On the Complexity Analysis of Randomized Block-Coordinate Descent Methods

Zhaosong Lu and Lin Xiao

Presented By Yunzhe Qiu and Zekun Liu

Outlines

Introduction

- Randomized block-coordinate descent (RBCD) methods
- Related works and contributions

Technical preliminaries

Randomized block-coordinate descent

- Convergence rate of expected values
- High probability complexity bound
- Comparison with Richtarik and Takac (2014)

Accelerated randomized coordinate decent

- Comparison with Nestrov (2012)
- Randomized estimate sequence

Outlines

Introduction

- Randomized block-coordinate descent (RBCD) methods
- Related works and contributions

Technical preliminaries

Randomized block-coordinate descent

- Convergence rate of expected values
- High probability complexity bound
- Comparison with Richtarik and Takac (2014)

Accelerated randomized coordinate decent

- Comparison with Nestrov (2012)
- Randomized estimate sequence

Randomized block-coordinate descent (RBCD) methods (1/2)

- Motivation
  - Minimizing the sum of two convex functions:

$$\min_{x \in \Re^N} \{F(x) \stackrel{\text{def}}{=} f(x) + \Psi(x)\},$$

where f is differentiable on $\Re^N$, and Ψ has a block separable structure:

$$\Psi(x) = \sum_{i=1}^{n} \Psi_i(x_i),$$

where each x_i denotes a subvector of x with cardinality N_i, and each Ψ_i: $\Re^{N_i} \to \Re \cup \{+\infty\}$ is a closed convex function.

Randomized block-coordinate descent (RBCD) methods (2/2)

- Iteration
  - Given the current iterate x^k, the RBCD method picks a block $i \in \{1, \ldots, n\}$ uniformly at random.
  - Solving a block-wise proximal subproblem:

$$d_i(x^k) := \arg\min_{d_i \in \Re^{N_i}} \left\{ \langle \nabla_i f(x^k), d_i \rangle + \frac{L_i}{2} \|d_i\|^2 + \Psi_i(x_i^k + d_i) \right\},$$

    where $\nabla_i f(x)$ denotes the partial gradient of f with respect to x_i, and L_i is the Lipschitz constant of $\nabla_i f(x)$.
  - Setting the new iterate as

$$x_j^{k+1} = \begin{cases} x_i^k + d_i(x^k), & j = i \\ x_j^k, & j \neq i \end{cases}$$

Related works and contributions

- Related works
  - Nesterov (2012)
    - Studying RBCD methods for two special cases
    - Proposing an accelerated RBCD (ARCD) method for the problem with $\Psi \equiv 0$
  - Richtarik and Takac (2014)
    - Extending Nesterov's RBCD methods to the general form
    - Establishing a high-probability type of iteration complexity
- Contributions
  - Obtaining sharper convergence rates for the RBCD method and for the ARCD method in the case $\Psi \equiv 0$ by developing the randomized estimate sequence technique
  - Obtaining a better high-probability type of iteration complexity

Outlines

Introduction

- Randomized block-coordinate descent (RBCD) methods
- Related works and contributions

Technical preliminaries

Randomized block-coordinate descent

- Convergence rate of expected values
- High probability complexity bound
- Comparison with Richtarik and Takac (2014)

Accelerated randomized coordinate decent

- Comparison with Nestrov (2012)
- Randomized estimate sequence

Technical preliminaries (1/7)

- **Assumption 0**. *Problem* $\min_{x\in\Re^N}\{F(x) \stackrel{\text{def}}{=} f(x)+\Psi(x)\}$ *has a minimum* $(F^* > -\infty)$ *and a nonempty optimal solution set, denoted by* X^*,

$$\min_{x\in\Re^N}\{F(x) \stackrel{\text{def}}{=} f(x)+\Psi(x)\}$$

- Permutation
  - For any partition of $x \in \Re^N$ into $\{x_i \in \Re^{N_i}: i = 1, \dots, n\}$, there is an $N \times N$ permutation U partitioned as $U = [U_1 \cdots U_n]$, where $U_i \in \Re^{N\times N_i}$, such that

$$x = \sum_{i=1}^n U_i x_i, \quad \text{and} \quad x_i = U_i^T x, i = 1, \dots, n.$$

