ON THE PERFORMANCE OF THE STOCHASTIC FISTA

Adil Salim* *and Walid Hachem*†

$\star$ VCC, King Abdullah University of Science and Technology.
Thuwal 23955-6900, Kingdom of Saudi Arabia.
† CNRS / LIGM (UMR 8049), Université Paris-Est Marne-la-Vallée.
5, boulevard Descartes, Champs-sur-Marne, 77454, Marne-la-Vallée Cedex 2, France.

ABSTRACT

In the field of convex optimization, one is frequently lead to solve a composite minimization problem involving a smooth function F and a non smooth function G. In this context, it is well known that the classical proximal gradient algorithm can be accelerated using the Nesterov technique, which leads to the celebrated FISTA. This paper investigates a stochastic version of FISTA that can be applied to the case where the function F is intractable, but can be represented as an expectation. Although it is rather well known that in the stochastic case, the Nesterov acceleration does not bring a clear advantage on the long term over the averaged proximal gradient algorithm, it is demonstrated in this paper that it is beneficial during the first iterations. This argues in favor of the Nesterov acceleration in many situations in machine learning, where only a few algorithm iterations are required. The technique is based on the study of a Lyapounov function which is inspired by the ODE method applied to the Nesterov acceleration.

Index Terms— FISTA, Nesterov acceleration, Stochastic optimization

1. INTRODUCTION

Many applications in the fields of machine learning and signal processing [1, 2, 3], require the solution of the composite programming problem

$$\min_{x \in \mathsf{X}} F(x) + G(x) \tag{1}$$

where X is an Euclidean space, and F and G belong to the set $\Gamma_0(\mathsf{X})$ of convex, lower semi-continuous and proper functions. In these contexts, F often represents a smooth cost function and G a non smooth regularization term. To solve (1), the proximal gradient algorithm is a standard method whose iterations can be written

$$x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma \nabla F(x_n)), \tag{2}$$

where $\gamma > 0$ is a step size, and where $\mathrm{prox}_{\gamma G}(x) = \arg\min_{y \in \mathsf{X}} \gamma G(y) + \frac{1}{2}\|x - y\|^2$ is the so-called proximity operator of γG. This algorithm is known to converge to a minimizer $x_\star$ of $F + G$ at the rate $(F+G)(x_n) - (F+G)(x_\star) = \mathcal{O}(1/n)$. By means of an additional relaxation step, called Nesterov acceleration, the algorithm becomes the so-called Fast Iterative Shrinkage-Thresholding Algorithm (FISTA, [4, 5]) which achieves the convergence rate $\mathcal{O}(1/n^2)$ that cannot be improved by another first order method [6, 7]. The exploration of Nesterov's acceleration technique aroused many works, see [8, 9, 10, 11, 12] among others. One interesting approach is provided by Su *et al.* [8], that uses a dynamical systems technique to explain Nesterov acceleration. More precisely, the sequence of iterates of FISTA is seen here as the discretization of the solution of a non autonomous second order ordinary differential equation (ODE). This ODE has been extensively studied in the literature, see *e.g* [13, 14]. Using a Lyapunov function technique, it can be shown that $(F + G)(\mathsf{x}(t)) - (F + G)(x_\star) = \mathcal{O}(1/t^2)$ where $x_\star$ is a minimizer of $F + G$ and x is a solution to the differential equation. Inspired by this Lyapunov function technique, [8] uses a discrete Lyapunov function to rederive the $\mathcal{O}(1/n^2)$ convergence rate of FISTA.

As it is often the case in machine learning and signal processing, this paper considers the resolution of (1) under a stochastic first order oracle model. In this model, the function F is represented as an expectation $F(x) = \mathbb{E}_\xi(f(\xi, x))$, where ξ is a random variable (r.v) and $f(\xi, \cdot) \in \Gamma_0(\mathsf{X})$. Moreover, the distribution of ξ is revealed across time to the user through i.i.d copies $(\xi_n)_n$ of ξ. In many applications, the behavior of optimization algorithms in the first iterations is of particular interest. For example, in empirical risk minimization, there is no need to optimize below the so-called statistical error [15]. Besides, early stopping techniques [16] rely on finding a rough estimate of a minimizer. Therefore, we shall focus on the behavior of optimization algorithms, in the first order stochastic oracle model, specifically in the first iterations.

