Mirror Sinkhorn: Fast Online Optimization on Transport Polytopes

Marin Ballu [1] **Quentin Berthet** [2]

Abstract

Optimal transport is an important tool in machine learning, allowing to capture geometric properties of the data through a linear program on transport polytopes. We present a single-loop optimization algorithm for minimizing general convex objectives on these domains, utilizing the principles of Sinkhorn matrix scaling and mirror descent. The proposed algorithm is robust to noise, and can be used in an online setting. We provide theoretical guarantees for convex objectives and experimental results showcasing it effectiveness on both synthetic and real-world data.

1. Introduction

Optimal transport is a seminal problem in optimization (Monge, 1781), and an important topic in analysis (Villani, 2008). The discrete case is a linear program on the set of nonnegative matrices with fixed row and column sums (Kantorovich, 1942). This set forms the *transport polytope*, whose elements can be interpreted as the law of joint distributions for (X, Y) with marginal distributions μ, ν.

In machine learning, OT has recently gained in importance following the work of (Cuturi, 2013), presenting an entropic-regularized method and using the Sinkhorn algorithm to efficiently optimize it; see (Peyré et al., 2019) and references therein for an overview of this topic and its applications. The method can be used to solve the original OT problem (Altschuler et al., 2017), by setting a regularization parameter that depends on the cost matrix and desired precision level. We consider here the more general problem of convex optimization of **any convex function** on the transport polytope. This comprises several optimization problems, included but **not limited** to both optimal transport and its entropic regularized version. The problem for a quadratic function has been studied both for the purpose of registering point cloud (Grave et al., 2019) (related to Gromov-Wasserstein problems (Mémoli, 2011; Solomon et al., 2016)) and for computing euclidean projection on the Birkhoff polytope (Li et al., 2020). It appears in statistical inference on random permutations (Birdal & Simsekli, 2019). Inference on random permutations can be obtained by minimizing various other convex functions (Linderman et al., 2018). Optimisation on the transport polytope also arises when trying to both compute and minimize a Wasserstein distance or sum of Wasserstein distances of a set of parametrized distributions. This applies to the computation of Wasserstein estimators (Ballu et al., 2020; Bassetti et al., 2006), private learning (Boursier & Perchet, 2019), Wasserstein barycenters (Rabin et al., 2011; Agueh & Carlier, 2011; Cuturi & Doucet, 2014) and generative models. In the latter case, the distribution generated by a neural network is compared to the sample distribution with the 1-Wasserstein distance in WGAN (Arjovsky et al., 2017) and Wasserstein autoencoders (Tolstikhin et al., 2017), or with a regularized version of the Wasserstein distance with Sinkhorn divergences (Genevay et al., 2018).

Several algorithms that have been suggested to solve optimal transport use iterated Bregman projections (Benamou et al., 2015; Dvurechensky et al., 2018). To minimize general convex functions on transport polytopes, we extend the approach to a single loop iterated algorithm. Interpretations of Sinkhorn algorithm as mirror descent in the dual (Mishchenko, 2019) and the primal (Léger, 2021; Aubin-Frankowski et al., 2022) have been used to derive convergence rates and to extend Sinkhorn to this more general problem. These extensions are supported by an analysis, assuming smoothness and strong convexity of the objective. Streaming iterations of Sinkhorn have been proposed in (Mensch & Peyré, 2020), focusing solely on minimizing a regularized optimal transport problem. General convex optimization algorithms can theoretically achieve the best asymptotic rates for optimal transport. A ε-close optimal transport plan can be obtained with accelerated gradient descent schemes (Dvurechensky et al., 2018; Lin et al., 2019; Guo et al., 2020), or accelerated alternative minimisation (Guminov et al., 2021). These are only analyzed when given deterministic gradient updates, and their parameters depend on an imposed desired level of optimization precision ε.

Using the Sinkhorn algorithm to solve OT, as studied in

[*]Equal contribution [1]University of Cambridge, UK [2]Google DeepMind, Paris, France. Correspondence to: Quentin Berthet <qberthet@google.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

(Altschuler et al., 2017) requires to set a regularization parameter $\alpha > 0$, as a function of the desired optimization precision. Indeed, this algorithm is tied to an entropic-regularized version of this problem, whose solution γ^*_α is different: there is a regularization bias. We note that this has some advantages. In particular, the solution is a continuously differentiable function of the problem inputs (Peyré et al., 2019). This is part of a wide effort to create differentiable versions of discrete operators such as optimizers (Cuturi & Blondel, 2017; Berthet et al., 2020; Blondel et al., 2020; Vlastelica et al., 2019; Paulus et al., 2020), to ease their inclusion in end-to-end differentiable pipelines that can be trained with first-order methods in applications (Cordonnier et al., 2021; Kumar et al., 2021; Carr et al., 2021; Le Lidec et al., 2021; Baid et al., 2022; Llinares-López et al., 2021) and other optimization algorithms (Dubois-Taine et al., 2022). This regularized objective, as well as alternate regularizations (Blondel et al., 2018) fall within our framework, and can be optimized using our algorithm.

Our algorithm, which we call *Mirror Sinkhorn*, is based on the principles of mirror descent on the transport polytope (which requires an oracle to solve Bregman-regularized linear problems on this set), and the Sinkhorn algorithm which enforces normalization of rows and columns to satisfy the constraints. The use of multiple Sinkhorn steps for solving a mirror descent oracle has been proposed in (Alvarez-Melis et al., 2018), and the convergence of the resulting algorithm has been further analysed in (Xie et al., 2020) and later in (Aubin-Frankowski et al., 2022) for smooth and strongly convex objectives with respect to the entropy. We propose an analysis for a single step of Sinkhorn normalization between gradient updates.

Our contributions. The *Mirror Sinkhorn* algorithm takes stochastic gradients as input, is adaptive to a change of objective, and its parameters are independent of the required precision. In summary, we make the following contributions

- We introduce a **single-loop**, practically efficient algorithm for optimization on the transport polytope.
- We provide theoretical guarantees for the **performance** under various assumptions, including OT (linear).
- We show that this algorithm can be adapted to handle different scenarios such as stochastic gradients, online settings, and related tensor problems.

Notations. The standard Euclidean product for vectors and matrices is denoted by $\langle \cdot, \cdot \rangle$. For a positive integer n, we denote by $1_n = (1, \dots, 1)$ the vector of $\mathbb{R}^n$ with all ones, and by $[n]$ the set of integers from 1 to n, included. We denote by Δ_n the probability simplex, defined as

$$\Delta_n = \left\{ \nu \in \mathbb{R}^n : \nu \geq 0,\ \langle 1_n, \nu \rangle = 1 \right\}.$$

For two reals a, b we denote by $a \wedge b$ the minimum of a and b. We extend this notation to vectors, meaning the entrywise minimum. The analogue for probability matrices in $\mathbb{R}^{m \times n}$ (resp. probability tensors in $\mathbb{R}^{m_1 \times \dots \times m_d}$) is defined *mutatis mutandis* and denoted by $\Delta_{m \times n}$ (resp. $\Delta_{m_1 \times \dots \times m_d}$). The operator $\mathrm{Diag}(\cdot)$ yields a diagonal matrix in $\mathbb{R}^{m \times m}$ from a vector in $\mathbb{R}^m$, with diagonal entries $(\mathrm{Diag}(x))_{ii} = x_i$. We note the entrywise product for matrices $(A \odot B)_{ij} = A_{ij} B_{ij}$ and the entrywise division $(A \oslash B)_{ij} = A_{ij}/B_{ij}$. The notations $\exp(A)$ and $\log(A)$ are used for the entrywise exponential and natural logarithm functions, i.e. $(\exp(A))_{ij} = e^{A_{ij}}$, $(\log(A))_{ij} = \log(A_{ij})$. The transpose of a matrix A is noted $A^\top$. We denote by $\|\cdot\|_\infty$ and $\|\cdot\|_1$ respectively the entrywise ℓ_1 and ℓ_∞ norms on vectors and matrices.

2. Problem and methods

For two positive integers $m, n \geq 1$, let $\mu \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^n$ be two probability distributions on $[m]$ and $[n]$ respectively, represented as vectors, elements of the simplexes Δ_m and Δ_n. The transport polytope between μ and ν, denoted by $\mathcal{T}(\mu, \nu)$, is a subset of $\Delta_{m \times n}$, set of $m \times n$ probability matrices. It contains all probability matrices γ whose rows sum to μ and columns sum to ν. It can be interpreted as the set of couplings, i.e. joint distributions between two variables with fixed marginals μ and ν, defined as

$$\mathcal{T}(\mu, \nu) = \left\{ \gamma \in \mathbb{R}_+^{m \times n} :\ \gamma 1_n = \mu,\ \gamma^\top 1_m = \nu \right\}. \quad (1)$$

It is the intersection of the simplex $\Delta_{m \times n}$ with affine spaces given by $\mathcal{R}(\mu) = \{\gamma \in \mathbb{R}^{m \times n} : \gamma 1_n = \mu\}$ and $\mathcal{C}(\nu) = \{\gamma \in \mathbb{R}^{m \times n} : \gamma^\top 1_m = \nu\}$. We consider in this work the problem of convex optimisation on the transport polytope:

$$\min_{\gamma \in \mathcal{T}(\mu, \nu)} f(\gamma), \quad (2)$$

where $f : \Delta_{m \times n} \to \mathbb{R}$ is a real-valued differentiable convex function that is defined on the set of probability matrices. The optimal transport problem is the particular case of (2) where f is linear:

$$\min_{\gamma \in \mathcal{T}(\mu, \nu)} \langle C, \gamma \rangle. \quad (3)$$

2.1. Mirror descent and Sinkhorn.

A method that can be used to solve constrained convex optimization problems such as (2) is Mirror Descent (Beck & Teboulle, 2003). For a convex differentiable function $\psi : \Delta_{m \times n} \to \mathbb{R}$, we define a Bregman divergence

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle.$$

The function ψ plays the role of a barrier function, it is such that for every compact $K \subset \mathbb{R}$, its inverse image

$\psi^{-1}(K) = \{x \in \Delta_{m\times n} : \psi(x) \in K\}$ is in the interior of the polytope of constraints. We recall that $\mathcal{T}(\mu,\nu)$ is the intersection of the probability simplex $\Delta_{m\times n}$ with the affine spaces $\mathcal{R}(\mu)$ and $\mathcal{C}(\nu)$. A choice of function ψ that guarantees that the iterates are in the interior of the simplex $\Delta_{m\times n}$ is the negative entropy, defined by

$$\forall \gamma \in \mathbb{R}_+^{m\times n},\ \psi(\gamma) = -H(\gamma) = \langle \gamma, \log \gamma\rangle,$$

and the resulting Bregman divergence is the relative entropy

$$D_{\mathrm{KL}}(A,B) = \langle A, \log(A) - \log(B)\rangle + \langle B - A, 1_{m\times n}\rangle.$$

