The derivatives of Sinkhorn-Knopp converge

Edouard Pauwels, Samuel Vaiter

THE DERIVATIVES OF SINKHORN–KNOPP CONVERGE*

EDOUARD PAUWELS† AND SAMUEL VAITER‡

Abstract. We show that the derivatives of the Sinkhorn–Knopp algorithm, or iterative proportional fitting procedure, converge towards the derivatives of the entropic regularization of the optimal transport problem with a locally uniform linear convergence rate.

Key words. Optimal transport, Sinkhorn algorithm, Algorithmic differentiation

MSC codes. 65K10, 90B06, 40A30

1. Introduction. The optimal transport (OT) problem plays an increasing role in optimization and machine learning [26]. In particular, entropic regularization of OT has gained a lot of attraction by the existence of a simple and efficient algorithm introduced in [31], also known as matrix scaling or iterative proportional fitting procedure in the stochastic literature, see [28]. It is known that Sinkhorn–Knopp iterates converge linearly, with an explicit rate computable from the cost matrix, to the solution of entropic OT since the work of [16] introducing the use of the Hilbert metric.

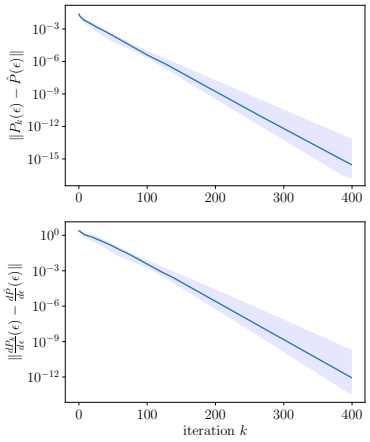

FIG. 1: Illustration of the linear convergence of the regularized transport plan $P_k(\theta)$ (2.5) of Sinkhorn-Knopp (SK_θ) and its derivatives $\frac{dP_k}{d\theta}(\theta)$ towards the derivative of the entropic optimal transport problem (OT_θ).

1.1. Differentiation of the Sinkhorn–Knopp algorithm. Among the different properties of Sinkhorn–Knopp, a striking one is its differentiability with respect to the inputs. Differentiating the iterates of the Sinkhorn–Knopp algorithm is a common routine in machine learning. It was first used by [1] for ranking with linear objective function. They proposed to use backpropagation through Sinkhorn–Knopp iterates with respect to the cost matrix, without discussion of the convergence of the Jacobian. It was later used for different applications, such as to compute Wasserstein barycenters casted as an optimization problem [6], where the backpropagation is performed with respect to the weight vector, for training generative models involving an OT loss as in [20, 17], to define differentiable sorting procedures [13] or to solve cluster assignments problems [8]. Popular libraries such as `POT` [15] or `OTT` [11] for computational optimal transport implement the backpropagation of Sinkhorn–Knopp. To mitigate the memory footprint required by backprogation, an alternative is to use implicit differentiation as discussed first by [24]

*Submitted July 22, 2022.

Funding: E. P. acknowledges the financial support of the AI Interdisciplinary Institute ANITI funding under the grant agreement ANR-19-PI3A-0004, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA9550-19-1-7026, and ANR MaSDOL 19-CE23-0017-01. S. V. acknowledges the support ANR GraVa, grant ANR-18-CE40-0005.

†IRIT, CNRS, Université de Toulouse, ANITI, Toulouse, France. Institut universitaire de France (IUF). (edouard.pauwels@irit.fr).

‡CNRS & Université Côte d'Azur, CNRS, LJAD, France (samuel.vaiter@cnrs.fr).

for computing the derivatives of Sinkhorn divergences. This approach was later used in [12, 14]. To the best of our knowledge, even though some these works justify the correctness of using automatic differentiation for a given iterate, *they do not consider the issue of the convergence of the derivatives* computed by automatic differentiation.

1.2. Convergence of algorithmic differentiation. The issue of the convergence of the derivatives of an algorithm was considered in the automatic differentiation community. The linear convergence of derivatives was studied in [18, 19] for piggyback recursion and in [9, Theorem 2.3] for backpropagation. More recently, convergence of the derivatives of gradient descent [25, 23], the Heavy-ball [25] method or non-smooth fixed point methods [5] were analyzed. All these analysis *require explicitly, or implicitly, that the (generalized) Jacobians are firmly nonexpansive*, *i.e.,* Lipschitz continuous with a constant strictly lesser than 1. Unfortunately, the derivatives of Sinkhorn–Knopp do not enjoy this property.

1.3. Contribution. We prove (Theorem 3.3) that the derivatives of the iterates of Sinkhorn–Knopp algorithm converge towards the derivative of the entropic regularization of optimal transport, with an explicit expression of the derivative and with a locally uniform linear convergence rate, provided that all functions entering problem definition are continuously differentiable.

1.4. Organization. Our paper is organized as follows. Section 2 introduces the parameterized entropic regularized optimal transport problem and the Sinkhorn–Knopp algorithm In Section 3, we state our main result stating the convergence of the derivatives of Sinkhorn-Knopp towards the derivatives of the regularized optimal transport with a locally uniform linear convergence rate. Section 4 provides the proof of our result. Section 5 contains important intermediate results to prove the lemma that allows us to obtain a linear rate for the convergence. Section 6 establishes miscellaneous lemmas that are used in the main proof.

1.5. Notations. The set of positive reals is denoted $\mathbb{R}_{>0}$, of nonnegative reals $\mathbb{R}_{\geq 0}$ and of nonzero reals $\mathbb{R}_{\neq 0}$. The simplex Δ^{n-1} is the set of vectors of $\mathbb{R}^n_{\geq 0}$ summing to 1

$$\Delta^{n-1} = \left\{ x \in \mathbb{R}^n \, : \, \sum_{i=1}^n x_i = 1 \text{ and } x_i \geq 0, \forall i \in \{1, \dots, n\} \right\}.$$

The identity matrix (of arbitrary size) is denoted I. For two vectors $x \in \mathbb{R}^n, y \in \mathbb{R}^n_{\neq 0}$, the *entry-wise* (Hadamard) division $\frac{x}{y}$ is defined as $\left(\frac{x}{y}\right)_i = x_i/y_i$, and the product $x \odot y$ is defined as $(x \odot y)_i = x_i y_i$, for all $i \in \{1, \dots, n\}$. The 1-vector $1_n \in \mathbb{R}^n$ is the vector only composed of 1's. When the context is clear, and to lighten the notations, $\frac{1}{x}$ for $x \in \mathbb{R}_{\neq 0}$ should be understood as $\frac{1_n}{x}$. Given a function $f : \mathbb{R} \to \mathbb{R}$, we extend its domain as $f : \mathbb{R}^p \to \mathbb{R}^p$ by applying it entrywise, that is for $x \in \mathbb{R}^n$, $f(x)_i = f(x_i)$, for all $i \in \{1, \dots, n\}$. Given $l \in \mathbb{N}_{>0}$ and a continuously differentiable function $F : \mathbb{R}^p \to \mathbb{R}^{n_1 \times \cdots \times n_l}$, we denote by $\frac{dF}{d\theta}(\theta) \in \mathbb{R}^{n_1 \times \cdots \times n_l \times p}$ its Jacobian matrix (or tensor) at $\theta \in \mathbb{R}^p$, *i.e.,*

$$\left(\frac{dF}{d\theta}(\theta)\right)_{i_1, \cdots, i_l, j} = \lim_{h \to 0} \frac{F_{i_1, \cdots, i_l}(\theta + h e_j) - F_{i_1, \cdots, i_l}(\theta)}{h},$$

where $(e_j)_{j=1,\dots,n}$ is the canonical basis of $\mathbb{R}^n$. Given a differentiable function $F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$, we denote by $J_F(x, \theta)$ the total derivative at $(x, \theta) \in \mathbb{R}^n \times \mathbb{R}^p$, that

is

$$J_F(x,\theta) = \left(\frac{\partial F(\cdot,\theta)}{\partial x}(x) \quad \frac{\partial F(x,\cdot)}{\partial \theta}(\theta) \right),$$

where $\frac{\partial F(\cdot,\theta)}{\partial x}(x)$ and $\frac{\partial F(x,\cdot)}{\partial \theta}(\theta)$ are the partial derivatives of F.

2. Entropic optimal transport and Sinkhorn–Knopp algorithm.

2.1. Entropic regularization. We consider a parametric formulation of the entropic OT[1]. The entropic regularization of optimal transport associated to the parameterized marginals $a : \mathbb{R}^p \to \Delta^{n-1} \cap \mathbb{R}^n_{>0}$ and $b : \mathbb{R}^p \to \Delta^{n-1} \cap \mathbb{R}^m_{>0}$ of level $\epsilon : \mathbb{R}^p \to \mathbb{R}_{>0}$ for the parameterized cost matrix $C : \mathbb{R}^p \to \mathbb{R}^{n\times m}$ reads for $\theta \in \mathbb{R}^p$

$$(\mathrm{OT}_\theta) \qquad \hat{P}(\theta) = \arg\min_{P \in U(\theta)} \mathcal{L}(P,\theta) \stackrel{\text{def.}}{=} \langle P, C(\theta)\rangle - \epsilon(\theta)\,\mathrm{Ent}(P),$$

where $\langle P, P'\rangle = \sum_{i,j} P_{i,j} P'_{i,j}$, $U(\theta)$ is the set of admissible couplings (also called transportation polytope)

$$U(\theta) \stackrel{\text{def.}}{=} \{P \in \mathbb{R}^{n\times m}_{\geq 0} \,:\, P1_m = a(\theta) \quad \text{and} \quad P^\top 1_n = b(\theta)\},$$

and Ent is the entropic regularization of the coupling matrix P defined as

$$\mathrm{Ent}(P) \stackrel{\text{def.}}{=} -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{i,j}\left(\log(P_{i,j}) - 1\right),$$

where $P_{i,j}\log(P_{i,j}) = 0$ if $P_{i,j} = 0$. Note that $\mathcal{L}_\theta = \mathcal{L}(\cdot,\theta)$ defined in (OT_θ) is $\epsilon(\theta)$-strongly convex, hence (OT_θ) has a unique minimizer[2] $\hat{P}(\theta) \in \mathbb{R}^{n\times m}_{>0}$. We assume that all functions entering problem definition are continuously differentiable.

2.2. Sinkhorn–Knopp algorithm. The Sinkhorn–Knopp algorithm is built upon the fact [30, Theorem 1] that the unique solution $\hat{P}(\theta)$ of (OT_θ) has the form for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,m\}$

$$\hat{P}(\theta)_{i,j} = u_i(\theta) K_{i,j}(\theta) v_j(\theta) \quad \text{where} \quad K_{i,j}(\theta) = \exp\left(-\frac{C_{i,j}(\theta)}{\epsilon(\theta)}\right) > 0, \tag{2.1}$$

for positive numbers $u_i(\theta), v_j(\theta)$, $i = 1,\dots,n$, and $j = 1,\dots,m$. The goal is thus to find positive vectors $u(\theta) \in \mathbb{R}^n_{>0}$ and $v(\theta) \in \mathbb{R}^m_{>0}$, such that

$$\mathrm{diag}(u(\theta))K(\theta)\mathrm{diag}(v(\theta))1_m = a(\theta) \quad \text{and} \quad \mathrm{diag}(v(\theta))K(\theta)^T\mathrm{diag}(u(\theta))1_n = b(\theta). \tag{2.2}$$

In its most elementary formulation, the Sinkhorn–Knopp algorithm, also called matrix scaling problem algorithm, has the following alternating updates,

$$u_{k+1}(\theta) = \frac{a(\theta)}{K(\theta)v_k(\theta)} \qquad \text{and} \qquad v_{k+1}(\theta) = \frac{b(\theta)}{K(\theta)^T u_{k+1}(\theta)}, \tag{2.3}$$

starting from a couple $(u_0(\theta), v_0(\theta)) \in \mathbb{R}^n_{>0} \times \mathbb{R}^m_{>0}$, see [32] for a discussion on initializations strategies. Even though in practice it is not necessary to evaluate it at each iteration, one can use (2.1) to form a current guess at iteration k as $\mathrm{diag}(u_k(\theta))K(\theta)\mathrm{diag}(v_k(\theta))$.

