«Optimization without borders», dedicated to Nesterov-65 and Protasov-50

About several ideas of Yu. Nesterov that have an impact on us recently

Alexander Gasnikov (MIPT, HSE, IITP)
gasnikov@yandex.ru

Sirius; *July 12, 2021*

The structure of the talk

1. Tensor methods
2. Coordinate methods
3. Gradient-free methods
4. Nesterov's conjugate gradients are primal-dual!
5. Accelerated Alternating minimization
6. Accelerated decentralized optimization for time-varying networks
7. Accelerated stochastic optimization
8. Accelerated methods with relatively Inexact gradient
9. Universal Mirror-Prox based on Nesterov's Universal method

Some Photos :)

Our books in Russian

Accelerated methods (including tensor ones) are one of the main subjects of this book! By writing this book we significantly based on our talks with Yurii.

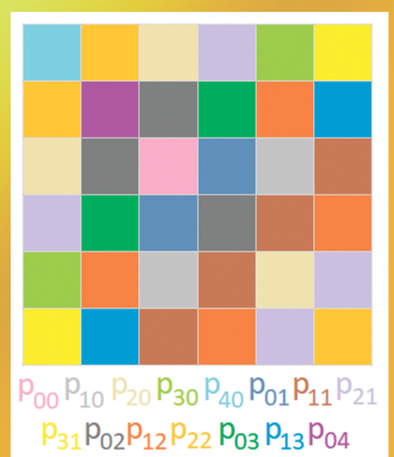
ISBN 978-5-7417-0776-0

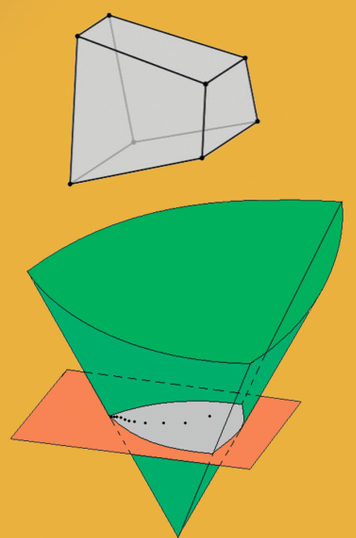
Е. А. Воронцова, Р. Хильдебранд,
А. В. Гасников, Ф. С. Стонякин
ВЫПУКЛАЯ ОПТИМИЗАЦИЯ

МФТИ

Е. А. Воронцова, Р. Хильдебранд,
А. В. Гасников, Ф. С. Стонякин

ВЫПУКЛАЯ
ОПТИМИЗАЦИЯ





Our books in Russian

https://opt.mipt.ru

Primal-dual methods, Universal method and Inexact oracle (in sense of Devolder-Glineur-Nesterov) are one of the main subjects of this book! By writing this book I significantly based on my talks with Yurii.

ISBN 978-5-4439-1380-3

А. В. ГАСНИКОВ СОВРЕМЕННЫЕ ЧИСЛЕННЫЕ МЕТОДЫ ОПТИМИЗАЦИИ

А. В. ГАСНИКОВ

СОВРЕМЕННЫЕ ЧИСЛЕННЫЕ МЕТОДЫ ОПТИМИЗАЦИИ

Метод универсального градиентного спуска

Our books in Russian

arXiv:2003.12160

Based on:

Gasnikov A., Dorn Y., Nesterov Y., Shpirko S. On the three-stage version of stable dynamic model // *arXiv:1405.7630.*

Gasnikov A. V., Gasnikova E. V., Nesterov Y. E. Dual methods for finding equilibriums in mixed models of flow distribution in large transportation networks // Computational Mathematics and Mathematical Physics. – 2018. – V. 58. – №. 9. – P. 1395-1403.

ISBN 978-5-7417-0737-1

А. В. Гасников, Е. В. Гасникова

Модели равновесного распределения транспортных потоков в больших сетях

МФТИ

А. В. Гасников, Е. В. Гасникова

МОДЕЛИ РАВНОВЕСНОГО РАСПРЕДЕЛЕНИЯ ТРАНСПОРТНЫХ ПОТОКОВ В БОЛЬШИХ СЕТЯХ

Optimization course for MIPT bachelor students of 3d year are significantly based on Nesterov's acceleration!

Методы оптимизации

https://opt.mipt.ru

📹 **Zoom** ✉️ **Telegram**

Московский Физико - Технический Институт

Лекции: **в зуме по пятницам 12.20 - 13.45**

О курсе

Осенний семестр охватывает выпуклый анализ, математическое программирование, являясь, в основном, глубоким теоретическим введением в мир оптимизации. Весенний семестр ориентируется на алгоритмы и предполагает плотную практическую работу. Актуальные новости о курсе рассылаются в телеграм чате.

профессор МФТИ, д.ф.-м.н. Александр Владимирович Гасников

профессор университета Гренобль-Альпы, PhD, Роланд Хильдебранд

Universal momentum acceleration

Polyak B.T. Some methods of speeding up the convergence of iteration methods // Comput. Math. Math. Phys. - 1964. - V. **4**:5. - P. 1–17

Nemirovski A. Orth-method for smooth convex optimization // Cybern. Soviet J. Comput. Syst. Sci. – 1982. – V. 2. – P. 937-947.