Under these conditions, a mirror descent algorithm for (2) would define its iterates by

$$\gamma_{t+1} = \underset{\mathcal{T}(\mu,\nu)}{\arg\min}\, \{\eta_t \langle \gamma, \nabla f(\gamma_t)\rangle + D_{\mathrm{KL}}(\gamma,\gamma_t)\} \qquad (4)$$

with η_t the step-size at time t. This update requires solving an entropic-regularised optimal transport problem at each iterate

$$\gamma_{t+1} = \underset{\mathcal{T}(\mu,\nu)}{\arg\min}\, \{\langle \gamma, C_t\rangle - \alpha_t H(\gamma)\}$$

with regularisation parameter $\alpha_t = 1/\eta_t$ and cost matrix $C_t = \nabla f(\gamma_t) - \log(\gamma_t)/\eta_t$. However, the oracle for this problem is not explicit. Solving this step approximately within a mirror descent loop has been used in Gromov-Wasserstein (Peyré et al., 2016). State of the art algorithms to solve this intermediate problem involve accelerated gradient descent schemes (Dvurechensky et al., 2018; Lin et al., 2019). A simple and popular algorithm to tackle this inner problem is the Sinkhorn matrix scaling algorithm (Sinkhorn, 1964; Cuturi, 2013). It consists in projecting alternatively an initial matrix $\gamma_1 = e^{-C/\alpha}$ on the marginal spaces with Bregman projections: if t is odd

$$\gamma_{t+1} = \underset{\gamma 1_n = \mu}{\arg\min}\, \{D_{\mathrm{KL}}(\gamma,\gamma_t)\}$$

and if t is even

$$\gamma_{t+1} = \underset{\gamma^\top 1_m = \nu}{\arg\min}\, \{D_{\mathrm{KL}}(\gamma,\gamma_t)\}.$$

Each of these iterates only relies on proportionately scaling the rows and columns of the matrix. An algorithm using several steps of the Sinkhorn algorithm to approximate mirror descent uses **nested loops**.

2.2. Mirror Sinkhorn.

The algorithm that we propose consists of alternating a step of entropic mirror descent and a step of Sinkhorn algorithm. Similarly to mirror descent, it can be written in a proximal form where the optimisation is performed on each marginal at a time: if t is even,

$$\gamma_{t+1} = \underset{\gamma 1_n = \mu}{\arg\min}\, \{\eta_t \langle \gamma, \nabla f(\gamma_t)\rangle + D_{\mathrm{KL}}(\gamma,\gamma_t)\}$$

if t is odd,

$$\gamma_{t+1} = \underset{\gamma^\top 1_m = \nu}{\arg\min}\, \{\eta_t \langle \gamma, \nabla f(\gamma_t)\rangle + D_{\mathrm{KL}}(\gamma,\gamma_t)\}.$$

Formally, Mirror Sinkhorn algorithm is defined in the following manner (see Figure 1 for an illustration).

Algorithm 1 Mirror Sinkhorn

Data: Initialise $\gamma_1 = \overline{\gamma}_1 = \mu\nu^\top$, define stepsize η_t.
for $1 \le t \le T$ **do**
  Update $\gamma'_{t+1} = \gamma_t \odot \exp(-\eta_t \nabla f(\gamma_t))$,
  if *t is even* **then**
    Rows: $\gamma_{t+1} = \mathrm{Diag}\left(\mu \oslash (\gamma'_{t+1} 1_n)\right) \gamma'_{t+1}$,
  else
    Cols: $\gamma_{t+1} = \gamma'_{t+1}\, \mathrm{Diag}\left(\nu \oslash ((\gamma'_{t+1})^\top 1_m)\right)$,
  end
  Update $\overline{\gamma}_{t+1} = \frac{t}{t+1}\overline{\gamma}_t + \frac{1}{t+1}\gamma_{t+1}$
end
Output: $\overline{\gamma}_T = \frac{1}{T}\sum_{t=1}^T \gamma_t$

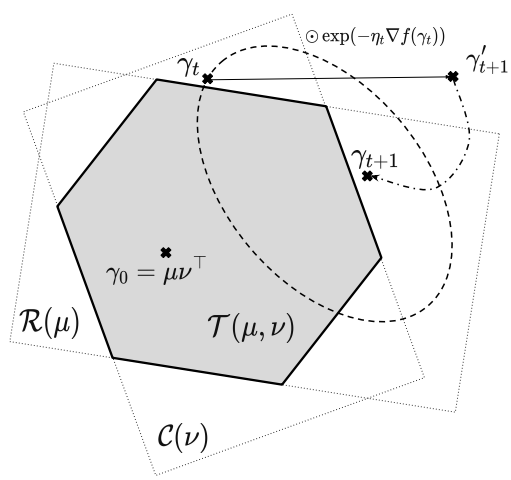

Figure 1. An illustration of the *Mirror Sinkhorn* algorithm: At each iteration, a gradient step on γ_t yields the intermediate γ'_{t+1}, which is projected by row or column scaling onto either $\mathcal{R}(\mu)$ or $\mathcal{C}(\nu)$.

Each step of the algorithm has a time complexity of $O(mn)$. Similarly to the Sinkhorn algorithm, this algorithm can be parallelized (Cuturi, 2013). Importantly, this a **single-loop** algorithm, in **stark contrast** with nested loop algorithms that rely on iterative algorithms within a loop of gradient updates. Our algorithms handles the affine constraints given by $\mathcal{R}(\mu)$ and $\mathcal{C}(\nu)$ by a **single and explicit normalization**. This makes the algorithm particularly easy to implement and efficient. Theoretical results providing convergence rates under different assumptions are provided in Section 3. Some settings require minor variations (use of stochastic

gradients, absence of running averaging). Even though we describe them in the corresponding sections, we also include them in the appendix for the sake of completeness.

Regularizing the marginals. One approach to guarantee that the constraints are satisfied is to add a positive regularization term that is minimized when the constraints are satisfied. The following proposition states that Algorithm 2.2 has the same iterates whether it is based on the regularized objective or the unregularized objective.

Proposition 2.1. *Let $f : \Delta_{m\times n} \to \mathbb{R}$ be a differentiable convex function, let $R_\mu : \Delta_m \to \mathbb{R}_+$ be a differentiable convex function such that $\nabla R_\mu(\mu') = 0$ if and only if $\mu' = \mu$ and let $R_\nu : \Delta_n \to \mathbb{R}_+$ be defined similarly. We define the following regularized objective*

$$f^{reg} : \gamma \mapsto f(\gamma) + R_\mu(\gamma 1_n) + R_\nu(\gamma^\top 1_m).$$

Let $(\gamma_t)_{t\geq 1}$ be the iterates of Mirror Sinkhorn with objective f and let $(\gamma_t^{reg})_{t\geq 1}$ be the iterates with objective f^{reg}. Then

$$\forall t \geq 1,\ \gamma_t^{reg} = \gamma_t.$$

Rounding for constraints satisfaction The output of our algorithm does not necessarily belong to $\mathcal{T}(\mu,\nu)$. This can easily be remedied by applying an elegant rounding algorithm by (Altschuler et al., 2017) to obtain a nearby matrix that belongs to the transport polytope.

Algorithm 2 Rounding algorithm (Altschuler et al., 2017)

Data: Matrix γ, target marginals μ, ν.
Normalise rows $\gamma' = \mathrm{Diag}\left(\frac{\mu}{\gamma 1_n} \wedge 1_m\right)\gamma$,
Normalise columns $\gamma'' = \gamma' \,\mathrm{Diag}\left(\frac{\nu}{(\gamma')^\top 1_m} \wedge 1_n\right)$,
Output: $\rho(\gamma) = \gamma'' + \left(\frac{\mu - \gamma'' 1_n}{\|\mu - \gamma'' 1_n\|_1}\right)(\nu - (\gamma'')^T 1_m)^\top$

One of the appeals of this function is that the ℓ^1 distance between its input and its output can be controlled by the ℓ^1 distance between their marginals

$$\|\rho(\gamma) - \gamma\|_1 \leq 2\left(\|\gamma 1_n - \mu\|_1 + \|\gamma^\top 1_m - \nu\|_1\right),$$

and that its output is in the transport polytope

$$\forall \gamma \in \mathbb{R}_+^{m\times n},\ \rho(\gamma) \in \mathcal{T}(\mu,\nu).$$

This algorithm has complexity $O(mn)$, which implies that it does not add asymptotic complexity if applied at the end of Algorithm 2.2. It is also parralelization-friendly.

To keep track of the constraint violation in the theoretical guarantees, we define $c(\cdot) : \Delta^{m\times n} \to \mathbb{R}$ by

$$c(\gamma) = \|\gamma 1_n - \mu\|_1 + \|\gamma^\top 1_m - \nu\|_1. \tag{5}$$

We also define for $\mu \in \Delta_m, \nu \in \Delta_n$ the constant δ appearing in our theoretical guarantees

$$\delta = \|\log \mu\|_\infty + \|\log \nu\|_\infty. \tag{6}$$

Use for optimal transport problem As described above, this algorithm can be used to tackle the OT problem. In order to provide an ε-close optimal transport plan, the Sinkhorn algorithm is initialized using the matrix $(e^{-C_{i,j}/\alpha})_{i,j}$ with α/ε being a constant which depends on the choice of μ and ν - see (Altschuler et al., 2017). The target error needs to be known at initialization. In contrast, the Mirror Sinkhorn algorithm's initialization and stepsizes do not depend on ε.

Another widely used method for improving the computation of optimal transport problems is to transform the problem into that of sliced Wasserstein distance, considering averages over lower-rank projections (Bonneel et al., 2015; Kolouri et al., 2019; Le et al., 2019; Nadjahi et al., 2021; Niles-Weed & Rigollet, 2022). There are some similarities with certain aspects of our method, e.g. when the use of lower-rank estimates of the gradient are used in a stochastic setting (see discussion in Section 3.2).

3. Theoretical guarantees

We present theoretical guarantees on the performance of the Mirror Sinkhorn algorithm in several settings. The proposed algorithm is evaluated under a variety of conditions, with small adaptations in each case. Proofs are in the Appendix.

3.1. Online setting

The Mirror Sinkhorn algorithm can also be used in an online setting, where there is not a unique function f but a stream of functions f_t and the performance of the algorithm is evaluated as a regret bound (Bubeck et al., 2012). The gradient can be replaced by $g_t = \nabla f_t(\gamma_t)$, to run the algorithm in this setting. The bounds on the worst-case regret shown in

Algorithm 3 Online Mirror Sinkhorn

Data: Initialise $\gamma_1 = \overline{\gamma}_1 = \mu\nu^\top$, define stepsize η_t, stream of loss functions (f_t).
for $1 \leq t \leq T$ **do**
  Update $\gamma'_{t+1} = \gamma_t \odot \exp\left(-\eta_t \nabla f_t(\gamma_t)\right)$,
  if *t is even* **then**
    Rows: $\gamma_{t+1} = \mathrm{Diag}\left(\mu \oslash (\gamma'_{t+1} 1_n)\right)\gamma'_{t+1}$,
  else
    Cols: $\gamma_{t+1} = \gamma'_{t+1}\,\mathrm{Diag}\left(\nu \oslash ((\gamma'_{t+1})^\top 1_m)\right)$,
  end
end
Output: γ_T

the following illustrate the claim that Mirror Sinkhorn is adaptive.

