Approximating Optimal Transport With Linear Programs

Kent Quanrud[1]

University of Illinois, Urbana-Champaign, Illinois, USA
http://web.engr.illinois.edu/~quanrud2
quanrud2@illinois.edu

── **Abstract** ──

In the regime of bounded transportation costs, additive approximations for the optimal transport problem are reduced (rather simply) to relative approximations for positive linear programs, resulting in faster additive approximation algorithms for optimal transport.

1 Introduction

For $\ell \in \mathbb{N}$, let $\Delta^\ell = \left\{ p \in \mathbb{R}^\ell_{\geq 0} : \|p\|_1 = 1 \right\}$ denote the convex set of probability distributions over $[\ell]$. In the **(discrete) optimal transport problem**, one is given two distributions $p \in \Delta^\ell$ and $q \in \Delta^k$ and a nonnegative matrix of **transportation costs** $C \in \mathbb{R}^{k \times \ell}_{\geq 0}$. The goal is to

$$\text{minimize } \sum_{i=1}^{k} \sum_{j=1}^{\ell} C_{ij} X_{ij} p_j \text{ over } X \in \mathbb{R}^{k \times \ell}_{\geq 0} \text{ s.t. } Xp = q \text{ and } X^t \mathbb{1} = \mathbb{1}. \tag{T}$$

We let (T) denote both the above optimization problem and its optimal value. (T) can be interpreted as the minimum cost of "transporting" a discrete distribution p to a target distribution q, where the cost of moving probability mass from one coordinate to another is given by C. (T) is sometimes called the *earth mover distance* between p and q, where one imagines p and q as each dividing the same amount of sand into various piles, and the goal is to rearrange the piles of sand of p into the piles of sand of q with minimum total effort.

Optimal transport (in much greater generality) is fundamental to applied mathematics [15, 16]. Computing (or approximating) the optimal transport matrix and its cost has many applications: we refer to recent work by Cuturi [8], Altschuler, Weed, and Rigollet [2] and Dvurechensky, Gasnikov, and Kroshnin [9] for further (and up-to-date) references.

Optimal transport is a linear program (abbr. LP) and can be solved exactly by linear program solvers. (T) can also be cast as a minimum cost flow problem, thereby solved combinatorially. The fastest exact algorithm runs in $\tilde{O}\left(k\ell\sqrt{k + \ell} \right)$ time via minimum cost flow [11]. Here and throughout $\tilde{O}(\cdot)$ hides polylogarithmic terms in k, ℓ.

There is recent interest, sparked by Cuturi [8], in obtaining *additive approximations* to (T) with running times that are *nearly linear* in the size of the cost matrix C. For $\delta > 0$, a matrix X is a **δ-additive approximation** if it is a feasible solution to (T) with cost at most a δ additive factor more than the optimal transport cost (T). A "nearly linear" running time is one whose dependence on k and ℓ is of the form $O(k\ell \operatorname{polylog}(k, \ell))$; i.e., linear in the input size up to polylogarithmic factors. Cuturi highlights applications in machine learning with large, high-dimensional datasets, for which a faster approximation algorithm may be preferable to a slower exact algorithm.

The first nearly linear time additive approximation was obtained recently by Altschuler et al. [2]. Their result combines a reduction to matrix scaling observed by Cuturi and an improved analysis for a classical matrix scaling algorithm due to Sinkhorn and Knopp [14] as applied to this setting (see also [4]). The bound has a cubic dependency on $\|C\|_\infty / \delta$, where $\|C\|_\infty = \max_{i,j} C_{ij}$ is the maximum value of any coordinate in C and is considered a lower order term. One factor of $1/\delta$ can be removed by recent advances in matrix scaling [7] (per Altschuler et al. [2]). A tighter analysis by Dvurechensky et al. [9] of the Sinkhorn-Knopp approach decreases the dependency on $\|C\|_\infty / \delta$ to the following.

▶ **Theorem 1** ([9]). *A δ-additive approximation to (T) can be computed in $\tilde{O}\left(k\ell \left(\frac{\|C\|_\infty}{\delta} \right)^2 \right)$ time.*

1.1 Results

The optimal transport cost can be approximated more efficiently as follows. Some of the results are parametrized by the quantity $\langle p, Cq \rangle$ instead of $\|C\|_\infty$. The quantity $\langle p, Cq \rangle$ is the average cost coefficient as sampled from the product distribution $p \times q$. Needless to say, the average cost $\langle p, Cq \rangle$ is at most the maximum cost $\|C\|_\infty$, and the relative difference may be arbitrarily large.

