Nested Sinkhorn Divergence To Compute
The Nested Distance

Alois Pichler*† Michael Weinhardt*

February 11, 2021

Abstract

The nested distance builds on the Wasserstein distance to quantify the difference of stochastic processes, including also the 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
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 [24] provide a comprehensive discussion of the Wasserstein distance and Villani [31] summarizes the optimal transport theory.

The nested distance is based on the Wasserstein distance. It has been introduced by Pflug [20] and generalizes and extends the theory from probability measures to stochastic processes, cf. Pflug and Pichler [21].

The nested distance is employed in multistage stochastic programming to describe the quality of an approximation. Multistage stochastic programming has applications in many sectors, e.g., the financial sector (Edirisinghe [10], Brodt [6]), in management science or in energy economics (Analui and Pflug [1], Beltrán et al. [3], Carpentier et al. [7, 8]). The prices, demands, etc., are
often modeled as a stochastic process $\xi = (\xi_0, \ldots, \xi_T)$ and the optimal values are rarely obtained analytically. For the numerical approach the stochastic process is replaced by a finite valued stochastic scenario process $\tilde{\xi} = (\tilde{\xi}_0, \ldots, \tilde{\xi}_T)$, which is a finite tree. Naturally, the approximation error should be minimized without unnecessarily increasing the complexity of the computational effort. Kirui et al. [14] provide a Julia package for generating scenario trees and scenario lattices for multistage stochastic programming. Maggioni and Pflug [19] provide guaranteed bounds and Horejšová et al. [13] 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 Bertsekas and Castanon [4], 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 [27] that shows a unique decomposition for non-negative matrices and ensures convergence of the associated iterative scheme. Cuturi [9] 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. [5] and Luise et al. [18], for image recognition and machine learning, cf. Kolouri et al. [15] and Genevay et al. [11], among many other applications.

Extending Sinkhorn’s algorithm to multistage stochastic programming has been proposed recently in Tran [30, 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. As well, 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 Section 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 Section 4 we regularize the nested distance and show the equality between two different approaches. Results and comparisons are visualized and discussed in Section 5. Section 6 summarizes and concludes the paper.

2 Preliminaries

This section recalls the definition of the nested distance and provides an example to illustrate the importance of the filtration. Throughout, we shall work 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 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 \( \Xi \). A function \( d: \mathcal{P} \times \mathcal{P} \to [0, \infty) \) 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) Strictness: if \( d(P_1, P_2) = 0 \), then \( P_1 = P_2 \).

Rachev [23] 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 \( \Xi \) is equipped with a metric \( d: \Xi \times \Xi \to \mathbb{R} \) so that \( (\Xi, d) \) is Polish.

**Definition 2.2** (Wasserstein distance). Let \( P \) and \( \tilde{P} \) be two probability measure on \( \Xi \) endowed with a distance \( d: \Xi \times \Xi \to \mathbb{R} \). The *Wasserstein distance* of order \( r \geq 1 \) is
\[
    d^r(P, \tilde{P}) := \inf_{\pi} \int_{\Xi \times \Xi} d(\xi, \tilde{\xi})^r \pi(d\xi, d\tilde{\xi}),
\]
where the infimum is over all probability measures \( \pi \) 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 \( \Xi \) and the Wasserstein distance is a distance on \( \mathcal{P} \) in the sense of Definition 2.1 above. However, in what follows any cost function \( c: \Xi \times \Xi \to \mathbb{R} \) could be considered instead of the distance \( d \) on \( \Xi \) (of course, \( c \) has to be measurable and the integral has to exist). The result might not be a distance in the sense of Definition 2.1. In what follows we will point to the differences.

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 \( \{\xi_i: i = 1, 2, \ldots, n\} \subset \Xi \) is finite. The Wasserstein distance \( d^r \) of two discrete measures \( P = \sum_{i=1}^n p_i \delta_{\xi_i} \) and \( \tilde{P} = \sum_{j=1}^n \tilde{p}_j \delta_{\tilde{\xi}_j} \) is the \( r \)-th root of the optimal
value of

\[
\begin{align*}
\text{minimize in } \pi & \sum_{i=1}^{n} \sum_{j=1}^{\tilde{n}} \pi_{ij} d_{ij}^l \\
\text{subject to} & \sum_{j=1}^{\tilde{n}} \pi_{ij} = p_i, \quad i = 1, \ldots, n, \\
& \sum_{i=1}^{n} \pi_{ij} = \tilde{p}_j, \quad j = 1, \ldots, \tilde{n} \\
& \pi_{ij} \geq 0,
\end{align*}
\]

(2.1)

where \(d_{ij} := d(\xi_i, \tilde{\xi}_j)\) is an \(n \times \tilde{n}\)-matrix collecting all distances. The optimal measure in (2.1) is denoted \(\pi^W\) and called an optimal transport plan. The convex, linear dual of (2.1) is

\[
\begin{align*}
\text{maximize in } \lambda \text{ and } \mu & \sum_{i=1}^{n} p_i \lambda_i + \sum_{j=1}^{\tilde{n}} \tilde{p}_j \mu_j \\
\text{subject to} & \lambda_i + \mu_j \leq d_{ij}^l \quad \text{for all } i = 1, \ldots, n \text{ and } j = 1, \ldots, \tilde{n}.
\end{align*}
\]

(2.2a)

(2.2b)

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

\[
\begin{align*}
\text{minimize in } x & \ c^T x \\
\text{subject to} & \ Ax = b, \\
& \ x \geq 0,
\end{align*}
\]

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

\[
A = \begin{pmatrix}
\mathbb{I}_n \otimes I_n \\
I_{\tilde{n}} \otimes \mathbb{I}_n
\end{pmatrix}
\]

with \(\mathbb{I} = (1, \ldots, 1)\).

2.2 The distance of stochastic processes

Let \((\Xi, \mathcal{F}, P)\) and \((\tilde{\Xi}, \tilde{\mathcal{F}}, \tilde{P})\) 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 \((\Xi, d)\) is a metric space. Without loss of generality we may set \(\Xi_t = \mathbb{R}\) for all \(t \in \{0, 1, \ldots, T\}\) and employ the \(\ell^1\)-distance, i.e., \(d(\xi, \tilde{\xi}) = \sum_{t=0}^{T} |\xi_t - \tilde{\xi}_t|\).

