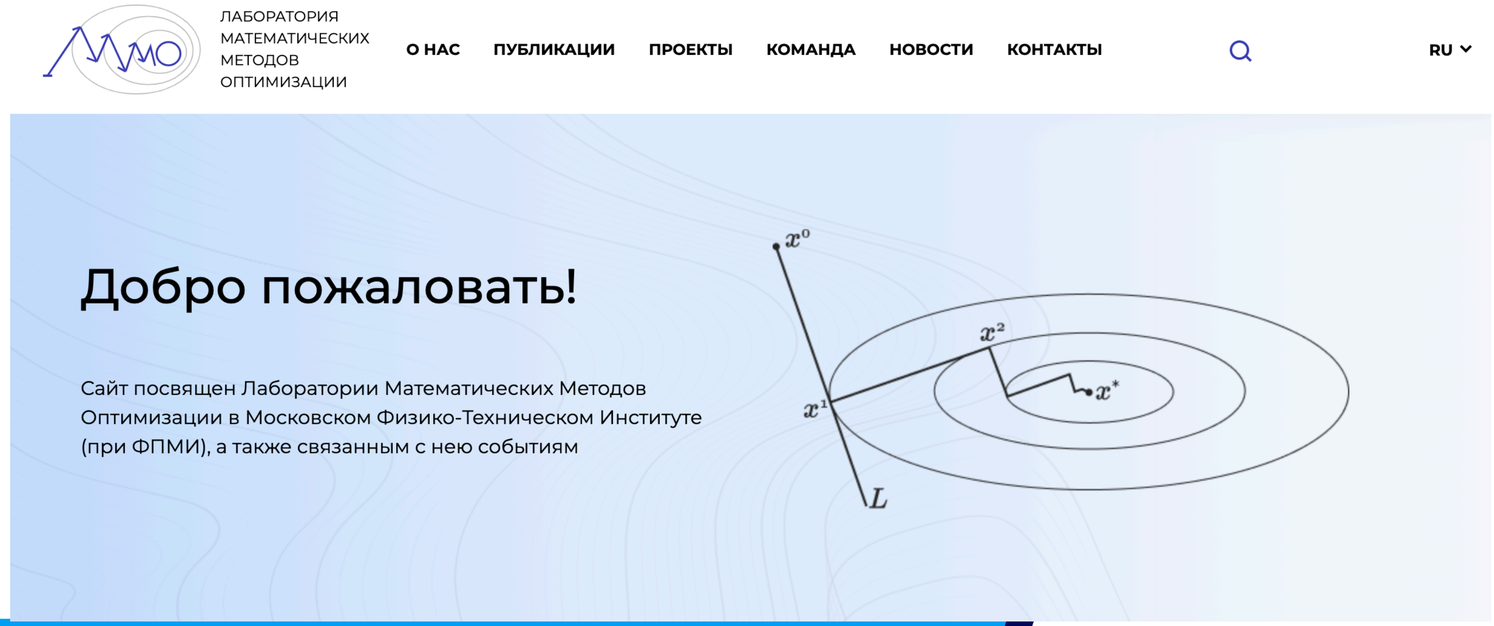
On Solving Minimization and Min-Max Problems by First-Order Methods with Relative Error in Gradients

Gasnikov Alexander
(Innopolis University, MIPT, Steklov's MI RAS)
gasnikov@yandex.ru

