International Conference On Machine Learning

Vladimir R. Kostic, Saverio Salzo and Massimilano Pontil, Istituto Italiano di Tecnologia, Genova, Italy

Batch Greenkhorn Algorithm for Entropic-Regularized Multimarginal Optimal Transport

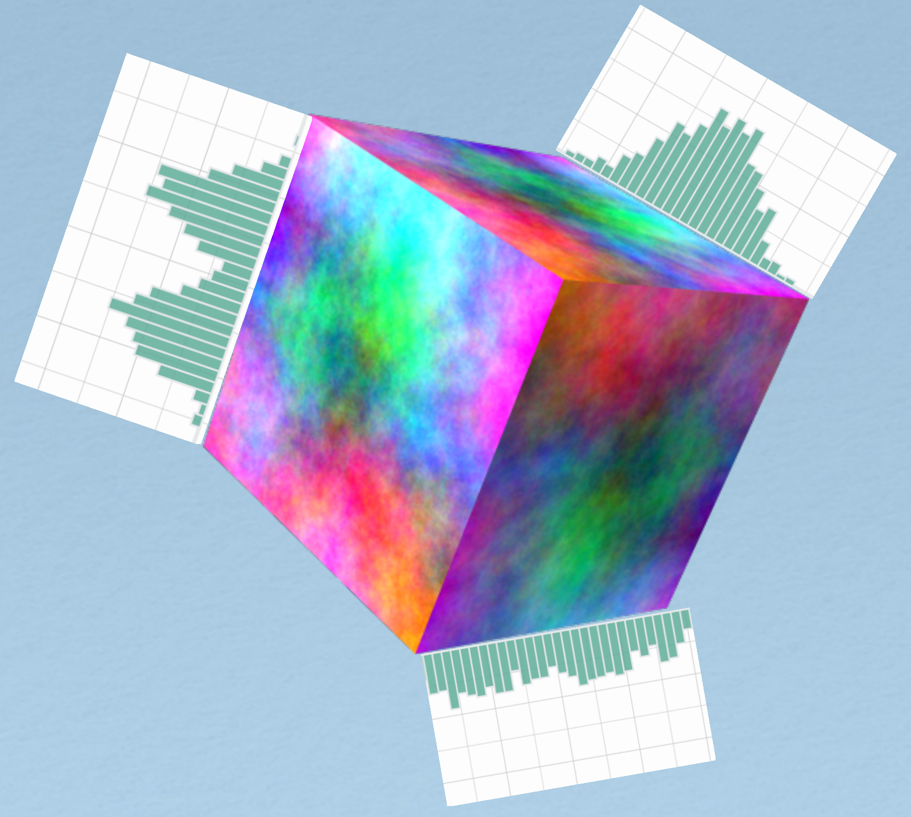
Linear Rate of Convergence and Iteration Complexity

ISTITUTO ITALIANO DI TECNOLOGIA
COMPUTATIONAL STATISTICS AND MACHINE LEARNING

MOT and RMOT : KL projections point of view

- Sinkhorn-type algorithms are power-horse of optimal transport!
- While many aspects of their convergence are understood, some questions remain open, especially in the multimarginal OT (MOT).

MOT : $\min_{\pi \in \Pi_+} \langle \mathsf{C}, \pi \rangle$



$\mathsf{a}_k \in \mathbb{R}_+^{n_k}$, $\|\mathsf{a}_k\|_1 = 1$ - given histograms, $k \in [m]$

$\mathsf{C} \in \mathbb{X}$ - given cost tensor

$\mathbb{X} = \mathbb{R}^{n_1 \times \cdots \times n_m}$ - vector space of m-dim tensors

$\mathsf{R}_k : \mathbb{X} \to \mathbb{R}^{n_k}$ - k-th push-forward operator

$\mathsf{R} : \mathbb{X} \to \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}, \qquad \mathsf{R}(\pi) = (\mathsf{R}_1(\pi), \ldots, \mathsf{R}_m(\pi))$

$\Pi_+ = \{\pi \in \mathbb{X}_+ \mid \mathsf{R}(\pi) = (\mathsf{a}_1, \ldots, \mathsf{a}_m)\}$ - transport polytope