Remark 2.5. The example depicted in Figure 1 illustrates that a simple application of the Wasserstein distance does not capture the different information (knowledge) available at the intermediate stage. Indeed, let \(\epsilon > 0\). The distance matrix of the trajectories is

\[
d = \begin{pmatrix}
\epsilon & 2 + \epsilon \\
2 & 0
\end{pmatrix}
\]
Figure 1: Two processes illustrating two different flows of information, cf. Heitsch et al. [12], Kovacevic and Pichler [16]

and the optimal transport plan is

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

It follows the Wasserstein distance according (2.1) is

\[ d = \sum_{i,j} d_{ij} \pi_{ij} = \varepsilon/2. \]

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 \( t = 0, 1, \ldots, T - 1 \), but only probabilities at the final stage \( t = T \), where all the information from intermediate stages are ignored. The information at the previous stage is encoded by the \( \sigma \)-algebra

\[ \mathcal{F}_t = \sigma \left( A_1 \times \cdots \times A_t \times \Xi_{t+1} \times \cdots \times \Xi_T : A_i \subset \Xi_i \text{ measurable} \right) \]

for \( t = 0, 1, \ldots, T \) (\( \hat{\mathcal{F}}_t \), resp.). The following generalization of the Wasserstein distance takes all conditional probabilities into account.

**Definition 2.6** (The nested distance). The nested distance of order \( r \geq 1 \) of two filtered probability spaces \( \mathbb{P} = (\Xi, (\mathcal{F}_t), \mathbb{P}) \) and \( \mathbb{P} = (\Xi, (\hat{\mathcal{F}}_t), \mathbb{P}) \), for which a distance \( d : \Xi \times \Xi \to \mathbb{R} \) is defined, is the optimal value of the optimization problem

\[
\begin{align*}
\text{minimize}_{\pi} & \quad \left( \int_{\Xi \times \Xi} d(\xi, \hat{\xi})^r \pi(d\xi, d\hat{\xi}) \right)^{1/r} \\
\text{subject to} & \quad \pi(A \times \Xi) \mid \mathcal{F}_t \otimes \hat{\mathcal{F}}_t = P(A) \mid \mathcal{F}_t, \quad A \in \mathcal{F}_t, \ t = 1, \ldots, T, \\
& \quad \pi(\Xi \times B) \mid \mathcal{F}_t \otimes \hat{\mathcal{F}}_t = \tilde{P}(B) \mid \hat{\mathcal{F}}_t, \quad B \in \hat{\mathcal{F}}_t, \ t = 1, \ldots, T,
\end{align*}
\]

where the infimum in (2.3) is among all bivariate probability measures \( \pi \in \mathcal{P}(\Xi \times \Xi) \) defined on \( \mathcal{F}_T \otimes \hat{\mathcal{F}}_T \). The optimal value of (2.1), the nested distance of order \( r \), is denoted by \( d^r(\mathbb{P}, \tilde{\mathbb{P}}) \).

For the discrete nested distance we use trees to model the whole space and filtration. We denote by \( \mathcal{N}_t \) (\( \hat{\mathcal{N}}_t \), resp.) 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 \( m < i \). The nested distance for
trees is the $r$-th root of the optimal value of
\begin{align}
\text{minimize} \quad & \sum_{i,j} \pi_{ij} \cdot d_{ij} \\
\text{subject to} \quad & \sum_{j \in J} \pi(i,j | i_t,j_t) = P(i | i_t), \quad i_t < i, j_t, \\
& \sum_{i \in I} \pi(i,j | i_t,j_t) = \tilde{P}(j | j_t), \quad j_t < j, i_t, \\
& \pi_{ij} \geq 0 \text{ and } \sum_{i,j} \pi_{ij} = 1,
\end{align}

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 on the same stage $t$. As usual for discrete measures, the conditional probabilities $\pi(i,j | i_t, j_t)$ are given by
\begin{align}
\pi(i,j | i_t, j_t) := \frac{\pi_{ij}}{\sum_{i' > i_t, j' > j_t} \pi_{i'j'}}.
\end{align}

**Remark 2.7.** Employing the definition (2.5) for $\pi(i,j | i_t, j_t)$ reveals that the problem (2.4) is indeed a linear program in $\pi$ (cf. (2.1)).

### 2.3 Rapid, nested computation of the process distance

This subsection addresses an advanced approach for solving the linear program (2.4). We first recall the tower property, which allows an important simplification of the constraints in (2.3).

**Lemma 2.8.** To compute the nested distance it is enough to condition on the immediately following $\sigma$-algebra: the conditions
\begin{align}
\pi(A \times \Xi | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \text{ for all } A \in \mathcal{F}_T
\end{align}
in (2.3) may be replaced by
\begin{align}
\pi(A \times \Xi | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \text{ for all } A \in \mathcal{F}_{t+1}.
\end{align}

**Proof.** The proof is based on the tower property of the expectation and can be found in [22, Lemma 2.43].

As a result of the tower property the full problem (2.4) can be calculated faster in a recursive way and the matrix for the constraints has not to be stored. We employ this result in an algorithm blow. For further details we refer to Pflug and Pichler [22, 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.