To solve Problem (1) in the first order oracle model, the proximal stochastic gradient algorithm consists in replacing at each iteration of the proximal gradient algorithm the true gradient $\nabla F(x_n)$ by a stochastic realization $\nabla_x f(\xi_{n+1}, x_n)$.

After averaging the iterates, this algorithm achieves the convergence rate $\mathcal{O}(1/\sqrt{n})$ in expectation [2] when used with a constant step size over n iterations (the step size depends on n). Although this convergence rate is optimal ([17]), this algorithm is known to be close to its deterministic $\mathcal{O}(1/n)$ behavior in the first iterations [2, 18]. In this work, the stochastic FISTA is studied. Each iteration of this algorithm is a proximal stochastic gradient step followed by a Nesterov acceleration step. It can be written

$$x_{n+1} = \operatorname{prox}_{\gamma_{n+1}G}(y_n - \gamma_{n+1}\nabla_y f(\xi_{n+1}, y_n)) \tag{3}$$

$$y_{n+1} = x_{n+1} + \frac{n}{n+r}(x_{n+1} - x_n). \tag{4}$$

where $r \geq 3$ and (γ_n) is a sequence of positive and non increasing step sizes. We provide convergence rates for the stochastic FISTA using a stochastic Lyapunov technique inspired from [8]. When used with a constant step size, the algorithm achieves the optimal $\mathcal{O}(1/\sqrt{n})$ convergence rate without averaging, and with decreasing step sizes stochastic FISTA achieves the rate $\mathcal{O}(\log(n)/\sqrt{n})$. More importantly, it will be shown that the stochastic FISTA is close to its $\mathcal{O}(1/n^2)$ deterministic behavior in the first iterations. Therefore, it is faster than the proximal stochastic gradient algorithm in the beginning of the algorithms. This fact is supported by numerical experiments on a logistic regression task.

A perturbed FISTA is also studied [3]. The paper [3] proves convergence rates for the perturbed FISTA in the case where the true gradient $\nabla F(y_n)$ is replaced at each iteration by a Monte Carlo approximation H_{n+1}. We will rather concentrate on the first order oracle model in which our Lyapunov type analysis allows to derive sharper bounds. In particular, we can better understand the behavior of the algorithm in its first iterations.

2. CONVERGENCE RATE

Consider a probability space $(\Xi, \mathscr{G}, \mu)$ and a r.v. ξ with distribution μ defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The mathematical expectation is denoted $\mathbb{E}$ and the variance $\mathbb{V}$. Consider a sequence of i.i.d copies (ξ_n) of ξ and a deterministic $x_0 = y_0 \in \mathsf{X}$. We posit the following assumptions.

Assumption 1. For every $s \in \Xi$, $f(s, \cdot)$ and G are in $\Gamma_0(\mathsf{X})$. Moreover, for every $x \in \mathsf{X}$, $f(\cdot, x)$ is measurable and μ-integrable.

Assumption 2. For every $s \in \Xi$, $f(s, \cdot)$ is differentiable, and there exists $\sigma \geq 0$ such that for every $x \in \mathsf{X}$,

$$\mathbb{V}_\xi(\|\nabla f(\xi, x)\|) \leq \sigma^2.$$

Assumption 3. The function $F(x) = \mathbb{E}_\xi(f(\xi, x))$ is differentiable and there exists $L > 0$ such that its gradient ∇F is L-Lipschitz continuous.

Assumption 4. The function $H = F + G$ admits a minimizer $x_\star$ and we denote $H(x_\star) = H_\star$.

Theorem 1. Consider the iterates (x_n) of the stochastic FISTA (3)-(4) and let assumptions 1–4 hold true. If $r > 3$ and if the sequence $(n\gamma_n)$ is square summable, then, almost surely (a.s.) the sequence $(n\gamma_n(H(x_n) - H_\star))$ is summable and the sequence $(n^2\gamma_n(H(x_n) - H_\star))$ is bounded.

