An Accelerated Directional Derivative Method for Smooth Stochastic Convex Optimization

Eduard Gorbunov

Moscow Institute of Physics and Technology

13 June, 2018
The Problem

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

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

where $\xi$ is a random vector with probability distribution $P(\xi)$, $\xi \in \mathcal{X}$,
The Problem

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

(1)

where \( \xi \) is a random vector with probability distribution \( P(\xi), \xi \in \mathcal{X} \), and for \( P \)-almost every \( \xi \in \mathcal{X} \), the function \( F(x, \xi) \) is closed
The Problem

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

where $\xi$ is a random vector with probability distribution $P(\xi), \xi \in \mathcal{X}$, and for $P$-almost every $\xi \in \mathcal{X}$, the function $F(x, \xi)$ is closed and $f$ is convex.
The Problem

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

where \( \xi \) is a random vector with probability distribution \( P(\xi), \xi \in \mathcal{X} \), and for \( P \)-almost every \( \xi \in \mathcal{X} \), the function \( F(x, \xi) \) is closed and \( f \) is convex. Moreover, we assume that, for \( P \) almost every \( \xi \), the function \( F(x, \xi) \) has gradient \( g(x, \xi) \), which is \( L(\xi) \)-Lipschitz continuous with respect to the Euclidean norm

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

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

Under this assumptions, \( \mathbb{E}_\xi [g(x, \xi)] = \nabla f(x) \) and

\[
\|\nabla f(x) - \nabla f(y)\|_2 \leq L_2 \|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n.
\]
The Problem

Under this assumptions, \( \mathbb{E}_\xi [g(x, \xi)] = \nabla f(x) \) and

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

Also we assume that

\[
\mathbb{E}_\xi \left[ \| g(x, \xi) - \nabla f(x) \|_2^2 \right] \leq \sigma^2. \quad (2)
\]
Finally, we assume that an optimization procedure, given a point $x \in \mathbb{R}^n$, direction $e \in S_2^1$, and $\zeta$ independently drawn from $P$, can obtain a noisy stochastic approximation $\tilde{f}'(x, \zeta, e)$ for the directional derivative $\langle g(x, \zeta), e \rangle$:

$$\tilde{f}'(x, \zeta, e) = \langle g(x, \zeta), e \rangle + \zeta(x, \zeta, e) + \eta(x, \zeta, e),$$

where $S_2^1$ is the Euclidean sphere or radius one with the center at the point zero and the values $\Delta \zeta, \Delta \eta$ are controlled and can be made as small as desired. Note that we use the smoothness of $F(\cdot, \zeta)$ to write the directional derivative as $\langle g(x, \zeta), e \rangle$, but we do not assume that the whole stochastic gradient $g(x, \zeta)$ is available.
The Problem

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

\[
\tilde{f}'(x, \xi, e) = \langle g(x, \xi), e \rangle + \eta(x, \xi, e) + \xi,
\]

\( E[\xi \eta(x, \xi, e)^2] \leq \Delta \xi, \quad \forall x \in \mathbb{R}^n, \quad \forall e \in S_2(1), \) a.s. in \( \xi \), (3)

where \( S_2(1) \) is the Euclidean sphere or radius one with the center at the point zero and the values \( \Delta \xi, \Delta \eta \) are controlled and can be made as small as it is desired.

Note that we use the smoothness of \( F(\cdot, \xi) \) to write the directional derivative as \( \langle g(x, \xi), e \rangle \), but we do not assume that the whole stochastic gradient \( g(x, \xi) \) is available.
Finally, we assume that an optimization procedure, given a point \( x \in \mathbb{R}^n \), direction \( e \in S_2(1) \) and \( \xi \) independently drawn from \( P \),
Finally, we assume that an optimization procedure, given a point \( x \in \mathbb{R}^n \), direction \( e \in S_2(1) \) and \( \xi \) independently drawn from \( P \), can obtain a noisy stochastic approximation \( \tilde{f}'(x, \xi, e) \) for the directional derivative \( \langle g(x, \xi), e \rangle \):

\[
\tilde{f}'(x, \xi, e) = \langle g(x, \xi), e \rangle + \varphi(x, \xi, e) + \eta(x, \xi, e),
\]

\[
E[\varphi(x, \xi, e)^2] \leq \Delta \varphi, \quad \forall x \in \mathbb{R}^n, \quad \forall e \in S_2(1),
\]

\[
|\eta(x, \xi, e)| \leq \Delta \eta, \quad \forall x \in \mathbb{R}^n, \quad \forall e \in S_2(1), \quad \text{a.s. in } \xi.
\]

