The Block Proximal Gradient Method

Instructor: Jin Zhang

Department of Mathematics
Southern University of Science and Technology

Fall 2023

Contents

Model

$$\min_{x_1 \in \mathbb{R}^{n_1}, \cdots, x_p \in \mathbb{R}^{n_p}} \left\{ F(x_1, \cdots, x_p) = f(x_1, \cdots, x_p) + \sum_{j=1}^{p} g_j(x_j) \right\}, \qquad (2)$$

where $f : \prod_{j=1}^{p} \mathbb{R}^{n_j} \to \overline{\mathbb{R}}$, $g_j : \mathbb{R}^{n_j} \to \overline{\mathbb{R}}$ for $j = 1, \cdots, p$.

Notations

- A vector $x \in \Pi_{i=1}^{p}\mathbb{R}^{n_i}$ can be written as $x = (x_1, \cdots, x_p) = (x_i)_{i=1}^{p}$. and define

$$\|x\| = \|(x_1, \cdots, x_p)\| = \sqrt{\sum_{i=1}^{p} \|x_i\|^2}$$

- The function $g : \Pi_{i=1}^{p}\mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ is defined by $g(x) = \sum_{i=1}^{p} g_i(x_i)$.
- The gradient of f with respect to the ith block $(i = 1, \cdots, p)$ is denoted by $\nabla_i f$ and it holds that

$$\nabla f(x) = (\nabla_1 f(x), \cdots, \nabla_p f(x)) = (\nabla_i f(x))_{i=1}^{p} .$$

- For any $i = 1, \cdots, p$ we define $\mathcal{U}_i : \mathbb{R}^{n_i} \to \Pi_{i=1}^{p}\mathbb{R}^{n_i}$ to be the linear transformation given by

$$\mathcal{U}_i(d) = (0, \cdots, 0, \underbrace{d}_{i\text{th block}}, 0, \cdots, 0), \text{ for all } d \in \mathbb{R}^{n_i}.$$

Assumption 1:

(A). $g_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ is proper closed and convex for any $i = 1, \cdots, p$.

(B). $f : \Pi_{i=1}^{p} \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ is proper and closed, dom$f$ is convex, dom$g \subseteq$ int(dom$f$), and f is differentiable over int(dom$f$).

(C). f is L_f-smooth over int(dom$f$)$(L_f > 0)$.

(D). There exist $L_1, \cdots, L_p > 0$ such that for any $i = 1, \cdots, p$ it holds that

$$\|\nabla_i f(x) - \nabla_i f\left(x + \mathcal{U}_i(d)\right)\| \leq L_i \|d\|$$

for all $x \in$ int(dom$f$) and $d \in \mathbb{R}^{n_i}$ for which $x + \mathcal{U}_i(d) \in$ int(dom$f$).

(E). The optimal set of problem (2) is nonempty and denoted by X^*. The optimal value of the problem is denoted by F_{opt}.

L_f: global Lipschitz constant
$L_1, L_2, \cdots, L_p$: block Lipschitz constants.

2.1 The Partial Gradient Mappings

Definition (partial prox-grad mapping and partial gradient mapping)

Suppose that f and $g_1, \cdots, g_p$ satisfy properties (A) and (B) of Assumption 1, $L > 0$, and let $i = 1, \cdots, p$. Then

- the ith partial prox-grad mapping is the operator $T_L^{f,g_i} : \text{int}(\text{dom} f) \to \mathbb{R}^{n_i}$ defined by

$$T_L^{f,g_i}(x) = \text{prox}_{\frac{1}{L} g_i}\left(x_i - \frac{1}{L}\nabla_i f(x)\right) \text{ for any } x \in \text{int}(\text{dom} f).$$

- the ith partial gradient mapping is the operator $G_L^{f,g_i} : \text{int}(\text{dom} f) \to \mathbb{R}^{n_i}$ defined by

$$G_L^{f,g_i}(x) = L\left(x_i - T_L^{f,g_i}(x)\right) \text{ for any } x \in \text{int}(\text{dom} f).$$

We set $T_L^i \equiv T_L^{f,g_i}$ and $G_L^i \equiv G_L^{f,g_i}$ when there's no ambiguity.

