More about Proximal Operator

Jingchang Liu

January 3, 2019

University of Science and Technology of China

Table of Contents

Introductions

Problems

Formulation

$$\min_{x\in\mathbb{R}^d} f(x) := g(x) + h(x),$$

where g and h are convex functions but only g is smooth.

Example: $\ell_1 - \textit{regularized least squares}$

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1.$$

Proximal gradient

$$x^{k+1} = \text{prox}_h^{\gamma}(x^k - \gamma\nabla g(x^k)),$$

where $\text{prox}_h^{\gamma}(y) = \arg\min_{x\in\mathbb{R}^d} \frac{1}{2\gamma}\|x - y\|^2 + h(x)$.

Algorithms

Accelerated proximal gradient

$$x^{k+1} = \mathrm{prox}_h^{\gamma}(y^k - \gamma \nabla g(y^k)),$$

where $y^k = x^k + \beta^k(x^k - x^{k-1})$, and the sequence (β^k) is chosen appropriately.

Convergence rates (convex, smooth)

- PG: $\mathcal{O}(1/k)$.
- Accelerated PG: $\mathcal{O}(1/k^2)$.
- Strongly convex: linear rate.

Cases

Exact proximal

- $h(x) = \gamma\|x\|_1$, soft-thresholding

$$\text{prox}_h^{\gamma}(y) = \text{sign}(y) \times \max\{|y| - \gamma, 0\}.$$

Inexact proximal

- Overlapping group lasso

$$h(w) = \lambda \sum_{k=1}^{K} \|w_{g_k}\|,$$

where g_k is a group (subset) of variables.

- Graph-guide lasso

$$h(w) = \lambda \sum_{\{k_1, k_2\} \in E} |x_{k_1} - x_{k_2}|,$$

where E is the set of edges for the graph defined on the d variates

Inexact PG

Inexact PG

Iterations

$$x^{k+1} = \mathrm{prox}_h^{\gamma}(x^k - \gamma\{\nabla g(x^k) + e^k\}),$$

where e^k is the error in the calculation of the gradient, and the proximal operator is solved inexactly:

$$\frac{1}{2\gamma}\|x^k - y\|^2 + h(x^k) \leq \epsilon^k + \min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2\gamma}\|x - y\|^2 + h(x) \right\}.$$

Propositions 1

Basic PG, convexity
For all $k \geq 1$, we have

$$f\left(\frac{1}{k}\sum_{i=1}^{k} x^i\right) - f(x^*) \leq \frac{L}{2k}\left(\|x^0 - x^*\| + 2A_k + \sqrt{2B_k}\right)^2,$$

with $A_k = \sum_{i=1}^{k}\left(\frac{\|e_i\|}{L} + \sqrt{\frac{2\epsilon_i}{L}}\right)$, $B_k = \sum_{i=1}^{k} \frac{\epsilon_i}{L}$.

Conclusion
$\mathcal{O}(1/k)$ convergence rate still holds when both $(\|e_k\|)$ and $(\|\sqrt{\epsilon_k}\|)$ are summable.

Proposition 2

Accelerated PG, convexity
$y^k = x^k + \frac{k-1}{k+2}(x^k - x^{k-1})$, for all $k \geq 1$, we have

$$f(x^k) - f(x^*) \leq \frac{2L}{(k+1)^2}\left(\|x^0 - x^*\|^2 + 2\tilde{A}_k + \sqrt{2\tilde{B}_k}\right)^2,$$

with $\tilde{A}_k = \sum_{i=1}^{k}\left(\frac{\|e_i\|}{L} + \sqrt{\frac{2\epsilon_i}{L}}\right)$, $\tilde{B}_k = \sum_{i=1}^{k}\frac{i^2\epsilon_i}{L}$.

Conclusion
$\mathcal{O}(1/k^2)$ convergence rate still holds when both $(k\|e_k\|)$ and $(k\|\sqrt{\epsilon_k}\|)$ are summable.

