Block Coordinate Descent

Dr. Dingzhu Wen

School of Information Science and Technology (SIST)
ShanghaiTech University

wendzh@shanghaitech.edu.cn

April 1, 2024

Overview

Bachground

Optimization Problem:

$$\min \; f(\mathbf{x}), \quad \text{s.t.} \; \mathbf{x} \in \mathbb{R}^N.$$

For a large-scale probelm, e.g., training deep neural networks, the computation complexity is extremely high.

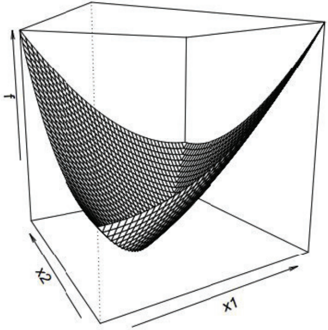
Qeustion: Can we decompose the problems into several low-complexity sub-problems?

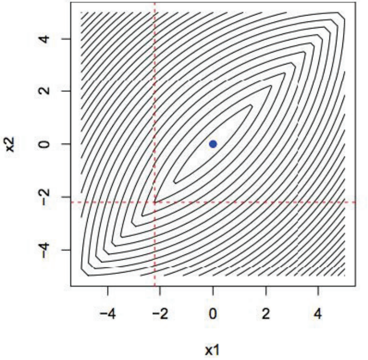
Background

Question: For an arbtrary convex, differentiable function $f : \mathbb{R}^N \rightarrow \mathbb{R}$, if we are at a point $\mathbf{x}$ such that $f(\mathbf{x})$ is minimized along each coordinate axis, then have we found a global minimizer? That is, does $f(\mathbf{x} + \delta \mathbf{e}_i)$ for all $(\delta, i) \Rightarrow f(\mathbf{x}) = \min_{\mathbf{z}} \ f(\mathbf{z})$. $\mathbf{e}_i$ is a basis vector with only one non-zero element, which equals to 1.

Answer: Yes.
Proof: Each coordinate axis has the 0 gradient.
Question: Same question if $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is convex but non-differentiable.
Answer: No.

Background





Background

Question: Same question, but $f : \mathbb{R}^N \to \mathbb{R}$ has the following structure:

$$f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{N} h_i(x_i),$$

where $g(\mathbf{x})$ is convex and differentiable, and $\{h_i(x_i)\}$ are convex but non-differentiable.

Answer: Yes.

Proof: $f(\mathbf{x} + \delta \mathbf{e}_i) = g(\mathbf{x} + \delta \mathbf{e}_i) + \sum_{j \neq i}^{N} h_j(x_j) + h(x_i + \delta)$.

Since $\mathbf{x}$ is optimal along the i-th axis, according to the subgradient optimality, we have

$$\begin{aligned} 0 \in \nabla_i g(\mathbf{x}) + \partial h_i(x_i) &\Leftrightarrow -\nabla_i g(\mathbf{x}) \in \partial h_i(x_i), \\ &\Leftrightarrow h_i(y_i) \geq h_i(x_i) - \nabla_i g(\mathbf{x})(y_i - x_i), \end{aligned}$$

Background

Proof (Continue):

$$\begin{aligned}
0 \in \nabla_i g(\mathbf{x}) + \partial h_i(x_i) &\Leftrightarrow -\nabla_i g(\mathbf{x}) \in \partial h_i(x_i), \\
&\Leftrightarrow h_i(y_i) \geq h_i(x_i) - \nabla_i g(\mathbf{x})(y_i - x_i), \ \forall \mathbf{y} \\
&\Leftrightarrow \nabla_i g(\mathbf{x})(y_i - x_i) + h_i(y_i) - h_i(x_i) \geq 0,
\end{aligned}$$

On the other hand, f is convex. Hence,

$$\begin{aligned}
f(\mathbf{y}) - f(\mathbf{x}) &= g(\mathbf{y}) - g(\mathbf{x}) + \sum_{i=1}^{N} \left[h(y_i) - h(x_i) \right], \\
&\geq \nabla g(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \sum_{i=1}^{N} \left[h(y_i) - h(x_i) \right], \\
&= \sum_{i=1}^{N} \left[\nabla_i g(\mathbf{x})(y_i - x_i) + h_i(y_i) - h_i(x_i) \right], \\
&\geq 0.
\end{aligned}$$

Coordinate Descent

Optimization Problem: $\min \quad f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{N} h_i(x_i)$,

where $g(\mathbf{x})$ is convex and differentiable, and $\{h_i(x_i)\}$ are convex but not necessarily differentiable.