Theorem 2. Assume that $r \geq 3$, and that assumptions 1–4 hold true. Then,

$$\begin{aligned}\mathbb{E}H(x_n) - H_\star \leq &\frac{\gamma_0(r-2)^2}{\gamma_n(n+r-2)^2}\left(H(x_0) - H(x_\star)\right) \\ &+ \frac{(r-1)^2}{2\gamma_n(n+r-2)^2}\|x_0 - x_\star\|^2 \\ &+ \sigma^2\sum_{k=1}^{n}\frac{\gamma_k^2}{2\gamma_n(1-L\gamma_k)}\frac{(k+r-2)^2}{(n+r-2)^2}.\end{aligned} \tag{5}$$

Moreover, the step sizes that minimize the upper bound can be written $\gamma_n = C/(n+r-2)^{3/2}$ where $C > 0$ is such that $\gamma_0 L \leq 0.1$. In this case

$$\begin{aligned}\mathbb{E}H(x_n) - H_\star \leq &\frac{\sqrt{r-2}(H(x_0) - H_\star)}{\sqrt{n+r-2}} + \frac{(r-1)^2\|x_0 - x_\star\|^2}{2C\sqrt{n+r-2}} \\ &+ \frac{5\sigma^2}{9}\frac{C\log(n+r-2)}{\sqrt{n+r-2}}.\end{aligned} \tag{6}$$

Theorem 3. Assume that $r \geq 3$, and that assumptions 1–4 hold true. If moreover the step size is constant $\gamma_n \equiv \gamma > 0$, then,

$$\begin{aligned}\mathbb{E}H(x_n) - H_\star \leq &\frac{(r-2)^2}{(n+r-2)^2}(H(x_0) - H_\star) \\ &+ \frac{(r-2)^2}{2\gamma(n+r-2)^2}\|x_0 - x_\star\|^2 \\ &+ \frac{2\sigma^2}{3}\frac{\gamma}{1-L\gamma}(n+r-1).\end{aligned} \tag{7}$$

If $G \equiv 0$, we have the slightly more precise result

$$\begin{aligned}\mathbb{E}F(x_n) - F_\star \leq &\frac{(r-2)^2}{(n+r-2)^2}(F(x_0) - F_\star) \\ &+ \frac{(r-2)^2}{2\gamma(n+r-2)^2}\|x_0 - x_\star\|^2 \\ &+ \frac{2\sigma^2}{3}\gamma(1+L\gamma)(n+r-1).\end{aligned} \tag{8}$$

Let $n \geq 1$ and assume the stochastic FISTA is run over n iterations with a constant step size γ. Then, the step size that minimizes the upper bound can be written $\gamma = C/(n+r-2)^{3/2}$ where $C > 0$. In this case, if $\gamma L \leq 0.1$, the n^{th} iterate of stochastic FISTA satisfies

$$\mathbb{E}H(x_n) - H_\star \leq \frac{(r-2)^2}{(n+r-2)^2}(H(x_0) - H_\star) + \frac{(r-2)^2\|y_0 - x_\star\|^2}{2C\sqrt{n+r-2}} + \frac{40}{27}\frac{C\sigma^2}{\sqrt{n+r-2}}. \tag{9}$$

3. DERIVATION OF THE CONVERGENCE RATES

This section is devoted to the proof of the theorems 1–3.

Lemma 4. Let assumptions 1–4 hold true. For every $n \geq 1$, consider the random variable (also called the stochastic Lyapunov function)

$$V_n := \frac{2\gamma_n(n+r-2)^2}{r-1}(H(x_n) - H_\star) + (r-1)\|z_n - x_\star\|^2$$

where $z_n := \frac{n+r-1}{r-1}y_n - \frac{n}{r-1}x_n$ and where x_n and y_n are defined by Equations (3)-(4). Then

$$\mathbb{E}_n(V_{n+1}) \leq V_n + \frac{2}{r-1}(\gamma_{n+1}n(n+r-1) - \gamma_n(n+r-2)^2)(H(x_n) - H_\star) + \frac{{\gamma_{n+1}}^2}{1-L\gamma_{n+1}}\frac{(n+r-1)^2}{r-1}\sigma^2.$$

where $\mathbb{E}_n$ denotes the conditional expectation with respect to the sigma-field $\sigma(\xi_1, \ldots, \xi_n)$. Moreover, if $G \equiv 0$,

$$\mathbb{E}_n(V_{n+1}) \leq V_n + \frac{2}{r-1}(\gamma_{n+1}n(n+r-1) - \gamma_n(n+r-2)^2)(F(x_n) - F_\star) + \frac{{\gamma_{n+1}}^2(1+L\gamma_{n+1})(n+r-1)^2}{r-1}\sigma^2 - \frac{{\gamma_{n+1}}^2(1-L\gamma_{n+1})(n+r-1)^2}{r-1}\|\nabla F(y_n)\|^2$$