▶ **Theorem 2.** *One can compute a δ-additive approximate transportation matrix X from p to q sequentially in either*

1. $\tilde{O}\left(k\ell \left(\frac{\langle q, Cp \rangle}{\delta} \right)^2 \right)$ *deterministic time or*
2. $\tilde{O}\left(k\ell \frac{\|C\|_\infty}{\delta} \right)$ *randomized time;*

or deterministically in parallel with

3. $\tilde{O}\left(\left(\frac{\langle q, Cp \rangle}{\delta} \right)^3 \right)$ *depth and $\tilde{O}\left(k\ell \left(\frac{\langle p, Cq \rangle}{\delta} \right)^2 \right)$ total work, or*
4. $\tilde{O}\left(\left(\frac{\|C\|_\infty}{\delta} \right)^2 \right)$ *depth and $\tilde{O}\left(k\ell \left(\frac{\|C\|_\infty}{\delta} \right)^2 \right)$ total work.*

The bounds are obtained rather simply by reducing to a variety of relative approximation algorithms for certain types of LPs. The reductions can be summarized briefly as follows.

A simple but important observation is that the transportation cost from p to q is bounded above by $\langle q, Cp \rangle$ (Lemma 3 below). Consequently, $(1 \pm \epsilon)$-multiplicative approximations to the value of (T), for $\epsilon = \delta / \langle q, Cp \rangle$, are δ-additive approximations as well.

The approximate LP solvers produce matrices X that certify the approximate value, but do not meet the constraints of (T) exactly. In particular, the approximations X transport $(1 - \epsilon)$-fraction of the mass, leaving ϵ-fraction behind. The remaining ϵ-fraction of probability mass is then transported by a simple oblivious transportation scheme.

Here the algorithms diverge into two types, depending on how to model (T) as an LP. The first approach takes (T) as is, which is a "positive LP". Positive LPs are a subclass of LPs where all coefficients and variables are nonnegative. Positive LPs can be approximated faster than general LPs can be solved. Applied to (T), the approximation algorithms for positive LPs produce what we call "$(1 - \epsilon)$-uniform transportation matrices", which not only transport all but an ϵ-fraction of the total mass, but transport all but an ϵ-fraction of each coordinate of p, and fill all but an ϵ-fraction of each coordinate of q. It is shown that $(1 - \epsilon)$-uniformly approximate transportation matrix can be altered into exact transportation matrices with an additional cost of about $\epsilon \langle q, Cp \rangle$.

The second approach reformulates (T) as a "packing LP". Packing LPs are a subclass of positive LPs characterized by having only packing constraints. The advantage of packing LPs is that they can be approximated slightly faster than the broader class of positive LPs. However, the approximate transportation matrices X produced by the packing LP are not uniformly approximate in the sense discussed above. Consequently, there is a larger cost of about $\epsilon \|C\|_\infty$ to extend X to an exact transportation matrix.

1.2 Additional background

There is a burgeoning literature on parametrized regimes of optimal transport. The many parametrized settings are beyond the scope of this note, and we refer again to [2, 9] for further discussion.

An important special case of the optimal transport problem (T) is where C is a metric or, more generally, the shortest path metric of an undirected weighted graph. This setting is equivalent to *uncapacitated minimum cost flow*. Let m denote the number of edges and n the number of vertices of the underlying graph. Recently, Sherman [13] proved that a $(1 + \epsilon)$-multiplicative approximation to (T) can be obtained in $\tilde{O}\left(m^{1+o(1)}/\epsilon^2\right)$ time. This translates to a δ-additive approximation in $\tilde{O}\left(m^{1+o(1)}(\langle q, Cp \rangle/\delta)^2\right)$ time. Remarkably, if the graph is sparse, then $\tilde{O}\left(m^{1+o(1)}\right)$ is much smaller than the explicit size of the shortest path metric, n^2 – let alone the time required to compute all pairs of shortest paths.

There are many applications where the cost matrix C is induced by some combinatorial or geometric context and may be specified more sparsely than as $O(k\ell)$ explicit coordinates. It is well known that some of the LP solvers used below as a black box, as well as other similar algorithms, can often be extended to handle such implicit matrices so long as one can provide certain simple oracles (e.g., [10, 17, 12, 5, 6]).

