Noname manuscript No.
(will be inserted by the editor)

On Convergence of Bregman Proximal Point Subgradient Method

Zhenghao Xu ·

the date of receipt and acceptance should be inserted later

1 Preliminaries

1.1 Bregman divergence

Definition 1 (Legendre function) [1] A function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called a Legendre function if h is proper, lsc, strictly convex and essentially smooth.

Definition 2 (Bregman divergence) The Bregman divergence $D_h(x, y)$ induced by Legendre function h is defined as

$$D_h(x, y) \coloneqq h(x) - h(y) - \langle \nabla h(y), x - y \rangle$$

for all $x \in \operatorname{dom} h$ and $y \in \operatorname{dom} \nabla h$.

Here we define the non-differentiable version of relatively smoothness.

Definition 3 (Relatively smooth) [2] Let h be a Legendre function and f (possibly non-differentiable) be a function with $\operatorname{dom} f \subseteq \operatorname{dom} h$. Function f is said to be L-smooth relative to h if

$$Lh - f \quad \text{is convex on } \operatorname{dom} f,$$

or equivalently,

$$f(y) \leq f(x) + \langle f'(x), y - x \rangle + LD_h(y, x)$$

for all $x \in \operatorname{dom} f$, $f'(x) \in \partial f(x)$ and $y \in \operatorname{dom} f$.

Address(es) of author(s) should be given

1.2 Problem setup

We consider the optimization problem(P)

$$\min_{x\in\mathbb{R}^d} f(x) \tag{P}$$

which satisfies following assumptions.
Assumption 1

- h is a Legendre function on $\mathbb{R}^d$,
- $f:\mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$ is a convex, proper, lsc function (possibly not differentiable) and is L-smooth relative to h,
- For any $\lambda>0$, $u,v\in\mathbb{R}^d$ and $x\in\operatorname{int}\operatorname{dom} h$, the problem

  $$\min_{u\in\mathbb{R}^d} f(u)+\frac{1}{\lambda}D_h(u,x)$$

  has a unique minimizer in dom ∇h,
- The problem (P) has nonempty solution.

We denote $\mathcal{F}_L(\mathbb{R}^d)$ as the class of function tuples satisfying above assumptions, namely

$$\mathcal{F}_L(\mathbb{R}^d) \coloneqq \left\{(f,h)\;\middle|\; f,h \text{ satisfy Assumption 1}\right\}.$$

1.3 Method

We use the following fixed step Bregman Proximal Point Subgradient method to solve (P).

Algorithm 1: Bregman Proximal Point Subgradient (BPPSG)

Input: f and h satisfying Assumption 1, $x_0\in\operatorname{int}\operatorname{dom} h$, step size $\lambda\in(0,1/L]$.
Output: x_N
for $k=0,1,\dots,N-1$ **do**

$$y_k=\operatorname*{arg\,min}_{u\in\mathbb{R}^d} f(u)+\frac{1}{\lambda}D_h(u,x_k) \tag{1}$$

$$x_{k+1}=\nabla h^*\left(\nabla h(y_k)-\lambda f'(y_k)\right),\quad f'(y_k)\in\partial(y_k) \tag{2}$$

From the optimality condition of (1) we obtain

$$0\in\partial f(y_k)+\frac{1}{\lambda}\left(\nabla h(y_k)-\nabla h(x_k)\right). \tag{3}$$

Let $f'(y_k)=-\frac{1}{\lambda}\left(\nabla h(y_k)-\nabla h(x_k)\right)$ and plug this into (2) we have

$$\nabla h(x_{k+1})=\nabla h(y_k)+\left(\nabla h(y_k)-\nabla h(x_k)\right). \tag{4}$$

With (3) we get

$$\begin{cases} \nabla h(y_0) = \nabla h(x_0) - f'(y_0), \\ \nabla h(y_{k+1}) = \nabla h(y_k) - \lambda \left(f'(y_k) + f'(y_{k+1})\right) \quad k = 0, 1, 2, \ldots. \end{cases} \tag{5}$$

2 Convergence rate through PEP

We analyze first the convergence of sequence $\{y_k\}$, then move on to $\{x_k\}$. We shall see that they converge at about the same speed.

2.1 PEP formulation

We construct the relaxed version of PEP, that is, we require h to be convex, not necessarily strictly convex.

