An Inexact Accelreated Proximal Gradient Method for Large Scale Robust TV Variation Problems

Author 1 Name, Author 2 Name [*]

September 1, 2025

This paper is currently in draft mode. Check source to change options.

Abstract

I read a lot of papers on Catalyst, and restart. And it just dawned on me on how simple the ideas can be, and I had identified a specific type of problem where the idea has practical advantage. This is note is a plan of our upcoming practical paper, with numerical experiments, applications and sweet theories.

2010 Mathematics Subject Classification: Primary 47H05, 52A41, 90C25; Secondary 15A09, 26A51, 26B25, 26E60, 47H09, 47A63. **Keywords:**

1 Introduction

Let's make a type of algorithm that has the niche for applications with a state of the art performance.

{ass:smooth-nsmooth-sum}

Assumption 1.1 We assume the following about (F, f, g, L):

(i) $f : \mathbb{R}^n \to \mathbb{R}$ is a convex, L Lipschitz smooth function but doesn't support any easy implementation of its proximal operator.

[*]Subject type, Some Department of Some University, Location of the University, Country. E-mail: `author.namee@university.edu`.

(ii) $g : \mathbb{R}^n \to \mathbb{R}$ is convex, proper, and closed, and its proximal operator can be easily implemented, and easy to obtain some element ∂g at all points of the domain.
(iii) The over all objective has $F = f + g$.

Under this assumption, we denote the proximal gradient operator of $F = f + g$ as $T_B(x) = \text{prox}_{B^{-1}g}(x - B^{-1}\nabla f(x))$. Note that by definition it has also:

$$
\begin{aligned}
T_B(x) &= \text{prox}_{B^{-1}g}\left(x - B^{-1}\nabla f(x)\right) \\
&= \operatorname*{argmin}_z \left\{ g(z) + \langle \nabla f(x), z\rangle + \frac{B}{2}\|z - x\|^2 \right\}.
\end{aligned}
$$

Definition 1.2 (A measure of error from proximal gradient evaluations) *Let (F, f, g, L) satisfies Assumption 1.1. For all $x, z \in \mathbb{R}^n$, define S:*

$$
S_B(z|x) = \partial \left[z \mapsto g(z) + \langle \nabla f(x), z\rangle + \frac{B}{2}\|z - x\|^2 \right](z).
$$

Observe:

(i) $S_B(z|x) = \partial g(z) + \nabla f(x) + B(z - x)$,
(ii) $\mathbf{0} \in S_B(T(x)|x)$,
(iii) $(S_B(\cdot|x))^{-1}(\mathbf{0})$ is a singleton by strong convexity.

Let's assume inexact evaluation of $\tilde{x} \approx T_B(x)$, then the error measure is the set $S_B(\tilde{x}|x)$. Assuming that we have accurate information on $\nabla f(x)$, then $\forall w \in S_B(\tilde{x}|x)\ \exists \tilde{v} \in \partial g(\tilde{x})$.

$$
w = \tilde{v} + \nabla f(x) + B(\tilde{x} - x).
$$

We want to control w in the implementations of inexact accelerated proximal gradient algorithm.

2 Key ideas we need to get right

{def:inxt-pg} **Definition 2.1 (inexact proximal gradient)** *Let (F, f, g, L) satisfies Assumption 1.1. Let $\epsilon \ge 0, B \ge 0$. We Define for all $x \in \mathbb{R}^n$ the inexact proximal gradient operator $T_B^{(\epsilon)}(x)$ to be such that if $\tilde{x} \in T_B^{(\epsilon)}(x)$ then, $\exists w \in S_B(\tilde{x}|x)$: $\|w\| \le \epsilon\|\tilde{x} - x\|$.*

The algorithm we will design must produce iterates in a way that satisfies the inexact proximal gradient operator define above. The following theorem will characterize a key inequality for convergence claim.

{thm:inxt-pg-ineq} **Theorem 2.2 (inexact over regularized proximal gradient inequality)** *Let (F, f, g, L) satisfies Assumption 1.1. Take $T_B^{(\epsilon)}$ as given in Definition 2.1. Let $\epsilon \ge 0$. For all $x \in \mathbb{R}^n$, if $\exists B \ge 0$ such that $\tilde{x} \in T_{B+\epsilon}^{(\epsilon)}(x)$ and, $D_f(\tilde{x}, x) \le \frac{B}{2}\|\tilde{x} - x\|^2$. Then for all $z, x \in \mathbb{R}^n$ it has:*

$$0 \le F(z) - F(\tilde{x}) + \frac{B+\epsilon}{2}\|z - x\|^2 - \frac{B}{2}\|z - \tilde{x}\|^2.$$