  - For any $x \in \Re^N$, the partial gradient of f with respect to x_i is defined as

$$\nabla_i f(x) = U_i^T \nabla f(x), i = 1, \dots, n.$$

- **Assumption 1**. *The gradient of function* f *is block-wise Lipschitz continuous with constants* L_i, *i.e.,,*

$$\|\nabla_i f(x + U_i h_i) - \nabla_i f(x)\| \le L_i \|h_i\|, \quad \forall h_i \in \mathbb{R}^{N_i}, i = 1, \dots, n, x \in \mathbb{R}^N.$$

Technical preliminaries (2/7)

- Norm $\|\cdot\|_L$ and norm $\|\cdot\|_L^*$
  - Define a pair of norms in the whole space $x \in \Re^N$:

$$\|x\|_L = \left(\sum_{i=1}^{n} L_i \|x_i\|^2\right)^{1/2}, \qquad \forall x \in \Re^N,$$

$$\|g\|_L^* = \left(\sum_{i=1}^{n} \frac{1}{L_i} \|g_i\|^2\right)^{1/2}, \qquad \forall g \in \Re^N.$$

- Convexity parameter
  - The convexity parameter of a convex function ϕ: $\Re^N \to \Re \cup \{+\infty\}$ with respect to the norm $\|\cdot\|_L$, denoted by μ_ϕ, is the largest $\mu \geq 0$ such that for all $x, y \in \mathrm{dom}\phi$,

$$\phi(y) \geq \phi(x) + \langle s, y - x\rangle + \frac{\mu}{2}\|y - x\|_L^2, \quad \forall s \in \partial\phi(x).$$

Clearly, ϕ is strongly convex if and only if $\mu_\phi > 0$.

Technical preliminaries (3/7)

- **Lemma 1**. *Suppose that* $\Phi(x) = \sum_{i=1}^{n} \Phi_i(x_i)$. *For any* $x, d \in \Re^N$, *if we pick* $i \in \{1, \ldots, n\}$ *uniformly at random, then*

$$\mathbf{E}_i[\Phi(x + U_i d_i)] = \frac{1}{n}\Phi(x + d) + \frac{n-1}{n}\Phi(x).$$

- *Proof.* Since each i is picked randomly with probability $1/n$, we have

$$\begin{aligned}\mathbf{E}_i[\Phi(x + U_i d_i)] &= \frac{1}{n}\sum_{i=1}^{n}\left(\Phi_i(x_i + d_i) + \sum_{j\neq i}\Phi_j(x_j)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Phi_i(x_i + d_i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\Phi_j(x_j) \\ &= \frac{1}{n}\Phi(x + d) + \frac{n-1}{n}\Phi(x).\end{aligned}$$

- For notational convenience, the author defines

$$H(x, d) \coloneqq f(x) + \langle \nabla f(x), d\rangle + \frac{1}{2}\|d\|_L^2 + \Psi(x + d).$$

Technical preliminaries (4/7)

- Assume $\mu_f \geq 0$ and $\mu_\Psi \geq 0$ with respect to the norm $\|\cdot\|_L$ respectively. Then $\mu_F \geq \mu_f + \mu_\Psi$. By Assumption 1, we have

  $$f(x + U_i h_i) \leq f(x) + \langle \nabla_i f(x), h_i \rangle + \frac{L_i}{2}\|h_i\|^2, \forall h_i \in \mathbb{R}^{N_i}, i = 1, \dots, n, x \in \mathbb{R}^N, \text{(*)}$$

  Which implies that $\mu_f \leq 1$.

- **Lemma 2**. *Suppose $x, d \in \Re^N$. If we pick $i \in \{1, \dots, n\}$ uniformly at random,*

$$\mathbf{E}_i[F(x + U_i d_i)] - F(x) \leq \frac{1}{n}\big(H(x, d) - F(x)\big).$$

- *Proof.* According to **Lemma 1** and inequality (*), we have

$$\begin{aligned}
\mathbf{E}_i[F(x + U_i d_i)] - F(x) &= \frac{1}{n}\big(F(x + d) - F(x)\big) \\
&\leq \frac{1}{n}\left(f(x) + \langle \nabla f(x), d \rangle + \frac{1}{2}\|d\|_L^2 + \Psi(x + d) - F(x)\right) \\
&= \frac{1}{n}\big(H(x, d) - F(x)\big).
\end{aligned}$$

Technical preliminaries (5/7)

- Block-wise composite gradient mapping
  - There exists a subgradient $s_i \in \partial\Psi_i\big(x_i + d_i(x)\big)$ such that
$$\nabla_i f(x) + L_i d_i(x) + s_i = 0$$
  - Let $d(x) = \sum_{i=1}^n U_i d_i(x)$, we then have

$$d_i(x) := \arg\min_{d_i \in \Re^{N_i}} \left\{ \langle \nabla_i f(x), d_i \rangle + \frac{L_i}{2}\|d_i\|^2 + \Psi_i(x_i + d_i) \right\} \Leftrightarrow d(x) = \arg\min_{d \in \Re^N} H(x, d)\,.$$

