discrete setting. If the probability measures $r \in \mathcal{P}(\mathcal{X})$ remains with finite support but s is replaced by a general probability measure $\nu \in \mathcal{P}(\mathcal{Y})$ with $\mathcal{Y}$ a Polish space, then for a non-negative and ν-integrable[10] cost function, the distributional limit law

$$\sqrt{n}\ (\mathcal{OT}_c(\hat{r}_n,\nu) - \mathcal{OT}_c(r,\nu)) \xrightarrow{\mathcal{D}} \sup_{f\in S_c(r,\nu)} \langle \mathbb{G}_r, f\rangle \tag{3.20}$$

remains valid. Applied to countable spaces or the semi-discrete setting, the unifying approach in (3.8) extends the results in (3.18), (3.19) and (3.20) to general but bounded cost functions.

Real line $\mathbb{R}$: Considerably earlier statements on distributional limit laws for the empirical OT cost focus on the real line $\mathbb{R}$ with Euclidean cost $c(x,y) = |x-y|^p$ and $p \geq 1$. In this setting, the OT cost is equal to the L_p-norm

$$\mathcal{OT}_{|\cdot|^p}(\mu,\nu) = \int_0^1 \left|F^{-1}(u) - G^{-1}(u)\right|^p \,\mathrm{d}u \tag{3.21}$$

between the quantile functions F^{-1} and G^{-1} for μ and ν, respectively. The work by del Barrio et al. (1999) considers the cost $c(x,y) = |x-y|$. Applying Fubini's theorem to the explicit formula in (3.21) and $p=1$, the $\mathcal{OT}_{|\cdot|}(\mu,\nu)$ cost is nothing but the L_1-distance

$$\mathcal{OT}_{|\cdot|}(\mu,\nu) = \int_{\mathbb{R}} |F(t) - G(t)|\,\mathrm{d}t \tag{3.22}$$

between the cumulative distribution functions F and G for μ and ν, respectively. If $\mu = \nu$, distributional limit laws for the empirical $\mathcal{OT}_{|\cdot|}(\hat{\mu}_n,\mu)$ cost lead to considerations of an underlying functional central limit statement involving the empirical cumulative distribution function $\hat{F}_n$. A necessary and sufficient condition for the weak convergence of $\sqrt{n}\left(\hat{F}_n(t) - F(t)\right) \xrightarrow{\mathcal{D}} \mathbb{B}(F(t))$ with $\mathbb{B}(t)$ a standard Brownian bridge in the Banach space[11] $L^1(\mathbb{R})$ is that

$$\int_{\mathbb{R}} \sqrt{F(t)\,(1-F(t))}\,\mathrm{d}t < +\infty. \tag{3.23}$$

Provided the cumulative distribution function F of μ satisfies (3.23), an immediate consequence of the weak convergence for $\sqrt{n}\left(\hat{F}_n(t) - F(t)\right)$ and the explicit formula

[10] A cost function c is ν-integrable if $\int_{\mathcal{Y}} c(x,y)\,\mathrm{d}\nu(y) < +\infty$ for all $x \in \mathcal{X}$.

[11] The weak convergence $\sqrt{n}\left(\hat{F}_n(t) - F(t)\right) \xrightarrow{\mathcal{D}} \mathbb{B}(F(t))$ in the Skorokhod space $\mathcal{D}(-\infty,+\infty)$ is known as *Donsker's Theorem* and holds without any assumptions on F.

(3.22) is the convergence in law of the empirical OT cost

$$\sqrt{n}\,\mathcal{OT}_{|\cdot|}(\hat{\mu}_n,\mu) \xrightarrow{\mathcal{D}} \int_{\mathbb{R}} |\mathbb{B}(F(t))|\,\mathrm{d}t. \tag{3.24}$$

On the converse, the random sequence $\sqrt{n}\,\mathcal{OT}_{|\cdot|}(\hat{\mu}_n,\mu)$ is stochastically bounded only if (3.23) holds (del Barrio et al., 1999, Thm. 2.1). Although the proof technique by del Barrio et al. (1999) relies on the explicit formula (3.22) in terms of the L_1-distance, the unifying approach presented in this thesis puts statement (3.24) into a dual perspective. First, the limiting random variable in (3.24) is equal in distribution to

$$\int_{\mathbb{R}} |\mathbb{B}(F(t))|\,\mathrm{d}t \overset{D}{=} \sup_{f\in\mathrm{Lip}_1(\mathbb{R})} \mathbb{G}_\mu(f), \tag{3.25}$$

the right-hand side being exactly the limiting random variable in (3.8) for a cost function $c(x,y) = |x-y|$. In particular, the class $\mathrm{Lip}_1(\mathbb{R})$ of Lipschitz functions arises due to Kantorovich-Rubinstein duality[12]. Second, a necessary and sufficient condition for $\mathrm{Lip}_1(\mathbb{R})$ to be μ-Donsker is given in terms of the tail condition

$$\sum_{j\in\mathbb{N}} \sqrt{\mathbb{P}\left(|X| > j\right)} < +\infty \tag{3.26}$$

with a random variable $X \sim \mu$ (Giné and Zinn, 1986, Thm. 1). In fact, condition (3.26) is equivalent to (3.23) (del Barrio et al., 1999, Sec. 2). In other words, the distributional limit law (3.24) is the statistical consequence of Kantorovich-Rubinstein duality with respect to a Lipschitz function class, with (3.23) being the necessary and sufficient condition for the underlying empirical process to weakly converge in the Banach space $l^\infty(\mathrm{Lip}_1(\mathbb{R}))$. Owing to the necessary assumption (**C1**) of bounded costs, for a probability measure $\mu \in \mathcal{P}(\mathbb{R})$ with bounded support, the unifying approach (3.8) leads to the class $\mathrm{BL}_1(\mathbb{R})$ of *bounded* Lipschitz functions. In this case, the tail condition (3.13) on the probability measure μ is necessary and sufficient for $\mathrm{BL}_1(\mathbb{R})$ to be μ-Donsker (Giné and Zinn, 1986, Thm. 2). Condition (3.13) is well-known in empirical process theory (van der Vaart and Wellner, 1996). A sufficient condition is given in terms of finite moments

$$\sum_{k\in\mathbb{Z}} \sqrt{\mathbb{P}\left(X \in [k,k+1)\right)} \le 2\sum_{k\in\mathbb{N}} \sqrt{\mathbb{P}(|X| \ge k)} \le 2\sqrt{\mathbb{E}_\mu\left[|X|^{2+\delta}\right]}\sum_{k\in\mathbb{N}} k^{-(1+\delta/2)},$$

[12] Recall the Kantorovich-Rubinstein dual statement in (1.6). On the real line $\mathbb{R}$ and cost $c(x,y) = |x-y|$ it holds that

$$\mathcal{OT}_{|\cdot|}(\mu,\nu) = \sup_{f\in\mathrm{Lip}_1(\mathbb{R})} \int_{\mathbb{R}} f(x)\,\mathrm{d}(\mu-\nu)(x).$$

the right-hand side being finite[13] if $\mathbb{E}_\mu\left[|X|^{2+\delta}\right] < +\infty$ for some $\delta > 0$.

Still in the univariate setting but for $p > 1$, the asymptotic limit law for the empirical $\mathcal{OT}_{|\cdot|^p}(\hat{\mu}_n, \mu)$ cost in (3.8) usually degenerates. Suppose that the support of $\mu \in \mathcal{P}(\mathbb{R})$ is equal to a bounded and closed interval. Then, the corresponding Kantorovich potential for the dual of $\mathcal{OT}_{|\cdot|^p}(\mu, \mu)$ is unique (Santambrogio, 2015, Prop. 7.18) and *trivial* (cf. (3.10) and surrounding explanations). As a consequence, result (3.9) leads to the degenerate statement $\sqrt{n}\mathcal{OT}_{|\cdot|^2}(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} \mathbb{G}_\mu(f) \sim \delta_0$ and demonstrates the fluctuation of $\mathcal{OT}_{|\cdot|^2}(\hat{\mu}_n, \mu)$ around zero to be of smaller order than $n^{-1/2}$. Indeed, a refined analysis by del Barrio et al. (2005) proves that

$$n\,\mathcal{OT}_{|\cdot|^2}(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} \int_0^1 \mathbb{B}_0^2(u)\,\mathrm{d}u. \tag{3.27}$$

Underlying statement (3.27) is the explicit representation (3.21) together with the weak convergence of the empirical quantile process $\sqrt{n}\left(F_n^{-1}(u) - F^{-1}(u)\right)$ towards $\mathbb{B}_0(u) = \mathbb{B}(u)/f(F^{-1}(u))$ with $\mathbb{B}(t)$ a standard Brownian bridge in the Banach space $L_2(0, 1)$. The latter weak convergence requires rather strong regularity conditions on the cumulative distribution function F (Csörgö and Horváth, 1993), in contrast to the mild integrability assumptions (3.23) for statement (3.24)[14] and $p = 1$. If μ is supported on the interval (a, b), then a sufficient condition for the distributional limit law (3.27) is

$$\sup_{a<x<b} \frac{F(x)(1 - F(x))}{f^2(x)} |f'(x)| < +\infty, \tag{3.28}$$

together with the integrability $\int_{\mathbb{R}} \frac{F(x)(1-F(x))}{f(x)}\,\mathrm{d}x < +\infty$. The work by Berthet and Fort (2019) considers asymptotic limit laws for $1 < p < 2$ with $n^{p/2}\mathcal{OT}_{|\cdot|^p}(\hat{\mu}_n, \mu)$ weakly converging to a limiting random variable following a non-degenerate limit law.

Qualitatively different distributional limit laws for the empirical OT cost on the real line are obtained for $\mu \neq \nu$. The limiting random variable typically follows a centered Gaussian distribution. The work by del Barrio et al. (2019) (see also Munk and Czado (1998); Freitag et al. (2007) for contributions on an empirical trimmed OT cost) prove that under sufficient regularity assumptions on the cumulative distributions functions F

[13] For Euclidean spaces $\mathbb{R}^d$ with $d \geq 2$, condition (3.15) holds if $\mathbb{E}_\mu\left[\|X\|_\infty^{2d+\delta}\right] < +\infty$ for some $\delta > 0$.

[14] In contrast to $p = 1$, even if $\mu \in \mathcal{P}(\mathbb{R})$ is compactly supported, it requires additional assumption on μ for the random sequence $n\,\mathcal{OT}_{|\cdot|^2}(\hat{\mu}_n, \mu)$ to be stochastically bounded (Bobkov and Ledoux, 2019). For general $p \geq 1$, a sufficient condition for $n^{p/2}\mathcal{OT}_{|\cdot|^p}(\hat{\mu}_n, \mu)$ to be stochastically bounded depends on the cumulative distribution function F and density f of μ, i.e.,

$$\int_{\mathbb{R}} \frac{[F(t)(1 - F(t))]^{p/2}}{f(t)^{p-1}}\,\mathrm{d}t < +\infty.$$

For $p > 1$, the finiteness of the integral obviously requires further necessary assumptions on the density.

and G for μ and ν, respectively, it holds that

$$\sqrt{n}\left(\mathcal{OT}_{|\cdot|^2}(\hat{\mu}_n,\nu)-\mathcal{OT}_{|\cdot|^2}(\mu,\nu)\right)\xrightarrow{\mathcal{D}} Z\sim\mathcal{N}\left(0,\sigma^2(F,G)\right). \tag{3.29}$$

The variance $\sigma^2(F,G)$ of the limiting random variable Z is explicit and depends on the derivative of the cost function and corresponding distribution and quantile functions. In particular, the asymptotic variance in (3.29) is equal to

$$\sigma^2(F,G)=4\int_0^1\left(\int_{F^{-1}(\frac{1}{2})}^{F^{-1}(t)}\left(s-G^{-1}(F(s))\right)\mathrm{d}s-\int_0^1\int_{F^{-1}(\frac{1}{2})}^{F^{-1}(s)}\left(u-G^{-1}(F(u))\right)\mathrm{d}u\,\mathrm{d}s\right)^2\mathrm{d}t. \tag{3.30}$$

For $\mu,\nu\in\mathcal{P}(\mathbb{R})$ with bounded support, the unifying approach (3.8) applies and yields, within the same setup as considered in del Barrio et al. (2019), that

$$\sqrt{n}\left(\mathcal{OT}_{|\cdot|^2}(\hat{\mu}_n,\nu)-\mathcal{OT}_{|\cdot|^2}(\mu,\nu)\right)\xrightarrow{\mathcal{D}}\mathbb{G}_\mu(f) \tag{3.31}$$

with f the unique Kantorovich potential for the dual of $\mathcal{OT}_{|\cdot|^2}(\mu,\nu)$. The limiting random variable $\mathbb{G}_\mu(f)$ is distributed according to the centered Gaussian distribution $\mathcal{N}\left(0,\mathrm{Var}_{X\sim\mu}[f(X)]\right)$. Notably, inherent in the assumed regularity conditions employed by del Barrio et al. (2019) is a nonvanishing density for μ inside the bounded, connected interior of its support. This assumption imposes a unique Kantorovich potential f (Santambrogio, 2015, Prop. 7.18) and consequentially a centered Gaussian limit law in (3.31). A straightforward calculation[15] shows that the asymptotic variance $\sigma^2(F,G)$ in (3.29) is equal to $\mathrm{Var}_{X\sim\mu}[f(X)]$. In particular, statement (3.29) can also be obtained by an application of the unifying approach leading to (3.31). This details the connection between del Barrio et al. (2019) and the approach by Hundrieser et al. (2021c) presented here. Similar are recent results by Berthet et al. (2020) who consider more general *good*[16] cost functions. In short, if $\mu,\nu\in\mathcal{P}(\mathbb{R})$ are sufficiently different and their corresponding quantile functions regular enough, then $\sqrt{n}\,(\mathcal{OT}_c(\hat{\mu}_n,\nu)-\mathcal{OT}_c(\mu,\nu))$ asymptotically follows a centered Gaussian law as n tends to infinity. The regularity conditions again promote a unique Kantorovich potential and, in line with the unifying approach, limiting random variables following a centered Gaussian distribution are

[15] According to Santambrogio (2015, Sec. 1.3), if f is a Kantorovich potential for squared Euclidean cost, then $f'(x_0)=h'(x_0-y_0)$ with $h(x)=x^2$ for all (x_0,y_0) in the support of an optimal transport plan. Since $T(x)=G^{-1}(F(x))$ is the unique OT map, it follows that $2\left(s-G^{-1}(F(s))\right)=f'(s)$.

[16] According to Berthet et al. (2020), any cost function is considered as *good* if it allows for an explicit representation of the form

$$\mathcal{OT}_c(\mu,\nu)=\int_0^1 c\left(F^{-1}(u),G^{-1}(u)\right)\mathrm{d}u.$$

expected. Distributional limit laws for probability measures supported on the circle with qualitatively similar results as outlined for the real line are the content of the work by Hundrieser et al. (2021b).

Euclidean spaces $\mathbb{R}^{\mathbf{d}}$, $\mathbf{d} \geq 2$: Beyond the real line, distributional limit results for the empirical OT cost between probability measures supported on high-dimensional Euclidean spaces $\mathbb{R}^d$ with $d \geq 2$ are much more involved. To witness the fundamental problem in high-dimensional Euclidean spaces, suppose that μ is a uniform distribution on the unit cube $[0,1]^d$. As demonstrated by the unifying approach, weak convergence for the empirical $\mathcal{OT}_{\|\cdot\|}(\hat{\mu}_n, \mu)$ cost with $p = 1$ is strongly linked to μ-Donsker properties for $\mathrm{BL}_1(\mathbb{R}^d)$ the class of bounded Lipschitz functions emerging via Kantorovich-Rubinstein duality. As Strassen and Dudley (1969) point out, already for $d = 2$, the class $\mathrm{BL}_1(\mathbb{R}^2)$ is not even μ-Pregaussian and hence the Donsker property fails to hold (van der Vaart and Wellner, 1996, Sec. 2.1.2). Further, Ajtai et al. (1984) (see Bobkov and Ledoux (2021) for slight generalizations) prove that for μ the uniform distribution on the unit square ($d = 2$) and with high probability

$$c\left(\frac{\log(n)}{n}\right)^{1/2} \leq \mathcal{OT}_{\|\cdot\|}(\hat{\mu}_n, \mu) \leq C\left(\frac{\log(n)}{n}\right)^{1/2} \tag{3.32}$$

for c and C two universal constants. The bound (3.32) proves that the random sequence $\sqrt{n}\mathcal{OT}_{\|\cdot\|}(\hat{\mu}_n, \mu)$ is not stochastically bounded and hence cannot converge weakly. Similarly and for $d \geq 3$, the lower bound $\mathcal{OT}_{\|\cdot\|}(\hat{\mu}_n, \mu) \geq n^{-1/d}$ by Dudley (1969)[17] immediately implies the rate $\sqrt{n}$ to be of wrong order to obtain some non-degenerate limit law. Inherent to these observations is the by-now classical fact that the empirical OT cost suffers the *curse of dimensionality*. For $d > 2p$ and under no additional assumptions on μ, the expected empirical OT cost does not converge faster to zero than $\mathbb{E}_\mu\left[\mathcal{OT}_{\|\cdot\|^p}(\hat{\mu}_n, \mu)\right] \geq n^{-p/d}$ (Fournier and Guillin, 2015; Weed and Bach, 2019). Nonetheless, the empirical OT cost has excellent concentration properties around its expectation (Weed and Bach, 2019). Based on McDiarmid's bounded difference inequality, if $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ have support bounded on a set of diameter D, then it holds

$$\mathbb{P}\left(\left|\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu) - \mathbb{E}_\mu\left[\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu)\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{D^4}\right). \tag{3.33}$$

The concentration inequality in (3.33) implies that the random sequence given by $\sqrt{n}\left(\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu) - \mathbb{E}_\mu\left[\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu)\right]\right)$ is stochastically bounded. A fundamental step further has been taken by del Barrio and Loubes (2019). If $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ are probability

[17] The lower bound by Dudley (1969) does not require a probability statement and holds for any measure μ_n concentrated on n points with $n \geq n_0$ for some $n_0 \in \mathbb{N}$.

measures with positive density in the interior of their convex support and have finite moments of order $4+\delta$ for some $\delta > 0$, then it holds

$$\sqrt{n}\left(\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu) - \mathbb{E}_\mu\left[\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu)\right]\right) \xrightarrow{\mathcal{D}} \mathbb{G}_\mu(f) \tag{3.34}$$

with f the unique Kantorovich potential for the dual of $\mathcal{OT}_{\|\cdot\|^2}(\mu, \nu)$. The limiting random variable follows the centered Gaussian distribution given by $\mathcal{N}\left(0, \mathrm{Var}_{X\sim\mu}\left[f(X)\right]\right)$. The proof technique for statement (3.34) relies on the *Efron-Stein* variance inequality. In particular, a suitable chosen linear functional, motivated by Kantorovich duality (3.1) and based on the empirical measure, approximates the left hand side in (3.34) in L^2. An application of a standard central limit theorem applied to the linear functional concludes the proof. This approach crucially depends on the uniqueness of the Kantorovich potential that is enforced by the underlying regularity conditions on the probability measures[18]. Statement (3.34) is reminiscent to the main result in (3.9). Indeed, within the framework considered by del Barrio and Loubes (2019), but for probability measures with bounded support on $\mathbb{R}^d$ and $d = 2, 3$, statement (3.34) is included by the unifying approach. Notably, the central limit theorem in (3.9) is centered around the *population* quantity $\mathcal{OT}_{\|\cdot\|^2}(\mu, \nu)$, rather than the empirical expectation $\mathbb{E}_\mu\left[\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu)\right]$.
From a statistical perspective, centering around the population quantity is desirable and raises the fundamental question of analogous statements in (3.9) for measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $d \geq 4$ not covered by the unifying approach. However, for $d = 4$ and uniform measures on different unit balls centered around $x \neq y \in \mathbb{R}^4$, Chizat et al. (2020); Manole and Niles-Weed (2021) give upper and lower bounds on the bias

$$n^{-1/2} \lesssim \mathbb{E}_\mu\left[\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu)\right] - \mathcal{OT}_{\|\cdot\|^2}(\mu, \nu) \lesssim \log(n) n^{-1/2}. \tag{3.35}$$

From these bounds and the central limit law in (3.34), it follows that the random sequence $\sqrt{n}(\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu) - \mathcal{OT}_{\|\cdot\|^2}(\mu, \nu))$ either is not stochastically bounded or asymptotically follows a non-centered Gaussian distribution. Further and for $d \geq 5$, the bias is lower bounded by

$$\mathbb{E}_\mu\left[\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu)\right] - \mathcal{OT}_{\|\cdot\|^2}(\mu, \nu) \gtrsim n^{-2/d}. \tag{3.36}$$

This implies that the expectation of the empirical quantity in (3.34) can generally not be replaced by the population quantity in dimensions $d \geq 5$.
For $\mu = \nu$, the limit law in (3.34) degenerates which is in line with the main theory presented in this thesis. Indeed, under the required regularity conditions in del Barrio

[18] A similar proof strategy was employed by Mena and Niles-Weed (2019) to obtain central limit theorems for the empirical entropy regularized OT cost as discussed in Section 2.1 (see (2.37) for details).

and Loubes (2019), the unique Kantorovich potential f is trivial. Hence, the limiting random variable $\mathbb{G}_\mu(f) \sim \delta_0$ follows a Dirac distribution at zero. Apart from the squared Euclidean cost ($p = 2$), central limit theorems in the fashion of (3.34) with empirical expectation as centering constant but for more general convex cost functions are available in the work by del Barrio et al. (2021b). Overall, for Euclidean spaces $\mathbb{R}^d$ with dimension $d = 1, 2, 3$ and equipped with sufficiently regular and bounded cost function, an application of the unifying approach retrieves their main results with the notable exception of centering around the population quantity. For dimension $d \geq 5$, the random sequence $\sqrt{n}\left(\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu) - \mathcal{OT}_{\|\cdot\|^2}(\mu, \nu)\right)$ does not asymptotically follow a non-degenerate limit law. Indeed, in high-dimensional Euclidean regimes the random fluctuation of $\mathcal{OT}_{\|\cdot\|^2}(\hat{\mu}_n, \nu) - \mathcal{OT}_{\|\cdot\|^2}(\mu, \nu)$ around zero appears to be of different order than $n^{-1/2}$.

Extensions and alternative approaches: Beyond the plug-in approach based on empirical probability measures, there exist explicit central limit distributions for the empirical OT cost when employing different estimators. A particular case arises for Gaussian measures (more generally for elliptical distributions) with support even on high-dimensional Euclidean spaces and squared Euclidean cost. Crucial to these asymptotic statements is an observations similar to the univariate case. There exists a closed formula related to the *Bures metric* between positive semi-definite matrices: If $\mu = \mathcal{N}(\theta_\mu, \Sigma_\mu)$ and $\nu = \mathcal{N}(\theta_\nu, \Sigma_\nu)$ are two Gaussian distributions on the Euclidean space $\mathbb{R}^d$, then it holds

$$\mathcal{OT}_{\|\cdot\|^2}\left(\mathcal{N}(\theta_\mu, \Sigma_\mu), \mathcal{N}(\theta_\nu, \Sigma_\nu)\right) = \|\theta_\mu - \theta_\nu\|^2 + \operatorname{tr}\left(\Sigma_\mu + \Sigma_\nu - 2\left(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}}\right)^{\frac{1}{2}}\right). \tag{3.37}$$

Rather than employing the empirical measure, the Gaussian measure is approximated by a suitable estimator $\hat{\theta}_{\mu,n}$ for the mean and $\hat{\Sigma}_{\mu,n}$ for the covariance matrix, respectively. Employing standard central limit theorems for these estimators combined with a delta method, Rippl et al. (2016) prove the weak convergence of the random sequence

$$\sqrt{n}\left(\mathcal{OT}_{\|\cdot\|^2}\left(\mathcal{N}(\hat{\theta}_{\mu,n}, \hat{\Sigma}_{\mu,n}), \mathcal{N}(\theta_\nu, \Sigma_\nu)\right) - \mathcal{OT}_{\|\cdot\|^2}\left(\mathcal{N}(\theta_\mu, \Sigma_\mu), \mathcal{N}(\theta_\nu, \Sigma_\nu)\right)\right) \tag{3.38}$$

to a certain limiting random variable as the sample size n tends to infinity. The limiting random variable follows a centered Gaussian distribution. For Gaussian measures $\mu = \nu$, the limit distribution is degenerate and equal to a Dirac at zero. When scaling with rate n instead of $\sqrt{n}$, the random sequence $n\,\mathcal{OT}_{\|\cdot\|^2}\left(\mathcal{N}(\hat{\theta}_{\mu,n}, \hat{\Sigma}_{\mu,n}), \mathcal{N}\left(\theta_\mu, \Sigma_\mu\right)\right)$ weakly converges to a non-degenerate but also non-normal limit distribution.

Further, a recent trend aiming to overcome the curse of dimensionality for empirical OT based quantities employs smoothed and wavelet based estimators (Weed and Berthet, 2019; Deb et al., 2021). Different to the lower bound on the bias in (3.36), Manole et al. (2021) prove that if $\mu \neq \nu$ are probability measures in $\mathcal{P}([0,1]^d)$) and absolutely continuous with sufficiently regular density g and h, respectively, then the bias is of order

$$\mathbb{E}_\mu \left[\mathcal{OT}_{\|\cdot\|^2}(\hat{g}_n, h) \right] - \mathcal{OT}_{\|\cdot\|^2}(g,h) = o\left(n^{-1/2}\right). \tag{3.39}$$

The estimator $\hat{g}_n$ for the density g is a certain boundary corrected wavelet density estimator. Together with the approach by del Barrio et al. (2019), based on the Efron-Stein variance inequality, it concludes the central limit theorem

$$\sqrt{n}\left(\mathcal{OT}_{\|\cdot\|^2}(\hat{g}_n, h) - \mathcal{OT}_{\|\cdot\|^2}(g,h) \right) \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}\left(0, \sigma^2\right). \tag{3.40}$$

The variance of the asymptotic distribution is equal to $\sigma^2 = \mathrm{Var}_{X\sim\mu}\left[f(X)\right]$ with f the unique Kantorovich potential of the corresponding dual problem for $\mathcal{OT}_{\|\cdot\|^2}(g,h)$. For $\mu = \nu$, the limit law for (3.40) degenerates and a more refined analysis remains open. For usual Euclidean cost, $p = 1$ and $\mu = \nu$, a non-degenerate limit law is obtained for the empirical *Gaussian smoothed* OT cost

$$\mathcal{OT}^\sigma_{\|\cdot\|}(\hat{\mu}_n, \mu) := \mathcal{OT}_{\|\cdot\|}(\hat{\mu}_n * \mathcal{N}(0, \sigma I_d), \mu * \mathcal{N}(0, \sigma I_d))$$

and parameter $\sigma > 0$ as defined by Goldfeld and Greenewald (2020). The work by Sadhu et al. (2021) proves that if

$$\sum_{k\in\mathbb{Z}^d} \|k\| \sqrt{\mathbb{P}\left(X \in [k, k+1)\right)} < +\infty \tag{3.41}$$

for a random vector $X \sim \mu$, then the weak convergence for the empirical Gaussian smoothed OT cost holds,

$$\sqrt{n}\mathcal{OT}^\sigma_{\|\cdot\|}(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} \sup_{f\in \mathrm{Lip}_1(\mathbb{R}^d)} |\mathbb{G}_\mu(f * \phi_\sigma)|, \tag{3.42}$$

with ϕ_σ the density of the isotropic d-dimensional Gaussian distribution with parameter $\sigma > 0$. The limiting random variable follows a non-degenerate limit law. The proof technique is strongly reminiscent to the approach by Hundrieser et al. (2021c). Indeed, assumption (3.41) on μ, a weighted version of (3.15), guarantees the smooth function class $\mathcal{F}_\sigma := \left\{ f * \phi_\sigma \mid f \in \mathrm{Lip}_1(\mathbb{R}^d) \right\}$ appearing in the dual formulation of $\mathcal{OT}^\sigma_{\|\cdot\|}(\mu,\mu)$ to be μ-Donsker. The weak convergence then follows by an application of the continuous mapping theorem with limiting random variable equal to $\sup_{f\in\mathcal{F}_\sigma} |\mathbb{G}_\mu(f)|$ and $\mathbb{G}_\mu$ a μ-Brownian bridge process defined on $l^\infty(\mathcal{F}_\sigma)$. This is to be compared with statements

(3.24) and (3.25) only valid on the real line ($d = 1$), while the Gaussian smoothing allows for similar results in high-dimensional Euclidean spaces $d > 1$. The case $\mu \neq \nu$ is dealt with a functional delta method and leads to non-degenerate limit laws.

Conclusion: Distributional limit laws for the empirical OT cost are qualitatively distinguished between[19] the *null* $\mu = \nu$ and the *alternative* case $\mu \neq \nu$. For the latter, the obtained limiting random variables often follow a centered Gaussian distribution. The null $\mu = \nu$ is more challenging and the limit laws usually degenerate to a Dirac at zero. This observation is in line with the general theory presented in this thesis. Indeed, for $\mu = \nu$ degeneracy of the limit law follows from unique and trivial Kantorovich potentials, whereas Gaussian limits for $\mu \neq \nu$ appear as a consequence of unique but non-trivial Kantorovich potentials. For bounded and sufficiently regular cost functions the unifying approach (3.8) encompasses known distributional limit laws on countable discrete spaces as well as on the real line. An interesting and *novel* statement is obtained on Euclidean spaces $\mathbb{R}^d$ with dimension $d = 2, 3$. While previous distributional limit laws are centered around empirical expectations, the asymptotic statements obtained in this thesis are centered around the population quantity. Beyond these low-dimensional Euclidean settings, the distributional limit laws in (3.8) generally fail to hold when employing empirical measures in high-dimensional Euclidean regimes ($d \geq 5$).

Overall, the unifying approach (3.8) presented in this thesis points towards two fundamental principles underlying all distributional limit laws for the empirical OT cost: *Kantorovich duality* and *weak convergence of underlying empirical processes*. Indeed, common to all central limit type results are certain assumptions on the probability measures. In view of the unifying approach, these assumption arise naturally from empirical process theory to guarantee weak convergence of the empirical process $\sqrt{n}(\hat{\mu}_n - \mu)$ in a suitable normed space $l^\infty(\mathcal{F}_c)$. The function class $\mathcal{F}_c$ stems from Kantorovich duality that is the main machinery behind all proof techniques for distributional limit laws of the empirical OT cost covered by the literature. Overall, there is a noteworthy trade-off inherent in distributional limit laws for the empirical OT cost: Kantorovich duality requires $\mathcal{F}_c$ to be sufficiently rich, while the empirical process only converges weakly if $\mathcal{F}_c$ is not too complex. This phenomenon essentially restricts the asymptotic statements to low-dimensional Euclidean spaces. The restriction is the analogue of the curse of dimensionality in distributional limit laws for the empirical OT cost when centering around the population quantity. Recently proposed alternatives employ different esti-

[19] Explicit limiting results also depend on the underlying sampling situation, e.g., the two-sample case where simultaneously μ and ν are estimated by their respective empirical measures. Analogously to the presentation of the main results, the discussion is tailored to the one-sample case.

mators for the probability measures or Gaussian smoothing leading to non-degenerate asymptotic statements beyond low-dimensional Euclidean spaces. The unifying approach based on a functional delta method is able to cope with these approaches as long as underlying empirical processes weakly converge in suitable normed spaces. Lastly, it appears reasonable to conjecture the main result in (3.8) to also hold for *unbounded* cost functions, although the presented theory leaves open a thorough treatment for this.

Bibliography

Ajtai, M., Komlós, J., and Tusnády, G. (1984). On optimal matchings. *Combinatorica*, 4(4):259–264.

Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, pages 1964–1974.

Ambrosio, L. (2003). Lecture notes on optimal transport problems. In *Mathematical Aspects of Evolving Interfaces*, volume 1812, pages 1–52. Springer.

Ambrosio, L. and Pratelli, A. (2003). Existence and stability results in the L_1 theory of optimal transportation. In *Optimal Transportation and Applications*, volume 1813, pages 123–160. Springer.

Beale, E. M. (1955). On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):173–184.

Bercu, B. and Bigot, J. (2021). Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *The Annals of Statistics*, 49(2):968–987.

Berthet, P. and Fort, J.-C. (2019). Weak convergence of empirical Wasserstein type distances. *preprint arXiv:1911.02389*.

Berthet, P., Fort, J.-C., and Klein, T. (2020). A central limit theorem for Wasserstein type distances between two distinct univariate distributions. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(2):954 – 982.

Bertsekas, D. P. and Castanon, D. A. (1989). The auction algorithm for the transportation problem. *Annals of Operations Research*, 20(1):67–96.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, Massachusetts.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Bigot, J., Cazelles, E., and Papadakis, N. (2019). Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2):5120–5150.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Birge, J. R. and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.

Birkhoff, G. (1946). Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán, Serie A*, 5:147–154.

Bobkov, S. and Ledoux, M. (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society.

Bobkov, S. G. and Ledoux, M. (2021). A simple Fourier analytic proof of the AKT optimal matching theorem. *The Annals of Applied Probability*, 31(6):2567–2584.

Bonnans, J. F. and Shapiro, A. (2013). *Perturbation analysis of optimization problems*. Springer Science & Business Media.

Bonnotte, N. (2013). *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris 11.

Brenier, Y. (1987). Décomposition polaire et réarrangement monotone des champs de vecteurs. *Comptes Rendus de l'Académie des Sciences Paris Séries I Mathematics*, 305:805–808.

Burkard, R. E. and Cela, E. (1999). Linear assignment problems and extensions. In *Handbook of Combinatorial Optimization*, volume A, pages 75–149. Springer.

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

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

Cottle, R., Johnson, E., and Wets, R. (2007). George B. Dantzig (1914–2005). *Notices of the American Mathematical Society*, 54(3):344–362.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Csörgö, M. and Horváth, L. (1993). *Weighted approximations in probability and statistics*. John Wiley & Sons.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300.

Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, volume 32, pages 685–693. Proceedings of Machine Learning Research.

Dantzig, G. B. (1949). Programming in a linear structure. *Econometrica*, 17:73–74.

Dantzig, G. B. (1955). Linear programming under uncertainty. *Management Science*, 1(3-4):197–206.

De Loera, J., Rambau, J., and Santos, F. (2010). *Triangulations: Structures for algorithms and applications*, volume 25. Springer Science & Business Media.

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

del Barrio, E., Giné, E., and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27(2):1009–1071.

del Barrio, E., Giné, E., and Utzet, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189.

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

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

del Barrio, E., Gordaliza, P., and Loubes, J.-M. (2019). A central limit theorem for L_p transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8(4):817–849.

del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951.

Dobrić, V. and Yukich, J. E. (1995). Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118.

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

Dudley, R. M. (2014). *Uniform central limit theorems*, volume 142. Cambridge University Press.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95(1):125–140.

Eichhorn, A. and Römisch, W. (2007). Stochastic integer programming: Limit theorems and confidence intervals. *Mathematics of Operations Research*, 32(1):118–135.

Evans, L. C. and Gangbo, W. (1999). Differential equations methods for the Monge-Kantorovich mass transfer problem. *Memoirs American Mathematical Society*, 137(653).

Ewbank, J. B., Foote, B. L., and Kumin, H. L. (1974). A method for the solution of the distribution problem of stochastic linear programming. *SIAM Journal on Applied Mathematics*, 26(2):225–238.

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

Ferguson, A. R. and Dantzig, G. B. (1956). The allocation of aircraft to routes–An example of linear programming under uncertain demand. *Management Science*, 3(1):45–73.

Fiacco, A. V. (1983). *Introduction to sensitivity and stability analysis in nonlinear programming*, volume 165. Academic Press.

Flamary, R., Lounici, K., and Ferrari, A. (2019). Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation. *preprint arXiv:1905.10155*.

Ford, L. R. and Fulkerson, D. R. (1956). Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404.

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

Freitag, G., Czado, C., and Munk, A. (2007). A nonparametric test for similarity of marginals–With applications to the assessment of population bioequivalence. *Journal of Statistical Planning and Inference*, 137(3):697–711.

Gal, T. (1985). On the structure of the set bases of a degenerate point. *Journal of Optimization Theory and Applications*, 45(4):577–589.

Gal, T. (1986). Shadow prices and sensitivity analysis in linear programming under degeneracy. *Operations-Research-Spektrum*, 8(2):59–71.

Gal, T. and Greenberg, H. J. (2012). *Advances in sensitivity analysis and parametric programming*, volume 6. Springer Science & Business Media.

Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1574–1583. Proceedings of Machine Learning Research.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617. Proceedings of Machine Learning Research.

Giné, E. and Zinn, J. (1986). Empirical processes indexed by Lipschitz functions. *The Annals of Probability*, 14(4):1329–1338.

Goldfeld, Z. and Greenewald, K. (2020). Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pages 3327–3337. Proceedings of Machine Learning Research.

Greenberg, H. J. (1986). An analysis of degeneracy. *Naval Research Logistics Quarterly*, 33(4):635–655.

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

Hundrieser, S., Klatt, M., and Munk, A. (2021a). Limit distributions and sensitivity analysis for entropic optimal transport on countable spaces. *preprint arXiv:2105.00049*.

Hundrieser, S., Klatt, M., and Munk, A. (2021b). The statistics of circular optimal transport. In *Directional Statistics for Innovative Applications*. Springer. To appear [preprint arXiv:2103.15426].

Hundrieser, S., Klatt, M., Staudt, T., and Munk, A. (2021c). A unifying approach to central limit theorems for empirical optimal transport. *In preparation*.

Hütter, J.-C. and Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194.

Jansen, B., De Jong, J., Roos, C., and Terlaky, T. (1997). Sensitivity analysis in linear programming: just be careful! *European Journal of Operational Research*, 101(1):15–28.

Kantorovich, L. V. (1942). On the transfer of masses. *Doklady Akademii Nauk USSR*, 37(2):227–229.

Klatt, M., Munk, A., and Zemel, Y. (2020a). Limit laws for empirical optimal solutions in stochastic linear programs. *preprint arXiv:2007.13473*.

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

Klee, V. and Minty, G. J. (1972). How good is the simplex algorithm. *Inequalities*, 3(3):159–175.

Klee, V. and Witzgall, C. (1968). Facets and vertices of transportation polytopes. *Mathematics of the Decision Sciences*, 1:257–282.

Koltai, T. and Terlaky, T. (2000). The difference between the managerial and mathematical interpretation of sensitivity analysis results in linear programming. *International Journal of Production Economics*, 65(3):257–274.

Koopmans, T. C. (1951). Efficient allocation of resources. *Econometrica: Journal of the Econometric Society*, 19(4):455–465.

Kovačević, M., Stanojević, I., and Šenk, V. (2015). On the entropy of couplings. *Information and Computation*, 242:369–382.

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

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. (2019). Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, pages 12283–12294.

Luenberger, D. G. and Ye, Y. (1984). *Linear and nonlinear programming*, volume 2. Springer.

Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. (2018). Differential properties of Sinkhorn approximation for learning with Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 31, pages 5864–5874.

Mallows, C. L. (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 43(2):508–515.

Manne, A. S. (1953). *Notes on parametric linear programming*. Rand Corporation.

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

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

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32, page 4543–4553.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):223–241.

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

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

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory, Volume II: Applications*. Springer Science & Business Media.

Rachev, S. T., Stoyanov, S. V., and Fabozzi, F. J. (2011). *A probability metrics approach to financial risk measures*. John Wiley & Sons.

Ramachandran, D. and Rüschendorf, L. (1995). A general duality theorem for marginal problems. *Probability Theory and Related Fields*, 101(3):311–319.

Rippl, T., Munk, A., and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109.

Römisch, W. (2004). Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*, volume 16, pages 1575–1583. New York: Wiley.

Rubinstein, R. Y. and Shapiro, A. (1993). *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method*, volume 13. Wiley.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Rüschendorf, L. and Rachev, S. T. (1990). A characterization of random variables with minimum L_2-distance. *Journal of Multivariate Analysis*, 32(1):48–54.

Ruszczyński, A. and Shapiro, A. (2003). Stochastic programming models. *Handbooks in Operations Research and Management Science*, 10:1–64.

Saaty, T. and Gass, S. (1954). Parametric objective function (part 1). *Journal of the Operations Research Society of America*, 2(3):316–319.

Sadhu, R., Goldfeld, Z., and Kato, K. (2021). Limit distribution theory for the smooth 1-Wasserstein distance with applications. *preprint arXiv:2107.13494*.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkhäuser, New York*, 55(58-63):94.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2021). *Lectures on stochastic programming: modeling and theory*. Society for Industrial and Applied Mathematics.

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879.

Smith, C. S. and Knott, M. (1987). Note on the optimal transportation of distributions. *Journal of Optimization Theory and Applications*, 52(2):323–329.

Sommerfeld, M. (2017). *Wasserstein distance on finite spaces: Statistical inference and algorithms*. PhD thesis, Georg-August-Universität Göttingen.

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

Sommerfeld, M., Schrieber, J., Zemel, Y., and Munk, A. (2019). Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23.

Spielman, D. A. and Teng, S.-H. (2004). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the Association for Computing Machinery*, 51(3):385–463.

Strassen, V. and Dudley, R. (1969). The central limit theorem and ε-entropy. In *Probability and Information Theory*, volume 89, pages 224–231. Springer.

Sturmfels, B. and Venturello, L. (2020). Optimal transport to a variety. In *Mathematical Aspects of Computer and Information Sciences*, volume 11989, page 364. Springer Nature.

Sudakov, V. N. (1979). *Geometric problems in the theory of infinite-dimensional probability distributions*, volume 141. American Mathematical Society.

Talagrand, M. (1992). Matching random samples in many dimensions. *The Annals of Applied Probability*, 2(4):846–856.

Tameling, C., Sommerfeld, M., and Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5):2744–2781.

Tameling, C., Stoldt, S., Stephan, T., Naas, J., Jakobs, S., and Munk, A. (2021). Colocalization for super-resolution microscopy via optimal transport. *Nature Computational Science*, 1(3):199–211.

Tolstoí, A. (1930). Methods of finding the minimal total kilometrage in cargo transportation planning in space. *TransPress of the National Commissariat of Transportation*, 1:23–55.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Varadarajan, V. S. (1958). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26.

Vaserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

Vershik, A. (2002). L.V. Kantorovich and linear programming. In *Leonid Vital'evich Kantorovich: A Man and a Scientist*, volume 1, pages 130–152.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Walkup, D. and Wets, R. (1969). Lifting projections of convex polyhedra. *Pacific Journal of Mathematics*, 28(2):465–475.

Weed, J. (2018). An explicit analysis of the entropic penalty in linear programming. In *Conference on Learning Theory*, volume 75, pages 1841–1855. Proceedings of Machine Learning Research.

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

Weed, J. and Berthet, Q. (2019). Estimation of smooth densities in Wasserstein distance. In *Conference on Learning Theory*, volume 99, pages 3118–3119. Proceedings of Machine Learning Research.

Wets, R. (1980). The distribution problem and its relation to other problems in stochastic programming. *Stochastic Programming, Academic Press, London*, pages 245–262.

Addenda

The following addenda contain the three articles A, B and C that form the basis of this thesis. An introductory summary lists each article's reference and abstract.

Limit Laws for Empirical Optimal Solutions in Random Linear Programs
Marcel Klatt, Yoav Zemel, and Axel Munk
preprint available, arXiv:2007.13473 (2020)

Abstract We consider a general linear program in standard form whose right-hand side constraint vector is subject to random perturbations. For the corresponding random linear program, we characterize under general assumptions the random fluctuations of the empirical optimal solutions around their population quantities after standardization by a distributional limit theorem. Our approach is geometric in nature and further relies on duality and the collection of dual feasible basic solutions. The limiting random variables are driven by the amount of degeneracy inherent in linear programming. In particular, if the corresponding dual linear program is degenerate the asymptotic limit law might not be unique and is determined from the way the empirical optimal solution is chosen. Furthermore, we include consistency and convergence rates of the Hausdorff distance between the empirical and the true optimality sets as well as a limit law for the empirical optimal value involving the set of all dual optimal basic solutions.
Our analysis is motivated from statistical optimal transport that is of particular interest here and follows by a simple application of our general theory. The corresponding limit distribution is usually non-Gaussian which stands in strong contrast to recent finding for empirical entropy regularized optimal transport solutions.

