An Accelerated Directional Derivative Method for Smooth Stochastic Convex Optimization

Eduard Gorbunov[1], Pavel Dvurechensky[2], Alexander Gasnikov[1]

[1]Moscow Institute of Physics and Technology, Moscow, Russia
[2]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

Introduction

Consider the following optimization problem

$$\min_{x\in\mathbb{R}^n}\left\{f(x) := \mathbb{E}_\xi[F(x,\xi)] = \int_{\mathcal{X}} F(x,\xi)dP(x)\right\}, \qquad (1)$$

where ξ — random vector with probability distribution $P(\xi)$, $\xi\in\mathcal{X}$, $F(x,\xi)$ — closed a.s. in ξ, f — convex,

$$\|g(x,\xi)-g(y,\xi)\|_2 \leqslant L(\xi)\|x-y\|_2,\ \forall x,y\in\mathbb{R}^n,\ \text{a.s. in } \xi,$$

and $L_2 := \sqrt{\mathbb{E}_\xi[L(\xi)^2]} < +\infty$. Under this assumptions, $\mathbb{E}_\xi[g(x,\xi)] = \nabla f(x)$ and

$$\|\nabla f(x)-\nabla f(y)\|_2 \leqslant L_2\|x-y\|_2,\ \forall x,y\in\mathbb{R}^n.$$

Also we assume that

$$\mathbb{E}_\xi\left[\|g(x,\xi)-\nabla f(x)\|_2^2\right] \leqslant \sigma^2. \qquad (2)$$

Finally, we assume that an optimization procedure, given a point $x\in\mathbb{R}^n$, direction $e\in S_2(1)$ and ξ independently drawn from P, can obtain a noisy stochastic approximation $\widetilde{f}'(x,\xi,e)$ for the directional derivative $\langle g(x,\xi),e\rangle$:

$$\begin{aligned}\widetilde{f}'(x,\xi,e) &= \langle g(x,\xi),e\rangle + \zeta(x,\xi,e) + \eta(x,\xi,e),\\ \mathbb{E}_\xi\left[\zeta(x,\xi,e)^2\right] &\leqslant \Delta_\zeta,\ \forall x\in\mathbb{R}^n, \forall e\in S_2(1),\\ |\eta(x,\xi,e)| &\leqslant \Delta_\eta,\ \forall x\in\mathbb{R}^n, \forall e\in S_2(1),\ \text{a.s. in } \xi.\end{aligned}$$

We choose a *prox-function* $d(x)$ which is continuous, convex on $\mathbb{R}^n$ and is 1-strongly convex on $\mathbb{R}^n$ with respect to $\|\cdot\|_p$, $p\in[1,2]$. We define also the corresponding *Bregman divergence* $V[z](x) = d(x)-d(z)-\langle\nabla d(z), x-z\rangle$, $x,z\in\mathbb{R}^n$. Moreover,

$$\mathbb{E}_e\|e\|_q^2 \leq \rho_n, \qquad (3)$$

$$\mathbb{E}_e\left[\langle s,e\rangle^2\|e\|_q^2\right] \leq \frac{6\rho_n}{n}\|s\|_2^2,\quad \forall s\in\mathbb{R}^n, \qquad (4)$$

where $\rho_n = \min\{q-1, 16\ln n - 8\}n^{\frac{2}{q}-1}$, $n\geqslant 8$ and $s\in\mathbb{R}^n$.

New methods

Algorithm 1. Accelerated Randomized Directional Derivative (ARDD) method.

Input: x_0 — *starting point;* $N\geqslant 1$ — *number of iterations;* m — *batch size.*

Output: *point* y_N

$y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$
for $k = 0,\ldots, N-1$ **do**
  $\alpha_{k+1} \leftarrow \frac{k+2}{96n^2\rho_nL_2}$, $\tau_k \leftarrow \frac{1}{48\alpha_{k+1}n^2\rho_nL_2} = \frac{2}{k+2}$
  Generate $e_{k+1}\in RS_2(1)$ *independently from previous iterations and* ξ_i, $i=1,\ldots,m$ *– independent realizations of* ξ.
  Calculate
  $$\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m}\sum_{i=1}^m \widetilde{f}'(x_{k+1},\xi_i,e_{k+1})e_{k+1}.$$
  $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k)y_k$.
  $y_{k+1} \leftarrow x_{k+1} - \frac{1}{2L_2}\widetilde{\nabla}^m f(x_{k+1})$.
  $$z_{k+1} \leftarrow \operatorname*{argmin}_{z\in\mathbb{R}^n}\left\{\alpha_{k+1}n\left\langle\widetilde{\nabla}^m f(x_{k+1}), z-z_k\right\rangle + V[z_k](z)\right\}.$$
end for

Theorem 1 [1]. *Let ARDD method be applied to solve problem (1). Then*

$$\begin{aligned}\mathbb{E}[f(y_N)]-f(x^*) \leqslant\ & \frac{384\Theta_pn^2\rho_nL_2}{N^2} + \frac{4N}{nL_2}\cdot\frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2\\ &+\frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2}+2\Delta_\eta\right)\\ &+\frac{N^2}{12n\rho_nL_2}\left(\frac{\sqrt{\Delta_\zeta}}{2}+2\Delta_\eta\right)^2,\end{aligned} \qquad (5)$$

where $\Theta_p = V[z_0](x^*)$ *is defined by the chosen proximal setup and* $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\ldots,e_N,\xi_{1,1},\ldots,\xi_{N,m}}[\cdot]$.

