Batch Greenkhorn Algorithm for Entropic-Regularized Multimarginal Optimal Transport: Linear Rate of Convergence and Iteration Complexity

Vladimir R. Kostić [1 2] **Saverio Salzo** [1] **Massimiliano Pontil** [1 3]

Abstract

In this work we propose a batch multimarginal version of the Greenkhorn algorithm for the entropic-regularized optimal transport problem. This framework is general enough to cover, as particular cases, existing Sinkhorn and Greenkhorn algorithms for the bi-marginal setting, and greedy MultiSinkhorn for the general multimarginal case. We provide a comprehensive convergence analysis based on the properties of the iterative Bregman projections method with greedy control. Linear rate of convergence as well as explicit bounds on the iteration complexity are obtained. When specialized to the above mentioned algorithms, our results give new convergence rates or provide key improvements over the state-of-the-art rates. We present numerical experiments showing that the flexibility of the batch can be exploited to improve performance of Sinkhorn algorithm both in bi-marginal and multimarginal settings.

1. Introduction

Over the recent years the field of optimal transport (OT) (Villani, 2008) has received significant attention in machine learning and data science, as it provides natural and powerful tools to compare probability distributions. In this paper we study a general class of OT problems known as multimarginal optimal transport (MOT), whereby several probability distributions are coupled together in order to compute a measure of their association, see e.g. (Pass, 2015; Benamou et al., 2016). While bi-marginal OT is well established in the scientific community, and machine learning in particular, only recently MOT is receiving increasing interest due to its applications, ranging from density functional theory in quantum chemistry, to fluid dynamics, to economics, to image processing, among others, see (Peyré & Cuturi, 2019) and references therein. Particularly in machine learning, MOT is gaining relevance for generative adversarial networks (GANs) (Cao et al., 2019), domain adaptation (He et al., 2019), Wasserstein barycenters (Agueh & Carlier, 2011), clustering (Bento & Mi, 2021), Bayesian inference of joint distributions (Frogner & Poggio, 2019) and multi-dynamics reinforcement learning (Cohen et al., 2021).

In this paper, we focus on the discrete formulation of MOT, which consists in minimizing a linear cost function over all the joint distributions with $m \geq 2$ prescribed finitely supported marginals. It is well known that addressing directly the MOT problem is computationally expensive. Furthermore, unlike the bi-marginal case, MOT is NP-Hard for certain costs, even approximately (Altschuler & Boix-Adserà, 2021). To overcome this issue, regularization techniques have been widely considered. The key insight is to add a strongly convex regularizer to the MOT objective. In this respect, a popular choice is entropic-regularization, which, in the bi-marginal case, leads to the well-known Sinkhorn algorithm (Cuturi, 2013). While the convergence properties of Sinkhorn have been studied in detail (Peyré & Cuturi, 2019), computational solutions for regularized multimarginal optimal transport (RMOT) are less developed. The principal objective of this work is to propose a new and flexible algorithmic framework with strong convergence guarantees which can be effective for both ROT and RMOT problems.

Related work. Two popular frameworks for analyzing algorithmic solutions for entropic-regularized OT problems are iterative Bregman projections (IBP) and alternating dual minimization (ADM). The multimarginal version of Sinkhorn algorithm was first proposed by (Benamou et al., 2015) in finite dimension and its convergence was established by viewing it as a special case of *cyclic* IBP, whose global convergence is well-known (Bregman, 1967). However, this approach does not ensure any rate of convergence. More recently, using the alternating minimization framework, this result was extended to infinite dimension (DiMarino & Gerolin, 2020) and even a global linear rate of

[1]Istituto Italiano di Tecnologia, Via Melen 83, 16152 Genova, Italy [2]Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia [3]Department of Computer Science, University College London, Gower Street WC1E 6BT,London, United Kingdom. Correspondence to: Vladimir R. Kostić <vladimir.kostic@iit.it>, Saverio Salzo <saverio.salzo@iit.it>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

convergence was obtained (Carlier, 2021).

The ADM approach was exploited also to obtain explicit computational complexity bounds. Specifically, in the bi-marginal setting state-of-the-art complexity bounds were proved by Dvurechensky et al. (2018) for the Sinkhorn algorithm and by Lin et al. (2021) for the Greenkhorn algorithm of Altschuler et al. (2017). In the multi-marginal setting, Lin et al. (2020) proposed two solutions for approaching RMOT: the (greedy) multimarginal Sinkhorn algorithm, which iteratively projects on a *greedily* selected marginal, and the accelerated multimarginal Sinkhorn algorithm, which improves the convergence by incorporating Nesterov's estimate sequences. In addition, in (Tupitsa et al., 2020) a Nesterov's momentum acceleration was used to build a competitive algorithm against the greedy multimarginal Sinkhorn. Both papers (Lin et al., 2020; Tupitsa et al., 2020) derived sublinear computational complexity bounds for the considered RMOT algorithms, but no linear convergence rates were studied. Moreover, such accelerated versions for large scale problems, while having desirable theoretical properties, do suffer from longer computing times compared to greedy MultiSinkhorn (Lin et al., 2020).

We conclude this discussion by remarking that in the context of bi-marginal optimal transport, various algorithmic strategies such as simulated annealing ("ε-scaling") and multiscale algorithms have been used with much success to accelerate basic Sinkhorn algorithm, see (Flamary et al., 2021) and references therein. Furthermore , in (Feydy et al., 2019) implementations exploiting symbolic (kernel) matrices are developed allowing efficient large-scale OT computations.

Contribution. We propose a batch multimarginal version of the Greenkhorn algorithm, that greedily selects at each iteration a marginal and a batch of its components. It covers, as special cases, Sinkhorn and Greenkhorn for the bi-marginal setting, and (greedy) MultiSinkhorn of (Lin et al., 2020) for RMOT. For this new general algorithm, we established (1) *global linear rate of convergence* in Kulback-Leibler (KL) distance and (2) *iteration complexity* in the ℓ^1 distance from the given marginals. Moreover, in the two important special cases of *Batch Greenkhorn for bi-marginal ROT* and (greedy) *MultiSinkhorn*, we provide a global linear rate and an iteration complexity bound that depend explicitly on the problem data. These results offer new insights on the asymptotic behaviour of those algorithms or improve the state-of-the-art. Particularly, we remove a logarithmic factor in the known complexity results for Sinkhorn, Greenkhorn and MultiSinkhorn. Our theoretical contributions are summarized in Table 1. There, we point out the explicit rate for the greedy MultiSinkhorn algorithm which is strictly better than the one, recently derived by Carlier (2021), for cyclic multimarginal Sinkhorn. Notably, our rate scales better with the number of marginals m, which we also confirm in our numerical experiments below.

At last, we stress that our theoretical analysis is not based on the dual alternating minimization approach used in the state-of-the-art analyses, but it rather relies on the geometry of the KL distance and the framework of iterative Bregman projections.

Paper organization. In Sec. 2 we recall the formulation of RMOT as a Bregman projection problem. In Sec. 3 we present the proposed batch Greenkhorn algorithm for RMOT. Our main results are presented in Sec. 4. Numerical experiments for the proposed method on both bi-marginal and multimarginal optimal transport are presented in Sec. 5.

Notation. We denote by $n_1, \dots, n_m \in \mathbb{N}$ the sizes of the given m marginals. For every $n \in \mathbb{N}$, we set $[n] := \{1, 2, \dots, n\}$ and denote the unit simplex by $\Delta_n = \{x \in \mathbb{R}^n_+ \,|\, \|x\|_1 = 1\}$. For the sake of brevity we set $\mathcal{J} = [n_1] \times \cdots \times [n_m]$ and we denote by $j = (j_1, \dots, j_m)$ a general multi-index in $\mathcal{J}$. We set $\mathbb{X}_+ = \{\pi \in \mathbb{X} \,|\, \pi_j \geq 0, \text{ for every } j \in \mathcal{J}\}$ and $\mathbb{X}_{++} = \{\pi \in$

Table 1: Convergence results on Sinkhorn-type algorithms for entropic-regularized OT (ROT) and entropic-regularized multimarginal OT (RMOT): global linear convergence (GL) measured in KL-divergence and iteration complexity bounds (IC). To ease the presentation we assume $\|C\|_\infty = 1$. In the following m, n, η, ε and τ are the number marginals, the number of atoms of marginal distributions, the regularization parameter, the tolerance and the batch size, respectively.

Algorithm (problem)	Convergence type	Current best	Our result	
Sinkhorn (ROT)	(GL)	$1 - \frac{1}{2}e^{-24/\eta}$ (Carlier, 2021)	$\big(1 - e^{-17/\eta}\big)^2$	Theorem 4.5
	(IC)	$\mathcal{O}\big(\frac{1/\eta + \log n}{\varepsilon}\big)$ (Dvurechensky et al., 2018)	$\mathcal{O}\big(\frac{1}{\eta\varepsilon}\big)$	
Greenkhorn (ROT)	(IC)	$\mathcal{O}\big(\frac{1/\eta + \log n}{\varepsilon}\big)$ (Lin et al., 2021)	$\mathcal{O}\big(\frac{1}{\eta\varepsilon}\big)$	Theorem 4.4
BatchGreenkhorn (ROT)	(GL)	$\times$	$\big(1 - \frac{e^{-20/\eta}}{2n/\tau - 1}\big)^{2n/\tau}$	Theorem 4.4
	(IC)	$\times$	$\mathcal{O}\big(\frac{1}{\eta\varepsilon} n/\tau\big)$	
MultiSinkhorn (RMOT)	(GL)	$\times$	$\big(1 - \frac{e^{-(12m-7)/\eta}}{m-1}\big)^m$	Theorem 4.5
	(IC)	$\mathcal{O}\big(\frac{m(1/\eta + \log n)}{\varepsilon}\big)$ (Lin et al., 2020)	$\mathcal{O}\big(\frac{m}{\eta\varepsilon}\big)$	

$\mathbb{X} \,|\, \pi_j > 0, \text{ for every } j \in \mathcal{J}\}$. We need also to consider multi-indexes without the k-th index, so we define $\mathcal{J}_{-k} = [n_1] \times \cdots \times [n_{k-1}] \times [n_{k+1}] \times \cdots \times [n_m]$ and j_{-k} a general multi-index in $\mathcal{J}_{-k}$. We will identify $\mathcal{J}$ with $\mathcal{J}_{-k} \times [n_k]$ via the mapping $j \leftrightarrow (j_{-k}, j_k)$ with $j_{-k} = (j_1, \dots, j_{k-1}, j_{k+1}, \dots, j_m)$. The space $\mathbb{X}$ is an Euclidean space endowed with the standard scalar product and norm

$$\langle \pi, \pi' \rangle = \textstyle\sum_{j\in\mathcal{J}} \pi_j \pi'_j, \quad \|\pi\|^2 = \textstyle\sum_{j\in\mathcal{J}} \pi_j^2, \ \ \forall\, \pi, \pi' \in \mathbb{X}.$$

For every π, π' we denote by $\pi \odot \pi' \in \mathbb{X}$, the Hadamard product of π and π', that is, $(\pi \odot \pi')_j = \pi_j \pi'_j$. Moreover, if $\mathsf{v}_1 = (v_{1,j_1})_{j_1\in[n_1]} \in \mathbb{R}^{n_1}, \dots, \mathsf{v}_m = (v_{m,j_m})_{j_m\in[n_m]} \in \mathbb{R}^{n_m}$, we set $\oplus_{k=1}^m \mathsf{v}_k \in \mathbb{X}$ and $\otimes_{k=1}^m \mathsf{v}_k \in \mathbb{X}$ such that

$$\Big(\bigoplus_{k=1}^m \mathsf{v}_k\Big)_j = \sum_{k=1}^m v_{k,j_k} \quad \text{and} \quad \Big(\bigotimes_{k=1}^m \mathsf{v}_k\Big)_j = \prod_{k=1}^m v_{k,j_k},$$

respectively. For a function $\phi\colon \mathbb{X} \to \,]-\infty, +\infty]$, its Fenchel conjugate is the function defined by $\phi^*(\gamma) = \sup_{\pi\in\mathbb{X}} \langle \pi, \gamma\rangle - \phi(\pi)$. Finally, as usual, the Dirac measure at x is denoted by δ_x.

2. Entropic RMOT as the Bregman projection problem

In this section we formally introduce the discrete multimarginal optimal transport problem and its entropic-regularized version, emphasizing its connection with the Bregman projection problem.

Let $\mathsf{a}_k \in \Delta_{n_k}$, $k \in [m]$ be prescribed histograms. MOT consists in solving the linear program

$$\min_{\pi\in\Pi(\mathsf{a}_1,\dots,\mathsf{a}_m)} \langle \mathsf{C}, \pi\rangle, \tag{1}$$

where $\mathsf{C} \in \mathbb{X}$ is a given cost tensor and $\Pi(\mathsf{a}_1, \dots, \mathsf{a}_m)$, called *transport polytope*, is the convex set of nonnegative tensors in $\mathbb{X}$ whose marginals are $\mathsf{a}_1, \dots, \mathsf{a}_m$. Specifically,

$$\Pi(\mathsf{a}_1, \dots, \mathsf{a}_m) = \big\{\pi \in \mathbb{X}_+ \,\big|\, \mathsf{R}_k(\pi) = \mathsf{a}_k, \forall k \in [m]\big\}, \tag{2}$$

where for all $k \in [m]$, $\mathsf{R}_k\colon \mathbb{X} \to \mathbb{R}^{n_k}$ is the *k-th push-forward projection operator*, defined as

$$\mathsf{R}_k(\pi)_{j_k} = \sum_{j_{-k}\in\mathcal{J}_{-k}} \pi_{(j_{-k},j_k)}, \quad \forall j_k \in [n_k], \tag{3}$$

representing the operation of taking the k-th marginal of multi-dimensional histograms.

As noted in the introduction, problem (1) may be hard to solve and an effective alternative is solving a regularized version using the negative entropy (Cuturi, 2013): for a cost tensor $\mathsf{C} \in \mathbb{X}_+$, the entropic-regularized multimarginal optimal transport (RMOT) problem consists in computing

$$\pi^\star = \operatorname*{arg\,min}_{\pi\in\Pi(\mathsf{a}_1,\dots,\mathsf{a}_m)} \langle \mathsf{C}, \pi\rangle + \eta \mathsf{H}(\pi), \tag{4}$$

where $\mathsf{H}(\pi) := \sum_j \pi_j(\log \pi_j - 1)$, $\pi \in \operatorname{dom}\mathsf{H} = \mathbb{X}_+$ and $\eta > 0$ is a regularization parameter.

Now, recall that the Kulback-Leibler (KL) divergence $\mathsf{KL}\colon \mathbb{X} \times \mathbb{X} \to [0, +\infty]$ is defined as

$$\mathsf{KL}(\pi, \pi') = \begin{cases} \displaystyle\sum_{j\in\mathcal{J}} \pi_j\Big(\log\frac{\pi_j}{\pi'_j} - 1\Big) + \pi'_j, \ \pi\in\mathbb{X}_+, \pi'\in\mathbb{X}_{++}, \\ +\infty \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

and can be interpreted as the *Bregman distance* associated to the negative entropy H, that is, $\mathsf{KL}(\pi, \pi') = \mathsf{H}(\pi) - \mathsf{H}(\pi') - \langle \pi - \pi', \nabla\mathsf{H}(\pi')\rangle$. Then, for an arbitrary closed convex set $\mathcal{C} \subset \mathbb{X}$ such that $\mathcal{C} \cap \mathbb{X}_{++} \neq \varnothing$ and a point $\pi \in \mathbb{X}_{++}$, the *Bregman projection* of π onto $\mathcal{C}$ with respect to H is

$$\mathcal{P}_{\mathcal{C}}(\pi) := \operatorname*{arg\,min}_{\gamma\in\mathcal{C}} \mathsf{KL}(\gamma, \pi). \tag{5}$$

This is also called the *KL projection* of π onto $\mathcal{C}$, while its *KL distance* to $\mathcal{C}$ is defined as $\mathsf{KL}_{\mathcal{C}}(\pi) := \mathsf{KL}(\mathcal{P}_{\mathcal{C}}(\pi), \pi)$.

Taking this into account, it is easy to recognize that (4) can be rewritten as

$$\pi^\star = \operatorname*{arg\,min}_{\mathsf{R}_k(\pi)=\mathsf{a}_k, k\in[m]} \mathsf{KL}(\pi, \xi), \tag{6}$$

where $\xi = \nabla\mathsf{H}^*(-\mathsf{C}/\eta) = \exp(-\mathsf{C}/\eta) \in \mathbb{X}_{++}$ is the *Gibbs kernel* tensor. Note that, since $\operatorname{dom}\mathsf{H} = \mathbb{X}_+$, in passing to (6), the positiveness constraint embodied in problem (4) has been absorbed in the entropy function H. Overall, we can conclude that problem (4) is equivalent to the computation of the KL projection of the Gibbs kernel ξ onto the affine set $\{\pi \in \mathbb{X} \,|\, (\forall\, k \in [m])\ \mathsf{R}_k(\pi) = \mathsf{a}_k\}$.

3. Greedy KL projections for entropic RMOT

In this section, we study the entropic-regularized MOT problem (4), in the equivalent form of the KL projection problem (6). We will represent the underlying constraint set as an intersection of possibly simple affine sets, so that the problem becomes accessible via iterative Bregman projections. This will lead to a new batch version of Greenkhorn algorithm.

To that purpose, recalling definition (3), we first introduce the linear operator

$$\mathsf{R}\colon \mathbb{X} \to \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}, \ \ \mathsf{R}(\pi) = (\mathsf{R}_1(\pi), \dots, \mathsf{R}_m(\pi)), \tag{7}$$

and the affine set

$$\Pi := \{\pi \in \mathbb{X} \mid \mathsf{R}(\pi) = (\mathsf{a}_1, \dots, \mathsf{a}_m)\}, \tag{8}$$

which defines the constraints in (6). Then, we observe that it is possible to prove (see equation (41) in Appendix A) that

$$\mathcal{P}_\Pi(\xi) = \mathcal{P}_\Pi(\xi \odot \otimes_{k=1}^m \mathsf{a}_k). \tag{9}$$

Therefore, here we will equivalently target the computation of the KL projection of the normalized Gibbs kernel $\xi \odot \otimes_{k=1}^m \mathsf{a}_k$ onto Π.

Now we proceed by decomposing the affine set Π into an intersection of simpler affine sets. More precisely, we will rewrite the set (8) as an intersection of affine sets, obtained via specific *sketching*, on which KL projections have closed forms. Thus, for each $k \in [m]$ (which refers to the k-th marginal) and each batch $L \subset [n_k]$ we consider the *canonical injection*

$$\mathsf{S}_{(k,L)} \colon \mathbb{R}^L \to \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$$

of $\mathbb{R}^L$ into $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$, meaning that for each $\mathsf{u} = (u_{j_k})_{j_k \in L} \in \mathbb{R}^L$, $\mathsf{S}_{(k,L)}\mathsf{u}$ is the vector of $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$ obtained from the completion of u with zero entries. Then, since the adjoint operator $\mathsf{S}^*_{(k,L)} \colon \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m} \to \mathbb{R}^L$ is the standard projection, we can define

$$\mathsf{R}_{(k,L)} := \mathsf{S}^*_{(k,L)}\mathsf{R} \colon \mathbb{X} \to \mathbb{R}^L \tag{10}$$

and the set

$$\begin{aligned} \Pi_{(k,L)} &:= \{\pi \in \mathbb{X} \,|\, \mathsf{S}^*_{(k,L)}\mathsf{R}(\pi) = \mathsf{S}^*_{(k,L)}(\mathsf{a}_1, \ldots, \mathsf{a}_m)\} \\ &= \{\pi \in \mathbb{X} \,|\, \mathsf{R}_{(k,L)}(\pi) = \mathsf{a}_{k|L}\} \\ &= \{\pi \in \mathbb{X} \,|\, (\mathsf{R}_k(\pi))_{|L} = \mathsf{a}_{k|L}\}. \end{aligned} \tag{11}$$

Note that in the definition of $\Pi_{(k,L)}$ it is required that the k-th marginal of π is equal to a_k only on the components in L. In the end, given $\tau = (\tau_k)_{1 \le k \le m}$ a vector of batch sizes, we set

$$\mathcal{I}(\tau) = \{(k, L) \,|\, k \in [m], L \subset [n_k] \,|\, |L| \le \tau_k\}$$

and obtain

$$\Pi = \bigcap_{(k,L) \in \mathcal{I}(\tau)} \Pi_{(k,L)}. \tag{12}$$

This is the announced decomposition of the set Π into the intersection of possibly simpler affine sets.

