Convergence of Randomized Block Coordinate Gradient Descent

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : December, 31, 2017
Last update : July 25, 2019

Overview

Problem and block notation

Unconstrained optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \ : \ \mathbf{x} \in \mathbb{R}^n$$

- $\mathbf{x}$ is splitted into s blocks

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_s \end{bmatrix}$$

- Each block $\mathbf{x}_i$ is a vector in $\mathbb{R}^{n_i}$, with $n_1 + n_2 + ... + n_s = n$.
- Mathematically, the relationships between $\mathbf{x}$ and $\mathbf{x}_i$ can be descried using block matrix

$$\mathbf{I}_n = \left[\mathbf{U}_1 \,|\, \mathbf{U}_2 \,|\, ... \,|\, \mathbf{U}_s \right]$$

where $\mathbf{U}_i$ is $n \times n_i$ matrix that : all the elements are zero, "special" diagonal elements are 1.
Using $\mathbf{U}_i$, we have

$$\mathbf{x}_i = \mathbf{U}_i^\top \mathbf{x}, \qquad \mathbf{x} = \bigoplus_i \mathbf{U}_i \mathbf{x}_i.$$

Example of using $\mathbf{U}_i$

$n = 6$, $s = 3$ with $n_1 = 1$ (first index), $n_2 = 2$ (indices 2-3), $n_3 = 3$ (indices 4-6) :

$$\mathbf{x} = \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} \qquad \mathbf{U}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{U}_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{U}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Then we have :

- $[\mathbf{U}_1|\mathbf{U}_2|\mathbf{U}_3] = \mathbf{I}_6$
- $\mathbf{x}_1 = \mathbf{U}_1^\top \mathbf{x} = a$, $\mathbf{x}_2 = \mathbf{U}_2^\top \mathbf{x} = \begin{bmatrix} b \\ c \end{bmatrix}$, $\mathbf{x}_3 = \mathbf{U}_3^\top \mathbf{x} = \begin{bmatrix} d \\ e \\ f \end{bmatrix}$
- $\mathbf{x} = \mathbf{U}_1\mathbf{x}_1 \oplus \mathbf{U}_2\mathbf{x}_2 \oplus \mathbf{U}_3\mathbf{x}_3$

Problem setting

Unconstrained optimization problem :

$$\min_{\mathbf{x}} f(\mathbf{x})$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and convex
- f is β-smooth : for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, there is a scalar $\beta > 0$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta \|\mathbf{x} - \mathbf{y}\|_2$$

- f is component-wise β_i-smooth : for all $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^{n_i}$, $i = 1, 2, ..., s$, there is a scalar $\beta_i > 0$ such that

$$\|\nabla_i f(\mathbf{x}_i) - \nabla_i f(\mathbf{y}_i)\|_2 \leq \beta_i \|\mathbf{x}_i - \mathbf{y}_i\|_2$$

  where ∇_i is the partial gradient on coordinate indicated by i
- If we let $\mathbf{h} = \mathbf{y} - \mathbf{x}$, the component-wise β_i-smoothness condition can be re-written as follows : for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{h}_i \in \mathbb{R}^{n_i}$, $i = 1, 2, ..., s$, there is a scalar $\beta_i > 0$ such that

$$\|\nabla_i f(\mathbf{x}_i + \mathbf{h}_i) - \nabla_i f(\mathbf{x}_i)\|_2 \leq \beta_i \|\mathbf{h}_i\|_2$$

Further notation and assumptions

Define the following :

- $\beta_{\max}$ as the largest β_i among all the β_i
- $\beta_{\min}$ as the smallest β_i among all the β_i
- We can assume $\beta_i \leq \beta \leq \sum_i \beta_i \leq s\beta_{\max}$ for all $i = 1, ...s$

Other assumptions and notation

- Assumes there is a minimizer $\mathbf{x}^*$ that minimizes f
- $f^* = f(\mathbf{x}^*)$ is the minimum value of f.
- $R_0 = R(\mathbf{x}_0) = \max\{\|\mathbf{x} - \mathbf{x}^*\|_2 : f(\mathbf{x}) \leq f(\mathbf{x}_0)\} < \infty$.
  That is, there is a finite R_0 that the level set for f defined by a given point $\mathbf{x}_0$ is bounded.

