Inverse problem in mathematical modelling of computer networks

Gasnikov Alexander

(MIPT, IITP RAS) avgasnikov@gmail.com

QUASILINEAR EQUATIONS, INVERSE PROBLEMS AND THEIR APPLICATIONS

Moscow Institute of Physics and Technology, Dolgoprudny

13 September 2016

In the problem of traffic demand matrix estimation the goal is to recover traffic demand matrix represented as a vector $x \geq 0$ from known route matrix A (the element $\left[A\right]_{i,j}$ is equal 1 iff the demand with number j goes through link with number i and equals 0 otherwise) and link loads b (amount of traffic which goes through every link). This leads to the problem of finding the solution of linear system $Ax = b$. Also we assume that we have some $x_g \geq 0$ which reflects our prior assumption about x. Thus we consider x to be a projection of x_g on a simplex-type set $\left\{x \geq 0: \quad Ax = b\right\}$

$$\min_{\substack{Ax=b\\ x\geq 0}}\left\{g\left(x\right):=\left\|x-x_g\right\|_2^2\right\}=\min_{\substack{\left\|Ax-b\right\|_2^2\leq 0\\ x\geq 0}} g\left(x\right).$$

Slater's relaxation of this problem leads to the problem (denote x_* the solution of this problem)

$$\left\|x-x_g\right\|_2^2\to\min_{\substack{\left\|Ax-b\right\|_2^2\leq\varepsilon^2\\ x\geq 0}}.$$

This problem can be reduced to the problem (unfortunately without explicit dependence $\bar{\lambda}\left(\varepsilon\right)$)

$$\tilde{f}\left(x\right)=\left\|x-x_g\right\|_2^2+\bar{\lambda}\left\|Ax-b\right\|_2^2\to\min_{x\geq 0},$$

where $\bar{\lambda}$ – dual multiplier to the convex inequality $\|Ax-b\|_2^2 \le \varepsilon^2$. One might expect that $\bar{\lambda} \gg \|x_* - x_g\|_2^2 \big/ \varepsilon^2$, but in reality $\bar{\lambda}$ can be chosen much smaller ($\bar{\lambda} \sim \varepsilon^{-1} - \varepsilon^{-2}$) if we restrict ourselves only by approximate solution. Let’s reformulate the problem

$$f(x) = \|Ax-b\|_2^2 + \lambda \|x - x_g\|_2^2 \to \min_{x \ge 0},$$

where $\lambda = \bar{\lambda}^{-1}$. The last two problem statements can be considered as problems of Bayesian parameter estimation: One measure the vector b with some random error $\xi \in N\left(0, \sigma^2 I\right)$ and tries to find such vector $x \ge 0$ that satisfies the linear system

$Ax = b$ assuming that this vector is also random with prior distribution $x \in N\left(x_g, \tilde{\sigma}^2 I\right)$. So the model of the data is the following.

$$b = Ax + \xi,\ \xi \in N\left(0, \sigma^2 I\right), \text{prior on } x \in N\left(x_g, \tilde{\sigma}^2 I\right).$$

So the functions $\tilde{f}\left(x\right)$ and $f\left(x\right)$ introduced above now within multipliers are minus log likelihood for this model (with $\bar{\lambda} = \tilde{\sigma}^2 / \sigma^2,\ \lambda = \sigma^2 / \tilde{\sigma}^2$) . Hence one can consider e.g. the second minimization problem as a Bayesian estimation problem or as a Penalized Maximum Likelihood Estimation (V. Spokoiny, 2012).

In this talk we consider not only Euclidian projection. The second natural choice is Kullback–Leibler "projection". We also consider a problem of finding a sparse solution of the system $Ax = b$ which leads to LASSO-type problem.

The main result of the work is overview of modern approaches for the numerical solution of the mentioned above problems. The main practical motivation for us IP-traffic analysis. We also slightly generalize some known results.

Consider the following problem (instead of $\varphi(x)$ we can considered many other sum-type function of scalar product of rows some sparse matrix and x – most of the results below can be generalized in this direction)

$$\boxed{f(x) = \underbrace{\frac{1}{2}\|Ax - b\|_2^2}_{\varphi(x)} + g(x) \to \min_{x \in Q},} \qquad (1)$$

where A – is a matrix of size $m \times n$ with elements equal 0 or 1.

