Coordinate Gradient Descent

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft : November, 20, 2018
Last update : November 20, 2018

Overview

1. Problem setting
2. CD with exact component minimization
3. CD using gradient update

Problem setting : quadratic problem

$(\mathcal{P})$: given full rank $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^n$, find $\mathbf{x} \in \mathcal{C} \subset \mathbb{R}^n$ by solving

$$\mathbf{x} := \arg\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{p}^\top \mathbf{x} + c.$$

where $\mathbf{Q} = \mathbf{A}^\top \mathbf{A}, \; \mathbf{p} = \mathbf{A}^\top \mathbf{b}$.

These slides : on using coordinate descent to solve $(\mathcal{P})$.

Full gradient of f is $\nabla f(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$.

Full gradient is used when we update the whole vector $\mathbf{x}$ in the form of $\mathbf{x} = \mathbf{x} + t\nabla f(\mathbf{x})$.

Suppose now we are interested in updating the i^{th} component of $\mathbf{x}$.
i.e., we don't care about the update of the j^{th} component of $\mathbf{x}$ with $j \neq i$.

That is, we consider a sub-problem of $(\mathcal{P})$, such sub-problem is the minimization problem $(\mathcal{P})$ that only focus on x_i.

Component-wise gradient update

Now we are interested in the component-wise update

$$x_i = \text{Update}(x_i; f, \mathbf{x}_{\neq i}),$$

where $\mathbf{x}_{\neq i}$ is vector $\mathbf{x}$ without the i^{th} component.

Such equation means we use the information of f and other components to update x_i.

There are various possible formulations of Update(.). Suppose we use *gradient descent* on $\mathbf{x}_i$, then the *component-wise gradient update* will be

$$x_i = x_i - t_i \nabla_i f(x_i; \mathbf{x}_{\neq i}),$$

where $\nabla_i f(x_i; \mathbf{x}_{\neq i})$ is the partial gradient with the form

$$\nabla_i f(x_i; \mathbf{x}_{\neq i}) = \mathbf{A}_i^\top(\mathbf{A}\mathbf{x} - \mathbf{b}).$$

Note $\mathbf{x}$ is a vector and x_i is an element of $\mathbf{x}$, which is a scalar. Note $\nabla_i f(x_i; \mathbf{x}_{\neq i}) = \mathbf{A}_i^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$ is also scalar. This can be seen as follows :

- $\mathbf{A}_i$ is the i^{th} column of $\mathbf{A}$, which is a vector
- $\mathbf{A}\mathbf{x} - \mathbf{b}$ is a vector
- $\mathbf{A}_i^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$ is the dot product between vectors, so the result is a scalar

Component-wise expression of gradient

Now we only care about x_i, we can express $\nabla_i f(x_i; \mathbf{x}_{\neq i}) = \mathbf{A}_i^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$ as a function of x_i :

$$\begin{aligned} \nabla_i f(x_i; \mathbf{x}_{\neq i}) &= \mathbf{A}_i^\top(\mathbf{A}_i\mathbf{x} - \mathbf{b}) \\ &= \mathbf{A}_i^\top(\mathbf{A}_i x_i \oplus \mathbf{A}_{\neq i}\mathbf{x}_{\neq i}) - \mathbf{b}), \end{aligned}$$

where $\mathbf{A}_{\neq i}$ is matrix $\mathbf{A}$ without the i^{th} column. The notation $\mathbf{A}x_i \oplus \mathbf{A}_{\neq i}\mathbf{x}_{\neq i}$ denotes the block splitting of vector using block matrix $\mathbf{U}_i$ that are $n \times n_i$ matrices that all the elements are zero, except the diagonal elements are 1 and

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

Any way since the partial gradient is a scalar so we have

$$\nabla_i f(x_i; \mathbf{x}_{\neq i}) = \mathbf{A}_i^\top \mathbf{A}_i x_i + \underbrace{\mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} - \mathbf{A}_i^\top \mathbf{b}}_{\text{constants for } x_i}.$$

Component Descent using exact minimization

We want to minimize $f = \dfrac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ component by component.

We just see

$$\nabla_i f(x_i; \mathbf{x}_{\neq i}) = \mathbf{A}_i^\top \mathbf{A}_i x_i + \mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} - \mathbf{A}_i^\top \mathbf{b}.$$

Using 1st order optimality condition (Fermat's rule), we have

$$\nabla_i f(x_i; \mathbf{x}_{\neq i}) = 0$$

We have

$$\mathbf{A}_i^\top \mathbf{A}_i x_i + \mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} - \mathbf{A}_i^\top \mathbf{b} = 0$$

Rearrange

$$x_i = \frac{-\mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} + \mathbf{A}_i^\top \mathbf{b}}{\mathbf{A}_i^\top \mathbf{A}_i}$$

Coordinate Descent with exact component minimization

Algorithm 1: CD with exact component minimization (CD-Ex)

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**

Pick i

Perform exact update

$$x_i = \frac{-\mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} + \mathbf{A}_i^\top \mathbf{b}}{\mathbf{A}_i^\top \mathbf{A}_i}$$

end

Observation : repeated computations of $\mathbf{A}_i^\top \mathbf{A}_{\neq i}, \mathbf{A}_i^\top \mathbf{b}$ and $\mathbf{A}_i^\top \mathbf{A}_i$ inside the loop should be taken out !

