Discrete Optimal Transport with Independent Marginals is #**P**-Hard

Bahar Taşkesen[†], Soroosh Shafieezadeh-Abadeh[‡], Daniel Kuhn[†], and Karthik Natarajan[§]

[†]Risk Analytics and Optimization Chair, EPFL Lausanne
bahar.taskesen,daniel.kuhn@epfl.ch
[‡]Tepper School of Business, CMU
sshafiee@andrew.cmu.edu
[§]Engineering Systems and Design, Singapore University of Technology and Design
karthik_natarajan@sutd.edu.sg

Abstract

We study the computational complexity of the optimal transport problem that evaluates the Wasserstein distance between the distributions of two K-dimensional discrete random vectors. The best known algorithms for this problem run in polynomial time in the maximum of the number of atoms of the two distributions. However, if the components of either random vector are independent, then this number can be exponential in K even though the size of the problem description scales linearly with K. We prove that the described optimal transport problem is #**P**-hard even if all components of the first random vector are independent uniform Bernoulli random variables, while the second random vector has merely two atoms, and even if only approximate solutions are sought. We also develop a dynamic programming-type algorithm that approximates the Wasserstein distance in pseudo-polynomial time when the components of the first random vector follow arbitrary independent discrete distributions, and we identify special problem instances that can be solved exactly in strongly polynomial time.

1. Introduction

Optimal transport theory is closely intertwined with probability theory and statistics [Boucheron et al., 2013, Villani, 2008] as well as with economics and finance [Galichon, 2016], and it has spurred fundamental research on partial differential equations [Benamou and Brenier, 2000, Brenier, 1991]. In addition, optimal transport problems naturally emerge in numerous application areas spanning machine learning [Arjovsky et al., 2017, Carriere et al., 2017, Rolet et al., 2016], signal processing [Ferradans et al., 2014, Kolouri and Rohde, 2015, Papadakis and Rabin, 2017, Tartavel et al., 2016], computer vision [Rubner et al., 2000, Solomon et al., 2014, 2015] and distributionally robust optimization [Blanchet and Murthy, 2019, Gao and Kleywegt, 2016, Mohajerin Esfahani and Kuhn, 2018]. For a comprehensive survey of modern applications of optimal transport theory we refer to [Kolouri et al., 2017, Peyré and Cuturi, 2019]. Historically, the first optimal transport problem was formulated by Gaspard Monge as early as in 1781 [Monge, 1781]. Monge's formulation aims at finding a measure-preserving map that minimizes some notion of transportation cost between two probability distributions, where all probability mass at a given origin location must be transported to the same target location. Due to this restriction, an optimal transportation map is not guaranteed to exist in general, and Monge's problem could be infeasible. In 1942, Leonid Kantorovich formulated a convex relaxation of Monge's problem by introducing the notion of a transportation plan that allows for mass splitting [Kantorovich, 1942]. Interestingly, an optimal transportation plan always exists. This paradigm shift has served as a catalyst for significant progress in the field.

In this paper we study Kantrovich's optimal transport problem between two discrete distributions

$$\mu = \sum_{i\in\mathcal{I}} \mu_i \delta_{\boldsymbol{x}_i} \quad \text{and} \quad \nu = \sum_{j\in\mathcal{J}} \nu_j \delta_{\boldsymbol{y}_j},$$

on $\mathbb{R}^K$, where $\boldsymbol{\mu} \in \mathbb{R}^I$ and $\boldsymbol{\nu} \in \mathbb{R}^J$ denote the probability vectors, whereas $\boldsymbol{x}_i \in \mathbb{R}^K$ for $i \in \mathcal{I} = \{1, \dots, I\}$ and $\boldsymbol{y}_j \in \mathbb{R}^K$ for $j \in \mathcal{J} = \{1, \dots, J\}$ represent the discrete support points of μ and ν, respectively. Throughout the paper we assume that μ and ν denote the probability distributions of two K-dimensional discrete random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Given a transportation cost function $c : \mathbb{R}^K \times \mathbb{R}^K \to [0, +\infty]$, we define the optimal transport distance between the discrete distributions μ and ν as

$$W_c(\mu, \nu) = \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} c(\boldsymbol{x}_i, \boldsymbol{y}_j)\pi_{ij}, \tag{1}$$

where $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\boldsymbol{\pi} \in \mathbb{R}_+^{I\times J} : \boldsymbol{\pi}\mathbf{1} = \boldsymbol{\mu},\ \boldsymbol{\pi}^\top \mathbf{1} = \boldsymbol{\nu}\}$ denotes the polytope of probability matrices $\boldsymbol{\pi}$ with marginal probability vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. Thus, every $\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ defines a discrete probability distribution

$$\pi = \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \pi_{ij} \delta_{(\boldsymbol{x}_i, \boldsymbol{y}_j)}$$

of $(\boldsymbol{x}, \boldsymbol{y})$ under which $\boldsymbol{x}$ and $\boldsymbol{y}$ have marginal distributions μ and ν, respectively. Distributions with these properties are referred to as transportation plans. If there exists $p \geq 1$ such that $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^K$, then $W_c(\mu, \nu)^{1/p}$ is termed the p-th Wasserstein distance between μ and ν. The optimal transport problem (1) constitutes a linear program that admits a strong dual linear program of the form

$$\begin{aligned} \max \quad & \boldsymbol{\mu}^\top \boldsymbol{\psi} + \boldsymbol{\nu}^\top \boldsymbol{\phi} \\ \text{s.t.} \quad & \boldsymbol{\psi} \in \mathbb{R}^I,\ \boldsymbol{\phi} \in \mathbb{R}^J \\ & \psi_i + \phi_j \leq c(\boldsymbol{x}_i, \boldsymbol{y}_j) \quad \forall i \in \mathcal{I}, j \in \mathcal{J}. \end{aligned}$$

Strong duality holds because $\boldsymbol{\pi} = \boldsymbol{\mu}\boldsymbol{\nu}^\top$ is feasible in (1) and the optimal value is finite. Both the primal and the dual formulations of the optimal transport problem can be solved exactly using the simplex algorithm [Dantzig, 1951], the more specialized network simplex algorithm [Orlin, 1997] or the Hungarian algorithm [Kuhn, 1955]. Both problems can also be addressed with dual ascent methods [Bertsimas and Tsitsiklis, 1997], customized auction algorithms [Bertsekas, 1981, 1992] or interior point methods [Karmarkar, 1984, Lee and Sidford, 2014, Nesterov and Nemirovskii, 1994]. More recently, the emergence of high-dimensional optimal transport problems in machine learning has spurred the development of efficient approximation algorithms. Many popular approaches for approximating the optimal transport distance between two discrete distributions rely on solving a regularized variant of problem (1). For instance, when augmented with an entropic regularizer, problem (1) becomes amenable to greedy methods such as the Sinkhorn algorithm [Sinkhorn, 1967, Cuturi, 2013] or the related Greenkhorn algorithm [Abid and Gower, 2018, Altschuler et al., 2017, Chakrabarty and Khanna, 2020], which run orders of magnitude faster than the exact methods. Other promising regularizers that have attracted significant interest include the Tikhonov [Blondel et al., 2018, Dessein et al., 2018, Essid and Solomon, 2018, Seguy et al., 2018], Lasso [Li et al., 2016], Tsallis entropy [Muzellec et al., 2017] and group Lasso regularizers [Courty et al., 2016]. In addition, Newton-type methods [Blanchet et al., 2018, Quanrud, 2019], quasi-Newton methods [Blondel et al., 2018], primal-dual gradient methods [Dvurechensky et al., 2018, Guo et al., 2020, Jambulapati et al., 2019, Lin et al., 2019b,a], iterative Bregman projections [Benamou et al., 2015] and stochastic average gradient descent algorithms [Genevay et al., 2016] are also used to find approximate solutions for discrete optimal transport problems.

The existing literature mainly addresses optimal transport problems between discrete distributions that are specified by enumerating the locations and the probabilities of the underlying atoms. In this case, the worst-case time-complexity of solving the linear program (1) with an interior point algorithm, say, grows polynomially with the problem's input description. In contrast, we focus here on optimal transport problems between discrete distributions supported on a number of points that grows *exponentially* with the dimension K of the sample space even though these problems admit an input description that scales only *polynomially* with K. In this case, the worst-case time-complexity of solving the linear program (1) directly with an interior point algorithm grows exponentially with the problem's input description. More precisely, we henceforth assume that μ is the distribution of a random vector $\boldsymbol{x} \in \mathbb{R}^K$ with independent components. Hence, μ is uniquely determined by the specification of its K marginals, which can be encoded using $\mathcal{O}(K)$ bits. Yet, even if each marginal has only two atoms, μ accommodates already 2^K atoms. Optimal transport problems involving such distributions are studied by Çelik et al. [2021] with the aim to find a discrete distribution with independent marginals that minimizes the Wasserstein distance from a prescribed discrete distribution. While Çelik et al. [2021] focus on solving small instances of this nonconvex problem, our results surprisingly imply that even evaluating this problem's objective function is hard. In summary, we are interested in scenarios where the discrete optimal transport problem (1) constitutes a linear program with exponentially many variables and constraints. We emphasize that such linear programs are not necessarily hard to solve [Grötschel et al., 2012], and therefore a rigorous complexity analysis is needed. We briefly review some useful computational complexity concepts next.

Recall that the complexity class **P** comprises all decision problems (*i.e.*, problems with a Yes/No answer) that can be solved in polynomial time. In contrast, the complexity class **NP** comprises all decision problems with the property that each 'Yes' instance admits a certificate that can be verified in polynomial time. A problem is **NP**-hard if every problem in **NP** is polynomial-time reducible to it, and an **NP**-hard problem is **NP**-complete if it belongs to **NP**. In this paper we will mainly focus on the complexity class #**P**, which encompasses all counting problems associated with decision problems in **NP** [Valiant, 1979a,b]. Loosely speaking, an instance of a #**P** problem thus counts the number of distinct polynomial-time verifiable certificates of the corresponding **NP** instance. Consequently, a #**P** problem is at least as hard as its **NP** counterpart, and #**P** problems cannot be solved in polynomial time unless #**P** coincides with the class **FP** of polynomial-time solvable function problems. A Turing reduction from a function problem A to a function problem B is an algorithm for solving problem A that has access to a fictitious oracle for solving problem B in one unit of time. Note that the oracle plays the role of a subroutine and may be called several times. A polynomial-time Turing reduction from A to B runs in time polynomial in the input size of A. We emphasize that, even though each oracle call requires only one unit of time, the time needed for computing all oracle inputs and reading all oracle outputs is attributed to the runtime of the Turing reduction. A problem is #**P**-hard if every problem in #**P** is polynomial-time Turing reducible to it, and a #**P**-hard problem is #**P**-complete if it belongs to #**P** [Valiant, 1979b, Jerrum, 2003].

Several hardness results for variants and generalizations of the optimal transport problem have recently been discovered. For example, multi-marginal optimal transport and Wasserstein barycenter problems were shown to be **NP**-hard [Altschuler and Boix-Adsera, 2020, 2021], whereas the problem of computing the Wasserstein distance between a continuous and a discrete distribution was shown to be #**P**-hard even in the simplest conceivable scenarios [Taskesen et al., 2021]. In this paper, we focus on optimal transport problems between two discrete distributions μ and ν. We formally prove that such problems are also #**P**-hard when μ and/or ν may have independent marginals. Specifically, we establish a fundamental limitation of all numerical algorithms for solving optimal transport problems between discrete distributions μ and ν, where μ has independent marginals. We show that, unless **FP** = #**P**, it is not possible to design an algorithm that approximates $W_c(\mu, \nu)$ in time polynomial in the bit length of the input size (which scales only polynomially with the dimension K) and the bit length $\log_2(1/\varepsilon)$ of the desired accuracy $\varepsilon > 0$. This result prompts us to look for algorithms that output ε-approximations in *pseudo-polynomial time*, that is, in time polynomial in the input size, the magnitude of the largest number in the input and the inverse accuracy $1/\varepsilon$. It also prompts us to look for special instances of the optimal transport problem with independent marginals that can be solved in *weakly* or *strongly polynomial time*. An algorithm runs in weakly polynomial time if it computes $W_c(\mu, \nu)$ in time polynomial in the bit length of the input. Similarly, an algorithm runs in strongly polynomial time if it computes $W_c(\mu, \nu)$ in time polynomial in the bit length of the input and if, in addition, it requires a number of arithmetic operations that grows at most polynomially with the dimension of the input (*i.e.*, the number of input numbers).

The key contributions of this paper can be summarized as follows.

- We prove that the discrete optimal transport problem with independent marginals is #**P**-hard even if μ represents the uniform distribution on the vertices of the K-dimensional hypercube and ν has only two support points, and even if only approximate solutions of polynomial bit length are sought (see Theorem 3.3).

- We demonstrate that the discrete optimal transport problem with independent marginals can be solved in strongly polynomial time by a dynamic programming-type algorithm whenever both μ and ν are supported on a fixed bounded subset of a scaled integer lattice with a fixed scaling factor—even if μ represents an arbitrary distribution with independent marginals (see Theorem 4.1). The design of this algorithm reveals an intimate connection between optimal transport and the conditional value-at-risk arising in risk measurement.

- Using a rounding scheme to approximate μ and ν by distributions $\tilde{\mu}$ and $\tilde{\nu}$ supported on a scaled integer lattice with a sufficiently small grid spacing constant, we show that ε-accurate approximations of the optimal transport distance between μ and ν can always be computed in pseudo-polynomial time via dynamic programming (see Theorem 4.9). This result implies that the optimal transport problem with independent marginals is in fact #**P**-hard in the weak sense [Garey and Johnson, 1979, § 4].

Our complexity analysis complements existing hardness results for two-stage stochastic programming problems. Indeed, Dyer and Stougie [2006, 2015], Hanasusanto et al. [2016] and Dhara et al. [2021] show that computing optimal first-stage decisions of linear two-stage stochastic programs and evaluating the corresponding expected costs is hard if the uncertain problem parameters follow independent (discrete or continuous) distributions. This paper establishes similar hardness results for discrete optimal transport problems. Our paper also complements the work of Genevay et al. [2016], who describe a stochastic gradient descent method for computing ε-optimal transportation plans in $\mathcal{O}(1/\varepsilon^2)$ iterations. Their method can in principle be applied to the discrete optimal transport problems with independent marginals studied here. However, unlike our pseudo-polynomial time dynamic programming-based algorithm, their method is non-deterministic and does not output an approximation of the optimal transport distance $W_c(\mu, \nu)$.