  - We define the block-wise composite gradient mappings as
$$g_i(x) \overset{\text{def}}{=} -L_i d_i(x),\ i = 1, \ldots, n,$$
$$g(x) = \textstyle\sum_{i=1}^n U_i g_i(x).$$
  - Then we obtain
$$-\nabla_i f(x) + g_i(x) \in \partial\Psi_i\big(x_i + d_i(x)\big),$$
$$-\nabla f(x) + g(x) \in \partial\Psi\big(x + d(x)\big),$$
$$\langle g(x), d(x)\rangle = -\|d(x)\|_L^2 = -(\|g\|_L^*)^2.$$

Technical preliminaries (6/7)

- **Lemma 3**. *For any fixed $x, y \in \Re^N$, if we pick $i \in \{1, \ldots, n\}$ uniformly at random,*

$$\frac{1}{n}F(y) + \frac{n-1}{n}F(x) \geq \mathbf{E}_i\left[F\left(x + U_i d_i(x)\right)\right] + \frac{1}{n}\left(\langle g(x), y - x\rangle + \frac{1}{2}(\|g(x)\|_L^*)^2\right)$$
$$+\frac{1}{n}\left(\frac{\mu_f}{2}\|x - y\|_L^2 + \frac{\mu_\Psi}{2}\|x + d(x) - y\|_L^2\right).$$

- *Proof.* By convexity of f and Ψ and $-\nabla f(x) + g(x) \in \partial\Psi\left(x + d(x)\right)$,

$$H\left(x.d(x)\right) = f(x) + \langle \nabla f(x), d\rangle + \frac{1}{2}\|d\|_L^2 + \Psi(x + d)$$
$$\leq f(y) + \langle \nabla f(x), x - y\rangle - \frac{\mu_f}{2}\|x - y\|_L^2 + \langle \nabla f(x), d(x)\rangle + \frac{1}{2}\|d\|_L^2 + \Psi(y)$$
$$+\langle -\nabla f(x) + g(x), x + d(x) - y\rangle - \frac{\mu_\Psi}{2}\|x + d(x) - y\|_L^2$$
$$= F(y) + \langle g(x), x - y\rangle - \frac{1}{2}(\|g(x)\|_L^*)^2 - \frac{\mu_f}{2}\|x - y\|_L^2 - \frac{\mu_\Psi}{2}\|x + d(x) - y\|_L^2.$$

Technical preliminaries (7/7)

- **Corollary 1**. *Given $x \in \Re^N$, if we pick $i \in \{1, \dots, n\}$ uniformly at random, then*

$$F(x) - \mathbf{E}_i\left[F\left(x + U_i d_i(x)\right)\right] \geq \frac{1+\mu_\Psi}{2}(\|g(x)\|_L^*)^2 = \frac{1+\mu_\Psi}{2}\|d(x)\|_L^2.$$

- Corollary 1 also holds block-wise without taking expectation:

$$F(x) - F\left(x + U_i d_i(x)\right) \geq \frac{1+\mu_\Psi}{2} L_i \|d_i(x)\|^2.$$

- If we do not have knowledge on μ_f or μ_Ψ,
- **Corollary 2**. *For any fixed $x, y \in \Re^N$, if we pick $i \in \{1, \dots, n\}$ uniformly at random, then*

$$\frac{1}{n}F(y) + \frac{n-1}{n}F(x)$$

$$\geq \mathbf{E}_i\left[F\left(x + U_i d_i(x)\right)\right] + \frac{1}{n}\left(\langle g(x), y - x\rangle + \frac{1}{2}(\|g(x)\|_L^*)^2\right).$$

Outlines

Introduction

- Randomized block-coordinate descent (RBCD) methods
- Related works and contributions

Technical preliminaries

Randomized block-coordinate descent

- Convergence rate of expected values
- High probability complexity bound
- Comparison with Richtarik and Takac (2014)

Accelerated randomized coordinate decent

- Comparison with Nestrov (2012)
- Randomized estimate sequence

Randomized block-coordinate descent

- Algorithm

Algorithm: RBCD(x^0)

Repeat for $k = 0, 1, 2, \ldots$

1. Choose $i_k \in \{1, \ldots, n\}$ randomly with a uniform distribution.
2. Update $x^{k+1} = x^k + U_{i_k} d_{i_k}(x^k)$.

