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, Russia
[2]Weierstrass Institute for Applied Analysis and Stochastics, Germany

6 July, 2018

The Problem

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

The Problem

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

1. $f(x)$ — convex function

The Problem

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

1. $f(x)$ — convex function
2. $F(x,\xi)$ — closed function of x P-almost surely in ξ

The Problem

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

1. $f(x)$ — convex function
2. $F(x,\xi)$ — closed function of x P-almost surely in ξ
3. For P almost every ξ, the function $F(x,\xi)$ has gradient $g(x,\xi)$, which is $L(\xi)$-Lipschitz continuous with respect to the Euclidean norm

The Problem

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

1. $f(x)$ — convex function
2. $F(x,\xi)$ — closed function of x P-almost surely in ξ
3. For P almost every ξ, the function $F(x,\xi)$ has gradient $g(x,\xi)$, which is $L(\xi)$-Lipschitz continuous with respect to the Euclidean norm

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

The Problem

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

1. $f(x)$ — convex function
2. $F(x,\xi)$ — closed function of x P-almost surely in ξ
3. For P almost every ξ, the function $F(x,\xi)$ has gradient $g(x,\xi)$, which is $L(\xi)$-Lipschitz continuous with respect to the Euclidean norm

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

4. $L_2 := \sqrt{\mathbb{E}_\xi[L(\xi)^2]} < +\infty$

The Problem

Under this assumptions

1. $\mathbb{E}_{\xi}[g(x,\xi)] = \nabla f(x)$
2. $\|\nabla f(x) - \nabla f(y)\|_2 \leqslant L_2\|x-y\|_2,\ \forall x,y \in \mathbb{R}^n$

The Problem

Under this assumptions

1. $\mathbb{E}_{\xi}[g(x,\xi)] = \nabla f(x)$
2. $\|\nabla f(x) - \nabla f(y)\|_2 \leqslant L_2\|x-y\|_2,\ \forall x,y \in \mathbb{R}^n$

Also we assume that

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

The Problem

Finally, we assume that we have the following oracle.

The Problem

Finally, we assume that we have the following oracle.

1. Oracle:
   $x \in \mathbb{R}^n, e \in S_2(1) \to \widetilde{f}'(x, \xi, e) = \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e)$

The Problem

Finally, we assume that we have the following oracle.

1. Oracle:
   $x \in \mathbb{R}^n, e \in S_2(1) \to \widetilde{f}'(x, \xi, e) = \langle g(x, \xi), e\rangle + \zeta(x, \xi, e) + \eta(x, \xi, e)$
2. $\mathbb{E}_{\xi}\left[\zeta(x, \xi, e)^2\right] \leqslant \Delta_{\zeta},\ \forall x \in \mathbb{R}^n, \forall e \in S_2(1)$

The Problem

Finally, we assume that we have the following oracle.

1. Oracle:
   $x \in \mathbb{R}^n, e \in S_2(1) \to \widetilde{f}'(x, \xi, e) = \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e)$
2. $\mathbb{E}_\xi \left[\zeta(x, \xi, e)^2 \right] \leqslant \Delta_\zeta, \; \forall x \in \mathbb{R}^n, \forall e \in S_2(1)$
3. $|\eta(x, \xi, e)| \leqslant \Delta_\eta, \; \forall x \in \mathbb{R}^n, \forall e \in S_2(1)$

The Problem

Finally, we assume that we have the following oracle.

1. Oracle:
   $x \in \mathbb{R}^n, e \in S_2(1) \to \widetilde{f}'(x,\xi,e) = \langle g(x,\xi), e\rangle + \zeta(x,\xi,e) + \eta(x,\xi,e)$
2. $\mathbb{E}_{\xi}\left[\zeta(x,\xi,e)^2\right] \leqslant \Delta_\zeta,\ \forall x \in \mathbb{R}^n, \forall e \in S_2(1)$
3. $|\eta(x,\xi,e)| \leqslant \Delta_\eta,\ \forall x \in \mathbb{R}^n, \forall e \in S_2(1)$

Further we will use random vector from uniform distribution over the Euclidean sphere in $\mathbb{R}^n$ as e and denote it $e \in RS_2^n(1)$.