We assume that the matrix A is sparse with number of non-zero elements $nnz(A)$. Set $s = nnz(A)/n \ll m$, $\tilde{s} = sn/m$. By the solution of the problem (1) we will mean such vector (generally speaking a random vector if the method is randomized) x^N, that

$$\boxed{E\left[f\left(x^N\right)\right] - f_* \le \varepsilon^2.} \qquad (2)$$

Where the expectation is taken with respect to all randomness in the method.

Since all the situation we considered below can be treated in a strongly convex environment we assume that the high probability deviations bounds can be obtained from the Markov inequality

$$P\left(f\left(x^{N\left(\varepsilon^2\sigma\right)} \right) - f_* \geq \varepsilon^2\sigma/\sigma \right) \leq \frac{E\left[f\left(x^{N\left(\varepsilon^2\sigma\right)} \right) \right] - f_*}{\varepsilon^2\sigma/\sigma} \leq \sigma,$$

$$N\left(\varepsilon\right) \sim \sqrt{L/\lambda}\ln\left(LR^2/\varepsilon\right).$$

So we just have to make $\ln\left(\sigma^{-1}\right)$-times additional iterations to have $\geq 1-\sigma$ probability guarantee.

Possible cases for choice of $g(x)$ are:

1. (Ridge Regression / Tomogravity model)

$$g(x) = \lambda \left\| x - x^g \right\|_2^2,\ Q = \mathbb{R}_+^n;$$

2. (Mimimal mutual information model)

$$g(x) = \lambda \sum_{k=1}^{n} x_k \ln\left(x_k / x_k^g \right),\ x_k^g \in Q = S_n(R) = \left\{ x \geq 0 : \sum_{k=1}^{n} x_k = R \right\}.$$

3. (LASSO)

$$g(x) = \lambda \left\| x \right\|_1,\ Q = \mathbb{R}_+^n;$$

We use the following notations:

$\sigma_{\max}\left(A\right)$ – maximal eigenvalue of the matrix $A^T A$, note that

$$\sigma_{\max}\left(A\right) = \lambda_{\max}\left(A^T A\right) \le \operatorname{tr}\left(A^T A\right) = \sum_{k=1}^{n} \left\| A^{\langle k \rangle} \right\|_2^2 = nnz\left(A\right) = sn = \tilde{s}m,$$

where $A^{\langle k \rangle}$ – k-th column of matrix A;

$\max\limits_{k=1,\dots,n} \left\| A^{\langle k \rangle} \right\|_2^2 \le m;$

$R_2^2 = \frac{1}{2}\left\| x^0 - x_* \right\|_2^2$, where x^0 – starting point, x_* – solution of (1);

$\tilde{\mathrm{O}}(\quad) = \mathrm{O}(\quad)$ up to a logarithmic factor.

In the table below one can find complexity estimates (mathematical expectation of the total number of flops operations needed for finding solution of the problem (1) in sense (2)) for different algorithms applied to the problem (1) with different choice of $g\left(x\right)$. We marked by *star the situations which are new in some extent.

algorithm/ model	**Ridge Regression / Tomogravity model**	**Mimimal mutual information model**	**LASSO**
Conjugate Gradients Method and different modifications	$\min\left\{\begin{matrix}\tilde{O}\left(sn\sqrt{\frac{\sigma_{\max}\left(A\right)}{\lambda_1}}\right)\\ O\left(sn\sqrt{\frac{\sigma_{\max}\left(A\right)R_2^2}{\varepsilon^2}}\right)\end{matrix}\right\}$	Not applicable	Not applicable
Composite FGM (Nesterov, 2007)	$\min\left\{\begin{matrix}\tilde{O}\left(sn\sqrt{\frac{\sigma_{\max}\left(A\right)}{\lambda_1}}\right)\\ O\left(sn\sqrt{\frac{\sigma_{\max}\left(A\right)R_2^2}{\varepsilon^2}}\right)\end{matrix}\right\}$	$\min\left\{\begin{matrix}\tilde{O}\left(sn\sqrt{\frac{\max\limits_{k=1,\ldots,n}\left\Vert A^{\langle k\rangle}\right\Vert_2^2 R}{\lambda_2}}\right)\\ \tilde{O}\left(sn\sqrt{\frac{\max\limits_{k=1,\ldots,n}\left\Vert A^{\langle k\rangle}\right\Vert_2^2 R^2}{\varepsilon^2}}\right)\end{matrix}\right\}*$	$O\left(sn\sqrt{\frac{\sigma_{\max}\left(A\right)R_2^2}{\varepsilon^2}}\right)$

