New Algorithmic Directions in Optimal Transport and Applications for Product Spaces

Salman Beigi
Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
TEIAS, Khatam University, Tehran, Iran

Omid Etesami
EPFL, Campus Biotech, Geneva, Switzerland
Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Mohammad Mahmoody
Charlottesville, VA, USA

Amir Najafi
Sharif University of Technology, Tehran, Iran

Abstract

We consider the problem of optimal transport between two high-dimensional distributions μ, ν in $\mathbb{R}^n$ from a new algorithmic perspective, in which we are given a sample $x \sim \mu$ and we have to find a close $y \sim \nu$ while running in poly(n) time, where n is the size/dimension of x, y. In other words, we are interested in making the running time bounded in dimension of the spaces rather than bounded in the total size of the representations of the two distributions. Our main result is a general algorithmic transport result between any product distribution μ and an arbitrary distribution ν of total cost $\Delta + \delta$ under ℓ_p^p cost; here Δ is the cost of the so-called Knothe–Rosenblatt transport from μ to ν, while δ is a computational error that goes to zero for larger running time in the transport algorithm. For this result, we need ν to be "sequentially samplable" with a "bounded average sampling cost" which is a novel but natural notion of independent interest. In addition, we prove the following.

- We prove an algorithmic version of the celebrated Talagrand's inequality for transporting the standard Gaussian distribution Φ^n to an arbitrary ν under the Euclidean-squared cost. When ν is Φ^n conditioned on a set $\mathcal{S}$ of measure ε, we show how to implement the needed sequential sampler for ν in expected time poly(n/ε), using membership oracle access to $\mathcal{S}$. Hence, we obtain an algorithmic transport that maps Φ^n to $\Phi^n|\mathcal{S}$ in time poly(n/ε) and expected Euclidean-squared distance $O(\log 1/\varepsilon)$, which is optimal for a general set $\mathcal{S}$ of measure ε.
- As corollary, we find the first *computational* concentration (Etesami et al. SODA 2020) result for the Gaussian measure under the Euclidean distance with a *dimension-independent* transportation cost, resolving a question of Etesami et al. More precisely, for any set $\mathcal{S}$ of Gaussian measure ε, we map most of Φ^n samples to $\mathcal{S}$ with Euclidean distance $O\left(\sqrt{\log 1/\varepsilon}\right)$ in time poly(n/ε).

2012 ACM Subject Classification Theory of computation → Algorithm design techniques; Theory of computation → Online algorithms; Mathematics of computing → Probabilistic algorithms; Theory of computation → Probabilistic computation

Keywords and phrases Optimal transport, Randomized algorithms, Concentration bounds

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2025.10

Related Version *Full Version*: `https://arxiv.org/abs/2509.21502`

Funding *Omid Etesami*: Thanks to Sabanci University and TEIAS for their support during part of this work.

1 Introduction

Optimal transport (OT) is a fundamental problem that arises in mathematics, science, and engineering, including differential geometry [17], transportation planning [40], economics [21], machine learning [34, 38], image registration [23], and seismic tomography [35]. In machine

learning, it has been used in unsupervised learning [46], as a measure of the cost of misclassification [20], to define the fairness of algorithms [11], in Wasserstein GANs [2], for transfer learning [14], and in diffusion generative models [47, 26].

In the optimal transport problem, we would like to transport samples from a source distribution μ to points in the target distribution ν with a minimum expected "transportation cost" $\mathsf{c}(x, y)$ of transporting $x \sim \mu$ to $y \sim \nu$. The study of this problem dates back to the work of Monge [33], who wanted the transportation mapping $A(x) = y$ to be deterministic. Kantorovich [25] reformulated the problem by allowing $A(x)$ to be a randomized (stochastic) mapping. In other words, we now look for a coupling π over the distributions μ, ν with minimum expected transportation cost $\mathbb{E}\,\mathsf{c}(x, y)$, using which we define the optimal cost of transporting μ to ν,

$$\mathsf{T}(\mu, \nu) = \min_{\pi \in \mathcal{C}} \mathop{\mathbb{E}}_{(x,y)\sim\pi} \mathsf{c}(x, y)$$

where $\mathcal{C}$ is the set of all couplings between μ, ν. OT is closely related to the notion of "Wasserstein metric" that generalizes OT using a parameter $p \geq 1$ and is the same for $p = 1$.

As a prominent example of the use of OT in mathematics, Talagrand [43] gave a bound on the optimal transport, under the ℓ_2^2 cost, of the n-dimensional Gaussian measure Φ^n to an arbitrary distribution ν based on the KL-divergence of ν from Φ^n. Using this, he derived an essentially optimal concentration of measure result, showing that for any set $\mathcal{S}$ of measure ε in Φ^n, almost all of the measure Φ^n is within ℓ_2^2 (minimum) distance $O(\ln 1/\varepsilon)$ from $\mathcal{S}$.

Computational OT. Computational aspects of OT have recently become extremely important on their own [38]. In the most common formulation of "computational OT", we would like to compute or estimate $\mathsf{T}(\mu, \nu)$ efficiently. Computing $\mathsf{T}(\mu, \nu)$ is a key tool, e.g., for applications that use OT to quantify a loss that allows one to know "how far" we are from a target goal [4, 6, 7]. A common approach to computing $\mathsf{T}(\mu, \nu)$ is to work with empirical sample sets $\mathcal{S} \sim \mu^m, \mathcal{T} \sim \nu^m$, and find the best OT between the empirical distributions $U_{\mathcal{S}}, U_{\mathcal{T}}$ that are uniform over $\mathcal{S}, \mathcal{T}$ (e.g., see [24, 32] and the references therein). This approximation converges to the quantity $\mathsf{T}(\mu, \nu)$ in the limit, and the OT between $U_{\mathcal{S}}, U_{\mathcal{T}}$ can be computed using the Hungarian algorithm for minimum weighted matching [29]. The popular iterative Sinkhorn algorithm solves a regularized version of the OT problem [42] but it also works with empirical sample sets, that is, i.i.d. samples from the distributions. Using empirical samples, one does not rapidly converge to the optimal OT in some elementary cases. For example, to transport the uniform distribution on the n-dimensional unit cube to itself, the OT between two poly(n)-size empirical versions of the original distribution is $\Theta(\sqrt{n})$ in ℓ_2^2 distance even though the actual OT cost is zero.

Statistical OT. The above approach of using empirical samples $\mathcal{S} \sim \mu^m, \mathcal{T} \sim \nu^m$ can in fact be used to approximate the *transport map* itself from μ to ν, as in [24, 32]. For example, Brenier's theorem [10, 28] asserts that under the ℓ_2^2 cost and suitable conditions, a unique Monge mapping achieves optimal transport, and one can aim at approximating this deterministic mapping. This approach is sometimes known as *statistical* optimal transport [13]. However, this approach needs exponential in n samples for n-dimensional distributions to achieve good approximate results. Some previous works like [24, 32] make improvements on this analysis by assuming further smoothness and structural conditions on the distributions but the curse of dimensionality basically remains intact. More importantly, to the best of our knowledge, no previous work models the algorithmic aspect of searching for the transport map by limiting its algorithm to run in polynomial time over the size of the input $x \sim \mu$.

1.1 Our Contributions

In a nutshell, our contributions are (1) formalizing a new theory of algorithmic transport, (2) obtaining initial results on algorithmic transport for the high-dimensional setting, and (3) obtaining applications for algorithmic transport, e.g., to algorithmic concentration of measure. Each of the items above has multiple aspects that are elaborated in the following.

Algorithmic Transport in Polynomial Time. The common computational OT formulation aims to compute or approximate the optimal transportation cost $\mathsf{T}(\mu, \nu)$, yet it does not answer the key question of how to algorithmically compute the *transport map* efficiently over the size of the given input sample. I.e., suppose that we are given a particular sample $x \sim \mu$ as input, and we would like to map it to $y \sim \nu$ as follows: (1) The mapping shall be computed in polynomial time over the size of the input $|x| = n$. (2) We would like to control the expected cost of the transportation. To point out the subtle distinction between our new algorithmic formulation and the traditional computational OT, in this work we use the term *algorithmic transport* to refer to the task of computing a (randomized) mapping A efficiently based on its input size $|x|$ (e.g., the dimension of x), such that $A(x) \sim \nu$, whenever $x \sim \mu$.

Algorithmic transport, when done optimally, can be used to approximate OT cost efficiently as well. In particular, when the transportation cost is bounded by a constant, using $k = \Theta(\varepsilon^{-2})$ independent samples $(x_1, y_1), \ldots, (x_k, y_k) \sim (x, A(x))^k$, the average $\mathbb{E}_i\, \mathsf{c}(x_i, y_i)$ gives an ε-approximation of the OT, with high probability. However, it is not clear how to do the reverse and obtain algorithmic transport from computational OT.

When μ, ν are one dimensional, for natural (convex) costs such as $\mathsf{c}(x, y) = |x - y|^p, p \geq 1$ one can find simple algorithms that simply use a "monotone" transportation plan [45]. Furthermore, when the distributions μ, ν have small domains of size k, one can use algorithms based on min-cost flows to find a full description of the OT from μ to ν in $\mathrm{poly}(k)$ time [37]. However, our focus is on the high-dimensional setting and finding $\mathrm{poly}(n)$-time computable mappings between distributions of dimension n with *super*-polynomial support. We ask:

> *If μ, ν are n-dimensional distributions, how can we find a $\mathrm{poly}(n)$-time computable transport map from $x \sim \mu$ to $y \sim \nu$ of a small/optimal cost?*

Formalizing and answering the question above in various contexts is the focus of our work. Our studies also bear similarities to the line of work on approximating the total variation distance [5, 16] as it coincides with OT under the Hamming distance.

Transport in High-Dimensional Setting. In this work, we approach the main question above through a study of so-called *causal* transports [30, 3] in high dimension, in which the transporting algorithm A produces $y = (y_1, \ldots, y_n)$ from $x = (x_1, \ldots, x_n)$ in an *online* manner: The algorithm A shall output y_i based on $x_{[i]} = (x_1, \ldots, x_i)$ and before receiving x_{i+1}. Hence we also refer to those transports simply as online. The so-called Knothe-Rosenblatt transport (KR transport for short) [27, 39] is an important online transport with two properties: (1) its reverse is also online, and (2) it follows a "greedy" approach in each round by using a monotone mapping of dimension one. Our motivations for studying online transports is twofold: (1) Despite being a *restriction* on how the transport is done, the "online lens" guides us towards algorithm development; (2) In our eyes, information-theoretic study of online algorithms is interesting. In particular, in Section 2.1, we prove that the KR transport is optimal among all online transports when the source distribution is a product.

Main Result: Algorithmic Transport from Product Distributions. Our main result (Theorem 28) is to design a poly(n)-time online algorithm that transports a generic product distribution $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ to any n-dimensional distribution ν, assuming that (1) the transportation cost satisfies $\mathsf{c}(x, y) = \sum_i \mathsf{c}_i(x_i, y_i)$, where $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n)$, and (2) the transporting algorithm A has oracle access to proper samplers for both μ, ν.

The algorithm is actually very simple: Given x, having determined $y_1, \ldots, y_{i-1}$, to determine y_i, it samples $k-1$ samples besides x_i according to μ_i. Similarly it samples k samples according to the conditional distribution of the ith coordinate of ν conditioned on the values of $y_1, \ldots, y_{i-1}$. Then it optimally matches the two sets of two k samples. The value of y_i is the match of x_i in this matching.