Firstly, notice that (5) relates to x_0 and $\{y_k\}_{k=0}^{N}$ only. We rename these variables with the minimizer of (P) as $\{u_i\}_{i\in I}$, where $I = \{*, 0, 1, \ldots, N+1\}$,

$$\begin{cases} u_* = \arg\min_u f(u), \\ u_0 = x_0, \\ u_i = y_{i-1}, \quad i = 1, \ldots, N+1. \end{cases}$$

We define variables $\{(f_i, g_i, h_i, s_i)\}_{i\in I} \subseteq (\mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d)^{|I|}$ such that

$$\begin{cases} f_i = f(u_i), \quad g_i = f'(u_i), \\ h_i = h(u_i), \quad s_i = \nabla h(u_i). \end{cases}$$

With the interpolation conditions introduced in [3], we obtain the relaxed PEP formulation after dropping some of the inequalities:

$$\max_{\substack{f_i, h_i \in \mathbb{R} \\ g_i, s_i, u_i \in \mathbb{R}^d}} \quad f_{N+1} - f_*$$

$$s.t. \quad \begin{cases} h_* - h_0 - \langle s_0, u_* - u_0 \rangle \leq R, \\ f_i - f_* + \langle g_i, u_* - u_i \rangle \leq 0, \quad 1 \leq i \leq N+1 \\ f_{i+1} - f_i + \langle g_{i+1}, u_i - u_{i+1} \rangle \leq 0, \quad 1 \leq i \leq N \\ f_0 - f_1 + \langle g_0, u_1 - u_0 \rangle \leq 0, \\ \frac{1}{\lambda} h_{i-1} - f_{i-1} - (\frac{1}{\lambda} h_i - f_i) + \left\langle \frac{1}{\lambda} s_{i-1} - g_{i-1}, u_i - u_{i-1} \right\rangle \leq 0, \quad 2 \leq i \leq N+1 \\ \frac{1}{\lambda} h_{i+1} - f_{i+1} - (\frac{1}{\lambda} h_i - f_i) + \left\langle \frac{1}{\lambda} s_{i+1} - g_{i-1}, u_i - u_{i+1} \right\rangle \leq 0, \quad 1 \leq i \leq N \\ \frac{1}{\lambda} h_0 - f_0 - (\frac{1}{\lambda} h_1 - f_1) + \left\langle \frac{1}{\lambda} s_0 - g_0, u_1 - u_0 \right\rangle \leq 0, \\ \frac{1}{\lambda} h_{N+1} - f_{N+1} - (\frac{1}{\lambda} h_* - f_*) + \left\langle \frac{1}{\lambda} s_{N+1} - g_{N+1}, u_* - u_{N+1} \right\rangle \leq 0, \\ s_1 = s_0 - \lambda g_1 \\ s_i = s_{i-1} - \lambda (g_i + g_{i-1}) \quad 2 \leq i \leq N+1, \end{cases} \tag{r-PEP-P}$$

where $R > 0$ is a predetermined constant and $\lambda \in (0, 1/L]$. Here we make use of the fact that $\frac{1}{\lambda}Lh - f$ is convex as long as $\frac{1}{\lambda} \geq L$. This comes from the additive conservation of convexity, and the fact that h is convex.

Above (r-PEP-P) problem can be formulated as a semidefinite programming and be solved. The resulting dual solution gives a proof of upper bound on BPPSG method.

2.2 Upper bound through duality

We derive the upper bound for sequence $\{y_k\}_{k\geq 0}$. With the notation defined in the last section, $y_N = u_{N+1}$.

Theorem 1 (BPPSG convergence rate) *Let $L > 0$, (f, h, ϕ) are a tuple of function satistying Assumption 1. Then the sequence $\{y_k\}_{k\geq 0}$ generated by Algorithm 1 with constant step size $\lambda \in (0, 1/L]$ satisfies for all $N > 0$,*

$$f(y_N) - f(x_*) \leq \frac{D_h(x_*, x_0)}{\lambda(2N+1)}. \tag{6}$$

Proof We make use of the convexity inequalities in (r-PEP-P) and perform weighted sum of these inequalities, where weights are given by the corresponding dual solutions of (r-PEP-P).

