Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints

Alexey Chernov[1], Pavel Dvurechensky[2,3(✉)], and Alexander Gasnikov[4,5]

[1] Moscow Institute of Physics and Technology, Dolgoprudnyi 141700, Moscow Oblast, Russia
alexmipt@mail.ru

[2] Weierstrass Institute for Applied Analysis and Stochastics, 10117 Berlin, Germany
pavel.dvurechensky@wias-berlin.de

[3] Institute for Information Transmission Problems, Moscow 127051, Russia

[4] Moscow Institute of Physics and Technology, Dolgoprudnyi 141700, Moscow Oblast, Russia

[5] Institute for Information Transmission Problems, Moscow 127051, Russia
gasnikov@yandex.ru

Abstract. In this paper, we consider a class of optimization problems with a strongly convex objective function and the feasible set given by an intersection of a simple convex set with a set given by a number of linear equality and inequality constraints. Quite a number of optimization problems in applications can be stated in this form, examples being entropy-linear programming, ridge regression, elastic net, regularized optimal transport, etc. We extend the Fast Gradient Method applied to the dual problem in order to make it primal-dual, so that it allows not only to solve the dual problem, but also to construct nearly optimal and nearly feasible solution of the primal problem. We also prove a theorem about the convergence rate for the proposed algorithm in terms of the objective function residual and the linear constraints infeasibility.

Keywords: Convex optimization · Algorithm complexity · Entropy-linear programming · Dual problem · Primal-dual method

1 Introduction

In this paper, we consider a constrained convex optimization problem of the following form

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

where E is a finite-dimensional real vector space, Q is a simple closed and 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, $f(x)$ is a

Y. Kochetov et al. (Eds.): DOOR 2016, LNCS 9869, pp. 391–403, 2016.
DOI: 10.1007/978-3-319-44914-2_31

ν-strongly convex function on Q with respect to some chosen norm $\|\cdot\|_E$ on E. The last means that, for any $x, y \in Q$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\nu}{2}\|x - y\|_E^2$, 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 $g \in E^*$ at $x \in E$ by $\langle g, x \rangle$.

Problem (P_1) captures a broad set of optimization problems arising in applications. The first example is the classical entropy-linear programming (ELP) problem [1] which arises in different fields, such as econometrics [2], modeling in science and engineering [3], especially in modeling of traffic flows [4] and IP traffic matrix estimation [5,6]. Other examples are the ridge regression problem [7] and the elastic net approach [8], which are used in machine learning. Finally, the problem class (P_1) covers problems of regularized optimal transport (ROT) [9] and regularized optimal partial transport (ROPT) [10], which recently have become popular in application to image analysis.

Classical balancing algorithms such as [9,11,12] are very efficient for solving ROT problems or special types of ELP problem, but they can deal only with linear equality constraints of a special type and their rate of convergence estimates are rather impractical [13]. In [10], the authors provide a generalization, but only for ROPT problems which are a particular case of Problem (P_1) with linear inequalities constraints of a special type, and no convergence rate estimates are provided. Unfortunately, the existing balancing-type algorithms for ROT and ROPT problems become very unstable when the regularization parameter is chosen very small, which is the case when one needs to calculate a good approximation to the solution of an optimal transport (OT) or an optimal partial transport (OPT) problem.

In practice, typical dimensions of the spaces E, H_1, H_2 range from thousands to millions, which makes it natural to use a first-order method to solve Problem (P_1). A common approach to solve such large-scale Problem (P_1) is to make the transition to the Lagrange dual problem and solve it by some first-order method. Unfortunately, the existing methods, which elaborate this idea, have at least two drawbacks. Firstly, the convergence analysis of the Fast Gradient Method (FGM) [14] can not be directly applied since it is based on the assumption of boundedness of the feasible set in both the primal and the dual problem, which does not hold for the Lagrange dual problem. A possible way to overcome this obstacle is to assume that the solution of the dual problem is bounded and add some additional constraints to the Lagrange dual problem in order to make the dual feasible set bounded. But, in practice, the bound for the solution of the dual problem is usually unknown. In [15], the authors use this approach with additional constraints and propose a restart technique to define the unknown bound for the optimal dual variable value. The authors consider classical ELP problems only with equality constraints and do not discuss any possibility of application of their technique to Problem (P_1) with inequality constraints. Secondly, it is important to estimate the rate of convergence not only in terms of the error in the solution of the Lagrange dual problem, as it is done in [16,17],

but also in terms of the objective residual in the primal problem[1] $|f(x_k) - Opt[P_1]|$ and the linear constraints infeasibility $\|A_1 x_k - b_1\|_{H_1}$, $\|(A_2 x_k - b_2)_+\|_{H_2}$, where vector v_+ denotes the vector with components $[v_+]_i = (v_i)_+ = \max\{v_i, 0\}$, x_k is the output of the algorithm on the k-th iteration, $Opt[P_1]$ denotes the optimal function value for Problem (P_1). Alternative approaches [18,19], based on the idea of the method of multipliers, and the quasi-Newton methods such as L-BFGS also do not allow to obtain the convergence rate for the primal problem residual and the linear constraints infeasibility.