The transportation cost of A turns out to be $\Delta + \delta$, where Δ is the optimal cost of online transports from μ to ν (which, as we will prove, will coincide with the KR transport [27, 39] in our settings of interest), and δ is a term that could be made smaller by choosing k larger. We show that the reverse transport from ν back to the product μ can be done algorithmically as well. This will be useful for deriving further algorithmic transports through composition.

Sequential Samplers. When it comes to the samplability conditions needed in our main results above, we merely require that we can sample from μ_i efficiently. However, for the non-product distribution ν, the samplability condition is stronger and we require that one can sample from ν_i conditioned on any previously sampled prefix $y_{[i-1]}$. We refer to such samplers as *sequential* samplers. A key quantity of interest is the complexity of iterative sampling of the coordinates $y_1, \ldots, y_n$ sequentially (conditioned on previous ones) till we obtain a full sample y. We would like to have samplers where the average complexity of this sequential generation is bounded. As it turns out, we can indeed bound such costs in our special cases of interest.

From a real-world application point of view, this notion of efficient sequential sampler is very natural in some generative models. This is indeed the case for transformer-based language models that autoregressively generate their tokens one by one, each conditioned on the previously sampled sequence of tokens [44, 18]. That is, the joint distributions produced by these generative models have sequential samplers of low expected cost, as they indeed generate their sequence of symbols in a reasonable time and in an online fashion.

Algorithmic Transport for the Standard Gaussian Distribution. One of the fundamental results in OT is Talagrand's transportation inequality for the n-dimensional Gaussian distribution Φ^n [43]. It is proved that for every distribution ν, $\mathsf{T}(\Phi^n, \nu) \le 2\mathsf{KL}(\nu, \Phi^n)$, in which the cost is measured in ℓ_2^2, i.e., $\mathsf{c}(x, y) = \sum_{i \in [n]} |x_i - y_i|^2$, and $\mathsf{KL}(\cdot, \cdot)$ denotes the Kullback–Leibler divergence. In this work, we lift this classical result to the algorithmic setting. Note that, as mentioned in [43], this bound is optimal *in general*, e.g., when ν is a shifted Φ^n, in which case our results converge to this optimal bound as well. In particular, we derive this result from our main result by proving the following two complementary claims:

- *Information theoretic:* We observe that Talagrand's bound of $2\mathsf{KL}(\nu, \Phi^n)$ upper bounds not only the best "offline" transport from the standard Gaussian Φ^n, but also the best *optimal online* transportation of Φ^n to ν. Namely, we show that $\Delta \le 2\mathsf{KL}(\nu, \Phi^n)$, where Δ is the optimal online transportation cost as defined above.
- *Computational:* We use results from [19] to show that the Gaussian distribution in one dimension has a small transportation cost to its empirical samples on average.

Transporting Standard Gaussian to Conditional Gaussian. We show that in a natural setting, where ν is the Gaussian distribution conditioned on an event $\mathcal{S}$ of Gaussian measure ε, such sequential samplers can be efficiently simulated using oracle access to membership tests in $\mathcal{S}$. In other words, we find an algorithmic oracle-aided transportation algorithm that *simultaneously* work for all such distributions $\nu = \Phi^n|\mathcal{S}$. Note that such distributions have $2\mathsf{KL}(\nu, \Phi^n) \leq 2\ln 1/\varepsilon$. We obtain algorithmic transports running in expected time $\mathrm{poly}(n/\varepsilon)$ that achieve transport cost that converges to the upper bound of Talagrand.

Dimension-Independent Computational Concentration for Gaussian Spaces. One of the applications of OT is to obtain concentration of measure (CoM) inequalities [22]: One shows that any set $\mathcal{S}$ of "sufficiently large" measure in a concentrated metric probability space (μ, d), where μ is a distribution and d is a distance metric, expands to cover most of the measure in μ, when we consider neighbors of $\mathcal{S}$ within a certain distance. Recently, a *computational* (algorithmic) variant of the CoM phenomenon has been introduced [31, 15], in which one aims to show that the reverse mapping can be computed efficiently from almost all of the points in μ back to $\mathcal{S}$ by moving the points within a bounded distance. Namely, given a typical sampled point $x \sim \mu$, we aim to algorithmically find a "close neighbor" $y \in \mathcal{S}$ of bounded distance $\mathrm{d}(x, y)$. The work of [15] obtained such results for various settings, but their work left open obtaining computational CoM with dimension-independent (optimal) distance for the basic and natural space of Gaussian distributions under the ℓ_2 distance. Using our oracle set-transportation result for Gaussian spaces mentioned above, we resolve this open question and obtain such an optimal and dimension-free bound (see Corollary 36).

Reductions for (Deriving New) Algorithmic Transport. Finally, considering the role of reductions in resolving algorithmic tasks, we also develop the (right) notion of algorithmic reductions for the goal of relating algorithms for (optimal) transport across different spaces. In particular, suppose μ_1, μ_2 are distributions and $\mathsf{c}_1, \mathsf{c}_2$ are two different transportation costs. In the full version we state conditions under which, we can automatically transform an algorithmic transport result from μ_1 to ν (under the cost c_1) to a similar result that transports μ_2 to ν (under the cost c_2) for specific distributions μ_1, μ_2 and arbitrary distribution ν. We then show how to realize such reductions when we transport uniform distributions over the unit cube and the unit sphere (to an arbitrary distribution) by reducing them to the case of transporting Gaussian distributions. Consequently, we obtain algorithmic transports from these distributions as well.

2 Basic Concepts

In this section, we define the key notions studied in this paper and prove their basic properties.

Notation. We let $[n] = \{1, \ldots, n\}$. We denote the source (initial) distribution as μ. When μ is distributed over $\mathbb{R}^n$, we say that μ has dimension n and by μ_i we denote the distribution of its ith coordinate. We usually denote $x \sim \mu$, where $x = (x_1, \ldots, x_n)$ and $x_i \sim \mu_i$. $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ means that μ is a product distribution. We use a similar notation for the target distribution ν. By $y \leftarrow A(x)$ we denote the process of running a probabilistic algorithm A on input x to obtain output y. When μ is a distribution, A^μ denotes an oracle algorithm A that has access to fresh samples from μ, and when $\mathcal{S}$ is a set, $A^{\mathcal{S}}$ denotes a similar situation where the oracle responds membership in $\mathcal{S}$. For vector $(v_1, \ldots, v_n)$, by $v_{[i]}$ we denote the prefix vector $(v_1, \ldots, v_i)$. When a distribution μ of dimension n with marginals $\mu_1, \ldots, \mu_n$ is clear from the context, by $\mu_i|x_{[i-1]}$, we denote the distribution of μ_i conditioned on having sampled x_j from μ_j for all $j < i$. For further clarity on the

underlying joint distribution, we might sometimes use $\mu_i|_\mu x_{[i-1]}$ instead. By $\mu(\mathcal{S})$ or $\Pr_\mu[\mathcal{S}]$ we denote the probability of event $\mathcal{S}$ under the distribution μ. Whenever it is clear from the context, for an outcome x, we use $\mu(x)$ to either denote the probability of the outcome x or the density of μ at x depending on whether μ is discrete or continuous. By $\mathrm{Supp}(\mu)$ we denote the support set of μ, which for the discrete and continuous cases can be defined as $\{x \mid \mu(x) > 0\}$. When $\mathrm{Supp}(\mu) \cup \mathrm{Supp}(\nu) \subseteq \mathcal{S}$, their Kullback–Leibler (KL) divergence is denoted as $\mathsf{KL}(\nu, \mu) = \sum_{x \in \mathcal{S}} \nu(x) \log(\nu(x)/\mu(x))$ with the natural logarithm basis. In the preceding definition and generally throughout this paper, we use the summation notation corresponding to discrete distributions; the corresponding results for continuous distributions replace sums with proper integrals. For $p \geq 1$, the ℓ_p-norm and ℓ_p-distance over $\mathbb{R}^n$ are defined as $\ell_p(x) = \|x\|_p = \left(\sum_{i \in [n]} |x_i|^p\right)^{1/p}$, and $\ell_p(x, y) = \ell_p(x - y)$.

Transportation Costs. In the following, all transportation *costs*, usually denoted as c, are functions $\mathsf{c} \colon \mathbb{R}^{2n} \mapsto \mathbb{R}$ with non-negative outputs that model the cost of transporting $x \sim \mu$ to $y \sim \nu$. We always assume c to be lower semi-continuous but do not assume c to be symmetric or satisfy the triangle inequality; we state these conditions, whenever needed.

▶ **Definition 1** (Coupling and Optimal Transportation Cost). *We say that a distribution π over pairs with marginals π_1, π_2 is a coupling of μ, ν if $\pi_1 = \mu, \pi_2 = \nu$. If for every $x \sim \mu$, there is a unique y such that $(x, y) \in \mathrm{Supp}(\pi)$, then we call this a deterministic (Monge) transport from μ to ν. For a cost c, the transport cost of a coupling π of μ, ν is defined as*

$$\mathsf{T}_\mathsf{c}(\pi) = \mathop{\mathbb{E}}_{(x,y)\sim\pi} \mathsf{c}(x, y).$$

We refer to $\mathsf{T}^{1/p}_{\mathsf{c}^p}(\pi)$ as the (Wasserstein) p-cost of π under c. If $\mathcal{C}(\mu, \nu)$ denotes the set of all couplings between μ, ν, the (Kantorovich) optimal transportation cost *for (μ, ν) is defined as*

$$\mathsf{T}_\mathsf{c}(\mu, \nu) = \inf_{\pi \in \mathcal{C}(\mu,\nu)} \mathsf{T}_\mathsf{c}(\pi).$$

The infimum in Definition 1 for defining the optimal transportation costs turns out to be a minimum as c is lower-semi continuous [1].

▶ **Definition 2** (Algorithmic Transport). *For distributions μ, ν, algorithm A is a* transport *from distribution μ to distribution ν if A is a (probabilistic) algorithm such that $A(x) \sim \nu$ whenever $x \sim \mu$. By π_A we denote the coupling created by A. For a transportation cost c the transportation cost of A is defined as $\mathsf{T}_\mathsf{c}(A) = \mathsf{T}_\mathsf{c}(\pi_A)$.*

Computational Model. In Definition 2, we need to either work with discrete distributions with samples of finite length, or when the distributions are continuous we need to work with the generalization of *algorithms working with real numbers* as formalized in [8, 9].

We now define an algorithmic variant of so-called *causal* transport [30] with a discrete time [3], We call it "online" to emphasize on the algorithmic aspect a la *online learning* [41].

▶ **Definition 3** (Online (Algorithmic) Transport). *For distributions μ, ν of dimension n, we call a (probabilistic and perhaps computationally unbounded) algorithm A an* online transport *algorithm from μ to ν if it forms a transport from μ to ν, while it makes its decisions in an online way. Namely, A has an internal iterating process (for simplicity also denoted by A) that reads $(x_1, \ldots, x_n) \sim \mu$ coordinate by coordinate while holding an internal state, initially $s_0 = \varnothing$. In the ith iteration, we have $(s_i, y_i) \leftarrow A(s_{i-1}, x_i)$, and at the end we output $(y_1, \ldots, y_n) \sim \nu$. We also let $\mathcal{C}^{\mathrm{OnT}}(\mu, \nu)$ to be the set of all couplings that can be obtained by online algorithms and for a transport cost c obtain the optimal online transportation cost as*

$$\mathsf{T}^{\mathrm{OnT}}_\mathsf{c}(\mu, \nu) = \inf_{\pi \in \mathcal{C}^{\mathrm{OnT}}(\mu,\nu)} \mathsf{T}_\mathsf{c}(\pi).$$

To contrast and emphasize on a transport not being necessarily online, we refer to (potentially) non-online transports as *offline* transports.