Empirical Regularized Optimal Transport: Statistical Theory and Applications
Marcel Klatt, Carla Tameling, and Axel Munk
SIAM Journal on Mathematics of Data Science 2(2):419 – 443 *(2020)*
`https://epubs.siam.org/doi/pdf/10.1137/19M1278788`

Abstract We derive limit distributions for various empirical regularized optimal transport quantities between probability distributions supported on a finite metric space and show their bootstrap consistency. In particular, we prove that the empirical regularized transport plan itself asymptotically follows a Gaussian law. The theory includes the Boltzmann–Shannon entropy regularization and hence a limit law for the widely applied Sinkhorn divergence. Our approach is based on parametric optimization techniques for the regularized transport problem in conjunction with a statistical delta method. The asymptotic results are investigated in Monte Carlo simulations. We further discuss computational consequences and statistical applications, e.g., confidence bands for colocalization analysis of protein interaction networks based on regularized optimal transport.

A Unifying Approach to Central Limit Theorems for Optimal Transport
Shayan Hundrieser, Marcel Klatt, Thomas Staudt, and Axel Munk
in preparation

Abstract We provide a unifying approach to *central limit theorems* (CLTs) for empirical optimal transport (OT). At the heart of our theory is Kantorovich duality representing OT as a supremum over a function class $\mathcal{F}_c$ for an underlying sufficiently regular cost function c. In this regard, OT is considered as a functional defined on $l^\infty(\mathcal{F}_c)$ the Banach space of bounded functionals from $\mathcal{F}_c$ to $\mathbb{R}$. We prove its *Hadamard directional differentiability* and conclude via a *functional delta method* that necessitates weak convergence of an underlying empirical process in $l^\infty(\mathcal{F}_c)$. The latter can be dealt with *empirical process theory* and requires $\mathcal{F}_c$ to be *Donsker*. We provide sufficient conditions depending on the dimension of the ground space, the underlying cost function and the probability measures under consideration such that a Donsker property holds. Overall, our approach reveals a noteworthy trade-off inherent in CLTs for empirical OT: Kantorovich duality requires $\mathcal{F}_c$ to be sufficiently rich, while the empirical processes only converges weakly if $\mathcal{F}_c$ is not too complex. The limit distribution of the empirical OT cost is characterized as a supremum of a Gaussian process. In particular, we discuss when the limit distribution is centered normal or degenerates to a Dirac measure at zero. Our approach covers the situation when at least one of the measures has discrete or low dimensional support ($d \leq 3$). This encompasses well-known results for discrete and semi-discrete transport. Moreover, in contrast to recent contributions on distributional limit laws for empirical OT on Euclidean spaces which require centering around its expectation, the CLTs obtained here are centered around the *population* quantity which is well-suited for statistical applications.

CHAPTER A

Limit Laws for Empirical Optimal Solutions in Random Linear Programs

Limit Laws for Empirical Optimal Solutions in Random Linear Programs

Marcel Klatt [*] Yoav Zemel [†] Axel Munk [*‡]

Abstract

We consider a general linear program in standard form whose right-hand side constraint vector is subject to random perturbations. For the corresponding random linear program, we characterize under general assumptions the random fluctuations of the empirical optimal solutions around their population quantities after standardization by a distributional limit theorem. Our approach is geometric in nature and further relies on duality and the collection of dual feasible basic solutions. The limiting random variables are driven by the amount of degeneracy inherent in linear programming. In particular, if the corresponding dual linear program is degenerate the asymptotic limit law might not be unique and is determined from the way the empirical optimal solution is chosen. Furthermore, we include consistency and convergence rates of the Hausdorff distance between the empirical and the true optimality sets as well as a limit law for the empirical optimal value involving the set of all dual optimal basic solutions.

Our analysis is motivated from statistical optimal transport that is of particular interest here and follows by a simple application of our general theory. The corresponding limit distribution is usually non-Gaussian which stands in strong contrast to recent finding for empirical entropy regularized optimal transport solutions.

Keywords: Limit law, Linear programming, Optimal transport, Sensitivity analysis
MSC 2020 subject classification 90C05, 90C15, 90C31, 62E20, 49N15

1 Introduction

Linear programs arise naturally in many applications and have become ubiquitous in topics such as operations research, control theory, economics, physics, mathematics and statistics (see the textbooks by Bertsimas & Tsitsiklis (1997), Luenberger & Ye (2008), Dantzig (2016), Galichon (2018) and the references therein). Their solid mathematical foundation dates back to the mid-twentieth century, to mention the seminal works of Kantorovich (1939), Hitchcock (1941), Dantzig (1948) and Koopmans (1949), and its algorithmic computation is an active topic of research until today[1]. A linear program in standard form writes

$$\nu(b) \coloneqq \min_{x\in\mathbb{R}^d} c^T x \quad \text{s.t.} \quad Ax = b,\, x \geq 0, \tag{P$_b$}$$

with $(A, b, c) \in \mathbb{R}^{m\times d} \times \mathbb{R}^m \times \mathbb{R}^d$ and matrix A of full rank $m \leq d$. For the purpose of the paper, the lower subscript b in (P$_b$) emphasizes the dependence on the vector b. Associated

[*] Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen
[†] Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB
[‡] Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen

[1] For a detailed historical account see Vershik (2002) and Cottle et al. (2007) with particular focus on Kantorovich's and Dantzig's fundamental contributions to the field, respectively.

with the *primal* program (P_b) is its corresponding *dual* program

$$\max_{\lambda \in \mathbb{R}^m} \quad b^T \lambda \qquad \text{s.t.} \quad \lambda^T A \leq c^T. \tag{D_b}$$

At the heart of linear programming and fundamental to our work is the observation that if the primal program (P_b) attains a finite value, the optimum is attained at one of a finite set of candidates termed *basic* solutions. Each basic solution (possibly infeasible) is identified by a basis $I \subset \{1, \ldots, d\}$ indexing m linearly independent columns of the constraint matrix A. The bases I also defines a basic solution for the dual (D_b). In fact, the simplex algorithm (Dantzig, 1948) is specifically designed to move from one primal *feasible* basic solution to another while checking if the corresponding basis induces a dual feasible basic solution. Shortly after first algorithmic approaches and theoretical results became available, the need to incorporate uncertainty in the parameters has become apparent (see Dantzig (1955); Beale (1955); Ferguson & Dantzig (1956) for early contributions). In fact, apart from its relevance in numerical stability issues, in many applications the parameters reflect practical needs (budget, prices or capacities) but are not available exactly. This has opened a wealth of approaches to account for randomness in linear programs. Common to all formulations is their general assumption that some parameters in (P_b) are random and follow a known probability distribution. Important contributions in this regard are *chance constrained linear programs*, *two- and multiple-stage programming* as well as the theory of *stochastic linear programs* (see Shapiro et al. (2021) for a general overview). Specifically relevant to this paper is the so-called *distribution problem* characterizing the distribution of the random variable $\nu(X)$, where the right-hand side b (and possibly A and c) in (P_b) is replaced by a random variable X following a specific law (Ewbank et al., 1974; Wets, 1980).

In this paper, we take a different route and focus on statistical aspects on the standard linear program (P_b) if the right-hand side b is replaced by a consistent estimator b_n indexed by $n \in \mathbb{N}$, e.g., based on n observations. Different to aforementioned attempts, we only assume the random quantity $r_n(b_n - b)$ to converge weakly (denoted by $\xrightarrow{D}$) to some limit law G as n tends to infinity[2]. Our main goal is to characterize the asymptotic distributional limit of the empirical optimal solution

$$x^\star(b_n) \in \operatorname*{arg\,min}_{Ax = b_n,\, x \geq 0} \; c^T x \tag{1.1}$$

around its population quantities after proper standardization. For the sake of exposition, suppose that $x^\star(b)$ in (1.1) is unique[3]. The main results in Theorem 3.1 and Theorem 3.3 state that under suitable assumptions on (P_b) it holds, as n tends to infinity, that

$$r_n \left(x^\star(b_n) - x^\star(b) \right) \xrightarrow{D} M(G), \tag{1.2}$$

where $M : \mathbb{R}^m \to \mathbb{R}^d$ is given in Theorem 3.1. The function M in (1.2) is possibly random, and its explicit form is driven by the amount of degeneracy present in the primal and dual optimal solutions. The simplest case occurs if $x^\star(b)$ is non-degenerate. The function M is then a linear transformation depending on the corresponding unique optimal basis, so that the limit law $M(G)$ is Gaussian if G is Gaussian. If $x^\star(b)$ is degenerate but all dual optimal (basic) solutions for (D_b) are non-degenerate, then M is a sum of deterministic linear transformations defined on closed and convex cones indexed by the collection of dual

[2] A prototypical example is the standard central limit theorem, whereby under suitable assumptions $r_n = \sqrt{n}$ and G is a Gaussian random vector on $\mathbb{R}^m$.

[3] The main results in Theorem 3.1 and 3.3 hold more generally and beyond the uniqueness assumptions.

optimal bases. Specifically, the number of summands in M is equal to the number of dual optimal basic solutions for (D$_b$). A more complicated situation arises if both $x^\star(b)$ and some dual optimal basic solutions are degenerate. In this case, the function M is still a sum of linear transformations defined on closed and convex cones, but these transformations are potentially random and indexed by certain *subsets* of the set of optimal bases. The latter setting reflects the complex geometric and combinatorial nature in linear programs under degeneracy.

Let us mention at once that limiting distributions for the empirical optimal solution in the form of (1.2) have been studied for a long time in a more general setting of (potentially) non-linear optimisation problems; see for example Dupačová (1987); Dupačová & Wets (1988); Shapiro (1991, 1993, 2000); King & Rockafellar (1993). Regularity assumptions such as strong convexity of the objective function near the (unique) optimizer allow for either explicit asymptotic expansions of optimal values and optimal solutions or applications of implicit function theorems and generalizations thereof. These conditions usually do not hold for the linear programs considered in this paper.

To the best of our knowledge, our results are the first that cover limit laws for empirical optimal solutions to standard linear programs even beyond the non-degenerate case and without assuming uniqueness of optimizers. However, our proof technique relies on well-known concepts from parametric optimization and sensitivity analysis for linear programs (Guddat et al., 1974; Greenberg, 1986; Ward & Wendell, 1990; Hadigheh & Terlaky, 2006). Indeed, our approach is based on a careful study of the collection of dual optimal bases. An early contribution in this regard is the *basis decomposition theorem* by Walkup & Wets (1969a) analyzing the behavior of $\nu(b)$ in (P$_b$) as a function of b (see also Remark 3.2). Each dual feasible basis defines a so called *decision region* over which the optimal value $\nu(b)$ is linear. The integration over the collection of all these regions yields closed form expressions for the distribution problem (Bereanu, 1963; Ewbank et al., 1974). Further, related stability results are also found in the work by Walkup & Wets (1967); Böhm (1975); Bereanu (1976) and Robinson (1977). In algebraic geometry, decision regions are closely related to *cone-triangulations* of the primal feasible optimization region (Sturmfels & Thomas, 1997; De Loera et al., 2010). We emphasize that rather than working with decision regions directly, our analysis is tailored to cones of feasible perturbations. In particular, we are interested in regions capturing feasible directions as our problem settings is based on the random perturbation $\sqrt{n}\,(b_n - b)$. These regions turn out to be closed, convex cones and appear as indicator functions in the (random) function M in (1.2).

Our proof technique allows to recover some related known results for random linear programs (see Section 3). These include convergence of the optimality sets in Hausdorff distance (Proposition 3.6), and a limit law for the optimal value

$$r_n(\nu(b_n) - \nu(b)) \xrightarrow{D} \max_{\substack{\lambda(I) \text{ dual optimal} \\ \text{basic solution for } (\mathrm{D}_b)}} G^T\lambda(I), \tag{1.3}$$

as n tends to infinity. Indeed, (1.3) is a simple consequence of general results in constrained optimization (Shapiro, 2000; Bonnans & Shapiro, 2000), and the optimality set convergence follows from Walkup & Wets (1969b).

Our statistical analysis for random linear programs in standard form is motivated by recent findings in statistical optimal transport (OT). More precisely, while there exists a thorough theory for limit laws on empirical OT costs on discrete spaces (Sommerfeld & Munk, 2018; Tameling et al., 2019), related statements for their empirical OT solutions remain open. An exception is Klatt et al. (2020), who provide limit laws for empirical *(entropy) regularized* OT solutions, thus modifying the underlying linear program to be strictly convex, non-linear and most importantly non-degenerate in the sense that every regularized OT solution is strictly positive in each coordinate. Hence, an implicit function theorem approach in conjunction with a delta method allows concluding for Gaussian limits in this case. This stands in stark contrast to the *non-regularized* OT considered in this paper, where the degenerate case is generic rather than the exception for most practical situations. Only if the OT solution is unique and non-degenerate, then we observe a Gaussian fluctuation on the support set, i.e., on all entries with positive values. If the OT solution is degenerate (or not unique), then the asymptotic limit law (1.2) is usually not Gaussian anymore. Degeneracy in OT easily occurs as soon as certain subsets of demand and supply sum up to the same quantity. In particular, we encounter the largest degree of degeneracy if individual demand is equal to individual supply. Additionally, we obtain necessary and sufficient conditions on the cost function in order for the dual OT to be non-degenerate. These may be of independent interest, and allow to prove almost sure uniqueness results for quite general cost functions.

Our distributional results can be viewed as a basis for uncertainty quantification and other statistical inference procedures concerning solutions to linear programs. For brevity, we mention such applications in passing and do not elaborate further on them, leaving a detailed study of statistical consequences such as testing or confidence statements as an important avenue for further research.

The outline of the paper is as follows. We recap basics for linear programming in Section 2 also introducing deterministic and stochastic assumptions for our general theory. Our main results are summarized in Section 3, followed by their proofs in Section 4. The assumptions are discussed in more detail in Section 5. Section 6 focuses on OT and gives limit laws for empirical OT solutions.

2 Preliminaries and Assumptions

This section introduces notation and assumptions required to state the main results of the paper. Along the way, we recall basic facts of linear programming and refer to Bertsimas & Tsitsiklis (1997) and Luenberger & Ye (2008) for details.

Linear Programs and Duality. Let the columns of a matrix $A \in \mathbb{R}^{m\times d}$ be enumerated by the set $[d] \coloneqq \{1,\dots,d\}$. Consider for a subset $I \subseteq [d]$ the sub-matrix $A_I \in \mathbb{R}^{m\times|I|}$ formed by the corresponding columns indexed by I. Similarly, $x_I \in \mathbb{R}^{|I|}$ denotes the coordinates of $x \in \mathbb{R}^d$ corresponding to I. By full rank of A in (D_b), there always exists an index set I with cardinality m such that $A_I \in \mathbb{R}^{m\times m}$ is one-to-one. An index set I with that property is termed *basis* and induces a *primal* and *dual basic solution*

$$x(I,b) \coloneqq \mathrm{Aug}_I\left[(A_I)^{-1}b\right] \in \mathbb{R}^d, \quad \lambda(I) \coloneqq (A_I)^{-T}c_I \in \mathbb{R}^m,$$

respectively. Herein, and in order to match dimensions (a solution for (P_b) has dimension d instead of $m \le d$) the linear operator $\mathrm{Aug}_I\colon \mathbb{R}^m \to \mathbb{R}^d$ augments zeroes in the coordinates that are not in I. If $\lambda(I)$ (resp. $x(I,b)$) is feasible for (D_b) (resp. (P_b)) then it constitutes a *dual* (resp. *primal*) *feasible basic solution* with *dual* (resp. *primal*) *feasible basis* I. Moreover, $\lambda(I)$ (resp. $x(I,b)$) is termed *dual* (resp. *primal*) *optimal basic solution* if it is feasible and optimal for (D_b) (resp. (P_b)). Indeed, as long as (D_b) admits a feasible (optimal) solution then there exists a dual feasible (optimal) basic solution and vice versa for (P_b). At the heart of linear programming is the *strong duality* statement.

Fact 2.1. *Consider the primal linear program* (P_b) *and its dual* (D_b).

(i) If either of the linear programs (P_b) *or* (D_b) *has a finite optimal solution, so does the other and the corresponding optimal values are the same.*

(ii) If for a basis $I \subseteq [d]$ the vector $\lambda(I)$ is dual feasible and $x(I,b)$ is primal feasible, then both are primal and dual optimal basic solutions, respectively.

(iii) If (P_b) *and* (D_b) *are feasible then there always exists a basis $I \subseteq [d]$ such that $x(I,b)$ and $\lambda(I)$ are primal and dual optimal basic solutions, respectively.*

We introduce the feasibility and optimality set for the primal (P_b) by

$$\mathcal{P}(b) \coloneqq \left\{x \in \mathbb{R}^d \mid Ax = b, x \geq 0\right\}, \quad \mathcal{P}^\star(b) \coloneqq \left\{x^\star \in \mathcal{P}(b) \mid c^T x^\star = \inf_{x \in \mathcal{P}(b)} c^T x\right\}, \tag{2.1}$$

respectively. Notably, in our theory to follow A and c are generally assumed to be fixed and only the dependence of these sets with respect to parameter b is emphasized. We introduce our first assumption:

$$\textit{The set } \mathcal{P}^\star(b) \textit{ is non-empty and bounded.} \tag{\textbf{A1}}$$

In view of the strong duality statement in Fact 2.1, solving a linear program might be carried out focusing on the collection of all dual feasible bases. We partition this collection into two subsets depending on their feasibility for the primal program.

Remark 2.2 (Splitting of the Bases Collection)**.** *Let $I_1, \ldots, I_N$ enumerate all dual feasible bases, and let $1 \leq K \leq N$ be such that*

$$x(I_k, b) \textit{ is feasible}, 1 \leq k \leq K; \qquad x(I_k, b) \textit{ is infeasible}, \ K < k \leq N.$$

Notably, by Fact 2.1 the primal basic solution $x(I_k, b)$ is optimal for all $k \leq K$. Recall that the convex hull $\mathcal{C}(x_1, \ldots, x_K)$ of a collection of points $\{x_1, \ldots, x_K\} \subset \mathbb{R}^d$ is the set of all possible convex combinations of them.

Fact 2.3. *Consider the primal linear program* (P_b) *and assume* (**A1**) *holds. Then for any right hand side $\tilde{b} \in \mathbb{R}^m$ either one of the following statements is correct.*

(i) The feasible set $\mathcal{P}\left(\tilde{b}\right)$ is empty.

(ii) The optimality set $\mathcal{P}^\star\left(\tilde{b}\right)$ is non-empty, bounded and equal to the convex hull

$$\mathcal{P}^\star\left(\tilde{b}\right) = \mathcal{C}\left(\left\{x(I,\tilde{b}) \mid I \textit{ primal and dual feasible basis for } \left(P_{\tilde{b}}\right) \textit{ and } \left(D_{\tilde{b}}\right)\right\}\right).$$

The restriction of the convex hull to basic solutions induced by primal *and* dual optimal bases in Fact 2.3 is well-known. A straightforward argument is based on the simplex method that if set up with appropriate pivoting rules always terminates. If (**A1**) holds and there exists a unique basis ($K = 1$) then the primal program attains a unique solution. Uniqueness of solutions to linear programs is related to *degeneracy* of corresponding dual solutions. A dual feasible basic solution $\lambda(I)$ is degenerate if more than m of the d inequalities $\lambda(I)^T A \leq c^T$ hold as equalities. Similarly, a primal feasible basic solution $x(I,b)$ is degenerate if less than m of its coordinates are nonzero.

Fact 2.4. *Consider the linear program* (P_b) *and its dual* (D_b)*.*

(i) If (P_b) *(resp.* (D_b)*) has a non-degenerate optimal basic solution, then* (D_b) *(resp.* (P_b)*) has a unique solution.*

(ii) If (P$_b$) *(resp.* (D$_b$)*) has a unique non-degenerate optimal basic solution, then* (D$_b$) *(resp.* (P$_b$)*) has a unique non-degenerate optimal solution.*

(iii) If (P$_b$) *(resp.* (D$_b$)*) has a unique degenerate optimal basic solution, then* (D$_b$) *(resp.* (P$_b$)*) has multiple solutions.*

For a proof of Fact 2.4, we refer to Gal & Greenberg (2012, Lemma 6.2) in combination with *strict complementary slackness* from Goldman & Tucker (1956, Corollary 2A) stating that for feasible primal and dual linear program there exists a pair (x, λ) of primal and dual optimal solution such that either $x_j > 0$ or $\lambda^T A_j < c_j$ for all $1 \leq j \leq d$. In addition to uniqueness statements, many results in linear programming simplify when degeneracy is excluded. Related to degeneracy but slightly weaker is the assumption

$$\lambda(I_j) \neq \lambda(I_k), \qquad 1 \leq j < k \leq K. \tag{\textbf{A2}}$$

Indeed, if $\mathcal{P}^\star(b)$ is non-empty and bounded assumption (**A2**) characterizes non-degeneracy of all dual basic solution.

Lemma 2.5. *Suppose assumption* (**A1**) *holds. Then assumption* (**A2**) *is equivalent to non-degeneracy of all dual optimal basic solutions.*

To see that (**A1**) is necessary, let $D_m \in \mathbb{R}^{m\times m}$ with $m \geq 2$ be the identity matrix. Suppose that $A = (D_m, -D_m) \in \mathbb{R}^{m\times 2m}$, $c \in \mathbb{R}^{2m}_+$ is strictly positive except that $c_1 = c_{m+1} = 0$, and $b = (1, 0, 0, \dots, 0) \in \mathbb{R}^m$. Then there are $K = 2^{m-1}$ optimal bases defining K distinct (degenerate) dual solutions, so that assumption (**A2**) holds but dual degeneracy fails. Note that $\mathcal{P}^\star(b)$ is unbounded and contains the optimal ray (b^T, b^T).

Random Linear Programs. Introducing randomness in problems (P$_b$) and (D$_b$), we suppose to have incomplete knowledge of $b \in \mathbb{R}^m$, and replace it by a (consistent) estimator b_n, e.g., based on a sample of size n independently drawn from a distribution with mean b. This defines empirical primal and dual counterparts (P$_{b_n}$) and (D$_{b_n}$), respectively. We allow the more general case that only the first $m_0 \in \{0, \dots, m\}$ coordinates of b are unknown[4] and assume the existence of a sequence of random vectors $b_n = (b_n^{m_0}, [b]_{m-m_0})^T \in \mathbb{R}^{m_0} \times \mathbb{R}^{m-m_0}$ converging to b at rate $r_n^{-1} \to 0$ as n tends to infinity

$$\begin{gathered} G_n := r_n(b_n - b) \xrightarrow{D} G = \begin{pmatrix} G^{m_0} \\ 0_{m-m_0} \end{pmatrix} \in \mathbb{R}^{m_0} \times \mathbb{R}^{m-m_0}, \\ \textit{with } G^{m_0} \in \mathbb{R}^{m_0} \textit{ absolutely continuous,} \end{gathered} \tag{\textbf{B1}}$$

where $\xrightarrow{D}$ denotes convergence is distribution. In a typical central limit theorem type scenario, $r_n = \sqrt{n}$ and G^{m_0} is a centred Gaussian random vector in $\mathbb{R}^{m_0}$, assumed to have a non-singular covariance matrix. Assumption (**B1**) implies that $b_n \to b$ in probability. In order to avoid pathological cases, we impose the last assumption that asymptotically an optimal solution $x^\star(b_n)$ for the primal (P$_{b_n}$) exists

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{P}^\star(b_n) \neq \varnothing\right) = 1. \tag{\textbf{B2}}$$

Further discussions on the assumptions are deferred to Section 5.

[4] One may assume at first reading that $m_0 = m$; the additional generality will turn useful for the one-sample case naturally arising in optimal transport in Section 6.

3 Main Results

According to Fact 2.3, in presence of (**A1**) any optimal solution $x^\star(b_n) \in \mathcal{P}^\star(b_n)$ takes the form

$$x^\star(b_n) = \sum_{k\in\mathcal{K}} [\alpha_n^{\mathcal{K}}]_k x(I_k, b_n) := \alpha_n^{\mathcal{K}} \otimes x(I_{\mathcal{K}}, b_n),$$

where $\mathcal{K}$ is a non-empty subset of $[N] := \{1,\dots,N\}$ and $\alpha_n^{\mathcal{K}} \in \mathbb{R}^N$ is a random vector in the (essentially $|\mathcal{K}|$-dimensional) unit simplex $\Delta_{\mathcal{K}} := \left\{\alpha \in \mathbb{R}_+^N \mid \|\alpha\|_1 = 1, \alpha_k = 0 \ \forall k \notin \mathcal{K}\right\}$. The main result of the paper states the following asymptotic behaviour for the empirical optimal solution.

Theorem 3.1. *Suppose assumptions* (**A1**), (**B1**), *and* (**B2**) *hold, and let $x^\star(b_n) \in \mathcal{P}^\star(b_n)$ be any (measurable) choice of an optimal solution. Further, assume that for all non-empty $\mathcal{K} \subseteq [K]$, the random vectors $\left(\alpha_n^{\mathcal{K}}, G_n\right)$ converge jointly in distribution as n tends to infinity to $(\alpha^{\mathcal{K}}, G)$ on $\Delta_{|\mathcal{K}|} \times \mathbb{R}^m$. Then there exist closed convex cones $H_1^{m_0}, \dots, H_K^{m_0} \subseteq \mathbb{R}^{m_0}$ and random vectors $Y_n \in \mathcal{P}^\star(b)$ such that*

$$r_n\left(x^\star(b_n) - Y_n\right) \xrightarrow{D} M(G) := \sum_{\mathcal{K}} \mathbb{1}_{\left\{G^{m_0} \in H_{\mathcal{K}}^{m_0} \setminus \cup_{k\in[K]\setminus\mathcal{K}} H_k^{m_0}\right\}} \alpha^{\mathcal{K}} \otimes x(I_{\mathcal{K}}, G) \in \mathbb{R}^d,$$

where the sum runs over non-empty subsets $\mathcal{K}$ of $[K]$ and $H_{\mathcal{K}}^{m_0} = \cap_{k\in\mathcal{K}} H_k^{m_0}$.

Remark 3.2. *Underlying Theorem 3.1 is the well-known approach of partitioning $\mathbb{R}^m$ into (closed convex) cones. Indeed, the union of the closed convex cones*

$$\widetilde{H}_k = \{b \in \mathbb{R}^m : I_k \text{ is an optimal basis for } (P_b) \text{ and } (D_b)\}, \qquad k = 1, \dots, N,$$

is the feasibility set $A_+ := \{Ax : x \geq 0\} \subseteq \mathbb{R}^m$ and on each cone the optimal solution is an affine function of b (e.g., Walkup & Wets, 1969a; Guddat et al., 1974). The cones $\widetilde{H}_k$ depend only on A and c. In contrast, our cones $H_k^{m_0}$ also depend on b and define directions of perturbations of b that keep $\lambda(I_k)$ optimal for the perturbed problem for a given $k \leq K$. Assume for simplicity that $m_0 = m$ and write H_k instead of $H_k^{m_0}$. If $b = 0$, then $K = N$ and cones coincide $H_k = \widetilde{H}_k$, but otherwise H_k is a strict super-set of $\widetilde{H}_k$ as the corresponding representation (4.2) of $\widetilde{H}_k$ requires non-negativity on all coordinates. This is also in line with the observation that there are fewer (K) cones H_k than there are $\widetilde{H}_k$, namely N, and the union of the H_k's is a space that is at least as large as A_+ (since $b + A_+ \subseteq A_+$ because $b \in A_+$), typically $\mathbb{R}^m$. As an extreme example, suppose that (P_b) has a unique non-degenerate optimal solution $x(I_1, b)$. Then $K = 1$ and $H_1 = \mathbb{R}^m$ but the $\widetilde{H}_k$'s are strict subsets of $\mathbb{R}^m$ unless $N = 1$.

In Section 5, we discuss sufficient conditions for the joint distributional convergence of the random vector $\left(\alpha_n^{\mathcal{K}}, G_n\right)$. In short, if we use any linear program solver, such joint distributional convergence appears to be reasonable. If the optimal basis is unique ($K = 1$) with $x^\star(b) = x(I_1, b)$ non-degenerate, then $\lambda(I_1)$ is non-degenerate, and the proof shows that $H_k^{m_0} = \mathbb{R}^{m_0}$. The distributional limit theorem then takes the simple form

$$r_n(x^\star(b_n) - x^\star(b)) \xrightarrow{D} x(I_1, G) \in \mathbb{R}^d.$$

In general, when $K > 1$, the number of summands in the limiting random variable in Theorem 3.1 might grow exponentially in K. In between these two cases is the situation that assumption (**A2**) holds, which implies all dual optimal basic solutions for (D_b) are non-degenerate (see Lemma 2.5). The limiting random variable then simplifies, as the subsets $\mathcal{K}$ must be singletons.

Theorem 3.3. *Suppose assumptions (**A1**), (**A2**), and (**B2**) hold, and that* $r_n(b_n - b) \xrightarrow{D} G$. *Then any*[5] *(measurable) choice of* $x^\star(b_n) \in \mathcal{P}^\star(b_n)$ *satisfies*

$$r_n\left(x^\star(b_n) - Y_n\right) \xrightarrow{D} \sum_{k=1}^{K} \mathbb{1}_{\left\{G^{m_0} \in H_k^{m_0} \setminus \cup_{j<k} H_j^{m_0}\right\}}\, x\left(I_k, G\right) \in \mathbb{R}^d$$

with the closed and convex cones $H_k^{m_0}$ *as given in Theorem 3.1.*

Remark 3.4. *With respect to Theorem 3.1, assumption (**B1**) is weakened in Theorem 3.3 as absolute continuity of* G *(or* G^{m_0}*) is not required. Indeed, it can be arbitrary, and Theorem 3.3 thus accommodates, e.g., Poisson limit distributions. The proof shows that if* G *is absolutely continuous (i.e.,* $m_0 = m$*) then the indicator functions of* $G \in H_k^m \setminus \cup_{j<k} H_j^m$ *simplify to* $G \in H_k^m$*, because intersections* $H_k^m \cap H_j^m$ *have Lebesgue measure zero. The distributional limit theorem then reads as*

$$r_n\left(x^\star(b_n) - Y_n\right) \xrightarrow{D} \sum_{k=1}^{K} \mathbb{1}_{\left\{G^m \in H_k^m\right\}}\, x\left(I_k, G\right) \in \mathbb{R}^d.$$

We conclude this section by giving two further consequences of our proof techniques: a limit law for the objective value $\nu(b)$ for (P_b), and convergence in probability of optimality sets. Since the former is well-known and holds in more general, infinite-dimensional convex programs, we omit the proof details and instead refer to Shapiro (2000); Bonnans & Shapiro (2000) and results by Sommerfeld & Munk (2018); Tameling et al. (2019) tailored to OT.

Proposition 3.5. *Under assumptions (**A1**), (**B1**), and (**B2**) it holds that*

$$r_n[\nu(b_n) - \nu(b)] \xrightarrow{D} \max_{k \in [K]} G^T \lambda(I_k).$$

Another consequence of our bases driven approach underlying the proof of Theorem 3.1 is that the convergence of the Hausdorff distance

$$d_H\left(\mathcal{P}^\star(b_n), \mathcal{P}^\star(b)\right) \coloneqq \max\left\{\sup_{x \in \mathcal{P}^\star(b_n)} \inf_{y \in \mathcal{P}^\star(b)} \|x - y\|,\ \sup_{x \in \mathcal{P}^\star(b)} \inf_{y \in \mathcal{P}^\star(b_n)} \|x - y\|\right\}$$

between $\mathcal{P}^\star(b_n)$ and $\mathcal{P}^\star(b)$ is of order $O_{\mathbb{P}}(r_n^{-1})$. A different and considerably shorter argument relies on Walkup & Wets (1969b) and proves the following result.

Proposition 3.6. *Suppose assumptions (**A1**) and (**B2**) hold. If* $\|b_n - b\| = O_{\mathbb{P}}(r_n^{-1})$*, then it follows that* $d_H\left(\mathcal{P}^\star(b_n), \mathcal{P}^\star(b)\right) = O_{\mathbb{P}}(r_n^{-1})$.

We also refer to the work by Robinson (1977) for a similar result when the primal and dual optimality sets are both bounded.

4 Proofs for the Main Results

To simplify the notation, we assume that all random vectors in the paper are defined on a common generic probability space $(\Omega, \mathcal{F}, \mathbb{P})$. This is no loss of generality by the Skorokhod representation theorem.

[5] There is no need to assume joint distributional convergence of $(\alpha_n^{\mathcal{K}}, G_n)$ as in Theorem 3.1.

Preliminary steps. Recall from Remark 2.2 that bases $I_1, \dots, I_K$ are feasible for (P_b) and (D_b) and hence optimal. The bases $I_{K+1}, \dots, I_N$ are only feasible for (D_b) but not for (P_b). For a set $\mathcal{K} \subseteq [N]$ define the events, i.e., subsets of the underlying probability space

$$A_n^{\mathcal{K}} := \left\{ \begin{matrix} x(I_k, b_n(\omega)) \geq 0 \\ A^T \lambda(I_k) \leq c \end{matrix} \quad \iff k \in \mathcal{K} \right\} \subseteq \Omega,$$

$$B_n^{\mathcal{K}} := \left\{ \begin{matrix} x(I_k, b_n(\omega)) \geq 0 \\ A^T \lambda(I_k) \leq c \end{matrix} \quad \impliedby k \in \mathcal{K} \right\} \subseteq \Omega.$$

By strong duality (Fact 2.1 (ii)), the set $A_n^{\mathcal{K}}$ is the event that the bases indexed by $\mathcal{K}$ are precisely those that are optimal for (P_{b_n}) and (D_{b_n}). We have $A_n^{\mathcal{K}} \subseteq B_n^{\mathcal{K}}$, and $B_n^{\mathcal{K}} \subseteq B_n^{\{k\}}$ for all $k \in \mathcal{K}$. We start with two important observations, the first stating that only subsets of $[K]$ asymptotically matter.

Lemma 4.1. *Suppose that* $b_n \xrightarrow{D} b$.

(i) It holds that $\mathbb{P}\left(B_n^{\{k\}}\right) \to 0$ *as* $n \to \infty$ *for all* $k > K$.

(ii) If assumptions (**A1**) *and* (**B2**) *hold, then with high probability* $\mathcal{P}^\star(b_n)$ *is bounded and non-empty.*

Proof. For (i), observe that for $k > K$ there exists an index $i \in [d]$ such that $x_i(I_k, b) < 0$. The same inequality holds for b_n if sufficiently close to b, which happens with high probability. For (ii), non-emptiness with high probability follows from assumption (**B2**), so we only prove boundedness. Indeed, assumption (**A1**) implies that the recession cone $\{x \geq 0 \mid Ax = 0, c^T x = 0\}$ is trivial and equals $\{0\}$. This property does not depend on b_n, which yields the result. □

The event $A_n^{\varnothing}$ is equivalent to (P_{b_n}) being either infeasible or unbounded, and this has probability $o(1)$ by (**B2**). Combining this with the previous lemma and the sets $(A_n^{\mathcal{K}})_{\mathcal{K}}$ forming a partition of the probability space Ω, we deduce

$$x^\star(b_n) = \sum_{\varnothing \subset \mathcal{K} \subseteq [K]} \mathbb{1}_{A_n^{\mathcal{K}}}(\omega)\, \alpha_n^{\mathcal{K}}(\omega) \otimes x(I_{\mathcal{K}}, b_n(\omega)) + o_{\mathbb{P}}(1),$$

where $\mathbb{1}_A(\omega)$ denotes the usual indicator function of the set A. Defining the random vector

$$Y_n = \sum_{\varnothing \subset \mathcal{K} \subseteq [K]} \mathbb{1}_{A_n^{\mathcal{K}}}(\omega)\, \alpha_n^{\mathcal{K}}(\omega) \otimes x(I_{\mathcal{K}}, b)$$

that lies in $\mathcal{P}^\star(b)$ (because $\mathcal{K} \subseteq [K]$), we obtain

$$r_n[x^\star(b_n) - Y_n] = \sum_{\varnothing \subset \mathcal{K} \subseteq [K]} \mathbb{1}_{A_n^{\mathcal{K}}}(\omega)\, \alpha_n^{\mathcal{K}}(\omega) \otimes x(I_{\mathcal{K}}, G_n(\omega)) + o_{\mathbb{P}}(1). \tag{4.1}$$

We next investigate the indicator functions $\mathbb{1}_{A_n^{\mathcal{K}}}(\omega)$ appearing in (4.1). Omitting the dependence of b_n on ω, we rewrite

$$B_n^{\mathcal{K}} = \bigcap_{k \in \mathcal{K}} \bigcap_{i \in I_k} \{x_i(I_k, b_n) \geq 0\} = \bigcap_{k \in \mathcal{K}} \bigcap_{i \in [d]} \{x_i(I_k, G_n) \geq -r_n x_i(I_k, b)\}.$$

At the last internal intersection in the above display we can, with high probability, restrict to those i in the primal degeneracy set $\mathrm{DP}_k := \{i \in I_k \mid x_i(I_k, b) = 0\}$. Indeed, for $i \notin I_k$, the inequality reads $0 \geq 0$, whereas for $i \in I_k \setminus \mathrm{DP}_k$ the right-hand side goes to $-\infty$ and the left-hand side is bounded in probability. In other words $\mathbb{P}(B_n^{\mathcal{K}}) = o(1) + \mathbb{P}(G_n^{m_0} \in \cap_{k\in\mathcal{K}} H_k^{m_0})$, where

$$H_k^{m_0} = \{g^{m_0} \in \mathbb{R}^{m_0} : [x(I_k, (g^{m_0}, 0_{m-m_0}))]_{\mathrm{DP}_k} \geq 0\}. \tag{4.2}$$

For $\varnothing \subset \mathcal{K} \subseteq [K]$ define $H_{\mathcal{K}}^{m_0} = \cap_{k\in\mathcal{K}} H_k^{m_0}$, and write

$$A_n^{\mathcal{K}} = \left(B_n^{\mathcal{K}} \setminus \bigcup_{k\in[K]\setminus\mathcal{K}} B_n^{\{k\}} \right) \setminus \bigcup_{k>K} B_n^{\{k\}},$$

where the union over $k > K$ can be neglected by Lemma 4.1. Thus we conclude that

$$\mathbb{1}_{A_n^{\mathcal{K}}} = \mathbb{1}_{\{G_n^{m_0} \in H_{\mathcal{K}}^{m_0} \setminus \cup_{k\in[K]\setminus\mathcal{K}} H_k^{m_0}\}} + o_{\mathbb{P}}(1). \tag{4.3}$$

With these preliminary statements at our disposal, we are ready to prove the main result.

Theorem 3.1. The goal is to replace $G_n^{m_0}$ by G^{m_0} in the indicator function in (4.3) at the limit as n tends to infinity. By the Portmanteau theorem (Billingsley, 1999, Theorem 2.1) and elementary arguments[6] it suffices to show that the m_0-dimensional boundary of each $H_k^{m_0}$ has Lebesgue measure zero. This is indeed the case, as they are convex sets. Define the function $T^{\mathcal{K}} : \mathbb{R}^{|\mathcal{K}|} \times \mathbb{R}^m \to \mathbb{R}^d$ by

$$T^{\mathcal{K}}(\alpha, v) = \sum_{k\in\mathcal{K}} \mathbb{1}_{\{v_{[m_0]} \in H_{\mathcal{K}}^{m_0} \setminus \cup_{k\in[K]\setminus\mathcal{K}} H_k^{m_0}\}} \alpha_k x(I_k, v).$$

This function is continuous for all $\alpha \in \mathbb{R}^{\mathcal{K}}$ and all vectors $v \in \mathbb{R}^m$ such that $v_{[m_0]} \notin \partial[H_{\mathcal{K}}^{m_0} \setminus \cup_{k\in[K]\setminus\mathcal{K}} H_k^{m_0}]$. In particular, the continuity set is of full measure with respect to $(\alpha^{\mathcal{K}}, G)$. As there are finitely many possible subsets $\mathcal{K}$ denoted by $\mathcal{K}_1, \ldots, \mathcal{K}_B$, the function $T = (T^{\mathcal{K}_1}, \ldots, T^{\mathcal{K}_B}) : \mathbb{R}^{\sum_{i=1}^B |\mathcal{K}_i|} \times \mathbb{R}^m \to (\mathbb{R}^d)^B$ defined by

$$T\left(\alpha^{\mathcal{K}_1}, \ldots, \alpha^{\mathcal{K}_B}, v\right) = \left(T^{\mathcal{K}_1}(\alpha^{\mathcal{K}_1}, v), \ldots, T^{\mathcal{K}_B}(\alpha^{\mathcal{K}_B}, v)\right)$$

is continuous G-almost surely. The continuous mapping theorem together with the assumed joint distributional convergence of the random vector $(\alpha_n^{\mathcal{K}}, G_n)$ yield that

$$\sum_{\varnothing\subset\mathcal{K}\subseteq[K]} T^{\mathcal{K}}\left(\alpha_n^{\mathcal{K}}, G_n\right) \xrightarrow{D} \sum_{\varnothing\subset\mathcal{K}\subseteq[K]} T^{\mathcal{K}}\left(\alpha^{\mathcal{K}}, G\right)$$

which finishes the proof for Theorem 3.1. □

Theorem 3.3. With high probability (**A1**) and (**A2**) hold for b_n (by Lemma 4.1 for the former and trivially for the latter), which implies that $\mathcal{P}^\star(b_n)$ is a singleton (Lemma 2.5 and Fact 2.4). Hence, regardless of the choice of $\alpha_n^{\mathcal{K}}$, it holds that $\mathbb{1}_{A_n^{\mathcal{K}}} x^\star(b_n) = x(I_{\min\mathcal{K}}, b_n)$. In particular, we may assume without loss of generality that $\alpha_n^{\mathcal{K}}$ are deterministic and do not depend on n. Thus the joint convergence in Theorem 3.1 holds, and (4.1) simplifies to

$$r_n[x^\star(b_n) - Y_n] = \sum_{\varnothing\subset\mathcal{K}\subseteq[K]} \mathbb{1}_{A_n^{\mathcal{K}}}(\omega)\, x(I_{\min\mathcal{K}}, G_n) + o_{\mathbb{P}}(1) = x(I_{K(\omega)}, G_n(\omega)) + o_{\mathbb{P}}(1),$$

[6] Letting $A_k = H_k^{m_0}$ and $B_k = \mathbb{R}^{m_0} \setminus A_k$, it holds $\partial B_k = \partial A_k$ and thus $\partial(\bigcap_{k\in\mathcal{K}} A_k \cap \bigcap_{k\in[K]\setminus\mathcal{K}} B_k) \subseteq \bigcup_{k=1}^K \partial A_k$.

where $K(\omega)$ is the minimal $k \leq K$ such that $B_n^{\{k\}}$ holds. Since $B_n^{\{k\}}$ is asymptotically $\{G \in H_k^{m_0}\}$, Theorem 3.3 follows. Let us now show that $M(G)$ simplifies to $\sum_{k=1}^K \mathbb{1}_{\{G \in H_k^m\}} x(I_k, G)$ if G is absolutely continuous. It suffices to show that intersections $H_j^m \cap H_k^m$ with $j < k \leq K$ have Lebesgue measure zero. If $v \in H_j^m \cap H_k^m$ then there exists $\eta > 0$ such that $x(I_j, b + \eta v) \geq 0$ and $x(I_k, b + \eta v) \geq 0$. Since $\lambda(I_k)$ and $\lambda(I_j)$ are dual feasible, they must be optimal with respect to $b + \eta v$. Thus it holds

$$0 = \frac{1}{\eta}(b + \eta v)^T[\lambda(I_k) - \lambda(I_j)] = v^T[\lambda(I_k) - \lambda(I_j)].$$

By (**A2**) the vector $\lambda(I_k) - \lambda(I_j)$ is nonzero and hence v is contained in its orthogonal complement, which indeed has Lebesgue measure zero. □

Proposition 3.6. Let $K = \mathbb{R}_+^d$ and define the linear map $\tau : \mathbb{R}^d \to \mathbb{R}^{m+1}$ by $\tau(x) = (Ax, c^t x)$. For each b such that the linear program is feasible, let $v_b \in \mathbb{R}$ be the optimal objective value. If τ is injective, then the optimality sets are singletons and the result holds trivially. We thus assume that τ is not injective, and observe that

$$K \cap \tau^{-1}\{(b, v_b)\} = \{x \geq 0 : Ax = b, c^t x = v_b\} = \mathcal{P}^\star(b).$$

Since K is a polyhedron and τ is neither identically zero (A has full rank) nor injective, we can apply the main theorem of Walkup & Wets (1969b). We obtain

$$d_H(\mathcal{P}^\star(b_n), \mathcal{P}^\star(b)) \leq B\sqrt{\|b - b_n\|^2 + |v_b - v_{b_n}|^2} = O_{\mathbb{P}}(r_n^{-1}), \qquad B = B(A, c) < \infty,$$

because the optimal values satisfy $v_b - v_{b_n} = O_{\mathbb{P}}(r_n^{-1})$ by Proposition 3.5. □

5 On the Assumptions

We start collecting some well-known facts from parametric optimization (see Walkup & Wets (1969a); Guddat et al. (1974) for details). To this end, denote the dual feasible set by $\mathcal{N} \coloneqq \{\lambda \in \mathbb{R}^m \mid A^t\lambda \leq c\}$. Further, define the set of feasible parameters by $\mathcal{M} \coloneqq \{b \in \mathbb{R}^m \mid \mathcal{P}(b) \neq \varnothing\}$ and $\mathcal{M}^\star \coloneqq \{b \in \mathbb{R}^m \mid \mathcal{P}^\star(b) \neq \varnothing\}$ the solution set.