6
Algorithm 1: Nested computation of the nested distance $d^r(P, \tilde{P})$ of two tree-processes $P$ and $\tilde{P}$

**Input:** for all combinations of leaf nodes $i \in N_T$ and $j \in \tilde{N}_T$ with predecessors $(i_0, i_1, \ldots, i_{T-1}, i)$ and $(j_0, j_1, \ldots, j_{T-1}, j)$ set

$$d^r_T(i, j) := d((\xi_{i_0}, \xi_{i_1}, \ldots, \xi_i), (\xi_{j_0}, \xi_{j_1}, \ldots, \xi_j))$$

**Output:** the optimal transport plan at the leaf nodes $i \in N_T$ and $j \in \tilde{N}_T$ is

$$\pi(i, j) = \pi_1(i_1 | i_0, j_0) \cdots \pi_{T-1}(i, j | i_{T-1}, j_{T-1}).$$

for $t = T - 1$ down to $0$ and every combination of inner nodes $i \in N_t$ and $j \in \tilde{N}_t$ do

solve the linear programs

$$\begin{align*}
\text{minimize}_{\pi} & \quad \sum_{i' \in i_t^+, j' \in j_t^+} \pi(i', j' | i_t, j_t) \cdot d^r_{t+1}(i', j') \\
\text{subject to} & \quad \sum_{j' \in j_t^+} \pi(i', j' | i_t, j_t) = p(i | i_t), \quad i' \in i_t^+, \\
& \quad \sum_{i' \in i_t^+} \pi(i', j' | i_t, j_t) = \tilde{p}(j' | j_t), \quad j' \in j_t^+, \\
& \quad \pi(i', j' | i_t, j_t) \geq 0
\end{align*}$$

and denote its optimal value by $d^r_t(i_t, j_t)$.

**Result:** The nested distance is $d^r(P, \tilde{P}) := d^r_0(0, 0)$

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

**Definition 3.2 (Sinkhorn divergence).** The **Sinkhorn divergence** is obtained by the optimization problem

$$\begin{align*}
\text{minimize}_{\pi} & \quad \sum_{i,j} \pi_{ij} \frac{d_{ij}^2}{\lambda} - \frac{1}{\lambda} H(\pi) \\
\text{subject to} & \quad \sum_{j} \pi_{ij} = p_i, \quad i = 1, \ldots, n, \\
& \quad \sum_{i} \pi_{ij} = \tilde{p}_j, \quad j = 1, \ldots, \tilde{n}, \\
& \quad \pi_{ij} > 0 \quad \text{for all } i, j
\end{align*}$$

where $d$ is a distance or a cost matrix and $\lambda > 0$ is a regularization parameter. With $\pi^S$ being the
optimal transport in (3.1a)–(3.1b) we denote the Sinkhorn divergence by
\[ d^r_S := \sum_{i,j} \pi_{ij}^S d^r_{ij} \]
and the Sinkhorn divergence including the entropy by
\[ de^r_S := \sum_{i,j} \pi_{ij}^S d^r_{ij} - \frac{1}{\lambda} H(\pi^S). \]

We may mention here that we avoid the term Sinkhorn distance since for all \( \lambda \neq 0 \) the Sinkhorn divergence \( d^r_S \) is strictly positive and \( de^r_S \) can be negative for small \( \lambda \) which violates the axioms of a distance given in Definition 2.1 above (particularly (i), (iii) and (iv)). Strict positivity of \( d^r_S \) can be forced by a correction term, the so-called Sinkhorn Loss (see Bigot et al. [5, Definition 2.3]) or by employing the cost matrix \( d \cdot 1_{p \neq \tilde{p}} \) instead.

**Remark 3.3.** The strict inequality constraint (3.1b) is not a restriction. Indeed, the mapping \( \varphi(x) \) 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
\[ de^r_S \leq d^r_W \leq d^r_S. \] (3.2)

*Proof.* Recall that \( \pi \log \pi \leq 0 \) for all \( \pi \leq 1 \) and thus it holds that \( \sum_{i,j} \pi_{ij} d^r_{ij} + \frac{1}{\lambda} \sum_{i,j} \pi_{ij} \log \pi_{ij} \leq \sum_{i,j} \pi_{ij} d^r_{ij} \), for all \( \pi \in (0, 1]^{n \times n} \). It follows that
\[ \min_{\pi} \sum_{i,j} \pi_{ij} d^r_{ij} + \frac{1}{\lambda} \sum_{i,j} \pi_{ij} \log \pi_{ij} \leq \min_{\pi} \sum_{i,j} \pi_{ij} d^r_{ij} \]
and thus the first inequality. The remaining inequality is clear by the definition of the Wasserstein distance. \( \square \)

Both Sinkhorn divergences \( d^r_S \) and \( de^r_S \) approximate the Wasserstein distance \( d^r_W \), and we have convergence for \( \lambda \to \infty \) to \( d^r_W \). The following proposition provides precise bounds.

**Proposition 3.5.** For every \( \lambda > 0 \) we have
\[ 0 \leq d^r_S - d^r_W \leq \frac{1}{\lambda} \left(H(\pi^S) - H(\pi^W)\right) \] (3.3)
and
\[ 0 \leq d^r_W - de^r_S \leq \frac{1}{\lambda} H(\pi^S) \leq \frac{1}{\lambda} H(p \cdot \tilde{p}^\top) \] (3.4)
with \( p = (p_1, \ldots, p_n) \) and \( \tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_n) \), respectively.
**Proof.** The first inequalities follow from (3.2) and from optimality of \( \pi^S \) in the inequality
\[
d^r_S - \frac{1}{\lambda} H(\pi^S) \leq d^w_S - \frac{1}{\lambda} H(\pi^W).
\]
The latter again with (3.2) and \( d^r_S - d^w_S = \frac{1}{\lambda} H(\pi^S) \). Finally, by the log sum inequality, \( H(\pi) \leq H(p \cdot \bar{p}^\top) \) for every measure \( \pi \) with marginals \( p \) and \( \bar{p} \).