The remainder of this paper is structured as follows. In Section 2 we review a useful #**P**-hardness result for a counting version of the knapsack problem. By reducing this problem to the optimal transport problem with independent marginals, we prove in Section 3 that the latter problem is also #**P**-hard even if only approximate solutions are sought. In Section 4 we develop a dynamic programming-type algorithm that computes approximations of the optimal transport distance in pseudo-polynomial time, and we identify special problem instances that can be solved exactly in strongly polynomial time.

Notation. We use boldface letters to denote vectors and matrices. The vectors of all zeros and ones are denoted by $\mathbf{0}$ and $\mathbf{1}$, respectively, and their dimensions are always clear from the context. The calligraphic letters $\mathcal{I}, \mathcal{J}, \mathcal{K}$ and $\mathcal{L}$ are reserved for finite index sets with cardinalities I, J, K and L, that is, $\mathcal{I} = \{1, \dots, I\}$ etc. We denote by $\|\cdot\|$ the 2-norm, and for any $\boldsymbol{x} \in \mathbb{R}^K$ we use $\delta_{\boldsymbol{x}}$ to denote the Dirac distribution at $\boldsymbol{x}$.

2. A Counting Version of the Knapsack Problem

Counting the number of feasible solutions of a 0/1 knapsack problem is a seemingly simple but surprisingly challenging task. Formally, the problem of interest is stated as follows.

#Knapsack

Instance. A list of items with weights $w_k \in \mathbb{Z}_+$, $k \in \mathcal{K}$, and a capacity $b \in \mathbb{Z}_+$.

Goal. Count the number of subsets of the items whose total weight is at most b.

The #Knapsack problem is known to be #**P**-complete [Dyer et al., 1993], and thus it admits no polynomial-time algorithm unless **FP** = #**P**. Dyer et al. [1993] discovered a randomized sub-exponential time algorithm that provides almost correct solutions with high probability by sampling feasible solutions using a random walk. By relying on a rapidly mixing Markov chain, Morris and Sinclair [2004] then developed the first fully polynomial randomized approximation scheme. Later, Dyer [2003] interweaved dynamic programming and rejection sampling approaches to obtain a considerably simpler fully polynomial randomized approximation scheme. However, randomization remains essential in this approach. Deterministic dynamic programming-based algorithms were developed more recently by Gopalan et al. [2011], and Štefankovič et al. [2012]. In the next section we will demonstrate that a certain class of discrete optimal transport problems with independent marginals is at least as hard as the #Knapsack problem.

3. Optimal Transport with Independent Marginals

Consider now a variant of the optimal transport problem (1), where the discrete multivariate distribution $\mu = \otimes_{k \in \mathcal{K}} \mu_k$ is a product of K independent univariate marginal distributions $\mu_k = \sum_{l \in \mathcal{L}} \mu_k^l \delta_{x_k^l}$ with support points $x_k^l \in \mathbb{R}$ and corresponding probabilities μ_k^l for every $l \in \mathcal{L}$. This implies that μ accommodates a total of $I = L^K$ support points. The assumption that each μ_k, $k \in \mathcal{K}$, accommodates the same number L of support points simplifies notation but can be imposed without loss of generality. Indeed, the probability of any unneeded support point can be set to zero. The other discrete multivariate distribution $\nu = \sum_{j \in \mathcal{J}} \nu_j \delta_{\boldsymbol{y}_j}$

has no special structure. Assume for the moment that all components of the support points as well as all probabilities of μ_k, $k \in \mathcal{K}$, and ν are rational numbers and thus representable as ratios of two integers, and denote by U the maximum absolute numerical value among all these integers, which can be encoded using $\mathcal{O}(\log_2 U)$ bits. Thus, the total number of bits needed to represent the discrete distributions μ and ν is bounded above by $\mathcal{O}(\max\{KL, J\} \log_2 U)$. Note that this encoding does *not* require an explicit enumeration of the locations and probabilities of the $I = L^K$ atoms of the distribution μ. It is well known that the linear program (1) can be solved in polynomial time by the ellipsoid method, for instance, if μ is encoded by such an inefficient exhaustive enumeration, which requires up to $\mathcal{O}(\max\{I, J\} \log_2 U)$ input bits. Thus, the runtime of the ellipsoid method scales at most polynomially with I, J and $\log_2 U$. As $I = L^K$ grows exponentially with K, however, this does *not* imply tractability of the optimal transport problem at hand, which admits an efficient encoding that scales only linearly with K. In the remainder of this paper we will prove that the optimal transport problem with independent maringals is #**P**-hard, and we will identify special problem instances that can be solved efficiently.

In order to prove #**P**-hardness, we focus on the following subclass of optimal transport problems with independent marginals, where μ is the uniform distribution on $\{0, 1\}^K$, and ν has only two support points.

#Optimal Transport

Instance. Two support points $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^K$, $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$, and a probability $t \in [0, 1]$.

Goal. For μ denoting the uniform distribution on $\{0, 1\}^K$ and $\nu = t\delta_{\boldsymbol{y}_1} + (1 - t)\delta_{\boldsymbol{y}_2}$, compute an approximation $\widetilde{W}_c(\mu, \nu)$ of $W_c(\mu, \nu)$ for $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ such that the following hold.

(i) The bit length of $\widetilde{W}_c(\mu, \nu)$ is polynomially bounded in the bit length of the input $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$.

(ii) We have $|\widetilde{W}_c(\mu, \nu) - W_c(\mu, \nu)| \leq \overline{\varepsilon}$, where

$$\overline{\varepsilon} = \frac{1}{4I} \min \Big\{ \big| \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p \big| : i \in \mathcal{I},\ \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p \neq 0 \Big\}$$

with $I = 2^K$ and $\boldsymbol{x}_i$, $i \in \mathcal{I}$, representing the different binary vectors in $\{0, 1\}^K$.

We first need to show that the #Optimal Transport problem is well-posed, that is, we need to ascertain the existence of a sufficiently accurate approximation that can be encoded in a polynomial number of bits. To this end, we first prove that the maximal tolerable approximation error $\overline{\varepsilon}$ is not too small.

Lemma 3.1. There exists $\varepsilon \in (0, \overline{\varepsilon}]$ whose bit length is polynomially bounded in the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$.

Proof. Note first that encoding an instance of the #Optimal Transport problem requires at least K bits because the K coordinates of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ need to be enumerated. Note also that, by the definition of $\overline{\varepsilon}$, there exists an index $i^\star \in \mathcal{I}$ with $\overline{\varepsilon} = \frac{1}{4I} \big| \|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_2\|^p \big|$. We prove the claim first under the assumption that p is even, that is $p = 2q$ for some $q \in \mathbb{N}$. In this case, we have

$$\overline{\varepsilon} = \frac{1}{4I} \left| \Big((\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1)^\top (\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1) \Big)^q - \Big((\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1)^\top (\boldsymbol{x}_{i^\star} - \boldsymbol{y}_2) \Big)^q \right|.$$

Recalling that p is not an input of the #Optimal Transport problem and that q must therefore be treated as a constant, the absolute value in the last expression can be computed in polynomial time because it involves only $\mathcal{O}(K)$ additions and multiplications. Similarly, the evaluation of the denominator $4I = 2^{K+2}$ requires $\mathcal{O}(K)$ multiplications. Overall, $\overline{\varepsilon}$ can therefore be computed in polynomial time, which trivially implies that the bit length of $\overline{\varepsilon}$ is polynomially bounded. We may thus set $\varepsilon = \overline{\varepsilon}$.

Assume now that p is odd, that is, $p = 2q - 1$ for some $q \in \mathbb{N}$. In this case $\|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1\|^p$ and $\|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_2\|^p$ may be irrational numbers that cannot be encoded with any finite number of bits even if the vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ have only rational entries. Thus, $\overline{\varepsilon}$ may also be irrational, in which case we need to construct $\varepsilon < \overline{\varepsilon}$. To simplify notation, we henceforth use the shorthands $a = \|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1\|^2$ and $b = \|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_2\|^2$, which can be computed in polynomial time using $\mathcal{O}(K)$ additions and multiplications. If $a, b \geq 1$, then we have

$$\overline{\varepsilon} = \frac{1}{2^{K+2}} \left| a^{p/2} - b^{p/2} \right| = \frac{1}{2^{K+2}} \left| \frac{a^p - b^p}{a^{p/2} + b^{p/2}} \right| > \frac{1}{2^{K+2}} \left| \frac{a^p - b^p}{a^q + b^q} \right| \triangleq \varepsilon > 0,$$

where the first inequality holds because $p/2 < q$. The tolerance ε constructed in this way can be computed via $\mathcal{O}(K)$ additions and multiplications, and therefore its bitlengh is polynomially bounded. If $a \geq 1 \geq b$, $a \leq 1 \leq b$ or $a, b \leq 1$, then ε can be constructed in a similar manner. Details are omitted for brevity. □

Lemma 3.1 readily implies that for any instance of the #Optimal Transport problem there exists an approximate optimal transport distance $\widetilde{W}_c(\mu, \nu)$ that satisfies both conditions (i) as well as (ii). For example, we could construct $\widetilde{W}_c(\mu, \nu)$ by rounding the exact optimal transport distance $W_c(\mu, \nu)$ to the nearest multiple of ε. By construction, this approximation differs from $W_c(\mu, \nu)$ at most by ε, which is itself not larger than $\overline{\varepsilon}$. In addition, this approximation trivially inherits the polynomial bit length from ε. We emphasize that, in general, $\widetilde{W}_c(\mu, \nu)$ cannot be set to the exact optimal transport distance $W_c(\mu, \nu)$, because $W_c(\mu, \nu)$ may be irrational and thus have infinite bit length. However, Corollary 3.5 below implies that if p is even, then $\widetilde{W}_c(\mu, \nu) = W_c(\mu, \nu)$ satisfies both conditions (i) as well as (ii).

Note that the #Optimal Transport problem is parametrized by p. The following example shows that if p was treated as an input parameter, then the problem would have exponential time complexity.

Example 3.2. Consider an instance of the #Optimal Transport problem with $K = 1$, $y_1 = 1$, $y_2 = 2$ and $t = 0$. In this case we have $\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, $\nu = \delta_2$ and $\overline{\varepsilon} = \frac{1}{8}$. An elementary analytical calculation reveals that $W_c(\mu, \nu) = \frac{1}{2}(1 + 2^p)$. The bit length of any $\overline{\varepsilon}$-approximation $\widetilde{W}_c(\mu, \nu)$ of $W_c(\mu, \nu)$ is therefore bounded below by $\log_2(\frac{1}{2}(1 + 2^p) - \frac{1}{8}) \geq p - 1$, which grows exponentially with the bit length $\log_2(p)$ of p. Note that the time needed for computing $\widetilde{W}_c(\mu, \nu)$ is at least as large as its own bit length irrespective of the algorithm that is used. If p was an input parameter of the #Optimal Transport problem, the problem's worst-case time complexity would therefore grow at least exponentially with its input size.

The following main theorem shows that the #Optimal Transport problem is hard even if $p = 2$.

Theorem 3.3 (Hardness of #Optimal Transport)**.** #Optimal Transport is #**P**-hard even if $p = 2$.

We prove Theorem 3.3 by reducing the #Knapsack problem to the #Optimal Transport problem via a polynomial-time Turing reduction. To this end, we fix an instance of the #Knapsack problem with input $\boldsymbol{w} \in \mathbb{Z}_+^K$ and $b \in \mathbb{Z}_+$, and we denote by $\nu_t = t\delta_{\boldsymbol{y}_1} + (1-t)\delta_{\boldsymbol{y}_2}$ the two-point distribution with support points $\boldsymbol{y}_1 = \boldsymbol{0}$ and $\boldsymbol{y}_2 = 2b\boldsymbol{w}/\|\boldsymbol{w}\|^2$, whose probabilities are parameterized by $t \in [0, 1]$. Recall also that μ is the uniform distribution on $\{0, 1\}^K$, that is, $\mu = \frac{1}{I}\sum_{i \in \mathcal{I}} \delta_{\boldsymbol{x}_i}$, where $I = 2^K$ and $\{\boldsymbol{x}_i : i \in \mathcal{I}\} = \{0, 1\}^K$. Without loss of generality, we may assume that the support points of μ are ordered so as to satisfy

$$\|\boldsymbol{x}_1 - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_1 - \boldsymbol{y}_2\|^p \leq \|\boldsymbol{x}_2 - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_2 - \boldsymbol{y}_2\|^p \leq \cdots \leq \|\boldsymbol{x}_I - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_I - \boldsymbol{y}_2\|^p.$$

Below we will demonstrate that computing $W_c(\mu, \nu_t)$ approximately is at least as hard as solving the #Knapsack problem, which amounts to evaluating the cardinality of $\mathcal{I}(\boldsymbol{w}, b) = \{\boldsymbol{x} \in \{0, 1\}^K : \boldsymbol{w}^\top \boldsymbol{x} \leq b\}$.

