Convergence rates of accelerated proximal gradient algorithms under independent noise

Tao Sun[*] Roberto Barrio[†] Hao Jiang[‡] Lizhi Cheng[§]

October 24, 2017

Abstract

We consider an accelerated proximal gradient algorithm for the composite optimization with "independent errors" (errors little related with historical information) for solving linear inverse problems. We present a new inexact version of FISTA algorithm considering deterministic and stochastic noises. We prove some convergence rates of the algorithm and we connect it with the current existing catalyst framework for many algorithms in machine learning. We show that a catalyst can be regarded as a special case of the FISTA algorithm where the smooth part of the function vanishes. Our framework gives a more generic formulation that provides convergence results for the deterministic and stochastic noise cases and also to the catalyst framework. Some of our results provide simpler alternative analysis of some existing results in literature, but they also extend the results to more generic situations.

1 Introduction

Linear inverse problems have received a lot of attention last few years as they are widely applied to many areas such as signal processing [2, 31], imaging sciences (image deblurring problem [4, 21]) and computational statistics [18], to name a few. Inverse problems involve estimating data or parameters from incomplete or noisy observations, sometimes due to physical limitations of the measurement devices. Therefore, solutions to inverse problems are non-unique, and so, we must exploit the underlying structure of the desired solution set to pose a suitable approximate solution.

Basically, a linear inverse problem is usually described as

$$Ax = b + w, \tag{1.1}$$

where $A \in \mathbb{R}^{M\times N}$ is known, $b \in \mathbb{R}^N$ is observed measured data, and $w \in \mathbb{R}^N$ is an unknown additive noise vector (in some situations, like in stability theory analysis, it can be considered as a perturbation vector).

In this paper we are interested in the study of methods that permits to get the unknown vector $x \in \mathbb{R}^M$. If A is nonsingular, an intuitive method is just using least squares (LS) approach [5], i.e., solving the following data error minimization problem

$$x_{LS} \in \arg\min_{x} \|Ax - b\|_2^2. \tag{1.2}$$

[*] Department of Mathematics and System Science, National University of Defense Technology, Changsha, 410073, Hunan, China. Email: nudttaosun@gmail.com;nudtsuntao@163.com

[†] Departamento de Matemática Aplicada and IUMA, University of Zaragoza, E-50009 Zaragoza, Spain, Email: rbarrio@unizar.es

[‡] College of Computer, National University of Defense Technology, Changsha, Hunan, China, 410073, Email: haojiang@nudt.edu.cn

[§] Department of Mathematics and System Science& The State Key Laboratory for High Performance Computation, National University of Defense Technology, Changsha, Hunan, China, 410073, Email: clzcheng@nudt.edu.cn

However, in many applications, A is unfortunately singular. For example, in compressed sensing [11], A is the sensing matrix whose number of rows is smaller than the number of columns. And in the image blurring problem, A presents the blurring operator, which is also singular. It means that in this situation the LS solution x_{LS} will be infinitely undetermined and LS method will fail in finding the solution. In other cases A is nonsingular but close to it (ill-conditioned), and again, x_{LS}, although unique, has a huge error norm and is thus meaningless.

In general, it is hard to get the solution x if A is singular (or highly ill-conditioned). But the good news is the unknown vector is not "totally unknown" and we can use the underlying structure. In fact, some qualitative information is usually detected in these problems, like sparsity.

Another way to deal with ill-conditioned or singular problems is by means of regularization methods to stabilize the solution. The basic idea of regularization is to replace the original problem with a well-conditioned problem whose solution approximates the original solution. Recently a method that is commonly used (specially in signal processing literature) is the ℓ_1 regularization problem [8, 14]

$$\min_x \left\{ \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \right\}, \tag{1.3}$$

where $\|\cdot\|_1$ stands for the l_1 norm. This model is extended by the nuclear norm for solving the matrix completion problem [6]. The success of the ℓ_1 or nuclear norm regularization has attracted increasing attentions in optimization and inverse problems community for solving the following general model

$$\min_x \{F(x) = f(x) + g(x)\}, \tag{1.4}$$

where f is smooth and convex, and g is convex.

Some of the state-of-art methods for solving problem (1.4) are the Iterative Shrinkage-Thresholding Algorithms (ISTA) [7, 8, 13, 15, 16]. The convergence rate of ISTA is well known as $\mathcal{O}(1/k)$ where k is the iteration counter. In the nice paper [4], the authors propose a Fast Iterative Shrinkage-Thresholding Algorithm (FISTA), which improved the convergence rate from $\mathcal{O}(1/k)$ to $\mathcal{O}(1/k^2)$ while keeping the simplicity of ISTA. If g vanishes, FISTA will reduce to the Nesterov gradient method [23]. Plenty of variants have also been developed in recent years [3, 4, 30, 32]. The part of the algorithm that demands a higher computational cost is minimizing the convex function $Q_r(x, y^k)$ (see Eqn (2.1) below). For the ℓ_1 regularized problem, the minimizer usually enjoys a closed form; however, for many other situations, the minimizing may be expensive and inexact. This paper focuses on an inexact version of FISTA

The contribution of the paper is three-fold: –1– We consider a new inexact version of FISTA. Such algorithm is modeled basing on practical algorithms when the function $Q_r(x, y^k)$ is hard to minimize. We prove the convergence rates of this inexact algorithm under different cases; –2– We consider the expectation form of the error which has not been mentioned before in literature; –3– We build a connection between this inexact algorithm with a catalyst algorithm for various first-order optimization algorithms. Based on our results, the catalyst algorithm is still convergent under a much larger noise.

This paper is organized as follows: In Section 2, we present the new algorithm and the basic notations and preliminaries; in Section 3, the convergence rates are proved under deterministic and stochastic errors; Section 4 connects our results with a general catalyst method; some numerical tests are shown in Section 5; and finally, Section 6 gives some conclusions.

2 FISTA algorithm and preliminaries

The Fast Iterative Shrinkage-Thresholding Algorithms (FISTA) [3, 4, 30, 32] for solving (1.4) are based on the minimization of the functional

$$Q_r(z, y^k) := \langle z - y^k, \nabla f(z) \rangle + \frac{r}{2}\|z - y^k\|_2^2 + g(z). \tag{2.1}$$

Most of the differences between different inexact FISTA methods are in the "inexact ways" used, i.e.,

$$x^{k+1} \approx \arg\min_x Q_r(x, y^k).$$

In paper [28], the authors introduced "two errors": a process error $\bar{e}^k$ (is a vector) and the function values error, i.e., $Q_r(x^{k+1}, y^k) - \min Q_r(x, y^k) \leq \varepsilon_k$ (is a number). They proved the convergence for an accelerated algorithm (different from FISTA) under several assumptions on both $\bar{e}^k$ and ε_k. Paper [24] is devoted to $x^{k+1} = \arg\min_x Q_r(x, y^k) + e^k$ (we stress that $\bar{e}^k$ in [28] is different with e^k in [24]). The convergence is proved under several assumptions, which are related with historical information on the noise e^k. This is also to say that the noise is *not independent*. In paper [17], a slightly different criterion $F(x^{k+1}) \leq \arg\min_x Q_r(x, y^k) + \varepsilon_k$ is posed. Another approach on the inexactness is using the ϵ-subdifferential [33]. And in [10], an inexact first order oracle is also proposed for minimizing $Q_r(x, y^k)$ approximately.

In this paper, we consider the following inexact way (also used in [28])

$$Q_r(x^{k+1}, y^k) - \min Q_r(x, y^k) \leq \varepsilon_k$$

for FISTA. We remark that [28] is devoted to an accelerated forward-backward algorithm but not FISTA, although we use the same function values error. We consider such scheme because the inexactness usually comes from the fact the subproblem $\min_x Q_r(x, y^k)$ does not enjoy a closed form solution, i.e., cannot be solved exactly in a very few computations. Thus, we may use other algorithms to minimize $Q_r(x, y^k)$. Besides, most of the convergence rates in literature of these algorithms are built on the function values, and thus, we may obtain

$$Q_r(x^{k+1}, y^k) - \min Q_r(x, y^k) \leq \mathcal{O}(j_k^{-\gamma}), \tag{2.2}$$

where j_k are the steps for minimizing $Q_r(x, y^k)$ and $\gamma > 0$. Moreover, we also consider the stochastic error because we may use a stochastic method to solve the subproblem. In addition, we are also interested in connecting this algorithm with a general catalyst method for a class of first-order optimization algorithms.

2.1 New FISTA algorithm

In this paper, we consider a new inexact FISTA scheme to solve problem (1.4) given by (following the philosophy of [4])

New FISTA with constant stepsize

Input : Initial conditions $x_1 \in \mathbb{R}^N$, constant $r \in \mathbb{R}^+$ of $Q_r(z, y^k)$.