Proof. By Definition 2.1, $T_{B+\epsilon}^{(\epsilon)}(x)$ minimizes a $h(z) = z \mapsto g(z) + \langle \nabla f(x), z\rangle + \frac{B+\epsilon}{2}\|x - z\|^2$ to produce $\tilde{x}$ so that $w \in S_{B+\epsilon}(\tilde{x}|x) = \partial h(x)$. h is $B + \epsilon$ strongly convex by convexity of g. Since $w \in \partial h(\tilde{x})$, it has subgradient inequality through strong convexity:

$$(\forall z \in \mathbb{R}^n)\ \frac{B+\epsilon}{2}\|z - \tilde{x}\|^2 \le h(z) - h(\tilde{x}) - \langle w, z - \tilde{x}\rangle.$$

This means for all $z \in \mathbb{R}^n$:

$$\begin{aligned}
&\frac{B+\epsilon}{2}\|\tilde{x} - z\|^2 \\
&\le g(z) + \langle \nabla f(x), z\rangle + \frac{B+\epsilon}{2}\|z - x\|^2 - \left(g(\tilde{x}) + \langle \nabla f(x), \tilde{x}\rangle + \frac{B+\epsilon}{2}\|\tilde{x} - x\|^2\right) \\
&\quad - \langle w, z - \tilde{x}\rangle \\
&= \left(g(z) - g(\tilde{x}) + \frac{B+\epsilon}{2}\|z - x\|^2 - \frac{B+\epsilon}{2}\|\tilde{x} - x\|^2 - \langle w, z - \tilde{x}\rangle\right) \\
&\quad + \langle \nabla f(x), z - x + x - \tilde{x}\rangle \\
&\underset{(1)}{=} \left(g(z) - g(\tilde{x}) + \frac{B+\epsilon}{2}\|z - x\|^2 - \frac{B+\epsilon}{2}\|\tilde{x} - x\|^2 - \langle w, z - \tilde{x}\rangle\right) \\
&\quad - D_f(z, x) + f(z) + D_f(\tilde{x}, x) - f(\tilde{x}) \\
&= (F(z) - F(\tilde{x}) - \langle w, z - \tilde{x}\rangle) + \left(\frac{B+\epsilon}{2}\|z - x\|^2 - D_f(z, x)\right) \\
&\quad + \left(D_f(\tilde{x}, x) - \frac{B+\epsilon}{2}\|\tilde{x} - x\|^2\right) \\
&\underset{(2)}{\le} \frac{B+\epsilon}{2}\|z - x\|^2 - D_f(z, x) + \left(\frac{B+\epsilon}{2}\|z - x\|^2 - \frac{\epsilon}{2}\|\tilde{x} - x\|^2\right) \\
&\le F(z) - F(\tilde{x}) + \|w\|\|z - \tilde{x}\| + \frac{B+\epsilon}{2}\|z - x\|^2 - \frac{\epsilon}{2}\|\tilde{x} - x\|^2 \\
&\underset{(3)}{\le} F(z) - F(\tilde{x}) + \epsilon\|x - \tilde{x}\|\|z - \tilde{x}\| + \frac{B+\epsilon}{2}\|z - x\|^2 - \frac{\epsilon}{2}\|\tilde{x} - x\|^2.
\end{aligned}$$

At (1), we used:

$$
\begin{aligned}
&\langle \nabla f(x), z - x\rangle - \langle \nabla f(x), \tilde{x} - x\rangle \\
&= -D_f(z, x) + f(z) - f(x) + D_f(\tilde{x}, x) - f(\tilde{x}) + f(x) \\
&= f(z) + f(\tilde{x}) - D_f(z, x) + D_f(\tilde{x}, x).
\end{aligned}
$$

At (2), we had f convex as the assumption, hence $D_f(z, x) \le 0$. We also had the assumption that B makes $D_f(\tilde{x}, x) \le \frac{B}{2}\|\tilde{x} - x\|^2$, this simplies the third term from the previous line into $-\frac{\epsilon}{2}\|x - \tilde{x}\|^2$. At (3), we applied the assumed inequality $\|w\| \le \epsilon\|x - \tilde{x}\|\|z - \tilde{x}\|$. Continuing:

$$
\begin{aligned}
0 &\le \left(F(z) - F(\tilde{x}) + \frac{B+\epsilon}{2}\|z - x\|^2 - \frac{B+\epsilon}{2}\|z - \tilde{x}\|^2\right) + \epsilon\|\tilde{x} - x\|\|z - \tilde{x}\| - \frac{\epsilon}{2}\|\tilde{x} - x\|^2 \\
&\underset{(4)}{\le} F(z) - F(\tilde{x}) + \frac{B+\epsilon}{2}\|z - x\|^2 - \frac{B}{2}\|z - \tilde{x}\|^2.
\end{aligned}
$$