Lemma 3.4. If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ for some $p \geq 1$, then the optimal transport distance $W_c(\mu, \nu_t)$ is continuous, piecewise affine and convex in $t \in [0, 1]$. Moreover, it admits the closed-form formula

$$W_c(\mu, \nu_t) = \frac{1}{I}\sum_{i=1}^{\lfloor tI \rfloor} \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p + \frac{1}{I}\sum_{i=\lfloor tI \rfloor + 1}^{I} \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p \\ + \frac{(tI - \lfloor tI \rfloor)}{I}\left(\|\boldsymbol{x}_{\lfloor tI \rfloor + 1} - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_{\lfloor tI \rfloor + 1} - \boldsymbol{y}_2\|^p\right). \tag{2}$$

Proof. For any fixed $t \in [0, 1]$, the discrete optimal transport problem (1) satisfies

$$\begin{aligned} W_c(\mu, \nu_t) &= \min_{\pi \in \Pi(\mu, \nu_t)} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^p \pi_{ij} \\ &= \begin{cases} \min\limits_{\boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathbb{R}^I_+} & t\sum_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p q_{1,i} + (1-t)\sum_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p q_{2,i} \\ \text{s.t.} & t\boldsymbol{q}_1 + (1-t)\boldsymbol{q}_2 = \boldsymbol{1}/I, \; \boldsymbol{1}^\top \boldsymbol{q}_1 = 1, \; \boldsymbol{1}^\top \boldsymbol{q}_2 = 1. \end{cases} \end{aligned}$$

237 The second equality holds because the transportation plan can be expressed as

$$\pi = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \pi_{ij} \delta_{(\boldsymbol{x}_i, \boldsymbol{y}_j)} = t \cdot q_1 \otimes \delta_{\boldsymbol{y}_1} + (1-t) \cdot q_2 \otimes \delta_{\boldsymbol{y}_2},$$

with $q_j = \sum_{i \in \mathcal{I}} q_{j,i} \delta_{\boldsymbol{x}_i}$ representing the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y} = \boldsymbol{y}_j$ under π for every $j = 1, 2$. This is a direct consequence of the law of total probability. By applying the variable transformations $\boldsymbol{q}_1 \leftarrow tI\boldsymbol{q}_1$ and $\boldsymbol{q}_2 \leftarrow (1-t)I\boldsymbol{q}_2$ to eliminate all bilinear terms, we then find

$$W_c(\mu, \nu_t) = \begin{cases} \min\limits_{\boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathbb{R}^I_+} & \frac{1}{I}\sum_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p q_{1,i} + \frac{1}{I}\sum_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p q_{2,i} \\ \text{s.t.} & \boldsymbol{1}^\top \boldsymbol{q}_1 = tI, \; \boldsymbol{1}^\top \boldsymbol{q}_2 = (1-t)I, \; \boldsymbol{q}_1 + \boldsymbol{q}_2 = \boldsymbol{1}. \end{cases} \tag{3}$$

Observe that (3) can be viewed as a parametric linear program. By [Dantzig and Thapa, 2003, Theorem 6.6], its optimal value $W_c(\mu, \nu_t)$ thus constitutes a continuous, piecewise affine and convex function of t. It remains to be shown that $W_c(\mu, \nu_t)$ admits the analytical expression (2). To this end, note that the decision variable $\boldsymbol{q}_2$ and the constraint $\boldsymbol{q}_1 + \boldsymbol{q}_2 = \boldsymbol{1}$ in problem (3) can be eliminated by applying the substitution $\boldsymbol{q}_2 \leftarrow \boldsymbol{1} - \boldsymbol{q}_1$. Renaming $\boldsymbol{q}_1$ as $\boldsymbol{q}$ to reduce clutter, problem (3) then simplifies to

$$\begin{aligned} \min_{\boldsymbol{q} \in \mathbb{R}^I} \quad & \frac{1}{I}\sum_{i \in \mathcal{I}} \left(\|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p\right) q_i + \frac{1}{I}\sum_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p \\ \text{s.t.} \quad & \boldsymbol{1}^\top \boldsymbol{q} = tI, \; \boldsymbol{0} \leq \boldsymbol{q} \leq \boldsymbol{1}. \end{aligned} \tag{4}$$

Recalling that the atoms of μ are ordered such that $\|\boldsymbol{x}_1 - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_1 - \boldsymbol{y}_2\|^p \leq \cdots \leq \|\boldsymbol{x}_I - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_I - \boldsymbol{y}_2\|^p$, one readily verifies that problem (4) is solved analytically by

$$q_i^\star = \begin{cases} 1 & \text{if } i \leq \lfloor tI \rfloor \\ tI - \lfloor tI \rfloor & \text{if } i = \lfloor tI \rfloor + 1 \\ 0 & \text{if } i > \lfloor tI \rfloor + 1. \end{cases}$$

238 Substituting $\boldsymbol{q}^\star$ into (4) yields (2), and thus the claim follows. □

239 Lemma 3.4 immediately implies that the bit length of $W_c(\mu, \nu_t)$ is polynomially bounded.

Corollary 3.5. If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ and p is even, then the bit length of the optimal transport distance $W_c(\mu, \nu_t)$ grows at most polynomially with the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$.

Proof. The bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$ is finite if and only if all of its components are rational and thus representable as ratios of two integers. We denote by $U \in \mathbb{N}$ the maximum absolute value of these integers.

For ease of exposition, we assume first that $p = 2$ and $t = 1$. In addition, we use $D \in \mathbb{N}$ to denote the least common multiple of the denominators of the K components of $\boldsymbol{y}_1$. It is easy to see that $D \leq U^K$. By Lemma 3.4, the optimal transport distance $W_c(\mu, \nu_t)$ can thus be expressed as the average of the I quadratic terms $\|\boldsymbol{x}_i - \boldsymbol{y}_1\|^2 = \boldsymbol{x}_i^\top \boldsymbol{x}_i + 2\boldsymbol{x}_i^\top \boldsymbol{y}_1 + \boldsymbol{y}_1^\top \boldsymbol{y}_1$ for $i \in \mathcal{I}$. Each such term is equivalent to a rational number with denominator D^2 and a numerator that is bounded above by $K(1 + 2U + U^2)D^2$. Indeed, each component of $\boldsymbol{x}_i$ is binary, whereas each component of $\boldsymbol{y}_1$ can be expressed as a rational number with denominator D and a numerator with absolute value at most UD. By Lemma 3.4, $W_c(\mu, \nu_t)$ is thus representable as a rational number with denominator ID^2 and a numerator with absolute value at most $IK(1 + U)^2D^2$. Therefore, the number of bits needed to encode $W_c(\mu, \nu_t)$ is at most of the order

$$\mathcal{O}\left(\log_2(IKU^2D^2)\right) \leq \mathcal{O}\left(\log_2(2^K KU^2U^{2K})\right) = \mathcal{O}\left(K \log_2(U)\right),$$

where the inequality holds because $I = 2^K$ and $D \leq U^K$. As both K and $\log_2(U)$ represent lower bounds on the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$, we have thus shown that the bit length of $W_c(\mu, \nu_t)$ is indeed polynomially bounded in the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$. If p is any even number and t any rational probability, then the claim can be proved using similar—yet more tedious—arguments. Details are omitted for brevity. □

Corollary 3.5 implies that the optimal transport distance $W_c(\mu, \nu_t)$ is rational whenever p is an even integer and t is rational. Otherwise, $W_c(\mu, \nu_t)$ is generically irrational because the Euclidean norm of a vector $\boldsymbol{v} = (v_1, \ldots, v_K)$ is irrational unless $(v_1, \ldots, v_K, \|\boldsymbol{v}\|)$ is proportional to a Pythagorean $(K+1)$-tuple, where the inverse proportionality factor is itself equal to the square of an integer. We will now show that the cardinality of the set $\mathcal{I}(\boldsymbol{w}, b)$ can be computed by solving the univariate minimization problem

$$\min_{t \in [0,1]} W_c(\mu, \nu_t). \tag{5}$$

This insight is formalized in the next lemma.

Lemma 3.6. If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ for some $p \geq 1$, then $t^\star = |\mathcal{I}(\boldsymbol{w}, b)|/I$ is an optimal solution of problem (5). If in addition each component of $\boldsymbol{w}$ is even and b is odd, then $t^\star$ is unique.

Proof. From the proof of Lemma 3.4 we know that the optimal transport distance $W_c(\mu, \nu_t)$ coincides with the optimal value of (3). Thus, problem (5) can be reformulated as

$$\begin{array}{cl} \min\limits_{\substack{t \in [0,1] \\ \boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathbb{R}^I_+}} & \dfrac{1}{I} \sum\limits_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p q_{1,i} + \dfrac{1}{I} \sum\limits_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p \, q_{2,i} \\ \text{s.t.} & \mathbf{1}^\top \boldsymbol{q}_1 = tI, \;\; \mathbf{1}^\top \boldsymbol{q}_2 = (1-t)I, \;\; \boldsymbol{q}_1 + \boldsymbol{q}_2 = \mathbf{1}. \end{array} \tag{6}$$

Note that the decision variable t as well as the two normalization constraints for $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are redundant and can thus be removed without affecting the optimal value of (6). In other words, there always exists $t \in [0, 1]$ such that $\mathbf{1}^\top \boldsymbol{q}_1 = tI$ and $\mathbf{1}^\top \boldsymbol{q}_2 = (1 - t)I$. Hence, (6) simplifies to

$$\begin{array}{cl} \min\limits_{\boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathbb{R}^I_+} & \dfrac{1}{I} \sum\limits_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p q_{1,i} + \dfrac{1}{I} \sum\limits_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p \, q_{2,i} \\ \text{s.t.} & \boldsymbol{q}_1 + \boldsymbol{q}_2 = \mathbf{1}. \end{array} \tag{7}$$

Next, introduce the disjoint index sets

$$\begin{aligned}\mathcal{I}_0 &= \{i \in \mathcal{I} : \|\boldsymbol{x}_i - \boldsymbol{y}_1\| = \|\boldsymbol{x}_i - \boldsymbol{y}_2\|\}\\ \mathcal{I}_1 &= \{i \in \mathcal{I} : \|\boldsymbol{x}_i - \boldsymbol{y}_1\| < \|\boldsymbol{x}_i - \boldsymbol{y}_2\|\}\\ \mathcal{I}_2 &= \{i \in \mathcal{I} : \|\boldsymbol{x}_i - \boldsymbol{y}_1\| > \|\boldsymbol{x}_i - \boldsymbol{y}_2\|\},\end{aligned}$$

which form a partition of $\mathcal{I}$. Using these sets, optimal solution of problem (7) can be expressed as

$$q^\star_{1,i} = \begin{cases}\theta_i & \text{if } i \in \mathcal{I}_0\\ 1 & \text{if } i \in \mathcal{I}_1\\ 0 & \text{if } i \in \mathcal{I}_2\end{cases} \quad \text{and} \quad q^\star_{2,i} = \begin{cases}1-\theta_i & \text{if } i \in \mathcal{I}_0\\ 0 & \text{if } i \in \mathcal{I}_1\\ 1 & \text{if } i \in \mathcal{I}_2\end{cases} \tag{8}$$

Therefore, we have

$$\min_{t\in[0,1]} W_c(\mu,\nu_t) = \frac{1}{I}\sum_{i\in\mathcal{I}} \min\left\{\|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p, \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p\right\}.$$

Any minimizer $(\boldsymbol{q}_1^\star, \boldsymbol{q}_2^\star)$ of (7) gives thus rise to a minimizer $(t^\star, \boldsymbol{q}_1^\star, \boldsymbol{q}_2^\star)$ of (6), where $t^\star = (\mathbf{1}^\top \boldsymbol{q}_1^\star)/I$. Moreover, the minimizers of (5) are exactly all numbers of the form $t^\star = (\mathbf{1}^\top \boldsymbol{q}_1^\star)/I$ corresponding to the minimizer $(\boldsymbol{q}_1^\star, \boldsymbol{q}_2^\star)$ of (7). In view of (8), this observation allows us to conclude that

$$\operatorname*{argmin}_{t\in[0,1]} W_c(\mu,\nu_t) = [|\mathcal{I}_1|/I, |\mathcal{I}_0 \cup \mathcal{I}_1|/I]\,. \tag{9}$$

By the definitions of $\mathcal{I}(\boldsymbol{w}, b)$, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, it is further evident that

$$|\mathcal{I}(\boldsymbol{w},b)| = \left|\left\{i\in\mathcal{I} : \boldsymbol{w}^\top\boldsymbol{x}_i \le b\right\}\right| = \left|\left\{i\in\mathcal{I} : \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^2 \le \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^2\right\}\right| = |\mathcal{I}_1 \cup \mathcal{I}_0|.$$

Therefore, we may finally conclude that

$$|\mathcal{I}(\boldsymbol{w},b)|/I \in \operatorname*{argmin}_{t\in[0,1]} W_c(\mu,\nu_t).$$

Assume now that each component of $\boldsymbol{w}$ is even and b is odd. In this case, there exists no $\boldsymbol{x} \in \{0,1\}^K$ that satisfies $\boldsymbol{x}^\top\boldsymbol{w} = b$ and consequentially $\mathcal{I}_0$ is empty. Consequently, the interval of minimizers in (9) collapses to the singleton $|\mathcal{I}_1|/I = |\mathcal{I}(\boldsymbol{w},b)|/I$. This observation completes the proof. □

Armed with Lemmas 3.4 and 3.6, we are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. Select an instance of the #KNAPSACK problem with input $\boldsymbol{w} \in \mathbb{Z}_+^K$ and $b \in \mathbb{Z}_+$. Throughout this proof we will assume without loss of generality that each component of $\boldsymbol{w}$ is even and that b is odd. Indeed, if this was not the case, we could replace $\boldsymbol{w}$ with $\boldsymbol{w}' = 2\boldsymbol{w}$ and b with $b' = 2b+1$. It is easy to verify that the two instances of the #KNAPSACK problem with inputs $(\boldsymbol{w}, b)$ and $(\boldsymbol{w}', b')$ have the same solution. In addition, the bit length of $(\boldsymbol{w}', b')$ is polynomially bounded in the bit length of $(\boldsymbol{w}, b)$.

Given $\boldsymbol{w}$ and b, define the distributions μ and ν_t for $t \in [0,1]$ as well as the set $\mathcal{I}(\boldsymbol{w}, b)$ in the usual way. From Lemma 3.4 we know that $W_c(\mu,\nu_t)$ is continuous, piecewise affine and convex in t. The analytical formula (2) further implies that $W_c(\mu,\nu_t)$ is affine on the interval $[(i-1)/I, i/I]$ with slope a_i/I, where

$$a_i = W_c(\mu,\nu_{i/I}) - W_c(\mu,\nu_{(i-1)/I}) \qquad \forall i \in \mathcal{I}. \tag{10}$$

Thus, (5) constitutes a univariate convex optimization problem with a continuous piecewise affine objective function. As each component of $\boldsymbol{w}$ is even and b is odd, Lemma 3.6 implies that $t^\star = |\mathcal{I}(\boldsymbol{w},b)|/I$ is the unique minimizer of (5). Therefore, the given instance of the #Knapsack problem can be solved by solving (5) and multiplying its unique minimizer $t^\star$ with I.

In the following we will first show that if we had access to an oracle that computes $W_c(\mu,\nu_t)$ exactly, then we could construct an algorithm that finds $t^\star$ and the solution $t^\star I$ of the #Knapsack problem by calling the oracle $2K$ times (Step 1). Next, we will prove that if we had access to an oracle that solves the #Optimal Transport problem and thus outputs only approximations of $W_c(\mu,\nu_t)$, then we could extend the algorithm from Step 1 to a polynomial-time Turing reduction from the #Knapsack problem to the #Optimal Transport problem (Step 2). Step 2 implies that #Optimal Transport is #**P**-hard.