Step 0 : Take $y_1 = x_1$ and $t_0 = 1$.

Step k $(k \geq 1)$:

$$\left\{\begin{array}{rcl} \text{solve } x^{k+1} & \approx & \operatorname{argmin}_z Q_r(z, y^k) \text{ with rule } \mathcal{R}_d \text{ or } \mathcal{R}_s \\ t_{k+1} & = & \dfrac{1 + \sqrt{1 + 4t_k^2}}{2} \\ y^{k+1} & = & x^{k+1} + \left(\dfrac{t_k - 1}{t_{k+1}}\right)(x^{k+1} - x^k), \end{array}\right. \tag{2.3}$$

where $Q_r(z, y^k) := \langle z - y^k, \nabla f(z) \rangle + \frac{r}{2}\|z - y^k\|_2^2 + g(z)$.

Here, $\mathcal{R}_d$ and $\mathcal{R}_s$ stand for the different types of errors (deterministic or stochastic depending on the problem):

1. $\mathcal{R}_d$: $Q_r(x^{k+1}, y^k) - \min Q_r(z, y^k) \leq \varepsilon_k$.
2. $\mathcal{R}_s$: $\mathbb{E}[Q_r(x^{k+1}, y^k) - \min Q_r(z, y^k) \mid \psi^k] \leq \varepsilon_k$ and ψ^k is the sigma algebra generated in solving x^{k+1}.

One interesting property of the scheme is related with the time or stepsize sequence $(t_k)_{k \geq 1}$ generated by the method.

Lemma 1. *The sequence* $(t_k)_{k\geq 1}$ *generated by the scheme (2.3) satisfies the bounds*

$$\frac{k+1}{2} \leq t_k \leq k+1. \tag{2.4}$$

Proof. In fact, we can easily see that given $t_0 = 1$, $t_1 = (1+\sqrt{5})/2 \approx 1.61803$ and so $1/2 \leq t_1 \leq 2$. Now using direct induction we have

$$\begin{aligned} t_k &= \frac{1+\sqrt{1+4t_{k-1}^2}}{2} \leq \frac{1+\sqrt{1+4t_{k-1}^2}}{2} \\ &\leq \frac{1+\sqrt{1+4t_{k-1}+4t_{k-1}^2}}{2} \leq \frac{1+2t_{k-1}+1}{2} \leq t_{k-1}+1. \end{aligned} \tag{2.5}$$

Thus, we obtain $t_k \leq k+1$. Similarly, we have

$$t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2} \geq \frac{1+2t_{k-1}}{2} \geq \frac{1}{2} + t_{k-1}. \tag{2.6}$$

Therefore, we have $t_k \geq \frac{k+1}{2}$. □

2.2 Preliminaries

Now, we introduce several definitions as well as some useful properties in variational and convex analysis (see for more details the excellent monographes [22, 23, 25, 26]).

The *subgradient set* of a function J at x is the set given by

$$\partial J(x) := \{v \mid J(y) \geq J(x) + \langle v, y-x\rangle, \forall y \in \mathrm{dom}(J)\},$$

and we say that v is a *subgradient vector*. Note that if J is convex and differentiable, then its gradient at x is a subgradient and so $\partial J(x) = \{\nabla J(x)\}$. But a subgradient can exist even when J is not differentiable at x. Let J be a convex function, we say that J is *gradient-Lipschitz* with constant L_J if J is differentiable and

$$\|\nabla J(x) - \nabla J(y)\|_2 \leq L_J \|x-y\|_2,$$

and we say J is *strongly convex* with constant ν_J if for any $x, y \in \mathrm{dom}(J)$ and $v \in \partial J(x)$

$$J(y) \geq J(x) + \langle v, y-x\rangle + \frac{\nu_J}{2}\|y-x\|_2^2.$$

We collect several basic but useful lemmas, the proofs can be found in [23].

Lemma 2. *If J is strongly convex with constant ν, for any $x \in$* dom(J)

$$J(x) - J(x^*) \geq \frac{v}{2}\|x-x^*\|_2^2, \tag{2.7}$$

where x^* is the minimizer of J.

Lemma 3. *If J is gradient-Lipschitz with constant L_J, for any $x, y \in$* dom(J)*,*

$$J(y) \leq J(x) + \langle \nabla J(x), y-x\rangle + \frac{L_J}{2}\|y-x\|_2^2. \tag{2.8}$$

We denote that a sequence $(a_k)_{k\geq 1} \in \ell^1$ if $\sum_{i=1}^{\infty} |a_k| < +\infty$.

Lemma 4. *Assume the nonnegative number sequences* $(\eta_k)_{k\geq 1}, (\delta_k)_{k\geq 1}$ *and* $(\xi_k)_{k\geq 1}$ *satisfying*

$$\xi_{k+1} \leq (1+\eta_k)\xi_k + \delta_k, \tag{2.9}$$

then we have

$$\xi_{k+1} \leq e^{\sum_{i=1}^{k}\eta_i}(\xi_1 + \sum_{i=1}^{k}\delta_i). \tag{2.10}$$

If further $(\eta_k)_{k\geq 1} \in \ell^1, (\delta_k)_{k\geq 1} \in \ell^1$*, then* $(\xi_k)_{k\geq 1}$ *is convergent.*

Proof. As we have nonnegative number sequences $1+\eta_k \leq e^{\eta_k}$, so

$$\begin{aligned} \xi_{k+1} &\leq (1+\eta_k)\xi_k + \delta_k \\ &\leq e^{\eta_k}\xi_k + \delta_k \\ &\leq e^{\eta_k+\eta_{k-1}}\xi_{k-1} + e^{\eta_k}\delta_{k-1} + \delta_k \\ &\vdots \\ &\leq e^{\sum_{i=1}^{k}\eta_i}\xi_1 + e^{\sum_{i=1}^{k}\eta_i}\cdot\sum_{i=1}^{k}\delta_i. \end{aligned} \tag{2.11}$$

Thus Eqn. (2.10) is obtained.

If further $(\eta_k)_{k\geq 1} \in \ell^1, (\delta_k)_{k\geq 1} \in \ell^1$, i.e., $\sum_{i=1}^{k}\eta_i < +\infty$ and $\sum_{i=1}^{k}\delta_i < +\infty$ for any $k \in \mathbb{Z}_+$. That is also to say $\sum_{i=1}^{k}\eta_i$ and $\sum_{i=1}^{k}\delta_i$ are bounded. With (2.11), $(\xi_k)_{k\geq 1}$ is bounded, and we denote the bound $M > 0$. Thus,

$$\xi_{k+1} - \xi_k \leq \eta_k\xi_k + \delta_k \leq M\eta_k + \delta_k. \tag{2.12}$$

That is also

$$0 \leq \left(\xi_{k+1} + \sum_{i=k+1}^{+\infty}(M\eta_i + \delta_i)\right) \leq \left(\xi_k + \sum_{i=k}^{+\infty}(M\eta_i + \delta_i)\right). \tag{2.13}$$

Therefore, the new nonnegative sequence $(\xi_k + \sum_{i=k}^{+\infty}(M\eta_i + \delta_i))_{k\geq 1}$ is monotonically decreasing. That means $(\xi_k + \sum_{i=k}^{+\infty}(M\eta_i + \delta_i))_{k\geq 1}$ converges. Noting that $\lim_k \sum_{i=k}^{+\infty}(M\eta_i + \delta_i) = 0$, then $(\xi_k)_{k\geq 1}$ also converges. □

3 Convergence rates of the new FISTA algorithm

In this section we provide convergence results of the new FISTA method (2.3) under deterministic and stochastic errors.

From now on, we suppose that the smooth-convex function f is gradient-Lipschitz with constant L and we note that the functional $Q_r(z, y^k)$ is strongly convex with constant r. We denote by x^* the solution of the minimization problem (1.4) and therefore $\min F = F(x^*)$, and we also denote

$$\tilde{x}^{k+1} := \arg\min_z Q_r(z, y^k). \tag{3.1}$$

3.1 Deterministic noise

In the case of having a deterministic noise, we use the deterministic error function $\mathcal{R}_d$.

First, we introduce some technical lemmas that help us in the proof of the main convergence theorem.