[1] We recover the standard formulation letting a, b, C, ϵ be constant functions.

[2] The (strict) positivity follows from assumptions $a(\theta) > 0$ and $b(\theta) > 0$. Indeed, $P = a(\theta)b(\theta)^T$ is feasible for (OT_θ), with strictly positive entries, therefore Slater's qualification condition holds for (OT_θ) and the required form follows from necessary and sufficient KKT conditions for the (attained) optimum, see for example [10, Lemma 2]

2.3. Reduced formulation of Sinkhorn–Knopp. We will analyse an equivalent version of (2.3) by considering a single iterate u and performing the change of variable $x = \log(u)$. Given an initilization $x_0(\theta) \in \mathbb{R}^n$ is continuously differentiable, this results in rewriting (2.3) as the recursion in the "log-domain"

$$x_{k+1}(\theta) = F(x_k(\theta), \theta) \tag{SK$_\theta$}$$

where

$$F(x,\theta) \stackrel{\text{def.}}{=} \log(a(\theta)) - \log\left(K(\theta)\left(\frac{b(\theta)}{K(\theta)^T e^x}\right)\right).$$

Note that this formulation is close to the dual formulation of (OT$_\theta$) as explained in [26, Remark 4.22], but we will not need duality results along this paper.

We will work under the following standing assumption

Assumption 2.1 (Data are continuously differentiable). Let $\Omega \subseteq \mathbb{R}^p$ be a connected open set. The data in problem (OT$_\theta$), *i.e.,* $C\colon \Omega \to \mathbb{R}^{n\times m}$, $a\colon \Omega \to \Delta^{n-1} \cap \mathbb{R}^n_{>0}$, $b\colon \Omega \to \Delta^{m-1} \cap \mathbb{R}^n_{>0}$, $\epsilon\colon \Omega \to \mathbb{R}_{>0}$, and initialization $x_0\colon \Omega \to \mathbb{R}^n$, are all continuously differentiable functions on Ω.

It is possible to get back to the scaling factors $u_k(\theta)$ and $v_k(\theta)$ from the reduced variable $x_k(\theta)$ as

$$u_k(\theta) = e^{x_k(\theta)} \quad \text{and} \quad v_k(\theta) = \frac{b(\theta)}{K(\theta)^T e^{x_k(\theta)}}.$$

Using the relationship (2.1), the optimal coupling matrix can be approximated as

$$P(x,\theta) = \mathrm{diag}(e^x) K(\theta) \mathrm{diag}\left(\frac{b(\theta)}{K(\theta)^T e^x}\right), \tag{2.4}$$

and we construct transport plan estimates associated to each iterate, for all $k \in \mathbb{N}$,

$$P_k(\theta) = P(x_k(\theta), \theta). \tag{2.5}$$

It is known that $P_k(\theta)$ converges linearly [16] to the optimal transport plan $\hat{P}(\theta)$ for (OT$_\theta$). The next paragraph is dedicated to study the linear convergence of the reduced variable $x_k(\theta)$.

2.4. Linear convergence of the centered reduced iterates. It is known that $u_k(\theta)$ converges to a limit $\bar{u}(\theta)$, with a linear rate in the Hilbert metric [16], see also [26, Theorem 4.2], whereas we are concerned with the convergence of the reduced iterates in the "log-domain". In order to study the convergence of $(x_k)_{k\in\mathbb{N}}$, let us introduce the linear map L_{center} which associates to x its centered version:

$$L_{\text{center}}\colon \begin{cases} \mathbb{R}^n & \to \mathbb{R}^n \\ x & \mapsto x - \left(\frac{1}{n}\sum_{i=1}^n x_i\right) 1_n. \end{cases} \tag{2.6}$$

To analyze the convergence rate of Sinkhorn-Knopp algorithm, it is standard to use the Hilbert projective metric [4] defined on $\mathbb{R}^n_{>0}$ as

$$d_{\mathcal{H}}(u,u') = \|\log(u) - \log(u')\|_{\text{var}},$$

where $\|x\|_{\mathrm{var}}$ is the variation seminorm of $x \in \mathbb{R}^n$ defined as

$$\|x\|_{\mathrm{var}} = \max_{i=1,\dots,n} x_i - \min_{i=1,\dots,n} x_i. \tag{2.7}$$

The next lemma shows the (local) linear convergence in ℓ^2 norm of the centered reduced variable $L_{\mathrm{center}}(x_k(\theta))$.

LEMMA 2.2 (Local linear convergence of $L_{\mathrm{center}}(x_k(\theta))$). *The centered reduced variable $L_{\mathrm{center}}(x_k(\theta))$ converges linearly to $L_{\mathrm{center}}(\bar{x}(\theta))$,* i.e., *there exists $c : \Omega \to \mathbb{R}_{>0}$ and $\rho : \Omega \to (0,1)$ continuous such that all $k \in \mathbb{N}$ and $\theta \in \Omega$,*

$$\|L_{\mathrm{center}}(x_k(\theta)) - L_{\mathrm{center}}(\bar{x}(\theta))\| \le c(\theta)\rho(\theta)^k.$$

Furthermore, $\theta \to L_{\mathrm{center}}(\bar{x}(\theta))$ is continuous on Ω.

Proof. We combine the linear convergence result on $u_k(\theta)$ of [16] with Lemma 6.3, following the suggestion of [26, Remark 4.12].

We clarify below how to combine these arguments. We first show that the linear convergence of $u_k(\theta)$ is such that for all $\theta \in \Omega$ there exists $c(\theta) > 0$ and $\rho(\theta) \in (0,1)$ such that for all $k \in \mathbb{N}$

$$d_{\mathcal{H}}(u_k(\theta), \bar{u}(\theta)) \le c(\theta)\rho(\theta)^k,$$

and the mapping c and ρ are continuous. Indeed, [16, Theorem 4] ensures that for all $k \in \mathbb{N}$,

$$d_{\mathcal{H}}(u_k(\theta), \bar{u}(\theta)) + d_{\mathcal{H}}(v_k(\theta), \bar{v}(\theta)) \le \frac{\kappa^2(K(\theta))^k}{1-\kappa^2(K(\theta))}(d_{\mathcal{H}}(u_0(\theta), \bar{u}(\theta)) + d_{\mathcal{H}}(v_0(\theta), \bar{v}(\theta))),$$

where $\kappa(K)$ is the contraction ratio defined for $K \in \mathbb{R}^{n\times m}_{>0}$ as

$$\kappa(K) = \frac{\vartheta(K)^{1/2}-1}{\vartheta(K)^{1/2}+1} < 1 \quad \text{and} \quad \vartheta(K) = \max_{i,j,k,l} \frac{K_{i,k}K_{j,l}}{K_{j,k}K_{i,l}}.$$

Remark that P_k and $\hat{P}(\theta)$ enjoy the relation

$$P_k = \mathrm{diag}\left(\frac{u_k(\theta)}{\bar{u}(\theta)}\right)\hat{P}(\theta)\mathrm{diag}\left(\frac{v_k(\theta)}{\bar{v}(\theta)}\right)$$

and $d_{\mathcal{H}}(\frac{u_k(\theta)}{\bar{u}(\theta)}, 1_n) = d_{\mathcal{H}}(u_k(\theta), \bar{u}(\theta))$. Using [26, Theorem 4.2], we deduce that

$$\begin{aligned} d_{\mathcal{H}}(u_k(\theta), \bar{u}(\theta)) &\le \frac{\kappa^2(K(\theta))^k}{(1-\kappa^2(K(\theta)))^2}\left(d_{\mathcal{H}}(P(x_0(\theta),\theta)1_m, a) + d_{\mathcal{H}}(P(x_0(\theta),\theta)^T 1_n, b)\right), \\ &= c(\theta)\rho(\theta)^k, \end{aligned}$$

where

$$\begin{aligned} c(\theta) &= \kappa^2(\theta)\frac{d_{\mathcal{H}}(P(x_0(\theta),\theta)1_m, a(\theta)) + d_{\mathcal{H}}(P(x_0(\theta),\theta)^T 1_n, b(\theta))}{(1-\kappa^2(K(\theta)))^2}, \\ \rho(\theta) &= \kappa^2(\theta). \end{aligned}$$

Since $K(\theta) > 0$ and continuous, we have that $\theta \mapsto \kappa^2(\theta)$ is continuous, and since $\theta \mapsto x_0(\theta)$ is assumed to be continuous on Ω, $\theta \mapsto d_{\mathcal{H}}(P(x_0(\theta),\theta)$ is also continuous.

Thus, $c(\theta)$ and $\rho(\theta)$ depend continuously on the initial condition x_0 and problem data (a, b, K, ϵ) which are all continuous functions of θ. Therefore the linear convergence is actually locally uniform in θ.

To conclude the proof, we need to remark that the Hilbert projective metric on u corresponds to the variation seminorm after the change of variable $x = \log(u)$ so that for all $k \in \mathbb{N}$ and all $\theta \in \Omega$,

$$\|x_k(\theta) - \bar{x}(\theta)\|_{\mathrm{var}} = d_{\mathcal{H}}(u_k(\theta), \bar{u}(\theta)),$$

and Lemma 6.3 provides

$$\|L_{\mathrm{center}}(x_k(\theta)) - L_{\mathrm{center}}(\bar{x}(\theta))\|_\infty \leq \|x_k(\theta) - \bar{x}(\theta)\|_{\mathrm{var}},$$

which is the claimed result.

Regarding the continuity, let $\theta_0 \in \Omega$, for all $\theta \in \Omega$ and all $k \in \mathbb{N}$, we have

$$\begin{aligned} d_{\mathcal{H}}(\bar{u}(\theta), \bar{u}(\theta_0)) &\leq d_{\mathcal{H}}(\bar{u}(\theta), u_k(\theta)) + d_{\mathcal{H}}(u_k(\theta), u_k(\theta_0)) + d_{\mathcal{H}}(u_k(\theta_0), \bar{u}(\theta_0)) \\ &\leq c(\theta)\rho(\theta)^k + c(\theta_0)\rho(\theta_0)^k + d_{\mathcal{H}}(u_k(\theta), u_k(\theta_0)). \end{aligned}$$

We may choose k such that the first two terms are as small as desired uniformly for θ in a neighborhood of θ_0. The last term is continuous in θ and evaluates to 0 for $\theta = \theta_0$ so that reducing the neighborhood if necessary allows to choose it as small as desired, which proves continuity. ∎

Note that Lemma 2.2 does not imply the linear convergence of $(x_k(\theta))_{k\in\mathbb{N}}$. As we will see later in Lemma 4.4, this is not an issue to our objective – proving the convergence of the derivatives of (SK_θ) – because $x_k(\theta)$ and $L_{\mathrm{center}}(x_k(\theta))$ have the same "differentiable" properties.

