Universitat Pompeu Fabra Barcelona

University Pompeu Fabra

Artificial Intelligence and Machine Learning

# A proximal formulation of regularized optimal transport

*Author:*
Joaquim Ortiz de Haro

*Supervisor:*
Gergely Neu

Research Report

Created: July 25, 2019
Last modified: February 4, 2020

# Contents

# 1 Reading guide

This report summarizes 5 months of work as a research assistant in the Artificial Intelligence and Machine Learning group at UPF (University Pompeu Fabra), under the supervision of Gergely Neu.

The goal of the project was to analyze the new algorithms that use entropic regularization to solve the optimal transport problem. This optimal transport formulation framework is becoming popular in Machine Learning as a tool to compute distances between probability distributions.

This document is divided in five parts:

- Section 2: an introduction with a brief description of optimal transport.
- Section 3: a review of the state of the art results and publications in this field. We also reproduce the more important proofs, to point out ideas and techniques that have an influence on our work.
- Section 4: we propose a new framework to analyze the proximal version of regularized optimal transport. Although this algorithm has a good practical performance, it comes without any guarantee or theoretical analysis. In this report, we present a new complexity analysis for the proximal formulation based on a follow the regularized leader setting and new bounds on the number of Sinkhorn iterations.
- Section 5: a short overview of other ideas that we have explored during this project. Again, the goal is to find new ideas and frameworks to do a tight analysis of the proximal version of regularized optimal transport.
- Section 6: a conclusion with a short recap of the contributions, highlighting the missing pieces and a discussion of future research directions.

# 2 Introduction

Computing distances between probability distributions is a key problem in statistical machine learning, with applications to supervised and unsupervised learning [6]. This distance can be defined in several ways. For example, the total variation distance, the Kullback Leibler divergence, the Jensen-Shannon divergence and the earth mover distance (also called Wasserstein-1).

From these alternatives, the earth mover distance provides a more powerful geometric description to compare probabilities and shows a better performance in machine learning tasks. However, computing the distance is more computationally expensive.

The earth mover distance is the cost of moving mass from one shape to another optimally. It is an instance of an optimal transport problem (with a particular choice of cost matrix) and requires solving a linear program, which is expensive in the large settings of machine learning. In fact, the cost of computing optimal transport distances scales at least in $O(n^3 log(n))$ when comparing two histograms of dimension $n$ [17]. As a comparison, the Kullback Leibler divergence between to histograms of dimension $n$ is computed in $O(n)$.

In this report, we will focus on the efficient computation of the optimal transport distance (which generalizes the earth mover distance).

We now formalize the optimal transport problem in statistics. We denote the simplex as $\Delta^n = \{x \in \mathbb{R}^n_+ : \sum x_i = 1\}$. The optimal transport for a given cost matrix $C \in \mathbb{R}^{n \times n}$ plan between two distributions $r \in \Delta^n$ and $c \in \Delta^n$ is a joint distribution $P \in \Delta^{n \times n}$ with marginals $r$ and $c$ with minimum cost $\sum C_{ij} P_{ij}$. Through all the document, we assume that marginals $r, c$ have the same dimension $r, c \in \Delta^n$ for simplicity, but results generalize for $r \in \Delta^n$ and $c \in \Delta^m$. Let $\mathbf{1} \in \mathbb{R}^n$ be a column vector with all entries equal to one.

The optimal transport problem $L_c(r, c)$ is:

$$L_c(r, c) \stackrel{\text{def}}{=} \min_{P \in \mathcal{U}(r,c)} \langle C, P \rangle = \min_{P \in \mathcal{U}(r,c)} \sum_{i,j} C_{ij} P_{ij} \tag{2.1}$$

$$\mathcal{U}(r, c) = \{P \in \mathbb{R}^{n \times n}_+ : \ P\mathbf{1} = r, \ P^T \mathbf{1} = c\} \tag{2.2}$$

Its dual problem is a direct consequence of general strong duality for linear programs.

$$L_c(r, c) = \max_{(f,g) \in R(C)} \langle f, r \rangle + \langle g, c \rangle \tag{2.3}$$

$$R(C) \stackrel{\text{def}}{=} \{(f, g) \in \mathbb{R}^n \times \mathbb{R}^n : f_i + g_j \leq C_{ij}, \ \forall (i, j) \in [n] \times [n]\} \tag{2.4}$$

Optimal transport defines a distance between probability measures as soon as the cost matrix fulfils some properties. In fact, the earth mover distance corresponds to $C_{ij} = d_{ij}$ with $d_{ij}$ a distance function. For example, with 1-d histograms $C_{ij} = |i - j|$. We refer the reader to [18] for an updated and extensive reference to optimal transport in the context of statistics.

# 3 Regularized optimal transport

A recent approach [9] to solve the optimal transport problem (2.1) is to regularize the original problem with the convex entropic term $R(P) = \sum P_{ij} \log P_{ij} = -H(P)$, where $H(P)$ is the entropy of the joint distribution. The regularized problem becomes a matrix scaling problem, and can be efficiently solved with the Sinkhorn algorithm [22]. The solution of the regularized problem $P_\eta$ is an approximate solution of the solution original linear problem $P^*$.

$$P^* = \underset{P \in \mathcal{U}(r,c)}{\arg\min} \langle C, P \rangle \tag{3.1}$$

$$P_\eta = \underset{P \in \mathcal{U}(r,c)}{\arg\min} \langle C, P \rangle + \eta R(P) \tag{3.2}$$

The solution of the regularized program converges to the solution of the original unregularized problem as $\eta$ tends to zero. The convergence rate is, at least, linear in the regularization parameter.

$$\langle C, P_\eta \rangle - \langle C, P^* \rangle \leq \eta (R(P^*) - R(P_\eta)) \leq \eta G \tag{3.3}$$

$$G = \max_{x,y \in \Delta^{n \times n}} R(x) - R(y) \tag{3.4}$$

Moreover, under certain conditions of the regularization parameter and the cost structure, the convergence is exponential [24].

This approximate solution offers a good trade off between computation and accuracy. Moreover, it also defines a better distance in some inference tasks.

The good performance of this approach, drew attention for new practical and theoretical work. On one hand, Wasserstein distance is now applied in supervised and unsupervised learning tasks. From a theoretical perspective, the computational complexity of solving optimal transport via Sinkhorn iteration has been analyzed providing theoretical guarantees. Moreover, alternative methods based on convex optimization algorithms (e.g. gradient and mirror descent) and variations of Sinkhorn algorithm have been proposed and analyzed. Apart from the papers explicitly cited in this report, we point the reader to [14], [4], [19], [7], [4], [3], [10], [16], [13] for an overview of recent ideas and approaches to solve and analyze the optimal transport problem.

We now give a short summary of the most important results in the complexity analysis of solving optimal transport via Sinkhorn iterations. The complete derivation of lemmas and proofs is found in [5], [12].

## 3.1 Approximate solution to optimal transport

In approximate optimal transport, our goal is to find a matrix $\hat{P} \in \mathcal{U}(r, c)$ such that

$$\langle \hat{P}, C \rangle \leq \min_{P \in \mathcal{U}(r,c)} \langle P, C \rangle + \epsilon \tag{3.5}$$

To find $\hat{P}$, we first solve the regularized program via the Sinkhorn algorithm

$$\min_{P \in \mathcal{U}(r,c)} \langle C, P \rangle + \eta R(P) \tag{3.6}$$

The Sinkhorn algorithm is an iterative algorithm, see algorithm 4, where columns and rows are iteratively rescaled until the desired level of constraint violation is reached.

The input for Sinkhorn algorithm is the matrix $K$ with $k_{ij} = e^{-C_{ij}/\eta}$. Let $P_s$ be the output of the algorithm, which fulfills:

$$\|P_s \mathbf{1} - r\|_1 + \left\|P_s^T \mathbf{1} - c\right\|_1 \leq \varepsilon' \tag{3.7}$$

Finally, $\hat{P}$ is obtained by rounding $P_s$ to $U(r, c)$ with algorithm 2.

**Algorithm 1** Sinkhorn's algorithm

**Input:** $K, \varepsilon',\ u_0,\ v_0$.
let $B(u, v) = \mathrm{Diag}(u)\ K\ \mathrm{Diag}(v)$
**repeat**
  **if** $k \bmod 2 = 0$ **then**
    $u_{k+1} = u_k + \ln r - \ln\left(B\left(u_k, v_k\right) \mathbf{1}\right)$
  **else**
    $v_{k+1} = v_k + \ln c - \ln\left(B\left(u_k, v_k\right)^T \mathbf{1}\right)$
  **end if**
**until** $\|B\left(u_k, v_k\right) \mathbf{1} - r\|_1 + \|B\left(u_k, v_k\right)^T \mathbf{1} - c\|_1 \leq \varepsilon'$
**Output:** $P_s = B(u_k, v_k)$

**Algorithm 2** Rounding

**Input:** $F,\ \mathcal{U}(r, c)$
$X = \mathrm{Diag}(x)$ with $x_i = \min(r_i / r_i(F), 1)$
$F' = XF$
$Y = \mathrm{Diag}(y)$ with $y_j = \min(c_j / c_j(F'), 1)$
$F'' = F'Y$
$err_r = r - r(F'')$
$err_c = c - c(F'')$
$G = F'' + err_r\ err_c^T / \|err_r\|_1$
**Output:** $\hat{P} = G$

## 3.2 Lagrange dual problem and Sinkhorn algorithm

The dual problem of (3.6) plays a central role in the analysis of the complexity of Sinkhorn algorithm.

