Bregman Method from PEP Perspective

Zhenghao Xu

Zhejiang University

August 1, 2021

Table of Contents

List

- R. A. Dragomir, A. B. Taylor, A. D'Aspremont, and J. Bolte, "Optimal complexity and certification of bregman first-order methods," (2019)
- M. Teboulle, "A simplified view of first order methods for optimization," (2018)
- H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications," (2017)
- Other materials about Bregman method...

Table of Contents

Setup

In [Dragomir et al., 2019], the objective function only has differentiable part. (No nonsmooth proximal map)
The problem is set up on the framework of *relatively-smooth* optimization.

$$\min_{x \in C} f(x) \tag{P}$$

In [Bauschke et al., 2017], there is a nonsmooth term.

Assumptions

$$\min_{x \in C} f(x) \tag{P}$$

- h is proper, closed, *strictly convex* and continuously differentiable
- f is proper, closed, convex and continuously differentiable
- well-posedness (existence of minimizer, uniqueness of subproblem, ...)
- f is smooth *relative* to h, or say, $Lh - f$ is convex

Assumptions

$$\min_{x \in C} f(x) \tag{P}$$

- h is proper, closed, *strictly convex* and continuously differentiable
- f is proper, closed, convex and continuously differentiable
- well-posedness (existence of minimizer, uniqueness of subproblem, ...)
- f is smooth *relative* to h, or say, $Lh - f$ is convex

The last one is a generalization of common L-smooth condition

Relatively-smooth

Following statements are equivalent: [Teboulle, 2018]

- f is L-smooth relative to h
- $Lh - f$ is convex
- $D_f(x, y) \leq LD_h(x, y)$
- $D_{Lh-f}(x, y) \geq 0$
- $f(x) \leq f(y) + \langle \nabla f(z), x - y \rangle + LD_h(x, z) - D_f(y, z)$ (Three Points Descent Lemma)
- $f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + LD_h(x, y)$

Method

Bregman Gradient / NoLips

$\lambda \in (0, \frac{1}{L}]$, $x_0 \in \operatorname{int}\, \operatorname{dom} h$
for $k = 0, 1, \ldots$ do

$$x_{k+1} = \arg\min_{u \in C} \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k)$$

Method

Bregman Gradient / NoLips

$\lambda \in (0, \frac{1}{L}]$, $x_0 \in \mathrm{int}\ \mathrm{dom} h$
for $k = 0, 1, \ldots$ do

$$x_{k+1} = \arg\min_{u \in C} \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k)$$

with *Mirror Map* $\nabla h^*(y) = \arg\max_{u \in \mathbb{R}^n} \langle u, y \rangle - h(u)$, we have

Method

Bregman Gradient / NoLips

$\lambda \in (0, \frac{1}{L}]$, $x_0 \in \operatorname{int}\,\operatorname{dom} h$
for $k = 0, 1, \ldots$ do

$$x_{k+1} = \arg\min_{u \in C} \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k)$$

with *Mirror Map* $\nabla h^*(y) = \arg\max_{u \in \mathbb{R}^n} \langle u, y \rangle - h(u)$, we have

$$\begin{aligned}
x_{k+1} &= \arg\min_{u \in C} \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k) \\
&= \arg\min_{u \in C} \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} (h(u) - h(x_k) - \langle \nabla h(x_k), u - x_k \rangle) \\
&= \arg\min_{u \in C} \langle \lambda \nabla f(x_k) - \nabla h(x_k), u \rangle + h(u) \\
&= \arg\max_{u \in C} \langle \nabla h(x_k) - \lambda \nabla f(x_k), u \rangle - h(u) \\
&= \nabla h^* [\nabla h(x_k) - \lambda \nabla f(x_k)].
\end{aligned}$$