Lemma 5. *If the constant $r > 0$, then*

$$\|x^{k+1} - \tilde{x}^{k+1}\|_2^2 \leq \frac{2\varepsilon_k}{r}. \tag{3.2}$$

Proof. Note that $Q_r(z, y^k)$ is strongly convex with constant r, and $\tilde{x}^{k+1}$ (3.1) is the minimizer of $Q_r(z, y^k)$. With Lemma 2 we have

$$Q_r(x^{k+1}, y^k) - Q_r(\tilde{x}^{k+1}, y^k) \geq \frac{r}{2}\|x^{k+1} - \tilde{x}^{k+1}\|_2^2. \tag{3.3}$$

Noting that

$$Q_r(x^{k+1}, y^k) - Q_r(\tilde{x}^{k+1}, y^k) = Q_r(x^{k+1}, y^k) - \min Q_r(x, y^k) \leq \varepsilon_k,$$

and then we obtain the result. □

Lemma 6. *If the constants satisfy $r > L$, then we have*

$$\begin{aligned} F(\tilde{x}^{k+1}) &+ \frac{r}{2}\|\tilde{x}^{k+1} - y^k\|_2^2 - (F(x^{k+1}) + \frac{r}{2}\|x^{k+1} - y^k\|_2^2) \\ &\geq -\tau\varepsilon_k - \frac{r-L}{2}\|y^k - \tilde{x}^{k+1}\|_2^2, \end{aligned} \tag{3.4}$$

where $\tau = 1 + \dfrac{L^2}{r(r-L)}$.

Proof. Using the convexity of f,

$$0 \leq \left\langle \nabla f(x^{k+1}), x^{k+1} - \tilde{x}^{k+1} \right\rangle + f(\tilde{x}^{k+1}) - f(x^{k+1}). \tag{3.5}$$

From the definition of $\mathcal{R}_d$

$$\begin{aligned} -\varepsilon_k &\leq Q_r(\tilde{x}^{k+1}, y^k) - Q_r(x^{k+1}, y^k) \\ &= \left[\langle \tilde{x}^{k+1} - y^k, \nabla f(\tilde{x}^{k+1})\rangle + \frac{r}{2}\|\tilde{x}^{k+1} - y^k\|_2^2 + g(\tilde{x}^{k+1})\right] \\ &\quad - \left[\langle x^{k+1} - y^k, \nabla f(x^{k+1})\rangle + \frac{r}{2}\|x^{k+1} - y^k\|_2^2 + g(x^{k+1})\right]. \end{aligned} \tag{3.6}$$

Adding (3.5) to (3.6) side by side, we have

$$\begin{aligned} -\varepsilon_k \leq &\left(F(x^{k+1}) + \frac{r}{2}\|x^{k+1} - y^k\|_2^2\right) - \left(F(\tilde{x}^{k+1}) + \frac{r}{2}\|\tilde{x}^{k+1} - y^k\|_2^2\right) \\ &+ \left\langle \nabla f(x^{k+1}) - \nabla f(\tilde{x}^{k+1}), y^k - \tilde{x}^{k+1} \right\rangle. \end{aligned} \tag{3.7}$$

Now, using the Schwarz inequality $\langle a, b\rangle \leq \frac{\|a\|_2^2}{2\mu} + \frac{\mu}{2}\|b\|_2^2$ with $\mu = r - L$, $a = \nabla f(x^{k+1}) - \nabla f(\tilde{x}^{k+1})$ and $b = y^k - \tilde{x}^{k+1}$,

$$\begin{aligned} &\left\langle \nabla f(x^{k+1}) - \nabla f(\tilde{x}^{k+1}), y^k - \tilde{x}^{k+1} \right\rangle \\ &\leq \frac{1}{2(r-L)}\|\nabla f(x^{k+1}) - \nabla f(\tilde{x}^{k+1})\|_2^2 + \frac{r-L}{2}\|y^k - \tilde{x}^{k+1}\|_2^2 \\ &\leq \frac{L^2}{2(r-L)}\|x^{k+1} - \tilde{x}^{k+1}\|_2^2 + \frac{r-L}{2}\|y^k - \tilde{x}^{k+1}\|_2^2. \end{aligned} \tag{3.8}$$

Combining (3.7) and (3.8), we have

$$\begin{aligned} -\varepsilon_k \leq &\left(F(x^{k+1}) + \frac{r}{2}\|x^{k+1} - y^k\|_2^2\right) - \left(F(\tilde{x}^{k+1}) + \frac{r}{2}\|\tilde{x}^{k+1} - y^k\|_2^2\right) \\ &+ \frac{L^2}{2(r-L)}\|x^{k+1} - \tilde{x}^{k+1}\|_2^2 + \frac{r-L}{2}\|y^k - \tilde{x}^{k+1}\|_2^2. \end{aligned} \tag{3.9}$$

By rearrangement of the above inequality (3.9) and using Lemma 5, we obtain the result. □

Lemma 7. *If the constants satisfy $r > L$, then we have the bounds*

$$F(x^k) - F(x^{k+1}) \geq \frac{r}{2}\|y^k - x^{k+1}\|_2^2 + r\langle y^k - x^k, \tilde{x}^{k+1} - y^k\rangle - \tau\varepsilon_k, \tag{3.10}$$

$$F(x^*) - F(x^{k+1}) \geq \frac{r}{2}\|y^k - x^{k+1}\|_2^2 + r\langle y^k - x^*, \tilde{x}^{k+1} - y^k\rangle - \tau\varepsilon_k, \tag{3.11}$$

where $\tau = 1 + \frac{L^2}{r(r-L)}$.

Proof. With Lemma 3, we have

$$-f(\tilde{x}^{k+1}) \geq -f(y^k) - \big\langle \tilde{x}^{k+1} - y^k, \nabla f(y^k)\big\rangle - \frac{L}{2}\|\tilde{x}^{k+1} - y^k\|_2^2. \tag{3.12}$$

Subtracting $g(\tilde{x}^{k+1})$ to both sides

$$-F(\tilde{x}^{k+1}) \geq -f(y^k) - \big\langle \tilde{x}^{k+1} - y^k, \nabla f(y^k)\big\rangle - \frac{L}{2}\|\tilde{x}^{k+1} - y^k\|_2^2 - g(\tilde{x}^{k+1}).$$

Now, using the convexity of f and g, we have

$$\begin{array}{rcl} f(x^k) - f(y^k) & \geq & \langle x^k - y^k, \nabla f(y^k)\rangle, \\ g(x^k) - g(\tilde{x}^{k+1}) & \geq & \langle x^k - \tilde{x}^{k+1}, \bar{\nabla} g(\tilde{x}^{k+1})\rangle, \end{array}$$

where $\bar{\nabla} g(\tilde{x}^{k+1})$ is any vector of the set $\partial g(\tilde{x}^{k+1})$. Summing the inequalities yields

$$\begin{aligned} F(x^k) \geq & f(y^k) + g(\tilde{x}^{k+1}) + \langle x^k - y^k, \nabla f(y^k)\rangle \\ & + \langle x^k - \tilde{x}^{k+1}, \bar{\nabla} g(\tilde{x}^{k+1})\rangle. \end{aligned} \tag{3.13}$$

The definition of $\tilde{x}^{k+1}$ is based on the minimization of (2.1), and the optimization condition gives

$$-\nabla f(y^k) - r(\tilde{x}^{k+1} - y^k) \in \partial g(\tilde{x}^{k+1}). \tag{3.14}$$

Substituting (3.14) into (3.13), we obtain

$$\begin{aligned} F(x^k) \geq & f(y^k) + g(\tilde{x}^{k+1}) + \langle \tilde{x}^{k+1} - y^k, \nabla f(y^k)\rangle \\ & + r\langle \tilde{x}^{k+1} - x^k, \tilde{x}^{k+1} - y^k\rangle. \end{aligned} \tag{3.15}$$

Direct summation of (3.13) and (3.15) gives

$$F(x^k) - F(\tilde{x}^{k+1}) \geq (r - \frac{L}{2})\|y^k - \tilde{x}^{k+1}\|_2^2 + r\langle y^k - x^k, \tilde{x}^{k+1} - y^k\rangle. \tag{3.16}$$

Summing (3.16) and (3.4), we obtain the first inequality (3.10)

$$F(x^k) - F(x^{k+1}) \geq \frac{r}{2}\|x^{k+1} - y^k\|_2^2 + r\langle y^k - x^k, \tilde{x}^{k+1} - y^k\rangle - \tau\varepsilon_k. \tag{3.17}$$

On the other hand, with the convexities of f and g,

$$\begin{array}{rcl} f(x^*) - f(y^k) & \geq & \langle x^* - y^k, \nabla f(y^k)\rangle, \\ g(x^*) - g(\tilde{x}^{k+1}) & \geq & \langle x^* - \tilde{x}^{k+1}, \bar{\nabla} g(\tilde{x}^{k+1})\rangle. \end{array}$$