(Improved) CD with exact component minimization

What we need to do : pre-compute $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}^\top\mathbf{b}$ outside the loop.

If we let $\mathbf{G} = \mathbf{A}^\top\mathbf{A}$, then $\mathbf{A}_i^\top\mathbf{A}_i = G_{ii}$, and $\mathbf{A}_i^\top\mathbf{A}_{\neq i}$ is the i^{th} row of $\mathbf{G}$ without the i^{th} element, denote as $\mathbf{g}_{i,\neq i}$.

If we let $\mathbf{p} = \mathbf{A}^\top\mathbf{b}$, then $p_i = \mathbf{A}_i^\top\mathbf{b}$.

We have

Algorithm 2: (Improved) CD with exact component minimization (CD-Ex)

Result: A solution $\mathbf{x}$ that appoimately solves $\min\limits_{\mathbf{x}} f(\mathbf{x})$

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

Set $\mathbf{G} = \mathbf{A}^\top\mathbf{A}$ and $\mathbf{p} = \mathbf{A}^\top\mathbf{b}$

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

  Pick i

  Perform exact update

$$x_i = \frac{-\mathbf{g}_{i,\neq i}^\top\mathbf{x}_{\neq i} + p_i}{G_{ii}}.$$

end

CD using gradient update

The update sub problem can also be solved approximately.

Says we use gradient descent with step size t_i, the update

$$x_i = x_i - t_i \nabla_i f(x_i; \mathbf{x}_{\neq i}) = x_i - t_i \mathbf{A}_i^\top (\mathbf{A}\mathbf{x} - \mathbf{b}).$$

Similar to (full) gradient descent, one possible step size t_i is the inverse of the Lipschitz constant.

The function $f = \dfrac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is component-wise β_i-smooth : for all $i = 1, 2, ..., n$, there is a scalar $\beta_i > 0$ such that

$$\|\nabla_i f(a) - \nabla_i f(b)\|_2 \leq \beta_i |a - b|,$$

for any $a, b \in \mathbb{R}$.

Such $\beta_i = \|\mathbf{A}_i\|_2^2$, the L_2 norm of the i^{th} column of $\mathbf{A}$.
(Proof in next page)

Component-wise Lipschitz constant of f is $\|\mathbf{A}_i|_2^2$.

Aim : to show $\beta_i = \|\mathbf{A}_i\|_2^2$.

How : notice that

$$\nabla_i f(x_i; \mathbf{x}_{\neq i}) = \mathbf{A}_i^\top \mathbf{A}_i x_i + \underbrace{\mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} - \mathbf{A}_i^\top \mathbf{b}}_{\text{constants for } x_i}.$$

Therefore we have

$$\begin{aligned}
\nabla_i f(a; \mathbf{x}_{\neq i}) - \nabla_i f(b; \mathbf{x}_{\neq i}) &= \mathbf{A}_i^\top \mathbf{A}_i a - \mathbf{A}_i^\top \mathbf{A} b \\
&= \mathbf{A}_i^\top \mathbf{A}_i (a - b) \\
|\nabla_i f(a; \mathbf{x}_{\neq i}) - \nabla_i f(b; \mathbf{x}_{\neq i})| &= |\mathbf{A}_i^\top \mathbf{A}_i (a - b)| \\
&\overset{\text{c.s.}}{\leq} |\mathbf{A}_i^\top \mathbf{A}_i||a - b|
\end{aligned}$$

Therefore the component-wise Lipschitz constant is $|\mathbf{A}_i^\top \mathbf{A}_i| = \Big|\|\mathbf{A}_i\|_2^2\Big|$. As L_2 norm is always non-negative so we can drop out the absolute sign and have $\beta_i = \|\mathbf{A}_i\|_2^2$.

Component-wise Lipschitz constant of f is $\|\mathbf{A}_i\|_2^2$

Component-wise gradient update

$$x_i = x_i - t_i \nabla_i f(x_i; \mathbf{x}_{\neq i}).$$

Put $t_i = \dfrac{1}{\|\mathbf{A}_i\|_2^2}$ and $\nabla_i f = \mathbf{A}_i^\top \mathbf{A}_i x_i + \mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} - \mathbf{A}_i^\top \mathbf{b}$ We then have

$$\begin{aligned} x_i &= x_i - \frac{\mathbf{A}_i^\top \mathbf{A}_i x_i + \mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} - \mathbf{A}_i^\top \mathbf{b}}{\|\mathbf{A}_i\|_2^2} \\ &= \frac{-\mathbf{A}_i^\top \mathbf{A}_{\neq i}\mathbf{x}_{\neq i} + \mathbf{A}_i^\top \mathbf{b}}{\|\mathbf{A}_i\|_2^2}, \end{aligned}$$

which has the same form as the exact minimization !

