Annals of Operations Research (2022) 315:251–278
https://doi.org/10.1007/s10479-022-04698-0

ORIGINAL RESEARCH

Limit laws for empirical optimal solutions in random linear programs

Marcel Klatt[1] · **Axel Munk**[1,2] · **Yoav Zemel**[3]

Accepted: 22 March 2022 / Published online: 30 April 2022

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 distributional limit laws for empirical optimal transport plans follow 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

Mathematics Subject Classification 90C05 · 90C15 · 90C31 · 62E20 · 49N15

✉ Yoav Zemel
zemel@statslab.cam.ac.uk

Marcel Klatt
mklatt@mathematik.uni-goettingen.de

Axel Munk
munk@math.uni-goettingen.de

[1] Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany

[2] Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen, Germany

[3] Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, England

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 Dantzig, 1963; Bertsimas & Tsitsiklis, 1997; Luenberger & Ye, 2008; 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

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

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 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 and 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 (Tintner, 1960; Prékopa, 1966; Wets, 1980).

In this paper, we take a related route and focus on statistical aspects of 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

[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.

[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 $\mathbb{R}^m$.

the empirical optimal solution

$$x^\star(b_n) \in \arg\min_{Ax=b_n,\, x\geq 0} c^T x \tag{1}$$

around its population quantities after proper standardization. For the sake of exposition, suppose that $x^\star(b)$ in (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

$$r_n\left(x^\star(b_n) - x^\star(b)\right) \xrightarrow{D} M(G), \tag{2}$$

where $M : \mathbb{R}^m \to \mathbb{R}^d$ is given in Theorem 3.1. The function M in (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 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 (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), Shapiro (1993), Shapiro (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 $v(b)$ 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 $v(b)$ 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

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

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 (2).

Our proof technique allows to recover some related known results for random linear programs (see Sect. 3). These include convergence of the optimality sets in Hausdorff distance (Proposition 3.7), and a limit law for the optimal value

$$r_n(v(b_n) - v(b)) \xrightarrow{D} \max_{\substack{\lambda(I) \text{ dual optimal} \\ \text{basic solution for } (D_b)}} G^T \lambda(I), \tag{3}$$

as n tends to infinity. Indeed, (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 to conclude 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 (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 Sect. 2 also introducing deterministic and stochastic assumptions for our general theory. Our main results are summarized in Sect. 3, followed by their proofs in Sect. 4. The assumptions are discussed in more detail in Sect. 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 $[d] := \{1, \ldots, d\}$. 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*

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

respectively. Herein, and in order to match dimensions (a solution for (P_b) has dimension d instead of $m \leq d$) the linear operator $\mathrm{Aug}_I : \mathbb{R}^m \to \mathbb{R}^d$ 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

$$\mathcal{P}(b) := \left\{x \in \mathbb{R}^d \mid Ax = b, x \geq 0\right\}, \quad \mathcal{P}^\star(b) := \left\{x^\star \in \mathcal{P}(b) \mid c^T x^\star = \inf_{x\in\mathcal{P}(b)} c^T x\right\}, \tag{4}$$

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:

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

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

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*

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

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

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

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 $b = (1, 0, 0, \ldots, 0) \in \mathbb{R}^m$. 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) and (D_b), respectively. We allow the more general case that only the first $m_0 \in \{0, \ldots, m\}$ coordinates of b are unknown[4] and

[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 Sect. 6.

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

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

with $G^{m_0} \in \mathbb{R}^{m_0}$ *absolutely continuous*,

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

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

Further discussions on the assumptions are deferred to Sect. 5.

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 $[N] := \{1, \ldots, N\}$ 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* $H_1^{m_0}, \ldots, H_K^{m_0} \subseteq \mathbb{R}^{m_0}$ *and random vectors* $Y_n \in \mathcal{P}^\star(b)$ *such that*

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

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

$$\widetilde{H}_k = \{b \in \mathbb{R}^m : I_k \text{ is an optimal basis for } (\mathrm{P}_b) \text{ and } (\mathrm{D}_b)\}, \qquad k = 1, \ldots, N,$$

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 (7) 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 Sect. 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*