MOT and RMOT : KL projections point of view

- Sinkhorn-type algorithms are power-horse of optimal transport!
- Here we study entropic-regularised MOT (RMOT).

$$\text{RMOT} : \pi^\star = \arg\min_{\pi \in \Pi_+} \langle \mathsf{C}, \pi \rangle + \eta \mathsf{H}(\pi)$$

$$\pi^\star = \arg\min_{\substack{\mathsf{R}_k(\pi) = \mathsf{a}_k \\ k \in [m]}} \mathsf{KL}(\pi, \xi)$$

$\eta > 0$ - regularization parameter

$\mathsf{H}(\pi) = \sum_{j \in \mathcal{J}} \pi_j (\log \pi_j - 1)$

$\mathcal{J} := \{j = (j_1, \dots, j_m) | j_k \in [n_k], \forall k \in [m]\}$ - multiindices

$\mathsf{KL} : \mathbb{X} \times \mathbb{X} \to [0, +\infty]$ - Kulback-Leibler (KL) divergence

$$\mathsf{KL}(\pi, \gamma) = \begin{cases} \sum_{j \in \mathcal{J}} \pi_j \log \frac{\pi_j}{\gamma_j} - \pi_j + \gamma_j & \text{if } \pi \in \mathbb{X}_+, \gamma \in \mathbb{X}_{++} \\ +\infty & \text{otherwise,} \end{cases}$$

$\xi = \nabla \mathsf{H}^*(-\mathsf{C}/\eta) = \exp(-\mathsf{C}/\eta)$ - Gibbs kernel tensor

MOT and RMOT : KL projections point of view

- Sinkhorn-type algorithms are power-horse of optimal transport!
- We approach RMOT with the lenses of (greedy) Bregman projections.

$$\pi^\star = \mathcal{P}_\Pi(\xi)$$

Regularised optimal plan is Bregman projection of the kernel onto the affine set

$\Pi = \{\pi \in \mathbb{X} \mid \mathsf{R}(\pi) = (\mathsf{a}_1, \ldots, \mathsf{a}_m)\}$ - affine set

$\mathcal{P}_\mathcal{C}(\pi) := \arg\min_{\gamma\in\mathcal{C}} \mathsf{KL}(\gamma, \pi)$ - KL projection on $\mathcal{C}$

$\mathsf{KL}_\mathcal{C}(\pi) := \mathsf{KL}(\mathcal{P}_\mathcal{C}(\pi), \pi)$ - KL distance of π from $\mathcal{C}$

$\xi = \nabla \mathsf{H}^*(-\mathsf{C}/\eta) = \exp(-\mathsf{C}/\eta)$ - Gibbs kernel tensor

Greedy KL projections for entropic RMOT

$$\Pi_{(k,L)} := \{\pi \in \mathbb{X} \,|\, (\mathsf{R}_k(\pi))_{|L} = \mathsf{a}_{k|L}\}$$

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

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

for $t = 0, 1, \dots$
- choose $(k_t, L_t) \in \mathcal{I}(\tau)$
- $\pi^{t+1} = \mathcal{P}_{\Pi_{(k_t, L_t)}}(\pi^t)$