The Lagrangian of the regularized problem (3.6) is:

$$\mathcal{L}(P, \alpha, \beta) = \langle \alpha, r \rangle + \langle \beta, l \rangle + \langle C, P \rangle + \eta R(P) - \langle \alpha, P\mathbf{1} \rangle - \left\langle \beta, P^\top \mathbf{1} \right\rangle \tag{3.8}$$

Setting the gradient w.r.t. $P_{ij}$ to zero we find the structure of the solution.

$$C_{ij} + \eta \left(1 + \log \left(P_{ij}\right)\right) - \alpha_i - \beta_j = 0, \quad \forall i, j \in [n] \tag{3.9}$$

$$P_{ij} = e^{\frac{-C_{ij} + \alpha_i + \beta_j}{\eta} - 1}, \quad \forall i, j \in [n] \tag{3.10}$$

To simplify notation, we perform the change of variables.

$$u_i = \frac{\alpha_i}{\eta} - \frac{1}{2} \ , \ v_j = \frac{\beta_j}{\eta} - \frac{1}{2} \tag{3.11}$$

And solving $\max_{\alpha,\beta \in \mathbb{R}^n} \min_{P \in \mathbb{R}^{n \times n}} \mathcal{L}(P, \alpha, \beta)$ is equivalent to:

$$\min_{u,v \in \mathbb{R}^n} f(u, v) := \mathbf{1}^\top B(u, v)\mathbf{1} - \langle u, r \rangle - \langle v, l \rangle \tag{3.12}$$

Where $B(u, v) := \operatorname{diag}\left(e^u\right) e^{-\frac{C}{\eta}} \operatorname{diag}\left(e^v\right)$. Strong duality holds because the original problem is a convex optimization problem (with convex cost and affine constraints). The function $f(u, v)$ is used as a potential to analyze the behaviour and convergence of Sinkhorn algorithm. In fact, the iterations are a block coordinate descent on this potential.

## 3.3 Optimal transport via Sinkhorn iterations: Computational complexity

The goal of this section is to give a quick overview on the techniques and arguments used to bound the computational complexity, presented in [5], [12]. In particular, we bound the computational complexity to get an $\epsilon$-approx solution, see (3.5), in a problem with size $n$ (i.e $P \in \Delta^{n \times n}$ and $r, c \in \Delta^n$). The analysis also sets the values of the parameters $\eta$ and $\varepsilon'$ as a function of $\epsilon$ and $n$.

We start computing how many iterations of Sinkhorn algorithm we need to reach $\varepsilon'$ marginal violation. For this, we use the dual of the regularized problem as a potential function.

**Theorem 3.1.** *[12] Algorithm 4 outputs a matrix $B(u_k, v_k)$ satisfying $\|B(u_k, v_k)\mathbf{1} - r\|_1 + \|B(u_k, v_k)^T \mathbf{1} - c\|_1 \leq \varepsilon'$ in the number of iterations $k$ satisfying $k \leq 2 + \frac{4R}{\varepsilon'}$*

This constant $R$ plays a central role in the analysis and is defined as:

$$R = \max_i(u_i^0 - u_i^*) + \min_i(u_i^0 - u_i^*) \tag{3.13}$$

$$R \leq -\ln(e^{-\|C\|_\infty/\eta} \min_{ij}\{r_i, c_j\}) \tag{3.14}$$

*Sketch of the proof*: the key is to find a lower bound for the improvement of potential in two consecutive Sinkhorn iterations $\psi(u_k, v_k) - \psi(u_{k+1}, v_{k+1})$, where $\psi = f$, see (3.12).

$$\begin{aligned} &\psi(u_k, v_k) - \psi(u_{k+1}, v_{k+1}) = KL(r\|B_k\mathbf{1}) \\ &\geq \tfrac{1}{2}\|B_k\mathbf{1} - r\|_1^2 \geq \max\left\{\widetilde{\psi}(u_k, v_k)^2/(2R^2), (\varepsilon')^2/2\right\} \end{aligned} \tag{3.15}$$

Where $\widetilde{\psi}(u_k, v_k) = \psi(u_k, v_k) - \psi(u^*, v^*)$. The two estimates inside the max operator are combined to conclude that:

$$k \leq 2 + 4R/\varepsilon' \tag{3.16}$$

□

The output of the Sinkhorn algorithm $P_s$ is then rounded to $\mathcal{U}(r, c)$ using algorithm 2 and we obtain the approximate solution $\hat{P}$ for the original problem. We now analyze the computational complexity and compute the values for the regularization parameter $\eta$ and marginal violation $\varepsilon'$.

We highlight three key ideas:

– Algorithm 2 rounds a matrix $A \in \mathcal{U}(a_1, a_2)$ to $B \in \mathcal{U}(b_1, b_2)$ such that:

$$\|A - B\|_1 \leq 2(\|a_1 - b_1\|_1 + \|a_2 - b_2\|_1) \tag{3.17}$$

– The Sinkhorn algorithm outputs $\hat{P} \in \mathcal{U}(r', c')$ such that $\|r - r'\| + \|c - c'\| < \varepsilon'$. Note that, because of the structure of the matrix, this is the optimal solution for the regularized optimal transport problem with the same cost matrix but modified marginals $\mathcal{U}(r', c')$.

– $-2\log n \leq R(X) \leq 0$, as $P \in \Delta^{n\times n}$

These ideas are combined to get the bound:

$$\langle \hat{P}, C\rangle \leq \min_{P\in\mathcal{U}_{r,c}} \langle P, C\rangle + \eta 2\log n + 4\varepsilon'\|C\|_\infty \tag{3.18}$$

We set $\eta = \epsilon/(4\log n)$ and $\varepsilon' = \epsilon/(8\|C\|_\infty)$ to obtain an $\epsilon$ approximate solution.

Substituting the value of $\eta$ and $\varepsilon$ into (3.16), and considering that each iteration takes $O(n^2)$ time, we find the computational complexity:

$$\widetilde{O}\left(\frac{n^2}{\epsilon^2}\right) \tag{3.19}$$

Where $\widetilde{O}$ hides logarithmic factors. The main drawback of this approach is the dependence on $\frac{1}{\epsilon^2}$. This is also observed in practice, where the number of Sinkhorn iterations strongly increases with small regularization parameters. Moreover, numerical precision errors have a negative impact on the performance and accuracy.

## 3.4 Proximal formulation

### 3.4.1 $\epsilon$-scaling

A well known heuristic to solve matrix scaling with Sinkhorn when the regularization is very small, i.e $K = e^{-C/\eta},\ \eta \to 0$ is to solve a sequence of subproblems, starting with a "big" $\eta$ and decreasing $\eta_t \geq \eta_{t+1}$ until the desired $\eta$ (reusing the scaling factors from the previous subproblem). This is known as $\epsilon$-scaling.

In practice, each subproblem is an instance of a proximal regularized optimal transport. Let $r_t = \eta_t/\eta_{t+1}$. These are equivalent formulations (see Lemma 4.7):

$$P_t = \underset{P \in \mathcal{U}(r,c)}{\arg\min} \langle C, P \rangle + \eta_t R(X) \tag{3.20}$$

$$P_t = \underset{P \in U(r,c)}{\arg\min} \langle C, P \rangle + r_t D(P, P_{t-1}) \tag{3.21}$$

$$P_t = \underset{P \in \mathcal{U}(r,c)}{\arg\min} \langle C - r_t \ln P_{t-1}, P \rangle + r_t R(P) \tag{3.22}$$

Where $D(P, P_{t-1}) = \sum P_{ij} \log P_{ij}/P^t_{ij}$ is the Kullback Divergence.

Note that (3.21) and (3.22) can also be solved with the Sinkhorn algorithm. Now, the kernel matrix is $P_{t-1} \cdot e^{-C/r_t}$ instead of just $e^{-C/\eta}$.

Although, this approach has better practical performance than solving directly only one problem with small $\eta$, there are not theoretical guarantees on the convergence, the convergence rate or the computation complexity. Moreover, there is no theoretical guidance on how to design an optimal sequence $\{\eta_t\}$ or to choose a precision for the subproblems solutions (in terms of marginal violation).

$\epsilon$- scaling for optimal transport has been studied in [20], with good practical performance but still without convergence guarantees. One of the ideas is that the dual variables $\alpha_t, \beta_t$ of (3.20) converge to $\alpha^*, \beta^*$, the optimal dual variables of the original unregularized problem. In $\epsilon$-scaling, this observation is used to "warmstart" the scaling factors from the values of the previous subproblem.

The main challenge in the analysis of $\epsilon$-scaling is that the subproblems are not solved exactly, as the Sinkhorn algorithm is an iterative algorithm.