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

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

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

If the optimal solution of the limiting problem is unique, Theorem 3.1 can be formulated in a set-wise sense. The Hausdorff distance between two closed nonempty sets $A, B \subseteq \mathbb{R}^d$ is

$$d_H(A, B) = \max\left(\sup_{a\in A}\inf_{b\in B}\|a - b\|, \sup_{b\in B}\inf_{a\in A}\|a - b\|\right) \in [0, \infty]. \tag{5}$$

The collection of closed subsets of $\mathbb{R}^d$ equipped with d_H is a metric space (with possibly infinite distance) and convergence in distribution is defined as usual by integrals of continuous real-valued bounded functions; see for example King (1989), where the delta method is developed in this context. Recall that $\mathcal{C}$ stands for convex hull.

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

Theorem 3.5 *Suppose assumptions (**A1**), (**B1**), and (**B2**) hold, and that $\mathcal{P}^\star(b) = \{x^\star(b)\}$ is a singleton. On the collection of closed subsets of $\mathbb{R}^d$ with the Hausdorff distance d_H it holds that*

$$r_n\left(\mathcal{P}^\star(b_n) - x^\star(b)\right) \xrightarrow{D} \sum_{\mathcal{K}} \mathbb{1}_{\left\{G^{m_0} \in H_{\mathcal{K}}^{m_0} \setminus \bigcup_{k\in[K]\setminus\mathcal{K}} H_k^{m_0}\right\}} \mathcal{C}(x(I_{\mathcal{K}}, G)),$$

where $H_k^{m_0}$ and $H_{\mathcal{K}}^{m_0}$ are as defined in Theorem 3.1.

We conclude this section by giving two further consequences of our proof techniques: a limit law for the objective value $v(b)$ 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 3.6 *Under assumptions (**A1**), (**B1**), and (**B2**) it holds that*

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

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

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

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 3.7 *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.

Preliminary steps. Recall from Remark 2.2 that bases $I_1, \ldots, I_K$ are feasible for (P_b) and (D_b) and hence optimal. The bases $I_{K+1}, \ldots, I_N$ 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

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

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

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) and (D_b). 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* $\mathcal{P}^*(b_n)$ *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 $A_n^{\emptyset}$ is equivalent to (P_b) 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

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

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

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

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

$$r_n[x^\star(b_n) - Y_n] = \sum_{\emptyset \subset \mathcal{K} \subseteq [K]} \mathbb{1}_{A_n^{\mathcal{K}}}(\omega)\, \alpha_n^{\mathcal{K}}(\omega) \otimes x(I_{\mathcal{K}}, G_n(\omega)) + o_{\mathbb{P}}(1). \tag{6}$$

We next investigate the indicator functions $\mathbb{1}_{A_n^{\mathcal{K}}}(\omega)$ appearing in (6). 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

$$H_k^{m_0} = \{g^{m_0} \in \mathbb{R}^{m_0} : [x(I_k, (g^{m_0}, 0_{m-m_0})]_{\mathrm{DP}_k} \geq 0\}. \tag{7}$$

For $\emptyset \subset \mathcal{K} \subseteq [K]$ 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

$$\mathbb{1}_{A_n^{\mathcal{K}}} = \mathbb{1}_{\left\{G_n^{m_0} \in H_{\mathcal{K}}^{m_0} \setminus \bigcup_{k \in [K] \setminus \mathcal{K}} H_k^{m_0}\right\}} + o_{\mathbb{P}}(1). \tag{8}$$

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

Proof (Theorem 3.1) The goal is to replace $G_n^{m_0}$ by G^{m_0} in the indicator function in (8) 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 $T^{\mathcal{K}} : \mathbb{R}^{|\mathcal{K}|} \times \mathbb{R}^m \to \mathbb{R}^d$

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

This function is continuous for all $\alpha \in \mathbb{R}^{\mathcal{K}}$ and all vectors $v \in \mathbb{R}^m$ such that $v_{[m_0]} \notin \partial[H_{\mathcal{K}}^{m_0} \setminus \bigcup_{k\in[K]\setminus\mathcal{K}} H_k^{m_0}]$. 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 $T = \left(T^{\mathcal{K}_1}, \ldots, T^{\mathcal{K}_B}\right) : \mathbb{R}^{\sum_{i=1}^B |\mathcal{K}_i|} \times \mathbb{R}^m \to (\mathbb{R}^d)^B$ 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