The running time (2) of Theorem 2 was obtained independently by Blanchet, Jambulapati, Kent, and Sidford [3], by a similar reduction to packing LPs. Blanchet et al. [3] also get the running time (2) via matrix scaling, more in the spirit of the preceding works [8, 2, 9].

1.3 Organization

The rest of this note is organized as follows. Section 2 outlines a simple and crude approximation algorithm for (T), which is used to repair approximate transportation matrices, and to upper bound (T). Section 3 applies positive LP solvers to approximate (T), and leads to

the running times in Theorem 2 that depend on $\langle q, Cp \rangle$ and not $\|C\|_\infty$. Section 4 applies packing LP solvers to a reformulation of (T). This approach leads to the remaining running times in Theorem 2 that all depend on $\|C\|_\infty$.

2   Oblivious transport

The high-level idea is to use approximate LP solvers to transport most of p to q, and then transport the remaining probability mass with a cruder approximation algorithm. The second step always uses the following oblivious transportation scheme. The upper bound obtained below also provides a frame of reference for comparing additive and relative approximation factors, and is useful for bounding a binary search for the optimal value.

▶ **Lemma 3.** *For a distribution $q \in \Delta^k$, consider the matrix $X \in \mathbb{R}_{\geq 0}^{k \times \ell}$ with each column set to q; i.e., $X_{ij} = q_i$ for all i, j. For any $p \in \Delta^\ell$, X is a transportation matrix from p to q, with total cost $\langle q, Cp \rangle$.*

Proof. Fix $p \in \Delta^\ell$. For any $i \in [m]$,

$$\langle e_i, Xp \rangle = \sum_{j=1}^{\ell} X_{ij} p_j = q_i \sum_{j=1}^{\ell} p_j \stackrel{(1)}{=} q_i$$

since (1) p is a distribution. For any $j \in [n]$, we have

$$\langle \mathbb{1}, Xe_j \rangle = \sum_{i=1}^{k} X_{ij} = \sum_{i=1}^{k} q_i \stackrel{(2)}{=} 1.$$

since (2) q is a distribution. Thus X is a transportation matrix from q to p. The transportation cost of X is

$$\sum_{i=1}^{k} \sum_{j=1}^{\ell} C_{ij} X_{ij} p_j = \sum_{i=1}^{k} \sum_{j=1}^{\ell} C_{ij} q_i p_j = \langle q, Cp \rangle,$$

as desired.   ◀

3   Reduction to mixed packing and covering

Our first family of approximation algorithms, which obtain the bounds in Theorem 2 that are relative to $\langle q, Cp \rangle$, observe that optimal transport lies in the following class of LPs. A **mixed packing and covering program** is a problem of any of the forms

$$\{\text{find } x, \ \max \langle v, p \rangle, \ \text{or } \min \langle v, p \rangle\} \text{ over } x \in \mathbb{R}_{\geq 0}^n \text{ s.t. } Ax \leq b \text{ and } Cx \geq d, \tag{PC}$$

where $A \in \mathbb{R}_{\geq 0}^{m_1 \times n}$, $b \in \mathbb{R}_{\geq 0}^{m_1}$, $C \in \mathbb{R}_{\geq 0}^{m_2 \times n}$, and $d \in \mathbb{R}_{\geq 0}^{m_2}$, and $v \in \mathbb{R}_{\geq 0}^n$ all have nonnegative coefficients. We let N denote the total number of nonzeroes in the input. For $\epsilon > 0$, an **ϵ-relative approximation** to (PC) is either (a) a certificate that (PC) is either infeasible, or (b) a nonnegative vector $x \in \mathbb{R}_{\geq 0}^n$ such that $Ax \leq (1 + \epsilon)b$ and $Cx \geq (1 - \epsilon)d$ and, when there is a linear objective and the linear program is feasible, within a $(1 \pm \epsilon)$-multiplicative factor of the optimal value. Relative approximations to positive LPs can be obtained with nearly-linear dependence on N, and polynomial dependency on $\dfrac{1}{\epsilon}$, as follows.