The proof of this lemma is postponed to the appendix 5.1. Recalling that $r \geq 3$ and that (γ_n) is positive and non-increasing,

$$\begin{aligned}&\gamma_{n+1}n(n+r-1) - \gamma_n(n+r-2)^2 \\ &\leq \gamma_{n+1}n(n+r-1) - \gamma_{n+1}(n+r-2)^2 \\ &= -\gamma_{n+1}(1 + (r-3)(n+r-1)).\end{aligned}$$

Applying Robbins-Siegmund lemma [19] to the inequality of lemma 4 we get that if $r > 3$ and if $n\gamma_n$ is square summable, then $n\gamma_n(H(x_n) - H_\star)$ is summable and $n^2\gamma_n(H(x_n) - H_\star)$ is bounded a.s. Hence, theorem 1 is proven. Moreover, taking the expectation in the inequality of lemma 4, and iterating we get

$$\mathbb{E}(V_n) \leq V_0 + \sum_{k=0}^{n-1}\frac{\gamma_{k+1}^2}{1-L\gamma_{k+1}}\frac{(k+r-1)^2}{r-1}\sigma^2$$

Hence,

$$\mathbb{E}H(x_n) - H_\star \leq \frac{(r-1)V_0}{2\gamma_n(n+r-2)^2} + \sigma^2\sum_{k=1}^{n}\frac{\gamma_k^2}{2\gamma_n(1-L\gamma_k)}\frac{(k+r-2)^2}{(n+r-2)^2}. \tag{10}$$

Due to a lack of space, we do not enter into details in this part of the proof. If $\gamma_n = C/(n+r-2)^{3/2}$ where $C > 0$ is such that $\gamma_1 L \leq 0.1$, it is easily seen that

$$\mathbb{E}H(x_n) - H_\star \leq \frac{(r-1)V_0}{2C\sqrt{n+r-2}} + \frac{5\sigma^2}{9}\frac{C\log(n+r-2)}{\sqrt{n+r-2}}. \tag{11}$$

Hence, theorem 2 is proven. Finally, consider a constant step size $\gamma_n \equiv \gamma \in (0, 0.1/L)$. Then, we have for every $n \geq 1$,

$$\mathbb{E}H(x_n) - H_\star \leq \frac{(r-1)V_0}{2\gamma(n+r-2)^2} + \frac{2\sigma^2}{3}\frac{\gamma}{1-L\gamma}(n+r-1) \tag{12}$$

Note that if $G \equiv 0$, the second term a the right hand side is replaced by $\frac{2\sigma^2}{3}\gamma(1+L\gamma)(n+r-1)$.

Finally, assume that a fixed step size γ is taken over $n \geq 1$ iterations of the algorithm. Setting $\gamma = C/(n+r-2)^{3/2}$ where $C > 0$ is such that $\gamma L \leq 0.1$, we have,

$$\mathbb{E}H(x_n) - H_\star \leq \frac{(r-2)^2}{(n+r-2)^2}(H(x_0) - H_\star) + \frac{(r-2)^2\|x_0 - x_\star\|^2}{2C\sqrt{n+r-2}} + \frac{40}{27}\frac{C\sigma^2}{\sqrt{n+r-2}}, \tag{13}$$

and theorem 3 is proven.

4. BEHAVIOR IN THE FIRST ITERATIONS

In this section, we show that stochastic FISTA is faster than the proximal stochastic gradient algorithm in the first iterations. This is because both algorithms are close to their deterministic behavior in the first iterations, and the FISTA has convergence rate $\mathcal{O}(1/n^2)$ which is faster that the $\mathcal{O}(1/n)$ of the proximal gradient algorithm. We start by the following result that is the counterpart of theorem 3 for the proximal stochastic gradient algorithm.

Theorem 5.

The proof of this result can be found in [2] under the additional assumption that $(x_n)_n$ is a.s a bounded sequence.The proof of this theorem is postponed to Appendix 5.2

Verifier cela

4.1. Upper bounds

We first perform an analysis based on the upper bounds in theorems 3 and 5. This approach is ligitimate by the fact that these bounds are sharp. If the step size $\gamma \leq 0.1/L$ is constant, the last term at the right hand side of (7) or (8) is lower than the other terms at the beginning of the algorithm (if n is enough small). Loosely speaking, we have in this case $\mathbb{E}H(x_n) - H_\star \leq K/(n+r-2)^2$. For the proximal stochastic gradient algorithm, we rather have $\mathbb{E}H(\overline{x_n}) - H_\star \leq K/n$.