We now define a class of couplings that is closely related to online transport.

▶ **Definition 4** (Online Coupling). *Suppose π is a coupling between n-dimensional distributions μ, ν, and π_i is the corresponding marginal coupling between μ_i, ν_i. We call π an* online coupling *if for all $z = (x_{[i-1]}, y_{[i-1]}) \in \mathrm{Supp}(\pi_{[i-1]})$, $\pi_i|z$ is a coupling of $\mu_i|x_{[i-1]}$ (according to μ) and $\nu_i|y_{[i-1]}$ (according to ν). If $\mathcal{C}^{\mathrm{OnC}}(\mu,\nu)$ denotes the set of all online couplings between μ, ν, for a transport cost c we obtain the optimal online coupling cost between μ, ν as*

$$\mathsf{T}_{\mathsf{c}}^{\mathrm{OnC}}(\mu,\nu) = \inf_{\pi \in \mathcal{C}^{\mathrm{OnC}}(\mu,\nu)} \mathsf{T}_{\mathsf{c}}(\pi).$$

We now show how to characterize online couplings using online transports.

▶ **Proposition 5.** *A coupling π between μ, ν is* online *if and only if it can be obtained through both an online transport from μ to ν and an online transport from ν to μ.*

▶ **Definition 6.** *We call the cost function c over $\mathbb{R}^n \times \mathbb{R}^n$* linear *over $\mathsf{c}_1, \ldots, \mathsf{c}_n$, if $\mathsf{c}(x,y) = \mathsf{c}_1(x_1,y_1) + \cdots + \mathsf{c}_n(x_n,y_n)$, for all $x = (x_1,\ldots,x_n), y = (y_1,\ldots,y_n)$.*

Greedy Coupling. One might wonder how we can compute/approximate $\mathsf{T}_{\mathsf{c}}^{\mathrm{OnC}}(\mu,\nu)$. One approach is to use greedy methods, by trying to use an optimal coupling in each round. This is formalized in the following definition in settings with dedicated costs for each round. We will then discuss when this method succeeds in Theorem 10. More generally, we define locally-optimal couplings, even when they are not online.

▶ **Definition 7** (Locally Optimal and Greedy Couplings). *Suppose the cost function c over $\mathbb{R}^{2n}$ is linear over $\mathsf{c}_1, \ldots, \mathsf{c}_n$. A coupling π between μ, ν is* locally optimal, *if for every $z_{[i-1]} \in \mathrm{Supp}(\pi_{[i-1]})$, it holds that $\pi_i|z_{[i-1]}$ is an OT between $\mu_i|z_{[i-1]}, \nu_i|z_{[i-1]}$; i.e.,*

$$\mathsf{T}_{\mathsf{c}_i}(\pi_i|z_{[i-1]}) = \mathsf{T}_{\mathsf{c}_i}(\mu_i|z_{[i-1]}, \nu_i|z_{[i-1]}).$$

When π is an online coupling as well, the above condition simplifies to $\mathsf{T}_{\mathsf{c}_i}(\pi_i|z_{[i-1]}) = \mathsf{T}_{\mathsf{c}_i}(\mu_i|x_{[i-1]}, \nu_i|y_{[i-1]})$ in which case we call π greedy. *For $\mathcal{C}^{\mathrm{G}}(\mu,\nu)$ denoting the set of all greedy couplings from μ to ν, we define*

$$\mathsf{T}_{\mathsf{c}}^{\mathrm{G}}(\mu,\nu) = \sup_{\pi \in \mathcal{C}^{\mathrm{G}}(\mu,\nu)} \mathsf{T}_{\mathsf{c}}(\pi).$$

▶ Remark 8 (Greedy vs. Knothe-Rosenblatt Transports). Greedy couplings are closely related to Knothe-Rosenblatt (KR for short) transports [27, 39]. Specifically, for a greedy coupling π, when the cost functions c_i are convex, for any $z_{[i-1]} \sim \pi_{[i-1]}$, the locally optimal coupling $\pi_i|z_{[i-1]}$ could be obtained by simply using the unique monotone mapping [12]. Hence, KR coupling is a special case of greedy couplings and cover many interesting cases in this class. For example, when the cost function c is ℓ_p^p for $p \geq 1$, then $\mathsf{T}_{\mathsf{c}}^{\mathrm{G}}(\mu,\nu)$ equals the cost of the KR coupling between μ and ν. However, due to the generality of greedy couplings (e.g., for non-monotone costs) we define and use greedy transports.

Lambda and Delta Cost Functions. We now define two functions that play key roles in our analysis of the cost of online transports. The first (Lambda) function depends on a coupling, while the second one (Delta) depends on the two distributions that are coupled. As we prove later in Proposition 12, Lambda is a parameter that lower bounds the cost of any coupling. Delta is the optimal online transport from a product distribution to another one.

▶ **Definition 9** (The Lambda and Delta Functions). *For a coupling π of dimension n between distributions μ, ν of dimension n, and a cost function c that is linear over $\mathsf{c}_1, \ldots, \mathsf{c}_n$, we define the* Lambda functions *as*

$$\Lambda_{\mathsf{c}}(\pi) = \mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\mu_i|z_{[i-1]}, \nu_i|z_{[i-1]}).$$

We also define the Delta function *between distributions μ, ν of dimension n as*

$$\Delta_{\mathsf{c}}(\mu, \nu) = \mathop{\mathbb{E}}_{y\sim\nu} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\mu_i, \nu_i|y_{[i-1]}).$$

Note that the coupling π in Definition 9 does not have to be online. Furthermore, the definition of $\Lambda(\cdot)$ does depend on the order of the coordinates of the n-dimension distributions.

2.1 Online Coupling and Transport from Products

We end this section by stating a theorem showing that, whenever μ is product, any online coupling that is "locally optimal" in the sense that given history $z = (x_{[i-1]}, y_{[i-1]})$ it finds (an arbitrary) optimal transport between $(\mu_i), (\nu_i|y_{[i-1]})$, finds an optimum online coupling between μ, ν as well as an optimal online transport from μ to ν. This theorem does not assume convexity of the costs. As stated in Remark 8, for convex transportation costs, greedy algorithms can be instantiated using the KR transform.

▶ **Theorem 10** (Optimal Online Coupling and Transport from Products). *If $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ is product and the cost function c is linear over $\mathsf{c}_1, \ldots, \mathsf{c}_n$, then*

$$\mathsf{T}_{\mathsf{c}}^{\mathrm{OnT}}(\mu, \nu) = \mathsf{T}_{\mathsf{c}}^{\mathrm{OnC}}(\mu, \nu) = \mathsf{T}_{\mathsf{c}}^{\mathrm{G}}(\mu, \nu) = \Delta_{\mathsf{c}}(\mu, \nu).$$

Before proving Theorem 10 we prove some basic tools that are used in the proof. The first lemma that we state can be obtained from a simple application of the linearity of expectation.

▶ **Lemma 11** (Cost Splitting). *Let π be a coupling between distributions μ, ν of dimensions n, and let π_i be the corresponding coupling between the marginals μ_i, ν_i. Suppose c is linear over $\mathsf{c}_1, \ldots, \mathsf{c}_n$, and ω is an n-dimensional distribution that is arbitrarily correlated with π. Then,*

$$\mathsf{T}_{\mathsf{c}}(\pi) = \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\pi_i) = \mathop{\mathbb{E}}_{z\sim\omega} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\pi_i|\omega_{[i-1]} = z_{[i-1]}).$$

In particular, we can choose $\omega = \nu$, $\omega = \mu$, or $\omega = \pi$ as special cases.

We now prove some basic properties of the two functions, showing how to use it and how to characterize it in some special settings. In summary, Lambda function lower bounds the transportation of every coupling, while Delta will play a key role in characterizing the transportation cost for product distributions.

▶ **Proposition 12** (Properties of Lambda and Delta Functions). *Suppose π couples μ, ν and c is linear. The Lambda function satisfies the following properties.*

1. Lower Bound: *For all π, $\Lambda_{\mathsf{c}}(\pi) \le \mathsf{T}_{\mathsf{c}}(\pi)$, and the equality holds iff π is locally optimal.*
2. Online Transports from Products: *If π is an online transport and $\mu = \mu_1 \otimes \cdots \otimes \mu_n$, then*

   $$\Lambda_{\mathsf{c}}(\pi) \ge \Delta_{\mathsf{c}}(\mu, \nu).$$

3. Online Coupling for Products: *If π is an online* coupling, *and μ is product then*

   $$\Lambda_{\mathsf{c}}(\pi) = \Delta_{\mathsf{c}}(\mu, \nu).$$

Proof of Proposition 12. We prove the claims in order.

1. By letting $\omega = \pi$ in Lemma 11, we get

$$\mathsf{T}_{\mathsf{c}}(\pi) = \mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\pi_i|z_{[i-1]}) \geq \mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\mu_i|z_{[i-1]}, \nu_i|z_{[i-1]}) = \Lambda_{\mathsf{c}}(\pi),$$

   where the inequality follows from the fact that $\mathsf{T}_{\mathsf{c}_i}(\cdot,\cdot)$ minimizes the transportation cost.
2. We first claim that, in this case, for every $z_{[i-1]} = (x_{[i-1]}, y_{[i-1]}) \sim \pi_{[i-1]}$, we have $\mu_i|z_{[i-1]} = \mu_i$. This is true, because (1) $(\mu_i|x_{[i-1]}, y_{[i-1]}) = (\mu_i|x_{[i-1]})$ and the fact that π is an online transport, and (2) $(\mu_i|x_{[i-1]}) = \mu_i$ by the fact that μ is a product. Therefore,

$$\Lambda_{\mathsf{c}}(\pi) = \mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\mu_i|z_{[i-1]}, \nu_i|z_{[i-1]}) = \mathop{\mathbb{E}}_{z=(x,y)\sim\pi} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\mu_i, \nu_i|z_{[i-1]}).$$

   We now use the right hand side. In analyzing the right hand side, we first use Lemma 11 (using $\omega = \pi$) and then sample x, y in reverse order,

$$\mathop{\mathbb{E}}_{(x,y)\sim\pi} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\mu_i, \nu_i|z_{[i-1]}) = \sum_{i\in[n]} \mathop{\mathbb{E}}_{y_{[i-1]}\sim\nu_{[i-1]}} \mathop{\mathbb{E}}_{x_{[i-1]}\sim\nu_{[i-1]}|y_{[i-1]}} \mathsf{T}_{\mathsf{c}_i}(\mu_i, \nu_i|y_{[i-1]}, x_{[i-1]}),$$

   where for each $i \in [n]$, we sample $(x_{[i-1]}, y_{[i-1]}) \sim \pi_{[i-1]}$ by first sampling $y_{[i-1]}$ and then sampling $x_{[i-1]}$ conditioned on $y_{[i-1]}$. Now, for every $y_{[i-1]} \sim \nu_{[i-1]}$, we claim that

$$\mathop{\mathbb{E}}_{x_{[i-1]}\sim\nu_{[i-1]}|y_{[i-1]}} \mathsf{T}_{\mathsf{c}_i}(\mu_i, \nu_i|y_{[i-1]}, x_{[i-1]}) \geq \mathsf{T}_{\mathsf{c}_i}(\mu_i, \nu_i|y_{[i-1]}).$$

   This claim follows from Part 2 of Proposition 19 and the fact that the average of $\nu_i|y_{[i-1]}, x_{[i-1]}$ over the choice of $x_{[i-1]} \sim \nu_{[i-1]}|y_{[i-1]}$ is $\nu_i|y_{[i-1]}$.
3. When the coupling π is further an online coupling, then the equality holds, because $(\nu_i|y_{[i-1]}, x_{[i-1]}) = (\nu_i|y_{[i-1]})$, and the last inequality above becomes an equality. ◀

Proof of Theorem 10. It is enough to prove the following two claims.

