# An Accelerated Directional Derivative Method for Smooth Stochastic Convex Optimization

Eduard Gorbunov

Moscow Institute of Physics and Technology

13 June, 2018

# The Problem

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

# The Problem

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

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

# The Problem

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

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

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

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

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

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

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

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

# The Problem

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

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

## The Problem

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

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

# The Problem

Finally, we assume that an optimization procedure, given a point $x \in \mathbb{R}^n$,

# The Problem

Finally, we assume that an optimization procedure, given a point $x \in \mathbb{R}^n$, direction $e \in S_2(1)$

# The Problem

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 $P$,

# The Problem

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 $\widetilde{f}'(x, \xi, e)$ for the directional derivative $\langle g(x, \xi), e \rangle$:

# The Problem

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 $\widetilde{f}'(x, \xi, e)$ for the directional derivative $\langle g(x, \xi), e \rangle$:

$$
\begin{aligned}
\widetilde{f}'(x, \xi, e) &= \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e), \\
\mathbb{E}_\xi \left[ \zeta(x, \xi, e)^2 \right] &\leqslant \Delta_\zeta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \\
|\eta(x, \xi, e)| &\leqslant \Delta_\eta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \text{ a.s. in } \xi,
\end{aligned} \tag{3}
$$

# The Problem

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 $\widetilde{f}'(x, \xi, e)$ for the directional derivative $\langle g(x, \xi), e\rangle$:

$$\begin{aligned} \widetilde{f}'(x,\xi,e) &= \langle g(x,\xi), e\rangle + \zeta(x,\xi,e) + \eta(x,\xi,e), \\ \mathbb{E}_\xi\left[\zeta(x,\xi,e)^2\right] &\leqslant \Delta_\zeta,\ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \\ |\eta(x,\xi,e)| &\leqslant \Delta_\eta,\ \forall x \in \mathbb{R}^n, \forall e \in S_2(1),\ \text{a.s. in } \xi, \end{aligned} \tag{3}$$

where $S_2(1)$ is the Euclidean sphere or radius one with the center at the point zero

# The Problem

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 $\widetilde{f}'(x, \xi, e)$ for the directional derivative $\langle g(x, \xi), e \rangle$:

$$\begin{aligned} \widetilde{f}'(x, \xi, e) &= \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e), \\ \mathbb{E}_\xi \left[ \zeta(x, \xi, e)^2 \right] &\leqslant \Delta_\zeta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \\ |\eta(x, \xi, e)| &\leqslant \Delta_\eta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \text{ a.s. in } \xi, \end{aligned} \tag{3}$$

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.

## The Problem

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 $\widetilde{f}'(x, \xi, e)$ for the directional derivative $\langle g(x, \xi), e \rangle$:

$$\begin{aligned} \widetilde{f}'(x, \xi, e) &= \langle g(x, \xi), e \rangle + \zeta(x, \xi, e) + \eta(x, \xi, e), \\ \mathbb{E}_\xi \left[ \zeta(x, \xi, e)^2 \right] &\leqslant \Delta_\zeta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \\ |\eta(x, \xi, e)| &\leqslant \Delta_\eta, \ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \text{ a.s. in } \xi, \end{aligned} \tag{3}$$

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 $p \in [1, 2]$.

# 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* $V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$, $x, z \in \mathbb{R}^n$.

## 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 $d$,

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

# 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

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

# 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

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

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.

## 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 $\rho_n = \min\{q-1, 16\ln n - 8\} n^{\frac{2}{q}-1}$,*

$$\mathbb{E}_e \|e\|_q^2 \le \rho_n, \tag{7}$$

# 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 $n \geqslant 8$ and $\rho_n = \min\{q-1, 16\ln n - 8\} n^{\frac{2}{q}-1}$,*

$$\mathbb{E}_e \|e\|_q^2 \leq \rho_n, \tag{7}$$

$$\mathbb{E}_e \left( \langle s, e \rangle^2 \|e\|_q^2 \right) \leq \frac{6\rho_n}{n} \|s\|_2^2, \quad \forall s \in \mathbb{R}^n. \tag{8}$$