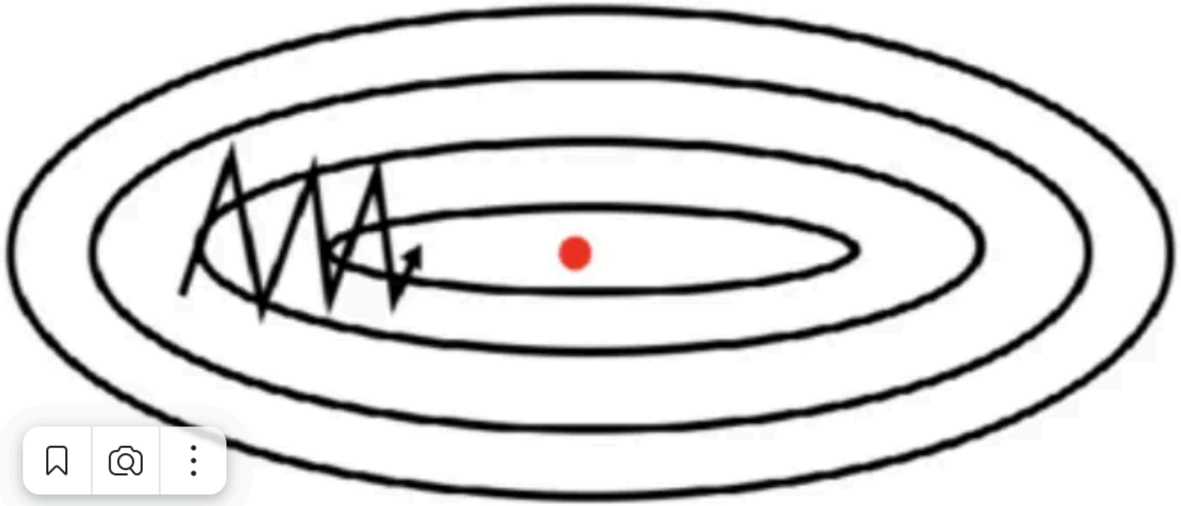
April 2, 2025

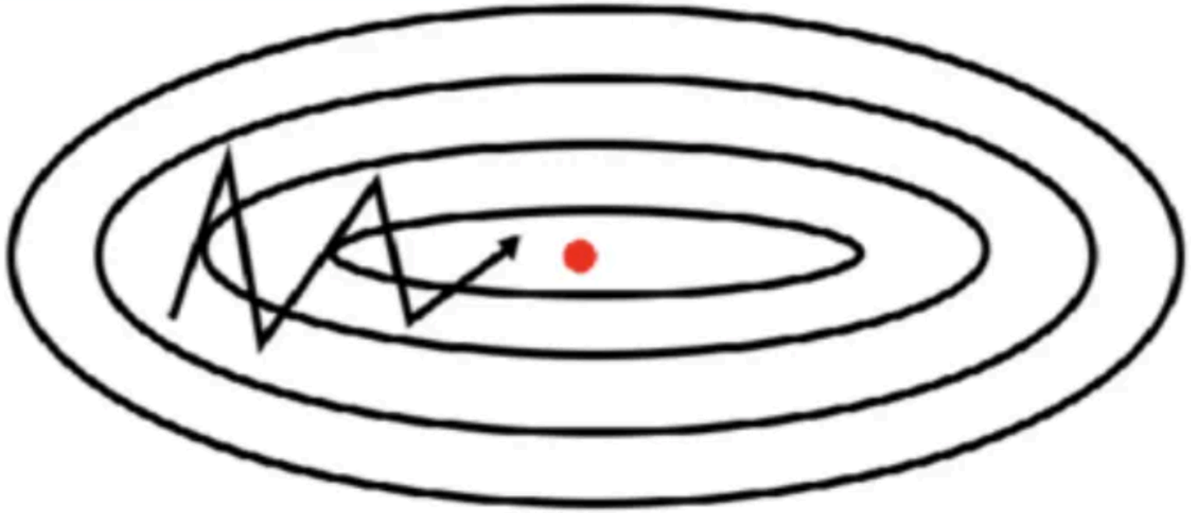
Web-cite of our Lab

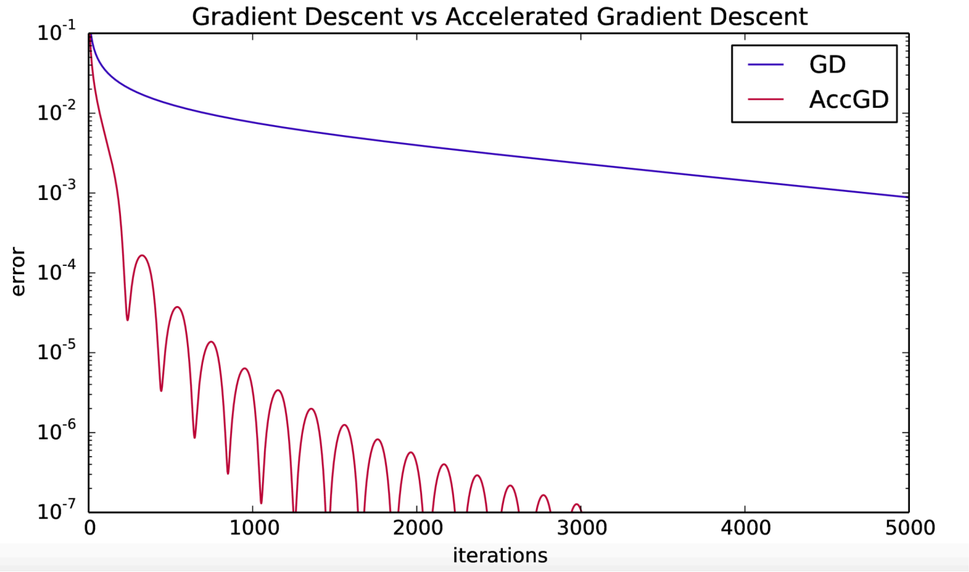


Web-cite: https://labmmo.ru/

Gelfand-Tsetlin's Acceleration







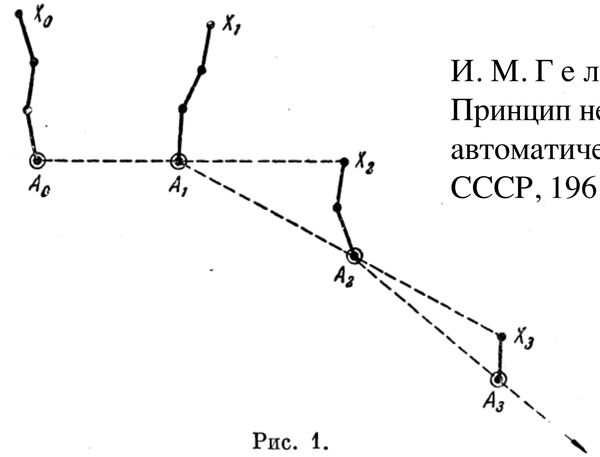


Рис. 1.

И. М. Г е л ь ф а н д, М. Л. Ц е т л и н. Принцип нелокального поиска в системах автоматической оптимизации. Докл. АН СССР, 1961, 137, № 2, 295—298.

