Computational Management Science (2022) 19:269–293
https://doi.org/10.1007/s10287-021-00415-7

ORIGINAL PAPER

The nested Sinkhorn divergence to learn the nested distance

Alois Pichler[1] · Michael Weinhardt[1]

Received: 10 February 2021 / Accepted: 13 September 2021 / Published online: 27 September 2021

Abstract
The nested distance builds on the Wasserstein distance to quantify the difference of stochastic processes, including also the evolution of information modelled by filtrations. The Sinkhorn divergence is a relaxation of the Wasserstein distance, which can be computed considerably faster. For this reason we employ the Sinkhorn divergence and take advantage of the related (fixed point) iteration algorithm. Furthermore, we investigate the transition of the entropy throughout the stages of the stochastic process and provide an entropy-regularized nested distance formulation, including a characterization of its dual. Numerical experiments affirm the computational advantage and supremacy.

Keywords Nested distance · Optimal transport · Sinkhorn divergence · Entropy

Mathematics Subject Classification 90C08 · 90C15 · 60G07

1 Introduction

The Wasserstein distance, also known as Monge–Kantorovich distance, is used in optimal transport theory to describe and characterize optimal transitions between probability measures. They are characterized by the lowest (or cheapest) average costs to fully transfer a probability measure into another. The costs are most typically proportional to the distance of locations to be connected. Rachev and Rüschendorf (1998) provide a comprehensive discussion of the Wasserstein distance and Villani (2009) summarizes the optimal transport theory.

The nested distance generalizes and extends the theory from probability measures to stochastic processes. It is based on the Wasserstein distance and has been introduced

Alois Pichler: DFG, German Research Foundation—Project-ID 416228727—SFB 1410.

✉ Alois Pichler
alois.pichler@math.tu-chemnitz.de

[1] Faculty of Mathematics, University of Technology, Chemnitz, 90126 Chemnitz, Germany

by Pflug (2009), cf. also Pflug and Pichler (2012). The nested distance quantifies the distance of stochastic processes and multistage stochastic programs are continuous with respect to the nested distance. Multistage stochastic programming has applications in many sectors, e.g., the financial sector (Edirisinghe 2005; Brodt 1983), in management science or in energy economics (Analui and Pflug 2014; Beltrán et al. 2017; Carpentier et al. 2012, 2015). The prices, demands, etc., are often modeled as a stochastic process $\xi = (\xi_0, \dots, \xi_T)$ and the optimal values are rarely obtained analytically. For the numerical approach the stochastic process is replaced by a finite valued stochastic scenario process $\tilde{\xi} = (\tilde{\xi}_0, \dots, \tilde{\xi}_T)$, which is a finite tree. Naturally, the approximation error should be minimized without unnecessarily increasing the complexity of the computational effort. Kirui et al. (2020) provide a Julia package for generating scenario trees and scenario lattices for multistage stochastic programming. Maggioni and Pflug (2019) provide guaranteed bounds and Horejšová et al. (2020) investigate corresponding reduction techniques.

This paper addresses the *Sinkhorn divergence* in place of the Wasserstein distance. This pseudo-distance is also called *Sinkhorn distance* or *Sinkhorn loss*. In contrast to the exact implementation of Bertsekas and Castanon (1989), e.g., Sinkhorn divergence corresponds to a regularization of the Wasserstein distance, which is strictly convex and which allows to improve the efficiency of the computation by applying Sinkhorn's (fixed-point) iteration procedure. The relaxation itself is similar to the modified objective of interior-point methods in numerical optimization. A cornerstone is the theorem by Sinkhorn (1967) that shows both a unique decomposition for non-negative matrices and ensures convergence of the associated iterative scheme. Cuturi (2013) has shown the potential of the Sinkhorn divergence and made it known to a wider audience. Nowadays, Sinkhorn divergence is used in statistical applications, cf. Bigot et al. (2019) and Luise et al. (2018), for image recognition and machine learning, cf. Kolouri et al. (2017) and Genevay et al. (2018), among many other applications.

Extending Sinkhorn's algorithm to multistage stochastic programming has been proposed recently in Tran (2020, Section 5.2.3, pp. 97–99), where a numerical example indicating computational advantages is also given. This paper resumes this idea and assesses the entropy relaxed nested distance from theoretical perspective. We address its approximating properties and derive its convex conjugate, the dual. Moreover, numerical tests included confirm the computational advantage regarding the simplicity of the implementation as well as significant gains in speed.

Outline of the paper The following Sect. 2 introduces the notation and provides the definitions to discuss the nested distance. Additionally, the importance of the filtration and the complexity of the computation is shown. Section 3 introduces the Sinkhorn divergence and derive its dual. In Sect. 4 we regularize the nested distance and show the equality between two different approaches. Results and comparisons are visualized and discussed in Sect. 5. Section 6 summarizes and concludes the paper.

2 Preliminaries

This section recalls the definition of distances generally, of the Wasserstein distance and nested distance and provides an example to highlight the impact of information available, which is increasing gradually over time and stages. Throughout, we shall base our exposition on a probability space $(\Xi, \mathcal{F}, P)$.

2.1 Wasserstein distance

The Wasserstein distance is a distance for probability measures. It is the building block for the process distance, which involves information in addition and its regularized version, which we address here, the Sinkhorn divergence. The Sinkhorn divergence is not a distance in itself. To point out the differences we highlight the defining elements.

Definition 2.1 (*Distance of measures*) Let $\mathcal{P}$ be a set of probability measures on Ξ. A function $d\colon \mathcal{P}\times\mathcal{P} \to [0, \infty)$ is called *distance*, if it satisfies the following conditions:

(i) Nonnegativity: for all $P_1, P_2 \in \mathcal{P}$,

$$d(P_1, P_2) \geq 0;$$

(ii) Symmetry: for all $P_1, P_2 \in \mathcal{P}$,

$$d(P_1, P_2) = d(P_2, P_1);$$

(iii) Triangle inequality: for all P_1, P_2 and $P_3 \in \mathcal{P}$,

$$d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2);$$

(iv) Definiteness: if $d(P_1, P_2) = 0$, then $P_1 = P_2$.

Rachev (1991) presents a huge variety of probability metrics. Here, we focus on the Wasserstein distance, which allows a generalization for stochastic processes. For this we assume that the sample space $\Xi = \mathbb{R}^d$ is equipped with a metric d.

Definition 2.2 (*Wasserstein distance*) Let P and $\tilde{P}$ be two probability measure on Ξ endowed with a distance $d\colon \Xi \times \Xi \to \mathbb{R}$ with finite moment of order r. The *Wasserstein distance* of order $r \geq 1$ is

$$d^r(P, \tilde{P}) := \inf_{\pi} \iint_{\Xi\times\Xi} d(\xi, \tilde{\xi})^r \, \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}),$$

where the infimum is over all probability measures π on $\Xi \times \Xi$ with marginals P and $\tilde{P}$, respectively.

Remark 2.3 (**Distance versus cost functions**) The definition of the Wasserstein distance presented here starts with a distance d on Ξ and the Wasserstein distance is a distance on $\mathcal{P}$ in the sense of Definition 2.1 above. We mention that the literature occasionally develops the theory for cost functions $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ instead of the distance d. Also, the results presented below extend to cost functions in place of the distance on the underlying space.

In a discrete framework, probability measures are of the form $P = \sum_{i=1}^n p_i\,\delta_{\xi_i}$ with $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$ and the support of P ($\{\xi_i : i = 1, 2, \dots, n\} \subset \Xi$) is finite. By Definition 2.1, the Wasserstein distance d^r of two discrete measures $P = \sum_{i=1}^n p_i\,\delta_{\xi_i}$ and $\tilde{P} = \sum_{j=1}^{\tilde{n}} \tilde{p}_j\,\delta_{\tilde{\xi}_j}$ is the r-th root of the optimal value of

$$\begin{aligned} \text{minimize}_{\text{ in }\pi}\ & \sum_{i=1}^{n}\sum_{j=1}^{\tilde{n}} \pi_{ij}\, d_{ij}^r \\ \text{subject to } & \sum_{j=1}^{\tilde{n}} \pi_{ij} = p_i, && i = 1, \dots, n, \\ & \sum_{i=1}^{n} \pi_{ij} = \tilde{p}_j, && j = 1, \dots \tilde{n} \text{ and} \\ & \pi_{ij} \geq 0, && \end{aligned} \tag{2.1}$$

where

$$d_{ij} := d(\xi_i, \tilde{\xi}_j) \tag{2.2}$$

is an $n \times \tilde{n}$-matrix collecting all distances. The optimal measure in (2.1) is denoted π^W and called an optimal transport plan. The convex, linear dual of (2.1) is

$$\text{maximize}_{\text{ in }\lambda\text{ and }\mu} \sum_{i=1}^{n} p_i\,\lambda_i + \sum_{j=1}^{\tilde{n}} \tilde{p}_j\,\mu_j \tag{2.3a}$$

$$\text{subject to } \lambda_i + \mu_j \leq d_{ij}^r \ \text{ for all } i = 1, \dots n \text{ and } j = 1, \dots \tilde{n}. \tag{2.3b}$$