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

$\tau = (\tau_k)_{1 \le k \le m}$ - vector of batch sizes

admissible choices

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

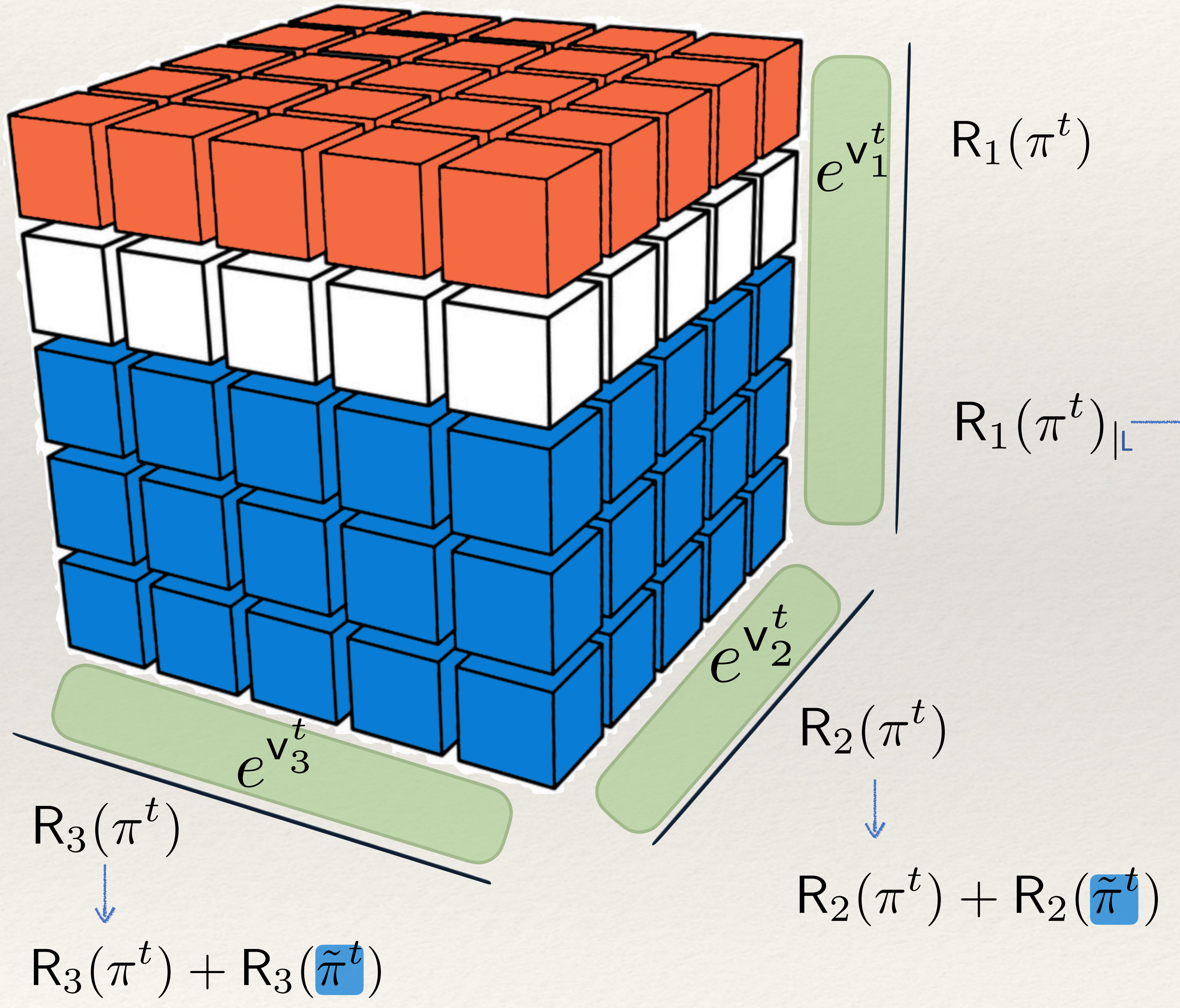
$$\Pi = \bigcap_{(k,L)\in\mathcal{I}(\tau)} \Pi_{(k,L)}$$

BatchGreenkhorn:

- formulation that allows convergence analysis
- efficient implementations are possible

BatchGreenkhorn algorithm

Full marginal step = Sinkhorn | **Batch step = BatchGreenkhorn** | **Coordinate step = Greenkhorn**



$$\pi^0 = \xi \odot \bigotimes_{k\in[m]} a_k$$

$$\pi^t = \underbrace{\exp\Big(-C/\eta}_{\text{kernel}} + \underbrace{\bigoplus_{k\in[m]} v_k^t}_{\text{potentials}}\Big) \odot \underbrace{\bigotimes_{k\in[m]} a_k}_{\text{normalization}}$$