Adding them together, we have

$$F(x^*) \geq f(y^k) + g(\tilde{x}^{k+1}) + \langle x^* - y^k, \nabla f(y^k)\rangle + \langle x^* - \tilde{x}^{k+1}, \bar{\nabla} g(\tilde{x}^{k+1})\rangle. \tag{3.18}$$

Substituting (3.14) into (3.18) gives

$$F(x^*) - F(\tilde{x}^{k+1}) \geq \left(r - \frac{L}{2}\right) \|y^k - \tilde{x}^{k+1}\|_2^2 + r\langle y^k - x^*, \tilde{x}^{k+1} - y^k\rangle. \tag{3.19}$$

Similarly, with (3.4), we have

$$F(x^*) - F(x^{k+1}) \geq \frac{r}{2}\|x^{k+1} - y^k\|_2^2 + r\langle y^k - x^*, \tilde{x}^{k+1} - y^k\rangle - \tau\varepsilon_k, \tag{3.20}$$

that gives the second inequality (3.11). □

Lemma 8. *Given a nonnegative sequence $(s_k)_{k\geq 0} \to 0$. Let the constants satisfy $r > L$ and the sequence $(x^k)_{k\geq 0}$ be generated by the inexact FISTA (2.3) with the deterministic error function $\mathcal{R}_d$, then we have*

$$F(x^k) - F(x^*) \leq \frac{r\left(\dfrac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \displaystyle\sum_{i=1}^{k} C\frac{t_i^2\varepsilon_i}{s_i}\right) \cdot \mathrm{e}^{\sum_{i=1}^{k} s_i}}{2t_{k-1}^2}, \tag{3.21}$$

where $C = \frac{2}{r} + 2\max_k\{\frac{s_k\tau}{r}\}$.

Proof. We denote

$$F^k := F(x^k) - F(x^*).$$

By (3.10) × $(t_k - 1)$ + (3.11),

$$\begin{aligned}\frac{2[(t_k - 1)F^k - t_kF^{k+1}]}{r} &\geq t_k\|x^{k+1} - y^k\|_2^2 \\ &\quad + 2\langle\tilde{x}^{k+1} - y^k, t_ky^k - (t_k - 1)x^k - x^*\rangle - \frac{2\tau t_k\varepsilon_k}{r}.\end{aligned} \tag{3.22}$$

With $t_{k-1}^2 = t_k^2 - t_k$, (3.22) × t_k yields

$$\begin{aligned}\frac{2[t_{k-1}^2F^k - t_k^2F^{k+1}]}{r} &\geq \|t_kx^{k+1} - t_ky^k\|_2^2 \\ &\quad + 2t_k\langle\tilde{x}^{k+1} - y^k, t_ky^k - (t_k - 1)x^k - x^*\rangle - \frac{2\tau t_k^2\varepsilon_k}{r}.\end{aligned} \tag{3.23}$$

Substituting $a = t_kx^{k+1} - (t_k - 1)x^k - x^*$ and $b = t_ky^k - (t_k - 1)x^k - x^*$ into identity

$$\|a - b\|_2^2 + 2\langle a - b, b\rangle = \|a\|_2^2 - \|b\|_2^2, \tag{3.24}$$

then we have:

$$\begin{aligned}&\|t_kx^{k+1} - t_ky^k\|_2^2 + 2t_k\langle\tilde{x}^{k+1} - y^k, t_ky^k - (t_k - 1)x^k - x^*\rangle \\ &\quad= \|t_kx^{k+1} - t_ky^k\|_2^2 + 2t_k\langle x^{k+1} - y^k, t_ky^k - (t_k - 1)x^k - x^*\rangle \\ &\qquad + 2t_k\langle\tilde{x}^{k+1} - x^{k+1}, t_ky^k - (t_k - 1)x^k - x^*\rangle \\ &\quad\overset{(3.24)}{=} \|t_kx^{k+1} - (t_k - 1)x^k - x^*\|_2^2 - \|t_ky^k - (t_k - 1)x^k - x^*\|_2^2 \\ &\qquad + 2t_k\langle\tilde{x}^{k+1} - x^{k+1}, t_ky^k - (t_k - 1)x^k - x^*\rangle \\ &\quad= \|t_kx^{k+1} - (t_k - 1)x^k - x^*\|_2^2 - \|t_{k-1}x^k - (t_{k-1} - 1)x^{k-1} - x^*\|_2^2 \\ &\qquad + 2t_k\langle\tilde{x}^{k+1} - x^{k+1}, t_{k-1}x^k - (t_{k-1} - 1)x^{k-1} - x^*\rangle.\end{aligned} \tag{3.25}$$

In the third identity, we use the fact $t_k y^k = t_k x^k + (t_{k-1} - 1)(x^k - x^{k-1})$. If we denote $w^k = \|t_{k-1}x^k - (t_{k-1} - 1)x^{k-1} - x^*\|_2^2$, (3.23) can be rewritten as

$$\begin{aligned}\frac{2t_k^2 F^{k+1}}{r} + w^{k+1} &\leq \frac{2t_{k-1}^2 F^k}{r} + w^k + \frac{2\tau t_k^2 \varepsilon_k}{r} \\ &\quad + 2t_k \langle x^{k+1} - \tilde{x}^{k+1}, t_{k-1}x^k - (t_{k-1} - 1)x^{k-1} - x^* \rangle \\ &\leq \frac{2t_k^2 F^k}{r} + w^k + \frac{2\tau t_k^2 \varepsilon_k}{r} + s_k w^k + \frac{t_k^2}{s_k}\|x^{k+1} - \tilde{x}^{k+1}\|_2^2, \qquad (3.26)\end{aligned}$$

where we use the Schwarz inequality $2\langle a, b\rangle \leq \mu\|a\|^2 + \frac{1}{\mu}\|b\|^2$ with $a = x^{k+1} - \tilde{x}^{k+1}$, $t_{k-1}x^k - (t_{k-1} - 1)x^{k-1} - x^*$ and $\mu = \frac{t_k}{s_k}$,

$$\begin{aligned}&2t_k \langle x^{k+1} - \tilde{x}^{k+1}, t_{k-1}x^k - (t_{k-1} - 1)x^{k-1} - x^* \rangle \\ &\leq \frac{t_k^2}{s_k}\|x^{k+1} - \tilde{x}^{k+1}\|_2^2 + s_k\|t_{k-1}x^k - (t_{k-1} - 1)x^{k-1} - x^*\|_2^2 \\ &= \frac{t_k^2}{s_k}\|x^{k+1} - \tilde{x}^{k+1}\|_2^2 + s_k w^k. \qquad (3.27)\end{aligned}$$

Obviously,

$$s_k w^k \leq s_k \left(\frac{2t_k^2 F^k}{r} + w^k\right);$$

and with Lemma 5,

$$\frac{2t_k^2 F^{k+1}}{r} + w^{k+1} \leq (1 + s_k)\left(\frac{2t_{k-1}^2 F^k}{r} + w^k\right) + \frac{2\tau t_k^2 \varepsilon_k}{r} + \frac{2t_k^2}{rs_k}\varepsilon_k. \qquad (3.28)$$

Denoting

$$\xi_k := \frac{2t_{k-1}^2 F^k}{r} + w^k,$$

then, we have

$$\begin{aligned}\xi_{k+1} &\leq (1 + s_k)\xi_k + \left[\frac{2\tau t_k^2 \varepsilon_k}{r} + \frac{2t_k^2}{rs_k}\varepsilon_k\right] \\ &\leq (1 + s_k)\xi_k + C\frac{t_k^2 \varepsilon_k}{s_k}, \qquad (3.29)\end{aligned}$$

where $C = \frac{2}{r} + 2\max_k\{\frac{s_k \tau}{r}\} > 0$. Applying Lemma 4 to (3.29), we obtain

$$\frac{2t_{k-1}^2 F^k}{r} \leq \xi_k \leq \mathrm{e}^{\sum_{i=1}^k s_i}\left(\frac{F(x^1) - F(x^*)}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^k C\frac{t_i^2 \varepsilon_i}{s_i}\right). \qquad (3.30)$$

That gives us the Eq. (3.21). □

Now we can state the main theorem about the convergence rates for the deterministic case.

Theorem 1. *Assume $\varepsilon_k \sim \mathcal{O}(1/k^\alpha)$, and let the constants satisfy $r > L$ and the sequence $(x^k)_{k\geq 0}$ be generated by the new inexact FISTA (2.3) with the deterministic error function $\mathcal{R}_d$, then we have the following results*

$$F(x^k) - \min F = \begin{cases} \mathcal{O}\left(\frac{1}{k^2}\right), & \textit{if } \alpha > 4, \\ \mathcal{O}\left(\frac{\ln^3 k}{k^2}\right), & \textit{if } \alpha = 4, \\ \mathcal{O}\left(\frac{1}{k^{\frac{\alpha}{2}-1}}\right), & \textit{if } 2 < \alpha < 4. \end{cases} \qquad (3.31)$$

where $\min F = F(x^*)$.