Let f and $g_1, \cdots, g_p$ satisfy properties (A) and (B) of Assumption 1, $L > 0$. If $g_i \equiv 0$ for some $i = 1, \cdots, p$, then

$$G_L^{f,g_i}(x) = \nabla_i f(x) \text{ for any } x \in \mathrm{int}(\mathrm{dom} f).$$

Lemma

Suppose that f and $g_1, \cdots, g_p$ satisfy properties (A) and (B) of Assumption 1, $L > 0$, and let $i = 1, \cdots, p$. Then for any $x \in \mathrm{int}(\mathrm{dom} f)$,

$$T_L(x) = \left(T_L^1(x), \cdots, T_L^p(x)\right)$$
$$G_L(x) = \left(G_L^1(x), \cdots, G_L^p(x)\right)$$

Theorem (stationary condition)

Let f and $g_1, \cdots, g_p$ satisfy properties (A) and (B) of Assumption 1. Then

(a). $x^* \in$ *dom*(g) *is a stationary point of problem (2) if and only if*

$$-\nabla_i f(x^*) \in \partial g_i(x_i^*), \quad i = 1, \cdots, p;$$

(b). *for any p positive numbers $M_1, \cdots, M_p > 0$, $x^* \in$ dom(g) is a stationary point of problem (2) if and only if*

$$G^i_{M_i}(x^*) = 0, \quad i = 1, \cdots, p.$$

Theorem (monotonicity of the partial gradient mapping)

Suppose that f and $g_1, \cdots, g_p$ satisfy properties (A) and (B) of Assumption 1, and let $i = 1, \cdots, p$. Suppose that $L_1 \geq L_2 > 0$. Then

$$\left\|G^i_{L_1}(x)\right\| \geq \left\|G^i_{L_2}(x)\right\|$$

and

$$\frac{\left\|G^i_{L_1}(x)\right\|}{L_1} \leq \frac{\left\|G^i_{L_2}(x)\right\|}{L_2}$$

for any $x \in \mathrm{int}(\mathrm{dom} f)$.

2.2 The Block Descent Lemma

Lemma: block descent lemma

Let $f : \Pi_{i=1}^{p} \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ be proper with $\mathrm{dom}(f)$ is convex. Assume that f is differentiable over $\mathrm{int}(\mathrm{dom}(f))$. Let $i = 1, \cdots, p$. Suppose that there exists $L_i > 0$ for which

$$\|\nabla_i f(y) - \nabla_i f\left(y + \mathcal{U}_i(d)\right)\| \leq L_i \|d\|$$

for any $y \in \mathrm{int}(\mathrm{dom} f)$ and $d \in \mathbb{R}^{n_i}$ for which $y + \mathcal{U}_i(d) \in \mathrm{int}(\mathrm{dom} f)$. Then

$$f\left(x + \mathcal{U}_i(d)\right) \leq f(x) + \langle \nabla_i f(x), d \rangle + \frac{L_i}{2} \|d\|^2$$

for any $x \in \mathrm{int}(\mathrm{dom} f)$ and $d \in \mathbb{R}^{n_i}$ for which $x + \mathcal{U}_i(d) \in \mathrm{int}(\mathrm{dom} f)$.

Lemma: (block sufficient decrease lemma).

Suppose that f and $g_1, \cdots, g_p$ satisfy properties (A) and (B) of Assumption 1, and Let $i = 1, \cdots, p$. Suppose that there exists $L_i > 0$ for which

$$\|\nabla_i f(y) - \nabla_i f\left(y + \mathcal{U}_i(d)\right)\| \leq L_i \|d\|$$

for all $y \in \operatorname{int}(\operatorname{dom} f)$ and $d \in \mathbb{R}^{n_i}$ for which $y + \mathcal{U}_i(d) \in \operatorname{int}(\operatorname{dom} f)$. Then the following inequality holds:

$$F(x) - F\left(x + \mathcal{U}_i\left(T^i_{L_i}(x) - x_i\right)\right) \geq \frac{1}{2L_i}\left\|G^i_{L_i}(x)\right\|^2$$

for all $x \in \operatorname{int}(\operatorname{dom} f)$.