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

which completes the proof of Theorem 3.1. □

Proof (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 (6) simplifies to

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

where $K(\omega)$ is the minimal $k \le K$ 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 $j < k \le K$ have Lebesgue measure zero. If $v \in H_j^m \cap H_k^m$ then there exists $\eta > 0$ such that $x(I_j, b + \eta v) \ge 0$ and $x(I_k, b + \eta v) \ge 0$. 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. □

Proof (Theorem 3.5) We consider the optimality sets $\mathcal{P}^\star(b_n)$ as elements of the power set in $\mathbb{R}^d$ endowed with the Hausdorff distance d_H.[7] Then, for all $\mathcal{K} \subseteq [N]$ the mapping $v \mapsto$

[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 $\partial(\bigcap_{k\in\mathcal{K}} A_k \cap \bigcap_{k\in[K]\setminus\mathcal{K}} B_k) \subseteq \bigcup_{k=1}^K \partial A_k$.

[7] To include the empty set define $d_H(\emptyset, \emptyset) = 0$ and $d_H(\emptyset, A) = 1$ for all nonempty A.

$\mathcal{C}(x(I_{\mathcal{K}}), v)$ is Lipschitz since without loss of generality $\mathcal{K} \neq \emptyset$ and

$$d_H(\mathcal{C}(x(I_{\mathcal{K}}, u), \mathcal{C}(x(I_{\mathcal{K}}, v)) \leq \|u - v\| \max_{k \in \mathcal{K}} \|A_{I_k}^{-1}\|_\infty.$$

It follows that

$$\mathcal{P}^\star(b_n) = \sum_{\mathcal{K} \subseteq [N]} \mathbb{1}_{A_n^{\mathcal{K}}} \mathcal{C}(x(I_{\mathcal{K}}, b_n)), \qquad (\mathcal{C}(\emptyset) = \emptyset)$$

is a measurable random subset of $\mathbb{R}^d$. According to Fact 2.3 in presence of (**A1**) and the preceding computations

$$r_n \mathcal{P}^\star(b_n) = o_{\mathbb{P}}(1) + \sum_{\emptyset \neq \mathcal{K} \subseteq [K]} \mathbb{1}_{\left\{G_n^{m_0} \in H_{\mathcal{K}}^{m_0} \setminus \bigcup_{k \in [K] \setminus \mathcal{K}} H_k^{m_0}\right\}} \mathcal{C}(x(I_{\mathcal{K}}, r_n b_n)).$$

If $\mathcal{P}^\star(b) = \{x^\star(b)\}$ is a singleton, then $\mathcal{C}(x(I_{[K]}, b) = \{x^\star(b)\}$ and therefore

$$\begin{aligned} r_n(\mathcal{P}^\star(b_n) - x^\star(b)) = o_{\mathbb{P}}(1) + & \sum_{\emptyset \neq \mathcal{K} \subseteq [K]} \mathbb{1}_{\left\{G_n^{m_0} \in H_{\mathcal{K}}^{m_0} \setminus \bigcup_{k \in [K] \setminus \mathcal{K}} H_k^{m_0}\right\}} \mathcal{C}(x(I_{\mathcal{K}}, G_n)) \\ \to & \sum_{\emptyset \neq \mathcal{K} \subseteq [K]} \mathbb{1}_{\left\{G^{m_0} \in H_{\mathcal{K}}^{m_0} \setminus \bigcup_{k \in [K] \setminus \mathcal{K}} H_k^{m_0}\right\}} \mathcal{C}(x(I_{\mathcal{K}}, G)) \end{aligned}$$

by the continuous mapping theorem. □

Proof (Proposition 3.7) 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 3.6. □

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 $\mathcal{N} := \left\{\lambda \in \mathbb{R}^m \mid A^t\lambda \leq c\right\}$. Further, define the set of feasible parameters by $\mathcal{M} := \{b \in \mathbb{R}^m \mid \mathcal{P}(b) \neq \emptyset\}$ and $\mathcal{M}^\star := \{b \in \mathbb{R}^m \mid \mathcal{P}^\star(b) \neq \emptyset\}$ 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 $\mathcal{P}^\star(b) \neq \emptyset$ then (**A1**) is equivalent to $\mathcal{P}^\star(b)$ being bounded. The latter property is independent on b and equivalent to the set $\{x \in \mathbb{R}^d \mid Ax = 0,\ x \geq 0,\ c^t x = 0\}$ 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 $\mathcal{P}^\star(b) \neq \emptyset$ 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 Sect. 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[8] 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 $(\alpha_n^{\mathcal{K}}, G_n)$ jointly converges[9] in distribution to some limit random variable $(\alpha^{\mathcal{K}}, G)$ 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

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

[9] 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.