Step 1. Assume now that we have access to an oracle that computes $W_c(\mu,\nu_t)$ exactly. In addition, introduce an array $\boldsymbol{a} = (a_0, a_1, \ldots, a_I)$ with entries a_i, $i \in \mathcal{I}$, defined as in (10) and with $a_0 = -\infty$. Thus, each element of $\boldsymbol{a}$ can be evaluated with at most two oracle calls. The array $\boldsymbol{a}$ is useful because it contains all the information that is needed to solve the univariate convex optimization problem (5). Indeed, as $W_c(\mu,\nu_t)$ is a convex piecewise linear function with slope a_i/I on the interval $[i/I,(i-1)/I]$, the array $\boldsymbol{a}$ is sorted in ascending order, and the unique minimizer $t^\star$ of (5) satisfies

$$|\mathcal{I}(\boldsymbol{w},b)| = t^\star I = \max\left\{i \in \mathcal{I}\cup\{0\} : a_i \leq 0\right\}. \tag{11}$$

In other words, counting all elements of the set $\mathcal{I}(\boldsymbol{w},b)$ and thereby solving the #Knapsack problem is equivalent to finding the maximum index $i \in \mathcal{I}\cup\{0\}$ that meets the condition $a_i \leq 0$. The binary search method detailed in Algorithm 1 efficiently finds this index. Binary search methods are also referred to as half-interval search or bisection algorithms, and they represent iterative methods for finding the largest number within a sorted array that is smaller or equal to a given threshold (0 in our case). Algorithm 1 first checks whether the number in the middle of the array is non-positive. Depending on the outcome, either the part of the array to the left or to the right of the middle element may be discarded because the array is sorted. This procedure is repeated until the array collapses to the single element corresponding to the sought number. As the length of the array is halved in each iteration, the binary search method applied to an array of length I returns the solution in $\log_2 I = K$ iterations [Cormen et al., 2009, § 12].

Algorithm 1 Binary search method

Input: An array $\boldsymbol{a} \in \mathbb{R}^I$ with $I = 2^K$ sorted in ascending order

```
Initialize n̲ = 0 and n̄ = I
for k = 1, ..., K do
    Set n ← (n̄ + n̲)/2
    if a_n ≤ 0 then n̲ ← n else n̄ ← n
end for
if a_n̲ ≤ 0 then n ← n̲ else n ← n̄
```

Output: n

One can use induction to show that, in any iteration k of Algorithm 1, n is given by a multiple of 2^{K-k} and represents indeed an eligible index. Similarly, in any iteration k we have $\overline{n} - \underline{n} = 2^{K-k+1}$.

Step 2. Assume now that we have only access to an oracle that solves the #Optimal Transport problem, which merely returns an approximation $\widetilde{W}_c(\mu,\nu_t)$ of $W_c(\mu,\nu_t)$. Setting $\widetilde{a}_0 = -\infty$ and

$$\widetilde{a}_i = \widetilde{W}_c(\mu,\nu_{i/I}) - \widetilde{W}_c(\mu,\nu_{(i-1)/I}) \qquad \forall i \in \mathcal{I}, \tag{12}$$

we can then introduce a perturbed array $\widetilde{\boldsymbol{a}} = (\widetilde{a}_0, \widetilde{a}_1, \ldots, \widetilde{a}_I)$ which provides an approximation for $\boldsymbol{a}$. In the following we will prove that, even though $\widetilde{\boldsymbol{a}}$ is no longer necessarily sorted in ascending order, the sign of $\widetilde{a}_i$ coincides with the sign of a_i for every $i \in \mathcal{I}$. Algorithm 1 therefore outputs the exact solution $|\mathcal{I}(\boldsymbol{w}, b)|$ of the #Knapsack problem even if its input $\boldsymbol{a}$ is replaced with $\widetilde{\boldsymbol{a}}$. To see this, we first note that

$$a_i = \frac{1}{I}\left(\|\boldsymbol{x}_i - \boldsymbol{y}_1\|^2 - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^2\right) \qquad \forall i \in \mathcal{I}, \tag{13}$$

which is an immediate consequence of the analytical formula (2) for $W_c(\mu, \nu_t)$. We emphasize that (13) has only theoretical relevance but cannot be used to evaluate a_i in practice because it relies on our assumption that the support points $\boldsymbol{x}_i$, $i \in \mathcal{I}$, are ordered such that $\|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p$ is non-decreasing in i. Indeed, there is no efficient algorithm for ordering these 2^K points in practice. Using (13), we then find

$$\overline{\varepsilon} = \frac{1}{4}\min_{i\in\mathcal{I}}\{|a_i| : a_i \neq 0\} = \frac{1}{4}\min_{i\in\mathcal{I}}|a_i|,$$

where the first equality follows from the definition of $\overline{\varepsilon}$, and the second equality holds because each component of $\boldsymbol{w}$ is even and b is odd, which implies that $\|\boldsymbol{x}_i - \boldsymbol{y}_1\| \neq \|\boldsymbol{x}_i - \boldsymbol{y}_2\|$ and thus $a_i \neq 0$ for all $i \in \mathcal{I}$. The last formula for $\overline{\varepsilon}$ immediately implies that $|a_i| \geq 2\overline{\varepsilon}$ for all $i \in \mathcal{I}$. Together with the estimate

$$|\widetilde{a}_i - a_i| \leq \left|\widetilde{W}_c(\mu, \nu_{i/I}) - W_c(\mu, \nu_{i/I})\right| + \left|\widetilde{W}_c(\mu, \nu_{(i-1)/I}) - W_c(\mu, \nu_{(i-1)/I})\right| \leq 4\overline{\varepsilon},$$

this implies that $\widetilde{a}_i$ has indeed the same sign as a_i for every $i \in \mathcal{I}$. As the execution of Algorithm 1 depends on the input array only through the signs of its components, Algorithm 1 with input $\widetilde{\boldsymbol{a}}$ computes indeed the exact solution $|\mathcal{I}(\boldsymbol{w}, b)|$ of the #Knapsack problem. If the perturbed slope $\widetilde{a}_n$ in line 4 of Algorithm 1 is evaluated via (12) by calling the #Optimal Transport oracle twice, then Algorithm 1 constitutes a Turing reduction from the #**P**-hard #Knapsack problem to the #Optimal Transport problem.

To prove that the #Optimal Transport problem is #**P**-hard, it remains to be shown that if any oracle call requires unit time, then the Turing reduction constructed above runs in polynomial time in the bit length of $(\boldsymbol{w}, b)$. This is indeed the case because Algorithm 1 calls the #Optimal Transport oracle only $2K$ times in total and because all other operations can be carried out efficiently. In particular, the time needed for reading the oracle outputs is polynomially bounded in the size of $(\boldsymbol{w}, b)$. Indeed, the bit length of $\widetilde{W}_c(\mu, \nu_{i/I})$ is polynomially bounded in the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, i/I)$ thanks to the definition of the #Optimal Transport problem, and the time needed for computing $(\boldsymbol{y}_1, \boldsymbol{y}_2, i/I)$ is trivially bounded by a polynomial in the bit length of $(\boldsymbol{w}, b)$ for any $i \in \mathcal{I}$. These observations complete the proof. □

We emphasize that the Turing reduction derived in the proof of Theorem 3.3 can be implemented without knowing the accuracy level $\overline{\varepsilon}$ of the #Optimal Transport oracle. This is essential because $\overline{\varepsilon}$ is defined as the minimum of exponentially many terms, and we are not aware of any method to compute it efficiently. Without such a method, a Turing reduction relying on $\overline{\varepsilon}$ could not run in polynomial time.

Remark 3.7 (Polynomial-Time Turing Reductions)**.** Recall that a polynomial-time Turing reduction from problem A to problem B is a Turing reduction that runs in polynomial time in the input size of A under the hypothetical assumption that there is an oracle for solving B in unit time. The time needed for computing oracle inputs and reading oracle outputs is attributed to the Turing reduction and is not absorbed in the oracle. Thus, a Turing reduction can run in polynomial time only if the oracle's output size is guaranteed to be polynomially bounded. The existence of a polynomial-time Turing reduction from A to B implies that if there was an efficient algorithm for solving B, then we could solve A in polynomial time (this operationalizes the assertion that "A is not harder than B"). One could use this implication as an alternative definition, that is, one could define a polynomial-time Turing reduction as a Turing reduction that runs in polyonomial time provided that the oracle runs in polynomial time. In our opinion, this alternative definition would be perfectly reasonable. However, it is not equivalent to the original definition by Valiant [1979b], which compels us to ascertain that the oracle output has polynomial size irrespective of the oracle's actual runtime. Instead, the alternative definition directly refers to the oracle's actual runtime. In that it conditions on oracles that run in polynomial time, it immediately guarantees that their outputs have polynomial size. In short, the original definition requires the bit length of the oracle's output to be polynonmially bounded for *every* oracle that solves B (which requires a proof), whereas the alternative definition requires such a bound only for oracles that solve B in polynomial time (which requires no proof). As Theorem 3.3 relies on the original definition of a polynomial-time Turing reduction, we had to introduce condition (ii) in the definition of the #Optimal Transport problem. We consider the differences between the original and alternative definitions of polynomial-time Turing reductions as pure technicalities, but discussing them here seems relevant for motivating our formulation of the #Optimal Transport problem.

Assume now that p is an even number, and consider any instance of the #Optimal Transport problem. In this case, all coefficients of the linear program (1) are rational, and thus $W_c(\mu, \nu_t)$ is a rational number that can be computed in finite time (*e.g.*, via the simplex algorithm). From Corollary 3.5 we further know that $W_c(\mu, \nu_t)$ has polynomially bounded bit length. Thus, $\widetilde{W}_c(\mu, \nu_t) = W_c(\mu, \nu_t)$ satisfies both properties (i) and (ii) that are required of an admissible approximation of the optimal transport distance. Nevertheless, Theorem 3.3 asserts that computing $W_c(\mu, \nu_t)$ approximately is already #**P**-hard. This trivially implies that computing $W_c(\mu, \nu_t)$ *exactly* is also #**P**-hard.

4. Dynamic Programming-Type Solution Methods

We now return to the generic optimal transport problem with independent marginals, where μ is representable as $\otimes_{k\in\mathcal{K}}\mu_k$, the marginals of μ constitute arbitrary univariate distributions supported on L points, and ν constitutes an arbitrary multivariate distribution supported on J points. This problem class covers all instances of the #Optimal Transport problem, and by Theorem 3.3 it is therefore #**P**-hard even if only approximate solutions are sought. In fact, *any* problem class that is rich enough to contain all instances of the #Optimal Transport problem is #**P**-hard. We will now demonstrate that particular instances of the optimal transport problem with independent marginals can be solved in polynomial or pseudo-polynomial time by a dynamic programming-type algorithm even though the distribution μ involves exponentially many atoms and the linear program (1) has exponential size. Throughout this discussion we call $\mathcal{N} \subseteq \mathbb{R}$ a one-dimensional regular grid with cardinality N if there exist $\hat{s}_1, \ldots, \hat{s}_N \in \mathbb{R}$ and a grid spacing constant $d > 0$ such that $\hat{s}_{i+1} = \hat{s}_i + d$ for all $i = 1, \ldots, N-1$ and $\mathcal{N} = \{\hat{s}_1, \ldots, \hat{s}_N\}$. We say that a set $\mathcal{M} \subseteq \mathbb{R}$ spans the one-dimensional regular grid $\mathcal{N}$ if $\mathcal{M} \subseteq \mathcal{N}$, $\min \mathcal{M} = \min \mathcal{N}$ and $\max \mathcal{M} = \max \mathcal{N}$.

Theorem 4.1 (Dynamic Programming-Type Algorithm for Optimal Transport Problems with Independent Marginals)**.** Suppose that $\mu = \otimes_{k\in\mathcal{K}}\mu_k$ is a product of K independent univariate distributions of the form $\mu_k = \sum_{l\in\mathcal{L}} \mu_k^l \delta_{x_k^l}$ and that $\nu_t = t\delta_{\boldsymbol{y}_1} + (1-t)\delta_{\boldsymbol{y}_2}$ is a two-point distribution. If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$ and if $\mathcal{M} = \{x_k^l(y_{1,k} - y_{2,k}) : k \in \mathcal{K}, l \in \mathcal{L}\}$ is spanned by a regular one-dimensional grid $\mathcal{N}$ with (known) cardinality N, then the optimal transport distance between μ and ν_t can be computed exactly by a dynamic programming-type algorithm using $\mathcal{O}(KL\log_2(KL) + KLN + K^2N^2)$ arithmetic operations. If all problem parameters are rational and representable as ratios of two integers with absolute values at most U, then the bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L, N and $\log_2(U)$.

Assuming that $\mathcal{M}$ is spanned by some regular one-dimensional grid $\mathcal{N}$, Theorem 4.1 establishes an upper bound on the number of arithmetic operations needed to solve the optimal transport problem with independent marginals. We will see that the proof of Theorem 4.1 is constructive in that it develops a concrete dynamic programming-type algorithm that attains the indicated upper bound (see Algorithm 2). However, this bound depends on the cardinality N of the grid $\mathcal{N}$, and Theorem 4.1 does not relate N to K, L or U. More importantly, it provides no guidelines for constructing $\mathcal{N}$ or even proving its existence.

Remark 4.2 (Existence of $\mathcal{N}$)**.** If all support points of μ and ν have rational components, then a regular one-dimensional grid $\mathcal{N}$ satisfying the assumptions of Theorem 4.1 is guaranteed to exist. In general, however, its cardinality scales exponentially with K and L, implying that the dynamic programming-type algorithm of Theorem 4.1 is inefficient. To see this, assume that for all $k \in \mathcal{K}$, $l \in \mathcal{L}$ and $j \in \{1, 2\}$ there exist integers $a_{k,l}, c_{j,k} \in \mathbb{Z}$ and $b_{k,l}, d_{j,k} \in \mathbb{N}$ such that $x_k^l = a_{k,l}/b_{k,l}$ and $y_{j,k} = c_{j,k}/d_{j,k}$. Thus, we have

$$x_k^l(y_{1,k} - y_{2,k}) = \frac{a_k^l(c_{1,k}d_{2,k} - c_{2,k}d_{1,k})}{b_{k,l}d_{1,k}d_{2,k}} \quad k \in \mathcal{K}, \ \forall l \in \mathcal{L},$$

which implies that all elements of $\mathcal{M}$ can be expressed as rational numbers with common denominator $D = \prod_{k\in\mathcal{K},l\in\mathcal{L}} b_{k,l}d_{1,k}d_{2,k}$. Clearly, $\mathcal{M}$ is therefore spanned by a regular one-dimensional grid $\mathcal{N}$ with grid spacing constant $d = D^{-1}$ and cardinality $N = D(\max \mathcal{M} - \min \mathcal{M}) + 1$. If U denotes as usual an upper bound on the absolute values of the integers $a_{k,l}$, $b_{k,l}$, $c_{j,k}$ and $d_{j,k}$ for all $k \in \mathcal{K}$, $l \in \mathcal{L}$ and $j \in \{1, 2\}$, then we have $D \leq U^{3KL}$, and all elements of $\mathcal{M}$ have absolute values of at most $2U^3$. The cardinality of $\mathcal{N}$ therefore satisfies $N \leq 4U^{3(KL+1)} + 1$. This reasoning suggests that, in the worst case, the dynamic programming-type algorithm of Theorem 4.1 may require up to $\mathcal{O}(K^2U^{3(KL+1)})$ arithmetic operations.