As a results of the representation (12), one can envisage to approach the computation of the KL projection of $\xi \odot \otimes_{k=1}^m \mathsf{a}_k$ onto Π by applying the procedure of *iterative Bregman projections* (IBP) (Bregman, 1967). This leads to the following algorithm. Let $\pi^0 = \xi \odot \otimes_{k=1}^m \mathsf{a}_k = e^{-\mathsf{C}/\eta} \odot \otimes_{k=1}^m \mathsf{a}_k \in \mathrm{int}(\mathrm{dom}\,\mathsf{H}) = \mathbb{X}_{++}$ and define the sequence π^t recursively as follows

$$\begin{aligned} &\text{for } t = 0, 1, \ldots \\ &\quad \text{choose } (k_t, L_t) \in \mathcal{I}(\tau), \\ &\quad \pi^{t+1} = \mathcal{P}_{\Pi_{(k_t,L_t)}}(\pi^t). \end{aligned} \tag{13}$$

Since the generalized Pythagoras theorem for Bregman projections (see equation (40) in Appendix A) yields that $\mathsf{KL}_\Pi(\pi^t) = \mathsf{KL}_{\Pi_{(k,L)}}(\pi^t) + \mathsf{KL}_\Pi(\mathcal{P}_{\Pi_{(k,L)}}(\pi^t))$, in (13) one may choose the sets in a *greedy* manner as

$$(k_t, L_t) = \operatorname*{arg\,max}_{(k,L) \in \mathcal{I}(\tau)} \mathsf{KL}_{\Pi_{(k,L)}}(\pi^t), \tag{14}$$

so that

$$(k_t, L_t) = \operatorname*{arg\,min}_{(k,L) \in \mathcal{I}(\tau)} \mathsf{KL}_\Pi(\mathcal{P}_{\Pi_{(k,L)}}(\pi^t)) \text{ and} \tag{15}$$

$$\mathsf{KL}_\Pi(\pi^{t+1}) = \min_{(k,L) \in \mathcal{I}(\tau)} \mathsf{KL}_\Pi(\mathcal{P}_{\Pi_{(k,L)}}(\pi^t)). \tag{16}$$

This means that the next iterate is chosen, among the possible projections, as the one which is the closest to the target set Π. Notable examples of existing algorithms that fit in this framework are (greedy) multimarginal Sinkhorn of (Lin et al., 2020) and bi-marginal Greenkhorn of (Altschuler et al., 2017).

We emphasize that the above greedy strategy typically leads to the best performance, provided that it can be implemented efficiently. In the following proposition and subsequent remark we show that the projection onto the sets $\Pi_{(k,L)}$ can be computed in a closed form and that the greedy choice of the sets $\Pi_{(k,L)}$'s can indeed be implemented efficiently. The proof is postponed to Appendix B.

Proposition 3.1. *For every $\pi \in \mathbb{X}_+$, $k \in [m]$ and $L \subset [n_k]$,*

$$\mathcal{P}_{\Pi_{(k,L)}}(\pi) = \pi \odot \exp\left(\mathsf{R}^*_{(k,L)}(\bar{\mathsf{u}})\right), \tag{17}$$

where

$$\bar{\mathsf{u}} = \log \frac{\mathsf{a}_{k|L}}{\mathsf{R}_k(\pi)_{|L}}, \tag{18}$$

and, consequently, for every $j \in \mathcal{J}$,

$$(\mathcal{P}_{\Pi_{(k,L)}}(\pi))_j = \pi_j \times \begin{cases} \dfrac{a_{k,j_k}}{\mathsf{R}_k(\pi)_{j_k}} & \text{if } j_k \in L, \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

Moreover,

$$\mathsf{KL}_{\Pi_{(k,L)}}(\pi) = \mathsf{KL}(\mathsf{a}_{k|L}, \mathsf{R}_k(\pi)_{|L}). \tag{20}$$

Remark 3.2. It follows from (20) that the greedy choice described above can be implemented by computing m vectors of sizes n_k and then choosing k_t among m as the index of the vector that has the maximal sum of the largest τ_k components. More formally one lets $\mathsf{d}_k = (\mathsf{KL}(a_{k,1}, (\mathsf{R}_k(\pi^t))_1), \ldots, \mathsf{KL}(a_{k,n_k}, (\mathsf{R}_k(\pi^t))_{n_k})) \in \mathbb{R}^{n_k}$ and considers the vector $\mathsf{d}_{k\downarrow} \in \mathbb{R}^{n_k}$ which has the components of d_k arranged in a decreasing order. Then $k_t = \arg\max_{k \in [m]} \left(\sum_{j_k=1}^{\tau_k} (\mathsf{d}_{k\downarrow})_{j_k}\right)$ and L_t corresponds to the indexes of the largest τ_{k_t} components of d_{k_t}.

We conclude this section by observing that to ensure better numerical stability, especially for small regularization parameters η, it is more convenient to work with the dual variables $\nabla \mathsf{H}(\pi) = \log(\pi)$. More precisely, according to (17), we have that each iteration of the IBP algorithm (13) can be parameterized as

$$\pi^t = \exp\Big(-\mathsf{C}/\eta + \bigoplus_{k=1}^m \mathsf{v}_k^t \Big) \odot \bigotimes_{k=1}^m \mathsf{a}_k, \ t \in \mathbb{N}, \quad (21)$$

and one can implement the algorithm by updating only the dual variables $\mathsf{v}_k^t = (v_{k,j}^t)_{1\le j\le n_k} \in \mathbb{R}^{n_k}$, $k \in [m]$, which are also known as *potentials* (see Proposition B.1). Thus, in the end, the IBP algorithm (13)-(14) can be written in the form of Algorithm 1.

Algorithm 1 `BatchGreenkhorn`$(\mathsf{a}_1, \dots, \mathsf{a}_m, \mathsf{C}, \eta, \tau)$

Initialization: $\mathsf{v}_k^0 = 0$, $\mathsf{r}_k^0 = \mathsf{R}_k(\exp(-\mathsf{C}/\eta) \odot \otimes_{k=1}^m \mathsf{a}_k)$
for $t = 0, 1, \dots$ **do**
  Compute $(k_t, L_t) = \underset{(k,L)\in\mathcal{I}(\tau)}{\arg\max}\ \mathsf{KL}(\mathsf{a}_{k|L}, \mathsf{r}_{k|L}^t)$
  Set $\mathsf{v}_k^{t+1} = \mathsf{v}_k^t$, $k \in [m]$ and update
    $\mathsf{v}_{k_t\,|L_t}^{t+1} \leftarrow \mathsf{v}_{k_t\,|L_t}^{t+1} + \log(\mathsf{a}_{k_t|L_t}) - \log(\mathsf{r}_{k_t|L_t}^t)$
  for $k \in [m]$ **do**
    $\mathsf{r}_k^{t+1} = \mathsf{R}_k(\exp(-\mathsf{C}/\eta + \oplus_{k=1}^m \mathsf{v}_k^{t+1}) \odot \oplus_{k=1}^m \mathsf{a}_k)$
  end for
end for
Output: $\pi^t = \exp(-\mathsf{C}/\eta + \oplus_{k=1}^m \mathsf{v}_k^t) \odot \otimes_{k=1}^m \mathsf{a}_k$

A natural issue of `BatchGreenkhorn` is that of choosing the batch size. As extreme cases we have full batch ($\tau_k = n_k$), which yields the (greedy) MultiSinkhorn algorithm proposed in (Lin et al., 2020), and $\tau_k = 1$, which, in the bi-marginal case, is known as the Greenkhorn algorithm (Altschuler et al., 2017). In this respect we observe that, since the greedy selection step in Algorithm 1 can be efficiently implemented, as discussed in Remark 3.2, the largest computational cost lies in the computation of the marginals r_k^{t+1}. For simplicity, let us assume that $n_k = n$ and $\tau_k = \tau$, for every $k \in [m]$. Then, computing it naively yields $\mathcal{O}(mn^m)$ operations, but indeed it can be done more efficiently in $\mathcal{O}(\tau n^{m-1})$ as we show in Appendix B. This way n/τ iterations with batch size τ have the same computational cost of one iteration with a full batch n and consequently mn/τ iterations of `BatchGreenkhorn` with batch sizes τ corresponds to one cycle of cyclic multimarginal Sinkhorn.

4. Convergence theory for Batch Greenkhorn algorithm

Results on the convergence of general IBP are typically without any rates (Bregman, 1967; Censor & Lent, 1981; Censor & Reich, 1996), with the notable exception of (Kostic & Salzo, 2021) where local linear rate for the greedy IBP was studied. In the following we prove the global linear convergence of Algorithm 1 and derive the explicit dependence of the rate on the given data in two important cases. Moreover, we provide an analysis of the iteration complexity.

Based on the properties of the operators R and $\mathsf{R}_{(k,L)}$, defined in (7) and (10) respectively, we can derive the main results using the properties of KL as Bregman divergence. The proofs are given in Appendix C.

Theorem 4.1 (Global linear convergence). *Algorithm 1 converges linearly. More precisely, if $(\mathsf{v}_k^t)_{k\in[m]}$ are generated by Algorithm 1, then the primal iterates given by (21) converge linearly in KL divergence to $\pi^\star$ given by (6), i.e. for all $t \in \mathbb{N}$*

$$\mathsf{KL}(\pi^\star, \pi^t) \le \left(1 - \frac{e^{-(2\|\mathsf{C}\|_\infty/\eta + 3M_1)}}{b_\tau - 1}\right)^t \mathsf{KL}(\pi^\star, \pi^0), \quad (22)$$

where $b_\tau = \sum_{k\in[m]} \lceil n_k/\tau_k \rceil$, and $0 < M_1 < +\infty$ is a constant independent of the batch sizes that satisfies $\max\left\{ \|\bigoplus_{k=1}^m \mathsf{v}_k^\star\|_\infty, \|\bigoplus_{k=1}^m \mathsf{v}_k^t\|_\infty \right\} \le M_1$ for $t \in \mathbb{N}$.

Theorem 4.2 (Iteration complexity). *Let $\varepsilon > 0$ and suppose that $\eta > \varepsilon$. For Algorithm 1, the number of iterations required to reach the stopping criterion $d_t^\infty := \max_{k\in[m]} \|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1 \le \varepsilon$ satisfies*

$$t \le 2 + \max_{k\in[m]} \left\lceil \frac{n_k}{\tau_k} \right\rceil \frac{5M_2}{\varepsilon}(2 + M_2\eta), \quad (23)$$

where $0 < M_2 < +\infty$ is a constant independent of the batch sizes such that $\sum_{k\in[m]} \|\mathsf{v}_k^\star - \mathsf{v}_k^t\| \le M_2$, for all $t \in \mathbb{N}$.

Remark 4.3. *The constants M_1 and M_2 considered in the above theorems always exist (see the proofs in Appendix C), but we are not able to derive an expression explicitly depending on the problem data for them that is valid for any m and $(\tau_k)_{k\in[m]}$. However, in the following subsections, we show the important cases $m = 2$ and $(\tau_k)_{k\in[m]}$ arbitrary and $m > 2$ and $(\tau_k)_{k\in[m]} = (n_k)_{k\in[m]}$, for which we do provide explicit dependence on the problem data.*

Concerning Theorem 4.2, we note that in literature the stopping criteria normally considered concerns the quantity $d_t := \sum_{k\in[m]} \|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1$, rather than d_t^∞. Moreover, the assumptions usually demand $n_k = n$ and $\tau_k = \tau$, for every $k \in [m]$. In this setting $b_\tau = mn/\tau$ and, since $d_t \le m d_t^\infty$, according to the bound (23), to achieve $d_t \le \varepsilon$ the following number of iterations is required

$$t \le 2 + \frac{nm}{\tau}\frac{5M_2}{\varepsilon}(2 + M_2\eta).$$

Hence, in terms of normalized cycles $T = t/b_\tau$ we have

$$T \le 1 + \frac{5M_2}{m\varepsilon}(2 + M_2), \quad (24)$$

which we stress is independent on the batch-size and the dimension n. Additionally, the rate (22) w.r.t. the normalized cycles become

$$\mathsf{KL}(\pi^\star, \pi^{b_\tau T}) \leq \left[\left(1 - \frac{e^{-(2\|\mathsf{C}\|_\infty/\eta + 3M_1)}}{b_\tau - 1}\right)^{b_\tau}\right]^T \mathsf{KL}(\pi^\star, \pi^0).$$

We conclude this general discussion with a remark on the computational complexity of the proposed algorithm in terms of arithmetic operations. If we denote by $\mathcal{O}_\xi$ the number of arithmetic operations needed to compute a full marginal using the Gibbs kernel ξ, when we factor this number out of the total computational complexity, what remains is the number of iterations normalized with respect to the full batch, meaning, $\bar{t} = \tau t/n$. Therefore, according to (23) the total computational complexity is given by

$$\left[2 + \frac{5M_2}{\varepsilon}(2 + M_2)\right]\mathcal{O}_\xi. \tag{25}$$

We note that in the worst case $\mathcal{O}_\xi$ is of the order of n^m, but this can be significantly reduced for costs with specific structure such as the graphical one (Haasler et al., 2020) or the separable one (Benamou et al., 2015; Peyré & Cuturi, 2019).

The following two subsections give explicit convergence results for two significant special cases of BatchGreenkhorn. The proofs consist essentially in computing the constants M_1 and M_2 in Theorems 4.1 and 4.2 respectively. Details are in Appendix C.

4.1. Bi-marginal BatchGreenkhorn

Theorem 4.4. *Suppose that $m = 2$. Then the algorithm* $\texttt{BatchGreenkhorn}(\mathsf{a}_1, \mathsf{a}_2, \mathsf{C}, \eta, \tau)$ *converges linearly with the global rate*

$$\mathsf{KL}(\pi^\star, \pi^t) \leq \left(1 - \frac{e^{-20\|\mathsf{C}\|_\infty/\eta}}{b_\tau - 1}\right)^t \mathsf{KL}(\pi^\star, \pi^0). \tag{26}$$

Moreover, when $\eta > \varepsilon$, the number of iterations required to reach the stopping criterion $d_t^\infty \leq \varepsilon$ satisfies

$$t \leq 2 + \max_{k\in[m]}\left\lceil \frac{n_k}{\tau_k}\right\rceil \frac{15\|\mathsf{C}\|_\infty(2 + 3\|\mathsf{C}\|_\infty)}{\eta\varepsilon}. \tag{27}$$

The results given in the above theorem are entirely novel when $1 < \tau_k < n_k$, $k = 1, 2$. In the case $\tau_1 = \tau_2 = 1$, we obtain Greenkhorn algorithm by Altschuler et al. (2017). Then Theorems 4.4 proves global linear rate of convergence and also improves the complexity bound given in (Lin et al., 2021). Indeed, assuming for simplicity that $n_1 = n_2 = n$ and defining $\mathsf{a}_{\min} = \min_{k\in[m], j\in[n]} \mathsf{a}_{k,j}$, Lin et al. (2021) shows that Greenkhorn algorithm meets the stopping criteria $d_t \leq \varepsilon$ in the following number of iterations

$$t \leq 2 + 112n\frac{\|\mathsf{C}\|_\infty/\eta + \log n + 2\log(\mathsf{a}_{\min}^{-1})}{\varepsilon},$$

whereas (27) gives $t \leq 2 + 15n\|\mathsf{C}\|_\infty(2 + 3\|\mathsf{C}\|_\infty)/(\eta\varepsilon)$. We see that our result removes the logarithmic factors in the bound. Moreover, our result in terms of normalized cycles yields $T \leq 1 + 15\|\mathsf{C}\|_\infty(2 + 3\|\mathsf{C}\|_\infty)/(\eta\varepsilon)$, which is, in addition, independent on the dimension n.

4.2. MultiSinkhorn

The following theorem provides new insights into the greedy algorithm proposed by Lin et al. (2020).

Theorem 4.5. *Suppose that for all $k \in [m]$ $\tau_k = n_k$. Then* $\texttt{BatchGreenkhorn}(\mathsf{a}_1, \ldots, \mathsf{a}_m, \mathsf{C}, \eta, \tau)$, *converges linearly with the global rate*

$$\mathsf{KL}(\pi^\star, \pi^t) \leq \left(1 - \frac{e^{-(12m-7)\|\mathsf{C}\|_\infty/\eta}}{m - 1}\right)^t \mathsf{KL}(\pi^\star, \pi^0). \tag{28}$$

Moreover, the number of iterations required to reach the stopping criterion $d_t^\infty \leq \varepsilon$ satisfies

$$t \leq 1 + \frac{8(4m - 3)\|\mathsf{C}\|_\infty}{\eta\varepsilon}. \tag{29}$$

Concerning the linear convergence rate, we notice that when the batch is full ($\tau_k = n_k$, $k \in [m]$), cyclic Sinkhorn of (Benamou et al., 2015; Carlier, 2021) and (greedy) MultiSinkhorn algorithm generally differ (unless $m = 2$). Our results shows (for the first time) that the rate of convergence of MultiSinkhorn algorithm is strictly better than that of the cyclic Sinkhorn algorithm obtained in (Carlier, 2021). Indeed in (Carlier, 2021), the following rate was provided

$$\mathsf{KL}(\pi^\star, \pi^{mT}) \leq \left(1 - \frac{e^{-8(2m-1)\|\mathsf{C}\|_\infty/\eta}}{m}\right)^T \mathsf{KL}(\pi^\star, \pi^0),$$

where T counts the number of cycles, each one consisting of m KL projections on the given marginals. Whereas, (28) shows that for MultiSinkhorn

$$\mathsf{KL}(\pi^\star, \pi^{mT}) \leq \left[\left(1 - \frac{e^{-(12m-7)\|\mathsf{C}\|_\infty/\eta}}{m - 1}\right)^m\right]^T \mathsf{KL}(\pi^\star, \pi^0),$$

which mainly gains an m-power in the rate.

As for the iteration complexity, in terms of normalized cycles $T = t/m$ and with the stopping criterion $d_{mT} \leq \varepsilon$, our result (29) yields $T \leq 1 + 8(4m - 3)\|\mathsf{C}\|_\infty/(\eta\varepsilon)$, which improves the existing one from (Lin et al., 2020), that is, $T \leq 1 + 2m(\|\mathsf{C}\|_\infty/\eta + \log(\mathsf{a}_{\min}^{-1}))/\varepsilon$. The latter bound contains the term $\log(\mathsf{a}_{\min}^{-1})$ which is at best $\log(n)$, and, hence, depends on the dimension of the problem. Our result removes this dependency. Additionally, we note that

by using our stopping criterion $d^{\infty}_{mT} \leq \varepsilon$, we obtain the complexity

$$T \leq 1 + \frac{32\|\mathsf{C}\|_{\infty}}{\eta\varepsilon},$$

which is, in addition, independent on the number of marginals m.

We end this section by briefly discussing the linear convergence rate when $m = 2$, for which MultiSinkhorn reduces to the classical Sinkhorn algorithm. In this case, our result shows better rate $(1-e^{-17\|\mathsf{C}\|_{\infty}/\eta})^2$ than the one that can be derived from (Carlier, 2021), that is, $1 - (1/2)e^{-24\|\mathsf{C}\|_{\infty}/\eta}$.

5. Numerical experiments

In this section we test the BATCHGREENKHORN algorithm on a number of OT problems relevant to machine learning in order to empirically check our theoretical findings. We treat the bi-marginal and multimarginal ($m \geq 3$) settings separately. We are interested in the following two aspects.

- Can the batch size be exploited to make BatchGreenkhorn faster than (cyclic) Sinkhorn?
- Does the (greedy) MultiSinkhorn perform better than the cyclic Sinkhorn, as our global rate predicts?

