Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization

Mark Schmidt, Nicolas Le Roux, Francis Bach

INRIA - SIERRA Project - Team

Laboratoire d'Informatique de l'École Normale Supérieure (CNRS/ENS/INRIA UMR 8548)

Motivation and Overview of Contribution

- We consider convex composite optimization problems:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x) := g(x) + h(x)$$

- g is smooth but h is non-smooth.
- g could be a *data-fitting term*, and h could be a *regularizer*.
- Convergence rates of different methods for this problem (for $\gamma < 1$):

Algorithm	Convex	Strongly-Convex
Sub-Gradient	$O(1/\sqrt{k})$	$O(1/k)$
Proximal-Gradient	$O(1/k)$	$O((1-\gamma)^k)$
Accelerated Proximal-Gradient	$O(1/k^2)$	$O((1-\sqrt{\gamma})^k)$

- Each proximal-gradient iteration requires the proximity operator:

$$\text{prox}_L(y) = \underset{x\in\mathbb{R}^d}{\arg\min} \frac{L}{2}\|x-y\|^2 + h(x)$$

Efficient to compute in some notable cases:
- *L1-regularization*
- *Disjoint Group L1-regularization*

- Can be approximated in other important scenarios:
  - *Total-variation regularization and graph-guided fused LASSO*
  - *Nuclear-norm regularization and other singular value penalties*
  - *Different formulations of overlapping group L1-regularization*
- We show that the convergence rates are preserved if the errors in the proximity and gradient decrease at appropriate rates.

Assumptions and Algorithm

- Assumptions:
  - g is convex and has L-Lipschitz continuous gradient:

$$||g'(x) - g'(y)|| \leq L||x-y||, \forall x, y$$

(if twice-differentiable, equivalent to $g''(x) \leq LI, \forall x$)

  - h is a lower semi-continuous proper convex function. (includes any real-valued convex function)
  - The function $f = g + h$ attains its minimum at some x^*.
- Algorithm:

$$x_k = \text{prox}_L[y_{k-1} - (1/L)(g'(y_{k-1}) + e_k)]$$

$$y_k = x_k + \beta_k(x_k - x_{k-1})$$

  - The basic proximal-gradient method sets $\beta_k = 0$.
  - The accelerated method chooses $\{\beta_k\}$ to improve the rate.
  - We have a gradient error e_k.
  - $\text{prox}_L[y]$ yields an ε_k-optimal solution of the proximal problem:

$$\frac{L}{2}\|x_k - y\|^2 + h(x_k) \leq \varepsilon_k + \min_{x\in\mathbb{R}^d}\left\{\frac{L}{2}||x-y||^2 + h(x)\right\}$$

- We also consider linear rates if g is μ-strongly convex, meaning $f(x) - \mu||x||^2$ is a convex function for $\mu > 0$. (if twice-differentiable equivalent to $g''(x) \geq \mu I, \forall x$)

Convergence Rates for Convex Objectives

Proposition 1. *If the sequences $\{\|e_k\|\}$ and $\{\sqrt{\varepsilon_k}\}$ are summable then the basic proximal-gradient method achieves*

$$f\left(\frac{1}{k}\sum_{i=1}^{k} x_i\right) - f(x_*) = O(1/k).$$

- For example, they could decrease as $O(1/k^{1+\delta})$ for $\delta > 0$.
- If they decrease as $O(1/k)$, then we get $O(\log^2 k/k)$.
- Bound also holds for the best iterate.

Proposition 2. *If the sequences $\{k\|e_k\|\}$ and $\{k\sqrt{\varepsilon_k}\}$ are summable then the accelerated proximal-gradient method achieves*

$$f(x_k) - f(x_*) = O(1/k^2),$$

with $\beta_k = (k-1)/(k+2)$.

- For example, they could decrease as $O(1/k^{2+\delta})$ for $\delta > 0$.
- If they decrease as $O(1/k^2)$, then we get $O(\log^2 k/k^2)$.
- The basic method does not necessarily dominate the accelerated method, because the accelerated method is more sensitive to errors.