### 3.4.2 Inexact mirror descent

In [23], the authors analyze the proximal algorithm inside the framework of inexact mirror descent for a linear function. The approximate solution of the subproblem is computed in terms of a bound on first order optimality. We reproduce here the most relevant results:

The goal is to minimize $c(P) = \langle C, P \rangle$ by solving a sequence of subproblems to $\delta$ first order optimality:

$$P_{k+1} = \underset{P \in \mathcal{U}(r,c)}{\overset{\delta}{\arg\min}} \{c(P) + \eta D(P, P_k)\} \tag{3.23}$$

First order optimality means:

$$\max_{P \in \mathcal{U}(r,c)} \langle \nabla l(P_{k+1}), P_{k+1} - P \rangle \leq \delta \tag{3.24}$$

with $l(P) = c(P) + \eta D(P, P_k)$

**Lemma 3.2.** *let $\overline{P} = \frac{1}{N} \sum P_k$, with $N = 2\eta D(P^*, P_1)/\epsilon$ iterations and $\delta \leq \frac{\epsilon}{2}$ precision. Then:*

$$\left\langle C, \overline{P} - P^* \right\rangle \leq \epsilon \tag{3.25}$$

*Proof:* Let $P_{k+1}$ be a solution with $\delta$ precision. This means that, $\forall P$ :

$$\langle C + \nabla D(P_{k+1}, P_k), P - P_{k+1} \rangle \geq -\delta \tag{3.26}$$

In particular,

$$\langle C + \nabla D(P_{k+1}, P_k), P^* - P_{k+1} \rangle \geq -\delta \tag{3.27}$$

Based on a basic inequality in mirror descent, see for example [25]:

$$c(P^*) + \eta D(P^*, P_k) \geq c(P_{k+1}) + \eta D(P_{k+1}, P_k) + \eta D(P^*, P_{k+1}) - \delta \tag{3.28}$$

With $D(P_{k+1}, P_k) \geq 0$:

$$c(P_{k+1}) - c(P^*) \leq \eta D(P^*, P_k) - \eta D(P^*, P_{k+1}) + \delta \tag{3.29}$$

Telescoping the divergence terms and denoting $\overline{P} = \frac{1}{T} \sum P_t$

$$c(\overline{P}) - c(P^*) \leq \frac{\eta \ D(P^*, P_1)}{N} + \delta \tag{3.30}$$

We are interested in a $\epsilon$ approx solution.

$$\frac{\eta \ D(P^*, P_1)}{N} + \delta \leq \epsilon \tag{3.31}$$

We can use $\delta \leq \frac{\epsilon}{2}$ and $N \geq \frac{2\eta D(P^*, P_1)}{\epsilon}$. □

Now we have to tune $\eta$, taking into account that it will influence on both: the number of mirror descent iterations and the number of Sinkhorn iterations that we need to solve each subproblem. In [23], they use $\eta = O(\|C\|_\infty)$ and also propose an adaptative scheme.

The next step is to relate first order optimality bound $\delta$ to the marginal violation $\varepsilon'$ of Sinkhorn stopping criteria. Overall, the complexity bound of this method is: $O(n^4/\epsilon^2)$. However, in practice it has a better performance than solving just one subproblem with a small regularization parameter.

# 4 Proximal optimal transport and Sinkhorn iterations as follow the regularized leader

## 4.1 Motivation and algorithm

The $\epsilon$-scaling heuristic consists on solving a sequence of proximal regularized optimal transport problems. In each round, the regularization is decreased and the new subproblem is solved (using last solution as an initial guess). Here, we would like to point out the strong connection of the $\epsilon$-scaling iterations with the follow the regularized leader (FTRL) algorithm.

FTRL is an online learning optimization algorithm, where the relative weights of the regularization term decreases through the rounds. In our particular setting, the linear function $f_t(P) = \langle C, P \rangle$ is constant.

Therefore, we propose to use the FTRL framework to do a theoretic analysis of a proximal optimal transport algorithm via Sinkhorn iterations. In practice, this algorithm is a formalization of the $\epsilon$-scaling heuristic.

The main advantage of this approach is that suboptimality of the subproblems solution can be measured directly as a gap in the cost function. One of the drawbacks of a previous proximal analysis based on mirror descent [23] (using first order optimality) is the requirement to bound the minimal entries of $P_{ij}$. However, the authors point out that this bound is not needed in practice.

We propose two variants of the algorithm. In this section $\arg\min^{\delta}$ denotes additive suboptimality in the cost function. In algorithm 3 each subproblem is solved in the relaxed feasible convex set $\tilde{\mathcal{U}}(r, c, \tilde{\epsilon})$ using the Sinkhorn algorithm and the rounding to $\mathcal{U}(r, c)$ is computed only once in the end.

$$\tilde{\mathcal{U}}(r, c, \tilde{\epsilon}) = \{P \in \mathbb{R}_+^{n \times n} : \ \|P\mathbf{1} - r\| \leq \tilde{\epsilon}, \ \left\|P^T\mathbf{1} - c\right\| \leq \tilde{\epsilon}\} \tag{4.1}$$

In algorithm 4 each subproblem is solved in $\mathcal{U}(r, c)$, which requires the Sinkhorn algorithm and a rounding step for each subproblem.

In the following section, we will analyze algorithm 4 for simplicity, although the arguments and proofs could also be extended to algorithm 3. A key idea is that the error introduced by this rounding step is linear on the marginal violation, so that it can be easily controlled and bounded.

**Algorithm 3** Optimal Transport FTRL algorithm, Only One Rounding

**Input:** $C, r, c, \delta, \tilde{\epsilon}$
**for** $t = 1, \ldots, T$ **do**
  $P_t = \arg\min^{\delta} \langle C, P \rangle + \frac{\eta}{t} H(P)$, s.t $P_t \in \tilde{\mathcal{U}}(r, c, \tilde{\epsilon})$
**end for**
$\overline{P} = \frac{1}{T} \sum P_t$
$\hat{P} = \text{Round}(\overline{P})$ , see algorithm 2
**Output:** $\hat{P}$

**Algorithm 4** Optimal Transport FTRL algorithm, Rounding each step

**Input:** $C, r, c$
**for** $t = 1, \ldots, T$ **do**
  $P_t = \arg\min^{\delta} \langle C, P \rangle + \frac{\eta}{t} H(P)$ s.t $P_t \in \mathcal{U}(r, c)$
**end for**
$\overline{P} = \frac{1}{T} \sum P_t$
$\hat{P} = \overline{P}$
**Output:** $\hat{P}$

## 4.2 Analysis overview

Our goal is to find which is the computational complexity to achieve a $\epsilon$-approximate solution, i.e. finding $P_\epsilon \in \mathcal{U}(r, c)$ such that:

$$\langle C, P_\epsilon \rangle - \langle C, P^* \rangle \leq \epsilon \tag{4.2}$$

$$P^* = \underset{P \in \mathcal{U}(r,c)}{\arg\min} \langle C, P \rangle \tag{4.3}$$

We set the approximate solution $P_\epsilon$ to be the average of the matrices computed by the proximal algorithm.

$$P_\epsilon = \overline{P}_T = \frac{1}{T} \sum_t P_t \tag{4.4}$$

Our new analysis has two important components:

- Bound on $\left\langle \overline{P}_T - P^*, C \right\rangle$ as function of the number of rounds $T$. We use an inexact follow the regularized leader framework.
- Relation between the inaccuracy solution of each subproblem and the Sinkhorn stopping criteria based on $L1$-norm of the marginal violation $\|P\mathbf{1} - r\|_1 + \left\|P^T\mathbf{1} - c\right\|_1$
- Bound on the number of Sinkhorn iterations $K_t$ necessary for solving each proximal subproblem (each $\epsilon$-scaling iteration). The number of Sinkhorn iterations will depend on the desired accuracy of the inexact solutions of the subproblems.

Each Sinkhorn iteration is computed in $O(n^2)$ computational time, with $n$ the size of the marginal vectors, i.e. $r, c \in \mathbb{R}^n$. Therefore, the computational time of the whole algorithm is:

$$n^2 \sum K_t \leq n^2 \times T \times K \tag{4.5}$$

where $K$ is an upper bound on the number of Sinkhorn iterations: $K_t \leq K \ \forall t$.

## 4.3 Bound on the regret with FTRL

In each iteration $t \in [1, \dots, T]$ of algorithm 4 , we are solving the subproblem (4.6). As pointed before, this is an instance of FTRL with a constant linear function $f_t(P) = \langle C, P \rangle$

$$\min_{P \in \mathcal{U}(r,c)} t \langle C, P \rangle + \eta R(P) \tag{4.6}$$

In fact, each iteration will be solved approximately, so that $P_{t+1}$ fulfils (4.7), with Sinkhorn iterations and a rounding step. We now analyze the relation of the error of the subproblems $\delta_t$ with the convergence rate and regret bound.

$$\langle C, P_{t+1} \rangle + \frac{\eta}{t} R(P_{t+1}) \leq \min_{P \in \mathcal{U}(r,c)} \langle C, P \rangle + \frac{\eta}{t} R(P) + \delta_t \tag{4.7}$$

One of the advantages of the follow the regularized leader approach is that the quality of the approximate solution is controlled with the additive error(i.e: $f \leq f^* + \delta$). The additive error is a natural measure of optimality in the approximated regularized optimal transport literature, and is directly related to the number of Sinkhorn iterations.

Here we present two different ways of analyzing the approximate follow the regularized leader for optimal transport. As we get two different upper bounds, we bound the regret as the minimum of the two different regrets. We define the regret as:

$$\text{Regret}_T = \sum_t^T \langle C, P_t - P^* \rangle = T \left\langle C, \overline{P}_T - P^* \right\rangle \tag{4.8}$$

Finally, note that the average regret $\text{Regret}/T$ is an upper bound to the convergence bound of $\left\langle C, \overline{P_T} - P^* \right\rangle$. This is due to convexity of the function $f_t(P) = \langle C, P \rangle$, which in this case is linear.

$$f\left(\overline{P}_T\right) - f\left(P^\star\right) \leq \frac{1}{T} \sum_{t=1}^{T} \left[f\left(P_t\right) - f\left(P^\star\right)\right] = \frac{\text{Regret}}{T} \tag{4.9}$$

### 4.3.1 Inexact FTRL: first bound

**Lemma 4.1.** *The regret of algorithm 4 is*

$$\text{Regret}_T \leq \eta G(1 + \log T) + T\delta + \|C\|_\infty \tag{4.10}$$

*Where $\eta$ is the user-chosen regularized constant, $G = \max R(a) - R(b)$ $s.t$ $a, b \in \Delta^{n \times n}$ and $\delta$ the approximation error of the subporblems defined as in* (4.7).

