A Stable Alternative to Sinkhorn's Algorithm for Regularized Optimal Transport

Pavel Dvurechensky[1,3](✉), Alexander Gasnikov[2,3], Sergey Omelchenko[2], and Alexander Tiurin[4]

[1] Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
pavel.dvurechensky@gmail.com
[2] Moscow Institute of Physics and Technology, Moscow, Russia
[3] Institute for Information Transmission Problems RAS, Moscow, Russia
[4] National Research University Higher School of Economics, Moscow, Russia

Abstract. In this paper, we are motivated by two important applications: entropy-regularized optimal transport problem and road or IP traffic demand matrix estimation by entropy model. Both of them include solving a special type of optimization problem with linear equality constraints and objective given as a sum of an entropy regularizer and a linear function. It is known that the state-of-the-art solvers for this problem, which are based on Sinkhorn's method (also known as RSA or balancing method), can fail to work, when the entropy-regularization parameter is small. We consider the above optimization problem as a particular instance of a general strongly convex optimization problem with linear constraints. We propose a new algorithm to solve this general class of problems. Our approach is based on the transition to the dual problem. First, we introduce a new accelerated gradient method with adaptive choice of gradient's Lipschitz constant. Then, we apply this method to the dual problem and show, how to reconstruct an approximate solution to the primal problem with provable convergence rate. We prove the rate $O(1/k^2)$, k being the iteration counter, both for the absolute value of the primal objective residual and constraints infeasibility. Our method has similar to Sinkhorn's method complexity of each iteration, but is faster and more stable numerically, when the regularization parameter is small. We illustrate the advantage of our method by numerical experiments for the two mentioned applications. We show that there exists a threshold, such that, when the regularization parameter is smaller than this threshold, our method outperforms the Sinkhorn's method in terms of computation time.

Submitted to the editors DATE. This research was funded by Russian Science Foundation (project 18-71-10108).

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-49988-4_28) contains supplementary material, which is available to authorized users.

A. Kononov et al. (Eds.): MOTOR 2020, LNCS 12095, pp. 406–423, 2020.
https://doi.org/10.1007/978-3-030-49988-4_28

Keywords: Smooth convex optimization · Linear constraints · First-order methods · Accelerated gradient descent · Algorithm complexity · Entropy-linear programming · Dual problem · Primal-dual method · Sinkhorn's fixed point algorithm · Entropy-regularized optimal transport · Traffic demand matrix estimation

1 Introduction

The main problem, we consider, is convex optimization problem of the following form

$$(P_1) \qquad \min_{x \in Q \subseteq E} \left\{ f(x) : A_1 x = b_1, A_2 x - b_2 \in -K \right\},$$

where E is a finite-dimensional real vector space, Q is a simple closed convex set, A_1, A_2 are given linear operators from E to some finite-dimensional real vector spaces H_1 and H_2 respectively, $b_1 \in H_1$, $b_2 \in H_2$ are given, $K \subseteq H_2$ is some cone, $f(x)$ is a γ-strongly convex function on Q with respect to some chosen norm $\| \cdot \|_E$ on E. The last means that, for any $x, y \in Q$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\gamma}{2} \|x - y\|_E^2$, where $\nabla f(x)$ is any subgradient of $f(x)$ at x and hence is an element of the dual space E^*. Also we denote the value of a linear function $\lambda \in E^*$ at $x \in E$ by $\langle \lambda, x \rangle$.

We are motivated to consider the described class of problems by two particular applications. The first one comes from transportation research and consists in recovering a matrix of traffic demands between city districts from the information on population and workplace capacities of each district. As it is shown in [23], a natural model of districts' population dynamics leads to an entropy-linear programming optimization (see (9) below for the precise formulation) problem for the traffic demand matrix estimation. In this case, the objective function in (P_1) is a sum of an entropy function and a linear function. It is important to note also that the entropy function is multiplied by a regularization parameter γ and the model is close to reality, when the regularization parameter is small. The same approach is used in IP traffic matrix estimation [54]. Close problems arise also in more complicated congestion traffic modelling [4].

The second application is the calculation of regularized optimal transport (ROT) between two probability measures introduced in [12]. The idea is to regularize the objective function in the classical optimal transport linear programming problem [31] by entropy of the transportation plan. This leads to the same type of problem with a regularization parameter as in the traffic demands matrix estimation. For the detailed problem statement, see (9). As it is argued in [13], for the case of discretization of continuous probability measures, entropy regularization allows to obtain a better approximation for the optimal transportation plan than the solution of the original linear programming problem. At the same time, the regularization parameter γ should be small. Otherwise, the solution of the regularized optimal transport problem will be a bad approximation for the original optimal transport problem. To sum up, in both applications, it is important to solve regularized problems with *small regularization parameter*.

The problem statement (P_1) covers many other applications besides mentioned above. For example, general entropy-linear programming (ELP) problem [20] arises in econometrics [24], modeling in science and engineering [32]. Such machine learning approaches as ridge regression [28] and elastic net [55] lead to the same type of problem.

1.1 Related Work

Sinkhorn's, RSA or Balancing Type Methods. Special types of Problem (P_1), such as traffic matrix estimation and regularized optimal transport, have efficient matrix-scaling-based solvers such as balancing algorithm, [8], Sinkhorn's method, [12,46], RAS algorithm [30]. Strong points of these algorithms are fast convergence in practice and easy parallel implementation. At the same time, these algorithms are suitable only for Problem (P_1) with special type of linear equality constraints. A generalization for a problem with a special type of linear inequalities constraints was suggested in [6], but without convergence rate estimates. Recently, [11] extended the approach of [12] for other special classes of entropy-minimization problems.

The problem of instability of the matrix-scaling approach for problems with small regularization parameter was addressed in [44], but the proposed techniques are less suitable for parallel computations than the initial algorithm. There is a proof of linear convergence of the Sinkhorn's method [21], but the theoretical bound is much worse than the rate in practice and theoretical rate is obtained in terms of convergence in a special metric, which is hard to interpret. The papers [2,18,27,35] analyse complexity of the Sinkhorn's algorithm to find an approximate solution to the regularized and non-regularized optimal transport problem. In particular, they show that the regularization parameter needs to be of the order of the desired accuracy, which can lead to the instability of the Sinkhorn's algorithm. An alternative matrix scaling algorithm was proposed in [1] together with theoretical analysis, but this method seems to be hard to implement in practice and no experimental results were reported.