Heavy-ball (momentum) method

ЖУРНАЛ
ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ и МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Том 4 Сентябрь 1964 Октябрь № 5

УДК 518:517.948

О НЕКОТОРЫХ СПОСОБАХ УСКОРЕНИЯ СХОДИМОСТИ ИТЕРАЦИОННЫХ МЕТОДОВ

Б. Т. ПОЛЯК

(Москва)

§ 2. Примеры, вычислительный аспект

Рассмотрим теперь более подробно некоторые двухшаговые методы и покажем, что они действительно дают ускорение сходимости по сравнению с соответствующими одношаговыми. А именно, мы изучим метод

$$x^{n+1} = x^n - \alpha P(x^n) + \beta (x^n - x^{n-1}) \qquad (9)$$

и его непрерывный аналог

$$\frac{d^2x}{dt^2} = \alpha_1 \frac{dx}{dt} + \alpha_2 P(x). \qquad (10)$$

Б. Т. Поляк, “О некоторых способах ускорения сходимости итерационных методов”, *Ж. вычисл. матем. и матем. физ.*, **4**:5 (1964), 791–803

Nesterov Y. E. A method of solving a convex programming problem with convergence rate O(1/k^2) // Doklady Akademii Nauk. – Russian Academy of Sciences, 1983. – V. 269. – №. 3. – P. 543-547.

d’Aspremont A. et al. Acceleration methods // Foundations and Trends® in Optimization. – 2021. – V. 5. – №. 1-2. – P. 1-245.

Nesterov's acceleration

$$\min_x f(x)$$

$$x^1 = x^0 - \frac{1}{L}\nabla f(x^0),$$

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

$$f(x^N) - f(x_*) \lesssim \frac{LR^2}{N^2}$$

For smooth convex problems if non-accelerated method converges after 10^6 iterations, then accelerated one after 10^3

```
optimizer = optim.SGD(model.parameters(), lr = 0.01, momentum=0.9)
optimizer = optim.Adam([var1, var2], lr = 0.0001)
```


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.

https://opt.mipt.ru/posobie.pdf

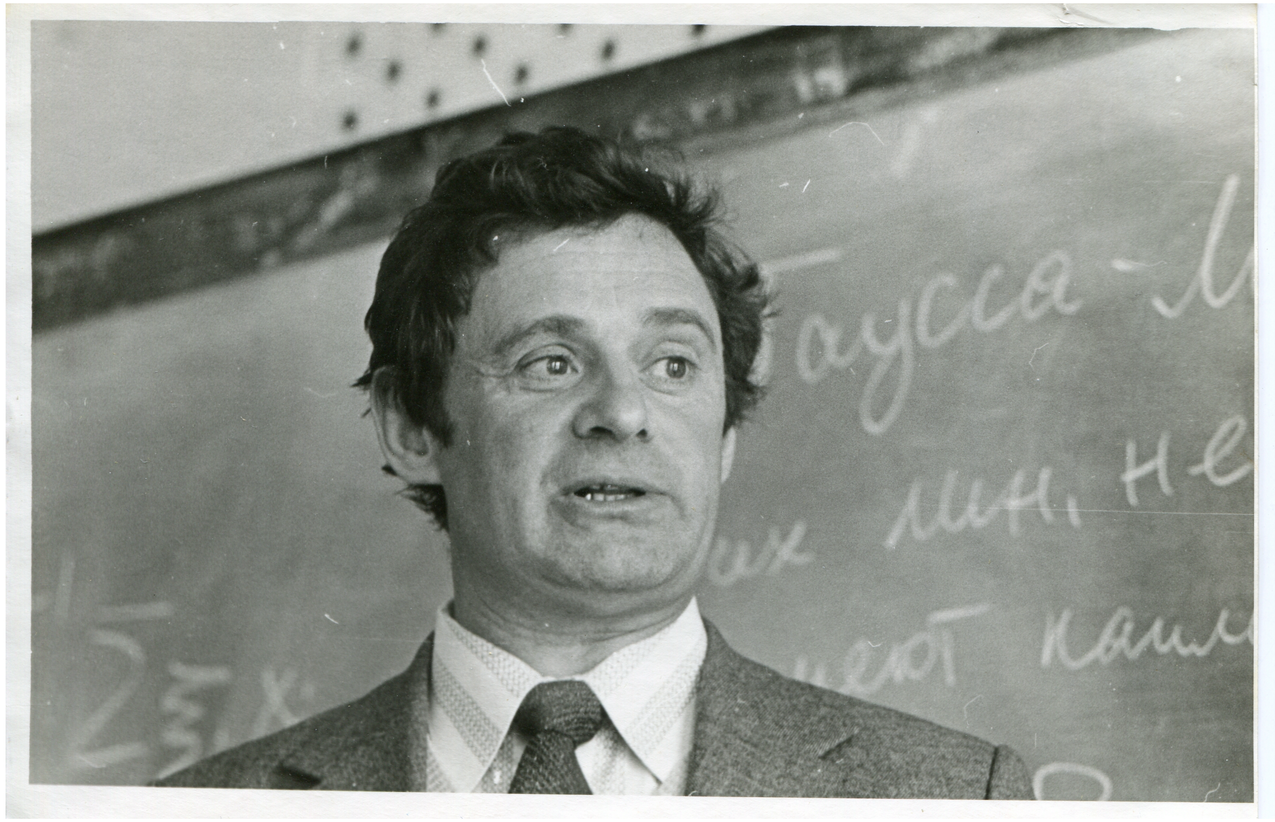
Неточность в градиенте и ранняя остановка

$$\min f(x),$$

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leqslant \delta$$

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2.$$

$$f(x^N) - f(x_*) \lesssim \frac{LR^2}{N^2} + R\delta$$



B. Poljak, Iterative algorithms for singular minimization problems, in Nonlinear Programming 4, Elsevier, 1981, pp. 147–166.