*Proof:* We start reviewing the standard FTRL regret bound, see [21]. (also [15], [8]).

$$\sum_{t=1}^{T} \left(f_t\left(P_t\right) - f_t(u)\right) \leq R(u) - R\left(P_1^*\right) + \sum_{t=1}^{T} \left(f_t\left(P_t\right) - f_t\left(P_{t+1}^*\right)\right), \quad \forall u \in S = \mathcal{U}(r,c) \tag{4.11}$$

where $S$ is our convex set and $P_t^*$ are the optimal solutions of the subproblems.

$$P_1^* = \arg\min R(P) \tag{4.12}$$

$$P_t^* = \operatorname*{argmin}_{P \in S} \sum_{i=1}^{t-1} f_i(P) + R(P) \quad \forall t = 2, 3, \ldots \tag{4.13}$$

In our case: $P_t$ is a $\delta_t$ approx solution of the subproblem and the linear function is constant the linear function is $f_t = \langle C, P \rangle \ \forall t$.

$$\langle C, P_t \rangle + \frac{1}{t-1} R(P_t) \leq \langle C, P_t^* \rangle + \frac{1}{t-1} R(P_t^*) + \delta_t \tag{4.14}$$

We rearrange sum of differences:

$$\sum_{t=1}^{T} \langle C, P_t \rangle - \left\langle C, P_{t+1}^* \right\rangle \leq \langle C, P_1 \rangle - \langle C, P_{T+1} \rangle + \sum_{t=2}^{T} \langle C, P_t \rangle - \langle C, P_t^* \rangle \tag{4.15}$$

Using the loose bound $R(P_t^*) - R(P_t) \leq G \, \forall t$ and $\langle C, P_1 - P_{T+1} \rangle \leq \|C\|_\infty$ and $\delta_t \leq \delta \, \forall t$

$$\sum_{t=1}^{T} \left( f_t \left( P_t \right) - f_t(u) \right) \leq \eta G + \|C\|_\infty + \sum_{t=2}^{T} \frac{\eta G}{t-1} + \delta_t \quad \forall u \in S \tag{4.16}$$

$$\text{Regret}_T \leq \eta G (1 + \log T) + T\delta + \|C\|_\infty \tag{4.17}$$

□

Note that the logarithm makes it difficult to explicitly choose $T$ as function of the error of the original problem.

$$\left\langle C, \overline{P_t} - P^* \right\rangle \leq \frac{1}{T} \text{Regret}_T \leq \epsilon \tag{4.18}$$

A naive solution is to use the trivial bound $\log T \leq \sqrt{T}$. With an approximate a priori knowledge of the value of $T$, more tight polynomial bounds can be proposed. This enables to explicitly find $T$ as a function of $\epsilon$.

For high values of $T$ (in the limit), where $\log T$ is upper bounded by a low degree polynomial, we could choose $T = O(\frac{1}{\epsilon})$ and $\delta = O(\epsilon)$.

### 4.3.2 Inexact FTRL: second bound

Now we present the second approach to deal with the approximate solutions inside FTRL.

**Lemma 4.2.** *The regret of algorithm 4 is*

$$\text{Regret} \leq \eta G + \|C\|_\infty + \tilde{\delta} \frac{T(T+1)}{2} \tag{4.19}$$

*Where $\eta$ is the learning rate, $G$ is the diameter of the set $\mathcal{U}(r, c)$, and $\tilde{\delta} \geq \delta_t$ is a bound on the approximation error of the subproblems as defined in 4.7.*

Proof: we start with a classical result in FTRL analysis (see [21]):

$$R(u) + \sum_{t=1}^{T} f_t(u) \geq R(w_1) + \sum_{t=1}^{T} f_t\left(P_{t+1}^*\right), \quad u \in S = \mathcal{U}(r,c) \tag{4.20}$$

where S is a convex set and $P_t^*$ has been defined in (4.13).

Now we extend this inequality (in the general setting of online learning with functions $f_1, f_2, \ldots$) to account for approximate solutions.

**Lemma 4.3.** *Suppose we choose $P_t$ such that.*

$$\sum_{i=1}^{t-1} f_i(P_t) + R(P_t) \leq \sum_{i=1}^{t-1} f_i(P_t^*) + R(P_t^*) + \lambda_t \tag{4.21}$$

*then*

$$R(u) + \sum_{t=1}^{T} f_t(u) \geq R(P_1) + \sum_{t=1}^{T} f_t\left(P_{t+1}\right) - \lambda_t \quad \forall u \in S \tag{4.22}$$

*Proof.* We proof the inequality by induction on t. The base cases $t = 1, 2$ follows from the definition of $P_1$, $P_2$ and that $R(P_1) \leq R(u) + \lambda_1 \forall u$. Now we assume the inequality holds for $t = T - 1$. We add $f_T(P_{T+1})$ to both sides and rearrange the inequalities.

$$f_T(P_{T+1}) + R(u) + \sum_{t=1}^{T-1} f_t(u) \geq R(P_1) + \sum_{t=1}^{T} (f_t(P_{t+1})) - \sum_{t=1}^{T-1} \lambda_t \tag{4.23}$$

This hold for all $u$, in particular for $u = P_{T+1}$. Finally, we use the definition of $P_{t+1}$ to show the inequality for $t = T$, which concludes the induction. □

We apply the lemma to get the regret $\forall u$:

$$\text{Regret}_T = \sum_{t=1}^{T} \left(f_t\left(P_t\right) - f_t(u)\right) \leq R(u) - R\left(P_1^*\right) + \sum_{t=1}^{T} f_t\left(P_t\right) - f_t\left(P_{t+1}\right) + \lambda_t \tag{4.24}$$

Using that $f_t(P) = \langle C, P \rangle$, the terms in the sum $\sum \langle C, P_t \rangle - \langle C, P_{t+1} \rangle$ telescope.

Note that each subproblem (4.7) is solved with precision $\delta_t = \lambda_t / t$, i.e. we control the error in terms of $\delta_t$, not directly $\lambda_t$. Using $\tilde{\delta} \geq \delta_t$, we bound $\lambda_t \leq t\ \tilde{\delta}$.

$$\text{Regret} = \eta G + \|C\|_\infty + \tilde{\delta} \frac{t(t+1)}{2} \tag{4.25}$$

□

Following this bound, we set $\eta = \tilde{O}(1), T = \tilde{O}(1/\epsilon)$, and $\tilde{\delta} = \tilde{O}(\epsilon^2)$. The notation $\tilde{O}(\cdot)$ hides logarithmic factors.

### 4.3.3 Inexact FTRL: tighter bounds?

As a future work, we believe that tighter bounds on the regret are possible. In particular, we would like to achieve a linear bound $T = O(\frac{1}{\epsilon})$ and a reasonable subproblem accuracy $\delta = O(\epsilon)$. This could be possible by using that $f_t(P) = \langle C, P \rangle$ in a more intelligent way. This requires to do a direct regret analysis on our specific setting and algorithm, instead of relying on the standard follow the regularized leader approach.

## 4.4 Relating the approximation error to the stopping criteria of Sinkhorn

In this section we compute how many Sinkhorn iterations we need to achieve accuracy $\delta_t$ in the subproblems

$$\langle C, P_{t+1} \rangle + \frac{\eta}{t} R(P_{t+1}) \leq \min_{P \in \mathcal{U}(r,c)} \langle C, P \rangle + \frac{\eta}{t} R(P) + \delta_t \tag{4.26}$$

First we relate the accuracy $\delta_t$ to the marginal violation $\rho$. This quantity defines the stopping criteria of the algorithm and is checked in $O(n^2)$ after each iteration.

$$\rho = \|P\mathbf{1} - r\|_1 + \left\|P^T\mathbf{1} - c\right\|_1 \tag{4.27}$$

**Theorem 4.4.** *([23], Theorem 8)To achieve a $\delta_t$ solution of* (4.26) *using Sinkhorn iterations (and a rounding step at the end), the marginal violation $\rho$ should be below*

$$\rho \leq \frac{\delta_t}{4\left(\|C\|_\infty + 2\frac{\eta}{t}\ln\frac{4\eta n^2}{\delta_t t}\right)} \leq \frac{\delta_t}{4\|C\|_\infty} \tag{4.28}$$

□

Finally, remember that $\delta_t$ is set as a function of the desired accuracy on the original problem $\langle C, P_\epsilon - P^* \rangle \leq \epsilon$, based on the FTRL analysis of algorithm 4.

## 4.5 A general bound of the number of Sinkhorn iterations

In this part, we relate the desired marginal violation $\rho$ with the number of Sinkhorn iterations. First we start reviewing some results from [12] and [23] for solving regularized optimal transport with Sinkhorn (4.29) (without proximal formulation). In then next section we will adapt the analysis to our particular setting of follow the regularized leader and $\epsilon$-scaling, see algorithm 4.

Let us consider the regularized optimal transport problem. Note that by setting $\gamma = \eta/t$ we recover the subproblems of the follow the regularized framework.

$$\min_{X \in \mathcal{U}(r,c)} \langle C, P \rangle + \gamma \sum_{i,j} P_{ij} \ln P_{ij} \tag{4.29}$$

The dual problem of (4.29) is equivalent to:

$$\min_{u,v \in \mathbb{R}^n} \langle \mathbf{1}B(u,v)\mathbf{1} \rangle - \langle u, p \rangle - \langle q, v \rangle \tag{4.30}$$

$$B(u,v) := \operatorname{diag}(e^u)\, e^{-C/\gamma} \operatorname{diag}(e^v) \tag{4.31}$$

where $(u, v)$ are a linear transformation of the Lagrange multipliers of the marginal constraints.