Nesterov Y. E. A method for solving the convex programming problem with convergence rate O (1/k^2) // Dokl. Akad. nauk USSR. – 1983. – V. 269. – P. 543-547.

$$x^{k+1} = x^k - \frac{1}{L}\nabla f\left(x^k + \frac{k-1}{k+2}(x^k - x^{k-1})\right) + \frac{k-1}{k+2}(x^k - x^{k-1}).$$

Nesterov's idea of acceleration is well combined with structural optimization, zero-order and high-order methods, primal-dual methods etc.

Nesterov Y. Smoothing technique and its applications in semidefinite optimization // Mathematical Programming. – 2007. – V. 110. – №. 2. – P. 245-259.

Nesterov Y. Efficiency of coordinate descent methods on huge-scale optimization problems // SIAM Journal on Optimization. – 2012. – V. 22. – №. 2. – P. 341-362.

Nesterov Y. Gradient methods for minimizing composite functions // Mathematical Programming. – 2013. – V. 140. – №. 1. – P. 125-161.

Nesterov Y., Spokoiny V. Random gradient-free minimization of convex functions // Foundations of Computational Mathematics. – 2017. – V. 17. – №. 2. – P. 527-566.

Nesterov Y. Implementable tensor methods in unconstrained convex optimization // Mathematical Programming. – 2019. – P. 1-27.

Nesterov Y. et al. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle // Optimization Methods and Software. – 2020. – P. 1-38.

Universal momentum acceleration

Recent monographs

Universal momentum acceleration

Recent surveys

Acceleration Methods

Alexandre d'Aspremont
CNRS & Ecole Normale Supérieure, Paris
aspremon@ens.fr

Damien Scieur
Samsung SAIT AI Lab & MILA, Montreal
damien.scieur@gmail.com

Adrien Taylor
INRIA & Ecole Normale Supérieure, Paris
adrien.taylor@inria.fr

First-Order Methods for Convex Optimization

Pavel Dvurechensky[a], Shimrit Shtern[b], Mathias Staudigl[c]

[a] *Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany*
[b] *Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel*
[c] *Maastricht University, Department of Data Science and Knowledge Engineering (DKE) and Mathematics Centre Maastricht (MCM), Paul-Henri Spaaklaan 1, 6229 EN Maastricht, The Netherlands*

1. Tensor method

Principal Idea of Yu. Nesterov (January, 2018): The following problem is convex and have almost the same complexity as Newton iteration when $p = 2,3$

Here $\|\nabla^p f(y) - \nabla^p f(x)\|_2 \leq M_p \|y - x\|_2$

$$\min_{y \in \mathbb{R}^n} \left\{ \sum_{r=0}^{p} \frac{1}{r!} [\ \nabla^r f(z)\ ]_{z=x} \underbrace{[y-x, \ldots, y-x]}_{r} + \frac{pM_p}{(p+1)!} \|y-x\|_2^{p+1} \right\}$$

Nesterov Y. Implementable tensor methods in unconstrained convex optimization // Mathematical Programming. – 2019. – P. 1-27.

There exists optimal (up to a log factor) acceleration (Monteiro-Svaiter, 2013; Nesterov, 2018 for $p = 2$ (book) and Gasnikov et al., 2019 for $p \geq 2$)!

Gasnikov A. et al. Near Optimal Methods for Minimizing Convex Functions with Lipschitz *p*-th Derivatives // Conference on Learning Theory. – PMLR, 2019. – P. 1392-1393.

Nesterov, Y. (2020). Inexact high-order proximal-point methods with auxiliary search procedure. CORE DP, 10, 2020.

1. Tensor method

Principal Idea of Yu. Nesterov (January, 2018): The following problem is convex and have almost the same complexity as Newton iteration when $p = 2,3$

Here $\|\nabla^p f(y) - \nabla^p f(x)\|_2 \leq M_p \|y - x\|_2$

$$\min_{y \in \mathbb{R}^n} \left\{ \sum_{r=0}^{p} \frac{1}{r!} [\ \nabla^r f(z)\ ]_{z=x} \underbrace{[y-x, \ldots, y-x]}_{r} + \frac{pM_p}{(p+1)!} \|y-x\|_2^{p+1} \right\}$$

Nesterov Y. Implementable tensor methods in unconstrained convex optimization // Mathematical Programming. – 2019. – P. 1-27.

To the best of my knowledge, Tensor methods is the main direction of current research of Yurii. So let's stop here and consider some details and vicinities!