Theorem 3.1. *For $(f_t)_{t\geq 1}$ a sequence of convex functions that are B-Lipschitz for the norm $\|\cdot\|_1$, let δ be as defined in Equation (6), $\eta_t = \frac{1}{B}\sqrt{\frac{\delta}{t}}$. Then, the iterates $(\gamma_t)_{t\geq 1}$ of Mirror Sinkhorn as described in Algorithm 3 satisfy the*

following regret bound:

$$\max_{\gamma\in\mathcal{T}(\mu,\nu)} \sum_{t=1}^{T} (f_t(\gamma_t) - f_t(\gamma)) \leq \frac{9B}{8}\sqrt{\delta T}(2+\log(T)),$$

and the constraints satisfy

$$\sum_{t=1}^{T} c(\gamma_t) \leq \frac{3}{2}\sqrt{\delta T}(2+\log(T)),$$

with $c(\cdot)$ as defined in Equation (5).

Rounding at each t, $\tilde{\gamma}_t = \rho(\gamma_t)$ with Alg. 2, it holds that

$$\max_{\gamma\in\mathcal{T}(\mu,\nu)} \sum_{t=1}^{T} (f_t(\tilde{\gamma}_t) - f_t(\gamma)) \leq \frac{9B}{8}\sqrt{\delta T}(2+\log(T)).$$

3.2. Stochastic setting

The Mirror Sinkhorn algorithm can be used both with deterministic and stochastic gradient updates on the function f. In the first case, as described in Algorithm 2.2, the updates are given by $\nabla f(\gamma_t)$, and in the stochastic case by g_t satisfying $\mathbb{E}[g_t|\gamma_t] = \nabla f(\gamma_t)$. This minor adaptation is fully described in Algorithm 4 in the appendix for completeness.

This setting is common in stochastic optimization and allows this algorithm to be used in learning tasks where the function f is a data-dependent loss, and the g_t are gradients of this loss for one data observation (or a mini-batch thereof). This also illustrates that the algorithm is not sensitive to noisy gradients. The complexity of each step of the stochastic algorithm can be further reduced subsampling the gradient. To summarize, this setting can be motivated in several situations:

- Gradient subsampling on a random set of indices S,
$$g_t = \frac{mn}{|S|}\sum_{(i,j)\in S} (\nabla f(\gamma_t))_{ij} e_{ij}\,.$$
- Gradient observation with random additive noise
$$g_t = \nabla f(\gamma_t) + \sigma Z_t\,.$$
such that $\mathbb{E}[Z_t] = 0$ and $\mathbb{E}\left[\|Z_t\|_\infty^2\right] < \infty$.
- For OT problems, the objective function is $f : \gamma \mapsto \langle \gamma, C\rangle$, and its gradient $\nabla f(\gamma)$ given by the cost matrix C for all γ. In the 2-Wasserstein distance (Euclidean cost), it is equivalent to having $C_i j = -\langle x_i, y_j\rangle$. If we observe (X_i), (Y_j) two independent families of random vectors in $\mathbb{R}^d$ such that $\mathbb{E}[X_i] = x_i \in \mathbb{R}^d$, $\mathbb{E}[Y_j] = y_j \in \mathbb{R}^d$, we can use
$$(g_t)_{ij} = -\langle X_i, Y_j\rangle\,,$$
as an example of $g_t = C_t$ of stochastic observation of the cost in an optimal transport problem.

All the results presented here can be directly applied to the deterministic case by setting $\sigma = 0$. All convergence bounds are anytime, meaning that the number of iterations T is not known when choosing the step-size. As a consequence, there is an additional logarithmic term in the rate of convergence. This term can be removed if the stepsize is chosen to be constant in t, but dependent on T.

Theorem 3.2. *For f convex and B-Lipschitz for the norm $\|\cdot\|_1$, and*

$$\mathbb{E}\left[\|g_t - \nabla f(\gamma_t)\|_\infty^2 | \gamma_t\right] \leq \sigma^2,$$

let $B_\sigma = \sqrt{B^2+\sigma^2}$, $\eta_t = \frac{1}{B_\sigma}\sqrt{\frac{\delta}{t}}$. Then, taking $\overline{\gamma}_T$ to be the output after T steps of the Mirror Sinkhorn as described in Algorithm 4 and $\tilde{\gamma}_T = \rho(\overline{\gamma}_T)$, it holds that

$$\mathbb{E}\left[f(\tilde{\gamma}_T) - f(\gamma^*)\right] \leq \frac{B_\sigma}{2}\sqrt{\frac{\delta}{T}}(1+\log(T)).$$

With constant stepsize $\eta_t = \frac{1}{B_\sigma}\sqrt{\frac{2\delta}{T}}$ we have

$$\mathbb{E}\left[f(\tilde{\gamma}_T) - f(\gamma^*)\right] \leq B_\sigma\sqrt{\frac{2\delta}{T}}.$$

3.3. Optimal Transport

The Mirror Sinkhorn algorithm can be applied to the Optimal Transport problem (OT) described in (3), when f is a linear form. The Sinkhorn algorithm is widely used to tackle this problem, but suffers from some limitations in its most common form. First, it requires to have access to the exact cost matrix at the start of the algorithm. In contrast, as shown above, we can apply Mirror Sinkhorn to a stream of random cost matrices (stochastic gradients of the linear problem) (see Algorithm 4, 5 for details). Further, the Sinkhorn algorithm solves an entropic-regularized version of OT (Cuturi, 2013) and as such suffers from a regularization bias for any fixed $\alpha > 0$ - which must be set small enough, as a function of the desired precision $\varepsilon > 0$ for OT (Altschuler et al., 2017). As noted above, this modification of the problem can be considered "a feature rather than a bug", as for any fixed $\alpha > 0$, the new solution γ_α^* has enviable properties, such as differentiability. Such regularized objectives can also be tackled by Mirror Sinkhorn, as described in Section 3.4

Applying Mirror Sinkhorn either with $\nabla f(\gamma) = C$ (deterministic case) or $g_t = C_t$ (stochastic case, as in Algorithm 4), with successive random cost matrices C_t, allows to solve OT, with theoretical guarantees given in the following result, a corollary of Theorem 3.2 (see Algorithm 5 in the appendix for full details).

Theorem 3.3. *Let $(C_t)_{t\geq 1}$ be a sequence of random cost matrices such that $\mathbb{E}[C_t] = C$, for a cost matrix $C \in$*

$\mathbb{R}^{m\times n}$ *satisfying* $\|C\|_\infty \leq 1$, *and* $\mathbb{E}\left[\|C_t - C\|_\infty^2\right] \leq \sigma^2$. *Setting* $\eta_t = \sqrt{\frac{\delta}{(1+\sigma^2)t}}$ *in Mirror Sinkhorn as described in Algorithm 5, for the output* $\overline{\gamma}_T$ *after* T *steps and* $\tilde{\gamma}_T = \rho(\overline{\gamma}_T)$, *it holds that*

$$\mathbb{E}\left[\langle C, \tilde{\gamma}_T\rangle\right] - \langle C, \gamma^*\rangle \leq \frac{9}{8}\sqrt{\frac{(1+\sigma^2)\delta}{T}}(2+\log(T)).$$

Constant stepsize. We also provide the result for constant stepsize, which is asymptotically the same as the rate for Sink/Greenkhorn (Lin et al., 2019), with optimized constant.

Theorem 3.4. *Let* $(C_t)_{t\geq 1}$ *be a sequence of random cost matrices with* $\mathbb{E}\left[C_t\right] = C$, *for* $C \in \mathbb{R}^{m\times n}$ *that satsifies* $\|C\|_\infty \leq 1$, *and* $\mathbb{E}\left[\|C_t - C\|_\infty^2\right] \leq \sigma^2$. *For* $\varepsilon > 0$, *and* $\eta_t = \varepsilon\sqrt{\frac{\delta}{1+\sigma^2}}$. *Then, taking the output* $\overline{\gamma}_T$ *after* T *steps of Mirror Sinkhorn as described in Algorithm 5 and* $\tilde{\gamma}_T = \rho(\overline{\gamma}_T)$, *it holds for* $T \geq 5(1+\sigma^2)\delta\varepsilon^{-2}$ *that*

$$\mathbb{E}\left[\langle C, \rho(\tilde{\gamma}_T)\rangle\right] - \langle C, \gamma^*\rangle \leq \varepsilon\,.$$

3.4. Strong Convexity and Smoothness

We consider here the specific case where f is both ℓ-strongly convex and L-smooth w. r. t. the relative entropy, i.e.

$$\ell D_{\mathrm{KL}}(\gamma', \gamma) \leq f(\gamma') - f(\gamma) - \langle \nabla f(\gamma), \gamma' - \gamma\rangle$$

and

$$f(\gamma') - f(\gamma) - \langle \nabla f(\gamma), \gamma' - \gamma\rangle \leq L D_{\mathrm{KL}}(\gamma', \gamma)$$

Under such assumptions, the algorithm converges at a faster rate, which fits with the results on the convergence of the Sinkhorn algorithm for an entropy-regularized objective.

Theorem 3.5. *For* f *an* ℓ*-smooth and* L*-strongly convex with respect to the relative entropy, let* $\eta_t = \frac{1}{\ell t}$. *Then, taking* $\overline{\gamma}_T$ *the output of Mirror Sinkhorn (Algorithm 2.2) after* T *steps , it holds for* $B = \|\nabla f(\gamma^*)\|_\infty$ *that*

$$f(\overline{\gamma}_T) - f(\gamma^*) + 2Bc(\overline{\gamma}_T) \leq \frac{(2B+L)^2}{8\ell T}(1+\log(T)),$$

with $c(\gamma) = \|\gamma 1_m - \mu\|_1 + \|\gamma^\top 1_n - \nu\|_1$.

Note that by definition of the relative entropy as Bregman divergence of the negative entropy $\gamma \mapsto -H(\gamma)$, the latter is 1-strongly convex and 1-smooth with respect to the former. Thus, any f of the form

$$f(\gamma) = \langle C, \gamma\rangle - \alpha H(\gamma)$$

is L-smooth and ℓ-strongly convex with $L = \ell = \alpha$. Theorem 3.5 therefore applies to entropic-regularized OT.