Method

Bregman Gradient / NoLips

$\lambda \in (0, \frac{1}{L}]$, $x_0 \in \mathrm{int}\ \mathrm{dom} h$
for $k = 0, 1, \ldots$ do

$$x_{k+1} = \arg\min_{u \in C} \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k)$$

with *Mirror Map* $\nabla h^*(y) = \arg\max_{u \in \mathbb{R}^n} \langle u, y \rangle - h(u)$, we have

$$\nabla h(x_{k+1}) = \nabla h(x_k) - \lambda \nabla f(x_k).$$

Definition 4 An algorithm $\mathcal{A}$ is called a Bregman first-order algorithm if, for a given problem instance $(f, h) \in \mathcal{B}_L$ and number of iterations $T \in \mathbb{N}$, it generates at each time step $t \in \{0, \ldots, T\}$, a set of primal points $\mathcal{X}_t$ and dual points $\mathcal{Y}_t$ from the following process:

1. Set $\mathcal{X}_0 = \{x_0\}$, where $x_0 \in \operatorname{int}\operatorname{dom} h$ is some initialization point, and $\mathcal{Y}_0 = \{\nabla f(x_0), \nabla h(x_0)\}$.
2. For each $t = 1, \ldots T$, perform one of the two following operations:
   - either call the **primal oracle** $(\nabla f, \nabla h)$ at some point x_t chosen such as
     $$x_t \in \operatorname{Span}(\mathcal{X}_{t-1}) \cap \operatorname{dom} \nabla h$$
     and update the dual set as
     $$\mathcal{Y}_t = \mathcal{Y}_{t-1} \cup \{\nabla f(x_t), \nabla h(x_t)\}.$$
   - Or call the **mirror oracle** ∇h^* at some dual point y_t chosen such as
     $$y_t \in \operatorname{Span}(\mathcal{Y}_{t-1})$$
     with
     $$\nabla h^*(y_t) = \operatorname*{argmin}_{u \in C} h(u) - \langle y_t, u \rangle$$
     and update the primal set as
     $$\mathcal{X}_t = \mathcal{X}_{t-1} \cup \{\nabla h^*(y_t)\}.$$
3. Output some point $x_T \in \operatorname{Span}(\mathcal{X}_T)$.

Convergence Result

Upper bound for NoLips: (proof inspired by PEP dual solution)
In NoLips, let $\lambda \in (0, L^{-1}]$, we have

$$f(x_N) - f(u) \leq \frac{D_h(x_0, u)}{\lambda N}.$$

Convergence Result

Upper bound for NoLips: (proof inspired by PEP dual solution)
In NoLips, let $\lambda \in (0, L^{-1}]$, we have

$$f(x_N) - f(u) \leq \frac{D_h(x_0, u)}{\lambda N}.$$

same rate for Bregman Proximal Gradient method [Teboulle, 2018]

Convergence Result

Upper bound for NoLips: (proof inspired by PEP dual solution)
In NoLips, let $\lambda \in (0, L^{-1}]$, we have

$$f(x_N) - f(u) \leq \frac{D_h(x_0, u)}{\lambda N}.$$

same rate for Bregman Proximal Gradient method [Teboulle, 2018]
Lower bound: (construction inspired by PEP solution with interpolation)

- for NoLips, numerical result: $\frac{D_h(x_0,u)}{\lambda N}$ is the lower bound
- for general Bregman method, constructed lower bound: $\frac{LD_h(x_0,x_*)}{N_1+N_2+1} \cdot (1 - \epsilon)$

Lower Bound

Construct "the worst function"

$$\hat{f}(x) = \max_{i=1,\dots,n} |x^{(i)} - 1 - \frac{\eta}{i}| = \|x - x_*\|_\infty$$

then approximate it to fit in the assumptions

$$f_\mu(x) = \min_{u \in \mathbb{R}^n} \hat{f}(u) + \frac{1}{2\mu}\|x - u\|^2 \tag{5}$$

$$\phi_\mu(t) = \begin{cases} t - \mu/2 & \text{if } t \geq \mu, \\ \frac{1}{2\mu}t^2 & \text{elsewhere.} \end{cases}$$

$$d_\mu(x) = \frac{\mu}{2}\|x\|^2 + \sum_{i=1}^{n} \phi_\mu(x^{(i)}),\ x \in \mathbb{R}^n. \tag{10}$$

$$h_\mu(x) = \frac{1}{L}\left(f_\mu(x) + d_\mu(x)\right). \tag{13}$$

each oracles involved discovers only one dimension per call.