Preliminaries

1. *Prox-function:* differentiable 1-strongly convex w.r.t. l_p-norm (where $1 \leqslant p \leqslant 2$) function $d : \mathbb{R}^n \to \mathbb{R}$.

Preliminaries

1. *Prox-function:* differentiable 1-strongly convex w.r.t. l_p-norm (where $1 \leqslant p \leqslant 2$) function $d : \mathbb{R}^n \to \mathbb{R}$.
2. *Bregman divergence* w.r.t. d is a function of two arguments:

$$V[z](x) \overset{\text{def}}{=} d(x) - d(z) - \langle \nabla d(z), x - z \rangle. \tag{3}$$

Preliminaries

1. *Prox-function:* differentiable 1-strongly convex w.r.t. l_p-norm (where $1 \leqslant p \leqslant 2$) function $d : \mathbb{R}^n \to \mathbb{R}$.
2. *Bregman divergence* w.r.t. d is a function of two arguments:

$$V[z](x) \overset{\text{def}}{=} d(x) - d(z) - \langle \nabla d(z), x - z \rangle. \qquad (3)$$

Note that from strong convexity of d follows

$$V[z](x) \geqslant \frac{1}{2}\|x - z\|_p^2, \quad x, z \in \mathbb{R}^n.$$

Key lemma

In our proofs of complexity bounds, we rely on the following lemma.

Key lemma

In our proofs of complexity bounds, we rely on the following lemma.

Lemma

Let $e \in RS_2(1)$, *i.e be a random vector uniformly distributed on the surface of the unit Euclidean sphere in* $\mathbb{R}^n$,

Key lemma

In our proofs of complexity bounds, we rely on the following lemma.

Lemma

Let $e \in RS_2(1)$*, i.e be a random vector uniformly distributed on the surface of the unit Euclidean sphere in* $\mathbb{R}^n$*,* $p \in [1,2]$ *and* q *be given by* $\frac{1}{p} + \frac{1}{q} = 1$*.*

Key lemma

In our proofs of complexity bounds, we rely on the following lemma.

Lemma

Let $e \in RS_2(1)$*, i.e be a random vector uniformly distributed on the surface of the unit Euclidean sphere in* $\mathbb{R}^n$*,* $p \in [1, 2]$ *and* q *be given by* $\frac{1}{p} + \frac{1}{q} = 1$*. Then, for* $n \geqslant 8$ *and* $\rho_n = \min\{q - 1, 16 \ln n - 8\} n^{\frac{2}{q} - 1}$,

Key lemma

In our proofs of complexity bounds, we rely on the following lemma.

Lemma

Let $e \in RS_2(1)$, i.e be a random vector uniformly distributed on the surface of the unit Euclidean sphere in $\mathbb{R}^n$, $p \in [1, 2]$ and q be given by $\frac{1}{p} + \frac{1}{q} = 1$. Then, for $n \geqslant 8$ and $\rho_n = \min\{q - 1, 16 \ln n - 8\} n^{\frac{2}{q} - 1}$,

$$\mathbb{E}_e \|e\|_q^2 \leq \rho_n, \tag{4}$$

Key lemma

In our proofs of complexity bounds, we rely on the following lemma.

Lemma

Let $e \in RS_2(1)$, i.e be a random vector uniformly distributed on the surface of the unit Euclidean sphere in $\mathbb{R}^n$, $p \in [1,2]$ and q be given by $\frac{1}{p} + \frac{1}{q} = 1$. Then, for $n \geqslant 8$ and $\rho_n = \min\{q-1, 16\ln n - 8\} n^{\frac{2}{q}-1}$,

$$\mathbb{E}_e \|e\|_q^2 \leq \rho_n, \tag{4}$$

$$\mathbb{E}_e \left(\langle s, e \rangle^2 \|e\|_q^2 \right) \leq \frac{6\rho_n}{n} \|s\|_2^2, \quad \forall s \in \mathbb{R}^n. \tag{5}$$