kernel — potentials — normalization

$$v_{1|L}^t \longrightarrow v_{1|L}^t + \log \frac{a_{1|L}}{R_1(\pi^t)_{|L}}$$

scalable update formulas + **cheap greedy decision**

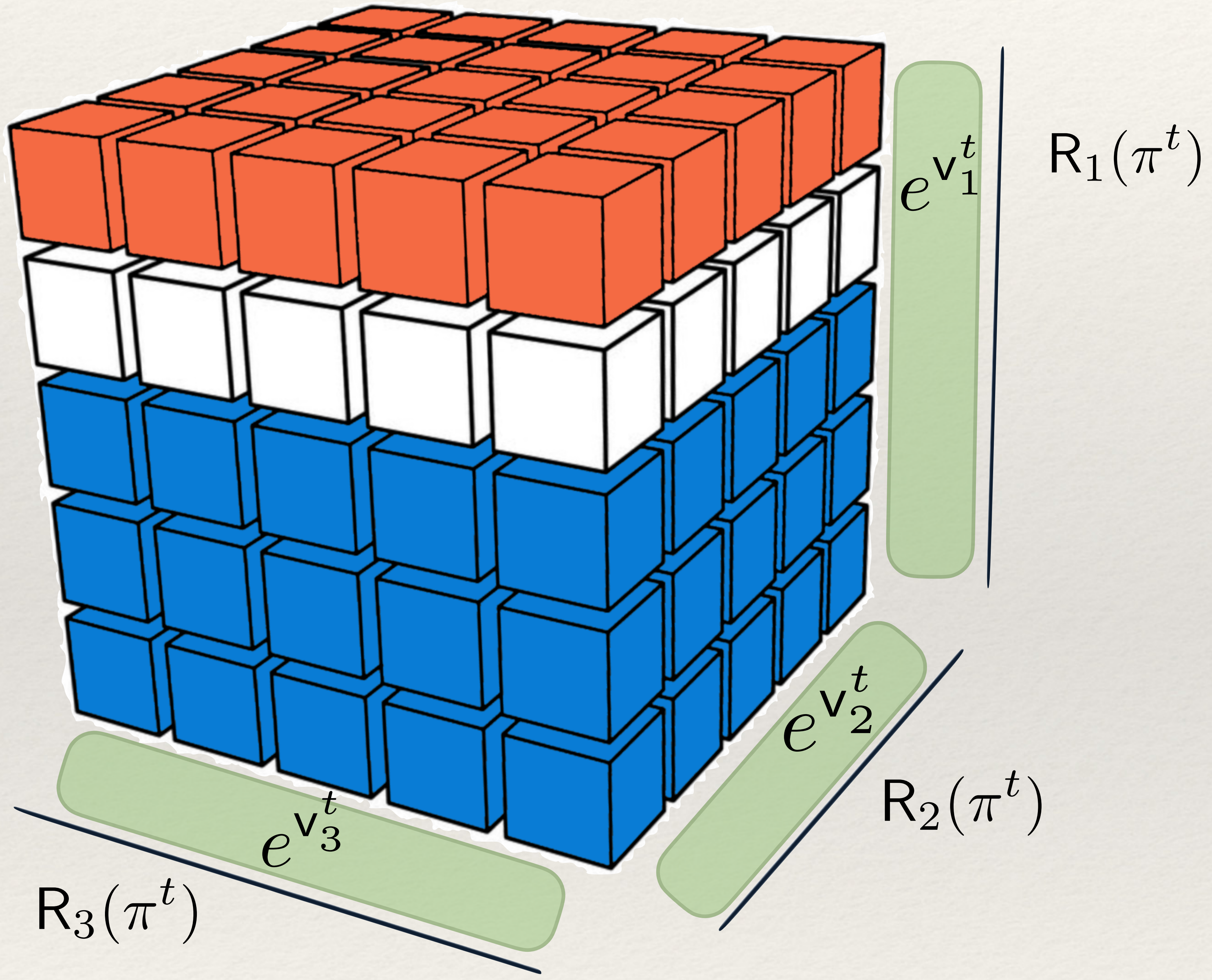
Key points:

- the cost of being greedy is linear in m and n
- dual and marginal updates can be done in no. of operations:

  ~ (full kernel) * (tau / n)
- we can compare w.r.t. normalised cycles T=1 pass of cyclic Sinkhorn

BatchGreenkhorn algorithm

Full marginal step = Sinkhorn **Batch step = BatchGreenkhorn** **Coordinate step = Greenkhorn**



Key tools for convergence theory:

- Pythagoras theorem $\mathrm{KL}_\Pi(\pi^{t+1}) = \mathrm{KL}_\Pi(\pi^t) - \mathrm{KL}(\pi^{t+1}, \pi^t)$

- symmetric Bregman decomposition

$$\mathrm{KL}(\pi^\star, \pi^t) + \mathrm{KL}(\pi^t, \pi^\star) = \langle \pi^\star - \pi^t, \log \frac{\pi^\star}{\pi^t} \rangle = \sum_{k \in [m]} \langle \pi^\star - \pi, \mathrm{R}_k^*(\mathrm{v}_k^\star - \mathrm{v}_k^t) \rangle$$

$$= \sum_{k \in [m]} \langle \mathrm{a}_k - \mathrm{R}_k(\pi^t), \mathrm{v}_k^\star - \mathrm{v}_k^t \rangle$$

- Pinsker inequality $\mathrm{KL}(\pi, \gamma) \geq \dfrac{3\|\pi - \gamma\|_1^2}{2\|\pi\|_1 + 4\|\gamma\|_1}$

- strong convexity of H and H* on bounded sets

Convergence results

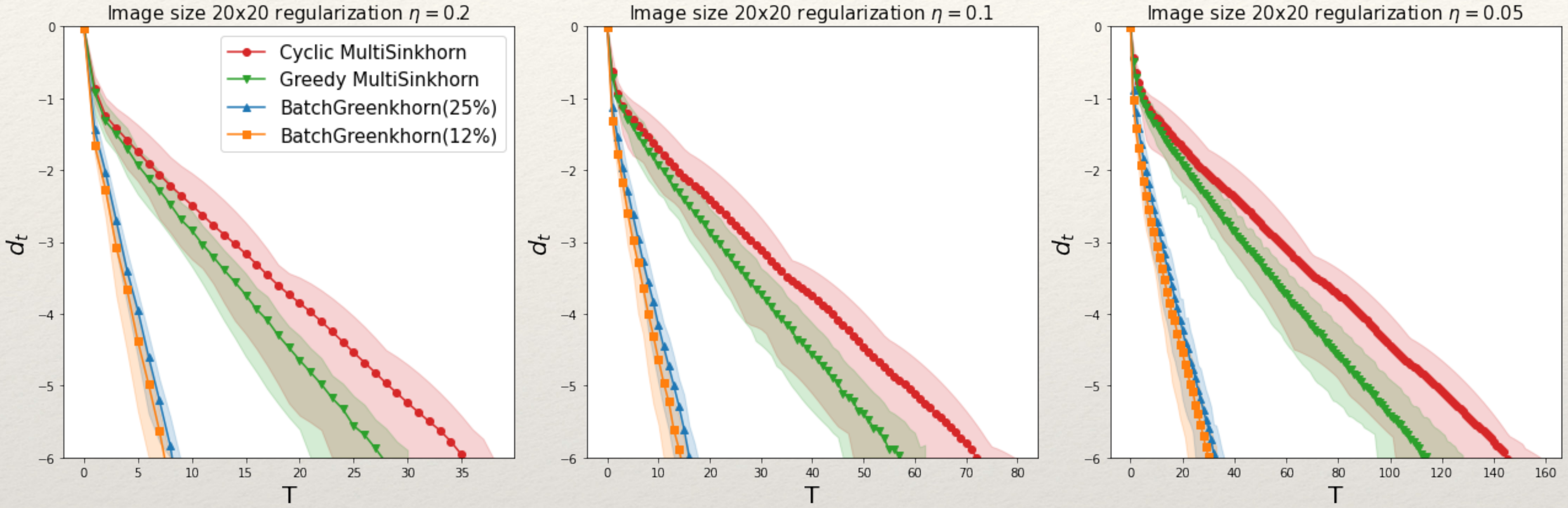
Algorithm (problem)	Convergence type	Current best	Our result	
Sinkhorn (ROT)	✔ (GL)	$1 - \frac{1}{2}e^{-24/\eta}$ (Carlier, 2021)	$\left(1 - e^{-17\|C\|_\infty/\eta}\right)^2$	Theorem 4.5
	✔ (IC)	$\mathcal{O}\left(\frac{\|C\|_\infty/\eta + \log n}{\varepsilon}\right)$ (Dvurechensky et al., 2018)	$\mathcal{O}\left(\frac{\|C\|_\infty}{\eta\varepsilon}\right)$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