Problem formulation

$$\min_{x\in\mathbb{R}^d}\{F(x) := f(x) + g(x)\}$$

where f, g - convex.

$$D^k f(x)[h]^k = \sum_{i_1,\ldots,i_d\geq 0:\ \sum_{j=1}^d i_j = k} \frac{\partial^k f(x)}{\partial x_1^{i_1}\ldots\partial x_d^{i_d}} h_1^{i_1}\cdot\ldots\cdot h_d^{i_d},$$

$$\|D^k f(x)\| = \max_{\|h\|\leq 1} \left\|D^k f(x)[h]^k\right\|.$$

$$\|D^p f(x) - D^p f(y)\| \leq L_{p,f}\|x-y\|.$$

$$\Omega_p(f,x;y) = f(x) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(x)\,[y-x]^k, y\in\mathbb{R}^d.$$

$$|f(y) - \Omega_p(f,x;y)| \leq \frac{L_{p,f}}{(p+1)!}\|y-x\|^{p+1}.$$

Main Algorithm

Algorithm 1 Accelerated Methaalgorithm (AM) (AM(x_0,f,g,p,H,K))

1: **Input:** $p \in \mathbb{N}$, $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^d \to \mathbb{R}$, $H > 0$.

2: $A_0 = 0, y_0 = x_0$.

3: **for** $k = 0$ **to** $k = K - 1$

4: Define $\lambda_{k+1} > 0$ and $y_{k+1} \in \mathbb{R}^d$ from

$$\frac{1}{2} \leq \lambda_{k+1} \frac{H\|y_{k+1} - \tilde{x}_k\|^{p-1}}{p!} \leq \frac{p}{p+1},$$

where

$$y_{k+1} = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(\tilde{x}_k) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(\tilde{x}_k)\,[y - \tilde{x}_k]^k + g(y) + \frac{H}{(p+1)!}\|y - \tilde{x}_k\|^{p+1} \right\}$$

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1}A_k}}{2}\,,\ A_{k+1} = A_k + a_{k+1}\,,$$

$$\tilde{x}_k = \frac{A_k}{A_{k+1}} y_k + \frac{a_{k+1}}{A_{k+1}} x_k.$$

5: $x_{k+1} := x_k - a_{k+1}\nabla f(y_{k+1}) - a_{k+1}\nabla g(y_{k+1})$.

6: **end for**

7: **return** y_K

Bubeck S. et al. Near-optimal method for highly smooth convex optimization // Conference on Learning Theory. – PMLR, 2019. – P. 492-507.

Gasnikov A. V. et al. Accelerated Meta-Algorithm for Convex Optimization Problems // Computational Mathematics and Mathematical Physics. – 2021. – V. 61. – №. 1. – P. 17-28.

Main Theorem

Theorem. Let y_k – output of AM(x_0,f,g,p,H,k) after k iterations under $p \geq 1$ and $H \geq (p+1)L_{p,f}$. Then

$$F(y_k) - F(x_*) \leq \frac{c_p H R^{p+1}}{k^{\frac{3p+1}{2}}} \quad \text{(RC)}$$

where $c_p = 2^{p-1}(p+1)^{\frac{3p+1}{2}}/p!$, $R = \|x_0 - x^*\|$.

Moreover, if $p \geq 2$ for ε: $F(y_k) - F(x_*) \leq \varepsilon$ at each iteration of AM we have to solve auxiliary problem (AP) on (λ_{k+1}, y_{k+1}) at most $O\left(\ln\left(\varepsilon^{-1}\right)\right)$ times.

Reminder:

$$\Omega_p(f, x; y) = f(x) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(x) \left[y - x\right]^k, y \in \mathbb{R}^d.$$

$$y_{k+1} = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \Omega_p(f, \tilde{x}_k; y) + g(y) + \frac{H}{(p+1)!} \|y - \tilde{x}_k\|^{p+1} \right\} \quad \text{(AP)}$$

Main Drawback

The main theorem assumes that we have to solve (AP) exactly!

$$y_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ \widetilde{\Omega}^k(y) := \Omega_p(f, \tilde{x}_k; y) + g(y) + \frac{H}{(p+1)!} \|y - \tilde{x}_k\|^{p+1} \right\} \quad \text{(AP)}$$

Is it possible to relax this requirement? YES!

Let us use instead of (AP) the following (practical) criteria:

$$\left\| \nabla \widetilde{\Omega}^k(\tilde{y}_{k+1}) \right\| \leq \frac{1}{4p(p+1)} \|\nabla F(\tilde{y}_{k+1})\|.$$

In this case the main theorem holds true with minor correction:

$$F(y_k) - F(x_*) \leq \frac{12}{5} \frac{c_p H R^{p+1}}{k^{\frac{3p+1}{2}}} \quad \text{(RC)}$$

How to solve (AP) when $g \equiv 0$?

$$\min_{x \in \mathbb{R}^d} \{F(x) := f(x)\}$$

$$y_{k+1} = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(\tilde{x}_k) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(\tilde{x}_k) [y - \tilde{x}_k]^k + \frac{H}{(p+1)!} \|y - \tilde{x}_k\|^{p+1} \right\} \quad \text{(AP)}$$

Nesterov, Y., & Polyak, B. T. (2006). Cubic regularization of Newton method and its global performance. Mathematical Programming, 108(1), 177-205. $p = 2$ is implementable!

Nesterov, Y. (2019). Implementable tensor methods in unconstrained convex optimization. Mathematical Programming, 1-27. $p = 3$ is implementable!