Remark 2.4 The problem (2.1) can be written as linear optimization problem

$$\begin{aligned} \text{minimize}_{\text{ in }x}\ & c^\top x \\ \text{subject to } & Ax = b, \\ & x \geq 0, \end{aligned}$$

where $x = (\pi_{11}, \pi_{21}, \dots, \pi_{n\tilde{n}})^\top$, $c = (d_{11}, d_{21}, \dots, d_{n\tilde{n}})^\top$, $b = (p_1, \dots, p_n, \tilde{p}_1, \dots, \tilde{p}_{\tilde{n}})^\top$ and A is the matrix

$$A = \begin{pmatrix} \mathbb{1}_{\tilde{n}} \otimes I_n \\ I_{\tilde{n}} \otimes \mathbb{1}_n \end{pmatrix}$$

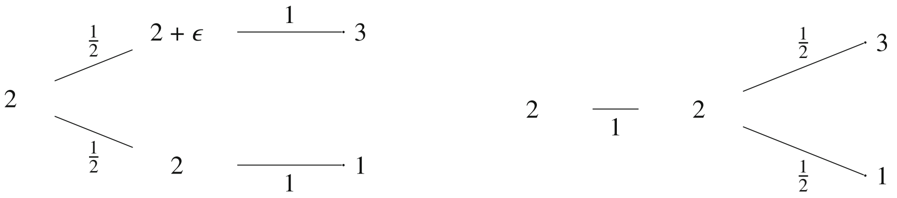


Fig. 1 Two processes illustrating two different flows of information, cf. Heitsch et al. (2006), Kovacevic and Pichler (2015). The arcs of the stochastic tree display the transition probabilities

with $\mathbb{1} = (1, \dots, 1)$; here, $\otimes$ denotes the Kronecker product.

2.2 The distance of stochastic processes

Be two probability spaces. We now consider two stochastic processes with realizations $\xi, \tilde{\xi} \in \Xi$ and $\Xi{:=}\Xi_0 \times \Xi_1 \times \cdots \times \Xi_T$. There are many metrics d such that (Ξ, d) is a metric space. Without loss of generality we may set $\Xi_t = \mathbb{R}$ for all $t \in \{0, 1, \dots, T\}$ and employ the ℓ^1-distance, i.e., $d(\xi, \tilde{\xi}) = \sum_{t=0}^{T} |\xi_t - \tilde{\xi}_t|$. As in (2.1) above, the distance matrix d_{ij} collects the distances of scenarios for discrete measures, cf. (2.2).

A stochastic process with finitely many states (i.e., outcomes) for $t \in \{0, 1, \dots, T\}$ is a *scenario tree*. Scenario trees are frequently employed in optimization under uncertainty to model the random outcome in the evolution of a process which describes the random price, say, of an underlying asset. They are convenient, because they can be implemented in computers to assess each individual trajectory as possible realization of the stochastic process.

The Figs. 1 and 3 depict such scenario tree, they indicate the transition probabilities in addition.

Remark 2.5 Figure 1 illustrates that the Wasserstein distance does not capture the different information (knowledge) available at the intermediate stage. Indeed, with $\epsilon > 0$, the matrix collecting the distances of the trajectories taken from both trees is

$$d = \begin{pmatrix} \epsilon & 2+\epsilon \\ 2 & 0 \end{pmatrix}$$

and the optimal transport plan for the Wasserstein distance is

$$\pi = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The Wasserstein distance, according (2.1), is $d = \sum_{i,j} d_{ij}\, \pi_{ij} = {}^{\epsilon}\!/_{2}$, where a small value for ϵ indicates that the processes are similar.

However, the information available at stage 1 is very distinct in both trees in Fig. 1. When observing $2 + \epsilon$ at stage 1 in the first tree it is certain that the next observation is 3, and it will be 1 when observing 2. In contrast, the second process does not provide any certainty whether the result will be 1 or 3 after observing 2 at the first stage.

We conclude from the preceding remark that the Wasserstein distance is not suitable to distinguish stochastic processes with different flows of information. The reason is that this approach does not involve conditional probabilities at stages $t = 0, 1, \dots, T-1$, but only probabilities at the final stage $t = T$, where all the information available at intermediate stages is ignored.

We follow the usual convention and express information, which is accessible, by corresponding sets. The information available at every stage t includes information from preceding stages, which have been revealed, but excludes information from later, future stages. For this reason the sets

$$A_1 \times \cdots \times A_t \times \Xi_{t+1} \times \cdots \times \Xi_T, \quad A_{t'} \subset \Xi_{t'} \text{ measurable},$$

encode the information available at stage t, they constitute the σ-algebra $\mathcal{F}_t$ ($\tilde{\mathcal{F}}_t$, resp.). The following generalization of the Wasserstein distance takes this flow of increasing information explicitly into account. We state the definition involving general probability measures here, although we work with discrete measures only in what follows.

Definition 2.6 (*The nested distance*) The *nested distance* of order $r \geq 1$ of two filtered probability spaces $\mathbb{P} = (\Xi, (\mathcal{F}_t), P)$ and $\tilde{\mathbb{P}} = (\tilde{\Xi}, (\tilde{\mathcal{F}}_t), \tilde{P})$ with finite moment of order r with respect to the distance $d\colon \Xi \times \tilde{\Xi} \to \mathbb{R}$ is the optimal value of the optimization problem

$$\text{minimize}_{\text{ in } \pi} \left(\iint_{\Xi \times \tilde{\Xi}} d(\xi, \tilde{\xi})^r \, \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}) \right)^{1/r} \tag{2.4}$$

$$\text{subject to } \pi(A \times \tilde{\Xi} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A \mid \mathcal{F}_t) \text{ a.s. for every } A \in \mathcal{F}_t,\ t = 1, \dots, T, \tag{2.5}$$

$$\pi(\Xi \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \tilde{P}(B \mid \tilde{\mathcal{F}}_t) \text{ a.s. for every } B \in \tilde{\mathcal{F}}_t,\ t = 1, \dots, T, \tag{2.6}$$

where the infimum is among all bivariate probability measures $\pi \in \mathcal{P}(\Xi \times \tilde{\Xi})$. The optimal value of (2.1), the nested distance of order r, is denoted by $\boldsymbol{d}^r(\mathbb{P}, \tilde{\mathbb{P}})$.

For discrete stochastic processes we use trees to model the whole space and filtration. In the stochastic tree, $\mathcal{N}_t$ ($\tilde{\mathcal{N}}_t$, resp.) denotes the set of all nodes at the stage t. Furthermore, a predecessor m of the node i, not necessarily the immediate predecessor, is indicated by $m \prec i$. Here, we may replace the conditional probabilities in (2.5) and (2.6) by the genuine transition probabilities. The arcs of the tree in Fig. 1 exemplarily display these transition probabilities.

Algorithm 1: Nested computation of the nested distance $\boldsymbol{d}^r(\mathbb{P},\tilde{\mathbb{P}})$ of two tree-processes $\mathbb{P}$ and $\tilde{\mathbb{P}}$.

```
Input: for all combinations of leaf nodes $i \in \mathcal{N}_T$ and $j \in \tilde{\mathcal{N}}_T$ with predecessors
    $(i_0, i_1, \dots, i_{T-1}, i)$ and $(j_0, j_1, \dots, j_{T-1}, j)$ set
    $\boldsymbol{d}_T^r(i,j) := d\Big((\xi_0, \xi_{i_1}, \dots, \xi_i),\ (\tilde{\xi}_0, \tilde{\xi}_{j_1}, \dots, \tilde{\xi}_j)\Big)^r$
Output: the optimal transport plan at the leaf nodes $i \in \mathcal{N}_T$ and $j \in \tilde{\mathcal{N}}_T$ is
    $\pi(i,j) = \pi_1(i_1, j_1 \mid i_0, j_0) \cdot \dots \cdot \pi_{T-1}(i, j \mid i_{T-1}, j_{T-1})$.
for $t = T-1$ down to 0 and every combination of inner nodes $i' \in \mathcal{N}_t$ and $j' \in \tilde{\mathcal{N}}_t$ do
    solve the linear programs
        $\text{minimize}_{\text{ in } \pi} \sum_{i' \in i_t+,\ j' \in j_t+} \pi(i', j' \mid i_t, j_t) \cdot \boldsymbol{d}_{t+1}^r(i', j')$
        $\text{subject to} \sum_{j' \in j_t+} \pi(i', j' \mid i_t, j_t) = P(i' \mid i_t), \quad i' \in i_t+,$
        $\sum_{i' \in i_t+} \pi(i', j' \mid i_t, j_t) = \tilde{P}(j' \mid j_t), \quad j' \in j_t+,$
        $\pi(i', j' \mid i_t, j_t) \geq 0 \qquad (2.9)$
    and denote its optimal value by $\boldsymbol{d}_t^r(i_t, j_t)$.
Result: The nested distance is $\boldsymbol{d}^r(\mathbb{P},\tilde{\mathbb{P}}) := \boldsymbol{d}_0^r(0,0)$.
```

The nested distance for stochastic trees is the r-th root of the optimal value of

$$
\begin{aligned}
\text{minimize}_{\text{ in } \pi} \ & \sum_{i,j} \pi_{ij} \cdot d_{ij}^r \\
\text{subject to} \ & \sum_{j \succ j_t} \pi(i, j \mid i_t, j_t) = P(i \mid i_t), && i_t \prec i,\ j_t, \\
& \sum_{i \succ i_t} \pi(i, j \mid i_t, j_t) = \tilde{P}(j \mid j_t), && j_t \prec j,\ i_t, \\
& \pi_{ij} \geq 0 \text{ and } \sum_{i,j} \pi_{ij} = 1,
\end{aligned}
\tag{2.7}
$$

where $i \in \mathcal{N}_T$ and $j \in \tilde{\mathcal{N}}_T$ are the leaf nodes and $i_t \in \mathcal{N}_t$ as well as $j_t \in \tilde{\mathcal{N}}_t$ are nodes at the same stage t and $P(i \mid i_t) := \frac{P(i)}{P(i_t)}$ is the conditional probability. Analogously, the conditional probabilities $\pi(i, j \mid i_t, j_t)$ are

$$
\pi(i, j \mid i_t, j_t) := \frac{\pi_{ij}}{\sum_{i' \succ i_t, j' \succ j_t} \pi_{i'j'}}. \tag{2.8}
$$

Remark 2.7 Employing the definition (2.8) for $\pi(i, j \mid i_t, j_t)$ reveals that the problem (2.7) is indeed a *linear* program in π (cf. (2.1)).