Our contributions in this work are the following. We extend the Fast Gradient Method [14,20], applied to the dual problem, in order to make it primal-dual, so that it allows not only to solve the dual problem, but also to construct nearly optimal and nearly feasible solution to the primal problem (P_1). We also equip our method with a stopping criterion, which allows an online control of the quality of the approximate primal-dual solution. Unlike [9,10,15–19], we provide the estimates for the rate of convergence in terms of the primal objective residual $|f(x_k) - Opt[P_1]|$ and the linear constraints infeasibility $\|A_1 x_k - b_1\|_{H_1}$, $\|(A_2 x_k - b_2)_+\|_{H_2}$. In the contrast to the estimates in [14], our estimates do not rely on the assumption that the feasible set of the dual problem is bounded. At the same time, our approach is applicable for the wider class of problems defined by (P_1) than the approaches in [9,15]. In the computational experiments, we show that our approach allows to solve ROT problems more efficiently than the algorithms of [9,10,15] when the regularization parameter is small.

2 Preliminaries

2.1 Notation

For any finite-dimensional real vector space E, we denote by E^* its dual. We denote the value of a linear function $g \in E^*$ at $x \in E$ by $\langle g, x \rangle$. Let $\|\cdot\|_E$ denote some norm on E and $\|\cdot\|_{E,*}$ denote the norm on E^*, which is dual to $\|\cdot\|_E$

$$\|g\|_{E,*} = \max_{\|x\|_E \leq 1} \langle g, x \rangle.$$

In the special case of a Euclidean space E, we denote the standard Euclidean norm by $\|\cdot\|_2$. Note that, in this case, the dual norm is also Euclidean. 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

$$\langle u, Ax \rangle = \langle A^T u, x \rangle, \quad \forall u \in E_2^*, \quad x \in E_1.$$

[1] The absolute value here is crucial since x_k may not satisfy linear constraints and, hence, $f(x_k) - Opt[P_1]$ could be negative.

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

$$\|\nabla f(x) - \nabla f(y)\|_{E,*} \leq L\|x - y\|_E.$$

We characterize the quality of an approximate solution to Problem (P_1) by three quantities $\varepsilon_f, \varepsilon_{eq}, \varepsilon_{in} > 0$ and say that a point $\hat{x}$ is an $(\varepsilon_f, \varepsilon_{eq}, \varepsilon_{in})$-solution to Problem (P_1) if the following inequalities hold

$$|f(\hat{x}) - Opt[P_1]| \leq \varepsilon_f, \quad \|A_1\hat{x} - b_1\|_2 \leq \varepsilon_{eq}, \quad \|(A_2\hat{x} - b_2)_+\|_2 \leq \varepsilon_{in}, \tag{1}$$

where $Opt[P_1]$ denotes the optimal function value for Problem (P_1) and, for any vector v, the vector v_+ denotes the vector with components $[v_+]_i = (v_i)_+ = \max\{v_i, 0\}$. Also, for any $t \in R$, we denote by $\lceil t \rceil$ the smallest integer greater than or equal to t.

2.2 Dual Problem

Let us denote $\Lambda = \{\lambda = (\lambda^{(1)}, \lambda^{(2)})^T \in H_1^* \times H_2^* : \lambda^{(2)} \geq 0\}$. The Lagrange dual problem to Problem (P_1) is

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

We rewrite Problem (D_1) in the equivalent form of a minimization problem.