In the reminder of this section, we have that $n_k = n$ for every $k \in [m]$, $\tau \in [n]$ and will take cyclic Sinkhorn as a baseline. We introduce the quantity

$$d^{\infty}(\pi^{Tmn/\tau}) := \max_{k\in[m]} \|\mathsf{a}_k - \mathsf{R}_k(\pi^{Tmn/\tau})\|_1, \qquad (30)$$

and we measure the performance of the different algorithms using two different metrics: (1) the *competitive ratio of the residuals* $\rho_\tau(T) := d^{\infty}(\pi^{Tmn/\tau})/d^{\infty}(\widehat{\pi}^{T})$ where $\pi^{Tmn/\tau}$ is the output of BatchGreenkhorn, while $\widehat{\pi}^{T}$ is the output of (cyclic) Sinkhorn, and (2) the *speedup in computational time* σ_τ, defined as the ratio between BatchGreenkhorn time and (cyclic) Sinkhorn time, for achieve the precision 10^{-6}, that is, $d^{\infty}(\pi^{Tmn/\tau}) \leq 10^{-6}$.

In our experiments we plot the first measure against the number of normalized cycles T and the second measure vs the relative batch size τ/n and the relative regularization parameter $\eta/\|\mathsf{C}\|_{\infty}$. The algorithms are implemented in Pytorch and the tests were run on a single Tesla M40 GPU. The codes are available at `www.github.com/CSML-IIT-UCL/RMOT`. We show several results of the experiments, while in Appendix D a more detailed report is provided.

Bi-marginal setting. In this setup we test BatchGreenkhorn and Sinkhorn on the task of computing entropic-regularized 2-Wasserstein distance between large point-clouds. In order to treat these large scale OT problems we implement both algorithms using KeOps library (Charlier et al., 2021) achieving a linear memory footprint suitable to GPUs. First, we consider a random pair of objects from ModelNet10 dataset (Wu et al., 2015) that contains CAD models as pre-aligned shapes from 10 categories. Then, in ten random trials we sample point-clouds of size n from each one as marginal distributions, and test the algorithms for different regularization parameters η. The results are reported in Fig. 1. The first two rows show that with one only exception BatchGreenkhorn always outperforms cyclic Sinkhorn in both metrics described above. The last row shows for sizes $n = 30,000$ (Left) and $n = 50,000$ (Right) and different regularization parameters, the speedup factor. In both cases the best overall performance is achieved at $\tau \approx 5000$. Additionally, in Appendix D for this example we illustrate that, in fact, BatchGreenkhorn can be further accelerated using simulated annealing strategy from (Feydy et al., 2019) developed for the classical Sinkhorn. We note, however, that integrating simulated annealing and greedy strategies is a delicate matter, which requires further investigation and can be an interesting future research direction.

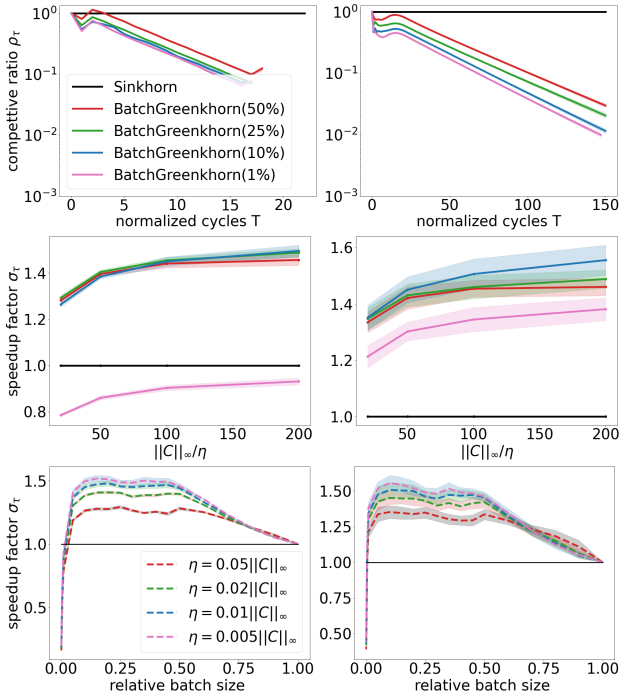

Figure 1: Bi-marginal ModelNet10 experiment. First row shows competitive ratios $\rho_\tau(T)$ for size $n = 50000$ and two relative regularization parameters $\eta = 0.05\|\mathsf{C}\|_{\infty}$ (Left) and $\eta = 0.005\|\mathsf{C}\|_{\infty}$ (Right). Second row shows the speed up σ_τ vs. relative batch size τ/n for two sizes $n = 30,000$ (Left) and $n = 50,000$ (Right). In all plots, the mean is bold and $\pm$ standard deviation is shaded.

As a second experiment, we evaluate the performance of algorithms on the task of label-to-label distance considered in (Alvarez-Melis & Fusi, 2020) for CIFAR-10 training dataset (Krizhevsky, 2009) that consists of 50,000 32x32

color images in 10 labeled classes. This leads to the computation of 45 entropic-regularized 2-Wasserstein distances between point-clouds of dimension 3072. Given a tolerance $\varepsilon = 10^{-6}$, for moderately small regularization $\eta = 0.04\|\mathsf{C}\|_\infty$ BatchGreenkhorn(12.5%) algorithm completed the task in 41.98 min compared to Sinkhorn which took 65.32 min, see Table 4 in Appendix D. This shows again the practical benefit of using the batch in this problem.

Multi-marginal setting. In this setup, due to the lack of KeOps equivalent for higher-order tensors, we implement the algorithms by pre-computing the kernel tensor and then storing it on GPU. Clearly, this implies strong memory limits on the size of the problem one can tackle. In fact, for large scale problems (moderate $m \geq 3$ and moderate n), the computation of the kernel tensor actually greatly dominates the computation time of all algorithms. However,we still provide some preliminary results on the role of greediness and the batch size in RMOT.

First, we consider a synthetic experiment in order to validate that with growing number of marginals m greedy MultiSinkhorn converges progressively faster than the cyclic Sinkhorn, as our theoretical analyses in Sec. 4.2 predicts. For each $m \in \{3, 6, 9, 12\}$ we randomly generate m 3-point clouds in 1D of the same form (x_1, x_2, x_3) where $x_i \sim \mathcal{N}(-2 + i, 0.1)$, $i = 1, 2, 3$. As a ground cost we use the square of Euclidean distance and we test both algorithms for ten trials and show competitive ratio $\rho_\tau(T)$, where $\tau = n = 3$ (full batch) for every m in Fig. 2.

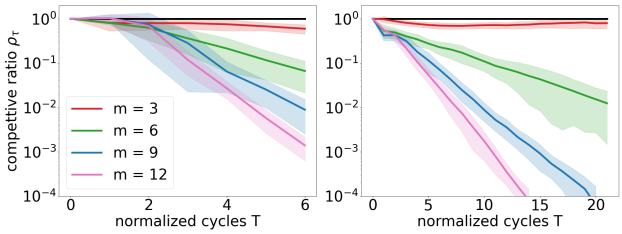

Figure 2: Comparison of convergence rates for (greedy) MultiSinkhorn and cyclic Sinkhorn on m synthetically generated 3-point-clouds for two relative regularization parameters $\eta = 0.1\|\mathsf{C}\|_\infty$ (Left) and $\eta = 0.05\|\mathsf{C}\|_\infty$ (Right).

Next, we follow the proposal of (Frogner & Poggio, 2019) where given m discrete distributions, one can formulate the MAP estimation of their joint distribution via Bayesian inference as entropic RMOT for an appropriately chosen prior. From ModelNet3D dataset, we randomly pick $m = 6$ objects, for which we wish to infer the joint distribution. Since our data comes from the aligned 3D objects, using the square of the Euclidean distance as a cost in the prior is a suitable choice. For two small sample sizes, $n = 8$ and $n = 16$, we construct point-clouds from the objects and run cyclic Sinkhorn, (greedy) MultiSinkhorn and BatchGreenkhorn

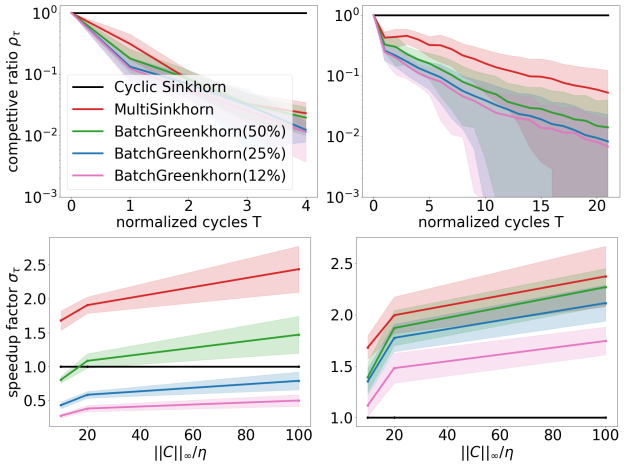

Figure 3: Multi-marginal ModelNet10 experiment. First row: competitive ratios $\rho_\tau(T)$ for size $n = 16$ and two relative regularization parameters $\eta = 0.1\|\mathsf{C}\|_\infty$ (Left) and $\eta = 0.01\|\mathsf{C}\|_\infty$ (Right). Second row: speedup factor vs relative regularization parameters for two sizes $n = 8$ (Left) and $n = 16$ (Right).

with three batch sizes. Fig. 3 shows the results for ten random trials. We observe that w.r.t. distance to the transport polytope, BatchGreenkhorn outperforms cyclic Sinkhorn for every batch size, while in terms of computational time full batch, i.e. (greedy) MultiSinkhorn, performs the best. This is not surprising since the size of each marginal is too small (12.5% relative batch size for $n = 8$ gives multimarginal Greenkhorn, i.e. $\tau = 1$) for the batch to have an effect.

6. Conclusions, limitations and future work

We presented *batch Greenkhorn*, a new algorithm for solving multimarginal entropic-regularized optimal transport problems, which is a natural extension of the Greenkhorn algorithm. It greedily selects at each iteration a marginal and batch of its components. We study the convergence of the algorithm in the framework of the iterative Bregman projections method, providing novel linear rate of convergence as well as iteration complexity bounds. We made a comprehensive comparison with existing results showing the improvements of our methodology over the state-of-the-art. In addition, we presented numerical experiments that illustrate how the new flexibility of batch can be exploited in practice to speed up the Sinkhorn algorithm. A problem which remains open is that of deriving bounds on the dual variables with an explicit dependence on the given problem data for $m \geq 3$ when the batch is not full. According to our general Theorems 4.1 and 4.2, this will allow to have explicit global linear rate and iteration complexity in all possible cases. Additional research directions are the extensions of our analysis to infinite dimensions and general convex regularizers, implementing batch Greenkhorn with structured costs, and analyze the impact of parallel computations.

Acknowledgements

We would like to thank the reviewers for all their valuable comments and suggestions, which helped us to improve the quality of the manuscript.

References

Agueh, M. and Carlier, G. Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Altschuler, J., Weed, J., and Rigollet, P. Near-Linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1961–1971, 2017.

Altschuler, J. M. and Boix-Adserà, E. Hardness results for Multimarginal Optimal Transport Problems. *Discrete Optimization*, 42:100669, Nov 2021.

Alvarez-Melis, D. and Fusi, N. Geometric dataset distances via optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21428–21439, 2020.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, pp. A1111–A1138, 2015. doi: 10.1137/141000439.

Benamou, J.-D., Carlier, G., and Nenna, L. A Numerical Method to Solve Multi-marginal Optimal Transport Problems with Coulomb Cost. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pp. 577–601. Springer, 2016.

Bento, J. and Mi, L. Multi-Marginal Optimal Transport Defines a Generalized Metric, 2021. arXiv: 2001.11114.

Bregman, L. The Relaxation Method of Finding the Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Cao, J., Mo, L., Zhang, Y., Jia, K., Shen, C., and Tan, M. Multi-marginal Wasserstein GAN. In *Advances in Neural Information Processing Systems*, pp. 1774–1784, 2019.

Carlier, G. On the Linear Convergence of the Multimarginal Sinkhorn Algorithm, 2021. hal-03176512.

Censor, Y. and Lent, A. An Iterative Row-action Method for Interval Convex Programming. *Journal of Optimization Theory and Applications*, 34:321–353, 1981.

Censor, Y. and Reich, S. Iterations of Paracontractions and Firmaly Nonexpansive Operators with Applications to Feasibility and Optimization. *Optimization*, 37:323–339, 1996.

Charlier, B., Feydy, J., Glaunès, J. A., Collin, F.-D., and Durif, G. Kernel operations on the gpu, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.

Cohen, S., Terenin, A., Pitcan, Y., Amos, B., Deisenroth, M. P., and Kumar, K. S. S. Sliced multi-marginal optimal transport, 2021. arXiv: 2102.07115.

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

Di-Marino, S. and Gerolin, A. An Optimal Transport Approach for the Schrödinger Bridge Problem and Convergence of Sinkhorn Algorithm. *J. Sci. Comput.*, 85:27, 2020.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational Optimal Transport: Complexity by Accelerated Gradient Descent is Better Than by Sinkhorn's Algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1367–1376, 2018.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Frogner, C. and Poggio, T. Fast and Flexible Inference of Joint Distributions from their Marginals. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2002–2011, 2019.

Haasler, I., Singh, R., Zhang, Q., Karlsson, J., and Chen, Y. Multi-marginal Optimal Transport and Probabilistic Graphical Models, 2020. arXiv: 2006.14113.

He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, 28:5464–5478, 2019.

Kostic, V. and Salzo, S. The Method of Bregman Projections in Deterministic and Stochastic Convex Feasibility Problems, 2021. arXiv: 2101.01704.

Krizhevsky, A. Learning multiple layers of features from tiny images, 2009. Technical report.

Lin, T., Ho, N., Cuturi, M., and Jordan, M. I. On the Complexity of Approximating Multimarginal Optimal Transport, 2020. arXiv: 1910.00152.

Lin, T., Ho, N., and Jordan, M. I. On the Efficiency of Sinkhorn and Greenkhorn and Their acceleration for Optimal Transport, 2021. arXiv:1906.01437.

Pass, B. Multi-marginal Optimal Transport: Theory and Applications. *ESAIM: M2AN*, 49(6):1771–1790, 2015.

Peyré, G. and Cuturi, M. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.

Tupitsa, N., Dvurechensky, P., Gasnikov, A., and Uribe, C. A. Multimarginal Optimal Transport by Accelerated Alternating Minimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 6132–6137, 2020.

Villani, C. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3d shapenets: A deep representation for volumetric shapes. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912–1920, 2015.

Appendices

This supplementary material is organized as follows:

- In Appendix A we provide some basic facts on Bregman divergences.
- Appendix B contains the proofs of the results stated in Section 3, notably Proposition 3.1, and gives more information on the implementation of Algorithm 1.
- In Appendix C we provide the proof of the main results of Sec. 4, concerning the linear convergence and iteration complexity of Algorithm 1.
- Finally, in Appendix D contains more extensive report on numerical experiments presented in Section 5.

For the reader's convenience all results presented in the main body of the paper are restated in this supplementary material.

A. Bregman divergences and Bregman projections

In this section we recall few facts on Bregman distance and Bregman projections onto affine sets. In the following X is an Euclidean space and $\phi\colon X \to]-\infty,+\infty]$ is an extended-real valued function. The set of minimizers of the function ϕ is denoted by $\arg\min_{x\in X}\phi(x)$, the *domain* of ϕ is $\operatorname{dom}\phi := \{x \in X \,|\, \phi(x) < +\infty\}$ and ϕ is *proper* when $\operatorname{dom}\phi \neq \varnothing$. The function ϕ is *convex* if $\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y)$ for all $x, y \in \operatorname{dom}\phi$ and $t \in [0,1]$. If the above inequality is strict when $0 < t < 1$ and $x \neq y$, the function is *strictly convex*. The function ϕ is *closed* if the sublevel sets $\{x \in X \,|\, \phi(x) \leq t\}$ are closed in X for any $t \in \mathbb{R}$. For a convex function $\phi\colon X \to]-\infty,+\infty]$, we denote by ϕ^* its *Fenchel conjugate*, that is, $\phi^*\colon X \to]-\infty,+\infty]$, $\phi^*(y) := \sup_{x\in X}\{\langle x, y\rangle - \phi(x)\}$. The conjugate of a convex function is always closed and convex, and if ϕ is proper closed and convex, then $(\phi^*)^* = \phi$.

A proper closed and convex function ϕ is *essentially smooth* if it is differentiable on $\operatorname{int}(\operatorname{dom}\phi) \neq \varnothing$, and $\|\nabla\phi(x_n)\| \to +\infty$ whenever $x_n \in \operatorname{int}(\operatorname{dom}\phi)$ and $x_n \to x \in \operatorname{bdry}(\operatorname{dom}\phi)$. The function ϕ is *essentially strictly convex* if $\operatorname{int}(\operatorname{dom}\phi^*) \neq \varnothing$ and is strictly convex on every convex subset of $\operatorname{dom}\partial\phi$. A *Legendre* function is a proper closed and convex function which is also essentially smooth and essentially strictly convex. A function is Legendre if and only if its conjugate is so. Moreover, if ϕ is a Legendre function, then $\nabla\phi\colon \operatorname{int}(\operatorname{dom}\phi) \to \operatorname{int}(\operatorname{dom}\phi^*)$ and $\nabla\phi^*\colon \operatorname{int}(\operatorname{dom}\phi^*) \to \operatorname{int}(\operatorname{dom}\phi)$ are bijective, inverses of each other, and continuous. Given a Legendre function ϕ, the *Bregman distance* associated to ϕ is the function $D_\phi\colon X \times X \to [0,+\infty]$ such that

$$D_\phi(x,y) = \begin{cases} \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle & \text{if } y \in \operatorname{int}(\operatorname{dom}\phi) \\ +\infty & \text{otherwise.} \end{cases} \tag{31}$$

Fact A.1. *Let ϕ be a Legendre function on X. Then the following hold*

(i) $(\forall\, \pi, \gamma \in \operatorname{dom}\phi) \quad D_\phi(\pi,\gamma) + D_\phi(\gamma,\pi) = \langle \pi - \gamma, \nabla\phi(\pi) - \nabla\phi(\gamma)\rangle$.

(ii) $(\forall\, \pi, \gamma \in \operatorname{dom}\phi) \quad D_\phi(\pi,\gamma) = D_{\phi^*}(\nabla\phi(\gamma), \nabla\phi(\pi))$.

(iii) *If ϕ is twice differentiable, then*
$(\forall\, \pi, \gamma \in \operatorname{dom}\phi)(\exists\, \xi \in [\pi,\gamma]) \qquad D_\phi(\pi,\gamma) = \frac{1}{2}\langle \nabla^2\phi(\xi)(\pi-\gamma), \pi-\gamma\rangle$,
where $[\pi,\gamma] = \{(1-\alpha)\pi + \alpha\gamma \,|\, \alpha \in [0,1]\}$ is the segment with end points π and γ.