Proof. Case –1– ($\alpha > 4$): We set $s_i = 1/i^{\frac{\alpha-2}{2}}$; and it is easy to see that $(s_i)_{i\geq 0} \in \ell^1$ and then $\mathrm{e}^{\sum_{i=1}^k s_i} < +\infty$. Noting that $\frac{i}{2} \leq t_{i-1} \leq i$, then,

$$\frac{t_i^2 \varepsilon_i}{s_i} \sim \mathcal{O}\left(\frac{1}{i^{\frac{\alpha-2}{2}}}\right) \in \ell^1.$$

Thus, we have

$$\frac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^k C\frac{t_i^2\varepsilon_i}{s_i} < +\infty.$$

Therefore

$$\left(\frac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^k C\frac{t_k^2\varepsilon_k}{s_k}\right) \cdot \mathrm{e}^{\sum_{i=1}^k s_k} < +\infty. \tag{3.32}$$

From Lemma 8, we have the result

$$F(x^k) - \min F \sim \mathcal{O}\left(\frac{1}{t_{k-1}^2}\right) \sim \mathcal{O}\left(\frac{1}{k^2}\right). \tag{3.33}$$

Case –2– ($\alpha = 4$): Now, we set $s_i = 1/(i \ln i)$; and then we have

$$\begin{aligned}\sum_{i=2}^k \frac{1}{i\ln i} &= \sum_{i=2}^k \int_0^1 \frac{1}{i\ln i}dt \leq \sum_{i=2}^k \int_i^{i+1} \frac{1}{t\ln t}dt \\ &= \sum_{i=2}^k (\ln\ln(i+1) - \ln\ln i) \\ &= \ln\ln(k+1) - \ln\ln 2 \sim \mathcal{O}(\ln\ln k).\end{aligned} \tag{3.34}$$

That means $\mathrm{e}^{\sum_{i=1}^k s_i} = \mathcal{O}(\ln k)$. With the fact $\frac{i}{2} \leq t_{i-1} \leq i$, we have

$$\frac{t_i^2\varepsilon_i}{s_i} \sim \mathcal{O}(\frac{\ln i}{i}).$$

That also indicates

$$\sum_{i=1}^k C\frac{t_i^2\varepsilon_i}{s_i} \sim \mathcal{O}(\ln^2 k).$$

Thus, we obtain

$$\frac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^k C\frac{t_i^2\varepsilon_i}{s_i} \sim \mathcal{O}(\ln^2 k).$$

Combining the above equations

$$\left(\frac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^k C\frac{t_k^2\varepsilon_k}{s_k}\right) \cdot \mathrm{e}^{\sum_{i=1}^k s_k} \sim \mathcal{O}(\ln^3 k). \tag{3.35}$$

And from Lemma 8, we have the result

$$F(x^k) - \min F \sim \mathcal{O}\left(\frac{\ln^3 k}{t_{k-1}^2}\right) \sim \mathcal{O}\left(\frac{\ln^3 k}{k^2}\right). \tag{3.36}$$

Case –3– ($2 < \alpha < 4$): We set $s_i = 1/i^{\frac{\alpha}{2}}$; and then $(s_i)_{i\geq 0} \in \ell^1$ and $\mathrm{e}^{\sum_{i=1}^k s_i} < +\infty$. Considering $\frac{i}{2} \leq t_{i-1} \leq i$, we have

$$\frac{t_i^2 \varepsilon_i}{s_i} \sim \mathcal{O}\left(\frac{1}{i^{2+\frac{\alpha}{2}}}\right).$$

That also indicates

$$\sum_{i=1}^k C\frac{t_i^2 \varepsilon_i}{s_i} \sim \mathcal{O}\left(\frac{1}{k^{1+\frac{\alpha}{2}}}\right).$$

Thus, we have

$$\frac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^k C\frac{t_i^2 \varepsilon_i}{s_i} \sim \mathcal{O}(\frac{1}{k^{1+\frac{\alpha}{2}}}).$$

Combining the above equations

$$\left(\frac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^k C\frac{t_k^2 \varepsilon_k}{s_k}\right) \cdot \mathrm{e}^{\sum_{i=1}^k s_k} \sim \mathcal{O}\left(\frac{1}{k^{1+\frac{\alpha}{2}}}\right). \tag{3.37}$$

From Lemma 8, we have the result

$$F(x^k) - \min F \sim \mathcal{O}\left(\frac{1}{k^{1+\frac{\alpha}{2}} t_{k-1}^2}\right) \sim \mathcal{O}\left(\frac{1}{k^{\frac{\alpha}{2}-1}}\right). \tag{3.38}$$

□

The ε_k stands for a bound of the noise in each iteration. The parameter α depends on the algorithm and the inner loop iteration used to solve the subproblem (2.2). We recall the error in the subproblem (2.2) in previous analysis, where γ in (2.2) is due to the algorithm chose for solving the subproblem. For instance, if we select $j_k = k$ in (2.2), then γ is actually α; as another choice, if j_k is set as $j_k = k^2$, then, $\alpha = 2\gamma$; similarly, if $j_k = k^{\frac{1}{2}}$, we have $\alpha = \frac{\gamma}{2}$.

Although our results look worse than those of Corollary 3.7 presented in [24], we remark that our results are under totally different settings: –1– our error ε_k is based on the function values and can be controlled; –2– the terms (s_k) in [24] are not errors, but auxiliary parameters. The errors in [24] are given by [Eqns. (18), (19), [24]], which are under very strong assumptions and hard to verify.

3.2 Stochastic noise

Up to our knowledge, previous literature in convergence analysis is always focusing on the use of deterministic noise. This is because deterministic optimization methods are used for solving the subproblem $\min_z Q_r(z, y^k)$. Actually, we can also use iterative stochastic algorithms for solving the subproblem (2.2). And then, the inexact solver for the subproblems will lead to stochastic noise. Note that the (conditional) expectation of the stochastic noise is bounded, but the noise itself may be unbounded (like Gaussian noise). Thus, the case of stochastic noise should be discussed separately.

In the case of having a stochastic noise, we consider the stochastic error function $\mathcal{R}_S$.

First, we introduce some technical lemmas that help us in the proof of the main convergence theorem. As the global procedure in this case is similar as in the previous subsection with the deterministic noise, we will just detail some steps.

We denote by χ^k the sigma algebra generated by $x^0, x^1, \ldots, x^k$ and $\psi^0, \psi^1, \ldots, \psi^k$, i.e.,

$$\chi^k := \sigma(x^0, x^1, \ldots, x^k, \psi^0, \psi^1, \ldots, \psi^k).$$

Obviously, we have $\psi^k \subseteq \chi^k$.

Lemma 9. *If the constant $r > 0$, then*

$$\mathbb{E}(\|x^{k+1} - \tilde{x}^{k+1}\|_2^2 \mid \chi^k) \leq \frac{2\varepsilon_k}{r}. \tag{3.39}$$

Proof. First we note that

$$\mathbb{E}[Q_r(x^{k+1}, y^k) - \min Q_r \mid \psi^k] = \mathbb{E}[Q_r(x^{k+1}, y^k) - Q_r(\tilde{x}^{k+1}, y^k) \mid \psi^k] \leq \varepsilon_k$$

Taking the conditional expectation on both sides over χ^k, and with the law of iterated expectations [1], we then derive

$$\mathbb{E}\left[\mathbb{E}[Q_r(x^{k+1}, y^k) - Q_r(\tilde{x}^{k+1}, y^k) \mid \psi^k] \,\middle|\, \chi^k\right] = \mathbb{E}[Q_r(x^{k+1}, y^k) - Q_r(\tilde{x}^{k+1}, y^k) \mid \psi^k] \leq \varepsilon_k. \tag{3.40}$$

Now, taking again the conditional expectation of (3.3) on both sides over χ^k and using (3.40), we then derive the result. □

Lemma 10. *Given a nonnegative sequence $(s_k)_{k\geq 0} \to 0$. Let the constants satisfy $r > L$ and the sequence $(x^k)_{k\geq 0}$ be generated by the inexact FISTA (2.3) with the stochastic error function $\mathcal{R}_s$, then we have*

$$\mathbb{E}\big(F(x^k) - \min F\big) \leq \frac{r\left(\frac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^{k} C\frac{t_i^2\varepsilon_i}{s_i}\right) \cdot \mathrm{e}^{\sum_{i=1}^{k} s_i}}{2t_{k-1}^2}, \tag{3.41}$$

where $C = \frac{2}{r} + 2\max_k\{\frac{s_k\tau}{r}\}$.