Lower Bound

Theorem 2 (Lower complexity bound for $\mathcal{B}_L$) *Let $N \geq 1$, a precision $\epsilon \in (0,1)$ and let $x_0 \in \mathbb{R}^{2N+1}$ be a starting point. Then, there exist functions $(f,h) \in \mathcal{B}_L(\mathbb{R}^{2N+1})$ such that for any Bregman gradient method $\mathcal{A}$ satisfying Definition 4 and initialized at x_0, the output $\overline{x}$ returned after performing at most N calls to each one of the primal and mirror oracles satisfies*

$$f(\overline{x}) - \min_{\mathbb{R}^{2N+1}} f \geq \frac{LD_h(x_*, x_0)}{2N+1} \cdot (1-\epsilon).$$

$2N$ is the total number of oracle calls, including *primal oracle* (∇f, ∇h) and *mirror oracle* (∇h^*)
can be replaced by $N_1 + N_2$, N_1 primal oracles and N_2 mirror oracles.

PEP Formulation

Assumptions on function class can be expressed as

$$\mathcal{B}_L(\mathbb{R}^n) = \left\{ f, h : \mathbb{R}^n \to \mathbb{R} \;\middle|\; \begin{array}{l} f \text{ is convex, differentiable and has at least one minimizer,} \\ h \text{ is strictly convex and differentiable,} \\ Lh - f \text{ is convex,} \\ \forall \lambda > 0,\ \forall x, p \in \mathbb{R}^n,\ \text{the function } u \mapsto \langle p, u - x \rangle + \frac{1}{\lambda} D_h(u, x) \\ \quad \text{has a unique minimizer.} \end{array} \right\},$$

We may construct PEP:

$$\begin{array}{ll} \text{maximize} & \big(f(x_N) - f(x_*)\big)/D_h(x_*, x_0) \\ \text{subject to} & (f, h) \in \mathcal{B}_L(\mathbb{R}^n), \\ & x_* \text{ is a minimizer of } f, \\ & x_1, \ldots, x_N \text{ are generated from } x_0 \text{ by Algorithm 1 with step size } \lambda, \end{array} \tag{PEP}$$

PEP Discretization

$$
\begin{array}{ll}
\text{maximize} & \big(f(x_N) - f(x_*)\big)/D_h(x_*, x_0) \\
\text{subject to} & (f,h) \in \mathcal{B}_L(\mathbb{R}^n), \\
& x_* \text{ is a minimizer of } f, \\
& x_1, \dots, x_N \text{ are generated from } x_0 \text{ by Algorithm 1 with step size } \lambda,
\end{array}
\tag{PEP}
$$

$\Longleftrightarrow$

$$
\begin{array}{ll}
\text{maximize} & f_N - f_* \\
\text{subject to} & f_i = f(x_i), g_i = \nabla f(x_i), \\
& h_i = h(x_i), s_i = \nabla h(x_i), \quad \text{for all } i \in I \text{ and some } (f,h) \in \mathcal{B}_L(\mathbb{R}^n), \\
& g_* = 0, \\
& s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1 \dots N-1\}, \\
& h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1,
\end{array}
\tag{PEP}
$$

PEP Interpolation

$$\mathcal{B}_L(\mathbb{R}^n) = \left\{ f, h : \mathbb{R}^n \to \mathbb{R} \;\middle|\; \begin{array}{l} f \text{ is convex, differentiable and has at least one minimizer,} \\ h \text{ is strictly convex and differentiable,} \\ Lh - f \text{ is convex,} \\ \forall \lambda > 0,\ \forall x, p \in \mathbb{R}^n,\ \text{the function } u \mapsto \langle p, u - x\rangle + \frac{1}{\lambda} D_h(u, x) \\ \quad \text{has a unique minimizer.} \end{array} \right\},$$