(Cyclic) Coordinate Descent:

- Initialize $\mathbf{x}_0 \in \mathbb{R}^N$
- **For** $k = 1, 2, ...$
  - $x_{i,k} = \arg\min_{x_i} \; f(x_{1,k}, ..., x_{i-1,k}, x_i, x_{i+1,k-1}, ..., x_{N,k-1})$,
- **End for** (Unitl Convergence)

Convergence: Bounded, closed convex function with a monotonic sequence.

Coordinate Descent

Some Practical Notes:

- Order of cycle through coordinates is arbitrary, can use any permutation of $\{1, 2, ..., n\}$.
- Can everywhere replace individual coordinates with blocks of coordinates. For example, we can always update a group of coordinates at the same time.(So called Block Coordinate Descent)
- "One-at-a-time" update scheme is critical, and "all-at-once" scheme does not necessarily converge.
- The analogy for solving linear systems: Gauss-Seidel versus Jacobi method.

Coordinate Descent: Linear Regression

Linear Regression: $\min_{\mathbf{w}} \quad f(\mathbf{w}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$.
For the i-th model parameter:

$$\begin{aligned}\frac{\partial f}{\partial w_i} = 0 &\Leftrightarrow \mathbf{x}_i^T(\mathbf{X}\mathbf{w} - \mathbf{y}) = 0,\\ &\Leftrightarrow \mathbf{x}_i^T \mathbf{x}_i w_i + \mathbf{x}_i^T(\mathbf{X}_{-i}\mathbf{w}_{-i} - \mathbf{y}) = 0,\\ &\Leftrightarrow w_i = \frac{\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}_{-i}\mathbf{w}_{-i})}{\mathbf{x}_i^T \mathbf{x}_i},\end{aligned}$$

where $\mathbf{x}_i$ is the i-th column of $\mathbf{X}$, $\mathbf{X}_{-i}$ is the matrix $\mathbf{X}$ with $\mathbf{x}_i$ removed, $\mathbf{w}_{-i}$ is $\mathbf{w}$ with w_i removed.
Remark: The computational cost (in terms of flops) for 1 cycle of coordinate descent is $\mathcal{O}(MN)$ with M being the number of samples. In each iteration, the computational cost is $\mathcal{O}(M)$ (Same as SGD).

Block Coordinate Descent

Optimization Problem:

$$\min \quad f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{N} h_i(x_i),$$

where $g(\mathbf{x})$ is multi-convex (generally non-convex) and differentiable, and $\{h_i(x_i)\}$ are convex but not necessarily differentiable.
Non-convexity and non-smoothness cause

- Tricky convergence analysis,
- Expensive updates to all variables simultaneously.

Goal: To develop an efficient algorithm with simple update and global convergence (of course, to a stationary point)

Block Coordinate Descent

Optimization Problem:

$$\min \quad f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{N} h_i(x_i),$$

Algorithm 1: Block coordinate descent

- **Initialization**: Choose $(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_s)$
- **For** $k = 1, 2, ...,$ **do**
  - **For** $i = 1, 2, ..., s$ **do**
    - Update $\mathbf{x}_{i,k}$ with all other blocks fixed.
  - **End for**
  - **If** stopping criterion is satisfied **then**
    - Return $(\mathbf{x}_{1,k}, \mathbf{x}_{2,k}, ..., \mathbf{x}_{s,k})$.
  - **End if**
- **End for**

Scheme 1: Block Minimization

The most-often used update:

$$\mathbf{x}_{i,k} = \arg\min_{\mathbf{x}_{i,k}} \quad f(\mathbf{x}_{1,k}, \dots, \mathbf{x}_{i-1,k}, \mathbf{x}_i, \mathbf{x}_{i+1,k-1}, \dots, \mathbf{x}_{s,k-1}).$$

Existing results for differentiable convex f:

- Differentiable f and bounded level set $\rightarrow$ objective converges to optimal value;
- Further with strict convexity $\rightarrow$ sequence converges.