3.5. Tensor Case

The Mirror Sinkhorn algorithm can be applied to a generalization involving probability tensors with multiple marginal constraints. Let $m_1, \ldots, m_d \geq 1$, be d positive integers. As noted in the definitions, the set of probability tensors (nonnegative tensors with entries summing to 1) is denoted by $\Delta_{m_1\times\cdots\times m_d}$. For $\mu_1 \in \Delta^{m_1}, \ldots, \mu_d \in \Delta^{m_d}$ probability vectors, the multiple transport polytope with marginals $\mu_1, \ldots, \mu_d$ is the set of tensors defined by

$$\mathcal{T}(\mu_1, \ldots, \mu_d) = \left\{\gamma \in \mathbb{R}_+^{m_1\times\cdots\times m_d} : \forall k,\, S_k(\gamma) = \mu_k\right\},$$

where $S_k(\gamma)$ is the sum of γ across all dimensions but k:

$$\forall 1 \leq i_k \leq m_k,\ (S_k(\gamma))_{i_k} = \sum_{j:j_k=i_k} \gamma_{j_1,\ldots,j_d}.$$

For a convex function $f : \Delta_{m_1\times\cdots\times m_d} \to \mathbb{R}$, the optimization problem described in equation (2) generalizes to

$$\min_{\gamma\in\mathcal{T}(\mu_1,\ldots,\mu_d)} f(\gamma).$$

The Mirror Sinkhorn algorithm can also be adapted to tackle these problems, choosing at each step the dimension along which there is the largest constraint variation to normalize (see Algorithm 6 in the Appendix for a full description).

Theorem 3.6. *For* $f : \Delta_{m_1\times\cdot\times m_d} \to \mathbb{R}$ *convex and* B*-Lipschitz with respect to the norm* $\|\cdot\|_1$, *and with* $\delta = \sum_{k=1}^d \|\log\mu_k\|_\infty$, $\eta_t = \frac{1}{B}\sqrt{\frac{\delta}{t}}$. *Taking* $\overline{\gamma}_T$ *the output after* T *steps of Algorithm 6, it holds that*

$$f(\overline{\gamma}_T) - f(\gamma^*) \leq \frac{9B}{8}\sqrt{\frac{\delta}{T}}\,(2+\log(T))\ ,$$

and the constraints violations are bounded as follows

$$\sum_{k=1}^d \|S_k(\overline{\gamma}_T) - \mu_k\|_1 \leq \frac{3}{2}\sqrt{\frac{\delta}{T}}\,(2+\log(T))\,.$$

4. Experiments

4.1. Optimal transport

We present the performance of Mirror Sinkhorn on a linear objective, i.e. the optimal transport problem

$$\min_{\gamma\in\mathcal{T}(\mu,\nu)} \langle C, \gamma\rangle\,.$$

We take $m, n = 100$, random $C \in \mathbb{R}^{m\times n}$ with independent off-diagonal coefficients in $[0, 1]$ and zero diagonal, $\mu = \nu$ random, so that the optimizer and the optimal value of the problem is known, and equal to 0 (with a different C for each run).

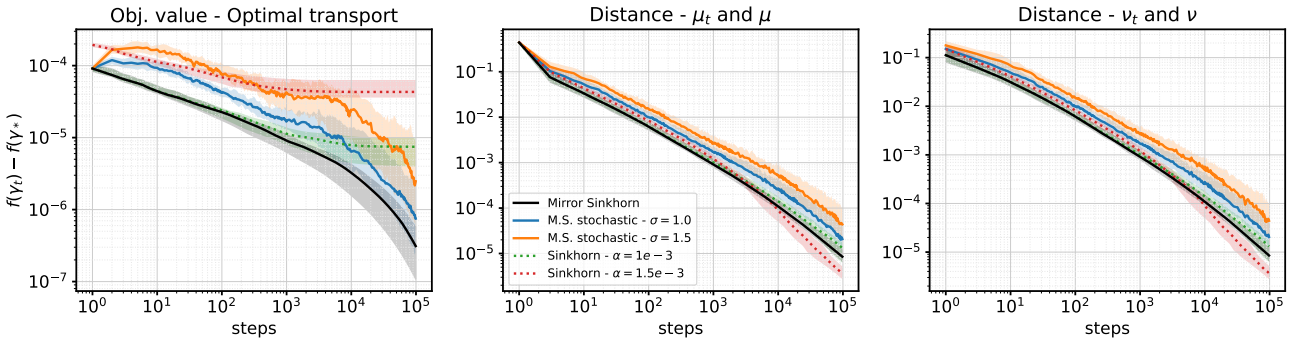

Figure 2. Performance of the Mirror Sinkhorn algorithm on the optimal transport problem, i.e. when f is linear. Shown are exact gradient oracles (black) and stochastic (blue, orange). The experiment is described in Section 4.1. Our proposed algorithm is compared to the Sinkhorn algorithm (red, green, dotted lines) with several values of regularization parameter. **Left:** The value of $f(\gamma_t)$ as a function of t - Note $f(\gamma^*) = 0$. **Center / Right:** The ℓ_1 distance between the two marginals (μ_t, ν_t) of γ_t and constraints (μ, ν) at any step.

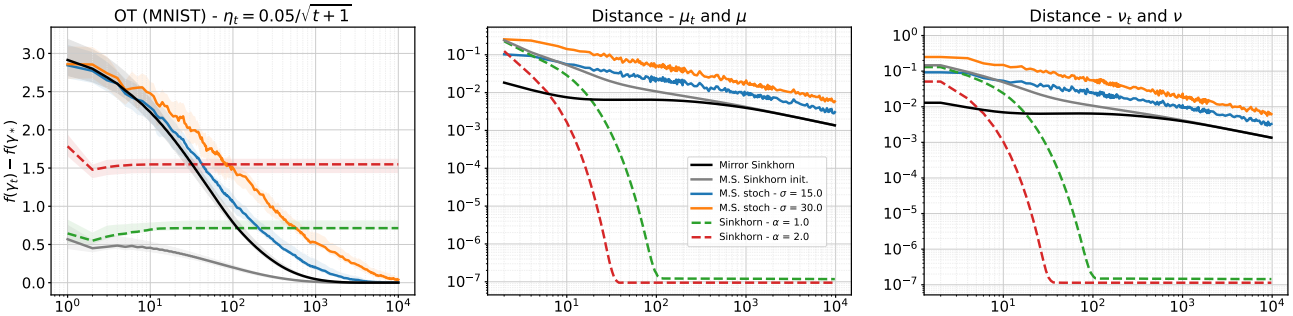

Figure 3. Performance of Mirror Sinkhorn on OT for MNIST data (Altschuler et al., 2017). See Figure 2 and Section 4.1 for details.

We compare the performance of our algorithm to that of Sinkhorn's algorithm for different values of regularization parameter $\alpha > 0$, running for $T = 10^5$ steps. The number of steps is chosen very large to illustrate the convergence rate, but this is not required to obtain a low optimization error, as shown in these experiments. We run this experiment both with $C_t = \nabla f(\gamma_t) = C$ (where the gradient is exact), and $C_t = C + \sigma Z_t$, where Z_t has i.i.d coefficients (stochastic gradients). These are reported in Figure 2, for 32 runs, with median and 10th-90th percentile. Our conclusion is that Mirror Sinkhorn is a fast and efficient algorithm to solve the optimal transport problem. In particular, it does not suffer from a regularization bias, and $\langle C, \gamma_t \rangle$ converges to its optimal value. There are two main advantages compared to using Sinkhorn for the optimal transport are: First, it is adaptive, there is no need to have a fixed desired precision, and to derive an corresponding regularization parameter. Second, it is an online algorithm, that can run on a stream of stochastic observations of C. This is not possible for some of the algorithms to either solve OT or its entropic regularized counterpart, including the Sinkhorn algorithm.

A consequence of the convergence of γ_t to γ^* is the slow progress for the two marginal constraints: as shown in our results, the violation in the constraint is polynomially decreasing in t, rather than the linear convergence (i.e. exponentially decreasing in t) of the Sinkhorn algorithm (Birkhoff, 1957; Carlier, 2022). This phenomenon is particularly visible when the entropic regularization parameter $\alpha > 0$ is higher: this yields solutions that are further from the boundary. The linear convergence driven by a constant quantifying the distance of γ^*_α from the boundary of $\mathcal{T}(\mu, \nu)$, which therefore increases with α. This is visible in Figure 2, for the two different values of α, highlighting the inherent trade-off between regularization bias and constraint violation.

We also include an illustration of our method on two datasets used in (Altschuler et al., 2017), following their experimental setup: we use as instances of OT random pairs from MNIST (10 in total), and simulated SQUARES data consisting of pairs of 20×20 images with a light square of random size on a dark background (also 10 in total). We report the results of this experiment, and the comparison of Sinkhorn with our algorithm in various setting, in Figure 3 (for MNIST) and Figure 7 (in Appendix for SQUARES), we also include for ease of comparison, Mirror Sinkhorn with the initialization of the Sinkhorn algorithm (in gray). We note that the well-understood slower convergence of our algorithm over the marginals is of a much smaller order than the gain in function objective.

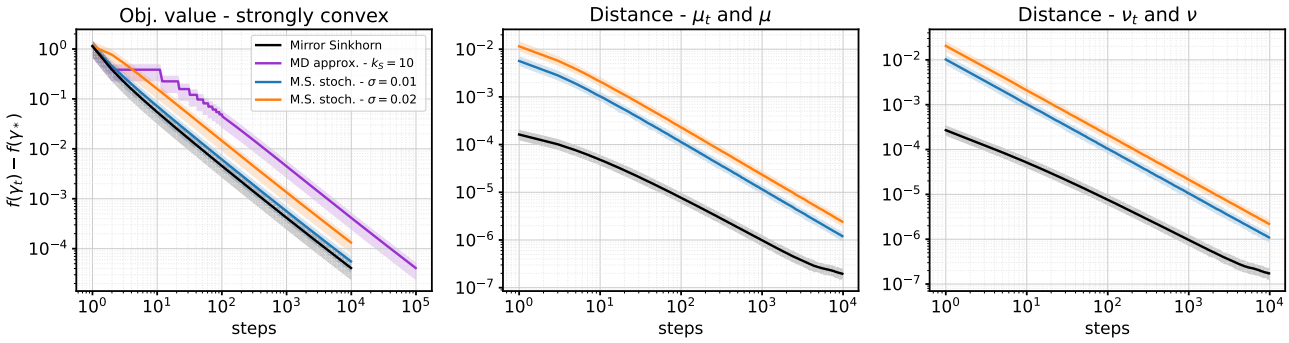

Figure 4. Performance of the Mirror Sinkhorn algorithm on strongly convex objectives, with exact gradient oracles (black) and stochastic gradients (blue and orange). The experiment is described in Section 4.2. This is compared with an approximation of mirror descent (purple) with $k_S = 10$ steps of Sinkhorn projection at each gradient update. **Left:** The value of $f(\gamma_t) - f(\gamma^*)$ as a function of t. **Center / Right:** The ℓ_1 distance between the marginals (μ_t, ν_t) of γ_t and constraints (μ, ν) before each normalization.

