Bregman Method from PEP Perspective

Zhenghao Xu

Zhejiang University

October 17, 2021

Table of Contents

Progress

Finished:

- evidence of improvement (numerical)

In progress:

- optimal fixed step size (dual matrix)
- valid one-step analysis
- toolbox for explicit dual matrix (in Python)

Table of Contents

Formulation

Problem:

$$\min_{x \in \mathbb{R}^d} \quad F(x) \triangleq f(x) + \phi(x) \tag{P}$$

Assumptions:

- f is *convex*, *proper*, *lsc* and continuously differentiable.
- ϕ is *convex*, *proper*, *lsc* (possibly nonsmooth).
- h is a Legendre kernel function.
- f is L-smooth relative to h, i.e., $Lh - f$ convex.[Bauschke et al., 2017]
- f, h more restricted.
- bounded initial distance and well-posedness of problem and method.

Formulation

Problem:

$$\min_{x \in \mathbb{R}^d} \quad F(x) \triangleq f(x) + \phi(x) \tag{P}$$

Method: Bregman Proximal Gradient

$$x_{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \phi(x) + LD_h(x, x_k) \} \tag{BPG}$$

Case 1: h - SC & Smooth

Fix $\phi(x) \equiv 0$. Assume: h is σ_h-strongly convex and L_h-smooth. Through PEP we obtain following result:

$$(f_1 - f_*) + \frac{L \left\| \nabla h(x_1) - \nabla h(x_*) - \sqrt{L_h \sigma_h}(x_1 - x_*) \right\|^2}{2(L_h - \sigma_h)}$$

$$\leq (1 - \frac{1}{\sqrt{L_h/\sigma_h} + 1}) L D_h(x_*, x_0).$$

Upper bound can be achieved when $\nabla h(x_1) - \nabla h(x_*) = \sqrt{L_h \sigma_h}(x_1 - x_*)$.

$$f_1 - f_* \leq (1 - \frac{1}{\sqrt{L_h/\sigma_h} + 1}) L D_h(x_*, x_0).$$

Numerically verified! Worst case (right hand side) can be reached.

Case 1: h - SC & Smooth

How is it derived?

- PEP gives
  - convexity of f between x_* and x_0: $1-c$;
  - S & SC of h between x_* and x_1: L;
  - convexity of $Lh - f$ between x_1 and x_0: $1-c$;
  - convexity of $Lh - f$ between x_1 and x_*: c.
- Summing up inequalities and we have

$$
\begin{aligned}
&(f_1 - f_*) + \frac{L\left\|\nabla h(x_1) - \nabla h(x_*) - \sqrt{L_h\sigma_h}(x_1 - x_*)\right\|^2}{2(L_h - \sigma_h)} \\
&- \left(\frac{L(cL_h + (1-c)\sigma_h - \sqrt{L_h\sigma_h})}{L_h - \sigma_h}\right)(D_h(x_*, x_1) + D_h(x_1, x_*)) \\
&\quad \leq (1-c)LD_h(x_*, x_0).
\end{aligned}
$$

Derivation

$$
\begin{aligned}
&(f_1 - f_*) + \frac{L \left\| \nabla h(x_1) - \nabla h(x_*) - \sqrt{L_h \sigma_h}(x_1 - x_*) \right\|^2}{2(L_h - \sigma_h)} \\
&- \left(\frac{L(cL_h + (1-c)\sigma_h - \sqrt{L_h \sigma_h})}{L_h - \sigma_h} \right) (D_h(x_*, x_1) + D_h(x_1, x_*)) \\
&\quad \leq (1-c) L D_h(x_*, x_0).
\end{aligned}
$$

To eliminate the negative term on LHS, solve

$$
\frac{L(cL_h + (1-c)\sigma_h - \sqrt{L_h \sigma_h})}{L_h - \sigma_h} = 0,
$$

which gives

$$
c = \frac{1}{\sqrt{L_h/\sigma_h} + 1}.
$$

Attempt 0

So we hope that given $\{c_i\}_{i=1}^m$ and corresponding inequalities from N step PEP, we are able to establish some kind of result in the form

$$\begin{aligned} & f_N - f_* + \underbrace{\sum \|\dots\|^2}_{I} - \underbrace{\sum E_i(c_1,\dots,c_m)(D_h(\dots))}_{II} \\ & \leq C(c_1,\dots,c_m)\frac{LD_h(x_*,x_0)}{N}, \end{aligned}$$

so that by solving equations $\{E_i(c_1,\dots,c_m)=0\}$ we get the $\{c_i\}$ and also the upper bound.