If we have $D^2 f(\tilde{x}_k) = \nabla^2 f(\tilde{x}_k)$ then the complexity to solve (AP) by using automatic differentiation and gradient descent in relative smoothness assumption one can solve auxiliary problem with the complexity

$$\tilde{O}\left(T_{\nabla f(x)} + d^2 + d^3\right) \text{ a.o.}$$

Nesterov Y. Inexact basic tensor methods. – 2019/23. – CORE Preprint. $p = 2$

If we don't want to calculate $D^2 f(\tilde{x}_k) = \nabla^2 f(\tilde{x}_k)$ and want to solve (AP) with precision δ (in function), then the complexity will be $O\left(T_{\nabla f(x)} \delta^{-\frac{1}{6}}\right)$.

How to solve (AP) when $g \equiv 0$?

(Super)Hyper-fast Second-order method

$$y_{k+1} = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(\tilde{x}_k) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(\tilde{x}_k) \left[y - \tilde{x}_k \right]^k + \frac{H}{(p+1)!} \|y - \tilde{x}_k\|^{p+1} \right\} \qquad \text{(AP)}$$

Nesterov, Y. (2019). Implementable tensor methods in unconstrained convex optimization. Mathematical Programming, 1-27.

Nesterov, Y. (2020). Superfast second-order methods for unconstrained convex optimization. CORE DP, 7, 2020.

For $p = 2,3$ (AP) has almost the same complexity according to the developed method $\tilde{O}\left(T_{\nabla f(x)} + d^2 + d^3\right)$ a.o. and we really need only the first and the second order oracle in both cases! So if we have 3-d order smoothness, we'd better to choose in AM $p = 3$, but to solve (AP) by using second-order information.

Nesterov, Y. (2020). Inexact high-order proximal-point methods with auxiliary search procedure. CORE DP, 10, 2020.

Applications of AM

Composite optimization: g is prox-friendly

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1), 183-202. p = 1

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. Mathematical Programming, 140(1), 125-161. (CORE Preprint, 2007) p = 1

Catalyst: $f \equiv 0$, $p = 1$. Accelerated proximal envelop, $H = L_1^g$

Lin, H., Mairal, J., & Harchaoui, Z. (2015). A universal catalyst for first-order optimization. In Advances in neural information processing systems (pp. 3384-3392).

Accelerated gradient sliding: g isn't prox-friendly, we apply AM for (AP)

Lan, G., & Ouyang, Y. (2016). Accelerated gradient sliding for structured convex optimization. arXiv preprint arXiv:1609.04905. p = 1

Kamzolov, D., Gasnikov, A., & Dvurechensky, P. (2020). On the optimal combination of tensor optimization methods. arXiv preprint arXiv:2002.01004. p = 2, 3 (implementable variants)

Accelerated methods for composite saddle point problem: $p = 1$

Alkousa, M., Dvinskikh, D., Stonyakin, F., & Gasnikov, A. (2019). Accelerated methods for composite non-bilinear saddle point problem. arXiv preprint arXiv:1906.03620.

Lin, T., Jin, C., & Jordan, M. (2020). Near-optimal algorithms for minimax optimization. arXiv preprint arXiv:2002.02417.

Accelerated gradient sliding

$$\min_{x \in \mathbb{R}^d} \left\{ F(x) := f(x) + g(x) \right\}$$

But if g is not prox-friendly, what should we do?

Idea: To apply Accelerated Algorithm (AM) $p = 1$ for next problem

$$y_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ f(\tilde{x}_k) + \langle \nabla f(\tilde{x}_k), y - \tilde{x}_k \rangle + g(y) + \frac{L}{2} \|y - \tilde{x}_k\|^2 \right\}$$

This problem is L-strongly convex. So we should apply proper restarted version of AM. In this case the complexity splits ($L_f \leq L_g$):

$$O\left(\sqrt{\frac{L_f R^2}{\varepsilon}}\right) \quad \nabla f \text{ calls} \qquad\qquad \tilde{O}\left(\sqrt{\frac{L_g R^2}{\varepsilon}}\right) \quad \nabla g \text{ calls}$$

Lan, G. First-order and Stochastic Optimization Methods for Machine Learning. Springer, 2020

Data Science applications

$$\min_{x \in \mathbb{R}^d} \left\{ F(x) := f(x) + g(x) \right\}$$

$$f(x) := \frac{1}{m} \sum_{k=1}^{m} f_k(x)$$

If g is prox-friendly, then Composite AM requires

$$O\left(m \sqrt{\frac{L_f R^2}{\varepsilon}} \right) \quad \nabla f_k \text{ calls}$$

But this is not an optimal bound. Optimal bound will be (variance reduction)

$$O\left(m + \sqrt{m \frac{\max L_k R^2}{\varepsilon}} \right) \quad \nabla f_k \text{ calls}$$

If g is not prox-friendly (Kernel SVM) it's an open problem. Solution:

Dvinskikh D. M. et al. Accelerated Gradient Sliding for Minimizing a Sum of Functions // Doklady Mathematics. – Pleiades Publishing, 2020. – V. 101. – №. 3. – P. 244-246.