Scheme 1: Block Minimization

The most-often used update:

$$\mathbf{x}_{i,k} = \arg\min_{\mathbf{x}_{i,k}} \; f(\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{i-1,k}, \mathbf{x}_i, \mathbf{x}_{i+1,k-1}, \ldots, \mathbf{x}_{s,k-1}).$$

Existing results for non-differentiable convex f:

- Non-differentiable f can cause stagnation at a non-critical point;
- Non-smooth part is separable $\rightarrow$ subsequence convergence.

Scheme 1: Block Minimization

Example of non-convex f:

May cycle or stagnate at a non-critical point (Powell'73):

$$F(x_1, x_2, x_3) = -x_1x_2 - x_2x_3 - x_3x_1 + \sum_{i=1}^{3}[(x_i - 1)_+^2 + (-x_i - 1)_+^2]$$

Each $F(x_i)$ has the form $(-a)x_i + [(x_i - 1)_+^2 + (-x_i - 1)_+^2]$

it minimizer $x_i^* = sign(a)(1 + 0.5\,|a|)$

Scheme 1: Block Minimization

Example of non-convex f:

Starting from$(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$ with$\epsilon \geq 0$, minimizing F over $x_1, x_2, x_3, x_1, x_2, x_3, \ldots$ produces:

$$
\begin{aligned}
&\xrightarrow{x_1} (1+\tfrac{1}{8}\epsilon, 1+\tfrac{1}{2}\epsilon, -1-\tfrac{1}{4}\epsilon) \quad \xrightarrow{x_2} (1+\tfrac{1}{8}\epsilon, -1-\tfrac{1}{16}\epsilon, -1-\tfrac{1}{4}\epsilon)\\
&\xrightarrow{x_3} (1+\tfrac{1}{8}\epsilon, -1-\tfrac{1}{16}\epsilon, 1+\tfrac{1}{32}\epsilon) \quad \xrightarrow{x_1} (-1-\tfrac{1}{64}\epsilon, -1-\tfrac{1}{16}\epsilon, 1+\tfrac{1}{32}\epsilon)\\
&\xrightarrow{x_2} (-1-\tfrac{1}{64}\epsilon, 1+\tfrac{1}{128}\epsilon, 1+\tfrac{1}{32}\epsilon) \quad \xrightarrow{x_3} (-1-\tfrac{1}{64}\epsilon, 1+\tfrac{1}{128}\epsilon, -1-\tfrac{1}{256}\epsilon)
\end{aligned}
$$

Scheme 1: Block Minimization

Remedies for non-convex f:

- f is differentiable and strictly quasiconvex over each block $\Rightarrow$ The limit point is a critical point (either non-differentiable or gradient equals to 0),
- f is pseudoconvex ($\nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \geq 0 \Rightarrow f(\mathbf{y}) \geq f(\mathbf{x}), \forall(\mathbf{x},\mathbf{y})$) over every two blocks and non-differentiable part is separable $\Rightarrow$ The limit point is a critical point.

No global convergence is guaranteed.

Scheme 2: Block Proximal Descent

Add an regularization term, $\frac{L_{i,k-1}}{2}\|x_i - x_{i,k-1}\|_2^2$

$$\mathbf{x}_{i,k} = \arg\min_{\mathbf{x}_{i,k}} f(\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{i-1,k}, \mathbf{x}_i, \mathbf{x}_{i+1,k-1}, \ldots, \mathbf{x}_{s,k-1}) + \frac{L_{i,k-1}}{2}\|x_i - x_{i,k-1}\|_2^2.$$

Convergence results require fewer assumptions on f:

- f is convex $\Rightarrow$ objective converges to optimal value.
- f is non-convex $\Rightarrow$ limit point is stationary.