Lemma 5.1. *If for some $b_0 \in \mathcal{M}$ the set $\mathcal{P}(b_0)$ is bounded (resp. unbounded) then $\mathcal{P}(b)$ is bounded (resp. unbounded) for all $b \in \mathcal{M}$. Similarly, if for some $b_0 \in \mathcal{M}^\star$ the set $\mathcal{P}^\star(b_0)$ is bounded (resp. unbounded) then $\mathcal{P}^\star(b)$ is bounded (resp. unbounded) for all $b \in \mathcal{M}^\star$. Moreover, it holds that*

(i) the set $\mathcal{M}$ is non-empty and equal to an m-dimensional convex cone.

(ii) if the dual set $\mathcal{N}$ is non-empty then it holds that $\mathcal{M} = \mathcal{M}^\star$.

(iii) if the dual set $\mathcal{N}$ is non-empty and bounded then $\mathcal{M} = \mathcal{M}^\star = \mathbb{R}^m$.

The following discussion on the assumptions is a consequence of Lemma 5.1. We first collect sufficient conditions for assumption (**A1**).

Corollary 5.2 (Sufficiency for (**A1**))**.** *The following statements hold.*

*(i) If $\mathcal{N}$ is non-empty and $\mathcal{P}(b)$ is bounded for some $b \in \mathcal{M}$ then assumption (**A1**) holds for all $b \in \mathcal{M}$.*

*(ii) If $\mathcal{N}$ is non-empty, bounded and $\mathcal{P}^\star(b)$ is bounded for some $b \in \mathbb{R}^m$ then assumption (**A1**) holds for all $b \in \mathbb{R}^m$.*

Certainly, if $\mathcal{P}^\star(b) \neq \varnothing$ then (**A1**) is equivalent to $\mathcal{P}^\star(b)$ being bounded. The latter property is independent on b and equivalent to the set $\left\{x \in \mathbb{R}^d \mid Ax = 0,\, x \geq 0,\, c^t x = 0\right\}$ being empty. A sufficient condition for that is boundedness of $\mathcal{P}(b)$ that can be easily checked in certain settings.

Lemma 5.3 (Sufficiency for $\mathcal{P}(\mathbf{b})$ bounded)**.** *Suppose that A has non-negative entries and no column of A equals $0 \in \mathbb{R}^m$. Then $\mathcal{P}(b)$ is bounded (possibly infeasible) for all $b \in \mathbb{R}^m$.*

It is noteworthy that if the dual feasible set $\mathcal{N}$ is non-empty and bounded, then $\mathcal{P}^\star(b) \neq \varnothing$ for all $b \in \mathbb{R}^m$, but $\mathcal{P}(b)$ is necessarily unbounded (Clark, 1961). Thus, $\mathcal{N}$ is unbounded under the conditions of Lemma 5.3. We emphasize that assumption (**A2**) is neither easy to verify nor expected to hold for most structured linear programs. Indeed, under (**A1**) assumption (**A2**) is equivalent to all dual basic solutions being non-degenerate (Lemma 2.5). However, degeneracy in linear programs is often the case rather the exception (Bertsimas & Tsitsiklis, 1997). Notably, if (**A2**) and (**A1**) are satisfied the set $\mathcal{P}^\star(b)$ is singleton.
The assumption (**B1**) has to be checked for each particular case and can usually be verified by an application of the central limit theorem (for a particular example see Section 6). Assumption (**B2**) is obviously necessary for the limiting distribution to exist. If the dual feasible set $\mathcal{N}$ is non-empty and bounded and (**B1**) holds then (**B2**) is always satisfied. A more refined statement is the following.

Lemma 5.4 (Sufficiency for (**B2**))**.** *Consider the set $\mathcal{P}(b_0)$ assumed to be non-empty. Then $\mathcal{P}(b)$ is non-empty for all b sufficiently close to b_0 if*

(i) the set $\mathcal{P}(b_0)$ contains a non-degenerate feasible basic solution.

(ii) Slater's constraint qualification[7] *holds.*

In particular, if the dual feasible set $\mathcal{N}$ is non-empty and (**B1**) *holds then both conditions* (i) *and* (ii) *are sufficient for* (**B2**).

Joint convergence. Our goal here is to state useful conditions such that the random vector $\left(\alpha_n^{\mathcal{K}}, G_n\right)$ jointly converges[8] in distribution to some limit random variable $\left(\alpha^{\mathcal{K}}, G\right)$ on the space $\Delta_{|\mathcal{K}|} \times \mathbb{R}^m$. By assumption (**B1**), $G_n \to G$ in distribution, and a necessary condition for the joint distributional convergence of $(\alpha_n^{\mathcal{K}}, G_n)$ is that $\alpha_n^{\mathcal{K}}$ has a distributional limit $\alpha^{\mathcal{K}}$. There is no reason to expect $\alpha_n^{\mathcal{K}}$ and G_n to be independent, as discussed at the end of this section. We give a weaker condition than independence that is formulated in terms of the conditional distribution of $\alpha_n^{\mathcal{K}}$ given G_n (or, equivalently, given $b_n = b + G_n/r_n$). These conditions are natural in the sense that if $b_n = g$, then the choice of solution $x^\star(g)$, as encapsulated by the $\alpha_n^{\mathcal{K}}$'s, is determined by the specific linear program solver in use.
Treating conditional distributions rigorously requires some care and machinery. Let $\mathcal{Z} = \mathcal{Z}^{\mathcal{K}} = \Delta_{|\mathcal{K}|} \times \mathbb{R}^m$ and for $\varphi : \mathcal{Z} \to \mathbb{R}$ denote

$$\|\varphi\|_\infty = \sup_z |\varphi(z)|, \qquad \|\varphi\|_{\mathrm{Lip}} = \sup_{z_1 \neq z_2} \frac{|\varphi(z_1) - \varphi(z_2)|}{\|z_1 - z_2\|}, \qquad \|\varphi\|_{\mathrm{BL}} = \|\varphi\|_\infty + \|\varphi\|_{\mathrm{Lip}}.$$

We say that φ is bounded Lipschitz if it belongs to $\mathrm{BL}(\mathcal{Z}) = \{\varphi : \mathcal{Z} \to \mathbb{R} \mid \|\varphi\|_{\mathrm{BL}} \leq 1\}$. The bounded Lipschitz metric

$$\mathrm{BL}(\mu_1, \mu_2) \coloneqq \sup_{\varphi \in \mathrm{BL}(\mathcal{Z})} \left| \int_{\mathcal{Z}} \varphi(z)\, d(\mu_1 - \mu_2)(z) \right| \tag{5.1}$$

[7] The feasible set $\mathcal{P}(b_0)$ contains a positive element $x \in (0, \infty)^d$.

[8] Recall that the $\alpha_n^{\mathcal{K}}$ represent random weights (summing up to one) for each optimal basis I_k, $k \in \mathcal{K}$ for the case that $A_n^{\mathcal{K}}$ occurs, i.e., that several bases yield primal optimal solutions and hence any convex combination is also optimal.

is well-known to metrize convergence in distribution of (probability) measures on $\mathcal{Z}$ (Dudley, 2002, Theorem 11.3.3). According to the disintegration theorem (see Kallenberg (1997, Theorem 5.4), Dudley (2002, Section 10.2) or Chang & Pollard (1997) for details), we may write the joint distribution of $\left(\alpha_n^{\mathcal{K}}, b_n\right)$ as an integral of conditional distributions $\mu_{n,g}^{\mathcal{K}}$ that represent the distribution of $\alpha_n^{\mathcal{K}}$ given that $b_n = g$. More precisely, $g \mapsto \mu_{n,g}^{\mathcal{K}}$ is measurable from $\mathbb{R}^m$ to the metric space of probability measures on $\Delta_{|\mathcal{K}|}$ with the bounded Lipschitz metric, so that for any $\varphi \in \mathrm{BL}(\mathcal{Z})$ it holds that

$$\mathbb{E}\varphi(\alpha_n^{\mathcal{K}}, b_n) = \mathbb{E}\psi_n(b_n), \qquad \psi_n(g) = \int_{\Delta_{|\mathcal{K}|}} \varphi(\alpha, g) d\mu_{n,g}^{\mathcal{K}}(\alpha),$$

where $\psi_n : \mathbb{R}^m \to \mathbb{R}$ is a measurable function. The joint distribution of $(\alpha_n^{\mathcal{K}}, G_n)$ is determined by the collection of expectations

$$\mathbb{E}\varphi(\alpha_n^{\mathcal{K}}, G_n) = \mathbb{E}\psi_n(G_n) = \mathbb{E}\psi_n\left(r_n(b_n - b)\right), \qquad \varphi \in \mathrm{BL}(\mathcal{Z}).$$

Our sufficient condition for joint convergence is given by the following lemma. It is noteworthy that the spaces $\mathbb{R}^m$ and $\Delta_{|\mathcal{K}|}$ can be replaced with arbitrary Polish spaces, and even more general spaces, as long as the disintegration theorem is valid.

Lemma 5.5. *Let $\{\mu_g^{\mathcal{K}}\}_{g \in \mathbb{R}^m}$ be a collection of probability measures on $\Delta_{|\mathcal{K}|}$ such that the map $g \mapsto \mu_g^{\mathcal{K}}$ is continuous at G-almost any g, and suppose that $\mu_{n,g}^{\mathcal{K}} \to \mu_g^{\mathcal{K}}$ uniformly with respect to the bounded Lipschitz metric* BL*. Then $(\alpha_n^{\mathcal{K}}, G_n)$ converges in distribution to a random vector $(\alpha^{\mathcal{K}}, G)$ satisfying*

$$\mathbb{E}\varphi(\alpha^{\mathcal{K}}, G) = \mathbb{E}_G \int_{\Delta_{|\mathcal{K}|}} \varphi(\alpha, G) d\mu_G^{\mathcal{K}}(\alpha) \coloneqq \mathbb{E}\psi(G)$$

for any continuous bounded function $\varphi \in \mathrm{BL}(\mathcal{Z})$ (this determines the distribution of the random vector $(\alpha^{\mathcal{K}}, G)$ completely). Moreover, if $\mathcal{L}$ denotes the distribution of a random vector, then the rate of convergence can be quantified as

$$\mathrm{BL}(\mathcal{L}[(\alpha_n^{\mathcal{K}}, G_n)], \mathcal{L}[(\alpha^{\mathcal{K}}, G)]) \le \sup_g \mathrm{BL}(\mu_{n,g}^{\mathcal{K}}, \mu_g^{\mathcal{K}}) + (1 + L)\mathrm{BL}(\mathcal{L}[G_n], \mathcal{L}[G]),$$

where $L \coloneqq \sup_{g_1 \neq g_2} \mathrm{BL}(\mu_{g_1}^{\mathcal{K}}, \mu_{g_2}^{\mathcal{K}}) / \|(g_1 - g_2)\| \in [0, \infty]$. The supremum with respect to g can be replaced by an essential supremum.

The conditions of Lemma 5.5 (and hence the joint convergence in Theorem 3.1) will be satisfied in many practical situations. For example, given b_n and an initial basis for the simplex method, its output is determined by the pivoting rule (for a general overview see Terlaky & Zhang (1993) and references therein). Deterministic pivoting rules lead to degenerate conditional distributions of $\alpha_n^{\mathcal{K}}$ given $b_n = g$, whereas random pivoting rules may lead to non-degenerate conditional distributions. In both cases these conditional distributions do not depend on n at all, but only on the input vector g. In particular, the uniform convergence in Lemma 5.5 is trivially fulfilled (the supremum is equal to zero). It is reasonable to assume that these conditional distributions depend continuously on g except for some boundary values that are contained in a lower-dimensional space (which will have measure zero under the absolutely continuous random vector G).

6 Optimal Transport

Optimal transport (OT) dates back to the French mathematician and engineer Monge (1781). Roughly speaking, it seeks to transport objects from one collection of locations to another in the most economical manner. Apart from the work of Appell (1887), much of the progress of OT began in the mid-twentieth century, firstly due to its practical relevance in economics. Indeed, much of the theory of linear programming including the simplex algorithm has been motivated by findings for OT with early contributions by Hitchcock (1941); Kantorovich (1942); Dantzig (1948) and Koopmans (1951). Since then a surprisingly rich theory has emerged with important contributions by Kantorovich & Rubinstein (1958); Zolotarev (1976); Sudakov (1979); Kellerer (1984); Rachev (1985); Brenier (1987); Smith & Knott (1987); McCann (1997); Jordan et al. (1998); Ambrosio et al. (2008) and Lott & Villani (2009), among many others. We also refer to the excellent monographs by Rachev & Rüschendorf (1998); Villani (2008) and Santambrogio (2015) for further details. In fact, OT has recently gained renewed interest especially as computational progress paves the way to explore novel fields of applications such as imaging (Rubner et al., 2000; Solomon et al., 2015), machine learning (Frogner et al., 2015; Arjovsky et al., 2017; Peyré & Cuturi, 2019), and statistical data analysis (Chernozhukov et al., 2017; Sommerfeld & Munk, 2018; del Barrio et al., 2019; Panaretos & Zemel, 2019).

On a finite space $\mathcal{X} = \{x_1, \dots, x_N\}$ equipped with some underlying cost $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ OT between two probability measures $r, s \in \Delta_N \coloneqq \{r \in \mathbb{R}^N \,|\, \mathbf{1}_N^T r = 1,\, r_i \geq 0\}$ is equal to the linear program

$$OT(r,s) = \min_{\pi \in \Pi(r,s)} \sum_{i,j=1}^{N} c_{ij}\pi_{ij}, \tag{OT}$$

where $c_{ij} = c(x_i, x_j)$ and the set $\Pi(r,s)$ denotes all non-negative matrices with row and column sum equal to r and s, respectively. OT comprises the challenge to find an optimal solution termed *OT coupling* $\pi^\star(r,s)$ between r and s such that the integrated cost is minimal among all possible couplings. We denote by $\Pi^\star(r,s)$ the set of all OT couplings. The dual problem is

$$\max_{\alpha,\beta \in \mathbb{R}^N} \quad r^T\alpha + s^T\beta \quad \text{s.t.} \quad \alpha_i + \beta_j \leq c_{ij},\ \forall\, i,j \in [N]. \tag{DOT}$$

In our context reflecting many practical situations (**?**), the measures r and s are unknown and need to be estimated from data. To this end, we assume to have access to independent and identically distributed (i.i.d.) $\mathcal{X}$-valued random variables $X_1, \dots, X_n \sim r$, where a reasonable proxy for the measure r is its empirical version $\hat{r}_n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. As an illustration of our general theory, we focus on limit theorems that asymptotically $(n \to \infty)$ characterize the fluctuations of an estimated coupling $\pi^\star(\hat{r}_n, s)$ around $\pi^\star(r,s)$. For the sake of readability, we focus primarily on the one sample case, where only r is replaced by $\hat{r}_n$ but include a short account on the case that both measures are estimated.

A few words regarding the assumptions from Section 2 in the OT context are in order. Assumption (**A1**) always holds, since $\Pi(r,s) \subseteq [0,1]^{N^2}$ is bounded and contains the independence coupling rs^T. Assumption (**A2**) that according to Lemma 2.5 is equivalent to all dual feasible basic solutions for (DOT) being non-degenerate, however, does not always hold. Sufficient conditions for (**A2**) to hold in OT are given in Subsection 6.1. Concerning the probabilistic assumptions, we notice that (**B2**) always holds as for any (possibly random) pair of measures $(\hat{r}_n, s)$ the set $\Pi(\hat{r}_n, s)$ is non-empty and bounded. Assumption (**B1**) is easily verified by an application of the multivariate central limit theorem. Indeed, the multinomial process of empirical frequencies $\sqrt{n}(r - \hat{r}_n)$ converges weakly to a centered Gaussian random vector $G(r) \sim \mathcal{N}(0, \Sigma(r))$ with covariance

$$\Sigma(r) \coloneqq \begin{bmatrix} r_1(1-r_1) & -r_1r_2 & \dots & -r_1r_N \\ -r_1r_2 & r_2(1-r_2) & \dots & -r_2r_N \\ \vdots & & \ddots & \vdots \\ -r_1r_N & -r_2r_N & \dots & r_N(1-r_N) \end{bmatrix}. \tag{6.1}$$

Notably, $\Sigma(r)$ is singular and $G(r)$ fails to be absolutely continuous with respect to Lebesgue measure. A slight modification allows to circumvent this issue. The constraint matrix in OT,

$$A = \begin{pmatrix} \mathbf{1}_N^T & & \\ & \ddots & \\ & & \mathbf{1}_N^T \\ \mathbf{I}_N & \dots & \mathbf{I}_N \end{pmatrix} \in \mathbb{R}^{2N \times N^2}, \tag{6.2}$$

has rank $2N-1$. Letting $r_\dagger = r_{[N-1]} \in \mathbb{R}^{N-1}$ denote the first $N-1$ coordinates of $r \in \mathbb{R}^N$ and $A_\dagger \in \mathbb{R}^{(2N-1)\times N^2}$ denote A with its N-th row removed, it holds that

$$\Pi(r,s) = \left\{ \pi \in \mathbb{R}^{N^2} \mid A_\dagger \pi = \begin{bmatrix} r_\dagger \\ s \end{bmatrix}, \, \pi \geq 0 \right\}. \tag{6.3}$$

The limiting random variable for $\sqrt{n}(r_\dagger - \hat{r}_{\dagger_n})$, as n tends to infinity, is equal to $G(r_\dagger)$ following an absolutely continuous distribution if and only if $r_\dagger > 0$ and $\|r_\dagger\|_1 < 1$. Equivalently, r is in the relative interior of Δ_N (denoted $\mathrm{ri}(\Delta_N)$), i.e., $0 < r \in \Delta_N$. Under this condition (**A1**), (**B1**) and (**B2**) hold and from the main result in Theorem 3.1 we immediately deduce the limiting distribution of optimal OT couplings.

Theorem 6.1 (Distributional limit law for OT couplings)**.** *Consider the optimal transport problem* (OT) *between two probability measures $r, s \in \mathrm{ri}(\Delta_N)$ and let $\hat{r}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure derived by i.i.d. random variables $X_1, \dots, X_n \sim r$. If sample size n tends to infinity, then there exists a sequence $\pi_n^\star(r,s) \in \Pi^\star(r,s)$ such that*

$$\sqrt{n}\left(\pi^\star(\hat{r}_n, s) - \pi_n^\star(r,s)\right) \xrightarrow{D} \sum_{\mathcal{K}} \mathbb{1}_{\{G(r_\dagger) \in H_{\mathcal{K}} \setminus \cup_{k \notin \mathcal{K}} H_k\}} \, \alpha^{\mathcal{K}} \otimes \pi(I_{\mathcal{K}}, [G(r_\dagger), 0_N]) \tag{6.4}$$

with $G(r_\dagger) = (G^1(r_\dagger), 0)$. If further assumption (**A2**) *holds, then $\Pi^\star(r,s) = \{\pi^\star(r,s)\}$ and*

$$\sqrt{n}\left(\pi^\star(\hat{r}_n, s) - \pi^\star(r,s)\right) \xrightarrow{D} \sum_{k=1}^{K} \mathbb{1}_{\{G(r_\dagger) \in H_k\}} \, \pi(I_k, [G(r_\dagger), 0_N]).$$

Remark 6.2. *The two sample case presents an additional challenge. By the multivariate central limit theorem we have for $\min(m,n) \to \infty$ and $\frac{m}{n+m} \to \lambda \in (0,1)$ that*

$$\sqrt{\frac{nm}{n+m}} \left(\begin{bmatrix} \widehat{r_\dagger}_n \\ \hat{s}_m \end{bmatrix} - \begin{bmatrix} r_\dagger \\ s \end{bmatrix} \right) \xrightarrow{D} G_\lambda\left(r_\dagger, s\right) \coloneqq \left(\sqrt{\lambda} G^1(r_\dagger), \sqrt{1-\lambda} G^2(s) \right) \tag{6.5}$$

with $G^1(r_\dagger)$ and $G^2(s)$ independent and the compound limit law following a centered Gaussian distribution with block diagonal covariance matrix, where the two blocks are given by (6.1)*, respectively. However, the limit law fails to be absolutely continuous. Nevertheless, the distributional limit theorem for OT couplings remains valid in this case and there exists a sequence $\pi_{n,m}^\star(r,s) \in \Pi^\star(r,s)$ such that*

$$\sqrt{\frac{nm}{n+m}} \left(\pi^\star(\hat{r}_n, \hat{s}_m) - \pi_{n,m}^\star(r,s)\right) \xrightarrow{D} \sum_{\mathcal{K}} \mathbb{1}_{\{G_\lambda(r_\dagger, s) \in H_{\mathcal{K}} \setminus \cup_{k \notin \mathcal{K}} H_k\}} \, \alpha^{\mathcal{K}} \otimes \pi(I_{\mathcal{K}}, G_\lambda\left(r_\dagger, s\right)).$$

We provide further details in Appendix B*.*

We emphasize that once a limit law for the OT coupling is available, one can derive limit laws for sufficiently smooth functionals thereof. As examples let us mention the OT curve (Klatt et al., 2020) and OT geodesics (McCann, 1997). The details are omitted for brevity and instead, we provide an illustration of the distributional limit theorem (Theorem 6.1).

Example 6.3. *We consider a ground space $\mathcal{X} = \{x_1 < x_2 < x_3\} \subset \mathbb{R}$ consisting of $N = 3$ points with cost $c = (0, |x_1 - x_2|, |x_1 - x_3|, |x_2 - x_1|, 0, |x_2 - x_3|, |x_3 - x_1|, |x_3 - x_2|, 0) \in \mathbb{R}^9$ for which OT then reads as*

$$\min_{\pi \in \mathbb{R}^9} \quad c^T \pi \quad s.t. \quad A_\dagger \pi = \begin{bmatrix} r_\dagger \\ s \end{bmatrix}, \pi \geq 0$$

with constraint matrix $A_\dagger \in \mathbb{R}^{5\times 9}$. A basis I is a subset of cardinality five out of the column index set $\{1, \ldots, 9\}$ such that $(A_\dagger)_I$ is of full rank. For OT it is convenient to think of a feasible solution in terms of a transport matrix $\pi \in \mathbb{R}^{3\times 3}$ with π_{ij} encoding mass transport from source i to destination j. For instance, the basis $I = \{1, 2, 3, 5, 9\}$ corresponds to the transport scheme

$$TS(\{1,2,3,5,9\}) \coloneqq \begin{pmatrix} * & * & * \\ & * & \\ & & * \end{pmatrix},$$

where each possible non-zero entry is marked by a star and specific values depend on the measures r and s. In particular, to basis I corresponds the (possibly infeasible) basic solution $\pi(I, (r_\dagger, s)) = (A_\dagger)_I^{-1}(r_\dagger, s)$ that we illustrate in terms of its transport scheme by

$$\pi\left(I, (r_\dagger, s)\right) = \begin{pmatrix} s_1 & s_2 - r_2 & r_1 + r_2 - s_1 - s_2 \\ & r_2 & \\ & & s_1 + s_2 + s_3 - r_1 - r_2 \end{pmatrix} = \begin{pmatrix} s_1 & s_2 - r_2 & s_3 - r_3 \\ & r_2 & \\ & & r_3 \end{pmatrix},$$

where $r = (r_\dagger, 1 - \|r_\dagger\|_1) \in \mathbb{R}^3$ and the second equality employs that r and s sum up to one. Obviously, $\pi(I, (r_\dagger, s))$ is feasible if and only if $s_2 \geq r_2$ and $s_3 \geq r_3$. Suppose that the measures are equal $r = s$. Then the transport problem attains a unique solution supported on the diagonal, i.e., all the mass remains at its current location. A straightforward computation yields $K = 8$ primal and dual optimal bases

$$TS(I_1) = \begin{pmatrix} * & * & \\ & * & \\ & * & * \end{pmatrix}, \quad TS(I_2) = \begin{pmatrix} * & & \\ * & * & * \\ & & * \end{pmatrix}, \quad TS(I_3) = \begin{pmatrix} * & & \\ * & * & \\ * & & * \end{pmatrix}, \quad TS(I_4) = \begin{pmatrix} * & & * \\ & * & * \\ & & * \end{pmatrix},$$
$$TS(I_5) = \begin{pmatrix} * & * & * \\ & * & \\ & & * \end{pmatrix}, \quad TS(I_6) = \begin{pmatrix} * & & \\ & * & \\ * & * & * \end{pmatrix}, \quad TS(I_7) = \begin{pmatrix} * & & \\ * & * & \\ & * & * \end{pmatrix}, \quad TS(I_8) = \begin{pmatrix} * & * & \\ & * & * \\ & & * \end{pmatrix}.$$

For example, the transport scheme $TS(I_1)$ corresponds to basis $I_1 = \{1, 2, 5, 8, 9\}$ and induces an invertible matrix A_{I_1}. Omitting the superscript $m_0 = 5$ for clarity, the respective closed convex cones H_k for $1 \leq k \leq K$ as defined in (4.2) are

$$\begin{aligned}
H_1 &= \left\{v \in \mathbb{R}^5 \mid v_1 \geq v_3,\, v_1 + v_2 \leq v_3 + v_4\right\}, & H_2 &= \left\{v \in \mathbb{R}^5 \mid v_1 \leq v_3,\, v_1 + v_2 \geq v_3 + v_4\right\}, \\
H_3 &= \left\{v \in \mathbb{R}^5 \mid v_2 \geq v_4,\, v_1 + v_2 \leq v_3 + v_4\right\}, & H_4 &= \left\{v \in \mathbb{R}^5 \mid v_1 \geq v_3,\, v_2 \geq v_4\right\}, \\
H_5 &= \left\{v \in \mathbb{R}^5 \mid v_2 \leq v_4,\, v_1 + v_2 \geq v_3 + v_4\right\}, & H_6 &= \left\{v \in \mathbb{R}^5 \mid v_1 \leq v_3,\, v_2 \leq v_4\right\}, \\
H_7 &= \left\{v \in \mathbb{R}^5 \mid v_1 \leq v_3,\, v_1 + v_2 \leq v_3 + v_4\right\}, & H_8 &= \left\{v \in \mathbb{R}^5 \mid v_1 \geq v_3,\, v_1 + v_2 \geq v_3 + v_4\right\}.
\end{aligned}$$

*Each of these cones is an intersection of two proper half-spaces, respectively. Some of these cones exhibit non-trivial intersections and in particular (**A2**) fails to hold. Such* *cases arise for the pairs* $\{I_3, I_7\}$, $\{I_6, I_7\}$, $\{I_4, I_8\}$ *and* $\{I_5, I_8\}$. *The intersections of the corresponding cones are given by*

$$H_3 \cap H_7 = \left\{ v \in \mathbb{R}^5 \mid v_2 \geq v_4,\, v_1 + v_2 \leq v_3 + v_4 \right\}, \qquad H_6 \cap H_7 = \left\{ v \in \mathbb{R}^5 \mid v_1 \leq v_3,\, v_2 \leq v_4 \right\},$$
$$H_5 \cap H_8 = \left\{ v \in \mathbb{R}^5 \mid v_2 \leq v_4,\, v_1 + v_2 \geq v_3 + v_4 \right\}, \qquad H_4 \cap H_8 = \left\{ v \in \mathbb{R}^5 \mid v_1 \geq v_3,\, v_2 \geq v_4 \right\}.$$

The weak convergence in (6.5) *and together with OT for* $p = 1$ *and* $r = s$ *then leads to the distributional limit law for OT couplings*

$$M(\mathbf{G}) = \sum_{\mathcal{K} \in \{\{1\},\{2\},\{3,7\},\{6,7\},\{4,8\},\{5,8\}\}} \mathbb{1}_{\{\mathbf{G} \in H_{\mathcal{K}} \setminus \cup_{k \notin \mathcal{K}} H_k\}} \alpha^{\mathcal{K}} \otimes x(I_{\mathcal{K}}, G).$$

Although $K = 8$, *there are only four distinct dual solutions:* $\lambda(I_1)$, $\lambda(I_2)$, $\lambda(I_7)$ *and* $\lambda(I_8)$.

6.1 Degeneracy and Uniqueness in Optimal Transport

This subsection provides sufficient conditions for assumption (**A2**) to hold. In view of Lemma 2.5 and since for OT assumption (**A1**) is always satisfied, assumption (**A2**) is equivalent to non-degeneracy of all dual optimal basic solutions. Notably, this implies uniqueness of the OT coupling. Conversely, if for a given cost the OT coupling is unique for all $r, s \in \Delta_N$, then (**A2**) holds. We begin with a sufficient criterion for (**A2**) only depending on the cost.

Lemma 6.4. *Suppose that the following holds for the cost function* c. *For any* $n \geq 2$ *and any family of indices* $\{(i_k, j_k)\}_{1 \leq k \leq n}$ *with all* i_k *pairwise different and all* j_k *pairwise different it holds that*

$$\sum_{k=1}^{n} c_{i_k j_k} \neq \sum_{k=1}^{n} c_{i_k j_{k-1}}, \qquad j_0 \coloneqq j_n. \tag{6.6}$$

Then all dual basic solutions are non-degenerate. In particular, (**A2**) *holds and the optimal OT coupling is unique for any pair of measures* $r, s \in \Delta_N$.

We are unaware of an explicit reference for condition (6.6) that is reminiscent to the well-known *cyclic monotonicity property* (Rüschendorf, 1996). Further, (6.6) can be thought of as dual to the condition of Klee & Witzgall (1968) that for any proper subsets $A, B \subset [N]$ not both empty

$$\sum_{i \in A} r_i \neq \sum_{j \in B} s_j \tag{6.7}$$

that guarantees every *primal* basic solution to be non-degenerate. Notably, (6.6) is satisfied for OT on the real line with cost $c(x, y) = |x - y|^p$ and measures with at least $N = 3$ support points if and only if $p > 0$ and $p \neq 1$. If the underlying space involves too many symmetries, such as a regular grid with cost defined by the underlying grid structure, it typically fails to hold. An alternative condition that ensures (**A2**) is the *strict Monge condition* that the cost c satisfies

$$c_{ij} + c_{i'j'} < c_{ij'} + c_{i'j}, \qquad \forall\, i < i', j < j', \tag{6.8}$$

possibly after relabelling the indices (Dubuc et al., 1999). This translates to easily interpretable statements on the real line.

Lemma 6.5. *Let* $\mathcal{X} \coloneqq \{x_1 < \ldots < x_N\}$ *be a set of* N *distinct ordered points on the real line. Suppose that the cost takes the form* $c(x, y) = f(|x - y|)$ *with* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *such that* $f(0) = 0$ *and either*

$$(i)\; f \text{ is strictly convex}, \qquad (ii)\; f \text{ is strictly concave}.$$

Then assumption (**A2**) *holds.*

The first statement follows by employing the Monge condition (see also McCann (1999, Proposition A2) for an alternative approach). The second case is more delicate, and indeed, the description of the unique optimal solution is more complicated (see Appendix B). In fact, in both cases the unique transport coupling can be computed by the Northwest corner algorithm (Hoffman, 1963). Typical costs covered by Lemma 6.5 are $c(x,y) = |x-y|^p$ for any $p > 0$ with $p \neq 1$. Indeed, for $p = 0$ or $p = 1$, uniqueness often fails (see Remark 6.7). In a general linear program (P_b), the set of costs c for which (**A2**) fails to hold has Lebesgue measure zero (e.g., Bertsimas & Tsitsiklis, 1997). Here we provide a result in the same flavour for OT.

Proposition 6.6. *Let μ and ν be absolutely continuous on $\mathbb{R}^D$, with $D \geq 2$, and let $c(x,y) = \|x-y\|_q^p$, where $p \in \mathbb{R} \setminus \{0\}$ and $q \in (0,\infty]$ are such that if $p = 1$ then $q \notin \{1,\infty\}$. For probability vectors $r,s \in \Delta_N$ define the probability measures $r(\mathbf{X}) = \sum_{k=1}^N r_k \delta_{X_k}$ and $s(\mathbf{Y}) = \sum_{k=1}^N s_k \delta_{Y_k}$ with two independent collections of i.i.d. $\mathbb{R}^D$-valued random variables $X_1,\dots,X_N \sim \mu$ and $Y_1,\dots,Y_N \sim \nu$. Then (6.6) holds almost surely for the optimal transport (OT). In particular, with probability one for any $r,s \in \Delta_N$ and pair of marginals $r(\mathbf{X})$ and $s(\mathbf{Y})$, the corresponding optimal transport coupling is unique.*

See Wang et al. (2013) for a related result for $p = q = 2$ and *fixed* marginals r,s. Note that Proposition (6.6) includes the Coulomb case ($p = -1$) that has applications in physics (Cotar et al., 2013). As the proof details, the result is valid for piece-wise analytic (non-constant) functions.

Remark 6.7 (Non-uniqueness)**.** *Let μ be uniform on $[0,1]^D$ and ν be uniform on $[1,2]^D + (2,0,0,\dots,0)$. Then, with probability one, all transport couplings bear the same cost if $p = 0$ or if $p = 1$ and $q \in \{1,\infty\}$. Thus, for $N \geq 2$ uniqueness fails.*

Acknowledgements M. Klatt and A. Munk gratefully acknowledge support from the DFG Research Training Group 2088 *Discovering structure in complex data: Statistics meets Optimization and Inverse Problems* and *CRC 1456 Mathematics of Experiment*. Y. Zemel was supported in part by Swiss National Science Foundation Grant 178220, and in part by a U.K. Engineering and Physical Sciences Research Council programme grant.

References

Ambrosio, L., Gigli, N., & Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Appell, P. (1887). Mémoire sur les déblais et les remblais des systemes continus ou discontinus. *Mémoires présentes par divers Savants à l'Académie des Sciences de l'Institut de France*, 29, 1–208.

Arjovsky, M., Chintalah, S., & L, B. (2017). Wasserstein generative adversarial networks. *Proceedings of Machine Learning Research*, 70, 214–223.

Beale, E. M. (1955). On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2), 173–184.

Bereanu, B. (1963). Decision regions and minimum risk solutions in linear programming. In *Colloquium on applications of mathematics to economics, Budapest*, pages 37–42.

Bereanu, B. (1976). The continuity of the optimum in parametric programming and applications to stochastic programming. *Journal of Optimization Theory and Applications*, 18(3), 319–333.

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

Billingsley, P. (1999). *Convergence of Probability Measures*. New York: Wiley, 2nd edition.

Böhm, V. (1975). On the continuity of the optimal policy set for linear programs. *SIAM Journal on Applied Mathematics*, 28(2), 303–306.

Bonnans, J. F. & Shapiro, A. (2000). *Perturbation Analysis of Optimization Problems*. Springer Science & Business Media.

Brenier, Y. (1987). Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305, 805–808.

Brualdi, R. A. (2006). *Combinatorial Matrix Classes*, volume 13. Cambridge University Press.

Chang, J. T. & Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51(3), 287–317.

Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1), 223–256.

Clark, F. E. (1961). Remark on the constraint sets in linear programming. *The American Mathematical Monthly*, 68(4), 351–352.

Cotar, C., Friesecke, G., & Klüppelberg, C. (2013). Density functional theory and optimal transportation with Coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4), 548–599.

Cottle, R., Johnson, E., & Wets, R. J.-B. (2007). George B. Dantzig (1914–2005). *Notices of the AMS*, 54(3), 344–362.

Dang, N. V. (2015). Complex powers of analytic functions and meromorphic renormalization in QFT. *preprint arXiv:1503.00995*.

Dantzig, G. B. (1948). Programming in a linear structure. *Bulletin of the American Mathematical Society*, 54(11), 1074–1074.

Dantzig, G. B. (1955). Linear programming under uncertainty. *Management Science*, 1, 197 – 206.

Dantzig, G. B. (2016). *Linear programming and extensions*. Princeton university press.

De Loera, J. A., Rambau, J., & Santos, F. (2010). *Triangulations Structures for Algorithms and Applications*. Springer.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., & Mayo-Íscar, A. (2019). Robust clustering tools based on optimal transportation. *Statistics and Computing*, 29, 139–160.

Dubuc, S., Kagabo, I., & Marcotte, P. (1999). A note on the uniqueness of solutions to the transportation problem. *INFOR: Information Systems and Operational Research*, 37(2), 141–148.

Dudley, R. M. (2002). *Real Analysis and Probability*, volume 74. Cambridge University Press.

Dupačová, J. (1987). Stochastic programming with incomplete information: a surrey of results on postoptimization and sensitivity analysis. *Optimization*, 18(4), 507–532.

Dupačová, J. & Wets, R. J.-B. (1988). Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *The Annals of Statistics*, 16(4), 1517–1549.

Ewbank, J. B., Foote, B. L., & Kumin, H. L. (1974). A method for the solution of the distribution problem of stochastic linear programming. *SIAM Journal on Applied Mathematics*, 26(2), 225–238.

Ferguson, A. R. & Dantzig, G. B. (1956). The allocation of aircraft to routes: An example of linear programming under uncertain demand. *Management Science*, 3(1), 45–73.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., & Poggio, T. A. (2015). Learning with a Wasserstein loss. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28*, pages 2053–2061. Curran Associates, Inc.

Gal, T. & Greenberg, H. J. (2012). *Advances in sensitivity analysis and parametric programming*, volume 6. Springer Science & Business Media.

Galichon, A. (2018). *Optimal transport methods in economics.* Princeton University Press.

Gangbo, W. & McCann, R. J. (1996). The geometry of optimal transportation. *Acta Mathematica*, 177(2), 113–161.

Goldman, A. J. & Tucker, A. W. (1956). Theory of linear programming. *Linear inequalities and related systems*, 38, 53–97.

Greenberg, H. J. (1986). An analysis of degeneracy. *Naval Research Logistics Quarterly*, 33(4), 635–655.

Guddat, J., Hollatz, H., & Bank, B. (1974). Theorie der linearen parametrischen Optimierung. *Berlin, Akademic-Verlag.*

Hadigheh, A. G. & Terlaky, T. (2006). Sensitivity analysis in linear optimization: Invariant support set intervals. *European Journal of Operational Research*, 169(3), 1158–1175.

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

Hoffman, A. J. (1963). On simple linear programming problems. In *Proceedings of Symposia in Pure Mathematics*, volume 7, pages 317–327.

Jordan, R., Kinderlehrer, D., & Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1), 1–17.

Kallenberg, O. (1997). *Foundations of Modern Probability.* Springer-Verlag, 2nd edition.

Kantorovich, L. V. (1939). Mathematical methods in the organization and planning of production. *Publication House of the Leningrad State University*, 6, 336–422.

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

Kantorovich, L. V. & Rubinstein, G. S. (1958). On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7), 52–59.

Kellerer, H. G. (1984). Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 67(4), 399–432.

King, A. J. & Rockafellar, R. T. (1993). Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18(1), 148–162.

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

Klee, V. & Witzgall, C. (1968). Facets and vertices of transportation polytopes. *Mathematics of the Decision Sciences*, 1, 257–282.

Koopmans, T. C. (1949). Optimum utilization of the transportation system. *Econometrica: Journal of the Econometric Society*, 17, 136–146.

Koopmans, T. C. (1951). Efficient allocation of resources. *Econometrica: Journal of the Econometric Society*, 19(4), 455–465.

Lott, J. & Villani, C. (2009). Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, 169(3), 903–991.

Luenberger, D. G. & Ye, Y. (2008). *Linear and Nonlinear Programming.* New York: Springer.

McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in Mathematics*, 128(1), 153–179.

McCann, R. J. (1999). Exact solutions to the transportation problem on the line. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 455(1984), 1341–1380.

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

Panaretos, V. M. & Zemel, Y. (2019). Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Applications*, 6, 405–431.

Peyré, G. & Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6), 355–607.

Rachev, S. T. (1985). The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4), 647–676.

Rachev, S. T. & Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory, Volume II: Applications.* New York: Springer.

Robinson, S. M. (1977). A characterization of stability in linear programming. *Operations Research*, 25(3), 435–447.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2), 99–121.

Rüschendorf, L. (1996). On c-optimal random variables. *Statistics & Probability Letters*, 27(3), 267–270.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians.* Basel: Birkhäuser.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1), 169–186.

Shapiro, A. (1993). Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4), 829–845.

Shapiro, A. (2000). Statistical inference of stochastic optimization problems. In *Probabilistic constrained optimization*, pages 282–307. Springer.

Shapiro, A., Dentcheva, D., & Ruszczynski, A. (2021). *Lectures on stochastic programming: modeling and theory.* SIAM.

Smith, C. S. & Knott, M. (1987). Note on the optimal transportation of distributions. *Journal of Optimization Theory and Applications*, 52(2), 323–329.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., & Guibas, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4), 1–11.

Sommerfeld, M. & Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Methodological)*, 80(1), 219–238.

Sturmfels, B. & Thomas, R. R. (1997). Variation of cost functions in integer programming. *Mathematical Programming*, 77(2), 357–387.

Sudakov, V. N. (1979). *Geometric problems in the theory of infinite-dimensional probability distributions*, volume 141. American Mathematical Soc.

Tameling, C., Sommerfeld, M., & Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5), 2744–2781.

Terlaky, T. & Zhang, S. (1993). Pivot rules for linear programming: a survey on recent theoretical developments. *Annals of Operations Research*, 46(1), 203–233.

Vershik, A. (2002). L.V. Kantorovich and linear programming. In *Leonid Vital'evich Kantorovich: a man and a scientist*, volume 1, pages 130–152.

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

Walkup, D. W. & Wets, R. J.-B. (1967). Continuity of some convex-cone-valued mappings. *Proceedings of the American Mathematical Society*, 18(2), 229–235.

Walkup, D. W. & Wets, R. J.-B. (1969a). Lifting projections of convex polyhedra. *Pacific Journal of Mathematics*, 28(2), 465–475.

Walkup, D. W. & Wets, R. J.-B. (1969b). A Lipschitzian characterization of convex polyhedra. *Proceedings of the American Mathematical Society*,, pages 167–173.

Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., & Rohde, G. K. (2013). A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2), 254–269.

Ward, J. E. & Wendell, R. E. (1990). Approaches to sensitivity analysis in linear programming. *Annals of Operations Research*, 27(1), 3–38.

Wets, R. J.-B. (1980). The distribution problem and its relation to other problems in stochastic programming. In *Stochastic Programming*, pages 245–262. London: Academic Press.

Zolotarev, V. M. (1976). Metric distances in spaces of random variables and their distributions. *Mathematics of the USSR-Sbornik*, 30(3), 373.

