Stopping rules for accelerated gradient methods with additive noise in gradient and influence of relative inexactness.

Vasin Artem

Moscow Institute of Physics and Technology, MIPT

14 april 2021

Basic assumptions and problem description

We consider convex optimization problem on a convex (not necessarily bounded) set $Q \subset \mathbb{R}^n$

$$\min_{x \in Q} f(x)$$

We will use such designations:

$$(\forall x \in Q)\ \|\nabla f(x) - \tilde{\nabla} f(x)\|_2 \leqslant \delta,$$
$$R = \|x_{start} - x^*\|_2,$$
$$(\forall x, y \in Q)\ \|\nabla f(x) - \nabla f(y)\|_2 \leqslant L\|x - y\|_2$$

We will also consider the strongly convex case:

$$(\forall x, y \in Q)\ f(x) + \langle \tilde{\nabla} f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2 \leqslant f(y)$$

Also will study the effect of relative imprecision:

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

Algorithm description

Algorithm 1 $STM(L, \mu, \tau, x_{start}), \quad Q \subseteq \mathbb{R}^n$

Input: Starting point x_{start}, number of steps N
Set $\tilde{x}_0 = x_{start}$,
Set $A_0 = \frac{1}{L}$,
Set $\alpha_0 = \frac{1}{L}$,
$\psi_0(x) = \frac{1}{2}\|x - \tilde{x}_0\|_2^2 + \alpha_0 \left(f(\tilde{x}_0) + \langle \tilde{\nabla} f(\tilde{x}_0), x - \tilde{x}_0 \rangle + \frac{\mu}{2}\|x - \tilde{x}_0\|_2^2 \right)$,
Set $z_0 = \mathrm{argmin}_{y \in Q}\, \psi_0(y)$,
Set $x_0 = z_0$.
for $k = 1 \ldots N$ **do**

$\alpha_k = \frac{1+\mu_\tau A_{k-1}}{2L} + \sqrt{\frac{1+\mu_\tau A_{k-1}}{4L^2} + \frac{A_{k-1}}{1+\mu_\tau A_{k-1}}}$,

$A_k = A_{k-1} + \alpha_k$,

$\tilde{x}_k = \frac{A_{k-1}x_{k-1} + \alpha_k z_{k-1}}{A_k}$,

$\psi_k(x) = \psi_{k-1}(x) + \alpha_k \left((f(\tilde{x}_k) + \langle \tilde{\nabla} f(\tilde{x}_k), x - \tilde{x}_k \rangle + \frac{\mu}{2}\|x - \tilde{x}_k\|_2^2 \right)$,

$z_k = \mathrm{argmin}_{y \in Q}\, \psi_k(y)$,

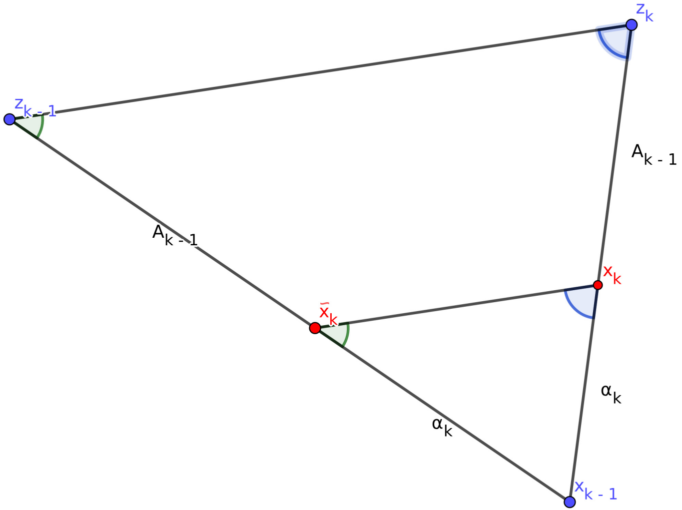
$x_k = \frac{A_{k-1}x_{k-1} + \alpha_k z_k}{A_k}$.

end for
Output: x_N.