But seems failed!

- might have multiple ways to express I and II.
- attempted ways gives unimproved result. $(C(c_1,\dots,c_m)=1)$

Attempt 1

Attempt 1: generalize the one-step worst case
$(\nabla h(x_1) - \nabla h(x_*) = \sqrt{L_h \sigma_h}(x_1 - x_*))$ to every step.

Theorecitally it provides a lower bound on the upper bound (if feasible).

Not plausible! - numerical result with constraints

$$\begin{cases} \nabla h(x_i) - \nabla h(x_*) = \sqrt{L_h \sigma_h}(x_i - x_*), & \forall i = 1, \ldots, N, \\ \nabla h(x_i) - \nabla h(x_{i-1}) = \sqrt{L_h \sigma_h}(x_i - x_{i-1}), & \forall i = 1, \ldots, N, \end{cases}$$

is near to optimal.
But gap still exists (about 1e-2, relatively significant).

Attempt 2

Attempt 2: find potential-like inequality in the form of

$$f(x_{k+1}) - f_* \le a_k(f(x_k) - f_*) + b_k D_h(x_*, x_k).$$

or equivalently,

$$f(x_1) - f_* \le a(f(x_0) - f_*) + bD_h(x_*, x_0).$$

The bounded initial divergence constraint in PEP is replaced by

$$f(x_0) - f_* + \frac{b}{a} D_h(x_*, x_0) \le R.$$

Uncertain! - there's always terms like $D_h(x_0, x_1)$ and can't get the worst case solution like in $N = 1$ case.

Attempt 3

Attempt 3: add the symmetric coefficient constraints on D_h. That is,

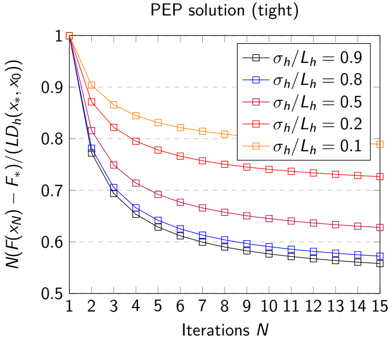
$$D_h(x, y) \leq \alpha D_h(y, x).$$

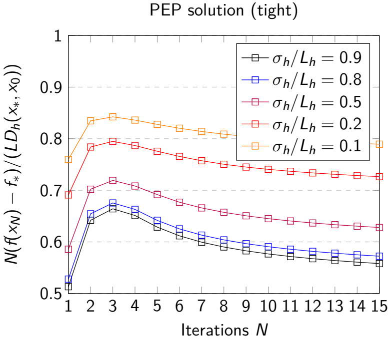
Uncertain!

- when α is large, no effect on result;
- when α is close to 1, it dominates the result, regardless of L_h/σ_h, the curve is similar.

Numerical Observation

Compare proximal case ($\phi \not\equiv 0$) and smooth case ($\phi \equiv 0$)





So far we only proved the smooth case when $N = 1$.

BPG and BG does not share the same behavior at the beginning.

Numerical Observation

Some results related to PEP formulation are weird.

Supposedly, if f is L-smooth relative to h, and h is L_h-smooth, then f should be LL_h-smooth.
But adding LL_h-smooth interpolation condition directly on constraints of f gives better (smaller) numerical result of PEP.

Theoretically the combination of former two conditions are stronger, but seemingly weaker direct interpolation condition is more "restrictive".

Table of Contents

Acceleration

Acceleration in [Lin et al., 2020, Lin et al., 2019] is not that clear.

Especially the relationship with estimate sequence? (Lack of intuition.)

Bibliography I

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.

Lin, T., Ho, N., Chen, X., Cuturi, M., and Jordan, M. I. (2020).
Fixed-support wasserstein barycenters: Computational hardness and fast algorithm.
arXiv: Computational Complexity.

Lin, T., Ho, N., and Jordan, M. I. (2019).
On the efficiency of sinkhorn and greenkhorn and their acceleration for optimal transport.