Key lemma: intuition

The last inequality for $q = \infty$ could be rewritten (without loss of generality assume that $\|s\|_2 = 1$) as follows:

$$\mathbb{E}_e \left[\langle s, e \rangle^2 \|e\|_\infty^2 \right] \lesssim \frac{1}{n} \cdot \frac{\ln n}{n} \quad \forall s \in S_2(1).$$

Key lemma: intuition

The last inequality for $q = \infty$ could be rewritten (without loss of generality assume that $\|s\|_2 = 1$) as follows:

$$\mathbb{E}_e \left[\langle s, e\rangle^2 \|e\|_\infty^2\right] \lesssim \frac{1}{n} \cdot \frac{\ln n}{n} \quad \forall s \in S_2(1).$$

It could be obtained using phenomenon of concentration of measure. It turns out (см. A. Blum, J. Hopcroft, R. Kannan, *Foundations of Data Science*; K. Ball, *An elementary introduction to modern convex geometry*; V. A. Zorich, *Mathematical analysis in natural science problems*) that with probability $\geqslant 1 - \frac{2}{c} e^{-\frac{c^2}{2}}$ the following inequality holds $|\langle l, e\rangle| \leqslant \frac{c}{\sqrt{n-1}}$, where l — some arbitrary fixed vector.

Key lemma: intuition

The last inequality for $q = \infty$ could be rewritten (without loss of generality assume that $\|s\|_2 = 1$) as follows:

$$\mathbb{E}_e\left[\langle s, e\rangle^2 \|e\|_\infty^2\right] \lesssim \frac{1}{n} \cdot \frac{\ln n}{n} \quad \forall s \in S_2(1).$$

It could be obtained using phenomenon of concentration of measure. It turns out (см. A. Blum, J. Hopcroft, R. Kannan, *Foundations of Data Science*; K. Ball, *An elementary introduction to modern convex geometry*; V. A. Zorich, *Mathematical analysis in natural science problems*) that with probability $\geqslant 1 - \frac{2}{c} e^{-\frac{c^2}{2}}$ the following inequality holds $|\langle l, e\rangle| \leqslant \frac{c}{\sqrt{n-1}}$, where l — some arbitrary fixed vector. Putting $c = 10$ and $l = s$ we get that with *big* probability $\langle s, e\rangle^2 \leqslant \frac{100}{n}$; and putting $c = 2\sqrt{\ln n}$ and vectors l directed along coordinate axis one can obtain that with probability $\geqslant 1 - \frac{1}{n\sqrt{n}}$ the following inequality holds $\|e\|_\infty^2 \leqslant \frac{4\ln n}{n}$.

Accelerated Randomized Directional Derivative Method

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**
  $\alpha_{k+1} \leftarrow \frac{k+2}{96n^2\rho_n L_2}$, $\tau_k \leftarrow \frac{1}{48\alpha_{k+1}n^2\rho_n L_2} = \frac{2}{k+2}$.
  Generate $e_{k+1} \in RS_2(1)$ independently from previous iterations and ξ_i, $i = 1, \ldots, m$ – independent realizations of ξ.
  $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k)y_k$.
  Calculate

$$\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m}\sum_{i=1}^{m} \widetilde{f}'(x_{k+1}, \xi_i, e_{k+1})e_{k+1}.$$

  $y_{k+1} \leftarrow x_{k+1} - \frac{1}{2L_2}\widetilde{\nabla}^m f(x_{k+1})$.
  $z_{k+1} \leftarrow \operatorname*{argmin}_{z\in\mathbb{R}^n}\left\{\alpha_{k+1}n\left\langle\widetilde{\nabla}^m f(x_{k+1}),\, z - z_k\right\rangle + V[z_k](z)\right\}$.
end for
return y_N

Complexity of ARDD

Theorem

Let ARDD method be applied to solve problem (1)*.*

Complexity of ARDD

Theorem

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

$$\begin{aligned}\mathbb{E}[f(y_N)] - f(x^*) \leqslant & \frac{384\Theta_p n^2 \rho_n L_2}{N^2} + \frac{4N}{nL_2}\cdot\frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2 \\ & + \frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) \\ & + \frac{N^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2,\end{aligned} \tag{6}$$