The proof itself is purely arithmetic. We use the notations defined in the last section. The nonzero dual solution and corresponding constraint inequalities involved in the proof:

- $c^{(1)}_{N,i} = \frac{2}{2N+1}, (i = 1, \ldots, N+1)$ for convexity of f between u_* and u_i:
$$f_i - f_* + \langle g_i, u_* - u_i \rangle \leq 0,$$
- $c^{(2)}_{N,i} = \frac{2i-1}{2N+1}, (i = 1, \ldots, N)$ for convexity of f between u_i and u_{i+1}:
$$f_{i+1} - f_i + \langle g_{i+1}, u_i - u_{i+1} \rangle \leq 0,$$
- $c^{(3)}_{N,N} = \frac{1}{2N+1}$ for convexity of f between u_1 and u_0
$$f_0 - f_1 + \langle g_0, u_1 - u_0 \rangle \leq 0,$$
- $c^{(4)}_{N,i} = \frac{2i-1}{4N+2}, (i = 2, \ldots, N+1)$ for convexity of $\frac{1}{\lambda}h - f$ between u_i and u_{i-1}:
$$\frac{1}{\lambda}h_{i-1} - f_{i-1} - (\frac{1}{\lambda}h_i - f_i) + \left\langle \frac{1}{\lambda}s_{i-1} - g_{i-1}, u_i - u_{i-1} \right\rangle \leq 0,$$
- $c^{(5)}_{N,i} = \frac{2i-1}{4N+2}, (i = 1, \ldots, N)$ for convexity of $\frac{1}{\lambda}h - f$ between u_i and u_{i+1}:
$$\frac{1}{\lambda}h_{i+1} - f_{i+1} - (\frac{1}{\lambda}h_i - f_i) + \left\langle \frac{1}{\lambda}s_{i+1} - g_{i+1}, u_i - u_{i+1} \right\rangle \leq 0,$$

– $c^{(6)}_{N,N} = \frac{1}{2N+1}$ for convexity of $\frac{1}{\lambda}h - f$ between u_1 and u_0:

$$\frac{1}{\lambda}h_0 - f_0 - (\frac{1}{\lambda}h_1 - f_1) + \left\langle \frac{1}{\lambda}s_0 - g_0, u_1 - u_0 \right\rangle \leq 0,$$

– $c^{(7)}_{N,N} = \frac{1}{2N+1}$ for convexity of $\frac{1}{\lambda}h - f$ between u_* and u_{N+1}:

$$\frac{1}{\lambda}h_{N+1} - f_{N+1} - (\frac{1}{\lambda}h_* - f_*) + \left\langle \frac{1}{\lambda}s_{N+1} - g_{N+1}, u_* - u_{N+1} \right\rangle \leq 0.$$