3. Derivatives of Sinkhorn–Knopp algorithm and their convergence.

3.1. Derivatives of the transport plan. Remark that for all $(x, \theta) \in \mathbb{R}^n \times \Omega$, $P(x, \theta)$ is an $n \times m$ matrix. Hence, $P(x, \cdot)$ is a map from $\mathbb{R}^p$ to $\mathbb{R}^{n\times m}$ and $P(\cdot, \theta)$ is a map from $\mathbb{R}^n$ to $\mathbb{R}^{n\times m}$. Thus, we identify its partial derivatives with third-order tensors:

$$\begin{aligned} &\frac{\partial P(\bar{x}(\theta), \theta)}{\partial x} \in \mathbb{R}^{n\times m\times n}, \\ &\frac{\partial P(\bar{x}(\theta), \theta)}{\partial \theta} \in \mathbb{R}^{n\times m\times p}. \end{aligned} \tag{3.1}$$

Left multiplication by these derivatives is considered as follows, for arguments of compatible size: for any $c \in \mathbb{R}^n$, $\frac{\partial P(\bar{x}(\theta),\theta)}{\partial x} c \in \mathbb{R}^{n\times m}$ and for any $M \in \mathbb{R}^{n\times q}$, for some $q \in \mathbb{N}$, $\frac{\partial P(\bar{x}(\theta),\theta)}{\partial x} M \in \mathbb{R}^{n\times m\times q}$, both operations being compatible with the usual identification of vectors as single rows in $\mathbb{R}^{n\times 1}$. This multiplication is assumed to be compatible with the rules of differential calculus, for example, if $v\colon \mathbb{R}^p \to \mathbb{R}^n_{>0}$ is C^1, then we have the identity, for any $\theta \in \mathbb{R}^p$,

$$\frac{\partial}{\partial\theta} P(v(\theta), \theta) = \frac{\partial P(v(\theta), \theta)}{\partial x} \frac{dv(\theta)}{d\theta} + \frac{\partial P(\bar{x}(\theta), \theta)}{\partial \theta} \in \mathbb{R}^{n\times m\times p}. \tag{3.2}$$

The operation is also invariant with order of products, if $M = uv^T$, then

$$\frac{\partial P(\bar{x}(\theta), \theta)}{\partial x} M = \frac{\partial P(\bar{x}(\theta), \theta)}{\partial x} \left(uv^T\right) = \left(\frac{\partial P(\bar{x}(\theta), \theta)}{\partial x} u\right) v^T.$$

3.2. Spectral pseudo-inverse. In order to explicitly describe the derivative of $\hat{P}(\theta)$, we will use the following notion of pseudo-inverse of a diagonalizable matrix.

DEFINITION 3.1 (Spectral pseudo-inverse [29, 3]). *Given a diagonalizable matrix $M \in \mathbb{R}^{n\times n}$, let $M = QDQ^{-1}$ its diagonalization, where $Q \in \mathbb{R}^{n\times n}$ is invertible and $D \in \mathbb{R}^{n\times n}$ is diagonal. The* spectral pseudo-inverse *of M is given by $M^\sharp = QD^\dagger Q^{-1}$ where $\dagger$ denotes Moore-Penrose pseudo-inverse.*

The Moore-Penrose $D^\dagger$ pseudo-inverse of a diagonal matrix $D \in \mathbb{R}^{n\times n}$ is given by $(D^\dagger)_{ii} = (D_{ii})^{-1}$ if $(D_{ii}) \neq 0$ and 0 otherwise. The key property of the spectral pseudo-inverse is that it preserves the eigenspaces of M, contrary to the more standard Moore–Penrose pseudo-inverse.

LEMMA 3.2 (Eigenspaces presevation of spectral pseudo-inverse [29]). *Let $M \in \mathbb{R}^{n\times n}$ a diagonalizable matrix. Then, M and $M^\sharp$ have the same kernel and the remaining eigenspaces are the same with inverse eigenvalues.*

Note that this definition and result are defined even for non-diagonalizable matrices in [29] using its Jordan reduced form, but for the sake of our results, we only need this property for diagonalizable matrices.

3.3. Main result. Our contribution is the following.

THEOREM 3.3 (The derivatives of Sinkhorn–Knopp converge). *Under Assumption 2.1, let $\bar{x}(\theta)$ the limit of Sinkhorn–Knopp iterations (SK_θ) initialized by $x_0(\theta)$ for all $\theta \in \Omega$.*

Then, the optimal coupling matrix $\hat{P}$ is continuously differentiable and its derivative $\frac{d\hat{P}(\theta)}{d\theta} \in \mathbb{R}^{n\times m\times p}$ is given by

$$\frac{d\hat{P}(\theta)}{d\theta} = \frac{\partial P(\bar{x}(\theta),\theta)}{\partial x}(I - A(\theta))^\sharp B(\theta) + \frac{\partial P(\bar{x}(\theta),\theta)}{\partial \theta}$$

where $A(\theta)$, $B(\theta)$ are the components of the total derivative of F at $(\bar{x}(\theta),\theta)$, i.e.,

$$[A(\theta)\ B(\theta)] = J_F(\bar{x}(\theta),\theta),$$

F (resp. P) is defined in (SK_θ) (resp. (2.4)), and partial derivatives of P are described in Section 2.3. Here $\sharp$ denotes the spectral pseudo-inverse of a diagonalizable matrix (Definition 3.1).

Furthermore, P_k is continuously differentiable for all k and the sequence of derivatives $\frac{dP_k}{d\theta}$ converges at a linear rate, locally uniformly in θ. In particular, for all $\theta \in \Omega$,

$$\boxed{\lim_{k\to+\infty} \frac{dP_k}{d\theta}(\theta) = \frac{d\hat{P}}{d\theta}(\theta).}$$

Remark 3.4 (Relation to previous works). The differentiability of the Sinkhorn–Knopp iterations is an elementary and well-known fact, used for example in [1], the new contribution here being that the derivatives converge toward the derivative of entropic regularization (OT_θ). Using an alternative formulation (in the context of implicit differentiation), [14] proves the differentiability of the entropic regularization of OT (first part of Theorem 3.3), and obtained an alternative expression of the derivative. They do not however prove the convergence of the derivatives, that is the main concern of our work and the expression for the derivative in Theorem 3.3 was not mentioned in previous literature, to our knowledge.

If F was a strict contraction mapping, applying [18, Proposition 1] would be sufficient to conclude and obtain the same expression as in Theorem 3.3 with an inverse instead of the spectral pseudo-inverse. This is unfortunately not the case, and a more refined analysis is necessary to obtain the convergence. The main intuition behind this analysis is that Sinkhorn iterations are equivariant with respect to scaling of $u = \exp(x)$, and the optimal solution P in (2.4) is invariant with respect to the same scaling. In terms of derivative, it produces a lack of invertibility of $\frac{\partial F(x,\theta)}{\partial x}$ but the corresponding direction does not depend on (x, θ), and precisely lies in the kernel of $\frac{\partial P(x,\theta)}{\partial x}$ for all (x, θ). This "alignment" allows to maintain an overall convergence of derivatives. Section 4 is dedicated to prove this intuition rigorously.

Remark 3.5 (Limitations of our result). Despite the generality of Theorem 3.3, we would like to point out two limitations:

1. We do *not* have any guarantees for the convergence of the derivatives of the iterates $x_k(\theta)$, $k \in \mathbb{N}$. Said otherwise, we have guarantees for the derivatives of the optimal transport plan P_k, not for the derivatives of the scaling factors u_k, v_k, or the derivatives of the reduced variable x_k.
2. Inspecting the proof of Theorem 3.3, the linear convergence factor is a $(\bar{\rho})^{\frac{1}{2}}$ where $\bar{\rho}$ is an upper bound on both the linear convergence factor of the iterates (Lemma 2.2) and the second largest eigenvalue of $\frac{\partial F}{\partial x}$ at the solution, call it λ. Classical discrete dynamical system arguments (see [26, Remark 4.5] on local linear convergence) suggest that the linear convergence factor of the iterates is asymptotically of order λ. Taking this into consideration, our proof suggest an asymptotic linear convergence factor of the order $\sqrt{\lambda}$ for the derivatives, a factor strictly greater than that of the sequence. This discrepancy is a consequence of Lemma 5.2 which we use for simplicity of the presentation which requires a *non-asymptotic* analysis to ensure uniformity in θ. However, removing uniformity, this could be improved to obtain pointwise an asymptotic linear convergence factor arbitrarily close to λ using Lemma 6.4 instead, combined with arguments outlined in [26, Remark 4.5].

Remark 3.6 (Application to automatic differentiation of Sinkhorn–Knopp). Given $k \in \mathbb{N}$ and $\dot{\theta} \in \mathbb{R}^p$, forward automatic differentiation [33] allows to evaluate $\dot{P}_k = \frac{dP_k(\theta)}{d\theta}\dot{\theta} \in \mathbb{R}^{n\times m}$, *e.g.,* Jacobian-Vector Products (JVP), just by implementing (SK_θ). Similarly, given $\bar{w}_k \in \mathbb{R}^{n\times m}$, the reverse mode of automatic differentiation [22], also called backpropagation, computes $\bar{\theta}_k^T = \bar{w}_k^T \frac{dP_k(\theta)}{d\theta} \in \mathbb{R}^p$, *e.g.,* a Vector-Jacobian Product (VJP). Using a similar argument as in [5], it is possible, thanks to Theorem 3.3, to prove the convergence of these quantities. Note that in practice, the object of interest is not necessarily P_k by itself, but its composition by another function, *e.g.,* $\langle C(\theta), P_k(\theta)\rangle$ to compute the primal Sinkhorn divergence, $\langle C(\theta), P_k(\theta)\rangle - \mathrm{Ent}(P_k(\theta))$ to compute the OT loss, a sum of similar terms when dealing with Wasserstein barycenters [2], or any function $L(P_k(\theta))$ where $L : \mathbb{R}^{n\times m} \to \mathbb{R}^k$ is a continuously differentiable function. Applying our result (Theorem 3.3) and the chain rule leads to a similar result.

Remark 3.7 (Differentiation with respect a, b, C or ϵ). Theorem 3.3 is presented with an abstract parameterization of the problem with variable $\theta \in \mathbb{R}^q$. Choosing different values for θ allows to obtain derivatives of P_k for $k \in \mathbb{N}$ as well as $\hat{P}$ with respect to the original transport problem data: a, b, C or ϵ. These are typically evaluated numerically by algorithmic differentiation, but one could get closed form

expressions in simple cases. For example choosing $\theta = a$, we have

$$\frac{\partial F(x,\theta)}{\partial a} = \operatorname{diag}\left(\frac{1}{a}\right).$$

Similarly, setting $\theta = b$, we have

$$\frac{\partial F(x,\theta)}{\partial b} = -\operatorname{diag}\left(\frac{1}{K\frac{b}{K^T e^x}}\right) K \operatorname{diag}\left(\frac{1}{K^T e^x}\right).$$

Similarly one could compute derivatives with respect to the cost matrix C or ϵ, but the corresponding expressions become more complicated, and the use of automatic differentiation alleviates this difficulty in practice.