Complexity of ARDD

Theorem

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

$$\mathbb{E}[f(y_N)] - f(x^*) \leqslant \frac{384\Theta_p n^2 \rho_n L_2}{N^2} + \frac{4N}{nL_2}\cdot\frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2, \tag{6}$$

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

Complexity of ARDD

	$p = 1$	$p = 2$
N	$O\left(\sqrt{\frac{n\ln nL_2\Theta_1}{\varepsilon}}\right)$	$O\left(\sqrt{\frac{n^2L_2\Theta_2}{\varepsilon}}\right)$
m	$O\left(\max\left\{1, \sqrt{\frac{\ln n}{n}}\cdot\frac{\sigma^2}{\varepsilon^{3/2}}\cdot\sqrt{\frac{\Theta_1}{L_2}}\right\}\right)$	$O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon^{3/2}}\cdot\sqrt{\frac{\Theta_2}{L_2}}\right\}\right)$
Δ_ζ	$O\left(\min\left\{n(\ln n)^2L_2^2\Theta_1, \frac{\varepsilon^2}{n\Theta_1}, \frac{\varepsilon^{\frac{3}{2}}}{\sqrt{n\ln n}}\cdot\sqrt{\frac{L_2}{\Theta_1}}\right\}\right)$	$O\left(\min\left\{n^3L_2^2\Theta_2, \frac{\varepsilon}{n\Theta_2}, \frac{\varepsilon^{\frac{3}{2}}}{n}\cdot\sqrt{\frac{L_2}{\Theta_2}}\right\}\right)$
Δ_η	$O\left(\min\left\{\sqrt{n}\ln nL_2\sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n\Theta_1}}, \frac{\varepsilon^{\frac{3}{4}}}{\sqrt[4]{n\ln n}}\cdot\sqrt[4]{\frac{L_2}{\Theta_1}}\right\}\right)$	$O\left(\min\left\{n^{\frac{3}{2}}L_2\sqrt{\Theta_2}, \frac{\varepsilon}{\sqrt{n\Theta_2}}, \frac{\varepsilon^{\frac{3}{4}}}{\sqrt{n}}\cdot\sqrt[4]{\frac{L_2}{\Theta_2}}\right\}\right)$
O-le calls	$O\left(\max\left\{\sqrt{\frac{n\ln nL_2\Theta_1}{\varepsilon}}, \frac{\sigma^2\Theta_1\ln n}{\varepsilon^2}\right\}\right)$	$O\left(\max\left\{\sqrt{\frac{n^2L_2\Theta_2}{\varepsilon}}, \frac{\sigma^2\Theta_2 n}{\varepsilon^2}\right\}\right)$

Table: ARDD parameters for the cases $p = 1$ and $p = 2$.

Randomized Directional Derivative Method

Algorithm 2 Randomized Directional Derivative (RDD) method

Input: x_0 — starting point; $N \geqslant 1$ — number of iterations; m — batch size.
Output: point $\bar{x}_N$.

1: **for** $k = 0, \ldots, N-1$ **do**
2: $\alpha \leftarrow \frac{1}{48 n \rho_n L_2}$.
3: Generate $e_{k+1} \in RS_2(1)$ independently from previous iterations and ξ_i, $i = 1, \ldots, m$ – independent realizations of ξ.
4: $x_{k+1} \leftarrow \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \alpha n \left\langle \widetilde{\nabla}^m f(x_k), x - x_k \right\rangle + V[x_k](x) \right\}$.
5: Calculate

$$\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m} \sum_{i=1}^{m} \widetilde{f}'(x_{k+1}, \xi_i, e_{k+1}) e_{k+1}.$$

6: **end for**
7: **return** $\bar{x}_N \leftarrow \frac{1}{N} \sum\limits_{k=0}^{N-1} x_k$

Complexity of RDD

Theorem

Let RDD method be applied to solve problem (1).