A Omitted Proofs

Lemma 2.5. Dual non-degeneracy obviously implies (**A2**), so we only show the converse (in presence of (**A1**)). Suppose that $\lambda(I_j)$ is degenerate $1 \le j \le K$. Then the index set L of active constraints in the dual, i.e., the set of indices such that $[\lambda^T(I_j)A]_l = c_l$, is such that $I_j \subset L$. Let $Pos_j \subseteq I_j$ be the set of positive entries of the optimal primal basic solution $x(I_j, b)$. Then

$$L = I_j \sqcup L \setminus I_j = Pos \sqcup I_j \setminus Pos \sqcup L \setminus I_j \,.$$

Since the columns of A_{I_j} form a basis of $\mathbb{R}^m$, each other column a_z writes

$$a_z = \sum_{i \in Pos} y_i^z a_i + \sum_{s \in I_j \setminus Pos} y_s^z a_s , \qquad z \in L \setminus I_j \,. \tag{A.1}$$

Suppose there exists some index $z \in L \setminus I_j$ and $s \in I_j \setminus Pos$ such that $y_s^z \neq 0$. Then we can define a new basis $\widetilde{I} \coloneqq I_j \setminus \{s\} \cup \{z\}$ such that $\lambda(\widetilde{I}) = \lambda(I_j)$ and as $Pos_j \subseteq \widetilde{I}$ we conclude that $x(\widetilde{I}, b) = x(I_j, b)$. This contradicts (**A2**). Hence $y_s^z = 0$ for all $s \in I_j \setminus Pos_j$ in (A.1). Now suppose that $y_i^z > 0$ for some $i \in Pos_j$, so that $i_0 \in \arg\min_{i|y_i^z>0} \frac{x_i}{y_i^z}$ is well defined and the minimum is strictly positive. Expressing a_{i_0} as a linear combination of $A_{\{z\}\cup Pos_j \setminus \{i_0\}}$, we find that

$$b = \sum_{i \in Pos} x_i a_i = \sum_{\substack{i \in Pos \\ i \neq i_0}} x_i a_i + x_{i_0} \left(\frac{1}{y_{i_0}^z} a_z - \sum_{\substack{i \in Pos \\ i \neq i_0}} \frac{y_i^z}{y_{i_0}^z} a_i \right) = \frac{x_{i_0}}{y_{i_0}^z} a_z + \sum_{\substack{i \in Pos \\ i \neq i_0}} \left(x_i - \frac{x_{i_0} y_i^z}{y_{i_0}^z} \right) a_i = \sum_{i \in \widetilde{I}} \widetilde{x}_i a_i$$

for some proper choice of $\widetilde{x}_i$. By definition of i_0 we find that $\widetilde{x}_i$ are non-negative, so that $\widetilde{I}$ is a primal and dual optimal basis. Moreover, $\lambda(\widetilde{I}) = \lambda(I_j)$ that again contradicts (**A2**). We deduce for the representation in (A.1) that $y_i^z \le 0$. Consider the vector

$$w \coloneqq \mathrm{Aug}_{\{z\}}(1) - \mathrm{Aug}_{Pos}(y^z) \,.$$

By definition $w \geq 0$, $Aw = 0$. and $c^T w = 0$, so that $w \neq 0$ is a primal optimal ray, in contradiction of (**A1**). In total we see that if any basis I_j for $1 \le j \le K$ yields a degenerate

dual basic solution we can modify basis I_j to some I_i with $i \neq j$ and $1 \le i \le K$ such that $\lambda(I_j) = \lambda(I_j)$.
It is in principle possible that $\lambda(I_l)$ is dual optimal $K+1 \le l \le N$ but $x(I_l, b)$ is not primal optimal. Let us show that this cannot happen under assumption (**A2**). Consider any optimal primal basic solution $x(I_j, b)$ for $1 \le j \le K$ and denote by Pos_j its positivity set. Optimality of $\lambda(I_l)$ implies that its active set L contains $I_l \cup Pos_j$. As I_l is not a primal optimal basis, it holds that $Pos_j \notin I_l$, so that $|L| > m$ and $\lambda(I_l)$ is degenerate. But then we can modify basis I_l to some primal and dual optimal basis I_i for $1 \le i \le K$ such that $\lambda(I_i) = \lambda(I_l)$ is degenerate, in contradiction with (**A2**). Hence, any optimal dual basic solution is non-degenerate and induced by some primal and dual optimal basis I. □

Lemma 5.5. We need to show that for any $\varphi \in \mathrm{BL}(\mathcal{Z})$ we have that

$$\begin{aligned}|\mathbb{E}[\varphi(\alpha_n^{\mathcal{K}}, G_n)] - \mathbb{E}[\varphi(\alpha^{\mathcal{K}}, G)]| &= |\mathbb{E}\psi_n(G_n) - \mathbb{E}\psi(G)| \\ &\le |\mathbb{E}[\psi_n(G_n) - \psi(G_n)]| + |\mathbb{E}[\psi(G_n) - \psi(G)]|\end{aligned}$$

vanishes as $n \to \infty$. To bound the first term notice that for any fixed g it holds that $\|\alpha \mapsto \varphi(\alpha, g)\|_{\mathrm{BL}(\Delta_{|\mathcal{K}|})} \le \|\varphi\|_{\mathrm{BL}(\mathcal{Z})} \le 1$, so

$$|\psi_n(g) - \psi(g)| \le \left| \int_{\Delta_{|\mathcal{K}|}} \varphi(\alpha, g) d[\mu_{n,g}^{\mathcal{K}} - \mu_g^{\mathcal{K}}](\alpha) \right| \le \mathrm{BL}(\mu_{n,g}^{\mathcal{K}}, \mu_g^{\mathcal{K}}).$$

Hence, we find $\mathbb{E}|\psi_n(G_n) - \psi(G_n)| \le \sup_g \mathrm{BL}(\mu_{n,g}^{\mathcal{K}}, \mu_g^{\mathcal{K}})$ that tends to zero by assumption. Notice that the supremum can be an essential supremum, i.e., taken on set of full measure with respect to both (G_n) and (G) instead of the whole of $\mathbb{R}^m$. For the second term observe that $\|\psi\|_\infty \le \|\varphi\|_\infty$ and that

$$\begin{aligned}|\psi(g_1) - \psi(g_2)| &= \left| \int_{\Delta_{|\mathcal{K}|}} [\varphi(\alpha, g_1) - \varphi(\alpha, g_2)] d\mu_{g_1}^{\mathcal{K}}(\alpha) + \int_{\Delta_{|\mathcal{K}|}} \varphi(\alpha, g_2) d[\mu_{g_2}^{\mathcal{K}} - \mu_{g_1}^{\mathcal{K}}](\alpha) \right| \\ &\le \|\varphi\|_{\mathrm{Lip}} \|g_1 - g_2\| + \mathrm{BL}(\mu_{g_2}^{\mathcal{K}}, \mu_{g_1}^{\mathcal{K}}).\end{aligned}$$

Hence, we conclude that

$$\|\psi\|_{\mathrm{BL}(\mathbb{R}^m)} \le \|\varphi\|_{\mathrm{BL}(\mathcal{Z})} + L \le 1 + L.$$

Dividing ψ by its bounded Lipschitz norm, we find

$$\mathbb{E}|\psi(G_n) - \psi(G)| \le \|\psi\|_{\mathrm{BL}(\mathbb{R}^m)} \mathrm{BL}(\mathcal{L}(G_n), \mathcal{L}(G)) \le (1+L)\mathrm{BL}(\mathcal{L}(G_n), \mathcal{L}(G)).$$

This completes the proof for the quantitative statement. Joint convergence still follows if $g \mapsto \mu_g^{\mathcal{K}}$ is only continuous G-almost surely (but not Lipschitz). In fact, ψ is still continuous and bounded G-almost surely so that $\mathbb{E}\psi(G_n) \to \mathbb{E}\psi(G)$. Therefore, $\mathbb{E}\varphi(\alpha_n^{\mathcal{K}}, G_n) \to \mathbb{E}\varphi(\alpha^{\mathcal{K}}, G)$ for all $\varphi \in \mathrm{BL}(\mathcal{Z})$, which implies that $(\alpha_n^{\mathcal{K}}, G_n) \to (\alpha^{\mathcal{K}}, G)$ in distribution. □

B Optimal Transport

Theorem 6.1, two-sample. The only part where absolute continuity of $G = (G^1(r_\dagger), G^2(s))$ was required is when showing that the boundaries of the cones defined in (4.2) have zero probability with respect to G. We shall show that this is still the case, despite the singularity of $G^2(s)$.

The cones under consideration take the form

$$H_k = \bigcap_{i \in \mathrm{DP}_k} \left\{ v \in \mathbb{R}^{2N-1} \mid [\pi(I_k, v)]_i \geq 0 \right\}$$

(where $\pi(I_k, v)$ is viewed as an N^2-dimensional vector), and their boundaries satisfy

$$\partial H_k \subseteq \bigcup_{i \in \mathrm{DP}_k} \left\{ v \in \mathbb{R}^{2N-1} \mid [\pi(I_k, v)]_i = 0 \right\} \subseteq \bigcup_{i \in I_k} \left\{ v \in \mathbb{R}^{2N-1} \mid [\pi(I_k, v)]_i = 0 \right\}.$$

Let $w = (v_{[N-1]}, -\sum_{j=1}^{N-1} v_j, v_{\{N,...2N-1\}}) \in \mathbb{R}^{2N}$ be the augmented vector corresponding to v. In view of Brualdi (2006, Corollary 8.1.4), there exist $\mathsf{R}_{i,k} \subset \{1, \ldots, N\}$ and $\mathsf{S}_{i,k} \subset \{N+1, \ldots, 2N\}$, not both empty, such that

$$[\pi(I_k, v)]_i = \pm \left(\sum_{j \in \mathsf{R}_{i,k}} w_j - \sum_{l \in \mathsf{S}_{i,k}} w_l \right) = \begin{cases} \pm \left(\sum_{j \in \mathsf{R}_{i,k}} v_j - \sum_{l \in \mathsf{S}_{i,k}} v_{l-1} \right) & N \notin \mathsf{R}_{i,k} \\ \pm \left(-\sum_{j \in \mathsf{R}_{i,k}} v_j - \sum_{l \in \mathsf{S}_{i,k}} v_{l-1} \right) & N \in \mathsf{R}_{i,k}. \end{cases}$$

It suffices to show that for any pair of sets $\mathsf{R} \subseteq \{1, \ldots, N-1\}$ and $\mathsf{S} \subset \{1, \ldots, N\}$ that are not both empty,

$$\mathbb{P} \left(\sum_{j \in \mathsf{R}} G^1(r_\dagger)_i = \pm \sum_{l \in \mathsf{S}} G^2(s)_l \right) = 0.$$

Recall that $G^1(r_\dagger)$ is independent of $G^2(s)$ and admits a density on $\mathbb{R}^{N-1}$. Hence, when R is non-empty, the above probability is indeed zero. If R is empty, then S is a non-empty proper subset of $[N]$. Since $s \in \mathrm{ri}(\Delta_N)$, the kernel of $\Sigma(s)$ is the span of the vector of ones. Hence the distribution of $\sum_{k \in \mathsf{S}_i} G^2(s)_k$ is absolutely continuous, so it vanishes with probability zero. This completes the proof. □

Lemma 6.5. By elementary arguments the cost $c(x_i, x_j) = f(|x_i - x_j|)$ satisfies the strict Monge condition (6.8) when f is strictly convex. We thus only consider the case where f is strictly concave. Since it was assumed non-negative, finite, and with $f(0) = 0$, it must be that f is continuous and strictly increasing. Clearly, the optimal cost between μ and ν is finite. According to Gangbo & McCann (1996, Proposition 2.9), all the common mass must stay in place. Hence, we may assume that μ and ν are mutually singular. We prove a more general result, from which Lemma 6.5 follows immediately. □

Lemma B.1. *Let μ and ν be mutually singular and both supported on a finite union of intervals. Let f be finite, strictly concave, and strictly increasing on the supports of μ and ν. Then the optimal coupling between μ and ν with respect to the cost function $c(x, y) = f(|x - y|)$ is unique.*

Remark B.2. *If μ and ν have finite support, the assumption is satisfied. We believe that the statement is true for an arbitrary pair of measures μ and ν, but the above formulation is sufficient as in the context of the present μ and ν are anyway finitely supported. For example, the support could contain countably many intervals as long as there is "clear" starting point a_0 below; but M could be infinite.*

Proof. There is nothing to prove if $\mu = \nu = 0$, so we assume $\mu \neq \nu$. It follows from the assumptions that there exists a finite sequence of $M + 1 \geq 3$ real numbers

$$-\infty \leq a_0 < a_1 < a_2 < a_3 < .... < a_M \leq \infty$$

such that (interchanging μ and ν if necessary)

$$\mu([a_0,a_1]\cup[a_2,a_3]\cup[a_4,a_5]\cup\dots)=1;$$
$$\mu([a_1,a_2]\cup[a_3,a_4]\cup[a_5,a_6]\cup\dots)=1.$$

Let $m_0=\mu([a_0,a_1])$ and suppose that $m_0\le\nu([a_1,a_2])$. Define the quantile

$$a^\star=\inf\{a:\nu[a_1,a]\ge m_0\}\in[a_1,a_2].$$

We now claim that in any optimal coupling π between μ and ν, the μ-mass of $[a_0,a_1]$ must go to $[a_1,a^\star]$. Indeed, suppose that a positive μ-mass from $[a_0,a_1]$ goes strictly beyond $a^\star$. Then some mass from the support of μ but not in $[a_0,a_1]$ has to go to $[a_1,a^\star]$. Such a coupling gives positive measure to the set

$$[a_0,a_1]\times[a^\star+\epsilon,\infty)\bigcap[a_2,\infty]\times[a_1,a^\star]$$

for some $\epsilon>0$. Strict monotonicity of the cost function makes this sub-optimal, since this coupling entails sending mass from x_1 to y_1 and from x_2 to y_2 with $x_1<y_2<\min(x_2,y_1)$, (see Gangbo & McCann (1996, Theorem 2.3) for a rigorous proof). Hence the claim is proved. Let μ_1 be the restriction of μ to $[a_0,a_1]$ and ν_1 be the restriction of ν to $[a_1,a^\star]$ with mass m_0, namely $\nu_1(B)=\nu(B)$ if $B\subseteq[a_1,a^\star)$, $\nu_1(\{a^\star\})=m_0-\nu([a_1,a^\star))$ and $\nu(B)=0$ if $B\cap[a_1,a^\star]=\varnothing$. By definition of $a^\star$, ν_1 is a measure (i.e., $\nu_1(\{a^\star\})\ge0$) and ν_1 and μ_1 have the same total mass m_0. Each of these measures is supported on an interval and these intervals are (almost) disjoint. Strict concavity of the cost function entails that any optimal coupling between μ_1 and ν_1 must be non-increasing (in a set-valued sense). Since there is only one such coupling, the coupling is unique.

By the preceding paragraph and the above claim, we know that π must be non-increasing from $[a_0,a_1]$ to $[a_1,a^\star]$, which determines π uniquely on that part. After this transport is carried out, we are left with the measures $\nu-\nu_1$ and $\mu-\mu_1$, where the latter is supported on one less interval, namely the interval $[a_0,a_1]$ disappears.

If instead $\mu_0([a_0,a_1])>\nu([a_1,a_2])$, we can use the same construction with

$$a^\star=\inf\{a:\mu([a_0,a]\ge\nu([a_1,a_2])\}\in[a_0,a_1],$$

and the interval $[a_1,a_2]$ will disappear. We then merge $[a^\star,a_1]$ with $[a_2,a_3]$, that is

$$\mu-\mu_1 \text{ is supported on } [a^\star,a_3]\cup[a_4,a_5]\cup\dots,$$
$$\nu-\nu_1 \text{ is supported on } [a_3,a_4]\cup[a_5,a_6]\cup\dots.$$

If $\mu([a_0,a_1])=\nu([a_1,a_2])$ then both the intervals $[a_0,a_1]$ and $[a_1,a_2]$ disappear when considering $\mu-\mu_1$ and $\nu-\nu_1$. In all three cases we can continue inductively and construct π in a unique way. Since there are finitely many intervals, the procedure is guaranteed to terminate. Thus π is unique. □

Lemma 6.4. Let I_k be a dual feasible basis inducing a dual solution (α,β). Every such basis induces a graph $G(I_k)$ in the sense that if $(i,j)\in I_k$ then the i-th support point of the measure r is connected to the j-th support point of measure s, i.e., $(i,j)\in G(I_k)$. By definition of dual feasible basis it holds that $\alpha_i+\beta_j=c_{ij}$. In fact, such a basis induces a tree structure between all support points of r and all support points of s (Peyré & Cuturi, 2019, Section 3.4).

In order to exclude that $\lambda(I_k)=\lambda(I_l)$ for $k\neq l$ we proceed as follows. Since $G(I_k)\neq G(I_l)$, there exists at least one edge (i,j) in $G(I_l)\setminus G(I_k)$. By definition if $(\widetilde{\alpha},\widetilde{\beta})$ is the feasible dual solutions induced by I_l, then $\widetilde{\alpha}_i + \widetilde{\beta}_j = c_{ij}$. To conclude $\lambda(I_k) = \lambda(I_l)$ it suffices to prove that $\alpha_i + \beta_j \neq c_{ij}$. To see this, notice that adding edge (i,j) to $G(I_k)$ creates a cycle. In particular, after proper relabelling there exists a path of the form

$$(i = i_1, j_1, i_2, j_2, \dots, i_n, j_n = j)$$

such that $(i_l, j_l) \in G(I_k)$ as well as $(i_{l+1}, j_l) \in G(I_k)$ for all $1 \leq l \leq n-1$. Recall further that by definition if edge $(i_k, j_k) \in G(I_k)$ then $\alpha_{i_k} + \beta_{j_k} = c_{i_k, j_k}$. By the summability assumtion (6.6) it follows

$$0 \neq \sum_{k=1}^{n} c_{i_k j_k} - \sum_{k=1}^{n} c_{i_k j_{k-1}} = \sum_{k=1}^{n} (\alpha_{i_k} + \beta_{j_k}) - \sum_{k=2}^{n} (\alpha_{i_k} + \beta_{j_{k-1}}) - c_{ij} = \alpha_i + \beta_j - c_{ij},$$

that gives $\alpha_i + \beta_j \neq c_{ij}$. □

Proposition 6.6. According to Lemma 6.4 all dual feasible basic solutions for (DOT) are non-degenerate if there exists no family of indices $\{(i_k, j_k)\}$ for $n \geq 2$ with all i_k pairwise different and all j_k pairwise different such that

$$\sum_{k=1}^{n} \|X_{i_k} - Y_{j_k}\|_q^p = \sum_{k=1}^{n} \|X_{i_k} - Y_{j_{k-1}}\|_q^p, \quad Y_{j_0} \coloneqq Y_{j_n}. \tag{B.1}$$

It suffices to prove that (B.1) holds with probability zero for fixed n. For the sake of notational simplicity, we choose the first $n \leq N$ random locations $(\mathbf{X}, \mathbf{Y}) \coloneqq (X_1, \dots, X_n, Y_1, \dots, Y_n)$. We denote by $(\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_n, y_1, \dots, y_n) \in \left(\mathbb{R}^D\right)^{2n}$ and define the set

$$A \coloneqq \left\{ (\mathbf{x}, \mathbf{y}) \in \left(\mathbb{R}^D\right)^{2n} \mid \sum_{k=1}^{n} \|x_k - y_{k-1}\|_q^p - \|x_k - y_k\|_q^p = 0 \right\}.$$

We need to show that $\mathbb{P}((\mathbf{X}, \mathbf{Y}) \in A) = 0$. Set $e_i \in \mathbb{R}^D$ to be the ith unit vector and consider the closed set

$$B \coloneqq \bigcup_{i \in [D], k \in [n]} \left\{ (\mathbf{x}, \mathbf{y}) \in \left(\mathbb{R}^D\right)^{2n} \mid \langle x_k, e_i \rangle \in \{\langle y_{k-1}, e_i \rangle, \langle y_k, e_i \rangle\},\ y_0 \coloneqq y_n \right\}.$$

Define the function $f \colon (\mathbb{R}^D)^{2n} \setminus B \to \mathbb{R}$ with $f(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} \|x_k - y_{k-1}\|_q^p - \|x_k - y_k\|_q^p$. We can rewrite

$$\mathbb{P}((\mathbf{X}, \mathbf{Y}) \in A) \leq \mathbb{P}\left((\mathbf{X}, \mathbf{Y}) \in f^{-1}(0)\right) + \mathbb{P}((\mathbf{X}, \mathbf{Y}) \in B).$$

The second term on the right-hand side is zero since by independence and absolute continuity the high-dimensional vector $(\mathbf{X}, \mathbf{Y})$ has a Lebesgue density and the set B lives in dimension less than $2Dn$. It remains to discuss $\mathbb{P}\left((\mathbf{X}, \mathbf{Y}) \in f^{-1}(0)\right)$. The open set $(\mathbb{R}^D)^{2n} \setminus B$ on which f is defined can be partitioned into finitely many (less than 6^{nD}) open connected components $U_1, \dots, U_L$ according to the signs of $\langle x_k - y_k, e_i \rangle$ and $\langle x_k - y_{k-1}, e_i \rangle$. On each such component $f_{|U_i}$ is analytic. It follows that $\mathbb{P}\left((\mathbf{X}, \mathbf{Y}) \in f_{|U_l}^{-1}(\{0\})\right) = 0$ unless $f_{|U_l}$ is identically zero (Dang, 2015, Lemma 1.2). To exclude the latter possibility, consider for any point $(\mathbf{x}, \mathbf{y}) \in U_l$ and $\epsilon \in \mathbb{R}$ the function

$$f_{|U_l}(\epsilon) = f_{|U_l}(x_1 + \epsilon e_i, x_2, \dots, x_n, y_1, \dots, y_n)$$

with derivative at $\epsilon = 0$ given by

$$\frac{\partial f_{|U_l}}{\partial \epsilon}_{|\epsilon=0} = p\left(\|x_1 - y_1\|_q^{p-q} \frac{|x_{1_i} - y_{1_i}|^q}{(x_{1_i} - y_{1_i})} - \|x_1 - y_n\|_q^{p-q} \frac{|x_{1_i} - y_{n_i}|^q}{(x_{1_i} - y_{n_i})} \right), \tag{B.2}$$

where x_{i_j} denotes the jth entry of the ith vector. If this derivative is nonzero, then clearly f is not identically zero. If the derivative is zero then we shall show that there exists another point in U_l for which this derivative is nonzero. Since U_l is open, we can add δe_j to y_n for small δ and any $1 \leq j \leq D$. If $p \neq q$ then, taking $j \neq i$ (which is possible because $D \geq 2$) only modifies the term $\|x_1 - y_n\|$ in (B.2), and for small δ the derivative will not be zero. If $p = q \neq 1$ then the norms do not appear in (B.2) and taking $j = i$ would yield a nonzero derivative. Hence, if p and q are not both equal to one, f is not identically zero on each piece U_l, which is what we needed to prove. A similar idea works in case $q = \infty$ and $p \neq 1$.

The argument only depends on the positions of the random support points of the probability measures $r = \sum_{k=1}^{n} r_k \delta_{X_k}$ and $s = \sum_{k=1}^{n} s_k \delta_{Y_k}$ and hence is uniform in their probability weights. Recall further Proposition 2.4 that if the dual problem admits a non-degenerate optimal solution the primal optimal solution is unique. We conclude that almost surely the optimal transport coupling is unique. □

CHAPTER B

Empirical Regularized Optimal Transport: Statistical Theory and Applications

Empirical Regularized Optimal Transport: Statistical Theory and Applications

Marcel Klatt * Carla Tameling * Axel Munk *†

Abstract

We derive limit distributions for various empirical regularized optimal transport quantities between probability distributions supported on a finite metric space and show their bootstrap consistency. In particular, we prove that the empirical regularized transport plan itself asymptotically follows a Gaussian law. The theory includes the Boltzmann–Shannon entropy regularization and hence a limit law for the widely applied Sinkhorn divergence. Our approach is based on parametric optimization techniques for the regularized transport problem in conjunction with a statistical delta method. The asymptotic results are investigated in Monte Carlo simulations. We further discuss computational consequences and statistical applications, e.g., confidence bands for colocalization analysis of protein interaction networks based on regularized optimal transport.

Keywords Bootstrap, limit law, protein networks, regularized optimal transport, sensitivity analysis, Sinkhorn divergence

MSC 2010 subject classification Primary: 62E20, 62G20, 65C60 Secondary: 90C25, 90C31, 90C59

1 Introduction

The theory of optimal transport (OT) has a long history in physics, mathematics, economics and related areas (Galichon, 2016; Kantorovich, 1942; Monge, 1781; Rachev & Rüschendorf, 1998; Villani, 2008). Recently, also OT based *data analysis* has become popular in many areas of application, among others, in computer science (Balikas et al., 2018; Schmitz et al., 2018), mathematical imaging (Adler et al., 2017; Ferradans et al., 2014; Rubner et al., 2000), machine learning (Arjovsky et al., 2017; Lu et al., 2017; Sommerfeld et al., 2019) and statistics (Chernozhukov et al., 2017; del Barrio et al., 1999; Evans & Matsen, 2012; Munk & Czado, 1998; Panaretos & Zemel, 2018; Sommerfeld & Munk, 2018).

This work focuses on certain statistical aspects of *regularized* optimal transport (ROT) based data analysis. Throughout the following, let the ground space $\mathcal{X} = \{x_1, \dots, x_N\}$ be finite (e.g. representing spatial locations) and equipped with a metric $d\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, where $\mathbb{R}_{\geq 0}$ denotes the non-negative and $\mathbb{R}_{>0}$ ($\mathbb{R}_{<0}$) the positive (negative) reals. Each probability distribution on $\mathcal{X}$ is represented as an element in Δ_N the N-dimensional simplex of vectors $r \in \mathbb{R}^N$ such that $\sum_{i=1}^N r_i = 1$, $r_i \geq 0$. For the sake of exposition, we implicitly assume that $r, s \in \Delta_N$ share the same support, in particular, are of the same length. Moreover, for $p \geq 1$ the cost to transport one unit from x_i to x_j is represented in

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

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

a vector $c_p \in \mathbb{R}^{N^2}$ with entries $c_{p(i-1)N+j} := d^p(x_i, x_j)$ defined by the underlying metric d. These assumptions can be easily generalized (see Remark 2.5). Determining the OT between the probability distributions $r, s \in \Delta_N$ on $\mathcal{X}$ then amounts to solve the standard linear program

$$\begin{aligned} \min_{\pi \in \mathbb{R}^{N^2}} \quad & \langle c_p, \pi \rangle \\ \text{subject to} \quad & A\pi = \begin{bmatrix} r \\ s \end{bmatrix}, \pi \geq 0\,. \end{aligned} \tag{1.1}$$

The coefficient matrix $A := \begin{bmatrix} I_{N\times N} \otimes \mathbb{1}_{1\times N} \\ \mathbb{1}_{1\times N} \otimes I_{N\times N} \end{bmatrix} \in \mathbb{R}^{2N\times N^2}$, where $\otimes$ denotes the Kronecker product, encodes the marginal constraints for any transport plan $\pi \in \mathbb{R}^{N^2}$ so that considered as a $N \times N$ matrix its row and column sums are equal to r and s, respectively. Besides its non-negativity constraint, the linear program (1.1) consists of $2N$ linear equality constraints in N^2 unknowns. An optimal solution of (1.1) denoted as $\pi_p(r, s)$ (not necessarily unique) is known as an *optimal transport plan*. The quantity $W_p(r, s) := \langle c_p, \pi_p(r, s) \rangle^{1/p}$ is referred to as OT distance, Kantorovich distance (Kantorovich, 1942; Vershik, 2013), earth mover's distance (Levina & Bickel, 2001), or p-th Wasserstein distance (Villani, 2008) between the probability distributions r and s.

Despite its conceptual appeal and practical success in various applications, the routine use of OT based data analysis is still lacking as many real world applications suffer from the computational burden to solve (1.1). The zoo of OT solvers is diverse and classical examples include the Hungarian method (Kuhn, 1955), the auction algorithm (Bertsekas & Castanon, 1989) or the transportation simplex (Luenberger et al., 1984). However, their respective (average) runtime is still too demanding. For example, the auction algorithm is known to require $\mathcal{O}(N^3 \log(N))$ elementary operations on average, which delimits its use already for problems of intermediate size. There has been made certain progress to overcome this numerical obstacle. For instance, exploiting properties of the l_1-distance as a specific instance for d, on a regular grid the OT problem (1.1) can be stated in only $\mathcal{O}(N)$ unknowns as it suffices to consider transport between neighbouring grid points. This results in an algorithm with average time complexity $\mathcal{O}(N^2)$ (Ling & Okada, 2007). For more general distances Gottschlich & Schuhmacher (2014) have introduced the shortlist method, which has been shown empirically to have a runtime of magnitude $\mathcal{O}(N^{2.5})$. Furthermore, multiscale approaches (Gerber & Maggioni, 2017), subsampling techniques (Sommerfeld et al., 2019) and sparse approximate methods have recently been developed, including Schmitzer (2016) who computes OT via a sequence of sparse problems. However, solving already moderately sized problems in reasonable time is still a challenging issue, e.g., in two or three-dimensional imaging, where N is of magnitude $\sim 10^6$-10^8 (Schrieber et al., 2017), and large scale problems, such as temporal-spatial image or network analysis seem currently out of reach. This encouraged the development of various surrogates for the OT distance which are computationally better accessible. We mention thresholding of the full distance leading to graph sparsification (Pele & Werman, 2009), relaxation (Ferradans et al., 2014) and regularized OT distances (Dessein et al., 2018; Essid & Solomon, 2018; Lorenz et al., 2021). Among the most prominent proposals for the latter approach is the entropy regularization of OT (Cuturi, 2013; Peyré et al., 2019). Instead of solving the linear program (1.1), the entropy regularization approach asks to solve for a positive regularization parameter $\lambda > 0$ the *regularized* OT (ROT) problem

$$\begin{aligned} \min_{\pi \in \mathbb{R}^{N^2}} \quad & \langle c_p, \pi \rangle + \lambda f(\pi) \\ \text{subject to} \quad & A\pi = \begin{bmatrix} r \\ s \end{bmatrix} . \end{aligned} \tag{1.2}$$

The function $f \colon \mathbb{R}^{N^2}_{\geq 0} \to \mathbb{R}$ is the negative Boltzmann-Shannon entropy

$$f(\pi) \coloneqq \sum_{i=1}^{N^2} \pi_i \log(\pi_i) - \pi_i + 1 \tag{1.3}$$

with the convention $0 \log(0) = 0$. The ROT problem (1.2) is a strictly convex optimization program and hence has a *unique* optimal solution $\pi_{p,\lambda,f}(r,s)$ denoted as *entropy regularized optimal transport plan*. Moreover and different to (1.1), the regularization in (1.2) avoids the non-negativity constraint $\pi \geq 0$ as entropy regularization already enforces a *dense* structure for the ROT plan, i.e., all its entries are positive. This property, long known and favoured, for instance, in traffic prediction (Wilson, 1969) is key for our sensitivity analysis for ROT (Theorem 2.3) that might be of independent interest. The entropy ROT distance, also known as p-th Sinkhorn divergence (Cuturi, 2013), is then defined as

$$W_{p,\lambda,f}(r,s) \coloneqq \langle c_p, \pi_{p,\lambda,f}(r,s) \rangle^{1/p} . \tag{1.4}$$

The major benefit of entropy regularization is of algorithmic nature. The entropy ROT plan is approximated by the Sinkhorn-Knopp algorithm initially introduced by Sinkhorn (1964) that has a linear convergence rate and only requires $\mathcal{O}(N^2)$ operations in each step (Altschuler et al., 2017; Cuturi, 2013). It has been argued that as the regularization parameter $\lambda > 0$ decreases to zero in (1.2) this approximates a solution of the initial OT problem and hence serves as a good proxy. Nevertheless, high accuracy is computationally hindered in small regularization regimes (Benamou et al., 2015; Schmitzer, 2019) as the runtime of these algorithms scale with λ^{-2} (Altschuler et al., 2017; Dvurechensky et al., 2018).

Among others, the results of this paper complement these *computational* findings for ROT. We will show a substantially different *statistical* behaviour of the regularized ($\lambda > 0$) compared to non-regularized OT ($\lambda = 0$) when the probability distributions r and s are estimated from data, hence randomly perturbed. To this end and as often typical in applications, the underlying population distribution r (resp. s or both) is estimated from given data by its empirical version

$$\hat{r}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \left(\hat{s}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{Y_i} \right) \tag{1.5}$$

derived by a sample of $\mathcal{X}$-valued random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} r$ ($Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} s$). We investigate the fluctuation of the estimated ROT plan $\pi_{p,\lambda,f}(\hat{r}_n, s)$ and its distance $W_{p,\lambda,f}(\hat{r}_n, s)$ around their population version $\pi_{p,\lambda,f}(r,s)$ and $W_{p,\lambda,f}(r,s)$, respectively. More precisely, we prove (with proper $\sqrt{n}$-standardization) central limit theorems as the sample size n approaches infinity. Our main result (Theorem 3.1) concerns the empirical ROT plan itself. Denoting by $\xrightarrow{D}$ convergence in distribution, we find that for $\lambda > 0$ and sample size n approaching infinity

$$\sqrt{n} \left\{ \pi_{p,\lambda,f}(\hat{r}_n, s) - \pi_{p,\lambda,f}(r,s) \right\} \xrightarrow{D} \mathcal{N}_{N^2} \left(0, \Sigma_{p,\lambda,f}(r|s) \right) , \tag{1.6}$$

that is the limit in distribution is given by a N^2-dimensional Gaussian law. The covariance $\Sigma_{p,\lambda,f}(r|s)$ (see (3.1) for details) depends on λ, the Hessian of the regularization function f and the probability distributions r and s. From this, central limit theorems for the ROT distance are obtained (Theorem 3.2) whose empirical version properly centred and scaled is approximated for $n \to \infty$ by a centred Gaussian law

$$\sqrt{n}\,\{W_{p,\lambda,f}(\hat{r}_n, s) - W_{p,\lambda,f}(r,s)\} \xrightarrow{D} \mathcal{N}_1\left(0, \sigma^2_{p,\lambda,f}(r|s)\right). \tag{1.7}$$

The results hold true for $r = s$ and $r \neq s$ and are valid for a broad class of regularizers f in (1.2). These will be denoted as *proper regularizers* (Definition 2.2) and subsume various regularization methods of OT (Dessein et al., 2018) including the widely applied entropy ROT in (1.3). In particular, our theory covers limit laws for the empirical entropy ROT plan and consequentially for the empirical Sinkhorn divergence.

Parallel to our work, central limit theorems for the optimal *value* of entropy regularization in (1.2) and variations of that, e.g., the Sinkhorn loss (Genevay et al., 2018), are derived in Bigot et al. (2019) with a technique introduced in Sommerfeld & Munk (2018) for (non-regularized) OT distances. More recently, Mena & Niles-Weed (2019) analysed entropy ROT distances for general sub-Gaussian probability distributions on $\mathbb{R}^d$. They obtained a central limit theorem for a continuous counterpart of the optimal value in (1.2). In contrast to finite spaces as treated here, the centring constant for the sample law in Mena & Niles-Weed (2019) is given by an unknown empirical constant which depends on the underlying measure and makes it difficult for immediate use for data analysis. Notably, their technique of proof is completely different to ours and based on the Efron-Stein inequality that already turned out to be useful for central limit theorems for non-regularized OT distances for absolutely continuous distributions (Del Barrio et al., 2019). In contrast to Bigot et al. (2019); Mena & Niles-Weed (2019), our work demonstrates that proper regularization for OT on finite spaces allows for a general statistical analysis of the corresponding ROT *plan* and not only for an optimal *value*. This paves the way to a statistical analysis for general functionals of the ROT plan including the results by Bigot et al. (2019) as a specific instance (see Remark 3.4). We argue in Section 7 that compared to any single value OT based distance (*regularized* or not), the ROT *plan* encodes more structural information across scales and hence can serve as a more informative tool for spatial data analysis. This will provide the foundation for our colocalization analysis of protein networks (see Section 7.1.2).

Our findings for the empirical ROT distance highlight a substantial difference to related limit laws for *non-regularized* transport (Sommerfeld & Munk, 2018; Tameling et al., 2019). More specifically, for $r = s$ the empirical (non-regularized) transport distance $W_p(\hat{r}_n, r)$ does not follow asymptotically a Gaussian law, in contrast to (1.7) and for $p \geq 1$ it holds that

$$n^{1/2p}\, W_p(\hat{r}_n, r) \xrightarrow{D} \left\{ \max_{u \in \Phi^*} \langle G, u \rangle \right\}^{1/p}, \tag{1.8}$$

as $n \to \infty$ (Sommerfeld & Munk, 2018). Here, Φ^* is the set of optimal dual solutions for (1.1) and G is a centred N-dimensional Gaussian random vector with covariance matrix

$$\Sigma(r) := \begin{pmatrix} r_1(1-r_1) & -r_1 r_2 & \dots & -r_1 r_N \\ -r_2 r_1 & r_2(1-r_2) & \dots & -r_2 r_N \\ \vdots & \vdots & \ddots & \vdots \\ -r_N r_1 & -r_N r_2 & \dots & r_N(1-r_N) \end{pmatrix}. \tag{1.9}$$

This different limit behaviour provides some insight into the above mentioned computational difficulties to approximate the non-regularized OT distance by the Sinkhorn divergence as λ tends to zero (see Schmitzer (2019) for a discussion). In contrast to Sommerfeld & Munk (2018); Tameling et al. (2019), our approach relies on a sensitivity analysis which builts on parametric optimization techniques for the *non*-linear optimization problem (1.2) and, in particular, is based on the linear independence constraint qualification (LICQ) in combination with an application of the implicit function theorem (IFT). Let us finally stress that the limit distribution for the ROT plan in (1.7) highlights another major difference to the non-regularized OT plan. Remarkably, such a result is entirely unknown for the non-regularized OT plan, to the best of our knowledge.

The outline of this paper is as follows. As a prerequisite for our main methodology and results we provide in Section 2 all necessary terminology from convex optimization. We derive sensitivity results for ROT plans which might be of interest by itself as they describe the stability of ROT plans when perturbing the boundary conditions given by r and s. Parallel to our work, such a sensitivity result was partially obtained by Luise et al. (2018). However, their proof is carried out on the dual formulation of (1.2) and is limited to entropy regularization. Section 3 is dedicated to distributional limit results stated in Theorem 3.1 and Theorem 3.2. Moreover, applying recent results by Weed (2018), we obtain rates for the regularizer $\lambda(n)$ tending to zero and depending on the sample size in order to recover asymptotically the limit laws for (non-regularized) OT (see Section 3.2). Our proof for (1.6) and (1.7) is based on a statistical delta method, and as a byproduct we obtain consistency of the n out of n bootstrap in Section 4. This is again in notable contrast to the (non-regularized) empirical OT distance, where the n out of n bootstrap is known to fail (Sommerfeld & Munk, 2018). In Section 5 we investigate in a Monte Carlo study the approximation of the empirical Sinkhorn divergence sample distribution by its theoretical limit law. In addition, we analyse the influence of the amount of regularization λ to the accuracy of the normal approximation, compare our results to asymptotic distributions for the non-regularized OT distance and investigate the bootstrap empirically. Finally, in Section 7 we apply our methodology to the statistical analysis of colocalization for protein interaction networks in cells. To this end, we provide confidence bands (Theorem 7.1) for a measure of protein proximity based on the estimated ROT plan between two such protein distributions. Depending on a resampling method to reduce computational complexity, we propose a protocol to analyse data sets that are beyond the scope of computational feasibility of current state of the art solvers.

2 Sensitivity Analysis for Regularized Transport Plans

As a prerequisite for our statistical analysis of empirical ROT quantities, we provide in this section a *sensitivity analysis* of (1.2). More precisely, we investigate how the corresponding optimal solution, i.e., the ROT plan

$$\pi_{p,\lambda,f}(r,s) = \operatorname*{arg\,min}_{\pi\in\Pi(r,s)} \langle c_p, \pi\rangle + \lambda f(\pi) \tag{2.1}$$

changes when perturbing the marginal constraints given by $r, s \in \Delta_N$, where the set of feasible solutions is denoted as $\Pi(r,s)$ (see also Remark 2.1 below). Within the context of *parametric optimization*, which accounts for (possibly) large deviations of certain parameters, our focus is on small (local) variations usually termed as sensitivity analysis (see Bonnans & Shapiro (2013); Gal & Greenberg (2012) and references therein). Our approach is based on replacing the optimization problem by equations characterizing necessary and sufficient optimality conditions for the ROT plan. An application of the implicit function theorem (IFT) then yields the desired result. For a general overview on such an approach we refer to Fiacco (1984) and the references therein. We start with an easy but important observation.

Remark 2.1. *The ROT formulation in (1.2) and (2.1) can be stated in terms of only $2N-1$ equality constraints instead of $2N$. In fact, since the total supply equals the total demand ($r, s \in \Delta_N$) any one of the equality constraints is redundant. In the sequel, we define for $r, s \in \Delta_N$ the feasible set for ROT as*

$$\Pi(r,s) \coloneqq \left\{\pi \in \mathbb{R}^{N^2}_{\geq 0} \mid A_\star \pi = [r, s_\star]^T\right\}$$

with coefficient matrix $A_\star$ and vector $s_\star$. The subscript star denotes the deletion of the last row of the matrix A and the last entry of the vector $s \in \Delta_N$, respectively. This reduction allows for linearly independent constraints described by $A_\star$ or equivalently full rank of $A_\star^T$ (Luenberger et al., 1984) but requires some caution as we treat r and s asymmetrically.

In our analysis we consider general regularizers f in (2.1) that are of Legendre type (Rockafellar, 1970), which means that f is a closed proper convex function on $\mathbb{R}^{N^2}$ that is essentially smooth and strictly convex on the interior of its domain (Dessein et al., 2018). Recall that a function f is essentially smooth if it is differentiable on the interior of its domain and for every sequence $(x_k)_{k\in\mathbb{N}} \subset \text{int}(\text{dom } f)$ converging to a boundary point of $\text{int}(\text{dom } f)$ it holds that $\lim_{k\to\infty} \|\nabla f(x_k)\| = +\infty$.

Definition 2.2 (Proper regularizer)**.** *Let $f\colon \mathbb{R}^{N^2} \to \mathbb{R}\cup\{+\infty\}$ be twice continuously differentiable on the interior of its domain with positive definite Hessian $\nabla^2 f$. Moreover, assume for f and its Fenchel conjugate f^* that*

1. *f is of Legendre type,*
2. *$\mathbb{R}^{N^2}_{<0} \subset$ dom f^*,*
3. *$(0,1)^{N^2} \subseteq$ dom f,*
4. *dom $f \subseteq \mathbb{R}^{N^2}_{\geq 0}$.*

Then f is said to be a proper regularizer.

Examples for proper regularizers are discussed more carefully in the next subsection. Notice that by strict convexity, for each proper regularizer f, regularization parameter $\lambda > 0$ and $p \geq 1$ the ROT plan in (2.1) is *unique.*

Theorem 2.3 (Sensitivity for the ROT plan)**.** *For f a proper regularizer, $r_0, s_0 \in \Delta_N$ and $\lambda > 0$ consider the ROT plan (2.1). Then there exists a neighbourhood $\mathcal{U} \subseteq \Delta_N \times (\Delta_N)_\star$ of $(r_0, s_{0\star})$ and a unique continuously differentiable function $\phi_{p,\lambda,f}\colon \mathcal{U} \to \mathbb{R}^{N^2}$ such that $\phi_{p,\lambda,f}(r_0, s_{0\star}) = \pi_{p,\lambda,f}(r_0, s_0)$. Moreover, the ROT plan is parametrized by $\phi_{p,\lambda,f}$ for all $(r, s_\star) \in \mathcal{U}$ with derivative at $(r_0, s_{0\star})$ given by*

$$\nabla\phi_{p,\lambda,f}(r_0, s_{0\star}) = \left[\nabla^2 f(\pi_{p,\lambda,f}(r_0,s_0))\right]^{-1} A_\star^T \left[A_\star \left[\nabla^2 f(\pi_{p,\lambda,f}(r_0,s_0))\right]^{-1} A_\star^T\right]^{-1},$$

a matrix of dimension $N^2 \times (2N-1)$.

Proof. The proof relies on the basic sensitivity theorem (Fiacco, 1984, Thm. 3.2.2) considering optimization problems of the form

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x,\epsilon) \\ \text{s.t.}\ \ &g_i(x,\epsilon) \geq 0,\ i = 1,\dots,m, \\ &h_j(x,\epsilon) = 0,\ j = 1,\dots,p \end{aligned} \tag{2.2}$$

with parameter $\epsilon \in \mathbb{R}^k$ considered to be perturbed and associated Lagrangian for $\mu \in \mathbb{R}^p$ and $\nu \in \mathbb{R}^m$ given by $\mathcal{L}(x, \mu, \nu, \epsilon) = f(x, \epsilon) - \sum_{j=1}^{p} \mu_j h_j(x, \epsilon) + \sum_{i=1}^{m} \nu_i g_i(x, \epsilon)$. Notice that the ROT problem (1.2) can be restated as

$$\begin{aligned} &\min_{\pi \in \mathbb{R}^{N^2}} \langle c_p, \pi \rangle + \lambda f(\pi) \\ &\text{s.t. } \ h(\pi, \epsilon) = \mathbf{0}\,, \end{aligned} \tag{2.3}$$

where $\epsilon = (r_0, s_{0\star}) \in \mathbb{R}^{2N-1}$, $\mathbf{0} \in \mathbb{R}^{2N-1}$ denotes the zero vector and the linear function $h(\pi, \epsilon) = A_\star \pi - \epsilon^T$ encodes the marginal conditions. In particular, we do not need to consider inequality constraints as non-negativity is already enforced by the proper regularizer in the objective function. Hence, for $\mu \in \mathbb{R}^{2N-1}$ the corresponding Lagrangian simplifies to

$$\mathcal{L}(\pi, \mu, \epsilon) = \langle c_p, \pi \rangle + \lambda f(\pi) - h(\pi, \epsilon)^T \mu\,.$$

In order to obtain sensitivity statements for the optimal solution of (2.3), i.e., the ROT plan $\pi_{p,\lambda,f}$, when perturbing ϵ, it remains to check the assumptions in Fiacco (1984, Thm. 3.2.2) in case of convex optimization problems with linear equality constraints. For the sake of readability these assumptions are recalled for the general problem (2.2) with linear equality constraints only and then checked for the reformulation (2.3) of the ROT.