At (4), we use some algebra:

$$
\begin{aligned}
&\epsilon\|\tilde{x} - x\|\|z - \tilde{x}\| - \frac{\epsilon}{2}\|\tilde{x} - x\|^2 \\
&= \epsilon\|\tilde{x} - x\|\|z - \tilde{x}\| - \frac{\epsilon}{2}\|\tilde{x} - x\|^2 - \frac{\epsilon}{2}\|z - \tilde{x}\|^2 + \frac{\epsilon}{2}\|z - \tilde{x}\|^2 \\
&= -\epsilon(\|x - \tilde{x}\| - \|z - \tilde{x}\|)^2 + \frac{\epsilon}{2}\|z - \tilde{x}\|^2 \\
&\le \frac{\epsilon}{2}\|z - \tilde{x}\|^2.
\end{aligned}
$$

■

2.1 The accelerated proximal gradient algorithm

{def:inxt-apg}

Definition 2.3 (accelerated inexact proximal gradient algorithm) *Let*

(i) $(\alpha_k)_{k\ge 0}$ *be a sequence in* $(0, 1]$.
(ii) *Let* $(B_k)_{k\ge 0}$ *be a non-negative sequence.*
(iii) *Let* (F, f, g, L) *be given by Assumption 1.1.*
(iv) *Let* $(\epsilon_k)_{k\ge 0}$ *be a non-negative sequence that is the error schedule.*

Initialize with any (x_{-1}, v_{-1}). *For these given parameters, an algorithm is a type of accelerated proximal gradient if it generates* $(y_k, x_k, v_k)_{k\ge 0}$ *such that for* $k \ge 0$*:*

$$
\begin{aligned}
&y_k = \alpha_k v_{k-1} + (1 - \alpha_k)x_{k-1}, \\
&x_k \in T^{(\epsilon_k)}_{B_k+\epsilon_k} y_k : D_f(x, \tilde{x}) \le (1/2)\|x - \tilde{x}\|^2, \\
&v_k = x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}).
\end{aligned}
$$

3 convergence rates results

{ass:apg-cnvg} We will now show that Algorithms satisfying Definition 2.3 has desirable convergence rate.

Assumption 3.1 (convergence assumptions) Let (F, f, g, L) satisfies Assumption 1.1 and in addition assume that F admits a set of non-empty minimizers X^+.

{lemma:inxt-apg-onestep} **Lemma 3.2 (inexact one step convergence claim)**
Let (F, f, g, L, X^+) satisfies Assumption 3.1. Suppose that an algorithm satisfies optimizes the given $F = f + g$ also satisfying Definition 2.3. Then for the generated iterates $(y_k, x_k, v_k)_{k\geq 0}$, it has for all $k \geq 1$:

$$\begin{aligned}&F(\bar{x}) - F(x_k) - \frac{B_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2\\ &\leq \max\left(1 - \alpha_k, \frac{\alpha_k(B_k + \epsilon_k)}{\alpha_{k-1}^2 B_{k-1}}\right)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 B_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2\right).\end{aligned}$$

Proof. Let $\bar{x} \in X^+$, making it a minimizer of F. Define $z_k := \alpha_k\bar{x} + (1 - \alpha_k)x_{k-1}$. It can be verified that:

{lemma:inxt-apg-onestep-a}

$$\begin{aligned}z_k - x_k &= \alpha_k(\bar{x} - v_k),\\ z_k - y_k &= \alpha_k(\bar{x} - v_{k-1}).\end{aligned} \tag{a}$$

Because from Definition 2.3 it has for all $k \geq 1$:

$$\begin{aligned}z_k - x_k &= \alpha_k\bar{x} + (1 - \alpha_k)x_{k-1} - x_k\\ &= \alpha_k\bar{x} + (x_{k-1} - x_k) - \alpha_k x_{k-1}\\ &= \alpha_k\bar{x} - \alpha_k v_k,\\ z_k - y_k &= \alpha_k\bar{x} + (1 - \alpha_k)x_{k-1} - y_k\\ &= \alpha_k\bar{x} - \alpha_k v_{k-1}.\end{aligned}$$