Note that we use the smoothness of \( F(\cdot, \xi) \) to write the directional derivative as \( \langle g(x, \xi), e \rangle \), but we do not assume that the whole stochastic gradient \( g(x, \xi) \) is available.
Finally, we assume that an optimization procedure, given a point $x \in \mathbb{R}^n$, direction $e \in S_2(1)$ and $\xi$ independently drawn from $P$, can obtain a noisy stochastic approximation $\tilde{f}'(x, \xi, e)$ for the directional derivative $\langle g(x, \xi), e \rangle$:

$$\tilde{f}'(x, \xi, e) = \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e),$$

$$\mathbb{E}_\xi [\zeta(x, \xi, e)^2] \leq \Delta_\zeta, \quad \forall x \in \mathbb{R}^n, \forall e \in S_2(1),$$

$$|\eta(x, \xi, e)| \leq \Delta_\eta, \quad \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \text{ a.s. in } \xi,$$  \hspace{1cm} (3)
Finally, we assume that an optimization procedure, given a point \( x \in \mathbb{R}^n \), direction \( e \in S_2(1) \) and \( \xi \) independently drawn from \( P \), can obtain a noisy stochastic approximation \( \tilde{f}'(x, \xi, e) \) for the directional derivative \( \langle g(x, \xi), e \rangle \):

\[
\tilde{f}'(x, \xi, e) = \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e),
\]

\[
\mathbb{E}_{\xi} [\zeta(x, \xi, e)^2] \leq \Delta_{\zeta}, \quad \forall x \in \mathbb{R}^n, \forall e \in S_2(1),
\]

\[
|\eta(x, \xi, e)| \leq \Delta_{\eta}, \quad \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \text{ a.s. in } \xi,
\]

where \( S_2(1) \) is the Euclidean sphere or radius one with the center at the point zero.
Finally, we assume that an optimization procedure, given a point \( x \in \mathbb{R}^n \), direction \( e \in S_2(1) \) and \( \xi \) independently drawn from \( P \), can obtain a noisy stochastic approximation \( \tilde{f}'(x, \xi, e) \) for the directional derivative \( \langle g(x, \xi), e \rangle \):

\[
\tilde{f}'(x, \xi, e) = \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e),
\]

\[
\mathbb{E}_{\xi} \left[ \zeta(x, \xi, e)^2 \right] \leq \Delta_\zeta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \quad (3)
\]

\[
|\eta(x, \xi, e)| \leq \Delta_\eta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \text{ a.s. in } \xi,
\]

where \( S_2(1) \) is the Euclidean sphere or radius one with the center at the point zero and the values \( \Delta_\zeta, \Delta_\eta \) are controlled and can be made as small as it is desired.
Finally, we assume that an optimization procedure, given a point \( x \in \mathbb{R}^n \), direction \( e \in S_2(1) \) and \( \xi \) independently drawn from \( P \), can obtain a noisy stochastic approximation \( \tilde{f}'(x, \xi, e) \) for the directional derivative \( \langle g(x, \xi), e \rangle \):

\[
\tilde{f}'(x, \xi, e) = \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e),
\]

\[
\mathbb{E}_\xi [\zeta(x, \xi, e)^2] \leq \Delta_\zeta, \quad \forall x \in \mathbb{R}^n, \forall e \in S_2(1),
\]

\[
|\eta(x, \xi, e)| \leq \Delta_\eta, \quad \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \text{ a.s. in } \xi, \tag{3}
\]

where \( S_2(1) \) is the Euclidean sphere or radius one with the center at the point zero and the values \( \Delta_\zeta, \Delta_\eta \) are controlled and can be made as small as it is desired. Note that we use the smoothness of \( F(\cdot, \xi) \) to write the directional derivative as \( \langle g(x, \xi), e \rangle \), but we do not assume that the whole stochastic gradient \( g(x, \xi) \) is available.
Preliminaries

We choose a *prox-function* $d(x)$ which is continuous, convex on $\mathbb{R}^n$.
Preliminaries

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 $\| \cdot \|_p$,
Preliminaries

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 $\| \cdot \|_p$, where $\| \cdot \|_p$ is a vector $l_p$-norm with $p \in [1, 2]$. 