Remark 4.2 guarantees that a regular one-dimensional grid $\mathcal{N}$ satisfying the assumptions of Theorem 4.1 exists whenever the input bit length of the optimal transport problem with independent marginals is finite. However, Remark 4.2 also reveals that the algorithm of Theorem 4.1 may be highly inefficient in general. Remark 4.3 below discusses special conditions under which this algorithm is of practical interest.

Remark 4.3 (Efficiency of the Dynamic Programming-Type Algorithm)**.** The algorithm of Theorem 4.1 is efficient on problem instances that display the following properties.

(i) If $\mathcal{M}$ is spanned by a regular one-dimensional grid whose cardinality N grows only polynomially with K and L but is independent of U, then the number of arithmetic operations required by the algorithm of Theorem 4.1 grows polynomially with K and L but is independent of U, and the bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L and $\log_2(U)$. Hence, the algorithm runs in *strongly polynomial time* on a Turing machine.

(ii) If $\mathcal{M}$ is spanned by a regular one-dimensional grid whose cardinality N grows polynomially with K, L and $\log_2(U)$, then the number of arithmetic operations required by the algorithm of Theorem 4.1 as well as the bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L and $\log_2(U)$. Hence, the algorithm runs in *weakly polynomial time* on a Turing machine.

(iii) If $\mathcal{M}$ is spanned by a regular one-dimensional grid whose cardinality grows polynomially with K, L and U (but exponentially with $\log_2(U)$), then the number of arithmetic operations required by the

algorithm of Theorem 4.1 grows polyonomially with K, L and U, and the bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L and $\log_2(U)$. Hence, the algorithm runs in *pseudo-polynomial time* on a Turing machine.

Before proving Theorem 4.1, we recall the definition of the Conditional Value-at-Risk (CVaR) by Rockafellar and Uryasev [2002]. Specifically, if the random vector $\boldsymbol{x}$ is governed by the probability distribution μ, then the CVaR at level $t \in (0,1)$ of any Borel measurable loss function $\ell(\boldsymbol{x})$ is defined as

$$\mathrm{CVaR}_t[\ell(\boldsymbol{x})] = \inf_{\beta \in \mathbb{R}} \ \beta + \frac{1}{t} \, \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\max\{\ell(\boldsymbol{x}) - \beta, 0\} \right].$$

Here, the minimization problem over β is solved by the Value-at-Risk (VaR) at level t [Rockafellar and Uryasev, 2002, Theorem 10], which is defined as the left $(1-t)$-quantile of the loss distribution, that is,

$$\mathrm{VaR}_t[\ell(x)] = \inf \left\{ \tau \in \mathbb{R} : \mu[\ell(x) \le \tau] \ge 1 - t \right\}.$$

The proof of Theorem 4.1 also relies on the following lemma.

Lemma 4.4 (Minkowski sums of regular one-dimensional grids)**.** If $\mathcal{N}$ is a one-dimensional regular grid with cardinality N and grid spacing constant $d > 0$, then the k-fold Minkowski sum $\sum_{i=1}^k \mathcal{N}$ of $\mathcal{N}$ is another one-dimensional regular grid with cardinality $k(N-1)+1$ and the same grid spacing constant d.

Proof. Any regular one-dimensional grid with cardinality N and grid spacing constant $d > 0$ is representable as the image of $\{1, \ldots, N\}$ under the affine transformation $f(s) = \hat{s}_1 - d + ds$, where $\hat{s}_1$ denotes the smallest element of $\mathcal{N}$. It is immediate to see that the k-fold Minkowski sum of $\mathcal{N}$ is another one-dimensional regular grid with grid spacing constant d. In addition, the cardinality of this Minkowski sum satisfies

$$\left| \sum_{i=1}^k \mathcal{N} \right| = \left| \sum_{i=1}^k f(\{1, \ldots, N\}) \right| = \left| f\left(\sum_{i=1}^k \{1, \ldots, N\} \right) \right| = |f(\{k, \ldots, kN\})| = |\{k, \ldots, kN\}| = k(N-1)+1,$$

where the second equality holds because f is affine and because the cardinality of any set is invariant under translations. Thus, the claim follows. □

Proof of Theorem 4.1. Throughout this proof we exceptionally assume that each arithmetic operation can be performed in unit time irrespective of the bit lengths of the involved operands. We emphasize that everywhere else in the paper, however, time is measured in the standard Turing machine model of computation. Throughout this proof we further set $I = L^K$ and denote as usual by $\boldsymbol{x}_i$, $i \in \mathcal{I}$, the I different support points of μ. Then, the optimal transport distance between μ and ν_t can be expressed as

$$\begin{aligned} W_c(\mu, \nu_t) &= \min_{\boldsymbol{\pi} \in \Pi(\mu, \nu_t)} \ \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \left(\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{y}_j\|^2 - 2\boldsymbol{x}_i^\top \boldsymbol{y}_j \right) \pi_{ij} \\ &= \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\|\boldsymbol{x}\|^2 \right] + \mathbb{E}_{\boldsymbol{y} \sim \nu_t} \left[\|\boldsymbol{y}\|^2 \right] - 2 \max_{\boldsymbol{\pi} \in \Pi(\mu, \nu_t)} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \boldsymbol{x}_i^\top \boldsymbol{y}_j \pi_{ij}. \end{aligned} \tag{14}$$

The two expectations in (14) can be evaluated in $\mathcal{O}(KL)$ arithmetic operations because

$$\mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\|\boldsymbol{x}\|^2 \right] = \sum_{k \in \mathcal{K}} \mathbb{E}_{x_k \sim \mu_k} \left[(x_k)^2 \right] = \sum_{k \in \mathcal{K}} \sum_{l \in \mathcal{L}} \mu_k^l (x_k^l)^2 \text{ and } \ \mathbb{E}_{\boldsymbol{y} \sim \nu_t} \left[\|\boldsymbol{y}\|^2 \right] = t\|\boldsymbol{y}_1\|^2 + (1-t)\|\boldsymbol{y}_2\|^2,$$

and it is easy to verify that their bit lengths are polynomially bounded in K, L and $\log_2(U)$. Moreover, as in the proof of Lemma 3.6, the maximization problem in (14) simplifies to

$$\max_{\boldsymbol{\pi}\in\Pi(\mu,\nu_t)} \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \boldsymbol{x}_i^\top \boldsymbol{y}_j \pi_{ij} = \begin{cases} \max\limits_{\boldsymbol{q}_1,\boldsymbol{q}_2\in\mathbb{R}^I_+} & t\sum\limits_{i\in\mathcal{I}} \boldsymbol{x}_i^\top \boldsymbol{y}_1 q_{1,i} + (1-t)\sum\limits_{i\in\mathcal{I}} \boldsymbol{x}_i^\top \boldsymbol{y}_2 q_{2,i} \\ \text{s.t.} & \mathbf{1}^\top \boldsymbol{q}_1 = 1,\ \mathbf{1}^\top \boldsymbol{q}_2 = 1 \\ & t q_{1,i} + (1-t) q_{2,i} = \mu[\boldsymbol{x}=\boldsymbol{x}_i] \quad \forall i\in\mathcal{I}. \end{cases}$$

$$= \sum_{i\in\mathcal{I}} \boldsymbol{x}_i^\top \boldsymbol{y}_2\, \mu[\boldsymbol{x}=\boldsymbol{x}_i] + \begin{cases} \max\limits_{\boldsymbol{q}\in\mathbb{R}^I_+} & \sum\limits_{i\in\mathcal{I}} \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2) q_i \\ \text{s.t.} & \mathbf{1}^\top \boldsymbol{q} = t \\ & q_i \le \mu[\boldsymbol{x}=\boldsymbol{x}_i] \quad \forall i\in\mathcal{I}, \end{cases} \tag{15}$$

where the second equality follows from the variable substitution $\boldsymbol{q} \leftarrow t\boldsymbol{q}_1$ and the subsequent elimination of $\boldsymbol{q}_2$ by using the equations $(1-t)q_{2,i} = \mu[\boldsymbol{x}=\boldsymbol{x}_i] - q_i$ for all $i\in\mathcal{I}$. Observe next that the first sum in (15) can again be evaluated using $\mathcal{O}(KL)$ arithmetic operations because

$$\sum_{i\in\mathcal{I}} \boldsymbol{x}_i^\top \boldsymbol{y}_2\, \mu[\boldsymbol{x}=\boldsymbol{x}_i] = \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{x}^\top \boldsymbol{y}_2\right] = \sum_{k\in\mathcal{K}} \mathbb{E}_{x_k\sim\mu_k}\left[x_k y_{2,k}\right] = \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}} x_k^l \mu_k^l y_{2,k},$$

and the bit length of this sum is polynomially bounded in K, L and $\log_2(U)$. For $t=0$, the optimal value of the maximization problem in (15) vanishes. For $t=1$, on the other hand, the problem's optimal solution satisfies $q_i = \mu[\boldsymbol{x}=\boldsymbol{x}_i]$ for all $i\in\mathcal{I}$. By using now standard arguments, one readily verifies that the corresponding optimal value can once again be computed in $\mathcal{O}(KL)$ arithmetic operations and has polynomially bounded bit length in K, L and $\log_2(U)$. In the remainder of the proof we may thus assume that $t\in(0,1)$. To solve the maximization problem in (15) in this generic case, we first reformulate it as

$$\max\left\{\sum_{i\in\mathcal{I}} \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)\,\mu[\boldsymbol{x}=\boldsymbol{x}_i]\, q_i \ :\ \mathbf{0}\le\boldsymbol{q}\le\mathbf{1},\ \sum_{i\in\mathcal{I}}\mu[\boldsymbol{x}=\boldsymbol{x}_i]\, q_i = t\right\} \tag{16}$$

by applying the variable substitution $q_i \leftarrow q_i/\mu[\boldsymbol{x}=\boldsymbol{x}_i]$. By assumption, there exists a regular one-dimensional grid $\mathcal{N}$ with cardinality N such that $x_k^l(y_{1,k}-y_{2,k})\in\mathcal{N}$ for every $k\in\mathcal{K}$ and $l\in\mathcal{L}$. This readily implies that $\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)\in\mathcal{N}_K = \sum_{k=1}^K\mathcal{N}$. In the following, we introduce an auxiliary random variable s supported on $\mathcal{N}_K$, and we show that problem (16) is equivalent to

$$\max\left\{t\,\mathbb{E}_{s\sim\eta}[s] \ :\ \eta\in\mathcal{P}(\mathcal{N}_K),\ \eta[s=\hat{s}] \le \frac{1}{t}\mu\left[\boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}\right]\ \forall\hat{s}\in\mathcal{N}_K\right\}, \tag{17}$$

which optimizes over a family of possible distributions η of s. To prove that the optimal value of (17) is at least as large as that of (16), we fix an arbitrary $\boldsymbol{q}$ feasible in (16) and construct a probability distribution η feasible in (17) that has the same objective function value as $\boldsymbol{q}$. Specifically, we define η through

$$\eta[s=\hat{s}] = \frac{1}{t}\sum_{i\in\mathcal{I}}\mu\left[\boldsymbol{x}=\boldsymbol{x}_i,\ \boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}\right] q_i \quad \forall\hat{s}\in\mathcal{N}_K$$

412 and note that $\eta[s=\hat{s}]\ge 0$ for every $\hat{s}\in\mathcal{N}_K$ because $\boldsymbol{q}\ge\mathbf{0}$. In addition, we have

$$\sum_{\hat{s}\in\mathcal{N}_K}\eta[s=\hat{s}] = \frac{1}{t}\sum_{i\in\mathcal{I}}\sum_{\hat{s}\in\mathcal{N}_K}\mu\left[\boldsymbol{x}=\boldsymbol{x}_i,\ \boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}\right] q_i = \frac{1}{t}\sum_{i\in\mathcal{I}}\mu\left[\boldsymbol{x}=\boldsymbol{x}_i\right] q_i = 1,$$

where the second equality follows from the law of total probability, and the third equality holds because $\boldsymbol{q}$ must satisfy the last constraint in (16). This guarantees that $\eta\in\mathcal{P}(\mathcal{N}_K)$. The other constraints in (17) are trivially satisfied by the construction of η and because $\boldsymbol{q} \leq \mathbf{1}$. Finally, the objective function value of η in (17) coincides with the objective function value of $\boldsymbol{q}$ in (16) because

$$
\begin{aligned}
t\,\mathbb{E}_{s\sim\eta}[s] = t \sum_{\hat{s}\in\mathcal{N}_K} \hat{s}\,\eta[s=\hat{s}] &= t \sum_{\hat{s}\in\mathcal{N}_K} \hat{s}\,\frac{1}{t}\sum_{i\in\mathcal{I}} \mu\left[\boldsymbol{x}=\boldsymbol{x}_i,\ \boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}\right] q_i \\
&= \sum_{\hat{s}\in\mathcal{N}_K}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)\,\mu\left[\boldsymbol{x}=\boldsymbol{x}_i,\ \boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}\right] q_i \\
&= \sum_{i\in\mathcal{I}} \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)\,\mu[\boldsymbol{x}=\boldsymbol{x}_i]\,q_i.
\end{aligned}
$$

As $\boldsymbol{q}$ was chosen arbitrarily, we have shown that the optimal value of (17) is at least as large as that of (16).