The weighted sum is

$$\begin{aligned}
&\sum_{i=1}^{N+1} \frac{2}{2N+1}\left(f_i - f_* + \langle g_i, u_* - u_i\rangle\right) \\
&+ \sum_{i=1}^{N} \frac{2i-1}{2N+1}\left(f_{i+1} - f_i + \langle g_{i+1}, u_i - u_{i+1}\rangle\right) \\
&+ \frac{1}{2N+1}\left(f_0 - f_1 + \langle g_0, u_1 - u_0\rangle\right) \\
&+ \sum_{i=2}^{N+1} \frac{2i-1}{4N+2}\left(\frac{1}{\lambda}h_{i-1} - f_{i-1} - (\frac{1}{\lambda}h_i - f_i) + \left\langle \frac{1}{\lambda}s_{i-1} - g_{i-1}, u_i - u_{i-1}\right\rangle\right) \\
&+ \sum_{i=1}^{N} \frac{2i-1}{4N+2}\left(\frac{1}{\lambda}h_{i+1} - f_{i+1} - (\frac{1}{\lambda}h_i - f_i) + \left\langle \frac{1}{\lambda}s_{i+1} - g_{i+1}, u_i - u_{i+1}\right\rangle\right) \\
&+ \frac{1}{2N+1}\left(\frac{1}{\lambda}h_0 - f_0 - (\frac{1}{\lambda}h_1 - f_1) + \left\langle \frac{1}{\lambda}s_0 - g_0, u_1 - u_0\right\rangle\right) \\
&+ \frac{1}{2N+1}\left(\frac{1}{\lambda}h_{N+1} - f_{N+1} - (\frac{1}{\lambda}h_* - f_*) + \left\langle \frac{1}{\lambda}s_{N+1} - g_{N+1}, u_* - u_{N+1}\right\rangle\right) \\
=& f_{N+1} - f_* - \frac{1}{\lambda(2N+1)}(h_* - h_0) + \frac{1}{2N+1}\left\langle \frac{1}{\lambda}s_0, -u_0\right\rangle \\
&+ \underbrace{\left\langle \sum_{i=1}^{N+1} \frac{2}{2N+1}g_i + \frac{1}{2N+1}\left(\frac{1}{\lambda}s_{N+1} - g_{N+1}\right), u_* \right\rangle}_{I} \\
&+ \underbrace{\sum_{i=2}^{N} \frac{2i-1}{4N+2}\left\langle \frac{1}{\lambda}s_{i-1} - g_{i-1} + \frac{1}{\lambda}s_{i+1} + g_{i+1} - \frac{2}{\lambda}s_i, u_i \right\rangle}_{II} \\
&+ \frac{1}{4N+2}\left\langle \frac{1}{\lambda}s_2 + g_2 - \frac{3}{\lambda}s_1 - g_1 + \frac{2}{\lambda}s_0, u_1 \right\rangle \\
&+ \frac{1}{2}\left\langle \frac{1}{\lambda}s_N - g_N - \frac{1}{\lambda}s_{N+1} - g_{N+1}, u_{N+1} \right\rangle \\
:=& E \leq 0.
\end{aligned}$$

By equation (5), written with new notations as

$$\begin{cases} s_1 = s_0 - g_1, \\ s_{k+1} = s_k - \lambda\left(g_k + g_{k+1}\right) \quad k = 1, \ldots, N, \end{cases}$$

we have

$$\frac{1}{\lambda} s_{i-1} - g_{i-1} + \frac{1}{\lambda} s_{i+1} + g_{i+1} - \frac{2}{\lambda} s_i = 0, \quad \forall 2 \leq i \leq N, \tag{7}$$

$$s_{N+1} + \lambda g_{N+1} + 2\lambda \sum_{i=1}^{N} g_i = s_0. \tag{8}$$

Thus $II = 0$ and

$$I = \frac{1}{\lambda(2N+1)} \left\langle s_0, u_* \right\rangle.$$

The weighted sum of inequalities becomes

$$\begin{aligned} E =& f_{N+1} - f_* - \frac{1}{\lambda(2N+1)}(h_* - h_0) \\ &+ \frac{1}{\lambda(2N+1)} \left\langle s_0, u_* - u_0 \right\rangle \\ &+ \frac{1}{4N+2} \left\langle -\frac{2}{\lambda} s_1 - 2g_1 + \frac{2}{\lambda} s_0, u_1 \right\rangle \\ &+ \frac{1}{2} \left\langle \frac{1}{\lambda} s_N - g_N - \frac{1}{\lambda} s_{N+1} - g_{N+1}, u_{N+1} \right\rangle \\ =& f_{N+1} - f_* - \frac{1}{\lambda(2N+1)} D_h(u_*, u_0) \leq 0. \end{aligned}$$

Replace the symbol back and we obtain the convergence result

$$f(y_N) - f(x_*) \leq \frac{1}{\lambda(2N+1)} D_h(x_*, x_0). \tag{9}$$

Analogously, we can perform PEP on the sequence $\{x_i\}$ and obtain its convergence rate. The result is similar.

Proposition 1 (BPPSG convergence rate (take II)) *Let $L > 0$, (f, h, ϕ) are a tuple of function satistying Assumption 1. Then the sequence $\{x_k\}_{k\geq 0}$ generated by Algorithm 1 with constant step size $\lambda \in (0, 1/L]$ satisfies for all $N > 0$,*

$$f(x_N) - f(x_*) \leq \frac{D_h(x_*, x_0)}{2\lambda N}. \tag{10}$$

References

1. R. Rockafellar. Convex analysis. In *Princeton Landmarks in Mathematics and Physics*, 1970.
2. Heinz H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42:330–348, 2017.
3. Adrien B. Taylor, J. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.