RCD APPROX / ALPHA (Richtarick, 2013)	$\min\left\{\begin{array}{l}\tilde{O}\left(sn\sqrt{\frac{s}{\lambda_1}}\right)^* \\ O\left(sn\sqrt{\frac{sR_2^2}{\varepsilon^2}}\right)\end{array}\right\}$	Not applicable	$O\left(sn\sqrt{\frac{sR_2^2}{\varepsilon^2}}\right)$
Dual RCA	$\min\left\{\begin{array}{l}\tilde{O}\left(\tilde{s}m\sqrt{\frac{\tilde{s}}{\lambda_1}}\right) \\ \tilde{O}\left(\tilde{s}m\sqrt{\frac{\tilde{s}R_2^2}{\varepsilon^2}}\right)\end{array}\right\}^*$	$\min\left\{\begin{array}{l}\tilde{O}\left(mn\sqrt{\frac{R}{\lambda_2}}\right) \\ \tilde{O}\left(mn\sqrt{\frac{R^2}{\varepsilon^2}}\right)\end{array}\right\}^*$	Not applicable

Remark 1 (FGM for composite problems). Minimal mutual information model in strongly convex case (case 2) fits well to the framework of composite optimization since one can consider $g(x) = \lambda \sum_{k=1}^{n} x_k \ln\left(x_k / x_k^g\right)$ as the composite term. It is sufficient that $g(x)$ is λ-strongly convex in 1-norm. Smoothness of $g(x)$ is not required. We use composite FGM method with 1-norm in primal x-space and prox-function $d(x) = \ln n + \sum_{k=1}^{n} x_k \ln x_k$ in not strongly convex case and

$$d(x) = \frac{1}{2(a-1)}\|x\|_a^2 \text{ with } a = \frac{2\log n}{2\log n - 1}$$

in strongly convex case. In not strongly convex case we can calculate the new point according to explicit formulas (exponential weighting), because prox-term and composite one both are entropy-type in strongly convex case this also can be done effective. Note that in not strongly convex case we have to use $R^2 \ln n$ instead of R^2. In strongly convex case we also have to replace R by $R\ln n$. Here we have an example when non-Euclidian prox-structure in strongly convex cases gave more benefits than Euclidian one.

Remark 3 (Estimates for FGM-type methods). Now we try to explain how these estimates were obtained. First two rows of the table have the following form

$$\underbrace{\tilde{O}(sn)}_{\substack{\text{costs of one iteration, the main} \\ \text{part is calculation of full gradient,} \\ \text{that is calculation of } A^T\cdot(Ax)}} \cdot \min\left\{ \underbrace{\tilde{O}\left(\sqrt{L_p/\tilde{\lambda}_p}\right)}_{\substack{\text{the number of iterations,} \\ \text{according to the FGM in} \\ \text{strongly convex case}}}, \underbrace{O\left(\sqrt{L_p R_p^2/\tilde{\varepsilon}}\right)}_{\substack{\text{the number of iterations,} \\ \text{according to the FGM in} \\ \text{non strongly convex case}}} \right\},$$

where $\tilde{\varepsilon}$ – precision we’d like to have. We use $\tilde{\varepsilon} = \varepsilon^2$ (see (2)); R_p^2 – Bregman divergence between the starting point and the solution of problem (1) in case when we choose p-norm in primal x-space (for example in $p = 2$ we have R_2^2, introduced above); $\tilde{\lambda}_p$ – constant of strongly convexity $f\left(x\right)$ in p-norm (in case 1 $p=2$, $\tilde{\lambda}_2=\lambda$ and in case 2 $p=1$, $\tilde{\lambda}_1=\lambda/R$);$L_p$ – Lipschitz constant of the gradient of $\varphi\left(x\right)$:

$$L_p=\max_{x\in Q}\max_{\|h\|_p\le 1}\left\langle h,\left\|\frac{\partial^2\varphi}{\partial x_i\partial x_j}\right\|h\right\rangle=\max_{\|h\|_p\le 1}\left\langle h,A^TAh\right\rangle=\max_{\|h\|_p\le 1}\|Ah\|_2^2$$

In cases 1, 3 $p=2$, $L_2=\lambda_{\max}\left(A^TA\right)\overset{def}{=}\sigma_{\max}\left(A\right)$ and in case 2 $p=1$, $L_1=\max\limits_{k=1,\dots,n}\left\|A^{\langle k\rangle}\right\|_2^2$.