Remark 3.8 (Numerical illustration). Figure 1 illustrates a simple example where C is an Euclidean cost matrix between two point clouds X, Y in $\mathbb{R}^2$ of size $n_X = 100$ and $n_Y = 50$. The starting point cloud X follows a uniform law in the square $[-1/2, 1/2]$ and the target Y a uniform law on a circle inscribed in the square. The marginals are two uniform histograms $a = 1_n/n$ and $b = 1_m/m$. Sinkhorn–Knopp algorithm (SK_θ) is automatically differentiated with the Python library `jax` [7] with respect to the parameter ϵ, and we record the median of 10 trials for $\epsilon = 10^{-2}$. The blue filled area represents the first and last deciles. We run the algorithm for a high number of iterations N_{it} and display both

$$\left\| P_k(\epsilon) - \hat{P}(\epsilon) \right\| \quad \text{and} \quad \left\| \frac{dP_k}{d\epsilon}(\epsilon) - \frac{d\hat{P}}{d\epsilon}(\epsilon) \right\|.$$

Note we assume here that $P_{N_{\mathrm{it}}}(\epsilon)$ (resp. $\frac{dP_{N_{\mathrm{it}}}}{d\epsilon}(\epsilon)$) is close enough the optimal solution $\hat{P}(\epsilon)$ (resp. $\frac{d\hat{P}}{d\epsilon}(\epsilon)$) such that it is a good proxy. In particular, we ran (SK_θ) up to machine precision.

4. Proof of Theorem 3.3. Before diving into the proof, we are going to provide important spectral properties of the Jacobian of the transport plan (Section 4.1), then introduce a proxy G of the Jacobian of F that is a firmly nonexpansive mapping in contrast of $\frac{dF}{dx}$ (Section 4.2) and finally rewrite (3.2) thanks to G (Section 4.3).

4.1. Eigendecomposition of the transport plan and Jacobian. The following lemma provides important properties of the Jacobians of P and F as a function of x. Here θ is fixed and we look at properties of the derivative with respect to x, hence the dependency in θ does not appear.

LEMMA 4.1 (Expression of the Jacobian of $F(x)$). *Let $x \in \mathbb{R}^n$.*

1. *We have $\frac{dP(x)}{dx} 1_n = 0_{n\times m}$, where the product is described in Section 2.3.*
2. *The Jacobian $\frac{dF(x)}{dx}$ of F reads*

$$\begin{aligned}\frac{dF(x)}{dx} &= \operatorname{diag}\left(\frac{1}{K\left(\frac{b}{K^T e^x}\right)}\right) K \operatorname{diag}\left(\frac{b}{\left(K^T e^x\right)^2}\right) K^T \operatorname{diag}\left(e^x\right) \\ &= \operatorname{diag}\left(e^{F(x)}\right) \operatorname{diag}\left(\frac{1}{a \odot e^x}\right) P(x) \operatorname{diag}\left(\frac{1}{b}\right) P^T(x).\end{aligned}$$

Proof.

1. We note that $P(x+\lambda 1_n) = P(x)$ for all $\lambda \in \mathbb{R}$ so that $(P(x+\lambda 1_n) - P(x))/\lambda = \lambda \frac{dP(x)}{dx} 1_n + o(\lambda) = 0$. This implies that $\frac{dP(x)}{dx} 1_n = 0$.
2. The first expression is a direct computation observing that if $f : \mathbb{R}^n \to \mathbb{R}^n$ is an entry-wise function, then $J_f(x) = \mathrm{diag}(f'(x))$ where f' is again applied entry-wise. Indeed, we have for $x \in \mathbb{R}^n$, $\frac{de^x}{dx} = \mathrm{diag}(e^x)$, which in turns gives $\frac{dK^T e^x}{dx} = K^T \mathrm{diag}(e^x)$. Then, we obtain the derivatives of the ratio

$$\frac{d\frac{b}{K^T e^x}}{dx}(x) = -\mathrm{diag}\left(\frac{b}{(K^T e^x)^2}\right) K^T \mathrm{diag}(e^x).$$

Similarly, since K is a linear operator,

$$\frac{d\left(K\frac{b}{K^T e^x}\right)}{dx}(x) = -K\mathrm{diag}\left(\frac{b}{(K^T e^x)^2}\right) K^T \mathrm{diag}(e^x).$$

Finally, since $\frac{d\log(g(x))}{dx} = \frac{dg(x)}{dx} \odot \frac{1}{g(x)}$, for a differentiable $g : \mathbb{R}^n \to \mathbb{R}^n$, we obtain that

$$\frac{dF(x)}{dx} = \mathrm{diag}\left(\frac{1}{K\left(\frac{b}{K^T e^x}\right)}\right) K\mathrm{diag}\left(\frac{b}{(K^T e^x)^2}\right) K^T \mathrm{diag}\left(e^x\right).$$

The second expression uses the definition of P in (2.4). Observe that

$$\mathrm{diag}\left(\frac{b}{(K^T e^x)^2}\right) K^T \mathrm{diag}\left(e^x\right) = \mathrm{diag}\left(\frac{1}{K^T e^x}\right) P^T(x), \tag{4.1}$$

and (using the fact that diagonal matrices commute)

$$\mathrm{diag}(e^{F(x)}) = \mathrm{diag}\left(\frac{1}{K\left(\frac{b}{K^T e^x}\right)}\right) \mathrm{diag}(a). \tag{4.2}$$

Observe now that,

$$K\mathrm{diag}\left(\frac{1}{K^T e^x}\right) = \mathrm{diag}\left(\frac{1}{e^x}\right) P(x) \mathrm{diag}\left(\frac{1}{b}\right). \tag{4.3}$$

Combining (4.1), (4.2) and (4.3) gives the result. ∎

Remark 4.2. If $x = F(x)$, at a fixed point solution, the Jacobian expression in Lemma 4.1 can be simplified as follows

$$\frac{dF(x)}{dx} = \mathrm{diag}\left(\frac{1}{a}\right) P(x) \mathrm{diag}\left(\frac{1}{b}\right) P^T(x).$$

We have the following result on the eigenvalues and eigenvectors of $\frac{dF}{dx}$.

LEMMA 4.3 (Eigendecomposition of $\frac{dF}{dx}$). *If b and K have positive entries, then for any x, $\frac{dF(x)}{dx}$ is diagonalisable on $\mathbb{R}$. 1 is an eigenvalue with multiplicity 1 and the other eigenvalue have modulus strictly smaller than 1. Furthermore, one has the following eigenvectors:*

$$\frac{dF(x)}{dx} 1_n = 1_n$$
$$\left(\frac{dF(x)}{dx}\right)^T \frac{a \odot e^x}{e^{F(x)}} = \frac{a \odot e^x}{e^{F(x)}}$$

Proof. Fix $x \in \mathbb{R}^n$ and let

$$S = \operatorname{diag}\left(\frac{1}{K\left(\frac{b}{K^T e^x}\right)}\right), \; M = K \operatorname{diag}\left(\frac{b}{(K^T e^x)^2}\right) K^T, \text{ and } T = \operatorname{diag}\left(e^x\right).$$

The matrices S and T are diagonal with positive entries and M is symmetric such that $SMT = \frac{dF(x)}{dx}$. Setting $A = (TS^{-1})^{1/2}$, we have, using the fact that diagonal matrix commute

$$\begin{aligned} ASMTA^{-1} &= T^{\frac{1}{2}} S^{-\frac{1}{2}} SMTS^{\frac{1}{2}} T^{-\frac{1}{2}} \\ &= T^{\frac{1}{2}} S^{\frac{1}{2}} MS^{\frac{1}{2}} T^{\frac{1}{2}}, \end{aligned}$$

and therefore $A\frac{dF(x)}{dx}A^{-1}$ is real symmetric, hence diagonalisable with real eigenvalues. As a consequence, $\frac{dF(x)}{dx}$ being similar to $A\frac{dF(x)}{dx}A^{-1}$ it has the same property. It is an easy calculation to check that $\frac{dF(x)}{dx}1_n = 1_n$. Indeed, $T1_n = e^x$, and since $\operatorname{diag}(y)x = y \odot x$ for $x, y \in \mathbb{R}^n$, we have that $Me^x = K\frac{b}{K^T e^x}$ and then $SK\frac{b}{K^T e^x} = 1_n$. Hence, the multiplicity of the eigenvalue 1 as well as properties of the remaining eigenvalue is a consequence of Perron–Frobenius theorem [21, Theorem 8.2.8 and Theorem 8.3.4] applied to the stochastic matrix $\frac{dF(x)}{dx}$.

Let us prove the last identity. We have

$$\begin{aligned} e^{F(x)} &= \frac{a}{K\left(\frac{b}{K^T e^x}\right)}, \\ P(x)1_m &= \operatorname{diag}(e^x) K \left(\frac{b}{K^T e^x}\right) = \frac{a \odot e^x}{e^{F(x)}}, \\ P(x)^T 1_n &= \frac{b}{K^T e^x} \odot K^T e^x = b, \end{aligned}$$

from which we deduce

$$\begin{aligned} &\left(\frac{dF(x)}{dx}\right)^T \frac{a \odot e^x}{e^{F(x)}} \\ &= P(x)\operatorname{diag}\left(\frac{1}{b}\right) P^T(x) \operatorname{diag}\left(\frac{e^{F(x)}}{(a \odot e^x)}\right) \frac{a \odot e^x}{e^{F(x)}} \\ &= P(x)\operatorname{diag}\left(\frac{1}{b}\right) P^T(x) 1_n \\ &= P(x) 1_m \\ &= \frac{a \odot e^x}{e^{F(x)}}. \end{aligned}$$

This concludes the proof. ∎

4.2. Reduced partial Jacobian of F. For any $(x, \theta) \in \mathbb{R}^n \times \mathbb{R}^p$, we set

$$\begin{aligned} \alpha(x, \theta) &= 1_n^T \left(\frac{a(\theta) \odot e^x}{e^{F(x,\theta)}}\right) \\ v(x, \theta) &= \frac{1}{\alpha(x, \theta)} \frac{a(\theta) \odot e^x}{e^{F(x,\theta)}}. \end{aligned} \tag{4.4}$$

For any x, θ consider furthermore the block decomposition of the total derivative of F, $[A(x,\theta)\ B(x,\theta)] = J_F(x,\theta)$ and set

$$G(x,\theta) = A(x,\theta) - 1_n v(x,\theta)^T. \tag{4.5}$$

We call G the reduced partial Jacobian of F, note that G is continuously differentiable on $\mathbb{R}^n \times \Omega$. From Lemma 4.3, we have that 1_n is an eigenvector of $A(x,\theta)$ and $v(x,\theta)$ is an eigenvector of $A(x,\theta)^T$, both with eigenvalue 1, which has multiplicity 1, with $1_n^T v(x,\theta) = 1$. Therefore Lemma 6.1 ensures that the matrix $G(x,\theta)$ is diagonalisable in the same basis as $A(x,\theta)$ with the same eigenvalues, except eigenvalue 1 which is set to 0, and therefore its spectral radius is strictly less than 1. Later in the proof, we will study a recurrence involving A (which is not a contraction), and we will use an equivalent recurrence involving G (which is a contraction). Note that the functions J_F, P, A, B, G are continuously differentiable on $\mathbb{R}^n \times \Omega$.

The following lemma shows that J_F and G are invariant by the centering operation L_{center}, and more generally by translation of $\lambda 1_n$.

LEMMA 4.4 (Invariance by centering). *For all $\lambda \in \mathbb{R}$, $x \in \mathbb{R}^n$, and $\theta \in \Omega$, we have,*

$$\begin{aligned} F(x + \lambda 1_n, \theta) &= F(x,\theta) + \lambda 1_n, \\ J_F(x + \lambda 1_n, \theta) &= J_F(x,\theta), \\ v(x + \lambda 1_n, \theta) &= v(x,\theta), \\ G(x + \lambda 1_n, \theta) &= G(x,\theta). \end{aligned}$$

In particular, $J_F(L_{\text{center}}(x),\theta) = J_F(x,\theta)$ and $G(L_{\text{center}}(x),\theta) = G(x,\theta)$ where L_{center} is the centering operator introduced in Lemma 2.2.