2.3 Rapid, nested computation of the process distance

This subsection addresses an advanced approach for solving the linear program (2.7). We first recall the tower property, which allows an important simplification of the constraints in (2.4).

Lemma 2.8 *To compute the nested distance it is enough to condition on the immediately following σ-algebra: the conditions*

$$\pi\left(A \times \Xi \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right) \textit{ for all } A \in \mathcal{F}_T$$

in (2.4) *may be replaced by*

$$\pi\left(A \times \Xi \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right) \textit{ for all } A \in \mathcal{F}_{t+1}.$$

Proof The proof is based on the tower property of the expectation and can be found in Pflug and Pichler (2014, Lemma 2.43). □

As a result of the tower property the full problem (2.7) can be calculated faster in a recursive way and the matrix for the constraints does not have to be stored. We employ this result in an algorithm below. For further details we refer to Pflug and Pichler (2014, Chapter 2.10.3). The collection of all direct successors of node i_t (j_t, resp.) is denoted by i_t+ (j_t+, resp.).

3 Sinkhorn divergence

In what follows we consider the entropy-regularization of the Wasserstein distance (2.1) and characterize its dual. Moreover, we recall Sinkhorn's algorithm, which allows and provides a considerably faster implementation. These results are combined then to accelerate the computation of the nested distance.

3.1 Entropy-regularized Wasserstein distance

Interior point methods add a logarithmic penalty to the objective to force the optimal solution of the modified problem into the strict interior. The Sinkhorn distance proceeds similarly. The regularizing term $H(x) := -\sum_{i,j} x_{ij} \log x_{ij}$ is added to the cost function in problem (2.1). This has shown beneficiary in other problem settings as well.

Remark 3.1 The mapping $\varphi(y) := y \log y$ is convex and negative for $y \in (0, 1)$ with continuous extensions $\varphi(0) = \varphi(1) = 0$ so that $H(x) \geq 0$, provided that all $x_{ij} \in [0, 1]$.

Definition 3.2 (*Sinkhorn divergence*) The *Sinkhorn divergence* is the objectove of the optimization problem

$$\text{minimize}_{\text{ in } \pi} \sum_{i=1}^{n} \sum_{j=1}^{\tilde{n}} \pi_{ij}\, d_{ij}^r - \frac{1}{\lambda} H(\pi) \tag{3.1a}$$

$$\begin{aligned} \text{subject to } \sum_{j=1}^{\tilde{n}} \pi_{ij} &= p_i, & i &= 1, \ldots, n, \\ \sum_{i=1}^{n} \pi_{ij} &= \tilde{p}_j, & j &= 1, \ldots, \tilde{n}, \\ \pi_{ij} &> 0 & \text{for all } & i, j, \end{aligned} \tag{3.1b}$$

where d is a distance or a cost matrix and $\lambda > 0$ is a regularization parameter. With π^S being the optimal transport in (3.1a)–(3.1b) we denote the Sinkhorn divergence by

$$d_S^r := \sum_{i=1}^{n} \sum_{j=1}^{\tilde{n}} \pi_{ij}^S\, d_{ij}^r$$

and the Sinkhorn divergence including the entropy by

$$de_S^r := \sum_{i=1}^{n} \sum_{j=1}^{\tilde{n}} \pi_{ij}^S\, d_{ij}^r - \frac{1}{\lambda} H\big(\pi^S\big).$$

We may mention here that we avoid the term Sinkhorn *distance* since for all $\lambda > 0$, the Sinkhorn divergence d_S^r is strictly positive and de_S^r can be negative for small λ which violates the axioms of a distance given in Definition 2.1 above (particularly (i), (iii) and (iv)). Strict positivity of d_S^r can be forced by a correction term, the so-called Sinkhorn loss [see Bigot et al. (2019, Definition 2.3)] or by employing the cost matrix $d \cdot \mathbb{1}_{p \neq \tilde{p}}$ instead.

Remark 3.3 The strict inequality constraint (3.1b) is not a restriction. Indeed, the mapping $\varphi(\cdot)$ defined in Remark 3.1 has derivative $\varphi'(0) = -\infty$ and thus it follows that every optimal measure satisfies the strict inequality $\pi_{ij} > 0$ for $\lambda > 0$.

We have the following inequalities.

Proposition 3.4 (Comparison of Sinkhorn and Wasserstein) *It holds that*

$$de_S^r \leq d^r \leq d_S^r. \tag{3.2}$$

Proof Recall that $\pi \log \pi \leq 0$ for all $\pi \in (0,1)$ and thus it holds that $\sum_{i=1}^{n} \sum_{j=1}^{\tilde{n}} \pi_{ij}\, d_{ij}^r + \frac{1}{\lambda} \sum_{i=1}^{n} \sum_{j=1}^{\tilde{n}} \pi_{ij} \log \pi_{ij} \leq \sum_{i=1}^{n} \sum_{j=1}^{\tilde{n}} \pi_{ij}\, d_{ij}^r$ for all $\pi \in (0,1]^{n \times \tilde{n}}$. It follows that

$$\min_{\pi} \sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\pi_{ij}\, d_{ij}^r + \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\pi_{ij}\log\pi_{ij} \le \min_{\pi} \sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\pi_{ij}\, d_{ij}^r$$

and thus the first inequality. The remaining inequality is clear by the definition of the Wasserstein distance. □

Both Sinkhorn divergences d_S^r and de_S^r approximate the Wasserstein distance d^r, and we have convergence for $\lambda \to \infty$ to d^r. The following proposition provides precise bounds.

Proposition 3.5 *For every $\lambda > 0$ we have*

$$0 \le d_S^r - d^r \le \frac{1}{\lambda}\Big(H(\pi^S) - H(\pi^W)\Big) \tag{3.3}$$

and

$$0 \le d^r - de_S^r \le \frac{1}{\lambda}H(\pi^S) \le \frac{1}{\lambda}H(p\cdot\tilde p^\top) \tag{3.4}$$

with $p = (p_1, \dots, p_n)$ and $\tilde p = (\tilde p_1, \dots, \tilde p_{\tilde n})$, respectively.

Proof The first inequalities follow from (3.2) and from optimality of π^S in the inequality

$$d_S^r - \frac{1}{\lambda}H(\pi^S) \le d^r - \frac{1}{\lambda}H(\pi^W).$$

The latter again with (3.2) and $d_S^r - de_S^r = \frac{1}{\lambda}H\big(\pi^S\big)$. Finally, by the log sum inequality, $H(\pi) \le H\big(p\cdot\tilde p^\top\big)$ for every measure π with marginals p and $\tilde p$. □

Remark 3.6 As a consequence of the log sum inequality we obtain as well that $H(\pi^S) \le \log n + \log\tilde n$. The inequalities (3.3) and (3.4) thus give strict upper bounds in comparing the Wasserstein distance and the Sinkhorn divergence.

Alternative definitions There exist alternative concepts of the Sinkhorn divergence which we want to mention here. The first alternative definition involves the Kullback–Leibler divergence $D_{KL}(\pi \mid P\otimes\tilde P)$, which is defined as

$$D_{KL}(\pi \mid P\otimes\tilde P) := -\sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\pi_{ij}\log\frac{\pi_{ij}}{p_i\,\tilde p_j} = H(P) + H(\tilde P) - H(\pi),$$

where the latter equality is justified provided that π has marginal measures P and $\tilde P$. The Sinkhorn divergence (in the alternative definition) is the r-th root of the optimal value of

$$\begin{aligned}
&\text{minimize}_{\text{ in }\pi} \sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\pi_{ij}\, d_{ij}^r && && \text{(3.5a)}\\
&\text{subject to} \sum_{j=1}^{\tilde n}\pi_{ij}=p_i, && i=1,\dots,n, &&\\
&\sum_{i=1}^{n}\pi_{ij}=\tilde p_j, && j=1,\dots,\tilde n, &&\\
&\pi_{ij}>0 && \text{and} &&\\
&D_{KL}(\pi \mid P\otimes\tilde P)\le\alpha && \text{for all } i, j, && \text{(3.5b)}
\end{aligned}$$

where $\alpha \geq 0$ is the regularization parameter. For each α in (3.5b) we have by the duality theory a corresponding λ in (3.1a) such that the optimal values coincide. Let $\alpha > 0$ and π^{KL} be the solution to problem (3.5a)–(3.5b) with Lagrange multipliers β and γ. Then the optimal value of problem (3.5a) equals d_S^r from (3.1a) with

$$\lambda = -\frac{\log(\pi_{ij}^{KL})+1}{d_{ij}+\beta_i+\gamma_j}$$

for any $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,\tilde n\}$. For further information and illustration we refer to Cuturi (2013, Section 3).