To prove the converse inequality, we fix an arbitrary η feasible in (17) and construct a $\boldsymbol{q}$ feasible in (16) that has the same objective function value as η. Specifically, we define $\boldsymbol{q}$ through

$$
q_i = \frac{\eta[s=\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)]}{\mu[\boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)]}\,t.
$$

It is clear that $q_i \geq 0$, and the constraints of (17) for $\hat{s} = \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)$ imply that $q_i \leq 1$ for all $i \in \mathcal{I}$. Thus, we have $\mathbf{0} \leq \boldsymbol{q} \leq \mathbf{1}$. In addition, $\boldsymbol{q}$ also satisfies the last constraint in (16) because

$$
\begin{aligned}
\sum_{i\in\mathcal{I}} \mu[\boldsymbol{x}=\boldsymbol{x}_i]\,q_i &= \sum_{i\in\mathcal{I}} \mu\left[\boldsymbol{x}=\boldsymbol{x}_i,\ \boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)\right] q_i \\
&= \sum_{i\in\mathcal{I}} \frac{\mu[\boldsymbol{x}=\boldsymbol{x}_i,\ \boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)]\,\eta[s=\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)]}{\mu[\boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)]}\,t \\
&= \sum_{\hat{s}\in\mathcal{N}_K} \sum_{\substack{i\in\mathcal{I}:\\ \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}}} \frac{\mu[\boldsymbol{x}=\boldsymbol{x}_i,\ \boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}]\,\eta[s=\hat{s}]}{\mu[\boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}]}\,t = \sum_{\hat{s}\in\mathcal{N}_K} \eta[s=\hat{s}]\,t = t,
\end{aligned}
$$

where the second equality follows from the definition of $\boldsymbol{q}$, and the third equality holds because for every $i \in \mathcal{I}$ there exists a unique $\hat{s} \in \mathcal{N}_K$ with $\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2) = \hat{s}$. The last equality holds because $\eta \in \mathcal{P}(\mathcal{N}_K)$. Finally, the objective function value of $\boldsymbol{q}$ in (16) coincides with the objective function value of η in (17) because

$$
\begin{aligned}
\sum_{i\in\mathcal{I}} \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)\,\mu[\boldsymbol{x}=\boldsymbol{x}_i]\,q_i &= t\sum_{i\in\mathcal{I}} \frac{\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)\,\mu[\boldsymbol{x}=\boldsymbol{x}_i]\,\eta[s=\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)]}{\mu[\boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)]} \\
&= t\sum_{\hat{s}\in\mathcal{N}_K} \sum_{\substack{i\in\mathcal{I}:\\ \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}}} \frac{\hat{s}\,\mu[\boldsymbol{x}=\boldsymbol{x}_i]\,\eta[s=\hat{s}]}{\mu[\boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)=\hat{s}]} = t\sum_{\hat{s}\in\mathcal{N}_K} \hat{s}\,\eta[s=\hat{s}] = t\,\mathbb{E}_{s\sim\eta}[s].
\end{aligned}
$$

In summary, we have thus shown that (16) is equivalent to (17).

The inequality constraints in (17) express that η must be absolutely continuous with respect to the marginal distribution of $\ell(\boldsymbol{x}) = \boldsymbol{x}^\top(\boldsymbol{y}_1-\boldsymbol{y}_2)$ under μ and that the corresponding probability density function does not exceed $1/t$ μ-almost surely. By [Föllmer and Schied, 2004, Theorem 4.47], the optimal value of problem (17) therefore coincides with the t-fold multiple of the CVaR of $\ell(\boldsymbol{x})$ at level t. Assume from now on without loss of generality that $\mathcal{N}_K = \{\hat{s}_{K,1}, \ldots, \hat{s}_{K,|\mathcal{N}_K|}\}$ and that the elements of $\mathcal{N}_K$ are sorted in ascending order, that is, $\hat{s}_{K,1} < \cdots < \hat{s}_{K,|\mathcal{N}_K|}$. Also, denote by n_t the unique index satisfying

$$
\sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] \geq 1-t > \sum_{n=1}^{n_t-1} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]. \tag{18}
$$

By [Rockafellar and Uryasev, 2002, Proposition 8], the optimal value of problem (17) thus equals

$$
t \cdot \mathrm{CVaR}_t[\ell(\boldsymbol{x})] = \left(\sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] - (1-t)\right)\hat{s}_{K,n_t} + \sum_{n=n_t+1}^{|\mathcal{N}_K|} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]\hat{s}_{K,n}. \tag{19}
$$

In summary, we have reduced the task of computing the optimal value of problem (17) to computing the CVaR of $\ell(\boldsymbol{x})$ at level t, which amounts to evaluating a sum of $\mathcal{O}(|\mathcal{N}_K|)$ terms. We will now prove that evaluating this sum requires $\mathcal{O}(K^2L^2 + K^2N^2)$ arithmetic operations. To this end, we first show that the grid points $\hat{s}_{K,n}$, $n = 1, \ldots, |\mathcal{N}_K|$, can be computed in time $\mathcal{O}(K^2L^2 + KN)$ (Step 1), then we show that the probabilities $\mu[\ell(\boldsymbol{x}) - \hat{s}_{K,n}]$, $n = 1, \ldots, |\mathcal{N}_K|$, can be computed recursively in time $\mathcal{O}(K^2N^2)$ (Step 2), and finally we use these ingredients to compute the right hand side of (19) in time $\mathcal{O}(KN)$ (Step 3).

Step 1. By assumption, the one-dimensional regular grid $\mathcal{N}$ has known cardinality N and spans $\mathcal{M} = \{x_k^l(y_{1,k} - y_{2,k}) : k \in \mathcal{K},\, l \in \mathcal{L}\}$. To compute all elements of $\mathcal{N}$, we first compute all elements of $\mathcal{M}$ in time $\mathcal{O}(KL)$ and sort them in non-decreasing order in time $\mathcal{O}(KL\log_2(KL))$ using merge sort, for example. As $\mathcal{M}$ spans $\mathcal{N}$, the minimum and the maximum of $\mathcal{M}$ coincide with the minimum $\hat{s}_1$ and the maximum $\hat{s}_N$ of $\mathcal{N}$, respectively. Given $\hat{s}_1$ and $\hat{s}_N$, we can then compute the grid spacing constant $d = (\hat{s}_N - \hat{s}_1)/(N-1)$ as well as the elements $\hat{s}_n = \hat{s}_1 + d(n-1)$, $n = 1, \ldots, N$, of $\mathcal{N}$, which requires $\mathcal{O}(N)$ arithmetic operations. The bit lengths of all numbers computed so far are bounded by a polynomial in $\log_2(U)$ and $\log_2(N)$.

It is easy to see that $\mathcal{N}_K = \sum_{k=1}^K \mathcal{N}$ is also a one-dimensional regular grid that has the same grid spacing constant as $\mathcal{N}$ and whose minimum $\hat{s}_{K,1} = K\hat{s}_1$ can be computed in constant time. The elements of $\mathcal{N}_K$ are then obtained by computing $\hat{s}_{K,n} = \hat{s}_{K,1} + d(n-1)$ for all $n = 1, \ldots, |\mathcal{N}_K|$, where $|\mathcal{N}_K| = K(N-1)+1$ thanks to Lemma 4.4. This computation requires $\mathcal{O}(KN)$ arithmetic operations, and the bit lengths of all involved numbers are still bounded by a polynomial in $\log_2(U)$ and $\log_2(N)$. This completes Step 1.

Step 2. We now show that the probabilities $\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]$ for $n = 1, \ldots, |\mathcal{N}_K|$ can be calculated recursively in time $\mathcal{O}(K^2N^2)$. To this end, we introduce the partial sums $\ell_k(\boldsymbol{x}) = \sum_{m=1}^k x_m(y_{1,m} - y_{2,m})$ for every $k \in \mathcal{K}$ and note that $\ell_K(\boldsymbol{x}) = \ell(\boldsymbol{x})$. For every $k \in \mathcal{K}$, the range of the function $\ell_k(\boldsymbol{x})$ is a subset of the one-dimensional regular grid $\mathcal{N}_k = \sum_{k'=1}^k \mathcal{N}$. The law of total probability then implies that

$$\mu[\ell_k(\boldsymbol{x}) = \hat{s}] = \sum_{\hat{s}' \in \mathcal{N}} \mu\left[\ell_{k-1}(\boldsymbol{x}) = \hat{s} - \hat{s}',\; x_k(y_{1,k} - y_{2,k}) = \hat{s}'\right] \quad \forall k \in \mathcal{K}\backslash\{1\},\; \forall \hat{s} \in \mathcal{N}_k,$$

where $\hat{s}_1, \ldots, \hat{s}_N$ denote as usual the elements of $\mathcal{N}$, and where $\mu[\ell_1(\boldsymbol{x}) = \hat{s}] = \sum_{i=1}^N \mu_k[x_1(y_{1,k} - y_{2,k}) = \hat{s}_i]$ for all $\hat{s} \in \mathcal{N}_1$. As $\ell_k(\boldsymbol{x}) = \ell_{k-1}(\boldsymbol{x}) + x_k(y_{1,k} - y_{2,k})$, $\ell_{k-1}(\boldsymbol{x})$ is constant in $x_k, \ldots, x_K$ and the components of $\boldsymbol{x}$ are mutually independent under the product distribution $\mu = \otimes_{k\in\mathcal{K}}\mu_k$, we thus have

$$\mu[\ell_k(\boldsymbol{x}) = \hat{s}] = \sum_{\hat{s}' \in \mathcal{N}} \mu\left[\ell_{k-1}(\boldsymbol{x}) = \hat{s} - \hat{s}'\right] \times \mu_k\left[x_k(y_{1,k} - y_{2,k}) = \hat{s}'\right] \quad \forall k \in \mathcal{K}\backslash\{1\},\; \forall \hat{s} \in \mathcal{N}_k. \tag{20}$$

The marginal probabilities $\mu_k[x_k(y_{1,k} - y_{2,k}) = \hat{s}']$ for all $k \in \mathcal{K}$ and $\hat{s}' \in \mathcal{N}$ can be pre-computed in time $\mathcal{O}(KLN)$. Given $\mu[\ell_{k-1}(\boldsymbol{x}) = \hat{s}]$, $\hat{s} \in \mathcal{N}_{k-1}$, each probability $\mu[\ell_k(\boldsymbol{x}) = \hat{s}]$, $\hat{s} \in \mathcal{N}_k$, can then be computed in time $\mathcal{O}(N)$ by using (20). As $|\mathcal{N}_k| = \mathcal{O}(kN)$ for every $k \in \mathcal{K}$ thanks to Lemma 4.4, each iteration $k \in \mathcal{K}$ of the the dynamic programming-type recursion (20) requires at most $\mathcal{O}(KN^2)$ arithmetic operations. Finally, as there are $\mathcal{O}(K)$ iterations in total, the sought probabilities $\mu[\ell_K(\boldsymbol{x}) = \hat{s}]$, $\hat{s} \in \mathcal{N}_K$, can be computed in time $\mathcal{O}(K^2N^2)$. An elementary calculation further shows that the bit lengths of these probabilities are bounded by a polynomial in K, N and $\log_2(U)$. This completes Step 2.

Step 3. As all terms appearing in the sum on the right hand side of (19) have been pre-computed in Steps 1 and 2, the sum itself can now be evaluated in time $\mathcal{O}(KN)$ thanks to Lemma 4.4. Note that the critical index n_t defined in (18) can also be computed in time $\mathcal{O}(KN)$. The bit lengths of all numbers involved in these calculations are bounded by a polynomial in K, N and $\log_2(U)$. This completes Step 3.

In summary, the time required for evaluating the CVaR in (19) totals $\mathcal{O}(KL\log_2(KL) + KLN + K^2N^2)$, which matches the overall time required for all calculations described in Steps 1, 2 and 3. This computation time dominates the time $\mathcal{O}(KL)$ spent on all preprocessing steps, and thus the claim follows. □

The dynamic programming-type procedure developed in the proof of Theorem 3.3 is summarized in Algorithm 2. This procedure outputs the optimal transport distance between μ and ν_t (denoted by W_c). In addition, Algorithm 2 can be used for constructing the optimal transportation plan from μ to ν_t.

Algorithm 2 Optimal Transport with Independent Marginals

Input: $\{\mu_k^l\}_{k\in\mathcal{K},l\in\mathcal{L}}$, $\{x_k^l\}_{k\in\mathcal{K},l\in\mathcal{L}}$, $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^K$, t, N

Initialize $\hat{s}_1 = \min\limits_{k\in\mathcal{K},l\in\mathcal{L}} x_k^l(y_{1,k} - y_{2,k})$ and $\hat{s}_N = \max\limits_{k\in\mathcal{K},l\in\mathcal{K}} x_k^l(y_{1,k} - y_{2,k})$

Set $d = (\hat{s}_N - \hat{s}_1)/(N-1)$ and $\hat{s}_n = \hat{s}_1 + d(n-1)$ $\forall n = 1, \dots, N$

Compute $\mu_k[x_k(y_{1,k} - y_{2,k}) = \hat{s}_n]$ $\forall k \in \mathcal{K}$ and $n \in \mathcal{N}$

Set $\mu[\ell_1(\boldsymbol{x}) = \hat{s}_{1,n}] = \sum\limits_{\hat{s}'\in\mathcal{N}} \mu_1[x_1(y_{1,1} - y_{2,1}) = \hat{s}']$ $\forall n = 1, \dots, N$

for $k = 2, \dots, K$ **do**

  for $n = 1, \dots, k(N-1)+1$ **do**

    $\hat{s}_{k,n} = k\hat{s}_1 + d(n-1)$

    $\mu[\ell_k(\boldsymbol{x}) = \hat{s}_{k,n}] = \sum\limits_{\hat{s}'\in\mathcal{N}} \mu[\ell_{k-1}(\boldsymbol{x}) = \hat{s}_{k,n} - \hat{s}'] \times \mu_k[x_k(y_{1,k} - y_{2,k}) = \hat{s}']$

  end for

end for

Find the index $n_t \in \{1, \dots, K(N-1)+1\}$ satisfying (18)

Set

$$\text{CVaR} = \frac{1}{t}\left[\left(\sum_{n=1}^{n_t} \mu[\ell_K(\boldsymbol{x}) = \hat{s}_{K,n}] - 1 + t\right)\hat{s}_{K,n_t} - 2\sum_{n=n_t+1}^{K(N-1)+1} \mu[\ell_K(\boldsymbol{x}) = \hat{s}_{K,n}]\hat{s}_{K,n}\right]$$