**Remark 3.6.** As a consequence of the log sum inequality we obtain as well that \( \log n \leq \log 2 \). 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 definitions of the Sinkhorn divergence which we want to mention here. The first alternative definition involves the Kullback–Leibler divergence \( D_{KL}(\pi \mid P \otimes \bar{P}) \), which is defined as
\[
D_{KL}(\pi \mid P \otimes \bar{P}) := -\sum_{i,j} \pi_{ij} \log \frac{\pi_{ij}}{p_i \bar{p}_j} = H(P) + H(\bar{P}) - H(\pi),
\]
where the latter equality is justified provided that \( \pi \) has marginal measures \( P \) and \( \bar{P} \). The Sinkhorn divergence (in the alternative definition) is the \( r \)-th root of the optimal value of
\[
\begin{align*}
\text{minimize}_{\pi} & \quad \sum_{i,j} \pi_{ij} d^r_{ij} \\
\text{subject to} & \quad \sum_{j} \pi_{ij} = p_i, \quad i = 1, \ldots, n, \\
& \quad \sum_{i} \pi_{ij} = \bar{p}_j, \quad j = 1, \ldots, \tilde{n}, \\
& \quad \pi_{ij} > 0 \quad \text{and} \\
& \quad D_{KL}(\pi \mid P \otimes \bar{P}) \leq \alpha \quad \text{for all} \ i, j,
\end{align*}
\]
where \( \alpha \geq 0 \) is the regularization parameter. For each \( \alpha \) in (3.5b) we have by the duality theory a corresponding \( \lambda \) in (3.1a) such that the optimal values coincide. Let \( \alpha > 0 \) and \( \pi^{KL} \) be the solution to problem (3.5a)–(3.5b) with Lagrange multipliers \( \beta \) and \( \gamma \). Then the optimal value of problem (3.5a) equals \( d^r_S \) from (3.1a) with
\[
\lambda = \frac{\log(\pi^{KL}_{ij}) + 1}{d_{ij} + \beta_i + \gamma_j}
\]
for any \( i \in \{1, \ldots, n\} \) and \( j \in \{1, \ldots, \tilde{n}\} \). For further information and illustration we refer to Cuturi [9, Section 3].

A further, potential definition employs a different entropy regularization is given by
\[
\tilde{H}(\pi) = -\sum_{i,j} \pi_{ij} \cdot (\log \pi_{ij} - 1).
\]
Luise et al. [18] use this definition for Sinkhorn approximation for learning with Wasserstein distance and proof an exponential convergence. This definition leads to a similar matrix decomposition and iterative algorithm 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

\[ L(\pi; \beta, \gamma) := \sum_{i,j} \pi_{ij} d_{ij} + \frac{1}{\lambda} \sum_{i,j} \pi_{ij} \log \pi_{ij} + \beta^T (p - \pi \cdot 1) + (\bar{p} - \pi^T \cdot \pi)^T \gamma \]  

(3.6)

of the problem (3.2). The partial derivatives are

\[ \frac{\partial L}{\partial \pi_{ij}} = \frac{1}{\lambda} (\log \pi_{ij} + 1) + d_{ij} - \beta_i - \gamma_j = 0, \]  

(3.7)

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

\[ \pi_{ij}^* = \exp \left( -\lambda (d_{ij} - \beta_i - \gamma_j) - 1 \right) \]

(3.8)

By inserting \( \pi_{ij}^* \) in the Lagrangian function \( L \) we get the convex dual function

\[ d(\beta, \gamma) := \inf_{\pi} L(\pi; \beta, \gamma) = L(\pi^*; \beta, \gamma) \]

\[ = \sum_{i,j} d_{ij} \cdot e^{-\lambda(d_{ij} - \beta_i - \gamma_j)} - \frac{1}{\lambda} \sum_{i,j} e^{-\lambda(d_{ij} - \beta_i - \gamma_j)} \cdot (e^{-\lambda(d_{ij} - \beta_i - \gamma_j)} + 1) \]

\[ + \sum_i \beta_i \left( p_i - \sum_j e^{-\lambda(d_{ij} - \beta_i - \gamma_j)} \right) + \sum_j \gamma_j \left( \bar{p}_j - \sum_i e^{-\lambda(d_{ij} - \beta_i - \gamma_j)} \right) \]

\[ = - \sum_{i,j} \left( \beta_i + \gamma_j + \frac{1}{\lambda} \right) e^{-\lambda(d_{ij} - \beta_i - \gamma_j)} + \sum_i \beta_i p_i + \sum_j \gamma_j \bar{p}_j \]

\[ - \sum_i \beta_i \left( \sum_j e^{-\lambda(d_{ij} - \beta_i - \gamma_j)} \right) - \sum_j \gamma_j \left( \sum_i e^{-\lambda(d_{ij} - \beta_i - \gamma_j)} \right) \]

\[ = \sum_i \beta_i p_i + \sum_j \gamma_j \bar{p}_j - \frac{1}{\lambda} \sum_{i,j} e^{-\lambda(d_{ij} - \beta_i - \gamma_j)}. \]

The dual problem thus is

maximize in \( \beta, \gamma \)

\[ \sum_i \beta_i p_i + \sum_j \gamma_j \bar{p}_j - \frac{1}{\lambda} \sum_{i,j} e^{-\lambda(d_{ij} - \beta_i - \gamma_j)}. \]

subject to \( \beta \in \mathbb{R}^n, \ \gamma \in \mathbb{R}^\ell. \)

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

maximize in \( \beta, \gamma \)

\[ \sum_i \beta_i p_i + \sum_j \gamma_j \bar{p}_j \]

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

(3.9a)

(3.9b)
Remark 3.7. We deduce from (3.9b) that \(-\lambda \left( d_{ij} - \beta_i - \gamma_j \right) - 1 \leq 0\), or
\[
\beta_i + \gamma_j \leq d_{ij} + \frac{1}{\lambda} \quad \text{for all } i, j
\] (3.10)
provided that \(\lambda > 0\). It is thus apparent that (3.9a)–(3.9b) is a relaxation of problem (2.2a)–(2.2b) 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 \(\pi^*\) in (3.8), the gradients are
\[
\frac{\partial L}{\partial \beta_i} = p_i - \sum_{j=1}^{n} \pi_{ij} = p_i - \sum_{j=1}^{n} e^{-\lambda (d_{ij} - \beta_i - \gamma_j)} = 0
\] (3.11)
and
\[
\frac{\partial L}{\partial \gamma_j} = \tilde{p}_j - \sum_{i=1}^{n} \pi_{ij} = \tilde{p}_j - \sum_{i=1}^{n} e^{-\lambda (d_{ij} - \beta_i - \gamma_j)} = 0
\] (3.12)
so that the equations
\[
\beta_i = \frac{1}{\lambda} \log \left( \frac{p_i}{\sum_{j=1}^{n} e^{-\lambda (d_{ij} - \gamma_j)}} \right) \quad \text{and} \quad \gamma_j = \frac{1}{\lambda} \log \left( \frac{\tilde{p}_j}{\sum_{i=1}^{n} e^{-\lambda (d_{ij} - \beta_i)}} \right)
\]
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
\[
\tilde{\beta}_i = \frac{p_i}{\sum_{j=1}^{n} e^{-\lambda d_{ij} \tilde{\gamma}_j}} \quad \text{and} \quad \tilde{\gamma}_j = \frac{\tilde{p}_j}{\sum_{i=1}^{n} \tilde{\beta}_i e^{-\lambda d_{ij}}},
\] (3.13)
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 (3.13) can be used to determine \(\tilde{\beta}\) and \(\tilde{\gamma}\) alternately. Indeed, from (3.11) and (3.12) we infer that \(\pi^*\) is a doubly stochastic matrix and Sinkhorn’s theorem (cf. Sinkhorn [27], Sinkhorn and Knopp [28]) for the matrix decomposition ensures that iterating (3.13) converges and the vectors \(\tilde{\beta}\) and \(\tilde{\gamma}\) are unique up to a scalar. Algorithm 2 summarizes the individual steps again.