(iv) *Suppose that ϕ is twice differentiable on* $\operatorname{int}(\operatorname{dom}\phi)$. *Then*

$$\big(\forall \pi \in \operatorname{int}(\operatorname{dom}\phi),\ \nabla^2\phi(\pi) \text{ is invertible}\big) \Leftrightarrow \big(\phi^* \text{ is twice differentiable}\big). \tag{32}$$

Fact A.2. *Let ϕ be a Legendre function on X and ϕ^* be it's Fenchel conjugate. If* $\operatorname{dom}\phi^*$ *is open, then the following hold*

(i) *For every $\pi \in \operatorname{int}(\operatorname{dom}\phi)$, the sublevel sets of $D_\phi(\pi,\cdot)$ are compact.*

(ii) *For every* $\pi \in \operatorname{int}(\operatorname{dom}\phi)$*, and every sequence* $(\gamma_k)_{k\in\mathbb{N}}$ *in* $\operatorname{int}(\operatorname{dom}\phi)$

$$D_\phi(\pi,\gamma_k)\to 0 \;\Rightarrow\; \gamma_k\to\pi. \tag{33}$$

Let $\mathcal{C}\subset X$ be an affine set, represented as follows

$$A\colon X\to Y,\qquad b\in\operatorname{Im}(A),\qquad \mathcal{C}:=\{\pi\in X\,|\,A\pi=b\}, \tag{34}$$

for some linear operator A between X and another Euclidean space Y. Given a Legendre function $\phi\colon X\to\,]-\infty,+\infty]$ and $\pi\in\operatorname{int}(\operatorname{dom}\phi)$, the *Bregman projection* of π onto $\mathcal{C}$ is defined as the unique solution, denoted by $\mathcal{P}^\phi_{\mathcal{C}}(\pi)$, of the optimization problem

$$\min_{\gamma\in\mathcal{C}} D_\phi(\gamma,\pi)=\min_{\gamma\in\mathcal{C}}\phi(\gamma)-\phi(\pi)-\langle\gamma-\pi,\nabla\phi(\pi)\rangle \tag{35}$$

and the optimal value defines the *Bregman distance from* π *to* $\mathcal{C}$ and is denoted by $D^\phi_{\mathcal{C}}(\pi)$. The dual of the above problem is

$$\min_{\lambda\in Y}\phi^*(\nabla\phi(\pi)+A^*\lambda)-\phi^*(\nabla\phi(\pi))-\langle b,\lambda\rangle \tag{36}$$

and strong duality holds, meaning that

$$\min_{\gamma\in\mathcal{C}} D_\phi(\gamma,\pi)=-\min_{\lambda\in Y}\big[\phi^*(\nabla\phi(\pi)+A^*\lambda)-\phi^*(\nabla\phi(\pi))-\langle b,\lambda\rangle\big]. \tag{37}$$

Moreover, the following KKT conditions hold for a couple $(\pi^\star,\lambda^\star)$ solving the primal and dual problem above

$$\pi^\star\in\operatorname{int}(\operatorname{dom}\phi),\quad A\pi^\star=b\quad\text{and}\quad\nabla\phi(\pi)+A^*\lambda^\star=\nabla\phi(\pi^\star). \tag{38}$$

Note that the KKT conditions characterizes the projection, so that

$$\pi^\star=\mathcal{P}^\phi_{\mathcal{C}}(\pi)\;\Leftrightarrow\;\big(\pi^\star\in\operatorname{int}(\operatorname{dom}\phi),\;A\pi^\star=b,\;\text{and}\;\nabla\phi(\pi)-\nabla\phi(\pi^\star)\in\operatorname{Im}(A^*)\big). \tag{39}$$

Finally we mention the *generalized Pythagoras theorem*. If $\mathcal{C}_1$ is an affine set such that $\mathcal{C}\subset\mathcal{C}_1$, then, for every $\pi\in\operatorname{int}(\operatorname{dom}\phi)$ it holds

$$D^\phi_{\mathcal{C}}(\pi)=D^\phi_{\mathcal{C}_1}(\pi)+D^\phi_{\mathcal{C}}(\mathcal{P}^\phi_{\mathcal{C}_1}(\pi)). \tag{40}$$

Moreover, in this case $\mathcal{P}^\phi_{\mathcal{C}}(\pi)=\mathcal{P}^\phi_{\mathcal{C}}(\mathcal{P}^\phi_{\mathcal{C}_1}(\pi))$, and

$$(\forall\gamma\in\operatorname{int}(\operatorname{dom}\phi))\quad\nabla\phi(\gamma)-\nabla\phi(\pi)\in\operatorname{Im}(A^*)\Longleftrightarrow\mathcal{P}_{\mathcal{C}}(\gamma)=\mathcal{P}_{\mathcal{C}}(\pi). \tag{41}$$

In the following we let $\phi\colon\mathbb{R}\to\,]-\infty,+\infty[$ be the (negative) *Boltzmann-Shannon entropy*, that is,

$$\phi(t)=\begin{cases}t\log t-t & \text{if } t>0\\ 0 & \text{if } t=0\\ +\infty & \text{if } t<0.\end{cases}$$

It is clear that $\phi^*(s)=\exp(s)$. We define the Bregman distance associated to ϕ

$$\begin{aligned}D_\phi(s,t)&=\begin{cases}\phi(s)-\phi(t)-\phi'(t)(s-t) & \text{if } t>0\\ +\infty & \text{otherwise.}\end{cases}\\ &=\begin{cases}s\log\frac{s}{t}-s+t & \text{if } t>0\\ +\infty & \text{otherwise,}\end{cases}\end{aligned} \tag{42}$$

which is nothing but the *Kullback-Leibler divergence* on $\mathbb{R}$.

Proposition A.3. *Let* $M>0$*. The following hold.*

(i) *The function ϕ is strongly convex on the interval $]0, M]$ with modulus of strong convexity equal to $1/M$. Moreover, for every $a > 0$ and $s, t \in \mathbb{R}_{++}$, $D_\phi(s,t) = aD_\phi(s/a, t/a)$.*

(ii) *The function ϕ^* is strongly convex on the interval $[-M, +\infty[$ with modulus of strong convexity equal to $\exp(-M)$.*

Proof. (i): It follows from the fact that the second derivative of ϕ is $\phi''(t) = 1/t$, which is bounded from below away from zero on the interval $]0, M]$ by the constant $1/M$. the second part follows directly from the definition (42).

(ii): It follows from the fact that the second derivative of $\exp$ is bounded from below away from zero on the interval $[-M, +\infty[$ by $\exp(-M)$. □

The negative entropy and the Kullback-Leibler divergence on $\mathbb{X}$ are

$$\mathsf{H}(\gamma) = \sum_j \phi(\gamma_j) \quad \text{and} \quad \mathsf{KL}(\gamma, \pi) = \sum_j D_\phi(\gamma_j, \pi_j). \tag{43}$$

Lemma A.4. *Let $\pi, \gamma, \alpha \in \mathbb{X}_{++}$ and suppose that*

$$0 < M_{\min} \le \min_j \frac{\min\{\pi_j, \gamma_j\}}{\alpha_j} \le \max_j \frac{\max\{\pi_j, \gamma_j\}}{\alpha_j} \le M_{\max}.$$

Then, setting $\mathsf{A} = \alpha \odot (\cdot) \colon \mathbb{X} \to \mathbb{X}$ (which is a positive diagonal operator), we have

$$\mathsf{KL}(\pi, \gamma) \ge \max\left\{ \frac{M_{\min}}{2} \|\log \pi - \log \gamma\|_{\mathsf{A}}^2, \frac{1}{2M_{\max}} \|\pi - \gamma\|_{\mathsf{A}^{-1}}^2 \right\}. \tag{44}$$

Proof. It follows from Proposition A.3(i) that, for $a > 0$ and $s, t > 0$ such that $s/a, t/a \le M$, we have $D_\phi(s,t) = aD_\phi(s/a, t/a) \ge a(2M)^{-1}|s/a - t/a|^2 = (2M)^{-1}a^{-1}|s-t|^2$. Thus, since $\gamma_j/\alpha_j, \pi_j/\alpha_j \le M_{\max}$, we have

$$\mathsf{KL}(\pi, \gamma) = \sum_j D_\phi(\pi_j, \gamma_j) \ge \frac{1}{2M_{\max}} \sum_j \frac{1}{\alpha_j} |\pi_j - \gamma_j|^2 = \frac{1}{2M_{\max}} \|\pi - \gamma\|_{\mathsf{A}^{-1}}^2.$$

Now, it follows from Proposition A.3(ii) that for every $a, s, t > 0$ such that $s/a, t/a \ge e^{-M}$ we have $\log(s/a), \log(t/a) \ge -M$ and hence $D_\phi(s,t) = aD_\phi(s/a, t/a) = aD_{\phi^*}(\log(s/a), \log(t/a)) \ge a(e^{-M}/2)|\log s - \log t|^2$. Therefore, since $\pi_j/\alpha_j, \gamma_j/\alpha_j \ge M_{\min}$, we have

$$\mathsf{KL}(\pi, \gamma) = \sum_j D_\phi(\pi_j, \gamma_j) \ge \frac{M_{\min}}{2} \sum_j \alpha_j |\log \pi_j - \log \gamma_j|^2 = \frac{M_{\min}}{2} \|\log \pi - \log \gamma\|_{\mathsf{A}}^2.$$ □

B. BatchGreenkhorn algorithm and its implementation

Here we provide proofs of the results in Sec. 3.

Proposition 3.1. *For every $\pi \in \mathbb{X}_+$, $k \in [m]$ and $L \subset [n_k]$,*

$$\mathcal{P}_{\Pi_{(k,L)}}(\pi) = \pi \odot \exp\left(\mathsf{R}^*_{(k,L)}(\bar{\mathsf{u}})\right), \tag{17}$$

where

$$\bar{\mathsf{u}} = \log \frac{\mathsf{a}_{k|L}}{\mathsf{R}_k(\pi)_{|L}}, \tag{18}$$

and, consequently, for every $j \in \mathcal{J}$,

$$(\mathcal{P}_{\Pi_{(k,L)}}(\pi))_j = \pi_j \times \begin{cases} \dfrac{a_{k,j_k}}{\mathsf{R}_k(\pi)_{j_k}} & \text{if } j_k \in L, \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

Moreover,

$$\mathsf{KL}_{\Pi_{(k,L)}}(\pi) = \mathsf{KL}(\mathsf{a}_{k|L}, \mathsf{R}_k(\pi)_{|L}). \tag{20}$$

Proof. It follows from (11) that

$$\Pi_{(k,L)} = \Big\{ \pi \in \mathbb{X} \,\big|\, \mathsf{R}_{(k,L)}(\pi) = \mathsf{a}_{k|L} \Big\}. \tag{45}$$

Then the first equality in (17) follows directly form the KKT conditions (38), which in this case yields

$$\mathsf{R}_{(k,L)}(\bar{\pi}) = \mathsf{a}_{k|L} \quad \text{and} \quad \bar{\pi} = \nabla \mathsf{H}^* \big(\nabla \mathsf{H}(\pi) + \mathsf{R}^*_{(k,L)}(\bar{\mathsf{u}}) \big), \tag{46}$$

where, according to (36), the dual parameter $\bar{\mathsf{u}} \in \mathbb{R}^L$ solves the minimization problem in (18). Now, since for every $j \in \mathcal{J}$, $(\nabla \mathsf{H}(\pi))_j = \log(\pi_j)$ and $(\nabla \mathsf{H}^*(\gamma))_j = \exp(\gamma_j)$, then (46) gives

$$\bar{\pi}_j = \exp(\log(\pi_j) + \mathsf{R}^*_{(k,L)}(\bar{\mathsf{u}}))_j) = \pi_j \exp\big((\mathsf{R}^*_{(k,L)}(\bar{\mathsf{u}}))_j \big)$$

and the second equality in (17) follows.

Now, let $J^{(k)}_L \colon \mathbb{R}^L \to \mathbb{R}^{n_k}$ be the canonical injection of $\mathbb{R}^L$ into $\mathbb{R}^{n_k}$. Then, recalling the definition of $\mathsf{R}_{(k,L)}$ in (10), we have $\mathsf{R}_{(k,L)} = J^{(k)*}_L \mathsf{R}_k$ and hence $\mathsf{R}^*_{(k,L)} = \mathsf{R}^*_k J^{(k)}_L$, where $\mathsf{R}^*_k \colon \mathbb{R}^{n_k} \to \mathbb{X}$ acts as $(\mathsf{R}^*_k \mathsf{v})_j = \mathsf{v}_{j_k}$. Therefore, for every $j \in \mathcal{J}$,

$$(\mathsf{R}^*_{(k,L)} \bar{\mathsf{u}})_j = (\mathsf{R}^*_k J^{(k)}_L \bar{\mathsf{u}})_j = (J^{(k)}_L \bar{\mathsf{u}})_{j_k} = \begin{cases} \bar{u}_{j_k} & \text{if } j_k \in L \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\bar{\pi}_j = \pi_j \times \begin{cases} e^{\bar{u}_{j_k}} & \text{if } j_k \in L \\ 1 & \text{otherwise.} \end{cases} \tag{47}$$

On the other hand, since $\mathsf{a}_{k|L} = J^{(k)*}_L \mathsf{R}_k \bar{\pi}$, by (47), we derive that, for every $j_k \in L$,

$$a_{k,j_k} = (\mathsf{R}_k \bar{\pi})_{j_k} = \sum_{j_{-k} \in \mathcal{J}_{-k}} \bar{\pi}_{(j_{-k}, j_k)} = e^{\bar{u}_j} (\mathsf{R}_k \pi)_j, \tag{48}$$

so that $e^{\bar{u}_{j_k}} = a_{k,j_k} / (\mathsf{R}_k \pi)_{j_k}$. Hence now (19) follows from (47). Concerning the formula for the distance, by (19), we have that,

$$\begin{aligned} D^{\phi}_{\Pi_{(k,L)}}(\pi) &= D^{\phi}(\bar{\pi}, \pi) \\ &= \sum_{j \in \mathcal{J}} \bar{\pi}_j \log\Big(\frac{\bar{\pi}_j}{\pi_j}\Big) - \bar{\pi}_j + \pi_j \\ &= \sum_{j_k \in L} \sum_{j_{-k} \in \mathcal{J}_{-k}} \bar{\pi}_{(j_{-k}, j_k)} \log\Big(\frac{a_{k,j_k}}{(\mathsf{R}_k \pi)_{j_k}}\Big) - \bar{\pi}_{(j_{-k}, j_k)} + \pi_{(j_{-k}, j_k)} \\ &= \sum_{j_k \in L} a_{k,j_k} \log\Big(\frac{a_{k,j_k}}{(\mathsf{R}_k \pi)_{j_k}}\Big) - a_{k,j_k} + (\mathsf{R}_k \pi)_{j_k} \\ &= \mathsf{KL}(\mathsf{a}_{k|L}, \mathsf{R}_k(\pi)_{|L}), \end{aligned}$$

which completes the proof. □

The following proposition justifies equation (21) and Algorithm 1.

Proposition B.1. *Let $(\pi^t)_{t \in \mathbb{N}}$ be defined according to algorithm (13). Then, we have*

$$(\forall\, t \in \mathbb{N}) \qquad \pi^t = \exp\Big(-\mathsf{C}/\eta + \bigoplus_{k=1}^{m} \mathsf{v}^t_k \Big) \odot \bigotimes_{k=1}^{m} \mathsf{a}_k, \tag{49}$$

where $\mathsf{v}^t_k = (v^t_{k,j})_{1 \le j \le n_k} \in \mathbb{R}^{n_k}$ and

$$\mathsf{v}^{t+1}_k = \delta_{k,k_t} J^{(k_t)}_{L_t} \mathsf{u}^t + \mathsf{v}^t_k, \quad \mathsf{u}^t = \log \mathsf{a}_{k_t|L_t} - \log(\mathsf{R}_{k_t} \pi^t)_{|L_t}, \tag{50}$$

$J^{(k_t)}_{L_t} \colon \mathbb{R}^{L_t} \to \mathbb{R}^{n_{k_t}}$ is the canonical injection, δ_{k,k_t} is the Kronecker symbol, and v^0_k, $k \in [m]$ are arbitrary.

Proof. By definition of π^{t+1} we have

$$\pi^{t+1} = \mathcal{P}_{\Pi_{(k_t,L_t)}}(\pi^t), \quad \Pi_{(k_t,L_t)} = \big\{\pi \in \mathbb{X} \,|\, J_{L_t}^{(k_t)*}\mathsf{R}_{k_t}\pi = \mathsf{a}_{k_t|L_t}\big\}. \tag{51}$$

Then it follows from Proposition 3.1 that

$$\nabla\mathsf{H}(\pi^{t+1}) = \nabla\mathsf{H}(\pi^t) + \mathsf{R}_{k_t}^* J_{L_t}^{(k_t)}\mathsf{u}^t \tag{52}$$

with $\mathsf{u}^t \in \mathbb{R}^{L_t}$. Therefore, applying the above equation recursively, we get

$$\begin{aligned}
\nabla\mathsf{H}(\pi^t) &= \nabla\mathsf{H}(\pi^0) + \sum_{s=0}^{t-1} \mathsf{R}_{k_s}^* J_{L_s}^{(k_s)}\mathsf{u}^s \\
&= \nabla\mathsf{H}(\pi^0) + \sum_{s=0}^{t-1}\sum_{k=1}^{m} \mathsf{R}_k^*\delta_{k,k_s} J_{L_s}^{(k_s)}\mathsf{u}^s \\
&= \nabla\mathsf{H}(\pi^0) + \sum_{k=1}^{m}\sum_{s=0}^{t-1} \mathsf{R}_k^*\delta_{k,k_s} J_{L_s}^{(k_s)}\mathsf{u}^s \\
&= \nabla\mathsf{H}(\pi^0) + \sum_{k=1}^{m} \mathsf{R}_k^*\Big(\sum_{s=0}^{t-1} \delta_{k,k_s} J_{L_s}^{(k_s)}\mathsf{u}^s\Big).
\end{aligned}$$

Then if we set $\mathsf{v}_k^t = \sum_{s=0}^{t-1}\delta_{k,k_s} J_{L_s}^{(k_s)}\mathsf{u}^s \in \mathbb{R}^{n_k}$, we have (recalling that $\pi_0 = e^{-\mathsf{C}/\eta} \odot \otimes_{k=1}^m \mathsf{a}_k$)

$$\pi_j^t = \exp\Big(\log(\pi_j^0) + \sum_{k=1}^{m}(\mathsf{R}_k^*\mathsf{v}_k^t)_j\Big) = \exp\Big(-\mathsf{C}_j/\eta + \sum_{k=1}^{m} v_{k,j_k}^t\Big)\prod_{k=1}^{m} a_{k,j_k}.$$

Moreover, it is clear that

$$\mathsf{v}_k^{t+1} = \sum_{s=0}^{t}\delta_{k,k_s} J_{L_s}^{(k_s)}\mathsf{u}^s = \delta_{k,k_t} J_{L_t}^{(k_t)}\mathsf{u}^t + \mathsf{v}_k^t$$

Finally, it follows from (18) that, for every $j_{k_t} \in L_t$, $e^{u_{j_{k_t}}^t} = a_{k_t,j_{k_t}}/(\mathsf{R}_k\pi^t)_{j_{k_t}}$, so that

$$(\forall\, j_{k_t} \in L_t) \quad u_{j_{k_t}}^t = \log a_{k_t,j_{k_t}} - \log\big((\mathsf{R}_k\pi^t)_{j_{k_t}}\big).$$

The statement follows. □

Next we give more detailed implementation of the batch Greenkhorn given in Algorithm 1.