For the case $p = 1$, we choose the following prox-function $d(x) = e^{n(\frac{\alpha - 1}{\alpha})\left(2 - \frac{1}{\alpha}\right)}\frac{\|x\|_2^2}{\alpha}$, $\alpha = 1 + \frac{1}{\ln n}$ (5)

and, for the case $p = 2$, we choose the prox-function to be the squared Euclidean norm $d(x) = \frac{1}{2}\|x\|_2^2$ (6).
Preliminaries

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 $\| \cdot \|_p$, where $\| \cdot \|_p$ is a vector $l_p$-norm with $p \in [1, 2]$. We define also the corresponding *Bregman divergence* $V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$, $x, z \in \mathbb{R}^n$. 

Eduard Gorbunov (MIPT) 13 June, 2018 5 / 16
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 \( \| \cdot \|_p \), where \( \| \cdot \|_p \) is a vector \( l_p \)-norm with \( p \in [1, 2] \). We define also the corresponding *Bregman divergence* \( V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle \), \( x, z \in \mathbb{R}^n \). Note that, by the strong convexity of \( d \),

\[
V[z](x) \geq \frac{1}{2} \| x - z \|_p^2, \quad x, z \in \mathbb{R}^n. \tag{4}
\]
Preliminaries

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 $\| \cdot \|_p$, where $\| \cdot \|_p$ is a vector $l_p$-norm with $p \in [1, 2]$. We define also the corresponding *Bregman divergence* $V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$, $x, z \in \mathbb{R}^n$. Note that, by the strong convexity of $d$,

$$V[z](x) \geq \frac{1}{2} \| x - z \|_p^2, \quad x, z \in \mathbb{R}^n. \quad (4)$$

For the case $p = 1$, we choose the following prox-function

$$d(x) = \frac{e n^{(\kappa - 1)(2 - \kappa) / \kappa} \ln n}{2} \| x \|_\kappa^2, \quad \kappa = 1 + \frac{1}{\ln n}. \quad (5)$$
Preliminaries

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 \( \| \cdot \|_p \), where \( \| \cdot \|_p \) is a vector \( l_p \)-norm with \( p \in [1, 2] \). We define also the corresponding *Bregman divergence* \( V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle \), \( x, z \in \mathbb{R}^n \). Note that, by the strong convexity of \( d \),

\[
V[z](x) \geq \frac{1}{2} \| x - z \|_p^2, \quad x, z \in \mathbb{R}^n. \tag{4}
\]

For the case \( p = 1 \), we choose the following prox-function

\[
d(x) = \frac{e^{n(\kappa - 1)(2 - \kappa) / \kappa} \ln n}{2 \| x \|_\kappa^2}, \quad \kappa = 1 + \frac{1}{\ln n} \tag{5}
\]

and, for the case \( p = 2 \), we choose the prox-function to be the squared Euclidean norm

\[
d(x) = \frac{1}{2} \| x \|_2^2. \tag{6}
\]
Key lemma

In our proofs of complexity bounds, we rely on the following 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 $\mathbb{R}^n$, 

\[ E_e \|e\|_2^q \leq \rho_n, \]  

(7)  

\[ E_e (\langle s, e \rangle^2 \|e\|_2^q) \leq 6 \rho_n n \|s\|_2^2, \quad \forall s \in \mathbb{R}^n. \]  

(8)
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
\[
\frac{1}{p} + \frac{1}{q} = 1.
\]
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 $n \geq 8$ 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 $n \geq 8$ and $\rho_n = \min\{q - 1, 16 \ln n - 8\} n^{\frac{2}{q} - 1}$,

$$\mathbb{E}_e \|e\|_q^2 \leq \rho_n,$$  \hspace{1cm} (7)
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 \( n \geq 8 \) and \( \rho_n = \min\{q - 1, 16 \ln n - 8\} n^{\frac{q-1}{q}} \),

\[
\mathbb{E} \left\| e \right\|_q^2 \leq \rho_n, \tag{7}
\]

\[
\mathbb{E} \left( \langle s, e \rangle^2 \left\| e \right\|_q^2 \right) \leq \frac{6 \rho_n}{n} \left\| s \right\|_2^2, \quad \forall s \in \mathbb{R}^n. \tag{8}
\]
Algorithm 1 Accelerated Randomized Directional Derivative (ARDD) method

Input: \( x_0 \) — starting point; \( N \geq 1 \) — number of iterations; \( m \) — batch size.