Function class $\mathcal{B}_\mathrm{L}$ is not easy to interpolate.
We turn to interpolate its restricted version and relaxed version.
restricted:

$$\underline{\mathcal{B}_L}(\mathbb{R}^n) = \mathcal{B}_L(\mathbb{R}^n) \cap \{(f, h) : \mathbb{R}^n \to \mathbb{R} \,|\, f \text{ and } Lh - f \text{ are strictly convex}\}$$

relaxed:

$$\overline{\mathcal{B}_L}(\mathbb{R}^n) = \{(f, h) : \mathbb{R}^n \to \mathbb{R} \,|\, f \text{ and } Lh - f \text{ are convex}\}.$$

we get

$$\underline{\mathcal{B}_L}(\mathbb{R}^n) \subset \mathcal{B}_L(\mathbb{R}^n) \subset \overline{\mathcal{B}_L}(\mathbb{R}^n).$$

$$\underline{\mathcal{B}_L}(\mathbb{R}^n) = \mathcal{B}_L(\mathbb{R}^n) \cap \{(f,h) : \mathbb{R}^n \to \mathbb{R} \,|\, f \text{ and } Lh - f \text{ are strictly convex}\}$$

$\Longrightarrow$

$$\begin{aligned}
\text{maximize } & f_N - f_* \\
\text{subject to } & f_i = f(x_i), g_i = \nabla f(x_i), \\
& h_i = h(x_i), s_i = \nabla h(x_i), \quad \text{for all } i \in I \text{ and some } (f,h) \in \underline{\mathcal{B}_L}(\mathbb{R}^n), \\
& g_* = 0, \\
& s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1 \ldots N-1\}, \\
& h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1, \\
& x_i \neq x_j \quad \text{for } i \neq j \in I,
\end{aligned} \tag{$\underline{\text{PEP}}$}$$

while

$$\overline{\mathcal{B}_L}(\mathbb{R}^n) = \{(f,h) : \mathbb{R}^n \to \mathbb{R} \,|\, f \text{ and } Lh - f \text{ are convex}\}.$$

$\Longrightarrow$

$$\begin{aligned}
\text{maximize } & f_N - f_* \\
\text{subject to } & f_i = f(x_i), g_i \in \partial f(x_i), \\
& h_i = h(x_i), s_i \in \partial h(x_i), \\
& Ls_i - g_i \in \partial(Lh - f)(x_i) \quad \text{for all } i \in I \text{ and some } (f,h) \in \overline{\mathcal{B}_L}(\mathbb{R}^n), \\
& g_* = 0, \\
& s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1 \ldots N-1\}, \\
& h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1,
\end{aligned} \tag{$\overline{\text{PEP}}$}$$

Interpolable Conditions

Theorem 3 (Smooth strongly convex interpolation, [37]) *Let I be a finite index set, $\{(x_i, f_i, g_i)\}_{i\in I} \in (\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n)^{|I|}$ and $0 \le \mu \le L \le +\infty$. The following statements are equivalent:*

(i) There exists a proper closed convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that f is μ-strongly convex, has a L-Lipschitz continuous gradient and

$$f_i = f(x_i),\ g_i \in \partial f(x_i) \quad \forall i \in I.$$

(ii) For every $i, j \in I$ we have

$$f_i - f_j - \langle g_j, x_i - x_j\rangle \ge \frac{1}{2L}\|g_i - g_j\|^2 + \frac{\mu}{2(1-\mu/L)}\|x_i - x_j - \frac{1}{L}(g_i - g_j)\|^2.$$

Proposition 3 (Differentiable and strictly convex interpolation) *Let I be a finite index set and $\{(x_i, f_i, g_i)\}_{i\in I} \in (\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n)^{|I|}$. The following statements are equivalent:*