Randomized Block Coordinate Gradient Descent Algorithm

RBCGD :

- Uniform probability is used for selecting the component to update
- Gradient descent update is used for component update

Algorithm 1: RBCGD

Result: A solution $\mathbf{x}$ that approimately solves $\min_{\mathbf{x}} f(\mathbf{x})$

Initialization pick initial point $\mathbf{x}_0 \in \mathbb{R}^n$

while *stopping condition is not met* **do**

Random indexing : pick i_k as $\mathbb{P}(i_k = j) = \frac{1}{s}$

Gradient update : update selected coordinate x_k as

$$\mathbf{x}_{i_k} = \mathbf{x}_{i_k} - t_{i_k} \nabla_{i_k} f(\mathbf{x}^k),$$

or, using the notation of matrix $\mathbf{U}_{i_k}$,

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_{i_k} \mathbf{U}_{i_k} \nabla_{i_k} f(\mathbf{x}^k)$$

end

Convergence of RBCGD ... 1/7

Theorem For convex and β-smooth f, the RBCGD algorithm with constant step size $t_{i_k} = \frac{1}{\beta}$ will produce solution $\mathbf{x}$ at the k iteration such that

$$\mathbb{E}[f(\mathbf{x}^k) - f^*] \leq \frac{2\beta s R(\mathbf{x}_0)^2}{k}.$$

i.e., on average, the RBCGD has convergence rate $\mathcal{O}(\frac{1}{k})$.
Note. As $\mathbb{E}f^* = f^*$, so $\mathbb{E}[f(\mathbf{x}^k) - f^*] = \mathbb{E}f(\mathbf{x}^k) - f^*$

Proof. First we have

$$f \text{ is } \beta\text{-smooth} \quad f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \qquad (1)$$

$$\text{RBCGD update} \qquad \mathbf{x}^{k+1} = \mathbf{x}^k - t_{i_k}\mathbf{U}_{i_k}\nabla_{i_k} f(\mathbf{x}^k) \qquad (2)$$

Put $\mathbf{y} = \mathbf{x}^{k+1}$ and $\mathbf{x} = \mathbf{x}^k$ into (1), then use (2) for $\mathbf{x}^{k+1} - \mathbf{x}^k$, we get

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - t_{i_k}\nabla f(\mathbf{x}^k)^\top \mathbf{U}_{i_k}\nabla_{i_k} f(\mathbf{x}^k) + \frac{\beta}{2}\|t_{i_k}\mathbf{U}_{i_k}\nabla_{i_k} f(\mathbf{x}^k)\|_2^2$$

Convergence of RBCGD ... 2/7

Put $t_{i_k} = \frac{1}{\beta}$, with $\nabla f(\mathbf{x}^k)^\top \mathbf{U}_{i_k} \nabla_{i_k} f(\mathbf{x}^k) = \|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2$, we have

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{1}{\beta}\|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2 + \frac{1}{2\beta}\|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2 = f(\mathbf{x}^k) - \frac{1}{2\beta}\|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2$$

Subtract f^* from both side

$$f(\mathbf{x}^{k+1}) - f^* \leq f(\mathbf{x}^k) - f^* - \frac{1}{2\beta}\|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2$$

Take $\mathbb{E}_{i_k}$ on both side

$$\mathbb{E}\Big(f(\mathbf{x}^{k+1}) - f^*\Big) \leq \mathbb{E}\Big(f(\mathbf{x}^k) - f^* - \frac{1}{2\beta}\|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2\Big)$$

Only the terms $\mathbf{x}^{k+1} = \mathbf{x}^k - t_{i_k}\mathbf{U}_{i_k}\nabla_{i_k} f(\mathbf{x}^k)$ and $\nabla_{i_k} f(\mathbf{x}^k)$ depend on i_k

$$\mathbb{E}_{i_k} f(\mathbf{x}^{k+1}) - f^* \quad \leq \quad f(\mathbf{x}^k) - f^* - \frac{1}{2\beta}\mathbb{E}_{i_k}\|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2 \qquad (3)$$

Convergence of RBCGD ... 3/7

As index i_k is chosen with uniform probability $\frac{1}{s}$, so

$$\mathbb{E}_{i_k}\|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2 \;=\; \sum_{j=1}^{s}\|\nabla_j f(\mathbf{x}^k)\|_2^2\frac{1}{s}$$