4.2. Strongly convex optimization

We consider minimization of a smooth and strongly convex objective f over $\mathcal{T}(\mu, \nu)$ for randomly chosen μ, ν of respective sizes m, n. We present an experiment on f which is a sum of several strongly convex objectives minimized at a common $\gamma^* \in \mathcal{T}(\mu, \nu)$ randomly chosen in its interior (with a different $\gamma*$ for each run). This allows to plot $f(\gamma_t) - f(\gamma^*)$ since the latter term is known. We take $m = 50$, $n = 60$ on this illustrative example, and run our algorithm on $T = 10^4$ gradient update steps (T is very large for illustration purposes), with a stepsize regime proportional to $1/(t+1)$. Our algorithm is evaluated both with exact and stochastic gradient updates. In the latter case, the stochastic gradients are derived from the exact gradient by adding independent noise, allowing the impact of gradient noise to be measured.

The results are represented in Figure 4 for 1024 independent runs, with median and 10th-90th percentile. They empirically confirm the speed of convergence of Theorem 3.5. Our method is compared to an approximation of mirror descent, using a nested loop for the proximal step. We use a modified Mirror Sinkhorn algorithm with k_S steps of alternating row/column normalization at each gradient update.

Here we take $k_S = 10$, and observe that surprisingly, the convergence is of similar order as that of Mirror Sinkhorn (which can be interpreted as having $k_S = 1$). Compared to the algorithm that we propose, using a nested-loop algorithm doing several steps of normalization to mimick mirror descent yields a significant slowdown, with a multiplicative factor on the number of algorithmic steps. Comparing the results of our algorithm with this approach by comparing instead at each gradient update, using multiple normalization steps did not yield significant improvement over our approach, even with higher value for k_S.

4.3. Point cloud registration - single cell data

We consider the problem of point cloud registration, i.e. mapping unmatched data sources, each composed of n data points X and Y, potentially in different dimensions d_X and d_Y, based on common geometry, related to Gromov-Wasserstein and similar problems (Mémoli, 2011; Djuric et al., 2012; Peyré et al., 2016; Solomon et al., 2016). In the case where $d_X = d_Y$, this problem can be formulated as *Wasserstein-Procrustes* (Grave et al., 2019) where a convex relaxation for the problem of assigning the n points of X to those of Y by finding a permutation of rows and columns that matches two Gram matrices K_X and K_Y of X and Y. This relaxation can be generalized to any setting where two such matrices (similarity, or distances) can be formulated. It is written as

$$\min_{\gamma \in \mathcal{T}(\mu,\nu)} \|K_X \gamma - \gamma K_Y\|_2^2 \,,$$

for μ, ν uniform on $[n]$. This quadratic problem can be regularized by adding a $-\lambda\|\gamma\|_2^2$ penalty. For $\lambda > 0$ small enough, the objective function remains convex on $\mathcal{T}_{\mu,\nu}$. It is in general possible to consider this problem to any two matrices K_X and K_Y that represent pairwise correlation or distance between two sets of points, even if they are in different spaces, as in the Gromov-Wasserstein problem. We apply our algorithm to this functional, on single-cell measurement data. In this experiment, X represents genetic expression and Y chromatin accessibility. The SNARE-seq data (Chen et al., 2019) consists of 1047 vectors in dimension 10 and 19 respectively. The ground-truth matching is known, allowing for interpretability of the result. Following the process in (Demetci et al., 2020), a k-NN connectivity graph is constructed from the correlations between data points. The matrix of pairwise shortest path distances on this graph between each pair of nodes is then used for K_X and K_Y. Normalization and maximal capping is applied.

We minimize this functional by taking $\lambda = 3$, with a k-

NN graph taken for $k = 5$. We recall that in this case, $n = 1,047$. We are applying a step-size regime proportional to $1/(t+1)$, for $T = 10^5$ steps. We analyze the relevance of this optimization problem to the motivating task of identifying the matching: we predict a binary matrix by thresholding γ_T at c/n for a small constant c, and compare it to the ground truth permutation matrix. This allows to count the number of predicted, and of correct matches by comparing to the known correspondence. The results for the objective value and the corresponding prediction accuracy are reported in Figure 5, and the distance between the marginals presented in Figure 6 in Appendix B.

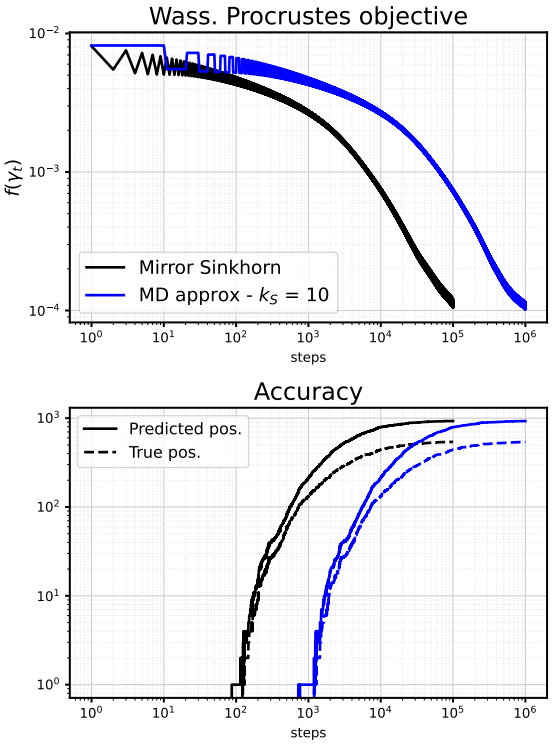

Figure 5. Performance of the Mirror Sinkhorn algorithm (black) on the Wasserstein Procrustes objective, for the experiment described in Section 4.3. This is compared with an approximation of mirror descent (blue) with $k_S = 10$ steps of Sinkhorn projection at each gradient update. **Top:** The value of $f(\gamma_t)$ as a function of t. **Bottom**: For the predicted assignment matrix, thresholded version of γ_t, the number (compared to $n = 1,047$) of predicted positives (solid line) and of true positives (dashed line) for both algorithms.

5. Concluding remarks

We introduce an algorithm for convex optimization problems on transport polytopes, providing theoretical guarantees and empirical evidence for its performance on a wide range of problems.

Several questions are left open and could be interesting directions for future research. In the case of the optimal transport problem, there are natural connections between our approach and annealing strategy for Sinkhorn (see, e.g. Peyré et al., 2019, Remark 4.9). This is related to the question of taking several steps of Sinkhorn to approximate mirror descent in our comparisons, rather than $k_S = 1$, for nonlinear convex objectives. We did not find in our experimental results any significant improvement when taking $k_S > 1$, but a significant computational overhead (due to the nested loops).

Other algorithmic approaches for OT and its entropic-regularized version have focused on optimization in the dual, as well as multimarginal, unbalanced, opr partial versions. They have a different perspective (and no online aspect) and are focused on the linear objective (Dvurechensky et al., 2018; Dvurechenskii et al., 2018; Lin et al., 2019; Le et al., 2022), and our results, in particular Theorem 3.4, can be seen in this larger context.

As noted above, our analysis of the tensor setting is important, as it allows to treat the problem of multi-marginal optimal transport (MOT) - for linear objective functions f, and more generally for us, of any convex objective on this set. One of the main applications of the MOT setting is that of Wasserstein barycenters, and our work allows us to cover both these problems and generalizations (Pass, 2015; Lin et al., 2022; Bigot et al., 2019).

Our results apply mostly on smooth (i.e. with Lipschitz gradients) convex functions, with additional results for strongly convex functions. Since results exist in the online learning literature on learning with Lipschitz losses (Hazan et al., 2007), it would also be a possible direction for research. Another interesting direction would be in nonconvex optimization, and the possiblity of approximating stationary points, on which there is an existing literature for mirror descent (Zhang & He, 2018). Finally, since our approach focuses on finite-dimensional iterates, semi-discrete approaches are not directly extendable to our setting. An extension in this direction, as considered by (Genevay et al., 2016), would also be possible.

Acknowledgements

MB was supported by the Cantab Capital Institute for Mathematics of Information during this work. The authors would like to thank Jonathan Niles-Weed for discussions on this problem, and Laetitia Meng-Papaxanthos for pointing them towards the SNARE-seq data.

References

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

Altschuler, J., Niles-Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pp. 1964–1974, 2017.

Alvarez-Melis, D., Jaakkola, T., and Jegelka, S. Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pp. 1771–1780. PMLR, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Aubin-Frankowski, P.-C., Korba, A., and Léger, F. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. *arXiv preprint arXiv:2206.08873*, 2022.

Baid, G., Cook, D. E., Shafin, K., Yun, T., Llinares-López, F., Berthet, Q., Belyaeva, A., Töpfer, A., Wenger, A. M., Rowell, W. J., et al. Deepconsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nature Biotechnology*, pp. 1–7, 2022.

Ballu, M., Berthet, Q., and Bach, F. Stochastic optimization for regularized wasserstein estimators. In *International Conference on Machine Learning*, pp. 602–612. PMLR, 2020.

Bassetti, F., Bodini, A., and Regazzini, E. On minimum kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302, 2006.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J.-P., and Bach, F. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020.

Bigot, J., Cazelles, E., and Papadakis, N. Data-driven regularization of wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755, 2019.

Birdal, T. and Simsekli, U. Probabilistic permutation synchronization using the riemannian structure of the birkhoff polytope. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11105–11116, 2019.

Birkhoff, G. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pp. 880–889. PMLR, 2018.

Blondel, M., Teboul, O., Berthet, Q., and Djolonga, J. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pp. 950–959. PMLR, 2020.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Boursier, E. and Perchet, V. Private learning and regularized optimal transport. *arXiv preprint arXiv:1905.11148*, 2019.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Carlier, G. On the linear convergence of the multimarginal sinkhorn algorithm. *SIAM Journal on Optimization*, 32 (2):786–794, 2022.

Carr, A. N., Berthet, Q., Blondel, M., Teboul, O., and Zeghidour, N. Self-supervised learning of audio representations from permutations with differentiable ranking. *IEEE Signal Processing Letters*, 28:708–712, 2021.

Chen, S., Lake, B. B., and Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12): 1452–1457, 2019.

Cordonnier, J.-B., Mahendran, A., Dosovitskiy, A., Weissenborn, D., Uszkoreit, J., and Unterthiner, T. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2351–2360, 2021.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Cuturi, M. and Blondel, M. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pp. 894–903. PMLR, 2017.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, 2020.

Djuric, N., Grbovic, M., and Vucetic, S. Convex kernelized sorting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 893–899, 2012.

Dubois-Taine, B., Bach, F., Berthet, Q., and Taylor, A. Fast stochastic composite minimization and an accelerated frank-wolfe algorithm under parallelization. *arXiv preprint arXiv:2205.12751*, 2022.