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

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

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

$$\text{BL}(\mu_1, \mu_2) := \sup_{\varphi \in \text{BL}(\mathcal{Z})} \left| \int_{\mathcal{Z}} \varphi(z)\, d\,(\mu_1 - \mu_2)\,(z) \right| \tag{9}$$

is well-known to metrize convergence in distribution of (probability) measures on $\mathcal{Z}$ Dudley (1966) [Theorems 6 and 8]. 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 $\varphi \in \text{BL}(\mathcal{Z})$ 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

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

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*

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

for any continuous bounded function $\varphi \in \text{BL}(\mathcal{Z})$ (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

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

where $L := \sup_{g_1 \neq g_2} \text{BL}(\mu_{g_1}^{\mathcal{K}}, \mu_{g_2}^{\mathcal{K}})/\|(g_1 - g_2)\| \in [0, \infty]$. 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 $\mathcal{X} = \{x_1, \ldots, x_N\}$ equipped with some underlying cost $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ OT between two probability measures $r, s \in \Delta_N := \{r \in \mathbb{R}^N \mid \mathbf{1}_N^T r = 1,\ r_i \geq 0\}$ 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 (Tameling et al., 2021), 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 $\hat{r}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. 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 Sect. 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 Subsect. 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

$$\Sigma(r) := \begin{bmatrix} r_1(1-r_1) & -r_1 r_2 & \dots & -r_1 r_N \\ -r_1 r_2 & r_2(1-r_2) & \dots & -r_2 r_N \\ \vdots & & \ddots & \vdots \\ -r_1 r_N & -r_2 r_N & \dots & r_N(1-r_N) \end{bmatrix}. \tag{10}$$

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,

$$A = \begin{pmatrix} \mathbf{1}_N^T & & \\ & \ddots & \\ & & \mathbf{1}_N^T \\ \mathbf{I}_N & \dots & \mathbf{I}_N \end{pmatrix} \in \mathbb{R}^{2N \times N^2}, \tag{11}$$

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

$$\Pi(r, s) = \left\{ \pi \in \mathbb{R}^{N^2} \mid A_\dagger \pi = \begin{bmatrix} r_\dagger \\ s \end{bmatrix},\ \pi \geq 0 \right\}. \tag{12}$$

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*

$$\sqrt{n}\left(\pi^\star(\hat{r}_n, s) - \pi_n^\star(r, s)\right) \xrightarrow{D} \sum_{\mathcal{K}} \mathbb{1}_{\{G(r_\dagger) \in H_\mathcal{K} \setminus \cup_{k \notin \mathcal{K}} H_k\}}\, \alpha^\mathcal{K} \otimes \pi(I_\mathcal{K}, [G(r_\dagger), 0_N]) \tag{13}$$

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

$$\sqrt{\frac{nm}{n+m}}\left(\begin{bmatrix}\widehat{r_\dagger}_n\\ \hat{s}_m\end{bmatrix} - \begin{bmatrix}r_\dagger\\ s\end{bmatrix}\right) \xrightarrow{D} G_\lambda\left(r_\dagger, s\right) := \left(\sqrt{\lambda}G^1(r_\dagger), \sqrt{1-\lambda}G^2(s)\right) \tag{14}$$

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 (10), 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 $\pi^\star_{n,m}(r,s) \in \Pi^\star(r,s)$ such that

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

We provide further details in Appendix 1.