(i) *Linear independence constraint qualification (LICQ) (Fiacco, 1984, p.24 (CQ3)): The LICQ condition is fulfilled at the optimal solution x for (2.2) if $\nabla_x h_j(x, \epsilon)$ for all j are linearly independent.*
Notice that $\nabla_\pi h(\pi, \epsilon) = A_\star$ and recall Remark 2.1 that by full rank of $A_\star$ implies LICQ to hold at any feasible solution, in particular, at the ROT plan $\pi_{p,\lambda,f}$.

(ii) *Second order sufficiency condition (Fiacco, 1984, Lem. 3.2.1): If the functions defining problem (2.2) are twice continuously differentiable in a neighbourhood of the optimal solution x, then the second order sufficiency conditions is satisfied at x if there exists a vector $\mu \in \mathbb{R}^p$ such that the first-order Karush-Kuhn-Tucker conditions hold, i.e.,*

$$\begin{aligned} \nabla_x \mathcal{L}(x, \mu, \epsilon) &= 0 \\ h_j(x, \epsilon) &= 0,\ j = 1, \ldots, p \end{aligned}$$

and further, if $z^T \nabla_x^2 \mathcal{L}(x, \mu, \epsilon) z > 0$ for all vectors $z \neq 0$ such that $\nabla h_j(x, \epsilon) z = 0$ for $j = 1, \ldots, p$.
By definition of a proper regularizer the unique ROT plan $\pi_{p,\lambda,f}$ for $\epsilon = (r_0, s_{0\star})$ is contained in the positive orthant $\mathbb{R}^{N^2}_{>0}$ (see Dessein et al. (2018) for details). The objective and the constraint function h defining problem (2.3) are at least twice continuously differentiable at $(\pi_{p,\lambda,f}, \epsilon)$. Moreover, Slater's qualification constraint (Fiacco, 1984, p.24 (CQ2)), i.e., the function h is (affine) linear and the ROT problem is feasible, is fulfilled. For convex optimization problems with linear equality constraints this implies the existence of a vector μ such that $\pi_{p,\lambda,f}$ fulfils the first-order Karush-Kuhn-Tucker conditions

$$\begin{aligned} \nabla_\pi \mathcal{L}\left(\pi_{p,\lambda,f}, \mu, \epsilon\right) &= 0\,, \\ h\left(\pi_{p,\lambda,f}, \epsilon\right) &= 0\,. \end{aligned}$$

Further, notice that for all vectors $z \neq 0$ it holds that $z^T \nabla_\pi^2 \mathcal{L}(\pi, \mu, \epsilon) z = \lambda z^T \nabla_\pi^2 f z > 0$ as by assumption on the regularizer $\nabla^2 f$ is positive definite and $\lambda > 0$. Hence, the second order sufficiency condition is satisfied for the ROT.

An application of Fiacco (1984, Thm. 3.2.2) based on the IFT to the first-order Karush-Kuhn-Tucker conditions yields the statement of our claim. More precisely, there exists a neighbourhood $\mathcal{U}$ around $(r_0, s_{0\star})$ and a continuously differentiable function $y\colon \mathcal{U} \to \mathbb{R}^{N^2} \times \mathbb{R}^{2N-1}$ with $y(r, s_\star) = (\phi_{p,\lambda,f}(r, s_\star), \gamma_{p,\lambda,f}(r, s_\star))$. In particular, the function $\phi_{p,\lambda,f}\colon \mathcal{U} \to \mathbb{R}^{N^2}$ parametrizes the ROT plan for all $(r, s_\star) \in \mathcal{U}$. It remains to compute the derivative of $\phi_{p,\lambda,f}$ at $\epsilon = (r_0, s_{0\star})$. It holds that

$$\nabla y(\epsilon) = \begin{bmatrix} \nabla \phi_{p,\lambda,f}(\epsilon) \\ \nabla \gamma_{p,\lambda,f}(\epsilon) \end{bmatrix} = \begin{bmatrix} \left[\nabla_\pi^2 \mathcal{L}(\pi_{p,\lambda,f}, \mu, \epsilon)\right]^{-1} A_\star^T \left[A_\star \left[\nabla_\pi^2 \mathcal{L}(\pi_{p,\lambda,f}, \mu, \epsilon)\right]^{-1} A_\star^T\right]^{-1} \\ -\left[A_\star \left[\nabla_\pi^2 \mathcal{L}(\pi_{p,\lambda,f}, \mu, \epsilon)\right]^{-1} A_\star^T\right]^{-1} \end{bmatrix},$$

where we applied Fiacco (1984, equalities (4.2.8)) in the second equality. Notice again that $\nabla_\pi^2 \mathcal{L} = \lambda \nabla_\pi^2 f$ which finishes the proof. □

Remark 2.4 (Sensitivity for non-regularized OT plans)**.** *Besides the uniqueness issue in (non-regularized) OT, a similar strategy for OT plans* ($\lambda = 0$) *does not work. Our approach relies on the LICQ assumption that for (non-regularized) OT requires additional knowledge about all binding constraints in the non-negativity constraints* $\pi \geq 0$ *for the OT plan. In particular, the LICQ assumption fails to hold under* degeneracy *in linear programming (Luenberger et al., 1984), i.e., if* $\pi_p(r, s)$ *contains more than* $2N - 1$ *zeroes, e.g., for* $r = s$*.*

Remark 2.5 (General cost, varying number of support points)**.** *In the course of this paper, we typically assume that the cost comes from the underlying metric of the ground space. However, Theorem 2.3 holds for any cost vector* c *and formally also the subsequent statistical analysis in Theorem 3.1 and Theorem 3.2. Nevertheless, in order to exclude degeneracy of the limit, e.g., if the cost is given by a constant, we have specified the theorems to* c *coming from a distance. Further, the result implicitly covers different number of support points of the probability distributions* r *and* s*, e.g., the probability distribution* r *does not need to have full support. In such a case, we simply delete the zero entries and our sensitivity result holds for the reduced problem.*

Remark 2.6 (Influence curve)**.** *The result from Theorem 2.3 allows for a general analysis to quantify the specific influence of a particular point and its mass to the ROT plan or the corresponding divergence in* (1.4)*. For the latter quantity denoted by* $\phi(r, s) \coloneqq W_{p,\lambda,f}^p(r, s)$*, we might in particular be interested in the rate of change that is caused along a perturbation in direction* $\delta_{x_i} - r$*. Here, we abuse notation and consider* δ_{x_i} *as the* i*-th canonical unit vector with entry equal to one at index* i*. More precisely, we can calculate the* influence curve *(see Rieder (2012); Van der Vaart (2000) for a concrete definition and applications in robust statistics) given by*

$$\begin{aligned} \phi'(\delta_{x_i} - r, s) &= \frac{d}{dt}\Big|_{t=0} W_{p,\lambda,f}^p((1-t)r + t\delta_{x_i}, s) = \frac{d}{dt}\Big|_{t=0} \langle c_p, \pi_{p,\lambda,f}(r - t(r - \delta_{x_i}), s)\rangle \\ &= c_p^T \left[\nabla^2 f(\pi_{p,\lambda,f}(r, s))\right]^{-1} A_\star^T \left[A_\star \left[\nabla^2 f(\pi_{p,\lambda,f}(r, s))\right]^{-1} A_\star^T\right]^{-1} (r - \delta_{x_i}, \mathbf{0})^T, \end{aligned}$$

where $\mathbf{0}$ *indicates the zero vector of dimension* $N - 1$*.*

Remark 2.7 (Limitation to finite spaces)**.** *Our method of proof of Theorem 2.3 does not extend to the countable nor to the continuous formulation of ROT. Already for the countable case existence and uniqueness of a ROT plan is not straightforward. For example, the optimal value of ROT* (1.2) *can easily be infinity if the marginals are not restricted* *to have finite p-th moment and finite entropy (Kovačević et al., 2015). Further, our approach in Tameling et al. (2019) for the treatment of OT distance on countable spaces is not applicable as entropy is not differentiable in this case.*

2.1 Proper Regularizers

The class of proper regularizers in Definition 2.2 is rather rich and some common ones are the negative Boltzmann-Shannon entropy (1.3) or l_p quasi norms ($0 < p < 1$) defined as $f(\pi) = \sum_{i=1}^{N^2} \pi_i^p$ for $\pi \in \mathbb{R}_{\geq 0}$, among others. For further examples that have also been the focus of recent research (Dessein et al., 2018) we refer to Appendix A.

Example 2.8 (Sinkhorn divergence and optimal value function)**.** *For f the negative Boltzmann-Shannon entropy, $p \geq 1$ and $\lambda > 0$ in (1.2), the gradient of the parametrization for the entropy ROT plan $\pi_\lambda(r,s)$ (we suppress additional indexing by p and f here) is given by*

$$\nabla_{(r,s)}\pi_\lambda(r,s) = D\, A_\star^T \left[A_\star\, D\, A_\star^T\right]^{-1} ,$$

where $D \in \mathbb{R}^{N^2\times N^2}$ is a diagonal matrix with diagonal given by the entropy ROT plan $\pi_\lambda(r,s)$. Further, let us denote by

$$S_\lambda(r,s) := \langle c_p, \pi_\lambda(r,s)\rangle + \lambda f(\pi_\lambda(r,s)) \tag{2.4}$$

the optimal value of the ROT problem (1.2). The gradient $\nabla_{(r,s)}S_\lambda$ has recently gained quite some popularity (see e.g. Bigot et al. (2019); Feydy et al. (2019); Luise et al. (2018)). Notice that $S_\lambda(r,s) = S_\lambda(\pi_\lambda(r,s))$ and hence by our result in Theorem 2.3 and an application of the multivariate chain rule we obtain $\nabla_{(r,s)}S_\lambda$ precisely. For this, we denote by

$$\left(\alpha_\lambda^{(r,s)}, \beta_\lambda^{(r,s)}\right) \in \underset{(\alpha,\beta)\in\mathbb{R}^N\times\mathbb{R}^{N-1}}{argmax} \langle r,\alpha\rangle + \langle s_\star,\beta\rangle - \lambda \sum_{i,j=1}^{N,N-1}\left(\exp\left(\frac{\alpha_i+\beta_j-c_{ij}}{\lambda}\right)-1\right) \tag{2.5}$$

optimal dual solutions for the corresponding dual problem (Peyré et al., 2019) of (1.2). In fact, the entropy ROT plan can be recovered from optimal dual solutions by

$$\pi_\lambda(r,s) = \exp\left(\frac{1}{\lambda}\left(\left(\alpha_\lambda^{(r,s)}, \beta_\lambda^{(r,s)}\right)^T A_\star - c_p\right)\right)$$

(Peyré et al., 2019, Prop. 4.4), where exp *(and* log *in the following) is meant to be coordinate-wise. Recall the lower subscript star as we delete the last constraint in (1.2) (Remark 2.1). Differentiation of the objective function (2.4) with respect to π and applying the previous display yields*

$$\nabla_\pi S_\lambda(\pi_\lambda(r,s)) = c_p + \lambda\log(\pi_\lambda(r,s)) = \left(\alpha_\lambda^{(r,s)}, \beta_\lambda^{(r,s)}\right)^T A_\star\,.$$

We conclude that

$$\begin{aligned}\nabla_{(r,s)}S_\lambda(r,s) &= \nabla_\pi S_\lambda(\pi_\lambda(r,s))\,\nabla_{(r,s)}\pi_\lambda(r,s)\\ &= \left(\alpha_\lambda^{(r,s)}, \beta_\lambda^{(r,s)}\right)^T A_\star D A_\star^T \left[A_\star D A_\star^T\right]^{-1} = \left(\alpha_\lambda^{(r,s)}, \beta_\lambda^{(r,s)}\right)^T .\end{aligned}$$

We would like to stress that not all common regularizers for OT are proper regularizers.

Example 2.9 (A counterexample)**.** *For f the l_p norm $1 \leq p < +\infty$, Theorem 2.3 does not hold. In fact, f is not a proper regularizer and allows for a sparse ROT plan (Blondel et al., 2018). In particular, we cannot guarantee that its corresponding ROT satisfies the linear independence constraint qualification (see discussion after Theorem 2.3). Hence, our proof strategy for l_p ($p \geq 1$) ROT plans fails.*

3 Distributional Limits

For two probability distributions $r, s \in \Delta_N$, parameter $\lambda > 0$, $p \geq 1$ and proper regularizer f an estimator for $\pi_{p,\lambda,f}(r,s)$ in (2.1) is given by its empirical counterpart $\pi_{p,\lambda,f}(\hat{r}_n, s)$ with $\hat{r}_n$ the empirical distribution of the i.i.d. sample $X_1, \ldots, X_n$ in (1.5). The next theorem states a Gaussian limit distribution for the empirical ROT plan. Since the sensitivity result in Theorem 2.3 holds regardless of $r = s$ or $r \neq s$ and as the ROT plan is always dense, we do not derive any substantially difference regarding statistical limit behaviour in either of these cases.

Theorem 3.1. *Let $r, s \in \Delta_N$ be two probability distributions on the finite metric space $(\mathcal{X}, d)$ and let $\hat{r}_n$ be the empirical version given in (1.5) derived by $X_1, \ldots, X_n \overset{i.i.d.}{\sim} r$. Then, as $n \to \infty$, it holds that*

$$\sqrt{n}\,\{\pi_{p,\lambda,f}(\hat{r}_n, s) - \pi_{p,\lambda,f}(r,s)\} \xrightarrow{D} \mathcal{N}_{N^2}\left(0, \Sigma_{p,\lambda,f}(r|s)\right)$$

with covariance matrix

$$\Sigma_{p,\lambda,f}(r|s) = \nabla_r\, \phi_{p,\lambda,f}(r, s_\star)\, \Sigma(r)\, \left[\nabla_r\, \phi_{p,\lambda,f}(r, s_\star)\right]^T , \tag{3.1}$$

where $\Sigma(r)$ is defined in (1.9) and $\nabla_r\, \phi_{p,\lambda,f}(r, s_\star)$ are the partial derivatives of $\phi_{p,\lambda,f}$ with respect to r as given in Theorem 2.3.

Proof. Since the vector $n\hat{r}_n$ follows a multinomial distribution with probability vector r, the multivariate central limit theorem yields $\sqrt{n}(\hat{r}_n - r) \xrightarrow{D} \mathcal{N}_N(0, \Sigma(r))$. The multivariate delta method in conjunction with the previous sensitivity analysis for regularized transport plans concludes the statement. More precisely, we obtain that

$$\begin{aligned}\sqrt{n}\,\{\pi_{\lambda,f}(\hat{r}_n, s) - \pi_{\lambda,f}(r,s)\} &= \sqrt{n}\,\{\phi_{p,\lambda,f}(\hat{r}_n, s_\star) - \phi_{p,\lambda,f}(r, s_\star)\} \\ &\xrightarrow{D} \mathcal{N}_{N^2}(0, \nabla_r\, \phi_{p,\lambda,f}(r, s_\star)\, \Sigma(r)\, \left[\nabla_r\, \phi_{p,\lambda,f}(r, s_\star)\right]^T) \,.\end{aligned}$$

□ □

Based on Theorem 3.1 it is easy to conclude limit distributions for the empirical counterpart of the ROT *distance* (1.4). Here, s (which might be equal to r) plays the role of a fixed reference probability distribution to be compared empirically with the probability distribution r. We again do not derive any substantially different distributional limit behaviour between the cases $r = s$ and $r \neq s$. This is in notably contrast to the non-regularized OT (see the discussion in Section 1 and Section 3.2).

Theorem 3.2. *Under the assumptions of Theorem 3.1, as $n \to \infty$, it holds that*

$$\sqrt{n}\,\{W_{p,\lambda,f}(\hat{r}_n, s) - W_{p,\lambda,f}(r,s)\} \xrightarrow{D} \mathcal{N}_1\left(0, \sigma^2_{p,\lambda,f}(r|s)\right) \tag{3.2}$$

with variance

$$\sigma^2_{p,\lambda,f}(r|s) = \gamma^T\, \Sigma_{p,\lambda,f}(r|s)\, \gamma \,, \tag{3.3}$$

where γ is the gradient of the function $\pi \mapsto \langle c_p, \pi\rangle^{1/p}$ evaluated at the regularized transport plan $\pi_{p,\lambda,f}(r,s)$, and $\Sigma_{p,\lambda,f}(r|s)$ is the covariance matrix from Theorem 3.1. Standardizing the l.h.s. in (3.2) by the square root of the empirical variance $\sigma^2_{p,\lambda,f}(\hat{r}_n|s)$ results in a standard normal limit distribution.

Proof. Let $G \sim \mathcal{N}_{N^2}(0, \Sigma_{\lambda,f}(r|s))$ be an N^2-dimensional Gaussian random vector according to the limit distribution in Theorem 3.1. The multivariate delta method yields

$$\sqrt{n}\left\{W_{p,\lambda,f}(\hat{r}_n, s) - W_{p,\lambda,f}(r, s)\right\} = \sqrt{n}\left\{\langle c, \pi_{p,\lambda,f}(\hat{r}_n, s)\rangle^{\frac{1}{p}} - \langle c, \pi_{p,\lambda,f}(r, s)\rangle^{\frac{1}{p}}\right\} \xrightarrow{D} \gamma^T G\,.$$

The variance of the random variable $\gamma^T G$ is $\sigma^2_{p,\lambda,f}(r|s) = \gamma^T \Sigma_{p,\lambda,f}(r|s)\,\gamma$. In particular, $\Sigma_{p,\lambda,f}(r|s)$ is continuous in r. Consequently, by the strong law of large numbers $\hat{r}_n \xrightarrow{a.s.} r$ and the continuous mapping theorem we have that $\sigma^2_{p,\lambda,f}(\hat{r}_n|s) \xrightarrow{a.s.} \sigma^2_{p,\lambda,f}(r|s)$ which together with Slutzky's theorem (Van der Vaart, 2000, Lemma 2.8) concludes the statement. □

As a corollary, we immediately obtain limit distributions for the empirical entropy ROT plan and the empirical Sinkhorn divergence.

Corollary 3.3 (Sinkhorn transport and Sinkhorn divergence). *Consider the negative Boltzmann-Shannon entropy f in (1.3). Then the statements in Theorem 3.1 and Theorem 3.2 remain valid. Note, that the gradient in the corresponding covariance matrix (3.1) is given by Example 2.8.*

Remark 3.4 (Smooth ROT functionals). *From Theorem 3.1 we easily derive asymptotic distributions for other sufficiently smooth functionals of the ROT plan. For instance, assume $\lambda > 0$ and $p \geq 1$ to be fixed and let f be the Boltzmann-Shannon entropy. For notational convenience we suppress the dependence on p and f here. Consider the corresponding ROT problem (1.2) and its optimal objective value $S_\lambda(r, s)$ in (2.4). In conjunction with the sensitivity analysis from Example 2.8 and an application of the statistical delta method we conclude for sample size $n \to \infty$ that*

$$\sqrt{n}\left\{S_\lambda(\hat{r}_n, s) - S_\lambda(r, s)\right\} \xrightarrow{D} \left\langle G, \alpha_\lambda^{(r,s)}\right\rangle\,,$$

where $G \sim \mathcal{N}_N\left(0, \Sigma(r)\right)$ and $\alpha_\lambda^{(r,s)}$ in (2.5). This result has been obtained independently in Bigot et al. (2019). An identical argument for the case $r \neq s$ can be applied to obtain limit laws for the Sinkhorn loss (Genevay et al., 2018) defined by

$$L_\lambda(r, s) \coloneqq S_\lambda(r, s) - \frac{1}{2}\left\{S_\lambda(r, r) - S_\lambda(s, s)\right\}\,. \tag{3.4}$$

It then follows for sample size $n \to \infty$ that

$$\sqrt{n}\left\{L_\lambda(\hat{r}_n, s) - L_\lambda(r, s)\right\} \xrightarrow{D} \left\langle G, \alpha_\lambda^{(r,s)} - \frac{1}{2}\left(\alpha_\lambda^{(r,r)} + \beta_\lambda^{(r,r)}\right)\right\rangle$$

with $\beta_\lambda^{(r,r)}$ augmented from $\mathbb{R}^{N-1}$ to $\mathbb{R}^N$ by a coordinate set to be zero (the augmentation is due to deletion of the last constraint as discussed in Remark 2.1). An interesting case arises for the Sinkhorn loss $L_\lambda(r, s)$ and $r = s \in \Delta_N$. As detailly explained by Bigot et al. (2019), it holds that $\nabla L_\lambda(r, r) = 0$. A first-order expansion yields that $\sqrt{n}L_\lambda(\hat{r}_n, r)$ converges to a point mass at zero, i.e., a degenerate limit law useless for statistical inference. However, a second-order expansion based on a perturbation analysis for the dual solutions provides a non degenerate asymptotic limit of $nL_\lambda(\hat{r}_n, r)$. This can be represented as a weighted sum of independent χ_1^2 random variables. The weights of this sum are then given by the eigenvalues of the Hessian $\nabla^2 L_\lambda(r, r)$ (Bigot et al., 2019, Thm. 2.8).

3.1 Estimating Both Probability Distributions

Theorem 3.1 and Theorem 3.2 also hold true if we estimate both distributions r and s by their empirical versions $\hat{r}_n$ and $\hat{s}_m$ in (1.5) derived by two collections of $\mathcal{X}$-valued random variables $X_1, \dots, X_n \overset{i.i.d.}{\sim} r$ and independently $Y_1, \dots, Y_m \overset{i.i.d.}{\sim} s$, respectively. Note that we treat r and s asymmetrically (see Remark 2.1). Hence, the underlying multinomial process is based on the reduced multinomial vector $(r, s_\star)$ rather than (r, s). The scaling rate is given by $\sqrt{{nm}/{n+m}}$ such that $n \wedge m \to +\infty$ and ${m}/{n+m} \to \delta \in (0,1)$. For instance, the limit distribution for the empirical ROT plan with parameters $\lambda > 0$, $p \geq 1$ and proper regularizer f then reads as

$$\sqrt{\frac{nm}{n+m}} \left\{\pi_{p,\lambda,f}(\hat{r}_n, \hat{s}_m) - \pi_{p,\lambda,f}(r,s)\right\} \xrightarrow{D} \mathcal{N}_{N^2}\left(0, \Sigma_{p,\lambda,f}(r,s)\right) .$$

The variance $\Sigma_{p,\lambda,f}(r,s)$ is different to $\Sigma_{p,\lambda,f}(r|s)$ from Theorem 3.1. More precisely, given the covariance matrix of the reduced multinomial process $\Sigma(\delta, r, s_\star) = \begin{bmatrix} \delta\,\Sigma(r) & 0 \\ 0 & (1-\delta)\,\Sigma(s_\star) \end{bmatrix}$, we find that

$$\Sigma_{p,\lambda,f}(r,s) = \nabla\phi_{p,\lambda,f}(r, s_\star)\, \Sigma(\delta, r, s_\star)\, \nabla\phi_{p,\lambda,f}(r,s_\star)^T . \tag{3.5}$$

Similar the limit distribution for the empirical ROT distance now reads as

$$\sqrt{\frac{nm}{n+m}} \left\{W_{p,\lambda,f}(\hat{r}_n, \hat{s}_m) - W_{p,\lambda,f}(r,s)\right\} \xrightarrow{D} \mathcal{N}_1\left(0, \sigma^2_{p,\lambda,f}(r,s)\right) .$$

The variance $\sigma^2_{p,\lambda,f}(r,s)$ is again different to $\sigma^2_{p,\lambda,f}(r|s)$ from Theorem 3.2 and given by $\sigma^2_{p,\lambda,f}(r,s) = \gamma^T\, \Sigma_{p,\lambda,f}(r,s)\, \gamma$, where we recall the definition of γ from Theorem 3.2.

3.2 Comparison to (non-regularized) Optimal Transport

The ROT distance approximates the OT distance as $\lambda > 0$ tends to zero (Dessein et al., 2018). Especially for Boltzmann-Shannon entropy regularization, the speed of convergence has been the topic of quite some work dating back at least to Cominetti & San Martín (1994) (see Weed (2018) for a recent contribution and further references). In fact, for general regularizers f such that $\mathbb{R}^{N^2}_{\geq 0} \subset \mathrm{dom}(f)$ a similar argument as in Weed (2018, Prop. 1) yields that

$$\left| W^p_{p,\lambda,f}(r,s) - W^p_p(r,s) \right| \leq \lambda R_f(r,s)$$

with $R_f(r,s) := \max_{\pi, \tilde{\pi} \in \Pi(r,s)} f(\pi) - f(\tilde{\pi})$ the general regularization radius of $\Pi(r,s)$, the set of all feasible OT plans between marginals r and s. For finite spaces with N points R_f can usually be bounded by a quantity independent of r, s. For instance, for f the Boltzmann-Shannon entropy we obtain $R_f \leq 2\log(N)$ and hence that

$$\sup_{r,s \in \Delta_N} \left| W^p_{p,\lambda,f}(r,s) - W^p_p(r,s) \right| \leq 2\lambda \log(N) . \tag{3.6}$$

By a similar argument as in Bigot et al. (2019, Theorem 2.10) we decompose

$$\begin{aligned} &\sqrt{n}\left\{W^p_{p,\lambda(n),f}(\hat{r}_n, s) - W^p_{p,\lambda(n),f}(r,s)\right\} \\ =&\sqrt{n}\left\{W^p_{p,\lambda(n),f}(\hat{r}_n, s) - W^p_p(\hat{r}_n, s)\right\} + \sqrt{n}\left\{W^p_p(r,s) - W^p_{p,\lambda(n),f}(r,s)\right\} \\ &+ \sqrt{n}\left\{W^p_p(\hat{r}_n, s) - W^p_p(r,s)\right\} . \end{aligned} \tag{3.7}$$

For $\lambda(n) = o\left(n^{-1/2}\right)$ the first two terms converge to zero by (3.6) while the third term asymptotically follows the limit law for the (non-regularized) OT distance (Sommerfeld & Munk, 2018). Hence, the limit law of the empirical Sinkhorn divergence in (3.7) is the same as for its non-regularized counterpart. Note that this holds for $W^p_{p,\lambda,f}$ and in fact fails for $W_{p,\lambda,f}$ in the case $r = s$. Our numerical analysis indicates that for the Sinkhorn divergence even much slower rates for $\lambda > 0$ to zero are sufficient to approximate the limit law for the OT distance (for some selective numerical results see Figure 2 in Section 5). In fact, for fixed marginals r and s exponentially fast convergence is possible (Weed, 2018). Nevertheless, as our analysis requires uniformity in the marginals the sub-optimality gap appears as a constant which can be arbitrarily small, in general.

4 Bootstrap

The (non-regularized) OT distance on finite spaces defines a functional that is only *directionally* Hadamard differentiable, i.e., has a *non-linear* derivative with respect to r and s. Hence, the (naive) n out of n bootstrap method to approximate the distributional limits for the (non-regularized) empirical OT distance fails (Sommerfeld & Munk, 2018). However, our arguments underlying the proof of Theorem 3.1 for ROT are based on the usual delta method for *linear* derivatives. As a byproduct we obtain that for the ROT plan and for the ROT distance the n out of n bootstrap is consistent. More precisely, conditionally on the data $X_1, \dots, X_n$, the law of the multinomial empirical bootstrap process $\sqrt{n}\{\hat{r}^*_n - \hat{r}_n\}$ is an asymptotically consistent estimator of the law for the multinomial empirical process $\sqrt{n}\{\hat{r}_n - r\}$ (Van der Vaart & Wellner, 1996, Theorem 3.6.1). Here, $\hat{r}^*_n = \frac{1}{n}\sum_{i=1}^n \delta_{X^*_i}$ is the empirical bootstrap estimator for $\hat{r}_n$ derived by a sample $X^*_1, \dots, X^*_n \overset{i.i.d.}{\sim} \hat{r}_n$. Such conditional weak convergence can be formulated in terms of the bounded Lipschitz metric, that is

$$\sup_{h \in \mathrm{BL}_1(\mathbb{R}^N)} \left|\mathbb{E}[h(\sqrt{n}\{\hat{r}^*_n - \hat{r}_n\})|X_1, \dots, X_n] - \mathbb{E}[h(\sqrt{n}\{\hat{r}_n - r\})]\right| \tag{4.1}$$

converges to zero in probability, where

$$\mathrm{BL}_1\left(\mathbb{R}^N\right) := \left\{ f \colon \mathbb{R}^N \to \mathbb{R} \,\Big|\, \sup_{x \in \mathbb{R}^N} |f(x)| \leq 1,\ |f(x_1) - f(x_2)| \leq \|x_1 - x_2\| \right\}$$

is the set of all bounded functions with Lipschitz constant at most one. Combined with the consistency of the delta method for the bootstrap (Van der Vaart & Wellner, 1996, Theorem 3.9.11), this proves the consistency for the empirical bootstrap ROT plan. The statements again hold true for $r = s$ and $r \neq s$.

Theorem 4.1 (n out of n bootstrap)**.** *Under the assumptions of Theorem 3.1 the n out of n bootstrap for the regularized optimal transport plan is consistent, that is*

$$\begin{aligned}\sup_{h \in BL_1\left(\mathbb{R}^{N^2}\right)} \Big|&\mathbb{E}[h(\sqrt{n}\,\{\pi_{p,\lambda,f}(\hat{r}^*_n, s) - \pi_{p,\lambda,f}(\hat{r}_n, s)\}\,|X_1, \dots, X_n] \\ &- \mathbb{E}[h(\sqrt{n}\,\{\pi_{p,\lambda,f}(\hat{r}_n, s) - \pi_{p,\lambda,f}(r, s)\})]\Big| \xrightarrow{\mathbb{P}} 0\,.\end{aligned}$$

This holds as well for the regularized optimal transport distance

$$\begin{aligned}\sup_{h \in BL_1(\mathbb{R})} \Big|&\mathbb{E}[h(\sqrt{n}\,\{W_{p,\lambda,f}(\hat{r}^*_n, s) - W_{p,\lambda,f}(\hat{r}_n, s)\}\,|X_1, \dots, X_n] \\ &- \mathbb{E}[h(\sqrt{n}\,\{W_{p,\lambda,f}(\hat{r}_n, s) - W_{p,\lambda,f}(r, s)\})]\Big| \xrightarrow{\mathbb{P}} 0\,.\end{aligned}$$

Analogously, the bootstrap consistency is valid if $Y_1^*, \dots, Y_m^* \overset{i.i.d.}{\sim} \hat{s}_m^*$ independently to $X_1^*, \dots, X_n^* \overset{i.i.d.}{\sim} \hat{r}_n^*$ (see Section 3). Identical arguments can be applied to show that the n out of n bootstrap for the Sinkhorn loss $L_\lambda(r, s)$ in (3.4) for $r \neq s \in \Delta_N$ is consistent.

4.1 The m out of n Bootstrap for the Sinkhorn Loss

It is well known that the n out of n bootstrap usually fails for vanishing first-order derivatives (see Chen & Fang (2019); Rippl et al. (2016); Shao (1994) and references therein). This case arises for the Sinkhorn loss $L_\lambda(r, s)$ in (3.4) for $r = s \in \Delta_N$ (see Remark 3.4). A potential remedy is the so called Babu-Bootstrap correction (Babu, 1984) as suggested by Bigot et al. (2019). However, although the empirical performance of the Babu Bootstrap for the Sinkhorn loss is well investigated, it still misses theoretical guarantees. We like to propose an alternative bootstrap procedure. In fact, the m out of n bootstrap for $m = m(n)$ such that ${}^m/_n \to 0$ remains consistent for the Sinkhorn loss under first order degeneracy. The proof immediately follows from Shao (1994, Thm. 1).

Theorem 4.2 (m out of n bootstrap)**.** *Under the assumptions of Theorem 3.1 and the case $r = s \in \Delta_N$ the m out of n bootstrap is consistent for the Sinkhorn loss* (3.4)*. More precisely, for $m = m(n)$ such that $m = o(n)$ as $n \to \infty$ it holds that*

$$\sup_{h \in BL_1(\mathbb{R})} \left| \mathbb{E}[h(n\{L_\lambda(\hat{r}_m^*, r) - L_\lambda(\hat{r}_n, r)\}) | X_1, \dots, X_n] - \mathbb{E}[h(nL_\lambda(\hat{r}_n, r))] \right| \xrightarrow{\mathbb{P}} 0 \,.$$

5 Simulations

We illustrate our distributional limit results in Monte Carlo simulations. As an illustrative example, we investigate the speed of convergence for the empirical Sinkhorn divergence ($p = 1$) to its limit distribution (Corollary 3.3) in the one-sample case in both settings $r = s$ and $r \neq s$. As the Sinkhorn divergence approximates the (non-regularized) OT distance, we also compare for small regularization parameters the finite sample distribution of the Sinkhorn divergence with the limit laws for the (non-regularized) optimal transport distance (OT distance) in (1.8). All simulations were performed using R (R Core Team, 2016). The Sinkhorn divergences are calculated with the R-package *Barycenter* (Klatt, 2018).

Remark 5.1 (Computation of (empirical) variances)**.** *For $p = 1$ it holds that $\sigma^2_{1,\lambda,f}(r|s) = c_1^T \, \Sigma_{1,\lambda,f}(r|s) \, c_1$. According to Example 2.8 and Corollary 3.3, the computation of the variance involves the computation of*

$$\begin{aligned} \Sigma_{1,\lambda,f}(r|s) &= \nabla_r \phi_{1,\lambda,f}(r, s_\star) \, \Sigma(r) \; [\nabla_r \phi_{1,\lambda,f}(r, s_\star)]^T \\ &= D\, A_\star^T \left[A_\star \, D \, A_\star^T\right]^{-1}_{[1:N]} \Sigma(r) \, \left[A_\star \, D \, A_\star^T\right]^{-1}_{[N:1]} A_\star \, D \,, \end{aligned}$$

where the subscript notation $[1:N]$ ($[N:1]$) denotes the first N columns (rows) of the corresponding matrix. Recall that D is equal to a diagonal matrix with diagonal given by the entropy ROT plan $\pi_{1,\lambda,f}(r, s)$ and $A_\star$ is the coefficient matrix in (1.1) *reduced by its last row. Besides calculating the entropy ROT plan, the computation of the variance faces matrix inversion. However, the matrix $A_\star \, D \, A_\star^T$ is symmetric and positive definite by full rank of $A_\star$. Moreover, it posses a block structure given by $A_\star \, D \, A_\star^T = \begin{bmatrix} R & \Pi_{1,\lambda,f} \\ \Pi_{1,\lambda,f}^T & S_\star \end{bmatrix}$, where $\Pi_{1,\lambda,f}$ denotes the matrix version of the entropy ROT plan reduced by its last column, and* *we set $R := diag(r)$ and $S_\star := diag(s_\star)$. Hence, we can apply a block wise inversion and obtain*

$$\left[A_\star\, D\, A_\star^T\right]^{-1}_{[1:N]} = \begin{bmatrix} \left[R - \Pi_{1,\lambda,f}\, S_\star^{-1}\, \Pi_{1,\lambda,f}^T\right]^{-1} \\ -S_\star^{-1}\, \Pi_{1,\lambda,f}^T \left[R - \Pi_{1,\lambda,f}\, S_\star^{-1}\, \Pi_{1,\lambda,f}^T\right]^{-1} \end{bmatrix}.$$

We consider the finite metric space $\mathcal{X}$ to be an equidistant two-dimensional $L \times L$ grid on $[0,1] \times [0,1]$ and the cost $c \in \mathbb{R}^{L^4}$ consisting of the euclidean distance ($p = 1$) between the pixels on the grid. Note that different grid sizes for fixed regularization parameter λ correspond to different amounts of regularization. As recommended by Cuturi (2013), we let the amount of regularization depend on the median distance $q_{50}(c)$ between the pixels on the grid. More precisely, we define the regularization parameter by

$$\lambda := \lambda_0\, q_{50}(c)\,, \tag{5.1}$$

where $\lambda_0 > 0$ is a parameter that we vary for different simulations. The probability distributions $r, s \in \Delta_{L^2}$ on $\mathcal{X}$ are generated as two independent realizations of a Dirichlet random variable $\mathrm{Dir}(\boldsymbol{\alpha})$ with concentration parameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha) \in \mathbb{R}^{L \times L}$. The choice $\alpha = 1$ corresponds to a uniform distribution on the probability simplex Δ_{L^2}.

5.1 Speed of Convergence

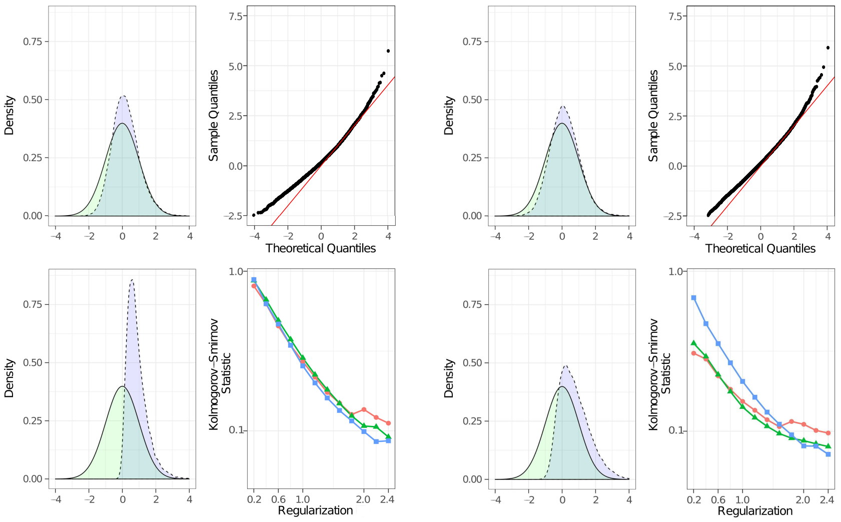
We first generate for grid size $L = 10$ probability distributions r on $\mathcal{X}$ as independent realizations of a $\mathrm{Dir}(\mathbf{1})$ random variable. Given such a distribution, we fix the amount of regularization to $\lambda = 2\, q_{50}(c)$, that is $\lambda_0 = 2$ in (5.1). We then sample $n = 25$ observations according to this probability distribution r and compute

$$\sqrt{\frac{n}{\sigma^2_{1,\lambda,f}(\hat{r}_n \mid r)}}\, \left\{W_{1,\lambda,f}(\hat{r}_n, r) - W_{1,\lambda,f}(r, r)\right\}\,,$$

referred to as a Sinkhorn sample. This is repeated 20,000 times and simulates the scenario when the data generating probability distribution r coincides with the probability distribution to be compared. Similarly, we consider the same set up in the case $r \neq s$ when we simulate independently a second distribution $s \sim \mathrm{Dir}(\mathbf{1})$. The finite sample distributions are then compared to their theoretical limit distributions which by Theorem 3.2 are standard normal distributions.

We demonstrate the results by kernel density estimators and corresponding Q-Q-plots in the first row of Figure 1 (a) and (b). We observe that the finite sample distribution is already well approximated for small sample size ($n = 25$) by the theoretical limit distribution. However, the amount of entropy regularization ($\lambda_0 = 2$) added to OT is rather large. We find that for sample size $n = 25$ the smaller the regularization λ_0 the worse the approximation by the theoretical Gaussian limit law. This is depicted in the second row of Figure 1 (a) and (b) where we analyse for small regularization parameters the Kolmogorov-Smirnov distance (maximum absolute difference between empirical cdf and cdf of the limit law). Note, that the theoretical limit law approximates the finite sample distribution slightly better in the case $r \neq s$.

We additionally investigate the speed of convergence of the finite sample distribution in the small regularization regime when the sample size is large ($n \gg 25$). The approximation is clearly driven by the amount of regularization. This is illustrated for different regularization parameters λ_0 and large sample size in Appendix B. As a benchmark, for $\lambda_0 = 0.2$ we already require $n = 5{,}000$ samples to observe an accurate approximation by the limit distribution from Theorem 3.2.



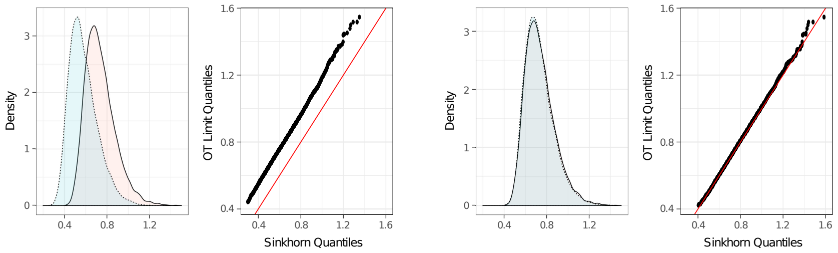
(a) $r = s$

(b) $r \neq s$

Figure 1: (a) Finite sample accuracy of the limit law in the one sample case for $\mathbf{r = s}$. First row: Finite sample density (dashed line) of the empirical Sinkhorn divergence for $n = 25$ on a regular grid of size $L = 10$ with regularization parameter $\lambda_0 = 2$ compared to its limit law (standard Gaussian, solid line). The finite sample density has been estimated with a kernel density estimator with Gaussian kernel and Silverman's rule to select bandwidth. On the right the corresponding Q-Q-plot, where perfect fit is indicated by the red solid line. Second row: L.h.s. same setting as above, $\lambda_0 = 0.6$. R.h.s. Finite sample accuracy in dependence on λ_0: The Kolmogorov-Smirnov distance on a logarithmic scale averaged over five realizations of a $\text{Dir}(\mathbf{1})$ distribution between the finite sample distribution ($n = 25$) of the empirical Sinkhorn divergence and the standard normal distribution as a function of the regularization λ_0.
(b) Finite sample accuracy of the limit law in the one sample case for $\mathbf{r \neq s}$. Same scenario as in (a) whereas here the probability distribution r to be sampled is not equal to the fixed reference probability distribution s.

Motivated by the inaccurate approximation for small sample size in the small regularization regime and the fact that the Sinkhorn divergence converges to the OT distance as $\lambda_0 \searrow 0$, we compare the finite sample distribution to the (non-regularized) OT limit law by Sommerfeld & Munk (2018) for the case $r = s$. From Figure 2 we see that the finite sample distribution for $\lambda_0 = 0.1$ and $\lambda_0 = 0.05$ and small sample size ($n = 25$) is approximated by the OT limit law in (1.8).

Finally, we simulate the (naive) n out of n plug-in bootstrap approximation from Section 4. In fact, the finite bootstrap sample distribution is a good approximation of the finite sample distribution. However, as before, the speed of convergence to the corresponding limit distribution is driven by the amount of regularization. Further details and illustrations for the bootstrap approximation are deferred to Appendix B.



(a) $\lambda_0 = 0.1$ **(b)** $\lambda_0 = 0.05$

Figure 2: Comparison to OT limit law in the one-sample case for $\mathbf{r} = \mathbf{s}$. Comparison of the finite sample distribution ($n = 25$) of the empirical Sinkhorn divergence on an equidistant grid of size $L = 10$ for regularization parameter $\lambda_0 = 0.1$ (a) and $\lambda_0 = 0.05$ (b) to the OT limit law in (1.8) (Sommerfeld & Munk, 2018, Theorem 1). Kernel density estimator for the Sinkhorn sample (dotted line) and the OT sample (solid line) with corresponding Q-Q-plot on the right. The OT distance limit distribution is approximated by a sample of size $20,000$ implemented in the R-package *otinference* (Sommerfeld, 2017).

In summary, the finite sample distribution converges to its theoretical limit law. The accuracy of the approximation is driven by the regularization parameter λ_0. For large regularization added to OT, the limit law serves as a good approximation to the finite sample distribution, already for small sample sizes and independent of the size of the ground space. As λ_0 decreases, accuracy of approximation decreases which is consistent with our theoretical findings in Section 3.2.