Dvurechenskii, P., Dvinskikh, D., Gasnikov, A., Uribe, C., and Nedich, A. Decentralize and randomize: Faster algorithm for wasserstein barycenters. *Advances in Neural Information Processing Systems*, 31, 2018.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International Conference on Machine Learning*, pp. 1367–1376, 2018.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pp. 3440–3448, 2016.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Grave, E., Joulin, A., and Berthet, Q. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1880–1890, 2019.

Guminov, S., Dvurechensky, P., Tupitsa, N., and Gasnikov, A. On a combination of alternating minimization and nesterov's momentum. In *International Conference on Machine Learning*, pp. 3886–3898. PMLR, 2021.

Guo, W., Ho, N., and Jordan, M. Fast algorithms for computational optimal transport and wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pp. 2088–2097. PMLR, 2020.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Kantorovich, L. V. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pp. 199–201, 1942.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Kumar, A., Brazil, G., and Liu, X. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8973–8983, 2021.

Le, K., Nguyen, H., Nguyen, K., Pham, T., and Ho, N. On multimarginal partial optimal transport: Equivalent forms and computational complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 4397–4413. PMLR, 2022.

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. Tree-sliced variants of wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Le Lidec, Q., Laptev, I., Schmid, C., and Carpentier, J. Differentiable rendering with perturbed optimizers. *Advances in Neural Information Processing Systems*, 34: 20398–20409, 2021.

Léger, F. A gradient descent perspective on sinkhorn. *Applied Mathematics & Optimization*, 84(2):1843–1855, 2021.

Li, X., Sun, D., and Toh, K.-C. On the efficient computation of a generalized jacobian of the projector over the birkhoff polytope. *Mathematical Programming*, 179(1):419–446, 2020.

Lin, T., Ho, N., and Jordan, M. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pp. 3982–3991. PMLR, 2019.

Lin, T., Ho, N., Cuturi, M., and Jordan, M. I. On the complexity of approximating multimarginal optimal transport. *Journal of Machine Learning Research*, 23(65): 1–43, 2022.

Linderman, S., Mena, G., Cooper, H., Paninski, L., and Cunningham, J. Reparameterizing the birkhoff polytope for variational permutation inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 1618–1627. PMLR, 2018.

Llinares-López, F., Berthet, Q., Blondel, M., Teboul, O., and Vert, J.-P. Deep embedding and alignment of protein sequences. *bioRxiv*, 2021.

Mémoli, F. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Mensch, A. and Peyré, G. Online sinkhorn: Optimal transport distances from sample streams. *Advances in Neural Information Processing Systems*, 33:1657–1667, 2020.

Mishchenko, K. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv preprint arXiv:1909.06918*, 2019.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Nadjahi, K., Durmus, A., Jacob, P. E., Badeau, R., and Simsekli, U. Fast approximation of the sliced-wasserstein distance using concentration of random projections. *Advances in Neural Information Processing Systems*, 34: 12411–12424, 2021.

Niles-Weed, J. and Rigollet, P. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4): 2663–2688, 2022.

Pass, B. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 49(6):1771–1790, 2015.

Paulus, M., Choi, D., Tarlow, D., Krause, A., and Maddison, C. J. Gradient estimation with stochastic softmax tricks. *Advances in Neural Information Processing Systems*, 33: 5691–5704, 2020.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pp. 2664–2672. PMLR, 2016.

Peyré, G., Cuturi, M., et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.

Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Vlastelica, M., Paulus, A., Musil, V., Martius, G., and Rolínek, M. Differentiation of blackbox combinatorial solvers. *arXiv preprint arXiv:1912.02175*, 2019.

Xie, Y., Wang, X., Wang, R., and Zha, H. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pp. 433–453. PMLR, 2020.

Zhang, S. and He, N. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

Appendix

A. Main proofs

Proof of Proposition 2.1. We first note

$$(\nabla f^{\text{reg}}(\gamma))_{ij} = (\nabla f^{\text{reg}}(\gamma))_{ij} + (\nabla R_\mu(\gamma 1_n))_i + (\nabla R_\nu(\gamma^\top 1_m))_j. \tag{7}$$

Let $t \geq 1$ be even, suppose that the iterate γ_t^{reg} obtained when replacing f by f^{reg} in the Algorithm 2.2 is identical to γ_t. Then

$$\begin{aligned}
(\gamma'_{t+1})^{\text{reg}}_{ij} &= (\gamma'_{t+1})_{ij} \exp\left(-\eta_t(\nabla R_\mu(\gamma_t 1_n))_i - \eta_t(\nabla R_\nu(\gamma_t^\top 1_m))_j\right), && (8)\\
&= (\gamma'_{t+1})_{ij} \exp\left(-\eta_t(\nabla R_\mu(\gamma_t 1_n))_i\right), && (9)\\
& && (10)
\end{aligned}$$

since $\gamma_t^\top 1_m = \nu$. The normalization of rows discards the gradient of the regularizer:

$$(\gamma'_{t+1} 1_m)^{\text{reg}}_i = (\gamma'_{t+1} 1_m)_i \exp\left(-\eta_t(\nabla R_\mu(\gamma_t 1_n))_i\right) \tag{11}$$

so

$$\begin{aligned}
(\gamma_{t+1})^{\text{reg}}_{ij} &= \frac{\mu_i}{(\gamma'_{t+1} 1_m)^{\text{reg}}_i} (\gamma'_{t+1})^{\text{reg}}_{ij}, && (12)\\
&= \frac{\mu_i}{(\gamma'_{t+1} 1_m)_i \exp\left(-\eta_t(\nabla R_\mu(\gamma_t 1_n))_i\right)} (\gamma'_{t+1})_{ij} \exp\left(-\eta_t(\nabla R_\mu(\gamma_t 1_n))_i\right), && (13)\\
&= \frac{\mu_i}{(\gamma'_{t+1} 1_m)_i} (\gamma'_{t+1})_{ij}, && (14)\\
&= (\gamma_{t+1})_{ij}. && (15)
\end{aligned}$$

The reasoning is the same for t odd, and the initialization is true $\gamma_1 = \mu\nu^\top = \gamma_1^{\text{reg}}$. So $\gamma_t = \gamma_t^{\text{reg}}$ for all $t \geq 1$. □

Theorem A.1. *If f is B-Lipschitz for the norm $\|\cdot\|_1$, let $\delta = \|\log \mu\|_\infty + \|\log \nu\|_\infty$ and let the stepsize be $\eta_t = \frac{1}{B}\sqrt{\frac{\delta}{t}}$ in the algorithm 2.2, then the output $\overline{\gamma}_T$ after T steps verifies*

$$f(\overline{\gamma}_T) - f(\gamma^*) \leq \frac{9B}{8}\sqrt{\frac{\delta}{T}}(2 + \log(T)), \tag{16}$$

and the constraints are close to be verified

$$c(\overline{\gamma}_T) \leq \frac{3}{2}\sqrt{\frac{\delta}{T}}(2 + \log(T)), \tag{17}$$

with

$$c(\overline{\gamma}_T) = \|\overline{\gamma}_T 1_n - \mu\|_1 + \|(\overline{\gamma}_T)^\top 1_m - \nu\|_1. \tag{18}$$

With constant step-size $\eta_t = \frac{1}{B}\sqrt{\frac{\delta}{T}}$ we have

$$f(\overline{\gamma}_T) - f(\gamma^*) \leq \frac{17B}{8}\sqrt{\frac{\delta}{T}} \tag{19}$$

and

$$c(\overline{\gamma}_T) \leq 2\sqrt{\frac{\delta}{T}}. \tag{20}$$

Proof. With Lemma A.3 and Pinsker's inequality:

$$D_{\text{KL}}(\gamma^*, \gamma_t) - D_{\text{KL}}(\gamma^*, \gamma_{t+1}) \geq 2\|\gamma_{t+1} - \gamma_t\|_1^2 + \eta_t \langle \nabla f(\gamma_t), \gamma_{t+1} - \gamma^* \rangle, \tag{21}$$

We use the fact that f is Lipschitz and convex:

$$f(\gamma_t) - f(\gamma^*) \leq \langle \nabla f(\gamma_t), \gamma_{t+1} - \gamma^* \rangle + B\|\gamma_{t+1} - \gamma_t\|_1 \tag{22}$$

then

$$D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) \geq \eta_t(f(\gamma_t) - f(\gamma^*)) + 2\|\gamma_{t+1} - \gamma_t\|_1^2 - \eta_t B\|\gamma_{t+1} - \gamma_t\|_1. \tag{23}$$

Given that $\gamma^* = \arg\min_{\gamma \in \mathcal{T}(\mu,\nu)} f(\gamma)$,

$$\begin{aligned} f(\gamma_t) - f(\gamma^*) &\geq f(\gamma_t) - f(\rho(\gamma_t)) && (24)\\ &\geq -B\|\gamma_t - \rho(\gamma_t)\|_1 && (25)\\ &\geq -2Bc(\gamma_t) && (26)\\ &\geq -2B\|\gamma_{t+1} - \gamma_t\|_1 && (27) \end{aligned}$$

by Lemma A.4 and Lemma A.5. So for $B' \geq 2B$

$$0 \leq \eta_t(f(\gamma_t) - f(\gamma^*) + B'c(\gamma_t)) \leq D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) + (B + B')\eta_t\|\gamma_{t+1} - \gamma_t\|_1 - 2\|\gamma_{t+1} - \gamma_t\|_1^2. \tag{28}$$

Moreover

$$0 \leq \eta_T(f(\gamma_t) - f(\gamma^*) + B'c(\gamma_t)) \leq \eta_t(f(\gamma_t) - f(\gamma^*) + B'c(\gamma_t)), \tag{29}$$

so

$$0 \leq \eta_T(f(\gamma_t) - f(\gamma^*) + B'c(\gamma_t)) \leq D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) + \frac{(B + B')^2}{8}\eta_t^2. \tag{30}$$

Thus with $B' = 2B$,

$$0 \leq \eta_t(f(\gamma_t) - f(\gamma^*) + Bc(\gamma_t)) \leq D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) + \frac{9B^2\eta_t^2}{8}. \tag{31}$$

We average over $1 \leq t \leq T$,

$$\frac{1}{T}\sum_{t=1}^{T}(f(\gamma_t) - f(\gamma^*) + Bc(\gamma_t)) \leq \frac{1}{T\eta_T}D_{\mathrm{KL}}(\gamma^*, \gamma_1) + \frac{9B^2}{8}\frac{\sum_{k=1}^{T}\eta_t^2}{T\eta_T}. \tag{32}$$

We remark $D_{\mathrm{KL}}(\gamma^*, \gamma_1) \leq \delta$. We use the convexity of c and f to extend the bound to the final iterate of the algorithm

$$f(\overline{\gamma}_T) - f(\gamma^*) + 2Bc(\overline{\gamma}_T) \leq \frac{\delta}{T\eta_T} + \frac{9B^2}{8}\frac{\sum_{k=1}^{T}\eta_t^2}{T\eta_T}. \tag{33}$$