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

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

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 $\{1,\dots,9\}$ 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

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

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

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

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 (7) are

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

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

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

The weak convergence in (14) 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*

$$\sum_{k=1}^{n} c_{i_k j_k} \neq \sum_{k=1}^{n} c_{i_k j_{k-1}}, \quad j_0 := j_n. \tag{15}$$

*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 (15) that is reminiscent to the well-known *cyclic monotonicity property* (Rüschendorf, 1996). Further, (15) 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

$$\sum_{i \in A} r_i \neq \sum_{j \in B} s_j \tag{16}$$

that guarantees every *primal* basic solution to be non-degenerate. Notably, (15) 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

$$c_{ij} + c_{i'j'} < c_{ij'} + c_{i'j}, \quad \forall\, i < i', j < j', \tag{17}$$

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

Lemma 6.5 *Let $\mathcal{X} := \{x_1 < \ldots < x_N\}$ 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*

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

*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 1). 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 $X_1, \ldots, X_N \sim \mu$ and $Y_1, \ldots, Y_N \sim \nu$. Then (15) 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 $[1,2]^D + (2,0,0,\ldots,0)$. 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.

A Omitted proofs

Proof (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 $1 \leq j \leq K$. 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

$$a_z = \sum_{i \in Pos} y_i^z a_i + \sum_{s \in I_j \setminus Pos} y_s^z a_s, \qquad z \in L \setminus I_j \,. \tag{18}$$

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 $\widetilde{I} := I_j \setminus \{s\} \cup \{z\}$ 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 (18).

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

$$\begin{aligned} b = \sum_{i \in Pos} x_i a_i &= \sum_{\substack{i \in Pos \\ i \neq i_0}} x_i a_i + x_{i_0} \left(\frac{1}{y_{i_0}^z} a_z - \sum_{\substack{i \in Pos \\ i \neq i_0}} \frac{y_i^z}{y_{i_0}^z} a_i \right) \\ &= \frac{x_{i_0}}{y_{i_0}^z} a_z + \sum_{\substack{i \in Pos \\ i \neq i_0}} \left(x_i - \frac{x_{i_0} y_i^z}{y_{i_0}^z} \right) a_i = \sum_{i \in \widetilde{I}} \widetilde{x}_i a_i \end{aligned}$$

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 (18) that $y_i^z \leq 0$. Consider the vector

$$w := \text{Aug}_{\{z\}}(1) - \text{Aug}_{Pos}(y^z)\,.$$

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 $1 \leq j \leq K$ yields a degenerate dual basic solution we can modify basis I_j to some I_i with $i \neq j$ and $1 \leq i \leq K$ such that $\lambda(I_j) = \lambda(I_j)$.

It is in principle possible that $\lambda(I_l)$ is dual optimal $K+1 \leq l \leq N$ 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 $1 \leq j \leq K$ 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 $Pos_j \nsubseteq I_l$, 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 $1 \leq i \leq K$ 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. □

Proof (Lemma 5.5) We need to show that for any $\varphi \in \text{BL}(\mathcal{Z})$ we have that

$$\begin{aligned}|\mathbb{E}[\varphi(\alpha_n^{\mathcal{K}}, G_n)] - \mathbb{E}[\varphi(\alpha^{\mathcal{K}}, G)]| &= |\mathbb{E}\psi_n(G_n) - \mathbb{E}\psi(G)| \\ &\leq |\mathbb{E}[\psi_n(G_n) - \psi(G_n)]| + |\mathbb{E}[\psi(G_n) - \psi(G)]|\end{aligned}$$

vanishes as $n \to \infty$. To bound the first term notice that for any fixed g it holds that $\|\alpha \mapsto \varphi(\alpha, g)\|_{\text{BL}(\Delta_{|\mathcal{K}|})} \leq \|\varphi\|_{\text{BL}(\mathcal{Z})} \leq 1$, so

$$|\psi_n(g) - \psi(g)| \leq \left| \int_{\Delta_{|\mathcal{K}|}} \varphi(\alpha, g) d[\mu_{n,g}^{\mathcal{K}} - \mu_g^{\mathcal{K}}](\alpha) \right| \leq \text{BL}(\mu_{n,g}^{\mathcal{K}}, \mu_g^{\mathcal{K}}).$$