6 Reducing Computational Complexity by Resampling

Dvurechensky et al. (2018) analyse algorithms to compute the entropy ROT that yield ϵ-approximates to the OT distance, i.e. $W^p_{p,\lambda,f}(r,s) \leq W^p_p(r,s) + \epsilon$. These methods are usually based on matrix scaling of the underlying $N \times N$ distance matrix in order to find the entropy ROT plan $\pi_{p,\lambda,f}(r,s)$ in (2.1). With increasing number of support points N of the underlying distributions, matrix scaling becomes computational infeasible mainly because of the high memory demand to store and scale the distance matrix.
In order to maintain computational feasibility, we modify an idea introduced by Sommerfeld et al. (2019), i.e. we use a resampling scheme of the underlying probability distribution. The subsequent data example in Section 7 requires to deal with probability distributions with support on up to $N = 300,000$ points (images represented by normalized gray scale pixel intensities). This requires storing a distance matrix with entries $300,000^2 = 9 \cdot 10^{12}$ which is far beyond the storage capacity of any standard laptop. In order to stay within the scope of computational feasibility, we resample from these

probability distributions $n = 50,000$ data points which reduces the required storage to $50,000^2 = 2.5 \cdot 10^9$ entries. Note that while the number of support points of the resampled probability distributions decreases just by a sixth, the memory demand for the corresponding distance matrix is reduced by three orders of magnitude. This is crucial as it keeps matrix scaling feasible.

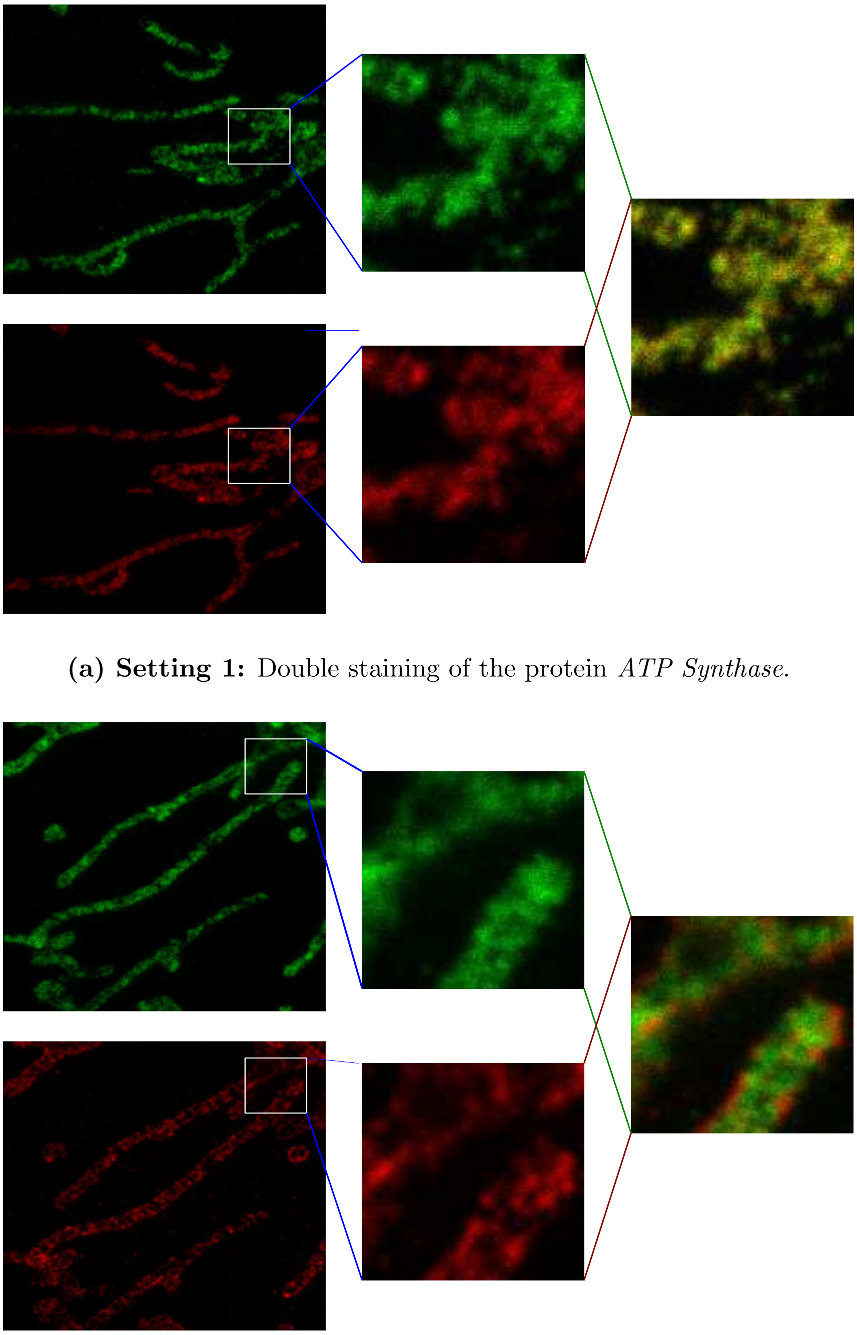
7 Colocalization Analysis

Complex protein interaction networks are at the core of almost all cellular processes. To gain insight into this interaction a valuable tool is colocalization analysis of images generated by fluorescence microscopy aiming to quantify the spatial proximity of proteins (Adler & Parmryd, 2010; Wang et al., 2019; Zinchuk & Grossenbacher-Zinchuk, 2014). With the advent of super-resolution microscopy (nanoscopy) nowadays protein structures at a size of a single protein can be discerned. This challenges conventional colocalization methods, e.g., based on pixel intensity correlation as there is no spatial overlap due to blurring anymore (Xu et al., 2016). The aim of this section is to highlight the benefit of our theory for colocalization in nanoscopy. To this end, we introduce a certain functional of the regularized transport plan as a novel measure for spatial proximity, particularly suited for highly localized image structures as obtained from super resolution microscopy. In the following, we provide a brief illustration. A full analysis on various biological images and a comparison with existing methods is beyond the scope of this paper and will be published separately.

In Figure 3 2-colour-STED (see Hell (2007); Sahl et al. (2017) for more background on this super resolution technique) images (recorded at Jakobs' lab, Department of NanoBiophotonics, Max-Planck Institute for Biophysical Chemistry, Göttingen) are displayed for illustrative purposes. These are generated by inserting two different fluorophore markers which emit photons at different wavelength (two colours) and reading them out, after excitation, with a stimulated emission depletion (STED) laser beam (Hell, 2007). Figure 3 (a) shows 2-colour-STED images which were generated by attaching the two markers to the protein *ATP Synthase*. Figure 3 (b) displays the second case in which the markers are attached to two different proteins *ATP Synthase* and *MIC60*.

We aim to quantify the spatial proximity of *ATP Synthase* and *MIC60* (Figure 3 (b)) and to set this in relation with the highest empirically possible colocalization represented by the double staining of the protein *ATP Synthase* (Figure 3 (a)). The overlay of the channels from setting 2 in Figure 3 (b) already indicates that *ATP Synthase* and *MIC60* are located in different regions as there are only small areas which are yellow. In contrast, and as expected for the highest empirically possible colocalization, the yellow areas in the overlay of the two channels from the double staining in Figure 3 (a) are more pronounced. In the following, we illustrate that the ROT plan (2.1) provides a useful tool to measure colocalization in super-resolution images as it describes the (regularized) optimal matching between the two protein intensity distributions. We stress that the idea to use OT for image comparison (Rubner et al., 2000) and for colocalization is not new, however, previous attempts (Zaritsky et al., 2017) focus on the one-dimensional pixel intensity distribution of grey values which does not capture the spatial information.

The set of pixels define the ground space $\mathcal{X} = \{x_1, \ldots, x_N\}$ with $N = N_x \cdot N_y$, where N_x, N_y are the number of pixels in x- and y- direction, respectively. The pixels colour intensities are understood (after normalization) as discrete probability distributions $r, s \in \Delta_N$ on an equidistant grid in $[0, N_x \cdot l] \times [0, N_y \cdot l]$, where l is the pixel size. The cost to transport pixel intensities from one pixel to the other is given by the squared euclidean



(a) **Setting 1:** Double staining of the protein *ATP Synthase*.

(b) **Setting 2:** Staining of the proteins *ATP Synthase* (green) and *MIC60* (red).

Figure 3: STED images for colocalization analysis. Exemplary STED images of the two colocalization scenarios. Left: Images of a green and a red channel. Image size 666 × 666 pixels, pixel size = 15nm. Middle: Zoom ins (128 × 128 pixels). Right: Overlay of zoomed in images.

distance $c_{(i-1)N+j} = \|x_i - x_j\|^2$ and $c_{\max} := \operatorname{diam}(\mathcal{X})$ is the maximal distance on the ground space $\mathcal{X}$. We introduce the *regularized colocalization measure RCol* which is based on the ROT plan $\pi_{p,\lambda,f}(r,s)$ and defined for $t \in [0, c_{\max}]$ by

$$\mathrm{RCol} := \mathrm{RCol}(\pi_{p,\lambda,f}(r,s))(t) = \sum_{i=1}^{N^2} \pi_{p,\lambda,f}(r,s)_i \mathbb{1}\{c_i \leq t\}\,. \tag{7.1}$$

Intuitively, $\mathrm{RCol}(\pi_{p,\lambda,f}(r,s))(t)$ is the proportion of pixel intensities which is transported on scales smaller or equal to t in the (regularized) optimal matching of the two intensity distributions with respect to some amount of regularization specified by λ. In contrast to the ROT distance (which is just a real value) the RCol-curves provide insight on how much mass on each spatial scale has to be moved in order to match the protein structures. This is biologically relevant, as proteins to be matched above a certain scale will not undergo a chemical reaction. The function $\mathrm{RCol}(\pi)$ constitutes an element in $\mathcal{D}[0, c_{\max}]$ the space of all *càdlàg* functions (Billingsley, 2013) on $[0, c_{\max}]$ equipped with the supremum norm $\|f\|_\infty := \sup_{t\in[0,c_{\max}]} |f(t)|$.

Theorem 7.1. *Let $\widehat{RCol}_n := RCol(\pi_{p,\lambda,f}(\hat{r}_n, s))$ be the empirical regularized colocalization. Under the assumptions of Theorem 3.1, as $n \to \infty$ it holds that*

$$\sqrt{n}\left\{\widehat{RCol}_n - RCol\right\} \xrightarrow{D} RCol(G)\,,$$

where G is the centred Gaussian random variable in $\mathbb{R}^{N^2}$ with covariance $\Sigma_{p,\lambda,f}(r|s)$ given in (3.1).

Proof. The map $\pi \mapsto \mathrm{RCol}(\pi)$ is linear and 1-Lipschitz. Hence, according to Theorem 3.1, the continuous mapping theorem (Van der Vaart, 2000, Theorem 18.11) and

$$\sqrt{n}\left\{\widehat{\mathrm{RCol}}_n - \mathrm{RCol}\right\} = \mathrm{RCol}(\sqrt{n}\left\{\pi_{p,\lambda,f}(\hat{r}_n, s) - \pi_{p,\lambda,f}(r,s)\right\})$$

the assertion follows. □

This result provides *approximate uniform confidence bands (CBs)* for the regularized colocalization. For $\alpha \in (0,1)$, let $\mathfrak{u}_{1-\alpha}$ be the $1-\alpha$ quantile from the distribution of $\|\mathrm{RCol}(G)\|_\infty$. Hence,

$$\mathcal{I}_n := \left[-\frac{\mathfrak{u}_{1-\alpha}}{\sqrt{n}} + \widehat{\mathrm{RCol}}_n,\ \frac{\mathfrak{u}_{1-\alpha}}{\sqrt{n}} + \widehat{\mathrm{RCol}}_n\right] \tag{7.2}$$

is a $1-\alpha$ approximate uniform CB for the regularized colocalization RCol. More precisely, it holds that $\lim_{n\to\infty} \mathbb{P}(\mathrm{RCol} \in \mathcal{I}_n) = 1-\alpha$. In our subsequent data example we require to estimate both probability distributions $r, s \in \Delta_N$ (see Section 3.1). The confidence band (7.2) naturally extends to this case. Defining for $n, m \in \mathbb{N}$ the two sample empirical version of the regularized colocalization $\widehat{\mathrm{RCol}}_{n,m} := \mathrm{RCol}(\pi_{p,\lambda,f}(\hat{r}_n, \hat{s}_m))$, we have, as the sample size $n \wedge m \to +\infty$ and $m/n+m \to \delta \in (0,1)$, that

$$\sqrt{\frac{nm}{n+m}}\left\{\widehat{\mathrm{RCol}}_{n,m} - \mathrm{RCol}\right\} \xrightarrow{D} \mathrm{RCol}(G)\,. \tag{7.3}$$

Now, the random variable G is centred Gaussian with covariance matrix $\Sigma_{p,\lambda,f}(r,s)$ given in (3.5). In particular, for equal sample size $n = m$ the corresponding two sample CB reads as

$$\mathcal{I}_{n,n} := \left[-\frac{\sqrt{2}\mathfrak{u}_{1-\alpha}}{\sqrt{n}} + \widehat{\mathrm{RCol}}_{n,n},\ \frac{\sqrt{2}\mathfrak{u}_{1-\alpha}}{\sqrt{n}} + \widehat{\mathrm{RCol}}_{n,n}\right]\,, \tag{7.4}$$

with $\mathfrak{u}_{1-\alpha}$ according to the supremum of the r.h.s. of (7.3).

7.1 Bootstrap Confidence Bands for Colocalization

We denote by $\widehat{\mathrm{RCol}}^*_{n,n}$ a bootstrap version of the regularized colocalization based on the empirical bootstrap estimators $\hat{r}^*_n$ and $\hat{s}^*_n$ derived by bootstrapping from the empirical distributions $\hat{r}_n$ and $\hat{s}_n$ (see Section 4). Recall that RCol is Lipschitz and therefore, according to Theorem 4.1 and an application of the continuous mapping theorem for the bootstrap (Kosorok, 2008, Proposition 10.7), we deduce that

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R})}\Bigg|\mathbb{E}\left[h\left(\left\|\sqrt{n/2}\left\{\widehat{\mathrm{RCol}}^*_{n,n}-\widehat{\mathrm{RCol}}_{n,n}\right\}\right\|_\infty\right)\big|X_1,\dots,X_n,Y_1,\dots,Y_n\right]$$
$$-\,\mathbb{E}\left[h\left(\left\|\sqrt{n/2}\left\{\widehat{\mathrm{RCol}}_{n,n}-\mathrm{RCol}\right\}\right\|_\infty\right)\right]\Bigg| \xrightarrow{\mathbb{P}} 0\,.$$

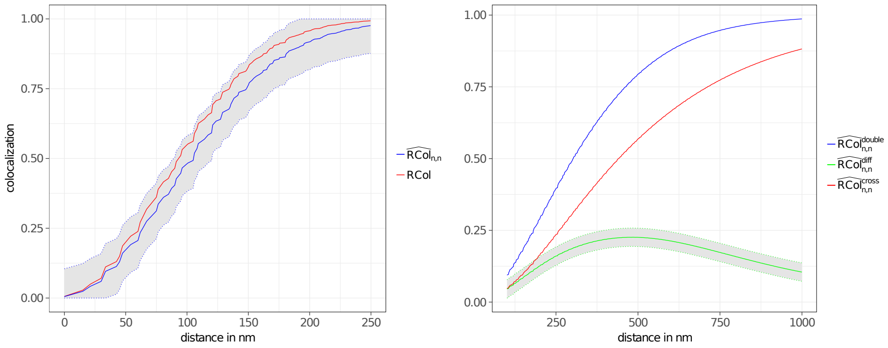
Hence, the quantile $\mathfrak{u}_{1-\alpha}$ is consistently approximated by its bootstrap analogue, say $\mathfrak{u}^*_{1-\alpha}$. This yields a bootstrap approximation of the CB in (7.4).

Remark 7.2 (Reducing Computational Complexity by Resampling)**.** *The evaluation of RCol relies on the computation of the corresponding ROT plan. Algorithms for the ROT plan are usually based on matrix scaling of the underlying $N \times N$ distance matrix (see Peyré et al. (2019) and references therein). For increasing number of support points N of the underlying probability distributions r, s these algorithms become computationally challenging. Resampling from the distributions r and s (see also Sommerfeld et al. (2019)) alleviates this problem. For instance, the subsequent data example requires to deal with probability distributions supported on up to $N = 300,000$ points (images represented by normalized gray scale pixel intensities). Matrix scaling requires a storage capacity for $300,000^2 = 9\cdot 10^{12}$ entries which is far beyond the storage capacity of standard computers. In contrast, resampling $n = 50,000$ data points only demands storage for $2.5 \cdot 10^9$ entries which is available on any standard laptop. Notice that while the number of support points is only decreased by a sixth, the memory demand is reduced by three orders of magnitude and consequently keeps matrix scaling feasible.*

7.1.1 Validation of Bootstrap and Resampling on Real Data

The goal of the following analysis is to investigate the validity of the bootstrap CBs in combination with the resampling scheme (see Remark 7.2). To this end, we consider different pairs of zoom ins (128×128 sections) of the STED images in Figure 3 including the pairs depicted in the middle of each setting. For these instances we are still able to calculate the full regularized colocalization RCol without resampling. Our goal is to validate that RCol is covered by the bootstrap confidence bands. Statistically speaking after fixing a significance level α, we are interested how close the empirical coverage probability is to the claimed nominal coverage of $1-\alpha$. The regularizer f is the entropy (1.3) with amount of regularization given by (5.1) for specified $\lambda_0 > 0$. We resample $n = 2000$ times according to the intensity distribution of the 128×128 image, calculate $B = 100$ bootstrap replications and set the significance level for the CBs in (7.4) to $\alpha = 0.05$. As an illustrative example, Figure 4 (a) demonstrates a case where RCol is covered by the bootstrap CB.

To investigate how well the empirical coverage probability approximates the nominal coverage probability of $1-\alpha = 0.95$, we repeat our approach 100 times and report how often RCol is covered by the bootstrap CB. The result is given in Table 1. For bootstrap replications $B = 100$ and rather large amount of regularization $\lambda_0 \in \{0.5,\ 1,\ 2\}$ we are close to the nominal coverage probability. For small regularization $\lambda_0 = 0.01$ we obtain a slightly smaller empirical coverage probability than desired. This observation is consistent with our empirical simulations for the ROT distance in Section 5.



(a) **(b)**

Figure 4: Entropy regularized colocalization analysis. (a) Setting two: Staining of ATP Synthase and MIC60 for the zoom in (128×128 image). The sampled regularized colocalization ($\lambda_0 = 0.01$) (solid blue line, resampling $n = 2000$) with bootstrap confidence bands (gray area between dashed blue lines) based on the n out of n bootstrap with $B = 100$ replications and $\alpha = 0.05$. Red solid line: Population regularized colocalization. **(b) Resampled colocalization analysis.** Sampled regularized ($\lambda_0 = 0.01$) colocalization (double staining of ATP Synthase, blue solid line; staining of ATP Synthase and MIC60, red solid line; resampling $n = 50,000$) based on the images on the left in Figure 3 together with their difference (solid green line) with bootstrap confidence band (gray area between dashed green lines, n out of n bootstrap, $\alpha = 0.05$, $B = 100$ bootstrap replications).

Table 1: Validation of the bootstrap confidence bands. Empirical coverage probability. Resampling $n = 2000$, bootstrap replications $B = 100$, regularization λ_0 and nominal coverage $1 - \alpha = 0.95$.

Regularization λ_0	Empirical coverage probability
2	0.98
1	0.97
0.5	0.93
0.01	0.88

7.1.2 Empirical Colocalization Analysis of the STED Data

We apply our resampling scheme on the full sized images (666×666 pixels) to evaluate the spatial proximity for each of the two settings, i.e., double staining of *ATPS* (Figure 3 (a)) and staining of *ATPS* and *MIC60* (Figure 3 (b)). We expect the regularized colocalization for setting one (double staining of *ATP Synthase*), say $\text{RCol}^{\text{double}}$, to be large in small distance regimes as most of the transport of pixel intensities should be carried out on small distances. For the regularized colocalization in setting two (staining of *ATP Synthase* and *MIC60*), say $\text{RCol}^{\text{cross}}$, we should observe that there is a significant amount of pixel intensities transported over larger distances resulting in a colocalization that is rather small in the small distance regimes.

Recall that in our validation setup we resampled $n = 2000$ times out of pixel intensities represented by 128×128 images. To achieve comparable accuracy, we require here for 666×666 pixel images a resampling scheme based on $n = 50,000$ resamples of the underlying pixel intensity distributions. We then calculate their sampled regularized colocalization, denoted as $\widehat{\text{RCol}}^{\text{double}}_{n,n}$ and $\widehat{\text{RCol}}^{\text{cross}}_{n,n}$, respectively. In order to compare them, we propose to check their difference $\widehat{\text{RCol}}^{\text{diff}}_{n,n} := \widehat{\text{RCol}}^{\text{double}}_{n,n} - \widehat{\text{RCol}}^{\text{cross}}_{n,n}$, especially in the small distance regime. As before, we obtain CBs by bootstrapping.

The results are presented in Figure 4 (b). We observe that the sampled regularized colocalization $\widehat{\mathrm{RCol}}_{n,n}^{\mathrm{double}}$ (solid blue line) is larger than $\widehat{\mathrm{RCol}}_{n,n}^{\mathrm{cross}}$ (solid red line). Considering their difference $\widehat{\mathrm{RCol}}_{n,n}^{\mathrm{diff}}$ (solid green line) together with the bootstrap CB for the difference (gray area between dashed green lines, $\alpha = 0.05$) demonstrates that the difference is significantly positive at all spatial scales below 1000nm. In fact, our resampling approach for the regularized colocalization analysis reveals that double staining of *ATP Synthase* is significantly more colocalized than staining of *ATP Synthase* and *MIC60* on all relevant spatial scales as biologically expected.

Acknowledgments

The authors are grateful to S. Jakobs and S. Stoldt for providing us with the data in Section 7 as well as to an AE and a referee for helpful comments.
Support by the Volkswagen Foundation and the DFG Research Training Group 2088 *Discovering structure in complex data: Statistics meets Optimization and Inverse Problems* is gratefully acknowledged.

References

Adler, J. & Parmryd, I. (2010). Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytometry A*, 77A(8), 733–742.

Adler, J., Ringh, A., Öktem, O., & Karlsson, J. (2017). Learning to solve inverse problems using Wasserstein loss. *preprint arXiv:1710.10898.*

Altschuler, J., Weed, J., & Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems* (pp. 1961–1971).

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214–223).

Babu, G. J. (1984). Bootstrapping statistics with linear combinations of chi-squares as weak limit. *Sankhyā: The Indian Journal of Statistics, Series A*, (pp. 85–93).

Balikas, G., Laclau, C., Redko, I., & Amini, M.-R. (2018). Cross-lingual document retrieval using regularized Wasserstein distance. In *European Conference on Information Retrieval* (pp. 398–410).: Springer.

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

Bertsekas, D. P. & Castanon, D. A. (1989). The auction algorithm for the transportation problem. *Annals of Operations Research*, 20(1), 67–96.

Bigot, J., Cazelles, E., & Papadakis, N. (2019). Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2), 5120–5150.

Billingsley, P. (2013). *Convergence of Probability Measures.* John Wiley & Sons.

Blondel, M., Seguy, V., & Rolet, A. (2018). Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics* (pp. 880–889).

Bonnans, J. F. & Shapiro, A. (2013). *Perturbation analysis of optimization problems.* Springer Science & Business Media.

Chen, Q. & Fang, Z. (2019). Inference on functionals under first order degeneracy. *Journal of Econometrics*, 210(2), 459–481.

Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1), 223–256.

Cominetti, R. & San Martín, J. (1994). Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1-3), 169–187.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems* (pp. 2292–2300).

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

Del Barrio, E., Loubes, J.-M., et al. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2), 926–951.

Dessein, A., Papadakis, N., & Rouas, J.-L. (2018). Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1), 590–642.

Dvurechensky, P., Gasnikov, A., & Kroshnin, A. (2018). Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International Conference on Machine Learning* (pp. 1366–1375).

Essid, M. & Solomon, J. (2018). Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4), A1961–A1986.

Evans, S. N. & Matsen, F. A. (2012). The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3), 569–592.

Ferradans, S., Papadakis, N., Peyré, G., & Aujol, J.-F. (2014). Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3), 1853–1882.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., & Peyré, G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 2681–2690).: PMLR.

Fiacco, A. V. (1984). *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming.* Academic Press, Inc., by Springer-Verlag.

Gal, T. & Greenberg, H. J. (2012). *Advances in sensitivity analysis and parametric programming*, volume 6. Springer Science & Business Media.

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

Genevay, A., Peyre, G., & Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics* (pp. 1608–1617).

Gerber, S. & Maggioni, M. (2017). Multiscale strategies for computing optimal transport. *The Journal of Machine Learning Research*, 18(1), 2440–2471.

Gottschlich, C. & Schuhmacher, D. (2014). The shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *PloS one*, 9(10), e110214.

Hell, S. W. (2007). Far-field optical nanoscopy. *Science*, 316(5828), 1153–1158.

Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37 (pp. 199–201).

Klatt, M. (2018). *Barycenter: Regularized Wasserstein Distances and Barycenters.* R package version 1.3.2, https://CRAN.R-project.org/package=Barycenter.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer.

Kovačević, M., Stanojević, I., & Šenk, V. (2015). On the entropy of couplings. *Information and Computation*, 242, 369–382.

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

Levina, E. & Bickel, P. (2001). The earth mover's distance is the Mallows distance: Some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2 (pp. 251–256).: IEEE.

Ling, H. & Okada, K. (2007). An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5), 840–853.

Lorenz, D. A., Manns, P., & Meyer, C. (2021). Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, 83(3), 1919–1949.

Lu, Y., Chen, L., & Saidi, A. (2017). Optimal transport for deep joint transfer learning. *preprint arXiv:1709.02995.*

Luenberger, D. G., Ye, Y., et al. (1984). *Linear and Nonlinear Programming*, volume 2. Springer.

Luise, G., Rudi, A., Pontil, M., & Ciliberto, C. (2018). Differential properties of Sinkhorn approximation for learning with Wasserstein distance. In *Advances in Neural Information Processing Systems* (pp. 5864–5874).

Mena, G. & Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris.*

Munk, A. & Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1), 223–241.

Panaretos, V. M. & Zemel, Y. (2018). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1).

Pele, O. & Werman, M. (2009). Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision* (pp. 460–467).

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6), 355–607.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rachev, S. T. & Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media.

Rieder, H. (2012). *Robust Asymptotic Statistics*, volume 1. Springer Science & Business Media.

Rippl, T., Munk, A., & Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151, 90–109.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2), 99–121.

Sahl, S. J., Hell, S. W., & Jakobs, S. (2017). Fluorescence nanoscopy in cell biology. *Nature reviews Molecular cell biology*, 18(11), 685.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., Peyré, G., & Starck, J.-L. (2018). Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1), 643–678.

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

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

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

Shao, J. (1994). Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society*, 122(4), 1251–1262.

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2), 876–879.

Sommerfeld, M. (2017). *otinference: Inference for Optimal Transport.* R package version 0.1.0, https://CRAN.R-project.org/package=otinference.

Sommerfeld, M. & Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 219–238.

Sommerfeld, M., Schrieber, J., Zemel, Y., & Munk, A. (2019). Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105), 1–23.

Tameling, C., Sommerfeld, M., & Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5), 2744–2781.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge university press.

Van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Vershik, A. M. (2013). Long history of the Monge-Kantorovich transportation problem. *The Mathematical Intelligencer*, 35(4), 1–9.

Villani, C. (2008). *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Wang, S., Arena, E. T., Becker, J. T., Bement, W. M., Sherer, N. M., Eliceiri, K. W., & Yuan, M. (2019). Spatially adaptive colocalization analysis in dual-color fluorescence microscopy. *IEEE Transactions on Image Processing*.

Weed, J. (2018). An explicit analysis of the entropic penalty in linear programming. In *Conference On Learning Theory* (pp. 1841–1855).

Wilson, A. G. (1969). The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, (pp. 108–126).

Xu, L., Rönnlund, D., Aspenström, P., Braun, L. J., Gad, A. K. B., & Widengren, J. (2016). Resolution, target density and labeling effects in colocalization studies - suppression of false positives by nanoscopy and modified algorithms. *FEBS J.*, 283(5), 882–898.

Zaritsky, A., Obolski, U., Gan, Z., Reis, C. R., Kadlecova, Z., Du, Y., Schmid, S. L., & Danuser, G. (2017). Decoupling global biases and local interactions between cell biological variables. *Elife*, 6, e22323.

Zinchuk, V. & Grossenbacher-Zinchuk, O. (2014). Quantitative colocalization analysis of fluorescence microscopy images. *Current Protocols in Cell Biology*, 62(1), 4.19.1–4.19.14.

A Proper Regularizers

Various examples for proper regularizers that have also been the focus of recent research, e.g. by Dessein et al. (2018), can be found in Table 2. Moreover, we give their Hessians which are required for the sensitivity analysis (Theorem 2.3 in the main paper).

Table 2: Proper regularizers.

Regularizer f	dom f/f^*	$\nabla^2 f$
Boltzmann-Shannon entropy		
$\sum_{i=1}^{N^2} \log(\pi_i)\pi_i - \pi_i + 1$	$\mathbb{R}^{N^2}_{\geq 0}/\mathbb{R}^{N^2}$	$\mathrm{diag}\big(\frac{1}{\pi}\big)$
Burg entropy		
$\sum_{i=1}^{N^2} \pi_i - \log(\pi_i) - 1$	$\mathbb{R}^{N^2}_{>0}/(-\infty, 1)^{N^2}$	$\mathrm{diag}\big(\frac{1}{\pi^2}\big)$
Fermi-Dirac entropy		
$\sum_{i=1}^{N^2} \log(\pi_i)\pi_i + (1-\pi_i)\log(1-\pi_i)$	$[0,1]^{N^2}/\mathbb{R}^{N^2}$	$\mathrm{diag}\Big(\frac{1}{(1-\pi)\pi}\Big)$
β-potentials $(0 < \beta < 1)$		
$\frac{1}{\beta(\beta-1)} \sum_{i=1}^{N^2} \pi_i^\beta - \beta\pi_i + \beta - 1$	$\mathbb{R}^{N^2}_{\geq 0}/\Big(-\infty, \frac{1}{1-\beta}\Big)^{N^2}$	$\mathrm{diag}\big(\pi^{\beta-2}\big)$
l_p quasi norms $(0 < p < 1)$		
$-\sum_{i=1}^{N^2} \pi_i^p$	$\mathbb{R}^{N^2}_{\geq 0}/\mathbb{R}^{N^2}_{\leq 0}$	$p(1-p)\mathrm{diag}\big(\pi^{p-2}\big)$

B Simulations

We demonstrate that the approximation of the finite sample distribution by the limit law is driven by the amount of regularization $\lambda > 0$ added to OT. Recall that $\lambda := \lambda_0\, q_{50}(c)$, where $\lambda_0 > 0$ is a parameter that we vary for different simulations and $q_{50}(c)$ is the median distance between the pixels on the grid. As illustrated in Figure 5, we observe in our simulations good approximation results for rather large regularization whereas for small regularization the approximation accuracy decreases. This can only be compensated if the sample size n severely increases to several thousands. In general, the approximation is better for $r \neq s$ as for the case of equal distributions $r = s$.

We simulate the (naive) n out of n plug-in bootstrap approximation from Section 4 in the one-sample case for $r = s$. For a grid with $L = 10$ and cost given by the euclidean distance as before, we simulate $r \sim \mathrm{Dir}(\mathbf{1})$ and generate 20,000 realizations

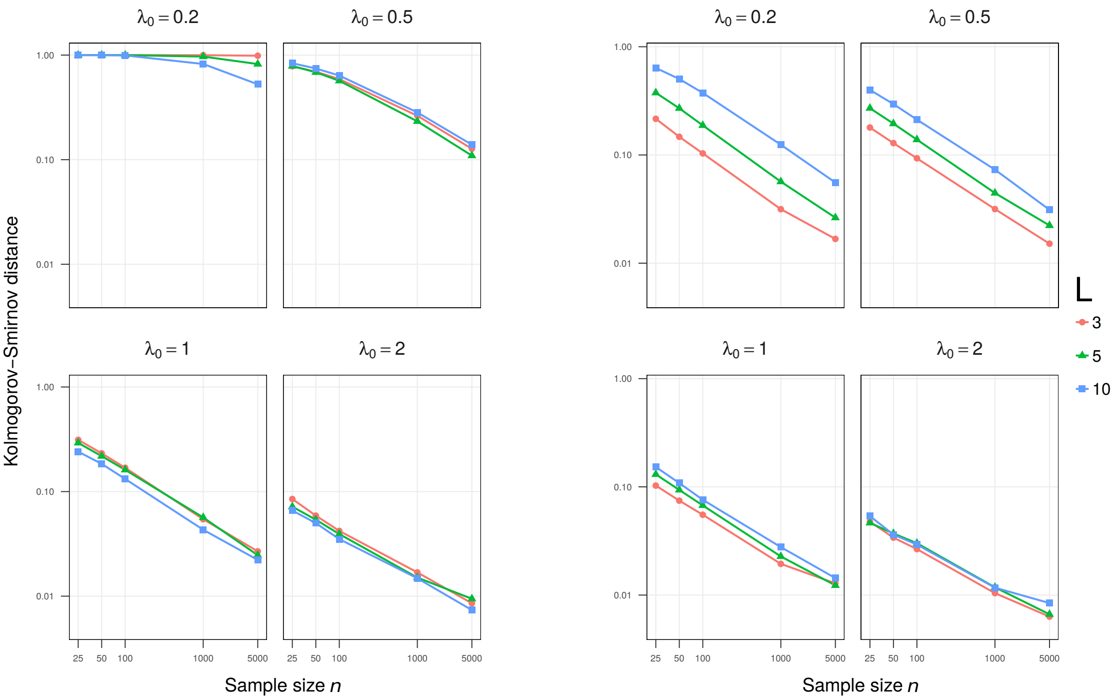
$$\sqrt{n}\,\{W_{1,\lambda,f}(\hat{r}_n, r) - W_{1,\lambda,f}(r,r)\}\ , \tag{B.1}$$

where we set the sample size $n = 100$ and as before vary λ_0 (see (5.1) in the main text). Moreover, for fixed empirical distribution $\hat{r}_n$ and each λ_0 we generate $B = 500$ bootstrap replications

$$\sqrt{n}\,\{W_{1,\lambda,f}(\hat{r}_n^*, r) - W_{1,\lambda,f}(\hat{r}_n, r)\} \tag{B.2}$$

by drawing independently with replacement $n = 100$ times according to $\hat{r}_n$. We then compare the finite sample distributions again by kernel density estimators. The results are depicted in Figure 6.

As already explained in the main text, the finite bootstrap sample distribution is a good approximation of the finite sample distribution. However, as before, the speed of convergence to the corresponding limit distribution is driven by the amount of regularization. Similar results hold for the two-sample case (not displayed).



(a) $r = s$

(b) $r \neq s$

Figure 5: Kolmogorov-Smirnov distance in the one-sample case for r = s (a) and r ≠ s (b). The Kolmogorov-Smirnov distance between the finite Sinkhorn divergence sample distribution and its theoretical normal distribution for $r = s$ (left) and $r \neq s$ (right) as a function of the sample size $n \in \{25, 50, 100, 1000, 5000\}$ for different grid sizes $L \times L$ and different regularization parameters λ_0. The distances are averaged over five pairs of realizations of a Dir($\mathbf{1}$) distribution. The axes are given on a logarithmic scale.

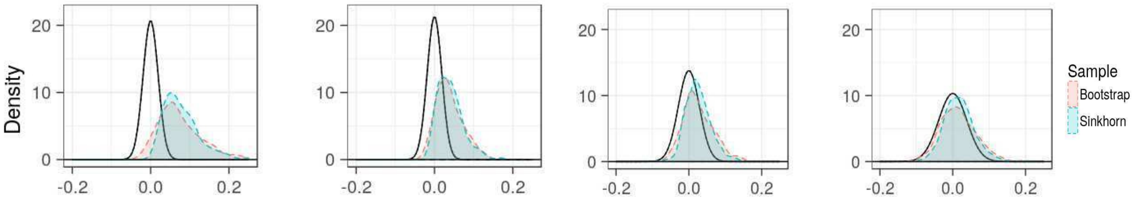


Figure 6: Bootstrap in the one-sample case for $r = s$. Illustration of the (naive) n out of n plug-in bootstrap approximation ($n = 100$) for different regularization parameters λ_0 on a grid of size $L = 10$. The density in blue (resp. red) is obtained by a kernel density estimator (Gaussian kernel with bandwidth given by Silverman's rule) of a Sinkhorn divergence sample (20,000 realizations) (B.1) (resp. bootstrap sample (B.2), $B = 500$ replications). The density of the corresponding Gaussian limit is depicted in black.

CHAPTER C

A Unifying Approach to Central Limit Theorems for Empirical Optimal Transport

A Unifying Approach to Central Limit Theorems for Empirical Optimal Transport

Shayan Hundrieser [*][†] Marcel Klatt [*] Thomas Staudt [*][†] Axel Munk [*][†][‡]

Abstract

We provide a unifying approach to *central limit theorems* (CLTs) for empirical optimal transport (OT). At the heart of our theory is Kantorovich duality representing OT as a supremum over a function class $\mathcal{F}_c$ for an underlying sufficiently regular cost function c. In this regard, OT is considered as a functional defined on $\ell^\infty(\mathcal{F}_c)$ the Banach space of bounded functionals from $\mathcal{F}_c$ to $\mathbb{R}$. We prove its *Hadamard directional differentiability* and conclude via a *functional delta method* that necessitates weak convergence of an underlying empirical process in $\ell^\infty(\mathcal{F}_c)$. The latter can be dealt with *empirical process theory* and requires $\mathcal{F}_c$ to be *Donsker*. We provide sufficient conditions depending on the dimension of the ground space, the underlying cost function and the probability measures under consideration such that a Donsker property holds. Overall, our approach reveals a noteworthy trade-off inherent in CLTs for empirical OT: Kantorovich duality requires $\mathcal{F}_c$ to be sufficiently rich, while the empirical processes only converges weakly if $\mathcal{F}_c$ is not too complex.

The limit distribution of the empirical OT cost is characterized as a supremum of a Gaussian process. In particular, we discuss when the limit distribution is centered normal or degenerates to a Dirac measure at zero. Our approach covers the situation when at least one of the measures has discrete or low dimensional support ($d \leq 3$). This encompasses well-known results for discrete and semi-discrete transport. Moreover, in contrast to recent contributions on distributional limit laws for empirical OT on Euclidean spaces which require centering around its expectation, the CLTs obtained here are centered around the *population* quantity which is well-suited for statistical applications.

Keywords: Central limit theorem, optimal transport, regularity theory, empirical processes, bootstrap, dual potential

MSC 2020 subject classification Primary: 60B12, 60F05, 60G15, 62E20, 62F40; Secondary: 90C08, 90C25, 90C31

1 Introduction

Comparing probability distributions is a fundamental task in statistics, probability theory, machine learning, data analysis and related fields. From this viewpoint, in addition to longstanding mathematical interest, optimal transport (OT) based metrics recently achieved great attention. A major reason is that OT metrics and related similarity measures not only allow comparing general probability distributions, but can also be designed to respect the metric structure of the underlying ground space. This often results in visually

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

[†] Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, Germany

[‡] Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen

appealing and well interpretable outcomes which explains the advancement of OT based data analysis throughout various disciplines, ranging from economics (Galichon, 2016) to statistics (Panaretos & Zemel, 2019), machine learning (Peyré & Cuturi, 2019), signal and image processing (Bonneel et al., 2011; Kolouri et al., 2017) and biology (Schiebinger et al., 2019; Tameling et al., 2021), among others.

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

$$\mathrm{OT}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi(x,y). \tag{1.1}$$

The set $\Pi(\mu,\nu)$ denotes the collection of probability measures on $\mathcal{X}\times\mathcal{Y}$ such that their marginal distributions coincide with μ and ν, respectively. Loosely speaking, OT in (1.1) comprises the challenge to transform the measure μ into the measure ν in a cost optimal way. Under mild assumptions on the cost function, OT_c in (1.1) enjoys the dual formulation

$$\mathrm{OT}_c(\mu,\nu) = \sup_{f\in\mathcal{F}_c} \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} f^c(y)\,\mathrm{d}\nu(y) \tag{1.2}$$

formally known as *Kantorovich duality*. The function class $\mathcal{F}_c$ depends on the underlying cost and $f^c(y) = \inf_{x\in\mathcal{X}} c(x,y) - f(x)$ is the *c-conjugate* for any $f\in\mathcal{F}_c$. Any function f attaining the supremum in (1.2) is termed *Kantorovich potential* and the set

$$S_c(\mu,\nu) := \left\{ f\in\mathcal{F}_c \,\middle|\, \mathrm{OT}_c(\mu,\nu) = \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} f^c(y)\,\mathrm{d}\nu(y)\right\} \tag{1.3}$$

denotes the collection of all Kantorovich potentials. We refer to the monographs by Rachev & Rüschendorf (1998a,b), Villani (2003, 2008), and Santambrogio (2015) for comprehensive treatment, and to Section 2 for further details. In a statistical application the measures μ and ν are estimated from data. We therefore assume to have access to realizations of independent and identically distributed (i.i.d.) random variables $X_1,\dots,X_n\sim\mu$ and the *empirical measure* $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ serves as a proxy for μ. The *population quantity* $\mathrm{OT}_c(\mu,\nu)$ is then estimated by its empirical *plug-in estimator* $\mathrm{OT}_c(\hat{\mu}_n,\nu)$ and poses the question of its statistical performance.

To this end, *central limit theorems* (CLTs) for empirical OT are fundamental as they capture the asymptotic fluctuation of the plug-in estimators around their population quantities after proper standardization. General results in the literature can be broadly distinguished between the *null* $\mu=\nu$ and the *alternative* $\mu\neq\nu$. A well studied setting in this regard is the p-th order *Wasserstein distance*[1] $\mathrm{OT}_p^{1/p}(\mu,\nu)$ on $\mathbb{R}^d$ that arises by choosing the cost $c(x,y) := \|x-y\|^p$ and probability measures with finite p-th moments in (1.1). First analyses have been devoted to the real line ($d=1$) for which the p-th order Wasserstein distance is equal the L^p distance between the quantile functions of the measures. Under the null $\mu=\nu$ and $p=1$, early contributions by del Barrio et al. (1999) (see also Mason 2016) provide necessary and sufficient conditions on the probability measures such that the random quantity $\sqrt{n}\mathrm{OT}_1(\hat{\mu}_n,\mu)$ weakly converges towards an integral of a suitable Brownian bridge (see Example 5.3). A weak limit for $p=2$ is included in Munk & Czado (1998) and del Barrio et al. (2005). The regime $p\in(1,2)$ was analyzed only recently by

[1] To alleviate notation, we write $\mathrm{OT}_p(\mu,\nu)$ for probability measures $\mu,\nu\in\mathcal{P}(\mathbb{R}^d)$ supported on Euclidean spaces and cost function equal to $c(x,y) = \|x-y\|^p$ for some $p\geq 1$. The Wasserstein distance, commonly denoted by $W_p(\mu,\nu)$, is then equal to $\mathrm{OT}_p^{1/p}(\mu,\nu)$. Analogously, the set of Kantorovich potentials in (1.3) is denoted by $S_p(\mu,\nu)$ in this case.

Berthet & Fort (2019). For $p > 2$, central limit theorems are not immediately available but the work by Bobkov & Ledoux (2019) indicates that similar results can be obtained from general quantile process theory (Csörgö & Horváth, 1993). Under the alternative $\mu \neq \nu$ for $p \geq 1$, the random quantity $\sqrt{n}(\mathrm{OT}_p(\hat{\mu}_n, \nu) - \mathrm{OT}_p(\mu, \nu))$ often asymptotically follows a centered Gaussian distribution (Munk & Czado, 1998; Berthet et al., 2020; del Barrio et al., 2019; Berthet & Fort, 2019). Similar in spirit are recent contributions by Hundrieser et al. (2021a) for measures supported on the circle.