As $\nabla_j f(\mathbf{x}^k)$ is the j^{th} component of $\nabla f(\mathbf{x})$, so the sum $\sum_j \|\nabla_j f(\mathbf{x}^k)\|_2^2$ equals to $\|\nabla f(\mathbf{x})\|_2^2$

$$\mathbb{E}_{i_k}\|\nabla_{i_k} f(\mathbf{x}^k)\|_2^2 \;=\; \frac{1}{s}\|\nabla f(\mathbf{x}^k)\|_2^2$$

Put this back to (3)

$$\mathbb{E}_{i_k} f(\mathbf{x}^{k+1}) - f^* \;\leq\; f(\mathbf{x}^k) - f^* - \frac{1}{2\beta s}\|\nabla f(\mathbf{x}^k)\|_2^2$$

Take $\mathbb{E}$ on both side w.r.t. random variable $i_0, i_1, \ldots, i_k$

$$\mathbb{E} f(\mathbf{x}^{k+1}) - f^* \;\leq\; \mathbb{E} f(\mathbf{x}^k) - f^* - \frac{1}{2\beta s}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|_2^2$$

Let $\phi_k = \mathbb{E} f(\mathbf{x}^k) - f^*$, we get a compact expression

$$\phi_{k+1} \;\leq\; \phi_k - \frac{1}{2\beta s}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|_2^2$$

Convergence of RBCGD ... 4/7

Apply Jensen's inequality $-\mathbb{E}g(X) \leq -g(\mathbb{E}X)$

$$\phi_{k+1} \leq \phi_k - \frac{1}{2\beta s}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|_2^2 \leq \phi_k - \frac{1}{2\beta s}\left(\mathbb{E}\|\nabla f(\mathbf{x}^k)\|\right)^2 \tag{4}$$

What's next : bound $\left(\mathbb{E}\|\nabla f(\mathbf{x}^k)\|\right)^2$
f is convex $\implies f(\mathbf{y}) - f(\mathbf{x}) \leq \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$, put $\mathbf{y} = \mathbf{x}^k$, $\mathbf{x} = \mathbf{x}^*$:

$$f(\mathbf{x}^k) - f^* \leq \nabla f(\mathbf{x}^k)^\top(\mathbf{x}^k - \mathbf{x}^*)$$

By Cauchy–Schwarz inequality : $\mathbf{a}^\top\mathbf{b} \leq \|\mathbf{a}^\top\mathbf{b}\| \leq \|\mathbf{a}\|\|\mathbf{b}\|$,

$$f(\mathbf{x}^k) - f^* \leq \|\nabla f(\mathbf{x}^k)\|\|\mathbf{x}^k - \mathbf{x}^*\|$$

From assumption $R(\mathbf{x}_0) = \max\{\|\mathbf{x} - \mathbf{x}^*\|_2 : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$, as $f(\mathbf{x}^k) \leq f(\mathbf{x}_0)$, so $\|\mathbf{x}^k - \mathbf{x}^*\| \leq R(\mathbf{x}_0)$ and

$$f(\mathbf{x}^k) - f^* \leq R(\mathbf{x}_0)\|\nabla f(\mathbf{x}^k)\|$$

Take expectation

$$\mathbb{E}f(\mathbf{x}^k) - f^* \leq R(\mathbf{x}_0)\mathbb{E}\|\nabla f(\mathbf{x}^k)\| \iff \mathbb{E}\|\nabla f(\mathbf{x}^k)\| \geq \frac{\phi_k}{R(\mathbf{x}_0)}$$

Now we have an inequality involving $\mathbb{E}\|\nabla f(\mathbf{x}^k)\|$ to put into (4)!

Rearrange (4) :

$$\frac{1}{2\beta s}\left(\mathbb{E}\|\nabla f(\mathbf{x}^k)\|\right)^2 \quad \leq \quad \phi_k - \phi_{k+1}$$

Square the inequality : $\frac{\phi_k^2}{R(\mathbf{x}_0)^2} \leq \left(\mathbb{E}\|\nabla f(\mathbf{x}^k)\|\right)^2$, put it into (4) we have

$$\frac{1}{2\beta s}\frac{\phi_k^2}{R(\mathbf{x}_0)^2} \leq \frac{1}{2\beta s}\left(\mathbb{E}\|\nabla f(\mathbf{x}^k)\|\right)^2 \leq \phi_k - \phi_{k+1}$$