A further, possible definition employs a different entropy regularization given by

$$\tilde H(\pi) = -\sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\pi_{ij}\cdot(\log\pi_{ij}-1).$$

Luise et al. (2018) use this definition for Sinkhorn approximation for learning with Wasserstein distance and prove an exponential convergence. This definition leads to a similar matrix decomposition and iterative algorithm as described in the following sections.

3.2 Dual representation of Sinkhorn

We shall derive Sinkhorn's algorithm and its extension to the nested distance via duality. To this end consider the Lagrangian function

$$\begin{aligned}
L(\pi;\beta,\gamma) := &\sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\pi_{ij}\, d_{ij} + \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\pi_{ij}\log\pi_{ij} + \beta^\top(p-\pi\cdot\mathbb{1})\\
&+(\tilde p-\mathbb{1}^\top\cdot\pi)^\top\gamma \qquad \text{(3.6)}
\end{aligned}$$

of the problem (3.2). The partial derivatives are

$$\frac{\partial L}{\partial \pi_{ij}} = \frac{1}{\lambda}\left(\log \pi_{ij} + 1\right) + d_{ij} - \beta_i - \gamma_j = 0, \tag{3.7}$$

and it follows from (3.7) that the optimal measure has entries

$$\begin{aligned}\pi^*_{ij} &= \exp\left(-\lambda(d_{ij} - \beta_i - \gamma_j) - 1\right)\\ &= \operatorname{diag}\Big(\exp(\lambda\,\beta - 1/2)\Big)\cdot\exp(-\lambda\, d)\cdot\operatorname{diag}\Big(\exp(\lambda\,\gamma - 1/2)\Big).\end{aligned} \tag{3.8}$$

By inserting π^*_{ij} in the Lagrangian function L we get the convex dual function

$$\begin{aligned}
&d(\beta,\gamma) := \inf_{\pi} L(\pi;\beta,\gamma) = L(\pi^*;\beta,\gamma)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{\tilde n} d_{ij}\cdot e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1} - \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{\tilde n} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}\cdot\left(\lambda(d_{ij}-\beta_i-\gamma_j)+1\right)\\
&\quad+\sum_{i=1}^{n}\beta_i\left(p_i - \sum_{j=1}^{\tilde n} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}\right) + \sum_{j=1}^{\tilde n}\gamma_j\left(\tilde p_j - \sum_{i=1}^{n} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{\tilde n}\left(\beta_i+\gamma_j-\frac{1}{\lambda}\right)e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1} + \sum_{i=1}^{n}\beta_i p_i + \sum_{j=1}^{\tilde n}\gamma_j\tilde p_j\\
&\quad-\sum_{i=1}^{n}\beta_i\left(\sum_{j=1}^{\tilde n} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}\right) - \sum_{j=1}^{\tilde n}\gamma_j\left(\sum_{i=1}^{n} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}\right)\\
&= \sum_{i=1}^{n}\beta_i\, p_i + \sum_{j=1}^{\tilde n}\gamma_j\,\tilde p_j - \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{\tilde n} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}.
\end{aligned}$$

The dual problem thus is

$$\begin{aligned}&\text{maximize}_{\text{ in }\beta,\gamma}\ \sum_{i=1}^{n}\beta_i\, p_i + \sum_{j=1}^{\tilde n}\gamma_j\,\tilde p_j - \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{\tilde n} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}\\ &\qquad\text{subject to } \beta\in\mathbb{R}^n,\ \gamma\in\mathbb{R}^{\tilde n}.\end{aligned}$$

Due to $\sum_{i,j} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1} = 1$ we may write the latter problem as

$$\text{maximize}_{\text{ in }\beta,\gamma}\ \sum_{i=1}^{n}\beta_i\, p_i + \sum_{j=1}^{\tilde n}\gamma_j\,\tilde p_j \tag{3.9a}$$

$$\text{subject to } \sum_{i=1}^{n}\sum_{j=1}^{\tilde n} e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1} = 1 \text{ and } \beta\in\mathbb{R}^n,\ \gamma\in\mathbb{R}^{\tilde n}. \tag{3.9b}$$

Remark 3.7 We deduce from (3.9b) that $-\lambda\left(d_{ij}-\beta_i-\gamma_j\right)-1\leq 0$, or

$$\beta_i+\gamma_j\leq d_{ij}+\frac{1}{\lambda}\quad \text{for all } i,\, j \tag{3.10}$$

provided that $\lambda>0$. It is thus apparent that (3.9a)–(3.9b) is a relaxation of problem (2.3a)–(2.3b) together with the constraint (3.10). As well, observe that both problems coincide for $\lambda\to\infty$ in (3.9b).

3.3 Sinkhorn's algorithm

To derive Sinkhorn's algorithm we consider the Lagrangian function (3.6) again, but now for the remaining variables. Similar to π^* in (3.8), the gradients are

$$\frac{\partial L}{\partial\beta_i}=p_i-\sum_{j=1}^{\tilde n}\pi_{ij}=p_i-\sum_{j=1}^{\tilde n}e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}=0$$

and

$$\frac{\partial L}{\partial\gamma_j}=\tilde p_j-\sum_{i=1}^{n}\pi_{ij}=\tilde p_j-\sum_{i=1}^{n}e^{-\lambda(d_{ij}-\beta_i-\gamma_j)-1}=0$$

so that the equations

$$\beta_i=\frac{1}{\lambda}\log\left(\frac{p_i}{\sum_{j=1}^{\tilde n}e^{-\lambda(d_{ij}-\gamma_j)-1}}\right)\quad\text{and}\quad \gamma_j=\frac{1}{\lambda}\log\left(\frac{\tilde p_j}{\sum_{i=1}^{n}e^{-\lambda(d_{ij}-\beta_i)-1}}\right)$$

follow. To avoid the logarithm introduce $\tilde\beta_i{:=}e^{\lambda\,\beta_i-1/2}$ and $\tilde\gamma_j{:=}e^{\lambda\,\gamma_j-1/2}$ and rewrite the latter equations as

$$\tilde\beta_i=\frac{p_i}{\sum_{j=1}^{\tilde n}e^{-\lambda\,d_{ij}}\,\tilde\gamma_j}\quad\text{and}\quad \tilde\gamma_j=\frac{\tilde p_j}{\sum_{i=1}^{n}\tilde\beta_i\,e^{-\lambda\,d_{ij}}}, \tag{3.11}$$

while the optimal transition plan (3.8) is

$$\pi^*_{ij}=\tilde\beta_i\cdot e^{-\lambda\,d_{ij}}\cdot\tilde\gamma_j.$$

The simple starting point of Sinkhorn's iteration is that (3.11) can be used to determine $\tilde\beta$ and $\tilde\gamma$ alternately. Indeed, Sinkhorn's theorem (cf. Sinkhorn 1967; Sinkhorn and Knopp 1967) for the matrix decomposition ensures that iterating (3.11) converges and the vectors $\tilde\beta$ and $\tilde\gamma$ are unique up to a scalar. Algorithm 2 summarizes the individual steps again.

Algorithm 2: Sinkhorn's iteration

```
Input: distance matrix $d^r \in \mathbb{R}^{n\times\tilde{n}}_{\geq 0}$, probability vectors $p \in \mathbb{R}^n_{\geq 0}$, $\tilde{p} \in \mathbb{R}^{\tilde{n}}_{\geq 0}$, regularization
       parameter $\lambda > 0$, stopping criterion and a starting value $\tilde{\gamma} = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_{\tilde{n}})$
Output: $\tilde{\beta}$, $\tilde{\gamma}$ for $\operatorname{diag}(\tilde{\beta}) \cdot e^{-\lambda\, d^r} \cdot \operatorname{diag}(\tilde{\gamma})$
set
        $$k_{ij} := \exp\big(-\lambda\, d^r_{ij}\big).  \qquad (3.12)$$
while stopping criterion is not satisfied do
    for $i = 1$ to $n$ do
        $\tilde{\beta}_i \leftarrow \frac{p_i}{\sum_{j=1}^{\tilde{n}} k_{ij}\, \tilde{\gamma}_j}$
    for $j = 1$ to $\tilde{n}$ do
        $\tilde{\gamma}_j \leftarrow \frac{\tilde{p}_j}{\sum_{i=1}^{n} \tilde{\beta}_i\, k_{ij}}$
Result: The matrix $\pi^*_{ij} = \tilde{\beta}_i\, e^{-\lambda\, d^r_{ij}}\, \tilde{\gamma}_j = \tilde{\beta}_i\, k_{ij}\, \tilde{\gamma}_j$ solves the relaxed Wasserstein
        problem (3.1a)–(3.1b).
```

Remark 3.8 (**Central path**) We want to emphasize that for changing the regularization parameter λ it is note necessary to recompute all powers in (3.12). Indeed, increasing λ to $2 \cdot \lambda$, for example, corresponds to raising all entries in the matrix (3.12) to the power 2, etc.

Remark 3.9 (**Softmax**) The expression (3.11) resembles to what is known as the *Gibbs measure* and to the *softmax* in data science.