In any case, all the mentioned algorithms are designed for a special instance of Problem (P_1).

First-Order Methods for Constrained Problems. We consider Problem (P_1) in large-scale setting, when the natural choice is some first-order method. Due to the presence of linear constraints, the applicability of projected-gradient-type methods to the primal problem is limited. Thus, the most common approach involves construction of the dual problem and primal-dual updates during the algorithm progress. There are many algorithms of this type like ADMM [7,25] and other primal-dual methods [5,9,19], see the extensive review in [48]. As it is pointed in [48], these methods have the following drawbacks. They need the tractability assumption of the proximal operator for the function f and some additional assumptions. These methods don't have appropriate convergence rate characterization: if any, the rates are non-optimal and are either only for the dual problem or for some weighted sum of primal objective residual and linear constraints infeasibility. In [48], the authors themselves develop a good alternative, based only on the assumption of proximal tractability of the function f, but only for problems with linear equality constraints. This approach was further developed in [53] for more general types of constraints. The key feature of the algorithm developed there is its adaptivity to the unknown level of smoothness in the dual problem. Nevertheless, the provided stopping criterion, which is based on the prescribed number of iterations, requires to know all the smoothness parameters. Further, in [49], the authors propose algorithms with optimal rates of convergence for a more general class of problems, but, for the case of strongly convex f, they assume that it is strongly convex with respect to a Euclidean-type norm. Thus, their approach is not applicable to entropy minimization problems, which are our main focus.

An advanced ADMM with provable convergence rate with appropriate convergence characterization was proposed in [41], but only for the case of equality constraints and Lipschitz-smooth f, which does not cover the case of entropy minimization. A general primal-dual framework for unconstrained problems was proposed in [14], but it is not applicable in our setting. An adaptive to unknown Lipschitz constant algorithm for primal-dual problems was developed in [36], but the authors work with a different from our problem statement and the case of strongly convex objective is considered only in Euclidean setting, which also does not cover the case of entropy minimization.

Several recent algorithms [10,17,22,26,34,39,42] are based on the application of accelerated gradient method [37,38] to the dual problem and have optimal rates. At the same time, these works do not consider general types of constraints as in Problem (P_1). Also the proposed algorithms use, as an input parameter, an estimate of the Lipschitz constant of the gradient in the dual problem, which can be very pessimistic and lead to slow convergence.

The idea of primal-dual accelerated gradient methods turned out to be quite fruitful in the context of distributed decentralized optimization and it application to Wasserstein baeycenter problem [15,16,29,33,43,50,51].

1.2 Contributions

1. In contrast to the existing methods for constrained problems in [3,5,7,9,10,22,25,34,36,42,48,49,53], we propose an algorithm simultaneously for Problem (P_1) with general linear equality and cone constraints; with optimal rate of convergence in terms of both primal objective residual and constraints infeasibility; with adaptivity to the Lipschitz constant of the objective's gradient; with online stopping criterion, which does not require the knowledge of this Lipschitz constant; with ability to work with entropy function as f. The main difference with [18] is that here we consider more general cone constraints and consider not only regularized optimal transport problems as application.
2. In contrast to existing Sinkhorn's-algorithm-based algorithms for solving entropy-regularized optimal transport problems [1,2,6,8,12,30,44,46], we provide an algorithm simultaneously with provable convergence rate, easy implementability in practice and higher stability, when the regularization parameter is small.

3. In the experiments, we show that our algorithm is better than the Sinkhorn's method in situations of small regularization parameter in the primal problem, which means that the dual problem becomes less smooth problem.

The rest of the paper is organized as follows. In Sect. 2, we introduce notation, definition of approximate solution to Problem (P_1), main assumptions, and particular examples of (P_1) in applications. Section 3 is devoted to primal-dual algorithm for Problem (P_1) and its convergence analysis. Finally, in Sect. 4, we present the results of the numerical experiments for regularized optimal transport and traffic matrix estimation problems.

2 Preliminaries

For any finite-dimensional real vector space E, we denote by E^* its dual. We denote the value of a linear function $\lambda \in E^*$ at $x \in E$ by $\langle \lambda, x \rangle$. Let $\|\cdot\|_E$ denote some norm on E and $\|\cdot\|_{E,*}$ denote the norm on E^* which is dual to $\|\cdot\|_E$, i.e. $\|\lambda\|_{E,*} = \max_{\|x\|_E \leq 1} \langle \lambda, x \rangle$. In the special case, when E is a Euclidean space, we denote the standard Euclidean norm by $\|\cdot\|_2$. Note that, in this case, the dual norm is also Euclidean. For a cone $K \subseteq E$, the dual cone $K^* \subseteq E^*$ is defined as $K^* := \{\lambda \in E^* : \langle \lambda, x \rangle \geq 0 \quad \forall x \in K\}$. By $\partial f(x)$ we denote the subdifferential of a function $f(x)$ at a point x. Let E_1, E_2 be two finite-dimensional real vector spaces. For a linear operator $A : E_1 \to E_2$, we define its norm as follows

$$\|A\|_{E_1 \to E_2} = \max_{x \in E_1, u \in E_2^*} \{\langle u, Ax \rangle : \|x\|_{E_1} = 1, \|u\|_{E_2,*} = 1\}.$$

For a linear operator $A : E_1 \to E_2$, we define the adjoint operator $A^T : E_2^* \to E_1^*$ in the following way $\langle u, Ax \rangle = \langle A^T u, x \rangle, \quad \forall u \in E_2^*, \quad x \in E_1$. We say that a function $f : E \to \mathbb{R}$ has a L-Lipschitz-continuous gradient if it is differentiable and its gradient satisfies Lipschitz condition $\|\nabla f(x) - \nabla f(y)\|_{E,*} \leq L\|x - y\|_E, \quad \forall x, y \in E$. Note that, from this inequality, it follows that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_E^2, \quad \forall x, y \in E. \tag{1}$$

Also, for any $t \in \mathbb{R}$, we denote by $\lceil t \rceil$ the smallest integer greater than or equal to t.

We characterize the quality of an approximate solution to Problem (P_1) by three quantities $\varepsilon_f, \varepsilon_{eq}, \varepsilon_{in} > 0$.