Ignore the middle

$$\frac{1}{2\beta s}\frac{\phi_k^2}{R(\mathbf{x}_0)^2} \quad \leq \quad \phi_k - \phi_{k+1}$$

Re-arrange, divide both side by ϕ_k^2,

$$\frac{\phi_k - \phi_{k+1}}{\phi_k^2} \geq \frac{1}{2\beta s}\frac{1}{R(\mathbf{x}_0)^2} \tag{5}$$

Convergence of RBCGD ... 6/7

Trick : as $f(\mathbf{x}^k) \geq f(\mathbf{x}^{k+1})$, thus

$$\begin{aligned} \phi_k = f(\mathbf{x}^k) - f^* &\geq f(\mathbf{x}^{k+1}) - f^* = \phi_{k+1} \\ \phi_k^2 = \phi_k \phi_k &\geq \phi_k \phi_{k+1} \\ \frac{1}{\phi_k \phi_{k+1}} &\geq \frac{1}{\phi_k^2} \end{aligned}$$

So

$$\frac{\phi_k - \phi_{k+1}}{\phi_k \phi_{k+1}} \geq \frac{\phi_k - \phi_{k+1}}{\phi_k^2} \overset{(5)}{\geq} \frac{1}{2\beta s} \frac{1}{R(\mathbf{x}_0)^2}$$

Ignore the middle again gives

$$\frac{1}{\phi_{k+1}} - \frac{1}{\phi_k} \geq \frac{1}{2\beta s} \frac{1}{R(\mathbf{x}_0)^2}$$

Change iteration notation from $k+1$ to k we have

$$\frac{1}{\phi_k} - \frac{1}{\phi_{k-1}} \geq \frac{1}{2\beta s} \frac{1}{R(\mathbf{x}_0)^2}$$

Convergence of RBCGD ... 7/7

By recursion :

$$\begin{aligned}
\frac{1}{\phi_k} &\geq \frac{1}{\phi_{k-1}} + \frac{1}{2\beta s}\frac{1}{R(\mathbf{x}_0)^2} \\
\frac{1}{\phi_k} &\geq \frac{1}{\phi_{k-2}} + \frac{2}{2\beta s}\frac{1}{R(\mathbf{x}_0)^2} \\
&\vdots \\
\frac{1}{\phi_k} &\geq \frac{1}{\phi_0} + \frac{k}{2\beta s}\frac{1}{R(\mathbf{x}_0)^2}
\end{aligned}$$

As $\phi_0 = f(\mathbf{x}_0) - f^* \geq 0$ and $\frac{1}{\phi_0} \geq 0$, we get

$$\begin{aligned}
\frac{1}{\phi_k} &\geq \frac{k}{2\beta s R(\mathbf{x}_0)^2} \\
\phi_k &\leq \frac{2\beta s R(\mathbf{x}_0)^2}{k} \\
\mathbb{E} f(\mathbf{x}^k) - f^* &\leq \frac{2\beta s R(\mathbf{x}_0)^2}{k} \qquad \square
\end{aligned}$$

Last page - summary

For the unconstrained optimization problem $\min_{\mathbf{x}} f(\mathbf{x})$ that

- $\mathbf{x} \in \mathbb{R}^n$ is splitted into s blocks
- $f : \mathbb{R}^n \to \mathbb{R}$ is
  - continuous and convex
  - β-smooth and β_i-component-wise-smooth (here we assume $\beta_i = \beta \, \forall i$)
  - There is a finite $R(\mathbf{x}_0)$ that the level set of f defined by a given point $\mathbf{x}_0$ is bounded

Then the Randomized Block Coordinate Gradient Descent Algorithm :

- Picking index i_k as $\mathbb{P}(i_k = j) = \frac{1}{s}$
- Gradient update as $\mathbf{x}^{k+1} = \mathbf{x}^k - t_{i_k} \mathbf{U}_{i_k} \nabla_{i_k} f(\mathbf{x}^k)$
- Step size $t_{i_k} = \dfrac{1}{\beta}$

converges, in expectation at order $\mathcal{O}(\frac{1}{k})$ as

$$\mathbb{E}[f(\mathbf{x}^k) - f^*] \leq \frac{2\beta s R(\mathbf{x}_0)^2}{k}.$$

End of document