We replace $\eta_t = \frac{1}{B}\sqrt{\frac{\delta}{t}}$

$$f(\overline{\gamma}_T) - f(\gamma^*) + 2Bc(\overline{\gamma}_T) \leq \frac{9B}{8}\sqrt{\frac{\delta}{T}}\left(\frac{8}{9} + 1 + \log(T)\right). \tag{34}$$

We also infer from (30), with $B' = 5B$:

$$3B\eta_T c(\gamma_t) \leq D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) + \frac{9}{2}B^2\eta_t^2. \tag{35}$$

We conclude by summing as well. □

Theorem A.2. *If f is B-Lipschitz for the norm $\|\cdot\|_1$, let $\delta = \|\log\mu\|_\infty + \|\log\nu\|_\infty$ and let the stepsize be $\eta_t = \frac{1}{B}\sqrt{\frac{\delta}{t}}$ in the algorithm 2.2, then the output $\overline{\gamma}_T$ after T steps verifies*

$$f(\rho(\overline{\gamma}_T)) - f(\gamma^*) \leq \frac{9B}{8}\sqrt{\frac{\delta}{T}}(2 + \log(T)). \tag{36}$$

With constant step-size $\eta_t = \frac{1}{B}\sqrt{\frac{\delta}{T}}$ we have

$$f(\rho(\overline{\gamma}_T)) - f(\gamma^*) \leq \frac{17B}{8}\sqrt{\frac{\delta}{T}}. \tag{37}$$

Proof. The proof follows from (34)

$$f(\overline{\gamma}_T) - f(\gamma^*) + 2Bc(\overline{\gamma}_T) \leq B\sqrt{\frac{\delta}{T}}\left(3 + \log(T)\right), \tag{38}$$

and

$$f(\overline{\gamma}_T) - f(\gamma^*) + 2Bc(\overline{\gamma}_T) \geq f(\rho(\overline{\gamma}_T)) - f(\gamma^*) \tag{39}$$

by Lemma A.5 and the fact that f is B-Lipschitz. □

Lemma A.3. *The iterates of algorithm 2.2 verify*

$$D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) = D_{\mathrm{KL}}(\gamma_{t+1}, \gamma_t) + \eta_t \langle \nabla f(\gamma_t), \gamma_{t+1} - \gamma^* \rangle, \tag{40}$$

Proof. We assume t is odd, and write for any $\gamma \in \Delta_{m\times n}$:

$$\begin{aligned} D_{\mathrm{KL}}(\gamma, \gamma_t) - D_{\mathrm{KL}}(\gamma, \gamma_{t+1}) &= \langle \gamma, \log\left(\gamma_{t+1} \oslash \gamma_t\right)\rangle, & (41)\\ &= \left\langle \gamma, \log\left(\gamma_{t+1} \oslash \gamma'_{t+1}\right)\right\rangle + \left\langle \gamma, \log\left(\gamma'_{t+1} \oslash \gamma_t\right)\right\rangle, & (42)\\ &= \left\langle \gamma, \log\left(\gamma_{t+1} \oslash \gamma'_{t+1}\right)\right\rangle - \eta_t \langle \nabla f(\gamma_t), \gamma\rangle, & (43) \end{aligned}$$

we remark that

$$\begin{aligned} \left\langle \gamma^*, \log\left(\gamma_{t+1} \oslash \gamma'_{t+1}\right)\right\rangle &= \sum_i \sum_j \gamma^*_{ij} \log\left(\frac{\mu_i}{\sum_k (\gamma'_{t+1})_{ik}}\right), & (44)\\ &= \left\langle \gamma^* 1_n, \log\left(\mu \oslash (\gamma'_{t+1} 1_n)\right)\right\rangle, & (45)\\ &= \left\langle \mu, \log\left(\mu \oslash (\gamma'_{t+1} 1_n)\right)\right\rangle, & (46)\\ & & (47) \end{aligned}$$

and similarly

$$\begin{aligned} \left\langle \mu, \log\left(\mu \oslash (\gamma'_{t+1} 1_n)\right)\right\rangle &= \left\langle \gamma_{t+1} 1_n, \log\left(\mu \oslash (\gamma'_{t+1} 1_n)\right)\right\rangle, & (48)\\ &= \sum_i \sum_j (\gamma_{t+1})_{ij} \log\left(\frac{\mu_i}{\sum_k (\gamma'_{t+1})_{ik}}\right), & (49)\\ &= \left\langle \gamma_{t+1}, \log\left(\gamma_{t+1} \oslash \gamma'_{t+1}\right)\right\rangle. & (50) \end{aligned}$$

Thus for $\gamma = \gamma^*$ in (41):

$$D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) = \left\langle \gamma_{t+1}, \log\left(\gamma_{t+1} \oslash \gamma'_{t+1}\right)\right\rangle - \eta_t \langle \nabla f(\gamma_t), \gamma^*\rangle. \tag{51}$$

Now we consider $\gamma = \gamma_{t+1}$ in (41):

$$D_{\mathrm{KL}}(\gamma_{t+1}, \gamma_t) = \left\langle \gamma_{t+1}, \log\left(\gamma_{t+1} \oslash \gamma'_{t+1}\right)\right\rangle - \eta_t \langle \nabla f(\gamma_t), \gamma_{t+1}\rangle \tag{52}$$

which allows to conclude

$$D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) = D_{\mathrm{KL}}(\gamma_{t+1}, \gamma_t) + \eta_t \langle \nabla f(\gamma_t), \gamma_{t+1} - \gamma^*\rangle. \tag{53}$$

The reasoning is identical for t even. □

Algorithm 4 Stochastic Mirror Sinkhorn

Data: Initialise $\gamma_1 = \overline{\gamma}_1 = \mu\nu^\top$, define stepsize η_t.
for $1 \leq t \leq T$ **do**
  Sample g_t such that $\mathbb{E}\left[g_t|\gamma_t\right] = \nabla f(\gamma_t)$,
  Update $\gamma'_{t+1} = \gamma_t \odot \exp\left(-\eta_t g_t\right)$,
  if *t is even* **then**
    Rows: $\gamma_{t+1} = \mathrm{Diag}\left(\mu \oslash (\gamma'_{t+1} 1_n)\right) \gamma'_{t+1}$,
  else
    Cols: $\gamma_{t+1} = \gamma'_{t+1} \mathrm{Diag}\left(\nu \oslash ((\gamma'_{t+1})^\top 1_m)\right)$,
  end
  Update $\overline{\gamma}_{t+1} = \frac{t}{t+1}\overline{\gamma}_t + \frac{1}{t+1}\gamma_{t+1}$
end
Output: $\overline{\gamma}_T = \frac{1}{T}\sum_{t=1}^T \gamma_t$

Lemma A.4. *The iterates of algorithm 2.2 verify*

$$D_{\mathrm{KL}}(\gamma_{t+1}, \gamma_t) \geq 2\|\gamma_{t+1} - \gamma_t\|_1^2 \geq 2\max\{c(\gamma_t), c(\gamma_{t+1})\}. \tag{54}$$

Proof. The first inequality comes from Pinsker, the second from Jensen on the function $\|\cdot\|_1$:

$$\|\gamma_{t+1} - \gamma_t\|_1 \geq \max\{\|\gamma_{t+1}1_n - \gamma_t 1_n\|_1, \|(\gamma_{t+1})^\top 1_m - (\gamma_t)^\top 1_m\|_1\}. \tag{55}$$

We assume t is odd, then

$$\|\gamma_{t+1}1_n - \gamma_t 1_n\|_1 = \|\gamma_{t+1}1_n - \mu\|_1 = c(\gamma_{t+1}) \tag{56}$$

and

$$\|(\gamma_{t+1})^\top 1_m - (\gamma_t)^\top 1_m\|_1 = \|\nu - (\gamma_t)^\top 1_m\|_1 = c(\gamma_t). \tag{57}$$

The proof is symmetrical for t even. □

Proposition A.5. *(Altschuler et al., 2017) For all $\gamma \in \mathbb{R}_+^{m\times n}$*

$$\|\rho(\gamma) - \gamma\|_1 \leq 2c(\gamma) = 2\left(\|\gamma 1_m - \mu\|_1 + \|\gamma^\top 1_n - \nu\|_1\right), \tag{58}$$

moreover $\rho(\gamma) \in \mathcal{T}(\mu,\nu)$ and the runtime of the algorithm is $O(mn)$.

Proposition A.6. *Let $\ell \geq 0$. If f is ℓ-strongly convex with regards to the relative entropy, for any $\gamma \in \mathbb{R}_+^{m\times n}$,*

$$f(\gamma) - f(\gamma^*) \geq \ell D_{\mathrm{KL}}(\gamma^*, \gamma) - 2\|f(\gamma^*)\|_\infty c(\gamma). \tag{59}$$

This is also true for $\ell = 0$. If f is L-smooth with regards to the relative entropy, then for any $\gamma \in \mathbb{R}_+^{m\times n}$,

$$f(\gamma) - f(\gamma^*) \leq L D_{\mathrm{KL}}(\gamma^*, \gamma) + 2\|f(\gamma^*)\|_\infty c(\gamma). \tag{60}$$

Proof. By strong convexity

$$f(\gamma) - f(\gamma^*) \geq \ell D_{\mathrm{KL}}(\gamma^*, \gamma) + \langle \nabla f(\gamma^*), \gamma - \gamma^* \rangle \tag{61}$$

Since γ^* is the minimum of f on $\mathcal{T}(\mu,\nu)$,

$$\langle \nabla f(\gamma^*), \gamma^* \rangle = \langle \nabla f(\gamma^*), \rho(\gamma) \rangle, \tag{62}$$

so

$$f(\gamma) - f(\gamma^*) \geq \ell D_{\mathrm{KL}}(\gamma^*, \gamma) - \|f(\gamma^*)\|_\infty \|\gamma - \rho(\gamma)\|_1 \tag{63}$$

and we conclude with Lemma A.5.