Set

$$W_c = \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}} \mu_k^l (x_k^l)^2 + t\sum_{k\in\mathcal{K}} y_{1,k}^2 + (1-t)\sum_{k\in\mathcal{K}} y_{2,k}^2 - 2\sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}} x_k^l \mu_k^l y_{2,k} - 2t\cdot\text{CVaR}$$

Output: W_c

Remark 4.5 (Optimal Transportation Plan)**.** The critical index n_t computed by Algorithm 2 allows us to construct an optimal transportation plan $\boldsymbol{\pi}^\star \in \mathbb{R}_+^{I\times J}$ that solves the linear program (1), where $\pi_{i,j}^\star$ denotes the probability mass moved from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$ for every $i \in \mathcal{I}$ and $j \in \mathcal{J}$. To see this, note that the defining properties of n_t in (18) imply that $\text{VaR}_t[\ell(\boldsymbol{x})] = \hat{s}_{K,n_t}$ and $\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n_t}] > 0$. We may thus define $\boldsymbol{\pi}^\star$ via

$$\pi_{i,1}^\star = \begin{cases} \mu[\boldsymbol{x} = \boldsymbol{x}_i] & \text{if } \ell(\boldsymbol{x}_i) > \hat{s}_{K,n_t} \\ \dfrac{t - 1 + \sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]}{\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n_t}]} \times \mu[\boldsymbol{x} = \boldsymbol{x}_i] & \text{if } \ell(\boldsymbol{x}_i) = \hat{s}_{K,n_t} \\ 0 & \text{if } \ell(\boldsymbol{x}_i) < \hat{s}_{K,n_t} \end{cases}$$

and $\pi_{i,2}^\star = \mu[\boldsymbol{x} = \boldsymbol{x}_i] - \pi_{i,1}^\star$ for all $i \in \mathcal{I}$. By the first inequality in (18), we have $\boldsymbol{\pi}^\star \geq \mathbf{0}$. In addition, we trivially find $\pi_{i,1}^\star + \pi_{i,2}^\star = \mu[\boldsymbol{x} = \boldsymbol{x}_i]$ for all $i \in \mathcal{I}$, and we have

$$\sum_{i\in\mathcal{I}} \pi_{i,1}^\star = \sum_{\substack{i\in\mathcal{I}:\\ \ell(\boldsymbol{x}_i) > \hat{s}_{K,n_t}}} \mu[\boldsymbol{x} = \boldsymbol{x}_i] + \sum_{\substack{i\in\mathcal{I}:\\ \ell(\boldsymbol{x}_i) = \hat{s}_{K,n_t}}} \frac{t - 1 + \sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]}{\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n_t}]} \times \mu[\boldsymbol{x} = \boldsymbol{x}_i]$$

$$= \sum_{n=n_t+1}^{|\mathcal{N}_K|} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] + t - 1 + \sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] = t = 1 - \sum_{i\in\mathcal{I}} \pi^\star_{i,2}.$$

In summary, this shows that $\boldsymbol{\pi}^\star$ is feasible in the optimal transport problem (1). Finally, we have

$$\begin{aligned}
\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \pi^\star_{ij}\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 &= \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\|\boldsymbol{x}\|^2\right] + \mathbb{E}_{\boldsymbol{y}\sim\nu_t}\left[\|\boldsymbol{y}\|^2\right] - 2\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \boldsymbol{x}_i^\top \boldsymbol{y}_j \pi^\star_{ij} \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\|\boldsymbol{x}\|^2\right] + \mathbb{E}_{\boldsymbol{y}\sim\nu_t}\left[\|\boldsymbol{y}\|^2\right] - 2\,\mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{x}^\top \boldsymbol{y}_2\right] - 2\sum_{i\in\mathcal{I}} \ell(\boldsymbol{x}_i)\,\pi^\star_{i,1} \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\|\boldsymbol{x}\|^2\right] + \mathbb{E}_{\boldsymbol{y}\sim\nu_t}\left[\|\boldsymbol{y}\|^2\right] - 2\,\mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{x}^\top \boldsymbol{y}_2\right] - 2t\cdot \mathrm{CVaR}_t[\ell(\boldsymbol{x})],
\end{aligned}$$

where the first two equalities follow from (14) and (15), respectively, while the third equality exploits the definitions of $\boldsymbol{\pi}^\star$ and the CVaR. The last expression manifestly matches the output W_c of Algorithm 2. Hence, we may conclude that $\boldsymbol{\pi}^\star$ is indeed optimal in (1). Note that evaluating $\pi^\star_{ij}$ for a fixed $i \in \mathcal{I}$ and $j \in \mathcal{J}$ requires at most $\mathcal{O}(NK + KL)$ arithmetic operations provided that the critical index n_t and the probabilities $\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]$, $n \in \mathcal{N}_K$, are given. These quantities are indeed computed by Algorithm 2.

In the following we will identify special instances of the optimal transport problem with independent marginals that can be solved efficiently. Assume first that both μ and ν are supported on $\{0,1\}^K$. This implies that all marginals of μ represent independent Bernoulli distributions. Unlike in Theorem 3.3, however, these Bernoulli distributions may be non-uniform. The following corollary shows that, in this case, the optimal transport problem with independent marginals can be solved in strongly polynomial time.

Corollary 4.6 (Binary Support)**.** Suppose that all assumptions of Theorem 4.1 hold. If in addition $L = 2$, $x_k^1 = 0$ and $x_k^2 = 1$ for all $k \in \mathcal{K}$, and $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \{0,1\}^K$, then the optimal transport distance between μ and ν_t can be computed in strongly polynomial time.

Proof. Under the given assumptions, we have $\mathcal{M} = \{x_k^l(y_{1,k} - y_{2,k}) : k \in \mathcal{K},\, l \in \mathcal{L}\} \subseteq \{-1, 0, 1\}$. Hence, Theorem 4.1 applies with $\mathcal{N} \subseteq \{-1,0,1\}$ and $N \leq 3$, and therefore Algorithm 2 computes $W_c(\mu,\nu_t)$ using $\mathcal{O}(K^2)$ arithmetic operations. As N is constant in K, L and $\log_2(U)$, Remark 4.3 *(i)* implies that $W_c(\mu,\nu_t)$ can be computed in strongly polynomial time in the Turing machine model. □

By generalizing the proof of Corollary 4.6 in the obvious way, one can show that the optimal transport problem with independent marginals remains strongly polynomial-time solvable whenever μ and ν_t are supported on a (fixed) bounded subset of the scaled integer lattice $\mathbb{Z}^K/M$ for some (fixed) scaling factor $M \in \mathbb{N}$. If μ and ν_t are supported on a subset of $\mathbb{Z}^K/M$ that may grow with the problem's input size or if the scaling factor M may grow with the input size, then Algorithm 2 ceases to run in polynomial time. We now show, however, that Algorithm 2 stills run in pseudo-polynomial time in these cases.

Corollary 4.7 (Lattice Support)**.** Suppose that all assumptions of Theorem 4.1 hold. If there exists a positive integer $M \leq U$, such that $x_k^l \in \mathbb{Z}/M$ for all $k \in \mathcal{K}$ and $l \in \mathcal{L}$, while $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{Z}^K/M$, then the optimal transport distance between μ and ν_t can be computed in pseudo-polynomial time.

Proof. Under the given assumptions, we have $\mathcal{M} = \{x_k^l(y_{1,k} - y_{2,k}) : k \in \mathcal{K},\, l \in \mathcal{L}\} \subseteq \mathbb{Z}/M^2$. Therefore, $\mathcal{M}$ spans a one-dimensional regular grid $\mathcal{N} \subseteq \mathbb{Z}/M^2$ with grid spacing constant $d = 1/M^2$ and cardinality

$$\begin{aligned}
N &= (\max\mathcal{M} - \min\mathcal{M})/d \\
&= \max_{k\in\mathcal{K},\, l\in\mathcal{L}} \left\{ M x_k^l (M y_{1,k} - M y_{2,k}) \right\} - \min_{k\in\mathcal{K},\, l\in\mathcal{L}} \left\{ M x_k^l (M y_{1,k} - M y_{2,k}) \right\}.
\end{aligned} \tag{21}$$

Recall that $x_k^l = a_k^l/b_k^l$ for some $a_k^l \in \mathbb{Z}$ and $b_k^l \in \mathbb{N}$ with $|a_k^l|, |b_k^l| \leq U$ and that $M \leq U$. We may thus conclude that $|Mx_k^l| \leq U^2$ for all $k \in \mathcal{K}$ and $l \in \mathcal{L}$. Similarly, one can show that $|My_{1,k}| \leq U^2$ and $|My_{2,k}| \leq U^2$ for all $k \in \mathcal{K}$. By (21), we thus have $N \leq 4U^2$, which implies via Theorem 4.1 that Algorithm 2 computes $W_c(\mu, \nu_t)$ using $\mathcal{O}(KL\log_2(KL) + K^2U^4)$ arithmetic operations. We emphasize that the number of arithmetic operations thus grows polynomially with K, L and U but exponentially with $\log_2(U)$. By Remark 4.3 *(iii)*, $W_c(\mu, \nu_t)$ can therefore be computed in pseudo-polynomial time. □

So far we have discussed methods to solve the optimal transport problem with independent marginals *exactly*. In the remainder of this section we will show that *approximate* solutions can always be computed in pseudo-polynomial time. The following lemma provides a key ingredient for this argument.

Lemma 4.8 (Approximating Optimal Transport Distances)**.** Consider four discrete probability distributions $\mu = \sum_{i\in\mathcal{I}} \mu_i \delta_{\boldsymbol{x}_i}$, $\tilde{\mu} = \sum_{i\in\mathcal{I}} \mu_i \delta_{\tilde{\boldsymbol{x}}_i}$, $\nu = \sum_{j\in\mathcal{J}} \nu_j \delta_{\boldsymbol{y}_j}$ and $\tilde{\nu} = \sum_{j\in\mathcal{J}} \nu_j \delta_{\tilde{\boldsymbol{y}}_j}$ supported on a hypercube $[-U, U]^K$ for some $U > 0$. If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$ and there exists $\varepsilon \geq 0$ such that $\|\tilde{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_\infty \leq \varepsilon$ for all $i \in \mathcal{I}$ and $\|\tilde{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_\infty \leq \varepsilon$ for all $j \in \mathcal{J}$, then we have

$$|W_c(\mu, \nu) - W_c(\tilde{\mu}, \tilde{\nu})| \leq 8KU\varepsilon. \tag{22}$$

We emphasize that Lemma 4.8 holds for arbitrary discrete distributions μ, $\tilde{\mu}$, ν and $\tilde{\nu}$ provided that $\tilde{\mu}$ and $\tilde{\nu}$ are obtained by perturbing only the support points of μ and ν, respectively, but not the corresponding probabilities. In particular, the lemma holds even if μ and $\tilde{\mu}$ fail to represent product distributions with independent marginals, and even if ν and $\tilde{\nu}$ fail to represent two-point distributions. Note also that, by slight abuse of notation, μ_i, $i \in \mathcal{I}$, represent here the probabilties of the support points of μ and should not be confused with the univariate marginal distributions μ_k, $k \in \mathcal{K}$, in the rest of the paper.

Proof of Lemma 4.8. The elementary identity $|a^2 - b^2| = (a + b)|a - b|$ for any $a, b \in \mathbb{R}_+$ implies that

$$|W_c(\mu, \nu) - W_c(\tilde{\mu}, \tilde{\nu})| = \left(W_c(\mu, \nu_t)^{\frac{1}{2}} + W_c(\tilde{\mu}, \tilde{\nu})^{\frac{1}{2}}\right) \left|W_c(\mu, \nu)^{\frac{1}{2}} - W_c(\tilde{\mu}, \tilde{\nu})^{\frac{1}{2}}\right|. \tag{23}$$

By the definition of the optimal transport distance, the first term on the right-hand-side of (23) satisfies

$$\begin{aligned} W_c(\mu, \nu)^{\frac{1}{2}} + W_c(\tilde{\mu}, \tilde{\nu})^{\frac{1}{2}} &= \left(\min_{\pi\in\Pi(\mu,\nu)} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 \pi_{ij}\right)^{\frac{1}{2}} + \left(\min_{\tilde{\pi}\in\Pi(\tilde{\mu},\tilde{\nu})} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{y}}_j\|^2 \tilde{\pi}_{ij}\right)^{\frac{1}{2}} \\ &\leq 4\sqrt{K}U, \end{aligned}$$

where the inequality holds because π and $\tilde{\pi}$ are probability distributions and because

$$\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 \leq K\|\boldsymbol{x}_i - \boldsymbol{y}_j\|_\infty^2 \leq 4KU^2 \quad \text{and} \quad \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{y}}_j\|^2 \leq \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{y}}_j\|_\infty^2 \leq 4KU^2$$

for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$, taking into account that all support points of the four probability distributions μ, $\tilde{\mu}$, ν and $\tilde{\nu}$ fall into the hypercube $[-U, U]^K$. The second term on the right-hand-side of (23) satisfies

$$\begin{aligned} \left|W_c(\mu, \nu)^{\frac{1}{2}} - W_c(\tilde{\mu}, \tilde{\nu})^{\frac{1}{2}}\right| &\leq \left|W_c(\mu, \nu)^{\frac{1}{2}} - W_c(\tilde{\mu}, \nu)^{\frac{1}{2}}\right| + \left|W_c(\tilde{\mu}, \nu)^{\frac{1}{2}} - W_c(\tilde{\mu}, \tilde{\nu})^{\frac{1}{2}}\right| \\ &\leq W_c(\mu, \tilde{\mu})^{\frac{1}{2}} + W_c(\nu, \tilde{\nu})^{\frac{1}{2}} \\ &= \left(\min_{\pi^\mu\in\Pi(\mu,\tilde{\mu})} \sum_{i,i'\in\mathcal{I}} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_{i'}\|^2 \pi^\mu_{ii'}\right)^{\frac{1}{2}} + \left(\min_{\pi^\nu\in\Pi(\nu,\tilde{\nu})} \sum_{j,j'\in\mathcal{J}} \|\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_{j'}\|^2 \pi^\nu_{jj'}\right)^{\frac{1}{2}} \end{aligned}$$

$$\leq \left(\frac{1}{I}\sum_{i\in\mathcal{I}}\|\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i\|^2\right)^{\frac{1}{2}}+\left(\frac{1}{J}\sum_{j\in\mathcal{J}}\|\boldsymbol{y}_j-\tilde{\boldsymbol{y}}_j\|^2\right)^{\frac{1}{2}}\leq 2\sqrt{K}\varepsilon,$$

where the second inequality holds because the 2-Wasserstein distance is a metric and thus obeys the triangle inequality [Villani, 2008, § 6], whereas the third inequality holds because π^μ and π^ν with $\pi^\mu_{ii'}=\frac{1}{I}\delta_{ii'}$ for all $i,i'\in\mathcal{I}$ and $\pi^\nu_{jj'}=\frac{1}{J}\delta_{jj'}$ for all $j,j'\in\mathcal{J}$, respectively, are feasible transportation plans. Finally, the last inequality follows from our assumption that $\|\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i\|_\infty\leq\varepsilon$ and $\|\boldsymbol{y}_j-\tilde{\boldsymbol{y}}_j\|_\infty\leq\varepsilon$, which implies that

$$\|\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i\|^2\leq K\|\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i\|^2_\infty\leq K\varepsilon^2 \quad\text{and}\quad \|\boldsymbol{y}_j-\tilde{\boldsymbol{y}}_j\|^2\leq K\|\boldsymbol{y}_j-\tilde{\boldsymbol{y}}_j\|^2_\infty\leq K\varepsilon^2$$

for all $i\in\mathcal{I}$ and $j\in\mathcal{J}$. Substituting the above estimates back into (23) finally yields (22). □

We now address the approximate solution of optimal transport problems with independent marginals.