Algorithm 2. Randomized Directional Derivative (RDD) method.

Input: x_0 — *starting point;* $N\geqslant 1$ — *number of iterations;* m — *batch size.*

Output: *point* $\bar{x}_N$.

for $k = 0,\ldots, N-1$ **do**
  $\alpha \leftarrow \frac{1}{48n\rho_nL_2}$.
  Generate $e_{k+1}\in RS_2(1)$ *independently from previous iterations and* ξ_i, $i=1,\ldots,m$ *– independent realizations of* ξ.
  Calculate
  $$\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m}\sum_{i=1}^m \widetilde{f}'(x_{k+1},\xi_i,e_{k+1})e_{k+1}.$$
  $$x_{k+1} \leftarrow \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\alpha n\left\langle\widetilde{\nabla}^m f(x_k), x-x_k\right\rangle + V[x_k](x)\right\}.$$
end for

Theorem 2 [1]. *Let RDD method be applied to solve problem (1). Then*

$$\begin{aligned}\mathbb{E}[f(\bar{x}_N)]-f(x_*) \leqslant\ & \frac{384n\rho_nL_2\Theta_p}{N} + \frac{2}{L_2}\frac{\sigma^2}{m} + \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2\\ &+\frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{\sqrt{\Delta_\zeta}}{2}+2\Delta_\eta\right)\\ &+\frac{N}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2}+2\Delta_\eta\right)^2,\end{aligned} \qquad (6)$$

where $\Theta_p = V[z_0](x^*)$ *is defined by the chosen proximal setup and* $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\ldots,e_N,\xi_{1,1},\ldots,\xi_{N,m}}[\cdot]$.

Method	$p=1$	$p=2$
ARDD	$\tilde{O}\left(\max\left\{\sqrt{\frac{nL_2\Theta_1}{\varepsilon}}, \frac{\sigma^2\Theta_1}{\varepsilon^2}\right\}\right)$	$\tilde{O}\left(\max\left\{\sqrt{\frac{n^2L_2\Theta_2}{\varepsilon}}, \frac{\sigma^2\Theta_2 n}{\varepsilon^2}\right\}\right)$
RDD	$\tilde{O}\left(\max\left\{\frac{L_2\Theta_1}{\varepsilon}, \frac{\sigma^2\Theta_1}{\varepsilon^2}\right\}\right)$	$\tilde{O}\left(\max\left\{\frac{nL_2\Theta_2}{\varepsilon}, \frac{n\sigma^2\Theta_2}{\varepsilon^2}\right\}\right)$

Table 1. ARDD and RDD complexities for $p=1$ and $p=2$

Derivative-Free Optimization

We assume that an optimization procedure, given a pair of points $(x,y)\in\mathbb{R}^{2n}$, can obtain a pair of noisy stochastic realizations $(\widetilde{f}(x,\xi),\widetilde{f}(y,\xi))$ of the objective value f, where

$$\begin{aligned}\widetilde{f}(x,\xi) &= F(x,\xi)+\Xi(x,\xi),\\ |\Xi(x,\xi)| &\leqslant \Delta,\ \forall x\in\mathbb{R}^n,\ \text{a.s. in } \xi,\end{aligned} \qquad (7)$$

and ξ is independently drawn from P. Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^m\frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e\\ &= \left(\left\langle g^m(x,\vec{\xi}_m),e\right\rangle + \frac{1}{m}\sum_{i=1}^m(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \qquad (8)$$

where $e\in RS_2(1)$, ξ_i, $i=1,\ldots,m$ are independent realizations of ξ, m is the *batch size*, t is some small positive parameter which we call *smoothing parameter*, $g^m(x,\vec{\xi}_m) := \frac{1}{m}\sum_{i=1}^m g(x,\xi_i)$, and

$$\begin{aligned}\zeta(x,\xi_i,e) &= \frac{F(x+te,\xi_i)-F(x,\xi_i)}{t} - \langle g(x,\xi_i),e\rangle,\quad i=1,\ldots,m\\ \eta(x,\xi_i,e) &= \frac{\Xi(x+te,\xi_i)-\Xi(x,\xi_i)}{t},\quad i=1,\ldots,m.\end{aligned}$$

By Lipschitz smoothness of $F(\cdot,\xi)$, we have $|\zeta(x,\xi,e)|\leqslant\frac{L(\xi)t}{2}$ for all $x\in\mathbb{R}^n$ and $e\in S_2(1)$. Hence, $\mathbb{E}_\xi(\zeta(x,\xi,e))^2\leqslant\frac{L_2^2t^2}{4} =: \Delta_\zeta$ for all $x\in\mathbb{R}^n$ and $e\in S_2(1)$. At the same time, from (7), we have that $|\eta(x,\xi,e)|\leqslant\frac{2\Delta}{t} =: \Delta_\eta$ for all $x\in\mathbb{R}^n$, $e\in S_2(1)$ and a.s. in ξ. So, we can recover results from [2] using this technique.

Bibliography

[1] Pavel Dvurechensky, Alexander Gasnikov, and Eduard Gorbunov.
An accelerated directional derivative method for smooth stochastic convex optimization.
arXiv:1804.02394, 2018.

[2] Pavel Dvurechensky, Alexander Gasnikov, and Eduard Gorbunov.
An accelerated method for derivative-free smooth stochastic convex optimization.
arXiv:1802.09022, 2018.