**Lemma 4.5.** *(Lemma 3 of [23].) Sinkhorn iterations are a contraction $e^{u^t-u^*}$ and $e^{v^t-v^*}$ in Hilbert's projective metric.*

$$R_t := \begin{cases} \max_j \left(v_j^t - v_j^*\right) - \min_j \left(v_j^t - v_j^*\right), & t \bmod 2 = 0 \\ \max_i \left(u_i^t - u_i^*\right) - \min_i \left(u_i^t - u_i^*\right), & t \bmod 2 = 1 \end{cases} \tag{4.32}$$

*where $(u^*, v^*)$ is the optimal scaling variables. Then for any $t \geq 0$ it holds $R_{t+1} \leq R_t$.*

The proof of the following lemma is missing in the original paper. Here we show the derivation of this bound to get a good insight and to adapt it later to our proximal formulation.

**Lemma 4.6.** *(Claim without proof from [23]), using $u_0 = \log r,\ v_0 = \log c$.*

$$R_0 \leq \frac{\max C_{ij} - \min C_{ij}}{\eta} \tag{4.33}$$

*Proof:* The following ideas are based on **Lemma 1** of [12].

$$\sum_j e^{u_i^*} C_{ij} e^{v_j^*} = r_i \tag{4.34}$$

$$\sum_j e^{u_i^* - u_i^0} e^{\log r_i} C_{ij} e^{v^*} = r_i \tag{4.35}$$

With $\kappa = \min_{ij} e^{\frac{-C_{ij}}{\gamma}}$

$$e^{u_i^* - u_i^0} \kappa \sum_j e^{v_j^*} \leq 1 \ \forall i \tag{4.36}$$

$$\max_i e^{u_i^* - u_i^0} \leq \frac{1}{\kappa \sum_j e^{v_j^*}} \tag{4.37}$$

On the other hand, with $\tau = \max_{ij} e^{\frac{-C_{ij}}{\gamma}}$

$$e^{u_i^* - u_i^0} \tau \sum_j e^{v_j^*} \geq 1 \ \forall i \tag{4.38}$$

$$\min_i e^{u_i^* - u_i^0} \geq \frac{1}{\tau \sum_j e^{v_j^*}} \tag{4.39}$$

$$\max_i u_i^* - u_i^0 \leq \log \sum_j e^{v_j^*} - \log \kappa \tag{4.40}$$

$$-\min_i u_i^* - u_i^0 \leq -\log \sum_j e^{v_j^*} + \log \tau \tag{4.41}$$

Finally, we add both inequalities to get the desired bound. Note that we compute the difference as $u_i^0 - u_i^*$ using $\max_i u_i^* - u_i^0 = -\min_i u_i^0 - u_i^*$ $\square$

Recall (see **Theorem 1** of [12]) that the number of Sinkhorn iterations $k$ needed to get an $\epsilon$ error in $\|\cdot\|_1$ on the marginal violation $\rho$ is of the order of:

$$k = O\left(\frac{R_0}{\epsilon}\right) \tag{4.42}$$

Therefore, the quantity $R_0$ plays a central role in the complexity analysis of the algorithm.

## 4.6 Bounding the number of Sinkhorn iterations in the proximal setting

In this section we bound how many iterations we need to solve the subproblems of our follow the regularized leader framework. Recall that we are solving a sequence of problems:

$$P_t = \arg\min_{P \in \mathcal{U}} \langle C, P \rangle + \frac{\eta}{t} \sum P_{ij} \log P_{ij} \tag{4.43}$$

**Lemma 4.7.** *The subproblem:*

$$P_t = \arg\min_{P \in \mathcal{U}} \langle C, P \rangle + \frac{\eta}{t} \sum P_{ij} \log P_{ij} \tag{4.44}$$

*can be rewritten both as (same solution):*

$$P_t = \arg\min_{P \in \mathcal{U}} \langle C, P \rangle + \eta D(P, P_{t-1}) \tag{4.45}$$

$$P_t = \arg\min_{P \in \mathcal{U}} \langle C - \eta \log P_{t-1}, P \rangle + \eta \sum P_{ij} \log P_{ij} \tag{4.46}$$

*Proof*: Note that the solution of (4.44) is

$$P_t = \mathrm{Diag}(e^{u_t})\ e^{C/(\eta/t)}\ \mathrm{Diag}(e^{v_t}) \tag{4.47}$$

Which can be rewritten as ($\cdot$ means element-wise product):

$$P_t = \mathrm{Diag}(e^{u_t^p})\ \mathrm{Diag}(e^u_{t-1})\ \left(e^{C/(\eta/(t-1))}\ \cdot e^{C/\eta}\right)\ \mathrm{Diag}(e^{v_t^p})\ \mathrm{Diag}(e^v_{t-1}) \tag{4.48}$$

Noting that:

$$P_{t-1} = \mathrm{Diag}(e^u_{t-1})\ e^{C/(\eta/(t-1))}\ \mathrm{Diag}(e^u_{t-1}) \tag{4.49}$$

We rewrite:

$$P_t = \mathrm{Diag}(e^{u_t^p})\left(P_{t-1} \cdot e^{C/\eta}\right)\mathrm{Diag}(e^v_{t-1}) \tag{4.50}$$

We can commute matrix products, and use associativity with the element-wise product because matrices are diagonal. . □

Note that $P_{t-1}$ does not need to be the exact solution of the previous subproblem, as we only require that it has the given structure and it belongs to the feasible set.

In this report, we assume that Lemma 4.7 either still holds even with the rounding step, or induces a "small" error that does not increase the overall computational complexity.

**Assumption 4.8.** *Either:*
*- Lemma 4.7 still holds when considering the rounding step after Sinkhorn iterations of algorithm 4.*
*or:*
*- Lemma 4.7 does not hold exactly, but solving the "approximate equivalent" formulation induces a small suboptimallity additive error that does not increase the computational complexity.* □

We compute a first bound on how many iterations we need to solve (4.45) taking into account this equivalence, lemma 4.7.

**Lemma 4.9.** *In the proximal version of optimal transport,*

$$\min_{P \in \mathcal{U}(p,q)} \langle C, P \rangle + \gamma D(P, P_t) \tag{4.51}$$

*where $D(P, P^t)$ is the Bregman divergence. A general bound for $R_0^t$ is:*

$$R_0^t \leq \frac{\max C_{ij} - \min C_{ij}}{\gamma} - \log \min P_t \tag{4.52}$$

*Proof*: Problem (4.51) can be rewritten to recover the standard (non proximal) formulation.

$$\min_{P \in \mathcal{U}(p,q)} \langle C - \gamma \log P_t, P \rangle + \gamma \sum_{i,j} P_{ij} \ln P_{ij} \tag{4.53}$$

An upper bound of the number of iterations needed to solve (4.51) is obtained by considering $\overline{C} = C - \gamma \log P_t$ (4.53) and computing the value of $R_0^t$ in this case.

In this case,

$$R_0^t \leq \frac{\max C_{ij} - \min C_{ij} + \gamma \log \max P_t - \gamma \log \min P_t}{\gamma} \tag{4.54}$$

$$R_0^t \leq \frac{\max C_{ij} - \min C_{ij}}{\gamma} - \log \min P^t \tag{4.55}$$

□

With the proximal formulation, $\eta$ is a "big" regularization. We are interested in the logarithmic term of lemma 4.9:

$$\log \min P_{t-1} \tag{4.56}$$

We know analyze how small could be $\min P_{t-1}$ as a function of $t$. Recall that, in non-degenerate cases of the unregularized problem, $\lim_{t \to \infty} \min P_{t-1} = 0$.

**Lemma 4.10.** *For some $C_{ij}$, $\min P_{t-1}$ converges exponentially fast to zero, i.e.,*

$$\min P_t = O(e^{-t}) \tag{4.57}$$

*Proof*: We can show this with a toy example. $r = c = \frac{1}{n}\mathbf{1}$, $C_{ij} = 0 \ \forall (i,j) \neq (0,0)$ and $C_{00} = 1$. We can compute the analytical solution to show $P_{00}$ converges to 0 exponentially fast. For each subproblem, the cost only depends on $P_{00}$ and is a trade off between the $C_{00}$ and the regularization of $P_{00}$. □

Therefore, a naive application of the Lemma 4.9 and 4.10 gives the bound $R_o^t = O(t)$. This suggests that the number of iterations increases as t advances, which is not reflected on practice.

We now propose a tighter bound on $R_0^t$ specific for our framework, leveraging that A) we solve a sequence of subproblems decreasing the regularization and B) the sequence converges to the optimal solution of the unregularized problem.

The key idea is that, in lemmas 4.9, 4.10, the analysis was made for any $P_{ij}^{t-1}$ and $C_{ij}$, without considering the interaction between them. Now we want to use the relation

of $P_{ij}^{t-1}$ and $C_{ij}$, as $P_{ij}^{t-1}$ is the solution (an approximation) of the subproblem with same $C_{ij}$ and bigger regularization.

In fact, we can parametrize the sequence $P_t$ as a family of solutions for a parametrized problem with regularization $\frac{\eta}{t}$. For $t = 0$, the solution is the maximum entropy matrix with desired marginals $r, c$; that is $P_0 = rc^T$. For $t \to \infty$, the solution is $P^*$, which contains entries equal to zero.

The intuition is that, as $t$ increases, the small entries of $P_{ij}$ (those that will be zero at the optimum) always decrease (because they converge to zero from above).