Complexity of RDD

Theorem

Let RDD method be applied to solve problem (1). Then

$$\begin{aligned}\mathbb{E}[f(\bar{x}_N)] - f(x_*) \quad &\leqslant \frac{384n\rho_n L_2\Theta_p}{N} + \frac{2}{L_2}\frac{\sigma^2}{m} + \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 \\ &+ \frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) \\ &+ \frac{N}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2,\end{aligned} \tag{7}$$

Complexity of RDD

Theorem

Let RDD method be applied to solve problem (1). Then

$$
\begin{aligned}
\mathbb{E}[f(\bar{x}_N)] - f(x_*) \quad &\leqslant \frac{384 n\rho_n L_2 \Theta_p}{N} + \frac{2}{L_2}\frac{\sigma^2}{m} + \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 \\
&+ \frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) \\
&+ \frac{N}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2,
\end{aligned} \tag{7}
$$

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

Complexity of RDD

	$p = 1$	$p = 2$
N	$O\left(\frac{L_2\Theta_1 \ln n}{\varepsilon}\right)$	$O\left(\frac{nL_2\Theta_2}{\varepsilon}\right)$
m	$O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}\right)$	$O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}\right)$
Δ_ζ	$O\left(\min\left\{\frac{(\ln n)^2}{n}L_2^2\Theta_1, \frac{\varepsilon^2}{n\Theta_1}, \frac{\varepsilon L_2}{n}\right\}\right)$	$O\left(\min\left\{nL_2^2\Theta_2, \frac{\varepsilon^2}{n\Theta_2}, \frac{\varepsilon L_2}{n}\right\}\right)$
Δ_η	$O\left(\min\left\{\frac{\ln n}{\sqrt{n}}L_2\sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n\Theta_1}}, \sqrt{\frac{\varepsilon L_2}{n}}\right\}\right)$	$O\left(\min\left\{\sqrt{n}L_2\sqrt{\Theta_2}, \frac{\varepsilon}{\sqrt{n\Theta_2}}, \sqrt{\frac{\varepsilon L_2}{n}}\right\}\right)$
O-le calls	$O\left(\max\left\{\frac{L_2\Theta_1 \ln n}{\varepsilon}, \frac{\sigma^2\Theta_1 \ln n}{\varepsilon^2}\right\}\right)$	$O\left(\max\left\{\frac{nL_2\Theta_2}{\varepsilon}, \frac{n\sigma^2\Theta_2}{\varepsilon^2}\right\}\right)$

Table: RDD parameters for the cases $p = 1$ and $p = 2$.

ARDD and RDD

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: ARDD and RDD complexities for $p = 1$ and $p = 2$

ARDD and RDD

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: ARDD and RDD complexities for $p = 1$ and $p = 2$

Remark

Note that for $p = 1$ RDD gives dimensional independent complexity bounds.

Strongly convex case

We will additionally assume two facts.

1. Function $f(x)$ is μ_p-strongly convex $w.r.t.$ l_p-norm.

Strongly convex case

We will additionally assume two facts.

1. Function $f(x)$ is μ_p-strongly convex $w.r.t.$ l_p-norm.
2. There is such a constant Ω_p for our choice of prox-function $d(\cdot)$ that

Strongly convex case

We will additionally assume two facts.

1. Function $f(x)$ is μ_p-strongly convex *w.r.t.* l_p-norm.
2. There is such a constant Ω_p for our choice of prox-function $d(\cdot)$ that

$$x - \text{ such random point that } \mathbb{E}_x[\|x - x_*\|_p^2] \leqslant R_p^2$$

Strongly convex case

We will additionally assume two facts.

1. Function $f(x)$ is μ_p-strongly convex *w.r.t.* l_p-norm.
2. There is such a constant Ω_p for our choice of prox-function $d(\cdot)$ that

$$
\begin{gathered}
x - \text{ such random point that } \mathbb{E}_x[\|x - x_*\|_p^2] \leqslant R_p^2 \\
\Longrightarrow \mathbb{E}_x d\left(\frac{x-x_*}{R_p}\right) \leqslant \frac{\Omega_p}{2}
\end{gathered}
\tag{8}
$$

ARDD method for strongly convex functions (ARDDsc)

Algorithm 3 Accelerated Randomized Directional Derivative method for strongly convex functions (ARDDsc)

Input: x_0 — starting point s.t. $\|x_0 - x_*\|_p^2 \leq R_p^2$; $K \geqslant 1$ — number of iterations; μ_p – strong convexity parameter.