Output: point \( y_N \)

1. \( y_0 \leftarrow x_0, z_0 \leftarrow x_0 \)
2. for \( k = 0, \ldots, N - 1 \) do
3. \( \alpha_{k+1} \leftarrow \frac{k+2}{96n^2\rho_nL_2}, \tau_k \leftarrow \frac{1}{48\alpha_{k+1}n^2\rho_nL_2} = \frac{2}{k+2} \).
4. Generate \( e_{k+1} \in RS_2(1) \) independently from previous iterations and \( \xi_i, i = 1, \ldots, m \) — independent realizations of \( \xi \).
5. Calculate
\[
\nabla^m f(x_{k+1}) = \frac{1}{m} \sum_{i=1}^{m} \tilde{f}'(x_{k+1}, \xi_i, e_{k+1})e_{k+1}.
\]
6. \( x_{k+1} \leftarrow \tau_k z_k + (1 - \tau_k)y_k. \)
7. \( y_{k+1} \leftarrow x_{k+1} - \frac{1}{2L_2} \nabla^m f(x_{k+1}). \)
8. \( z_{k+1} \leftarrow \arg\min_{z \in \mathbb{R}^n} \left\{ \alpha_{k+1}n \langle \nabla^m f(x_{k+1}), z - z_k \rangle + V[z_k](z) \right\}. \)
9. end for
10. return \( y_N \)
Theorem

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

Theorem

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

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

(9)
Complexity of ARDD

Theorem

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

\[
\mathbb{E}[f(y_N)] - f(x^*) \leq \frac{384 \Theta_p n^2 \rho_n L_2}{N^2} + \frac{4N}{NL_2} \cdot \frac{\sigma^2}{m} + \frac{61N}{24L_2} \Delta_\zeta + \frac{122N}{3L_2} \Delta_\eta^2
\]
\[
+ \frac{12 \sqrt{2n \Theta_p}}{N^2} \left( \frac{\sqrt{\Delta_\zeta}}{2} + 2 \Delta_\eta \right)
\]
\[
+ \frac{N^2}{12n \rho_n L_2} \left( \frac{\sqrt{\Delta_\zeta}}{2} + 2 \Delta_\eta \right)^2,
\]

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

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

<table>
<thead>
<tr>
<th></th>
<th>$p = 1$</th>
<th>$p = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$O\left(\sqrt{\frac{n \ln n L_2 \Theta_1}{\varepsilon}}\right)$</td>
<td>$O\left(\sqrt{\frac{n^2 L_2 \Theta_2}{\varepsilon}}\right)$</td>
</tr>
<tr>
<td>$m$</td>
<td>$O\left(\max\left{1, \sqrt{\frac{\ln n}{n}} \cdot \frac{\sigma^2}{\varepsilon^{3/2}} \cdot \sqrt{\frac{\Theta_1}{L_2}}\right}\right)$</td>
<td>$O\left(\max\left{1, \frac{\sigma^2}{\varepsilon^{3/2}} \cdot \sqrt{\frac{\Theta_2}{L_2}}\right}\right)$</td>
</tr>
<tr>
<td>$\Delta_\zeta$</td>
<td>$O\left(\min\left{n(\ln n)^2 L_2 \Theta_1, \frac{\varepsilon}{n \Theta_1}, \frac{3 \sigma^2}{\varepsilon^{3/2} \ln n} \cdot \sqrt{\frac{L_2}{\Theta_1}}\right}\right)$</td>
<td>$O\left(\min\left{n^3 L_2^2 \Theta_2, \frac{\varepsilon}{n \Theta_2}, \frac{3 \sigma^2}{n \varepsilon^{3/2}} \cdot \sqrt{\frac{L_2}{\Theta_2}}\right}\right)$</td>
</tr>
<tr>
<td>$\Delta_\eta$</td>
<td>$O\left(\min\left{\sqrt{n \ln n L_2} \sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n \Theta_1}}, \frac{3 \sigma^2}{4 \varepsilon^{3/2} \ln n} \cdot 4 \sqrt{\frac{L_2}{\Theta_1}}\right}\right)$</td>
<td>$O\left(\min\left{n^3 L_2^2 \Theta_2, \frac{\varepsilon}{\sqrt{n \Theta_2}}, \frac{3 \sigma^2}{4 \varepsilon^{3/2} \sqrt{n}} \cdot 4 \sqrt{\frac{L_2}{\Theta_2}}\right}\right)$</td>
</tr>
<tr>
<td>O-le calls</td>
<td>$O\left(\max\left{\sqrt{\frac{n \ln n L_2 \Theta_1}{\varepsilon}}, \frac{\sigma^2 \Theta_1 \ln n}{\varepsilon^2}\right}\right)$</td>
<td>$O\left(\max\left{\sqrt{\frac{n^2 L_2 \Theta_2}{\varepsilon}}, \frac{\sigma^2 \Theta_2 n}{\varepsilon^2}\right}\right)$</td>
</tr>
</tbody>
</table>
**Algorithm 2** Randomized Directional Derivative (RDD) method