  - Define the observed realization of the random variable after k iterations

$$\xi_{k-1} \stackrel{\text{def}}{=} \{i_0, i_1, \ldots, i_{k-1}\}.$$

  - Define the distance between x_0 and the optimal solution set

$$R_0 \stackrel{\text{def}}{=} \min_{x^* \in X^*} \|x^0 - x^*\|_L ,$$

where X^* is the set of optimal solutions of problem

$$\min_{x \in \Re^N} \{F(x) \stackrel{\text{def}}{=} f(x) + \Psi(x)\} .$$

Convergence rate of expected values

- Improvement to Nestrov (2012)
  - Extending the function Ψ from the indicator function of a block-separable closed convex set to the general case by employing the block-wise composite gradient mapping
- **Theorem 1**. *Let F^* be the optimal value, $\{x^k\}$ be the sequence generated by the RBCD method, and $\mathbf{E}_{\xi_{-1}}[F(x^0)] = F(x^0)$. Then for any $k \geq 0$, the iterate x^k satisfies*

$$\mathbf{E}_{\xi_{k-1}}\left[F\left(x^k\right)\right] - F^* \leq \frac{n}{n+k}\left(\frac{1}{2}R_0^2 + F(x^0) - F^*\right).$$

Furthermore, if at least one of f and Ψ is strongly convex, i.e., $\mu_f + \mu_\Psi > 0$, then

$$\mathbf{E}_{\xi_{k-1}}\left[F\left(x^k\right)\right] - F^* \leq \left(1 - \frac{2\left(\mu_f + \mu_\Psi\right)}{n\left(1 + \mu_f + \mu_\Psi\right)}\right)^k \left(\frac{1+\mu_\Psi}{2}R_0^2 + F(x^0) - F^*\right).$$

High probability complexity bound

- **Theorem 2**. *Let* $\{x^k\}$ *be the sequence generated by the RBCD method. Let* $0 < \epsilon < F(x^0) - F^*$ *and* $\rho \in (0,1)$ *be chosen arbitrarily.*

  (i) For all $k \geq K$*, there holds*

  $$\boldsymbol{P}\left(F\left(x^k\right) - F^* \leq \epsilon\right) \geq 1 - \rho, \qquad (*)$$

  where $K := \frac{2nc}{\epsilon}\left(1 + \ln\left(\frac{R_0^2 + 2\left[F\left(x^0\right) - F^*\right]}{4c\rho}\right)\right) + 2 - n.$

  (ii) Furthermore, if at least one of f *and* Ψ *is strongly convex, i.e.,* $\mu_f + \mu_\Psi > 0$*, then inequality (*) holds when* $k \geq \widetilde{K}$*, where* $\widetilde{K} := \frac{n\left(1 + \mu_f + 2\mu_\Psi\right)}{2} \ln\left(\frac{(1+\mu_\Psi)R_0^2 + 2\left[F\left(x^0\right) - F^*\right]}{2\rho\epsilon}\right).$

- **Theorem 3**. *Let* $r = \ln(1/\rho)$*. Suppose we run the RBCD method starting with* x^0 *for* r *times independently, each time for the same number of iterations* k*. Let* $x_{(j)}^k$ *denote the output by the RBCD method at the* k*th iteration of the* j*th run. Then there holds:*

  $$\boldsymbol{P}\left(\min_{\mathbf{1} \leq \boldsymbol{j} < \boldsymbol{r}} F\left(x_{(j)}^k\right) - F^* \leq \epsilon\right) \geq 1 - \rho,$$

  For all $k \geq \underline{K}$*, where* $\underline{K} := \left\lceil \frac{en}{\epsilon}\left(\frac{1}{2}R_0^2 + F(x^0) - F^*\right)\right\rceil - n.$

Comparison with Richtarik and Takac (2014)