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

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

1: $y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$
2: **for** $k = 0, \ldots, N-1$ **do**
3: $\alpha_{k+1} \leftarrow \frac{k+2}{96n^2\rho_n L_2}$, $\tau_k \leftarrow \frac{1}{48\alpha_{k+1}n^2\rho_n L_2} = \frac{2}{k+2}$.
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
$$\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m}\sum_{i=1}^{m} \widetilde{f}'(x_{k+1}, \xi_i, e_{k+1})e_{k+1}.$$
6: $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k)y_k$.
7: $y_{k+1} \leftarrow x_{k+1} - \frac{1}{2L_2}\widetilde{\nabla}^m f(x_{k+1})$.
8: $z_{k+1} \leftarrow \operatorname*{argmin}_{z\in\mathbb{R}^n}\left\{\alpha_{k+1}n\left\langle\widetilde{\nabla}^m f(x_{k+1}), z - z_k\right\rangle + V[z_k](z)\right\}$.
9: **end for**
10: **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*

$$\begin{aligned}\mathbb{E}[f(y_N)] - f(x^*) \quad &\leqslant \frac{384\Theta_p n^2 \rho_n L_2}{N^2} + \frac{4N}{nL_2}\cdot\frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2 \\ &+ \frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) \\ &+ \frac{N^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2,\end{aligned} \tag{9}$$

# Complexity of ARDD

### Theorem

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

$$\mathbb{E}[f(y_N)] - f(x^*) \leqslant \frac{384\Theta_p n^2 \rho_n L_2}{N^2} + \frac{4N}{nL_2} \cdot \frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2, \tag{9}$$

*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

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

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: $\alpha \leftarrow \frac{1}{48n\rho_n L_2}$.
3: Generate $e_{k+1} \in RS_2(1)$ independently from previous iterations and $\xi_i$, $i = 1, \ldots, m$ – independent realizations of $\xi$.
4: Calculate

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

5: $x_{k+1} \leftarrow \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \alpha n \left\langle \widetilde{\nabla}^m f(x_k), x - x_k \right\rangle + V[x_k](x) \right\}$.
6: **end for**
7: **return** $\bar{x}_N \leftarrow \frac{1}{N}\sum_{k=0}^{N-1} x_k$

# 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*

$$\begin{aligned}\mathbb{E}[f(\bar{x}_N)] - f(x_*) &\leqslant \frac{384 n\rho_n L_2\Theta_p}{N} + \frac{2}{L_2}\frac{\sigma^2}{m} + \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 \\ &+ \frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) \\ &+ \frac{N}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2,\end{aligned} \tag{10}$$

## Complexity of RDD

### Theorem

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

$$\begin{aligned}\mathbb{E}[f(\bar{x}_N)] - f(x_*) \quad \leqslant\; & \frac{384 n\rho_n L_2 \Theta_p}{N} + \frac{2}{L_2}\frac{\sigma^2}{m} + \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 \\ & + \frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) \\ & + \frac{N}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2,\end{aligned} \tag{10}$$

*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

| | $p=1$ | $p=2$ |
|---|---|---|
| $N$ | $O\left(\frac{L_2\Theta_1 \ln n}{\varepsilon}\right)$ | $O\left(\frac{nL_2\Theta_2}{\varepsilon}\right)$ |
| $m$ | $O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}\right)$ | $O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}\right)$ |
| $\Delta_\zeta$ | $O\left(\min\left\{\frac{(\ln n)^2}{n}L_2^2\Theta_1, \frac{\varepsilon^2}{n\Theta_1}, \frac{\varepsilon L_2}{n}\right\}\right)$ | $O\left(\min\left\{nL_2^2\Theta_2, \frac{\varepsilon^2}{n\Theta_2}, \frac{\varepsilon L_2}{n}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{\frac{\ln n}{\sqrt{n}}L_2\sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n\Theta_1}}, \sqrt{\frac{\varepsilon L_2}{n}}\right\}\right)$ | $O\left(\min\left\{\sqrt{n}L_2\sqrt{\Theta_2}, \frac{\varepsilon}{\sqrt{n\Theta_2}}, \sqrt{\frac{\varepsilon L_2}{n}}\right\}\right)$ |
| 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.*