Devolder O. Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. – PhD thesis, 2013.

Vasin A. et al. Accelerated gradient methods with absolute and relative noise in the gradient // Optimization Methods and Software. – 2023. – P. 1-50.

Градиентные методы в условиях относительно неточных градиентов

$$\min f(x),$$

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \alpha \|\nabla f(x)\|_2,$$

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2.$$

$$f(x) - f(x_*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2,$$

Условие PL (Поляка-Лояcиевича)

$$x^{k+1} = x^k - \frac{1}{L}\widetilde{\nabla} f(x^k)$$

Если вместо PL-условия μ-сильная выпуклость

$$f(x^N) - f(x_*) \leq \left(1 - \frac{\mu}{L}\frac{(1-\alpha)^2}{(1+\alpha)^2}\right)^N \left(f(x^0) - f(x_*)\right). \quad \frac{(1-\alpha)^2}{(1+\alpha)^2} \to O\left(\frac{1-\alpha}{1+\alpha}\right).$$

Поляк Б.Т. Введение в оптимизацию. — М.: Наука, 1983.

De Klerk E., Glineur F., Taylor A. B. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions // Optimization Letters. – 2017. – V. 11. – P. 1185-1199.

Vasin A. et al. Accelerated gradient methods with absolute and relative noise in the gradient // Optimization Methods and Software. – 2023. – P. 1-50.

Ускоренные градиентные методы в условиях относительно неточных градиентов

$$\min f(x),$$

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \alpha \|\nabla f(x)\|_2,$$

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2.$$

Для получения ускоренной сходимости в μ-сильно выпуклом случае требуется

$$\alpha = O\left(\left(\frac{\mu}{L}\right)^{1/2}\right)$$

Для неускоренных методов

$$\alpha < 1$$

Vasin A. et al. Accelerated gradient methods with absolute and relative noise in the gradient // Optimization Methods and Software. – 2023. – P. 1-50.
Kornilov N. et al. Intermediate Gradient Methods with Relative Inexactness //arXiv preprint arXiv:2310.00506. – 2023.

Основные результаты (январь 2025)

On Solving Minimization and Min-Max Problems by First-Order Methods with Relative Error in Gradients

Artem Vasin [*1] **Valery Krivchenko** [*1] **Dmitry Kovalev** [2 3] **Fedyor Stonyakin** [1 4] **Nazari Tupitsa** [5]
Pavel Dvurechensky [6] **Mohammad Alkousa** [4 1] **Nikita Kornilov** [1 7] **Alexander Gasnikov** [4 1 8]

Почему естественно относительная неточность

As descrbied in (Overton, 2001) the approximation error of a positive real number x with floating-point numbers $\hat{x}$ is given by

$$|x - \hat{x}| \leqslant \delta |x|, \tag{1}$$

where $\delta = 2^{-p}$ is called machine delta, and a precision p is a number of bits used for the mantissa (or significant digits) in the floating-point representation. Indeed, we can denote a number in floating-point format as:

$$\begin{gathered} \hat{x} = (-1)^s B \cdot 2^{E - E_0}, \\ B = 1.b_0 b_1 \dots b_p, \ b_j \in \{0, 1\}, \ s \in \{0, 1\}. \end{gathered} \tag{2}$$

We call machine delta δ such gap between 1 and next larger floating point number. One can see, that $\delta = 0.0 \dots 1 = 2^{-p}$. Using notation from (Overton, 2001) we can define unit in the last place function: $\text{ulp}(\hat{x}) = 0.0 \dots 12^{E - E_0} = \delta 2^{E - E_0}$. Note, that for positive floating-point numbers $\hat{x}$, value $\hat{x} + \text{ulp}(\hat{x})$ will be next larger floating point number.

Гладкость и сильная выпуклость

Assumption 2.1. f is L-smooth and μ-strongly convex, i.e., $\forall x, y \in \mathbb{R}^d$

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad (7)$$

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2. \quad (8)$$

We also make the following assumption about the inexact gradients of the objective function f.

Гипотеза о неточности градиента

Assumption 2.2 (Relative noise)**.** We assume access to an inexact gradient $\widetilde{\nabla} f(x)$ of f with relative error, meaning that there exists $\alpha \in [0, 1)$, s.t.

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leqslant \alpha \|\nabla f(x)\|_2, \forall x \in \mathbb{R}^d. \tag{9}$$

Следствие предположения 2.2

Corollary 2.3. *Assumption 2.2 for all* $\alpha \in [0, 1)$ *implies the following*

$$\nu \|\nabla f(x)\|_2 \leqslant \|\widetilde{\nabla} f(x)\|_2 \leqslant \rho \|\nabla f(x)\|_2, \tag{10}$$

$$\langle \widetilde{\nabla} f(x), \nabla f(x) \rangle \geqslant \gamma \|\widetilde{\nabla} f(x)\|_2 \|\nabla f(x)\|_2, \tag{11}$$

where $\gamma = \sqrt{1 - \alpha^2}, \nu = 1 - \alpha, \rho = 1 + \alpha.$

Ускоренный метод

Algorithm 1 RE-AGM (Relative Error Accelerated Gradient Method).

Input: $(L, \mu, x_{\text{start}}, \alpha)$.