1. $\mathsf{T}^{\mathrm{G}}_{\mathsf{c}}(\mu,\nu) \leq \Delta_{\mathsf{c}}(\mu,\nu)$.
2. $\mathsf{T}^{\mathrm{OnT}}_{\mathsf{c}}(\mu,\nu) \geq \Delta_{\mathsf{c}}(\mu,\nu)$.

The reason is that we already know $\mathsf{T}^{\mathrm{OnT}}_{\mathsf{c}}(\mu,\nu) \leq \mathsf{T}^{\mathrm{G}}_{\mathsf{c}}(\mu,\nu)$ (as being greedy is a limitation), and so proving the two claims above would imply all the equalities of the theorem statement.

To prove the first claim, we observe that cost $\Delta_{\mathsf{c}}(\mu,\nu)$ can be achieved using (any) greedy algorithm that (by definition) optimally couples $\mu_i = \mu_i|x_{[i-1]}$ with $\nu_i|y_{[i-1]}$ in the ith step. In fact, *all* greedy coupling algorithms have the same cost $\Delta_{\mathsf{c}}(\mu,\nu)$ when one of the distributions is product.

To prove the second claim, let π be an online transport with cost $\mathsf{T}^{\mathrm{OnT}}_{\mathsf{c}}(\mu,\nu)$. Our claim follows from Parts 1 and 2 of Proposition 12, due to π being online and μ being a product.

$$\mathsf{T}^{\mathrm{OnT}}_{\mathsf{c}}(\mu,\nu) = \mathsf{T}_{\mathsf{c}}(\pi) \geq \Lambda_{\mathsf{c}}(\pi) \geq \Delta_{\mathsf{c}}(\mu,\nu).$$ ◀

3 Basic Tools

3.1 Composition and Triangle Inequalities

Multi-distribution Coupling and Composition. We now generalize the notion of coupling to more than two distributions and use it to define composition of (online) couplings.

▶ **Definition 13** (Multi-distribution Coupling). *A coupling π of $\mu_1, \ldots, \mu_n$ is a distribution over n-vectors such that the marginal of the ith coordinate is distributed as μ_i.*

▶ **Definition 14** (Composition of Couplings). *For coupling $\pi_{1,2}$ over μ_1, μ_2 and coupling $\pi_{2,3}$ over μ_2, μ_3, we define the composition $\pi_{1,3} = \pi_{2,3} \circ \pi_{1,2}$ of $\pi_{1,2}$ and $\pi_{2,3}$ as the marginal of the first and third coordinates of the (unique) coupling of μ_1, μ_2, μ_3 such that.*

1. *For $1 \le i < j \le 3$, the marginal distribution of (μ_i, μ_j) in $\pi_{1,2,3}$ is distributed as $\pi_{i,j}$.*
2. *In the coupling $\pi_{1,2,3}$, μ_1, μ_3 are independent, conditioned on $x_2 \sim \mu_2$.*

We now use Wasserstein p-cost, to state the following well-known triangle inequality.

▶ **Lemma 15** (Triangle Inequality for Wasserstein p-Costs). *Suppose a cost function c satisfies the triangle inequality (but not necessarily symmetry) and $p \ge 1$. Then, for every coupling π over μ_1, μ_2, μ_3 with marginal coupling $\pi_{i,j}, i < j$ over π_i, π_j, we have the following,*

$$\mathsf{T}^{1/p}_{\mathsf{c}^p}(\pi_{1,3}) \le \mathsf{T}^{1/p}_{\mathsf{c}^p}(\pi_{1,2}) + \mathsf{T}^{1/p}_{\mathsf{c}^p}(\pi_{2,3}).$$

The following proposition can be obtained from the triangle inequality of Lemma 15.

▶ **Proposition 16** (Triangle Inequality for Wasserstein p-Costs in Multi-Round Settings). *Let μ be a distribution over $\mathbb{R}^n$, and for every $i \in [n], x_{[i-1]} \in \mathrm{Supp}(\mu_{[i-1]})$ let $J(x_{[i-1]})$ be a distribution over* triples *of distributions over $\mathbb{R}$. Suppose c satisfies the triangle inequality and c^p is linear over $\mathsf{c}_1, \ldots, \mathsf{c}_n$ for $p \ge 1$. Then, the following holds.*

$$\left(\mathop{\mathbb{E}}_{x\sim\mu} \sum_{i\in[n]} \mathop{\mathbb{E}}_{(\nu_1,\nu_2,\nu_3)\sim J(x_{[i-1]})} \mathsf{T}_{\mathsf{c}_i}(\nu_1, \nu_3) \right)^{1/p}$$
$$\le \sum_{k\in[2]} \left(\mathop{\mathbb{E}}_{x\sim\mu} \sum_{i\in[n]} \mathop{\mathbb{E}}_{(\nu_1,\nu_2,\nu_3)\sim J(x_{[i-1]})} \mathsf{T}_{\mathsf{c}_i}(\nu_k, \nu_{k+1}) \right)^{1/p}$$

The following can be obtained from the definition of online transport and Lemma 15.

▶ **Lemma 17** (Properties of the Composition of Online Transports). *Consider an online transport $A_{1,2}$ from μ_1 to μ_2 with coupling $\pi_{1,2}$ and an online transport $A_{2,3}$ from μ_2 to μ_3 with coupling $\pi_{2,3}$. Let $\pi_{1,3} = \pi_{2,3} \circ \pi_{1,2}$ be the composed coupling. Then,*

1. *The coupling $\pi_{1,3}$ is an online coupling.*
2. *There is an algorithm $A_{1,3}$ that transports μ_1 to μ_3 as the coupling $\pi_{1,3}$, whose complexity is bounded by running $A_{1,2}$ followed by running $A_{2,3}$.*
3. *If the cost function c satisfies the triangle inequality, then for all $p \ge 1$ the following holds*

$$\mathsf{T}^{1/p}_{\mathsf{c}^p}(A_{1,3}) \le \mathsf{T}^{1/p}_{\mathsf{c}^p}(A_{1,2}) + \mathsf{T}^{1/p}_{\mathsf{c}^p}(A_{2,3}).$$

The first item in Lemma 17 and Proposition 5 together show that the composition of two online coupling is also an online coupling.

3.2 Transport Through Intermediate Distributions

In this section, we describe a method of transporting μ to ν (perhaps in an online and iterative way) through optimal transports between intermediate distributions in one dimension. We start with some definitions. We start by defining the notion of average for distributions and stating a general way of transporting through averages.

▶ **Definition 18** (Average Distribution). *Suppose M is a distribution over distributions. We define* the average *of M, denoted as $\mathbb{E}[M] = \mathbb{E}_{\mu'\sim M}[\mu'] = \mu$, to be the distribution μ of the random variable x that is sampled by first sampling $\mu' \sim M$ and then $x \sim \mu'$. Namely, μ is the distribution that $\mu(\mathcal{S}) = \mathbb{E}_{\mu'\sim M}\, \mu'(\mathcal{S})$ for all the events $\mathcal{S}$ defined over $\cup_{\mu'\in\mathrm{Supp}(M)} \mathrm{Supp}(\mu')$.*

▶ **Proposition 19** (Transport to Averages). *Suppose M is a distribution over distributions with average μ.*

1. *Suppose π is the following joint distribution. We first sample $\mu' \sim M$, then couple μ' with ν as $\pi_{\mu'}$, and then output a sample $(x, y) \sim \pi_{\mu'}$. Then, π is a coupling between μ, ν.*
2. $\mathbb{E}_{\mu' \sim M} \mathsf{T}_{\mathsf{c}}(\mu', \nu) \geq \mathsf{T}_{\mathsf{c}}(\mu', \nu)$.

Proof. Part 1 holds because the marginals of x and y have the marginals of μ, ν. Part 2 follows from Part 1 and picking $\pi_{\mu'}$ to be the optimal transport between μ', ν. ◀

The following definition states a way of finding a transport from μ to ν by working with alternative (intermediate) distributions that approximate μ, ν.

▶ **Definition 20** (Transport Through Intermediate Distributions). *Let μ, ν be distributions, c be a cost function, and J be a distribution over pairs of distributions. We say that algorithm A couples μ, ν* through (the intermediate distribution) *J, if the following conditions hold.*

1. *J produces marginals with averages μ, ν. I.e., $\mu = \mathbb{E}_{(\mu', \nu') \sim J} \mu'$ and $\nu = \mathbb{E}_{(\mu', \nu') \sim J} \nu'$.*
2. *Algorithm A first samples $(\mu', \nu') \sim J$, then finds some* optimal *transport π between μ', ν' according to c, and finally outputs $(x, y) \sim \pi$.*

▶ **Definition 21** (Conditioning and Composing Transports with Distributions). *Suppose μ', μ, ν are distributions and π is a transport from μ to ν. If $\mathrm{Supp}(\mu') \subseteq \mathrm{Supp}(\mu)$, then consider the following sampling process.*

1. *Sample $x \sim \mu'$.*
2. *Sample y from the ν-coordinate of π, conditioned on its μ-coordinate being x.*

Then, the notation $\pi|\mu'$ denotes the joint distribution of (x, y) and $\pi \sharp \mu'$ denotes the distribution of y. Additionally, if M is a distribution over distributions, then $N = \pi \sharp M$ denotes the distribution over distributions sampled by outputting $\nu' = \pi \sharp \mu'$ for $\mu' \sim M$.

Notation. Let $U_{k,\mu}$ be the distribution over distributions obtained by first sampling $\mathcal{X} \sim \mu^k$, and then outputting $\mu' = U_{\mathcal{X}}$. A simple observation is that $\mathbb{E}\, U_{k,\mu} = \mu$ for all k.