▶ **Lemma 4** ([17]). *Given an instance of (PC) and $\epsilon > 0$, one can compute a ϵ-relative approximation to (PC) in $\tilde{O}\big(N/\epsilon^2\big)$ deterministic time.*

▶ **Lemma 5** ([12]). *Given an instance of a mixed packing and covering problem (PC) and $\epsilon > 0$, one can compute a ϵ-relative approximation to (PC) deterministically in parallel in $\tilde{O}\big(1/\epsilon^3\big)$ depth and $\tilde{O}\big(N/\epsilon^2\big)$ total work.*

(T) is a minimization instance of mixed packing and covering that is always feasible. The role of nonnegative variables is played by the coordinates of the transportation matrix $X \in \mathbb{R}_{\geq 0}^{n \times n}$, with costs $p_j C_{ij}$ for each X_{ij}. The two equations $Xp = q$ and $X^t \mathbb{1} = \mathbb{1}$ each give rise to two sets of packing constraints, $Xp \leq q$ and $X^t \mathbb{1} \leq \mathbb{1}$, and two sets of covering constraints, $Xp \geq q$ and $X^t \mathbb{1} \geq \mathbb{1}$. We have $N = O(k\ell)$ nonzeroes, $m = 2(k + \ell)$ packing and convering constraints, and $n = k\ell$ variables.

An ϵ-relative approximation to (PC) is not necessarily a feasible solution to (T). To help characterize the difference, we define the following. For fixed $\epsilon > 0$ and two distributions $p \in \Delta^\ell$ and $q \in \Delta^k$, a **$(1 - \epsilon)$-uniform transportation matrix** from p to q is a nonnegative matrix $X \in \mathbb{R}_{\geq 0}^{k \times \ell}$ with $(1 - \epsilon)q \leq Xp \leq q$ and $(1 - \epsilon)\mathbb{1} \leq X^t \mathbb{1} \leq \mathbb{1}$.

▶ **Lemma 6.** *Given an instance of the optimal transport problem (T) and $\epsilon > 0$, a $(1 - \epsilon)$-uniform transport matrix with cost at most (T) can be computed*

1. *sequentially in $\tilde{O}\left(\dfrac{kl}{\epsilon^2}\right)$ time, and*
2. *in parallel in $\tilde{O}\left(\dfrac{1}{\epsilon^3}\right)$ depth and $\tilde{O}\left(\dfrac{kl}{\epsilon^2}\right)$ total work.*

Proof. By either Lemma 4 or Lemma 5, one can compute an ϵ-approximation X to (T) with the claimed efficiency. Then $(1 - \epsilon)X$ is a $(1 - \epsilon)^2$-uniform approximate transportation matrix.   ◀

▶ **Lemma 7.** *Given a $(1 - \epsilon)$-uniform approximate transportation matrix X, one can compute a transportation matrix U with cost at most an additive factor of $4\epsilon \langle q, Cp \rangle$ more than the cost of X, in linear time and work and with constant depth.*

Proof. We first scale down X slightly to a $(1 - 2\epsilon)$-uniform transportation matrix, and then augment the shrunken transportation matrix with the oblivious transportation scheme from Lemma 3. Clearly this can be implemented in linear time and work and in constant depth.

Let $Y = \left(1 - \dfrac{\epsilon}{1 - \epsilon}\right)X$. Then Y is a $(1 - 2\epsilon)$-uniform approximate transportation matrix from p to q. Let $p' = \big(I - \text{diag}\big(S^t \mathbb{1}\big)\big)p$ and $q' = q - Sp$. Since Y is $(1 - 2\epsilon)$-uniform, we have $p' \leq 2\epsilon p$ and $q' \leq 2\epsilon q$. p' represents the probability mass not yet transported by Y, and q' represents the probability mass not yet filled by Y, and we have

$$\langle \mathbb{1}, p' \rangle = \langle \mathbb{1}, q' \rangle.$$

Let α denote this common value. Then

$$\alpha = \langle \mathbb{1}, q' \rangle = 1 - \langle \mathbb{1}, Yx \rangle \geq 1 - \langle \mathbb{1}, Xp \rangle + \frac{\epsilon}{1 - \epsilon} \langle \mathbb{1}, Xp \rangle \overset{(3)}{\geq} \epsilon$$

because (3) X being $(1 - \epsilon)$-uniform implies $1 - \epsilon \leq \langle Xp, \mathbb{1} \rangle \leq 1$. Let Z be the matrix where each column is q'/α; by Lemma 3, Z is a transportation matrix from p'/α to q'/α. Let $Z' = Z(I - \text{diag}\big(Y^t \mathbb{1}\big))$. Then