Proof. Taking conditional expectation on both sides of (3.26), we have

$$\begin{aligned}\mathbb{E}\left(\frac{2t_k^2F^{k+1}}{r} + w^{k+1} \,\middle|\, \chi^k\right) &\leq \frac{2t_k^2F^k}{r} + w^k + \frac{2\tau t_k^2\varepsilon_k}{r}\\ &\quad + s_kw^k + \frac{t_k^2}{s_k}\mathbb{E}(\|x^{k+1} - \tilde{x}^{k+1}\|_2^2 \mid \chi^k)\\ &\leq \frac{2t_k^2F^k}{r} + w^k + \frac{2\tau t_k^2\varepsilon_k}{r} + s_kw^k + \frac{2t_k^2\varepsilon_k}{rs_k},\end{aligned} \tag{3.42}$$

where the last inequality is due to Lemma 10. Taking the total expectation, and using

$$\mathbb{E}\left(\mathbb{E}\Big(\frac{2t_k^2F^{k+1}}{r} + w^{k+1} \,\Big|\, \chi^k\Big)\right) = \mathbb{E}\left(\frac{2t_k^2F^{k+1}}{r} + w^{k+1}\right),$$

we have:

$$\mathbb{E}\left(\frac{2t_k^2F^{k+1}}{r} + w^{k+1}\right) \leq (1 + s_k)\mathbb{E}\left(\frac{2t_k^2F^k}{r} + w^k\right) + \frac{2\tau t_k^2\varepsilon_k}{r} + \frac{2t_k^2\varepsilon_k}{rs_k}. \tag{3.43}$$

Denoting

$$\xi_k := \mathbb{E}\left(\frac{2t_{k-1}^2F^k}{r} + w^k\right),$$

then, we have

$$\xi_{k+1} \leq (1 + s_k)\xi_k + C\frac{t_k^2\varepsilon_k}{s_k}, \tag{3.44}$$

where $C = \frac{2}{r} + 2\max_k\{\frac{s_k\tau}{r}\} > 0$. Applying Lemma 4 to (3.44), we obtain

$$\mathbb{E}\left(\frac{2t_{k-1}^2F^k}{r}\right) \leq \xi_k \leq \mathrm{e}^{\sum_{i=1}^{k} s_i}\left(\frac{F(x^1) - \min F}{r} + \|x^1 - x^*\|_2^2 + \sum_{i=1}^{k} C\frac{t_i^2\varepsilon_i}{s_i}\right). \tag{3.45}$$

That gives us the result. □

Similarly, we can state (the proof is similar to those of Theorem 1) the main theorem about the convergence rates for the stochastic case.

Theorem 2. *Assume $\varepsilon_k \sim \mathcal{O}(1/k^{\alpha})$, and let the constants satisfy $r > L$ and the sequence $(x^k)_{k\geq 0}$ be generated by the inexact FISTA (2.3) with the stochastic error function $\mathcal{R}_s$, then we have the following results*

$$\mathbb{E}\big(F(x^k) - \min F\big) = \begin{cases} \mathcal{O}\left(\frac{1}{k^2}\right), & \textit{if } \alpha > 4, \\ \mathcal{O}\left(\frac{\ln^3 k}{k^2}\right), & \textit{if } \alpha = 4, \\ \mathcal{O}\left(\frac{1}{k^{\frac{\alpha}{2}-1}}\right), & \textit{if } 2 < \alpha < 4. \end{cases} \tag{3.46}$$

Note that the convergence rate in the stochastic case is similar to the deterministic one (Theorem 1). The parameter α is also determined by the algorithm and the inner loop iteration used for the subproblem (2.2). The main difference between deterministic and stochastic cases lies on the different meanings of ε_k. In this stochastic settings, ε_k is a bound for the conditional expectation of the variable $(Q_r(x^{k+1}, y^k) - \min Q_r(z, y^k))$. But the variable itself may be unbounded; while for deterministic case, the ε_k is a bound for the error.

4 Connection with the unified catalyst for first-order optimization

In this section, we connect the inexact accelerated algorithm with a recent generic scheme for accelerating first-order optimization methods in the sense of Nesterov, the *catalyst algorithm* proposed in paper [19]. This catalyst algorithm is an accelerated framework which can be used for many existing accelerated algorithms such as SAG[27], SAGA[9], MISO[20], SDCA[29], SVRG[35] and some coordinate descent algorithm [34].

The catalyst algorithm is devoted to the following minimization problem

$$\min_x \{G(x) = \frac{1}{n}\sum_{i=1}^{n} h_i(x) + g(x)\}, \tag{4.1}$$

where $\frac{1}{n}\sum_{i=1}^{n} h_i(x)$ is the smooth convex part and g is a nonsmooth convex function. For a given minimization algorithm $\mathcal{M}$ (it can be SAG, SAGA, MISO, SCDA or SVRG), the catalyst performs in the iteration k-th as

$$\begin{cases} \text{solve } x^{k+1} & \approx & \text{argmin} H_k(x) = G(x) + \frac{r}{2}\|x - y^k\|_2^2 \text{ with} \\ & & \qquad H_k(x^{k+1}) - \min H_k \leq \varepsilon_k \text{ by } \mathcal{M}, \\ t_{k+1} & = & \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \\ y^{k+1} & = & x^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(x^{k+1} - x^k). \end{cases} \tag{4.2}$$

In fact, in the inexact FISTA using $\mathcal{R}_d$, if we set $f \equiv 0$ and $g(x) = G(x)$ (in this case, $L = 0$), we immediately obtain the convergence results using Theorem 1 as follows.

Theorem 3. *Assume $\varepsilon_k \sim \mathcal{O}(1/k^{\alpha})$, let the constant satisfy $r > 0$ and the sequence $(x^k)_{k\geq 0}$ be generated by the catalyst algorithm (4.2) with deterministic errors, then we have the same convergence results (Eq. (3.31)) as Theorem 1.*

In paper [19], the authors just present convergence results for the case $\varepsilon_k \sim \mathcal{O}(1/k^{\alpha})$ and $\alpha > 4$. Noting that the catalyst algorithm is just a special case of our algorithm, we can use directly our results to the catalyst. Based on Theorem 1, the catalyst still converges if $2 < \alpha \leq 4$ and the convergence rates are given in Theorem 3.

Moreover, the authors in [19] only consider the deterministic noised. Actually, algorithms like SAG, SAGA, MISO, SCDA and SVRG are stochastic algorithms, and using them to minimize H_k only derive stochastic errors. Thus, we also present the convergence rates for catalyst algorithm when the noise is stochastic, i.e., solve $x^{k+1} \approx \mathrm{argmin} H_k(x) = G(x) + \frac{r}{2}\|x - y^k\|_2^2$ with $\mathbb{E}(H_k(x^{k+1}) - \min H_k \mid \psi^k) \leq \varepsilon_k$ by $\mathcal{M}$. Similarly, using our results (Theorem 2), we have:

Theorem 4. *Assume $\varepsilon_k \sim \mathcal{O}(1/k^{\alpha})$, let the constant satisfy $r > 0$ and the sequence $(x^k)_{k\geq 0}$ be generated by the catalyst algorithm (4.2) with stochastic errors, then we have the same convergence results (Eq. (3.46)) as Theorem 2.*

Theorems 3 and 4 are direct applications of our previous findings. Compared with the convergence rates in [19], our results provide more information (more values of α are studied). Besides, as in distributed optimization community stochastic algorithms are quite popular to be used in the catalyst algorithm, we have presented the Theorem 4 which states rigorous convergence results with stochastic errors.

5 Numerical tests

In this section, we present some numerical examples to demonstrate our theoretical findings. Two tests are conducted for the following problem

$$\min_{x} H(x) := \frac{1}{2}\|b - Ax\|_2^2 + \|x\|_1. \tag{5.1}$$

Applying accelerated proximal gradient algorithms to problem (5.1), and considering the inexact vector, we then derive the FISTA scheme for the problem, i.e.,

$$\left\{\begin{array}{lcl} x^{k+1} & = & S_{h\lambda}(y^k - h\cdot A^{\top}(Ay^k - b)) + e_k \\ t_{k+1} & = & \dfrac{1+\sqrt{1+4t_k^2}}{2} \\ y^{k+1} & = & x^{k+1} + \left(\dfrac{t_k - 1}{t_{k+1}}\right)(x^{k+1} - x^k), \end{array}\right. \tag{5.2}$$

where h is the stepsize, and $S_{h\lambda}$ is the well-known soft-shrinkage thresholding operator, and $e_k = \upsilon/k^{\alpha}$ where υ is the deterministic noise or the Gaussian noise. The stepsize h is set in this numerical test as $h = 1/\|A\|_2^2$.