▶ **Proposition 22.** *If M is a distribution over distributions with average distribution μ, and if π is any transport from μ to ν, then the following holds.*

1. *$N = \pi \sharp M$ is a distribution over distributions with average ν.*
2. *For cost c, $\mathsf{T}_{\mathsf{c}}(\pi) = \mathbb{E}_{\mu' \sim M} \mathsf{T}_{\mathsf{c}}(\pi|\mu')$ in which $\pi|\mu'$ is defined in Definition 21.*
3. *$U_{k,\nu} = \pi \sharp U_{k,\mu}$, and if μ is samplable in time t_μ and coupling π is computable in time t_π, then one can sample the set $\mathcal{Y}, |\mathcal{Y}| = k$ that describes $U_{\mathcal{Y}} \sim U_{k,\nu}$ in time $k(t_\mu + t_\pi)$.*

Proof. For Part 1, observe that if we sample $x \sim \mu'$ for $\mu' \sim M$, by definition we get $x \sim \mu$, which means $y \sim \pi \sharp M$ will be sampled as $y \sim \nu$. For Part 2, $\mathbb{E}_{\mu' \sim M} \mathsf{T}_{\mathsf{c}}(\pi|\mu')$ also computes the cost of the same coupling π by breaking it into marginal costs based on how $x \sim \mu$ is sampled. For Part 3, let $(x, y) \sim \pi$. We first sample $(x_1, \ldots, x_k) = \mathcal{X} \sim \mu^k$ and then let $\mathcal{Y} = (y_1, \ldots, y_k)$ for $y_i \sim y|x = x_i$. It holds that x_is are independently sampled according to μ, and because π transports μ to ν, y_i's are also independently sampled according to ν. ◀

▶ **Lemma 23** (Multi-Round Algorithmic Coupling Through Intermediate Distributions). *Suppose cost function c satisfies the triangle inequality, and c^p is linear over $\mathsf{c}_1, \ldots, \mathsf{c}_n$ for $p \geq 1$. Let π, with marginals $\pi_1, \ldots, \pi_n$ be a transport from μ with marginals $\mu_1, \ldots, \mu_n$ to ν with marginals $\nu_1, \ldots, \nu_n$. For round $i \in [n]$ and previously sampled $z_{[i-1]} = (x_{[i-1]}, y_{[i-1]}) \in \mathrm{Supp}(\pi_{[i-1]})$, suppose $J(z_{[i-1]})$ is a distribution over pairs of distributions defined based on $z_{[i-1]}$, and $\sigma_{z_{[i-1]}}$ is an* optimal *transport from $\mu_i|z_{[i-1]}$ to $\nu_i|z_{[i-1]}$ under c_i. Suppose π can also be*

obtained using the following algorithm A in n rounds. In round $i \in [n]$ and for previously sampled $z_{[i-1]} = (x_{[i-1]}, y_{[i-1]}) \in \mathrm{Supp}(\pi_{[i-1]})$, A couples $\mu_i|z_{[i-1]}$ and $\nu_i|z_{[i-1]}$ through the intermediate distribution $J(z_{[i-1]})$ as defined in Definition 20 using the cost c_i. Then,

$$\mathsf{T}^{1/p}_{\mathsf{c}^p}(\pi) \leq \left(\mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathop{\mathbb{E}}_{(\mu'_i,\nu'_i)\sim J(z_{[i-1]})} \mathsf{T}_{\mathsf{c}_i}\left(\mu'_i, \sigma^{-1}_{z_{[i-1]}} \sharp \nu'_i\right) \right)^{1/p} + \Lambda^{1/p}_{\mathsf{c}^p}(\pi),$$

where σ^{-1} refers to the inverse coupling that changes the order of its marginals.

Proof of Lemma 23. The proof uses the triangle inequality for Wasserstein p-costs for the multi-round setting (Proposition 16).

For each $i \in [n]$ and $z_{[i-1]} \in \mathrm{Supp}(\pi_{[i-1]})$, consider the following sampling process $I(z_{[i-1]})$ that extends $J(z_{[i-1]})$ by outputting one more coordinate as well.

1. Sample $(\mu', \nu') \sim J(z_{[i-1]})$.
2. Let $\mu'' = \sigma^{-1}_{z_{[i-1]}} \sharp \nu'$.
3. Obtain $(\mu', \mu'', \nu') \sim I(z_{[i-1]})$.

It holds that $\mathsf{T}^{1/p}_{\mathsf{c}^p}(A) = \left(\mathbb{E}_{z\sim\pi} \sum_{i\in[n]} \mathbb{E}_{(\mu',\mu'',\nu')\sim I(z_{[i-1]})} \mathsf{T}_{\mathsf{c}_i}(\mu', \nu')\right)^{1/p}$, which is the left side of the inequality of Proposition 16, and the right side is:

$$\left(\mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathop{\mathbb{E}}_{(\mu',\mu'',\nu')\sim I(z_{[i-1]})} \mathsf{T}_{\mathsf{c}_i}(\mu', \mu'') \right)^{1/p} + \left(\mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathop{\mathbb{E}}_{(\mu',\mu'',\nu')\sim I(z_{[i-1]})} \mathsf{T}_{\mathsf{c}_i}(\mu'', \nu') \right)^{1/p}$$

The first term is exactly the first term on the right hand side of the inequality of the lemma. Therefore, all we have to do is to prove that

$$\mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathop{\mathbb{E}}_{(\mu',\mu'',\nu')\sim I(z_{[i-1]})} \mathsf{T}_{\mathsf{c}_i}(\mu'', \nu') \leq \mathop{\mathbb{E}}_{z\sim\pi} \sum_{i\in[n]} \mathsf{T}_{\mathsf{c}_i}(\sigma_{z_{[i-1]}}).$$

In fact, we prove this statement for *every* choice of z and i, so ignoring z, i we prove the claim:

$$\mathop{\mathbb{E}}_{(\mu'',\nu')\sim I} \mathsf{T}_{\mathsf{c}_i}(\mu'', \nu') \leq \mathop{\mathbb{E}}_{(\mu'',\nu')\sim I} \mathsf{T}_{\mathsf{c}_i}((\sigma^{-1}|\nu')^{-1}) = \mathsf{T}_{\mathsf{c}_i}(\sigma),$$

where the middle term is added for the proof.

We now prove both the inequality and the equality above through the steps below.

- *Equality:* Since the average of $\nu' \sim J$ is ν_i and σ^{-1} is a transport from ν_i to μ_i, if we define $\mathsf{c}'_i(y_i, x_i) = \mathsf{c}_i(x_i, y_i)$, then by Part 2 of Proposition 22 we have

$$\mathop{\mathbb{E}}_{(\mu'',\nu')\sim I} \mathsf{T}_{\mathsf{c}_i}((\sigma^{-1}|\nu')^{-1}) = \mathop{\mathbb{E}}_{(\mu'',\nu')\sim I} \mathsf{T}_{\mathsf{c}'_i}(\sigma^{-1}|\nu') = \mathsf{T}_{\mathsf{c}'_i}(\sigma^{-1}) = \mathsf{T}_{\mathsf{c}_i}(\sigma).$$

- *Inequality:* Again, using $\mathsf{c}'_i(y_i, x_i) = \mathsf{c}_i(x_i, y_i)$, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{(\mu'',\nu')\sim I} \mathsf{T}_{\mathsf{c}_i}(\mu'', \nu') &= \mathop{\mathbb{E}}_{(\mu'',\nu')\sim I} \mathsf{T}_{\mathsf{c}'_i}(\nu', \mu'') \\ &\leq \mathop{\mathbb{E}}_{(\mu'',\nu')\sim I} \mathsf{T}_{\mathsf{c}'_i}(\sigma^{-1}|\nu') = \mathop{\mathbb{E}}_{(\mu'',\nu')\sim I} \mathsf{T}_{\mathsf{c}_i}((\sigma^{-1}|\nu')^{-1}), \end{aligned}$$

where the inequality is due to the fact that $\mathsf{T}_{\mathsf{c}'_i}(\nu', \mu'')$ is the optimal cost. ◀

3.3 Borrowed Tools

The following can be obtained from the proofs in [43, 22] (see the full version). For $p = 2$, it gives the celebrated Talagrand's transportation inequality for Gaussian under ℓ_2.

▶ **Theorem 24** (Talagrand's Inequality for the Gaussian Measure). *If* $\mathsf{c}(x, y) = \ell_p^p(x, y)$, $p \in [1, 2]$, Φ_n *is the standard Gaussian and* ν *is an arbitrary distribution both in* $\mathbb{R}^n$, *then*

$$\mathsf{T}_{\mathsf{c}}^{\mathrm{OnT}}(\Phi_n, \nu) = \Delta_{\mathsf{c}}(\Phi_n, \nu) \le n^{1-p/2} \cdot (2\mathsf{KL}(\nu, \Phi_n))^{p/2}.$$

▶ **Definition 25** (Transports to Empirical). *For distributions* μ *and symmetric cost* c, *we let* $\mathsf{T}_{\mathsf{c},k}^{\mathrm{Em}}(\mu) = \mathbb{E}_{\mathcal{X}\sim\mu^k} \mathsf{T}_{\mathsf{c}}(U_{\mathcal{X}}, \mu)$ *denote the cost of transporting* μ *to an empirical set of size* k, *where* $U_{\mathcal{X}}$ *is the uniform distribution over the multi-set* $\mathcal{X}$.

The following lemma follows from [19] and known moments of the Gaussian distribution.

▶ **Lemma 26** (Original-to-Empirical Transport for the Normal Distribution). *Let* $p \ge 1$, c *be* ℓ_p^p, *and* $\mu = \mathcal{N}(0, 1)$ *is the normal distribution. Then, for a constant* C_p *depending on* p,

$$\mathsf{T}_{\mathsf{c},k}^{\mathrm{Em}}(\mu) \le C_p \cdot 2^{1+3p/2} \cdot \Gamma(p+1)^{\frac{p}{2p+1}} \cdot k^{-1/2}.$$

4 Algorithmic Transport for Products

In this section, we put together the tools from previous sections to derive algorithmic results about online transport for the setting that one of the source or target distributions is product. We then derive a corollary for the Gaussian measure. We first define sequential samplers.

▶ **Definition 27** (Sequential Sampler). *For a distribution* ν *in dimension* n *with marginals* $\nu_1, \ldots, \nu_n$, *we call* $\hat{\nu}$ *its* sequential sampler *for* ν, *if for* all $y_{[i-1]} \sim \nu_{[i-1]}$ *calling* $\hat{\nu}(y_{[i-1]})$ *returns an independent answer* $\hat{\nu}(y_{[i-1]}) \sim \nu_i | y_{[i-1]}$. *For queries* $y_{[i-1]} \notin \mathrm{Supp}(\nu_{[i-1]})$, *calling* $\hat{\nu}(y_{[i-1]})$ *returns* $\perp$. *We also assign a (sequential sampling) cost* $\mathsf{sc}_\nu(y_{[i-1]})$ *to query* $y_{[i-1]}$, *and call* $\mathsf{sc}_\nu = \mathbb{E}_y \sum_{i\in[n-1]} \mathsf{sc}_\nu(y_{[i-1]})$ *the* average (sequential sampling) cost *of* $\hat{\nu}$. *For an oracle-algorithm* A *calling (a potentially randomized) set* $\mathcal{Q}$ *of queries to* $\hat{\nu}$, *we define its* average total cost *of calling* $\hat{\nu}$ *as* $\mathsf{sc}_\nu^A = \mathbb{E}_{\mathcal{Q}} \sum_{a\in\mathcal{Q}} \mathsf{sc}_\nu(a)$.[1]

One natural way of using sc is to model sampling time, but it can model other costs as well. The average cost sc_ν of $\hat{\nu}$ is indeed the average total cost of the following simple algorithm A that uses sc_ν sequentially to obtain a full sample: Let $y_{[0]}$ be the empty string, and for $i \in [n]$, A let $y_i = \hat{\nu}(y_{[i-1]})$. Also, when μ is a product distribution, then $\hat{\mu}$ is nothing other than a direct way of sampling from independent distributions ν_i for all $i \in [n]$.

Before stating our main result, recall the notation for transport cost to empirical sets from Definition 25.

▶ **Theorem 28** (Main Result). *Suppose* $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ *and* ν *are distributions over* $\mathbb{R}^n$, *with sequential samplers* $\hat{\mu}, \hat{\nu}$ *and corresponding oracle cost functions* $\mathsf{sc}_\mu, \mathsf{sc}_\nu$. *Suppose the transportation cost function* c *is a metric (i.e., symmetric and satisfies the triangle inequality) and* c^p *is linear over symmetric costs* $\mathsf{c}_1, \ldots, \mathsf{c}_n$.[2] *Then, there is an algorithm* A_k, *parameterized by* k, *that uses oracle access to samplers* $\hat{\mu}, \hat{\nu}$ *and achieves the following:*

[1] Since $\mathsf{sc}_\nu(y_{[i-1]})$ naturally measures the (e.g., computational) cost of sampling a coordinate conditioned on previously sampled coordinates, for natural settings and independent ν_1, ν_2, the value of $\mathsf{sc}_\nu(y_{[1]})$ will be independent of $y_{[1]}$.