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

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

1. for $k = 0, \ldots, N - 1$ do
2. \hspace{1em} $\alpha \leftarrow \frac{1}{48n\rho_nL_2}$.
3. \hspace{1em} Generate $e_{k+1} \in RS_2(1)$ independently from previous iterations and $\xi_i, i = 1, \ldots, m$ — independent realizations of $\xi$.
4. \hspace{1em} Calculate
   \[
   \tilde{\nabla}^m f(x_{k+1}) = \frac{1}{m} \sum_{i=1}^{m} \tilde{f}'(x_{k+1}, \xi_i, e_{k+1})e_{k+1}.
   \]
5. \hspace{1em} $x_{k+1} \leftarrow \text{argmin}_{x \in \mathbb{R}^n} \left\{ \alpha n \left\langle \tilde{\nabla}^m f(x_k), x - x_k \right\rangle + V[x_k](x) \right\}$.
6. end for
7. return $\bar{x}_N \leftarrow \frac{1}{N} \sum_{k=0}^{N-1} x_k$.
Theorem

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

Theorem

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

\[ E[f(\bar{x}_N)] - f(x_*) \leq \frac{384n\rho_nL_2\Theta p}{N} + \frac{2}{L_2} \frac{\sigma^2}{m} + \frac{n}{12L_2} \Delta_\zeta + \frac{4n}{3L_2} \Delta^2_\eta 
+ \frac{8\sqrt{2n\Theta p}}{N} \left( \frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta \right)^2
+ \frac{N}{3L_2\rho_n} \left( \frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta \right)^2, \]  

(10)
Theorem

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

\[
\mathbb{E}[f(\bar{x}_N)] - f(x_*) \leq \frac{384n\rho_n L_2 \Theta_p}{N} + \frac{2}{L_2} \frac{\sigma^2}{m} + \frac{n}{12L_2} \Delta_\zeta + \frac{4n}{3L_2} \Delta_\eta^2 \\
+ \frac{8\sqrt{2n\Theta_p}}{N} \left( \frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta \right) \\
+ \frac{N}{3L_2 \rho_n} \left( \frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta \right)^2,
\]

where \( \Theta_p = V[z_0](x^*) \) is defined by the chosen proximal setup and \( \mathbb{E}[\cdot] = \mathbb{E}_{e_1,...,e_N,\xi_1,1,...,\xi_N,m}[\cdot] \).
Complexity of RDD

<table>
<thead>
<tr>
<th></th>
<th>$p = 1$</th>
<th>$p = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$O\left(\frac{L_2\Theta_1\ln n}{\epsilon}\right)$</td>
<td>$O\left(\frac{nL_2\Theta_2}{\epsilon}\right)$</td>
</tr>
<tr>
<td>$m$</td>
<td>$O\left(\max\left{1, \frac{\sigma^2}{\epsilon L_2}\right}\right)$</td>
<td>$O\left(\max\left{1, \frac{\sigma^2}{\epsilon L_2}\right}\right)$</td>
</tr>
<tr>
<td>$\Delta_\varsigma$</td>
<td>$O\left(\min\left{\frac{\left(\ln n\right)^2}{n} L_2^2 \Theta_1, \frac{\epsilon^2}{n \Theta_1}, \frac{\epsilon L_2}{n}\right}\right)$</td>
<td>$O\left(\min\left{nL_2^2\Theta_2, \frac{\epsilon^2}{n \Theta_2}, \frac{\epsilon L_2}{n}\right}\right)$</td>
</tr>
<tr>
<td>$\Delta_\eta$</td>
<td>$O\left(\min\left{\frac{\ln n}{\sqrt{n}} L_2 \sqrt{\Theta_1}, \frac{\epsilon}{\sqrt{n \Theta_1}}, \sqrt{\frac{\epsilon L_2}{n}}\right}\right)$</td>
<td>$O\left(\min\left{\sqrt{nL_2} \sqrt{\Theta_2}, \frac{\epsilon}{\sqrt{n \Theta_2}}, \sqrt{\frac{\epsilon L_2}{n}}\right}\right)$</td>
</tr>
<tr>
<td>O-le calls</td>
<td>$O\left(\max\left{\frac{L_2 \Theta_1 \ln n}{\epsilon}, \frac{\sigma^2 \Theta_1 \ln n}{\epsilon^2}\right}\right)$</td>
<td>$O\left(\max\left{\frac{nL_2 \Theta_2}{\epsilon}, \frac{n\sigma^2 \Theta_2}{\epsilon^2}\right}\right)$</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>$p = 1$</th>
<th>$p = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARDD</td>
<td>$\tilde{O} \left( \max \left{ \sqrt{\frac{nL_2 \Theta_1}{\varepsilon}}, \frac{\sigma^2 \Theta_1}{\varepsilon^2} \right} \right)$</td>
<td>$\tilde{O} \left( \max \left{ \sqrt{\frac{n^2 L_2 \Theta_2}{\varepsilon}}, \frac{\sigma^2 \Theta_2}{\varepsilon^2} n \right} \right)$</td>
</tr>
<tr>
<td>RDD</td>
<td>$\tilde{O} \left( \max \left{ \frac{L_2 \Theta_1}{\varepsilon}, \frac{\sigma^2 \Theta_1}{\varepsilon^2} \right} \right)$</td>
<td>$\tilde{O} \left( \max \left{ \frac{nL_2 \Theta_2}{\varepsilon}, \frac{n\sigma^2 \Theta_2}{\varepsilon^2} \right} \right)$</td>
</tr>
</tbody>
</table>