Click-prediction model

$$\min_{x\in\mathbb{R}^d}\left\{F(x) := f(x) + \frac{\mu}{2}\|x\|_2^2\right\}$$

$$f(x) := \frac{1}{m}\sum_{k=1}^{m} f_k(x) \qquad f_k(x) = \log\left(1 + \exp\left(-y_k\langle a_k, x\rangle\right)\right)$$

Let s be a maximal number of nonzero elements in a_k and $L_k = O\left(\|a_k\|^2\right)$, $k = 1,...,m$. The total complexity (arithmetic operations) of optimal first-order variance-reduced schemes will be (we consider $\mu \le \varepsilon$) $O\left(sm + s\sqrt{m\frac{\max L_k R^2}{\varepsilon}}\right)$ a.o.

Is it possible to solve this problem faster? If ε is small enough the answer is YES. For that we should use 2d order Hyperfast tensor methods!

Idea: For sum type problem Hessian calculation can be comparable with Hessian inversion.

We can choose m by applying Hendrikx et al., 2020

Hendrikx H. et al. Statistically preconditioned accelerated gradient method for distributed optimization // International Conference on Machine Learning. – PMLR, 2020. – P. 4203-4227.
Dvurechensky P. et al. Hyperfast Second-Order Local Solvers for Efficient Statistically Preconditioned Distributed Optimization // arXiv:2102.08246

Accelerated method for saddle-points of sum-type

$$\min_x \max_y f(x, y)$$

$$f(x, y) := \frac{1}{m} \sum_{k=1}^{m} f_k(x, y)$$

We assume that all f_k has L-Lipschitz gradient and f is μ_x-strongly convex in x and μ_y-strongly concave in y.

Optimal bound (that can be obtained as a proper combination of AM and Alacaoglu-Malitsky VR algorithm) looks like (for $m = 1$ this corresponds to Lin-Jin-Jordan, 2020 result):

$$\tilde{O}\left(\sqrt{m\left(\sqrt{m} + L/\mu_x\right)\left(\sqrt{m} + L/\mu_y\right)}\right) \qquad \nabla f_k \text{ calls}$$

Alacaoglu A., Malitsky Y. Stochastic variance reduction for variational inequality methods // arXiv:2102.08352.
Tominin V. et al. On Accelerated Methods for Saddle-Point Problems with Composite Structure // arXiv:2103.09344.
Luo L. et al. Near Optimal Stochastic Algorithms for Finite-Sum Unbalanced Convex-Concave Minimax Optimization // arXiv:2106.01761.

2. Coordinate methods

Full-gradient method

$\sqrt{\frac{LR^2}{\varepsilon}}$ - iteration complexity

$\#\nabla f(x)$ - complexity of each iteration (gradient computation cost)

Random coordinate method

$n\sqrt{\frac{\bar{L}R^2}{\varepsilon}}$ - iteration complexity ($n = \dim x$)

$O(n) + \#\nabla_i f(x)$ - complexity of each iteration // $O(n)$ can be improved sometimes!

For many interesting cases recalculation of $\nabla_i f(x)$ is n-times cheaper than calculation of $\nabla f(x)$. So it seems that coordinate descent method has the same wall-clock time complexity as full-gradient one. But the difference is in $\bar{L}$ (average Lipschitz gradient constant along axis) versus L (Lipschitz gradient constant at worth direction). $\bar{L}$ can be n-times smaller than L!

Nesterov Y., Stich S. U. Efficiency of the accelerated coordinate descent method on structured optimization problems // SIAM Journal on Optimization. – 2017. – V. 27. – №. 1. – P. 110-123.

2. Coordinate methods

So it seems that coordinate descent method has the better wall-clock time complexity than full-gradient one

Lee Y. T., Sidford A. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems // 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. – IEEE, 2013. – P. 147-156.

But this is not true in general case!

$$\min_{x \in \mathbb{R}^n} \quad \{f(x) = \gamma \ln \left(\sum_{j=1}^{m} \exp \left(\frac{[Ax]_j}{\gamma} \right) \right) - \langle b, x \rangle\}$$

Matrix A is sparse (NOTE: Nesterov-Stich consider non-sparse case!).

Recently, there've been developed an approach how to resolve this problem with logarithmic additional payment

Pasechnyuk D., Matyukhin V. On the Computational Efficiency of Catalyst Accelerated Coordinate Descent // International Conference on Mathematical Optimization Theory and Operations Research. – Springer, Cham, 2021. – P. 176-191.

3. Gradient-free methods

Rough idea: accelerated gradient-free method has n-times large iteration complexity in comparison with full-gradient methods

Nesterov Y., Spokoiny V. Random gradient-free minimization of convex functions // Foundations of Computational Mathematics. – 2017. – V. 17. – №. 2. – P. 527-566.

In general this result is optimal!

Recently, there's been developed a new type of acceleration that allows to change standard Nesterov-Spokoiny iteration complexity

$$n \cdot \sqrt{\frac{L_2 R_2^2}{\varepsilon}} \quad \text{to} \quad n^{1/2+1/q} \cdot \sqrt{\frac{L_2 R_p^2}{\varepsilon}}, \qquad \frac{1}{p} + \frac{1}{q} = 1$$

Dvurechensky P., Gorbunov E., Gasnikov A. An accelerated directional derivative method for smooth stochastic convex optimization // European Journal of Operational Research. – 2021. – V. 290. – №. 2. – P. 601-621.