Theorem 4.9 (Approximate Solutions of the Optimal Transport Problem with Independent Marginals)**.** Suppose that $\mu=\otimes_{k\in\mathcal{K}}\mu_k$ with $\mu_k=\sum_{l\in\mathcal{L}}\mu_k^l\delta_{x_k^l}$ for every $k\in\mathcal{K}$ and that $\nu_t=t\delta_{\boldsymbol{y}_1}+(1-t)\delta_{\boldsymbol{y}_2}$, and let $\varepsilon>0$ be an error tolerance. If $c(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|^2$ and if all probabilities and coordinates of the support points of μ and ν_t are representable as fractions of two integers with absolute values of at most U, then the optimal transport distance between μ and ν_t can be computed to within an absolute error of at most ε by a dynamic programming-type algorithm using $\mathcal{O}(KL\log_2(KL)+K^6U^8/\varepsilon^4)$ arithmetic operations. The bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L, $\log_2(U)$ and $\log_2(\frac{1}{\varepsilon})$.

Proof. In order to approximate $W_c(\mu,\nu_t)$ to within an absolute accuracy of ε, we define $M=\lceil 8KU/\varepsilon\rceil$ and map all support points of μ and ν to the nearest lattice points in $\mathbb{Z}^K/M$ to construct perturbed probability distributions $\tilde{\mu}$ and $\tilde{\nu}_t$, respectively. Specifically, we construct $\tilde{x}_k^l$ by rounding x_k^l to the nearest point in $\mathbb{Z}/M$ for every $k\in\mathcal{K}$ and $l\in\mathcal{L}$. This requires $\mathcal{O}(KL)$ arithmetic operations. We then define the perturbed marginal distributions $\tilde{\mu}_k=\sum_{l\in\mathcal{L}}\mu_k^l\delta_{\tilde{x}_k^l}$ for all $k\in\mathcal{K}$ and set $\tilde{\mu}=\otimes_{k\in\mathcal{K}}\tilde{\mu}_k$. In addition, we denote by $\tilde{\boldsymbol{x}}_i$, $i\in\mathcal{I}$, the I different support points of $\tilde{\mu}$. Here, it is imperative to use the same orderings for the support points of μ and $\tilde{\mu}$, which implies that $\|\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i\|_\infty\leq\frac{1}{M}\leq\frac{\varepsilon}{8KU}$ for all $i\in\mathcal{I}$ thanks to the construction of $\tilde{\mu}$. We further construct $\tilde{y}_{j,k}$ by rounding $y_{j,k}$ to the nearest points in $\mathbb{Z}/M$ for every $j\in\mathcal{J}=\{1,2\}$ and $k\in\mathcal{K}$, and we define $\tilde{\boldsymbol{y}}_j=(\tilde{y}_{j,1},\ldots,\tilde{y}_{j,K})$ for all $j\in\mathcal{J}$. This construction requires $\mathcal{O}(K)$ arithmetic operations and guarantees that $\|\boldsymbol{y}_j-\tilde{\boldsymbol{y}}_j\|_\infty\leq\frac{1}{M}\leq\frac{\varepsilon}{8KU}$ for all $j\in\mathcal{J}$. Finally, we introduce the perturbed two-point distribution $\tilde{\nu}_t=t\delta_{\tilde{\boldsymbol{y}}_1}+(1-t)\delta_{\tilde{\boldsymbol{y}}_2}$. All support points of μ and ν have rational coordinates that are representable as fractions of two integers with absolute values at most U. Therefore, μ and ν are supported on $[-U,U]^K$. Similarly, as U and M are integers, which implies that U is an integer multiple of $\frac{1}{M}$, and as all support points of $\tilde{\mu}$ and $\tilde{\nu}$ are obtained by mapping the support points of μ and ν to the nearest lattice points in $\mathbb{Z}^K/M$, respectively, the perturbed distributions $\tilde{\mu}$ and $\tilde{\nu}$ must also be supported on $[-U,U]^K$. Lemma 4.8 therefore guarantees that $|W_c(\mu,\nu_t)-W_c(\tilde{\mu},\tilde{\nu}_t)|\leq\varepsilon$.

In the remainder of the proof we will estimate the number of arithmetic operations needed to compute $W_c(\tilde{\mu},\tilde{\nu}_t)$. Note first that the coordinates of all support points of $\tilde{\mu}$ and $\tilde{\nu}_t$ are fractions of integers with absolute values of at most $\tilde{U}=MU$. To see this, recall that $x_k^l=a_k^l/b_k^l$ for some $a_k^l\in\mathbb{Z}$ and $b_k^l\in\mathbb{N}$ with $|a_k^l|,|b_k^l|\leq U$. Using 'round' to denote the rounding operator that maps any real number to its nearest integer, we can express $\tilde{x}_k^l$ as $\tilde{a}_k^l/\tilde{b}_k^l$ with $\tilde{a}_k^l=\text{round}(Mx_k^l)\in\mathbb{Z}$ and $\tilde{b}_k^l=M\in\mathbb{N}$. By construction, we have $|\tilde{a}_k^l|\leq MU=\tilde{U}$ and $\tilde{b}_k^l=M\leq\tilde{U}$ for all $k\in\mathcal{K}$ and $l\in\mathcal{L}$. Similarly, one can show that $\tilde{y}_{j,k}$ is representable as a fraction of two integers with absolute values of at most $\tilde{U}$ for all $j\in\mathcal{J}$ and $k\in\mathcal{K}$. As $M\leq\tilde{U}$,

$\tilde{\mu}$ and $\tilde{\nu}$ thus satisfy all assumptions of Corollary 4.7 with $\tilde{U}$ instead of U, respectively. From the proof of this corollary we may therefore conclude that $W_c(\tilde{\mu}, \tilde{\nu}_t)$ can be computed in $\mathcal{O}(KL\log_2(KL) + K^2\tilde{U}^4)$ arithmetic operations using Algorithm 2. As $\tilde{U} = MU = \mathcal{O}(KU^2/\varepsilon)$, this establishes the claim about the number of arithmetic operations. From the definitions of $\tilde{U}$ and M and from the analysis of Algorithm 2 in Theorem 4.1, it is clear that the bit lengths of all numbers computed by the proposed procedure are indeed polynomially bounded in K, L, $\log_2(U)$ and $\log_2(\frac{1}{\varepsilon})$. This observation completes the proof. □

Theorem 4.9 shows that an ε-approximation of $W_c(\mu, \nu_t)$ can be computed with a number of arithmetic operations that grows only polynomially with K, L, U and $\frac{1}{\varepsilon}$ but exponentially with $\log_2(U)$ and $\log_2(\frac{1}{\varepsilon})$. By Remark 4.3 *(iii)*, approximations of $W_c(\mu, \nu_t)$ can therefore be computed in pseudo-polynomial time.

Acknowledgements. This research was supported by the Swiss National Science Foundation under the NCCR Automation, grant agreement 51NF40_180545. The research of the second author is supported by an Early Postdoc.Mobility Fellowship, grant agreement P2ELP2_195149. The research of the last author was partially supported by the MOE Academic Research Fund Tier 2 grant T2MOE1906.

References

Brahim Khalil Abid and Robert Gower. Stochastic algorithms for entropy-regularized optimal transport problems. In *Artificial Intelligence and Statistics*, pages 1505–1512, 2018.

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*, pages 1964–1974, 2017.

Jason M Altschuler and Enric Boix-Adsera. Hardness results for multimarginal optimal transport problems. *arXiv:2012.05398*, 2020.

Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters are NP-hard to compute. *arXiv:2101.01100*, 2021.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

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

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

Dimitri P Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1): 152–171, 1981.

Dimitri P Bertsekas. Auction algorithms for network flow problems: A tutorial introduction. *Computational Optimization and Applications*, 1(1):7–66, 1992.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Jose Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards optimal running times for optimal transport. *arXiv:1810.07717*, 2018.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *Artificial Intelligence and Statistics*, pages 880–889, 2018.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Mathieu Carriere, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In *International Conference on Machine Learning*, pages 664–673, 2017.

Türkü Özlüm Çelik, Asgar Jamneshan, Guido Montúfar, Bernd Sturmfels, and Lorenzo Venturello. Wasserstein distance to independence models. *Journal of Symbolic Computation*, 104:855–873, 2021.

Deeparnab Chakrabarty and Sanjeev Khanna. Better and simpler error analysis of the Sinkhorn-Knopp algorithm for matrix scaling. *Mathematical Programming*, pages 1–13, 2020. Forthcoming.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 2009.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

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

George B Dantzig. Application of the simplex method to a transportation problem. *Activity Analysis and Production and Allocation*, pages 359–373, 1951.

George B Dantzig and Mukund N Thapa. *Linear Programming 2: Theory and Extensions*. Springer, 2003.

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the ROT mover's distance. *Journal of Machine Learning Research*, 19(1):590–642, 2018.

Anulekha Dhara, Bikramjit Das, and Karthik Natarajan. Worst-case expected shortfall with univariate and bivariate marginals. *INFORMS Journal on Computing*, 33(1):370–389, 2021.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International Conference on Machine Learning*, pages 1367–1376, 2018.

Martin Dyer. Approximate counting by dynamic programming. In *ACM Symposium on Theory of Computing*, pages 693–699, 2003.

Martin Dyer and Leen Stougie. Computational complexity of stochastic programming problems. *Mathematical Programming*, 106:423–432, 2006.

Martin Dyer and Leen Stougie. Erratum to: Computational complexity of stochastic programming problems. *Mathematical Programming*, 153:723–725, 2015.

Martin Dyer, Alan Frieze, Ravi Kannan, Ajai Kapoor, Ljubomir Perkovic, and Umesh Vazirani. A mildly exponential time algorithm for approximating the number of solutions to a multidimensional knapsack problem. *Combinatorics, Probability and Computing*, 2(3):271–284, 1993.

Montacer Essid and Justin Solomon. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, 2004.

Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv:1604.02199*, 2016.

Michael R Garey and David S Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

Parikshit Gopalan, Adam Klivans, Raghu Meka, Daniel Štefankovic, Santosh Vempala, and Eric Vigoda. An FPTAS for #knapsack and related counting problems. In *Foundations of Computer Science*, pages 817–826, 2011.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 2012.

Wenshuo Guo, Nhat Ho, and Michael I Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pages 2088–2097, 2020.

Grani A Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann. A comment on "computational complexity of stochastic programming problems". *Mathematical Programming*, 159(1-2):557–569, 2016.

Arun Jambulapati, Aaron Sidford, and Kevin Tian. A direct $\tilde{O}(1/e)$ iteration parallel algorithm for optimal transport. In *Advances in Neural Information Processing Systems*, pages 11359–11370, 2019.

Mark Jerrum. *Counting, sampling and integrating: Algorithms and complexity*. Springer Science & Business Media, 2003.

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

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *ACM Symposium on Theory of Computing*, pages 302–311, 1984.

Soheil Kolouri and Gustavo K Rohde. Transport-based single frame super resolution of very low resolution face images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2015.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34 (4):43–59, 2017.

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

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 *IEEE Symposium on Foundations of Computer Science*, pages 424–433, 2014.

Wuchen Li, Stanley Osher, and Wilfrid Gangbo. A fast algorithm for earth mover's distance based on optimal transport and l_1 type regularization. *arXiv:1609.07092*, 2016.

Tianyi Lin, Nhat Ho, and Michael I Jordan. On the efficiency of the Sinkhorn and Greenkhorn algorithms for optimal transport. *arXiv:1906.01437*, 2019a.

Tianyi Lin, Nhat Ho, and Michael I. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pages 3982–3991, 2019b.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Ben Morris and Alistair Sinclair. Random walks on truncated cubes and sampling 0-1 knapsack solutions. *SIAM Journal on Computing*, 34(1):195–226, 2004.

Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen. Tsallis regularized optimal transport and ecological inference. In *Association for the Advancement of Artificial Intelligence*, pages 2387–2393, 2017.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming.* SIAM, 1994.

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

Nicolas Papadakis and Julien Rabin. Convex histogram-based joint image segmentation with regularized optimal transport cost. *Journal of Mathematical Imaging and Vision*, 59(2):161–186, 2017.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Kent Quanrud. Approximating optimal transport with linear programs. In *Symposium on Simplicity in Algorithms*, pages 6:1–6:9, 2019.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *Artificial Intelligence and Statistics*, pages 630–638, 2016.

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

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *International Conference on Learning Representations*, 2018.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics*, 33(4):67, 2014.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66, 2015.

Daniel Štefankovič, Santosh Vempala, and Eric Vigoda. A deterministic polynomial-time approximation scheme for counting knapsack solutions. *SIAM Journal on Computing*, 41(2):356–366, 2012.

Guillaume Tartavel, Gabriel Peyré, and Yann Gousseau. Wasserstein loss for image synthesis and restoration. *SIAM Journal on Imaging Sciences*, 9(4):1726–1755, 2016.

Bahar Taskesen, Soroosh Shafieezadeh-Abadeh, and Daniel Kuhn. Semi-discrete optimal transport: Hardness, regularization and numerical solution. *arXiv:2103.06263*, 2021.

Leslie G Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979a.

Leslie G Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2): 189–201, 1979b.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2008.