(i) There exists a convex function $f : \mathbb{R}^n \to \mathbb{R}$ such that f is differentiable, strictly convex and

$$f_i = f(x_i),\ g_i = \nabla f(x_i) \quad \forall i \in I.$$

(ii) For every $i, j \in I$ we have

$$\begin{cases} f_i - f_j - \langle g_j, x_i - x_j\rangle > 0 & \text{if } x_i \ne x_j, \\ f_i = f_j \text{ and } g_i = g_j & \text{otherwise.} \end{cases} \tag{19}$$

SDP Representation

With Gram matrix notation

$$G = \begin{pmatrix} G^{xx} & G^{gx} & G^{sx} \\ G^{gx\top} & G^{gg} & G^{gs} \\ G^{sx\top} & G^{gs\top} & G^{ss} \end{pmatrix} \succeq 0$$

$$G_{ij}^{xx} = \langle x_i, x_j \rangle,\ G_{ij}^{gx} = \langle g_i, x_j \rangle,\ G_{ij}^{gs} = \langle g_i, s_j \rangle,\ G_{ij}^{gg} = \langle g_i, g_j \rangle, G_{ij}^{sx} = \langle s_i, x_j \rangle,\ G_{ij}^{ss} = \langle s_i, s_j \rangle, \quad i, j \in I.$$

$$F = (f_0, \ldots, f_N, f_*) \in \mathbb{R}^{N+2}, \quad H = (h_0, \ldots, h_N, h_*) \in \mathbb{R}^{N+2},$$

we convert PEPs into SDP version.

SDP Representation

$$
\begin{aligned}
\text{maximize } & f_N - f_* \\
\text{subject to } & f_i = f(x_i), g_i = \nabla f(x_i), \\
& h_i = h(x_i), s_i = \nabla h(x_i), \quad \text{for all } i \in I \text{ and some } (f,h) \in \underline{\mathcal{B}_L}(\mathbb{R}^n), \\
& g_* = 0, \\
& s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1 \dots N-1\}, \\
& h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1, \\
& x_i \neq x_j \quad \text{for } i \neq j \in I,
\end{aligned}
\tag{\underline{PEP}}
$$

$\iff$

$$
\begin{aligned}
\text{maximize } & f_N - f_* \\
\text{subject to } & f_i - f_j - G^{gx}_{ji} + G^{gx}_{jj} > 0, \\
& (Lh_i - f_i) - (Lh_j - f_j) - L(G^{sx}_{ji} - G^{sx}_{jj}) + G^{gx}_{ji} - G^{gx}_{jj} > 0 \quad \text{for } i \neq j \in I, \\
& G^{gg}_{**} = 0, \\
& G^{sx}_{i+1,j} = G^{sx}_{ij} - \lambda G^{gx}_{ij} \quad \text{for } i \in \{0 \dots N-1\}, j \in I, \\
& h_* - h_0 - G^{sx}_{0*} + G^{sx}_{00} = 1, \\
& G^{xx}_{ii} + G^{xx}_{jj} - 2G^{xx}_{ij} > 0 \quad \text{for } i \neq j \in I, \\
& G \succeq 0,
\end{aligned}
\tag{sdp-\underline{PEP}}
$$

SDP Representation

$$
\begin{aligned}
\text{maximize } & f_N - f_* \\
\text{subject to } & f_i = f(x_i), g_i \in \partial f(x_i), \\
& h_i = h(x_i), s_i \in \partial h(x_i), \\
& Ls_i - g_i \in \partial(Lh - f)(x_i) \quad \text{for all } i \in I \text{ and some } (f,h) \in \overline{\mathcal{B}_L}(\mathbb{R}^n), \\
& g_* = 0, \\
& s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1 \dots N-1\}, \\
& h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1,
\end{aligned}
\tag{$\overline{\text{PEP}}$}
$$