Convergence Rates for Strongly Convex Objectives

Proposition 3. *If the sequences $\{||e_k||\}$ and $\{\sqrt{\epsilon_k}\}$ are in $O(Q^k)$ for $Q < (1-\gamma)$ then the basic proximal-gradient method achieves*

$$||x_k - x_*|| = O((1-\gamma)^k).$$

- If they converge with $Q > (1 - \gamma)$, the rate is $O(Q^k)$.
- If they converge with $Q = (1 - \gamma)$, the rate is $O(k(1-\gamma)^k)$.

Proposition 4. *If the sequences $\{||e_k||^2\}$ and $\{\epsilon_k\}$ are in $O(Q^k)$ for $Q < (1-\sqrt{\gamma})$ then the accelerated proximal-gradient method achieves*

$$f(x_k) - f(x_*) = O((1-\sqrt{\gamma})^k),$$

with $\beta_k = (1-\sqrt{\gamma})/(1+\sqrt{\gamma})$.

- Here, the accelerated method is less sensitive to errors. (according to our analysis)
- But, the accelerated method requires knowing μ.
- Gives a convergence rate on the iterates because:

$$\frac{\mu}{2}||x_k - x_*||^2 \leq f(x_k) - f(x_*)$$

Numerical Experiments on Structured Sparsity Problem

- We consider the CUR-like factorization to approximate a matrix W:

$$\min_X \tfrac{1}{2}\|W - WXW\|_F^2 + \lambda_{\text{row}}\sum_{i=1}^{n_r}\|X^i\|_p + \lambda_{\text{col}}\sum_{j=1}^{n_c}\|X_j\|_p$$

- For appropriate p, yields sparse rows *and* sparse columns.
- Previous work used $p = \infty$, since no known exact algorithm for $p = 2$.
- We consider $p = 2$, and use the proximal-Dykstra algorithm to compute an ε_k-optimal solution to the proximal problem based on duality gap.
- We compared ways to terminate the proximal-Dykstra algorithm.
- We use a basic line-search to compute L.