Notations

The kth iteration $x^k = (x_1^k, \cdots, x_p^k)$.
The kth iteration involves p subiterations, which generate the following auxiliary subsequences:

$$
\begin{aligned}
x^k = x^{k,0} &= (x_1^k, x_2^k, x_3^k, \cdots, x_p^k), \\
x^{k,1} &= (x_1^{k+1}, x_2^k, x_3^k, \cdots, x_p^k), \\
x^{k,2} &= (x_1^{k+1}, x_2^{k+1}, x_3^k, \cdots, x_p^k), \\
&\vdots \\
x^{k+1} = x^{k,p} &= (x_1^{k+1}, x_2^{k+1}, x_3^{k+1}, \cdots, x_p^{k+1}).
\end{aligned}
$$

That is to say, the kth member of the ith auxiliary sequence is

$$
x^{k,i} = \sum_{j=1}^{i} \mathcal{U}_j \left(x_j^{k+1}\right) + \sum_{j=i+1}^{p} \mathcal{U}_j \left(x_j^k\right).
$$

The CBPG Method

Initialization: pick $x^0 = (x_1^0, \cdots, x_p^0) \in \text{int}(\text{dom} f)$.

General Step: for any $k = 0, 1, 2, \cdots$ execute the following steps:

- set $x^{k,0} = x^k$;
- for $i = 1, 2, \cdots, p$, compute

$$x^{k,i} = x^{k,i-1} + \mathcal{U}_i \left(T_{L_i}^i(x^{k,i-1}) - x_i^{k,i-1} \right);$$

- set $x^{k+1} = x^{k,p}$.

3. The CBPG Method
3.1 Convergence Analysis of the CBPG Method-The Nonconvex Case

Lemma: sufficient decrease of the CBPG method-version I

Suppose that Assumption 1 holds, and let $\{x^k\}_{k\geq 0}$ be the sequence generated by the CBPG method for solving problem (2). Then

(a). for all $k \geq 0$ and $j = 0, 1, \cdots, p-1$ it holds that

$$F(x^{k,j}) - F(x^{k,j+1}) \geq \frac{1}{2L_{j+1}} \left\| G_{L_{j+1}}^{j+1}(x^{k,j}) \right\|^2 = \frac{L_{j+1}}{2} \left\| x^{k,j} - x^{k,j+1} \right\|^2;$$

(b). for all $k \geq 0$,

$$F(x^k) - F(x^{k+1}) \geq \frac{L_{\min}}{2} \left\| x^k - x^{k+1} \right\|^2,$$

where $L_{\min} = \min_{i=1,2,\cdots,p} L_i$.

Corollary (monotocicity of the sequence generated by the CBPG method).

Suppose that Assumption 1 holds, and let $\{x^k\}_{k\geq 0}$ be the sequence generated by the CBPG method for solving problem (2). Then for any $k \geq 0$,

$$F(x^{k+1}) \leq F(x^k),$$

and the equality holds if and only if $x^k = x^{k+1}$.

Lemma: sufficient decrease of the CBPG method-version II.

Suppose that Assumption 1 holds and let $\{x^k\}_{k\geq 0}$ be the sequence generated by the CBPG method for solving problem (2). Then for any $k \geq 0$,

$$F(x^k) - F(x^{k+1}) \geq \frac{C}{p} \left\| G_{L_{\mathsf{min}}}(x^k) \right\|^2,$$

where

$$C = \frac{L_{\mathsf{min}}}{2\left(L_f + 2L_{\mathsf{max}} + \sqrt{L_{\mathsf{min}} L_{\mathsf{max}}}\right)^2}$$

and

$$L_{\mathsf{min}} = \min_{i=1,2,\cdots,p} L_i, \quad L_{\mathsf{max}} = \max_{i=1,2,\cdots,p} L_i.$$

Theorem (convergence of the CBPG method-nonconvex case)

