Resolving a Fundamental Challenge in Inexct Proximal Method: The Unknown Constant of the Error Bound

Author 1 Name, Author 2 Name *

September 11, 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 $Q \subseteq \mathbb{R}^n$, we use the notation: $\operatorname{dist}(x|Q) = \inf_{z\in Q} \|x - z\|$. When $Q \subseteq \mathbb{R}^n$ is closed, nonempty and convex, we denote the closest point projection to the set by $\Pi(x|Q)$. For any matrix A, we denote its kernel by $\ker A$, and its range $\operatorname{rng} A$.

Definte some matrix $A \in \mathbb{R}^{m\times n}$ and, let vector $b \in \mathbb{R}^m$ be such that $b \in \operatorname{rng} A$. Consider the following optimization problem:

{eqn:prblm-exmp}
$$\min_{x\in\mathbb{R}^n} \left\{ \lambda\|x\|_1 + \frac{1}{2}\operatorname{dist}(x|\ \{z : Az = b\})^2 \right\}. \tag{1.1}$$

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

This problem is not easy to solve because taking the gradient of the second function (denote $f(x) = \frac{1}{2}\operatorname{dist}(x|\{z : Ax = b\}))$ in (1.1) requires the left pseudo inverse of matrix A. Since the gradient is given by:

$$\nabla f(z) = z - A^{\dagger}(Az - b) = z - \Pi(z|\{x|Ax = b\}).$$

When A is sparse or large. Taking the gradient of the function is a fundamental challenge for numerical algorithms.

The difficulty doesn't stop here at all and, the next issue about error bound condition is worse. If we were to approximate $\nabla f(z)$ with $\tilde{\nabla} f(z)$ to minimize the error $\|\nabla f(z) - \tilde{\nabla} f(z)\|$ using some type of optimization algorithm that solves the projection problem approximately:

$$\tilde{z} \approx z^{+} = \Pi(z|\{x|Ax = b\}) = \underset{y}{\operatorname{argmin}} \left\{\frac{1}{2}\|y - z\|^2 : Ay = b\right\}.$$

This approach has a fundamental challenge because the approximation error is $\|\tilde{z} - z^{+}\|$. To estimate this quantity for inexact algorithm, in general would require some error bound conditions. In this case, let $\sigma_{\min}(A)$ be the minimal nonzero singular value of A, the error bound is

$$\sigma_{\min}(A)\|\tilde{z} - z^{+}\| \le \|A\tilde{z} - b\|.$$

Look, this error bound condition requires knowing $\sigma_{\min}(A)$ which is just as hard as looking for the inverse of A. A lot of the algorithm for estimating singular value are iterative method, or their specialized variants for sparse, or structured matrices. This is a fundamental challenge when applying inexact methods in general. To convince, consider changing the second function in the objective to $f(x) = (1/2)\operatorname{dist}(x|\{z : Ax \in \mathbb{R}^n_+\})^2$. In this case, the error bound condition is known as "Hoffman Error Bound", and lower bounding the constant is a combinatorics problem, hence, extremely difficult to obtain in practice.

Contributions of the paper (hopefully).

(i) We show that an accelerated proximal gradient method with inexact gradient evaluation can converge under a relative error conditions.
(ii) We show that we don't need to know the constant for the error bound condition and we can still get convergence for the algorithm.
(iii) We give outer loop complexity analysis for our algorithm, if the inner loop error bound condition exists, and asymptoptic convergence rate when it doesn't exist.

{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.

(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 $\tilde{x} \approx T_B(x)$ where ∇f is inexact. Assuming that we have the estimate $\tilde{\nabla} f(x)$ for $\nabla f(x)$, then $\exists v \in \partial g(\tilde{x})$.

$$
\begin{aligned}
& 0 = v + \tilde{\nabla} f(x) + B(\tilde{x} - x) \\
\iff & \nabla f(x) - \tilde{\nabla} f(x) = v + \nabla f(x) + B(\tilde{x} - x).
\end{aligned}
$$

This means $\nabla f(x) - \tilde{\nabla} f(x) \in S_B(\tilde{x}|x)$. 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* $k \ge 0$*:*

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

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 $z = z_k, \tilde{x} = x_k, x = y_k, \epsilon = \epsilon_k, B = B_k$:

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

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:

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

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

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

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

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

4 Motivations for applications

References