Hence, we find $\mathbb{E}|\psi_n(G_n) - \psi(G_n)| \leq \sup_g \text{BL}(\mu_{n,g}^{\mathcal{K}}, \mu_g^{\mathcal{K}})$ 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 $\|\psi\|_\infty \leq \|\varphi\|_\infty$ and that

$$\begin{aligned}|\psi(g_1) - \psi(g_2)| &= \left| \int_{\Delta_{|\mathcal{K}|}} [\varphi(\alpha, g_1) - \varphi(\alpha, g_2)] d\mu_{g_1}^{\mathcal{K}}(\alpha) + \int_{\Delta_{|\mathcal{K}|}} \varphi(\alpha, g_2) d[\mu_{g_2}^{\mathcal{K}} - \mu_{g_1}^{\mathcal{K}}](\alpha) \right| \\ &\leq \|\varphi\|_{\text{Lip}} \|g_1 - g_2\| + \text{BL}(\mu_{g_2}^{\mathcal{K}}, \mu_{g_1}^{\mathcal{K}}).\end{aligned}$$

Hence, we conclude that

$$\|\psi\|_{\text{BL}(\mathbb{R}^m)} \leq \|\varphi\|_{\text{BL}(\mathcal{Z})} + L \leq 1 + L.$$

Dividing ψ by its bounded Lipschitz norm, we find

$$\mathbb{E}|\psi(G_n) - \psi(G)| \leq \|\psi\|_{\text{BL}(\mathbb{R}^m)} \text{BL}(\mathcal{L}(G_n), \mathcal{L}(G)) \leq (1 + L)\text{BL}(\mathcal{L}(G_n), \mathcal{L}(G)).$$

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 $\varphi \in \text{BL}(\mathcal{Z})$, which implies that $(\alpha_n^{\mathcal{K}}, G_n) \to (\alpha^{\mathcal{K}}, G)$ in distribution. □

B Optimal transport

Proof (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 (7) 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 $w = (v_{[N-1]}, -\sum_{j=1}^{N-1} v_j, v_{\{N,\dots 2N-1\}}) \in \mathbb{R}^{2N}$ be the augmented vector corresponding to v. In view of Brualdi (2006)[Corollary 8.1.4], there exist $\mathsf{R}_{i,k} \subset \{1, \dots, N\}$ and $\mathsf{S}_{i,k} \subset \{N+1, \dots, 2N\}$, 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 $\mathsf{R} \subseteq \{1, \dots, N-1\}$ and $\mathsf{S} \subset \{1, \dots, N\}$ 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. □

Proof (Lemma 6.5) By elementary arguments the cost $c(x_i, x_j) = f(|x_i - x_j|)$ satisfies the strict Monge condition (17) 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

$$-\infty \leq a_0 < a_1 < a_2 < a_3 < \cdots < a_M \leq \infty$$

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 $m_0 \leq \nu([a_1, a_2])$. Define the quantile

$$a^\star = \inf\{a : \nu[a_1, a] \geq m_0\} \in [a_1, a_2].$$

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 $B \cap [a_1, a^\star] = \emptyset$. By definition of $a^\star$, ν_1 is a measure (i.e., $\nu_1(\{a^\star\}) \geq 0$) 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

$$a^\star = \inf\{a : \mu([a_0, a] \geq \nu([a_1, a_2])\} \in [a_0, a_1],$$

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. □

Proof (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

$$(i = i_1, j_1, i_2, j_2, \ldots, i_n, j_n = j)$$

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 (15) 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 $\alpha_i + \beta_j \neq c_{ij}$. □

Proof (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

$$\sum_{k=1}^{n} \|X_{i_k} - Y_{j_k}\|_q^p = \sum_{k=1}^{n} \|X_{i_k} - Y_{j_{k-1}}\|_q^p, \quad Y_{j_0} := Y_{j_n}. \tag{19}$$

It suffices to prove that (19) holds with probability zero for fixed n. For the sake of notational simplicity, we choose the first $n \leq N$ random locations $(\mathbf{X}, \mathbf{Y}) := (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. We denote by $(\mathbf{x}, \mathbf{y}) = (x_1, \ldots, x_n, y_1, \ldots, y_n) \in \left(\mathbb{R}^D\right)^{2n}$ and define the set

$$A := \left\{ (\mathbf{x}, \mathbf{y}) \in \left(\mathbb{R}^D\right)^{2n} \mid \sum_{k=1}^{n} \|x_k - y_{k-1}\|_q^p - \|x_k - y_k\|_q^p = 0 \right\}.$$