# 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

$$\widetilde{f}(x, \xi) = F(x, \xi) + \Xi(x, \xi), \quad |\Xi(x, \xi)| \leqslant \Delta, \ \forall x \in \mathbb{R}^n, \text{ a.s. in } \xi, \qquad (11)$$

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 $\nabla f(x)$

## Derivative-Free Optimization

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

$$\widetilde{\nabla}^m f^t(x) = \frac{1}{m}\sum_{i=1}^{m} \frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t} e = \left(\left\langle g^m(x,\vec{\xi_m}), e\right\rangle + \frac{1}{m}\sum_{i=1}^{m}(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right) e, \tag{12}$$

# Derivative-Free Optimization

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

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\ &= \left(\left\langle g^m(x,\vec{\xi_m}),e\right\rangle + \frac{1}{m}\sum_{i=1}^{m}(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{12}$$

where $e \in RS_2(1)$,

# Derivative-Free Optimization

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

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\ &= \left(\left\langle g^m(x,\vec{\xi_m}), e\right\rangle + \frac{1}{m}\sum_{i=1}^{m}(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{12}$$

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

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e\\ &= \left(\left\langle g^m(x,\vec{\xi_m}), e\right\rangle + \frac{1}{m}\sum_{i=1}^{m}(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{12}$$

where $e \in RS_2(1)$, $\xi_i$, $i = 1, \ldots, m$ are independent realizations of $\xi$, $m$ is the *batch size*,

# Derivative-Free Optimization

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

$$
\begin{aligned}
\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\
&= \left(\left\langle g^m(x,\vec{\xi_m}), e\right\rangle + \frac{1}{m}\sum_{i=1}^{m}(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,
\end{aligned}
\tag{12}
$$

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*,

# Derivative-Free Optimization

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

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^m \frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\ &= \left(\left\langle g^m(x,\vec{\xi}_m), e\right\rangle + \frac{1}{m}\sum_{i=1}^m(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{12}$$

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*, $g^m(x,\vec{\xi}_m) := \frac{1}{m}\sum_{i=1}^m g(x,\xi_i)$

# Derivative-Free Optimization

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

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\ &= \left(\left\langle g^m(x,\vec{\xi_m}),e\right\rangle + \frac{1}{m}\sum_{i=1}^{m}(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{12}$$

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

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

# Derivative-Free Optimization

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

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^m \frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\ &= \left(\left\langle g^m(x,\vec{\xi_m}), e\right\rangle + \frac{1}{m}\sum_{i=1}^m(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{12}$$

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

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

# Derivative-Free Optimization

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

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

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, $\mathbb{E}_\xi(\zeta(x,\xi,e))^2 \leqslant \frac{L_2^2t^2}{4} =: \Delta_\zeta$ for all $x\in\mathbb{R}^n$ and $e\in S_2(1)$.

# Derivative-Free Optimization

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

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, $\mathbb{E}_\xi(\zeta(x, \xi, e))^2 \leqslant \frac{L_2^2 t^2}{4} =: \Delta_\zeta$ for all $x \in \mathbb{R}^n$ and $e \in S_2(1)$. At the same time, from (11), we have that $|\eta(x, \xi, e)| \leqslant \frac{2\Delta}{t} =: \Delta_\eta$ for all $x \in \mathbb{R}^n$, $e \in S_2(1)$ and a.s. in $\xi$.

# Derivative-Free Optimization

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

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, $\mathbb{E}_\xi(\zeta(x,\xi,e))^2 \leqslant \frac{L_2^2 t^2}{4} =: \Delta_\zeta$ for all $x \in \mathbb{R}^n$ and $e \in S_2(1)$. At the same time, from (11), we have that $|\eta(x,\xi,e)| \leqslant \frac{2\Delta}{t} =: \Delta_\eta$ 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.