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

We denote

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

Note that the gradient of the function $\varphi(\lambda)$ is equal to (see e.g. [14])

$$\nabla \varphi(\lambda) = \begin{pmatrix} b_1 - A_1 x(\lambda) \\ b_2 - A_2 x(\lambda) \end{pmatrix}, \tag{3}$$

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

$$\max_{x \in Q} \left(-f(x) - \langle A_1^T \lambda^{(1)} + A_2^T \lambda^{(2)}, x \rangle \right). \tag{4}$$

It is important that $\nabla \varphi(\lambda)$ is Lipschitz-continuous (see e.g. [14]) with the constant

$$L = \frac{1}{\nu} \left(\|A_1\|^2_{E \to H_1} + \|A_2\|^2_{E \to H_2} \right). \tag{5}$$

Obviously, we have

$$Opt[D_1] = -Opt[P_2], \tag{6}$$

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

$$Opt[P_1] \geq Opt[D_1]. \tag{7}$$

2.3 Main Assumptions

We make the following two main assumptions

1. The problem (4) is simple in the sense that for any $x \in Q$ it has a closed form solution or can be solved very fast up to a machine precision.
2. 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

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

2.4 Examples of Problem (P_1)

In this subsection, we describe several particular problems, which can be written in the form of Problem (P_1).

Entropy-linear programming problem [1].

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

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

Regularized optimal transport problem [9].

$$\min_{X \in \mathbb{R}^{p \times p}_+} \left\{ \gamma \sum_{i,j=1}^{p} x_{ij} \ln x_{ij} + \sum_{i,j=1}^{p} c_{ij} x_{ij} : Xe = a_1, X^T e = a_2 \right\}, \tag{9}$$

where $e \in \mathbb{R}^p$ is the vector of all ones, $a_1, a_2 \in S_p(1)$, $c_{ij} \geq 0, i, j = 1, ..., p$ 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.

Regularized optimal partial transport problem [10].

$$\min_{X \in \mathbb{R}^{p \times p}_+} \left\{ \gamma \sum_{i,j=1}^{p} x_{ij} \ln x_{ij} + \sum_{i,j=1}^{p} c_{ij} x_{ij} : Xe \leq a_1, X^T e \leq a_2, e^T X e = m \right\},$$

where $a_1, a_2 \in \mathbb{R}^p_+$, $c_{ij} \geq 0, i, j = 1, ..., p$, $m > 0$ are given, $\gamma > 0$ is the regularization parameter and the inequalities should be understood component-wise.

3 Algorithm and Theoretical Analysis

We extend the Fast Gradient Method [14,20] in order to make it primal-dual, so that it allows not only to solve the dual problem (P_2), but also to construct a nearly optimal and nearly feasible solution to the primal problem (P_1). We also equip it with a stopping criterion, which allows an online control of the quality of

the approximate primal-dual solution. Let $\{\alpha_i\}_{i\geq 0}$ be a sequence of coefficients satisfying

$$\alpha_0 \in (0,1], \quad \alpha_k^2 \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1.$$

We define also $C_k = \sum_{i=0}^{k} \alpha_i$ and $\tau_i = \frac{\alpha_{i+1}}{C_{i+1}}$. Usual choice is $\alpha_i = \frac{i+1}{2}$, $i \geq 0$. In this case $C_k = \frac{(k+1)(k+2)}{4}$. Next, let us define Euclidean norm on $H_1^* \times H_2^*$ in a natural way

$$\|\lambda\|_2^2 = \|\lambda^{(1)}\|_2^2 + \|\lambda^{(2)}\|_2^2,$$

for any $\lambda = (\lambda^{(1)}, \lambda^{(2)})^T \in H_1^* \times H_2^*$. Unfortunately, we can not directly use the convergence results of [14,20] for the reason that the feasible set Λ in the dual problem (D_1) is unbounded and the constructed sequence $\hat{x}_k$ may possibly not satisfy the equality and inequality constraints.

ALGORITHM 1. Fast Primal-Dual Gradient Method

Input: The sequence $\{\alpha_i\}_{i\geq 0}$, Lipschitz constant L (5), accuracy $\tilde{\varepsilon}_f, \tilde{\varepsilon}_{eq}, \tilde{\varepsilon}_{in} > 0$.
Output: The point $\hat{x}_k$.
Choose $\lambda_0 = (\lambda_0^{(1)}, \lambda_0^{(2)})^T = 0$.
Set $k = 0$.
repeat
Find

$$\eta_k = (\eta_k^{(1)}, \eta_k^{(2)})^T = \arg\min_{\lambda \in \Lambda} \left\{ \varphi(\lambda_k) + \langle \nabla\varphi(\lambda_k), \lambda - \lambda_k \rangle + \frac{L}{2}\|\lambda - \lambda_k\|_2^2 \right\}.$$

$$\zeta_k = (\zeta_k^{(1)}, \zeta_k^{(2)})^T = \arg\min_{\lambda \in \Lambda} \left\{ \sum_{i=0}^{k} \alpha_i \left(\varphi(\lambda_i) + \langle \nabla\varphi(\lambda_i), \lambda - \lambda_i \rangle\right) + \frac{L}{2}\|\lambda\|_2^2 \right\}.$$