Figure 5 describes the position of the vertices. On the sides, not their lengths are marked, but the relationships in the corresponding sides in the similarity of triangles. In the case $Q = \mathbb{R}^n$, we can simplify the step of the algorithm by replacing it with:

$$z_k = z_{k-1} - \frac{\alpha_k}{1 + A_k \mu_\tau} \left(\tilde{\nabla} f(\tilde{x}_k) + \mu_\tau (z_{k-1} - \tilde{x}_k) \right).$$

Alogrithm description



Convergence comparison

We will use the following functions for experiments.

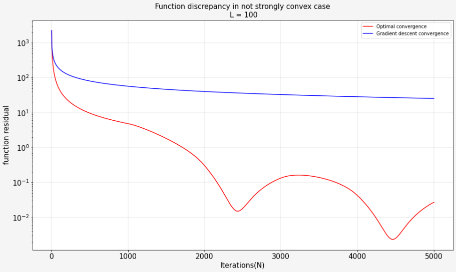
$$n = 2m + 1,$$
$$\left(\forall k \leqslant \frac{1}{2}(n-1)\right)$$
$$f(x_k) - f(x^*) \geqslant \frac{3}{32}\frac{LR^2}{(k+1)^2}$$
$$f(x) = \frac{L}{8}\left(x_1^2 + \sum_{k=1}^{2m}(x_k - x_{k+1})^2 + x_n^2\right) - \frac{L}{4}x_1,$$

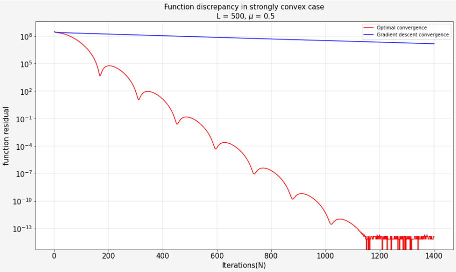
And for strongly convex case:

$$f(x) = \frac{\mu\left(\frac{L}{\mu} - 1\right)}{8}\left(x_1^2 + \sum_{k=1}^{\infty}(x_k - x_{k+1})^2 - 2x_1\right) + \frac{\mu}{2}\|x\|_2^2,$$
$$\chi = \frac{L}{\mu},$$
$$f(x_N) - f(x^*) \geqslant \frac{\mu}{2}\left(\frac{\sqrt{\chi} - 1}{\sqrt{\chi} + 1}\right)^{2N},\ N \geqslant 1,$$

Convergence comparison

Let us compare the convergence of gradient descent and the accelerated STM method.





$$\mu = 0 \Rightarrow f(x_k) - f(x^*) \leqslant \frac{4LR^2}{N^2},$$

$$\mu > 0 \Rightarrow f(x_k) - f(x^*) \leqslant LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{L}}N\right)$$

Convergence with additive noise

Remind the conception:

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

Also we will use:

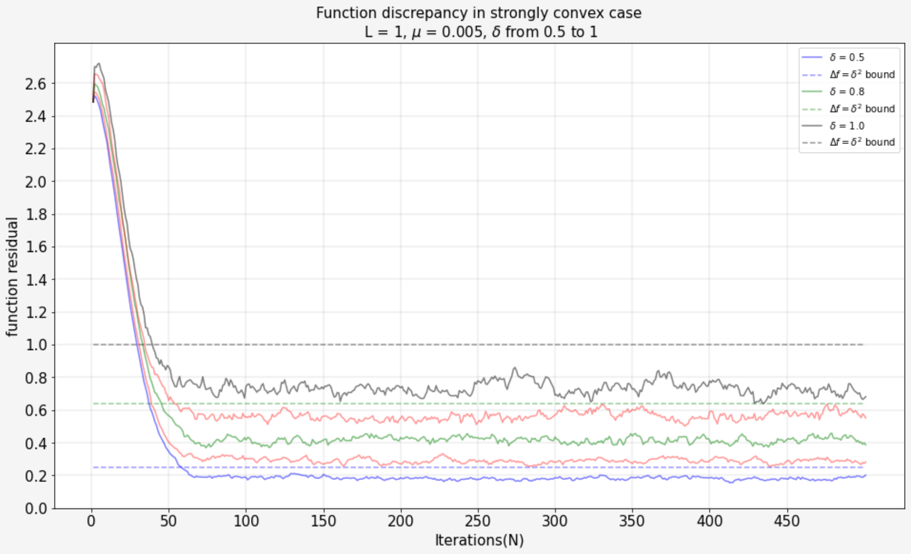
$$\tilde{R}_N = \max_{k \leqslant N}\{\|x_k - x^*\|_2, \|z_k - x^*\|_2, \|\tilde{x}_k - x^*\|_2\}$$