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

$$B := \bigcup_{i \in [D], k \in [n]} \{ (\mathbf{x}, \mathbf{y}) \in \left(\mathbb{R}^D\right)^{2n} \mid \langle x_k, e_i \rangle \in \{ \langle y_{k-1}, e_i \rangle, \langle y_k, e_i \rangle \}, \; y_0 := y_n \}.$$

Define the function $f : (\mathbb{R}^D)^{2n} \setminus B \to \mathbb{R}$ 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 $U_1, \ldots, U_L$ 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

$$f_{|U_l}(\epsilon) = f_{|U_l}(x_1 + \epsilon e_i, x_2, \ldots, x_n, y_1, \ldots, y_n)$$

with derivative at $\epsilon = 0$ given by

$$\frac{\partial f_{|U_l}}{\partial \epsilon}_{|\epsilon=0} = p\left(\|x_1 - y_1\|_q^{p-q} \frac{|x_{1_i} - y_{1_i}|^q}{(x_{1_i} - y_{1_i})} - \|x_1 - y_n\|_q^{p-q} \frac{|x_{1_i} - y_{n_i}|^q}{(x_{1_i} - y_{n_i})}\right), \tag{20}$$

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 (20), and for small δ the derivative will not be zero. If $p = q \neq 1$ then the norms do not appear in (20) 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. □

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., & Bottou, L. (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, pp 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* (Vol. 6). MA: Athena Scientific Belmont.

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

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

Bonnans JF, & 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,* (Vol. 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 RJB (2007) George B. Dantzig (1914–2005). Notices of the AMS *54*(3), 344–362
Dang NV (2015) Complex powers of analytic functions and meromorphic renormalization in QFT. preprint arXiv:1503.00995
Dantzig, G. B. (1963). *Linear Programming and Extensions*. Princeton University Press.
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.
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,* (Vol. 74). Cambridge University Press.
Dudley, R. (1966). Convergence of baire measures. *Studia Mathematica, 3*(27), 251–268.
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 TA (2015) Learning with a Wasserstein loss. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) *Advances in Neural Information Processing Systems* 28, Curran Associates, Inc., pp 2053–2061
Gal, T., & Greenberg, H. J. (2012). *Advances in Sensitivity Analysis and Parametric Programming* (Vol. 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. *Proceedings of Symposia in Pure Mathematics, 7,* 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* (2nd ed.). Springer-Verlag.
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. *Doklady Akademii Nauk USSR, 37,* 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. (1989). Generalized delta theorems for multivalued mappings and measurable selections. *Mathematics of Operations Research, 14*(4), 720–736.
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*, pp 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.
Prékopa, A. (1966). On the probability distribution of the optimum of a random linear program. *SIAM Journal on Control, 4*(1), 211–222.
Rachev ST, & Rüschendorf L (1998) Mass Transportation Problems: Volume I: Theory, Volume II: Applications. Springer, New York
Rachev, S. T. (1985). The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability and its Applications, 29*(4), 647–676.
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 and Probability Letters, 27*(3), 267–270.
Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Basel: Birkhäuser.
Shapiro A (2000) Statistical inference of stochastic optimization problems. In: *Probabilistic constrained optimization*, Springer, pp 282–307
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, 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., et al. (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* (Vol. 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.
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*(3), 199–211.
Terlaky, T., & Zhang, S. (1993). Pivot rules for linear programming: A survey on recent theoretical developments. *Annals of Operations Research, 46*(1), 203–233.
Tintner, G. (1960). A note on stochastic linear programming. *Econometrica: Journal of the Econometric Society* pp. 490–495.

Vershik, A. (2002). L.V. Kantorovich and linear programming. *Leonid Vital'evich Kantorovich: A man and a scientist, 1,* 130–152.
Villani, C. (2008). *Optimal Transport: Old and New*. Berlin: Springer.
Walkup, D.W., Wets, R.J.B. (1969b). A Lipschitzian characterization of convex polyhedra. *Proceedings of the American Mathematical Society* pp. 167–173.
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. (1969). Lifting projections of convex polyhedra. *Pacific Journal of Mathematics, 28*(2), 465–475.
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. *Stochastic Programming* (pp. 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.