A Primal-Dual First-Order Method for Minimization Problems with Linear Constraints

Alexey Chernov
Moscow Institute of Physics and Technology
alexmipt@mail.ru

Pavel Dvurechensky
Weierstrass Institute for Applied Analysis and Stochastics,
Institute for Information Transmission Problems
pavel.dvurechensky@gmail.com

Abstract

We consider a class of optimization problems with a strongly convex objective function. The feasible set in this class is given as an intersection of a simple convex set with a set given by a number of linear equality and inequality constraints. This class of problems often arises in applications covering the problems of entropy-linear programming, ridge regression, elastic net, regularized optimal transport, etc. We propose a method which can solve such problems with a given accuracy in terms of both the primal objective and the linear constraints infeasibility. Unlike existing methods it can deal with the case when no bound for the norm of any dual solution is available. We estimate the complexity of our method in terms of the number of iterations which is required to achieve the desired accuracy of the approximate solution.

1. Introduction

In this paper we deal with 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 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 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) was considered in [1]. This problem captures a broad set of optimization problems arising in applications. The first example is the classical entropy-linear programming (ELP) problem [2] which arises in many applications such as econometrics [3], modeling in science and engineering [4], especially in the modeling of traffic flows [5] and the IP traffic matrix estimation [6, 7]. Other examples are the ridge regression problem [8] and the elastic net approach [9] which are used in machine learning. Finally, the problem class (P_1) covers problems of regularized optimal transport (ROT) [10] and regularized optimal partial transport (ROPT) [11], which recently have become popular in application to the image analysis.

The classical balancing algorithms such as [10, 12, 13] are very efficient for solving ROT problems or special types of ELP problem, but they can deal only with linear equality constraints of special type and their rate of convergence estimates are rather impractical [14]. In [11] the authors provide a generalization but only for the 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 the 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 the optimal transport (OT) or the optimal partial transport (OPT) problem.

In practice the 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, existing methods which elaborate this idea have at least two drawbacks. Firstly, the conver-

gence analysis of the Fast Gradient Method (FGM) [15] 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 not known. In [16] the authors use this approach with additional constraints and propose a restart technique to define the unknown bound for the optimal dual variable value. Unfortunately, the authors consider only the classical ELP problem with only the equality constraints and it is not clear whether their technique can be applied for 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 at it is done in [17, 18] but also in terms of the error in the solution of the primal problem [1] $|f(x_k) - Opt[P_1]|$ and the linear constraints infeasibility $\|A_1x_k - b_1\|_{H_1}$, $\|(A_2x_k - b_2)_+\|_{H_2}$, 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 [19, 20] 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 approximate primal solution and the linear constraints infeasibility.

Finally the approach of [1] strongly relies on the assumption that a bound for the norm of some solution to the problem, dual to Problem (P_1) is available. Unfortunately this assumption does not hold for example for the ROT and ROPT problems.

Our contributions in this work are the following. We extend the approach of [1] in order to be able to solve Problem (P_1) in the case when no bound for the norm of any solution to the dual problem is available. Unlike [10, 11, 17, 18, 19, 20, 16] we provide the estimates for the rate of convergence in terms of the error in the solution of the primal problem $|f(x_k) - Opt[P_1]|$ and the linear constraints infeasibility $\|A_1x_k - b_1\|_{H_1}$, $\|(A_2x_k - b_2)_+\|_{H_2}$. In the contrast to the estimates in [15], 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 approaches in [10, 16].

[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.

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 $\|\cdot\|_E$

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

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. By $\partial f(x)$ we denote the subdifferential of the 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

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

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

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

Here $Opt[P_1]$ denotes the optimal function value for Problem (P_1) and 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 $\Lambda = \{\lambda = (\lambda^{(1)}, \lambda^{(2)})^T \in H_1^* \times H_2^* : \lambda^{(2)} \ge 0\}$. The Lagrange dual problem to Problem (P_1) is

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

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

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

We denote

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

Note that the gradient of the function $\varphi(\lambda)$ is equal to (see e.g. [15])

$$\nabla\varphi(\lambda)=\begin{pmatrix} b_1-A_1x(\lambda)\\ b_2-A_2x(\lambda)\end{pmatrix}, \tag{3}$$

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

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

Note that this gradient is Lipschitz-continuous (see e.g. [15]) with constant

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

It is obvious that

$$Opt[D_1]=-Opt[P_2]. \tag{5}$$

Here 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

$$Opt[P_1]\geq Opt[D_1]. \tag{6}$$

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

$$\|\lambda^{*(1)}\|_2\leq R_1^*<+\infty,\quad \|\lambda^{*(2)}\|_2\leq R_2^*<+\infty. \tag{7}$$

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 [2].