Remark B.2 (Implementation details on Algorithm 1)**.** *The most delicate part is to avoid recomputing the marginals* $\mathsf{R}_k(\pi^t) = \mathsf{r}_k^t = (r_{k,j_k}^t)_{j_k\in[n_k]}$, $k \in [m]$, *(step 5) that are necessary for making the greedy choice in step 3. Now, due to equation* (19) *in Proposition 3.1, we have that*

$$\pi_j^{t+1} = \pi_j^t \times \begin{cases} \dfrac{a_{k_t,j_{k_t}}}{\mathsf{R}_{k_t}(\pi^t)_{j_{k_t}}} & \text{if } j_{k_t} \in L_t, \\ 1 & \text{otherwise} \end{cases} \tag{53}$$

and hence

$$r_{k_t,j_{k_t}}^{t+1} = \begin{cases} a_{k_t,j_{k_t}} & j_{k_t} \in L_t, \\ r_{k_t,j_{k_t}}^t & \text{otherwise}. \end{cases} \tag{54}$$

To derive update formula for the other marginals, observe that for all $k \neq k_t$ *it follows from* (49) *that*

$$\begin{aligned}
r_{k,j_k}^{t+1} &= \sum_{j_{-k}\in\mathcal{J}_{-k}} \exp\Big(\log\pi^0_{(j_{-k},j_k)} + \sum_{h\neq k} v_{h,j_h}^{t+1} + v_{k,j_k}^{t+1}\Big) \\
&= \sum_{j_{-k}\in\mathcal{J}_{-k}} \exp\Big(\log\pi^0_{(j_{-k},j_k)} + \sum_{h\notin\{k,k_t\}} v_{h,j_h}^{t} + v_{k_t,j_{k_t}}^{t+1} + v_{k,j_k}^{t}\Big).
\end{aligned}$$

So, since according to (50), $v^{t+1}_{k,j_{k_t}} = v^t_{k,j_{k_t}}$ if $j_{k_t} \notin L_t$ and $v^{t+1}_{k,j_{k_t}} = v^t_{k,j_{k_t}} + \log(a_{k_t,j_{k_t}}/r^t_{k_t,j_{k_t}})$ if $j_{k_t} \in L_t$, we have

$$\begin{aligned} r^{t+1}_{k,j_k} &= \sum_{\substack{j_{-k}\in\mathcal{J}_{-k}\\ j_{k_t}\notin L_t}} \exp\Big(\log \pi^0_{(j_{-k},j_k)} + \sum_{h\neq k} v^t_{h,j_h} + v^t_{k,j_k}\Big) \\ &\quad + \sum_{\substack{j_{-k}\in\mathcal{J}_{-k}\\ j_{k_t}\in L_t}} \exp\Big(\log \pi^0_{(j_{-k},j_k)} + \sum_{h\neq k} v^t_{h,j_h} + v^t_{k,j_k}\Big)\frac{a_{k_t,j_{k_t}}}{r^t_{k_t,j_{k_t}}} \\ &= \sum_{\substack{j_{-k}\in\mathcal{J}_{-k}\\ j_{k_t}\in L_t}} \exp\Big(\log \pi^0_{(j_{-k},j_k)} + \sum_{h\neq k} v^t_{h,j_h} + v^t_{k,j_k}\Big)\Big(\frac{a_{k_t,j_{k_t}}}{r^t_{k_t,j_{k_t}}} - 1\Big) + r^t_{k,j_k}. \end{aligned}$$

Therefore, at each iteration $t \geq 0$ we will construct an auxiliary tensor $\widetilde{\pi}^t \in \mathbb{R}_+^{n_1\times\cdots\times n_{k_t-1}\times 1\times n_{k_t+1}\times\cdots\times n_m}$ by

$$\widetilde{\pi}^t_{j_1,\ldots,j_{k_t-1},1,j_{k_t+1},\ldots,j_m} = \sum_{j_{k_t}\in L_t} \exp\Big(\log \pi^0_{j_1,\ldots,j_{k_t-1},j_{k_t},j_{k_t+1},\ldots,j_m} + \sum_{k\in[m]} v^t_{k,j_k} \tag{55}$$

$$+ \log|a_{k_t,j_{k_t}} - r^t_{k_t,j_{k_t}}| - \log(r^t_{k_t,j_{k_t}})\Big)\operatorname{sgn}(a_{k_t,j_{k_t}} - r^t_{k_t,j_{k_t}}), \tag{56}$$

in order to obtain that for every $k \neq k_t$, $\mathsf{r}_k^{t+1} = \mathsf{r}_k^{t+1} + \mathsf{R}_k(\hat{\pi}^t)$. Hence, we can use $\hat{\pi}^t$ to efficiently update non-active marginals without recomputing them from scratch. Moreover, note that using (55)-(56) one avoids excessive numerical errors when $a_{k,j} \approx r^t_{k,j}$. These observations lead us to the following implementation of BATCHGREENKHORN.

Algorithm 2 `BatchGreenkhorn`$(\mathsf{a}_1,\ldots,\mathsf{a}_m,\mathsf{C},\eta,\tau)$

Input: $\mathsf{C}\in\mathbb{X}_+$, $\eta>0$, $(\mathsf{a}_1,\ldots,\mathsf{a}_m)$, $(\tau_1,\ldots,\tau_m)$, $1\leq\tau_k\leq n_k$, $\varepsilon>0$
Initialization: $\mathsf{v}_k^0 = 0$, $\mathsf{r}_k^0 = \mathsf{R}_k(\exp(-\mathsf{C}/\eta)\odot\otimes_{k=1}^m \mathsf{a}_k)$, $k\in[m]$
while $\sum_{k\in[m]}\|\mathsf{a}_k - \mathsf{r}_k^t\|_1 > \varepsilon$ **do**
  for $k\in[m]$ **do**
    Compute vectors p_k as $p_{k,j} := \mathsf{KL}(a_{k,j}, r^t_{k,j})$, for $j\in[n_k]$
    Take L'_k to be τ_k largest elements of p_k
  end for
  Choose the marginal with the best batch: $k_t \leftarrow \arg\max_{k\in[m]}\|\mathsf{p}_{k|L'_k}\|_1$ and $L_t = L'_{k_t}$
  Set $\mathsf{v}_k^{t+1} = \mathsf{v}_k^t$ and update $\mathsf{v}^{t+1}_{k_t\,|L_t} \leftarrow \mathsf{v}^{t+1}_{k_t\,|L_t} + \log(\mathsf{a}_{k_t|L_t}) - \log(\mathsf{r}^t_{k_t|L_t})$
  Set $\mathsf{r}_{k_t}^{t+1} = \mathsf{r}_{k_t}^t$ and update $\mathsf{r}^{t+1}_{k_t\,|L_t} = \mathsf{a}_{k_t|L_t}$
  for $k\in[m]\setminus\{k_t\}$ **do**
    Update $\mathsf{r}_k^{t+1} \leftarrow \mathsf{r}_k^t + \mathsf{R}_k(\widetilde{\pi}^t)$, where $\widetilde{\pi}^t$ is given by (55)–(56)
  end for
  Set $t\leftarrow t+1$
end while
Output: $\{\mathsf{v}_k^t\}_{k\in[m]}$

Remark B.3. *Let us assume that $\tau_k = \tau$ and $n_k = n$ for all $k\in[m]$ and that $m << n$. Then, we can conclude that the cost of one iteration of* BATCHGREENKHORN *is essentially determined by step 10 of Algorithm 2 which is performed in $\mathcal{O}(\tau n^{m-1})$ operations. Hence, one iteration of the* MULTISINKHORN *(i.e,* BATCHGREENKHORN *with a full batch $\tau = n$) has the same order of computational cost as n/τ iterations of* BATCHGREENKHORN *with a batch size τ. So, we can introduce the normalized iteration counter as $t = t_\tau n/\tau$, where t_τ is the iteration counter for the* `BatchGreenkhorn` *with a batch size τ.*

C. Convergence of Batch Greenkhorn algorithm

Here we provide proofs of the main results given in Sec. 4. We first set notation for the rest of the section. Given $k\in[m]$ and $L\subset[n_k]$, we denote by

$$J_k\colon \mathbb{R}^{n_k}\to\mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_k}\times\cdots\times\mathbb{R}^m, \quad \mathsf{v}_k\mapsto(0,\ldots,0,\mathsf{v}_k,0,\ldots,0), \tag{57}$$

the canonical injection of $\mathbb{R}^{n_k}$ into $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \times \cdots \times \mathbb{R}^{n_m}$ and by

$$J_L^{(k)} \colon \mathbb{R}^L \to \mathbb{R}^{n_k} \tag{58}$$

the canonical injection of $\mathbb{R}^L$ into $\mathbb{R}^{n_k}$.

We note that, referring to the operators R and R_k defined in (7) and (3), respectively, we have

$$\mathsf{R}^* \colon \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m} \to \mathbb{X}, \qquad \mathsf{R}^*(\mathsf{v}_1, \dots, \mathsf{v}_m) = \sum_{k=1}^m \mathsf{R}_k^*(\mathsf{v}_k) = \bigotimes_{k=1}^m \mathsf{v}_k, \tag{59}$$

where $(\mathsf{R}_k^*(\mathsf{v}_k))_{j_1,\dots,j_k,\dots,j_m} = v_{k,j_k}$. Indeed the second equality in (59) follows from the fact that for every $j \in \mathcal{J}$, $(\mathsf{R}^*(\mathsf{v}_1, \dots, \mathsf{v}_m))_j = \sum_{k=1}^m (\mathsf{R}_k^* \mathsf{v}_k)_j = \sum_{k=1}^m v_{k,j_k} = \big(\bigoplus_{k=1}^m \mathsf{v}_k\big)_j$. We note also that $\mathsf{R}_k = J_k^* \circ \mathsf{R}$, since J_k^* is the k-th canonical projection.

Then we provide a result concerning the properties of optimal potentials.

Lemma C.1. *Let $\pi^\star$ be the solution of RMOT given by (4). Then $\pi^\star = \mathcal{P}_\Pi(\xi \odot \otimes_{k=1}^m \mathsf{a}_k)$ and, for every $k \in [m]$, there exist $\mathsf{v}_k^\star = (v_{k,j}^\star)_{1 \le j \le n_k} \in \mathbb{R}^{n_k}$, such that*

$$\pi^\star = \exp\Big(-\frac{\mathsf{C}}{\eta} + \bigoplus_{k=1}^m \mathsf{v}_k^\star\Big) \odot \bigotimes_{k=1}^m \mathsf{a}_k, \tag{60}$$

and the $\mathsf{v}_k^\star$'s, can be chosen so that

$$\sum_{k \in [m]} \|\mathsf{v}_k^\star\|_\infty \le (4m-3)\frac{\|\mathsf{C}\|_\infty}{\eta}. \tag{61}$$

Moreover, if $m = 2$, then $\mathsf{v}_1^\star$ and $\mathsf{v}_2^\star$ can be chosen such that

$$\max_{k \in [m]} \|\mathsf{v}_k^\star\|_\infty \le \frac{3}{2}\frac{\|\mathsf{C}\|_\infty}{\eta}. \tag{62}$$

Proof. Since, by definition $\pi^\star = \mathcal{P}_\Pi(\xi)$, it easy to see, from the characterization of the projection given in (39), that

$$\pi^\star = \mathcal{P}_\Pi(\xi \odot \otimes_{k=1}^m \mathsf{a}_k) \;\Leftrightarrow\; \nabla \mathsf{H}(\xi \odot \otimes_{k=1}^m \mathsf{a}_k) - \nabla \mathsf{H}(\xi) \in \mathrm{Im}(\mathsf{R}^*).$$

Thus, since $\nabla \mathsf{H}(\xi \odot \otimes_{k=1}^m \mathsf{a}_k) - \nabla \mathsf{H}(\xi) = \log \otimes_{k=1}^m \mathsf{a}_k = \oplus_{k=1}^m \log \mathsf{a}_k = \mathsf{R}^*(\log \mathsf{a}_1, \dots, \log \mathsf{a}_m) \in \mathrm{Im}(\mathsf{R}^*)$, we have that $\mathcal{P}_\Pi(\xi \odot \otimes_{k=1}^m \mathsf{a}_k) = \mathcal{P}_\Pi(\xi) = \pi^\star$. Now, it follows from the KKT conditions (38) for the projection of $\xi \odot \otimes_{k=1}^m \mathsf{a}_k$ onto affine set Π, that

$$\pi^\star = \nabla \mathsf{H}^*\big(\nabla \mathsf{H}(\xi \odot \otimes_{k=1}^m \mathsf{a}_k) + \mathsf{R}^*(\mathsf{v}_1^\star, \dots, \mathsf{v}_m^\star)\big)$$

for some $(\mathsf{v}_1^\star, \dots, \mathsf{v}_m^\star) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$. Since $\nabla \mathsf{H}^* = \exp$ and $\nabla \mathsf{H} = \log$, (60) follows. Next, observe that for every $k \in [m]$, since $\mathsf{R}_k(\pi^\star) = \mathsf{a}_k$, using (60), we obtain that for every $j_k \in [n_k]$,

$$\exp(v_{k,j_k}^\star) \sum_{j_{-k} \in \mathcal{J}_{-k}} \exp(-\mathsf{C}_{(j_{-k},j_k)}/\eta + \sum_{h \ne k}^m v_{h,j_h}^\star) \prod_{h \ne k} a_{h,j_h} = 1. \tag{63}$$

Hence, the vectors $\mathsf{v}_1^\star, \dots, \mathsf{v}_m^\star$ solve a (discrete) Schrödinger system, and we can apply the results from (Carlier, 2021, Lemma 3.1) and (Di-Marino & Gerolin, 2020, Theorem 2.8) to obtain (61) and (62), respectively. □

Lemma C.2. *Let $\mathsf{A}, : \mathbb{X} \to \mathbb{X}$ and $\mathsf{A}_k \colon \mathbb{R}^{n_k} \to \mathbb{R}^{n_k}$ be diagonal and positive operators defined as $\mathsf{A}(\pi) = \pi \odot \bigotimes_{k=1}^m \mathsf{a}_k$ and $\mathsf{A}_k(\mathsf{v}_k) = \mathsf{v}_k \odot \mathsf{a}_k$, respectively. Then the following hold.*

(i) *For every $k \in [m]$ and every $\mathsf{v}_k \in \mathbb{R}^{n_k}$, $\|\mathsf{R}_k^* \mathsf{v}_k\|_{\mathsf{A}}^2 = \|\mathsf{v}_k\|_{\mathsf{A}_k}^2$*

(ii) *Let $(\mathsf{v}_1, \dots, \mathsf{v}_m) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$, if $\langle \mathsf{v}_k, \mathsf{a}_k \rangle = 0$ for every $k = 1, \dots, m-1$, then*

$$\|\mathsf{R}^*(\mathsf{v}_1, \dots, \mathsf{v}_m)\|_{\mathsf{A}}^2 = \sum_{k=1}^m \|\mathsf{R}_k^* \mathsf{v}_k\|_{\mathsf{A}}^2 = \sum_{k=1}^m \|\mathsf{v}_k\|_{\mathsf{A}_k}^2. \tag{64}$$

Proof. Let $k \in [m]$. Then, recalling that $(\mathsf{R}_k^*(\mathsf{v}_k))_j = v_{k,j_k}$, we have

$$\begin{aligned}\|\mathsf{R}_k^*\mathsf{v}_k\|_{\mathsf{A}}^2 &= \sum_{j\in\mathcal{J}} v_{k,j_k}^2 \prod_{\ell=1}^m a_{\ell,j_\ell} \\ &= \sum_{j_{-k}\in\mathcal{J}_{-k}} \sum_{j_k=1}^{n_k} v_{k,j_k}^2 a_{k,j_k} \prod_{\ell\neq k}^m a_{\ell,j_\ell} \\ &= \prod_{\ell\neq k} \Big(\sum_{\ell=1}^{n_\ell} a_{\ell,j_\ell}\Big) \sum_{j_k=1}^{n_k} v_{k,j_k}^2 a_{k,j_k}.\end{aligned}$$

Since $\sum_{\ell=1}^{n_\ell} a_{\ell,j_\ell} = 1$, we get $\|\mathsf{R}_k^*\mathsf{v}_k\|_{\mathsf{A}}^2 = \|\mathsf{v}_k\|_{\mathsf{A}_k}^2$ and (i) follows. Concerning (ii), equation (64) will follow if we prove that, for every $k, h \in [m]$ with $k \neq h$, we have that $\mathsf{R}_k^*(\mathsf{v}_k)$ and $\mathsf{R}_h^*(\mathsf{v}_h)$ are orthogonal in the metric $\langle\cdot,\cdot\rangle_{\mathsf{A}}$. Thus, let $k, h \in [m]$ and suppose (w.l.o.g.) that $k < h$. Then

$$\begin{aligned}\langle \mathsf{R}_k^*(\mathsf{v}_k), \mathsf{R}_h^*(\mathsf{v}_h)\rangle_{\mathsf{A}} &= \sum_{j\in\mathcal{J}} v_{k,j_k} v_{h,j_h} \prod_{\ell=1}^m a_{\ell,j_\ell} \\ &= \sum_{\substack{j_1,\ldots,j_{k-1},j_{k+1},\ldots,\\ j_{h-1},j_{h+1},\ldots,j_m}} \sum_{j_k=1}^{n_k}\sum_{j_h=1}^{n_h} v_{k,j_k} v_{h,j_h} a_{k,j_k} a_{h,j_h} \prod_{\ell\neq k,\ell\neq h}^m a_{\ell,j_\ell} \\ &= \Big(\sum_{j_k=1}^{n_k} v_{k,j_k}\Big)\Big(\sum_{j_h=1}^{n_h} v_{h,j_h}\Big) \prod_{\ell\neq k,\ell\neq h} \Big(\sum_{\ell=1}^{n_\ell} a_{\ell,j_\ell}\Big).\end{aligned}$$

Since $\sum_{\ell=1}^{n_\ell} a_{\ell,j_\ell} = 1$ and for $k < m$, $\sum_{j_k=1}^{n_k} v_{k,j_k} a_{k,j_k} = \langle \mathsf{v}_k, \mathsf{a}_k\rangle = 0$, we have $\langle \mathsf{R}_k^*(\mathsf{v}_k), \mathsf{R}_h^*(\mathsf{v}_h)\rangle_{\mathsf{A}} = 0$ and hence the statement (ii) follows.

□

Theorem 4.1 (Global linear convergence). *Algorithm 1 converges linearly. More precisely, if $(\mathsf{v}_k^t)_{k\in[m]}$ are generated by Algorithm 1, then the primal iterates given by (21) converge linearly in KL divergence to $\pi^\star$ given by (6), i.e. for all $t \in \mathbb{N}$*

$$\mathsf{KL}(\pi^\star, \pi^t) \leq \left(1 - \frac{e^{-(2\|\mathsf{C}\|_\infty/\eta + 3M_1)}}{b_\tau - 1}\right)^t \mathsf{KL}(\pi^\star, \pi^0), \tag{22}$$

where $b_\tau = \sum_{k\in[m]} \lceil n_k/\tau_k \rceil$, and $0 < M_1 < +\infty$ is a constant independent of the batch sizes that satisfies $\max\left\{\|\bigoplus_{k=1}^m \mathsf{v}_k^\star\|_\infty, \|\bigoplus_{k=1}^m \mathsf{v}_k^t\|_\infty\right\} \leq M_1$ for $t \in \mathbb{N}$.

Proof. We start by recalling the two formulas

$$\pi^t = \exp\Big(-\frac{\mathsf{C}}{\eta} + \mathsf{V}^t\Big) \odot \alpha \quad \text{and} \quad \pi^\star = \exp\Big(-\frac{\mathsf{C}}{\eta} + \mathsf{V}^\star\Big) \odot \alpha, \tag{65}$$

where $\alpha := \bigotimes_{k=1}^m \mathsf{a}_k$, $\mathsf{V}^t := \bigoplus_{k=1}^m \mathsf{v}_k^t$, and $\mathsf{V}^\star := \bigoplus_{k=1}^m \mathsf{v}_k^\star$. Moreover, since for every $(\lambda_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ such that $\sum_{k=1}^m \lambda_k = 0$, we have $\bigoplus_{k=1}^m (\mathsf{v}_k^t + \lambda_k) = \bigoplus_{k=1}^m \mathsf{v}_k^t$ and $\bigoplus_{k=1}^m (\mathsf{v}_k^\star + \lambda_k) = \bigoplus_{k=1}^m \mathsf{v}_k^\star$, we can choose the dual variables $(\mathsf{v}_k^t)_{k\in[n_k]}$ and $(\mathsf{v}_k^\star)_{k\in[n_k]}$ so that

$$(\forall\, k = 1, \ldots, m-1) \quad \langle \mathsf{v}_k^t, \mathsf{a}_k\rangle = 0 \quad \text{and} \quad \langle \mathsf{v}_k^\star, \mathsf{a}_k\rangle = 0. \tag{66}$$

First, observe that Pythagoras theorem yields that $\mathsf{KL}_\Pi(\pi^{t+1}) = \mathsf{KL}_\Pi(\pi^t) - \mathsf{KL}_{\Pi_{(k_t,L_t)}}(\pi^t) \leq \mathsf{KL}_\Pi(\pi^t)$, which implies that, for every $t \geq 0$, $D_{\mathsf{H}^*}(\log\pi^t, \log\pi^*) = \mathsf{KL}(\pi^\star, \pi^t) \leq \mathsf{KL}_\Pi(\pi^0) < +\infty$. Howevery, since H^* is a Legendre function, the sublevel sets of $D_{\mathsf{H}^*}(\cdot, \log\pi^*)$ are bounded, and hence the sequence $(\log\pi^t)_{t\in\mathbb{N}}$ is bounded in $\mathbb{X}$. Now, since the first of (65) yields that $\log\pi^t = -\mathsf{C}/\eta + \mathsf{V}^t + \log\alpha$, we have that also the sequence $(\mathsf{V}^t)_{t\in\mathbb{N}}$ is bounded in $\mathbb{X}$. Thus let $M_1 > 0$ be such that

$$\|\mathsf{V}^\star\|_\infty, \|\mathsf{V}^t\|_\infty \leq M_1 \quad (\forall\, t \in \mathbb{N}).$$