Remark 3.8 (Central path). We want to emphasize that for changing the regularization parameter \(\lambda\) it is not necessary to recompute all powers in (3.14). Indeed, increasing \(\lambda\) to \(2 \cdot \lambda\), for example, corresponds to raising all entries in the matrix (3.14) to the power 2, etc.

Remark 3.9 (Softmax). The expression (3.13) resembles to what is known as the Gibbs measure and to the softmax in data science.
Algorithm 2: Sinkhorn’s iteration

**Input:** distance matrix \( d^r \in \mathbb{R}^{n \times n}_0 \), probability vectors \( p, \tilde{p} \in \mathbb{R}^n_0 \), regularization parameter \( \lambda > 0 \), stopping criterion and a starting value \( \tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_n) \)

**Output:** \( \tilde{\beta}, \tilde{\gamma} \) for \( \text{diag}(\tilde{\beta}) \cdot e^{-\lambda d^r} \cdot \text{diag}(\tilde{\gamma}) \)

set \( k_{ij} := \exp \left( -\lambda d^r_{ij} \right) \).

\[ \text{(3.14)} \]

while stopping criterion is not satisfied do

for \( i = 1 \text{ to } n \) do

\[ \tilde{\beta}_i \leftarrow \frac{p_i}{\sum_{j=1}^{n} k_{ij} \tilde{\gamma}_j} \]

for \( j = 1 \text{ to } \tilde{n} \) 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 3.10** (Historical remark). In the literature, this approach is also known as matrix scaling (cf. Rote and Zachariasen [25]), RAS (cf. Bachem and Korte [2]) as well as Iterative Proportional Fitting (cf. Rüschendorf [26]). Kruithof [17] used the method for the first time in telephone forecasting. The importance of this iteration scheme for data science was probably observed in Cuturi [9, Algorithm 1] for the first time.

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 investigate 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 \( d^{(t)} \) be the matrix of incremental divergences of sub-trees at stage \( t \). Analogously to (2.6) we consider the conditional version of the problem (3.1a) and denote by \( \beta_{i,j:t} \) and \( \gamma_{j:i:t} \) the pair of
optimal Lagrange parameters associated with the problem

\[
\begin{align*}
\text{minimize}_{\pi, \alpha} & \quad \sum_{i' \in i+} \pi(i', j' \mid i_t, j_t) \cdot d^{(t+1)}(i', j') \\
\text{subject to} & \quad \sum_{j' \in j+} \pi(i', j' \mid i_t, j_t) = P(i' \mid i_t), \quad i' \in i+,
\end{align*}
\]

which is the Sinkhorn divergence. The optimal value is the new divergence \( d^{(t)}(i, j) \). Computing the nested distance recursively from \( t = T - 1 \) down to 0 we get

\[
\pi_{ij} = \pi_1(i_1, j_1 \mid i_0, j_0) \cdot \cdots \cdot \pi_{T-1}(i, j \mid i_{T-1}, j_{T-1}) = e^{-\lambda d^{(1)}_{i_0,j_0} - \beta_{i,j} - \gamma_{i,j}} \cdot \cdots \cdot e^{-\lambda d^{(T)}_{i_{T-1},j_{T-1}} - \beta_{i,j} - \gamma_{i,j}}
\]

\[
= e^{-T - \lambda \sum_{t=0}^{T-1} d^{(t+1)}_{i_t,j_t} - \beta_{i,j} - \gamma_{i,j}}
\]

where \( i \in N_T \) and \( j \in N_T \) 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

\[
\tilde{\beta}_{i,j} := \exp(\lambda \beta_{i,j} - 1/2) \quad \text{and} \quad \tilde{\gamma}_{i,j} := \exp(\lambda \gamma_{i,j} - 1/2).
\]

Combining the components it follows that

\[
\pi_{ij} = \exp\left(-T - \lambda \sum_{t=0}^{T-1} d^{(t+1)}_{i_t,j_t} - \beta_{i,j} - \gamma_{i,j}\right)
\]

\[
= \prod_{t=0}^{T-1} \tilde{\beta}_{i_t,j_t} \exp\left(-\lambda d^{(t+1)}_{i_t,j_t}\right) \tilde{\gamma}_{i_t,j_t},
\]

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),

\[13\]
resp.) coincides with the optimal transport plan given by

\[
\begin{align*}
\text{minimize in } \pi & \sum_{i,j} \pi_{ij} \cdot d_{ij}^r + \frac{1}{\lambda} \pi_{ij} \cdot \log(\pi_{ij}) \\
\text{subject to} & \sum_{j \neq j_1} \pi(i, j \mid i_1, j_1) = P(i \mid i_1), \quad i < i_1, j_1, \quad i \geq j_1, \\
& \sum_{i \neq i_1} \pi(i, j \mid i_1, j_1) = \tilde{P}(j \mid j_1), \quad j < j_1, i_1, \\
& \pi_{ij} > 0 \text{ and } \sum_{i,j} \pi_{ij} = 1.
\end{align*}
\] (4.3)