Randomized Block Coordinate Gradient Descent Algorithm

One way to select the index i is to select it by random.

Algorithm 3: Randomized Block Coordinate Gradient Descent (RBCGD)

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

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

$\mathbf{G} = \mathbf{A}^\top \mathbf{A}$, $\mathbf{p} = \mathbf{A}^\top \mathbf{b}$ **while** *stopping condition is not met* **do**

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

Gradient update : update selected coordinate x_k as

$$x_i = \frac{-\mathbf{g}_{i,\neq i}^\top \mathbf{x}_{\neq i} + p_i}{G_{ii}}$$

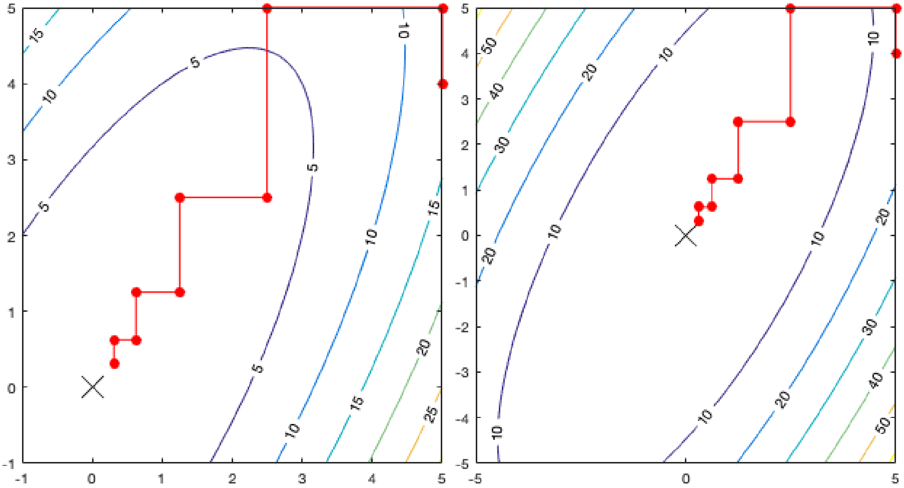
end

Convergence rate of this algorithm is, on average, $\mathcal{O}\left(\frac{1}{k}\right)$.
(For proof, see p 8-14 here)

Convergence rate of cyclic indexing is "similar" but much harder to obtain.

An example in $\mathbb{R}^2$

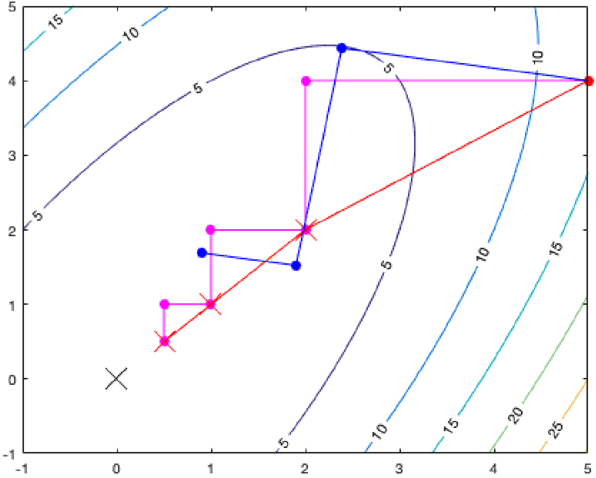
$\mathbf{x}^{\text{True}} = [0\ 0]^\top$, $\mathbf{b} = \mathbf{A}^{-1}\mathbf{x}^{\text{True}}$, $\mathbf{x}_0 = [5\ 4]^\top$, 9 iterations, $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$



CD vs GD

Same setting, 3 iterations.

Notations : Gradient Descent, CD (every n iterations), CD (all iterations)



For this specific example, CD wins.

Last page - summary

Problem $(\mathcal{P})$: given full rank $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^n$, solve

$$\mathbf{x} := \arg\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 = \frac{1}{2}\mathbf{x}^\top\mathbf{Q}\mathbf{x} - \mathbf{p}^\top\mathbf{x} + c.$$

The component-wise Lipschitz constant of f is L_2-norm squared of columns of $\mathbf{A}$.

CD algorithm (with exact component minimization or component-wise gradient descent)

$$x_i = \frac{-\mathbf{g}_{i,\neq i}^\top \mathbf{x}_{\neq i} + p_i}{G_{ii}},\ \text{where } \mathbf{G} = \mathbf{A}^\top\mathbf{A}, \mathbf{p} = \mathbf{A}^\top\mathbf{b}.$$

CD vs GD

End of document