A unifying analysis by Sommerfeld & Munk (2018); Tameling et al. (2019) is given for discrete metric spaces $\mathcal{X} = \{x_1, x_2, \ldots\}$ with metric cost $c(x_i, x_j) = d^p(x_i, x_j)$ and probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ which could be seen as a starting point for this paper. Based on a functional delta method the limiting random variable is specifically characterized in terms of a supremum over (infinite dimensional) Gaussian random vectors $\mathbb{G}_\mu \sim \mathcal{N}(0, \Sigma(\mu))$, namely

$$\sqrt{n}\left(\mathrm{OT}_c(\hat{\mu}_n, \nu) - \mathrm{OT}_c(\mu, \nu)\right) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \langle \mathbb{G}_\mu, f \rangle, \tag{1.4}$$

where the supremum is taken over $S_p(\mu, \nu)$ the set of optimal Kantorovich potentials in (1.3). Recently, del Barrio et al. (2021a) extended this to semi-discrete OT for which (1.4) remains valid even if the discrete measure ν is replaced by a general probability measure supported on some Polish space, e.g., absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$. Notably, ν is assumed to be fixed and only the discrete measure $\mu \in \mathcal{P}(\mathcal{X})$ is allowed to be replaced by its empirical counterpart.

Beyond the countable, one-dimensional and semi-discrete case, the asymptotic distributional behaviour for empirical OT becomes much more involved. Already for $d = 2$, precise CLTs remain elusive as highlighted by Ajtai et al. (1984) and Bobkov & Ledoux (2021) stating that the uniform distribution μ on the unit square fulfills $\mathrm{OT}_1(\hat{\mu}_n, \mu) \asymp (\log(n)/n)^{1/2}$ with high probability. On higher dimensions $d \geq 3$, it is well known that any absolutely continuous measure μ with compact support on $\mathbb{R}^d$ fulfills $\mathrm{OT}_1(\hat{\mu}_n, \mu) \gtrsim n^{-1/d}$ (Dudley, 1969; Dobrić & Yukich, 1995) which implies $\sqrt{n}\mathrm{OT}_1(\hat{\mu}_n, \mu)$ to diverge. Indeed, and for general $p \geq 1$, the literature on the convergence of empirical OT is vast and we therefore only give a selective view on recent papers biased towards our main results. Overall, slow convergence rates in the high-dimensional regime seem inevitable as the Wasserstein distances of any order $p \geq 1$ suffers the *curse of dimensionality* $\mathbb{E}\left[\mathrm{OT}_p^{1/p}(\hat{\mu}_n, \mu)\right] \asymp n^{-1/d}$ whenever $d > 2p$ and μ is Lebesgue absolutely continuous (Boissard & Le Gouic, 2014; Fournier & Guillin, 2015; Weed & Bach, 2019). This indicates the scaling rate $\sqrt{n}$ in (1.4) to be typically of wrong order in higher dimensional (Euclidean) spaces. However, when centering with the *empirical expectation*, concentration results demonstrate the random quantity $\sqrt{n}\left(\mathrm{OT}_2(\hat{\mu}_n, \nu) - \mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n, \nu)]\right)$ to be tight (Weed & Bach, 2019; Chizat et al., 2020). A fundamental step further has been taken by del Barrio & Loubes (2019) who obtain for probability measures with $4 + \delta$ finite moments ($\delta > 0$) and positive density in the interior of their convex support that

$$\sqrt{n}\left(\mathrm{OT}_2(\hat{\mu}_n, \nu) - \mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n, \nu)]\right) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mathrm{Var}_\mu[f(X)]\big). \tag{1.5}$$

The asymptotic variance $\mathrm{Var}_\mu[f(X)]$ is equal to the variance of the random variable $f(X)$ with $X \sim \mu$ and f the *unique* Kantorovich potential for (1.2). Under the null $\mu = \nu$, the asymptotic variance is equal to zero and the CLT in (1.5) degenerates, in contrast to the alternative $\mu \neq \nu$ that usually leads to a non-degenerate limit law. The approach by del Barrio & Loubes (2019) relies on approximating the empirical OT cost via (1.2) as a linear functional involving a unique Kantorovich potential for which a CLT immediately follows. Based on the Efron-Stein inequality this functional is shown to serve as a good L^2-approximizer for the random quantity $\sqrt{n}\,(\mathrm{OT}_2(\hat{\mu}_n,\nu) - \mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n,\nu)])$ which allows to conclude. These results have recently been extended by del Barrio et al. (2021b) to more general convex cost functions, including a CLT similar to (1.5) for OT_p under $p > 1$. Notably, the regularity conditions on the measures impose the Kantorovich potential to be unique and it remains unclear how to generalize their approach under non-uniqueness. More crucially, the statement requires centering around the empirical expectation which hinders its immediate use for statistical applications. It might be tempting to replace $\mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n,\nu)]$ by its population counterpart $\mathrm{OT}_2(\mu,\nu)$ in (1.5). On the real line ($d = 1$) and under sufficient regularity assumptions, this is possible (del Barrio et al., 2019). However, it remains a delicate issue for $d \geq 2$. By an observation in Manole & Niles-Weed (2021, Proof of Proposition 21), if μ and ν are uniform measures on two different balls of equal radius the bias is lower bounded by $\mathbb{E}[\mathrm{OT}_2(\hat{\mu}_n,\nu)] - \mathrm{OT}_2(\mu,\nu) \gtrsim n^{-2/d}$. This demonstrates the replacement of the centering with the population quantity in (1.5) to be invalid for $d \geq 5$. Nevertheless, employing a different estimator may allow for faster convergence rates for the bias in higher dimensions. Indeed, for probability measures μ, ν on the unit cube $[0,1]^d$ with sufficiently smooth densities Manole et al. (2021) propose a suitable wavelet estimator $\tilde{\mu}_n$ and prove the bias to be of order $|\mathbb{E}\,[\mathrm{OT}_2(\tilde{\mu}_n,\nu)] - \mathrm{OT}_2(\mu,\nu)| = o(n^{-1/2})$. Combined with a strategy as outlined by del Barrio et al. (2019), the random quantity $\sqrt{n}(\mathrm{OT}_2(\tilde{\mu}_n,\nu) - \mathrm{OT}_2(\mu,\nu))$ is shown to asymptotically follow a centered Gaussian distribution analogous to (1.5), which also degenerates under the null $\mu = \nu$. Despite this being an interesting result, CLTs for empirical OT costs based on general cost functions and centered around the population quantity remain largely open.

In this work, we provide a unifying approach to obtain CLTs for empirical OT that include certain aforementioned settings but go far beyond. In particular, our approach does not rely on discrete spaces, neither on explicit formulas involving quantiles ($d = 1$) nor unique Kantorovich potentials ($d \geq 2$). Our limit laws are reminiscent of the discrete case (1.4) but hold for considerably more general cost functions and probability measures. The limiting random variables are characterized as suprema of Gaussian processes $\mathbb{G}_\mu$ in $\ell^\infty(\mathcal{F}_c)$ indexed in Kantorovich potentials from $S_c(\mu,\nu)$ in (1.3). The CLTs are centered around the *population quantity*. Our main result (Theorem 2.2 in Section 2) states under certain assumptions that

$$\sqrt{n}\,(\mathrm{OT}_c(\hat{\mu}_n,\nu) - \mathrm{OT}_c(\mu,\nu)) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \mathbb{G}_\mu(f). \tag{1.6}$$

The distributional limit law (1.6) is valid whenever $\mathcal{F}_c$ in (1.7) (below) is μ-Donsker and Assumption **(A)**

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

combined with Assumption **(B1)** or **(B2)**

(B1) The space $\mathcal{X}$ is compact and $\{c(\cdot,y) \mid y \in \mathcal{Y}\}$ is equicontinuous[2] on $\mathcal{X}$,

(B2) The spaces $\mathcal{X}$ and $\mathcal{Y}$ are locally compact and $\{c(\cdot,y) \mid y \in \mathcal{Y}\}$ and $\{c(x,\cdot) \mid x \in \mathcal{X}\}$ are equicontinuous[2] on $\mathcal{X}$ and $\mathcal{Y}$, respectively,

are satisfied. In particular, under Assumption **(A)** the function class $\mathcal{F}_c$ used in (1.2) is chosen as a uniformly bounded class of c-concave functions on $\mathcal{X}$ (Villani, 2003, Remark

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

1.13)

$$\mathcal{F}_c := \left\{ f\colon \mathcal{X} \to \mathbb{R} \,\middle|\, \exists\, g\colon \mathcal{Y} \to \mathbb{R},\ -\|c\|_\infty \le g \le 0,\ f(x) = \inf_{y \in \mathcal{Y}} c(x,y) - g(y) \right\}. \tag{1.7}$$

Moreover, $\mathbb{G}_\mu$ in (1.6) is a tight centred Gaussian process and represents the weak limit of the empirical process $\sqrt{n}(\hat{\mu}_n - \mu)$ in $\ell^\infty(\mathcal{F}_c)$, the Banach space of bounded functionals from $\mathcal{F}_c$ to $\mathbb{R}$. The covariance of $\mathbb{G}_\mu$ specifically depends on μ and is equal to

$$\mathbb{E}\left[\mathbb{G}_\mu(f_1)\mathbb{G}_\mu(f_2)\right] = \int_{\mathcal{X}} f_1 f_2 \,\mathrm{d}\mu - \int_{\mathcal{X}} f_1 \,\mathrm{d}\mu \int_{\mathcal{X}} f_2 \,\mathrm{d}\mu \tag{1.8}$$

for two functions $f_1, f_2 \in \mathcal{F}_c$. If instead ν is estimated by its empirical version, then under the same conditions on the cost and the spaces, and if $\mathcal{F}_c^c$, the set of all c-conjugate functions from $\mathcal{F}_c$, is ν-Donsker, the CLT reads as

$$\sqrt{n}\left(\mathrm{OT}_c(\mu, \hat{\nu}_n) - \mathrm{OT}_c(\mu, \nu)\right) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \mathbb{G}_\nu(f^c), \tag{1.9}$$

where $\mathbb{G}_\nu$ is a centered Gaussian process in the Banach space $\ell^\infty(\mathcal{F}_c^c)$ with similar covariance as in (1.8) corresponding to the weak limit of $\sqrt{n}(\hat{\nu}_n - \nu)$.
Our proof technique relies on Kantorovich duality that represents the $\mathrm{OT}_c(\cdot,\cdot)$ cost as a functional from $\ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$ to the reals. We take upon this approach in Section 2 and prove that $\mathrm{OT}_c(\cdot,\cdot)$ is Hadamard directionally differentiable. An application of a general functional delta method (Dümbgen, 1993; Römisch, 2004) allows to conclude our main results on CLTs for empirical OT. This necessitates the empirical process $\sqrt{n}(\hat{\mu}_n - \mu)$ to converge weakly to some tight random element $\mathbb{G}_\mu$ in $\ell^\infty(\mathcal{F}_c)$. The latter can be dealt with *empirical process theory* and requires the function class $\mathcal{F}_c$ to be μ-*Donsker*. Our results naturally extend to the two sample case where both measures are estimated by their empirical versions, simultaneously. We further characterize normality (Theorem 3.2 in Section 3) and *degeneracy* (Theorem 4.2 in Section 4), i.e., when the limit law is equal to a Dirac measure which results from unique and trivial (almost surely constant) Kantorovich potentials, respectively. We emphasize that even if $\mu = \nu$, the CLTs might be non-degenerate, e.g., if μ has disconnected support, as for the discrete case in (1.4), where limit laws usually do not degenerate to a Dirac at zero (Tameling et al., 2019).
In light of the general CLT statement in (1.6), we discuss concrete settings for which the assumptions on the cost are satisfied and $\mathcal{F}_c$ is μ-Donsker (Section 5). The latter manifests itself in the complexity of $\mathcal{F}_c$ measured in terms of covering numbers (Van der Vaart & Wellner, 1996) as well as tail conditions on the measure μ. The covering numbers depend on properties of the cost and the underlying dimension of the ground space (see also Gangbo & McCann 1996; Chizat et al. 2020; Hundrieser et al. 2021b). The simplest case appears on finite spaces where $\mathcal{F}_c$ is even *universal* Donsker. On countable discrete spaces it requires the popular *Borisov-Dudley-Durst* summability condition on the measure μ. This is in line with the aforementioned results by Sommerfeld & Munk (2018) and Tameling et al. (2019), respectively (Corollary 5.1). Beyond discrete spaces, we employ well-known results in empirical process theory. More precisely and tailored to Euclidean spaces $\mathbb{R}^d$ with $d \le 3$, if the cost satisfies certain regularity conditions, then $\mathcal{F}_c$ is μ-Donsker provided the measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfies

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

This implies CLTs for empirical OT on the real line and general costs (Theorem 5.2) but also novel statements for dimension $d = 2, 3$ (Theorem 5.4). We highlight that the CLTs in (1.6) remain valid provided only one of the probability measures is supported on some low-dimensional *compact* Euclidean subset (Theorem 5.8). This is related to recent findings by Hundrieser et al. (2021b) discovering the *lower complexity adaptation* phenomenon that the convergence rate of the empirical OT cost under different population measures adapts to the measure with lower-dimensional support. Based on this observation, our CLTs in Theorem 5.8 encompass recent findings for the semi-discrete OT considered by del Barrio et al. (2021a). Nonetheless and contrary to their results, the CLT in (1.9) remains valid, even if $\mu \in \mathcal{P}(\mathcal{X})$ is fixed with finite support and the more general measure $\nu \in \mathcal{P}(\mathcal{Y})$ is replaced by its empirical counterpart. Apart from the proof of the main result, we postpone further technical proofs and auxiliary results on Kantorovich potentials to Appendix A.

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

2 Central Limit Theorem for Empirical Optimal Transport

Throughout this section, we consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a non-negative cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ such that Assumption **(A)** is fulfilled. Under this assumption the OT cost in (1.1) is finite for any two probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and enjoys Kantorovich duality (1.2). The set $\mathcal{F}_c$, over which the dual is optimized, is the collection of uniformly bounded *c-concave* functions on $\mathcal{X}$ defined as in (1.7) (for details see Villani 2003, Remark 1.13 and Appendix A.2). The supremum in (1.2) over $\mathcal{F}_c$ is attained, i.e., there exists a Kantorovich potential $f \in \mathcal{F}_c$, where we recall the set $S_c(\mu,\nu)$ in (1.3) of all Kantorovich potentials. Notably, as the cost function is continuous, any function $f \in \mathcal{F}_c$ and its c-conjugate f^c are upper semi-continuous and thus measurable on the Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively.

More general structural properties for c-concave functions $\mathcal{F}_c$ and its c-transformed class $\mathcal{F}_c^c := \{f^c \mid f \in \mathcal{F}_c\}$ are intrinsically linked to the cost function and the underlying Polish space. Hence, to guarantee (minimal) regularity properties of the set of Kantorovich potentials we impose Assumption **(B1)** or **(B2)**. Under either of these conditions the Kantorovich potentials $S_c(\mu,\nu)$ and $S_c^c(\mu,\nu)$ are continuous as well (Lemma A.1). Moreover, under compactness or local compactness of the underlying ground spaces $\mathcal{X}$ and $\mathcal{Y}$ as well as the equicontinuity condition of the cost function, we are able to employ a version of the *Arzelà-Ascoli* theorem (see proof of Theorem 2.2 and in particular Step 3) to the classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$.

Examples of locally compact spaces are Polish spaces but also include finite and countable spaces equipped with discrete topology. Assumptions **(A)** and **(B2)** are then fulfilled if the cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is uniformly bounded. Furthermore and for compact subsets

$\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, the Assumptions **(A)** and **(B1)** are fulfilled for costs $c(x,y) = \|x-y\|^p$ with $p \geq 1$ inherent in the popular Wasserstein distance. More examples in this regard are detailed in Section 5.

2.1 Main Result

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

Definition 2.1 (Van Der Vaart 1996)**.** A function class $\mathcal{F}$ of measurable, pointwise-bounded functions on $\mathcal{X}$ is called *μ-Donsker*, if the empirical process $\sqrt{n}(\hat{\mu}_n - \mu)$ weakly converges to a tight random element $\mathbb{G}_\mu$ in $\ell^\infty(\mathcal{F})$. Furthermore, $\mathcal{F}$ is *universal Donsker* if for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ the function class $\mathcal{F}$ is μ-Donsker.

In case $\mathcal{F}_c$ is μ-Donsker, the weak limit $\mathbb{G}_\mu$ is a mean-zero Gaussian process indexed over $\mathcal{F}_c$ with covariance for $f_1, f_2 \in \mathcal{F}_c$ as in (1.8). We now state our main result on the central limit laws for the empirical OT cost.

Theorem 2.2. *Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption* **(A)** *combined with* **(B1)** *or* **(B2)***. For two probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$, i.i.d. random variables $X_1, \dots, X_n \sim \mu$ and independently to that i.i.d. random variables $Y_1, \dots, Y_m \sim \nu$ with $n, m \in \mathbb{N}$ denote the empirical measures by $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_m := \frac{1}{m}\sum_{i=1}^m \delta_{Y_i}$.*

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

$$\sqrt{n}\left(\mathrm{OT}_c(\hat{\mu}_n, \nu) - \mathrm{OT}_c(\mu,\nu)\right) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \mathbb{G}_\mu(f). \tag{2.1a}$$

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

$$\sqrt{m}\left(\mathrm{OT}_c(\mu, \hat{\nu}_m) - \mathrm{OT}_c(\mu,\nu)\right) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \mathbb{G}_\nu(f^c). \tag{2.1b}$$

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

$$\sqrt{\frac{nm}{n+m}}\left(\mathrm{OT}_c(\hat{\mu}_n, \hat{\nu}_m) - \mathrm{OT}_c(\mu,\nu)\right) \xrightarrow{\mathcal{D}} \sup_{f \in S_c(\mu,\nu)} \left(\sqrt{\delta}\mathbb{G}_\mu(f) + \sqrt{1-\delta}\mathbb{G}_\nu(f^c)\right). \tag{2.1c}$$

Proof. Our proof is based on the Hadamard directional differentiability of the OT cost on the set of probability measures $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ with respect to the topology induced by $\ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$ (see the subsequent Theorem 2.5). Then, the statements in Theorem 2.2 follow from the functional delta method (Römisch, 2004, Theorem 1). More precisely, under settings (i) and (ii) it holds for $n \to \infty$ and $m \to \infty$ that

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

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

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

with $m/(n+m) \to \delta \in (0,1)$. Moreover, Theorem 2.5 below asserts that $\mathrm{OT}_c\colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ is Hadamard directionally differentiable at (μ,ν) with respect to $\ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c)$ tangentially to $\mathcal{P}(\tilde{\mathcal{X}}) \times \mathcal{P}(\tilde{\mathcal{Y}})$ with $\tilde{\mathcal{X}} = \mathrm{supp}(\mu)$ and $\tilde{\mathcal{Y}} = \mathrm{supp}(\nu)$ and $\mathcal{T}_\mu(\mathcal{P}(\tilde{\mathcal{X}})) = \mathrm{Cl}\left\{\frac{\mu'-\mu}{t} \text{ for } t>0, \mu' \in \mathcal{P}(\tilde{\mathcal{X}})\right\}$ and analogously for $\mathcal{T}_\nu(\mathcal{P}(\tilde{\mathcal{Y}}))$. Portmanteau's Theorem for closed sets (Van der Vaart & Wellner, 1996, Theorem 1.3.4 (iii)) proves that

$$\mathbb{P}\left(\mathbb{G}_\mu \in \mathcal{T}_\mu(\mathcal{P}(\tilde{\mathcal{X}}))\right) \geq \limsup_{n\to\infty} \mathbb{P}\left(\sqrt{n}(\hat{\mu}_n - \mu) \in \mathcal{T}_\mu(\mathcal{P}(\tilde{\mathcal{X}}))\right) = 1 \tag{2.2}$$

since $\mathrm{supp}(\mu_n) \subseteq \mathrm{supp}(\mu)$, and analogously for $\mathbb{G}_\nu$. The assertions in (2.1) now immediately follow by the functional delta method. □

We investigate specific settings that yield novel distributional limit results in Section 5. At this stage, we like to highlight that Theorem 2.2 links CLTs for the empirical OT cost to the study of the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ being Donsker, respectively.

Remark 2.3. Even if Assumptions **(B1)** and **(B2)** fail to hold, the first step in the proof of Theorem 2.5 still asserts Hadamard directional differentiability of the OT cost. However, in this case the derivative rather depends on the set of ε-approximate optimizer $S_c(\mu,\nu,\varepsilon) := \{f \in \mathcal{F}_c \mid \mu(f) + \nu(f^c) \geq \mathrm{OT}_c(\mu,\nu) - \varepsilon\}$ instead of the set of Kantorovich potentials $S_c(\mu,\nu)$. Hence, employing a functional delta method (Römisch, 2004, Theorem 1), we obtain that, as long as $\mathcal{F}_c$ is μ-Donsker, a CLT of the form

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

is valid. Assumptions **(B1)** and **(B2)** therefore serve to simplify Theorem 2.2 as the limit in ε vanishes. For more details we refer to Section 2.2.

Remark 2.4 (Bootstrap consistency)**.** It is well-known that the naive n-out-of-n bootstrap fails to be consistent if the functional is not linearly Hadamard differentiable (Dümbgen, 1993; Fang & Santos, 2019). Instead, under Hadamard directional differentiability the work by Dümbgen (1993, Proposition 2) asserts consistency of the m-out-of-n bootstrap for $m = o(n)$ which has immediate consequences for the approximation of quantiles for the empirical OT cost. We formalize this principle exemplary for the one-sample case. For $n,m \in \mathbb{N}$, consider i.i.d. samples $X_1,\dots,X_n \sim \mu$ with empirical measure $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and consider i.i.d. bootstrap samples $X_1^*,\dots,X_k^* \sim \hat{\mu}_n$ with corresponding (bootstrap) empirical measure $\hat{\mu}_{n,k}^* = \frac{1}{k}\sum_{i=1}^k \delta_{X_i^*}$. Then, it follows for $n,k\to\infty$ with $k = o(n)$ that

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

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

2.2 Hadamard Directional Differentiability

Based on Kantorovich duality (1.2), we consider the OT cost as an optimization problem over the set of c-concave functions mapping from the set of probability measures $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ into the reals. For a pair of probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, we prove Hadamard directional differentiability of the OT cost at (μ,ν) with respect to the Banach spaces $\ell^\infty(\mathcal{F}_c)\times\ell^\infty(\mathcal{F}_c^c)$. For a definition of this notion of differentiability, we refer to Römisch (2004). Moreover, since the support of empirical measures is always contained in the support of the underlying measure, we focus in the following only on differentiability *tangentially* to the set of probability measures whose support is contained in the support of μ and ν, respectively. This means that only those perturbations for μ and ν are considered which do not lead to an enlargement of the support. The corresponding derivative for the OT cost turns out to depend on the set of Kantorovich potentials $S_c(\mu,\nu)$ in (1.3).

Theorem 2.5 (Hadamard directional differentiability of OT cost)**.** *Suppose that for two Polish spaces $\mathcal{X}$, $\mathcal{Y}$ and the cost function $c\colon\mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$ the Assumption **(A)** combined with **(B1)** or **(B2)** are satisfied. Consider the function class $\mathcal{F}_c$ in* (1.7) *and two probability measures $(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})$ with their respective support $\tilde{\mathcal{X}}:=\operatorname{supp}(\mu)$ and $\tilde{\mathcal{Y}}:=\operatorname{supp}(\nu)$. Then, the OT cost functional*

$$\mathrm{OT}_c\colon\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})\subseteq\ell^\infty(\mathcal{F}_c)\times\ell^\infty(\mathcal{F}_c^c)\to\mathbb{R},\quad(\mu,\nu)\mapsto\sup_{f\in\mathcal{F}_c}\left(\mu(f)+\nu(f^c)\right)\tag{2.3}$$

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

$$\mathcal{D}^H_{|(\mu,\nu)}\mathrm{OT}_c\colon\mathcal{T}_\mu\left(\mathcal{P}(\tilde{\mathcal{X}})\right)\times\mathcal{T}_\nu\left(\mathcal{P}(\tilde{\mathcal{Y}})\right)\to\mathbb{R},\quad(\Delta_\mu,\Delta_\nu)\mapsto\sup_{f\in S_c(\mu,\nu)}\left(\Delta_\mu(f)+\Delta_\nu(f^c)\right).\tag{2.4}$$

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

Proof. The proof is inspired by Römisch (2004, Proposition 1) and Cárcamo et al. (2020, Theorem 2.1 and Corollary 2.2). Compared to their setting, the proof is specifically tailored to the OT cost in (1.2) and exploits properties of the function class $\mathcal{F}_c$. We divide the proof into four steps. The first two essentially prove the main result, whereas the last two are concerned with more technical details.

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

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

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

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

and the set of ε approximizers

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

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

$$\lim_{\varepsilon\searrow 0} \sup_{f\in S_c(\mu,\nu,\varepsilon)} (\varDelta_\mu(f) + \varDelta_\nu(f^c)) = \sup_{f\in S_c(\mu,\nu)} (\varDelta_\mu(f) + \varDelta_\nu(f^c)). \tag{2.5}$$

The left hand side in (2.5) is greater or equal to the right hand side since $S_c(\mu,\nu) \subseteq S_c(\mu,\nu,\varepsilon)$ for all $\varepsilon > 0$. For the converse, take a positive decreasing sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ with $\varepsilon_n \searrow 0$ and a sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}_c$ with $f_n \in S_c(\mu,\nu,\varepsilon_n)$ for which

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

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

$$\lim_{k\to\infty} (\varDelta_\mu(f_{n_k}) + \varDelta_\nu(f_{n_k}^c)) = (\varDelta_\mu(h) + \varDelta_\nu(h^c)) \tag{2.6}$$

the equality (2.5) follows from

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

To verify (2.6), let $\delta > 0$ and select $M \in \mathbb{N}$ such that $\|\varDelta_\mu - \varDelta_{\mu,M}\|_{\mathcal{F}_c} < \delta/4$. Pointwise convergence of f_{n_k} on $\operatorname{supp}(\mu)$ to $h + a$ combined with the uniform bound on $\mathcal{F}_c$ asserts by dominated convergence for μ_M, μ by $\operatorname{supp}(\mu_M) \subseteq \operatorname{supp}(\mu)$ existence of $K \in \mathbb{N}$ with

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

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

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

where we use in the second equality the definition of $\varDelta_{\mu,M}$ and in the last equality that $(\mu - \mu_M)(a) = 0$. Repeating the argument for $|\varDelta_\nu(f_{n_k}^c) - \varDelta_\nu(h^c)|$ yields (2.6).

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

> **Theorem 2.6** (McCoy & Ntantu 1988, Theorem 3.2.6). *If $\mathcal{X}$ is locally compact, then a set of continuous real-valued functions on $\mathcal{X}$ is compact in the compact-open topology if and only if it is closed, pointwise bounded and equicontinuous.*

Since $\{c(\cdot,y) \mid y \in \mathcal{Y}\}$ is equicontinuous, so is $\mathcal{F}_c$ and its closure $\mathrm{Cl}(\mathcal{F}_c)$ in the compact-open topology is as well. Moreover, since $\mathcal{F}_c$ is uniformly bounded by a constant, Theorem 2.6 asserts existence of a subsequence of f_n that converges uniformly on compact sets to a function $f \in \mathrm{Cl}(\mathcal{F}_c)$ which is also continuous. By restricting to a subsequence, we assume that f_n uniformly converges on compact sets to f. Under Assumption **(B1)**, this asserts that f_n uniformly converges on $\mathcal{X}$ to f and thus f_n^c uniformly converges on $\mathcal{Y}$ to f^c. Under setting **(B2)**, we cannot argue analogously and instead repeat the above argument and assume by local compactness of $\mathcal{Y}$ and equicontinuity of $\{c(x,\cdot) \mid x \in \mathcal{X}\}$ that f_n^c uniformly converges on compact sets of $\mathcal{Y}$ to some continuous function $g \in \mathrm{Cl}(\mathcal{F}_c^c)$.

Step 4. Structural properties of limits with general OT theory.
It remains to show that there exists a c-concave function $h \in S_c(\mu,\nu)$ and some $a \in \mathbb{R}$ such that $f = h + a$ on $\mathrm{supp}(\mu)$ and $g = h^c - a$ on $\mathrm{supp}(\nu)$. For this purpose, note that uniform convergence on compact sets implies pointwise convergence. Since f_n, f, f_n^c, g are all absolutely bounded by $\|c\|_\infty$, there exists by dominated convergence for any $\delta > 0$ some $N_1 \in \mathbb{N}$ such that $|\mu(f) - \mu(f_n)| + |\nu(g) - \nu(f_n^c)| \le \delta/2$ for all $n \ge N_1$. Further, there is $N_2 \in \mathbb{N}$ with $\varepsilon_{N_2} \le \delta/2$. Hence, for $n \ge \max(N_1, N_2)$ it follows from $f_n \in S_c(\mu,\nu,\varepsilon_n) \subseteq S_c(\mu,\nu,\delta/2)$ that

$$\mu(f) + \nu(g) \ge \mu(f_n) + \nu(f_n^c) - \delta/2 \ge \mathrm{OT}(\mu,\nu) - \delta,$$

which yields by choosing $\delta > 0$ arbitrarily small that $\mu(f) + \nu(g) \ge \mathrm{OT}(\mu,\nu)$. Moreover, for $(x,y) \in \mathcal{X} \times \mathcal{Y}$ we find that

$$f(x) + g(y) = \lim_{n\to\infty} f_n(x) + \lim_{n\to\infty} f_n^c(y) = \lim_{n\to\infty} f_n(x) + f_n^c(y) \le c(x,y), \tag{2.7}$$

which asserts by the dual formulation for the OT cost (Villani, 2008, Theorem 5.9) that $\mu(f) + \nu(g) = \mathrm{OT}_c(\mu,\nu)$. Upon defining $\tilde{h} := g^c$, we therefore obtain from (2.7) the inequalities $f \le \tilde{h}$ on $\mathcal{X}$ and $g \le \tilde{h}^c$ on $\mathcal{Y}$. Hence, it holds that

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

where the last inequality follows from $\tilde{h}(x) + \tilde{h}^c(y) \le c(x,y)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. We conclude that $f = \tilde{h}$ holds μ-almost surely and $f^c = \tilde{h}^c$ holds ν-almost surely, and continuity asserts equality on the whole support of μ and ν, respectively. Finally, we note by Villani (2003, Remark 1.13) that any c-concave Kantorovich potential $\tilde{h}$ can be suitably shifted by a constant $a \in \mathbb{R}$ such that $0 \le \tilde{h}(x) - a \le \|c\|_\infty$ for all $x \in \mathcal{X}$ and $-\|c\|_\infty \le (\tilde{h} - a)^c(y) = \tilde{h}^c(y) + a \le 0$ for all $y \in \mathcal{Y}$. In particular, the function $h := \tilde{h} - a$ lies in $S_c(\mu,\nu)$ and fulfills the asserted properties. □

Remark 2.7. The only setting for which the function $\tilde{h}$ in step four might not be contained in $S_c(\mu,\nu)$ occurs when the function g does not fulfill $-\|c\|_\infty \le g \le 0$. This could happen since c-conjugate potentials f^c for $f \in \mathcal{F}_c$ do not necessarily satisfy $-\|c\|_\infty \le f^c \le 0$.

3 Normal Limits under Unique Kantorovich Potentials

The set of Kantorovich potentials $S_c(\mu,\nu)$ in (1.3) and its c-conjugates $S_c^c(\mu,\nu)$ play a defining role for the limiting random variables in (2.1a), (2.1b), (2.1c) in Theorem 2.2. A particular setting arises if Kantorovich potentials are uniquely determined. Since any $f \in S_c(\mu,\nu)$ can be shifted arbitrarily and $f + a$ for $a \in \mathbb{R}$ is still a c-concave function that solves (1.2), uniqueness refers to $S_c(\mu,\nu)$ being a singleton up to an additive constant. Furthermore, for the purpose of this section, it suffices if uniqueness holds almost surely.

Definition 3.1 (Unique Kantorovich potentials)**.** The set of Kantorovich potentials $S_c(\mu,\nu)$ in (1.3) is said to be *almost surely unique* if the difference $f_1 - f_2$ for all $f_1, f_2 \in S_c(\mu,\nu)$ is constant μ-almost surely.

We like to mention that for a continuous cost function this notion of uniqueness is equivalent to ν-almost sure uniqueness (up to constant shifts) of c-conjugate Kantorovich potentials on the space $\mathcal{Y}$ (Staudt et al., 2021, Lemma 5). Under almost sure uniqueness of Kantorovich potentials all weak limits from (2.1) simplify to centered normal distributions.

Theorem 3.2 (Normal limits)**.** *Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$ satisfying Assumption **(A)** combined with **(B1)** or **(B2)**. Assume that the set $S_c(\mu,\nu)$ of Kantorovich potentials for $\mu\in\mathcal{P}(\mathcal{X})$ and $\nu\in\mathcal{P}(\mathcal{Y})$ is almost surely unique. Then, for any $f\in S_c(\mu,\nu)$ and empirical measures $\hat{\mu}_n, \hat{\nu}_m$ the following CLT is valid.*

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

$$\sqrt{n}\left(\mathrm{OT}_c(\hat{\mu}_n,\nu) - \mathrm{OT}_c(\mu,\nu)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{Var}_\mu[f(X)]), \quad X\sim\mu. \tag{3.1a}$$

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

$$\sqrt{m}\left(\mathrm{OT}_c(\mu,\hat{\nu}_m) - \mathrm{OT}_c(\mu,\nu)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{Var}_\nu[f^c(Y)]), \quad Y\sim\nu. \tag{3.1b}$$

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

$$\begin{aligned}&\sqrt{\frac{nm}{n+m}}\left(\mathrm{OT}_c(\hat{\mu}_n,\hat{\nu}_m) - \mathrm{OT}_c(\mu,\nu)\right)\\ &\qquad\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \delta\mathrm{Var}_\mu[f(X)] + (1-\delta)\mathrm{Var}_\nu[f^c(Y)]\right), \quad X\sim\mu, Y\sim\nu.\end{aligned} \tag{3.1c}$$

Proof. Under almost sure uniqueness of Kantorovich potentials on $\mathcal{X}$, we conclude by Staudt et al. (2021, Lemma 5) almost sure uniqueness on $\mathcal{Y}$. By Lemma A.1 the Kantorovich potentials are continuous on the supports of μ and ν. We therefore conclude that the set of Kantorovich potentials $S_c(\mu,\nu)$ and its c-conjugate counterparts $S_c^c(\mu,\nu)$ are (deterministically) unique up to a constant shift on the support of μ and ν, respectively. Following along the lines of step two of the proof for Theorem 2.5, we conclude that the directional Hadamard derivative is linear. Hence, by the functional delta method (Römisch, 2004) the weak limit for the empirical OT cost is normal with variance as stated in (3.1). □

A useful statistical application of Theorem 3.2 arises when the limit variance $\mathrm{Var}_\mu[f(X)]$ is consistently estimated from i.i.d. data. Indeed, under Assumptions **(A)** combined with **(B1)** or **(B2)** this holds, leading to the following pivotal limits.

Corollary 3.3 (Pivotal limit)**.** *Suppose the setting of Theorem 3.2. Then, for any measures $\mu\in\mathcal{P}(\mathcal{X})$ and $\nu\in\mathcal{P}(\mathcal{Y})$ such that the μ-almost surely unique Kantorovich potential $f\in S_c(\mu,\nu)$ is not μ-almost surely constant, it follows for $n\to\infty$ and any $f_n\in S_c(\hat{\mu}_n,\nu)$ that*

$$\sqrt{n}\,\frac{\mathrm{OT}_c(\hat{\mu}_n,\nu) - \mathrm{OT}_c(\mu,\nu)}{\sqrt{\mathrm{Var}_{\hat{\mu}_n}[f_n(X)]}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)\,. \tag{3.2}$$

Analogous statements hold for the weak limits in (3.1b), if $f^c\in S_c^c(\mu,\nu)$ is not ν-almost surely constant and in (3.1c) if at least f or f^c is not almost surely constant.

Proof of Corollary 3.3. We only show the claim for (3.2), the corresponding pivotal limits for (3.1b) and (3.1c) follow analogously. In view of Theorem 3.2 and Slutzky's lemma it suffices to show for $f_n \in S_c(\hat{\mu}_n, \nu)$ that $\mathrm{Var}_{\hat{\mu}_n}[f_n(X)]$ converges almost surely for $n \to \infty$ to $\mathrm{Var}_\mu[f(X)] > 0$ for $f \in S_c(\mu, \nu)$. Note that $S_c(\hat{\mu}_n, \nu) \subseteq S_c(\mu, \nu, 2\,\|\hat{\mu}_n - \mu\|_{\mathcal{F}_c})$ where $\|\hat{\mu}_n - \mu\|_{\mathcal{F}_c}$ tends to zero almost surely since $\mathcal{F}_c$ is μ-Donsker. Hence, following along steps three and four of the proof for Theorem 2.5, it follows that there exists a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ such that $(f_{n_k}, f^c_{n_k})$ converges pointwise for $k \to \infty$ to $(h + a, h^c - a)$ for some $h \in S_c(\mu, \nu)$ and some $a \in \mathbb{R}$. Since $\mathcal{F}_c$ is uniformly bounded by $\|c\|_\infty$, and since $\mathcal{F}_c$ as well as the element-wise squared function class $\mathcal{F}_c^2 = \{f^2 : f \in \mathcal{F}_c\}$ are both μ-Donsker (Van der Vaart & Wellner, 1996, Theorem 2.10.6), we therefore conclude that $\mathrm{Var}_{\hat{\mu}_{n_k}}[f_{n_k}(X)] \to \mathrm{Var}_\mu[h(X)]$ for $k \to \infty$. Finally, by almost sure uniqueness of Kantorovich potentials it holds that $\mathrm{Var}_\mu[h(X)] = \mathrm{Var}_\mu[f(X)]$. □

Uniqueness of Kantorovich potentials We close this section with a discussion on uniqueness of Kantorovich potentials as required in Theorem 3.2 and Corollary 3.3, an issue which has obtained less attention in the literature compared to the related questions of uniqueness for optimal transport plans.

On Euclidean spaces, Kantorovich potentials are unique when the cost function is differentiable and the probability measures have bounded support equal to the closure of a bounded open connected set (Santambrogio, 2015, Proposition 7.18). Generalizations for probability measures with unbounded support and strictly convex costs were obtained by del Barrio et al. (2021b, Theorem 2.4) and Bernton et al. (2021, Theorem B.2) relying on absolute continuity of the underlying measures together with connected support and negligible boundary. In such setting, one exploits the fact that the gradient of a Kantorovich potential at some point x is determined by the gradient of $c(\cdot, y)$ at x for any $(x, y) \in \mathrm{supp}(\pi)$, where π is an optimal transport plan (see Santambrogio 2015, Proposition 1.15). Under a connected support of μ, uniqueness of the gradient then implies uniqueness of the potential. In discrete settings, the theory of linear programming asserts uniqueness of dual optimizers if the measures are non-degenerate, i.e., if OT cannot be divided into separate sub-problems (Klee & Witzgall, 1968; Hung et al., 1986). The recent work by Staudt et al. (2021) elevates these uniqueness results to probability measures on Polish spaces with disconnected support and additionally outlines uniqueness guarantees on smooth manifolds.

4 Degenerate Limits under Trivial Kantorovich Potentials

An important special case for our CLTs emerges if Kantorovich potentials are not only almost surely unique but also constant. The asymptotic variance in Theorem 3.2 then is equal to zero and the limit law degenerates. For our theory it suffices if Kantorovich potentials are constant in an almost sure sense.

Definition 4.1 (Trivial Kantorovich potentials)**.** A Kantorovich potential $f \in S_c(\mu, \nu)$ is called *trivial* if f is constant μ-almost surely. The set of Kantorovich potentials $S_c(\mu, \nu)$ is said to be *trivial* if all Kantorovich potentials are trivial. Likewise, the same definition applies for conjugated potentials f^c related to ν and the set $S_c^c(\mu, \nu)$.

Simple examples for trivial Kantorovich potentials occur if one of the measures μ or ν is a Dirac measure or if the cost function is itself constant. More interesting examples will be discussed below.

Theorem 4.2 (Degenerate limits)**.** *Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption* ***(A)*** *combined with* ***(B1)*** *or* ***(B2)****. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be probability measures.*

(*i*) *Suppose that $\mathcal{F}_c$ is μ-Donsker. Then, the limit law* (2.1a) *degenerates to a Dirac measure at zero if and only if $S_c(\mu, \nu)$ is trivial.*

(*ii*) *Suppose that $\mathcal{F}_c^c$ is ν-Donsker. Then, the limit law* (2.1b) *degenerates to a Dirac measure at zero if and only if $S_c^c(\mu, \nu)$ is trivial.*

(*iii*) *Suppose that $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are μ- and ν-Donsker, respectively. Then, the limit law* (2.1c) *degenerates to a Dirac measure at zero if and only if $S_c(\mu, \nu)$ and $S_c^c(\mu, \nu)$ are trivial.*

In contrast to almost sure uniqueness of Kantorovich potentials, triviality of $f \in S_c(\mu, \nu)$ does not imply the triviality of f^c, meaning that an almost surely constant f can (and often will) have a c-conjugate f^c that is not almost surely constant (see also Remark 4.7). As it turns out, the existence of constant Kantorovich potentials f is intimately related to the existence of transport plans that act as projections onto the support of μ. To highlight the underlying geometric intuition, we introduce the following notion of projected measures.

Definition 4.3 (Projected measures)**.** For probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the probability measure μ is a *ν-projected measure* (*with respect to* c) denoted as $\mu \in P_c(\nu)$, if there exists a coupling $\pi \in \Pi(\mu, \nu)$ such that all $(x, y) \in \operatorname{supp}(\pi)$ satisfy

$$c(x, y) = \inf_{x' \in \operatorname{supp}(\mu)} c(x', y). \tag{4.1}$$

Analog definitions apply for μ-projected measures $\nu \in P_c(\mu)$.

It is evident that the coupling π in the definition above solves the OT problem in (1.1) between μ and ν. In fact, the proof of Theorem 4.4 shows that an equivalent way to characterize ν-projected measures is the equality

$$\mathrm{OT}_c(\mu, \nu) = \int_{\mathcal{Y}} \inf_{x \in \operatorname{supp}(\mu)} c(x, y) \, \mathrm{d}\nu(y).$$

Intuitively, a ν-projected measure is any measure μ that can be formed by projecting all points of $\operatorname{supp}(\nu)$ to some subset of $\mathcal{X}$ with respect to the cost function c. For example, $\mu \in P_c(\nu)$ always holds if $\mu = p_{\#}\nu$ for a cost projection $p: \mathcal{Y} \to \mathcal{X}$ that satisfies

$$p(y) \in \operatorname*{argmin}_{x \in \operatorname{supp}(\mu)} c(x, y). \tag{4.2}$$

An illustration for such a setting is given in Figure 1(*a*).

Theorem 4.4 (Existence of trivial Kantorovich potentials)**.** *Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption* ***(A)*** *combined with* ***(B1)*** *or* ***(B2)****. Then, for two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ trivial Kantorovich potentials $f \in S_c(\mu, \nu)$ exist if and only if $\mu \in P_c(\nu)$.*

Based on this characterization, we formulate several necessary and sufficient criteria for the existence of trivial Kantorovich potentials.

Corollary 4.5. *Under the assumptions of Theorem 4.4 trivial Kantorovich potentials $f \in S_c(\mu, \nu)$ exist, if one of the following conditions is satisfied.*

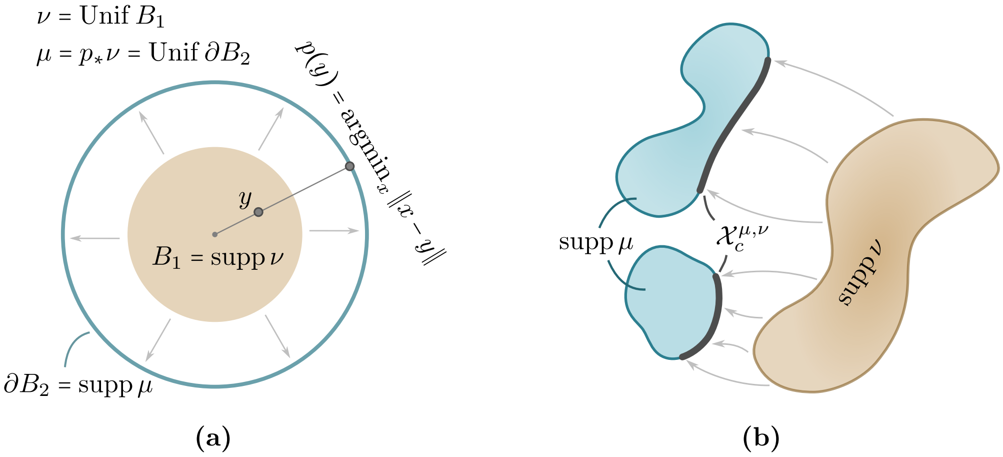

ν = Unif B_1
$\mu = p_*\nu$ = Unif ∂B_2
$p(y) = \text{argmin}_x \|x-y\|$
y
$B_1 = \text{supp}\,\nu$
$\partial B_2 = \text{supp}\,\mu$
(a)
supp μ
$\mathcal{X}_c^{\mu,\nu}$
supp ν
(b)

Figure 1: Trivial Kantorovich potentials. In **(a)**, μ is equal to the projection of ν onto the set $\tilde{\mathcal{X}} = \partial B_2$, the boundary of an Euclidean ball. According to Corollary 4.5 (ii), there exists a Kantorovich potential f that is constant on ∂B_2. Furthermore, due to Staudt et al. (2021, Theorem 2), the constant potential is also unique. In **(b)**, the projection $\mathcal{X}_c^{\mu,\nu}$ of all points $y \in \text{supp}(\nu)$ onto the support of μ under Euclidean costs is marked by black segments that are part of the boundary of $\text{supp}(\mu)$. Since $\text{supp}(\mu)$ is not contained in $\mathcal{X}_c^{\mu,\nu}$, Corollary 4.6 (iii) states that there cannot be a Kantorovich potential f that is constant on $\text{supp}(\mu)$.