4. Nesterov's conjugate gradients is primal-dual!

Algorithm 1 Accelerated Gradient Method with Small-Dimensional Relaxation (AGMsDR)

Output: x^k

1: Set $k = 0, A_0 = 0, x^0 = v^0, \psi_0(x) = V(x, x^0)$

2: **for** $k \geqslant 0$ **do**

3:

$$\beta_k = \underset{\beta\in[0,1]}{\operatorname{argmin}} f\left(v^k + \beta(x^k - v^k)\right), \quad y^k = v^k + \beta_k(x^k - v^k). \tag{4}$$

4: Option a), L is known,

$$x^{k+1} = \underset{x\in E}{\operatorname{argmin}} \left\{ f(y^k) + \langle \nabla f(y^k), x - y^k\rangle + \frac{L}{2}\|x - y^k\|^2 \right\}. \tag{5}$$

Find a_{k+1} from equation $\frac{a_{k+1}^2}{A_k + a_{k+1}} = \frac{1}{L}$.

Option b),

$$h_{k+1} = \underset{h\geqslant 0}{\operatorname{argmin}} f\left(y^k - h(\nabla f(y^k))^{\#}\right), \quad x^{k+1} = y^k - h_{k+1}(\nabla f(y^k))^{\#}. \tag{6}$$

Find a_{k+1} from equation $f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})}\|\nabla f(y^k)\|_*^2 = f(x^{k+1})$.

5: Set $A_{k+1} = A_k + a_{k+1}$.

6: Set $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k\rangle\}$.

7: $v^{k+1} = \operatorname{argmin}_{x\in E} \psi_{k+1}(x)$

8: $k = k + 1$

9: **end for**

$$\min_{Ax=0} f(x)$$

Dual problem

$$\min_{y} f^*(A^T y)$$

Line search is cheap for dual problem!

Nesterov Y. et al. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle // Optimization Methods and Software. – 2020. – P. 1-38.

4. Nesterov's conjugate gradients are primal-dual!

Why do we really need line search?

Answer: Line search allows to exploit spectrum of Hessian at the solution. For example [folklore result, known from Yu. Nesterov], if spectrum of quadratic goal function is uniform:

$$\left(\frac{L_2 R_2^2}{\varepsilon}\right)^{1/2}$$ For accelerated methods without line-search

$$\left(\frac{L_2 R_2^2}{\varepsilon}\right)^{1/6}$$ For conjugate gradients // (1/6)-super-acceleration

Note, that recently it was shown that special assumption on spectrum and step size policy allows to obtain (1/4)-super-acceleration without line-search.

Goujaud B. et al. Super-Acceleration with Cyclical Step-sizes // arXiv:2106.09687.

5. Accelerated Alternating minimization

Algorithm 1 Accelerated Alternating Minimization (AAM)

Input: Starting point x_0.

Output: x^k

1: Set $A_0 = 0$, $x^0 = v^0$.

2: **for** $k \geqslant 0$ **do**

3: Set $\beta_k = \underset{\beta \in [0,1]}{\operatorname{argmin}} f\left(x^k + \beta(v^k - x^k)\right)$

4: Set $y^k = x^k + \beta_k(v^k - x^k)$

5: Choose $i_k = \underset{i \in \{1,\dots,n\}}{\operatorname{argmax}} \|\nabla_i f(y^k)\|_2^2$ Minimization at i_k-block

6: Set $x^{k+1} = \underset{x \in S_{i_k}(y^k)}{\operatorname{argmin}} f(x)$

7: Find a_{k+1}, $A_{k+1} = A_k + a_{k+1}$ from

$$f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}} \|\nabla f(y^k)\|_2^2 = f(x^{k+1})$$

8: Set $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$

9: **end for**

Guminov S. et al. Accelerated Alternating Minimization, Accelerated Sinkhorn's Algorithm and Accelerated Iterative Bregman Projections // ICML, 2021.

6. Accelerated decentralized optimization for time-varying networks

$$\min_{Wx=0} \frac{1}{m} \sum_{i=1}^{m} f_i(x_i)$$

Matrix W is a Laplacian matrix of communication network on m nodes. $Wx = 0$ is equivalent to $x_1 = \ldots = x_m$. Assume that this matrix changes from iteration to iteration. On iteration k we have matrix Laplacian W_k. What is oracle and communication complexity of this problem?

$\sqrt{\frac{LR^2}{\varepsilon}}$ Oracle complexity per node (∇f_i)

$\max_k \chi_k \cdot \sqrt{\frac{LR^2}{\varepsilon}}$ Communication complexity ($W_k x$)

$$\chi_k = \lambda_{\max}(W_k) / \lambda^{+}_{\min}(W_k)$$

Kovalev D. et al. ADOM: Accelerated decentralized optimization method for time-varying networks // ICML, 2021.

Kovalev D. et al. Lower Bounds and Optimal Algorithms for Smooth and Strongly Convex Decentralized Optimization Over Time-Varying Networks // arXiv:2106.04469.