Set

$$\lambda_{k+1} = (\lambda_{k+1}^{(1)}, \lambda_{k+1}^{(2)})^T = \tau_k \zeta_k + (1 - \tau_k)\eta_k,$$

where $\tau_k = \frac{\alpha_{k+1}}{\sum_{i=0}^{k+1} \alpha_i}$.
Set

$$\hat{x}_k = \frac{1}{\sum_{i=0}^{k} \alpha_i} \sum_{i=0}^{k} \alpha_i x(\lambda_i) = (1 - \tau_{k-1})\hat{x}_{k-1} + \tau_{k-1} x(\lambda_k).$$

Set $k = k + 1$.
until $|f(\hat{x}_k) + \varphi(\eta_k)| \leq \tilde{\varepsilon}_f$, $\|A_1\hat{x}_k - b_1\|_2 \leq \tilde{\varepsilon}_{eq}$, $\|(A_2\hat{x}_k - b_2)_+\|_2 \leq \tilde{\varepsilon}_{in}$;

Theorem 1. *Let the assumptions listed in the Subsect. 2.3 hold and* $\alpha_i = \frac{i+1}{2}$, $i \geq 0$ *in Algorithm 1. Then Algorithm 1 stops after not more than*

$$N_{stop} = \max\left\{ \left\lceil \sqrt{\frac{8L(R_1^2 + R_2^2)}{\tilde{\varepsilon}_f}} \right\rceil, \left\lceil \sqrt{\frac{8L(R_1^2 + R_2^2)}{R_1\tilde{\varepsilon}_{eq}}} \right\rceil, \left\lceil \sqrt{\frac{8L(R_1^2 + R_2^2)}{R_2\tilde{\varepsilon}_{in}}} \right\rceil \right\} - 1$$

iterations. Moreover, after not more than

$$N = \max\left\{\left\lceil\sqrt{\frac{16L(R_1^2+R_2^2)}{\varepsilon_f}}\right\rceil, \left\lceil\sqrt{\frac{8L(R_1^2+R_2^2)}{R_1\varepsilon_{eq}}}\right\rceil, \left\lceil\sqrt{\frac{8L(R_1^2+R_2^2)}{R_2\varepsilon_{in}}}\right\rceil\right\} - 1$$

iterations of Algorithm 1, the point $\hat{x}_N$ will be an approximate solution to Problem (P_1) in the sense of (1).

Proof. From the complexity analysis of the FGM [14,20], one has

$$C_k\varphi(\eta_k) \le \min_{\lambda\in\Lambda}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda-\lambda_i\rangle\right) + \frac{L}{2}\|\lambda\|_2^2\right\}. \tag{10}$$

Let us introduce a set

$$\Lambda_R = \{\lambda = (\lambda^{(1)}, \lambda^{(2)})^T : \lambda^{(2)} \ge 0, \|\lambda^{(1)}\|_2 \le 2R_1, \|\lambda^{(2)}\|_2 \le 2R_2\},$$

where R_1, R_2 are given in (8). Then, from (10), we obtain

$$\begin{aligned} C_k\varphi(\eta_k) &\le \min_{\lambda\in\Lambda}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda-\lambda_i\rangle\right) + \frac{L}{2}\|\lambda\|_2^2\right\} \\ &\le \min_{\lambda\in\Lambda_R}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda-\lambda_i\rangle\right) + \frac{L}{2}\|\lambda\|_2^2\right\} \\ &\le \min_{\lambda\in\Lambda_R}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda-\lambda_i\rangle\right)\right\} + 2L(R_1^2+R_2^2). \end{aligned} \tag{11}$$

On the other hand, from the definition (2) of $\varphi(\lambda)$, we have

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

Combining this equality with (3), we obtain

$$\begin{aligned} &\varphi(\lambda_i) - \langle\nabla\varphi(\lambda_i), \lambda_i\rangle = \varphi(\lambda_i^{(1)}, \lambda_i^{(2)}) - \langle\nabla\varphi(\lambda_i^{(1)}, \lambda_i^{(2)}), (\lambda_i^{(1)}, \lambda_i^{(2)})^T\rangle \\ &= \langle\lambda_i^{(1)}, b_1\rangle + \langle\lambda_i^{(2)}, b_2\rangle - f(x(\lambda_i)) - \langle A_1^T\lambda_i^{(1)} + A_2^T\lambda_i^{(2)}, x(\lambda_i)\rangle \\ &- \langle b_1 - A_1x(\lambda_i), \lambda_i^{(1)}\rangle - \langle b_2 - A_2x(\lambda_i), \lambda_i^{(2)}\rangle = -f(x(\lambda_i)). \end{aligned}$$