Proof. We have for $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^n$,

$$\begin{aligned} F(x + \lambda 1_n, \theta) &= \log(a(\theta)) - \log\left(K(\theta)\left(\frac{b(\theta)}{K(\theta)^T e^{x+\lambda 1_n}}\right)\right) \\ &= \log(a(\theta)) - \log\left(K(\theta)\left(\frac{b(\theta)}{e^{\lambda} K(\theta)^T e^{x}}\right)\right) \\ &= \log(a(\theta)) - \log\left(e^{-\lambda} K(\theta)\left(\frac{b(\theta)}{K(\theta)^T e^{x}}\right)\right) \\ &= \log(a(\theta)) + \lambda 1_n - \log\left(K(\theta)\left(\frac{b(\theta)}{K(\theta)^T e^{x}}\right)\right) \\ &= F(x,\theta) + \lambda 1_n, \end{aligned}$$

which implies for all $\lambda \in \mathbb{R}$, $J_F(x + \lambda 1_n, \theta) = J_F(x,\theta)$. Observe now that

$$\begin{aligned} \frac{a(\theta) \odot e^{x+\lambda 1_n}}{e^{F(x+\lambda 1_n,\theta)}} &= \frac{a(\theta) \odot e^{\lambda} e^{x}}{e^{F(x,\theta)+\lambda 1_n}} \\ &= \frac{a(\theta) \odot e^{\lambda} e^{x}}{e^{\lambda} e^{F(x,\theta)}} \\ &= \frac{a(\theta) \odot e^{x}}{e^{F(x,\theta)}}. \end{aligned}$$

Thus, $\alpha(x + \lambda 1_n, \theta) = \alpha(x,\theta)$ and in turn, we get that $v(x + \lambda 1_n, \theta) = v(x,\theta)$.

To conclude, we have

$$\begin{aligned} G(x+\lambda 1_n,\theta) &= A(x+\lambda 1_n,\theta) - 1_n v(x+\lambda 1_n,\theta)^T \\ &= A(x,\theta) - 1_n v(x,\theta)^T = G(x,\theta), \end{aligned}$$

following the fact that $J_F(x+\lambda 1_n,\theta) = J_F(x,\theta)$, and in particular $A(x+\lambda 1_n,\theta) = A(x,\theta)$. ∎

4.3. Preliminary computation. We start with some computation and notations before providing the proof arguments. Setting for all $k \in \mathbb{N}$, and $\theta \in \mathbb{R}^p$, $[A_k(\theta)\ B_k(\theta)] = J_F(x_k(\theta),\theta)$ we have the piggyback recursion

$$\frac{dx_{k+1}(\theta)}{d\theta} = A_k(\theta)\frac{dx_k(\theta)}{d\theta} + B_k(\theta), \tag{4.6}$$

We have for all k and θ, using (3.2) for the total derivative of P,

$$\begin{aligned} \frac{dP_{k+1}(\theta)}{d\theta} &= \frac{\partial P(x_{k+1}(\theta),\theta)}{\partial x}\frac{dx_{k+1}(\theta)}{d\theta} + \frac{\partial P(x_{k+1}(\theta),\theta)}{\partial \theta} \\ &= \frac{\partial P(x_{k+1}(\theta),\theta)}{\partial x}\left(A_k(\theta)\frac{dx_k(\theta)}{d\theta} + B_k(\theta)\right) + \frac{\partial P(x_{k+1}(\theta),\theta)}{\partial \theta}. \end{aligned} \tag{4.7}$$

For all θ and all $k \in \mathbb{N}$, we have $A_k(\theta) = A(x_k(\theta),\theta)$, we set

$$G_k(\theta) = G(x_k(\theta),\theta) = A_k(\theta) - 1_n v(x_k(\theta),\theta)^T,$$

where G is defined in (4.5) and v is defined in (4.4). From Lemma 6.1, the matrix $G_k(\theta)$ is diagonalisable in the same basis as $A_k(\theta)$ with the same eigenvalues except eigenvalue 1 which is set to 0 and therefore its spectral radius is strictly less than 1.

From Lemma 4.1, we have $\frac{\partial P(x,\theta)}{\partial x} 1_n = 0_{n\times m}$ for all (x,θ) and therefore

$$\frac{\partial P(x,\theta)}{\partial x} G_k(\theta) = \frac{\partial P(x,\theta)}{\partial x} A_k(\theta) - \frac{\partial P(x,\theta)}{\partial x} 1_n v(x_k(\theta),\theta)^T = \frac{\partial P(x,\theta)}{\partial x} A_k(\theta).$$

Plugging this in (4.7), we obtain

$$\begin{aligned} \frac{dP_{k+1}(\theta)}{d\theta} &= \frac{\partial P(x_{k+1},\theta)}{\partial x}\left(A_k(\theta)\frac{dx_k}{d\theta} + B_k(\theta)\right) + \frac{\partial P(x_{k+1},\theta)}{\partial \theta} \\ &= \frac{\partial P(x_{k+1},\theta)}{\partial x}\left(G_k(\theta)\frac{dx_k}{d\theta} + B_k(\theta)\right) + \frac{\partial P(x_{k+1},\theta)}{\partial \theta}. \end{aligned}$$

This allows to rewrite the iterations equivalently, with $D_0 = \frac{dx_0}{d\theta}$, for all $k \geq 0$ and θ, using the product rule for partial derivatives of P defined in Section 2.3,

$$\begin{aligned} \frac{dP_k(\theta)}{d\theta} &= \frac{\partial P(x_k,\theta)}{\partial x} D_k(\theta) + \frac{\partial P(x_k,\theta)}{\partial \theta}, \\ D_{k+1}(\theta) &= G_k(\theta) D_k(\theta) + B_k(\theta). \end{aligned} \tag{4.8}$$

We are now ready to prove our main result.

Proof of Theorem 3.3.

Convergence of A_k, G_k and B_k. For all $\theta \in \Omega$, from Lemma 2.2, the centered iterates $(L_{\text{center}}(x_k(\theta)))_{k\in\mathbb{N}}$ converge with a linear rate to $L_{\text{center}}(\bar{x}(\theta))$ which is locally

uniform in θ. Furthermore F and G are infinitely differentiable jointly in $x \in \mathbb{R}^n$ and $\theta \in \Omega$ and therefore J_F and G are locally Lipschitz on $\mathbb{R}^n \times \Omega$.

We remark that for all θ, using Lemma 4.4

$$G_k(\theta) = G(x_k(\theta), \theta) = G(L_{\text{center}}(x_k(\theta)), \theta),$$

so that, as $k \to \infty$, $G_k(\theta)$ converges with a locally uniform linear rate to $G(\theta) := G(L_{\text{center}}(\bar{x}(\theta)), \theta) = G(\bar{x}(\theta), \theta)$. Similarly $B_k(\theta)$ converges with a locally uniform linear rate to $B(\theta) := B(\bar{x}(\theta), \theta)$ and $A_k(\theta)$ converges with a locally uniform linear rate to $A(\theta) := A(\bar{x}(\theta), \theta)$. Note that by Lemma 2.2, the map $\theta \mapsto L_{\text{center}}(\bar{x}(\theta))$ is continuous, so that A, G and B are continuous functions of θ.

For any θ $G(\theta)$ is digaonalizable with spectral radius strictly less than 1, the recursion on $D_k(\theta)$ should converge with a locally uniformly linear rate in θ. This assertion is a consequence of the following lemma which explicit the constants appearing in the linear rate for the matrix recursion.

LEMMA 4.5 (Explicit rate for linear convergence). *Let $\rho < 1$ and $\bar{G} \in \mathbb{R}^{n \times n}$ be diagonalisable on $\mathbb{R}$, with spectral radius smaller than ρ and and Q an invertible matrix which rows are made of an eigenbasis of $\bar{G}$. Let $\bar{B} \in \mathbb{R}^{n \times m}$. Let $(G_k)_{k \in \mathbb{N}}$ and $(B_k)_{k \in \mathbb{N}}$ be a sequence of matrices such that there exists a constant $c_1 > 0$ such that for all $k \in \mathbb{N}$,*

$$\|G_k - \bar{G}\|_{\text{op}} \leq c_1 \rho^{k+1}, \tag{4.9}$$

$$\|B_k - \bar{B}\| \leq c_1 \rho^{k+1}. \tag{4.10}$$

Then, for the recursion

$$D_{k+1} = G_k D_k + B_k,$$

setting $\bar{D} = (I - \bar{G})^{-1} \bar{B}$, there exists a continuous function $\text{const} \colon \mathbb{R}^5_{\geq 0} \times (0,1) \to \mathbb{R}_{\geq 0}$ *such that for all $k \in \mathbb{N}$,*

$$\|D_k - \bar{D}\| \leq \rho^{\frac{k}{2}} \text{const}(\|Q\|_{\text{op}}, \|Q^{-1}\|_{\text{op}}, c_1, \|D_0\|, \|\bar{B}\|, \rho).$$

Convergence of D_k. Let us explicit how Lemma 4.5 allows to prove convergence of $(D_k(\theta))_{k \in \mathbb{N}}$. Start with a fixed $\theta \in \Omega$, we first drop the dependency in θ for clarity. We have from Remark 4.2

$$A = \text{diag}\left(\frac{1}{a}\right) \hat{P} \text{diag}\left(\frac{1}{b}\right) \hat{P}^T.$$

Setting $S = \text{diag}\left(\frac{1}{\sqrt{a}}\right)$, we have that

$$S^{-1} A S = \text{diag}\left(\frac{1}{\sqrt{a}}\right) \hat{P} \text{diag}\left(\frac{1}{b}\right) \hat{P}^T \text{diag}\left(\frac{1}{\sqrt{a}}\right),$$

which is symmetric. Therefore, there is an orthogonal matrix $U($ and diagonal matrix E such that

$$S^{-1} A S = U E U^T,$$

and

$$A = SUEU^T S^{-1} = SUE(SU)^{-1}.$$

Set $Q = SU$, we have by submultiplicativity of $\|\cdot\|_{\mathrm{op}}$

$$\|Q\|_{\mathrm{op}} \leq \|U\|_{\mathrm{op}}\|S\|_{\mathrm{op}} = \|S\|_{\mathrm{op}} = \left\|\frac{1}{\sqrt{a}}\right\|_\infty.$$

Similarly $\|Q^{-1}\|_{\mathrm{op}} = \|\sqrt{a}\|_\infty$. From Lemma 6.1, Q diagonalizes both A and G.

Getting back the dependency in θ, we fix $\theta_0 \in \Omega$, and set for all $\theta \in \Omega$

$$\begin{aligned}\bar{D}&\colon \theta \mapsto (I - G(\theta))^{-1}B(\theta),\\ \bar{\rho}&\colon \theta \mapsto \max\{\rho(\theta), \|Q(\theta)^{-1}G(\theta)Q(\theta)\|_{\mathrm{op}}\} < 1,\end{aligned}$$

where $\rho(\theta) < 1$ is given in Lemma 2.2 and $\|Q(\theta)^{-1}G(\theta)Q(\theta)\|_{\mathrm{op}}$ is the largest eigenvalue, in absolute value, of $G(\theta)$, which is smaller than 1 and continuous with respect to θ. In particular, $\bar{\rho}$ is continuous.