Definition 1. *We say that a point $\hat{x}$ is an $(\varepsilon_f, \varepsilon_{eq}, \varepsilon_{in})$-solution to Problem (P_1) iff the following inequalities hold*

$$|f(\hat{x}) - Opt[P_1]| \leq \varepsilon_f, \quad \|A_1\hat{x} - b_1\|_2 \leq \varepsilon_{eq}, \quad \rho(A_2\hat{x} - b_2, -K) \leq \varepsilon_{in}. \tag{2}$$

Here $Opt[P_1]$ denotes the optimal function value for Problem (P_1),

$$\rho(A_2\hat{x} - b_2, -K) := \max_{\lambda^{(2)} \in K^*, \|\lambda^{(2)}\|_2 \leq 1} \langle \lambda^{(2)}, A_2\hat{x} - b_2 \rangle.$$

Note that the last inequality in (2) is a natural generalization of linear constraints infeasibility measure $\|(A_2x_k - b_2)_+\|_2$ for the case $K = \mathbb{R}^n_+$. Here the vector v_+ denotes the vector with components $[v_+]_i = (v_i)_+ = \max\{v_i, 0\}$.

The Lagrange dual problem to Problem (P_1) is

$$(D_1) \qquad \max_{\lambda \in \Lambda} \left\{ -\langle \lambda^{(1)}, b_1 \rangle - \langle \lambda^{(2)}, b_2 \rangle + \min_{x \in Q} \left(f(x) + \langle A_1^T \lambda^{(1)} + A_2^T \lambda^{(2)}, x \rangle \right) \right\}.$$

Here we denote $\Lambda = \{\lambda = (\lambda^{(1)}, \lambda^{(2)})^T \in H_1^* \times H_2^* : \lambda^{(2)} \in K^*\}$. It is convenient to rewrite Problem (D_1) in the equivalent form of a minimization problem

$$(P_2) \quad \min_{\lambda \in \Lambda} \left\{ \langle \lambda^{(1)}, b_1 \rangle + \langle \lambda^{(2)}, b_2 \rangle + \max_{x \in Q} \left(-f(x) - \langle A_1^T \lambda^{(1)} + A_2^T \lambda^{(2)}, x \rangle \right) \right\}.$$

It is obvious that

$$Opt[D_1] = -Opt[P_2], \tag{3}$$

where $Opt[D_1]$, $Opt[P_2]$ are the optimal function value in Problem (D_1) and Problem (P_2) respectively. The following inequality follows from the weak duality

$$Opt[P_1] \geq Opt[D_1]. \tag{4}$$

We denote

$$\begin{aligned} \varphi(\lambda) = \varphi(\lambda^{(1)}, \lambda^{(2)}) &= \langle \lambda^{(1)}, b_1 \rangle + \langle \lambda^{(2)}, b_2 \rangle \\ &+ \max_{x \in Q} \left(-f(x) - \langle A_1^T \lambda^{(1)} + A_2^T \lambda^{(2)}, x \rangle \right). \end{aligned} \tag{5}$$

Since f is strongly convex, $\varphi(\lambda)$ is a smooth function and its gradient is equal to (see e.g. [38])

$$\nabla \varphi(\lambda) = \begin{pmatrix} b_1 - A_1 x(\lambda) \\ b_2 - A_2 x(\lambda) \end{pmatrix}, \tag{6}$$

where $x(\lambda)$ is the unique solution of the strongly-convex problem

$$\max_{x \in Q} \left(-f(x) - \langle A_1^T \lambda^{(1)} + A_2^T \lambda^{(2)}, x \rangle \right). \tag{7}$$

Note that $\nabla \varphi(\lambda)$ is Lipschitz-continuous (see e.g. [38]) with constant

$$L \leq \frac{1}{\gamma} \left(\|A_1\|^2_{E \to H_1} + \|A_2\|^2_{E \to H_2} \right).$$

Previous works [10,22,34,42] rely on this quantity in the algorithm and use it to define the stepsize of the proposed algorithm. The drawback of this approach is that the above bound for the Lipschitz constant can be way too pessimistic. In this work, we propose an adaptive method, which has the same complexity bound, but is faster in practice due to the use of a “local” estimate for L in the stepsize definition.

We assume that the dual problem (D_1) has a solution $\lambda^* = (\lambda^{*(1)}, \lambda^{*(2)})^T$ and there exist some $R_1, R_2 > 0$ such that

$$\|\lambda^{*(1)}\|_2 \leq R_1 < +\infty, \quad \|\lambda^{*(2)}\|_2 \leq R_2 < +\infty. \tag{8}$$

It is worth noting that the quantities R_1, R_2 will be used only in the convergence analysis, but not in the algorithm itself.

To motivate the considered problem we describe two particular problems which can be written in the form of Problem (P_1).

Traffic demand matrix estimation, [52]**, and Regularized optimal transport problem,** [12].

$$\min_{X \in \mathbb{R}_+^{p \times p}} \left\{ \gamma \sum_{i,j=1}^{p} x_{ij} \ln x_{ij} + \sum_{i,j=1}^{p} c_{ij} x_{ij} : Xe = \mu, X^T e = \nu \right\}, \tag{9}$$

where $e \in \mathbb{R}^p$ is the vector of all ones, $\mu, \nu \in S_p(1) := \{x \in \mathbb{R}^p : \sum_{i=1}^p x_i = 1, x_i \geq 0, i = 1, ..., p\}$, $c_{ij} \geq 0, i, j = 1, ..., p$ are given, $\gamma > 0$ is the regularization parameter, X^T is the transpose matrix of X, x_{ij} is the element of the matrix X in the i-th row and the j-th column. This problem with small value of γ is our primary focus in this paper.

General entropy-linear programming problem, [20].

$$\min_{x \in S_n(1)} \left\{ \sum_{i=1}^{n} x_i \ln\left(x_i / \xi_i\right) : Ax = b \right\}$$

for some given $\xi \in \mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : x_i > 0, i = 1, ..., n\}$.