**Lemma 4.11.** *Let $\delta \in \mathbb{R}_{>0}$ be a threshold such that, if $P_{ij}^{t'} \leq \delta$ then $P_{ij}^{t} \leq \delta \quad \forall t \geq t'$. Then, $R_0^t$ for the sequence of subproblems* (4.43) *is bounded by:*

$$R_0^t \leq \log\left(1 + \frac{n\delta}{r^0}\right) + \frac{\|C\|_\infty}{\gamma} - \log\delta \tag{4.58}$$

*Proof:* Let $A_{ij}$ be the solution (exact or approximated) of the previous problem: $A_{ij} = P_{ij}^{t-1}$. $u^*$ and $v^*$ are the optimal scaling factors of the current subproblem (using the proximal formulation). We use $u_0^i = \ln r_i$ (see lemma 4.6) We now drop index $i$ from $u$ to ease the notation.

$$\sum_j e^{u^*-u_0}\ r_i\ A_{ij}\ e^{C_{ij}/\eta}\ e^{v^*} = r_i \quad \forall i \tag{4.59}$$

We define $\overline{A}_{ij} = \max\{A_{ij}, \delta\}$ and decompose $A$ as $A = \overline{A} - A^l$
Now, $\min \overline{A}_{ij} = \delta$, $\max A^l_{ij} = \delta$ and $A^l_{ij} = 0$ if $A_{ij} > \delta$. Let $J_i$ be the set of row indices such that $A_{ij} < \delta$ for a given row $i$.

$$\sum_j e^{u^*-u_0}\ A_{ij}\ e^{C_{ij}/\eta}\ e^{v^*} = 1 \quad \forall i \tag{4.60}$$

$$\sum_j e^{u^*-u_0}\ \overline{A_{ij}}\ e^{C_{ij}/\eta}\ e^{v^*} - \sum_{j\in J_i} e^{u^*-u_0}\ A^l_{ij}\ e^{C_{ij}/\eta}\ e^{v^*} = 1 \quad \forall i \tag{4.61}$$

We now compute a lower bound for the term $-\sum_{j\in J_i} e^{u^*-u_0}\ A^l_{ij}\ e^{C_{ij}/\eta}\ e^{v^*}$

$$-\sum_{j\in J_i} e^{u^*-u_0}\ A^l_{ij}\ e^{C_{ij}/\eta}\ e^{v^*} = \tag{4.62}$$

$$-\sum_{j\in J_i} e^{u^*}\frac{1}{r_i}\ A^l_{ij}\ e^{C_{ij}/\eta}\ e^{v^*} \geq -\sum_{j\in J_i}\frac{1}{r_i}\ A^l_{ij} \geq \tag{4.63}$$

$$-\sum_{j\in J_i}\frac{1}{r_i}\ \delta\ \geq\ -\frac{n\delta}{r^0} \tag{4.64}$$

Where we have used our assumption that if $P_{ij}^t \leq \delta$ then $P_{ij}^t \leq \delta$. This means that $e^{u_i^*} e^{\frac{C}{\eta}} e^{v_j^*} \leq 1 \quad \forall J_i$, because $e^{u_i^*} P_{ij} e^{\frac{C}{\eta}} e^{v_j^*} \leq P_{ij}$.

We denote $\min r_i = r^0$ and use that $A^l_{ij} \leq \delta$; $|J_i| \leq n$ (size of the set).
Using $\tilde{\kappa} = \min_{ij} \overline{A}_{ij}\ e^{\frac{-C_{ij}}{\gamma}} = \delta \min_{ij} e^{\frac{-C_{ij}}{\gamma}} = \delta\kappa$

$$e^{u_i^* - u_i^0} \tilde{\kappa} \sum_j e^{v_j^*} - \frac{n\delta}{r^0} \leq 1 \quad \forall i \tag{4.65}$$

Now we apply logarithm and take the maximum:

$$\max_i u_i^* - u_i^0 \leq \log\left(1 + \frac{n\delta}{r^0}\right) - \log\kappa - \log\delta + \log\sum_j e^{v_j^*} \tag{4.66}$$

On the other hand, to get an upper bound for $-\min_i u_i^* - u_i^0$ we use that $A_{ij} \leq 1$ to obtain same result as in (4.38).

$$-\min_i u_i^* - u_i^0 \leq -\log\sum_j e^{v_j^*} + \log\tau \tag{4.67}$$

with $\tau = \max_{ij} e^{\frac{-C_{ij}}{\gamma}}$. Adding the two bounds together ($\max_i u_i^* - u_i^0 = -\min_i u_i^0 - u_i^*$)

$$R_0^t = \max_i u_i^0 - u_i^* - \min_i u_i^0 - u_i^* \leq \log\left(1 + \frac{n\delta}{r^0}\right) + \frac{\|C\|_\infty}{\gamma} - \log\delta \tag{4.68}$$

□

With this analysis, $R_0^t$ depends only on $\log\delta$ and $\frac{\|C\|_\infty}{\gamma}$, which implies the complexity of solving subproblems does not increase for high $t$. At this point, we are interested to proof that $\delta$ has polynomial relation with $n$ (size of problem) and $\epsilon$ (the accuracy of the desired solution).

The value of $r_0$ is not a problem. When $r_i$ is very small, we can solve an alternative problem with bigger marginals so that $r^0 = O(\epsilon/n)$ and still get same accuracy, see **Theorem 2** [12].

In our case, assuming that $r^0 = O(\epsilon/n)$, we speculate that $\delta = O(\epsilon^2/n^2)$. Our intuition is as follows: we solve a sequence of subproblems starting with $P^0 = rc^T$ that converges to $P^*$. In the homotopy $P^0 = rc^T \to P^*$, some $P_{ij}$ go from $\epsilon^2/n^2 \to 0$. Other $P_{ij}$ increase, or increase at the beginning and decrease at the end. This conjecture would imply that, if $P_{ij} < \epsilon/n^2$ then it remains bounded by $\epsilon^2/n^2$. In the next section, we try to proof this assumption, that we formalize as:

**Assumption 4.12.** *It exists a threshold $\delta = O(\epsilon^2/n^2)$ such that:*

$$\textit{if} \quad P_{ij}^{t'} \leq \delta \quad \textit{then} \quad P_{ij}^t \leq \delta \quad \forall t \geq t'. \tag{4.69}$$

## 4.7 Notes about the evolution of $P_{ij}^t$

In this section we study in detail the assumption 4.12.

Although we did not find a definitive proof, we share our thoughts and progress. In this section, we will assume that we are solving each subproblem exactly. This is not true in our framework, where we focused on analyzing the error induced by inexact solutions.

However, computing the evolution of $P_{ij}^t$ in an exact setting is the first step towards the analysis of our approximate solution.

It is clear that the evolution of $P_{ij}^t$ is not monotone. At first the components that are above a weighted average get increased. At latter iteration, they can decrease, as the mass is accumulated at only the best components (low $C_{ij}$). On the other hand, for our formulation, we need to show that, if $P_{ij}$ goes beyond a threshold, it remains bounded by it.

We want to show that $P_{ij}$ will either:

- decrease
- increase
- remains constant
- first increase, then decrease (only two modes)

We start with proof for the case of linear cost optimization on the simplex with entropic regulation. This can be seen as a simpler model of the optimal transport problem with just one dimension, without the marginals constraints.

**Lemma 4.13.** *Let $P(t)$ be the solution of:*

$$\min_{P\in\Delta} \sum_i l_i P_i + \frac{\eta}{t} \sum_i P_i \log P_i \tag{4.70}$$

*Where $\Delta = \{P \in R^n_{\geq 0} : \sum_i P_i = 1\}$, $l \in \mathbb{R}^n_{\geq 0}$ , $\eta \in \mathbb{R}$ and $t \in \mathbb{N}$.*

*Then, $P_i(t)$ either decreases, increases, remains constant or first increases and then decreases.*

*Therefore, with $P(0) = \frac{1}{n}\mathbf{1}$, the threshold $\delta$ (defined as in assumption 4.12) equals $\frac{1}{n}$*

*Proof:* The solution of (4.70) has an analytical expression:

$$P_i = \frac{e^{-l_i t/\eta}}{\sum_j e^{-l_j t/\eta}} \tag{4.71}$$

Which can be rewritten as:

$$P_i = e^{-l_i t/\eta - \log \sum_j e^{-l_j t/\eta}} \tag{4.72}$$

We now analyze the evolution of $P_i$ as a function of $t$. $P_i(t)$ defines a smooth trajectory and we compute its derivative.

$$\frac{d}{dt}P_i(t) = P_i(t)\left(-l_i + \frac{\sum_j l_j e^{-l_j t/\eta}}{\sum_j e^{-l_j t/\eta}}\right)\frac{1}{\eta} = P_i(t)\left(-l_i + m(t)\right)\frac{1}{\eta} \tag{4.73}$$

The term $m(t)$ is a weighted average of the $l_i$

$$m(t) = \frac{\sum_j l_j e^{-l_j t/\eta}}{\sum_j e^{-l_j t/\eta}} \tag{4.74}$$

The sign of $\frac{d}{dt}P_i(t)$ depends on: $m(t) - l_i$. Note that $\lim_{t\to 0} m(t) = \frac{1}{n}\sum_j l_j$ and $\lim_{t\to\infty} m(t) = \min_i l_i$.