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

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 [10].

$$\min_{X\in\mathbb{R}^{p\times p}_+}\left\{\gamma\sum_{i,j=1}^p x_{ij}\ln x_{ij}+\sum_{i,j=1}^p c_{ij}x_{ij}:Xe=a_1,X^Te=a_2\right\}, \tag{8}$$

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 ith row and the jth column.

Regularized optimal partial transport problem [11].

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

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 use a restart technique with the method of [1] in order to deal with the unknown bounds R_1,R_2. Our method is the further extension of the Fast Gradient Method [15, 21]. Let $\{\alpha_i\}_{i\geq 0}$ be a sequence of coefficients satisfying

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

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 $C_k=\frac{(k+1)(k+2)}{4}$. Also we define the 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 $\lambda=(\lambda^{(1)},\lambda^{(2)})^T\in H_1^*\times H_2^*$.

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

$$\leq 4\sqrt{8L(R_1^2+R_2^2)}\max\left\{\sqrt{\frac{2}{\varepsilon_f}}\frac{R_1^*}{R_1}+\sqrt{\frac{R_1^*}{R_1^2\varepsilon_{eq}}}+ \sqrt{\frac{R_1^*}{R_1R_2\varepsilon_{in}}},\sqrt{\frac{2}{\varepsilon_f}}\frac{R_2^*}{R_2}+\sqrt{\frac{R_2^*}{R_1R_2\varepsilon_{eq}}}+\sqrt{\frac{R_2^*}{R_2^2\varepsilon_{in}}}\right\}.$$

ALGORITHM 1: Fast Primal-Dual Gradient Method

Input: The sequence $\{\alpha_i\}_{i\geq 0}$, accuracy $\varepsilon_f, \varepsilon_{eq}, \varepsilon_{in} > 0$, initial guess R_1, R_2

Output: The point $\hat{x}_k$.

Set $s = 0$.

repeat

Set

$$\tilde{\varepsilon}_{eq} = \min\left\{\frac{\varepsilon_f}{2^{s+1}R_1}, \varepsilon_{eq}\right\}, \quad \tilde{\varepsilon}_{in} = \min\left\{\frac{\varepsilon_f}{2^{s+1}R_2}, \varepsilon_{in}\right\}. \quad (9)$$

Set $\lambda_0 = (\lambda_0^{(1)}, \lambda_0^{(2)})^T = 0$.

Set $k = 0$.

Set

$$K = \max\left\{\left\lceil\sqrt{\frac{2^{3+2s}L(R_1^2+R_2^2)}{\varepsilon_f}}\right\rceil, \left\lceil\sqrt{\frac{2^{3+s}L(R_1^2+R_2^2)}{R_1\tilde{\varepsilon}_{eq}}}\right\rceil, \left\lceil\sqrt{\frac{2^{3+s}L(R_1^2+R_2^2)}{R_2\tilde{\varepsilon}_{in}}}\right\rceil\right\}.$$

repeat

Compute

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

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

Set

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

Set

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

Set $k = k+1$.

until $f(\hat{x}_k) + \varphi(\eta_k) \leq \varepsilon_f$, $\|A_1\hat{x}_k - b_1\|_2 \leq \tilde{\varepsilon}_{eq}$, $\|(A_2\hat{x}_k - b_2)_+\|_2 \leq \tilde{\varepsilon}_{in}$ *or* $k \geq K$;

Set $s = s+1$.

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

steps of the inner cycle of Algorithm 1 the point $\hat{x}_k$ will be an approximate solution to Problem (P_1) in the sense of (1).

Proof. Let us denote

$$\hat{s} = \max\left\{\left\lceil\log_2\frac{R_1^*}{R_1}\right\rceil, \left\lceil\log_2\frac{R_2^*}{R_2}\right\rceil\right\}.$$

Then $2^sR_1 \geq R_1^*$ and $2^sR_2 \geq R_2^*$ and the assumptions of Theorem 1 of [1] hold. Since the inner cycle of Algorithm 1 is the same as in Algorithm 1 of [1], according to the aforementioned theorem, we obtain that the stopping criterion in Algorithm 1 of this work fulfills for some $s \leq \hat{s}$. According to the same theorem the point $\hat{x}_k$ is an approximate solution to Problem (P_1) in the sense of (1).