In Figure 1 we illustrate and compare the upper bounds provided by theorems 5 and 3 (with $r = 3$). For different level σ^2 of noise, it is seen that the constant step proximal stochastic gradient algorithm is slower in the first iterations than the stochastic FISTA.

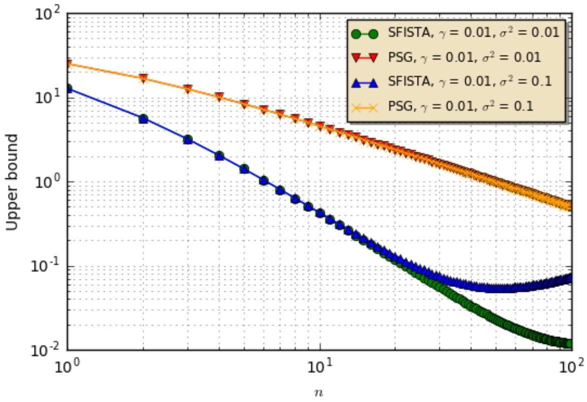

Fig. 1. The upper bounds as a function of n for the proximal stochastic gradient (PSG) algorithm (after averaging) and the stochastic FISTA (SFISTA)

4.2. Simulations

In this section, we consider $\lambda > 0$ and the minimization of $F(x) + G(x)$ where $F(x) = \mathbb{E}(f(\xi, x))$ is the cost function associated with the logistic regression and $G(x) = \lambda\|x\|_1$ (*i.e* a Lasso regularization). More precisely, the user is provided with i.i.d copies of a r.v. $\xi = (X, Y)$ online and $f(\xi, x) = \ell(Y\langle x, X\rangle)$ where $\ell(u) = \log(1 + \exp(-u))$. To solve this problem we compare the averaged proximal stochastic gradient algorithm [2] to the stochastic FISTA in Figure 2.

The regression task is performed over the Diabetic Retinopathy dataset[1] and the code is available on Github[2]. We simulate each algorithm with a constant step size ranging from 0.01 to 10.0 and plot the best performing curves. Each simulation is done ten times and the mean curve is represented over 100 iterations. The first and the last deciles are also showed in Figure 2 (the curves above and below the mean curves). We set the parameter λ to 0.1 and $r = 3.5$ for stochastic FISTA.

Despite exhibiting more instability, the stochastic FISTA is (as expected) faster than the proximal stochastic gradient in the first iterations for each step size.

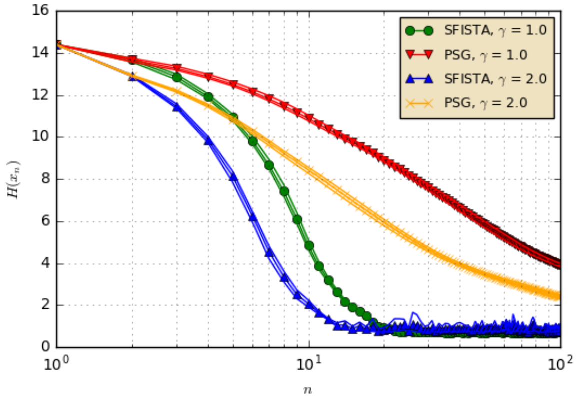

Fig. 2. The objective function $F + G$ as a function of n for each algorithm. PSG denotes the averages stochastic gradient and SFISTA denotes the stochastic FISTA.

5. APPENDIX

5.1. Proof of lemma 4

This appendix is devoted to the proof of lemma 4. For every $\gamma > 0$, consider the so-called Moreau's envelope of G defined by ${}^\gamma G(x) = \arg\min_{y \in \mathsf{X}} G(y) + \frac{1}{2\gamma}\|x - y\|^2$. It is known that ${}^\gamma G \in \Gamma_0(\mathsf{X})$ is differentiable with a $1/\gamma$-Lipschitz continuous gradient, and that $\nabla^\gamma G(x) = 1/\gamma(x - \mathrm{prox}_{\gamma G}(x)) \in \partial G(\mathrm{prox}_{\gamma G}(x))$ where ∂G denotes the subdifferential of G. Consider

$$T_\gamma(\xi, x) = \nabla f(\xi, x) + \nabla^\gamma G(x - \gamma \nabla f(\xi, x)).$$