$\iff$

$$
\begin{aligned}
\text{maximize } & f_N - f_* \\
\text{subject to } & f_i - f_j - G^{gx}_{ji} + G^{gx}_{jj} \geq 0, \\
& (Lh_i - f_i) - (Lh_j - f_j) - L(G^{sx}_{ji} - G^{sx}_{jj}) + G^{gx}_{ji} - G^{gx}_{jj} \geq 0 \quad \text{for } i,j \in I, \\
& G^{gg}_{**} = 0, \\
& G^{sx}_{i+1,j} = G^{sx}_{ij} - \lambda G^{gx}_{ij} \quad \text{for } i \in \{0 \dots N-1\}, j \in I, \\
& h_* - h_0 - G^{sx}_{0*} + G^{sx}_{00} = 1, \\
& G \succeq 0,
\end{aligned}
\tag{sdp-$\overline{\text{PEP}}$}
$$

(Question: is it sufficient for condition $s_{i+1} = s_i - \lambda g_i$ to be hold only on $\mathrm{Span}\{x_i\}_{i \in I}$?)

Tightness Guarantee

$$
\begin{aligned}
\text{maximize} \quad & f_N - f_* \\
\text{subject to} \quad & f_i - f_j - G^{gx}_{ji} + G^{gx}_{jj} > 0, \\
& (Lh_i - f_i) - (Lh_j - f_j) - L(G^{sx}_{ji} - G^{sx}_{jj}) + G^{gx}_{ji} - G^{gx}_{jj} > 0 \quad \text{for } i \neq j \in I, \\
& G^{gg}_{**} = 0, \\
& G^{sx}_{i+1,j} = G^{sx}_{ij} - \lambda G^{gx}_{ij} \quad \text{for } i \in \{0 \ldots N-1\}, j \in I, \\
& h_* - h_0 - G^{sx}_{0*} + G^{sx}_{00} = 1, \\
& G^{xx}_{ii} + G^{xx}_{jj} - 2G^{xx}_{ij} > 0 \quad \text{for } i \neq j \in I, \\
& G \succeq 0,
\end{aligned}
\tag{sdp-$\underline{\text{PEP}}$}
$$

$$
\begin{aligned}
\text{maximize} \quad & f_N - f_* \\
\text{subject to} \quad & f_i - f_j - G^{gx}_{ji} + G^{gx}_{jj} \geq 0, \\
& (Lh_i - f_i) - (Lh_j - f_j) - L(G^{sx}_{ji} - G^{sx}_{jj}) + G^{gx}_{ji} - G^{gx}_{jj} \geq 0 \quad \text{for } i, j \in I, \\
& G^{gg}_{**} = 0, \\
& G^{sx}_{i+1,j} = G^{sx}_{ij} - \lambda G^{gx}_{ij} \quad \text{for } i \in \{0 \ldots N-1\}, j \in I, \\
& h_* - h_0 - G^{sx}_{0*} + G^{sx}_{00} = 1, \\
& G \succeq 0,
\end{aligned}
\tag{sdp-$\overline{\text{PEP}}$}
$$

Only difference is the strictness of inequalities.

By topological argument, two problems have the same optimum.

Theorem 4 *The value of the performance estimation problem (PEP) for NoLips is equal to the value of the nonsmooth relaxation ($\overline{\text{PEP}}$), which can be computed by solving the semidefinite program (sdp-$\overline{\text{PEP}}$).*

Numerical Result

Table 1 Numerical value of the performance estimation problem (PEP) with $\lambda = 1$, $L = 1$. *Rel. error* denotes the relative error between val(PEP) and the theoretical bound of $1/N$ given by Theorem 1. *Primal feasibility* corresponds to the maximal absolute value of constraint violation returned by the MOSEK solver.

N	val(PEP)	Rel. error	Primal feasibility
1	1.000	1.8e-11	4.3e-10
2	0.500	1.8e-8	2.8e-9
3	0.333	1.8e-8	2.8e-9
4	0.250	4.9e-8	2.3e-8
5	0.200	1.8e-10	6.4e-11
10	0.100	6.4e-11	1.3e-11
20	0.050	1.1e-8	1.9e-10
50	0.020	6.5e-6	5.0e-7
100	0.01	7.2e-5	1.6e-6

- For any $\lambda \in (0, 1/L]$, $\mathrm{val}(PEP)$ is equal to $1/(\lambda N)$.
- For any $\lambda > 1/L$, $\mathrm{val}(PEP) = \infty$.