Fix a compact set $V \subset \Omega$ which contains θ_0 in its interior and a compact set $W \subset \mathbb{R}^n$ which contains $L_{\mathrm{center}}(x_k(\theta))$ for all $k \in \mathbb{N}$ and $\theta \in V$. We set $c_1 \colon \Omega \to \mathbb{R}_{\geq 0}$ such that $c_1 = Lc/\rho$ where $c\colon \Omega \to \mathbb{R}_{\geq 0}$ is the constant in Lemma 2.2 and L is a Lipschitz constant of J_F and G on $W \times V$ (recall that they are infinitely differentiable). Using Lemma 4.4, we have for all $\theta \in V$ and $k \in \mathbb{N}$,

$$\begin{aligned}\|J_F(x_k(\theta),\theta) - J_F(\bar{x}(\theta),\theta)\| &= \|J_F(L_{\mathrm{center}}(x_k(\theta)),\theta) - J_F(L_{\mathrm{center}}(\bar{x}(\theta)),\theta)\|\\ &\leq c_1(\theta)\bar{\rho}(\theta)^{k+1},\end{aligned}$$

and

$$\begin{aligned}\|G(x_k(\theta),\theta) - G(\bar{x}(\theta),\theta)\| &= \|G(L_{\mathrm{center}}(x_k(\theta)),\theta) - G(L_{\mathrm{center}}(\bar{x}(\theta)),\theta)\|\|\\ &\leq c_1(\theta)\bar{\rho}(\theta)^{k+1}.\end{aligned}$$

The largest eigenvalue of $G(\theta)$ is at most $\bar{\rho}(\theta)$ so that Lemma 4.5 applies, and we have for all $k \in \mathbb{N}$,

$$\begin{aligned}&\|D_k(\theta) - \bar{D}(\theta)\|\\ &\leq \bar{\rho}(\theta)^{\frac{k}{2}}\,\mathrm{const}\left(\left\|\frac{1}{\sqrt{a(\theta)}}\right\|_\infty, \left\|\sqrt{a(\theta)}\right\|_\infty, c_1(\theta), \left\|\frac{dx_0(\theta)}{d\theta}\right\|, \|B(\theta)\|, \bar{\rho}(\theta)\right),\end{aligned}$$

where $\mathrm{const}\colon \mathbb{R}^5_{\geq 0}\times(0,1)$ is continuous. All terms in the right hand side are continuous functions of θ, so that $D_k(\theta) \to \bar{D}(\theta) = (I - G(\theta))^{-1}B(\theta)$ at a locally uniform linear convergence rate.

Convergence of the derivatives of Sinkhorn–Knopp towards the derivatives of entropic regularization. For a fixed θ_0, the limit has to satisfy $\bar{D}(\theta_0) = G(\theta_0)\bar{D}(\theta_0) + B(\theta_0)$ since it is a fixed point of the limiting recursion. Therefore it is of the form

$$\begin{aligned}\bar{D}(\theta_0) &= (I - G(\theta_0))^{-1}B(\theta_0)\\ [A(\theta_0)\ B(\theta_0)] &= J_F(\bar{x}(\theta_0),\theta_0).\end{aligned}$$

Combining with local linear convergence of D_k, this shows that, using the recursion (4.8), as $k \to \infty$ uniformly in a neighborhood of θ_0

$$\lim_{k\to\infty} \frac{d}{d\theta}P_k(\theta) = \frac{\partial P(\bar{x}(\theta),\theta)}{\partial x}\bar{D}(\theta) + \frac{\partial P(\bar{x}(\theta),\theta)}{\partial \theta}. \tag{4.11}$$

Note that $P_k(\theta)$ converges pointwise towards $\hat{P}(\theta) = P(\bar{x}(\theta), \theta)$ which is a solution to problem (OT_θ). By local uniform convergence of derivatives and the fact that P_k are continuously differentiable, thanks to Lemma 6.2, we have that $\hat{P}$ is continuously differentiable and

$$\lim_{k\to\infty} \frac{dP_k(\theta)}{d\theta} = \frac{d\hat{P}(\theta)}{d\theta}.$$

Expression of the derivative. Finally, by construction of G in (4.5) and thanks to Lemma 6.1, we have for all x, θ, that $I - A(x,\theta)$ and $I - G(x,\theta)$ have the same eigenspaces all eigenvalues being nonzero except the one generated by 1_n for which corresponds to eigenvalue 0 for $I - A(x,\theta)$ and 1 for $I - G(x,\theta)$. Therefore, we have $(I - G(x,\theta)))^{-1} = (I - A(x,\theta))^\sharp + 1_n v(x,\theta)^T$, where $v(x,\theta)$ is the normalized eigenvector of $A(\theta)^T$ associated to eigenvalue 1 (see (4.4)). Recall that $\sharp$ denotes the spectral pseudo-inverse for diagonalisable matrices (Definition 3.1). From Lemma 4.1, we have $\frac{dP(x,\theta)}{dx} 1_n = 0$ for all (x,θ), therefore for all $\theta \in \Omega$,

$$\begin{aligned}
\frac{\partial P(\bar{x}(\theta),\theta)}{\partial x} \bar{D}(\theta) &= \frac{\partial P(\bar{x}(\theta),\theta)}{\partial x} (I - G(\bar{x}(\theta),\theta))^{-1} B(\theta), \\
&= \frac{\partial P(\bar{x}(\theta),\theta)}{\partial x} ((I - A(\theta))^\sharp + 1_n v(x,\theta)^T) B(\theta) \\
&= \frac{\partial P(\bar{x}(\theta),\theta)}{\partial x} (I - A(\theta))^\sharp B(\theta).
\end{aligned}$$

We have therefore that

$$\begin{aligned}
\frac{d\hat{P}(\theta)}{d\theta} &= \frac{\partial P(\bar{x}(\theta),\theta)}{\partial x} (I - A(\theta))^\sharp B(\theta) + \frac{\partial P(\bar{x}(\theta),\theta)}{\partial \theta}, \\
[A(\theta)\ B(\theta)] &= J_F(\bar{x}(\theta),\theta),
\end{aligned}$$

which concludes the proof. ∎

5. Proof of Lemma 4.5. We start with two lemmas on real sequences. The first one is a quantitative version of [27, Lemma 9, Chapter 2].

LEMMA 5.1 (Quantitative Gladyshev convergence). *Let $(\alpha_k)_{k\in\mathbb{N}}$ and $(\beta_k)_{k\in\mathbb{N}}$ be positive summable sequences and $(z_k)_{k\in\mathbb{N}}$ be a positive squence such that for all $k \in \mathbb{N}$*

$$z_{k+1} \leq (1 + \alpha_k) z_k + \beta_k.$$

Then for all $k \in \mathbb{N}$,

$$z_k \leq \exp\left(\sum_{i=0}^{+\infty} \alpha_i\right) \left(z_0 + \sum_{j=0}^{+\infty} \beta_j\right)$$

Proof. For all $k \in \mathbb{N}$, set

$$w_k = z_k \prod_{i=k}^{+\infty} (1 + \alpha_i) + \sum_{i=k}^{+\infty} \beta_i \prod_{j=i+1}^{+\infty} (1 + \alpha_j).$$

Remark that using concavity of logarithm $\prod_{i=0}^{+\infty}(1+\alpha_i) \leq \exp\left(\sum_{i=0}^{+\infty} \alpha_i\right)$, so that w_k is well defined. Remark also that $w_k \geq z_k$ for all k.

The sequence $(w_k)_{k\in\mathbb{N}}$ is decreasing. Indeed, we have for all $k \in \mathbb{N}$,

$$\begin{aligned} w_{k+1} &= z_{k+1} \prod_{i=k+1}^{+\infty} (1+\alpha_i) + \sum_{i=k+1}^{+\infty} \beta_i \prod_{j=i+1}^{+\infty} (1+\alpha_j) \\ &\leq ((1+\alpha_k)z_k + \beta_k) \prod_{i=k+1}^{+\infty} (1+\alpha_i) + \sum_{i=k+1}^{+\infty} \beta_i \prod_{j=i+1}^{+\infty} (1+\alpha_j) \\ &= z_k \prod_{i=k}^{+\infty} (1+\alpha_i) + \sum_{i=k}^{+\infty} \beta_i \prod_{j=i+1}^{+\infty} (1+\alpha_j) \\ &= w_k. \end{aligned}$$

Therefore, for all $k \in \mathbb{N}$,

$$\begin{aligned} z_k &\leq w_k \\ &\leq w_0 \\ &= v_0 \prod_{i=0}^{+\infty} (1+\alpha_i) + \sum_{i=0}^{+\infty} \beta_i \prod_{j=i+1}^{+\infty} (1+\alpha_j) \\ &\leq \prod_{i=0}^{+\infty} (1+\alpha_i) \left(v_0 + \sum_{i=0}^{+\infty} \beta_i \right) \leq \exp\left(\sum_{i=0}^{+\infty} \alpha_i \right) \left(v_0 + \sum_{i=0}^{+\infty} \beta_i \right), \end{aligned}$$

and the result follows. ∎

The following lemma specify Lemma 5.1 when α_k and β_k are geometric sequences.

LEMMA 5.2 (Application of Gladyshev's convergence to geometric sequences). *Let $\rho \in (0,1)$, $c > 0$, and $(\delta_k)_{k\in\mathbb{N}}$ be a positive sequence such that for all $k \in \mathbb{N}$,*

$$\delta_{k+1} \leq (\rho + c\rho^{k+1})\delta_k + c\rho^{k+1}. \tag{5.1}$$

Then, $(\delta_k)_{k\in\mathbb{N}}$ is a geometric sequence: for all $k \in \mathbb{N}$,

$$\delta_k \leq \rho^{\frac{k}{2}} \exp\left(\frac{c\sqrt{\rho}}{1-\rho} \right) \left(\delta_0 + \frac{c\sqrt{\rho}}{1-\sqrt{\rho}} \right),$$

Proof. Dividing (5.1) on both sides by $c\rho^{(k+1)/2}$, we have for all $k \in \mathbb{N}$,

$$\begin{aligned} \frac{\delta_{k+1}}{c\rho^{\frac{k+1}{2}}} &\leq \frac{\delta_k}{c\rho^{\frac{k-1}{2}}} + \frac{\delta_k}{c\rho^{\frac{k}{2}}} \frac{c\rho^{k+1} c\rho^{\frac{k}{2}}}{c\rho^{\frac{k+1}{2}}} + \rho^{\frac{k+1}{2}} \\ &= \frac{\sqrt{\rho}\delta_k}{c\rho^{\frac{k}{2}}} + \frac{\delta_k}{c\rho^{\frac{k}{2}}} c\rho^{k+\frac{1}{2}} + \rho^{\frac{k+1}{2}} \\ &\leq \frac{\delta_k}{c\rho^{\frac{k}{2}}} (1 + c\rho^{k+\frac{1}{2}}) + \rho^{\frac{k+1}{2}}. \end{aligned}$$