Then, recalling (65),

$$\frac{\pi^t}{\alpha} = \exp\Big(-\frac{\mathsf{C}}{\eta} + \mathsf{V}^t\Big) \geq \exp\Big(-\frac{\|\mathsf{C}\|_\infty}{\eta} - \|\mathsf{V}^t\|_\infty\Big) \geq \exp\Big(-\frac{\|\mathsf{C}\|_\infty}{\eta} - M_1\Big)$$

and

$$\frac{\pi^\star}{\alpha} = \exp\Big(-\frac{\mathsf{C}}{\eta} + \mathsf{V}^\star\Big) \geq \exp\Big(-\frac{\|\mathsf{C}\|_\infty}{\eta} - \|\mathsf{V}^\star\|_\infty\Big) \geq \exp\Big(-\frac{\|\mathsf{C}\|_\infty}{\eta} - M_1\Big)$$

and hence

$$\exp\Big(-\frac{\|\mathsf{C}\|_\infty}{\eta} - M_1\Big) \leq \min\Big\{\frac{\pi^t}{\alpha}, \frac{\pi^\star}{\alpha}\Big\} \quad (\forall\, t \in \mathbb{N}). \tag{67}$$

Let $t \in \mathbb{N}$, $k \in [m]$ and $L \subset [n_k]$. It follows from (19) that

$$(\forall\, j \in \mathcal{J}) \quad (\mathcal{P}_{\Pi_{(k,L)}}(\pi^t))_j = \pi^t_j \times \begin{cases} \dfrac{a_{k,j_k}}{\mathsf{R}_k(\pi)_{j_k}} & \text{if } j_k \in L, \\ 1 & \text{otherwise} \end{cases} \tag{68}$$

and hence

$$(\forall\, j \in \mathcal{J}) \quad \frac{(\mathcal{P}_{\Pi_{(k,L)}}(\pi^t))_j}{\alpha_j} \leq \frac{\pi^t_j}{\alpha_j} \max\Big\{1, \frac{a_{k,j_k}}{\mathsf{R}_k(\pi)_{j_k}}\Big\}. \tag{69}$$

Now, since $\mathsf{C} \geq 0$, we have

$$\frac{\pi^t}{\alpha} = \exp\Big(-\frac{\mathsf{C}}{\eta} + \mathsf{V}^t\Big) \leq \exp(\mathsf{V}^t) \leq \exp(\|\mathsf{V}\|_\infty) \leq \exp(M_1) \tag{70}$$

and

$$\begin{aligned}
\frac{\mathsf{R}_k(\pi)_{j_k}}{a_{k,j_k}} &= \frac{\sum_{j_{-k} \in \mathcal{J}_{-k}} \pi^t_{(j_{-k}, j_k)}}{a_{k,j_k}} \\
&= \frac{\sum_{j_{-k} \in \mathcal{J}_{-k}} \exp\big(-\mathsf{C}_{(j_{-k},j_k)}/\eta + \mathsf{V}^t_{(j_{-k},j_k)}\big) \prod_{h=1}^m a_{h,j_h}}{a_{k,j_k}} \\
&= \sum_{j_{-k} \in \mathcal{J}_{-k}} \exp\big(-\mathsf{C}_{(j_{-k},j_k)}/\eta + \mathsf{V}^t_{(j_{-k},j_k)}\big) \prod_{h \neq k}^m a_{h,j_h} \\
&\geq \exp\big(-\|\mathsf{C}\|_\infty/\eta - M_1\big) \sum_{j_{-k} \in \mathcal{J}_{-k}} \prod_{h \neq k}^m a_{h,j_h} \\
&= \exp\big(-\|\mathsf{C}\|_\infty/\eta - M_1\big) \prod_{h \neq k}^m \Big(\sum_{j_h=1}^{n_h} a_{h,j_h}\Big) \\
&= \exp\big(-\|\mathsf{C}\|_\infty/\eta - M_1\big),
\end{aligned} \tag{71}$$

since $\sum_{j_h=1}^{n_h} a_{h,j_h} = 1$. Therefore, by (69), (70), and (71),

$$\frac{\mathcal{P}_{\Pi_{(k,L)}}(\pi^t)}{\alpha} \leq \exp(M_1) \exp\big(\|\mathsf{C}\|_\infty/\eta + M_1\big) = \exp\big(\|\mathsf{C}\|_\infty/\eta + 2M_1\big)$$

and hence, recalling (70),

$$\max\Big\{\frac{\pi^t}{\alpha}, \frac{\mathcal{P}_{\Pi_{(k,L)}}(\pi^t)}{\alpha}\Big\} \leq \exp\big(\|\mathsf{C}\|_\infty/\eta + 2M_1\big). \tag{72}$$

We now prove that

$$\exp\big(-2\|\mathsf{C}\|_\infty - 3M_1\big) b_\tau^{-1} \mathsf{KL}_\Pi(\pi^t) = \mathsf{KL}_{\Pi_{(k_t,L_t)}}(\pi^t) = \max_{(k,L) \in \mathcal{I}(\tau)} \mathsf{KL}(\mathcal{P}_{\Pi_{(k,L)}}(\pi^t), \pi^t). \tag{73}$$

From this inequality it will follow, using the Pythagoras theorem $\mathsf{KL}_\Pi(\pi^{t+1}) + \mathsf{KL}_{\Pi_{(k_t,L_t)}}(\pi^t) = \mathsf{KL}_\Pi(\pi^t)$, that

$$\exp(-2\|\mathsf{C}\|_\infty - 3M_1)b_\tau^{-1}\mathsf{KL}_\Pi(\pi^t) = \mathsf{KL}_\Pi(\pi^t) - \mathsf{KL}_\Pi(\pi^{t+1}) \tag{74}$$

and hence

$$\mathsf{KL}_\Pi(\pi^{t+1}) \leq \big(1 - \exp(-2\|\mathsf{C}\|_\infty - 3M_1)b_\tau^{-1}\big)\mathsf{KL}_\Pi(\pi^t), \tag{75}$$

which gives the statement. Thus, it remains to prove (73). Let $t \in \mathbb{N}$ and, for the sake of brevity set

$$\pi := \pi^t, \quad \pi_k^h := \mathcal{P}_{\Pi_{(k,L_k^h)}}(\pi^t) \quad \mathsf{v}_k := \mathsf{v}_k^\star - \mathsf{v}_k^t, \quad \text{and} \quad \mathsf{v}_k^h := J_{L_k^h}^{(k)*}\mathsf{v}_k. \tag{76}$$

Let, for every $k \in [m]$, $(L_k^h)_{1\leq h\leq s_k}$ be a partition of $[n_k]$ made of non empty sets of cardinality exactly τ_k possibly except for the last one, such that $L_{k_{t-1}}^1 = L_{t-1}$, where $s_k = \lceil n_k/\tau_k \rceil$. Then, it follows from (65) that

$$\frac{\pi^\star}{\pi} = \frac{\exp(-\mathsf{C}/\eta)\exp(\mathsf{V}^\star)}{\exp(-\mathsf{C}/\eta)\exp(\mathsf{V}^t)} = \exp(\mathsf{V}^\star - \mathsf{V}^t). \tag{77}$$

Hence, recalling that $\mathsf{R}^*(\mathsf{v}_1,\dots,\mathsf{v}_m) = \sum_{k=1}^m \mathsf{R}_k^*\mathsf{v}_k = \bigoplus_{k=1}^m \mathsf{v}_k = \bigoplus_{k=1}^m(\mathsf{v}_k^\star - \mathsf{v}_k^t)$, we have

$$\begin{aligned}
\mathsf{KL}_\Pi(\pi) + \mathsf{KL}(\pi,\pi^\star) &= \mathsf{KL}(\pi^\star,\pi) + \mathsf{KL}(\pi,\pi^\star)\\
&= \langle \pi^\star - \pi, \log(\pi^\star/\pi)\rangle\\
&= \langle \pi^\star - \pi, \mathsf{V}^\star - \mathsf{V}^t\rangle\\
&= \langle \pi^\star - \pi, \mathsf{R}^*(\mathsf{v}_1,\dots,\mathsf{v}_m)\rangle\\
&= \sum_{k=1}^m \langle \pi^\star - \pi, \mathsf{R}_k^*(\mathsf{v}_k)\rangle.
\end{aligned} \tag{78}$$

Moreover, recalling that $\mathsf{R}_{(k,L_k^h)} = J_{L_k^h}^{(k)*} \circ \mathsf{R}_k^*$ and $\mathsf{v}_k^h = J_{L_k^h}^{(k)*}(\mathsf{v}_k)$, we have

$$\mathsf{R}_k^*(\mathsf{v}_k) = \mathsf{R}_k^*\bigg(\sum_{h=1}^{n_h} J_{L_k^h}^{(k)} \circ J_{L_k^h}^{(k)*}(\mathsf{v}_k)\bigg) = \sum_{h=1}^{n_h} \mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h) \tag{79}$$

and hence

$$\mathsf{KL}_\Pi(\pi) + \mathsf{KL}(\pi,\pi^\star) = \sum_{k=1}^m\sum_{h=1}^{s_k} \langle \pi^\star - \pi, \mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle. \tag{80}$$

Now, recalling the general definition of $\Pi_{(k,L)}$ in (11), since π_k^h and $\pi^\star$ both belong to $\Pi_{(k,L_k^h)}$, we have that $\pi^\star - \pi_k^h \in \mathrm{Ker}(\mathsf{R}_{(k,L_k^h)}) = \mathrm{Im}(\mathsf{R}_{(k,L_k^h)}^*)^\perp$ and hence

$$\langle \pi^\star - \pi, \mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle = \langle \pi^\star - \pi_k^h, \mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle + \langle \pi_k^h - \pi, \mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle = \langle \pi_k^h - \pi, \mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle$$

and hence

$$\begin{aligned}
\mathsf{KL}_\Pi(\pi) + \mathsf{KL}(\pi,\pi^\star) &= \sum_{k=1}^m\sum_{h=1}^{s_k} \langle \pi_k^h - \pi, \mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle\\
&= \sum_{k=1}^m\sum_{h=1}^{s_k} \langle \mathsf{A}^{-1}(\pi_k^h - \pi), \mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle_{\mathsf{A}},
\end{aligned} \tag{81}$$

where A is the positive diagonal operator defined in Lemma C.2. Now, it follows from (67), Lemma A.4, and Lemma C.2(i) that

$$\begin{aligned}
\mathsf{KL}(\pi^t,\pi^\star) &\geq (1/2)\exp(-\|\mathsf{C}\|_\infty/\eta - M_1)\|\log\pi^\star - \log\pi^t\|_{\mathsf{A}}^2\\
&= (1/2)\exp(-\|\mathsf{C}\|_\infty/\eta - M_1)\|\mathsf{V}^\star - \mathsf{V}^t\|_{\mathsf{A}}^2\\
&= (1/2)\exp(-\|\mathsf{C}\|_\infty/\eta - M_1)\|\mathsf{R}^*(\mathsf{v}_1,\dots,\mathsf{v}_m)\|_{\mathsf{A}}^2\\
&= (1/2)\exp(-\|\mathsf{C}\|_\infty/\eta - M_1)\sum_{k=1}^m \|\mathsf{v}_k\|_{\mathsf{A}_k}^2.
\end{aligned}$$

Moreover, recalling the definition of v_k^h in (76), since $\mathsf{v}_k = \sum_{h=1}^{s_k} J_{(k,L_k^h)}\mathsf{v}_k^h$ and $(J_{(k,L_k^h)}\mathsf{v}_k^h)_{h\in[s_k]}$ is a finite orthogonal sequence in $\mathbb{R}^{n_k}$ w.r.t. the metric $\langle\cdot,\cdot\rangle_{\mathsf{A}_k}$, we have

$$\|\mathsf{v}_k\|_{\mathsf{A}_k}^2 = \sum_{h=1}^{s_k}\|J_{(k,L_k^h)}\mathsf{v}_k^h\|_{\mathsf{A}_k}^2 = \sum_{h=1}^{s_k}\|\mathsf{R}_k^* J_{(k,L_k^h)}\mathsf{v}_k^h\|_{\mathsf{A}}^2 = \sum_{h=1}^{s_k}\|\mathsf{R}_{(k,L_k^h)}^*\mathsf{v}_k^h\|_{\mathsf{A}}^2, \tag{82}$$

where we used Lemma C.2(ii) applied to $J_{(k,L_k^h)}\mathsf{v}_k^h$ and the fact that, by definition, $\mathsf{R}_{(j,L_k^h)} = J_{L_k^h}^{(k)*}\mathsf{R}_k$. Overall we get that

$$\mathsf{KL}(\pi^t,\pi^\star) \geq (1/2)\exp(-\|\mathsf{C}\|_\infty/\eta - M_1)\sum_{k=1}^m\sum_{h=1}^{s_k}\|\mathsf{R}_{(k,L_k^h)}^*\mathsf{v}_k^h\|_{\mathsf{A}}^2$$

and hence (81) yields

$$\begin{aligned}
\mathsf{KL}_\Pi(\pi) &\leq \sum_{k=1}^m\sum_{h=1}^{s_k}\langle\mathsf{A}^{-1}(\pi_k^h-\pi),\mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle_{\mathsf{A}} - \mathsf{KL}(\pi,\pi^\star)\\
&\leq \sum_{k=1}^m\sum_{h=1}^{s_k}\langle\mathsf{A}^{-1}(\pi_k^h-\pi),\mathsf{R}_{(k,L_k^h)}^*(\mathsf{v}_k^h)\rangle_{\mathsf{A}} - \frac{1}{2}\exp(-\|\mathsf{C}\|_\infty/\eta - M_1)\|\mathsf{R}_{(k,L_k^h)}^*\mathsf{v}_k^h\|_{\mathsf{A}}^2\\
&\leq \frac{\exp(\|\mathsf{C}\|_\infty/\eta+M_1)}{2}\sum_{k=1}^m\sum_{h=1}^{s_k}\|\mathsf{A}^{-1}(\pi_k^h-\pi)\|_{\mathsf{A}}^2\\
&= \frac{\exp(\|\mathsf{C}\|_\infty/\eta+M_1)}{2}\sum_{k=1}^m\sum_{h=1}^{s_k}\|(\pi_k^h-\pi)\|_{\mathsf{A}^{-1}}^2,
\end{aligned}$$

where in the last inequality we used the Young-Fenchel inequality $\langle a,b\rangle_{\mathsf{A}} \leq \frac{\mu}{2}\|a\|_{\mathsf{A}}^2 + \frac{1}{2\mu}\|b\|_{\mathsf{A}}^2$. Now, recalling that we set $\pi = \pi^t$ and $\pi_k^h = \mathcal{P}_{\Pi_{(k,L_k^h)}}(\pi^t)$, it follows from (72) and Lemma A.4 that $\frac{1}{2}\|\pi_k^h-\pi\|_{\Theta^{-1}}^2 \leq \exp\big(\|\mathsf{C}\|_\infty/\eta + 2M_1\big)\mathsf{KL}(\pi_k^h,\pi)$, and consequently

$$\begin{aligned}
\mathsf{KL}_\Pi(\pi^t) &\leq \exp\big(2\|\mathsf{C}\|_\infty/\eta+3M_1\big)\sum_{k\in[m]}\sum_{h\in[s_k]}\mathsf{KL}(\pi_k^h,\pi)\\
&\leq \exp\big(2\|\mathsf{C}\|_\infty/\eta+3M_1\big)\Big(\sum_{k\in[m]}s_k - 1\Big)\max_{k\in[m]}\max_{h\in[s_k]}\mathsf{KL}(\pi_k^h,\pi)\\
&= \exp\big(2\|\mathsf{C}\|_\infty/\eta+3M_1\big)\Big(\sum_{k\in[m]}\lceil n_k/\tau_k\rceil - 1\Big)\max_{k\in[m]}\max_{h\in[s_k]}\mathsf{KL}(\mathcal{P}_{\Pi_{(k,L_k^h)}}(\pi^t),\pi^t)\\
&\leq (b_\tau-1)\exp\big(2\|\mathsf{C}\|_\infty/\eta+3M_1\big)\max_{(k,L)\in\mathcal{I}(\tau)}\mathsf{KL}(\mathcal{P}_{\Pi_{(k,L)}}(\pi^t),\pi^t),
\end{aligned}$$

where in the second inequality we used that, for $k = k_{t-1}$ and $h = 1$, $\pi_k^h = \mathcal{P}_{\Pi_{(k_{t-1},L_{t-1})}}(\pi^t) = \pi^t$ (since by definition $\pi^t \in \Pi_{(k_{t-1},L_{t-1})}$), so that $\mathsf{KL}(\pi_k^h,\pi) = 0$. This proves (73) and the proof is complete. □

We now provide a result concerning the convergence of numerical sequences which is critical to analyze the iteration complexity of the algorithm. This result has been first showed implicitly in (Dvurechensky et al., 2018). We provide here a more explicit version together with a complete proof for the reader's convenience.

Lemma C.3. *Let $M, C > 0$ and let $(\delta_t)_{t\in\mathbb{N}}$ and $(d_t^\infty)_{t\in\mathbb{N}}$ be two sequences of positive numbers such that, for every $t \in \mathbb{N}$,*

(i) $\delta_t - \delta_{t+1} \geq \left(\dfrac{d_t^\infty}{C}\right)^2$,

(ii) $\delta_t \leq M d_t^\infty$.

Let $\varepsilon > 0$ and set $\bar{t} = \min\{t \in \mathbb{N} \,|\, d_t^\infty \leq \varepsilon\}$. Then $\bar{t} \leq 1 + 2MC^2/\varepsilon$.

Proof. Items (i) and (ii) imply that

$$\delta_t - \delta_{t+1} \geq \left(\frac{\delta_t}{MC}\right)^2.$$

Therefore, since $\delta_t \geq \delta_{t+1}$, we have

$$\delta_t - \delta_{t+1} \geq \frac{\delta_t^2}{M^2C^2} \geq \frac{\delta_t \delta_{t+1}}{M^2C^2}$$

and hence, dividing by $\delta_t \delta_{t+1}$,

$$\frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \geq \frac{1}{M^2C^2}.$$

Thus,

$$\frac{1}{\delta_t} - \frac{1}{\delta_0} = \sum_{i=0}^{t-1} \left(\frac{1}{\delta_i} - \frac{1}{\delta_{i+1}}\right) \geq \frac{t}{M^2C^2}$$

and hence we get the following rate of convergence for the sequence $(\delta_t)_{t\in\mathbb{N}}$

$$\delta_t \leq \left(\frac{1}{\delta_0} + \frac{t}{M^2C^2}\right)^{-1}. \tag{83}$$

Now, we let $\delta \in]0, \delta_0]$. We wish to determine the number of iterations such that $\delta_t \leq \delta$. It follows from (83) that

$$\left(\frac{1}{\delta_0} + \frac{t}{M^2C^2}\right)^{-1} \leq \delta \;\Leftrightarrow\; \frac{1}{\delta_0} + \frac{t}{M^2C^2} \geq \frac{1}{\delta} \;\Leftrightarrow\; t \geq M^2C^2\left(\frac{1}{\delta} - \frac{1}{\delta_0}\right). \tag{84}$$

This means that if we take $t \geq M^2C^2(1/\delta - 1/\delta_0)$, we have $\delta_t \leq \delta$ as desired. So we set $t = \lfloor M^2C^2(1/\delta - 1/\delta_0)\rfloor + 1$. Then, we have $\delta_t \leq \delta$. Now we have to cases. Suppose that $t < \bar{t}$ and let $s \in \mathbb{N}$ be such that $t + s = \bar{t} - 1$. Then, for every $i = 0, \ldots, s$, since $t + i < \bar{t}$, we have $d_{t+i}^{\infty} > \varepsilon$ and hence, using (i),

$$\delta \geq \delta_t - \delta_{\bar{t}} = \sum_{i=0}^{s} \left(\frac{1}{\delta_{t+i}} - \frac{1}{\delta_{t+i+1}}\right) \geq \sum_{i=0}^{s} \frac{(d_{t+i}^{\infty})^2}{M^2C^2} \geq (s+1)\frac{\varepsilon^2}{M^2C^2}, \tag{85}$$

which implies that $s + 1 \leq C^2\delta/\varepsilon^2$. Overall we have

$$\bar{t} = t + s + 1 \leq \left\lfloor M^2C^2\left(\frac{1}{\delta} - \frac{1}{\delta_0}\right)\right\rfloor + 1 + C^2\frac{\delta}{\varepsilon^2} \leq 1 + \frac{M^2C^2}{\delta} - \frac{M^2C^2}{\delta_0} + \frac{C^2\delta}{\varepsilon^2}. \tag{86}$$

Note that this inequality is true for any $\delta \in]0, \delta_0]$. Now, suppose that $M\varepsilon \leq \delta_0$. Then we have

$$\bar{t} \leq 1 + \min_{\delta\in]0,\delta_0]} \left(\frac{M^2C^2}{\delta} + \frac{C^2\delta}{\varepsilon^2}\right) = 1 + 2\frac{MC^2}{\varepsilon}, \tag{87}$$

where the minimum is attained at $\delta = M\varepsilon \in]0, \delta_0]$. On the other hand, if $\delta_0 < M\varepsilon$, then the minimum on the right hand side of (86) is attained at $\delta = \delta_0$ and hence

$$\bar{t} \leq 1 + \frac{C^2\delta_0}{\varepsilon^2} \leq 1 + \frac{C^2M\varepsilon}{\varepsilon^2} = 1 + \frac{MC^2}{\varepsilon}. \tag{88}$$

In any case, the statement follows. □

Remark C.4. *The statement of Lemma C.3 is equivalent to the fact that the sequence* $(\min_{0\leq s<t} d_s)_{t\in\mathbb{N}}$ *converge to zero with rate* $\mathcal{O}(1/t)$*, i.e., that for every integer* $t > 1$*,*

$$\min_{0\leq s<t} d_s \leq \frac{2MC^2}{t-1}.$$

Next, we prove the main result on the iteration complexity.