**Table:** ARDD and RDD complexities for $p = 1$ and $p = 2
### ARDD and RDD

<table>
<thead>
<tr>
<th>Method</th>
<th>$p = 1$</th>
<th>$p = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARDD</td>
<td>$\tilde{O} \left( \max \left{ \frac{n L_2 \Theta_1}{\varepsilon}, \frac{\sigma^2 \Theta_1}{\varepsilon^2} \right} \right)$</td>
<td>$\tilde{O} \left( \max \left{ \sqrt{\frac{n^2 L_2 \Theta_2}{\varepsilon}}, \frac{\sigma^2 \Theta_2 n}{\varepsilon^2} \right} \right)$</td>
</tr>
<tr>
<td>RDD</td>
<td>$\tilde{O} \left( \max \left{ \frac{L_2 \Theta_1}{\varepsilon}, \frac{\sigma^2 \Theta_1}{\varepsilon^2} \right} \right)$</td>
<td>$\tilde{O} \left( \max \left{ \frac{n L_2 \Theta_2}{\varepsilon}, \frac{n \sigma^2 \Theta_2}{\varepsilon^2} \right} \right)$</td>
</tr>
</tbody>
</table>

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

### Remark

*Note that for $p = 1$ RDD gives *dimensional independent* complexity bounds.*
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 \((\tilde{f}(x, \xi), \tilde{f}(y, \xi))\) of the objective value \(f\), where

\[
\tilde{f}(x, \xi) = F(x, \xi) + \Xi(x, \xi), \quad |\Xi(x, \xi)| \leq \Delta, \; \forall x \in \mathbb{R}^n, \text{ a.s. in } \xi, \quad (11)
\]

and \(\xi\) is independently drawn from \(P\).
Derivative-Free Optimization

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

$$\tilde{\nabla f}(x) = \frac{1}{m} \sum_{i=1}^{m} \tilde{f}(x + t e, \zeta_i) - \tilde{f}(x, \zeta_i)$$

$$= \left( \langle g_m(x, \vec{\zeta}_m), e \rangle + \frac{1}{m} \sum_{i=1}^{m} \zeta(x, \zeta_i, e) + \eta(x, \zeta_i, e) \right) e,$$

where $e \in \mathbb{R}S^2(1)$, $\zeta_i, i = 1, ..., m$ are independent realizations of $\zeta$, $m$ is the batch size, $t$ is some small positive parameter which we call the smoothing parameter, $g_m(x, \vec{\zeta}_m) := \frac{1}{m} \sum_{i=1}^{m} g(x, \zeta_i)$, and $\zeta(x, \zeta_i, e) = F(x + t e, \zeta_i) - F(x, \zeta_i) - \langle g(x, \zeta_i), e \rangle,$ $i = 1, ..., m$. $\eta(x, \zeta_i, e) = \Xi(x + t e, \zeta_i) - \Xi(x, \zeta_i)$.
Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$
\widetilde{\nabla}^m f^t(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{\widetilde{f}(x+te_i, \xi_i) - \widetilde{f}(x, \xi_i)}{t} e
$$

$$
= \left( \left\langle g^m(x, \xi_m), e \right\rangle + \frac{1}{m} \sum_{i=1}^{m} (\zeta(x, \xi_i, e) + \eta(x, \xi_i, e)) \right) e,
$$