Setting for all $k \in \mathbb{N}$,

$$z_k = \frac{\delta_k}{c\rho^{\frac{k}{2}}}, \quad \alpha_k = c\rho^{k+\frac{1}{2}}, \quad \text{and} \quad \beta_k = \rho^{\frac{k+1}{2}},$$

we may apply Lemma 5.1 to obtain the result. Note that $\sum_{i=0}^{+\infty} \alpha_i = \frac{c\sqrt{\rho}}{1-\rho}$ and $\sum_{i=0}^{+\infty} \beta_i = \frac{\sqrt{\rho}}{1-\sqrt{\rho}}$, so that for all $k \in \mathbb{N}$

$$\begin{aligned}\frac{\delta_k}{c\rho^{\frac{k}{2}}} = z_k &\leq \exp\left(\sum_{i=0}^{+\infty} \alpha_i\right)\left(z_0 + \sum_{j=0}^{+\infty} \beta_j\right) \\ &= \exp\left(\frac{c\sqrt{\rho}}{1-\rho}\right)\left(\frac{\delta_0}{c} + \frac{\sqrt{\rho}}{1-\sqrt{\rho}}\right),\end{aligned}$$

which is the desired result. ∎

LEMMA 5.3 (Reduced perturbated convergence). *Let $\rho < 1$ and $\bar{G} \in \mathbb{R}^{n\times n}$ have operator norm smaller than ρ and $\bar{B} \in \mathbb{R}^{n\times m}$. Let $(G_k)_{k\in\mathbb{N}}$ and $(B_k)_{k\in\mathbb{N}}$ be a sequence of matrices such that there exists a constant $c_0 > 0$ such that for all $k \in \mathbb{N}$,*

$$\begin{aligned}\|G_k - \bar{G}\|_{\mathrm{op}} &\leq c_0\rho^{k+1}, \\ \|B_k - \bar{B}\| &\leq c_0\rho^{k+1}.\end{aligned}$$

Then for the recursion

$$D_{k+1} = G_k D_k + B_k,$$

setting $\bar{D} = (I - \bar{G})^{-1}\bar{B}$, we have

$$\begin{aligned}&\|D_k - \bar{D}\| \\ \leq\ & \rho^{\frac{k}{2}} \exp\left(c_0\sqrt{\rho}\frac{1+\|\bar{B}\|}{(1-\rho)^2}\right)\left(\|D_0\| + \frac{\|\bar{B}\|}{1-\rho} + \frac{c_0\sqrt{\rho}(1+\|B\|)}{(1-\sqrt{\rho})^2}\right).\end{aligned}$$

Proof. Note that $\bar{G}$ is invertible and it follows that the potential limit is $\bar{D} = (I - \bar{G})^{-1}\bar{B}$, as it is a fixed point of the limiting recursion, $\bar{D} = \bar{G}\bar{D} + \bar{B}$. We rewrite the recursion as follows

$$\begin{aligned}D_{k+1} - \bar{D} &= G_k D_k + B_k - \bar{G}\bar{D} - \bar{B} \\ &= G_k(D_k - \bar{D}) + (G_k - \bar{G})\bar{D} + B_k - \bar{B}.\end{aligned}$$

Setting for all $k \in \mathbb{N}$, $\delta_k = \|D_k - \bar{D}\|$, using the fact that $\|\cdot\|_{\mathrm{op}}$ is subordinate to $\|\cdot\|$, we have the recursion,

$$\begin{aligned}\delta_{k+1} &\leq \|G_k(D_k - \bar{D})\| + \|(G_k - \bar{G})\bar{D}\| + \|B_k - \bar{B}\| \\ &\leq \|G_k\|_{\mathrm{op}}\|(D_k - \bar{D})\| + \|(G_k - \bar{G})\|_{\mathrm{op}}\|\bar{D}\| + \|B_k - \bar{B}\| \\ &\leq (\rho + c_0\rho^{k+1})\delta_k + c_0\rho^{k+1}(\|\bar{D}\| + 1).\end{aligned}$$

Note that $\|\bar{D}\| = \|(I - \bar{G})^{-1}\bar{B}\| \leq \|(I - \bar{G})^{-1}\|_{\mathrm{op}}\|\bar{B}\| \leq \frac{\|\bar{B}\|}{1-\rho}$. Since

$$c_0 \leq c_0\frac{1+\|\bar{B}\|}{1-\rho} \quad \text{and} \quad c_0\left(1 + \frac{\|\bar{B}\|}{1-\rho}\right) \leq c_0\frac{1+\|\bar{B}\|}{1-\rho},$$

we apply Lemma 5.2 with $c = c_0\frac{1+\|\bar{B}\|}{1-\rho}$ and use the fact that $\frac{1}{1-\rho} \leq \frac{1}{1-\sqrt{\rho}}$ and $\|D_0 - \bar{D}\| \leq \|D_0\| + \frac{\|\bar{B}\|}{1-\rho}$. ∎

Proof of Lemma 4.5. Note that $\bar{G}$ is invertible and it follows that the potential limit is $\bar{D} = (I - \bar{G})^{-1}\bar{B}$, which satisfy $\bar{D} = \bar{A}\bar{D} + \bar{B}$. Since $\bar{G}$ is diagonalisable in the basis given by Q, there is a diagonal matrix E such that $\bar{G} = QEQ^{-1}$. We rewrite equivalently the recursion as follows

$$Q^{-1}D_{k+1} = Q^{-1}G_kQQ^{-1}D_k + Q^{-1}B_k,$$

and setting $\tilde{D}_k = Q^{-1}D_k$, $\tilde{G}_k = Q^{-1}G_kQ$ and $\tilde{B}_k = Q^{-1}B_k$ for all $k \in \mathbb{N}$, this reduces to

$$\tilde{D}_{k+1} = \tilde{G}_k\tilde{D}_k + \tilde{B}_k.$$

When $k \to \infty$, we have $\tilde{G}_k \to E$, which has operator norm at most ρ and $\tilde{B}_k \to Q^{-1}\bar{B}$. Set $\tilde{D}$ the fixed point of the limiting recursion for $\tilde{D}_k$,

$$\tilde{D} = (I - E)^{-1}Q^{-1}\bar{B} = Q^{-1}Q(I - E)^{-1}Q^{-1}\bar{B} = Q^{-1}(I - QEQ^{-1})^{-1}\bar{B} = Q^{-1}\bar{D}.$$

Furthermore for all $k \in \mathbb{N}$, we have the following bounds

$$\begin{aligned}
\|\tilde{G}_k - E\|_{\mathrm{op}} &= \|Q^{-1}(G_k - \bar{G})Q\|_{\mathrm{op}} \\
&\leq \|Q^{-1}\|_{\mathrm{op}}\|(G_k - \bar{G})\|_{\mathrm{op}}\|Q\|_{\mathrm{op}} && (\|\cdot\|_{\mathrm{op}} \text{ is submultiplicative}) \\
&\leq \left(c_1\|Q^{-1}\|_{\mathrm{op}}\|Q\|_{\mathrm{op}}\right)\rho^{k+1} && (\text{by hypothesis (4.9)}) \\
\|\tilde{B}_k - Q^{-1}\bar{B}\| &= \|Q^{-1}(B_k - \bar{B})\| \\
&\leq \|Q^{-1}\|_{\mathrm{op}}\|B_k - \bar{B}\| && (\|\cdot\|_{\mathrm{op}} \text{ is subordinate to } \|\cdot\|) \\
&\leq \left(c_1\|Q^{-1}\|_{\mathrm{op}}\right)\rho^{k+1}, && (\text{by hypothesis (4.10)}) \\
\|Q^{-1}\bar{B}\| &= \|Q^{-1}\|_{\mathrm{op}}\|\bar{B}\| \\
\|\tilde{D}_0\| &= \|Q^{-1}\|_{\mathrm{op}}\|D_0\|.
\end{aligned}$$

We apply Lemma 5.3 with

$$c_0 = c_1\|Q^{-1}\|_{\mathrm{op}}(1 + \|Q\|_{\mathrm{op}}),$$

which gives for all $k \in \mathbb{N}$,

$$\begin{aligned}
&\|D_k - \bar{D}\| \\
&= \|Q(\tilde{D}_k - \tilde{D})\| \\
&\leq \|Q\|_{\mathrm{op}}\|\tilde{D}_k - \tilde{D}\| \\
&\leq \rho^{\frac{k}{2}}\|Q\|_{\mathrm{op}}\exp\left(c_1\|Q^{-1}\|_{\mathrm{op}}(1 + \|Q\|_{\mathrm{op}})\sqrt{\rho}\frac{1 + \|Q^{-1}\|_{\mathrm{op}}\|\bar{B}\|}{(1-\rho)^2}\right) \\
&\qquad \times \|Q^{-1}\|_{\mathrm{op}}\left(\|D_0\| + \frac{\|\bar{B}\|}{1-\rho} + \frac{c_1(1 + \|Q\|_{\mathrm{op}})\sqrt{\rho}(1 + \|\bar{B}\|)}{(1-\sqrt{\rho})^2}\right),
\end{aligned}$$

which is the desired result ∎

6. Additional lemmas. In the following, we prove some technical, but important, lemmas used in the main proof.

LEMMA 6.1 (Reduced eigenspace). *Let $A \in \mathbb{R}^{n\times n}$ be diagonalisable. Let u be such that $Au = u$ and v such that $A^Tv = v$, and assume that eigenvalue 1 is simple, and that $uv^T = 1$. Then $\tilde{A} := A - uv^T$ and A have the same eigenspaces with the same eigenvalues, except eigenvalue 1 for A which is set to 0 for $\tilde{A}$.*

Proof. A of the form QDQ^{-1} for an invertible Q and a diagonal matrix D. Assume that the first diagonal entry of D is 1. Columns of Q form an eigenbasis and we may impose that the first column is u. Rows of Q^{-1} form an eigenbasis of A^T, set v_0 the vector corresponding to the first row. Since 1 is a simple eigenvalue, the corresponding eigenspace has dimension 1 and there exists $\alpha \neq 0$ such that $v = \alpha v_0$. We have $u^T v = 1$ by assumption and $u^T v_0 = 1$ because $Q^{-1}Q = I$, this shows that $\alpha = 1$ and therefore v is the first row of Q^{-1}.

We have $v^T u = 1$ and therefore $\tilde{A}u = Au - u = 0$. Let $\tilde{u}$ be a different column of Q corresponding to an eigenvector of A associated to eigenvalue d, we have $v^T\tilde{u} = 0$ so that $\tilde{A}\tilde{u} = A\tilde{u} = d\tilde{u}$. This concludes the proof. ∎

LEMMA 6.2 (Uniform convergence leads to continuous differentiable limit). *Let $U \subset \mathbb{R}^p$ be open and $(f_k)_{k\in\mathbb{N}}$ be a sequence of continuously differentiable functions from U to $\mathbb{R}$ converging pointwise to $\bar{f}\colon U \to \mathbb{R}$, such that ∇f_k converges pointwise, locally uniformly on U. Then $\bar{f}$ is continuously differentiable on U and $\nabla\bar{f} = \lim_{k\to\infty}\nabla f_k$.*

Proof. Let $g = \lim_{k\to\infty}\nabla f_k$ the pointwise limit. By local uniform convergence, g is continuous on U. Fix any $x \in U$ and any $v \in \mathbb{R}^n$ and set I a closed interval such that $x + tv \in U$ for all $t \in I$. The sequence of univariate functions $h_k\colon t \mapsto f_k(x + tv)$ is continuously differentiable and satisfy for all k and all $t \in I$

$$h_k'(t) = \langle \nabla f_k(x + tv), v\rangle .$$

The derivatives h_k' converge uniformly on I to $\langle g(x + tv), v\rangle$ which is continuous in t. Therefore the function $\bar{h}\colon t \mapsto \bar{f}(x + tv)$ is continuously differentiable with derivative given by $\langle g(x + tv), v\rangle$. Since $x \in U$ and $v \in \mathbb{R}^n$ were arbitrary, this implies that $\bar{f}$ admits continuous partial derivatives and it is therefore continuously differentiable with gradient g. ∎

LEMMA 6.3 (Centering). *For $x, x' \in \mathbb{R}^n$,*

$$\|L_{\text{center}}(x) - L_{\text{center}}(x')\|_\infty \leq \|x - x'\|_{\text{var}},$$

where L_{center} is defined in (2.6) and $\|\cdot\|_{\text{var}}$ is defined in (2.7).