Proposition 3

Basic PG, strong convexity
For all $k \geq 1$, we have

$$\|x^k - x^*\| \leq (1-\gamma)^k (\|x^0 - x^*\| + \tilde{A}_k),$$

with $\bar{A}_k = \sum_{i=1}^{k} (1-\gamma)^{-i} \left(\frac{\|e_i\|}{L} + \sqrt{\frac{2\epsilon_i}{L}} \right)$.

Conclusion
We obtain a linear rate, provided that $\|e_k\|$ and $\sqrt{\epsilon_k}$ decrease linearly to 0.

Proposition 4

Accelerated PG, strong convexity
$y^k = x^k + \frac{k-1}{k+2}(x^k - x^{k-1})$, for all $k \geq 1$, we have

$$f(x^k) - f(x^*) \leq (1-\sqrt{\gamma})^k \left(\sqrt{2(f(x^0) - f(x^*))} + \hat{A}_k \sqrt{\frac{2}{\mu}} + \sqrt{\hat{B}_k} \right)^2,$$

with $\hat{A}_k = \sum_{i=1}^{k} \left(\|e_i\| + \sqrt{2L\epsilon_i} \right) (1-\sqrt{\gamma})^{-i/2}$, $\hat{B}_k = \sum_{i=1}^{k} \epsilon_i (1-\sqrt{\gamma})^{-i}$.

Conclusion
We obtain a linear rate, provided that $\|e_k\|^2$ and ϵ_k decrease linearly to 0.

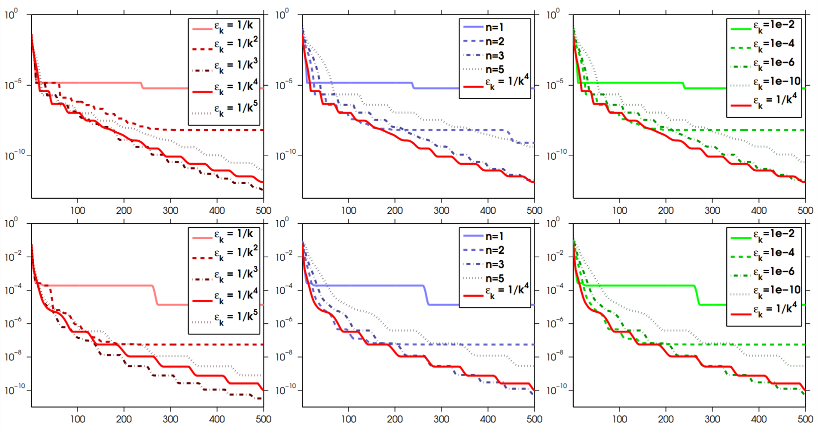

Figure 1: Objective function against number of proximal iterations for the accelerated proximalgradient method with different strategies for terminating the approximate proximity calculation. The top row is for the 9 Tumors data, the bottom row is for the Brain Tumor1 data.

Generation of proximal operator with a non-Euclidean distance measure

The proximal framework

Problem

$$\min_{x \in \mathbb{R}^d} \psi(x).$$

Proximal minimization

$$x^{k+1} = \text{prox}_{\psi}^{\gamma}(x^k), k = 0, 1, 2, \ldots$$

where

$$\text{prox}_{\psi}^{\gamma}(y) = \arg\min_{x \in \mathbb{R}^d} \frac{1}{2\gamma} \|x - y\|^2 + \psi(x).$$

Proximal iteration with non-Euclidean distance measure

$$x^{k+1} = \arg\min_{x} \left\{ \frac{1}{\lambda} D(x, x^k) + \psi(x) \right\}.$$

The Bregman distance

Definition: Legendre function
A function h, which is proper, lsc, strictly convex and essentially smooth will be called a Legendre function.

Definition: Bregman distance
Let h be a Legendre function. The Bregman distance associated to h, denoted by D_h is defined by

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle .$$

Properties

- Non-negative of D_h: h is convex if and only if $D_h(x, y) \geq 0$.
- Separability: if $h(x) = \sum_{j=1}^{n} h_j(x_j)$, then $D_h(x, y) = \sum_{j=1}^{n} D_{h_j}(x_j, y_j)$.
- Three Points Identity:

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

which is the generation of $\|z - x\|^2 - \|z - y\|^2 - \|x - y\|^2 = 2 < x - y, y - z >$.