The first test is done using *Synthetic Data*: the matrix A is generated by the Gaussian random variables. We present the function values $H(x^k)$ using the inexact FISTA algorithm, for three different values of α, and the case without noise (standard FISTA). Figure 1(a) presents the functions values of 500 iterations for different values of α for the deterministic noise case, where υ is generated by the Matlab codes `ones(.,.)`; while Figure 1(b) presents the functions values of 1000 iterations for different α for the Gaussian random noise case which is generated by Matlab codes `randn(.,.)`. From the figures, it is easy to observe that in the case of deterministic noise the convergence is always faster, but also that the convergence is faster increasing the value of the parameter α as shown theoretically in Theorem 1. In the stochastic case (random noise) all the algorithms behave similarly, with very small differences.

The second test is about image deblurring [4, 21]. Using wavelet analysis, an image f can be described as $f = Fx_f$, where F is the wavelet matrix. It is well known that x_f is sparse [12]. In our tests the blurring operator B is a Gaussian blurring one, and therefore, the noised image can be presented as

$$b = Bf + e = (BF)x_f + e,$$

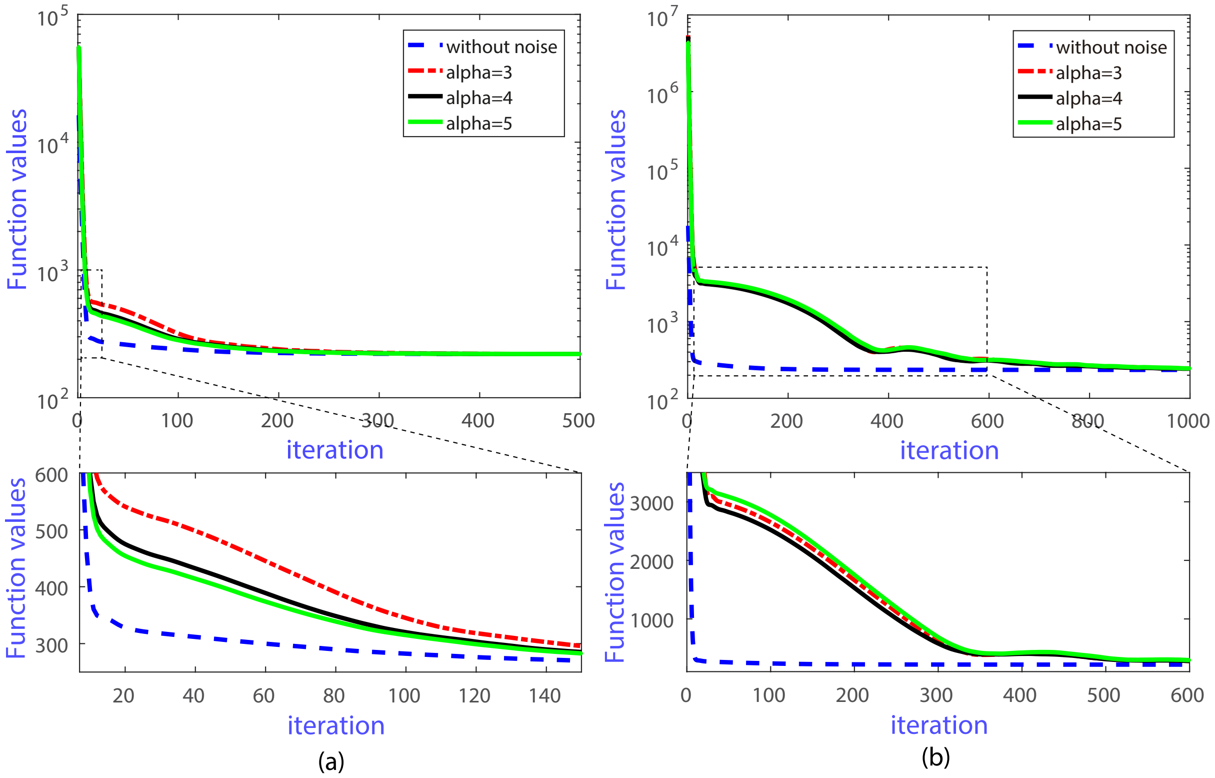
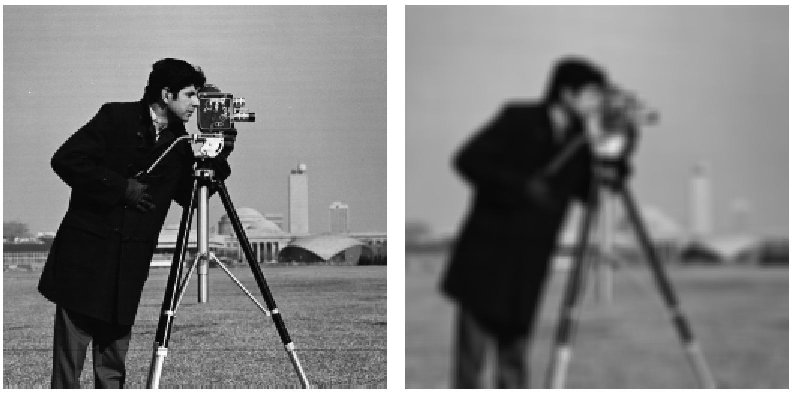
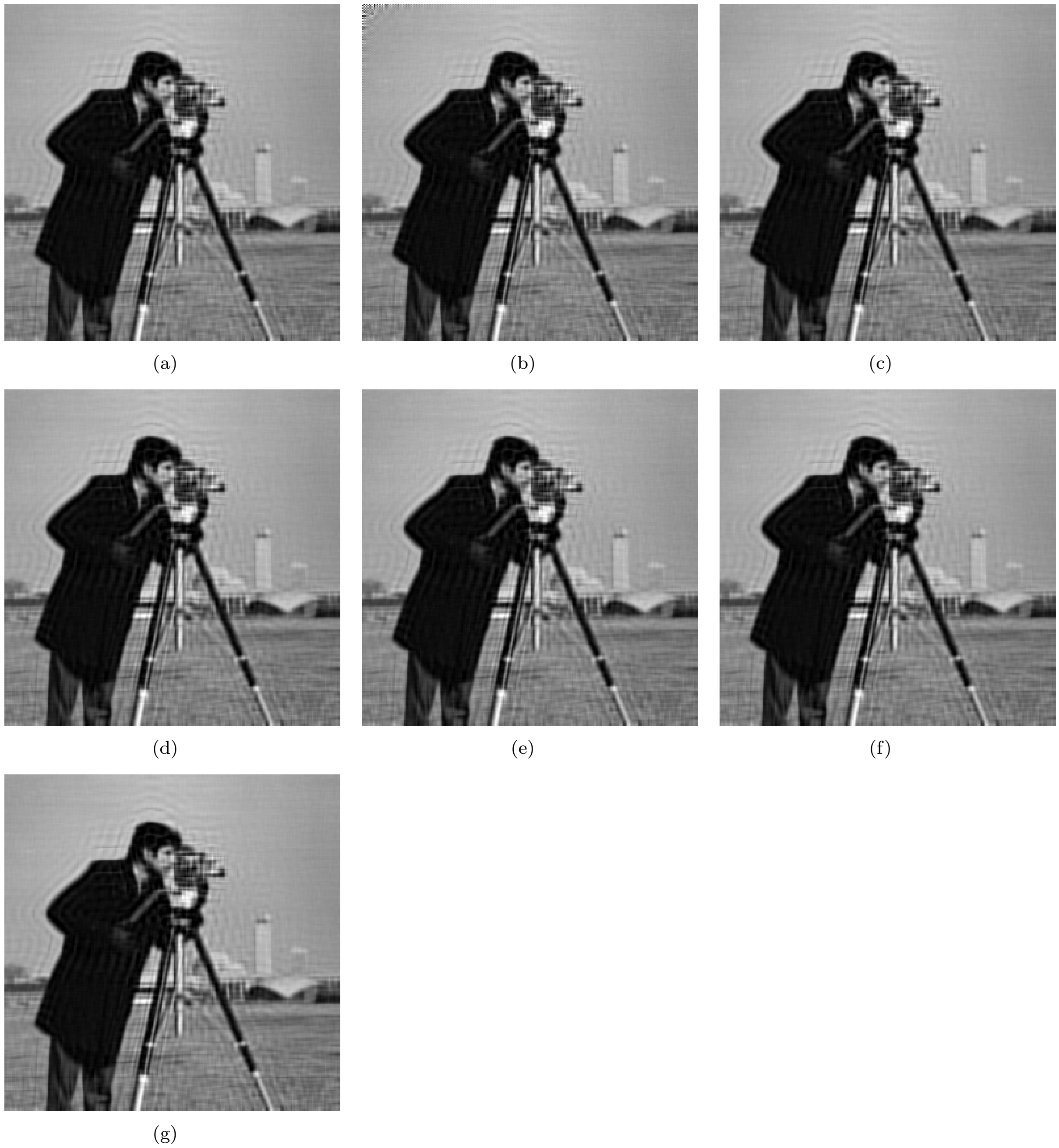


Figure 1: Functions values for different values of $\alpha = 3, 4$ and 5, and without any noise (a) Deterministic noise; (b) Stochastic noise.