Remark 3 (Estimates for Random Coordinate Descent (RCD) methods). Let’s compare FGM-type estimates to its RCD counterparts (see rows 2, 3 of the table)

$$\underset{\substack{\text{costs of one iteration,}\\ \text{the main part is recalculation}\\ \text{of component of gradient}}}{\tilde{O}(s)} \cdot \underset{\substack{\text{payment for}\\ \text{calculation only random}\\ \text{component of gradient}}}{n} \cdot \min\left\{ \underbrace{\tilde{O}\left(\sqrt{\bar{L}_p/\tilde{\lambda}_p}\right)}_{\substack{\text{the number of iterations,}\\ \text{according to the FGM in}\\ \text{strongly convex case}}}, \underbrace{O\left(\sqrt{\bar{L}_p R_p^2/\tilde{\varepsilon}}\right)}_{\substack{\text{the number of iterations,}\\ \text{according to the FGM in}\\ \text{non strongly convex case}}} \right\}$$

where $\bar{L}_p$ is, roughly speaking, the average Lipschitz constant of gradient of $\varphi(x)$:

$$\bar{L}_p^{1/2} = \max_{x\in Q}\frac{1}{n}\sum_{k=1}^{n}\left\langle e_k, \left\|\frac{\partial^2\varphi}{\partial x_i \partial x_j}\right\| e_k \right\rangle^{1/2} = \frac{1}{n}\sum_{k=1}^{n}\left\|A^{\langle k\rangle}\right\|_2 \le \sqrt{s}\,.$$

Here we considered only the case $p = 2$ and non strongly convex situations with separable composite and set Q. Generalization to non Euclidian set up or(and) non separable structure of composite term and set Q to the best of our knowledge hasn't been made until now.

If we assume that Q is formed by a few r affine restrictions (or some others separable convex inequalities), we can insert them with Lagrange multipliers in the goal function. Then we can solve new problem with fixed multipliers with the same complexity. At the same time we can consider the dual problem which has small dimension. To calculate (super-)subgradient of the goal function in the dual problem we have to solve primal problem ($\tilde{\varepsilon}$ solution in terms of function value of the primal problem gives us $\tilde{\varepsilon}$-subgradient of the dual problem). If we use ellipsoids method, we can find in a fast manner (since the dimension is small) solution of the dual problem with accuracy $\varepsilon = \mathrm{O}\left(\tilde{\varepsilon}\right)$ (in terms of dual function value). With appropriate choice of the method for the dual problem (ellipsoids method is proper) one can obtain the solution of the primal problem with the same accuracy in terms of primal function value ε.

So the main advantage of RCD methods consists in change of worth-case Lipschitz constant of gradient in complexity estimates to its average counterpart. This average Lipschitz constant can be much smaller, since (case 1) typically $s \ll \sigma_{\max}\left(A\right)$ since

$$\sigma_{\max}\left(A\right) = \max_{k=1,\ldots,n} \lambda_k\left(A^T A\right),\ \sum_{k=1}^{n} \lambda_k\left(A^T A\right) = \mathrm{tr}\left(A^T A\right) = sn.$$

Remark 4 (Estimates for Dual Random Coordinate Ascend (RCA) methods). First of all let’s form the dual problem. Denote by A_k – k-th row of matrix A, $\sigma_k(z) = \frac{1}{2}(z - b_k)^2$. Then we have (see Sion–Kakutani minimax theorem)

$$\min_{x\in Q}\left\{\frac{1}{2}\|Ax-b\|_2^2 + g(x)\right\} = \min_{x\in Q}\left\{\sum_{k=1}^{m}\sigma_k(A_k x) + g(x)\right\} =$$

$$= \min_{\substack{x\in Q\\ f=Ax}}\left\{\sum_{k=1}^{m}\sigma_k(f_k) + g(x)\right\} =$$