Suppose that Assumption 1 holds, and let $\{x^k\}_{k\geq 0}$ *be the sequence generated by the CBPG method for solving problem (2). Denote*

$$L_{min} = \min_{i=1,2,\cdots,p} L_i, \quad L_{max} = \max_{i=1,2,\cdots,p} L_i,$$

and

$$C = \frac{L_{min}}{2\left(L_f + 2L_{max} + \sqrt{L_{min}L_{max}}\right)^2}.$$

Then

(a). $G_{L_{min}}(x^k) \to 0$ *as* $k \to \infty$;

(b). $\min_{n=0,1,\cdots,k} \|G_{L_{min}}(x^n)\| \leq \frac{\sqrt{p(F(x^0)-F_{opt})}}{\sqrt{C(k+1)}}$;

(c). *all limit points of the sequence* $\{x^k\}_{k\geq 0}$ *are stationary points of problem (2).*

Assumption 2

(A). f is convex.

(B). For any $\alpha > 0$, there exists $R_\alpha > 0$ such that

$$\max_{x,x^*} \{\|x - x^*\| : F(x) \leq \alpha, x^* \in X^*\} \leq R_\alpha.$$

Lemma

Suppose that Assumption 1 and Assumption 2 hold, and let $\{x^k\}_{k\geq 0}$ be the sequence generated by the CBPG method for solving problem (2). Then for any $k \geq 0$,

$$F(x^k) - F(x^{k+1}) \geq \frac{L_{\mathsf{min}}}{2p\left(L_f + L_{\mathsf{max}}\right)^2 R^2}\left(F(x^{k+1}) - F_{\mathsf{opt}}\right)^2,$$

where $R = R_{F(x^0)}$, $L_{\mathsf{max}} = \max\limits_{j=1,2,\cdots,p} L_j$, and $L_{\mathsf{min}} = \min\limits_{j=1,2,\cdots,p} L_j$.

Lemma

Let $\{\alpha_k\}_{k\geq 0}$ be a nonnegative sequence of real numbers satisfying

$$\alpha_k - \alpha_{k+1} \geq \frac{1}{\gamma}\alpha_{k+1}^2, \quad k = 0, 1, \cdots,$$

for some $\gamma > 0$. Then for any $n \geq 2$,

$$\alpha_n \leq \max\left\{ \left(\frac{1}{2}\right)^{(n-1)/2} \alpha_0, \quad \frac{4\gamma}{n-1} \right\}.$$

In addition, for any $\epsilon > 0$, if $n \geq 2$ satisfies

$$n \geq \max\left\{ \frac{2}{\log 2}\left(\log \alpha_0 + \log \frac{1}{\epsilon}\right), \frac{4\gamma}{\epsilon} \right\} + 1,$$

then $\alpha_n \leq \epsilon$.

Theorem: ($O\left(\frac{1}{k}\right)$ rate of convergence of CBPG).

Suppose that Assumption 1 and Assumption 2 hold. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the CBPG method for solving problem (2). For any $k \geq 2$,

$$F(x^k) - F_{\text{opt}} \leq \max\left\{ \left(\frac{1}{2}\right)^{(k-1)/2} \left(F(x^0) - F_{\text{opt}}\right), \frac{8p\left(L_f + L_{\max}\right)^2 R^2}{L_{\min}(k-1)} \right\},$$

where $R = R_{F(x^0)}$, $L_{\max} = \max_{i=1,2,\cdots,p} L_i$, and $L_{\min} = \min_{i=1,2,\cdots,p} L_i$.

In addition, if $n \geq 2$ satisfies

$$n \geq \max\left\{ \frac{2}{\log 2}\left(\log\left(F(x^0) - F_{\text{opt}}\right) + \log\frac{1}{\epsilon}\right), \frac{8p\left(L_f + L_{\max}\right)^2 R^2}{L_{\min}\epsilon} \right\} + 1,$$

then $F(x^n) - F_{\text{opt}} \leq \epsilon$.

Finally, the following theorem shows that block Lipschitz continuity (Assumption 1(D)) implies that the function is L-smooth (Assumption 1(C)) if in addition f is differentiable over the entire space.