Output: point u_K.

1: Set $N_0 = \left\lceil \sqrt{\frac{8aL_2\Omega_p}{\mu_p}} \right\rceil$, where $a = 384n^2\rho_n$

2: **for** $k = 0, \ldots, K-1$ **do**

3: Set

$$m_k := \max\left\{1, \left\lceil \frac{8b\sigma^2 N_0 2^k}{L_2\mu_p R_p^2} \right\rceil\right\}, \quad R_k^2 := R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right), \text{ where } b = \frac{4}{n} \tag{9}$$

4: Set $d_k(x) = R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.

5: Run ARDD with starting point u_k and prox-function $d_k(x)$ for N_0 steps with batch size m_k.

6: Set $u_{k+1} = y_{N_0}$, $k = k+1$.

7: **end for**

8: **return** u_K

Complexity of ARDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and ARDDsc method be applied to solve this problem.

Complexity of ARDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and ARDDsc method be applied to solve this problem. Then

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta, \tag{10}$$

Complexity of ARDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and ARDDsc method be applied to solve this problem. Then

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta, \tag{10}$$

where $\Delta =$

$$\frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}{\Delta_\eta}^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2.$$

Complexity of ARDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and ARDDsc method be applied to solve this problem. Then

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta, \tag{10}$$

where $\Delta =$
$\frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}\Delta_\eta{}^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2.$
Moreover, under an appropriate choice of Δ_ζ and Δ_η s.t. $2\Delta \leqslant \varepsilon/2$, the oracle complexity to achieve ε-accuracy of the solution is

Complexity of ARDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and ARDDsc method be applied to solve this problem. Then

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta, \tag{10}$$

where $\Delta =$
$\frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}\Delta_\eta{}^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2.$
Moreover, under an appropriate choice of Δ_ζ and Δ_η s.t. $2\Delta \leqslant \varepsilon/2$, the oracle complexity to achieve ε-accuracy of the solution is

$$\widetilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2\Omega_p}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

Complexity of ARDDsc

	$p=1$
Δ_ζ	$O\left(\min\left\{\varepsilon\sqrt{\frac{L_2\mu_1}{n\ln n\Omega_1}},\ \varepsilon^2\frac{nL_2^2\Omega_1}{R_1^2\mu_1^2},\ \varepsilon\cdot\frac{\mu_1}{n\Omega_1}\right\}\right)$
Δ_η	$O\left(\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_1}{n\ln n\Omega_1}},\ \varepsilon\frac{\sqrt{n}\ln nL_2\sqrt{\Omega_1}}{R_1\mu_1},\ \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_1}{n\Omega_1}}\right\}\right)$
O-le calls	$\widetilde{O}\left(\max\left\{n^{\frac{1}{2}}\sqrt{\frac{L_2\Omega_1}{\mu_1}}\log_2\frac{\mu_1R_1^2}{\varepsilon},\ \frac{\sigma^2\Omega_1}{\mu_1\varepsilon}\right\}\right)$
	$p=2$
Δ_ζ	$O\left(\min\left\{\varepsilon\sqrt{\frac{L_2\mu_2}{n^2\Omega_2}},\ \varepsilon^2\frac{rn^3L_2^2\Omega_2}{R_2^2\mu_2^2},\ \varepsilon\cdot\frac{\mu_2}{n\Omega_2}\right\}\right)$
Δ_η	$O\left(\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_2}{n^2\Omega_2}},\ \varepsilon\frac{\sqrt{n^3}L_2\sqrt{\Omega_2}}{R_2\mu_2},\ \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_2}{n\Omega_2}}\right\}\right)$
O-le calls	$\widetilde{O}\left(\max\left\{n\sqrt{\frac{L_2\Omega_2}{\mu_2}}\log_2\frac{\mu_2R_2^2}{\varepsilon},\ \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}\right)$

Table: Algorithm 3 parameters for the cases $p=1$ and $p=2$.