For all $k \geq 0$, apply Theorem 2.2 with $z = z_k, \tilde{x} = x_k, x = y_k, \epsilon = \epsilon_k, B = B_k$:

$$
\begin{aligned}
0 &\leq F(z_k) - F(x_k) + \frac{B_k + \epsilon_k}{2}\|z_k - y_k\|^2 - \frac{B_k}{2}\|z_k - x_k\|^2 \\
&\underset{(1)}{\leq} \alpha_k F(\bar{x}) + (1-\alpha_k)F(x_{k-1}) - F(x_k) + \frac{B_k + \epsilon_k}{2}\|z_k - y_k\|^2 - \frac{B_k}{2}\|z_k - x_k\|^2 \\
&\underset{(a)}{=} \alpha_k F(\bar{x}) + (1-\alpha_k)F(x_{k-1}) - F(x_k) \\
&\quad + \frac{(B_k + \epsilon_k)\alpha_k^2}{2}\|\bar{x} - v_{k-1}\|^2 - \frac{B_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2 \\
&= F(\bar{x}) - F(x_k) + (1-\alpha_k)(F(x_{k-1}) - F(\bar{x})) \\
&\quad + \frac{(B_k + \epsilon_k)\alpha_k^2}{2}\|\bar{x} - v_{k-1}\|^2 - \frac{B_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2 \\
&= F(\bar{x}) - F(x_k) - \frac{B_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2 \\
&\quad + (1-\alpha_k)(F(x_{k-1}) - F(\bar{x})) + \frac{(B_k + \epsilon_k)\alpha_k^2}{2}\|\bar{x} - v_{k-1}\|^2 \\
&= F(\bar{x}) - F(x_k) - \frac{B_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2 \\
&\quad + (1-\alpha_k)(F(x_{k-1}) - F(\bar{x})) + \frac{(B_k + \epsilon_k)\alpha_k^2}{\alpha_{k-1}^2 B_{k-1}}\frac{\alpha_{k-1}^2 B_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2 \\
&= F(\bar{x}) - F(x_k) - \frac{B_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2 \\
&\quad + \max\left(1-\alpha_k, \frac{(B_k + \epsilon_k)\alpha_k^2}{\alpha_{k-1}^2 B_{k-1}}\right)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 B_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2\right).
\end{aligned}
$$

At (1) we used convexity of f which is assumed and it makes $f(z_k) \leq \alpha_k F(\bar{x}) + (1 - \alpha_k)F(x_{k-1})$ because $\alpha_k \in (0, 1]$ from Definition 2.3. ■

As a prelude, to derive the convergence rate we unroll the recurrence relation proved in the above lemma. It remains to create convergence criterions of the error relative sequence ϵ_k such that the original optimal convergence rate of $\mathcal{O}(1/k^2)$ the sequence remains unaffected. Let the sequence $(B_k)_{k\geq 0}$ be given as in Definition 2.3. We suggest the following using another sequence ρ_k given by for all $k \geq 1$:

$$
\rho_k := \frac{B_k + \epsilon_k}{B_{k-1}}\frac{B_{k-1}}{B_k} = \frac{B_k + \epsilon_k}{B_k}
$$

This means the following:

$$
\begin{aligned}
\max\left(1-\alpha_k, \frac{\alpha_k^2(B_k+\epsilon_k)}{\alpha_{k-1}^2 B_{k-1}}\right) &= \max\left(1-\alpha_k, \rho_k\frac{B_k\alpha_k^2}{B_{k-1}\alpha_{k-1}^2}\right) \\
&\le \max(1,\rho_k)\max\left(1-\alpha_k, \frac{B_k\alpha_k^2}{B_{k-1}\alpha_{k-1}^2}\right).
\end{aligned}
$$

If we consider $\rho_k \le (1+2/k^2)$, it has the ability to make

$$
\begin{aligned}
\prod_{k=1}^{n}\max\left(1-\alpha_k, \frac{\alpha_k^2(B_k+\epsilon_k)}{\alpha_{k-1}^2 B_{k-1}}\right) &\le \prod_{k=1}^{n}\max(1,\rho_k)\prod_{i=1}^{n}\max\left(1-\alpha_k, \frac{B_k\alpha_k^2}{B_{k-1}\alpha_{k-1}^2}\right) \\
&\le \prod_{k=1}^{n}\left(1+\frac{2}{k^2}\right)\prod_{i=1}^{n}\max\left(1-\alpha_k, \frac{B_k\alpha_k^2}{B_{k-1}\alpha_{k-1}^2}\right) \\
&\le 2\prod_{i=1}^{n}\max\left(1-\alpha_k, \frac{B_k\alpha_k^2}{B_{k-1}\alpha_{k-1}^2}\right).
\end{aligned}
$$