Theorem

Let $\phi : \Pi_{i=1}^{p} \mathbb{R}^{n_i} \to \mathbb{R}$ be real-valued convex satisfying the following assumptions:

(a). ϕ is differentiable over $\Pi_{i=1}^{p} \mathbb{R}^{n_i}$;

(b). there exist $L_1, \cdots, L_p > 0$ such that for any $i = 1, \cdots, p$ it holds that

$$\|\nabla_i \phi(x) - \nabla_i \phi\left(x + \mathcal{U}_i(d)\right)\| \leq L_i \|d\|$$

for all $x \in \Pi_{i=1}^{p} \mathbb{R}^{n_i}$ and $d \in \mathbb{R}^{n_i}$.

Then ϕ is L-smooth with $L = \sum_{i=1}^{p} L_i$.

Assumption 3

(A). $g_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper closed and convex for $i = 1, \cdots, p$.

(B). $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper closed and convex, dom$g \subseteq$ int(dom$f$), and f is differentiable over int(dom$f$).

(C). There exist $L_1, \cdots, L_p > 0$ such that for any $i = 1, \cdots, p$ it holds that

$$\|\nabla_i f(x) - \nabla_i f\left(x + \mathcal{U}_i(d)\right)\| \leq L_i \|d\|$$

for all $x \in$ int(dom$f$) and $d \in \mathbb{R}^{n_i}$ for which $x + \mathcal{U}_i(d) \in$ int(dom$f$).

(D). The optimal set of problem (2) is nonempty and denoted by X^*. The optimal value is denoted by F_{opt}.

The RBPG Method

Initialization: $x^0 = (x_1^0, \cdots, x_p^0) \in \text{int}(\text{dom} f)$.

General Step: for any $k = 0, 1, 2, \cdots$ execute the following steps:

- pick $i_k \in \{1, 2, \cdots, p\}$ randomly via a uniform distribution;
- $x^{k+1} = x^k + \mathcal{U}_{i_k}\left(T_{L_{i_k}}^{i_k}(x^k) - x_{i_k}^k\right) = x^k - \frac{1}{L_{i_k}}\mathcal{U}_{i_k}\left(G_{L_{i_k}}^{i_k}(x^k)\right)$.

Theorem (sufficient decrease of the RBPG method.)

Suppose that Assumption 3 holds, and let $\{x^k\}_{k\geq 0}$ be the sequence generated by the RBPG method. Then for any $k \geq 0$,

$$F(x^k) - F(x^{k+1}) \geq \frac{1}{2L_{i_k}}\left\|G_{L_{i_k}}^{i_k}(x^k)\right\|^2.$$

Notations:

- $\xi_{k-1} = \{i_0, i_1, \cdots, i_{k-1}\}$ is a multivariate random variable for any $k = 1, 2, \cdots$.
- We write $x \in \Pi_{i=1}^{p} \mathbb{R}^{n_i}$ as $x = (x_1, \cdots, x_p)$. Define the following weighted norm:

$$\|x\|_L = \sqrt{\sum_{i=1}^{p} L_i \|x_i\|^2}$$

  and its dual norm

$$\|x\|_{L,*} = \sqrt{\sum_{i=1}^{p} \frac{1}{L_i} \|x_i\|^2}.$$

- We define $\tilde{G}(x) = \left(G_{L_1}^{1}(x), \cdots, G_{L_p}^{p}(x)\right)$. Obviously, if $L_1 = \cdots = L_p = L$, then $\tilde{G}(x) = G_L(x)$.

Theorem ($O\left(\frac{1}{k}\right)$ rate of convergence of the RBPG method).

Suppose that Assumption 3 holds. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the RBPG method for solving problem (2). Let $x^* \in X^*$. Then for any $k \geq 0$,

$$E_{\xi_k}\left(F(x^{k+1})\right) - F_{\text{opt}} \leq \frac{p}{p+k+1}\left(\frac{1}{2}\left\|x^0 - x^*\right\|_L^2 + F(x^0) - F_{\text{opt}}\right).$$