- Comparison table to RBCD method

		Richtarik and Takac (2014) (α)	Lu and Xiao (2014) (β)	Comparison (β/α)
Convergence rate of expected value	General setting	$\frac{2nc(F(x^0)-F^*)}{k(F(x^0)-F^*)+2nc}$	$\frac{n}{n+k}\left(\frac{1}{2}R_0^2+F(x^0)-F^*\right)$	$\leq \frac{3}{4}$
	Special case	$\left(1-\frac{2(\mu_f+\mu_\Psi)}{n(1+\mu_f+\mu_\Psi)}\right)^k \left(\frac{1+\mu_\Psi}{2}R_0^2+F(x^0)-F^*\right)$	$\left(1-\frac{\mu_f+\mu_\Psi}{n(1+\mu_\Psi)}\right)^k (F(x^0)-F^*)$	For sufficient large k, $\ll 1$
High Probability complexity bound	General setting	$\frac{2nc}{\epsilon}\left(1+\ln\left(\frac{R_0^2+2[F(x^0)-F^*]}{4c\rho}\right)\right)+2-n$	$\frac{2nc}{\epsilon}\left(1+\ln\frac{1}{\rho}\right)+2-\frac{2nc}{F(x^0)-F^*}$	$\beta-\alpha \leq -\frac{2nc}{\epsilon}\ln\frac{4}{3}$
	Special case	$\frac{n(1+\mu_f+2\mu_\Psi)}{2}\ln\left(\frac{(1+\mu_\Psi)R_0^2+2[F(x^0)-F^*]}{2\rho\epsilon}\right)$	$\frac{n(1+\mu_\Psi)}{\mu_f+\mu_\Psi}\ln\left(\frac{F(x^0)-F^*}{\rho\epsilon}\right)$	For sufficient small ρ and ϵ, ≤ 1

Outlines

Introduction

- Randomized block-coordinate descent (RBCD) methods
- Related works and contributions

Technical preliminaries

Randomized block-coordinate descent

- Convergence rate of expected values
- High probability complexity bound
- Comparison with Richtarik and Takac (2014)

Accelerated randomized coordinate decent

- Comparison with Nestrov (2012)
- Randomized estimate sequence

Accelerated randomized coordinate descent (ARCD) (1/4)

- Problem definition
  - An unconstrained smooth minimization problem

$$\min_{x\in\Re^N} f(x),$$

where f is convex in $\Re^N$ with convexity parameter $\mu = \mu_f > 0$ with respect to the norm $\|\cdot\|_L$ and satisfies Assumption 1, and then $\mu < 1$.

Algorithm: ARCD(x^0)

Set $v^0 = x^0$, choose $\gamma_0 > 0$ arbitrarily, and repeat for $k = 0, 1, 2, \ldots$

1. Compute $\alpha_k \in (0, n]$ from the equation

$$\alpha_k^2 = \left(1 - \tfrac{\alpha_k}{n}\right)\gamma_k + \tfrac{\alpha_k}{n}\mu$$

and set

$$\gamma_{k+1} = \left(1 - \tfrac{\alpha_k}{n}\right)\gamma_k + \tfrac{\alpha_k}{n}\mu.$$

2. Compute y^k as

$$y^k = \frac{1}{\frac{\alpha_k}{n}\gamma_k + \gamma_{k+1}}\left(\tfrac{\alpha_k}{n}\gamma_k v^k + \gamma_{k+1}x^k\right).$$

3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random, and update

$$x^{k+1} = y^k - \tfrac{1}{L_{i_k}} U_{i_k}\nabla_{i_k} f(y^k).$$

4. Set

$$v^{k+1} = \tfrac{1}{\gamma_{k+1}}\left(\left(1 - \tfrac{\alpha_k}{n}\right)\gamma_k v^k + \tfrac{\alpha_k}{n}\mu y^k - \tfrac{\alpha_k}{L_{i_k}} U_{i_k}\nabla_{i_k} f(y^k)\right).$$

Accelerated randomized coordinate descent (ARCD) (2/4)

- Existence of α_k and γ_k
  - In the ARCD algorithm

$$\alpha_k^2 = \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu,$$

and

$$\gamma_k = \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu.$$

  - Let $\gamma > 0$ be arbitrarily given and define

$$h(\alpha) := \alpha^2 - \left(1 - \frac{\alpha}{n}\right)\gamma - \frac{\alpha}{n}\mu, \forall \alpha \geq 0.$$

We have

$$h(0) = -\gamma < 0, \qquad h(n) = n^2 - \mu \geq 0.$$

By continuity of h, there exists some $\alpha^* \in (0, n]$ such that $h(\alpha^*) = 0$.

Moreover, if $\mu = 0$, we have $0 < \alpha^* < n$.