Remark 3.10 (**Historical remark**) In the literature, this approach is also known as *matrix scaling* (cf. Rote and Zachariasen 2007), RAS (cf. Bachem and Korte 1979) as well as Iterative Proportional Fitting (cf. Rüschendorf 1995). Kruithof (1937) used the method for the first time in telephone forecasting. The importance of this iteration scheme for data science was probably observed in Cuturi (2013, Algorithm 1) for the first time.

Remark 3.11 We may refer to Altschuler et al. (2017) for a performance analysis including speed and convergence of the Sinkhorn algorithm. For a discussion of numerical stability of the algorithm we refer to Peyré and Cuturi (2019, Section 4.4).

4 Entropic transitions

This section extends the preceding sections and combines the Sinkhorn divergence and the nested distance by incorporating the regularized entropy $\frac{1}{\lambda} H(\pi)$ to the recursive nested distance Algorithm 1 and investigates its properties and consequences. We characterize the nested Sinkhorn divergence first. The main result is used to exploit duality.

4.1 Nested Sinkhorn divergence

Let $\boldsymbol{de}^{(t)}$ be the matrix of incremental divergences of sub-trees at stage t. Analogously to (2.9) we consider the conditional version of the problem (3.1a) and denote by $\beta_{i_t j_t}$ and $\gamma_{j_t i_t}$ the pair of optimal Lagrange parameters associated with the problem

$$\begin{aligned}
\text{minimize}_{\,\text{in}\,\pi} \quad & \sum_{i' \in i_t+,\, j' \in j_t+} \pi(i', j' \mid i_t, j_t) \cdot \boldsymbol{de}^{(t+1)}(i', j') \\
& \qquad + \frac{1}{\lambda} \pi(i', j' \mid i_t, j_t) \cdot \log \pi(i', j' \mid i_t, j_t) \\
\text{subject to} \quad & \sum_{j' \in j_t+} \pi(i', j' \mid i_t, j_t) = P(i' \mid i_t), \qquad\qquad i' \in i_t+, \\
& \sum_{i' \in i_t+} \pi(i', j' \mid i_t, j_t) = \tilde{P}(j' \mid j_t), \qquad\qquad j' \in j_t+, \\
& \pi(i', j' \mid i_t, j_t) > 0,
\end{aligned} \tag{4.1}$$

where $\pi(i', j' | i_t, j_t) = \exp\left(-\lambda\big(\boldsymbol{de}^{(t+1)}_{i_t j_t} - \beta_{i_t j_t} - \gamma_{j_t i_t}\big) - 1\right)$. The optimal value is the new divergence $\boldsymbol{de}^{(t)}(i_t, j_t)$. Computing the nested distance recursively from $t = T - 1$ down to 0 we get

$$\begin{aligned}
\pi_{ij} &= \pi_1(i_1, j_1 \mid i_0, j_0) \cdot \ldots \cdot \pi_{T-1}(i, j \mid i_{T-1}, j_{T-1}) \\
&= e^{-\lambda(\boldsymbol{de}^{(1)}_{i_0 j_0} - \beta_{i_0 j_0} - \gamma_{j_0 i_0}) - 1} \cdot \ldots \cdot e^{-\lambda(\boldsymbol{de}^{(T)}_{i_{T-1} j_{T-1}} - \beta_{i_{T-1} j_{T-1}} - \gamma_{j_{T-1} i_{T-1}}) - 1} \\
&= \exp\left(-T - \lambda \sum_{t=0}^{T-1} \boldsymbol{de}^{(t+1)}_{i_t j_t} - \beta_{i_t j_t} - \gamma_{j_t i_t}\right),
\end{aligned} \tag{4.2}$$

where $i \in \mathcal{N}_T$ and $j \in \tilde{\mathcal{N}}_T$ are the leaf nodes with predecessors $(i_0, i_1, \ldots, i_{T-1}, i)$ and $(j_0, j_1, \ldots, j_{T-1}, j)$. As above introduce

$$\tilde{\beta}_{i_t j_t} := \exp\left(\lambda\, \beta_{i_t j_j} - 1/2\right) \quad \text{and} \quad \tilde{\gamma}_{j_t i_t} := \exp\left(\lambda\, \gamma_{j_t i_t} - 1/2\right).$$

Combining the components it follows that

$$\begin{aligned}
\pi_{ij} &= \exp\left(-T - \lambda \sum_{t=0}^{T-1} \boldsymbol{de}^{(t+1)}_{i_t j_t} - \beta_{i_t j_t} - \gamma_{j_t i_t}\right) \\
&= \prod_{t=0}^{T-1} \tilde{\beta}_{i_t j_t} \exp\left(-\lambda\, \boldsymbol{de}^{(t+1)}_{i_t j_t}\right) \tilde{\gamma}_{j_t i_t},
\end{aligned}$$

where the product is the entry-wise product (Hadamard product).

The following theorem summarizes the relation of the nested distance with the Sinkhorn divergence.

Theorem 4.1 (Entropic relaxation of the nested distance) *The recursive solution* (4.1) ((4.2), resp.) *coincides with the optimal transport plan given by*

$$
\begin{aligned}
\textit{minimize}_{\textit{ in } \pi} \quad & \sum_{i=1}^{n}\sum_{j=1}^{\tilde n} \pi_{ij}\cdot d_{ij}^r + \frac{1}{\lambda}\pi_{ij}\cdot \log\left(\pi_{ij}\right) \\
\textit{subject to} \quad & \sum_{j \succ j_t +} \pi(i, j \mid i_t, j_t) = P(i \mid i_t), \qquad\qquad i_t \prec i, j_t, \\
& \sum_{i \succ i_t +} \pi(i, j \mid i_t, j_t) = \tilde P(j \mid j_t), \qquad\qquad j_t \prec j, i_t, \\
& \pi_{ij} > 0 \textit{ and } \sum_{i,j} \pi_{ij} = 1.
\end{aligned}
\tag{4.3}
$$

Proof First define $\pi := \prod_{t=1}^{T}\pi_t$, where π_t is the conditional transition probability, i.e., the solution at stage t and the matrices are multiplied element-wise (the Hadamard product) as in equation (4.2) above. It follows that

$$
\begin{aligned}
d^r\cdot\pi + \frac{1}{\lambda}\pi\log\pi &= d^r\cdot\prod_{t=1}^{T}\pi_t + \frac{1}{\lambda}\cdot\prod_{t=1}^{T}\pi_t\log\left(\prod_{t=1}^{T}\pi_t\right) \\
&= d^r\cdot\prod_{t=1}^{T}\pi_t + \frac{1}{\lambda}\cdot\prod_{t=1}^{T}\pi_t\cdot\sum_{t=1}^{T}\log\pi_t.
\end{aligned}
\tag{4.4}
$$

Observe that $\pi_t(A) = \mathbb{E}(1_A \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)$ (cf. Lemma (2.8)). Denote the r-distance of subtrees at stage t by $\boldsymbol{de}_t^r$. By linearity of the conditional expectation we have with (4.4) at the last stage

$$
\boldsymbol{de}_{T-1} = \mathbb{E}\left[\boldsymbol{de}_T^r + \frac{1}{\lambda}\log\pi_T \middle| \mathcal{F}_{T-1}\otimes\tilde{\mathcal{F}}_{T-1}\right]^{1/r}
$$

and from calculation in backward recursive way

$$
\begin{aligned}
\boldsymbol{de}_{T-2} &= \mathbb{E}\left[\boldsymbol{de}_{T-1}^r + \frac{1}{\lambda}\log\pi_{T-1} \middle| \mathcal{F}_{T-2}\otimes\tilde{\mathcal{F}}_{T-2}\right]^{1/r} \\
&= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{de}_T^r + \frac{1}{\lambda}\log\pi_T \middle| \mathcal{F}_{T-1}\otimes\tilde{\mathcal{F}}_{T-1}\right] + \frac{1}{\lambda}\log\pi_{T-1}\middle| \mathcal{F}_{T-2}\otimes\tilde{\mathcal{F}}_{T-2}\right]^{1/r} \\
&= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{de}_T^r + \frac{1}{\lambda}\log\pi_T + \frac{1}{\lambda}\log\pi_{T-1} \middle| \mathcal{F}_{T-1}\otimes\tilde{\mathcal{F}}_{T-1}\right] \middle| \mathcal{F}_{T-2}\otimes\tilde{\mathcal{F}}_{T-2}\right]^{1/r},
\end{aligned}
$$

where we have used the tower property of the conditional expectation in (4.5). By induction and the definition of $\boldsymbol{de}_t^r$ at stage t, it follows finally that

$$\begin{aligned}
\boldsymbol{de}_0 &= \mathbb{E}\left[\boldsymbol{de}_1^r + \frac{1}{\lambda}\log \pi_1 \middle| \mathcal{F}_0 \otimes \tilde{\mathcal{F}}_0\right]^{1/r} \\
&= \mathbb{E}\left[\mathbb{E}\left[\ldots \mathbb{E}\left[\boldsymbol{de}_T^r + \frac{1}{\lambda}\log \pi_T \middle| \mathcal{F}_{T-1} \otimes \tilde{\mathcal{F}}_{T-1}\right] \ldots \middle| \mathcal{F}_1 \otimes \tilde{\mathcal{F}}_1\right] + \frac{1}{\lambda}\log \pi_1 \middle| \mathcal{F}_0 \otimes \tilde{\mathcal{F}}_0\right]^{1/r} \\
&= \mathbb{E}\left[\mathbb{E}\left[\ldots \mathbb{E}\left[\boldsymbol{de}_T^r + \frac{1}{\lambda}\sum_{t=1}^{T}\log \pi_t \middle| \mathcal{F}_{T-1} \otimes \tilde{\mathcal{F}}_{T-1}\right] \ldots \middle| \mathcal{F}_1 \otimes \tilde{\mathcal{F}}_1\right] \middle| \mathcal{F}_0 \otimes \tilde{\mathcal{F}}_0\right]^{1/r} \qquad (4.5)
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}\left[\boldsymbol{de}_T^r + \frac{1}{\lambda}\sum_{t=1}^{T}\log \pi_t \middle| \mathcal{F}_0 \otimes \tilde{\mathcal{F}}_0\right]^{1/r} \\
&= \mathbb{E}\left[\boldsymbol{de}_T^r + \frac{1}{\lambda}\sum_{t=1}^{T}\log \pi_t\right]^{1/r}, \qquad (4.6)
\end{aligned}$$

where we have used the tower property of the conditional expectation again in (4.5). The assertion (4.3) of the theorem thus follows. □

Remark 4.2 The optimization problem in Theorem 4.1 considers all constraints as the full nested problem (2.7), only the objective differs. For this reason the optimal solution of (4.3) is feasible for the problem (2.7) and vice versa.