**Proof.** First define \( \pi := \prod_{t=1}^{T} \pi_t \), where \( \pi_t \) 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

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

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

\[
de_{T-1} = \mathbb{E} \left[ de^r_T + \frac{1}{\lambda} \log \pi_T \mid F_{T-1} \otimes \tilde{F}_{T-1} \right]^{\frac{1}{r}}
\]

and from calculation in backward recursive way

\[
de_{T-2} = \mathbb{E} \left[ de^r_{T-1} + \frac{1}{\lambda} \log \pi_{T-1} \mid F_{T-2} \otimes \tilde{F}_{T-2} \right]^{\frac{1}{r}}
\]

\[
= E \left[ \mathbb{E} \left[ de^r_T + \frac{1}{\lambda} \log \pi_T \mid F_{T-1} \otimes \tilde{F}_{T-1} \right] + \frac{1}{\lambda} \log \pi_{T-1} \mid F_{T-2} \otimes \tilde{F}_{T-2} \right]^{\frac{1}{r}}
\]

\[
= E \left[ \mathbb{E} \left[ de^r_T + \frac{1}{\lambda} \log \pi_T + \frac{1}{\lambda} \log \pi_{T-1} \mid F_{T-1} \otimes \tilde{F}_{T-1} \right] \mid F_{T-2} \otimes \tilde{F}_{T-2} \right]^{\frac{1}{r}}.
\]
Finally, it follows that
\[
de_0 = E \left[ \frac{1}{\lambda} \log \pi \left| \mathcal{F}_0 \otimes \tilde{\mathcal{F}}_0 \right] \right]^{\frac{1}{r}} \\
= E \left[ \ldots E \left[ \frac{1}{\lambda} \log \pi_T \left| \mathcal{F}_{T-1} \otimes \tilde{\mathcal{F}}_{T-1} \right| \ldots \right] \frac{1}{\lambda} \log \pi_1 \left| \mathcal{F}_0 \otimes \tilde{\mathcal{F}}_0 \right] \right]^{\frac{1}{r}} \\
= E \left[ \ldots E \left[ \frac{1}{\lambda} \sum_{t=1}^{T} \log \pi_t \left| \mathcal{F}_{T-1} \otimes \tilde{\mathcal{F}}_{T-1} \right| \ldots \right] \frac{1}{\lambda} \log \pi_0 \left| \mathcal{F}_0 \otimes \tilde{\mathcal{F}}_0 \right] \right]^{\frac{1}{r}} \\
= E \left[ \frac{1}{\lambda} \sum_{t=1}^{T} \log \pi_t \left| \mathcal{F}_0 \otimes \tilde{\mathcal{F}}_0 \right] \right]^{\frac{1}{r}} \\
= E \left[ \frac{1}{\lambda} \sum_{t=1}^{T} \log \pi_t \right]^{\frac{1}{r}} ,
\]
the assertion (4.3) of the theorem. \(\square\)

Remark 4.2. The optimization problem in Theorem 4.1 considers all constraints as the full nested problem (2.4), only the objective differs. For this reason the optimal solution of (4.3) is feasible for the problem (2.4) 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.,
\[
0 \leq d'_S - d'_W \leq \frac{1}{\lambda} \left( H(\pi^S) - H(\pi^W) \right) \quad \text{and} \quad 0 \leq d'_W - d'_S \leq \frac{1}{\lambda} H(\pi^S) \leq \frac{1}{\lambda} H(p \cdot p^T),
\]
where \(\pi^S\) (\(\pi^W\), resp.) is the optimal transport plan from (4.3) ((2.4), 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. \(\square\)

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 \(\bar{P}\) (\(\bar{P}\), resp.), i.e., \(m = \max \{|i+1|: i \in \mathcal{N}_t, t = 1, \ldots, T-1\}\). It holds that
\[
de'_S - d'_W \leq \frac{\log m + \log \tilde{m}}{\lambda} \cdot T.
\]
where \(T\) is the total number of stages.

Proof. Recall from Remark 3.6 that \(H(\pi^S) \leq \log (n \tilde{n}) = \log n + \log \tilde{n}\) for every conditional probability measures, where \(n\) and \(\tilde{n}\) are the number of immediate successors in both trees. The result follows with \(n \leq m^T\) (\(\tilde{n} \leq \tilde{m}^T\), resp.) and \(\log n \leq T \log m\) and the nested program (4.1). \(\square\)
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. (2.2a)–(2.2b) 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

\[
\begin{align*}
\text{minimize} & \quad \ln \pi \left( \int d(\xi, \tilde{\xi})^r \right)^{1/r} \\
\text{subject to} & \quad \pi(A \times \Xi | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t), \quad A \in \mathcal{F}_t, \ t = 1, \ldots, T, \quad (4.6a) \\
& \quad \pi(\Xi \times B | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(B | \tilde{\mathcal{F}}_t), \quad B \in \tilde{\mathcal{F}}_t, \ t = 1, \ldots, T. \quad (4.6b)
\end{align*}
\]

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

\[
\begin{align*}
\text{maximize} & \quad \ln M_0 \quad M_0 \\
\text{subject to} & \quad E\hat{\beta} = 0, \ \tilde{E}\hat{\gamma} = 0, \\
& \quad \sum_{\xi, \tilde{\xi}} \exp(-\lambda (d(\xi, \tilde{\xi})^r - \hat{\beta}(\xi) - \hat{\gamma}(\tilde{\xi}) - M_0) - 1) = 1, \\
& \quad \hat{\beta} \in \mathbb{R}^n, \ \hat{\gamma} \in \mathbb{R}^\tilde{n}.
\end{align*}
\]

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

\[
\begin{align*}
\text{proj}_t : L^1(\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) & \rightarrow L^1(\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\
\hat{\beta}(\xi) \cdot \hat{\gamma}(\tilde{\xi}) & \mapsto E(\hat{\beta} | \mathcal{F}_t)(\xi) \cdot \hat{\gamma}(\tilde{\xi})
\end{align*}
\]

and

\[
\begin{align*}
\text{proj}_t^\tilde{t} : L^1(\tilde{\mathcal{F}}_t \otimes \tilde{\mathcal{F}}_\tilde{t}) & \rightarrow L^1(\tilde{\mathcal{F}}_t \otimes \tilde{\mathcal{F}}_\tilde{t}) \\
\hat{\beta}(\xi) \cdot \hat{\gamma}(\tilde{\xi}) & \mapsto \hat{\beta}(\xi) \cdot E(\hat{\gamma} | \tilde{\mathcal{F}}_\tilde{t})(\tilde{\xi}).
\end{align*}
\]

We recall the following characterization of the measurability constraints (4.6a)–(4.6b) and refer to [22, Proposition 2.48] for its proof.