Bregman distance generated from some convex functions

Table 1: Bregman divergences generated from some convex functions.

Domain	$\varphi(\mathbf{x})$	$d_\varphi(\mathbf{x},\mathbf{y})$	Divergence
$\mathbb{R}$	x^2	$(x-y)^2$	Squared loss
$\mathbb{R}_+$	$x\log x$	$x\log(\frac{x}{y})-(x-y)$	
$[0,1]$	$x\log x+(1-x)\log(1-x)$	$x\log(\frac{x}{y})+(1-x)\log(\frac{1-x}{1-y})$	Logistic loss [3]
$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y}-\log(\frac{x}{y})-1$	Itakura-Saito distance
$\mathbb{R}$	e^x	$e^x-e^y-(x-y)e^y$	
$\mathbb{R}^d$	$\lVert\mathbf{x}\rVert^2$	$\lVert\mathbf{x}-\mathbf{y}\rVert^2$	Squared Euclidean distance
$\mathbb{R}^d$	$\mathbf{x}^T A\mathbf{x}$	$(\mathbf{x}-\mathbf{y})^T A(\mathbf{x}-\mathbf{y})$	Mahalanobis distance [4]
d-Simplex	$\sum_{j=1}^d x_j\log_2 x_j$	$\sum_{j=1}^d x_j\log_2(\frac{x_j}{y_j})$	KL-divergence
$\mathbb{R}_+^d$	$\sum_{j=1}^d x_j\log x_j$	$\sum_{j=1}^d x_j\log(\frac{x_j}{y_j})-\sum_{j=1}^d(x_j-y_j)$	Generalized I-divergence

Figure 2: Bregman distance generated from some convex functions.

Lipschitz-like Condition (LC)

Definition

$$\exists L > 0 \textit{ with } Lh - g \text{ convex.}$$

NoLips Descent Lemma
Take $L > 0$, the following statements are equivalent:

1. $Lh - g$ is convex, i.e. LC hold.
2. $D_g(x, y) \leq LD_h(x, y) \leftrightarrow D_{Lh-g}(x, y) \geq 0$.

Proof
Simply follows from the gradient inequality for the convex function $Lh - g$, and the fact that $0 \leq D_{Lh-g}(x, y) = LD_h(x, y) - D_g(x, y)$.

Linear Rate

Let $\{x^k\}$ be the sequence generated by Bregman proximal iterations with $\lambda = L^{-1}$, and assume that $g - \sigma h$ is convex for some $\sigma > 0$. Then, for any $n \geq 0$,

$$\psi(x^{n+1}) - \psi^* \leq \left(1 - \frac{\sigma}{L}\right)^{n+1} L D_h(x^*, x^0).$$

Composition

Formulation

$$\min_{x \in \mathbb{R}^d} f(x) := g(x) + \psi(x),$$

where g and h are convex functions but only g is smooth.

Bregman proximal gradient iterations

$$\begin{aligned} x^{k+1} &= \operatorname{prox}_{\psi}^{\gamma}(x^k - \gamma \nabla g(x^k)) \\ &= \arg\min_x \left\{ g(x^k) + \left\langle \nabla g(x^k), x - x^k \right\rangle + \frac{1}{\lambda} D_h(x, x^k) + \psi(x) \right\}. \end{aligned}$$

Previous iterations

$$x^{k+1} = \arg\min_x \left\{ g(x^k) + \left\langle \nabla g(x^k), x - x^k \right\rangle + \frac{1}{2\lambda} \|x - x^k\|^2 + \psi(x) \right\}$$

Conclusions

Conclusions

- Algorithms with inexact proximal calculation may also work under some conditions.
- Bregman distance based proximal iterations need more efforts to be studied.

Q & A