3 Primal-Dual Algorithm

In this section, we return to the primal-dual pair of problems (P_1)–(D_1). We apply Algorithm 1 in the supplementary of [18] to Problem (P_2) and incorporate in the algorithm a procedure, which allows to reconstruct also an approximate solution of Problem (P_1). The main novelty of this paper is the primal-dual analysis of this algorithm in the presence of inequality constraints. We choose Euclidean proximal setup, which means that we introduce euclidean norm $\|\cdot\|_2$ in the space of vectors λ and choose the prox-function $d(\lambda) = \frac{1}{2}\|\lambda\|_2^2$. Then, we have $V[\zeta](\lambda) = \frac{1}{2}\|\lambda - \zeta\|_2^2$.

Our primal-dual algorithm for Problem (P_1) is listed below as Algorithm 1. Note that, in this case, the set Λ has a special structure

$$\Lambda = \{\lambda = (\lambda^{(1)}, \lambda^{(2)})^T \in H_1^* \times H_2^* : \lambda^{(2)} \in K^*\}$$

as well as $\varphi(\lambda)$ and $\nabla\varphi(\lambda)$ are defined in (5) and (6) respectively. Thus, the step (10) of the algorithm can be written explicitly.

$$\zeta_{k+1}^{(1)} = \zeta_k^{(1)} + \alpha_{k+1}(A_1 x(\lambda_{k+1}) - b_1), \quad \zeta_{k+1}^{(2)} = \Pi_{K^*}\left(\zeta_k^{(2)} + \alpha_{k+1}(A_2 x(\lambda_{k+1}) - b_2)\right),$$

where $\Pi_{K^*}(\cdot)$ denotes euclidean projection on the cone K^*.

It is worth noting that, besides solution of the problem (7), the algorithm uses only matrix-vector multiplications and vector operations, which made it amenable for parallel implementation.

Algorithm 1. Primal-Dual Adaptive Similar Triangles Method (PDASTM)

```
Require: starting point $\lambda_0 = 0$, initial guess $L_0 > 0$, accuracy $\tilde{\varepsilon}_f, \tilde{\varepsilon}_{eq}, \tilde{\varepsilon}_{in} > 0$.
Set $k = 0$, $C_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
repeat
    Set $M_k = L_k/2$.
    repeat
        Set $M_k = 2M_k$, find $\alpha_{k+1}$ as the largest root of the equation $C_{k+1} := C_k + \alpha_{k+1} = M_k \alpha_{k+1}^2$.
        Calculate $\lambda_{k+1} = (\lambda_{k+1}^{(1)}, \lambda_{k+1}^{(2)})^T = (\alpha_{k+1}\zeta_k + C_k\eta_k)/C_{k+1}$.
        Calculate
            $\zeta_{k+1} = (\zeta_{k+1}^{(1)}, \zeta_{k+1}^{(2)})^T = \arg\min_{\lambda \in \Lambda} \left\{ \frac{1}{2}\|\lambda - \zeta_k\|_2^2 + \alpha_{k+1}(\varphi(\lambda_{k+1}) + \langle \nabla\varphi(\lambda_{k+1}), \lambda - \lambda_{k+1} \rangle) \right\}.$   (10)
        Calculate $\eta_{k+1} = (\eta_{k+1}^{(1)}, \eta_{k+1}^{(2)})^T = (\alpha_{k+1}\zeta_{k+1} + C_k\eta_k)/C_{k+1}$.
    until
        $\varphi(\eta_{k+1}) \leq \varphi(\lambda_{k+1}) + \langle \nabla\varphi(\lambda_{k+1}), \eta_{k+1} - \lambda_{k+1} \rangle + \frac{M_k}{2}\|\eta_{k+1} - \lambda_{k+1}\|_2^2.$   (11)
    Set $\hat{x}_{k+1} = \frac{1}{C_{k+1}} \sum_{i=0}^{k+1} \alpha_i x(\lambda_i) = (\alpha_{k+1} x(\lambda_{k+1}) + C_k \hat{x}_k)/C_{k+1}$.
    Set $L_{k+1} = M_k/2$, $k = k + 1$.
until $|f(\hat{x}_{k+1}) + \varphi(\eta_{k+1})| \leq \tilde{\varepsilon}_f$, $\|A_1\hat{x}_{k+1} - b_1\|_2 \leq \tilde{\varepsilon}_{eq}$, $\rho(A_2\hat{x}_{k+1} - b_2, -K) \leq \tilde{\varepsilon}_{in}$.
Ensure: The points $\hat{x}_{k+1}$, $\eta_{k+1}$.
```

Theorem 1. *Let the main assumptions hold. Then Algorithm 1 will stop not later than k equals to*

$$\max\left\{ \left\lceil \sqrt{\frac{16L(R_1^2 + R_2^2)}{\tilde{\varepsilon}_f}} \right\rceil, \left\lceil \sqrt{\frac{16L(R_1^2 + R_2^2)}{R_1\tilde{\varepsilon}_{eq}}} \right\rceil, \left\lceil \sqrt{\frac{16L(R_1^2 + R_2^2)}{R_2\tilde{\varepsilon}_{in}}} \right\rceil \right\}.$$

Moreover, no later than k equals to

$$\max\left\{ \left\lceil \sqrt{\frac{32L(R_1^2 + R_2^2)}{\varepsilon_f}} \right\rceil, \left\lceil \sqrt{\frac{16L(R_1^2 + R_2^2)}{R_1\varepsilon_{eq}}} \right\rceil, \left\lceil \sqrt{\frac{16L(R_1^2 + R_2^2)}{R_2\varepsilon_{in}}} \right\rceil \right\},$$

the point $\hat{x}_{k+1}$ generated by Algorithm 1 is an approximate solution to Problem (P_1) in the sense of (2) and $\|\hat{x}_{k+1} - x^*\| \leq \sqrt{\frac{2\varepsilon_f}{\gamma}}$, where x^* is a solution to Problem (P_1).