Unlike Euclidean case (where $\lambda = 2/L$ can still converge)

Analytical Upper Bound

- convexity of f, between u and x_i $(i = 0, \dots, k)$ with weights $\gamma_{*,i} = \frac{1}{k}$:

$$f(u) \geq f(x_i) + \langle \nabla f(x_i), u - x_i \rangle,$$

- convexity of f, between x_i and x_{i+1} $(i = 0, \dots, k-1)$ with weights $\gamma_{i,i+1} = \frac{i}{k}$:

$$f(x_i) \geq f(x_{i+1}) + \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle,$$

- convexity of $\frac{1}{\lambda} h - f$, between u and x_k with weight $\mu_{*,k} = \frac{1}{k}$:

$$\frac{1}{\lambda} h(u) - f(u) \geq \frac{1}{\lambda} h(x_k) - f(x_k) + \langle \frac{1}{\lambda} \nabla h(x_k) - \nabla f(x_k), u - x_k \rangle,$$

- convexity of $\frac{1}{\lambda} h - f$, between x_{i+1} and x_i $(i = 0, \dots, k-1)$ with weight $\mu_{i+1,i} = \frac{i+1}{k}$

$$\frac{1}{\lambda} h(x_{i+1}) - f(x_{i+1}) \geq \frac{1}{\lambda} h(x_i) - f(x_i) + \langle \frac{1}{\lambda} \nabla h(x_i) - \nabla f(x_i), x_{i+1} - x_i \rangle,$$

- convexity of $\frac{1}{\lambda} h - f$, between x_i and x_{i+1} $(i = 0, \dots, k-1)$ with weight $\mu_{i,i+1} = \frac{i}{k}$

$$\frac{1}{\lambda} h(x_i) - f(x_i) \geq \frac{1}{\lambda} h(x_{i+1}) - f(x_{i+1}) + \langle \frac{1}{\lambda} \nabla h(x_{i+1}) - \nabla f(x_{i+1}), x_i - x_{i+1} \rangle.$$

$\Longrightarrow$

$$f(x_N) - f(u) \leq \frac{D_h(x_0, u)}{\lambda N}.$$

Upper Bound under Other Criteria

– convexity of f, between x_* and x_i $(i = 0, \ldots, k)$ with weights $\gamma_{*,i} = \frac{2\lambda}{k(k-1)}$:

$$f(x_*) \geq f(x_i) + \langle \nabla f(x_i), x_* - x_i \rangle,$$

– optimality of x_* for each x_k with weight $\gamma_{k,*} = \frac{2\lambda}{k-1}$:

$$f(x_k) \geq f(x_*),$$

– convexity of $\frac{1}{\lambda}h - f$, between x_* and x_k with weight $\mu_{*,k} = \frac{2\lambda}{k(k-1)}$:

$$\tfrac{1}{\lambda}h(x_*) - f(x_*) \geq \tfrac{1}{\lambda}h(x_k) - f(x_k) + \langle \tfrac{1}{\lambda}\nabla h(x_k) - \nabla f(x_k), x_* - x_k \rangle,$$

– convexity of $\frac{1}{\lambda}h - f$, between x_{i+1} and x_i $(i = 0, \ldots, k-1)$ with weight $\mu_{i+1,i} = \frac{2\lambda(i+1)}{k(k-1)}$

$$\tfrac{1}{\lambda}h(x_{i+1}) - f(x_{i+1}) \geq \tfrac{1}{\lambda}h(x_i) - f(x_i) + \langle \tfrac{1}{\lambda}\nabla h(x_i) - \nabla f(x_i), x_{i+1} - x_i \rangle,$$

– definition of smallest residual among the iterates $(i = 1, \ldots, k)$ with weights $\tau_i = \frac{2(i-1)}{k(k-1)}$:

$$h(x_{i-1}) - h(x_i) - \langle \nabla h(x_i), x_{i-1} - x_i \rangle \geq \min_{1 \leq j \leq k} \{D_h(x_{j-1}, x_j)\}.$$

$\Longrightarrow$

$$\min_{1 \leq i \leq N} D_h(x_{i-1}, x_i) \leq \frac{2D_h(x_0, x_*)}{N(N-1)}.$$

Table of Contents

An acceleration scheme [Auslender and Teboulle, 2006]

Algorithm 2 Improved Interior Gradient Algorithm (IGA) [1]

Input: Functions f, h, initial point $x_0 \in \operatorname{int}\operatorname{dom} h$, step size λ.