7. Accelerated stochastic optimization

$$\min_{x \in Q} E_{\xi}[f(x, \xi)]$$

From recent works of Woodworth et al. 2021 we know that batch-parallelized accelerated gradient method is an optimal approach to solve smooth convex stochastic optimization problems in parallel and federated architectures (with and without overparametrization). But this conclusion was obtained without high-probability bounds. The problem in such bounds is the requirement that Q is compact (sometimes it's impossible to assume that, i.e. when solving dual problem by randomized methods Dvinskikh, 2021) and stochastic gradients are subgaussian. Based on clipping technique, developed in Nazin et al., 2019, Gorbunov et al., 2020 propose how to solve both these problems:

1) In all estimates we should take $R = \|x^0 - x_*\|_2$ instead of diam Q;
2) Heavy tails bounds can be improved to almost Hoeffding's concentration.

Nazin A., Nemirovsky A., Tsybakov A., Juditsky A. Algorithms of robust stochastic optimization based on mirror descent method // Autom. Remote Control, V. **80**:9. - 2019. - P. 1607–1627.
Gorbunov E., Danilova M., Gasnikov A. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping // NeurIPS, 2020.
Woodworth B. et al. The Min-Max Complexity of Distributed Stochastic Convex Optimization with Intermittent Communication // arXiv:2102.01583.
Woodworth B., Srebro N. An Even More Optimal Stochastic Optimization Algorithm: Minibatching and Interpolation Learning // arXiv:2106.02720.
Dvinskikh D. Decentralized Algorithms for Wasserstein Barycenters // PhD Thesis, WIAS, Berlin; arXiv preprint arXiv:2105.01587. – 2021.

8. Accelerated methods with relatively Inexact gradient

$$f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 - \delta_1 \|y - x\|_2 \leq f(y)$$
$$\leq f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 + \delta_2.$$

Devolder-Glineur-Nesterov, 2014

$$f(x_k) - f(x_*)$$
$$= O\left(\min\left\{\frac{LR^2}{k^2} + \tilde{R}\delta_1 + k\delta_2, LR^2 \exp\left(-\sqrt{\frac{\mu}{L}}\frac{k}{2}\right) + \tilde{R}\delta_1 + \sqrt{\frac{L}{\mu}}\delta_2\right\}\right)$$

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \alpha \|\nabla f(x)\|_2, \quad \alpha \in [0, 1).$$

$$\alpha \lesssim \left(\frac{\mu}{L}\right)^{3/4}, \alpha_k \lesssim \left(\frac{1}{k}\right)^{3/2}$$

In this case accelerated method doesn't feel inexactness. Note that non accelerated methods doesn't feel inexactness until $\alpha \to 1$ (B. Polyak)

Polyak B. T. Introduction to optimization. optimization software. Inc., Publications Division, New York. – 1987.
Devolder O., Glineur F., Nesterov Y. First-order methods of smooth convex optimization with inexact oracle // Mathematical Programming. – 2014. – V. 146. – №. 1. – P. 37-75.
Vasin A., Gasnikov A., Spokoiny V. Stopping rules for accelerated gradient methods with additive noise in gradient // arXiv:2102.02921.

9. Universal Mirror-Prox based on Nesterov's Universal method

Algorithm 6 Universal Mirror Prox

Input: General VI on a set $Q \subset E$ with operator $\Phi(\mathbf{z})$, accuracy $\varepsilon > 0$, initial guess $M_{-1} > 0$, prox-setup: $d(\mathbf{z}), V[\mathbf{z}](\mathbf{w})$.

1: Set $k = 0$, $\mathbf{z}^0 = \arg\min_{\mathbf{z} \in Q} d(\mathbf{z})$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Set $i_k = 0$
4: **repeat**
5: Set $M_k = 2^{i_k - 1} M_{k-1}$.
6: Calculate

$$\mathbf{w}^k = \arg\min_{\mathbf{z} \in Q} \left\{ \langle \Phi(\mathbf{z}^k), \mathbf{z} \rangle + M_k V[\mathbf{z}^k](\mathbf{z}) \right\}. \tag{72}$$

7: Calculate

$$\mathbf{z}^{k+1} = \arg\min_{\mathbf{z} \in Q} \left\{ \langle \Phi(\mathbf{w}^k), \mathbf{z} \rangle + M_k V[\mathbf{z}^k](\mathbf{z}) \right\}. \tag{73}$$

8: $i_k = i_k + 1$.
9: **until**

$$\langle \Phi(\mathbf{w}^k) - \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1} \rangle \leq \frac{M_k}{2} \left(\|\mathbf{w}^k - \mathbf{z}^k\|^2 + \|\mathbf{w}^k - \mathbf{z}^{k+1}\|^2 \right) + \frac{\varepsilon}{2}. \tag{74}$$

10: Set $k = k + 1$.
11: **end for**

Output: $\widehat{\mathbf{w}}^k = \frac{1}{k} \sum_{i=0}^{k-1} \mathbf{w}^i$.

Complexity: $2 \inf_{\nu \in [0,1]} \left(\frac{2L_\nu}{\varepsilon} \right)^{\frac{2}{1+\nu}} \cdot \max_{\mathbf{z} \in Q} V[\mathbf{z}_0](\mathbf{z})$

Nemirovskii A.S., Nesterov Yu.E. Optimal methods of smooth convex minimization // U.S.S.R. Comput. Math. Math. Phys., V. **25**:2. - 1985. P. 21–30.