[2] An example is $\mathsf{c} = \ell_p$.

1. $A_k^{\hat{\mu},\hat{\nu}}$ *transports* μ *to* ν *through an online coupling in time* $\mathrm{poly}(nk)$ *with p-cost*[3]
$$\mathsf{T}_{\mathsf{c}^p}^{1/p}(A_k^{\hat{\mu},\hat{\nu}}) \le \delta + \Delta$$
in which $\delta = 2\left(\sum_{i\in[n]} \mathsf{T}^{\mathrm{Em}}_{\mathsf{c}_i,k}(\mu_i)\right)^{1/p}$ *and* $\Delta = \Delta_{\mathsf{c}^p}^{1/p}(\mu,\nu)$ *as in Definition 9.*[4]
2. *The average total cost of A calling* $\hat{\mu},\hat{\nu}$ *is as follows.* $\mathsf{sc}_\nu^A \le k\cdot\mathsf{sc}_\nu$ *and* $\mathsf{sc}_\mu^A \le k\cdot\mathsf{sc}_\mu$.[5]
3. *There is an algorithm B that achieves the same as A does, but it transports* ν *back to* μ.

Proof. At a high level, we use an empirical variant of the greedy algorithm (which is related to the KR transport) to design the algorithm. The algorithm itself is quite simple; the bulk of the work goes into its analysis, which is quite delicate and uses many tools from Section 3.

The Transportation Algorithm A. The algorithm A works in n rounds. In round $i\in[n]$, given $x_i\sim\mu_i$ find $y_i\sim\nu_i|y_{[i-1]}$ as described below.

1. For $j\in[k]$, let $y_i^{(j)}\sim\hat{\nu}(y_{[i-1]})$ be independent samples forming the multi-set $\mathcal{Y}$ of size k.
2. Pick $t\leftarrow[k]$ at random. For all $j\in[k], j\neq t$, let $x_i^{(j)}\sim\mu$ be $k-1$ independent samples. Additionally, let $x_i^{(t)}=x_i$, and $\mathcal{X}$ be the multi-set $\left\{x_i^{(j)} \mid j\in[k]\right\}$ of size k.
3. Find the optimal transport between the two distributions $U_{\mathcal{X}}, U_{\mathcal{Y}}$ under the cost c_i (e.g., using the Hungarian method[6]) that is in the form of a matching between $\mathcal{X}$ and $\mathcal{Y}$.[7]
4. Output $y_i\in\mathcal{Y}$ that is matched with $x_i^{(t)}=x_i\in\mathcal{X}$.

We now analyze the algorithm A above.

Transportation. A's running time is clearly $\mathrm{poly}(kn)$. We now prove that A's algorithm produces an *online* coupling between μ,ν, by showing that in round i, it couples μ_i and $\nu_i|y_{[i-1]}$. It is simple to check that all the elements of $\mathcal{X}$ are distributed as μ_i and all the elements of $\mathcal{Y}$ are distributed as in $\nu_i|y_{[i-1]}$. At first, it might not be clear why y_i is distributed as $\nu_i|y_{[i-1]}$, because the matching algorithm might change its distribution by picking it adversarially. However, since the algorithm hides the index of x_i and statistically hides it among $\mathcal{X}$, the final "matched pair" (x_i,y_i) is a random edge of the optimal matching/transport. Therefore, y_i is also distributed accurately, and hence A is producing an online coupling.

More formally, we can choose $t\in[k]$ at random *after* the matching between $\mathcal{X},\mathcal{Y}$ is chosen. Moreover, the marginal distribution of $y_i^{(j)}$ is $\hat{\nu}(y_{[i-1]})$. Therefore, for every (even fixed) matching between $\mathcal{X},\mathcal{Y}$, picking t at random will lead to picking $y_i=y_i^{(j)}$ where j is the index of the sample in $\mathcal{Y}$ that is matched with the index t in $\mathcal{Y}$. Therefore, $y_i\sim\hat{\nu}(y_{[i-1]})$.

The Cost. To analyze the transportation cost we apply Lemma 23 from Section 3, which is stated in a more general form to better demonstrating the key ideas.

To apply Lemma 23, let $J(y_{[i-1]})$ return pair of distributions $(\mu_i'=U_{\mathcal{X}},\nu_i'=U_{\mathcal{Y}})$ that are constructed using independent sample multi-sets $\mathcal{X},\mathcal{Y}$ of size k, in order, from $\mu_i,\nu_i|y_{[i-1]}$. Finally, because the algorithm A finds an optimal transport between μ_i',ν_i', we will have the premises of Lemma 23 and conclude that

$$\mathsf{T}_{\mathsf{c}^p}^{1/p}(A_k^{\hat{\mu},\hat{\nu}}) \le \left(\mathop{\mathbb{E}}_{(x,y)\sim\pi}\sum_{i\in[n]}\mathop{\mathbb{E}}_{(U_{\mathcal{X}},U_{\mathcal{Y}})\sim J(y_{[i-1]})}\mathsf{T}_{\mathsf{c}_i}\left(U_{\mathcal{X}},\sigma_{z_{[i-1]}}^{-1}\sharp U_{\mathcal{Y}}\right)\right)^{1/p} + \Lambda_{\mathsf{c}^p}^{1/p}(\pi_A),$$

[3] See Definition 1.
[4] By Theorem 10, Δ_{c^p} is also equal to $\mathsf{T}^{\mathrm{OnT}}_{\mathsf{c}^p}(\mu,\nu)=\mathsf{T}^{\mathrm{OnC}}_{\mathsf{c}^p}(\mu,\nu)=\mathsf{T}^{\mathrm{G}}_{\mathsf{c}^p}(\mu,\nu)$.
[5] Note that because μ is a product distribution, if sc_μ models the computational cost of sampling from μ, then we would have $\mathsf{sc}_\mu=\sum_{i\in[n]}\mathsf{sc}_{\mu_i}$, where sc_{μ_i} models the computational cost of sampling from μ_i.
[6] This method can be implemented faster when the cost function is convex, in which case simply sorting $\mathcal{X},\mathcal{Y}$ gives us the optimal matching, as a monotone mapping.
[7] This can be proved, e.g., using the Birkhoff–von Neumann decomposition of doubly stochastic matrices.

in which $\sigma_{z_{[i-1]}}$ is an (optimal) transport from μ_i to $\nu_i|y_{[i-1]}$. (See Definition 21 for the $\sharp$ notation.) We now further simplify the summation above.

- Because $U_{\mathcal{X}}, U_{\mathcal{Y}}$ are empirical distributions from $\mu_i, \nu_i|y_{[i-1]}$, if we let $U_{\mathcal{X}} = \mu_i', U_{\mathcal{Y}} = \nu_i'$ in Proposition 22, by Part 3 we get $U_{\mathcal{X}'} = \sigma^{-1}_{z_{[i-1]}} \sharp U_{\mathcal{Y}}$ (see Definition 21 for the notation) in which $U_{\mathcal{X}'}$ is also an empirical distribution sampled from μ_i independently of $U_{\mathcal{X}}$. So, the first term of the right hand side in the inequality above simplifies to:

$$\left(\mathop{\mathbb{E}}_{(x,y)\sim\pi} \sum_{i\in[n]} \mathop{\mathbb{E}}_{\mathcal{X},\mathcal{X}'\sim\mu_i^k} \mathsf{T}_{\mathsf{c}_i}(U_{\mathcal{X}}, U_{\mathcal{X}'}) \right)^{1/p}$$

- Now, in the first term, both coordinates of $(x, y) \sim \pi$ are irrelevant to the summation.
- Since A is producing an *online* coupling the second term simplifies into $\Delta^{1/p}_{\mathsf{c}^p}(\mu, \nu) = \Lambda^{1/p}_{\mathsf{c}^p}(\pi_A)$, due to Part 3 of Proposition 12 and that μ is a product.
- Finally, by the triangle inequality of Proposition 16, the first term will become at most

$$2\left(\sum_{i\in[n]} \mathsf{T}^{\mathrm{Em}}_{k,\mathsf{c}_i}(\mu_i) \right)^{1/p} = 2\delta.$$

To apply Proposition 16, we let J_i to be the distribution over distributions that outputs the following triple of distributions (ν_1, ν_2, ν_3), where

$$\nu_1 = U_{\mathcal{X}}, \mathcal{X} \sim \mu_i^k, \nu_2 = \mu_i, \nu_3 = U_{\mathcal{X}'}, \mathcal{X}' \sim \mu_i^k.$$

Oracle Costs. In each round, A asks $k-1$ samples from μ_i and k samples from $\nu_i|y_{[i-1]}$. Furthermore, the previous samples $y_{[i-1]}$ are sampled according to $\nu_{[i-1]}$ itself, so the average total cost will be as claimed.

Inverse Transport. The reverse mapping uses the same algorithm for one dimension transport, but it maps $\nu_i|y_{[i-1]}$ to μ_i, and inspection shows its transportation and (expected) total oracle costs will be the same as that of A. ◀

4.1 Extending Transport to Conditional Distributions

In this subsection, we study how to use the main result of Theorem 28 and obtain transports from the same μ to a more rich set of distributions that can be obtained from ν by conditioning ν on an event $\mathcal{S}$ of not-so-small probability. Doing so would be extremely useful, when later on, we focus on transporting Gaussian distributions to the same distributions conditioned on an event $\mathcal{S}$. To prove this extension, we prove a general result about using sequential samplers for ν to obtain sequential samplers for $\nu|\mathcal{S}$.

▶ **Theorem 29** (Sequential Samplers for Event-conditioned Distributions)**.** *Suppose ν is an n-dimensional distribution that has a sequential sampler $\hat{\nu}$ with average cost sc_ν. Suppose $\mathcal{S}$ is an event of measure $\nu(\mathcal{S}) \geq \varepsilon$, and $\omega = \nu|\mathcal{S}$ is ν conditioned on $\mathcal{S}$. Then, there is an algorithm O that uses oracle $\hat{\nu}$ and a membership oracle $\mathcal{S}$ and achieves the following.*

1. *For all $y_{[i-1]} \sim \omega_{[i-1]}$, $O^{\mathcal{S},\hat{\nu}}(y_{[i-1]}) \sim \hat{\omega}(y_{[i-1]})$.*
2. *If we define $\mathsf{sc}_\omega(y_{[i-1]})$ be the average total cost of $O^{\mathcal{S},\hat{\nu}}(y_{[i-1]})$ querying $\hat{\nu}$, and if we define $\mathsf{sc}_\nu(i) = \mathbb{E}_{y\sim\nu}\, \mathsf{sc}_\nu(y_{[i-1]})$, then*

$$\mathsf{sc}_\omega \leq \frac{1}{\varepsilon} \sum_{i\in[n]} i \cdot \mathsf{sc}_\nu(i) \leq n \cdot \frac{\mathsf{sc}_\nu}{\varepsilon}.$$

3. *When iteratively sampling* $(y_1, \ldots, y_n) \sim \omega$*, the expected number of calls to* $\mathcal{S}$ *in round* i *is at most* $1/\varepsilon$*, making the total expected number of calls to* $\mathcal{S}$ *to be at most* n/ε*.*
4. *The running time of the iterative sampling of* $(y_1, \ldots, y_n) \sim \omega$*, relative to the provided oracles* $\hat{\nu}, \mathcal{S}$ *is at most* $O(n^2/\varepsilon)$*.*

In other words, one can use $O^{\mathcal{S},\hat{\nu}}$ to emulate a sequential sampler for $\omega = \nu|\mathcal{S}$ in such a way that the average cost of obtaining a full sequence $y \sim \nu|\mathcal{S}$ using n nested calls to the provided sequential sampler only goes up (at most) by a multiplicative factor $n/\Pr[\mathcal{S}]$.