Non-smooth terms must still be separable.

Scheme 3: Block Proximal Linear

Linearize g over block i and add $\frac{L_{i,k-1}}{2} \|x_i - x_{i,k-1}\|_2^2$

$$\mathbf{x}_{i,k} = \arg\min_{\mathbf{x}_{i,k}} < \nabla_i g(x_i), x_i - x_{i,k-1} > + r_i(x_i) + \frac{L_{i,k-1}}{2} \|x_i - x_{i,k-1}\|_2^2 .$$

Convergence results require fewer assumptions on f:

- Much easier than schemes 1 & 2; may have closed-form solutions for simple r_i.
- Used in randomized BCD for differentiable convex problems.
- The update is less greedy than schemes 1 & 2, causes more iterations, but may save total time.
- Empirically, the "relaxation" tend to avoid "shallow-puddle" local minima better than schemes 1 & 2.

Block Coordinate Gradient Descent

Optimization Problem: $\min \quad f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{N} h_i(x_i)$,

Algorithm 2: (Cyclic) Block coordinate gradient descent

- **Initialization**: Choose $\mathbf{x}_0$ ans set $k = 0$.
- **Repeat**
  - Choose index $i_k \in \{1, 2, , ..., N\}$.
  - $x_{i_k,k+1} = x_{i_k,k} - \eta_k \nabla_{i_k} f(x_{i_k})$ for some $\eta_k > 0$.
  - $x_{j,k+1} = x_{j,k+1}, \quad \forall j \neq i_k$.
  - $k = k + 1$.
- **Until** convergence.

Block Coordinate Gradient Descent

Optimization Problem: $\min \quad f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{N} h_i(x_i),$

Algorithm 3: Radomized block coordinate gradient descent

- **Initialization**: Choose $\mathbf{x}_0$ ans set $k = 0$.
- **Repeat**
  - Choose index i_k with uniform probability from $\{1, 2, , ..., N\}$, independently of choices at prior iterations.
  - $x_{i_k,k+1} = x_{i_k,k} - \eta_k \nabla_{i_k} f(x_{i_k})$ for some $\eta_k > 0$.
  - $x_{j,k+1} = x_{j,k+1}, \quad \forall j \neq i_k.$
  - $k = k + 1.$
- **Until** convergence.

Block Coordinate Sub-Gradient Descent

Optimization Problem: min $f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{N} h_i(x_i)$,

(Block) Coordinate Gradient Descent for Non-Differentiable $\{h_i(x_i)\}$

Algorithm 4: (Cyclic) Block coordinate subgradient method

- **Initialization**: Choose $\mathbf{x}_0$ ans set $k = 0$.
- **Repeat**
  - Choose index $i_k \in \{1, 2, , ..., N\}$.
  - $x_{i_k,k+1} = \mathrm{Prox}_{\eta_k,h}\left(x_{i_k,k} - \eta_k \nabla_{i_k} g(x_{i_k})\right)$ for some $\eta_k > 0$.
  - $x_{j,k+1} = x_{j,k+1}, \quad \forall j \neq i_k$.
  - $k = k + 1$.
- **Until** convergence.

Convergence Analysis

Convergence Analysis for Differentiable Objectives

Assumption

The objective function f is convex and uniformly L-smooth, and attains its minimum value f_* on a set $\mathcal{S}$. There is a finite R_0 such that the level set for f denoted by x_0 is bounded, that is,

$$\max_{\mathbf{x}_* \in \mathcal{S}} \max_{\mathbf{x}} \{\|\mathbf{x} - \mathbf{x}_*\| : f(\mathbf{x}) \leq f(\mathbf{x}_0)\} \leq R_0.$$

