An Accelerated Method for Derivative-Free Smooth Stochastic Convex Optimization

Eduard Gorbunov[1,2], Pavel Dvurechensky[2,3], Alexander Gasnikov[1,2]

[1]Moscow Institute of Physics and Technology, Moscow, Russia
[2]Institute for Information Transmission Problems of RAS, Moscow, Russia
[3]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

Introduction

Consider the following **unconstrained** 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 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, which we refer to as *oracle call*. Here ξ is independently drawn from P and

$$\widetilde{f}(x,\xi) = F(x,\xi)+\eta(x,\xi),\quad |\eta(x,\xi)|\leqslant\Delta,\ \forall x\in\mathbb{R}^n, \text{ a.s. in } \xi$$

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,

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

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

where $\rho_n = \min\{q-1, 16\ln n - 8\}n^{2/q-1}$, $n\geqslant 8$ and $s\in\mathbb{R}^n$. Here $q>0$ is such that $\frac{1}{p}+\frac{1}{q}=1$.

New Methods and Complexity Results

Algorithm 1. Accelerated Randomized Derivative-Free Directional Search (ARDFDS)

Input: x_0 — starting point, $N\geqslant 1$ — number of iterations, m — batch size, $t>0$ — smoothing parameter, $\{\alpha_k\}_{k=1}^N$ — stepsizes.

Output: point y_N

$y_0 \leftarrow x_0$, $z_0\leftarrow x_0$

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

$\tau_k\leftarrow\frac{2}{k+2}$

Generate $e_{k+1}\in RS_2(1)$ independently from previous iterations and ξ_i, $i=1,\ldots,m$ – independent realizations of ξ.

$x_{k+1}\leftarrow\tau_k z_k + (1-\tau_k)y_k$.

Calculate

$$\widetilde{\nabla}^m f^t(x_{k+1}) = \frac{1}{m}\sum_{i=1}^m\frac{(\widetilde{f}(x_{k+1}+te_{k+1},\xi_{k,i})-\widetilde{f}(x,\xi_{k,i}))e_{k+1}}{t}.$$

$y_{k+1}\leftarrow x_{k+1}-\frac{1}{2L_2}\widetilde{\nabla}^m f^t(x_{k+1})$.

$$z_{k+1}\leftarrow\underset{z\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\alpha_{k+1}n\left\langle\widetilde{\nabla}^m f^t(x_{k+1}), z-z_k\right\rangle + V[z_k](z)\right\}.$$

end for

Theorem 1. Let ARDFDS be applied to solve problem (1) with $\alpha_k = {}^{(k+1)}/_{(96n^2\rho_nL_2)}$. If we set $\Theta_p = V[z_0](x^*)$ which is defined by the chosen proximal setup, then $\forall n\geq 8$

$$\mathbb{E}[f(y_N)]-f(x^*)\leqslant\frac{384n^2\rho_nL_2\Theta_p}{N^2}+\frac{384N}{nL_2}\frac{\sigma^2}{m}+\frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{L_2t}{2}+\frac{2\Delta}{t}\right)+\frac{6N}{L_2}\left(L_2^2t^2+\frac{16\Delta^2}{t^2}\right)+\frac{N^2}{24n\rho_nL_2}\left(L_2^2t^2+\frac{16\Delta^2}{t^2}\right).$$

Feel free to contact us: eduard.gorbunov@phystech.edu, eduardgorbunov.github.io, pavel.dvurechensky@wias-berlin.de, gasnikov@yandex.ru.

Algorithm 2. Randomized Derivative-Free Directional Search (RDFDS).

Input: x_0 — starting point, $N\geqslant 1$ — number of iterations, m — batch size, $t>0$ — smoothing parameter, α — stepsize.

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

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

Generate $e_{k+1}\in RS_2(1)$ independently from previous iterations and ξ_i, $i=1,\ldots,m$ – independent realizations of ξ.