Remark 1. Note that the result of Theorem 1 can be reformulated as follows. For any $k \geq 1$, the output $(\hat{x}_k, \eta_k)$ of Algorithm 1 satisfies

$$
\begin{gathered}
-\frac{16L(R_1^2 + R_2^2)}{(k+1)^2} \leq f(\hat{x}_k) - Opt[P_1] \leq f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{16L(R_1^2 + R_2^2)}{(k+1)^2}, \\
\|A_1 \hat{x}_k - b_1\|_2 \leq \frac{16L(R_1^2 + R_2^2)}{R_1(k+1)^2}, \quad \rho(A_2 \hat{x}_k - b_2, -K) \leq \frac{16L(R_1^2 + R_2^2)}{R_2(k+1)^2}, \\
\|\hat{x}_k - x^*\|_E \leq \frac{8}{k+1} \sqrt{\frac{L(R_1^2 + R_2^2)}{\gamma}}.
\end{gathered}
$$

4 Numerical Experiments

In this section, we focus on the problem (9), which is motivated by important applications to traffic demand matrix estimation, [52], and regularized optimal transport calculation, [12]. We provide the results of our numerical experiments, which were performed on a PC with processor Intel Core i5-2410 2.3 GHz and 4 GB of RAM using pure Python 2.7 (without C code) under managing OS Ubuntu 14.04 (64-bits). Numpy.float128 data type with precision $1e{-}18$ and with max element $\approx 1.19e{+}4932$ was used. No parallel computations were used. We compare the performance of our algorithm with Sinkhorn's-method-based approach of [12], which is the state-of-the art method for problem (9). We use two types of cost matrix C and three types of vectors μ and ν.

Cost Matrix C. The first type of the cost matrix C is usually used in optimal transport problems and corresponds to 2-Wasserstein distance. Assume that we need to calculate this distance between two discrete measures μ, ν with finite support of size p. Then, the element c_{ij} of the matrix C is equal to Euclidean distance between the i-th point in the support of the measure μ and j-th point in the support of the measure ν. We will refer to this choice of the cost matrix as *Euclidean cost.* The second type the cost matrix C comes from traffic matrix estimation problem. Let's consider a road network of Manhattan type, i.e. districts present a $m \times m$ grid. We build a m^2 by m^2 matrix D of pairwise Euclidian distances processing the grid rows one by one and calculating euclidean distances from the current grid element to all the others elements of the grid. Then, as it suggested in [45], we form the cost matrix C as $C = \exp(-0.065D)$, where the exponent is taken elementwise. We will refer to this choice of the cost matrix as *Exp-Euclidean cost.*

To set a natural scale for the regularization parameter γ, we normalize in each case the matrix C dividing all its elements by the average of all elements.

Vectors μ and ν. The first type of vectors μ and ν is *normalized uniform random.* Each element of each vector is taken independently from the uniform distribution on $[0, 1]$ and then each vector is normalized so that each sums to 1, i.e.

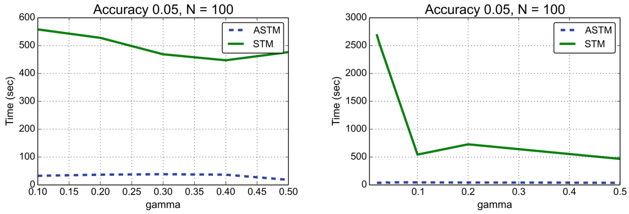
The second type of vectors is *random images*. The first $p/2$ elements of μ are normalized uniform random and the second $p/2$ elements are zero. For ν the situation is the opposite, i.e. the first $p/2$ elements are zero, and the second $p/2$ elements are normalized uniform random. In our preliminary experiments we found that the methods behave strange on vectors representing pictures from MNIST dataset. We supposed that the reason is that these vectors have many zero elements and decided to include the described random images to the experiments setting. Finally, the third type are vectors of intensities of *images* of handwritten digits from MNIST dataset. The size of each image is 28 by 28 pixels. Each image is converted to gray scale from 0 to 1 where 0 corresponds to black color and 1 corresponds to white, then each image is reshaped to a vector of length 784. In our experiments, we normalize these vectors to sum to 1.

Accuracy. We slightly redefine the accuracy of the solution and use relative accuracy with respect to the starting point, i.e.

$$\tilde{\varepsilon}_f = [\text{Accuracy}] \cdot |f(x(\lambda_0)) + \varphi(\eta_0)|, \quad \tilde{\varepsilon}_{eq} = [\text{Accuracy}] \cdot \|A_1 x(\lambda_0) - b_1\|_2,$$

where we used the fact that $\lambda_0 = \eta_0 = 0$ and there are no cone constraints in (9).

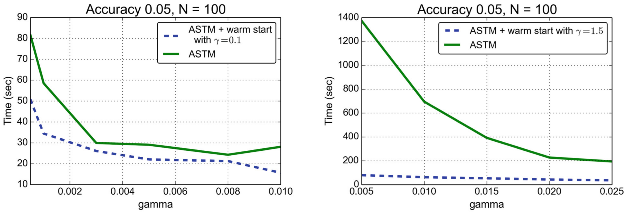
Adaptive vs Non-adaptive Algorithm. First, we show that the adaptivity of our algorithm with respect to the Lipschitz constant of the gradient of φ leads to faster convergence in practice. For this purpose, we use normalized uniform random vectors μ and ν and both types of cost matrix C. We compare our new Algorithm 1 with non-adaptive Similar Triangles Method (STM), which has cheaper iteration than the existing non-adaptive methods [10,22,34,42]. We choose $m = 10$, and, hence, $p = 100$, Accuracy is 0.05. For the Exp-Euclidean cost matrix C, we use $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and, for Euclidean cost matrix C, we use $\gamma \in \{0.02, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The results are shown in Fig. 1. In both cases our new Algorithm 1 is much faster than the STM. This effect was observed for other parameter values, so, in the following experiments, we consider PDASTM.



(a) (b)

Fig. 1. The perfomance of PDASTM vs STM, Accuracy 0.05, Exp-Euclidean C (left) and Euclidean C (right).

Warm Start. During our experiments on the images from MNIST dataset PDASTM worked worse than on the normalized uniform random vectors. Possible reason is the large number of zero elements in the former vectors (a lot of black pixels). So we decided to test the performance of the algorithms on the random images vectors μ and ν. Also we decided to apply the idea of warm start to force PDASTM to converge faster. As we know, Sinkhorn's method works very fast when γ is relatively large. Thus, we use it in this regime to find a good starting point for the PDASTM for the problem with small γ. Notably, the running time of Sinkhorn's method is small in comparison with time of ASTM running. We test the performance of PDASTM versus PDASTM with warm start on problems with Exp-Euclidean matrix C and $\gamma \in \{0.001, 0.003, 0.005, 0.008, 0.01\}$ and on problems with Euclidean matrix C and $\gamma \in \{0.005, 0.01, 0.015, 0.02, 0.025\}$. The results are in Fig. 2. Other parameters are stated in the figure. The experiments were run 7 times, the results were averaged. As we can see, warm start accelerates the PDASTM. Similar results were observed in other experiments, so, we made the final comparison between the Sinkhorn's method and PDASTM with warm start.