Let us now estimate the number of inner steps in our Algorithm 1. On each outer iteration Algorithm 1 makes not more than

$$K(s) = \max\left\{\left\lceil\sqrt{\frac{2^{3+2s}L(R_1^2+R_2^2)}{\varepsilon_f}}\right\rceil, \left\lceil\sqrt{\frac{2^{3+s}L(R_1^2+R_2^2)}{R_1\tilde{\varepsilon}_{eq}}}\right\rceil, \left\lceil\sqrt{\frac{2^{3+s}L(R_1^2+R_2^2)}{R_2\tilde{\varepsilon}_{in}}}\right\rceil\right\} \overset{(9)}{=}$$
$$\max\left\{\left\lceil\sqrt{\frac{2^{4+2s}L(R_1^2+R_2^2)}{\varepsilon_f}}\right\rceil, \left\lceil\sqrt{\frac{2^{3+s}L(R_1^2+R_2^2)}{R_1\varepsilon_{eq}}}\right\rceil, \left\lceil\sqrt{\frac{2^{3+s}L(R_1^2+R_2^2)}{R_2\varepsilon_{in}}}\right\rceil\right\} \quad (10)$$

inner steps. Then the total number of inner iterations is not more than

$$\sum_{s=0}^{\hat{s}}K(s) \leq 4\max\left\{\sqrt{\frac{16L(R_1^2+R_2^2)}{\varepsilon_f}}\frac{R_1^*}{R_1} + \sqrt{\frac{8L(R_1^2+R_2^2)R_1^*}{R_1^2\varepsilon_{eq}}} + \sqrt{\frac{8L(R_1^2+R_2^2)R_1^*}{R_1R_2\varepsilon_{in}}}, \sqrt{\frac{16L(R_1^2+R_2^2)}{\varepsilon_f}}\frac{R_2^*}{R_2} + \sqrt{\frac{8L(R_1^2+R_2^2)R_2^*}{R_1R_2\varepsilon_{eq}}} + \sqrt{\frac{8L(R_1^2+R_2^2)R_2^*}{R_2^2\varepsilon_{in}}}\right\}.$$

Here we used that

$$\sum_{s=0}^{\hat{s}}2^s \leq 2^{\hat{s}+1} \leq 4\max\left\{\frac{R_1^*}{R_1}, \frac{R_2^*}{R_2}\right\}$$

and

$$\sum_{s=0}^{\hat{s}}2^{s/2} \leq \frac{\sqrt{2}^{\hat{s}+1}}{\sqrt{2}-1} \leq 4\max\left\{\frac{R_1^*}{R_1}, \frac{R_2^*}{R_2}\right\}.$$

□

4. Discussion

We would like to point that the inner cycle of the proposed algorithm is the same as in [1]. Hence, the behavior of the Algorithm 1 in practice is the same as the one in [1]. The difference is that the new algorithm does not require the exact knowledge of bounds (7) for the norm of the dual solution in order to obtain a solution with a given accuracy in accordance with (1).

It may seem that the obtained convergence rate of $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ contradicts to the lower bound [23], which says that, in large-scale setting, the best complexity of a

first-order method to solve a system of equations $Ax = b$ to an accuracy ε is $\Omega(1/\varepsilon)$. The explanation is that we have made an additional assumption of simplicity of the problem (4).

It is interesting to compare the estimates of Theorem 1 with the approach described in [24], where the authors consider only equality constraints. The number of iterations of their method until it stops is close to ours and is

$$N = \max\left\{\sqrt{18LR^2/\varepsilon_{eq}}, \sqrt{18LR/\varepsilon_{in}}\right\}.$$

But this estimate does not give any information on the number of iterations until the desired accuracy of the solution is achieved.

Additionally, it is worth to compare results suggested with the approach of regularization of the dual function suggested in [16]. This approach also uses restarts technique and the iteration amount is

$$N_{reg} = \sqrt{2L(\varepsilon_{eq} + 2R\varepsilon_{in})/\varepsilon_{in}^2}\ln(4L\Delta_\phi(\varepsilon_{eq} + 2R\varepsilon_{in})/\varepsilon_{in}^2),$$

where Δ_ϕ is of the order of $f(x_0) - f^*$. As one can see, $N_{reg} \sim \sqrt{1/\varepsilon}\ln(1/\varepsilon)$ and, hence within a logarithmic factor worse than the estimate in Theorem 1. Also the approach of [16] does not guarantee that $f(\hat{x}_k) - f^* \geq -\varepsilon_{eq}$.