Theorem 4.2 (**Iteration complexity**). *Let $\varepsilon > 0$ and suppose that $\eta > \varepsilon$. For Algorithm 1, the number of iterations required to reach the stopping criterion $d_t^\infty := \max_{k\in[m]} \|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1 \leq \varepsilon$ satisfies*

$$t \leq 2 + \max_{k\in[m]} \left\lceil \frac{n_k}{\tau_k} \right\rceil \frac{5M_2}{\varepsilon}(2 + M_2\eta), \tag{23}$$

where $0 < M_2 < +\infty$ is a constant independent of the batch sizes such that $\sum_{k\in[m]} \|\mathsf{v}_k^\star - \mathsf{v}_k^t\| \leq M_2$, for all $t \in \mathbb{N}$.

Proof. For the sake of brevity let $\bar{b} = \max_{k\in[n]} \lceil n_k/\tau_k \rceil$ and set, for every $t \in \mathbb{N}$, $\delta_t := \mathsf{KL}_\Pi(\pi^t)$. Let $t \in \mathbb{N}$ be arbitrary. Recalling (78), we have that

$$\delta_t = \mathsf{KL}_\Pi(\pi^t) \leq \sum_{k\in[m]} \langle \pi^\star - \pi^t, \mathsf{R}_k^*(\mathsf{v}_k^\star - \mathsf{v}_k^t)\rangle = \sum_{k\in[m]} \langle \mathsf{a}_k - \mathsf{R}_k(\pi^t), \mathsf{v}_k^\star - \mathsf{v}_k^t\rangle,$$

which, using Hoölder inequality, yields

$$\delta_t \leq \sum_{k\in[m]} \|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1 \, \|\mathsf{v}_k^\star - \mathsf{v}_k^t\|_\infty \leq M_2 d_t^\infty. \tag{89}$$

Now, we prove

$$\delta_t - \delta_{t+1} \geq \min\left\{ \frac{(d_t^\infty)^2}{5\bar{b}}, \frac{\delta_t^2}{4M_2^2\bar{b}} \right\} \geq \frac{\delta_t^2}{5M_2^2\bar{b}}. \tag{90}$$

Let for every $k \in [m]$, $(L_k^h)_{1\leq h\leq [s_k]}$, $s_k := \lceil n_k/\tau_k \rceil$, be a partition of $[n_k]$ made of nonempty sets of cardinality exactly τ_k, except maybe for the last one, such that $L_{k_t}^1 = L_t$ (not that necessarily the cardinality of L_t is τ_{k_t}). Then, according to the greedy choice of (k_t, L_t) we have that

$$\bar{b}\,\mathsf{KL}_{\Pi_{(k_t,L_t)}}(\pi^t) \geq \max_{k\in[m]} s_k \max_{h\in[s_k]} \mathsf{KL}_{\Pi_{(k,L_k^h)}}(\pi^t) \geq \max_{k\in[m]} \sum_{h\in[s_k]} \mathsf{KL}_{\Pi_{(k,L_k^h)}}(\pi^t).$$

Thus, equation (20) of Proposition 3.1 yields

$$\bar{b}\,\mathsf{KL}_{\Pi_{(k_t,L_t)}}(\pi^t) \geq \max_{k\in[m]} \sum_{h\in[s_k]} \mathsf{KL}(\mathsf{a}_{k|L_k^h}, \mathsf{R}_k(\pi^t)_{|L_k^h}) = \max_{k\in[m]} \mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^t)). \tag{91}$$

Now Pinsker's inequality guaranties that, for every $k \in [m]$

$$\mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^t)) \geq \frac{\|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1^2}{\frac{2}{3}\|\mathsf{a}_k\|_1 + \frac{4}{3}\|\mathsf{R}_k(\pi^t)\|_1} = \frac{\|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1^2}{\frac{2}{3} + \frac{4}{3}\|\pi^t\|_1} \geq \frac{\|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1^2}{2 + \frac{4}{3}\|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1}, \tag{92}$$

where in the second inequality we used that $\|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1 \geq \|\mathsf{R}_k(\pi^t)\|_1 - \|\mathsf{a}_k\|_1 = \|\pi^t\|_1 - 1$. Thus, solving the quadratic inequality in $\|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1 \geq 0$ we can conclude that

$$\|\mathsf{a}_k - \mathsf{R}_k(\pi^t)\|_1 \leq \tfrac{2}{3}\mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^t)) + \sqrt{(\tfrac{2}{3}\mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^t)))^2 + 2\mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^t))}.$$

Therefore, if $\max_{k\in[m]} \mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^t)) \leq 1$, then $2 + 4d_t^\infty/3 \leq 5$, and consequently, $\max_{k\in[m]} \mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^t)) \geq (d_t^\infty)^2/5$, which, using Pythagoras theorem and (91), yields

$$\delta_t - \delta_{t+1} = \mathsf{KL}_\Pi(\pi^t) - \mathsf{KL}_\Pi(\pi^{t+1}) = \mathsf{KL}_{\Pi_{(k_t,L_t)}}(\pi^t) \geq \frac{(d_t^\infty)^2}{5\bar{b}}. \tag{93}$$

On the other hand, if $\max_{k\in[m]} \mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^t)) > 1$, it follows again from Pythagoras theorem and (91), that $\delta_t - \delta_{t+1} \geq 1/\bar{b}$. Moreover, since $\delta_t \leq \delta_0 \leq M_2 d_0^\infty \leq M_2(1 + \|\pi^0\|_1) \leq 2M_2$, we have that $1 \geq \delta_t^2/(4M_2^2)$, and (90) follows.

Now, similarly to what was done in the proof of Lemma C.3 we can derive from (90) that

$$\delta_t \leq \left(\frac{1}{\delta_0} + \frac{t}{5M_2^2\bar{b}}\right)^{-1}. \tag{94}$$

Thus, if we take $r = \lfloor 5M_2^2\bar{b}\rfloor + 1$ we have $\delta_r \leq 1$. Then, by (20) with $L = [n_k]$ and Pythagoras theorem we have that, for every $t \in \mathbb{N}$,

$$\mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^{r+t})) = \mathsf{KL}(\mathcal{P}_{\Pi_{(k,[n_k])}}(\pi^{r+t}), \pi^{r+t}) \leq \mathsf{KL}(\pi^\star, \pi^{r+t}) = \delta_{r+t} \leq \delta_r \leq 1.$$

Thus, $\max_{k\in[m]} \mathsf{KL}(\mathsf{a}_k, \mathsf{R}_k(\pi^{r+t})) \leq 1$ and for what we already saw,

$$\delta_{r+t} - \delta_{r+t+1} \geq \frac{d_{r+t}^2}{5\bar{b}}. \tag{95}$$

In the end the sequence $(\delta_{r+t})_{t\in\mathbb{N}}$ satisfies the two assumptions of Lemma C.3 with $C = \sqrt{5\bar{b}}$ and $M = M_2$. Thus, we can conclude that the smallest t so that $d_{r+t} \leq \varepsilon$ satisfies $t \leq 1 + 10M_2\bar{b}/\varepsilon$. Hence

$$r + t \leq 2 + \frac{10M_2\bar{b}}{\varepsilon} + 5M_2^2\bar{b} \leq 2 + \frac{10M_2\bar{b}}{\varepsilon} + \frac{5M_2^2\bar{b}\eta}{\varepsilon} = 2 + \frac{5M_2\bar{b}}{\varepsilon}(2 + M_2\eta). \qquad \square$$

The next two results are based on novel bounds on potentials that imply explicit dependence of constant $M > 0$ in the global rate (22) on the given data: $\mathsf{a}_1, \ldots, \mathsf{a}_m$, C and η.

Theorem 4.4. *Suppose that $m = 2$. Then the algorithm* $\texttt{BatchGreenkhorn}(\mathsf{a}_1, \mathsf{a}_2, \mathsf{C}, \eta, \tau)$ *converges linearly with the global rate*

$$\mathsf{KL}(\pi^\star, \pi^t) \leq \left(1 - \frac{e^{-20\|\mathsf{C}\|_\infty/\eta}}{b_\tau - 1}\right)^t \mathsf{KL}(\pi^\star, \pi^0). \tag{26}$$

Moreover, when $\eta > \varepsilon$, the number of iterations required to reach the stopping criterion $d_t^\infty \leq \varepsilon$ satisfies

$$t \leq 2 + \max_{k\in[m]} \left\lceil \frac{n_k}{\tau_k} \right\rceil \frac{15\|\mathsf{C}\|_\infty(2 + 3\|\mathsf{C}\|_\infty)}{\eta\varepsilon}. \tag{27}$$

Proof. Let v_k^t, $k \in [m]$, $t \geq 0$ be given by Algorithm 1. Then from Proposition B.1 we have that for every $t \geq 0$

$$\pi^{t+1} = \exp\Big(-\frac{\mathsf{C}}{\eta} + \bigoplus_{k=1}^m \mathsf{v}_k^{t+1}\Big) \odot \bigotimes_{k=1}^m \mathsf{a}_k, \tag{96}$$

with $\mathsf{v}_k^0 = 0$ and for $t \geq 0$, $k \in [m]$ and $j_k \in [n_k]$,

$$v_{k,j_k}^{t+1} = \begin{cases} v_{k,j_k}^t + \log(a_{k,j_k}) - \log(\mathsf{R}_k(\pi^t)_{j_k}) & k = k_t,\ j_k \in L_t, \\ v_{k,j_k}^t & \text{otherwise.} \end{cases}$$

So, to bound $\log \pi^t$, we will bound v_k^t, $k \in [m]$. Since $\mathsf{R}_{k_t}(\pi^{t+1})_{j_{k_t}} = a_{k_t,j_{k_t}}$ for all $j_{k_t} \in L_t$, (96) implies that

$$1 = \exp(v_{k_t,j_{k_t}}^{t+1}) \sum_{j_{-k_t}\in\mathcal{J}_{-k_t}} \exp\Big(-\mathsf{C}_{(j_{-k_t},j_{k_t})}/\eta + \sum_{k\neq k_t} v_{k,j_k}^t\Big) \prod_{k\neq k_t} a_{k,j_k},$$

and, hence,

$$\exp(-v_{k_t,j_{k_t}}^{t+1}) = \sum_{j_{-k_t}\in\mathcal{J}_{-k_t}} \exp\Big(-\mathsf{C}_{(j_{-k_t},j_{k_t})}/\eta + \sum_{k\neq k_t} v_{k,j_k}^t\Big) \prod_{k\neq k_t} a_{k,j_k}. \tag{97}$$

So, using (63) we obtain that for every $j_{k_t} \in L_t$

$$\exp(v_{k_t,j_{k_t}}^{t+1} - v_{k_t,j_{k_t}}^\star) = \frac{\displaystyle\sum_{j_{-k_t}\in\mathcal{J}_{-k_t}} \exp\Big(-\mathsf{C}_{(j_{-k_t},j_{k_t})}/\eta + \sum_{k\neq k_t} v_{k,j_k}^\star\Big)}{\displaystyle\sum_{j_{-k_t}\in\mathcal{J}_{-k_t}} \exp\Big(-\mathsf{C}_{(j_{-k_t},j_{k_t})}/\eta + \sum_{k\neq k_t} v_{k,j_k}^t\Big)},$$

while

$$\exp(v^\star_{k_t,j_{k_t}} - v^{t+1}_{k_t,j_{k_t}}) = \frac{\sum_{j_{-k_t}\in\mathcal{J}_{-k_t}} \exp\Big(-\mathsf{C}_{(j_{-k_t},j_{k_t})}/\eta + \sum_{k\neq k_t} v^t_{k,j_k}\Big)}{\sum_{j_{-k_t}\in\mathcal{J}_{-k_t}} \exp\Big(-\mathsf{C}_{(j_{-k_t},j_{k_t})}/\eta + \sum_{k\neq k_t} v^\star_{k,j_k}\Big)}.$$

However, since in general, $\alpha_i, \beta_i > 0$ implies that $(\sum_i \alpha_i)/(\sum_i \beta_i) \leq \max_i \alpha_i/\beta_i$, the last two equalities give

$$\exp(v^{t+1}_{k_t,j_{k_t}} - v^\star_{k_t,j_{k_t}}) \leq \max_{j_{-k_t}\in\mathcal{J}_{-k_t}} \exp\Big(\sum_{k\neq k_t}(v^\star_{k,j_k} - v^t_{k,j_k})\Big)$$

and

$$\exp(v^\star_{k_t,j_{k_t}} - v^{t+1}_{k_t,j_{k_t}}) \leq \max_{j_{-k_t}\in\mathcal{J}_{-k_t}} \exp\Big(\sum_{k\neq k_t}(v^t_{k,j_k} - v^\star_{k,j_k})\Big).$$

Hence, taking the logarithm we obtain that for every $j_{k_t} \in L_t$,

$$|v^{t+1}_{k_t,j_{k_t}} - v^\star_{k_t,j_{k_t}}| \leq \max_{j_{-k_t}\in\mathcal{J}_{-k_t}} \Big|\sum_{k\neq k_t}(v^\star_{k,j_k} - v^t_{k,j_k})\Big| \leq \max_{j_{-k_t}\in\mathcal{J}_{-k_t}} \sum_{k\neq k_t} |v^\star_{k,j_k} - v^t_{k,j_k}| = \sum_{k\neq k_t} \|\mathsf{v}^t_k - \mathsf{v}^\star_k\|_\infty.$$

Therefore, since $v^{t+1}_{k,j_k} = v^t_{k,j_k}$ if $k \neq k_t$ or $j_{k_t} \notin L_t$,

$$\max\Big\{\|\mathsf{v}^{t+1}_{k_t} - \mathsf{v}^\star_{k_t}\|_\infty, \sum_{k\neq k_t}\|\mathsf{v}^{t+1}_k - \mathsf{v}^\star_k\|_\infty\Big\} \leq \max\Big\{\|\mathsf{v}^t_{k_t} - \mathsf{v}^\star_{k_t}\|_\infty, \sum_{k\neq k_t}\|\mathsf{v}^t_k - \mathsf{v}^\star_k\|_\infty\Big\},$$

and, since $m = 2$,

$$\max_{k\in[m]}\|\mathsf{v}^{t+1}_k - \mathsf{v}^\star_k\|_\infty \leq \max_{k\in[m]}\|\mathsf{v}^t_k - \mathsf{v}^\star_k\|_\infty,$$

which implies, recalling that $\mathsf{v}^0 = 0$, that, for all $t \geq 0$, $\max_{k\in[m]}\|\mathsf{v}^t_k - \mathsf{v}^\star_k\|_\infty \leq \max_{k\in[m]}\|\mathsf{v}^\star_k\|_\infty$. Now, in view of (62) in Lemma C.1, we have

$$\max_{k\in[m]}\|\mathsf{v}^\star_k\|_\infty \leq \frac{3}{2}\frac{\|\mathsf{C}\|_\infty}{\eta} \tag{98}$$

and hence, since $\|\mathsf{v}^t_k\|_\infty \leq \|\mathsf{v}^t_k - \mathsf{v}^\star_k\|_\infty + \|\mathsf{v}^\star_k\|_\infty$, $\max_{k\in[m]}\|\mathsf{v}^t_k\|_\infty \leq 2\max_{k\in[m]}\|\mathsf{v}^\star_k\|_\infty \leq 3\|\mathsf{C}\|_\infty/\eta$. In the end, since for every $(\mathsf{v}_k)_{k\in[m]} \in \mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_m}$

$$\Big\|\bigoplus_{k=1}^m \mathsf{v}_k\Big\|_\infty = \max_{j\in\mathcal{J}}\Big|\sum_{k=1}^m v_{k,j_k}\Big| \leq \max_{j\in\mathcal{J}}\sum_{k=1}^m |v_{k,j_k}| = \sum_{k=1}^m \|\mathsf{v}_k\|_\infty \leq m\max_{k\in[m]}\|\mathsf{v}_k\|_\infty, \tag{99}$$

we can satisfy the boundedness assumptions on the dual variables of Theorem 4.1 with $M_1 = 6\|\mathsf{C}\|_\infty/\eta$ and (26) follows from (22). Concerning the iteration complexity, again by (98), since $\max_{k\in[m]}\|\mathsf{v}^t_k - \mathsf{v}^\star_k\|_\infty \leq \max_{k\in[m]}\|\mathsf{v}^\star_k\|_\infty$, we have $\sum_{k\in[m]}\|\mathsf{v}^t_k - \mathsf{v}^\star_k\|_\infty \leq 3\|\mathsf{C}\|_\infty/\eta$. So, using $\eta \geq \varepsilon$, (29) follows directly from (23) with $M_2 = 3\|\mathsf{C}\|_\infty/\eta$. □

Theorem 4.5. *Suppose that for all $k \in [m]$ $\tau_k = n_k$. Then* `BatchGreenkhorn`$(\mathsf{a}_1,\ldots,\mathsf{a}_m,\mathsf{C},\eta,\tau)$*, converges linearly with the global rate*

$$\mathsf{KL}(\pi^\star,\pi^t) \leq \left(1 - \frac{e^{-(12m-7)\|\mathsf{C}\|_\infty/\eta}}{m-1}\right)^t \mathsf{KL}(\pi^\star,\pi^0). \tag{28}$$

Moreover, the number of iterations required to reach the stopping criterion $d^\infty_t \leq \varepsilon$ satisfies

$$t \leq 1 + \frac{8(4m-3)\|\mathsf{C}\|_\infty}{\eta\varepsilon}. \tag{29}$$

Proof. Using the same notation as in the previous proof, we first show that

$$(\forall\, t\in\mathbb{N})(\forall\, k\in[m])(\forall\, j_k,\ell_k\in[n_k]) \quad v^t_{k,j_k} - v^t_{k,\ell_k} \leq 2\|\mathsf{C}\|_\infty/\eta. \tag{100}$$

Indeed, since for $t=0$, $v^t_{k,j_k} - v^t_{k,\ell_k} = 0$, we proceed by induction assuming that (100) holds for t and proving it for $t+1$. Noting that for every $k \in [m]$, every $j_{-k} \in \mathcal{J}_{-k}$ and every $j_k, \ell_k \in [n_k]$

$$\mathsf{C}_{(j_{-k},j_k)} - \mathsf{C}_{(j_{-k},\ell_k)} \leq 2\|\mathsf{C}\|_\infty,$$

and that $L_t = [n_{k_t}]$, from (97) we have that $v^{t+1}_{k_t,j_{k_t}} - v^{t+1}_{k_t,\ell_{k_t}} \leq 2\|\mathsf{C}\|_\infty/\eta$ holds for every $j_{k_t}, \ell_{k_t} \in [n_{k_t}]$. On the other hand for every $k \neq k_t$ $v^{t+1}_k = v^t_k$, which using the inductive hypothesis (100) yields $v^{t+1}_{k,j_k} - v^{t+1}_{k,\ell_k} \leq 2\|\mathsf{C}\|_\infty/\eta$. In any case (100) holds for $t+1$.