RDD for strongly convex functions

Algorithm 4 Randomized Directional Derivative method for strongly convex functions (RDDsc)

Input: x_0 — starting point s.t. $\|x_0 - x_*\|_p^2 \leq R_p^2$; $K \geqslant 1$ — number of iterations; μ_p – strong convexity parameter.

Output: point u_K.

1: Set $N_0 = \left\lceil \frac{8aL_2\Omega_p}{\mu_p} \right\rceil$, where $a = 384n\rho_n$.

2: **for** $k = 0, \ldots, K-1$ **do**

3: Set

$$m_k := \max\left\{1, \left\lceil \frac{8b\sigma^2 2^k}{L_2\mu_p R_p^2} \right\rceil\right\}, \quad R_k^2 := R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right), \text{ where } b = 2 \qquad (11)$$

4: Set $d_k(x) = R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.

5: Run RDD with starting point u_k and prox-function $d_k(x)$ for N_0 steps with batch size m_k.

6: Set $u_{k+1} = y_{N_0}$, $k = k+1$.

7: **end for**

8: **return** u_K

Complexity of RDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and RDDsc method be applied to solve this problem.

Complexity of RDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and RDDsc method be applied to solve this problem. Then

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta, \tag{12}$$

Complexity of RDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and RDDsc method be applied to solve this problem. Then

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta, \tag{12}$$

where

$$\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta{}^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2.$$

Complexity of RDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and RDDsc method be applied to solve this problem. Then

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta, \tag{12}$$

where

$$\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}{\Delta_\eta}^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2.$$

Moreover, under an appropriate choice of Δ_ζ and Δ_η s.t. $2\Delta \leqslant \varepsilon/2$, the oracle complexity to achieve ε-accuracy of the solution is

Complexity of RDDsc

Theorem

Let f in problem (1) be μ_p-strongly convex and RDDsc method be applied to solve this problem. Then

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta, \tag{12}$$

where

$$\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta{}^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2.$$

Moreover, under an appropriate choice of Δ_ζ and Δ_η s.t. $2\Delta \leqslant \varepsilon/2$, the oracle complexity to achieve ε-accuracy of the solution is

$$\widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2\Omega_p}{\mu_p}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

Complexity of RDDsc

	$p = 1$
Δ_ζ	$O\left(\min\left\{\frac{\varepsilon L_2}{n}, \varepsilon^2\frac{(\ln n)^2 L_2^2}{nR_1^2\mu_1^2}, \varepsilon\frac{\mu_1}{n\Omega_1}\right\}\right)$
Δ_η	$O\left(\min\left\{\sqrt{\frac{\varepsilon L_2}{n}}, \varepsilon\frac{\ln n L_2}{\sqrt{n}R_1\mu_1}, \sqrt{\varepsilon\frac{\mu_1}{n\Omega_1}}\right\}\right)$
O-le calls	$\widetilde{O}\left(\max\left\{\frac{L_2\Omega_1}{\mu_1}\log_2\frac{\mu_1 R_1^2}{\varepsilon}, \frac{\sigma^2\Omega_1}{\mu_1\varepsilon}\right\}\right)$
	$p = 2$
Δ_ζ	$O\left(\min\left\{\frac{\varepsilon L_2}{n}, \varepsilon^2\frac{nL_2^2}{R_2^2\mu_2^2}, \varepsilon\frac{\mu_2}{n\Omega_2}\right\}\right)$
Δ_η	$O\left(\min\left\{\sqrt{\frac{\varepsilon L_2}{n}}, \varepsilon\frac{\sqrt{n}L_2}{R_2\mu_2}, \sqrt{\varepsilon\frac{\mu_2}{n\Omega_2}}\right\}\right)$
O-le calls	$\widetilde{O}\left(\max\left\{\frac{nL_2\Omega_2}{\mu_2}\log_2\frac{\mu_2 R_2^2}{\varepsilon}, \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}\right)$

Table: Algorithm 4 parameters for the cases $p = 1$ and $p = 2$.