(12)
Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\tilde{\nabla}^m f^t(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{f}(x+te, \xi_i) - \tilde{f}(x, \xi_i)}{t} e$$

$$= \left( \langle g^m(x, \xi_m), e \rangle + \frac{1}{m} \sum_{i=1}^{m} (\zeta(x, \xi_i, e) + \eta(x, \xi_i, e)) \right) e,$$

(12)

where $e \in RS_2(1)$, 

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

\[
\tilde{\nabla}^m f^t(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{f}(x + te, \xi_i) - \tilde{f}(x, \xi_i)}{t} e \\
= \left( \left\langle g^m(x, \bar{\xi}_m), e \right\rangle + \frac{1}{m} \sum_{i=1}^{m} (\zeta(x, \xi_i, e) + \eta(x, \xi_i, e)) \right) e,
\]

where \( e \in RS_2(1), \xi_i, i = 1, ..., m \) are independent realizations of \( \xi \),

(12)
Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

\[
\tilde{\nabla}^m f^t(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{f}(x+te_i, \xi_i) - \tilde{f}(x, \xi_i)}{t} e
\]

\[
= \left( \left\langle g^m(x, \xi_m), e \right\rangle + \frac{1}{m} \sum_{i=1}^{m} (\zeta(x, \xi_i, e) + \eta(x, \xi_i, e)) \right) e,
\]

where $e \in RS_2(1)$, $\xi_i$, $i = 1, \ldots, m$ are independent realizations of $\xi$, $m$ is the batch size,
Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\tilde{\nabla}^m f^t(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{f}(x+te, \xi_i) - \tilde{f}(x, \xi_i)}{t} e$$

$$= \left( \langle g^m(x, \vec{\xi}_m), e \rangle + \frac{1}{m} \sum_{i=1}^{m} (\zeta(x, \xi_i, e) + \eta(x, \xi_i, e)) \right) e,$$

where $e \in RS_2(1)$, $\xi_i, i = 1, \ldots, m$ are independent realizations of $\xi$, $m$ is the batch size, $t$ is some small positive parameter which we call smoothing parameter,
Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$
\hat{\nabla}^m f^t(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{f}(x+te,\xi_i) - \tilde{f}(x,\xi_i)}{t} e
$$

$$
= \left( \left< g^m(x, \xi_m), e \right> + \frac{1}{m} \sum_{i=1}^{m} (\zeta(x, \xi_i, e) + \eta(x, \xi_i, e)) \right) e,
$$

where $e \in RS_2(1)$, $\xi_i$, $i = 1, \ldots, m$ are independent realizations of $\xi$, $m$ is the batch size, $t$ is some small positive parameter which we call smoothing parameter, $g^m(x, \xi_m) := \frac{1}{m} \sum_{i=1}^{m} g(x, \xi_i)$.
Based on these observations of the objective value, we form the following stochastic approximation of $\nabla f(x)$

$$\tilde{\nabla}^m f^t(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{f}(x+te,\xi_i)-\tilde{f}(x,\xi_i)}{t} e$$

$$= \left( \left\langle g^m(x, \xi_m), e \right\rangle + \frac{1}{m} \sum_{i=1}^{m} (\zeta(x, \xi_i, e) + \eta(x, \xi_i, e)) \right) e,$$

(12)

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

$$\zeta(x, \xi_i, e) = \frac{F(x+te,\xi_i)-F(x,\xi_i)}{t} - \left\langle g(x, \xi_i), e \right\rangle, \quad i = 1, \ldots, m$$
Derivative-Free Optimization

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

$$\tilde{\nabla}^m f^t(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{f}(x+te,\xi_i) - \tilde{f}(x,\xi_i)}{t} e$$

$$= \left( \langle g^m(x, \xi_m), e \rangle + \frac{1}{m} \sum_{i=1}^{m} (\zeta(x, \xi_i, e) + \eta(x, \xi_i, e)) \right) e,$$

where $e \in \mathbb{R}S_2(1)$, $\xi_i$, $i = 1, ..., m$ are independent realizations of $\xi$, $m$ is the batch size, $t$ is some small positive parameter which we call smoothing parameter, $g^m(x, \xi_m) := \frac{1}{m} \sum_{i=1}^{m} g(x, \xi_i)$, and