(a) (b)

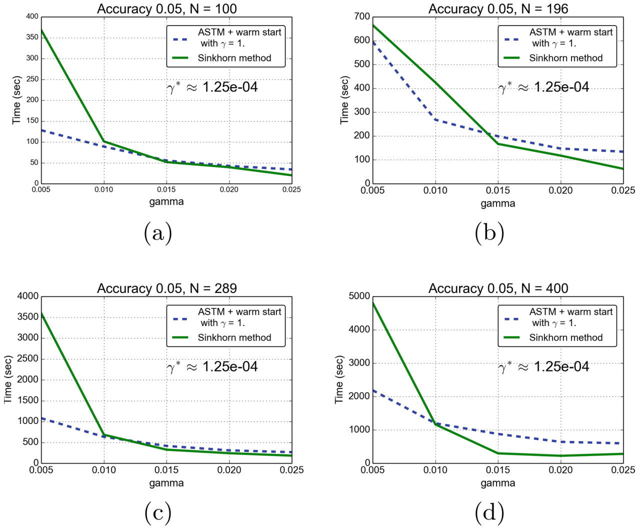
Fig. 2. The perfomance of PDASTM vs PDASTM with warm start, Accuracy 0.05, Exp-Euclidean C (left) and Euclidean C (right).

4.1 Sinkhorn's Method vs PDASTM with Warm Start

First we compare Sinkhorn's method and PDASTM with warm start on the problem with normalized uniform random vectors μ, ν and Euclidean cost matrix C with different values of $p \in \{100, 196, 289, 400\}$, Accuracy $\in \{0.01, 0.05, 0.1\}$, and $\gamma \in [0.005; 0.025]$. On each graph we point the value of γ used for generating a starting point for PDASTM with warm start by Sinkhorn's method. Each experiments was run 5 times and then the results were averaged. The results are shown on the Figs. 3, 4.

For the Exp-Euclidean cost matrix C, we performed the same experiments. For the space reasons, we provide the results on the Fig. 5 only for Accuracy 0.05. The results for other Accuracy values were similar.

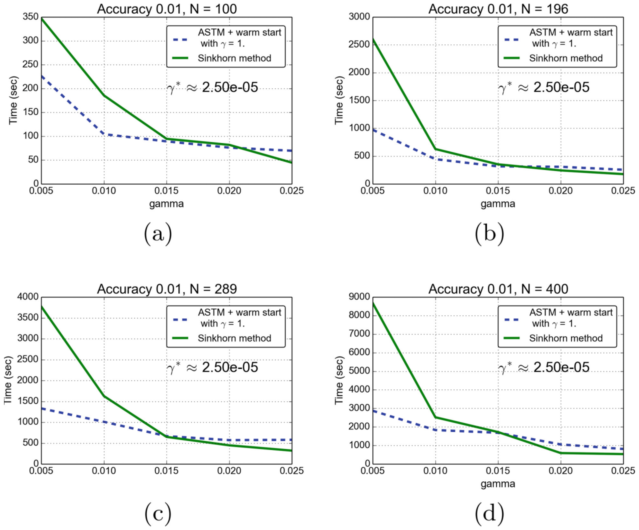
In another series of experiments we compare the performance of PDASTM with warm start and Sinkhorn's method on the problem with images from MNIST dataset and Euclidean cost matrix C. We run both algorithms for the same set of γ values for 5 pairs of images. The results are aggregated by γ and the performance is averaged for each γ. We take three values of Accuracy, $\{0.01, 0.05, 0.1\}$. The results are shown on the Fig. 6.



(a) (b)

(c) (d)

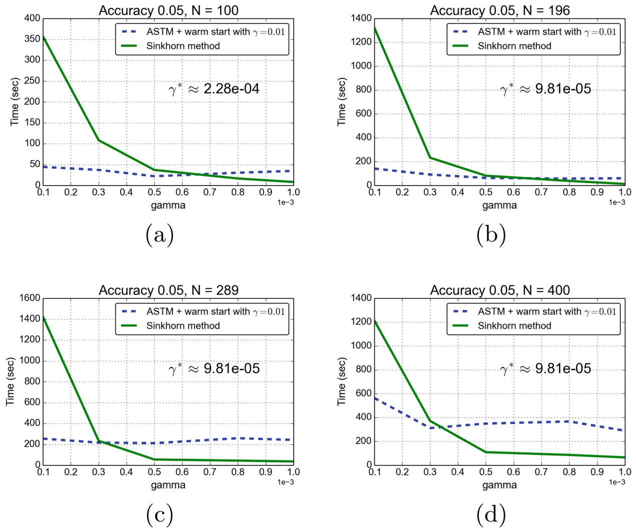
Fig. 3. The perfomance of PDASTM with warm start vs Sinkhorn's method, Accuracy 0.05, Euclidean cost matrix C.



(a) (b)

(c) (d)

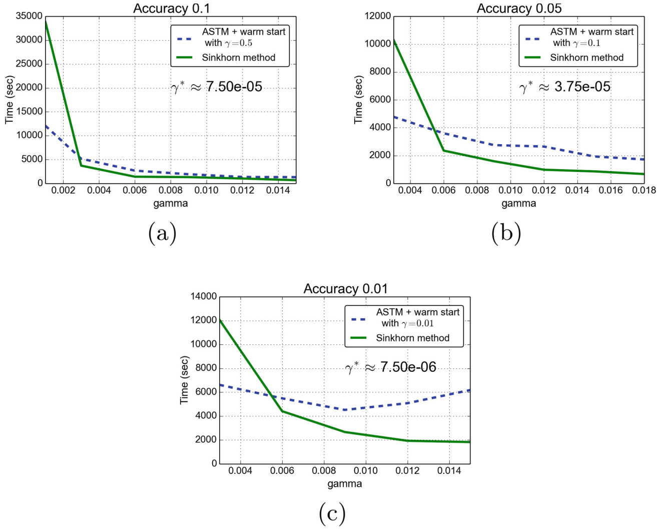
Fig. 4. The perfomance of PDASTM with warm start vs Sinkhorn's method, Accuracy 0.01, Euclidean cost matrix C.