**Proposition 4.5.** The measure \( \pi \) satisfies the marginal condition

\[
\pi(A \times \Xi | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t) \quad \text{for all} \ A \in \Xi
\]

if and only if

\[
E_\pi \beta = E_\pi \text{proj}_t \beta \quad \text{for all} \ \beta \in \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t.
\]

Moreover, \( \text{proj}_t(\beta) = E_\pi(\beta | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \) if \( \pi \) has marginal \( P \).
Theorem 4.6. The infimum or the nested distance including the entropy $d e^r (\mathcal{P}, \tilde{\mathcal{P}})$ of problem (4.3) equals the supremum of all numbers $M_0$ such that
\[ e^{-\lambda d(\xi, \xi') - M_T(\xi, \xi')) - 1} \in \mathcal{P}(\Xi \times \tilde{\Xi}), \quad (\xi, \xi') \in \Xi \times \tilde{\Xi}, \]
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
\[ M_t = M_0 + \sum_{s=1}^t \hat{\beta}_s + \hat{\gamma}_s \quad (4.7) \]
and the measurable functions $\hat{\beta}_t \in \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}$ and $\hat{\gamma}_t \in \mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_t$ satisfy $\text{proj}_{t-1}(\hat{\beta}_t) = 0$ and $\text{proj}_{t-1}(\hat{\gamma}_t) = 0$.

Proof. With Proposition 4.5 rewrite the dual problem as
\[
\inf_{\pi > 0} \sup_{M_0, \hat{\beta}, \hat{\gamma}_t} \mathbb{E}_{\pi} \left[ d^r + \frac{1}{\lambda} \log \pi \right] + M_0 \cdot (1 - \mathbb{E}_{\pi} \mathbb{1}) + \\
- \sum_{s=0}^{T-1} (\mathbb{E}_{\pi} f_{s+1} - \mathbb{E}_{\pi} \text{proj}_s(f_{s+1})) - \sum_{s=0}^{T-1} (\mathbb{E}_{\pi} g_{s+1} - \mathbb{E}_{\pi} \text{proj}_s(g_{s+1})),
\]
where the second line encodes the measurability constraints. By the minmax theorem (cf. Sion [29]) this is equivalent to
\[
\sup_{M_0, \hat{\beta}, \hat{\gamma}_t} M_0 + \inf_{\pi > 0} \mathbb{E}_{\pi} \left[ d^r + \frac{1}{\lambda} \log \pi - M_0 \cdot \mathbb{1} - \\
- \sum_{s=0}^{T-1} (f_{s+1} - \text{proj}_s(f_{s+1})) - \sum_{s=0}^{T-1} (g_{s+1} - \text{proj}_s(g_{s+1})). \right]
\]
The integral exists and the minimum is obtained by a probability measure
\[
\pi = \exp \left( -\lambda \left( d^r - \sum_{s=0}^{T-1} (f_{s+1} - \text{proj}_s(f_{s+1})) - \sum_{s=0}^{T-1} (g_{s+1} - \text{proj}_s(g_{s+1}) - M_0) \right) - 1 \right).
\]
Set $\tilde{\beta}_s := f_s - \text{proj}_{s-1}(f_s)$ and $\tilde{\gamma}_s := g_s - \text{proj}_{s-1}(g_s)$. Consequently, the problem reads
\[
\text{maximize}_{M_0} M_0 \\
\text{subject to } \exp \left[ -\lambda \left( d^r - \sum_{s=1}^{T} \tilde{\beta}_s - \sum_{s=1}^{T} \tilde{\gamma}_s - M_0 \right) - 1 \right] \in \mathcal{P}(\Xi \times \tilde{\Xi}) \\
\text{proj}_{t-1} (\tilde{\beta}_t) = 0, \text{proj}_{t-1} (\tilde{\gamma}_t) = 0,
\]
and thus the assertion. \(\square\)

The following corollary links the optimal probability measure and the stochastic process (4.7) for the optimal components $\tilde{\beta}$ and $\tilde{\gamma}$.

Corollary 4.7. The process $M_t$ in (4.7), for which the supremum is attained, is a martingale with respect to the optimal measure $\pi$.

Proof. The proof of [22, Theorem 2.49] applies with minor adaptions only. \(\square\)
5 Numerical results

(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)

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

The nested Sinkhorn divergence \( d'_S \) as well as \( de'_\xi \) depend on the regularization parameter \( \lambda \). We discuss this dependency, the error, speed of convergence and numerical issues in comparison to the non-regularized nested distance \( d'_W \).

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

Figure 2 displays the results. We see that the regularized nested distance \( d'_S \) (green) and \( de'_\xi \) (red) converge to the nested distance \( d'_W \) for increasing \( \lambda \). In contrast to \( d'_S \), the regularized nested distance including the entropy converges slower to \( d'_W \). The reason is that for larger \( \lambda \) the weight of the entropy in the cost function in (3.1a) decreases and the entropy of \( \pi^S \) and \( \pi^W \) coincide (cf. (4.5)). 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 \( d'_W \) and \( d'_S \) and \( de'_\xi \), respectively, is rapidly decreasing and insignificant for \( \lambda > 20 \). Moreover, the time displayed in Figure 2b does not depend on the regularization parameter \( \lambda \).

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.

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 $\pi^S$. This approach has no advantages.