The main idea in the proof is to use rejection sampling with a subtle analysis. Namely, $O^{\mathcal{S},\hat{\nu}}$ simply keeps using $\hat{\nu}$ to obtain full sequences multiple times until the sample sequence falls within the event $\mathcal{S}$. The full proof follows.

Proof of Theorem 29. For $v = (v_1, \ldots, v_n)$, let $v_{\geq i} = (v_i, \ldots, v_n)$ and $v = (v_{[i-1]}, v_{\geq i})$.

Our algorithm $O^{\mathcal{S},\hat{\nu}}(y_{[i-1]})$ samples from $\hat{\omega}(y_{[i-1]})$ as follows.

1. Sample from $\nu|y_{[i-1]}$ as follows: for $j = i, \ldots, n$ sample fresh values $y_j \sim \hat{\nu}(y_{[j-1]})$.
2. If $y = (y_{[i-1]}, y_{\geq i}) \in \mathcal{S}$, then output y_i; otherwise, go back to the previous step.

We refer to each execution of the two steps above (that has exactly one call to $\mathcal{S}$) a *trial*.

Part 1 follows from the fact that the above sampling process is a simple rejection sampling. To prove Part 2, let $H(y_{[i-1]})$ be a random variable that counts the number of trials, and let its expectation be

$$h(y_{[i-1]}) = \mathbb{E}[H(y_{[i-1]})] = \frac{1}{\Pr_{y\sim\nu|y_{[i-1]}}[y \in \mathcal{S}]}.$$

Also let $\overline{\mathsf{sc}}_\nu(y_{[i-1]}) = \mathbb{E}_{y\sim\nu|y_{[i-1]}} \sum_{j\geq i} \mathsf{sc}_\nu(y_{j-1})$. It can be observed that $\mathbb{E}_{y\sim\nu} \overline{\mathsf{sc}}_\nu(y_{[i-1]}) = \sum_{j\geq i} \mathsf{sc}_\nu(i)$. Using these notations, the oracle sampling cost of $\hat{\omega}(\cdot)$ at $y_{[i-1]}$ will be

$$\mathsf{sc}_\omega(y_{[i-1]}) = h(y_{[i-1]}) \cdot \overline{\mathsf{sc}}_\nu(y_{[i-1]}).$$

Therefore, the average cost of $\hat{\omega}$ will be

$$\mathsf{sc}_\omega = \mathop{\mathbb{E}}_{y\sim\omega} \sum_{i\in[n]} h(y_{[i-1]}) \cdot \mathsf{sc}_\omega(y_{[i-1]}) = \sum_{i\in[n]} \mathop{\mathbb{E}}_{y\sim\omega} h(y_{[i-1]}) \cdot \mathsf{sc}_\omega(y_{[i-1]}).$$

A subtle point is that, in the above sums the first half $y_{[i-1]}$ is sampled conditioned on $\mathcal{S}$, while the second half is done without such conditioning. We claim that for each i, we have

$$\mathop{\mathbb{E}}_{y\sim\omega} h(y_{[i-1]}) \cdot \mathsf{sc}_\omega(y_{[i-1]}) \leq \frac{1}{\varepsilon} \cdot \mathop{\mathbb{E}}_{y\sim\nu} \mathsf{sc}_\omega(y_{[i-1]}). \tag{1}$$

Note that if Eq. (1) holds, then we conclude Part 2, because we get:

$$\mathsf{sc}_\omega \leq \sum_{i\in[n]} \frac{1}{\varepsilon} \mathop{\mathbb{E}}_{y\sim\nu} \mathsf{sc}_\omega(y_{[i-1]}) = \frac{1}{\varepsilon} \cdot \sum_{i\in[n]} \sum_{j\geq i} \mathsf{sc}_\nu(i) = \frac{1}{\varepsilon} \sum_{i\in[n]} i \cdot \mathsf{sc}_\nu(i).$$

The following lemma proves Eq. (1) using $\mathcal{U} = \mathrm{Supp}(\nu_{[i-1]}), \mathcal{V} = \mathrm{Supp}(\nu_{\geq i})$, $\sigma = \nu$, $f(y) = \overline{\mathsf{sc}}_\nu(y_{[i-1]})$ and $\mathcal{S}$ as before.

▶ **Lemma 30** (Expected Cost of Two-Step Sequential Sampling)**.** *Suppose* σ *is distributed over* $\mathcal{U} \times \mathcal{V}$ *with margins* $\sigma_\mathcal{U}, \sigma_\mathcal{V}$*, and* $\mathcal{S} \subseteq \mathcal{U} \times \mathcal{V}$ *has probability* $\sigma(\mathcal{S}) = \varepsilon$*. Also, suppose* f *is a random variable defined over* σ *with average* $\bar{f}$*. Consider the following process: (1) Sample* $u \sim \sigma_\mathcal{U}|\mathcal{S}$*, which is the marginal distribution of* $\mathcal{U}$ *in* $\sigma|\mathcal{S}$ *and let* $\varepsilon_u = \Pr_{v\sim\sigma^u_\mathcal{V}}[(u, v) \in \mathcal{S}]$*, in which* $\sigma^u_\mathcal{V}$ *is the marginal distribution over* $\mathcal{V}$ *in* σ *conditioned on* $\sigma_\mathcal{U} = u$*. (2) Sample* $v \sim \sigma^u_\mathcal{V}|\mathcal{S}$*. Then,*

$$\mathop{\mathbb{E}}_{(u,v)} \frac{f(u, v)}{\varepsilon_u} \leq \frac{\bar{f}}{\varepsilon}.$$

Proof. We write the proof for the discrete setting. A similar proof holds in general. For each $u \sim \sigma_{\mathcal{U}}$, define $p_u = \Pr[\sigma_{\mathcal{U}} = u]$ and $f_u = \mathbb{E}_{v\sim\sigma_{\mathcal{V}}|u} f(u,v)$. We have $\varepsilon = \sum_{u\in\mathcal{U}} p_u\varepsilon_u$, and $q_u = \frac{p_u\cdot\varepsilon_u}{\varepsilon}$ is the probability we sample u in the sampling process of the lemma statement. Then, if we let $\mathcal{U}_{\mathcal{S}} = \{u \mid \varepsilon_u > 0\} = \mathrm{Supp}(\sigma_{\mathcal{U}}|\mathcal{S})$, we have

$$\mathop{\mathbb{E}}_{u} \frac{1}{\varepsilon_u} \mathop{\mathbb{E}}_{v|u} f(u,v) = \sum_{u\in\mathcal{U}_{\mathcal{S}}} \frac{q_u}{\varepsilon_u} f_u = \sum_{u\in\mathcal{U}} \frac{p_u}{\varepsilon} f_u \le \sum_{u\in\mathcal{U}_{\mathcal{S}}} \frac{p_u}{\varepsilon} f_u = \frac{\bar{f}}{\varepsilon}. \qquad \blacktriangleleft$$

To prove Part 3, using Lemma 30 and $f(u,v) = 1$, we conclude that the expected number of times we call the $\mathcal{S}$ oracle at each node $y_{[i-1]}$ is at most $1/\varepsilon$.

To prove Part 4 we can simply use a fake oracle sampling cost of $\hat{\mathsf{sc}}'_\nu(\cdot) = 1$. Then the claim about the running time follows from Part 2. $\blacktriangleleft$

Deriving corollaries. Using Theorem 29, we can derive more transportation results from Theorem 28 by conditioning ν on an arbitrary event $\mathcal{S}$ for which we have a membership oracle at hand. Note that the parameter Δ will change to a new value, but the key point is that we can control the cost of sequential samples from the new oracle, so long as we could do so for the initial oracle. Another interesting application of Theorem 29 is to transport a product distribution μ to $\mu|\mathcal{S}$ for an arbitrary event $\mathcal{S}$, obtaining the following corollary.[8]

▶ **Corollary 31.** *Suppose the assumptions of Theorem 28 hold. Then, we have the following:*

1. *There is an algorithm M_k such that, for all events $\mathcal{S}$ defined over μ, $M_k^{\mathcal{S},\hat{\mu}}$ transports μ to $\mu|\mathcal{S}$ in expected time* $\mathrm{poly}(nk/\varepsilon)$ *and p-cost $\mathsf{T}_{\mathsf{c}^p}^{1/p}(M_k^{\hat{\mu},\hat{\nu}}) \le \delta + \Delta$, in which δ is as in Theorem 28 and $\Delta = \Delta_{\mathsf{c}^p}^{1/p}(\mu,\mu|\mathcal{S})$.*
2. *There is an algorithm N_k such that, for all events $\mathcal{S}$ defined over ν, $N_k^{\mathcal{S},\hat{\mu},\hat{\nu}}$ transports μ to $\nu|\mathcal{S}$ in expected time* $\mathrm{poly}(nk/\varepsilon)$ *and p-cost $\mathsf{T}_{\mathsf{c}^p}^{1/p}(N_k^{\mathcal{S},\hat{\mu},\hat{\nu}}) \le \delta + \Delta$, in which δ is as in Theorem 28 and $\Delta = \Delta_{\mathsf{c}^p}^{1/p}(\mu,\nu|\mathcal{S})$. Moreover, $\mathsf{sc}_\nu^N \le n\cdot\mathsf{sc}_\nu^A/\varepsilon$, for A of Theorem 28.*

In both cases above, the expected number of calls to $\mathcal{S}$ is at most kn/ε, and the transportation can be reversed with the same upper bounds on the running time and oracle costs.

5 Algorithmic Transport for Gaussian

In this section we focus on cases where at least one of the two distributions involved in the transport is Gaussian. We first use the main result of Theorem 28 and derive an algorithmic variant of Talagrand's result [43] about transporting Gaussian measure to arbitrary distributions with bounded KL divergence from Gaussian. We then derive, as a corollary, a computational concentration result for the Gaussian source measure under the ℓ_2 distance. Finally, we focus on finding (optimal) online transports in cases where both the source and destination are Gaussians, but they could be arbitrary (non-product) Gaussians.

5.1 Algorithmic Variant of Talagrand's Transport for Gaussian

▶ **Theorem 32** (Algorithmic Version of Talagrand's Gaussian Transport Theorem)**.** *Let Φ^n be the standard Gaussian in dimension n and ν be an arbitrary distribution in $\mathbb{R}^n$. There is an algorithm A_k, with integer parameter k, such that whenever $A_k^{\hat{\nu}}$ is provided with a sequential sampler $\hat{\nu}$ for ν, the following properties hold.*

[8] In the next section we apply this idea to the special case of Gaussian distributions.