Set $h = \left(\frac{1-\alpha}{1+\alpha}\right)^{3/2} \frac{1}{L}$.

Set $\widehat{L} = \frac{1+\alpha}{(1-\alpha)^3} L$.

Set $u^0 = x_{\text{start}}$, $x^0 = u^0$.

Set $s = 1 + 2\alpha + 2\alpha^2$.

Set $m = 1 - 2\alpha$.

Set $q = \mu / \widehat{L}$.

for $k = 0, \ldots, N-1$ **do**

Solve the quadratic equation

$$m a_k^2 + (s-m) a_k - q = 0,$$

and take the largest root, i.e.,

$$a_k = \frac{(m-s) + \sqrt{(s-m)^2 + 4mq}}{2m}, \qquad (12)$$

$y^k = \frac{a_k u^k + x^k}{1 + a_k}$,

$u^{k+1} = (1 - a_k) u^k + a_k y^k - \frac{a_k}{\mu} \widetilde{\nabla} f(y^k)$,

$x^{k+1} = y^k - h \widetilde{\nabla} f(y^k)$.

end for

Output: x^N.

Ускоренный метод

Theorem 2.4. *Let Assumption 2.1 hold. If Assumption 2.2 holds with* $\alpha < \frac{\sqrt{2}-1}{18\sqrt{2}}\sqrt{\frac{\mu}{L}}$, *then Algorithm 1 with parameters* $(L, \frac{\mu}{2}, x^0, \alpha)$ *generates* x^N, *s.t.*

$$f(x^N) - f^* \leqslant LR^2\left(1 - \frac{1}{10\sqrt{2}}\sqrt{\frac{\mu}{L}}\right)^N,$$

where $R := \|x^0 - x^*\|_2$.

Algorithm 1 RE-AGM (Relative Error Accelerated Gradient Method).

Input: $(L, \mu, x_{\text{start}}, \alpha)$.
Set $h = \left(\frac{1-\alpha}{1+\alpha}\right)^{3/2}\frac{1}{L}$.
Set $\widehat{L} = \frac{1+\alpha}{(1-\alpha)^3}L$.
Set $u^0 = x_{\text{start}}$, $x^0 = u^0$.
Set $s = 1 + 2\alpha + 2\alpha^2$.
Set $m = 1 - 2\alpha$.
Set $q = \mu/\widehat{L}$.
for $k = 0, \ldots, N-1$ **do**
  Solve the quadratic equation

$$ma_k^2 + (s-m)a_k - q = 0,$$

  and take the largest root, i.e.,

$$a_k = \frac{(m-s) + \sqrt{(s-m)^2 + 4mq}}{2m}, \qquad (12)$$

  $y^k = \frac{a_k u^k + x^k}{1+a_k}$,
  $u^{k+1} = (1-a_k)u^k + a_k y^k - \frac{a_k}{\mu}\widetilde{\nabla} f(y^k)$,
  $x^{k+1} = y^k - h\widetilde{\nabla} f(y^k)$.
end for
Output: x^N.

Ускоренный метод

Theorem 2.5. *Let Assumption 2.1 hold. If Assumption 2.2 holds with* $\alpha = \frac{1}{3}\left(\frac{\mu}{L}\right)^{\frac{1}{2}-\tau}$, *where* $0 \leqslant \tau \leqslant \frac{1}{2}$, *then Algorithm 1 with parameters* $(L, \frac{\mu}{2}, x^0, \alpha)$ *generates* x^N, *s.t.*

$$f(x^N) - f^* \leqslant LR^2\left(1 - \frac{1}{10\sqrt{2}}\left(\frac{\mu}{L}\right)^{\frac{1}{2}+\tau}\right)^N,$$

where $R := \|x^0 - x^*\|_2$.

Algorithm 1 RE-AGM (Relative Error Accelerated Gradient Method).

Input: $(L, \mu, x_{\text{start}}, \alpha)$.
Set $h = \left(\frac{1-\alpha}{1+\alpha}\right)^{3/2}\frac{1}{L}$.
Set $\widehat{L} = \frac{1+\alpha}{(1-\alpha)^3}L$.
Set $u^0 = x_{\text{start}}$, $x^0 = u^0$.
Set $s = 1 + 2\alpha + 2\alpha^2$.
Set $m = 1 - 2\alpha$.
Set $q = \mu/\widehat{L}$.
for $k = 0, \ldots, N-1$ **do**
  Solve the quadratic equation

$$ma_k^2 + (s-m)a_k - q = 0,$$

  and take the largest root, i.e.,

$$a_k = \frac{(m-s) + \sqrt{(s-m)^2 + 4mq}}{2m}, \qquad (12)$$

  $y^k = \frac{a_k u^k + x^k}{1+a_k}$,
  $u^{k+1} = (1-a_k)u^k + a_k y^k - \frac{a_k}{\mu}\widetilde{\nabla} f(y^k)$,
  $x^{k+1} = y^k - h\widetilde{\nabla} f(y^k)$.
end for
Output: x^N.

Ускоренный метод

Remark 2.6. Substituting $\tau = 0$ and $\tau = \frac{1}{2}$ we obtain

1. If $\tau = 0$ and $\alpha = \frac{1}{3}\sqrt{\frac{\mu}{L}}$, then

$$f(x^N) - f^* \leqslant LR^2 \left(1 - \frac{1}{10\sqrt{2}}\sqrt{\frac{\mu}{L}}\right)^N.$$

This is the optimal convergence rate for the class of functions satisfying Assumption 2.1 meaning that for the error α up to $\frac{1}{3}\sqrt{\frac{\mu}{L}}$ our algorithm has optimal convergence rate, and, hence optimal iteration complexity.