$$(Y + Z')p = Yp + Zp' = Yx + q' = q,$$

and

$$(Y + Z')^t \mathbb{1} = Y^t \mathbb{1} + (I - \text{diag}(Y^t \mathbb{1}))Z^t \mathbb{1} = Y^t \mathbb{1} + (I - \text{diag}(Y^t \mathbb{1}))\mathbb{1}$$
$$= Y^t \mathbb{1} + \mathbb{1} - Y^t \mathbb{1} = \mathbb{1},$$

so $Y + Z'$ is a transportation matrix from p to q. The cost of Y is less than the cost of X, and the cost of Z' is at most

$$\sum_{i,j} C_{ij} Z'_{ij} p_j \overset{(4)}{=} \sum_{i,j} C_{ij} Z_{ij} p'_j \overset{(5)}{=} \frac{1}{\alpha} \sum_{i,j} C_{ij} q'_i p'_j$$
$$= \frac{1}{\alpha} \langle q', C p' \rangle \overset{(6)}{\leq} \frac{4\epsilon^2}{\alpha} \langle q, C p \rangle$$
$$\leq 4\epsilon \langle q, C p \rangle$$

by (4) definition of Z', (5) definition of Z, (6) $p' \leq 2\epsilon p$ and $q' \leq 2\epsilon q$, and (7) $\alpha \geq \epsilon$.   ◀

▶ **Theorem 8.** *One can deterministically compute a δ-additive approximation to (T)*

1. *sequentially in* $\tilde{O}\left(kl\left(\dfrac{\langle x, Cy \rangle}{\delta}\right)^2\right)$ *time, and*

2. *in parallel in* $\tilde{O}\left(\left(\dfrac{\langle x, Cy \rangle}{\delta}\right)^3\right)$ *depth and* $\tilde{O}\left(kl\left(\dfrac{\langle x, Cy \rangle}{\delta}\right)^2\right)$ *total work.*

Proof. Given $\delta > 0$, let $\epsilon = \dfrac{\delta}{4\langle q, Cp \rangle}$. We apply Lemma 6 to generate a $(1 - \epsilon)$-uniform transportation matrix X of cost at most (T) within the desired time/depth bounds. We then apply Lemma 7 to X to construct a transportation matrix U with cost at most (T) $+ 4\epsilon \langle q, Cp \rangle = $ (T) $+ \delta$, as desired.   ◀

4   Reduction to packing

A **(pure) packing LP** is a linear program of the form

$$\text{maximize } \langle c, p \rangle \text{ over } x \in \mathbb{R}^n_{\geq 0} \text{ over } Ax \leq b, \tag{P}$$

where $A \in \mathbb{R}^{m \times n}_{\geq 0}$, $b \in \mathbb{R}^m_{\geq 0}$, and $c \in \mathbb{R}^n_{\geq 0}$. For a fixed instance of (P), we let N denote the number of nonzeroes in A. For $\epsilon > 0$, a **$(1 - \epsilon)$-relative approximation** to (P) is a point $x \in \mathbb{R}^n_{\geq 0}$ such that $Ax \leq b$ and $\langle c, x \rangle$ is at least $(1 - \epsilon)$ times the optimal value of (P). $(1 - \epsilon)$-relative approximations packing LPs can be obtained slightly faster than ϵ-relative approximations to more general positive linear programs, as follows.

▶ **Lemma 9** ([1]). *Given an instance of the pure packing problem (P), and $\epsilon > 0$, a $(1 - \epsilon)$-multiplicative approximation to (P) can be computed in $\tilde{O}(N/\epsilon)$ randomized time.*

▶ **Lemma 10** ([12]). *Given an instance of the pure packing problem (P), and $\epsilon > 0$, a $(1 - \epsilon)$-multiplicative approximation to (P) can be computed deterministically in parallel in $\tilde{O}(1/\epsilon^2)$ depth and $\tilde{O}(N/\epsilon^2)$ total work.*

Consider the following LP reformulation of (T), that is parametrized by a value λ that specifies a desired transportation cost.