$$= \min_{\substack{x\in Q\\ f=Ax,\, f'}} \max_{y} \left\{ \left\langle f - f', y \right\rangle + \sum_{k=1}^{m} \sigma_k\left(f'_k \right) + g\left(x\right) \right\} =$$

$$= \max_{y\in\mathbb{R}^m} \left\{ -\max_{\substack{x\in Q\\ f=Ax}} \left\{ \left\langle f, y \right\rangle - g\left(x\right) \right\} - \max_{f'} \left\{ \left\langle f', y \right\rangle - \sum_{k=1}^{m} \sigma_k\left(f'_k \right) \right\} \right\} =$$

$$= \max_{y\in\mathbb{R}^m} \left\{ -\max_{x\in Q} \left(\left\langle -A^T y, x \right\rangle - g\left(x\right) \right) - \sum_{k=1}^{m} \max_{f'_k} \left(f'_k y_k - \sigma_k\left(f'_k \right) \right) \right\} =$$

$$= \max_{y\in\mathbb{R}^m} \left\{ -g^*\left(-A^T y \right) - \sum_{k=1}^{m} \sigma_k^*\left(y_k \right) \right\} = -\min_{y\in\mathbb{R}^m} \left\{ g^*\left(-A^T y \right) + \sum_{k=1}^{m} \sigma_k^*\left(y_k \right) \right\}$$

where we have explicit expressions for g^*, σ_k^* and its gradients. Moreover we have explicit dependence of feasible $\bar{x}(y) \in Q$. If y_* is an optimal solution of this dual problem, then $x_* = \bar{x}(y_*)$. Due to duality properties we also have that $\sum_{k=1}^{m} \sigma_k^*(y_k)$ is 1-strongly convex in 2-norm in dual y-space and $g^*\left(-A^T y\right)$ has Lipschitz constant of gradient in 2-norm equal to $\sigma_{\max}\left(A^T\right)/\lambda \le \tilde{s}m/\lambda = sn/\lambda$ in case 1 and $\max_{k=1,\dots,n} \left\|A^{\langle k \rangle}\right\|_2^2 R/\lambda$ in case 2. We can use RCD for the dual problem multiplies by "-1" and use the approach briefly described in the previous remark. It is worth noting that one has a possibility in case 1 to use recalculation at each iteration to obtain complexity of one iteration $\tilde{\mathrm{O}}\left(\tilde{s}\right)$. Unfortunately, in case 2 we can only obtain complexity of one iteration $\mathrm{O}\left(m\right)$. In this case we also have average Lipschitz constant ($A_{ij} \in \left\{0,1\right\}$ – $\left(i, j\right)$ element of matrix A)

$$\bar{L}^{1/2} \le 2 \max_{p \in S_n(1)} \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij}^2 p_j \right)^{1/2} \le 2.$$

Note that the dual problem is unconstrained. So this is one of the ways to work with not separable constraints in primal problem (but we have payment for that – dual functional isn't still separable, so in sparse case we have lack of possibility to use sparsity for accelerated methods).

We are interested in traffic applications in which typically $m \ll n$ (in the machine learning applications the situation is typically inverse $m \gg n$). So we should use primal RCD.

Remark 5 (Accuracy). In cases 1, 2 we expect to have such a situation for projection problem when estimates in strongly convex case seems to be close enough to estimates in not strongly convex (arguments at each min in the table above are close to each other). That is in some situations it doesn't matter to use strongly convexity or not – the rates of convergence up to a logarithmic factor are the same. But even in this situation there is a difference. The difference is the following: in the strongly convex case we have guarantee for convergence in argument, but in not strongly convex case we are able to use widely variety of prox-structures. One should also say that in real application we have to choose ε according to initial discrepancy. So we have to work with relative precision. This fact allows us to fix some level of the relative precision (we choose 0.01, i.e. 1%) and tie the stopping rule of the method to the performance of this criterion.