2. If $\tau = \frac{1}{2}$ and $\alpha = \frac{1}{3}$, then

$$f(x^N) - f^* \leqslant LR^2 \left(1 - \frac{1}{10\sqrt{2}}\frac{\mu}{L}\right)^N.$$

This convergence rate is the same as for gradient descent meaning that for values of the error α up to $\frac{1}{3}$ our algorithm still has linear convergence with the rate no worse than for the gradient descent.

Algorithm 1 RE-AGM (Relative Error Accelerated Gradient Method).

Input: $(L, \mu, x_{\text{start}}, \alpha)$.
Set $h = \left(\frac{1-\alpha}{1+\alpha}\right)^{3/2}\frac{1}{L}$.
Set $\widehat{L} = \frac{1+\alpha}{(1-\alpha)^3}L$.
Set $u^0 = x_{\text{start}}$, $x^0 = u^0$.
Set $s = 1 + 2\alpha + 2\alpha^2$.
Set $m = 1 - 2\alpha$.
Set $q = \mu/\widehat{L}$.
for $k = 0, \ldots, N-1$ **do**
  Solve the quadratic equation

$$ma_k^2 + (s-m)a_k - q = 0,$$

  and take the largest root, i.e.,

$$a_k = \frac{(m-s) + \sqrt{(s-m)^2 + 4mq}}{2m}, \qquad (12)$$

  $y^k = \frac{a_k u^k + x^k}{1+a_k}$,
  $u^{k+1} = (1-a_k)u^k + a_k y^k - \frac{a_k}{\mu}\widetilde{\nabla} f(y^k)$,
  $x^{k+1} = y^k - h\widetilde{\nabla} f(y^k)$.
end for
Output: x^N.

Седловая задача

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x, y)$$

Assumption 3.1. For given constants $(\mu_x, \mu_y, L_x, L_y, L_{xy})$ that satisfy $0 \leqslant \mu_x \leqslant L_x$, $0 \leqslant \mu_y \leqslant L_y$, $L_{xy} \geqslant 0$, we say that a differentiable function $f(x, y) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is (μ_x, μ_y)-strong-convex-strong-concave (SCSC) and has (L_x, L_y, L_{xy})-Lipschitz gradients (i.e., belongs to the class of functions denoted as $S_{\mu_x \mu_y L_x L_y L_{xy}}$) if $\forall x, x^0, x^1 \in \mathbb{R}^{d_x}, y, y^0, y^1 \in \mathbb{R}^{d_y}$

$$f(\cdot, y) : \mu_x - \text{strongly convex},$$
$$f(x, \cdot) : \mu_y - \text{strongly concave},$$
$$\|\nabla_x f(x^1, y) - \nabla_x f(x^0, y)\|_2 \leqslant L_x \|x^1 - x^0\|_2,$$
$$\|\nabla_y f(x, y^1) - \nabla_y f(x, y^0)\|_2 \leqslant L_y \|y^1 - y^0\|_2,$$
$$\|\nabla_x f(x, y^1) - \nabla_x f(x, y^0)\|_2 \leqslant L_{xy} \|y^1 - y^0\|_2,$$
$$\|\nabla_y f(x^1, y) - \nabla_y f(x^0, y)\|_2, \leqslant L_{xy} \|x^1 - x^0\|_2.$$

Седловая задача

We use the following additional notation that combines partial gradients into one operator: $g_z(z) = [g_x(z)^\top, -g_y(z)^\top]^\top$ where $g_x(z) = \nabla_x f(z)$, $g_y(z) = \nabla_y f(z)$, and $z = [x^\top, y^\top]^\top$.

Assumption 3.2. We say that convex-concave function $F(x, y)$ is a composite with bilinear coupling (i.e. belongs to the class of functions denoted as $K_{\mu_x \mu_y L_x L_y L_{xy}}$) if there exist such $f(x)$: μ_x-strongly-convex with L_x-Lipschitz gradient, $g(y)$: μ_y-strongly-convex with L_y-Lipschitz gradient and matrix A with largest singular value bounded by L_{xy} such that

$$F(x, y) = f(x) + y^\top A x - g(y).$$

Седловая задача

Assumption 3.3. We say that convex-concave function $f(x, y)$ has μ-strongly-monotone, L-Lipschitz gradient (i.e. belongs to the class of functions denoted as $S_{\mu,L}$) if for $\forall z^0, z^1 \in \mathbb{R}^d$:

$$\langle g_z(z^1) - g_z(z^0), z^1 - z^0 \rangle \geqslant \mu \|z^1 - z^0\|_2^2,$$
$$\|g_z(z^1) - g_z(z^0)\|_2 \leqslant L\|z^1 - z^0\|_2.$$

We also assume that an algorithm can use inexact partial gradients $\widetilde{\nabla}_x f(x, y)$ and $\widetilde{\nabla}_y f(x, y)$ that satisfy the following inequality for some $\alpha \in [0, 1)$

$$\|\widetilde{\nabla}_x f(x, y) - \nabla_x f(x, y)\|_2^2 + \|\widetilde{\nabla}_y f(x, y) - \nabla_y f(x, y)\|_2^2$$
$$\leqslant \alpha^2 \left(\|\nabla_x f(x, y)\|_2^2 + \|\nabla_y f(x, y)\|_2^2\right).$$

Седловая задача

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x, y)$$

Algorithm 2 Inexact Sim-GDA.