Accelerated randomized coordinate descent (ARCD) (3/4)

- Algorithm after notation simplification

Algorithm: ARCD(x^0)

Set $v^0 = x^0$, choose $\alpha_{-1} \neq 0$, and repeat for $k = 0, 1, 2, \ldots$

1. Compute $\alpha_k \in (0, n]$ from the equation

$$\alpha_k^2 = \left(1 - \frac{\alpha_k}{n}\right) \alpha_{k-1}^2 + \frac{\alpha_k}{n}\mu,$$

and set

$$\theta_k = \frac{n\alpha_k - \mu}{n^2 - \mu}, \qquad \beta_k = 1 - \frac{\mu}{n\alpha_k}.$$

2. Compute y^k as

$$y^k = \theta_k v^k + (1 - \theta_k) x^k.$$

3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random, and update

$$x^{k+1} = y^k - \frac{1}{L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k).$$

4. Set

$$v^{k+1} = \beta_k v^k + (1 - \beta_k) y^k - \frac{1}{\alpha_k L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k).$$

  - At each iteration k, the ARCD method generates y^k,$x^{k+1}$and v^{k+1}, where $x^{k+1}$and v^{k+1} depend on the realization of the random variable $\xi_k = \{i_0, i_1, \ldots, i_k\}$, while y^k depends on the realization of ξ_{k-1}.

Accelerated randomized coordinate descent (ARCD) (4/4)

- **Theorem 4**. *Let f^* be the optimal value of problem $\min_{x\in\Re^N} f(x)$, and $\{x^k\}$ be the sequence generated by the ARCD method. We define $\mathbf{E}_{\xi_{-1}}[f(x^0)] = f(x^0)$. Then for any $k \geq 0$, there holds:*

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)] - f^* \leq \lambda_k \left(f(x^0) - f^* + \frac{\gamma_0 R_0^2}{2} \right),$$

where $\lambda_0 = 1$ and $\lambda_k = \prod_{i=0}^{k-1}\left(1 - \frac{\alpha_i}{n}\right)$. In particular, if $\gamma_0 \geq \mu$, then

$$\lambda_k \leq \min\left\{ \left(1 - \frac{\sqrt{\mu}}{n}\right)^k, \left(\frac{n}{n + k\frac{\sqrt{\gamma_0}}{2}}\right)^2 \right\}.$$

Comparison with Nestrov (2012)(1/2)

- Comparison table of the convergence rate for the ARCD

	$\mu > 0$	$\mu = 0$
Nestrov (2014) ($\boldsymbol{\alpha}$)	$\mu\left[2R_0^2 + \frac{1}{n^2}(f(x^0) - f^*)\right]$ $\left[\left(1 + \frac{\sqrt{\mu}}{2n}\right)^{k+1} - \left(1 - \frac{\sqrt{\mu}}{2n}\right)^{k+1}\right]^{-2}$	$\left(\frac{n}{k+1}\right)^2\left[2R_0^2 + \frac{1}{n^2}(f(x^0) - f^*)\right]$
Lu and Xiao (2014) ($\boldsymbol{\beta}$)	$\min\left\{\left(1 - \frac{\sqrt{\mu}}{n}\right)^k, \left(\frac{n}{n + k\frac{\sqrt{\gamma_0}}{2}}\right)^2\right\}$ $\left(f(x^0) - f^* + \frac{\gamma_0 R_0^2}{2}\right)$	$\left(\frac{n}{n + k\frac{\sqrt{\gamma_0}}{2}}\right)^2\left(f(x^0) - f^* + \frac{\gamma_0 R_0^2}{2}\right)$

Comparison with Nestrov (2012)(2/2)

	$\mu > 0$	$\mu = 0$
Nestrov (2014) ($\boldsymbol{\alpha}$)	$O\left(\left(1+\frac{\sqrt{\mu}}{2n}\right)^{-2k}\right)$	$(2n^2R_0^2 + f(x_0) - f^*)/k^2$
Lu and Xiao (2014) ($\boldsymbol{\beta}$)	$O\left(\left(1-\frac{\sqrt{\mu}}{n}\right)^{k}\right)$	$\left(2n^2R_0^2 + \frac{4n^2}{\gamma_0}(f(x_0) - f^*)\right)/k^2$
Comparison	$\left(1+\frac{\sqrt{\mu}}{2n}\right)^{-2} > 1-\frac{\sqrt{\mu}}{n}$	When $\gamma_0 > 4n^2$, < 1

Randomized estimate sequence (1/7)

- Improvement to Nestrov (2004)
  - Extending the estimate sequence framework from for accelerated full gradient methods to for a RBCD setup.
- **Definition 1**. *Let $\phi_0(x)$ be a deterministic function and $\phi_k(x)$ be a random function depending on ξ_{k-1} for all $k \geq 1$, and $\lambda_k \geq 0$ for all $k \geq 0$. The sequence $\{\phi_k(x), \lambda_k\}_{k=0}^{\infty}$ is called a randomized estimate sequence of the function f(x) if*

$$\lambda_k \to 0$$

and for any $x \in \Re^N$ and all $k \geq 0$ we have

$$\mathbf{E}_{\xi_{k-1}}[\phi_k(x)] \leq (1-\lambda_k) f(x) + \lambda_k \phi_0(x),$$

where $\mathbf{E}_{\xi_{-1}}[\phi_0(x)] \stackrel{\text{def}}{=} \phi_0$.