Derivative-Free Optimization

Consider the following zeroth-order oracle.

Derivative-Free Optimization

Consider the following zeroth-order oracle.

1. Oracle: $(x, y) \to (\widetilde{f}(x, \xi), \widetilde{f}(y, \xi))$, where

$$\widetilde{f}(x, \xi) = F(x, \xi) + \Xi(x, \xi)$$

Derivative-Free Optimization

Consider the following zeroth-order oracle.

1. Oracle: $(x, y) \to (\widetilde{f}(x, \xi), \widetilde{f}(y, \xi))$, where

$$\widetilde{f}(x, \xi) = F(x, \xi) + \Xi(x, \xi)$$

2. $|\Xi(x, \xi)| \leqslant \Delta, \ \forall x \in \mathbb{R}^n$, a.s. in ξ

Derivative-Free Optimization

Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

Derivative-Free Optimization

Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^m \frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\ &= \left(\left\langle g^m(x,\vec{\xi_m}),e\right\rangle + \frac{1}{m}\sum_{i=1}^m(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{13}$$

Derivative-Free Optimization

Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\begin{aligned}
\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^{m} \frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t} e \\
&= \left(\left\langle g^m(x,\vec{\xi_m}), e\right\rangle + \frac{1}{m}\sum_{i=1}^{m}(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right) e,
\end{aligned} \tag{13}$$

1. $g^m(x,\vec{\xi_m}) := \frac{1}{m}\sum_{i=1}^{m} g(x,\xi_i)$

Derivative-Free Optimization

Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^m \frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e\\ &= \left(\left\langle g^m(x,\vec{\xi_m}),e\right\rangle + \frac{1}{m}\sum_{i=1}^m(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{13}$$

1. $g^m(x,\vec{\xi_m}) := \frac{1}{m}\sum_{i=1}^m g(x,\xi_i)$
2. $\zeta(x,\xi_i,e) = \frac{F(x+te,\xi_i)-F(x,\xi_i)}{t} - \langle g(x,\xi_i),e\rangle, \quad i=1,\dots,m$

Derivative-Free Optimization

Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^m \frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\ &= \left(\left\langle g^m(x,\vec{\xi_m}),e\right\rangle + \frac{1}{m}\sum_{i=1}^m(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{13}$$

1. $g^m(x,\vec{\xi_m}) := \frac{1}{m}\sum_{i=1}^m g(x,\xi_i)$
2. $\zeta(x,\xi_i,e) = \frac{F(x+te,\xi_i)-F(x,\xi_i)}{t} - \langle g(x,\xi_i),e\rangle, \quad i=1,\ldots,m$
3. $\eta(x,\xi_i,e) = \frac{\Xi(x+te,\xi_i)-\Xi(x,\xi_i)}{t}, \quad i=1,\ldots,m$

Derivative-Free Optimization

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

Derivative-Free Optimization

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

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 $e \in S_2(1)$.

Derivative-Free Optimization

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

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,

$$\mathbb{E}_\xi\left[\zeta(x,\xi,e)^2\right] \leqslant \frac{L_2^2t^2}{4} =: \Delta_\zeta, \quad \forall x\in\mathbb{R}^n, e\in S_2(1).$$

Derivative-Free Optimization

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

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,

$$\mathbb{E}_\xi\left[\zeta(x,\xi,e)^2\right] \leqslant \frac{L_2^2t^2}{4} =: \Delta_\zeta, \quad \forall x\in\mathbb{R}^n, e\in S_2(1).$$

At the same time, from $|\Xi(x,\xi)|\leqslant\Delta$, we have that

$$|\eta(x,\xi,e)| \leqslant \frac{2\Delta}{t} =: \Delta_\eta, \quad \forall x\in\mathbb{R}^n, e\in S_2(1), \text{ a.s. in } \xi$$

Thank you for your attention!
Questions?