Input: $(x^0 = x_{\text{start}}, y^0 = y_{\text{start}}, \eta_x, \eta_y)$.
for $k = 1, \ldots, N$ **do**
  $x^{k+1} = x^k - \eta_x \widetilde{\nabla}_x f(x^k, y^k)$,
  $y^{k+1} = y^k + \eta_y \widetilde{\nabla}_y f(x^k, y^k)$.
end for
Output: z^N.

Седловая задача

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x, y)$$

Algorithm 3 Inexact Alt-GDA.

Input: $(x^0 = x_{\text{start}}, y^0 = y_{\text{start}}, \eta_x, \eta_y)$.
for $k = 1, \ldots, N$ **do**
  $x^{k+1} = x^k - \eta_x \widetilde{\nabla}_x f(x^k, y^k)$,
  $y^{k+1} = y^k + \eta_y \widetilde{\nabla}_y f(x^{k+1}, y^k)$.
end for
Output: z^N.

Седловая задача

Consider convex-concave functions with μ-strongly-monotone, L-Lipschitz gradients, i.e. functions that belong to $S_{\mu,L}$. We can guarantee the following iteration complexity for Inexact Sim-GDA.

Theorem 3.4. *Let Inexact Sim-GDA Algorithm 2 be applied to Problem (5) with $f \in S_{\mu,L}$ (Assumption 3.3). Then, the sequence z^N generated by this algorithm linearly converges to the saddle point z^*. Moreover, the iteration complexity is*

$$\widetilde{\mathcal{O}}\left(\frac{L^2}{\mu^2}\frac{1}{(1-\alpha L/\mu)^2}\right),$$

where we omit logarithmic factors.

Theorem 3.5. *There exists such function $f \in S_{\mu,L}$ (Assumption 3.3) with $d_x = d_y = 1$ such that Inexact Sim-GDA with any constant step size η loses linear convergence when $\alpha \geqslant \frac{\mu}{L}$.*

Седловая задача

For exact Sim-GDA on $S_{\mu_x \mu_y L_x L_y L_{xy}}$ (Lee et al., 2024) obtained an iteration complexity estimate

$$\mathcal{O}\left(\frac{L_x}{\mu_x} + \frac{L_y}{\mu_y} + \frac{L_{xy}^2}{\mu_x \mu_y}\right)$$

Theorem 3.6. *Consider $f \in S_{\mu\mu LLL_{xy}}$ (Assumption 3.1). Inexact Sim-GDA retains linear convergence if*

$$\textit{case} \quad L_{xy}^2 \leqslant \frac{\mu L}{2}: \quad \alpha\left((1+\alpha)^2 + \alpha\right) < \frac{\mu}{2L} - \frac{L_{xy}^2}{2L^2},$$

$$\textit{case} \quad L_{xy}^2 > \frac{\mu L}{2}: \quad \alpha\left((1+\alpha)^2 + \alpha\right) < \frac{\mu^2}{8L_{xy}^2}.$$

Решение нелинейных уравнений

$$g(z) = 0$$

Assumption 4.1. For given constants (μ, L) that satisfy $0 \leqslant \mu \leqslant L$, we say that an operator $g(z) : \mathbb{R}^d \to \mathbb{R}^d$ is μ-quasi-strongly-monotone (QSM) and L-Lipschitz (i.e., belongs to the class denoted as $S_{\mu,L}$) if $\forall x, y \in \mathbb{R}^d$ the following holds

$$\|g(x) - g(y)\| \leq L\|x - y\|, \tag{14}$$

$$\|\widetilde{g}(z) - g(z)\|_2 \leqslant \alpha\|g(z)\|_2$$

$$\langle g(x), x - x^* \rangle \geq \mu\|x - x^*\|^2. \tag{15}$$

Решение нелинейных уравнений

Algorithm 4 Inexact ExtraGradient Method (EG).

Input: $(z^0 = z_{\text{start}}, \eta)$.
for $k = 1, \ldots, N$ **do**
  $z^{k+1/2} = z^k - \eta \widetilde{g}(z^k)$,
  $z^{k+1} = z^k - \eta \widetilde{g}(z^{k+1/2})$.
end for
Output: z^N.

Theorem 4.2. *For $g \in S_{\mu,L}$ (Assumption 4.1), $\exists \widehat{\alpha}$: $\widehat{\alpha} = \mathcal{O}\left(\sqrt{\mu/L}\right)$, s.t. if $\alpha < \widehat{\alpha}$ then inexact EG retains linear convergence with contraction factor of $1 - \frac{1}{2}\eta\mu$, $\eta \sim \frac{1}{L}$.*