$$\max \ \langle \mathbb{1}, Xp \rangle \text{ over } X \in \mathbb{R}_{\geq 0}^{k \times \ell}$$

$$\text{s.t. } \sum_{i=1}^{k} X_{ij} \leq 1 \text{ for all } j,$$

$$\sum_{j=1}^{\ell} X_{ij} p_j \leq q_i \text{ for all } i \in [m], \qquad (\mathsf{TP}(\lambda))$$

$$\sum_{i=1}^{k} \sum_{j=1}^{\ell} C_{ij} X_{ij} p_j \leq \lambda.$$

The advantage of $(\mathsf{TP}(\lambda))$ compared to (T) is that $(\mathsf{TP}(\lambda))$ is a packing LP, which as observed above can be solved slightly faster than a mixed packing and covering LP. The packing problem $(\mathsf{TP}(\lambda))$ has $m = O(k + \ell)$ packing constraints, $n = k\ell$ variables, and $N = O(k\ell)$ nonzeroes.

$(1 - \epsilon)$-approximations to $(\mathsf{TP}(\lambda))$ are not feasible solutions to (T) even for $\lambda = $ (T). To help characterize the difference, we define the following. For fixed $\epsilon > 0$ and two distributions $p \in \Delta^\ell$ and $q \in \Delta^k$, a **$(1 - \epsilon)$-transportation matrix from p to q** is a nonnegative matrix $X \in \mathbb{R}_{\geq 0}^{k \times \ell}$ with $Xp \leq q$, $X^t \mathbb{1} \leq \mathbb{1}$, and $\langle \mathbb{1}, Xp \rangle \geq 1 - \epsilon$.

▶ **Lemma 11.** *Consider an instance of the optimal transport problem (T), and let $\epsilon, \delta > 0$ be fixed parameters with $\epsilon \leq \dfrac{\delta}{\langle p, Cq \rangle}$. One can compute an $(1 - \epsilon)$-transportation matrix from p to q with cost at most (T) $+ \delta$*

1. *sequentially in $\tilde{O}\left(\dfrac{kl}{\epsilon}\right)$ randomized time, and*

2. *in parallel in $\tilde{O}\left(\dfrac{1}{\epsilon^2}\right)$ depth and $\tilde{O}\left(\dfrac{kl}{\epsilon^2}\right)$ total work.*

Proof. For fixed λ, either $\lambda \leq$ (T), or either $(1 - \epsilon)$-approximation algorithm from Lemma 9 or Lemma 10 returns $(1 - \epsilon)$-transportation matrices X from p to q with cost at most λ. We wrap the $(1 - \epsilon)$-relative approximation algorithms in a binary search for the smallest value, up to an additive factor of δ, that produces a $(1 - \epsilon)$-transportation matrix from p to q. Since $\lambda = $ (T) is sufficient, such a search returns a value of $\lambda \leq$ (T) $+ \delta$. By Lemma 3, the search can be bounded to the range $[0, \langle q, Cp \rangle]$. Thus the binary search needs at most $O\left(\log\left(\dfrac{\langle q, Cp \rangle}{\delta}\right)\right)$ iterations to identify such a value λ, for which we obtain the desired $(1 - \epsilon)$-transportation matrix. We can assume that $\dfrac{\langle p, Cq \rangle}{\delta}$ is at most poly(k, ℓ), since otherwise (T) can be solved exactly in poly$(k, \ell) \leq \dfrac{\langle p, Cq \rangle}{\delta} \leq \dfrac{1}{\epsilon}$ time. ◀

▶ **Lemma 12.** *Let X be a $(1 - \epsilon)$-transportation matrix from p to q. In $O(k\ell)$ time, one can extend X to a transportation matrix U with an additional cost of $\epsilon \|C\|_\infty$.*

Proof. We use the oblivious transportation scheme of Lemma 3 to transport the remaining ϵ-fraction of mass. It is straightforward to verify the additional transportation costs at most $\epsilon \|C\|_\infty$, as follows.