(a) (b)

Figure 2: (a) Original image; (b) Noised image.

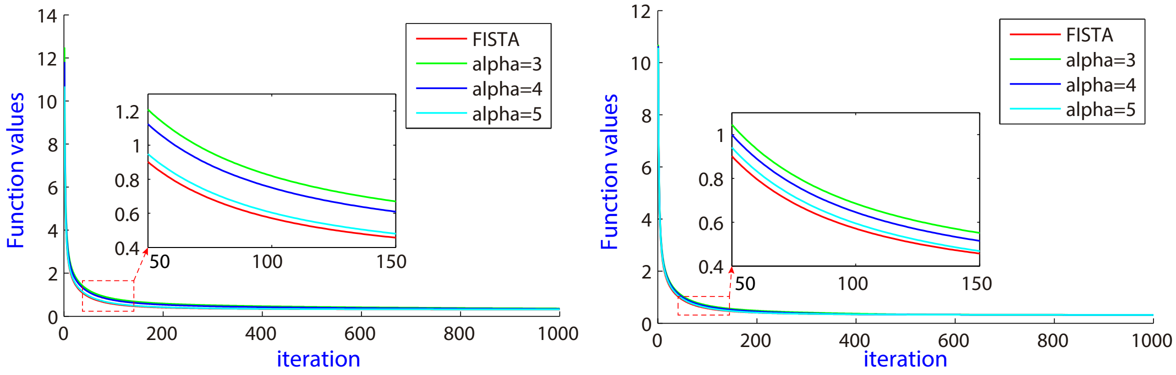


(a) (b) (c)

(d) (e) (f)

(g)

Figure 3: Deblurred images for different values of $\alpha = 3, 4$ and 5, and without any noise (a) Deblurring by FISTA; (b) Deblurring by inexact FISTA with deterministic noise and $\alpha = 3$; (c) Deblurring by inexact FISTA with deterministic noise and $\alpha = 4$; (d) Deblurring by inexact FISTA with deterministic noise and $\alpha = 5$; (e) Deblurring by inexact FISTA with stochastic noise and $\alpha = 3$; (f) Deblurring by inexact FISTA with stochastic noise and $\alpha = 4$; (g) Deblurring by inexact FISTA with stochastic noise and $\alpha = 5$.



(a)

(b)

Figure 4: Function values versus iteration. (a) Deterministic case; (b) Stochastic case.

where e is the noise. Thus, we can use the following model for the image deblurring

$$\min_{x} \frac{1}{2}\|(BF)x - b\|_2^2 + \sigma\|x\|_1, \tag{5.3}$$

where $\sigma > 0$ is the parameter.

In our tests we consider the cameraman test image. The codes of all algorithms are written entirely in MATLAB, and all the experiments are implemented under Windows 8 and MATLAB R2016a running on a laptop with an Intel Core i5 CPU (2.8 GHz) and 8 GB Memory. The scale of all images is 256×256. In this test, the blurring operator is generated as `B=fspecial('gaussian',15,5)`. And the wavelet F is chosen as the basic Haar wavelet. The maximum iteration number is set as 1000. Figure 2 shows the original and blurred images, and Figure 3 presents the results of the deblurring process by inexact FISTA algorithm with different values of α ($\alpha = 3, 4$ and 5) using deterministic and stochastic noises and without any noise (FISTA) to compare with. From the figures we observe that the inexact FISTA gives similar results as the case without noise, providing useful algorithms in noisy situations. Finally, Figure 4 reports the function values versus the iteration. In all simulations, both deterministic and stochastic noises in the algorithm are considered. We observe in the tests that the behaviour of the new inexact FISTA converges in all cases, being slightly faster when α grows, as shown in the theoretical analysis, and its behavior is close to the FISTA algorithm used without any noise.

From these preliminary tests, the new inexact FISTA algorithm presents a promising convergence behaviour to deblurring images with extra deterministic or stochastic noises.

6 Conclusion

In this paper, we study an inexact accelerated proximal gradient algorithm (a new inexact version of FISTA) considering both deterministic and stochastic error versions for solving linear inverse problems. We have shown that the catalyst framework can be regarded as a special case of the FISTA algorithm where the smooth part of the function vanishes. This new global FISTA framework provides generic convergence results, giving simpler proofs of the convergence in some cases studied previously, but also convergence results in cases not covered in literature, like the case of using stochastic errors or the catalyst framework. We present some preliminary numerical tests supporting the theoretical statements.

Acknowledgments

H.J. and L.C. have been supported by the National Science Foundation of China(No. 61402495). R.B. has been supported by the Spanish Research Project MTM2015-64095-P.

References

[1] Robert B Ash and Catherine Doleans-Dade. *Probability and measure theory.* Academic Press, 2000.

[2] Mingsian R. Bai, Chun Chung, Po-Chen Wu, Yi-Hao Chiang, and Chun-May Yang. Solution strategies for linear inverse problems in spatial audio signal processing. *Appl. Sci.*, 7:582, 2017.

[3] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] Åke Björck. *Numerical methods for least squares problems.* SIAM, 1996.

[6] Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[7] Antonin Chambolle, Ronald A De Vore, Nam-Yong Lee, and Bradley J Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7(3):319–335, 1998.

[8] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[9] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[10] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

[11] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[12] Paul Escande and Pierre Weiss. Sparse wavelet representations of spatially varying blurring operators. *SIAM Journal on Imaging Sciences*, 8(4):2976–3014, 2015.

[13] Mário AT Figueiredo and Robert D Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

[14] Mário AT Figueiredo, Robert D Nowak, and Stephen J Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007.

[15] Elaine T Hale, Wotao Yin, and Yin Zhang. A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing. *CAAM Technical Report TR07-07, Rice University, http://www.caam.rice.edu/ zhang/reports/tr0707.pdf*, 2007.

[16] Elaine T Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

[17] Kaifeng Jiang, Defeng Sun, and Kim-Chuan Toh. An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP. *SIAM Journal on Optimization*, 22(3):1042–1064, 2012.

[18] Jari Kaipio and Erkki Somersalo. Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504, 2007.

[19] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[20] Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

[21] A. Mohammad-Djafari. Inverse problems in imaging science: from classical regularization methods to state of the art bayesian methods. In *International Image Processing, Applications and Systems Conference*, pages 1–2, Nov 2014.

[22] Boris S Mordukhovich. *Variational analysis and generalized differentiation I: Basic theory*, volume 330. Springer Science & Business Media, 2006.

[23] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[24] Daniel Reem and Alvaro De Pierro. A new convergence analysis and perturbation resilience of some accelerated proximal forward–backward algorithms with errors. *Inverse Problems*, 33(4):044001, 2017.

[25] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[26] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[27] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

[28] Mark Schmidt, Nicolas L Roux, and Francis R Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

[29] Shai Shalev-Shwartz and Tong Zhang. Proximal stochastic dual coordinate ascent. *ArXiv:1211.2717*, 2012.

[30] Tao Sun and Lizhi Cheng. Reweighted fast iterative shrinkage thresholding algorithm with restarts for ℓ_1–ℓ_1 minimisation. *IET Signal Processing*, 10(1):28–36, 2016.

[31] Tao Sun, Hui Zhang, and Lizhi Cheng. Subgradient projection for sparse signal recovery with sparse noise. *Electronics Letters*, 50(17):1200–1202, 2014.

[32] Tao Sun, Hui Zhang, and Lizhi Cheng. Precondition techniques for accelerated linearized Bregman algorithms. *Pac. J. Optim*, 11(3):527–548, 2015.

[33] Silvia Villa, Saverio Salzo, Luca Baldassarre, and Alessandro Verri. Accelerated and inexact forward-backward algorithms. *SIAM Journal on Optimization*, 23(3):1607–1633, 2013.

[34] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[35] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.