(a) (b)

(c) (d)

Fig. 5. The perfomance of PDASTM with warm start vs Sinkhorn's method, Accuracy 0.05, Exp-Euclidean cost matrix C.



(a) (b)

(c)

Fig. 6. The perfomance of PDASTM with warm start vs Sinkhorn's method, Euclidean cost matrix C, MNIST dataset.

As we can see on all graphs, for small values of γ, namely, smaller than some threshold γ_0, our method outperforms the state-of-the-art Sinkhorn's method. Note that, from [38], it follows that, for very small values of γ, less than some threshold $\gamma* = \frac{\varepsilon}{4 \ln p}$, a good approximation of the solution to the problem (9) can be obtained by solution of the linear programming problem corresponding to $\gamma = 0$. We point these thresholds $\gamma*$ on the figures above. It should be noted that the threshold γ_0 is larger than $\gamma*$. This means that it is better to use our method, but not some method for linear programing problems.

Finally, we investigate the dependence of running time of PDASTM with warm start on the problem dimension p. As we can see from the Fig. 7, the dependence is close to quadratic, which was expected from the theoretical bounds. Also this dependence is close to that of the Sinkhorn's method.

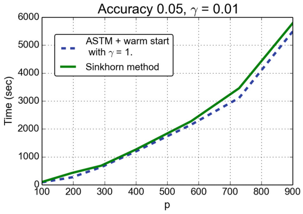


Fig. 7. Dependence of running time from the problem dimension p.

Conclusion

In this article, we propose a new adaptive accelerated gradient method for convex optimization problems and prove its convergence rate. We apply this method to a class of linearly constrained problems and show, how an approximate solution can be reconstructed. In the experiments, we consider two particular applied problems, namely, regularized optimal transport problem and traffic matrix estimation problem. The results of the experiments show that, in the regime of small regularization parameter, our algorithm outperforms the state-of-the-art Sinkhorn's-method-based approach. It would be interesting to extend the adaptive primal-dual methods for the stochastic setting [40] and for problems with inexact model of the objective [47].

References

1. Allen-Zhu, Z., Li, Y., Oliveira, R., Wigderson, A.: Much faster algorithms for matrix scaling. In: 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pp. 890–901 (2017). arXiv:1704.02315
2. Altschuler, J., Weed, J., Rigollet, P.: Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30, pp. 1961–1971. Curran Associates, Inc. (2017). arXiv:1705.09634

3. Anikin, A.S., Gasnikov, A.V., Dvurechensky, P.E., Tyurin, A.I., Chernov, A.V.: Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints. Comput. Math. Math. Phys. **57**(8), 1262–1276 (2017)
4. Baimurzina, D.R., et al.: Universal method of searching for equilibria and stochastic equilibria in transportation networks. Comput. Math. Math. Phys. **59**(1), 19–33 (2019). arXiv:1701.02473
5. Beck, A., Teboulle, M.: A fast dual proximal gradient algorithm for convex minimization and applications. Oper. Res. Lett. **42**(1), 1–6 (2014)
6. Benamou, J.D., Carlier, G., Cuturi, M., Nenna, L., Peyré, G.: Iterative Bregman projections for regularized transportation problems. SIAM J. Sci. Comput. **37**(2), A1111–A1138 (2015)
7. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
8. Bregman, L.: Proof of the convergence of Sheleikhovskii's method for a problem with transportation constraints. USSR Comput. Math. Math. Phys. **7**(1), 191–204 (1967)
9. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
10. Chernov, A., Dvurechensky, P., Gasnikov, A.: Fast primal-dual gradient method for strongly convex minimization problems with linear constraints. In: Kochetov, Y., Khachay, M., Beresnev, V., Nurminski, E., Pardalos, P. (eds.) DOOR 2016. LNCS, vol. 9869, pp. 391–403. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44914-2_31
11. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Scaling algorithms for unbalanced optimal transport problems. Math. Comput. **87**(314), 2563–2609 (2018). arXiv:1607.05816
12. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 2292–2300. Curran Associates, Inc. (2013)
13. Cuturi, M., Peyré, G.: A smoothed dual approach for variational Wasserstein problems. SIAM J. Imaging Sci. **9**(1), 320–343 (2016)
14. Dünner, C., Forte, S., Takáč, M., Jaggi, M.: Primal-dual rates and certificates. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning, ICML 2016, vol. 48. pp. 783–792. JMLR.org (2016)
15. Dvinskikh, D., Gorbunov, E., Gasnikov, A., Dvurechensky, P., Uribe, C.A.: On primal and dual approaches for distributed stochastic convex optimization over networks. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 7435–7440 (2019). https://doi.org/10.1109/CDC40024.2019.9029798. arXiv:1903.09844
16. Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C.A., Nedić, A.: Decentralize and randomize: faster algorithm for Wasserstein barycenters. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, NeurIPS 2018, pp. 10783–10793. Curran Associates, Inc. (2018). arXiv:1806.03915