Calculate

$$\widetilde{\nabla}^m f^t(x_{k+1}) = \frac{1}{m}\sum_{i=1}^m\frac{(\widetilde{f}(x_{k+1}+te_{k+1},\xi_{k,i})-\widetilde{f}(x,\xi_{k,i}))e_{k+1}}{t}.$$

$$x_{k+1}\leftarrow\underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\alpha n\left\langle\widetilde{\nabla}^m f^t(x_k), x-x_k\right\rangle+V[x_k](x)\right\}.$$

end for

Theorem 2. Let RDFDS with $\alpha=\frac{1}{48n\rho_nL_2}$ be applied to solve problem (1) and $\Theta_p = V[x_0](x^*)$. Then $\forall n\geq 8$

$$\mathbb{E}[f(\bar{x}_N)]-f(x_*)\leqslant\frac{384n\rho_nL_2\Theta_p}{N}+\frac{2\sigma^2}{L_2m}+\left(\frac{n}{6L_2}+\frac{N}{3L_2\rho_n}\right)\left(\frac{L_2^2t^2}{2}+\frac{8\Delta^2}{t^2}\right)+\frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{L_2t}{2}+\frac{2\Delta}{t}\right).$$

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

Table 1. ARDFDS and RDFDS complexities for $p=1$ and $p=2$

Method	$p=1$	$p=2$
ARDFDS	$\tilde{O}\left(\min\left\{\frac{\varepsilon^{3/2}}{\sqrt{L_2\Theta_1}n}, \frac{\varepsilon^2}{nL_2\Theta_1}\right\}\right)$	$\tilde{O}\left(\min\left\{\frac{\varepsilon^{3/2}}{n\sqrt{L_2\Theta_2}}, \frac{\varepsilon^2}{nL_2\Theta_2}\right\}\right)$
RDFDS	$\tilde{O}\left(\min\left\{\frac{\varepsilon}{n}, \frac{\varepsilon^2}{nL_2\Theta_1}\right\}\right)$	$\tilde{O}\left(\min\left\{\frac{\varepsilon}{n}, \frac{\varepsilon^2}{nL_2\Theta_2}\right\}\right)$

Table 2. The allowable noise level Δ for ARDFDS and RDFDS.

Method	Assumptions	Oracle complexity, $\widetilde{O}(\cdot)$	$p=1$	σ^2	δ
MD Duchi et al. (2015), Gasnikov et al. (2016), Shamir (2017)	bound. gr.	$\frac{n^{2/q}M_2^2R_p^2}{\varepsilon^2}$	✓	✓	✓
RSPGF Ghadimi & Lan (2013,2016)	bound. var.	$\max\left\{\frac{nL_2R_2^2}{\varepsilon}, \frac{n\sigma^2R_2^2}{\varepsilon^2}\right\}$	×	✓	×
RS Nesterov & Spokoiny (2017), Bogolubsky et al. (2017)		$\frac{nL_2R_2^2}{\varepsilon}$	×	×	✓
RDFDS [This paper]	bound. var.	$\max\left\{\frac{n^{2/q}L_2R_p^2}{\varepsilon}, \frac{n^{2/q}\sigma^2R_p^2}{\varepsilon^2}\right\}$	✓	✓	✓
AccRS Nesterov & Spokoiny (2017), Dvurechensky et al. (2017)		$n\sqrt{\frac{L_2R_2^2}{\varepsilon}}$	×	×	✓
ARDFDS [This paper]	bound. var.	$\max\left\{n^{1/2+1/q}\sqrt{\frac{L_2R_p^2}{\varepsilon}}, \frac{n^{2/q}\sigma^2R_p^2}{\varepsilon^2}\right\}$	✓	✓	✓

Table 3. Comparison of oracle complexity (total number of oracle calls) of different methods with two point feedback for convex optimization problems. In the column "Assumptions" we use "bound. gr." to define $\mathbb{E}_\xi\left[\|g(x,\xi)\|_2^2\right]\leq M_2^2$ and "bound. var." to define $\mathbb{E}\|g(x,\xi)-\nabla f(x)\|_2^2\leqslant\sigma^2$. Column $p=1$ corresponds to the support of non-Euclidean setup, column "σ^2" to the support of stochastic optimization methods, "δ" corresponds to the support of additional noise of an unknown nature.