✔ existing results improved

★ our new results

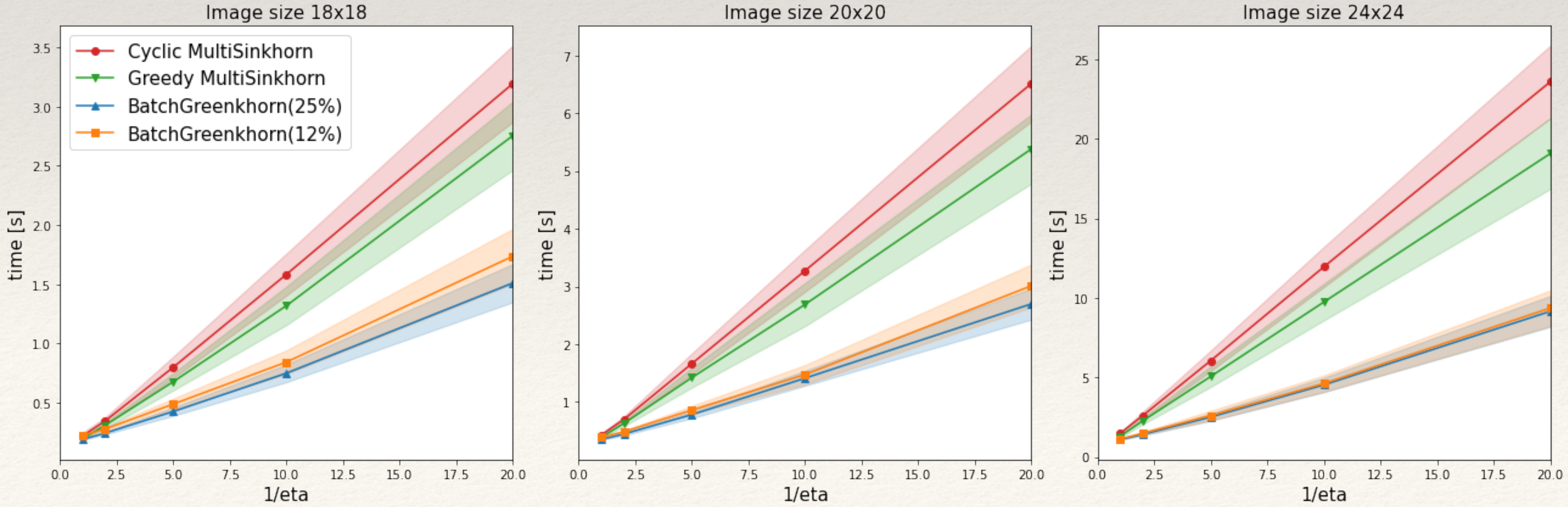
Performance w.r.t normalised iterations and time

Computation of free-support Wasserstein Barycenter of 3 histograms of image data



In this simple experiment RMOT for m=3 is solved for n = 400 (above) and n = 256, 400, 576 (bellow).

Above, we observe that w.r.t. normalised iterations the Sinkhorn algorithm is less efficient than the Batch Greenkhorn.



Bellow, we see that Batch Greenkhorn can even speed up (cyclic / greedy) Sinkhorn by tuning the batch size to exploit the adversarial effects of the convergence speed (iteration complexity) vs. parallelisation of kernel operations (computational complexity).

Contributions

We introduce and study *BatchGreenkhorn* as a new algorithmic framework for RMOT which comes with some theoretical and practical benefits:

- in bi-marginal OT it covers Sinkhorn and Greenkhorn, in RMOT it covers (greedy) MultiSinkhorn of Lin et al. (2020)
- we study convergence theory in primal iterates and provide global linear convergence rate and iteration complexity
- our results improve existing ones and fill some gaps in literature
- flexibility of the batch provides practical advantages

ICML
International Conference
On Machine Learning

Thank you!

ISTITUTO ITALIANO
DI TECNOLOGIA
COMPUTATIONAL STATISTICS
AND MACHINE LEARNING