### 6 Summary

Nested distance allows distinguishing trees by involving the information encoded in filtrations. In this paper we regularize the Wasserstein distance and introduce the Sinkhorn divergence to the nested distance. The tower property also applies for the regularization. We show that the nested divergence converges to the nested distance for increasing regularization parameter $\lambda \to \infty$.

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.
Table 1: Average distance and divergence with corresponding computational time in seconds on i5-3210M CPU. All states and probabilities are generated randomly. The regularization parameter is $\lambda = 20$ and $r = 1$.

| stages | Wasserstein $d_{W}^{t}$ | time | Sinkhorn $d_{S}^{t}$ | time | difference $d_{W}^{t} - d_{S}^{t}$ | time | 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×         |

7 Acknowledgement

We are thankful to Benoît Tran for pointing out further references, particularly on convergence, in his thesis [30] supervised by Marianne Akian and Jean-Philippe Chancelier.

References

[1] B. Analui and G. Ch. Pflug. On distributionally robust multiperiod stochastic optimization. *Computational Management Science*, 11(3):197–220, 2014. ISSN 1619-6988. doi:10.1007/s10287-014-0213-y. 1

[2] A. Bachem and B. Korte. On the RAS-algorithm. 23(2):189–198. doi:10.1007/bf02252097. 12

[3] F. Beltrán, W. de Oliveira, and E. C. Finardi. Application of scenario tree reduction via quadratic process to medium-term hydrothermal scheduling problem. 32(6):4351–4361. doi:10.1109/tpwrs.2017.2658444. 1

[4] D. P. Bertsekas and D. A. Castanon. The auction algorithm for the transportation problem. *Annals of Operations Research*, 20(1):67–96. doi:10.1007/bf02216923. 2

[5] J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2):5120–5150. doi:10.1214/19-EJS1637. 2, 8

[6] A. I. Brodt. Min-mad life: A multi-period optimization model for life insurance company investment decisions. *Insurance: Mathematics and Economics*, 2(2):91–102, 1983. 1

[7] P. Carpentier, J.-P. Chancelier, G. Cohen, M. De Lara, and P. Girardeau. Dynamic consistency for stochastic optimal control problems. *Annals of Operations Research*, 200(1):247–263, 2012. doi:10.1007/s10479-011-1027-8. 1
[8] P. Carpentier, J.-P. Chancelier, G. Cohen, and M. De Lara. *Stochastic Multi-Stage Optimization*. Springer International Publishing, 2015. doi:10.1007/978-3-319-18138-7.

[9] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 2013. 2, 9, 12

[10] N. C. P. Edirisinghe. Multiperiod portfolio optimization with terminal liability: Bounds for the convex case. *Computational Optimization and Applications*, 32(1-2):29–59. doi:10.1007/s10589-005-2053-8.

[11] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR. URL http://proceedings.mlr.press/v84/genevay18a.html.

[12] H. Heitsch, W. Römisch, and C. Strugarek. Stability of multistage stochastic programs. *SIAM Journal on Optimization*, 17(2):511–525, 2006.

[13] M. Horejšová, S. Vitali, M. Kopa, and V. Moriggia. Evaluation of scenario reduction algorithms with nested distance. *Computational Management Science*, 17(2):241–275, 2020. doi:10.1007/s10287-020-00375-4.

[14] K. B. Kirui, A. Pichler, and G. Ch. Pflug. ScenTrees.jl: A Julia package for generating scenario trees and scenario lattices for multistage stochastic programming. *Journal of Open Source Software*, 5(46):1912, 2020. doi:10.21105/joss.01912.

[15] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. 34(4):43–59. doi:10.1109/msp.2017.2695801.

[16] R. M. Kovacevic and A. Pichler. Tree approximation for discrete time stochastic processes: a process distance approach. *Annals of Operations Research*, pages 1–27, 2015. doi:10.1007/s10479-015-1994-2.

[17] R. Kruithof. Telefoonverkeersrekening. *De Ingenieur* 52, E15-E25, 1937. URL https://wwwhome.ewi.utwente.nl/~ptdeboer/misc/kruithof-1937-translation.html.

[18] G. Luise, A. Rudi, M. Pontil, and C. Ciliberto. Differential properties of Sinkhorn approximation for learning with Wasserstein distance. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*. URL https://arxiv.org/abs/1805.11897.

[19] F. Maggioni and G. Ch. Pflug. Guaranteed bounds for general non-discrete multistage risk-averse stochastic optimization programs. *SIAM Journal on Optimization*, 29(1):454–483, 2019. doi:10.1137/17M1140601.
[20] G. Ch. Pflug. Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization*, 20:1406–1420, 2009. doi:10.1137/080718401.

[21] G. Ch. Pflug and A. Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012. doi:10.1137/110825054.

[22] G. Ch. Pflug and A. Pichler. *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2014. ISBN 978-3-319-08842-6. doi:10.1007/978-3-319-08843-3. URL https://books.google.com/books?id=q_VWBQAQBAJ.

[23] S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. John Wiley and Sons, West Sussex, England, 1991. URL http://books.google.com/books?id=5grvAAAAMAAJ.

[24] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems Volume I: Theory, Volume II: Applications*, volume XXV of *Probability and its applications*. Springer, New York, 1998. doi:10.1007/b98893.

[25] G. Rote and M. Zachariasen. Matrix scaling by network flow. In N. Bansal, K. Pruhs, and C. Stein, editors, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, New Orleans, Louisiana, USA, January 7-9, 2007, pages 848–854. SIAM, 2007. URL http://dl.acm.org/citation.cfm?id=1283383.1283474.

[26] L. Rüschendorf. Convergence of the iterative proportional fitting procedure. 23(4):1160–1174. doi:10.1214/aos/1176324703.

[27] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. 74(4):402. doi:10.12307/2314570.

[28] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21:343–348, 1967. ISSN 0030-8730.

[29] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. URL https://projecteuclid.org/euclid.pjm/1103040253.

[30] D. N. B. Tran. *Programmation dynamique tropicale en optimisation stochastique multi-étapes*. PhD thesis, Université Paris-Est, 2020.

[31] C. Villani. *Optimal transport, old and new*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, 2009.