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

(ii) $\mu = p_\#\nu$ *for any well-defined c-projection* $p\colon \text{supp}(\nu) \to \tilde{\mathcal{X}}$ *onto a closed set* $\tilde{\mathcal{X}} \subset \mathcal{X}$.

Corollary 4.6. *Under the assumptions of Theorem 4.4 the set* $S_c(\mu,\nu)$ *does not contain trivial Kantorovich potentials if one of the following conditions is satisfied.*

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

(ii) $\text{int}(\text{supp}(\mu)) \notin \text{supp}(\nu)$ *for subsets* $\mathcal{X},\mathcal{Y} \subset V$ *of a normed linear space* $(V, \|\cdot\|)$ *with cost* $c(x,y) = h(\|x-y\|)$ *for strictly increasing* h,

(iii) $\text{supp}(\mu)$ *is not equal to the topological closure* $\mathcal{X}_c^{\mu,\nu} \subset \mathcal{X}$ *of the set*

$$\bigcup_{y \in \text{supp}(\nu)} \underset{x \in \text{supp}(\mu)}{\text{argmin}}\; c(x,y).$$

In each of these settings the limit law (2.1a) *in Theorem 2.2 is non-degenerate.*

Exchanging the roles of μ and ν, Theorem 4.4 as well as Corollaries 4.5 and 4.6 can equally be applied to f^c. Examples illustrating Corollary 4.5 (ii) and Corollary 4.6 (iii) in Euclidean settings are provided in Figure 1. In particular, Figure 1(a) highlights that there is a crucial difference between demanding constant Kantorovich potentials on all of $\mathcal{X}$ (which is not true for $\mathcal{X} = \mathbb{R}^2$ in this case) and only demanding them to be constant almost surely. This setting also serves as an example where $\mu \neq \nu$ and where every Kantorovich potential $f \in S_c(\mu,\nu)$ is trivial while f^c is not.

Remark 4.7 (Bi-triviality)**.** OT problems where both $f \in S_c(\mu,\nu)$ and its conjugate f^c are trivial have a special underlying geometry. If f is constant on $\text{supp}(\mu)$ and f^c is constant on $\text{supp}(\nu)$, then the optimality condition $f(x) + f^c(y) = c(x,y)$ for points (x,y) on the support of any optimal transport plan π implies that c is constant on $\text{supp}(\pi)$, i.e., all points are transported with the same cost. For an example of bi-triviality consider squared Euclidean costs and uniform distributions on centered spheres of positive radius. In particular, for different radii the measures are different and both Kantorovich potentials are trivial.

5 Examples

Elaborating the main result in Theorem 2.2, we focus in this section on concrete settings. The general setup requires Assumption **(A)** combined with **(B1)** or **(B2)** to hold. It then remains to verify the Donsker properties for the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ which then leads to CLTs for empirical OT.

5.1 Empirical Optimal Transport on Countable Discrete Spaces

Our general theory leads to a thorough analysis on CLTs for the OT cost on countable discrete spaces $\mathcal{X}$ and $\mathcal{Y}$ equipped with discrete topology. More precisely, for a non-negative cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ bounded by some constant $\|c\|_\infty < \infty$, the Assumptions **(A)** and **(B2)** hold trivially (recall Section 2). To prove that the function class $\mathcal{F}_c$ is μ-Donsker, we define $\mathcal{F}_c \mathbb{1}_x$ as the restriction of $\mathcal{F}_c$ to a fixed element $x \in \mathcal{X}$. Notably, each element in the latter function class is bounded by $\|c\|_\infty$ and we obtain

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

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

$$\sum_{x \in \mathcal{X}} \sqrt{\mu(x)} < \infty \tag{5.1}$$

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

Corollary 5.1 (Countable discrete spaces, Tameling et al. 2019)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be countable discrete spaces and $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a bounded cost function. Consider probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. If μ fulfills the Borisov-Dudley-Durst condition* (5.1)*, then the CLT in* (2.1a) *is valid. If ν fulfills* (5.1)*, then* (2.1b) *holds. In case both μ and ν fulfill* (5.1)*, then* (2.1c) *holds.*

5.2 Empirical Optimal Transport on Euclidean Spaces

From Theorem 2.2 we immediately derive CLTs for the empirical OT cost between probability measures supported on low dimensional Euclidean spaces $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ with $d \leq 3$ and sufficiently regular cost function. The proofs require the notion of covering and metric entropy for a real-valued function class $\mathcal{F}$ defined on $\mathcal{X}$ as introduced in the notation. Based on metric entropy bounds, empirical process theory provides tools to assess if a given function class is μ-Donsker or even universal Donsker (Van der Vaart & Wellner, 1996, Section 2.5). We first focus on the real line.

Theorem 5.2 ($d = 1$)**.** *Consider the Euclidean space $\mathbb{R}$ with cost $c\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ assumed to be bounded and (α, L)-Hölder for $\alpha \in (1/2, 1]$, i.e.,*

$$\left|c(x,y) - c(x',y')\right| \leq L\left(\left|x - x'\right|^\alpha + \left|y - y'\right|^\alpha\right) \quad \forall x, x', y, y' \in \mathbb{R}. \tag{5.2}$$

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

$$\sum_{k=-\infty}^{\infty} \sqrt{\mu\left([k,k+1)\right)} < \infty, \tag{5.3}$$

then the CLT in (2.1a) *is valid. If* $\nu \in \mathcal{P}(\mathbb{R})$ *fulfills* (5.3)*, then* (2.1b) *holds. In case both* μ *and* ν *fulfill* (5.3)*, then* (2.1c) *holds.*

Proof. By the assumptions imposed on the cost and since the setting focuses on the real line $\mathbb{R}$, the Assumptions **(A)** and **(B2)** hold. Based on the main Theorem 2.2, it remains to prove the Donsker properties of the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ in this case. By Lemma A.2 (i), the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are both contained in the class of real valued (α, L)-Hölder functions defined on $\mathbb{R}$ with $\alpha \in (1/2, 1]$. According to Van der Vaart & Wellner (1996, Example 2.10.25) with $M_k = L$ for all $k \in \mathbb{Z}$ this class is μ-Donsker if (5.3) is fulfilled. □

Example 5.3 (Kantorovich-Rubinstein limit law)**.** For a probability measure $\mu \in \mathcal{P}(\mathbb{R})$ with bounded support and Euclidean cost $c(x,y) = |x-y|$, Theorem 5.2 states that

$$\sqrt{n}\mathrm{OT}_1(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} \sup_{f \in S_1(\mu,\mu)} \mathbb{G}_\mu(f). \tag{5.4}$$

By Kantorovich-Rubinstein duality the set of Kantorovich potentials $S_1(\mu,\mu) = \mathrm{BL}_{\|c\|_\infty}(\mathbb{R})$ is equal to the set of Lipschitz functions uniformly bounded by $\|c\|_\infty$. In particular, the limit law (5.4) does not degenerate (unless of course μ is a Dirac measure). Notably, condition (5.3) is necessary and sufficient for $\mathrm{BL}_{\|c\|_\infty}(\mathbb{R})$ to be μ-Donsker (Giné & Zinn, 1986, Theorem 1). A refined statement, including probability measures with unbounded support, is given by del Barrio et al. (1999) (see also Mason 2016). If F_μ denotes the cumulative distribution function of μ such that

$$\int_{-\infty}^{\infty} \sqrt{F_\mu(t)(1-F_\mu(t))}\,\mathrm{d}t < \infty, \tag{5.5}$$

then, for $\mathbb{B}_\mu(t) = \mathbb{B}(F_\mu(t))$ with $\mathbb{B}(t)$ a standard Brownian bridge, a CLT of the form

$$\sqrt{n}\mathrm{OT}_1(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} \int_{-\infty}^{\infty} |\mathbb{B}_\mu(t)|\,\mathrm{d}t \tag{5.6}$$

is valid. The approach presented here yields a dual perspective on (5.6). Indeed, by suitably coupling the Gaussian processes $\mathbb{G}_\mu$ and $\mathbb{B}_\mu$ and approximating the respective random element in terms of an unsigned measure, an application of Fubini's theorem shows that

$$\int_{-\infty}^{\infty} |\mathbb{B}_\mu(t)|\,\mathrm{d}t \overset{\mathcal{D}}{=} \sup_{f \in \mathrm{Lip}(\mathbb{R})} \mathbb{G}_\mu(f). \tag{5.7}$$

The set $\mathrm{Lip}(\mathbb{R})$ arises due to Kantorovich-Rubinstein duality. Assumption (5.5) is necessary and sufficient for $\mathrm{Lip}(\mathbb{R})$ to be μ-Donsker[3]. In other words, the CLT (5.6) reflects the statistical consequences of Kantorovich-Rubinstein duality with respect to a Lipschitz function class. A similar statement holds for the circle and distributional limit derived thereon (Hundrieser et al., 2021a).

Theorem 2.2 also characterizes the limit law for the empirical OT cost beyond the real line. For Euclidean spaces with dimension $d = 2, 3$ we obtain the following novel results.

[3] Indeed, assumption (5.5) is equivalent to $\sum_{j=1}^{\infty} \sqrt{\mathbb{P}(|X| > j)} < \infty$ for $X \sim \mu$ and well-known to be necessary and sufficient for $\mathrm{Lip}(\mathbb{R})$ to be μ-Donsker (Giné & Zinn, 1986, Theorem 2).

Theorem 5.4 $(d = 2,3)$**.** *Consider the Euclidean space $\mathbb{R}^d$ for $d = 2,3$ with cost $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ assumed to be bounded and L-Lipschitz*[4]*. Further, suppose there exists some $\Lambda > 0$ such that for all $k \in \mathbb{Z}^d$ there exist $x_k, y_k \in [k, k+1)$ such that*

$$\begin{aligned} &c(\cdot, y) - \Lambda \|\cdot - x_k\|_2^2 \text{ is concave on } [k,k+1) \text{ for all } y \in \mathbb{R}^d, \\ &c(x, \cdot) - \Lambda \|\cdot - y_k\|_2^2 \text{ is concave on } [k,k+1) \text{ for all } x \in \mathbb{R}^d. \end{aligned} \tag{5.8}$$

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

$$\sum_{k \in \mathbb{Z}^d} \sqrt{\mu([k,k+1))} < \infty, \tag{5.9}$$

then the CLT in (2.1a) is valid. If $\nu \in \mathcal{P}(\mathbb{R}^d)$ fulfills (5.9), then (2.1b) holds. In case both μ and ν fulfill (5.9), then (2.1c) holds.

Proof. By the assumptions imposed on the cost and since the setting focuses on the Euclidean space $\mathbb{R}^d$ with $d \le 2,3$, the Assumptions **(A)** and **(B2)** hold. Based on the main Theorem 2.2, it remains to prove the Donsker properties of the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ in this case. For a convex bounded set $\Omega \subseteq \mathbb{R}^d$ and constants $K, L > 0$, let $BL_{K,L}^{\text{conc}}(\Omega)$ be the class of concave L-Lipschitz functions which are absolutely bounded by K. In order to prove that $\mathcal{F}_c$ is μ-Donsker if (5.9) holds, we employ Van der Vaart & Wellner (1996, Theorem 2.10.24). To this end, consider a partition $\mathbb{R}^d = \bigcup_{k \in \mathbb{Z}^d} [k, k+1)$ and define the function class $\mathcal{F}_{c,k} := \mathcal{F}_c \mathbb{1}_{[k,k+1)}$. We first verify that each class $\mathcal{F}_{c,k}$ is μ-Donsker. First note for $k \in \mathbb{Z}^d$ and any element $f \in \mathcal{F}_c$ that $f(\cdot) - \Lambda\|\cdot - x_k\|_2^2$ is bounded, Lipschitz and concave on $[k, k+1)$ (see Lemma A.2 (ii)). More precisely, for any $f \in \mathcal{F}_c$ and since $x_k \in [k, k+1)$ it follows

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

with $\kappa := (\|c\|_\infty + \Lambda d)$ and $l := (L + 2\Lambda d)$. According to Bronshtein (1976, Theorem 6)[5], we conclude for $\varepsilon > 0$ sufficiently small

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

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

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

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

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

[4] In particular, the cost satisfies (5.2) for $\alpha = 1$ and Euclidean norm.

[5] Bronshtein (1976) provides metric entropy bounds for *convex* bounded Lipschitz functions on a *cube* but of course they remain valid for concave functions.

After taking square roots the latter bound is integrable around zero for $d = 2, 3$ which yields the μ-Donsker property for $\mathcal{F}_{c,k}$ (Van Der Vaart, 1996, Theorem 2.5.2). The μ-Donsker property of the whole $\mathcal{F}_c$ now follows if

$$\sup_{n \in \mathbb{N}} \sum_{k \in \mathbb{Z}^d} \mathbb{E}\left[\left\| \sqrt{n}(\hat{\mu}_n - \mu) \right\|_{\mathcal{F}_{c,k}} \right] < \infty. \tag{5.10}$$

By standard chaining arguments each individual summand can be bounded by

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

Herein, the supremum runs over all finitely supported probability measures over $\mathbb{R}^d$ which leads to the claimed upper bound by our previous arguments. Summing over $k \in \mathbb{Z}^d$ yields (5.10) provided μ fulfills the summability condition (5.9). □

The summability constraints (5.3) and (5.9) are reminiscent to the Borisov-Dudley-Durst condition (5.1). Indeed, they naturally appear by partitioning the Euclidean space $\mathbb{R}^d = \bigcup_{k \in \mathbb{Z}^d} [k, k+1)$ and controlling the empirical process indexed over the respective function class restricted to individual partitions (Van der Vaart & Wellner, 1996, Theorem 2.10.24). For the latter, the proof of Theorem 5.4 exploits well-known metric entropy bounds for the class of α-Hölder and concave, Lipschitz functions, respectively. Notably, crucial to CLTs for dimension $d = 2, 3$ is condition (5.8) that enables suitable upper bounds for the metric entropy of $\mathcal{F}_c$ and $\mathcal{F}_c^c$.

Remark 5.5 (On the assumptions)**.** A few words regarding the required assumptions for our CLTs are in order.

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

(*ii*) Condition (5.8) is fulfilled if the cost function is twice continuously differentiable in both components with a uniform bound $K > 0$ on the Eigenvalues of its Hessian. In this setting the bound from (5.8) is valid for $\Lambda = K/2$.

(*iii*) The summability constraints (5.3) and (5.9) are well-known in the context of empirical process theory (Van der Vaart & Wellner, 1996, Section 2.10.4). A sufficient condition is given in terms of moments $\mathbb{E}\left[\|X\|_{\infty}^{2d+\delta} \right] < \infty$ for some $\delta > 0$ since

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

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

Example 5.6 ($\mathbb{R}^d$, $d \leq 3$, $p \geq 2$)**.** On Euclidean spaces $\mathbb{R}^d$ with $d \leq 3$ and any $p \geq 2$, the CLTs under the null $\mu = \nu$ for probability measures with bounded support lead to a degenerate limit law

$$\sqrt{n} \mathrm{OT}_p(\hat{\mu}_n, \mu) \xrightarrow{\mathcal{D}} 0, \tag{5.11}$$

whenever the support of μ is equal the closure of some connected open set. Indeed, under these conditions the set $S_p(\mu,\mu)$ is trivial (see discussion at the end of Section 3 and Corollary 4.5) which asserts a degenerate limit law. For $d=1$ and $p=2$, it is known that the fluctuation of $\mathrm{OT}_2(\hat{\mu}_n,\mu)$ is of smaller order than $n^{-1/2}$. Then, it holds by del Barrio et al. (2005) under additional regularity assumptions including μ having connected support that

$$n\mathrm{OT}_2(\hat{\mu}_n,\mu) \xrightarrow{\mathcal{D}} \int_0^1 \mathbb{B}_\mu^2(u)\,\mathrm{d}u, \tag{5.12}$$

where $\mathbb{B}_\mu = \mathbb{B}(u)/f_\mu(F_\mu^{-1}(u))$ with $\mathbb{B}$ a standard Brownian bridge, f_μ the density and F_μ^{-1} the quantile function of μ, respectively. If instead the support of μ is disconnected, then non-trivial Kantorovich potentials exist (see Staudt et al. 2021, Lemma 11) and hence for $p\geq 2$ the distributional limits for $\sqrt{n}\mathrm{OT}_p(\hat{\mu}_n,\mu)$ do not degenerate. Moreover, under the alternative $\mu\neq\nu\in\mathcal{P}(\mathbb{R}^d)$ with bounded support, non-degenerate limit laws

$$\sqrt{n}\left(\mathrm{OT}_p(\hat{\mu}_n,\nu)-\mathrm{OT}_p(\mu,\nu)\right) \xrightarrow{\mathcal{D}} \sup_{f\in S_p(\mu,\nu)} \mathbb{G}_\mu(f) \tag{5.13}$$

for $p\geq 2$ are generic rather than the exception. While uniqueness of Kantorovich potentials often holds (see Section 3), their triviality is linked to the underlying geometry of the corresponding measures' supports (Theorem 4.4) and appears limited to exotic settings.

Example 5.7 ($d\geq 4$)**.** Beyond the low dimensional setting $d\leq 3$ as covered by our unifying approach, already for $d=4$ and uniform measure supported on different unit balls around $x\neq y\in\mathbb{R}^d$ Chizat et al. (2020); Manole & Niles-Weed (2021) prove that

$$n^{-1/2} \leq \mathbb{E}\left[\mathrm{OT}_2(\hat{\mu}_n,\nu)\right]-\mathrm{OT}_2(\mu,\nu) \leq n^{-1/2}\log(n). \tag{5.14}$$

Combined with the CLT by del Barrio & Loubes (2019) in (1.5) this indicates either the random sequence $\sqrt{n}(\mathrm{OT}_2(\hat{\mu}_n,\nu)-\mathrm{OT}_2(\mu,\nu))$ to be not tight or at best to asymptotically follow a non-centered limit law. For $d\geq 5$, the asymptotic fluctuation of $\mathrm{OT}_2(\hat{\mu},\nu)-\mathrm{OT}_2(\mu,\nu)$ around zero is generally of different order than $n^{-1/2}$ and we refer to the discussion in the introduction for details. An exception occurs for different probability measures when one is supported on a low dimensional space as detailed in the following subsection.

5.3 Empirical Optimal Transport under Lower Complexity Adaptation

We highlight our CLTs for empirical OT in the realm of the *lower complexity adaptation*. This phenomenon, recently discovered by Hundrieser et al. (2021b), states that statistical rates to estimate the empirical OT cost between two probability measures is typically driven by the less complex measure, e.g., the one with lower dimensional support. In light of this principle, we emphasize that our CLTs extend beyond the low-dimensional Euclidean case provided that at least one measure has some low-dimensional compact support. The main result relies on the observation that the uniform metric entropies for $\mathcal{F}_c$ and $\mathcal{F}_c^c$ coincide in such settings (Hundrieser et al., 2021b, Theorem 3.4). Hence, under suitable bounds on the uniform metric entropy for only one of the function classes $\mathcal{F}_c$ or $\mathcal{F}_c^c$, it follows that *both* are universal Donsker (Definition 2.1). The following result makes use of this observation and is specifically tailored to settings where the complexity of Kantorovich potentials on $\mathcal{X}$ can be suitably controlled such that a universal Donsker property holds.

Theorem 5.8. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and suppose the cost $c\colon\mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$ is bounded and continuous. Further, assume one of the following three settings.*

(i) *The set* $\mathcal{X} = \{x_1, \ldots, x_K\}$ *is finite and equipped with discrete topology.*

(ii) *The set* $\mathcal{X} \subseteq \mathbb{R}$ *is a compact interval and there exist* $\alpha \in (1/2, 1]$ *and* $L > 0$ *such that* $c(\cdot, y)$ *is* (α, L)*-Hölder for all* $y \in \mathcal{Y}$*.*

(iii) *The set* $\mathcal{X} \subseteq \mathbb{R}^d$ *for* $d = 2, 3$ *is compact and convex, and there exist* $L > 0$ *and* $\Lambda > 0$ *such that* $c(\cdot, y)$ *is* L*-Lipschitz and* $c(\cdot, y) - \Lambda \|\cdot\|_2^2$ *is concave for all* $y \in \mathcal{Y}$*.*

Then, for arbitrary probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ *the weak limits from* (2.1) *hold.*

Proof. For the proof of Theorem 5.8, we make use of the following relationship between the covering numbers of the function class $\mathcal{F}_c$ in (1.7) and its c-conjugate $\mathcal{F}_c^c$.

Lemma 5.9 (Hundrieser et al. 2021b, Theorem 2.4)**.** *Let* $\mathcal{X}, \mathcal{Y}$ *be Polish spaces and consider a bounded cost function* $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$*. Suppose that* $\mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) < \infty$ *for* $\varepsilon > 0$ *sufficiently small. Then, it holds that*

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

Continuing with the proof of Theorem 5.8, Assumptions **(A)** and **(B1)** are fulfilled for all three settings. For setting (i), note that $\mathcal{F}_c \subseteq [-\|c\|_\infty, \|c\|_\infty]^{\mathcal{X}}$, which implies for all $\varepsilon > 0$ that

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty) \leq \log \mathcal{N}\left(\varepsilon, [-\|c\|_\infty, \|c\|_\infty]^{\mathcal{X}}, \|\cdot\|_\infty\right) \lesssim_{|\mathcal{X}|, \|c\|_\infty} \log\left(\lceil \varepsilon^{-1} \rceil\right).$$

For setting (ii), Lemma A.2 (i) asserts that any $f \in \mathcal{F}_c$ is also (α, L)-Hölder on $\mathcal{X}$. Hence, it follows by Van der Vaart & Wellner (1996, Theorem 2.7.1) for sufficiently small $\varepsilon > 0$ that

$$\log\left(\mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty)\right) \lesssim_{d, \alpha, L, \mathcal{X}, \|c\|_\infty} \varepsilon^{-d/\alpha}.$$

For setting (iii), note by Lemma A.2 (ii) that $\mathcal{F}_c - \Lambda\|\cdot\|_2^2 \subseteq BL_{\kappa,l}^{\text{conc}}(\mathcal{X})$ for $\kappa = (\|c\|_\infty + \Lambda d)$ and $l = (L + 2\Lambda d)$. Hence, invoking the upper bound by Bronshtein (1976, Theorem 6)[6] yields for sufficiently small $\varepsilon > 0$ that

$$\begin{aligned} \log\left(\mathcal{N}(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty)\right) &= \log\left(\mathcal{N}(\varepsilon, \mathcal{F}_c - \Lambda\|\cdot\|_2^2, \|\cdot\|_\infty)\right) \\ &\leq \log\left(\mathcal{N}(\varepsilon, BL_{\kappa,l}^{\text{conc}}(\mathcal{X}), \|\cdot\|_\infty)\right) \lesssim_{d, \Lambda, L, \mathcal{X}, \|c\|_\infty} \varepsilon^{-d/2}. \end{aligned} \tag{5.15}$$

By Lemma 5.9 identical bounds on the uniform metric entropy of $\mathcal{F}_c^c$ hold for all three settings. Since for all settings the square root of the uniform metric entropy is integrable with respect to ε around zero, it follows by Van der Vaart & Wellner (1996, Theorem 2.5.2) that both $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are universal Donsker. We thus conclude from Theorem 2.2 the CLTs for empirical OT (2.1) for arbitrary probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. □

We emphasize that in contrast to previous CLTs from Subsections 5.1 and 5.2, no summability conditions are necessary for μ and ν. Furthermore, no additional assumptions are imposed on the Polish space $\mathcal{Y}$. For instance, $\mathcal{Y}$ can be equal to a high- or even infinite-dimensional Polish vector space.

[6] The uniform metric entropy bound by Bronshtein (1976) is actually formulated only for convex Lipschitz functions defined on a cube. Nevertheless, since any convex, Lipschitz function on a bounded convex domain set can be extended to a convex function with identical Lipschitz modulus on a cube containing the initial domain (Yan, 2014, Theorem 4.1), the last inequality in (5.15) remains valid at the cost of a possibly larger constant independent of ε.

Remark 5.10 (Semi-discrete OT)**.** Statement (i) in Theorem 5.8 covers CLTs for semi-discrete OT (Aurenhammer et al., 1998; Mérigot, 2011; Hartmann & Schuhmacher, 2020), where one of the probability measures is assumed to be finitely supported. In particular, for Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, consider a finitely supported probability measure $\mu = \sum_{i=1}^{K} \mu_i \delta_{x_i} \in \mathcal{P}(\mathcal{X})$ and an arbitrary probability measure $\nu \in \mathcal{P}(\mathcal{Y})$. For a possibly unbounded continuous cost function, del Barrio et al. (2021a) recently derived a CLT for the one-sample case when the discrete measure μ is replaced by its empirical counterpart $\hat{\mu}_n$. Our theory complements their result insofar that we require the cost function to be bounded, but additionally allow sampling from the general measure ν and still obtain the results in Theorem 2.2. In particular, when the support of ν is connected, then Kantorovich potentials are almost surely unique (Staudt et al., 2021, Example 3). Hence, in this setting the corresponding weak limit is always centered normal, although the limit variance may degenerate to zero (recall Section 4). Even if the support of ν is not connected, under a suitable non-degeneracy condition of the optimal transport plan Kantorovich potentials still can be shown to be unique (Staudt et al., 2021, Theorem 1).

Remark 5.11 (CLTs for Wasserstein distance in high-dimensional spaces)**.** As a consequence of Theorem 5.8, we obtain CLTs for the empirical Wasserstein distance on $\mathbb{R}^d$ even beyond $d \leq 3$ as long as both probability measures μ and ν have bounded support and one of them exhibits a sufficiently low dimensional support. More precisely, if μ is supported on a finite set of points or a bounded line in $\mathbb{R}^d$ with $p \geq 1$ or in case it is supported on an affine sub-vector space of dimension $d' \leq 3$ with $p \geq 2$, our asymptotic results from (2.1) remain valid. Notably, for the latter case the asymptotic results still hold for $p \in [1, 2)$ if $\text{supp}(\nu)$ is disjoint from the convex hull of $\text{supp}(\mu)$ since the cost function $\| \cdot - y\|^p$, then is semi-concave on the latter set.

Remark 5.12 (Degeneracy of limit laws)**.** We like to comment on degeneracy of our limit laws in case of Euclidean spaces $\mathbb{R}^d$ with $d > 3$ and costs of the form $c(x, y) = h(\|x - y\|)$ for h strictly increasing. Indeed, for probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with bounded support, where the support of ν has non-empty interior while μ is concentrated on a sufficiently low-dimensional affine subspace, it follows by Corollary 4.6 (ii) that $S_c^c(\mu, \nu)$ cannot be trivial. Hence, when replacing the measure ν by its empirical measure $\hat{\nu}_m$ the limit laws do not degenerate. When instead sampling from μ, the limit law could indeed degenerate although this is rather the exception than the rule (see Theorem 4.4). We recall Figure 1(a), for an example where all Kantorovich potentials $S_c(\mu, \nu)$ are trivial, whereas $S_c^c(\mu, \nu)$ is not.

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

References

Ajtai, M., Komlós, J., & Tusnády, G. (1984). On optimal matchings. *Combinatorica*, 4(4), 259–264.

Aurenhammer, F., Hoffmann, F., & Aronov, B. (1998). Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1), 61–76.

Bernton, E., Ghosal, P., & Nutz, M. (2021). Entropic optimal transport: Geometry and large deviations. *preprint arXiv:2102.04397.*

Berthet, P. & Fort, J.-C. (2019). Weak convergence of empirical Wasserstein type distances. *preprint arXiv:1911.02389.*

Berthet, P., Fort, J.-C., & Klein, T. (2020). A central limit theorem for Wasserstein type distances between two distinct univariate distributions. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(2), 954 – 982.

Bobkov, S. & Ledoux, M. (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society.

Bobkov, S. G. & Ledoux, M. (2021). A simple fourier analytic proof of the AKT optimal matching theorem. *The Annals of Applied Probability*, 31(6), 2567–2584.

Boissard, E. & Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'Institut Henri Poincaré, Probabilités et statistiques*, volume 50, pages 539–563.

Bonneel, N., Van De Panne, M., Paris, S., & Heidrich, W. (2011). Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12.

Bronshtein, E. M. (1976). ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17(3), 393–398.

Cárcamo, J., Cuevas, A., & Rodríguez, L.-A. (2020). Directional differentiability for supremum-type functionals: Statistical applications. *Bernoulli*, 26(3), 2143 – 2175.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., & Peyré, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. In H. Larochelle, M. Ranzato, & others (Eds.), *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269.: Curran Associates, Inc.

Csörgö, M. & Horváth, L. (1993). *Weighted approximations in probability and statistics.* John Wiley & Sons.

del Barrio, E., Giné, E., & Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27(2), 1009–1071.

del Barrio, E., Giné, E., & Utzet, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1), 131–189.

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

del Barrio, E., González-Sanz, A., & Loubes, J.-M. (2021b). Central limit theorems for general transportation costs. *preprint arXiv:2102.06379.*

del Barrio, E., Gordaliza, P., & Loubes, J.-M. (2019). A central limit theorem for L_p transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8(4), 817–849.

del Barrio, E. & Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2), 926–951.

Dobrić, V. & Yukich, J. E. (1995). Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1), 97–118.

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

Dudley, R. M. (2014). *Uniform central limit theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95(1), 125–140.

Fang, Z. & Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1), 377–412.

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

Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.

Gangbo, W. & McCann, R. J. (1996). The geometry of optimal transportation. *Acta Mathematica*, 177(2), 113–161.

Giné, E. & Zinn, J. (1986). Empirical processes indexed by Lipschitz functions. *The Annals of Probability*, 14(4), 1329–1338.

Hartmann, V. & Schuhmacher, D. (2020). Semi-discrete optimal transport: a solution procedure for the unsquared Euclidean distance case. *Mathematical Methods of Operations Research*, 92(1).

Hundrieser, S., Klatt, M., & Munk, A. (2021a). The statistics of circular optimal transport. In A. SenGupta & B. Arnold (Eds.), *Directional Statistics for Innovative Applications*: Springer. To appear [preprint arXiv:2103.15426].

Hundrieser, S., Staudt, T., & Munk, A. (2021b). Empirical optimal transport between different meaures adapts to lower complexity. *In preparation.*

Hung, M. S., Rom, W. O., & Waren, A. D. (1986). Degeneracy in transportation problems. *Discrete applied mathematics*, 13(2-3), 223–237.

Klee, V. & Witzgall, C. (1968). Facets and vertices of transportation polytopes. *Mathematics of the decision sciences*, 1, 257–282.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., & Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4), 43–59.

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

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

Mason, D. M. (2016). A weighted approximation approach to the study of the empirical Wasserstein distance. In *High Dimensional Probability VII*, pages 137–154. Springer.

McCoy, R. & Ntantu, I. (1988). *Topological properties of spaces of continuous functions.* Berlin New York: Springer-Verlag.

Mérigot, Q. (2011). A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592.: Wiley Online Library.

Munk, A. & Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1), 223–241.

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

Peyré, G. & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6), 355–607.

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

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

Römisch, W. (2004). Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*, pages 1575–1583. New York: Wiley.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians: Calculus of variations, PDEs, and modeling.* Progress in Nonlinear Differential Equations and Their Applications. Springer.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4), 928–943.

Sommerfeld, M. & Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 219–238.

Staudt, T., Hundrieser, S., & Munk, A. (2021). On the uniqueness of Kantorovich potentials. *preprint arXiv:2201.08316.*

Tameling, C., Sommerfeld, M., & Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5), 2744–2781.

Tameling, C., Stoldt, S., Stephan, T., Naas, J., Jakobs, S., & Munk, A. (2021). Colocalization for super-resolution microscopy via optimal transport. *Nature Computational Science*, 1, 199–211.

Van Der Vaart, A. (1996). New donsker classes. *The Annals of Probability*, 24(4), 2128–2140.

Van der Vaart, A. & Wellner, J. (1996). *Weak convergence and empirical processes: With applications to statistics.* Springer Series in Statistics. Springer.

Villani, C. (2003). *Topics in optimal transportation.* Graduate Studies in Mathematics. American Mathematical Society.

Villani, C. (2008). *Optimal transport: old and new.* A Series of Comprehensive Studies in Mathematics. Springer.

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

Yan, M. (2014). Extension of convex function. *Journal of Convex Analysis*, 21(4), 965–987.

A Omitted Proofs and Auxiliary Results

A.1 Proofs for Section 4

Proof of Theorem 4.4. Let f be an almost surely constant Kantorovich potential. By Lemma A.1 the Kantorovich potential f is continuous and since Kantorovich potentials are conserved under constant shifts, we can without loss of generality assume that $f \equiv 0$ on $\operatorname{supp}(\mu)$. According to Villani (2008, Theorem 5.19), we find that f coincides on $\operatorname{supp}(\mu)$ with a Kantorovich potential $\tilde{f} \colon \operatorname{supp}(\mu) \to \mathbb{R}$ for the same transport problem but with the space $\mathcal{X}$ replaced by $\operatorname{supp}(\mu)$. In particular, $\tilde{f} \equiv 0$ is the zero function and $\tilde{f}^c(y) = \inf_{x \in \operatorname{supp}(\mu)} c(x,y)$ for $y \in \mathcal{Y}$ is its c-conjugate (which is measurable since c is continuous). Again by Villani (2008, Theorem 5.19), we conclude that

$$\mathrm{OT}_c(\mu,\nu) = \int_{\mathcal{X}} f \,\mathrm{d}\mu + \int_{\mathcal{Y}} f^c \,\mathrm{d}\nu = \int_{\operatorname{supp}(\mu)} \tilde{f} \,\mathrm{d}\mu + \int_{\mathcal{Y}} \tilde{f}^c \,\mathrm{d}\nu = \int_{\mathcal{Y}} \inf_{x \in \operatorname{supp}(\mu)} c(x,y) \,\mathrm{d}\nu(y).$$

Conversely, assume that condition (4.1) holds. We first show that the set

$$A := \bigcup_{y \in \mathcal{Y}} \operatorname*{argmin}_{x \in \operatorname{supp}(\mu)} c(x,y) = \left\{ x \in \operatorname{supp}(\mu) \,\middle|\, \exists\, y_x \in \mathcal{Y} \text{ such that } c(x,y_x) = \inf_{\tilde{x} \in \operatorname{supp}(\mu)} c(\tilde{x},y_x) \right\}$$

is dense in $\operatorname{supp}(\mu)$. Suppose the contrary, then there exists an $x_0 \in \operatorname{supp}(\mu)$ and a neighbourhood $U \subset \operatorname{supp}(\mu)$ of x_0 such that for all $(x,y) \in U \times \mathcal{Y}$ we have $c(x,y) > \inf_{\tilde{x} \in \operatorname{supp}(\mu)} c(\tilde{x},y)$. Since $\mu(U) > 0$ as a neighbourhood of the support point x_0, it follows that

$$\int_{U \times \mathcal{Y}} c(x,y) \,\mathrm{d}\pi(x,y) > \int_{U \times \mathcal{Y}} \inf_{x \in \operatorname{supp}(\mu)} c(x,y) \,\mathrm{d}\nu(y)$$

for any optimal transport plan π, which can be used to contradict (4.1). Next, we define the potential $\tilde{f} \colon \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ via

$$\tilde{f}(x) = \begin{cases} 0, & \text{if } x \in \operatorname{supp}(\mu), \\ -\infty, & \text{else.} \end{cases}$$

We note that $\tilde{f}^c(y) = \inf_{x \in \mathcal{X}} c(x,y) - \tilde{f}(x) = \inf_{x \in \operatorname{supp}(\mu)} c(x,y)$ and thus find

$$\int_{\mathcal{X}} \tilde{f} \,\mathrm{d}\mu + \int_{\mathcal{Y}} \tilde{f}^c \,\mathrm{d}\nu = \int \inf_{x \in \operatorname{supp}(\mu)} c(x,y) \,\mathrm{d}\nu = \mathrm{OT}_c(\mu,\nu)$$

by condition (4.1). Therefore, $\tilde{f}$ is an optimal dual solution. It remains to show that the corresponding Kantorovich potential f, defined as $\tilde{f}^{cc}$, is indeed constant on the support of μ. For $x \in A$, we observe

$$f(x) = \inf_{y \in \mathcal{Y}} \left(c(x,y) - \inf_{\tilde{x} \in \operatorname{supp}(\mu)} c(\tilde{x},y) \right) \le c(x,y_x) - \inf_{\tilde{x} \in \operatorname{supp}(\mu)} c(\tilde{x},y_x) = 0,$$

where y_x is chosen as in the definition of A. From the first equality, we also note that $f(x) \geq 0$, thus $f = 0$ on A. Since f is c-conjugate and therefore continuous, the density of the set $A \subset \operatorname{supp}(\mu)$ implies $f = 0$ on the full support. □

Proof of Corollary 4.5. By Lemma A.1 all Kantorovich potentials are continuous so that statements holding μ- or ν-almost surely, in fact, hold for each point in the respective supports. Under condition (i), we find $\mathrm{OT}_c(\mu,\nu) = 0$ and $\operatorname{argmin}_{x \in \operatorname{supp}(\mu)} c(x,y) = 0$ for each $y \in \operatorname{supp}(\nu)$, such that (4.1) holds. If condition (ii) holds, then $\pi = (p, \mathrm{id})_* \nu \in \Pi(\mu,\nu)$ is a transport plan. Due to the projection property, it must also be optimal and (4.1) follows. □

Proof of Corollary 4.6. If c is a positive definite cost and $\mu \neq \nu$, we find that $\mathrm{OT}_c(\mu,\nu) > 0$. At the same time, $\operatorname{supp}(\nu) \subset \operatorname{supp}(\mu)$ implies that the right hand side of (4.1) vanishes, so there cannot exist constant potentials under condition (i). To show that (ii) also leads to non-degeneracy, let $U := \operatorname{int}(\operatorname{supp}(\mu)) \setminus \operatorname{supp}(\nu)$, which is non-empty by assumption and open since $\operatorname{supp}(\nu)$ is closed. For any coupling $\pi \in \Pi(\mu,\nu)$, it holds that $\pi(U \times \operatorname{supp}(\nu)) = \mu(U) > 0$ (since U contains points in the support of μ), and so we always find $x \in U$ and $y \in \operatorname{supp}(\nu)$ such that $(x,y) \in \operatorname{supp}(\pi)$. Let $u = x - y$. As U is open, we find $\varepsilon > 0$ small enough that $x' := x - \varepsilon\, u \in U \subset \operatorname{supp}(\nu)$ as well. Then, employing the strict monotonicity of h,

$$c(x',y) = h(\|x'-y\|) = h\big((1-\varepsilon)\|x-y\|\big) < h(\|x-y\|) = c(x,y) \tag{A.1}$$

for $(x,y) \in \operatorname{supp}(\pi)$ and $x' \in \operatorname{supp}(\mu)$, so $\mu \notin P_c(\nu)$ follows by Definition 4.3. Applying Theorem 4.4 yields the claim. Finally, the fact that $\operatorname{supp}(\mu)$ equals the set $\mathcal{X}_c^{\mu,\nu}$ if (4.1) holds has been established in the proof of Proposition 4.4, which proves (iii). □

A.2 Regularity of Kantorovich Potentials

Assumptions imposed on the cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ translate to properties on the function class $\mathcal{F}_c$ and its c-conjugate $\mathcal{F}_c^c$. A trivial observation is boundedness for any $f \in \mathcal{F}_c$ as long as the cost is non-negative and bounded. Indeed, for any $f \in \mathcal{F}_c$ it holds that

$$-\|c\|_\infty \leq \inf_{y \in \mathcal{Y}} c(x,y) - \|c\|_\infty \leq f(x) \leq \inf_{y \in \mathcal{Y}} c(x,y) \leq \|c\|_\infty$$

and analogously for any $f \in \mathcal{F}_c^c$. We summarize additional findings in this regard in the following two statements (see also Gangbo & McCann (1996); Villani (2008); Santambrogio (2015); Staudt et al. (2021) for further details).

Lemma A.1 (Continuity of Kantorovich potentials)**.** *Consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and a cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying Assumption* **(A)** *combined with* **(B2)** *or* **(B2)***. Then, Kantorovich potentials $S_c(\mu,\nu)$ and $S_c^c(\mu,\nu)$ are continuous on $\mathcal{X}$ and $\mathcal{Y}$, respectively.*

Proof. Existence of Kantorovich potentials is guaranteed by Assumption **(A)**. Note that equicontinuity of the partially evaluated costs implies continuity of Kantorovich potentials since the modulus of continuity of a function class is preserved under pointwise infima or suprema (see Santambrogio 2015, Section 1.2 for details). This implies continuity of Kantorovich potentials on $\mathcal{X}$ under Assumption **(B1)**, whereas under Assumption **(B2)** continuity holds on both $\mathcal{X}$ and $\mathcal{Y}$. Moreover, under Assumption **(B1)** the space $\mathcal{X}$ is compact and hence by continuity of the costs, Staudt et al. (2021, Lemma 2) asserts that the Kantorovich potentials on $\mathcal{Y}$ are also continuous on the support of ν. □

As particular instances where structural properties of the cost function are inherited to Kantorovich potentials, we focus on the setting of Hölder smoothness and semi-concavity.

Lemma A.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces.*

(i) If for some $\alpha \in (0,1]$ and $L > 0$ the cost $c(\cdot, y)$ is (α, L)-Hölder continuous as a function in x for all $y \in \mathcal{Y}$, then any $f \in \mathcal{F}_c$ is (α, L)-Hölder continuous.

(ii) If for some $\lambda > 0$ the cost $c(\cdot, y)$ is λ-semi-concave as a function in x for all $y \in \mathcal{Y}$, then any $f \in \mathcal{F}_c$ is λ-semi-concave.

Note that reversing the roles of x and y leads to analogous results for the c-conjugate function class $\mathcal{F}_c^c$ if $\{c(x, \cdot) \mid x \in \mathcal{X}\}$ is uniformly Hölder smooth or semi-concave.

Proof. Suppose $c(\cdot, y)$ is (α, L)-Hölder continuous for any $y \in \mathcal{Y}$. Then, it follows that

$$f(x) = \inf_{y' \in \mathcal{Y}} c(x, y') - g(y') \le c(x, y) - g(y) \le c(x', y) - g(y) + L\|x - x'\|^\alpha.$$

Taking the infimum on the right hand side with respect to y yields

$$f(x) \le f(x') + L\|x - x'\|^\alpha.$$

Changing the roles of x and x' proves that $|f(x) - f(x')| \le L\|x - x\|^\alpha$ for any $f \in \mathcal{F}_c$. Suppose that $c(\cdot, y)$ is λ-semi-concave, i.e., $c(\cdot, y) - \lambda\| \cdot \|^2$ is concave for all $y \in \mathcal{Y}$ as a function of x. It then follows that

$$\begin{aligned} & f(tx + (1-t)x') - \lambda\|tx + (1-t)x'\|^2 \\ & \ge t\left(\inf_{y \in \mathcal{Y}} c(x, y) - \lambda\|x\|^2 - g(y)\right) + (1-t)\left(\inf_{y \in \mathcal{Y}} c(x', y) - \lambda\|x'\|^2 - g(y)\right) \\ & = t\left(f(x) - \lambda\|x\|^2\right) + (1-t)\left(f(x') - \lambda\|x'\|^2\right). \end{aligned}$$

Hence, any $f \in \mathcal{F}_c$ is itself a λ-semi-concave function. □