1. *For all $p \geq 1$ and ν, $A_k^{\hat{\nu}}$ transports Φ^n to ν in time $O(nk \log k)$ with p-cost at most*

$$\mathsf{T}_{\ell_p^p}^{1/p}(A_k^{\hat{\nu}}) \leq \Delta_{\ell_p^p}^{1/p}(\Phi^n, \nu) + \left(O_p(nk^{-1/2})\right)^{1/p}.$$

   For $p = 2$, by the Talagrand inequality of Theorem 24, we have $\Delta_{\ell_2^2}(\Phi^n, \nu) \leq 2\mathsf{KL}(\Phi^n, \nu)$.
2. *The average total oracle cost of $A_k^{\hat{\nu}}$ is at most $k \cdot \mathbb{E}_{y\sim\nu} \sum_{i\in[n]} \mathsf{sc}(\nu_i | y_{[i-1]})$.*
3. *There is an algorithm $B_k^{\hat{\nu}}$ that achieves the same as $A_k^{\hat{\nu}}$, but it transports ν back to Φ^n.*

▶ Remark 33 (Working with ℓ_p instead of ℓ_p^p). One might wonder what happens if we want to measure (and upper bound) transfer costs using ℓ_p rather than ℓ_p^p. However, this can be obtained using Jensen's inequality (or rather the monotonicity of Wasserstein p-costs for a fixed cost c). In particular, for every coupling π, we have $\mathsf{T}_{\ell_p}(\pi) \leq \mathsf{T}_{\ell_p^p}^{1/p}(\pi)$ for all $p \geq 1$. Hence Theorem 32 is stated in the stronger form already.

Proof of Theorem 32. The proof follows directly from Theorem 28 and Lemma 26. Namely, we use Corollary 26 to bound the term δ in Theorem 28 that upper bounds the transportation cost of empirical Gaussian from the Gaussian itself. One small point here is that, we will not need oracle samplers from the Gaussian itself, as we can use well-known sampling methods such as the Box-Muller method that generate such samples efficiently [36].[9] ◀

We now focus on a special case of interest, in which the target distribution ν is $\Phi^n|\mathcal{S}$ for an event $\mathcal{S}$ of probability $\Phi^n(\mathcal{S}) = \varepsilon$, and show that in this case, one can have a single online transportation algorithm that uniformly works for all $\mathcal{S}$ by merely accessing $\mathcal{S}$ through a membership oracle. We first define such uniform transportation algorithms.

▶ **Definition 34** (Oracle Set-Transport). *For distribution μ and transportation cost c, we say that (μ, c) has a* set-transport *of cost at most $\kappa(\cdot)$ for a non-increasing function $\kappa\colon [0,1] \mapsto [0,1]$, if for every event $\mathcal{S} \subseteq \mathrm{Supp}(\mu)$, it holds that $\mathsf{T}_{\mathsf{c}}(\mu, \mu|\mathcal{S}) \leq \kappa(\mu(\mathcal{S}))$. We further say that (μ, c) has an* oracle *set-transport of cost at most $\kappa(\cdot)$ if there is a single algorithm A such that with oracle membership queries for an arbitrary set $\mathcal{S}$ and sampling queries for μ, $A^{\mathcal{S},\mu}$ produces a transport of cost at most $\kappa(\mu(\mathcal{S}))$ from μ to $\mu|\mathcal{S}$.*

▶ **Theorem 35** (Oracle-Set Transport for Gaussian Measure). *Let Φ^n be the standard Gaussian in dimension n. There is an (online) oracle-set transport algorithm A_k for Φ^n such that:*
1. *For all $p \in [1,2]$ and $\mathcal{S}$ of measure $\Phi^n(\mathcal{S}) = \varepsilon$,*

$$\mathsf{T}_{\ell_p^p}^{1/p}(A_k^{\mathcal{S}}) \leq \kappa^{1/p}(\varepsilon) = n^{1/p-1/2}\sqrt{2\ln 1/\varepsilon} + \left(O_p(nk^{-1/2})\right)^{1/p},$$

   which is at most $(1+\gamma) \cdot n^{1/p-1/2}\sqrt{2\ln 1/\varepsilon}$, for sufficiently large $k = \mathrm{poly}(n, 1/\varepsilon, 1/\gamma)$.
2. *In expectation, $A_k^{\mathcal{S}}$ asks at most kn/ε queries to $\mathcal{S}$ and runs in $\mathrm{poly}(nk/\varepsilon)$.*
3. *There is an algorithm B_k that achieves the same, but $B_k^{\mathcal{S}}$ transports $\Phi^n|\mathcal{S}$ back to Φ^n.*

Proof of Theorem 35. To prove Theorem 35 we first use the first item of Corollary 31 where $\mu = \Phi^n$. This way, we already know that the running time of the transportation algorithm and its number of calls to $\mathcal{S}$ are bounded as stated.

Then, we need to bound both terms Δ, δ. To bound δ, we again use Corollary 26 as we did in the proof of Theorem 32. To bound Δ, we again use Corollary 24 and the well-known fact that $\mathsf{KL}(\mu|\mathcal{S}, \mu) \leq \ln 1/\varepsilon$ for $\mathcal{S}$ such that $\mu(\mathcal{S}) \geq \varepsilon$ (applied to $\mu = \Phi^n$). ◀

[9] In particular, given two independent and uniform $u_1, u_2 \sim [0,1]$, the sampling works as follows: $v_1 = \sqrt{-2\ln u_1}\cos(2\pi u_2), v_2 = \sqrt{-2\ln u_1}\sin(2\pi u_2)$ are independent samples $v_1, v_2 \sim \mathcal{N}(0,1)$.

Due to our transports being "reversible", one can obtain a variant of the result above that transports conditional distributions to conditional distributions through composition.

5.2 Dimension-Independent Computational Concentration for Gaussian

It is well-known that transportation inequalities can be used to derive concentration of measure results [22]. Recently, a *computational* variant of this phenomenon has been explored [31, 15], which bears similarities to how we make transportation algorithmic. In a computational concentration result, we need an algorithm that maps "most" of the sampled points from the space to any "sufficiently large" event $\mathcal{S}$, algorithmically. The "cost" of the concentration is (a worst-case) allowed distance d that the algorithm is allowed to move the points, and its error is the fraction of the sampled points that it fails to map to $\mathcal{S}$ withing the allowed distance d. The work of [15] obtained such results optimally for some settings (e.g., Gaussian under ℓ_1 distance), however they left open obtaining an optimal (dimension-free) computational concentration result for the Gaussian space under the ℓ_2 distance.

Using Theorem 35, we can resolve the question left open in [15] and derive such optimal computational concentration for the Gaussian space under ℓ_2 as a simple corollary to our algorithmic transport result. Theorem 36 below follows from Theorem 35 and the Markov inequality. Using $p = 2$ below implies the desired dimension-independent result.

▶ **Corollary 36** (Computational Concentration for Gaussian). *For all $\varepsilon, \delta, \lambda, p \in [1, 2]$, given oracle access to $\mathcal{S} \subseteq \mathbb{R}^n$, $A_k^{\mathcal{S}}(x)$ of Theorem 35 runs in $\mathrm{poly}(\frac{n}{\varepsilon\lambda})$-time and with probability $1 - \delta$ over $x \sim \Phi^n$, it finds a point $y \in \mathcal{S}$ of distance*

$$\ell_p(x, y) \leq \frac{(1+\lambda) \cdot n^{1/p-1/2} \sqrt{2 \ln 1/\varepsilon}}{\delta}.$$

References

1 Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

2 Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017. URL: `http://proceedings.mlr.press/v70/arjovsky17a/arjovsky17a.pdf`.

3 Julio Backhoff, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

4 Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. *Advances in Neural Information Processing Systems*, 32, 2019.

5 Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S Meel, Dimitrios Myrisiotis, A Pavan, and NV Vinodchandran. On approximating total variation distance. In *IJCAI*, 2023.

6 Jeremiah Birrell and Mohammadreza Ebrahimi. Adversarially robust deep learning with optimal-transport-regularized divergences. *arXiv preprint*, 2023. `doi:10.48550/arXiv.2309.03791`.

7 Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 47(2):1500–1529, 2022. `doi:10.1287/moor.2021.1178`.

8 Lenore Blum. *Complexity and real computation*. Springer Science & Business Media, 1998.

9 Mark Braverman. On the complexity of real functions. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 155–164. IEEE, 2005. `doi:10.1109/SFCS.2005.58`.

10 Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

11 Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness. In *Advances in Neural Information Processing Systems*, 2022.

12 Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From knothe's transport to brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010. doi:10.1137/080740647.

13 Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint*, 2024. arXiv:2407.18163.

14 Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3733–3742, 2017.

15 Omid Etesami, Saeed Mahloujifar, and Mohammad Mahmoody. Computational concentration of measure: Optimal bounds, reductions, and more. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 345–363. SIAM, 2020. doi:10.1137/1.9781611975994.21.

16 Weiming Feng, Heng Guo, Mark Jerrum, and Jiaheng Wang. A simple polynomial-time approximation algorithm for the total variation distance between two product distributions. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 343–347. SIAM, 2023. doi:10.1137/1.9781611977585.CH30.

17 Alessio Figalli and Cédric Villani. *Optimal Transport and Curvature*, pages 171–217. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi:10.1007/978-3-642-21861-3_4.

18 Nicolas Ford, Daniel Duckworth, Mohammad Norouzi, and George E Dahl. The importance of generation order in language modeling. *arXiv preprint*, 2018. arXiv:1808.07910.

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

20 Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A. Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, pages 2053–2061, 2015.

21 Alfred Galichon. The unreasonable effectiveness of optimal transport in economics. *Proceeding of the 2020 World Congress of the Econometric Society*, 2020.

22 Nathael Gozlan and Christian Léonard. Transport inequalities. a survey. *Markov Processes And Related Fields*, 16:635–736, 2010.

23 Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004. doi:10.1023/B:VISI.0000036836.66311.97.

24 Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2), 2021.

25 Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

26 Daegyu Kim et al. Improving diffusion-based generative models via approximated optimal transport. *arXiv preprint*, 2024. arXiv:2403.05069.

27 Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52, 1957.

28 Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49, 1984.

29 Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

30 Rémi Lassalle. Causal transport plans and their monge–kantorovich problems. *Stochastic Analysis and Applications*, 36(3):452–484, 2018.

31 Saeed Mahloujifar and Mohammad Mahmoody. Can adversarially robust learning leverage computational hardness? In *Algorithmic Learning Theory*, pages 581–609. PMLR, 2019. URL: `http://proceedings.mlr.press/v98/mahloujifar19a.html`.

32 Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024.

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

34 Paul Montesuma, Loic Ngolè Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

35 Ludovic Métivier, Romain Brossier, Jean Virieux, and Jesus de la Puente. Measuring the misfit between seismograms using an optimal transport distance. *Geophysical Journal International*, 205(1):345–377, 2016.

36 Raymond Edward Alan Christopher Paley and Norbert Wiener. *Fourier transforms in the complex domain*, volume 19. American Mathematical Soc., 1934.

37 Gabriel Peyré. Course notes on computational optimal transport. *Mathematical Tours*, 2024. URL: `https://mathematical-tours.github.io/`.

38 Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

39 Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

40 Filippo Santambrogio. Models and applications of optimal transport theory, 2009. Lecture Notes, Grenoble.

41 Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

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

43 Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.

44 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

45 Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

46 Hongkang Yang and Esteban G. Tabak. Clustering, factor discovery and optimal transport. *IMA Journal of Applied Mathematics*, 10(4):1353–1387, 2021. `doi:10.1093/imaiai/iaaa040`.

47 Yi-Zhuang You et al. Renormalization group flow, optimal transport and diffusion-based generative model. *Physical Review E*, 2024.