We show that $m(t)$ decreases monotonically. Then, $m(t) - l_i$ is either: a) always negative b) first positive and then negative or c) always positive (converging to zero). Note that the sign of $m(t) - l_i$ determines the sign of $\frac{d}{dt}P_i(t)$

We compute $\frac{d}{dt}m(t)$:

$$\frac{d}{dt}m(t) = \frac{1}{\eta}\frac{(-\sum_j l_j^2 e^{-l_j t/\eta})(\sum_j e^{-l_j t/\eta}) + (\sum_j l_j e^{-l_j t/\eta})(\sum_j l_j e^{-l_j t/\eta})}{\left(\sum_j e^{-l_j t/\eta}\right)^2} \tag{4.75}$$

The sign is determined only by the numerator. We multiply terms in the sum, gather terms together and use symmetry $i, j$.

$$-\sum_{i,j} l_j^2 e^{(-l_j - l_i)t/\eta} + \sum_{i,j} l_i l_j e^{(-l_j - l_i)t/\eta} = \tag{4.76}$$

$$\sum_{i,j}(l_i l_j - l_j^2)\ e^{(-l_j - l_i)t/\eta} = \tag{4.77}$$

$$\sum_{i\neq j}(-l_j^2 + 2l_i l_j - l_i^2)e^{(-l_j - l_i)t/\eta} = \tag{4.78}$$

$$\sum_{i\neq j} -(l_i - l_j)^2 e^{(-l_j - l_i)t/\eta} \leq 0 \tag{4.79}$$

□

After showing this lemma in the simplex case, we wonder whether we could apply the same strategy to proof the evolution of $P_{ij}(t)$ in optimal transport. Unfortunately, we do not have a definitive answer. We report our progress. To ease the notation, we assume $\eta = 1$. The optimization problems, parametrized by $t$ are:

$$\min \langle C, P\rangle + \frac{1}{t}\sum_{ij} P_{ij} \ln P_{ij} \tag{4.80}$$

The solution of (4.80) is

$$P_{ij}(t) = e^{C_{ij}t + \alpha_i t + \beta_j t} \tag{4.81}$$

with $\alpha_i = \alpha_i(t),\ \beta_i = \beta_i(t)$. As in Lemma 4.13, we are interested in the derivative.

$$\frac{d}{dt}P_{ij}(t) = P_{ij}(t)\left(-C_{ij} + \alpha_i + \beta_j + t\dot{\alpha}_i + t\dot{\beta}_j\right) \tag{4.82}$$

where $\dot{\alpha} = \frac{d}{dt}\alpha$

We compute $\alpha, \beta, \dot{\alpha}, \dot{\beta}$ using the Lagrange dual

$$\alpha(t), \beta(t) = \arg\min \sum_{ij} e^{(-C_{ij} + \alpha_i + \beta_j)t} - \ t < \alpha, r > - t < \beta, c > \tag{4.83}$$

$$\alpha(t), \beta(t) = \arg\min f(t, (\alpha(t), \beta(t))) \tag{4.84}$$

Because, $f()$ this is convex and there are no constraints, the solution $\alpha(t), \beta(t)$ fulfils:

$$\nabla f(t, \alpha(t), \beta(t)) = 0 \tag{4.85}$$

$$\frac{\partial}{\partial t}\nabla f(t, \alpha(t), \beta(t)) + \nabla^2 f(t, \alpha(t), \beta(t)) \cdot (\dot{\alpha}, \dot{\beta})^T = 0 \tag{4.86}$$

From this equation we could try to get information about the evolution of $(\dot{\alpha}, \dot{\beta})$. It is possible to compute some expressions in close form, for example:

$$\frac{\nabla f_i}{\partial t} = \sum_j P_{ij} \ln P_{ij} \tag{4.87}$$

$$\frac{\partial^2}{\partial \alpha_i \partial \beta_j} = t^2 P_{ij} \tag{4.88}$$

$$\frac{\partial^2 f}{\partial \alpha_i \partial \alpha_{i'}} = 0 \tag{4.89}$$

$$\frac{\partial^2 f}{\partial \alpha_i^2} = t^2 r_i \tag{4.90}$$

$$\tag{4.91}$$

Our strategy is to get $\dot{\alpha}, \dot{\beta}$ from equation (4.86) and then plug this result into (4.82), trying to prove that $\frac{d}{dt}P_{ij}(t)$ is either: positive, negative, zero, or first positive and later negative. One of our concerns is that the solution of the optimization (4.83) is defined up to an additive constant. This means that $\nabla^2 f$ has rank $n^2 - 1$. We are still unsure about the consequences of this observation.

## 4.8 Summary and next steps

In this last section of the chapter, we put together all the results to get the computational complexity of our algorithm to find an $\epsilon$ -approximate solution $\overline{P}_T$, :

$$\left\langle \overline{P}_T - P^*, C \right\rangle \leq \epsilon \tag{4.92}$$

**Lemma 4.14.** *The computational complexity of algorithm 4, is*

$$\tilde{O}\left(\frac{n^2}{\epsilon^3}\right) \tag{4.93}$$

*where $\tilde{O}(\cdot)$ hides the logarithmic factors.*

*Proof*: This final lemma is a combination of the results already presented in this section. Note that we consider true our (still unproved) assumptions 4.12 and 4.8. We recap here the main steps of the proof:

1. Using FTRL, we found the number of outer iterations $T$ and the required precision of the subproblems. Here we use the loose bound of lemma 4.2, although we believe we could also use the tighter non-explicit bound of lemma 4.1.

   Using $G = 2\log n$, we set $\eta = \|C\|_\infty / G$ so that $\eta = \tilde{O}(1)$. Using the average regret bound:

   $$\frac{2\|C\|_\infty}{T} + \frac{\delta T}{2} \leq \epsilon \tag{4.94}$$

   We set $T = O(1/\epsilon)$ and $\delta = O(\epsilon^2)$

2. With Theorem 4.4 we relate the accuracy of subproblems $\delta$ with the L1 marginal violation $\rho$ of Sinkhorn algorithm. The relation is linear so $\rho = O(\delta) = O(\epsilon^2)$.

3. The number of Sinkhorn iterations for solving subproblem $t$ is $k = O(R_0^t/\rho)$. We use our new bound of $R_o^t$ of lemma 4.11 with assumptions 4.8 and 4.12 to get $R_0^t = \tilde{O}(1)$ (constant in all subproblems), so $k = \tilde{O}(1/\epsilon^2)$

4. Each Sinkhorn iteration has complexity $n^2$

5. The final rounding step of the algorithm to get the desired marginals only induces a small bounded error that depends linearly on the marginal violation. This does not modify the overall complexity.

□

Here we summarize the missing parts of the analysis. Solving these issues will allow us to prove the convergence and "faster" rates for our algorithm 4, providing a theoretic analysis of the $\epsilon$-scaling heuristic.

- Decide if we need to round the inexact solution of each subproblem of the sequence. For the regret bound using FTRL, we are assuming that each matrix belongs to the feasible set, i.e $P^t \in \mathcal{U}(r, c)$. However, the study of the number of Sinkhorn iterations assumes that no rounding is applied, so that $P^t$ has the structure coming directly from Sinkhorn algorithm. One option to unify the criteria is to consider FTRL in a relaxed convex set $\tilde{\mathcal{U}}(r, c, \tilde{\epsilon})$ and apply the rounding algorithm to get $P \in \mathcal{U}(r, c)$ only once at the end. Remember that the error in the cost introduced by the rounding is "small" and bounded linearly by the $\|\cdot\|_1$ between the marginals. See also assumption 4.8.

$$\tilde{\mathcal{U}}(r, c, \tilde{\epsilon}) = \{P \in \Delta^{n\times n} : \|P\mathbf{1} - r\| + \left\|P^T\mathbf{1} - c\right\| \leq \tilde{\epsilon}\} \tag{4.95}$$

- We are still looking for a proof of Assumptions 4.12, 4.8, which are necessary to apply the tight bound on the number of iterations of Sinkhorn iteration. These assumptions should be proved in both cases, the exact and inexact setting, which could be challenging or even not possible.

- Getting a tighter and explicit bound of $\left\langle \overline{P}_T - P^*, C \right\rangle$, refining the analysis of FTRL or even using another framework. Given that the linear function is constant, we believe that achieving $T = O(1/\epsilon)$ is possible.

# 5 An alternative formulation: three more ideas

Here we present several short notes on different concepts and ideas that we also explored during these months. Even though we could not reach any definitive conclusion following these approaches, we hope future work gets inspired by these ideas and considerations.

## 5.1 Approximate projections

Again, our goal is to analyze the inexact proximal version of Sinkhorn, algorithm 5.

**Algorithm 5** Approximate Proximal regularized Optimal Transport

**Input:** $C, r, c, \delta$
let $D(P, P_t)$ be the Bregman divergence.
**for** $t = 1, \dots, T$ **do**
  $P_t = \arg\min^{\delta} \langle C, P \rangle + \eta D(P, P_{t-1})$, s.t $P_t \in \mathcal{U}(r, c)$
**end for**
$\overline{P}_T = \frac{1}{T} \sum P_t$
**Output:** $\overline{P}_T$

Now, we analyze the algorithm using the framework of mirror descent with approximate projections, following the approach in [11].

Recall the basic mirror descent formulation in online learning:

$$P_{t+1} = \underset{P \in \mathcal{U}}{\operatorname{argmin}} \ \eta \langle \ell_t, P \rangle + D_R(P, P_t) \tag{5.1}$$

Which is equivalent to a two step process, where $\Pi_R$ is a projection to the feasible set.

$$\begin{aligned} \tilde{P}_{t+1} &= \underset{P \in \Delta}{\operatorname{argmin}} \ \eta \langle \ell_t, P \rangle + D_R(P, P_t) \\ P_{t+1} &= \Pi_R \left( \tilde{P}_{t+1} \right) \end{aligned} \tag{5.2}$$

In case of inexact projections, $P_{t+1}$ is not the exact solution of (5.1), but a $c$-approximation.

$$\left\| P_{t+1} - \Pi_R \left( \tilde{P}_{t+1} \right) \right\| \leq c \tag{5.3}$$

Solving each subproblem (5.1) using the Sinkhorn iteration can be modelled as a c-approximate solution. Our goal is to relate the marginal violation of each iteration with this c-approximation, and therefore to the complexity of the algorithm.