Assuming no $B_k = 0$ then the error schedule $\rho_k \le (1+2/k^2)$ translates to

$$
\begin{aligned}
&\frac{B_k+\epsilon_k}{B_k} \le 1+\frac{2}{k^2} \\
\iff\ &\epsilon_k \le -B_k + B_k(1+2/k^2) \le \frac{2B}{k^2}.
\end{aligned}
$$

4 Motivations for applications

Optimization problem of the following type:

$$
\min_x \left\{f(Ax) + g(Dx)\right\},
$$

where

(i) Let $A \in \mathbb{R}^{m_1\times n}, D \in \mathbb{R}^{m_2\times n}$, where inverting $A^TA + \mu I$ exactly for some $\mu > 0$ is computationally intensive on computers.

(ii) $f : \mathbb{R}^{m_1} \to \mathbb{R}$ is L Lipschitz smooth and, convex but taking maybe prox_f is computationally intensive.

(iii) $g : \mathbb{R}^{m_2} \to \overline{\mathbb{R}}$ is convex, closed and proper.

Solving it using traditional method such as ADMM [1, chapter 6], or Chambolle Pock [2], would require the knowledge of $(A^TA + \mu I)^{-1}$, and f has prox that is simple to evaluate.

Inverting matrix exactly is never easy, especially the sparse and large matrix in image processing (The number of elements in the matrix scales by $\mathcal{O}(n^4)$ for a size $n \times n$ image for storage, and inverting it with bruteforce is (n^6)). People can do it a few time but it's not practical/simple to do it in every step of some iterative algorithms! Of course, we can use other smarter method such as Krylov Subspace method for block structured sparse matrices (or most recently, their randomized variants), but they will not be perfectly accurate. Our method however, it will have the advantage of being matrix inversion free.

Our proposed algorithm allows inexact evaluation of $\operatorname{prox}_{B^{-1}g\circ D}$. This proximal operator can be evaluated using popular algorithm such as Chambolle Pock, or any other efficient approximation algorithm and it will bypass the need of any matrix inversion involving D. Because look, the problem we solve for our algorithm is $\operatorname{prox}_{(B+\epsilon)^{-1}g\circ D}(x)$ which can be reformulated into:

$$\underset{y,z}{\operatorname{argmin}}\left\{g(y)+\frac{B+\epsilon}{2}\|z-x\|^2 : Az = y, y \in \mathbb{R}^{m_2}, z \in \mathbb{R}^n\right\}.$$

Solving this problem inexactly is not difficult. First order method such as Chambolle Pock will suffice, and it won't need expensive matrix inversion. And, taking the gradient of $\nabla f(Ax)$ doesn't need for $(A^TA+\mu I)^{-1}$. Our method is also compatible with enhancement such as restart for objective function satisfying the quadratic growth conditions for optimal linear convergence.

This is especially useful for problem where, A is massive and sparse without obvious structure. For example, a TV variational minimization approach for Robust MRI imaging that involves downsampling, blurring, or convolution [5, 4][3, Section 7.4]. Applications as such adds another matrix D, or additional linear constraints to improve the robustness of MRI image reconstruction. Other potential applications with such structural optimization objective include large regression problem where the penalization term is highly non-trivial.

Why we think this paper will fly? Because:

(i) The core theories part for the outer loop convergence described previously are quite simple and easy to understand for interested readers and, practitioners compared to previous results like in the Catalyst Frameworks or Inexact Accelerated Proximal Gradient Method.

(ii) I haven't found anyone that has good solutions in this type of problem which bypasses inverting D or A.

(iii) It will shine numerically if coupled with advanced approximation algorithm for evaluation the prox of $g \circ D$, for specific g relevant for applications.

(iv) Our theories give the flexibilities for whatever the practitioners want to do with $\operatorname{prox}_{B^{-1}g\circ D}$. They can specialize and improved based our theoretical frameworks.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, Foundations and Trends® in Machine Learning, 3 (2011), pp. 1–122. Publisher: Now Publishers, Inc.

[2] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging.* June 2010.

[3] ——, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319.

[4] S. H. Joshi, A. Marquina, S. J. Osher, I. Dinov, J. D. Van Horn, and A. W. Toga, *MRI resolution enhencement using total variation regularization*, Proceedings / IEEE International Symposium on Biomedical Imaging: from nano to macro. IEEE International Symposium on Biomedical Imaging, 2009 (2009), pp. 161–164.

[5] M. Zhang, M. Zhang, F. Zhang, A. Chaddad, and A. Evans, *Robust brain MR image compressive sensing via re-weighted total variation and sparse regression*, Magnetic Resonance Imaging, 85 (2022), pp. 271–286.