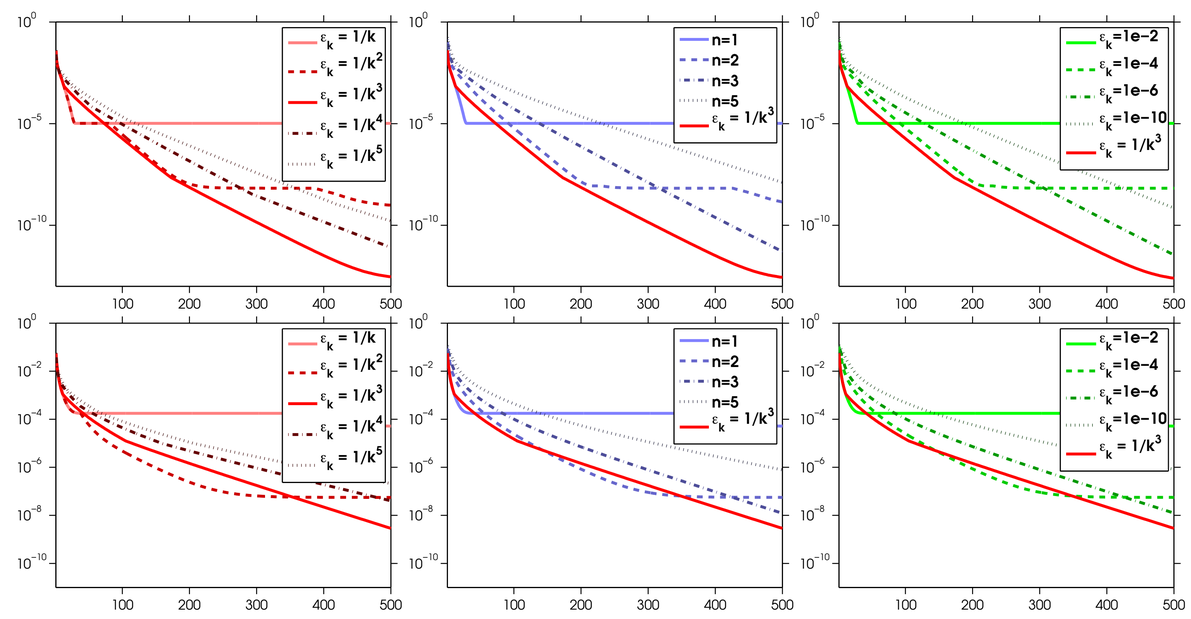

Figure 1: Objective function against number of proximal iterations for the proximal-gradient method with different strategies for terminating the approximate proximity calculation. The top row is for the *9_Tumors* data, the bottom row is for the *Brain_Tumor1* data.

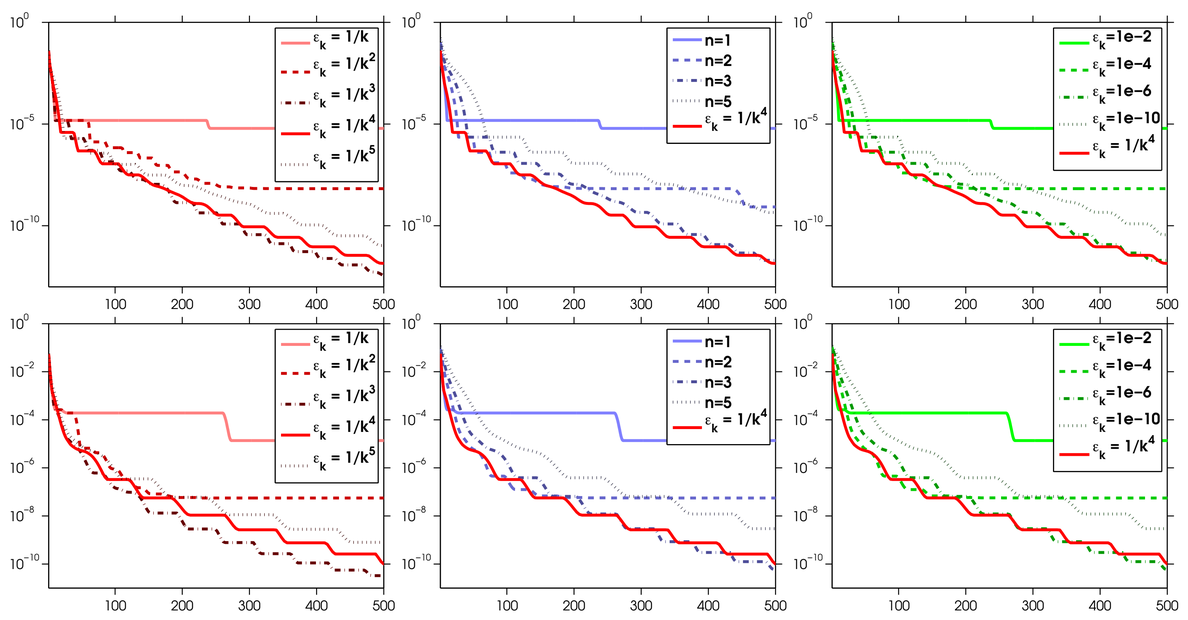

Figure 2: Objective function against number of proximal iterations for the *accelerated* proximal-gradient method with different strategies for terminating the approximate proximity calculation. The top row is for the *9_Tumors* data, the bottom row is for the *Brain_Tumor1* data.

Discussion

- Inexact proximal-gradient methods may be useful in other applications such as total-variation or nuclear-norm regularization.
- Our analysis allows errors in the gradient; this may be useful in undirected graphical models, kernel methods, and semi-definite programming.
- It would be interesting to extend our analysis to the case of an unknown L and μ.
- It would be interesting to explore methods for adaptively updating e_k and ε_k.
- It would be interesting to analyze the effects of errors on proximal-Newton methods.