Acknowledgements. Authors would like to thank A. Gasnikov for several fruitful discussions. The research was partially supported by RFBR, research project No. 15-31-20571 mol_a_ved.

References

[1] Chernov A., Dvurechensky P., Gasnikov A.: Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints // Preprint arXiv:1605.02970

[2] Fang, S.-C., Rajasekera J., Tsao H.-S.: Entropy optimization and mathematical programming. Kluwer's International Series (1997).

[3] Golan, A., Judge, G., Miller, D.: Maximum entropy econometrics: Robust estimation with limited data. Chichester, Wiley (1996).

[4] Kapur, J.: Maximum – entropy models in science and engineering. John Wiley & Sons, Inc. (1989).

[5] Gasnikov, A. et.al.: Introduction to mathematical modelling of traffic flows. Moscow, MCCME, (2013) (in russian)

[6] Rahman, M. M., Saha, S., Chengan, U. and Alfa, A. S.: IP Traffic Matrix Estimation Methods: Comparisons and Improvements// 2006 IEEE International Conference on Communications, Istanbul, p. 90–96 (2006)

[7] Zhang Y., Roughan M., Lund C., Donoho D.: Estimating Point-to-Point and Point-to-Multipoint Traffic Matrices: An Information-Theoretic Approach // IEEE/ACM Transactions of Networking, 13(5), p. 947–960 (2005)

[8] Hastie, T., Tibshirani, R., Friedman, R.: The Elements of statistical learning: Data mining, Inference and Prediction. Springer (2009)

[9] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2), p. 301–320 (2005)

[10] Cuturi, M.: Sinkhorn Distances: Lightspeed Computation of Optimal Transport // Advances in Neural Information Processing Systems, p. 2292–2300 (2013)

[11] Benamou, J.-D. , Carlier, G., Cuturi, M., Nenna, L., Peyre, G.: Iterative Bregman Projections for Regularized Transportation Problems //SIAM J. Sci. Comput., 37(2), p. A1111–A1138 (2015)

[12] 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 computational mathematics and mathematical physics, 7(3), p. 200–217 (1967)

[13] Bregman, L.: Proof of the convergence of Sheleikhovskii's method for a problem with transportation constraints // Zh. Vychisl. Mat. Mat. Fiz., 7(1), p. 147–156 (1967)

[14] Franklin, J. and Lorenz, J.: On the scaling of multidimensional matrices. // Linear Algebra and its applications, 114, 717–735 (1989)

[15] Nesterov, Yu.: Smooth minimization of non-smooth functions. // Mathematical Programming, Vol. 103, no. 1, p. 127–152 (2005)

[16] Gasnikov, A., Gasnikova, E., Nesterov, Yu., Chernov, A.: About effective numerical methods to solve entropy linear programming problem // Computational Mathematics and Mathematical Physics, , Vol. 56, no. 4, p. 514–524 (2016) http://arxiv.org/abs/1410.7719

[17] Polyak, R. A., Costa, J., Neyshabouri, J.: Dual fast projected gradient method for quadratic programming // Optimization Letters, 7(4), p.631–645 (2013)

[18] Necoara, I., Suykens, J.A.K.: Applications of a smoothing technique to decomposition in convex optimization // IEEE Trans. Automatic control, 53(11), p. 2674–2679 (2008).

[19] Goldstein, T., O'Donoghue, B., Setzer, S.: Fast Alternating Direction Optimization Methods. // Tech. Report, Department of Mathematics, University of California, Los Angeles, USA, May (2012)

[20] 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), p. 269–297 (2014)

[21] Devolder, O., Glineur, F., Nesterov, Yu.: First-order Methods of Smooth Convex Optimization with Inexact Oracle. Mathematical Programming 146(1–2), p. 37–75 (2014)

[22] Fletcher, R., Reeves, C. M.: Function minimization by conjugate gradients //Comput. J., 7, p. 149–154 (1964)

[23] Nemirovski, A., Yudin, D.: Problem complexity and method efficiency in optimization, John Wiley (1983)

[24] Anikin A., Gasnikov A., Dvurechensky P., Turin A., Chernov. A.: Dual approaches to the strongly convex simple function minimization problem under affine restrictions // Comp. Math. & Math. Phys. V. 57 (2017)