Summing these inequalities from $i = 0$ to $i = k$ with the weights $\{\alpha_i\}_{i=1,\dots k}$, we get, using the convexity of $f(\cdot)$,

$$\begin{aligned}
&\sum_{i=0}^{k} \alpha_i \left(\varphi(\lambda_i) + \langle \nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right)\\
&= -\sum_{i=0}^{k} \alpha_i f(x(\lambda_i)) + \sum_{i=0}^{k} \alpha_i \langle (b_1 - A_1 x(\lambda_i), b_2 - A_2 x(\lambda_i))^T, (\lambda^{(1)}, \lambda^{(2)})^T\rangle\\
&\le -C_k f(\hat{x}_k) + C_k \langle (b_1 - A_1\hat{x}_k, b_2 - A_2\hat{x}_k)^T, (\lambda^{(1)}, \lambda^{(2)})^T\rangle.
\end{aligned}$$

Substituting this inequality to (11), we obtain

$$\begin{aligned}
&C_k\varphi(\eta_k) \le -C_k f(\hat{x}_k)\\
&+ C_k \min_{\lambda\in\Lambda_R} \left\{\langle (b_1 - A_1\hat{x}_k, b_2 - A_2\hat{x}_k)^T, (\lambda^{(1)}, \lambda^{(2)})^T\rangle\right\} + 2L(R_1^2 + R_2^2).
\end{aligned}$$

Finally, since

$$\begin{aligned}
&\max_{\lambda\in\Lambda_R} \left\{\langle (-b_1 + A_1\hat{x}_k, -b_2 + A_2\hat{x}_k)^T, (\lambda^{(1)}, \lambda^{(2)})^T\rangle\right\}\\
&= 2R_1\|A_1\hat{x}_k - b_1\|_2 + 2R_2\|(A_2\hat{x}_k - b_2)_+\|_2,
\end{aligned}$$

we obtain

$$\varphi(\eta_k) + f(\hat{x}_k) + 2R_1\|A_1\hat{x}_k - b_1\|_2 + 2R_2\|(A_2\hat{x}_k - b_2)_+\|_2 \le \frac{2L(R_1^2 + R_2^2)}{C_k}. \quad (12)$$

Since $\lambda^* = (\lambda^{*(1)}, \lambda^{*(2)})^T$ is an optimal solution of Problem (D_1), we have, for any $x \in Q$,

$$Opt[P_1] \le f(x) + \langle \lambda^{*(1)}, A_1 x - b_1\rangle + \langle \lambda^{*(2)}, A_2 x - b_2\rangle.$$

Using the assumption (8) and that $\lambda^{*(2)} \ge 0$, we get

$$f(\hat{x}_k) \ge Opt[P_1] - R_1\|A_1\hat{x}_k - b_1\|_2 - R_2\|(A_2\hat{x}_k - b_2)_+\|_2. \quad (13)$$

Hence,

$$\begin{aligned}
&\varphi(\eta_k) + f(\hat{x}_k) = \varphi(\eta_k) - Opt[P_2] + Opt[P_2] + Opt[P_1] - Opt[P_1] + f(\hat{x}_k) \overset{(6)}{=}\\
&= \varphi(\eta_k) - Opt[P_2] - Opt[D_1] + Opt[P_1] - Opt[P_1] + f(\hat{x}_k) \overset{(7)}{\ge}\\
&\ge -Opt[P_1] + f(\hat{x}_k) \overset{(13)}{\ge} -R_1\|A_1\hat{x}_k - b_1\|_2 - R_2\|(A_2\hat{x}_k - b_2)_+\|_2. \quad (14)
\end{aligned}$$

This and (12) give

$$R_1\|A_1\hat{x}_k - b_1\|_2 + R_2\|(A_2\hat{x}_k - b_2)_+\|_2 \le \frac{2L(R_1^2 + R_2^2)}{C_k}. \quad (15)$$

Hence, we obtain

$$\varphi(\eta_k) + f(\hat{x}_k) \overset{(14),(15)}{\geq} -\frac{2L(R_1^2 + R_2^2)}{C_k}. \tag{16}$$

On the other hand, we have

$$\varphi(\eta_k) + f(\hat{x}_k) \overset{(12)}{\leq} \frac{2L(R_1^2 + R_2^2)}{C_k}. \tag{17}$$