References

[1]. *Hastie T., Tibshirani R., Friedman R.* The Elements of statistical learning: Data mining, Inference and Prediction. Springer, 2009. http://statweb.stanford.edu/~tibs/ElemStatLearn/

[2]. *Spokoiny V.G.* Penalized maximum likelihood estimation and effective dimension // e-print, 2012. arXiv:1205.0498

[3]. *Zhang Y., Roughan M., Duffield N., Greenberg A.* Fast Accurate Computation of Large-Scale IP Traffic Matrices from Link Loads // In ACM Sigmetrics. San Diego, CA, 2003. https://www.cs.utexas.edu/~yzhang/papers/tomogravity-sigm03.pdf

[4]. *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. 2004. V. 10. № 10. https://www.cs.utexas.edu/~yzhang/papers/mmi-ton05.pdf

[5]. *Polyak B.T.* Introduction to optimization. Hardcover, 1987.

[6]. *Nesterov Yu.* Gradient methods for minimizing composite functions // Math. Prog. 2013. V. 140. № 1. P. 125–161.

[7]. *Fercoq O., Richtarik P.* Accelerated, Parallel and Proximal Coordinate Descent // e-print, 2013. arXiv:1312.5799

[8]. *Qu Z., Richtarik P.* Coordinate Descent with Arbitrary Sampling I: Algorithms and Complexity // e-print, 2014. arXiv:1412.8060

[9]. *Shalev-Shwartz S., Zhang T.* Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization // In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014. P. 64–72. arXiv:1309.2375

[10]. *Zheng Q., Richtárik P., Zhang T.* Randomized dual coordinate ascent with arbitrary sampling // e-print, 2015. arXiv:1411.5873

[11]. *Beck A., Teboulle M.* A Fast Iterative Thresholding Algorithm for Linear Inverse Problem // SIAM Journal of Image Sciences. 2009. V. 2(1). P. 183–202.

[12]. *Nesterov Y.* Smooth minimization of non-smooth function // Math. Program. Ser. A. 2005. V. 103. № 1. P. 127–152.

[13]. *Juditsky A., Nesterov Yu.* Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // Stoch. System. 2014. V. 4. no. 1. P. 44–80. arXiv:1401.1792

[14]. *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2013.

http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf

[15]. *Nemirovski A., Onn S., Rothblum U.* Accuracy certificates for computational problems with convex structure // Mathematics of Operations Research. 2010. V. 35:1. P. 52–78.

[16]. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. CORE UCL, PhD thesis, March 2013.

[17]. *Gasnikov A., Dvurechensky P., Usmanova I.* About accelerated randomized methods // TRUDY MIPT. 2016. V. 8. no. 2. (in print) arXiv:1508.02182 [in Russian]

[18]. *Boyd S., Parikh N., Chu E., Peleato B., Eckstein J.* Distributed optimization and statistical learning via the alternating direction method of multipliers // Foundations and Trends in Machine Learning. 2011. V. 3(1). P. 1–122. http://stanford.edu/~boyd/papers.html

[19]. *Richtárik P.* http://www.maths.ed.ac.uk/~richtarik/

[20]. *Allen-Zhu Z., Orecchia L.* Linear coupling: An ultimate unification of gradient and mirror descent // e-print, 2014. arXiv:1407.1537

[21]. *Gasikov A.V., Gasnikova E.V., Nesterov Yu.E., Chernov A.V.* Entropy-linear programming // Comp. Math. and Math. Phys. 2016. V. 56. no. 4. P. 523–534. arXiv:1410.7719

[22]. *Nesterov Yu.* Universal gradient methods for convex optimization problems // CORE Discussion Paper 2013/63. 2013.

[23]. *Nesterov Yu.* Complexity bounds or primal-dual methods minimizing the model of objective function // CORE Discussion Paper 2015/3. http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2015_3web.pdf

[24]. *Nesterov Yu., Nemirovski A.* On first order algorithms for l_1 / nuclear norm minimization // Acta Numerica. 2013. V. 22. P. 509–575.

[25]. *Gasnikov A., Dvurechensky P., Nesterov Yu.* Stochastic gradient methods with inexact oracle // TRUDY MIPT. 2016. V. 8. no. 1. P. 41–91. arxiv:1411.4218 [in Russian]

[26]. *Vasiliev F.P.* Optimization methods. M. MCCME, 2011. Vol. 1. [in Russian]

[27]. *Juditsky A., Nemirovski A.* First order methods for nonsmooth convex large-scale optimization, I, II. In: Optimization for Machine Learning. Eds. S. Sra, S. Nowozin, S. Wright. MIT Press, 2012.