Set $z_0 = x_0$ and $t_0 = 1$.

for k = 0,1,... **do**

$y_k = (1 - \frac{1}{t_k})x_k + \frac{1}{t_k} z_k$

$z_{k+1} = \operatorname{argmin}\{\langle \nabla f(y_k), u - y_k \rangle + \frac{1}{t_k \lambda} D_h(u, z_k) \,|\, u \in \mathbb{R}^n\}$

$x_{k+1} = (1 - \frac{1}{t_k})x_k + \frac{1}{t_k} z_{k+1}$

$t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2.$

end for

Requires extra assumptions that f is L-smooth and h is σ-strongly convex. Theoretical bound:

$$f(x_N) - f_* \leq \frac{4\tilde{L}}{\sigma N^2} \left(D_h(x_*, x_0) + f(x_0) - f_*\right). \tag{23}$$

IGA in Relatively-smooth Case

Algorithm 2 Improved Interior Gradient Algorithm (IGA) [1]

Input: Functions f, h, initial point $x_0 \in \operatorname{int} \operatorname{dom} h$, step size λ.

Set $z_0 = x_0$ and $t_0 = 1$.

for k = 0,1,... **do**

$y_k = (1 - \frac{1}{t_k})x_k + \frac{1}{t_k} z_k$

$z_{k+1} = \operatorname{argmin} \{\langle \nabla f(y_k), u - y_k \rangle + \frac{1}{t_k \lambda} D_h(u, z_k) \mid u \in \mathbb{R}^n\}$

$x_{k+1} = (1 - \frac{1}{t_k})x_k + \frac{1}{t_k} z_{k+1}$

$t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2.$

end for

However, this method does not fit in the more general *relatively-smooth* case.

PEP value is unbounded!

Numerical result shows that IGA's corresponding PEP is unbounded for any sequence $\{t_k\}$ such that $t_{k_0} > 1$ for some k_0.

Towards Better Condition

The work of [Dragomir et al., 2019] shows that in the relatively-smooth case, $O(N^{-1})$ is not improvable. This condition is too loose.
On the other hand, L-smooth is too strong for functions having, like, $\log(\cdot)$ terms.

Towards Better Condition

The work of [Dragomir et al., 2019] shows that in the relatively-smooth case, $O(N^{-1})$ is not improvable. This condition is too loose.
On the other hand, L-smooth is too strong for functions having, like, $\log(\cdot)$ terms.
We need to find other conditions that can be accelerated.

- triangle scaling property: [Hanzely et al., 2021] (not yet)
- Holder continuous gradient: [Nesterov, 2015]

Bibliography I

Auslender, A. and Teboulle, M. (2006).
Interior gradient and proximal methods for convex and conic optimization.
SIAM J. Optim., 16:697–725.

Bauschke, H. H., Bolte, J., and Teboulle, M. (2017).
A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications.
Math. Oper. Res., 42:330–348.

Dragomir, R., Taylor, A. B., d'Aspremont, A., and Bolte, J. (2019).
Optimal complexity and certification of bregman first-order methods.
ArXiv, abs/1911.08510.

Bibliography II

Hanzely, F., Richtárik, P., and Xiao, L. (2021).
Accelerated bregman proximal gradient methods for relatively smooth convex optimization.
Comput. Optim. Appl., 79:405–440.

Nesterov, Y. (2015).
Universal gradient methods for convex optimization problems.
Mathematical Programming, 152:381–404.

Teboulle, M. (2018).
A simplified view of first order methods for optimization.
Mathematical Programming, 170:67–96.