Some drawbacks of this approach are:

- in the basic formulation, $l_t$ is a time-varying (adversarial) vector. In our case $l_t = l \ \forall t$
- there is not a clear way to leverage that $P_t$ converges to $P^*$, the solution of the unregularized problem.
- The gradient of the Bregman divergence is unbounded when the entries of $P$ approach to zero. In $P^*$, several components are equal to zero, which complicates the analysis.

## 5.2 Two player game

There is a recent interest in formulating optimization algorithms as two player zero sum games, see [2], [1]. With this approach, the authors reformulate conditional gradient descent (Frank Wolf) as a two player saddle point computation.

A basic idea of this formulation is that the duality gap of the saddle point computation is the sum of the regret of both players, see [2], [1].

To apply this algorithm to our setting, we have designed a two player framework that mimics Sinkhorn iterations and the proximal transport formulation. The key idea is to introduce the error induced by inexact projections in the framework and to compute how it gets accumulated and its impact on the duality gap.

In our case, the convex-concave function (two player game) is the Lagrangian:

$$g(P, y) = \mathcal{L}_{x\geq 0}(P, \alpha, \beta) = \langle C, P\rangle - \langle \beta, P\mathbf{1} - r\rangle - \langle \alpha, P\mathbf{1} - c\rangle \tag{5.4}$$

**Algorithm 6** Two player game optimal transport

**Input:** $C, r, c$

let $\alpha_t, \beta_t$ be estimates of the dual variables of (5.4), and $u_t, v_t$ the scaling factors of the current projection.

**for** $t = 1, \ldots, T$ **do**

(Player 1)

$$\alpha_t, \beta_t, u_t, v_t = \Pi(P_{t-1} \cdot e^{\frac{-C}{\eta}}) \tag{5.5}$$

(Player 2)

$$P_t = P_{t-1} \cdot e^{\frac{-C}{\eta}} \cdot e^{u_t + v_t} \tag{5.6}$$

**end for**

$\overline{P} = \frac{1}{T}\sum P_t$

**Output:** $\overline{P}$

**Lemma 5.1.** *Suppose that player 1 of algorithm 6 makes perfect projections* $\Pi$*, then average regret on the dual gap is* $O\left(\frac{1}{\sqrt{T}}\right)$

*Proof*: The regret on the dual gap is the sum of the regrets of the two players. Player 1 makes perfect projections:

$$P_t \in U_{r,c}\ \forall t,\ P_t\mathbf{1} - c = 0,\ P_t^T\mathbf{1} - r = 0 \tag{5.7}$$

The regret of the first player (it tries to maximize) is:

$$g(P_T, y^*) - \frac{1}{T}\sum g(P_t, y_t) = \langle C, P_T\rangle - \frac{1}{T}\sum \langle C, P_t\rangle = 0 \tag{5.8}$$

where $y = (\alpha, \beta)$. The regret of the second player (tries to minimize):

$$\frac{1}{T}\sum g(P_t, y_t) - g(P^*, y_t) = \frac{1}{T}\sum g_t(P_t) - \frac{1}{T}\sum g_t(P^*) = R_2 \tag{5.9}$$

Note that the second player is playing mirror descent on a sequence of linear loss functions: $l_t = -\frac{C}{\eta} + u_t + v_t$

Without assuming any smoothness in the vector $g_t$, the regret bound is the classical $O(\frac{1}{\sqrt{T}})$. Potentially, we believe that this could be refined to $O(\frac{1}{T})$ using the fact that the linear functions are not adversarial and $X_t \to X^*$, $\alpha_t \to \alpha^*$ and $\beta_t \to \beta^*$. □

The learning rate $\eta$ should be chosen to achieve a good trade off between the computation load of player 1 and the regret of player 2.

So far we have analyzed the perfect projection setting. But our goal is to analyze the approximate version, where Player 1 only uses Sinkhorn iterations and computes approximate solutions.

We provide here some short and informal thoughts:
The vectors $\alpha_t, \beta_t, u_t, v_t$ are approximations and now, after the update of player 2, $P_t$ does not have marginals $r, c$: $P_t\mathbf{1} - r \neq 0$ or $P_t^T\mathbf{1} - c \neq 0$.

In the approximate setting, the first player has also some regret. We try to bound $g(P_t, \alpha_t) - g(P_t, \alpha^*)$ (assume w.l.o.g that columns are scaled after Sinkhorn algorithm).

$$g(P_t, \alpha_t) - g(P_t, \alpha^*) \leq \langle \alpha_t - \alpha^*, P_t - r \rangle \tag{5.10}$$

One of the advantages of this formulation is that with the Sinkhorn algorithm we control directly the marginal violation $\|P_t\mathbf{1} - r\|_1$. Therefore, we could use directly the "loose" bound:

$$\langle \alpha_t - \alpha^*, P_t - r \rangle \leq \|\alpha_t - \alpha^*\|_\infty \|P_t\mathbf{1} - r\|_1 \tag{5.11}$$

In non degenerate cases, $\alpha_t \to \alpha$. However, to get a regret bound we still need a more precise description of the speed of convergence and a notion of the distance $\alpha_t \to \alpha^*$ as a function of $t$. Another difficulty is that $\alpha_t$ is an approximate solution of (5.5).

## 5.3 Potentials

Another idea to analyze the proximal regularized formulation of optimal transport is to find a potential that decreases in each iteration.

There is a well known potential function for analyzing Sinkhorn iterations with a fixed regularization, based on the dual function, see (4.30). However, it is still unclear how this potential could be extended to our proximal regularized setting.

Note that each iteration reduces the potential associated to the current regularization. However, this potential function is also changing each iteration, which makes the analysis more complicated.

# 6 Conclusion

The optimal transport distance is a powerful and useful metric to compare probability distributions in statistics and machine learning. In contrast to alternative distance measures, the main drawback is the computational complexity.

Adding entropic regularization to the original linear program, we can find approximate solutions much faster (reducing complexity from $O(n^3)$ to $O(n^2)$, where $n$ is the histogram size) using the Sinkhorn algorithm (iterative matrix scaling).

However, finding solutions with small error requires using small regularizations, which increases the number of Sinkhorn iterations. The dependence is $k = O(\epsilon^{-2})$, where $k$ is the number of Sinkhorn iterations and $\epsilon$ the approximate error.

In this report, we have proposed an algorithm based on proximal regularization. It consists on solving a sequence of problems decreasing the regularization, and resembles the well-known heuristic $\epsilon$-scaling.

Using a follow the regularized leader analysis and novel bounds for this iterative and proximal regularized setting, we bound the complexity of our algorithm as $O(n^2/\epsilon^3)$. However, this result depends on two still unproved assumptions, that we leave for future work.

We have not analyzed the experimental performance of the algorithm, but we believe it should be equivalent to $\epsilon$-scaling. This strategy has been shown to outperform the standard approach of solving only one scaling problem with very small regularization. Moreover, one of the advantages of our algorithm and $\epsilon$-scaling is that they are any-time algorithms, producing better solutions as more computational time is available.

# References


[1] Jacob D. Abernethy, Kevin A. Lai, Kfir Y. Levy, and Jun-Kun Wang. Faster rates for convex-concave games. *CoRR*, abs/1805.06792, 2018.

[2] Jacob D Abernethy and Jun-Kun Wang. On frank-wolfe and equilibrium computation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6584–6593. Curran Associates, Inc., 2017.

[3] Brahim Khalil Abid and Robert M Gower. Greedy stochastic algorithms for entropy-regularized optimal transport problems. *arXiv preprint arXiv:1803.01347*, 2018.

[4] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable sinkhorn distances via the nyström method. In *Advances in Neural Information Processing Systems*, pages 4429–4439, 2019.

[5] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

[6] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *ArXiv*, abs/1701.07875, 2017.

[7] Jose Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards optimal running times for optimal transport. *arXiv preprint arXiv:1810.07717*, 2018.

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

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

[10] Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

[11] Travis Dick, András György, and Csaba Szepesvári. Online learning in markov decision processes with changing cost sequences. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–512–I–520. JMLR.org, 2014.

[12] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. *CoRR*, abs/1802.04367, 2018.

[13] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448, 2016.

[14] Arun Jambulapati, Aaron Sidford, and Kevin Tian. A direct $\tilde{O}(1/\epsilon)$ iteration parallel algorithm for optimal transport. *arXiv preprint arXiv:1906.00618*, 2019.

[15] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. 2018.

[16] Tianyi Lin, Nhat Ho, and Michael I Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. *arXiv preprint arXiv:1901.06482*, 2019.

[17] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009.

[18] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11:355–607, 2018.

[19] Kent Quanrud. Approximating optimal transport with linear programs. *arXiv preprint arXiv:1810.05957*, 2018.

[20] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

[21] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[22] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[23] Fedor Stonyakin, Darina Dvinskikh, Pavel Dvurechensky, Alexey Kroshnin, Olesya Kuznetsova, Artem Agafonov, Alexander Gasnikov, Alexander Tyurin, César A Uribe, Dmitry Pasechnyuk, et al. Gradient methods for problems with inexact model of the objective. *arXiv preprint arXiv:1902.09001*, 2019.

[24] Jonathan Weed. An explicit analysis of the entropic penalty in linear programming. In *COLT*, 2018.

[25] Xinhua Zhang. Lecture notes: Bregman divergence and mirror descent.