Combining (15), (16), (17), we conclude

$$\begin{aligned} &\|A_1\hat{x}_k - b_1\|_2 \leq \frac{2L(R_1^2 + R_2^2)}{C_k R_1}, \\ &\|(A_2\hat{x}_k - b_2)_+\|_2 \leq \frac{2L(R_1^2 + R_2^2)}{C_k R_2}, \\ &|\varphi(\eta_k) + f(\hat{x}_k)| \leq \frac{2L(R_1^2 + R_2^2)}{C_k}. \end{aligned} \tag{18}$$

As we know, for the chosen sequence $\alpha_i = \frac{i+1}{2}, i \geq 0$, it holds that $C_k = \frac{(k+1)(k+2)}{4} \geq \frac{(k+1)^2}{4}$. Then, in accordance to (18), after given in the theorem statement number N_{stop} of the iterations of Algorithm 1, the stopping criterion is fulfilled and Algorithm 1 stops.

Now let us prove the second statement of the theorem. We have

$$\begin{aligned} &\varphi(\eta_k) + Opt[P_1] = \varphi(\eta_k) - Opt[P_2] + Opt[P_2] + Opt[P_1] \overset{(6)}{=} \\ &= \varphi(\eta_k) - Opt[P_2] - Opt[D_1] + Opt[P_1] \overset{(7)}{\geq} 0. \end{aligned}$$

Hence,

$$f(\hat{x}_k) - Opt[P_1] \leq f(\hat{x}_k) + \varphi(\eta_k). \tag{19}$$

On the other hand,

$$f(\hat{x}_k) - Opt[P_1] \overset{(13)}{\geq} -R_1\|A_1\hat{x}_k - b_1\|_2 - R_2\|(A_2\hat{x}_k - b_2)_+\|_2. \tag{20}$$

Note that, since the point $\hat{x}_k$ may not satisfy the equality and inequality constraints, one can not guarantee that $f(\hat{x}_k) - Opt[P_1] \geq 0$. From Equation (19), (20), we can see that if we set $\tilde{\varepsilon}_f = \varepsilon_f$, $\tilde{\varepsilon}_{eq} = \min\{\frac{\varepsilon_f}{2R_1}, \varepsilon_{eq}\}$, $\tilde{\varepsilon}_{in} = \min\{\frac{\varepsilon_f}{2R_2}, \varepsilon_{in}\}$, and run Algorithm 1 for N iterations, where N is given in the theorem statement, we obtain that (1) fulfills and $\hat{x}_N$ is an approximate solution to Problem (P_1) in the sense of (1). □

We point that other authors [9,10,15–19] do not provide the complexity analysis for their algorithms when the accuracy of the solution is defined by (1).

4 Preliminary Numerical Experiments

To compare our algorithm with the existing algorithms, we choose the problem (9) of regularized optimal transport [9], which is a special case of Problem (P_1). The first reason for this choice is that, despite insufficient theoretical analysis, the existing balancing-type methods for solving this class of problems are known to be very efficient in practice [9] and provide a kind of benchmark for any new method. The second reason is that ROT problem have recently become very popular in application to image analysis based on Wasserstein spaces geometry [9,10].

Our numerical experiments were carried out on a PC with CPU Intel Core i5 (2.5 Hgz), 2 Gb of RAM using Matlab 2012 (8.0). We compare proposed in this article Algorithm 1 (below we refer to it as FGM) with the following algorithms

- Applied to the dual problem (D_1), Conjugate Gradient Method in the Fletcher–Reeves form [21] with the stepsize chosen by one-dimensional minimization. We refer to this algorithm as CGM.
- The algorithm proposed in [15] and based on the idea of Tikhonov's regularization of the dual problem (D_1). In this approach the regularized dual problem is solved by the Fast Gradient Method [14]. We will refer to this algorithm as REG;