Notice as well that the tower property can be used in a forward calculation.

Similarly to Proposition 3.5 we have the following extension to the nested Sinkhorn divergence.

Corollary 4.3 *For the nested distance and the nested Sinkhorn divergence, the same inequalities as in Proposition* 3.5 *apply, i.e.,*

$$\begin{aligned}
0 \le \boldsymbol{d}_S^r - \boldsymbol{d}^r &\le \frac{1}{\lambda}\Big(H(\pi^S) - H(\pi^W)\Big) \quad \textit{and} \quad 0 \le \boldsymbol{d}^r - \boldsymbol{de}_S^r \le \frac{1}{\lambda}H(\pi^S) \\
&\le \frac{1}{\lambda}H(p \cdot p^\top),
\end{aligned}$$

where π^S (π^W, resp.) is the optimal transport plan from (4.3) ((2.7), resp.) *with discrete, unconditional probabilities p and $\tilde{p}$ at the final stage T.*

Proof The proof follows the lines of the proof of the Propositions 3.4 and 3.5. □

Moreover, we have the following general inequality that allows an error bound depending on the total T of stages.

Corollary 4.4 *Let* m *($\tilde{m}$, resp.) be the maximum number of immediate successors in the process* $\mathbb{P}$ *($\tilde{\mathbb{P}}$, resp.), i.e.,* $m = \max\{|i+|\colon i \in \mathcal{N}_t,\ t = 1, \dots, T-1\}$*. It holds that*

$$\boldsymbol{de}_S^r - \boldsymbol{d}^r \le \frac{\log m + \log \tilde{m}}{\lambda} \cdot T, \tag{4.7}$$

where T *is the total number of stages.*

Proof Recall from Remark 3.6 that $H(\pi^S) \le \log(n\,\tilde{n}) = \log n + \log \tilde{n}$ for every conditional probability measures, where n and $\tilde{n}$ are the number of immediate successors in both trees. The result follows with $n \le m^T$ ($\tilde{n} \le \tilde{m}^T$, resp.) and $\log n \le T \log m$ and the nested program (4.1). □

4.2 Nested Sinkhorn duality

The nested distance is of importance in stochastic optimization because of its dual, which is characterized by the Kantorovich–Rubinstein theorem, cf. (2.3a)–(2.3b) above. The nested distance allows for a characterization by duality as well. Here we develop the duality for the nested Sinkhorn divergence. In line with Theorem 4.1 we need to consider the problem

$$\text{minimize}_{\text{ in }\pi} \left(\iint \left(d(\xi,\tilde{\xi})^r + \frac{1}{\lambda}\log \pi(\xi,\tilde{\xi})\right)\pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi})\right)^{1/r}$$

$$\text{subject to } \pi(A \times \tilde{\Xi} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A \mid \mathcal{F}_t), \qquad A \in \mathcal{F}_t,\ t = 1, \dots, T, \tag{4.8a}$$

$$\pi(\Xi \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \tilde{P}(B \mid \tilde{\mathcal{F}}_t), \qquad B \in \tilde{\mathcal{F}}_t,\ t = 1, \dots, T. \tag{4.8b}$$

However, we first reformulate the problem (3.9a)–(3.9b). By translating the dual variables, $\beta\hat{:}= -\beta + \mathbb{E}\,\beta$ and $\gamma\hat{:}= -\gamma + \tilde{\mathbb{E}}\gamma$, and defining $M_0 := -\mathbb{E}\,\beta - \tilde{\mathbb{E}}\gamma$ we have the alternative representation

$$\begin{aligned}
&\text{maximize}_{\text{ in } M_0}\ M_0\\
&\qquad \text{subject to } \ \mathbb{E}\,\hat{\beta} = 0,\ \tilde{\mathbb{E}}\hat{\gamma} = 0,\\
&\qquad\qquad\qquad \sum_{\xi,\tilde{\xi}} \exp\left(-\lambda\left(d(\xi,\tilde{\xi})^r - \hat{\beta}(\xi) - \hat{\gamma}(\tilde{\xi}) - M_0\right) - 1\right) = 1,\\
&\qquad\qquad\qquad \hat{\beta} \in \mathbb{R}^n,\ \hat{\gamma} \in \mathbb{R}^{\tilde{n}}.
\end{aligned}$$

To establish the dual representation of the nested distance we introduce the projections

$$\begin{aligned}
\operatorname{proj}_t\colon L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T) &\to L^1(\mathcal{F}_t \otimes \tilde{\mathcal{F}}_T)\\
\hat{\beta} \otimes \hat{\gamma} &\mapsto \mathbb{E}(\hat{\beta} \mid \mathcal{F}_t) \otimes \hat{\gamma}
\end{aligned}$$

and

$$\begin{aligned}\tilde{\operatorname{proj}}_t \colon L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T) &\to L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_t)\\ \hat{\beta} \otimes \hat{\gamma} &\mapsto \hat{\beta} \otimes \mathbb{E}(\hat{\gamma} \mid \tilde{\mathcal{F}}_t),\end{aligned}$$

where $\beta \otimes \gamma$ is the function defined by $\big(\beta \otimes \gamma\big)(\xi, \eta) := \beta(\xi) \cdot \gamma(\eta)$ and where we note that the conditional expectation is a random variable itself.

We recall the following characterization of the measurability constraints (4.8a)–(4.8b) and refer to Pflug and Pichler (2014, Proposition 2.48) for its proof.

Proposition 4.5 *The measure π satisfies the marginal condition*

$$\pi(A \times \tilde{\Xi} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A \mid \mathcal{F}_t) \quad \textit{a.s. for all } A \in \Xi$$

if and only if

$$\mathbb{E}_\pi\, \beta = \mathbb{E}_\pi \operatorname{proj}_t \beta \quad \text{for all } \beta \textit{ measurable with respect to } \mathcal{F}_T \otimes \tilde{\mathcal{F}}_T.$$

Moreover, $\operatorname{proj}_t(\beta) = \mathbb{E}_\pi(\beta \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_T)$ *if π has marginal P.*

Theorem 4.6 *The infimum or the nested distance including the entropy $\boldsymbol{de}^r(\mathbb{P}, \tilde{\mathbb{P}})$ of problem* (4.3) *equals the supremum of all numbers M_0 such that*

$$e^{-\lambda(d(\xi,\tilde{\xi})^r - M_T(\xi,\tilde{\xi}))-1} \in \mathcal{P}(\Xi \times \tilde{\Xi}), \qquad (\xi, \tilde{\xi}) \in \Xi \times \tilde{\Xi},$$

where $\mathcal{P}(\Xi \times \tilde{\Xi})$ is a set of probability measures on $(\Xi \times \tilde{\Xi})$ and M_t is an $\mathbb{R}$-valued process on $\Xi \times \tilde{\Xi}$ of the form

$$M_t = M_0 + \sum_{s=1}^{t} \hat{\beta}_s + \hat{\gamma}_s \tag{4.9}$$

and the functions $\hat{\beta}_t$, measurable with respect to $\mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}$, and $\hat{\gamma}_t$, measurable with respect to $\mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_t$, satisfy $\operatorname{proj}_{t-1}(\hat{\beta}_t) = 0$ *and* $\tilde{\operatorname{proj}}_{t-1}(\hat{\gamma}_t) = 0$.