17. Dvurechensky, P., Gasnikov, A., Gasnikova, E., Matsievsky, S., Rodomanov, A., Usik, I.: Primal-dual method for searching equilibrium in hierarchical congestion population games. In: Supplementary Proceedings of the 9th International Conference on Discrete Optimization and Operations Research and Scientific School (DOOR 2016) Vladivostok, Russia, 19–23 September 2016, pp. 584–595 (2016). arXiv:1606.08988
18. Dvurechensky, P., Gasnikov, A., Kroshnin, A.: Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1367–1376 (2018). arXiv:1802.04367
19. Dvurechensky, P., Nesterov, Y., Spokoiny, V.: Primal-dual methods for solving infinite-dimensional games. J. Optim. Theory Appl. **166**(1), 23–51 (2015)
20. Fang, S.-C., Rajasekera, J. R., Tsao, H.-S. J.: Entropy Optimization and Mathematical Programming. Kluwer' International Series. Springer, Boston (1997)
21. Franklin, J., Lorenz, J.: On the scaling of multidimensional matrices. Linear Algebra Appl. **114**, 717–735 (1989). Special Issue Dedicated to Alan J. Hoffman
22. Gasnikov, A.V., Gasnikova, E.V., Nesterov, Y.E., Chernov, A.V.: Efficient numerical methods for entropy-linear programming problems. Comput. Math. Math. Phys. **56**(4), 514–524 (2016)
23. Gasnikov, A., Gasnikova, E., Mendel, M., Chepurchenko, K.: Evolutionary derivations of entropy model for traffic demand matrix calculation. Matematicheskoe Modelirovanie **28**(4), 111–124 (2016). (in Russian)
24. Golan, A., Judge, G., Miller, D.: Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley, Chichester (1996)
25. Goldstein, T., O'Donoghue, B., Setzer, S., Baraniuk, R.: Fast alternating direction optimization methods. SIAM J. Imaging Sci. **7**(3), 1588–1623 (2014)
26. Guminov, S.V., Nesterov, Y.E., Dvurechensky, P.E., Gasnikov, A.V.: Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. Dokl. Math. **99**(2), 125–128 (2019)
27. Guminov, S., Dvurechensky, P., Tupitsa, N., Gasnikov, A.: Accelerated alternating minimization, accelerated Sinkhorn's algorithm and accelerated Iterative Bregman Projections (2019). arXiv:1906.03622
28. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York (2001). https://doi.org/10.1007/978-0-387-21606-5
29. Jakovetić, D., Xavier, J., Moura, J.M.F.: Fast distributed gradient methods. IEEE Trans. Autom. Control **59**(5), 1131–1146 (2014)
30. Kalantari, B., Khachiyan, L.: On the rate of convergence of deterministic and randomized RAS matrix scaling algorithms. Oper. Res. Lett. **14**(5), 237–244 (1993)
31. Kantorovich, L.: On the translocation of masses. Doklady Acad. Sci. USSR (N.S.) **37**, 199–201 (1942)
32. Kapur, J.: Maximum – Entropy Models in Science and Engineering. Wiley, New York (1989)
33. Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Uribe, C.: On the complexity of approximating Wasserstein barycenters. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, Long Beach, California, USA, 09–15 June 2019, vol. 97, pp. 3530–3540. PMLR (2019). arXiv:1901.08686

34. Li, J., Wu, Z., Wu, C., Long, Q., Wang, X.: An inexact dual fast gradient-projection method for separable convex optimization with linear coupled constraints. J. Optim. Theory Appl. **168**(1), 153–171 (2016)
35. Lin, T., Ho, N., Jordan, M.: On efficient optimal transport: an analysis of greedy and accelerated mirror descent algorithms. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, Long Beach, California, USA, 09–15 June 2019, vol. 97, pp. 3982–3991. PMLR (2019)
36. Malitsky, Y., Pock, T.: A first-order primal-dual algorithm with linesearch. SIAM J. Optim. **28**(1), 411–432 (2018)
37. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Boston (2004)
38. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
39. Nesterov, Y., Gasnikov, A., Guminov, S., Dvurechensky, P.: Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. Optim. Methods Softw., 1–28 (2020). https://doi.org/10.1080/10556788.2020.1731747. arXiv:1809.05895
40. Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Spokoiny, V.: Adaptive gradient descent for convex and non-convex stochastic optimization (2019). arXiv:1911.08380
41. Ouyang, Y., Chen, Y., Lan, G., Eduardo Pasiliao, J.: An accelerated linearized alternating direction method of multipliers. SIAM J. Imaging Sci. **8**(1), 644–681 (2015)
42. Patrascu, A., Necoara, I., Findeisen, R.: Rate of convergence analysis of a dual fast gradient method for general convex optimization. In: 2015 54th IEEE Conference on Decision and Control (CDC), pp. 3311–3316 (2015)
43. Scaman, K., Bach, F., Bubeck, S., Lee, Y.T., Massoulié, L.: Optimal algorithms for smooth and strongly convex distributed optimization in networks. In: Precup, A., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, International Convention Centre, Sydney, Australia, 06–11 August 2017, pp. 3027–3036. PMLR (2017)
44. Schmitzer, B.: Stabilized sparse scaling algorithms for entropy regularized transport problems. SIAM J. Sci. Comput. **41**(3), A1443–A1481 (2019). arXiv:1610.06519
45. Shvetsov, V.I.: Mathematical modeling of traffic flows. Autom. Remote Control **64**(11), 1651–1689 (2003)
46. Sinkhorn, R.: Diagonal equivalence to matrices with prescribed row and column sums. II. Proc. Am. Math. Soc. **45**, 195–198 (1974)
47. Stonyakin, F.S., et al.: Gradient methods for problems with inexact model of the objective. In: Khachay, M., Kochetov, Y., Pardalos, P. (eds.) MOTOR 2019. LNCS, vol. 11548, pp. 97–114. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22629-9_8. arXiv:1902.09001
48. Tran-Dinh, Q., Cevher, V.: Constrained convex minimization via model-based excessive gap. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS 2014, pp. 721–729. MIT Press, Cambridge (2014)
49. Tran-Dinh, Q., Fercoq, O., Cevher, V.: A smooth primal-dual optimization framework for nonsmooth composite convex minimization. SIAM J. Optim. **28**(1), 96–134 (2018). arXiv:1507.06243

50. Tupitsa, N., Dvurechensky, P., Gasnikov, A., Uribe, C.A.: Multimarginal optimal transport by accelerated gradient descent (2020). arXiv:2004.02294
51. Uribe, C.A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Nedić, A.: Distributed computation of Wasserstein barycenters over networks. In: 2018 IEEE Conference on Decision and Control (CDC), pp. 6544–6549 (2018). arXiv:1803.02933
52. Wilson, A.: Entropy in Urban and Regional Modelling. Monographs in Spatial and Environmental Systems Analysis. Routledge, Abingdon (2011)
53. Yurtsever, A., Tran-Dinh, Q., Cevher, V.: A universal primal-dual convex optimization framework. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS 2015, pp. 3150–3158. MIT Press, Cambridge (2015)
54. Zhang, Y., Roughan, M., Lund, C., Donoho, D.L.: Estimating point-to-point and point-to-multipoint traffic matrices: an information-theoretic approach. IEEE/ACM Trans. Netw. **13**(5), 947–960 (2005)
55. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B **67**(2), 301–320 (2005)