Convergence Analysis

Convergence Analysis for Differentiable Objectives

Theorem

Suppose that the previous assumption holds. Suppose that $\eta_k = \dfrac{1}{L}$ *in* **Algorithm 3**. *Then for all* $k > 0$*, we have*

$$\mathbb{E}[f(\mathbf{x}_k)] - f_* \leq \frac{2NLR_0^2}{k}.$$

Convergence Analysis

Proof: Denote $\mathbf{e}_{i_k}$ as the vector with the i_k-th element being 1 and others being 0.

$$
\begin{aligned}
f(\mathbf{x}_{k+1}) &= f(\mathbf{x}_k - \eta_k \nabla_{i_k} f(x_{i_k})\mathbf{e}_{i_k}), \\
&\leq f(\mathbf{x}_k) - \eta_k [\nabla_{i_k} f(x_{i_k})]^2 + \frac{L\eta_k^2}{2}[\nabla_{i_k} f(x_{i_k})]^2, \quad (L\text{-smoothness}) \\
&= f(\mathbf{x}_k) - \eta_k \left(1 - \frac{L\eta_k}{2}\right)[\nabla_{i_k} f(x_{i_k})]^2, \\
&= f(\mathbf{x}_k) - \frac{1}{2L}[\nabla_{i_k} f(x_{i_k})]^2. \qquad (\eta_k = \frac{1}{L})
\end{aligned}
$$

Convergence Analysis

Then,based on equal probability and take expectation

$$
\begin{aligned}
\mathbb{E}_{i_k} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \frac{1}{2L}\mathbb{E}_{i_k}[\nabla_{i_k} f(x_{i_k})]^2, \\
&= f(\mathbf{x}_k) - \frac{1}{2LN}\sum_{i_k=1}^{N}[\nabla_{i_k} f(x_{i_k})]^2, \\
&= f(\mathbf{x}_k) - \frac{1}{2LN}\left\|\nabla f(\mathbf{x}_k)\right\|^2.
\end{aligned}
$$

Next, denote $\psi_k = \mathbb{E}_{i_{k-1}}\left[f(\mathbf{x}_k)\right] - f_*$,

$$
\psi_{k+1} = \psi_k - \frac{1}{2LN}\mathbb{E}_{i_k}\left[\left\|\nabla f(\mathbf{x}_k)\right\|^2\right].
$$

Convergence Analysis

According to Jensen's inequality,

$$\psi_{k+1} \leq \psi_k - \frac{1}{2LN}\mathbb{E}_{i_k}\left[\|\nabla f(\mathbf{x}_k)\|^2\right] \leq \psi_k - \frac{1}{2LN}\left[\mathbb{E}_{i_k}\left(\|\nabla f(\mathbf{x}_k)\|\right)\right]^2.$$

On the other hand, from convexity,

$$f(\mathbf{x}_k) - f_* \leq \nabla f(\mathbf{x}_k)^T(\mathbf{x}_k - \mathbf{x}_*) \leq \|\nabla f(\mathbf{x}_k)\| \|\mathbf{x}_k - \mathbf{x}_*\| \leq R_0 \|\nabla f(\mathbf{x}_k)\|.$$

It follows that

$$\|\nabla f(\mathbf{x}_k)\| \geq \frac{\psi_k}{R_0},$$

Convergence Analysis

Combining the two inequalities,

$$\psi_{k+1} \leq \psi_k - \frac{1}{2LN}\frac{\psi_k^2}{R_0^2}.$$

Thus,

$$\frac{1}{\psi_{k+1}} - \frac{1}{\psi_k} = \frac{\psi_k - \psi_{k+1}}{\psi_k \psi_{k+1}} \geq \frac{\psi_k - \psi_{k+1}}{\psi_k^2} \geq \frac{1}{2LNR_0^2}.$$

Recursively,

$$\frac{1}{\psi_k} \geq \frac{1}{\psi_0} + \frac{k}{2LNR_0^2} \geq \frac{k}{2LNR_0^2}.$$

This ends the proof.

Thank you!

wendzh@shanghaitech.edu.cn