In the presence of the additive noise STM convergence:

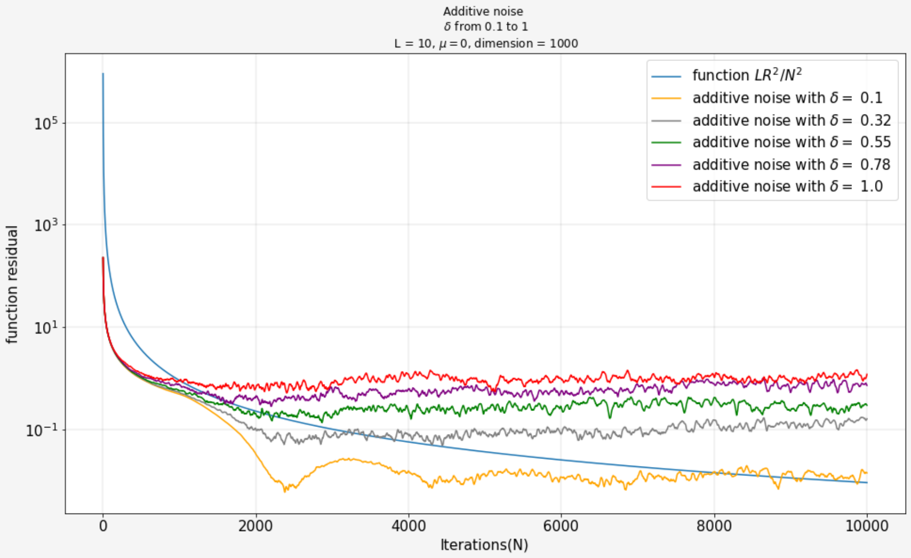
$$f(x_N) - f(x^*) \leqslant \frac{4LR^2}{N^2} + N\frac{\delta^2}{2L} + 3\delta\tilde{R}_N,$$

$$f(x_N) - f(x^*) \leqslant LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}N\right) + \left(1 + \sqrt{\frac{\mu}{2L}}\right)\left(\frac{\delta^2}{2L} + \frac{\delta}{\mu}\right)$$

Plots of convergence



Plots of convergence



Stopping rule for STM

Theorem

Assume that we know we value of $f(x^*)$ and such bound, R_*, that $\|x^*\| \leqslant R_*$. Using stopping rule $\forall \zeta > 0$

$$f(x_N) - f(x^*) \leqslant \frac{\delta^2}{2L} N + 3R_*\delta + \zeta$$

And it is guaranteed, that the criteria will reached in

$$N_{stop} = O\left(\frac{\sqrt{LR^2}}{\zeta}\right).$$

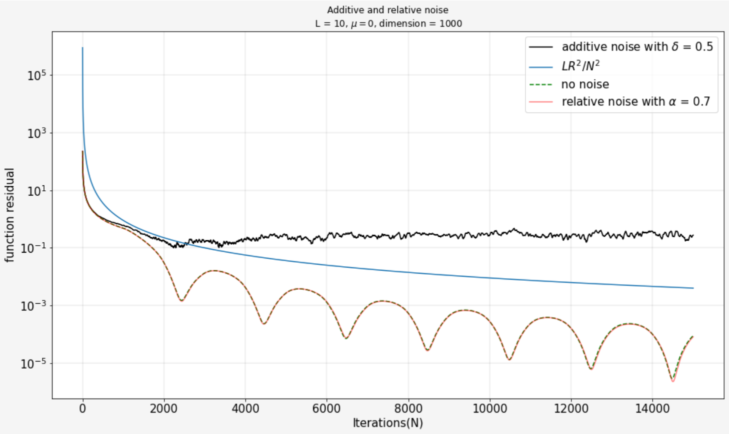
Choosing $\zeta \sim \varepsilon, \delta \sim \frac{\varepsilon}{R_*}$ we get estimation:

$$f(x_{N_{stop}}) - f(x^*) \leqslant \varepsilon,$$

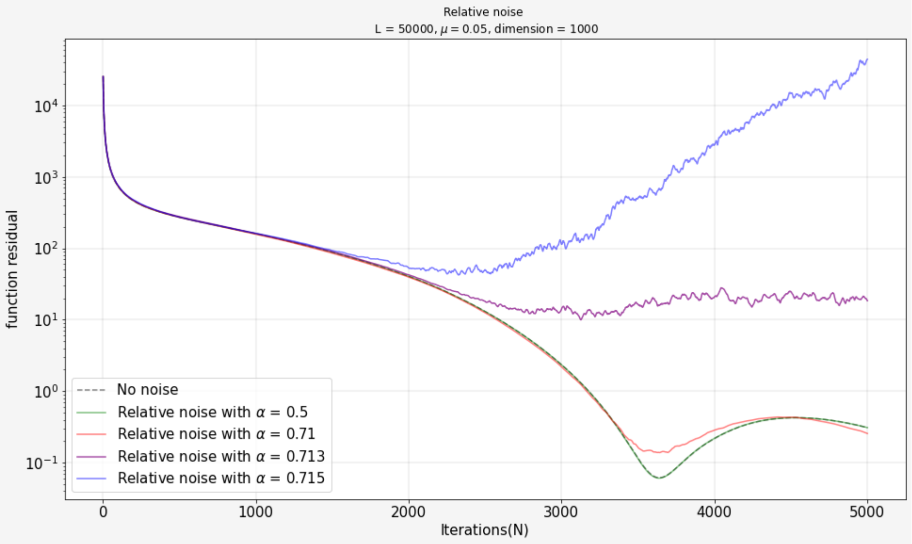
$$N_{stop} = O\left(\frac{\sqrt{LR^2}}{\varepsilon}\right)$$

Remind the conception:

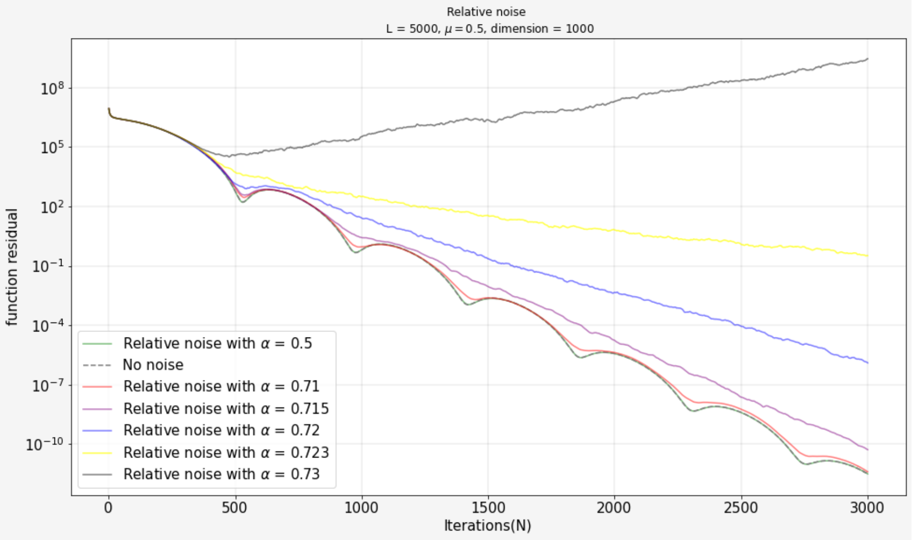
$$\|\nabla f(x) - \tilde{\nabla} f(x)\|_2 \leqslant \alpha \|\nabla f(x)\|_2$$