Let $p' = (I - \text{diag}(X^t \mathbb{1}))p$ and $q' = q - Xp$. p' represents the probability mass not yet transported by X, and q' represents the probability mass not yet transported by Y. Let $\alpha = \langle \mathbb{1}, p' \rangle = \langle \mathbb{1}, q' \rangle$ denote the residual probability mass. Since X is a $(1 - \epsilon)$-transportation matrix, we have

$$\alpha = \langle \mathbb{1}, q' \rangle = 1 - \langle \mathbb{1}, Xp \rangle \leq \epsilon. \tag{1}$$

Let Y be the matrix[2] where each column is set to q'/α; by Lemma 3, Y is a transportation matrix from p'/α to q'/α. Let $Y' = Y(I - \text{diag}(X^t \mathbb{1}))p$. By the same calculations as in the proof of Lemma 7, $X + Y'$ is a transportation matrix from q to p. The cost of Y' is

$$\sum_{ij} C_{ij} Y'_{ij} p_j \overset{(7)}{=} \frac{1}{\alpha} \langle q', C'p' \rangle \overset{(8)}{\leq} \frac{\|q'\|_1 \|p'\|_1}{\alpha} \|C'\|_\infty = \alpha \|C'\|_\infty \overset{(9)}{\leq} \epsilon \|C'\|_\infty$$

by (8) the proof of Lemma 7, (9) Cauchy-Schwartz, and (10) the above inequality (1).   ◄

▶ **Theorem 13.** *One can compute a δ-additive approximation to (T)*

1. *sequentially in $\tilde{O}\left(k\ell \dfrac{\|C\|_\infty}{\delta} \right)$ randomized time, and*

2. *in parallel with $\tilde{O}\left(\left(\dfrac{\|C\|_\infty}{\delta} \right)^2 \right)$ depth and $\tilde{O}\left(k\ell \left(\dfrac{\|C\|_\infty}{\delta} \right)^2 \right)$ total work.*

Proof. Let $\delta > 0$ be fixed. Let $\epsilon = \dfrac{\delta}{2\|C\|_\infty}$. By Lemma 11, we can compute a $(1 - \epsilon)$-transportation matrix X with cost at most $(\mathsf{T}) + \dfrac{\delta}{2}$. By Lemma 12, we can extend X to a transportation matrix U with additional cost of at most $\epsilon\|C\|_\infty = \dfrac{\delta}{2}$, for a total cost at most $(\mathsf{T}) + \delta$.   ◄

── **References** ──────────────

1 Zeyuan Allen Zhu and Lorenzo Orecchia. Nearly-Linear Time Positive LP Solver with Faster Convergence Rate. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 229–236, 2015.

2 Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1961–1971, 2017.

3 Jose Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards Optimal Running Times for Optimal Transport. *CoRR*, abs/1810.07717, 2018. `arXiv:1810.07717`.

4 Deeparnab Chakrabarty and Sanjeev Khanna. Better and Simpler Error Analysis of the Sinkhorn-Knopp Algorithm for Matrix Scaling. In *1st Symposium on Simplicity in Algorithms, SOSA 2018, January 7-10, 2018, New Orleans, LA, USA*, pages 4:1–4:11, 2018.

5 Chandra Chekuri and Kent Quanrud. Near-Linear Time Approximation Schemes for some Implicit Fractional Packing Problems. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 801–820, 2017.

───────────────

2 In fact, any transportation matrix from p' to q' will do.

6 Chandra Chekuri and Kent Quanrud. Randomized MWU for Positive LPs. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 358–377, 2018.

7 Michael B. Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix Scaling and Balancing via Box Constrained Newton's Method and Interior Point Methods. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 902–913, 2017.

8 Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2292–2300, 2013.

9 Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1366–1375, 2018.

10 Christos Koufogiannakis and Neal E. Young. A Nearly Linear-Time PTAS for Explicit Fractional Packing and Covering Linear Programs. *Algorithmica*, 70(4):648–674, 2014. Preliminary version in FOCS 2007.

11 Yin Tat Lee and Aaron Sidford. Path Finding Methods for Linear Programming: Solving Linear Programs in $\tilde{O}(\sqrt{rank})$ Iterations and Faster Algorithms for Maximum Flow. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 424–433, 2014.

12 Michael W. Mahoney, Satish Rao, Di Wang, and Peng Zhang. Approximating the Solution to Mixed Packing and Covering LPs in Parallel $\tilde{O}(\epsilon^{-3})$ time. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 52:1–52:14, 2016.

13 Jonah Sherman. Generalized Preconditioning and Undirected Minimum-Cost Flow. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 772–780, 2017.

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

15 Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

16 Cédric Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009.

17 Neal E. Young. Nearly Linear-Time Approximation Schemes for Mixed Packing/Covering and Facility-Location Linear Programs. *CoRR*, abs/1407.3015, 2014. `arXiv:1407.3015`.