Численные эксперименты

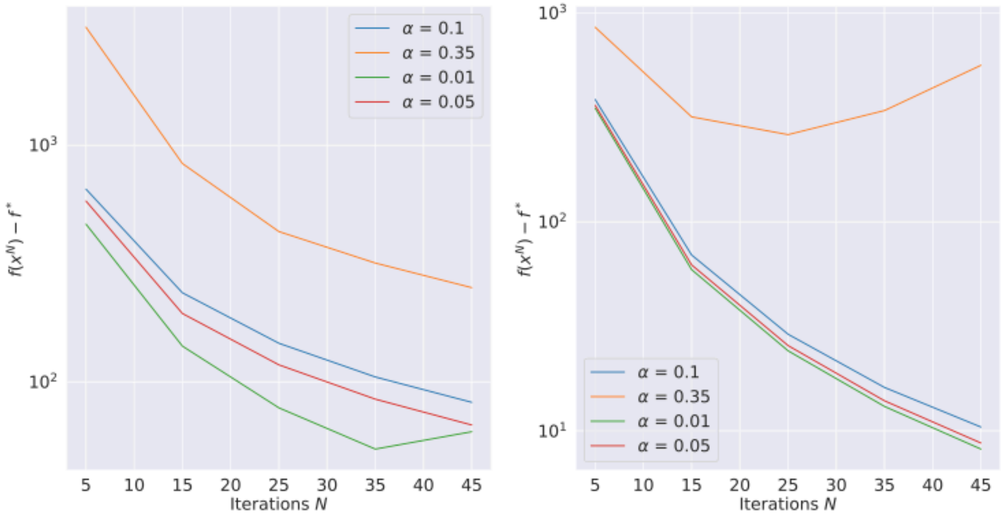


Figure 1. PEP comparison RE-AGM (left) and STM (right), $L = 100, \mu = 0.01$

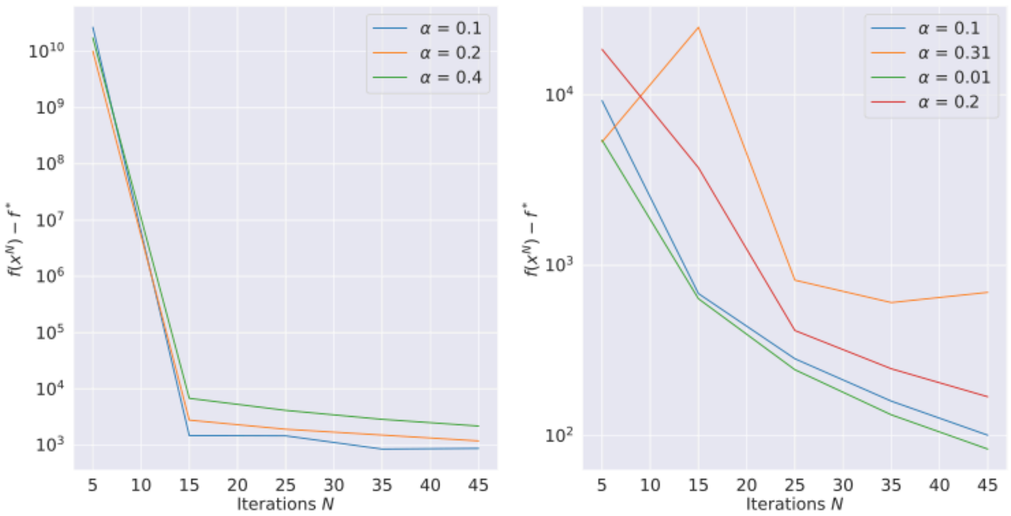


Figure 2. PEP comparison RE-AGM (left) and STM (right), $L = 1000, \mu = 0.001$

Пример доказательства нижней оценки

Theorem 3.5 *There exists such function $f \in S_{\mu,L}$ (Assumption 3.3) with $d_x = d_y = 1$ such that inexact Sim-GDA with any constant step size η loses linear convergence when $\alpha \geqslant \frac{\mu}{L}$.*

Proof of Theorem 3.5. Consider function

$$F(x, y) = \frac{\epsilon}{2}x^2 + xy - \frac{\epsilon}{2}y^2,$$

where $\epsilon > 0$. We make one step of exact Sim-GDA:

$$\begin{aligned} x^{k+1} &= x^k - \eta\left(y^k + \epsilon x^k\right) \\ y^{k+1} &= y^k + \eta\left(x^k - \epsilon y^k\right), \end{aligned}$$

and introduce operator $g(x^k, y^k) = \left[(y^k + \epsilon x^k)^\top, (-x^k + \epsilon y^k)^\top\right]^\top$:

$$z^{k+1} = z^k - \eta g(x^k, y^k).$$

Пример доказательства нижней оценки

The operator g is μ-strongly-monotone and L-Lipshitz with $\mu = \epsilon$ and $L = \sqrt{1+\epsilon^2}$. Let $\widetilde{g}(x^k, y^k)$ be a disturbed value of g in (x^k, y^k). Then

$$\|g(x^k, y^k) - \widetilde{g}(x^k, y^k)\| = \epsilon\sqrt{(x^k)^2 + (y^k)^2} \leqslant \frac{\epsilon}{\sqrt{1+\epsilon^2}}\|g(x^k, y^k)\| = \frac{\mu}{L}\|g(x^k, y^k)\|,$$

which means that g satisfies relative inexactness definition with $\alpha = \frac{\mu}{L}$. A step of inexact Sim-GDA with disturbed $\widetilde{g}$ leads to

$$\begin{aligned} x^{k+1} &= x^k - \eta y^k, \\ y^{k+1} &= y^k + \eta x^k. \end{aligned}$$

For any $\eta > 0$:

$$(x^{k+1})^2 + (y^{k+1})^2 = \left(1+\eta^2\right)\left((x^k)^2 + (y^k)^2\right) > (x^k)^2 + (y^k)^2,$$

and we have divergence. □