Randomized estimate sequence (2/7)

- **Lemma 4**. *Let x^* be the optimal solution to problem $\min_{x\in\Re^N} f(x)$ and f^* be the optimal value. Suppose that $\{\phi_k(x), \lambda_k\}_{k=0}^{\infty}$ is a randomized estimate sequence of the function f(x). Assume that $\{x^k\}$ is a sequence such that for each $k \geq 0$,*

$$\mathbf{E}_{\xi_{k-1}}\left[f\left(x^k\right)\right] \leq \min_x \mathbf{E}_{\xi_{k-1}}\left[\phi_k(x)\right],$$

where $\mathbf{E}_{\xi_{-1}}\left[f\left(x^0\right)\right] \stackrel{\text{def}}{=} f\left(x^0\right)$. Then we have

$$\mathbf{E}_{\xi_{k-1}}\left[\left(x^k\right)\right] - f^* \leq \lambda_k(\phi_0(x^*) - f^*) \to 0.$$

Randomized estimate sequence (3/7)

- Proof.

$$\mathbf{E}_{\xi_{k-1}}\left[f\left(x^k\right)\right] \leq \min_x \mathbf{E}_{\xi_{k-1}}[\phi_k(x)]$$

$$\leq \min_x\{(1-\lambda_k)f(x)+\lambda_k\phi_0(x)\}$$

$$\leq (1-\lambda_k)f(x)+\lambda_k\phi_0(x)$$

$$= f^* + \lambda_k\left(\phi_0(x^*)-f^*\right)$$

$$\lambda_k \to 0$$

Randomized estimate sequence (4/7)

- **Lemma 5**. *Let Assume that f satisfies Assumption 1 with convexity parameter $\mu \geq 0$. In addition, suppose that*
  - *$\phi_k(x)$ is an arbitrary deterministic function on $\Re^N$;*
  - *$\{y^k\}_{k=1}^{\infty}$ is a sequence in $\Re^N$ such that y^k depends on ξ_{k-1};*
  - *$\{\alpha_k\}_{k=1}^{\infty}$ is independent of ξ_k and satisfies $\alpha_k \in (0, n)$ for all $k \geq 0$ and $\sum_k^{\infty} \alpha_k = \infty$.*

  Then the pair of sequences $\{\phi_k(x)\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$ constructed by setting $\lambda_0 = 1$ and

$$\lambda_{k+1} = \left(1 - \frac{\alpha_k}{n}\right)\lambda_k,$$

$$\phi_{k+1}(x) = \left(1 - \frac{\alpha_k}{n}\right)\phi_k(x) + \alpha_k\left(\frac{1}{n}f(y^k) + \langle \nabla_{i_k} f(y^k), x_{i_k} - y_{i_k}^k \rangle + \frac{\mu}{2n}\|x - y^k\|_L^2\right),$$

  is a randomized estimate sequence of $f(x)$.

Randomized estimate sequence (5/7)

- Proof.

$$\ln \lambda_k = \sum_{i=0}^{k-1} \ln\left(1 - \frac{\alpha_i}{n}\right) \leq -\frac{1}{n}\sum_{i=0}^{k-1} \alpha_i \to -\infty$$

$$\mathbf{E}_{\xi_{-1}}[\phi_0(x)] = (1-\lambda_0)f(x) + \lambda_0\phi_0(x),$$

$$\begin{aligned}\mathbf{E}_{\xi_k}[\phi_{k+1}(x)] &= \mathbf{E}_{\xi_{k-1}}\left[\mathbf{E}_{i_k}\phi_{k+1}(x)\right] \\ &\leq \mathbf{E}_{\xi_{k-1}}\left[\left(1-\frac{\alpha_i}{n}\right)\phi_k(x) + \frac{\alpha_i}{n}f(x)\right]\end{aligned}$$

Randomized estimate sequence (6/7)

- **Lemma 6**. *Let* $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\|x - v^0\|_L^2$. *Then the randomized estimate sequence constructed in Lemma 5 preserves the canonical form of the functions, i.e., for all* $k \geq 0$,