Proof With Proposition 4.5 rewrite the dual problem as

$$\begin{aligned}&\inf_{\pi>0} \sup_{M_0, f_t, g_t} \mathbb{E}_\pi \left[d^r + \frac{1}{\lambda}\log \pi\right] + M_0 \cdot (1 - \mathbb{E}_\pi \mathbb{1}) +\\ &\quad - \sum_{s=0}^{T-1} \big(\mathbb{E}_\pi f_{s+1} - \mathbb{E}_\pi \operatorname{proj}_s(f_{s+1})\big) - \sum_{s=0}^{T-1} \big(\mathbb{E}_\pi g_{s+1} - \mathbb{E}_\pi \tilde{\operatorname{proj}}_s(g_{s+1})\big),\end{aligned}$$

where the second line encodes the measurability constraints. By the minmax theorem (cf. Sion 1958) this is equivalent to

$$\begin{aligned}\sup_{M_0, f_t, g_t} M_0 + \inf_{\pi>0} \mathbb{E}_\pi \Big[d^r &+ \frac{1}{\lambda}\log\pi - M_0\cdot \mathbb{1}\\ &- \sum_{s=0}^{T-1}\left(f_{s+1} - \operatorname{proj}_s(f_{s+1})\right) - \sum_{s=0}^{T-1}\left(g_{s+1} - \tilde{\operatorname{proj}}_s(g_{s+1})\right)\Big].\end{aligned}$$

The integral exists and the minimum is obtained by a probability measure

$$\pi = \exp\left(-\lambda\left(d^r - \sum_{s=0}^{T-1}(f_{s+1} - \operatorname{proj}_s(f_{s+1})) - \sum_{s=0}^{T-1}(g_{s+1} - \tilde{\operatorname{proj}}_s(g_{s+1}) - M_0\right) - 1\right).$$

Set $\beta_s \hat{:}= f_s - \operatorname{proj}_{s-1}(f_s)$ and $\gamma_s \hat{:}= g_s - \tilde{\operatorname{proj}}_{s-1}(g_s)$. Consequently, the problem reads

$$\begin{aligned}&\text{maximize}_{\text{ in } M_0}\ M_0\\ &\quad\text{subject to } \exp\left[-\lambda\left(d^r - \sum_{s=1}^{T}\hat{\beta}_s - \sum_{s=1}^{T}\hat{\gamma}_s - M_0\right) - 1\right] \in \mathcal{P}(\Xi\times\tilde{\Xi})\\ &\qquad\qquad \operatorname{proj}_{t-1}(\hat{\beta}_t) = 0,\ \tilde{\operatorname{proj}}_{t-1}(\hat{\gamma}_t) = 0,\end{aligned}$$

and thus the assertion. □

The following corollary links the optimal probability measure and the stochastic process (4.9) for the optimal components $\hat{\beta}$ and $\hat{\gamma}$.

Corollary 4.7 *The process M_t in (4.9), for which the supremum is attained, is a martingale with respect to the optimal measure π.*

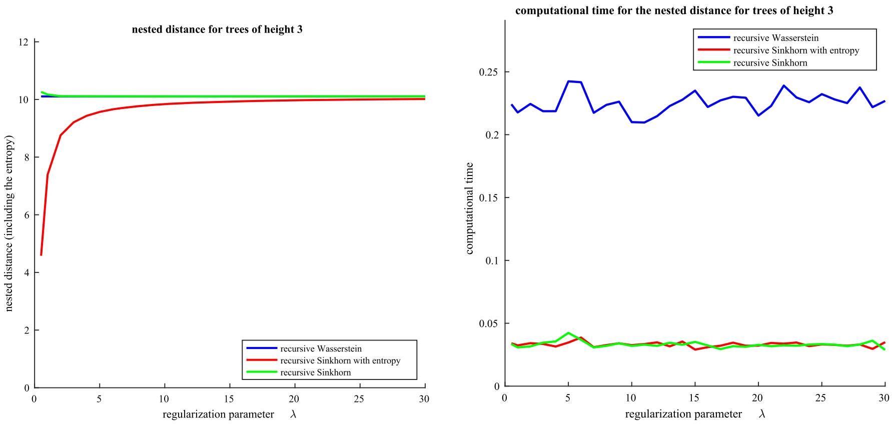
Proof The proof of Pflug and Pichler (2014, Theorem 2.49) applies with minor adaptions only. □

5 Numerical results

The nested Sinkhorn divergence $\boldsymbol{d}_S^r$ as well as $\boldsymbol{de}_S^r$ depend on the regularization parameter λ. We discuss this dependency, the error, speed of convergence and numerical issues in comparison to the non-regularized nested distance $\boldsymbol{d}^r$.

We compare Algorithms 1 and 2 with respect to the nested distance $\boldsymbol{d}^r$ and the nested Sinkhorn divergence with and without the entropy $\frac{1}{\lambda}H(\pi^S)$ as well as the required computational time for two finite valued stochastic scenario processes visualized in Fig. 3.

Figure 2 displays the results. We see that the regularized nested distance $\boldsymbol{d}_S^r$ (green) and $\boldsymbol{de}_S^r$ (red) converge to the nested distance $\boldsymbol{d}^r$ for increasing λ. In contrast to $\boldsymbol{d}_S^r$, the



(a) Nested distance (blue) and Sinkhorn divergence (green, red) for regularization parameter $\lambda \in \{0.5, 1, 2, \ldots, 30\}$

(b) computation time required for Algorithm 1 (blue) and Algorithm 2 (green, red)

Fig. 2 Results from computation of an arbitrary chosen processes given in Fig. 3 with $d(\xi_i, \tilde{\xi}_j) = |\xi_i - \tilde{\xi}_j|$ and $r = 1$

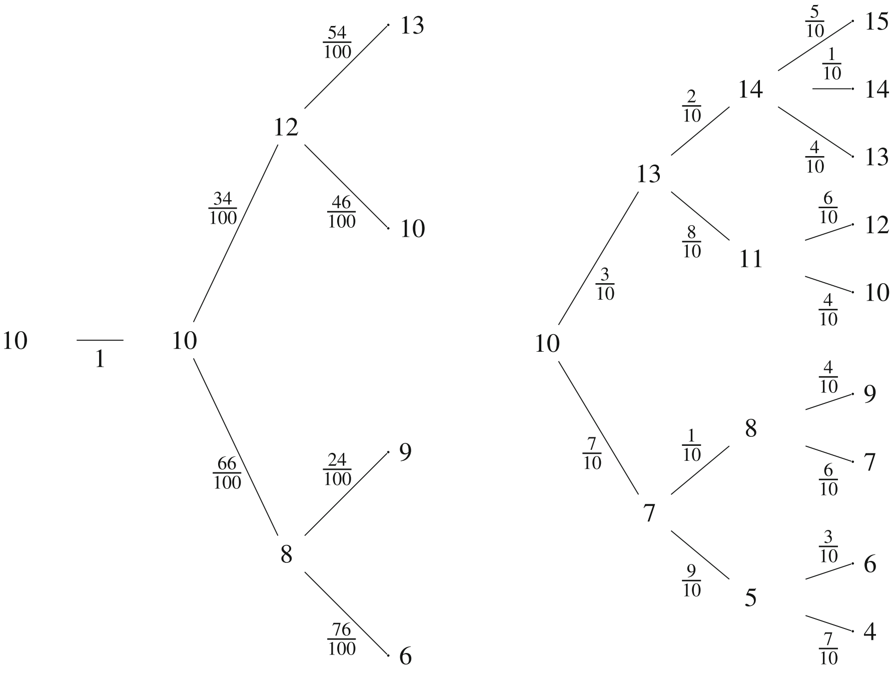


Fig. 3 Two arbitrary chosen processes with height $T = 3$

Table 1 Average distance and divergence with corresponding computational time in seconds on i5-3210M CPU

Stages	Wasserstein		Sinkhorn			Difference	Time
T	$\boldsymbol{d}^r$	Time	$\boldsymbol{d}^r_S$	$\boldsymbol{de}^r_S$	Time	$\boldsymbol{d}^r - \boldsymbol{de}^r_S$	Acceleration
1	1.8	0.06 s	1.81	1.75	0.006 s	0.06	10×
2	5.1	0.13 s	5.12	4.97	0.022 s	0.14	5.8×
3	5.8	0.50 s	5.81	5.66	0.062 s	0.15	8.1×
4	7.3	1.54 s	7.32	7.08	0.368 s	0.24	4.2×
5	10.1	10.29 s	10.05	9.72	2.873 s	0.35	3.6×

All states and probabilities are generated randomly. The regularization parameter is $\lambda = 20$ and $r = 1$

regularized nested distance including the entropy converges slower to $\boldsymbol{d}^r$. The reason is that for larger λ the weight of the entropy in the cost function in (3.1a) decreases and the entropy of π^S and π^W coincide (cf. (4.7)). Computing the distances with Sinkhorn's algorithm in recursive way, in contrast to solving the linear problem for the Wasserstein distance, is about six times faster. In addition, the required time for the regularized nested distance with and without the entropy varies much less by contrast with the computational time for the nested distance. Furthermore, the differences between $\boldsymbol{d}^r$ and $\boldsymbol{d}^r_S$ and $\boldsymbol{de}^r_S$, respectively, is rapidly decreasing and insignificant for $\lambda > 20$. Moreover, the time displayed in Fig. 2b does not depend on the regularization parameter λ.

The following two examples illustrate the computational accelerations.

Example 5.1 We now fix $\lambda = 20$ and vary the stages $T \in \{1, 2, 3, 4, 5\}$. The first finite tree has the branching structure [1 2 3 2 3 4] and the second tree has a simpler structure [1 2 2 1 3 2] (i.e., the first tree has 144 leaf nodes and the second tree 24). All states and probabilities in the trees are generated randomly.

Table 1 summarizes the results collected. We notice that the Sinkhorn algorithm is up to 10 times faster compared with the usual Wasserstein distance, although the speed advantage decreases for larger trees. The Sinkhorn algorithm also leads to small errors which increase marginally for trees with more stages.