Threshold $\overline{\alpha}$ value



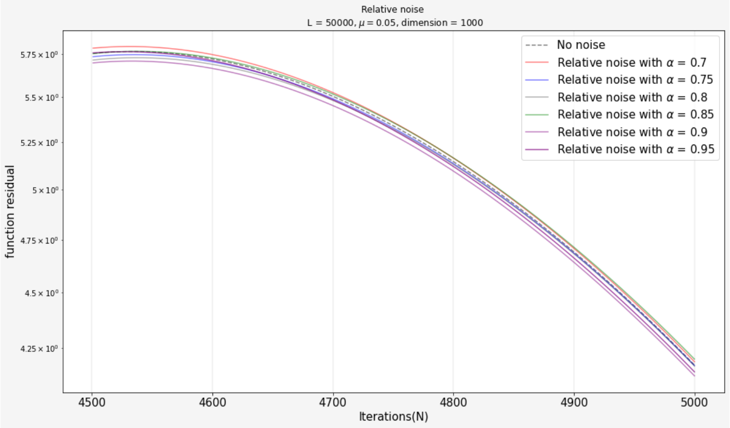
Threshold $\overline{\alpha}$ value



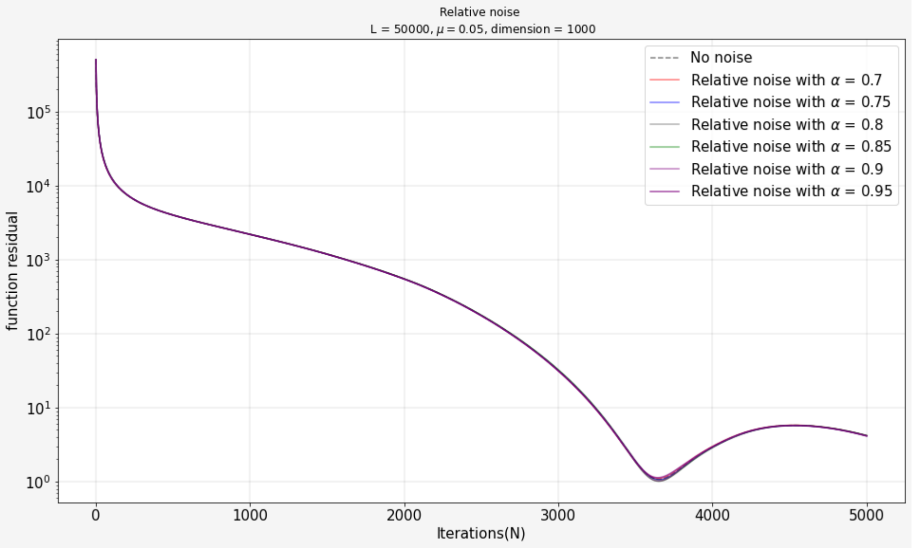
Boosting for relative inexactness

For boosting we can use gradient step in the end of each iteration:

$$x_k := x_k - \frac{1}{L}\tilde{\nabla} f(x_k)$$



Boosting for relative inexactness



Attempts in proofs

Theorem

If $\|\nabla f(\tilde{x}_k)\|_2 = O\left(\|\nabla f(x_k)\|_2\right)$ *and* $\mu > 0$. *Then choosing* $\alpha = O\left(\left(\frac{\mu}{L}\right)^{\frac{3}{4}}\right)$, *we get:*

$$f(x_k) - f(x^*) \leqslant \left(\frac{LR^2}{1-\alpha^2} + \frac{3L\alpha^2}{2\mu(1-\alpha^2)}\left(f(x_0) - f(x^*)\right)\right)\exp\left(-\frac{1}{6\sqrt{2}}\sqrt{\frac{\mu}{L}}k\right)$$

Also we can consider another condition

Theorem

If sequence $\tilde{x}_k$ *satisfies:*

$$f(x_k) - f(x^*) \leqslant O\left(LR^2 exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}k\right) + \left(1 + \sqrt{\frac{4L}{\mu}}\right)\left(\frac{\delta^2}{\mu} + \frac{\delta^2}{L}\right)\right)$$

We can obtain we same convergence for same α *bound.*