By smoothness,

$$f(\gamma) - f(\gamma^*) \leq L D_{\mathrm{KL}}(\gamma^*, \gamma) + \langle \nabla f(\gamma^*), \gamma - \gamma^* \rangle \tag{64}$$

so

$$f(\gamma) - f(\gamma^*) \leq L D_{\mathrm{KL}}(\gamma^*, \gamma) + \|f(\gamma^*)\|_\infty \|\gamma - \rho(\gamma)\|_1 \tag{65}$$

and we also conclude with Lemma A.5. □

A.1. Online Optimisation

Proof of Theorem 3.1. We follow the proof of Theorem A.1 up to (31)

$$0 \leq \eta_t(f_t(\gamma_t) - f_t + 2Bc(\gamma_t)) \leq D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) + \frac{9B^2\eta_t^2}{8}. \tag{66}$$

Then we sum over $1 \leq t \leq T$. Idem for the constraints. □

A.2. Stochastic Case

Proof of Theorem 3.2. We follow the proof of Theorem A.1 up to (31), where everything is true in expectations,

$$0 \leq \mathbb{E}\left[\eta_t(f(\gamma_t) - f + 2Bc(\gamma_t))\right] \leq \mathbb{E}\left[D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1})\right] + \frac{9B_\sigma^2\eta_t^2}{8}. \tag{67}$$

Then we note

$$f(\rho(\gamma_t)) - f(\gamma^*) \leq f(\gamma_t) - f + 2Bc(\gamma_t), \tag{68}$$

we sum over $1 \leq t \leq T$ and conclude with Jensen's inequality. □

A.3. Optimal Transport

Algorithm 5 Mirror Sinkhorn for Optimal Transport

Data: Initialise $\gamma_1 = \overline{\gamma}_1 = \mu\nu^\top$, define stepsize η_t, streaming stochastic cost matrices C_t.
for $1 \leq t \leq T$ **do**
  Update $\gamma'_{t+1} = \gamma_t \odot \exp\left(-\eta_t C_t\right)$,
  if *t is even* **then**
    Rows: $\gamma_{t+1} = \mathrm{Diag}\left(\mu \oslash (\gamma'_{t+1} 1_n)\right) \gamma'_{t+1}$,
  else
    Cols: $\gamma_{t+1} = \gamma'_{t+1} \mathrm{Diag}\left(\nu \oslash ((\gamma'_{t+1})^\top 1_m)\right)$,
  end
  Update $\overline{\gamma}_{t+1} = \frac{t}{t+1}\overline{\gamma}_t + \frac{1}{t+1}\gamma_{t+1}$
end
Output: $\overline{\gamma}_T = \frac{1}{T}\sum_{t=1}^T \gamma_t$

Theorem A.7. *Let $(C_t)_{t\geq 1}$ be a sequence of random cost matrices such that $\mathbb{E}\left[C_t\right] = C$ for a given cost matrix $C \in \mathbb{R}^{m\times n}$ that verifies $\|C\|_\infty \leq 1$, and $\mathbb{E}\left[\|C_t - C\|_\infty^2\right] \leq \sigma^2$. Let the stepsize be $\eta_t = \sqrt{\frac{\delta}{(1+\sigma^2)t}}$ in the algorithm 5, then the output $\overline{\gamma}_T$ after T steps verifies:*

$$\mathbb{E}\left[\langle C, \overline{\gamma}_T\rangle\right] - \langle C, \gamma^*\rangle \leq 2\sqrt{\frac{(1+\sigma^2)\delta}{T}}(1 + \log(T)), \tag{69}$$

and the constraints are close to be verified

$$\mathbb{E}\left[c(\overline{\gamma}_T)\right] \leq \sqrt{\frac{\delta}{T}}(2 + \log(T)), \tag{70}$$

with

$$c(\overline{\gamma}_T) = \|\overline{\gamma}_T 1_n - \mu\|_1 + \|(\overline{\gamma}_T)^\top 1_m - \nu\|_1. \tag{71}$$

Proof. This follows directly from Theorem A.1. □

Proof of Theorem 3.3. This follows from Theorem A.7 with the same proof as Theorem A.2. □

Proof of Theorem 3.4. This follows directly from the proofs of Theorem A.7 and of Theorem A.2. □

A.4. Strong Convexity

Proof of Theorem 3.5. By smoothness,

$$f(\gamma_{t+1}) - f(\gamma_t) \leq \langle \nabla f(\gamma_t), \gamma_{t+1} - \gamma_t \rangle + L D_{\mathrm{KL}}(\gamma_{t+1}, \gamma_t) \tag{72}$$

by strong convexity,

$$f(\gamma_t) - f(\gamma^*) \leq \langle \nabla f(\gamma_t), \gamma_t - \gamma^* \rangle - \ell D_{\mathrm{KL}}(\gamma^*, \gamma_t) \tag{73}$$

adding the two with Lemma A.3,

$$(1 - \ell\eta_t) D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) \geq (1 - L\eta_t) D_{\mathrm{KL}}(\gamma_{t+1}, \gamma_t) + \eta_t (f(\gamma_{t+1}) - f(\gamma^*)). \tag{74}$$

We reason as in the proof of Theorem A.1, assuming $\eta_t < 1/L$:

$$\eta_T \left(f(\gamma_{t+1}) - f(\gamma^*) + B' c(\gamma_{t+1}) \right) \leq (1 - \ell\eta_t) D_{\mathrm{KL}}(\gamma^*, \gamma_t) - D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) + \frac{(B')^2 \eta_t^2}{8(1 - L\eta_t)}. \tag{75}$$

Let $\eta_t = \frac{1}{\ell t}$, then

$$f(\gamma_{t+1}) - f(\gamma^*) + B' c(\gamma_{t+1}) \leq \ell(t-1) D_{\mathrm{KL}}(\gamma^*, \gamma_t) - \ell t D_{\mathrm{KL}}(\gamma^*, \gamma_{t+1}) + \frac{(B' + L)^2}{8\ell t}. \tag{76}$$

A sum up to T provides

$$\sum_{t=1}^{T} (f(\gamma_{t+1}) - f(\gamma^*) + B' \|\gamma_{t+1} - \gamma_t\|_1) + \ell T D_{\mathrm{KL}}(\gamma^*, \gamma_{T+1}) \leq \sum_{t=1}^{T} \frac{(B' + L)^2}{8\ell t}. \tag{77}$$

Finally we use Jensen and Lemma A.5 with $B' = 2B$. □

A.5. Tensor Problem

Algorithm 6 Tensor Mirror Sinkhorn

Data: Initialise $(\gamma_1)_{i_1 \dots i_d} = (\overline{\gamma}_1)_{i_1 \dots i_d} = \mu_{i_1} \cdots \mu_{i_d}$, define stepsize η_t.
for $1 \leq t \leq T$ **do**
  Update $\gamma'_{t+1} = \gamma_t \odot \exp\left(-\eta_t \nabla f(\gamma_t)\right)$,
  for $1 \leq k \leq d$ **do**
    Sum across all dimensions but k:

$$(S_k(\gamma'_{t+1}))_{i_k} = \sum_{j : j_k = i_k} (\gamma'_{t+1})_{j_1, \dots, j_d}.$$

    Constraint distance $c_k = D_{\mathrm{KL}}\left(\mu_k, S_k(\gamma'_{t+1})\right)$
  end
  Find furthest dimension $K = \arg\max_k c_k$,
  Normalize along dimension K:

$$(\gamma_{t+1})_{i_1 \cdots i_d} = \frac{(\mu_K)_{i_K}}{\left(S_K(\gamma'_{t+1})\right)_{i_K}} (\gamma'_{t+1})_{i_1 \cdots i_d}.$$

end
Output: $\overline{\gamma}_T = \frac{1}{T} \sum_{t=1}^{T} \gamma_t$

Proof of Theorem 3.6. The proof of Lemma A.3 is still valid for the tensor case. Lemma A.4 is modified with cyclical Bregman projections on the marginal spaces:

$$\sum_{t=1}^{T} D_{\mathrm{KL}}(\mu_{k(t)}, S_{k(t)}(\gamma_t)) \leq \delta + 2 \sum_{t=1}^{T} \eta_t \|\nabla f(\gamma_t)\|_\infty \tag{78}$$

with $k(t)$ the dimension chosen at iteration t. Then, since the choice of dimension to normalise is greedy,

$$\sum_{t=1}^{T} \max_{1 \leq k \leq d} D_{\mathrm{KL}}(\mu_k, S_k(\gamma_t)) \leq \delta + 2 \sum_{t=1}^{T} \eta_t \|\nabla f(\gamma_t)\|_\infty, \tag{79}$$

and we conclude with Jensen's inequality. □

B. Complementary experimental results

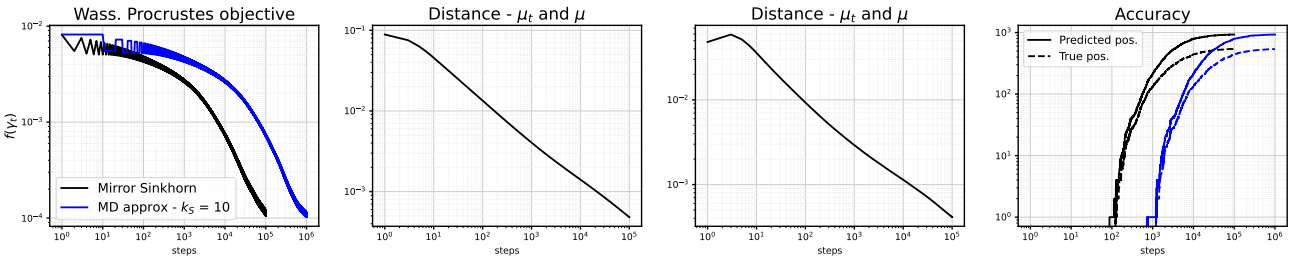

Figure 6. Performance of the Mirror Sinkhorn algorithm (black) on the Wasserstein Procrustes objective, for the experiment described in Section 4.3. This is compared with an approximation of mirror descent (blue) with $k_S = 10$ steps of Sinkhorn projection at each gradient update. **Left:** The value of $f(\gamma_t)$ as a function of t. **Centers:** The ℓ_1 distance between the two marginals (μ_t, ν_t) of γ_t and constraints (μ, ν) at any given time. **Right**: For the predicted assignment matrix, thresholded version of γ_t, the number (compared to $n = 1,047$) of predicted positives (solid line) and of true positives (dashed line) for both algorithms.

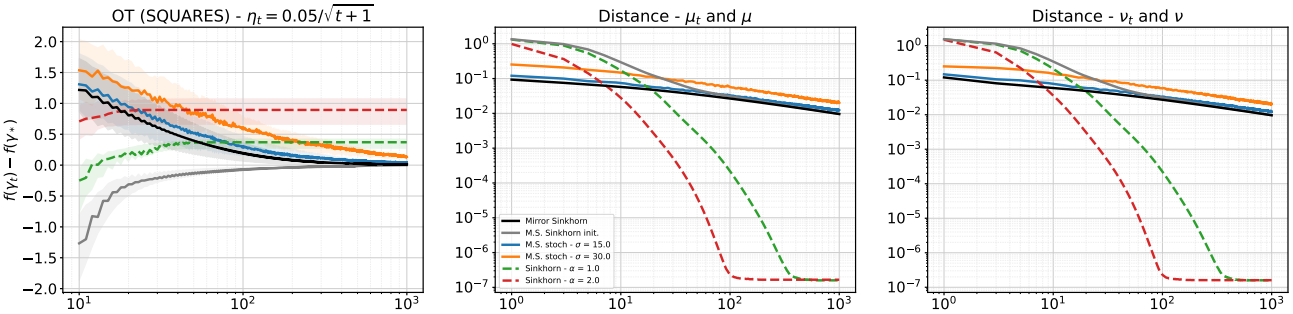

Figure 7. Performance of Mirror Sinkhorn on OT for SQUARES data (Altschuler et al., 2017). See Figure 2 and Section 4.1 for details.