Nesterov Y. Universal gradient methods for convex optimization problems // Mathematical Programming. – 2015. – V. 152. – №. 1. – P. 381-404.

Dvurechensky P. E. et al. Advances in low-memory subgradient optimization // Numerical Nonsmooth Optimization. – 2020. – P. 19-59.

Some photos

Special thanks to Svetlana Nesterova!

Marriage, USSR 1982

Yurii and Svetlana almost 40 years together!

IOWA, 1994

Yurii with sons

France, 1989

With main co-author Arkadi Nemirovski (First trip to Europe)

Chicago, 1990

With Arkadi Nemirovski (First trip to USA)

Chicago, 1990

With Arkadi Nemirovski (First trip to USA)

Chicago, 1990

With Arkadi Nemirovski (First trip to USA)

IOWA, 1990

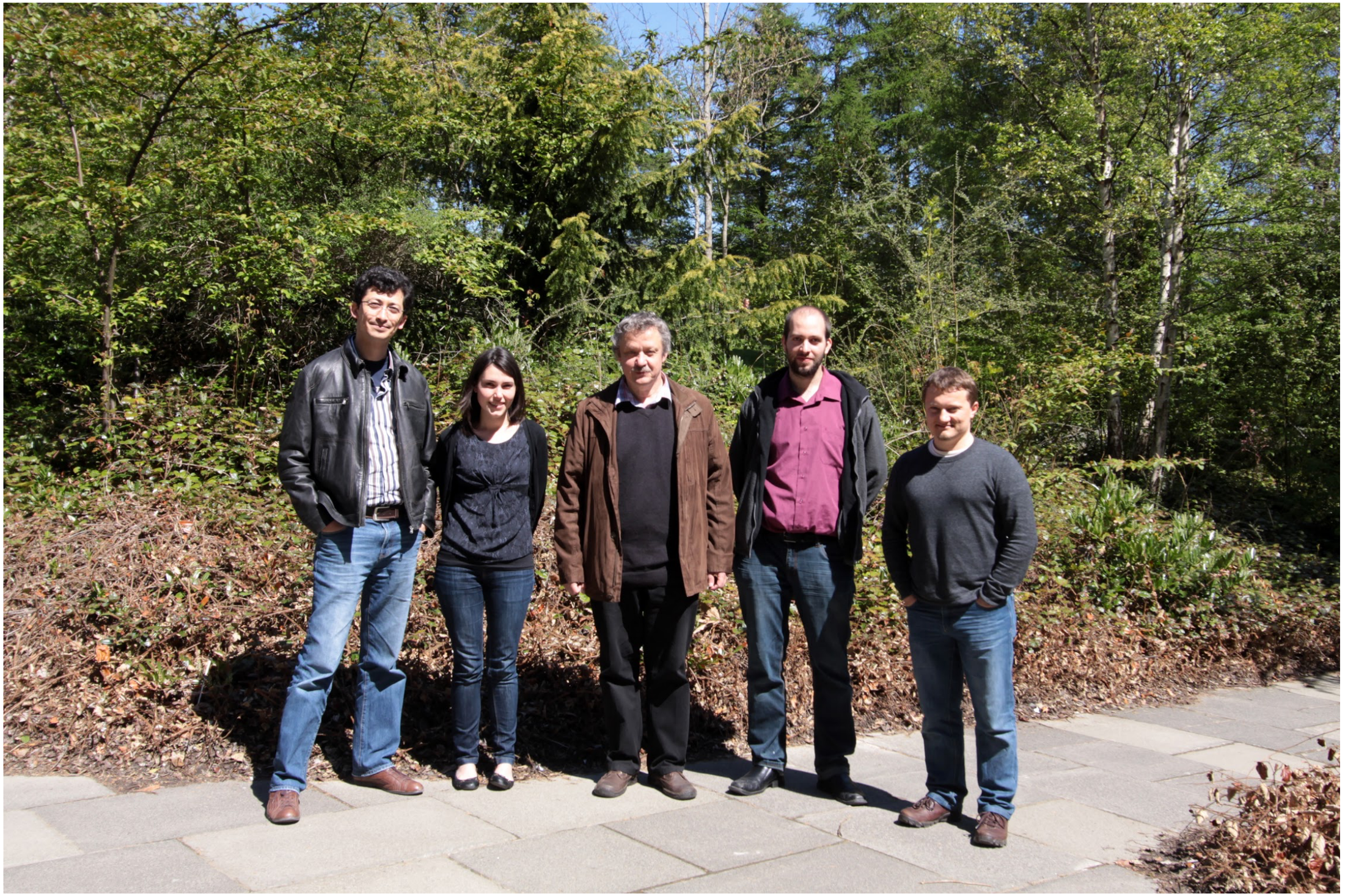
With Arkadi Nemirovski

SWISS, 1996

Luminy, 2007

With Polijaks

2013



With Peter Richtarik, Martin Takac et al.

EURO Gold, 2016

USA, 2017

With A. Nemirovksi and A. Shapiro

Now

At home with ACCOPT group members and family