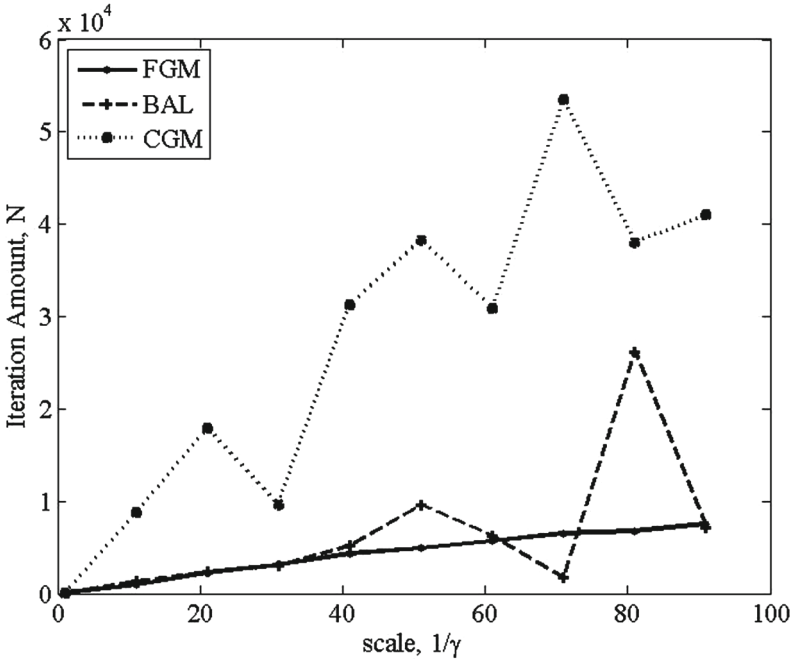

Fig. 1. Complexity of FGM, BAL and CGM as γ varies

– Balancing method [9,12] which is a special type of a fixed-point-iteration method for the system of the optimality conditions for ROT problem. It is referred below as BAL.

The key parameters of the ROT problem in the experiments are as follows

– $n := dim(E) = p^2$ – problem dimension, varies from 2^4 to 9^4;
– $m_1 := dim(H_1) = 2\sqrt{n}$ and $m_2 = dim(H_2) = 0$ – dimensions of the vectors b_1 and b_2 respectively;
– c_{ij}, $i, j = 1, ..., p$ are chosen as squared Euclidean pairwise distance between the points in a $\sqrt{p} \times \sqrt{p}$ grid originated by a 2D image [9,10];
– a_1 and a_2 are random vectors in $S_{m_1}(1)$ and $b_1 = (a_1, a_2)^T$;
– the regularization parameter γ varies from 0.001 to 1;
– the desired accuracy of the approximate solution in (1) is defined by its relative counterpart ε_f^{rel} and ε_g^{rel} as follows

$$\varepsilon_f = \varepsilon_f^{rel} \cdot f(x(\lambda_0)) \quad \varepsilon_{eq} = \varepsilon_g^{rel} \cdot \|A_1 x(\lambda_0) - b_1\|_2,$$

where λ_0 is the starting point of the algorithm. Note that $\varepsilon_{in} = 0$ since no inequality constraints are present in ROT problems.

Figure 1 shows the number of iterations for the FGM, BAL and CGM methods depending on the inverse of the regularization parameter γ. The results for

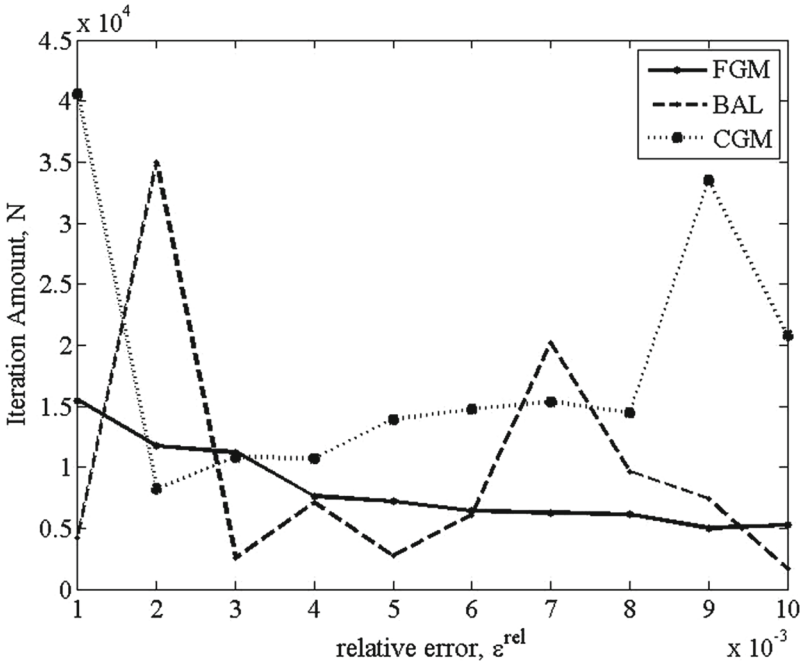

Fig. 2. Complexity of FGM, BAL and CGM as the desired relative accuracy varies

the REG are not plotted since this algorithm required one order of magnitude more iterations than the other methods. In these experiment we chose $n = 2401$ and $\varepsilon_f^{rel} = \varepsilon_g^{rel} = 0.01$. One can see that the complexity of the FGM (i.e. proposed Algorithm 1) depends nearly linearly on the value of $1/\gamma$, and that this complexity is smaller than that of the other methods when γ is small.