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}\|x - v^k\|_L^2,$$

where the sequences $\{\gamma_k\}$, $\{v^k\}$ *and* $\{v^k\}$ *are defined as follows:*

$$\gamma_{k+1} = \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu,$$

$$v^{k+1} = \frac{1}{\gamma_{k+1}}\left(\left(1 - \frac{\alpha_k}{n}\right)\gamma_k v^k + \frac{\alpha_k}{n}\mu y^k - \frac{\alpha_k}{L_{i_k}}U_{i_k}\nabla_{i_k} f(y^k)\right),$$

$$\phi_{k+1}^* = \left(1 - \frac{\alpha_k}{n}\right)\phi_k^* + \frac{\alpha_k}{n}f(y^k) - \frac{\alpha_k^2}{2\gamma_{k+1}L_{i_k}}\|\nabla_{i_k} f(y^k)\|_L^2$$
$$+ \frac{\alpha_k\left(1 - \frac{\alpha_k}{n}\right)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2n}\|y^k - v^k\|_L^2 + \langle\nabla_{i_k} f(y^k), v_{i_k}^k - y_{i_k}^k\rangle\right).$$

Randomized estimate sequence (7/7)

- Proof.

$$\nabla^2\phi_{k+1}(x) = \gamma_{k+1} diag\left(L_1 I_{N_i}, \cdots, L_n I_{N_n}\right)$$

$$\nabla\phi_{k+1}\left(v^{k+1}\right) = 0$$

$$\phi_{k+1}\left(y^k\right) = \phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\left\|y^k - v^{k+1}\right\|_L^2$$

ϕ_{k+1} is quadratic, thus

$$\phi_{k+1}(x) = \phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\left\|x - v^{k+1}\right\|_L^2$$

Proof of Theorem 4 (1/3)

- Set up

1. $\phi_0(x) = f(v^0) + \frac{\gamma_0}{2}\|x - v^0\|_L^2$

2. $\{y^k\}, \{\alpha_k\}$ *be generated in the ARCD method*

3. $\{\phi_k(x), \lambda_k\}_{k=0}^{\infty}$ *be the randomized estimate sequence of* $f(x)$ *generated as in Lemma 5 using* $\{y^k\}, \{\alpha_k\}$

Proof of Theorem 4 (2/3)

$$\mathbf{E}_{\xi_{k-1}}\left[f\left(x^k\right)\right] \leq \mathbf{E}_{\xi_{k-1}}\left[\left(\phi_k^* = \min_x \phi_k(x)\right)\right]$$

$$\mathbf{E}_{\xi_{k-1}}\left[f\left(x^k\right)\right] \leq \min_x \mathbf{E}_{\xi_{k-1}}\left[\phi_k(x)\right]$$

Proof of Theorem 4 (3/3)

Decay of λ_k (assume $\gamma_0 \geq \mu$)

- $\gamma_{k+1} = \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu \geq \mu$ $\qquad \alpha_k = \sqrt{\gamma_{k+1}} \geq \sqrt{\mu}$

Thus $\lambda_k = \prod_{i=0}^{k-1}\left(1 - \frac{\alpha_i}{n}\right) \leq \left(1 - \frac{\sqrt{\mu}}{n}\right)^k$

- $\gamma_{k+1} \geq \left(1 - \frac{\alpha_k}{n}\right)\gamma_k \geq \left(1 - \frac{\alpha_k}{n}\right)\gamma_0\lambda_k = \gamma_0\lambda_{k+1}$

$\alpha_k = \sqrt{\gamma_{k+1}} \geq \sqrt{\gamma_0\lambda_{k+1}}$

Since $\{\lambda_k\}$ is a decreasing sequence,

$$\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}} \geq \frac{\frac{\alpha_k}{n}}{2\sqrt{\lambda_{k+1}}} \geq \frac{\sqrt{\gamma_0}}{2n}, \lambda_0 = 1$$

We obtain $\frac{1}{\sqrt{\lambda_k}} \geq 1 + \frac{k\sqrt{\gamma_0}}{2n}$

Therefore $\lambda_k \leq \left(\frac{n}{n + k\frac{\sqrt{\gamma_0}}{2}}\right)^2$

References

- Y. Nesterov. Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Boston, 2004.
- Y. Nesterov. Eciency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization, 22(2): 341{362, 2012.
- P. Richtarik and M. Takac. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Mathematical Programming, 144(1-2):1-38, 2014.