Proof. Note that for $f \in \mathbb{R}^n$ and $a \in \mathbb{R}$, $\|f + a1_n\|_{\text{var}} = \|f\|_{\text{var}}$. Set $f = L_{\text{center}}(x) - L_{\text{center}}(x')$, we have $1_n^T f = \sum_{i=1}^n f_i = 0$ so that

$$\min_i f_i \leq \sum_{i=1}^n f_i = 0 \leq \max_i f_i.$$

This implies the following

$$\begin{aligned}
&\|f\|_\infty = \max_i |f_i| \\
&= \max_i \max\{f_i, -f_i\} \\
&= \max\{\max_i f_i, \max_i -f_i\} \\
&= \max\{\max_i f_i, -\min_i f_i\} \\
&\leq \max\{\max_i f_i - \min_i f_i, \max_i f_i - \min_i f_i\} \\
&= \|f\|_{\text{var}}.
\end{aligned}$$

Now $f = L_{\text{center}}(x) - L_{\text{center}}(x') = x - x' + 1_n \left(\frac{1}{n}\sum_{i=1}^n x'_i - \frac{1}{n}\sum_{i=1}^n x_i\right)$, so that $\|f\|_{\text{var}} = \|x - x'\|_{\text{var}}$ which concludes the proof. ∎

LEMMA 6.4. *Let $\rho \in (0,1)$, $c > 0$ and $(\delta_k)_{k\in\mathbb{N}}$ be a positive sequence such that for all $k \in \mathbb{N}$,*

$$\delta_{k+1} \leq (\rho + c\rho^{k+1})\delta_k + c\rho^{k+1}. \tag{6.1}$$

Then, for all $k \in \mathbb{N}$, such that $k \geq \frac{\rho}{1-\rho}$, we have

$$\delta_k \leq \rho^k \exp\left(1 + \frac{c}{1-\rho}\right)(\delta_0 + c(k+1)).$$

Proof. Fix $\alpha \in (0,1)$ to be chosen latter. Dividing (6.1) on both sides by $c\rho^{(k+1)/\alpha}$, we have for all $k \in \mathbb{N}$,

$$\begin{aligned}
\frac{\delta_{k+1}}{c\rho^{\alpha(k+1)}} &\leq \frac{\delta_k}{c\rho^{\alpha k}}\left(\frac{\rho c\rho^{\alpha k}}{c\rho^{\alpha(k+1)}} + \frac{c\rho^{k+1}c\rho^{\alpha k}}{c\rho^{\alpha(k+1)}}\right) + \frac{c\rho^{k+1}}{c\rho^{\alpha(k+1)}} \\
&= \frac{\delta_k}{c\rho^{\alpha k}}\left(\rho^{1-\alpha} + c\rho^{k+1-\alpha}\right) + \rho^{(k+1)(1-\alpha)} \\
&\leq \frac{\delta_k}{c\rho^{\alpha k}}\left(1 + c\rho^{k+1-\alpha}\right) + \rho^{(k+1)(1-\alpha)}.
\end{aligned}$$

Setting for all $k \in \mathbb{N}$,

$$z_k = \frac{\delta_k}{c\rho^{\alpha k}}, \quad \alpha_k = c\rho^{k+1-\alpha}, \quad \text{and} \quad \beta_k = \rho^{(k+1)(1-\alpha)},$$

we apply Lemma 5.1 to obtain the result. As $(\alpha_k)_{k\in\mathbb{N}}$ and $(\beta_k)_{k\in\mathbb{N}}$ are geometric sequences, we have $\sum_{i=0}^{+\infty}\alpha_i = \frac{c\rho^{1-\alpha}}{1-\rho} \leq \frac{c}{1-\rho}$ and $\sum_{i=0}^{+\infty}\beta_i = \frac{\rho^{1-\alpha}}{1-\rho^{1-\alpha}} \leq \frac{1}{1-\rho^{1-\alpha}}$, so that for all $k \in \mathbb{N}$,

$$\begin{aligned}
\frac{\delta_k}{c\rho^{\alpha k}} &= z_k \\
&\leq \exp\left(\sum_{i=0}^{+\infty}\alpha_i\right)\left(z_0 + \sum_{j=0}^{+\infty}\beta_j\right) \\
&= \exp\left(\frac{c}{1-\rho}\right)\left(\frac{\delta_0}{c} + \frac{1}{1-\rho^{1-\alpha}}\right).
\end{aligned}$$

Since α was arbitrary, the preceding holds for all $k \in \mathbb{N}$ and $\alpha \in (0,1)$. Fix $k \in \mathbb{N}$ such that $k > \frac{\rho}{1-\rho}$. Setting $\alpha = 1 + \log\left(1 + \frac{1}{k}\right)/\log(\rho)$, since $\rho \in (0,1)$, we have

$$0 = 1 + \log\left(1 + \frac{1-\rho}{\rho}\right)/\log(\rho) < \alpha < 1.$$

We have

$$\rho^{\alpha k} = \rho^k \rho^{k\log((k+1)/k)/\log(\rho)} = \rho^k\left(1 + \frac{1}{k}\right)^k \leq e\rho^k,$$

and

$$\frac{1}{1-\rho^{1-\alpha}} = \frac{1}{1-\rho^{-\log(1+1/k)/\log(\rho)}} = \frac{1}{1-\rho^{\log(k/(k+1))/\log(\rho)}} = \frac{1}{1-\frac{k}{k+1}} = k+1.$$

Therefore, for all $k \geq \frac{\rho}{1-\rho}$,

$$\delta_k \leq \rho^k \exp\left(1+\frac{c}{1-\rho}\right)(\delta_0 + c(k+1)),$$

proving our claim. ∎

Acknowledgments. E.P. and S.V. would like to thank Jérôme Bolte for fruitful and inspiring discussions. S.V. thanks Jeremy Cohen and Titouan Vayer for raising the issue of the convergence of the derivatives of the Sinkhorn–Knopp iterates during a seminar in Lyon.

REFERENCES

[1] R. P. Adams and R. S. Zemel, *Ranking via sinkhorn propagation*, 2011, https://arxiv.org/abs/1106.1925.

[2] M. Agueh and G. Carlier, *Barycenters in the wasserstein space*, SIAM Journal on Mathematical Analysis, 43 (2011), pp. 904–924.

[3] A. Ben-Israel and T. N. Greville, *Generalized inverses: theory and applications*, vol. 15, Springer Science & Business Media, 2003.

[4] G. Birkhoff, *Extensions of jentzsch's theorem*, Transactions of the American Mathematical Society, 85 (1957), pp. 219–227.

[5] J. Bolte, E. Pauwels, and S. Vaiter, *Automatic differentiation of nonsmooth iterative algorithms*, arXiv preprint arXiv:2206.00457, (2022).

[6] N. Bonneel, G. Peyré, and M. Cuturi, *Wasserstein barycentric coordinates: Histogram regression using optimal transport*, ACM Transactions on Graphics, 35 (2016), https://doi.org/10.1145/2897824.2925918.

[7] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, *JAX: composable transformations of Python+NumPy programs*, 2018, http://github.com/google/jax.

[8] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, *Unsupervised learning of visual features by contrasting cluster assignments*, in NeurIPS, vol. 33, 2020, pp. 9912–9924.

[9] B. Christianson, *Reverse accumulation and attractive fixed points*, Optimization Methods and Software, 3 (1994), pp. 311–326.

[10] M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, in NeurIPS, vol. 26, 2013.

[11] M. Cuturi, L. Meng-Papaxanthos, Y. Tian, C. Bunne, G. Davis, and O. Teboul, *Optimal transport tools (ott): A jax toolbox for all things wasserstein*, arXiv preprint arXiv:2201.12324, (2022).

[12] M. Cuturi, O. Teboul, J. Niles-Weed, and J.-P. Vert, *Supervised quantile normalization for low-rank matrix approximation*, in ICML, 2020.

[13] M. Cuturi, O. Teboul, and J.-P. Vert, *Differentiable ranking and sorting using optimal transport*, in NeurIPS, vol. 32, 2019.

[14] M. Eisenberger, A. Toker, L. Leal-Taixé, F. Bernard, and D. Cremers, *A unified framework for implicit sinkhorn differentiation*, in CVPR, 2022.

[15] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, *Pot: Python optimal transport*, Journal of Machine Learning Research, 22 (2021), pp. 1–8.

[16] J. Franklin and J. Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra and its Applications, 114-115 (1989), pp. 717–735, https://doi.org/https://doi.org/10.1016/0024-3795(89)90490-4.
[17] A. Genevay, G. Peyré, and M. Cuturi, *Learning generative models with sinkhorn divergences*, in AISTAT, vol. 84, 2018, pp. 1608–1617.
[18] J. C. Gilbert, *Automatic differentiation and iterative processes*, Optimization Methods and Software, 1 (1992), pp. 13–21, https://doi.org/10.1080/10556789208805503.
[19] A. Griewank, C. Bischof, G. Corliss, A. Carle, and K. Williamson, *Derivative convergence for iterative equation solvers*, Optimization Methods and Software, 2 (1993), pp. 321–355, https://doi.org/10.1080/10556789308805549.
[20] T. Hashimoto, D. Gifford, and T. Jaakkola, *Learning population-level diffusions with generative rnns*, in ICML, vol. 48, 2016, pp. 2417–2426.
[21] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, second ed., 2013.
[22] S. Linnainmaa, *Taylor expansion of the accumulated rounding error*, BIT Numerical Mathematics, 16 (1976), pp. 146–160.
[23] J. Lorraine, P. Vicol, and D. Duvenaud, *Optimizing millions of hyperparameters by implicit differentiation*, in AISTAT, vol. 108, 2020, pp. 1540–1552.
[24] G. Luise, A. Rudi, M. Pontil, and C. Ciliberto, *Differential properties of sinkhorn approximation for learning with wasserstein distance*, in NeurIPS, vol. 31, 2018.
[25] S. Mehmood and P. Ochs, *Automatic differentiation of some first-order methods in parametric optimization*, in AISTAT, 2020, pp. 1584–1594.
[26] G. Peyré and M. Cuturi, *Computational optimal transport*, Foundations and Trends in Machine Learning, 51 (2019), pp. 1–44, https://doi.org/10.1561/2200000073.
[27] B. Polyak, *Introduction to optimization*, in Optimization Software, Publications Division, Citeseer, 1987.
[28] L. Ruschendorf, *Convergence of the Iterative Proportional Fitting Procedure*, The Annals of Statistics, 23 (1995), pp. 1160 – 1174, https://doi.org/10.1214/aos/1176324703.
[29] J. E. Scroggs and P. L. Odell, *An alternate definition of a pseudoinverse of a matrix*, SIAM Journal on Applied Mathematics, 14 (1966), pp. 796–810.
[30] R. Sinkhorn, *A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices*, The Annals of Mathematical Statistics, 35 (1964), pp. 876 – 879.
[31] R. Sinkhorn, *Diagonal equivalence to matrices with prescribed row and column sums*, The American Mathematical Monthly, 74 (1967), pp. 402–405.
[32] J. Thornton and M. Cuturi, *Rethinking initialization of the sinkhorn algorithm*, 2022, https://doi.org/10.48550/ARXIV.2206.07630.
[33] R. E. Wengert, *A simple automatic derivative evaluation program*, Commun. ACM, 7 (1964), p. 463–464, https://doi.org/10.1145/355586.364791.