Figure 2 shows the number of iterations for the FGM, BAL and CGM methods depending on the relative error ε^{rel}. The results for the REG are not plotted since this algorithm required one order of magnitude more iterations than the other methods. In these experiment we chose $n = 2401$, $\gamma = 0.1$ and $\varepsilon_f^{rel} = \varepsilon_g^{rel} = \varepsilon^{rel}$. One can see that in half of the cases the FGM (i.e. proposed Algorithm 1) performs better or equally to the other methods.

5 Conclusion

This paper proposes a new primal-dual approach to solve a general class of problems stated as Problem (P_1). Unlike the existing methods, we managed to provide the convergence rate for the proposed algorithm in terms of the primal objective residual $|f(\hat{x}_k - Opt[P_1]|$ and the linear constraints infeasibility $\|A_1\hat{x}_k - b_1\|_2$, $\|(A_2\hat{x}_k - b_2)_+\|_2$. Our numerical experiments show that our algorithm performs better than existing methods for problems of regularized optimal transport, which are a special instance of Problem (P_1) for which there exist efficient algorithms.

Acknowledgements. The research by A. Gasnikov and P. Dvurechensky presented in Sect. 3 was conducted in IITP RAS and supported by the Russian Science Foundation grant (project 14-50-00150), the research by A. Gasnikov and P. Dvurechensky presented in Sect. 4 was partially supported by RFBR, research project No. 15-31-20571 mol_a_ved. The research by A. Chernov presented in Sect. 4 was partially supported by RFBR, research project No.14-01-00722-a.

References

1. Fang, S.-C., Rajasekera, J., Tsao, H.-S.: Entropy Optimization and Mathematical Programming. Kluwers International Series, Boston (1997)
2. Golan, A., Judge, G., Miller, D.: Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley, Chichester (1996)
3. Kapur, J.: Maximum entropy models in science and engineering. John Wiley & Sons Inc., New York (1989)
4. Gasnikov, A., et al.: Introduction to Mathematical Modelling of Traffic Flows. MCCME, Moscow (2013). (in russian)
5. Rahman, M.M., Saha, S., Chengan, U., Alfa, A.S.: IP traffic matrix estimation methods: comparisons and improvements. In: 2006 IEEE International Conference on Communications, Istanbul, pp. 90–96 (2006)
6. Zhang, Y., Roughan, M., Lund, C., Donoho, D.: Estimating point-to-point and point-to-multipoint traffic matrices: an information-theoretic approach. IEEE/ACM Trans. Networking **13**(5), 947–960 (2005)

7. Hastie, T., Tibshirani, R., Friedman, R.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, Heidelberg (2009)
8. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc.: Seri. B (Stat. Methodol.) **67**(2), 301–320 (2005)
9. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems, pp. 2292–2300 (2013)
10. Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., Peyre, G.: Iterative bregman projections for regularized transportation problems. SIAM J. Sci. Comput. **37**(2), A1111–A1138 (2015)
11. Bregman, L.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
12. 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)
13. Franklin, J., Lorenz, J.: On the scaling of multidimensional matrices. Linear Algebra Appl. **114**, 717–735 (1989)
14. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
15. Gasnikov, A., Gasnikova, E., Nesterov, Y., Chernov, A.: About effective numerical methods to solve entropy linear programming problem. Comput. Math. Math. Phys. **56**(4), 514–524 (2016). http://arxiv.org/abs/1410.7719
16. Polyak, R.A., Costa, J., Neyshabouri, J.: Dual fast projected gradient method for quadratic programming. Optim. Lett. **7**(4), 631–645 (2013)
17. Necoara, I., Suykens, J.A.K.: Applications of a smoothing technique to decomposition in convex optimization. IEEE Trans. Autom. Control **53**(11), 2674–2679 (2008)
18. Goldstein, T., O'Donoghue, B., Setzer, S.: Fast Alternating Direction Optimization Methods. Technical Report, Department of Mathematics, University of California, Los Angeles, USA, May (2012)
19. Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**(1), 269–297 (2014)
20. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**(1–2), 37–75 (2014)
21. Fletcher, R., Reeves, C.M.: Function minimization by conjugate gradients. Comput. J. **7**, 149–154 (1964)