$$\zeta(x, \xi_i, e) = \frac{F(x+te,\xi_i) - F(x,\xi_i)}{t} - \langle g(x, \xi_i), e \rangle, \quad i = 1, ..., m$$

$$\eta(x, \xi_i, e) = \frac{\Xi(x+te,\xi_i) - \Xi(x,\xi_i)}{t}, \quad i = 1, ..., m.$$
Derivative-Free Optimization

\[ \zeta(x, \xi_i, e) = \frac{F(x+te, \xi_i) - F(x, \xi_i)}{t} - \langle g(x, \xi_i), e \rangle, \quad i = 1, \ldots, m \]

\[ \eta(x, \xi_i, e) = \frac{\Xi(x+te, \xi_i) - \Xi(x, \xi_i)}{t}, \quad i = 1, \ldots, m. \]
\begin{align*}
\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{align*}

By Lipschitz smoothness of $F(\cdot, \xi)$, we have $|\zeta(x, \xi, e)| \leq \frac{L(\xi)t}{2}$ for all $x \in \mathbb{R}^n$ and $e \in S_2(1)$. 
\[ \zeta(x, \xi_i, e) = \frac{F(x + te, \xi_i) - F(x, \xi_i)}{t} - \langle g(x, \xi_i), e \rangle, \quad i = 1, \ldots, m \]

\[ \eta(x, \xi_i, e) = \frac{\Xi(x + te, \xi_i) - \Xi(x, \xi_i)}{t}, \quad i = 1, \ldots, m. \]

By Lipschitz smoothness of \( F(\cdot, \xi) \), we have \( |\zeta(x, \xi, e)| \leq \frac{L(\xi)t}{2} \) for all \( x \in \mathbb{R}^n \) and \( e \in S_2(1) \). Hence, \( \mathbb{E}_\xi(\zeta(x, \xi, e))^2 \leq \frac{L^2 t^2}{4} =: \Delta_\zeta \) for all \( x \in \mathbb{R}^n \) and \( e \in S_2(1) \).
\[
\begin{align*}
\zeta(x, \xi_i, e) & = \frac{F(x+te, \xi_i) - F(x, \xi_i)}{t} - \langle g(x, \xi_i), e \rangle, \quad i = 1, \ldots, m \\
\eta(x, \xi_i, e) & = \frac{\Xi(x+te, \xi_i) - \Xi(x, \xi_i)}{t}, \quad i = 1, \ldots, m.
\end{align*}
\]

By Lipschitz smoothness of \(F(\cdot, \xi)\), we have \(|\zeta(x, \xi, e)| \leq \frac{L(\xi)t}{2}\) for all \(x \in \mathbb{R}^n\) and \(e \in S_2(1)\). Hence, \(\mathbb{E}_\xi(\zeta(x, \xi, e))^2 \leq \frac{L^2 t^2}{4} =: \Delta_\zeta\) for all \(x \in \mathbb{R}^n\) and \(e \in S_2(1)\). At the same time, from (11), we have that \(|\eta(x, \xi, e)| \leq \frac{2\Delta}{t} =: \Delta_\eta\) for all \(x \in \mathbb{R}^n\), \(e \in S_2(1)\) and a.s. in \(\xi\).
Derivative-Free Optimization

\[ \zeta(x, \xi_i, e) = \frac{F(x+te, \xi_i) - F(x, \xi_i)}{t} - \langle g(x, \xi_i), e \rangle, \quad i = 1, \ldots, m \]

\[ \eta(x, \xi_i, e) = \frac{\Xi(x+te, \xi_i) - \Xi(x, \xi_i)}{t}, \quad i = 1, \ldots, m. \]

By Lipschitz smoothness of \( F(\cdot, \xi) \), we have \( |\zeta(x, \xi, e)| \leq \frac{L(\xi) t}{2} \) for all \( x \in \mathbb{R}^n \) and \( e \in S_2(1) \). Hence, \( \mathbb{E}_\xi (\zeta(x, \xi, e))^2 \leq \frac{L^2 t^2}{4} =: \Delta_\zeta \) for all \( x \in \mathbb{R}^n \) and \( e \in S_2(1) \). At the same time, from (11), we have that \( |\eta(x, \xi, e)| \leq \frac{2\Delta}{t} =: \Delta_\eta \) for all \( x \in \mathbb{R}^n, e \in S_2(1) \) and a.s. in \( \xi \).

So, we can use the same methods and analyze such problems in the same way.