Next, let $t \in \mathbb{N}$ and define the normalizing constants $\lambda^{t+1}_1, \dots \lambda^{t+1}_m \in \mathbb{R}$ as $\lambda^{t+1}_k := -\langle \mathsf{a}_k, \mathsf{v}^{t+1}_k \rangle$ for $k \neq k_t$, and $\lambda^{t+1}_{k_t} := -\sum_{k \neq k_t} \lambda^{t+1}_k$. Then denoting $\mathsf{u}^{t+1}_k := \mathsf{v}^{t+1}_k + \lambda^{t+1}_k$, $k \in [m]$, since $\sum_{k\in[m]} \lambda^{t+1}_k = 0$, we have $\bigoplus_{k=1}^m \mathsf{v}^{t+1}_k = \bigoplus_{k=1}^m \mathsf{u}^{t+1}_k$ and hence, recalling (49),

$$\pi^{t+1} = \exp\Big(-\mathsf{C}/\eta + \bigoplus_{k=1}^m \mathsf{u}^{t+1}_k \Big) \odot \bigotimes_{k=1}^m \mathsf{a}_k, \tag{101}$$

Moreover, from (100) we have that for every $k \neq k_t$ and every $j_k, \ell_k \in [n_k]$

$$u^{t+1}_{k,j_k} - u^{t+1}_{k,\ell_k} = v^{t+1}_{k,j_k} - v^{t+1}_{k,\ell_k} \leq 2\|\mathsf{C}\|_\infty/\eta,$$

which, using $\sum_{j\in[n_k]} a_{k,j} = 1$ and the fact that the λ^{t+1}_k's are chosen so that $\langle \mathsf{a}_k, \mathsf{u}^{t+1}_k \rangle = 0$ for all $k \neq k_t$, implies

$$-u^{t+1}_{k,\ell_k} = \sum_{j_k \in [n_k]} a_{k,j_k}(u^{t+1}_{k,j_k} - u^{t+1}_{k,\ell_k}) \leq 2\|\mathsf{C}\|_\infty/\eta, \quad \ell_k \in [n_k], \tag{102}$$

and

$$u^{t+1}_{k,j_k} = \sum_{\ell_k \in [n_k]} a_{k,\ell_k}(u^{t+1}_{k,j_k} - u^{t+1}_{k,\ell_k}) \leq 2\|\mathsf{C}\|_\infty/\eta, \quad j_k \in [n_k]. \tag{103}$$

Therefore, we have obtained that $\|\mathsf{u}^{t+1}_k\|_\infty \leq 2\|\mathsf{C}\|_\infty/\eta$ for $k \neq k_t$. On the other hand, similar to what was done in the proof of Theorem 4.4 we can derive that, for every $j_{k_t} \in L_t = [n_{k_t}]$,

$$\exp(-u^{t+1}_{k_t,j_{k_t}}) = \sum_{j_{-k_t} \in \mathcal{J}_{-k_t}} \exp\Big(-\mathsf{C}_{(j_{-k_t},j_{k_t})}/\eta + \sum_{k\neq k_t} u^t_{k,j_k} \Big) \prod_{k\neq k_t} a_{k,j_k}.$$

Since, recalling (102) and (103),

$$\exp(-(2m-1)\|C\|_\infty/\eta) \leq \exp\Big(-\mathsf{C}_{(j_{-k_t},j_{k_t})}/\eta + \sum_{k\neq k_t} u^t_{k,j_k} \Big) \leq \exp((2m-1)\|C\|_\infty/\eta),$$

and $\sum_{j_{-k_t} \in \mathcal{J}_{-k_t}} \prod_{k\neq k_t} a_{k,j_k} = 1$, we have

$$\exp(-(2m-1)\|C\|_\infty/\eta) \leq \exp(-u^{t+1}_{k_t,j_{k_t}}) \leq \exp((2m-1)\|C\|_\infty/\eta).$$

Therefore,

$$\exp\big(|u^{t+1}_{k_t,j_{k_t}}|\big) = \max\big\{ \exp\big(u^{t+1}_{k_t,j_{k_t}}\big), \exp\big(-u^{t+1}_{k_t,j_{k_t}}\big) \big\} \leq \exp((2m-1)\|C\|_\infty/\eta) \tag{104}$$

and hence

$$\|\mathsf{u}^{t+1}_{k_t}\|_\infty \leq (2m-1)\|\mathsf{C}\|_\infty/\eta.$$

Therefore, we have $\sum_{k\in[m]} \|\mathsf{u}^t_k\|_\infty \leq (4m-3)\|\mathsf{C}\|_\infty/\eta =: M$ and due to (101) and the computation (99), we can use $M_1 = M$ in Theorem 4.1 and get (28). Concerning iteration complexity, recalling (61) we have that $\sum_{k\in[m]} \|\mathsf{u}_k - \mathsf{v}^\star_k\|_\infty \leq 2(4m-3)\|\mathsf{C}\|_\infty/\eta$ and hence as done in (89) we have

$$\delta_t \leq \frac{2(4m-3)\|\mathsf{C}\|_\infty}{\eta} d^\infty_t.$$

Moreover, since $\|\pi^t\|_1 = 1$ and $\bar{b} = 1$, it follows from (91), (92) and (93) that

$$\delta_t - \delta_{t+1} \geq \frac{(d^\infty_t)^2}{2}. \tag{105}$$

Thus, the statement follows from Lemma C.3. □

D. Numerical Experiments

Here we provide a more extensive report on the numerical experiments presented in Section 5:

- Bi-marginal ModelNet10 experiment: for the pair of objects shown in the first column of Figure 4, we report the performance of BatchGreenkhorn and Sinkhorn in Figure 5. Three sizes n, four regularization parameters η and eighteen relative batch-sizes were tested.
- Bi-marginal ModelNet10 experiment with simulated annealing: to illustrate that BatchGreenkhorn can be further accelerated, we have implemented BatchGreenkhorn with a simulated annealing strategy in the same way as (Flamary et al., 2021) did for Sinkhorn. This has required a log-domain implementation of BatchGreenkhorn in small regularization parameter regimes. For the small size point clouds from first two objects from Figure 4 we compare BatchGreenkhorn(25%) and Sinkhorn to a baseline of GeomLoss library (Feydy et al., 2019) by setting *debias=True, backend='online'* and using annealing by setting $scaling$ parameter. Since our implementations are using the stopping criterion $d_\infty \leq \varepsilon$, we first run GeomLoss and check a posteriori the error ε on the marginals, and then we run Sinkhorn and BatchGreenkhorn with and without annealing till this precision ε is achieved. In Table 2 we show the computation times for size $n = 1000$, regularization parameter $\eta/\|\mathsf{C}\|_\infty = 10^{-3}$ and two scaling parameters. We note that the annealing strategy does accelerate convergence of BatchGreenkhorn. However, we stress that we did not optimize the scaling procedure for BatchGreenkhorn. In particular the overhead of computing during the iterations the stopping criterion has not been factor out and this affects more the BatchGreenkhorn than Sinkhorn, because of the greedy strategy. Moreover, here the setting is small size and we already observed in the previous experiments that in that scenario BatchGreenkhorn is not performing so well with respect to Sinkhorn.
- Label-to-label distance experiment: in Tables 3 and 4 we present total running time (min) needed for BatchGreenkhorn and Sinkhorn algorithms to compute 45 η-regularized OT problems with tolerance $\varepsilon = 10^{-6}$ for FashinMNIST and CIFAR10 datasets, respectively.
- Multi-marginal ModelNet10 experiment: for inferring the joint distribution of six objects shown in Figure 4, we report the performance of cyclic Sinkhorn, (greedy) MultiSinkhorn and BatchGreenkhorn in Figure 6.

Table 2: Run time in seconds for simulated annealing acceleration on ModelNet10 experiment.

Algorithm	$scaling = 0.75$ $\varepsilon = 6.75 \cdot 10^{-2}$	$scaling = 0.99$ $\varepsilon = 4.2 \cdot 10^{-3}$
GeomLoss	**0.51**	11.5
Sinkhorn	3.22	5.90
Sinkhorn with annealing	0.55	**2.49**
BatchGreenkhorn(25%)	3.87	13.86
BatchGreenkhorn(25%) with annealing	0.94	3.48

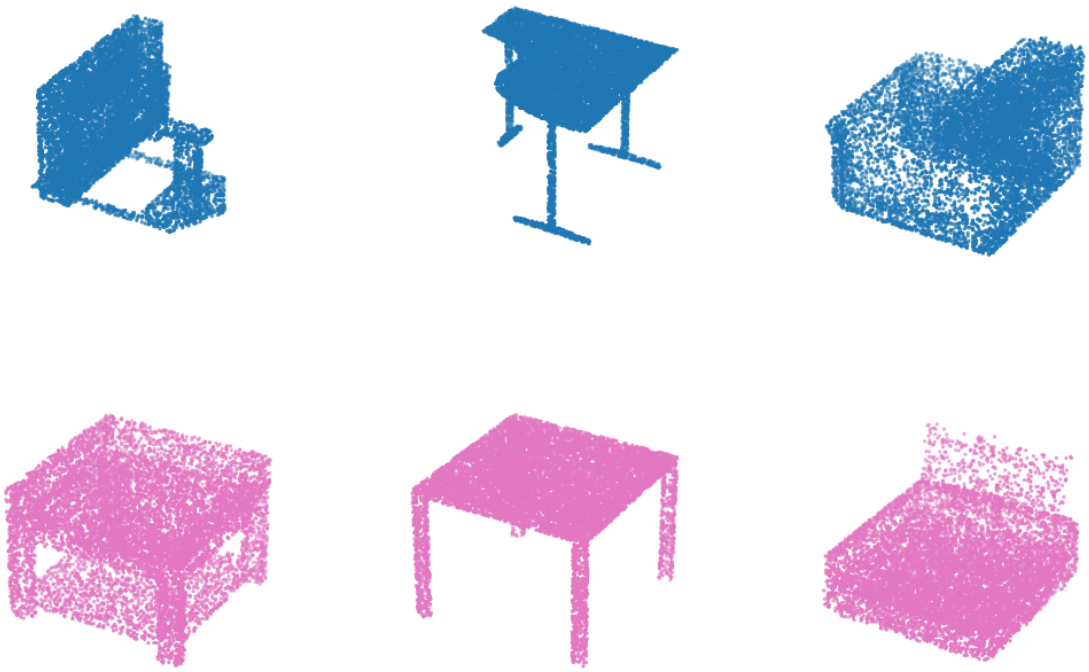

Figure 4: Six objects from ModelNet10 dataset used in the experiments.

Table 3: Total run time in minutes for the FashionMNIST dataset.

	$n = 2500$				$n = 5000$			
$\|C\|_\infty/\eta$	Sinkhorn	BatchGreenkhorn			Sinkhorn	BatchGreenkhorn		
		50%	25%	12.5%		50%	25%	12.5%
5	1.89	1.42	1.38	**1.37**	5.88	4.37	4.26	**4.21**
10	2.86	2.15	2.02	**2.00**	8.73	6.55	6.27	**6.16**
15	3.88	2.82	**2.71**	**2.71**	11.83	8.76	8.32	**8.22**
20	4.92	3.55	3.41	**3.39**	15.00	10.82	10.43	**10.26**
25	6.01	4.30	4.14	**4.12**	18.41	13.04	12.55	**12.52**

Table 4: Total run time in minutes for the CIFAR-10 dataset.

	$n = 2500$				$n = 5000$			
$\|C\|_\infty/\eta$	Sinkhorn	BatchGreenkhorn			Sinkhorn	BatchGreenkhorn		
		50%	25%	12.5%		50%	25%	12.5%
5	6.61	5.20	**4.96**	5.37	21.41	16.82	**15.75**	17.13
10	9.87	7.31	6.81	**6.65**	31.78	23.35	21.53	**20.92**
15	12.99	9.6	9.13	**8.91**	42.24	30.53	28.53	**28.09**
20	16.7	12.09	11.47	**11.11**	53.63	38.10	35.58	**35.03**
25	20.23	14.55	13.76	**13.35**	65.32	45.88	43.25	**41.98**

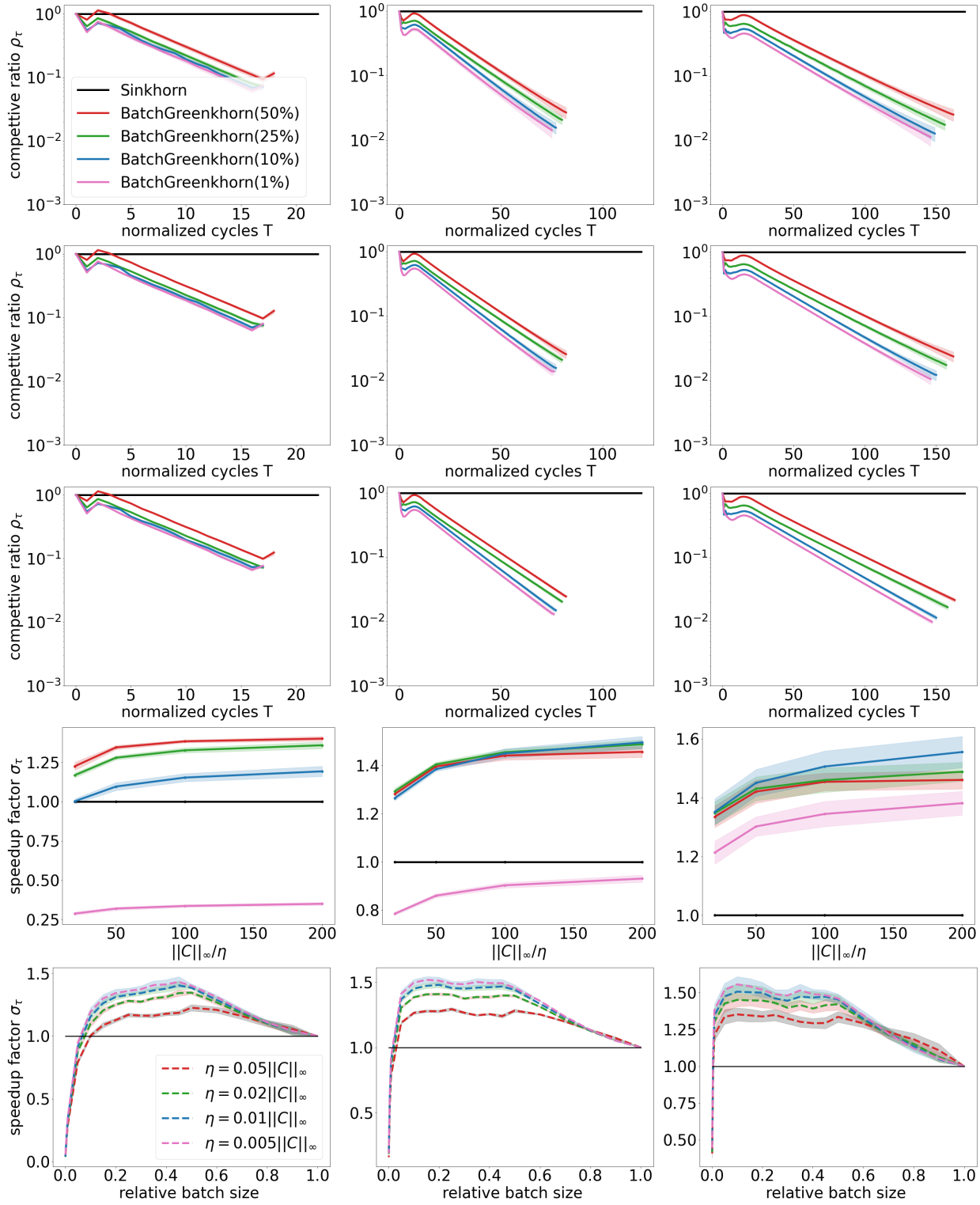

Figure 5: Bi-marginal ModelNet10 experiment : In the first three rows we show competitive ratios $\rho_\tau(T)$ for $n = 10000, 30000, 50000$ (rows) and relative regularization parameters $\eta/\|\mathsf{C}\|_\infty = 0.05, 0.02, 0.01, 0.005$ (columns). In the bottom two rows we show speedup factors σ_τ vs. relative regularization $\|\mathsf{C}\|_\infty/\eta$ (fourth row) and vs. relative batch size (fifth row) for $n = 10000, 30000, 50000$ (columns). In all plots, mean is bold and $\pm$ standard deviation is shaded.

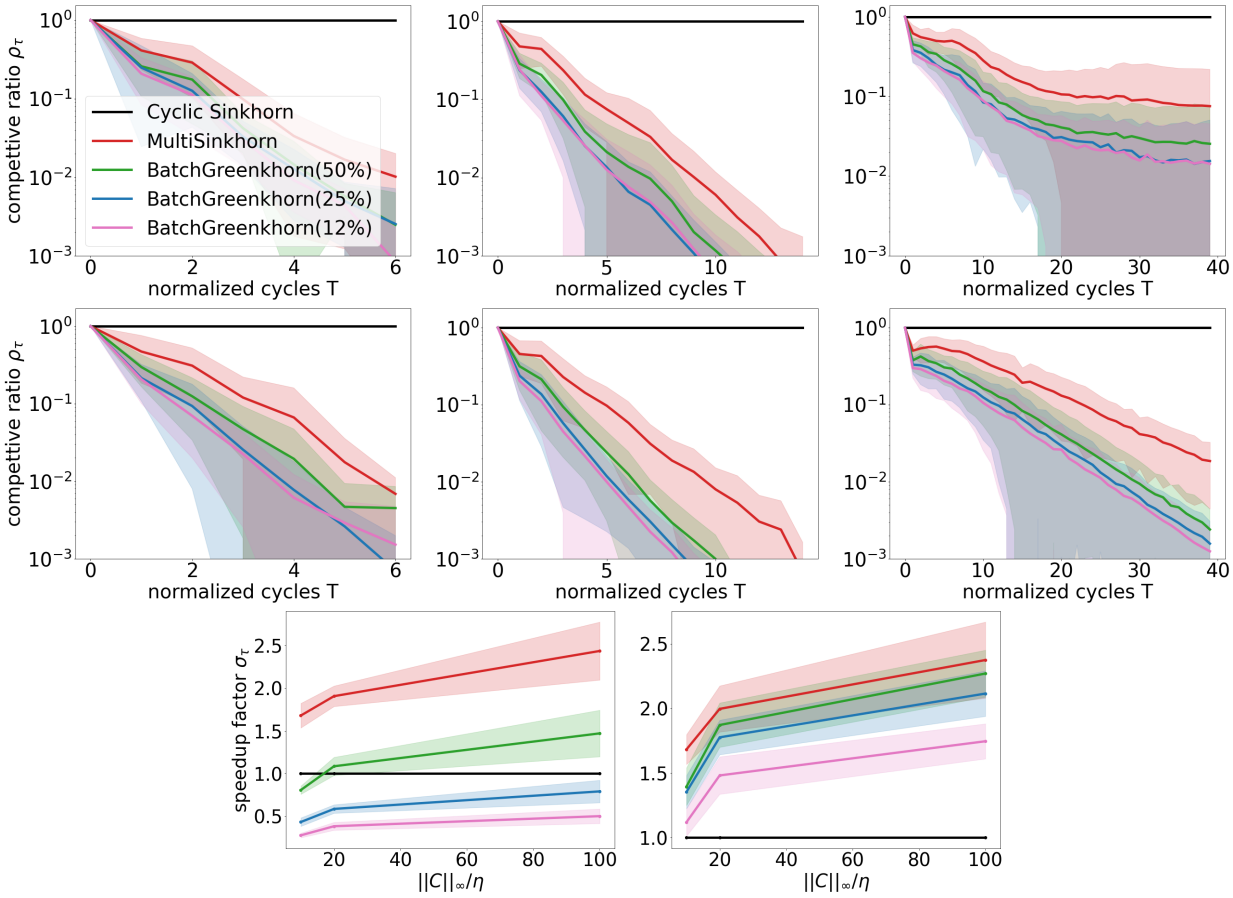

Figure 6: Multi-marginal ModelNet10 experiment : In the top two rows we show competitive ratios $\rho_\tau(T)$ for $n = 8$ and $n = 16$ (rows) and relative regularization parameters $\eta/\|\mathsf{C}\|_\infty = 0.1, 0.05, 0.01$ (columns). In the bottom row we show speedup factors σ_τ vs. relative regularization $\|\mathsf{C}\|_\infty/\eta$ for $n = 8$ and $n = 16$ (columns). In all plots, mean is bold and $\pm$ standard deviation is shaded.