Example 5.2 To provide an additional performance comparison we fix $\lambda = 20$ and vary $T \in \{1, 2, 3, 4, 5, 6\}$. The first finite tree has the structure [1 4 5 3 4 4 6] and the second tree [1 2 2 1 3 2 3]. This means that the first tree has 5760 leaf nodes while the second tree has only 72. All states and probabilities in the trees are generated randomly. Table 2 summarizes the results are summarized

Additionally, we tried to improve the speed by modifying the recursive algorithm. Instead of computing once from $T - 1$ down to 0 we computed from $T - 1$ down to 0 several times to achieve a convergence in the optimal transport plan π^S. This approach has no advantages.

Remark 5.3 The entries k_{ij} of the matrix (3.12) are small for $d_{ij} \neq 0$, particularly for $\lambda \gg 1$ (i.e., λ large) and $r \gg 1$. In this case, the entries of the vectors $\tilde{\beta}$ and $\tilde{\gamma}$

Table 2 Average distance and divergence (cf. Table 1)

Stages	Wasserstein		Sinkhorn			Difference	Time
T	d^r	Time	d_S^r	de_S^r	Time	$d^r - de_S^r$	Acceleration
1	2.3	0.05 s	2.3	2.2	0.008 s	0.09	5.8x
2	5.6	0.2 s	5.6	5.4	0.060 s	0.19	3.2x
3	6.3	1.4 s	6.3	6.1	0.233 s	0.22	5.8x
4	7.7	6.2 s	7.7	7.4	0.202 s	0.33	3.1x
5	11.4	58 s	11.3	10.8	17.16 s	0.58	3.4x
6	13.4	717 s	10.2	9.7	347.5 s	3.75	2.0x

Table 3 The nested distance and divergence for two trees with 3 stages and leaves and branching 2. The regularization parameter is $\lambda = 10$

Stages	Wasserstein	Sinkhorn	
Order r	d^r	d_S^r	de^r
1	0.8615	0.8737	0.6194
2	1.0532	1.0554	0.9348
3	1.2021	1.2026	1.1454
4	1.3188	1.3189	1.2925
5	1.4134	1.4134	1.4014

in Algorithm 2 can grow extraordinary high. For this reason, rescaling the vectors is necessary. Further, an adequate balance between λ, modelling the approximation quality, and acceleration desired is crucial in real applications. See also Remark 3.11 for the same issue.

Example 5.4 Table 3 investigates the approximation quality for varying orders r. The trees compared have 3 stages and each node branches into two directions. For large $r \gg 1$ it is important to recall Remark 5.3 here, but on the other side the approximation quality improves for increasing order r.

6 Summary

Stochastic processes with information evolving in finitely many stages and finitely many states are encoded in stochastic trees. The nested distance, which builds on the Wasserstein distance, allows distinguishing stochastic processes and stochastic trees.

In this paper we regularize the Wasserstein distance by employing the Sinkhorn divergence. This approach extends to multiple stages and allows introducing a nested Sinkhorn divergence for stochastic processes. We elaborate its properties and describe the accelerations, which can be achieved in this way.

In conclusion, we can summarize that the Sinkhorn divergence offers a good trade-off between the regularization error and the speed advantage. Further work should focus on defining a (nested) distance for neuronal networks and extending the implementation of Sinkhorn divergence in the Julia package for faster tree generation and computation.

Acknowledgements We are thankful to Benoît Tran for pointing out further references, particularly on convergence, in his thesis Tran (2020) supervised by Marianne Akian and Jean-Philippe Chancelier.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Altschuler J, Weed J, Rigollet P (2017) Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In: Proceedings of the 31st international conference on neural information processing systems, pp 1961–1971. Curran Associates Inc., arxiv:1705.09634

Analui B, Pflug GCh (2014) On distributionally robust multiperiod stochastic optimization. Comput Manag Sci 11(3):197–220. https://doi.org/10.1007/s10287-014-0213-y

Bachem A, Korte B (1979) On the RAS-algorithm. Computing 23(2):189–198. https://doi.org/10.1007/bf02252097

Beltrán F, de Oliveira W, Finardi EC (2017) Application of scenario tree reduction via quadratic process to medium-term hydrothermal scheduling problem. IEEE Trans Power Syst 32(6):4351–4361. https://doi.org/10.1109/tpwrs.2017.2658444

Bertsekas DP, Castanon DA (1989) The auction algorithm for the transportation problem. Ann Oper Res 20(1):67–96. https://doi.org/10.1007/bf02216923

Bigot J, Cazelles E, Papadakis N (2019) Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. Electron J Stat 13(2):5120–5150. https://doi.org/10.1214/19-EJS1637

Brodt AI (1983) Min-mad life: a multi-period optimization model for life insurance company investment decisions. Insurance Math Econ 2(2):91–102

Carpentier P, Chancelier J-P, Cohen G, De Lara M, Girardeau P (2012) Dynamic consistency for stochastic optimal control problems. Ann Oper Res 200(1):247–263. https://doi.org/10.1007/s10479-011-1027-8

Carpentier P, Chancelier J-P, Cohen G, De Lara M (2015) Stochastic multi-stage optimization. Springer International Publishing, Berlin. https://doi.org/10.1007/978-3-319-18138-7

Cuturi M (2013) Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in neural information processing systems

Edirisinghe NCP (2005) Multiperiod portfolio optimization with terminal liability: bounds for the convex case. Comput Optim Appl 32(1–2):29–59. https://doi.org/10.1007/s10589-005-2053-8

Genevay A, Peyré G, Cuturi M (2018) Learning generative models with Sinkhorn divergences. In: Storkey A, Perez-Cruz F (eds) Proceedings of the twenty-first international conference on artificial intelligence and statistics, volume 84 of proceedings of machine learning research, pp 1608–1617. PMLR http://proceedings.mlr.press/v84/genevay18a.html

Heitsch H, Römisch W, Strugarek C (2006) Stability of multistage stochastic programs. SIAM J Optim 17(2):511–525

Horejšová M, Vitali S, Kopa M, Moriggia V (2020) Evaluation of scenario reduction algorithms with nested distance. CMS 17(2):241–275. https://doi.org/10.1007/s10287-020-00375-4

Kirui KB, Pichler A, Pflug GCh (2020) ScenTrees.jl: a Julia package for generating scenario trees and scenario lattices for multistage stochastic programming. J Open Source Softw 5(46):1912. https://doi.org/10.21105/joss.01912

Kolouri S, Park SR, Thorpe M, Slepcev D, Rohde GK (2017) Optimal mass transport: signal processing and machine-learning applications. IEEE Signal Process Mag 34(4):43–59. https://doi.org/10.1109/msp.2017.2695801

Kovacevic RM, Pichler A (2015) Tree approximation for discrete time stochastic processes: a process distance approach. Ann Oper Res. https://doi.org/10.1007/s10479-015-1994-2

Kruithof R (1937) Telefoonverkeersrekening. De Ingenieur 52:E15-E25. https://wwwhome.ewi.utwente.nl/ptdeboer/misc/kruithof-1937-translation.html

Luise G, Rudi A, Pontil M, Ciliberto C (2018) Differential properties of Sinkhorn approximation for learning with Wasserstein distance. In: Advances in neural information processing systems 31 (NIPS 2018). arXiv:1805.11897

Maggioni F, Pflug GCh (2019) Guaranteed bounds for general non-discrete multistage risk-averse stochastic optimization programs. SIAM J Optim 29(1):454–483. https://doi.org/10.1137/17M1140601

Peyré G, Cuturi M (2019) Computational optimal transport: with applications to data science. Found Trends® Mach Learn. 11(5-6):355–607. https://doi.org/10.1561/2200000073

Pflug GCh (2009) Version-independence and nested distributions in multistage stochastic optimization. SIAM J Optim 20:1406–1420. https://doi.org/10.1137/080718401

Pflug ChG, Pichler A (2012) A distance for multistage stochastic optimization models. SIAM J Optim 22(1):1–23. https://doi.org/10.1137/110825054

Pflug GCh, Pichler A (2014) Multistage stochastic optimization. Springer series in operations research and financial engineering. Springer, Berlin. ISBN 978-3-319-08842-6. https://doi.org/10.1007/978-3-319-08843-3. https://books.google.com/books?id=q_VWBQAAQBAJ

Rachev ST (1991) Probability metrics and the stability of stochastic models. Wiley, West Sussex

Rachev ST, Rüschendorf L (1998) Mass transportation problems volume I: theory, volume II: applications, volume XXV of Probability and its applications. Springer, New York. https://doi.org/10.1007/b98893

Rote G, Zachariasen M (2007) Matrix scaling by network flow. In: Bansal N, Pruhs K, Stein C (eds) Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7–9. SIAM, pp 848–854. http://dl.acm.org/citation.cfm?id=1283383.1283474

Rüschendorf L (1995) Convergence of the iterative proportional fitting procedure. Ann Stat 23(4):1160–1174. https://doi.org/10.1214/aos/1176324703

Sinkhorn R (1967) Diagonal equivalence to matrices with prescribed row and column sums. Am Math Mon 74(4):402. https://doi.org/10.2307/2314570

Sinkhorn R, Knopp P (1967) Concerning nonnegative matrices and doubly stochastic matrices. Pacific J Math 21:343–348

Sion M (1958) On general minimax theorems. Pacific J Math 8(1):171–176

Tran DNB (2020) Programmation dynamique tropicale en optimisation stochastique multi-étapes. Ph.D. thesis, Université Paris-Est

Villani C (2009) Optimal transport, old and new, vol. 338. Grundlehren der Mathematischen Wissenschaften. Springer, Berlin