Accelerated Directional Search with non-euclidean prox-structure

Eduard Gorbunov

Moscow Institute of Physics and Technology

27 October, 2017

Formulation of the problem

Consider the problem

$$f(x) \rightarrow \min_{\mathbb{R}^n}, \tag{1}$$

where f(x) is convex and L-smooth with respect to $||\cdot||_2$ on $\mathbb{R}^n$, that is, for every $x, y \in \mathbb{R}^n$ it satisfies

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

.

Designations

Let $e \in RS_2^n(1)$, where $RS_2^n(1)$ is uniform distribution of vectors on n-dimensional Euclidean sphere. Instead of using $\nabla f(x)$ we will use its stochastic approximation $n\langle \nabla f(x), e\rangle e$. One can easily show that $\mathbb{E}_e[n\langle \nabla f(x), e\rangle e] = \nabla f(x)$.

Designations

Consider $d : \mathbb{R}^n \to \mathbb{R}$ (and call it *prox-function* or *distance generating function*) which is 1-strongly convex with respect to norm $||\cdot||_p$ $(1 \leqslant p \leqslant 2)$, for example, $d(x) = \frac{1}{2(a-1)}||x||_a^2$, where $a = \frac{2\log n}{2\log n - 1}$, for the case $p = 1$. The Bregman divergence is given as

$$V_z(y) \stackrel{\text{def}}{=} d(y) - d(z) - \langle \nabla d(z), y - z \rangle. \qquad (2)$$

Let $V_{x_0}(x^*) = \Theta$, where x_0 is some initial point and x^* is the minimizer of $f(x)$.

Designations

Let

$$\mathrm{Grad}_e(x) = x - \frac{1}{L}\langle \nabla f(x), e \rangle e, \tag{3}$$

which corresponds the gradient descent step with respect to 2-norm, and

$$\mathrm{Mirr}_e(x, z, \alpha) = \underset{y \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ \alpha \langle n \langle \nabla f(x), e \rangle e, y - z \rangle + V_z(y) \right\}, \tag{4}$$

which corresponds the mirror descent step with respect to 1-norm.

Method

Algorithm 1 ACDS

Require: f — convex and L-smooth with respect to $||\cdot||_2$ on $\mathbb{R}^n$; x_0 — some initial point; N — the number of iterations.

Ensure: y_N such that $\mathbb{E}_{e_1,e_2,\ldots,e_N}[f(y_N)] - f(x^*) \leqslant \frac{4\Theta L C_{n,p}}{N^2}$.

$y_0 \leftarrow x_0,\ z_0 \leftarrow x_0$
for $k = 0, \ldots, N-1$ **do**
  $\alpha_{k+1} \leftarrow \frac{k+2}{2LC_{n,p}},\ \tau_k \leftarrow \frac{1}{\alpha_{k+1} L C_{n,p}} = \frac{2}{k+2}$
  Generate $e_{k+1} \in RS_2^n(1)$ independently of previous iterations
  $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k) y_k$
  $y_{k+1} \leftarrow \mathrm{Grad}_{e_{k+1}}(x_{k+1})$
  $z_{k+1} \leftarrow \mathrm{Mirr}_{e_{k+1}}(x_{k+1},\ z_k,\ \alpha_{k+1})$
end for
return y_N

Convergence rate

Theorem

Let $f(x)$ is convex and L-smooth with respect to $||\cdot||_2$ on $\mathbb{R}^n$, $d(x)$ is 1-strongly convex with respect to norm $||\cdot||_p$ $(1 \leqslant p \leqslant 2)$, N is the number of iterations. Then ACDS outputs y_N satisfying

$$\mathbb{E}_{e_1,e_2,\dots,e_N}[f(y_N)] - f(x^*) \leqslant \frac{4\Theta L C_{n,p}}{N^2}, \tag{5}$$

where $\Theta = V_{x_0}(x^*)$, $C_{n,p} = \frac{4}{3}\min\{q-1, 4\ln n\} \cdot n^{\frac{2}{q}+1}$, $\frac{1}{q} + \frac{1}{p} = 1$.

Remark

Consider two special cases: $p = 2$ and $p = 1$. In the first situation ($p = 2$) one can obtain $C_{n,p} = n^2$ (without factor $\frac{4}{3}$). In the case when $p = 1$ we have $C_{n,p} = \frac{16}{3} n \ln n$.

Proof of this theorem one can find in arXiv preprint 1710.00162 (in Russian).

Parallel trajectories

Assume that we want to obtain such y that $f(y) - f(x^*) \leqslant 2\varepsilon$. In this case we could choose $N = \lceil \sqrt{\frac{4\Theta L C_{n,p}}{\varepsilon}} \rceil$ to guarantee $\mathbb{E}_{e_1,e_2,\ldots,e_N}[f(y_N)] - f(x^*) \leqslant \varepsilon \Leftrightarrow \mathbb{E}_{e_1,e_2,\ldots,e_N}[f(y_N) - f(x^*)] \leqslant \varepsilon$. By Markov's inequality

$$\mathbb{P}\{f(y_N) - f(x^*) \geqslant 2\varepsilon\} \leqslant \frac{\varepsilon}{2\varepsilon} = \frac{1}{2}. \quad (6)$$

It means that if we run $m = \lceil \log_2(\sigma^{-1}) \rceil$ independent realizations (trajectories) of ACDS we will obtain such $y_N^1, y_N^2, \ldots, y_N^m$ that

$$\mathbb{P}\{\min_{i=\overline{1,m}} f(y_N^i) - f(x^*) \geqslant 2\varepsilon\} \leqslant \left(\frac{1}{2}\right)^m \leqslant \sigma. \quad (7)$$

So with probability $1 - \sigma$ minimum among the values $f(y_N^1), f(y_N^2), \ldots, f(y_N^m)$ will satisfy required accuracy.

Discussion

In 2014 Z. Allen-Zhu and L. Orrechia proposed accelerated method based on the idea of coupling gradient and mirror descents. Their method uses gradient (no stochastic approximations) and after N iterations outputs y_N satisfying

$$f(y_N) - f(x^*) \leqslant \frac{4\Theta L}{N^2}. \tag{8}$$

In the case when $p = 1$ our method needs approximately $\frac{n}{\ln n}$ times less arithmetical operations under the assumption that $f(x)$ is defined by black-box model and its gradient is restored by $n+1$ values of $f(x)$.