With this notation, $\mathrm{prox}_{\gamma G}(x - \gamma\nabla f(\xi, x)) = x - \gamma T_\gamma(\xi, x)$. Using Assumption 3,

$$\begin{aligned} F(y - \gamma T_\gamma(\xi, y)) \leq & F(y) - \gamma\langle \nabla F(y), T_\gamma(\xi, y)\rangle \\ & + \frac{L}{2}\gamma^2\|T_\gamma(\xi, y)\|^2 \end{aligned}$$

for every $y \in \mathsf{X}, \gamma > 0, \xi \in \Xi$. Besides, using Assumption 1,

$$\begin{aligned} F(y) \leq & F(x) + \langle \nabla F(y), y - x\rangle \quad \text{and} \\ G(y - \gamma T_\gamma(\xi, y)) \leq & G(x) \\ & + \langle \nabla^\gamma G(y - \gamma\nabla f(\xi, y)), y - x - \gamma T_\gamma(\xi, y)\rangle. \end{aligned}$$

Summing the three last inequalities,

$$\begin{aligned} H(y - \gamma T_\gamma(\xi, y)) \leq & H(x) + \frac{L}{2}\gamma^2\|T_\gamma(\xi, y)\|^2 \qquad (14) \\ & + \langle \nabla F(y) + \nabla^\gamma G(y - \gamma\nabla f(\xi, y)), y - x\rangle \\ & - \gamma\langle \nabla F(y) + \nabla^\gamma G(y - \gamma\nabla f(\xi, y)), T_\gamma(\xi, y)\rangle. \end{aligned}$$

[1] http://www.it.lut.fi/project/imageret/diaretdb1/

[2] https://github.com/adil-salim/

Setting $x \equiv x_n$, $y \equiv y_n$, $\gamma \equiv \gamma_{n+1}$, $\xi \equiv \xi_{n+1}$ and denoting $\nabla^{\gamma_{n+1}} G \equiv \nabla^{\gamma_{n+1}} G(y_n - \gamma_{n+1} \nabla f(\xi_{n+1}, y_n))$

$$\begin{aligned} H(y_n - \gamma_{n+1} T_{\gamma_{n+1}}(\xi_{n+1}, y_n)) \leq H(x_n) \\ + \frac{L}{2} \gamma_{n+1}^2 \|T_{\gamma_{n+1}}(\xi_{n+1}, y_n)\|^2 \\ + \langle \nabla F(y_n) + \nabla^{\gamma_{n+1}} G, y_n - x_n \rangle \\ - \gamma_{n+1} \langle \nabla F(y_n) + \nabla^{\gamma_{n+1}} G, T_{\gamma_{n+1}}(\xi_{n+1}, y_n) \rangle. \end{aligned} \tag{15}$$

Setting $x \equiv x_\star$, $y \equiv y_n$, $\gamma \equiv \gamma_{n+1}$ and $\xi \equiv \xi_{n+1}$, we obtain

$$\begin{aligned} H(y_n - \gamma_{n+1} T_{\gamma_{n+1}}(\xi_{n+1}, y_n)) \leq H_\star \\ + \frac{L}{2} \gamma_{n+1}^2 \|T_{\gamma_{n+1}}(\xi_{n+1}, y_n)\|^2 \\ + \langle \nabla F(y_n) + \nabla^{\gamma_{n+1}} G, y_n - x_\star \rangle \\ - \gamma_{n+1} \langle \nabla F(y_n) + \nabla^{\gamma_{n+1}} G, T_{\gamma_{n+1}}(\xi_{n+1}, y_n) \rangle. \end{aligned} \tag{16}$$

Noting that

$$z_{n+1} = z_n - \gamma_{n+1} \frac{n+r-1}{r-1} T_{\gamma_{n+1}}(\xi_{n+1}, y_n), \tag{17}$$

the average $n/(n+r-1) \times$(15)+$(r-1)/(n+r-1) \times$(16) of the above inequalities leads to

$$\begin{aligned} H(x_{n+1}) - H_\star \leq \frac{n}{n+r-1}(H(x_n) - H_\star) \\ + \frac{r-1}{n+r-1} \langle \nabla F(y_n) + \nabla^{\gamma_{n+1}} G, z_{n+1} - x_\star \rangle \\ + \frac{L}{2} \gamma_{n+1}^2 \|T_{\gamma_{n+1}}(\xi_{n+1}, y_n)\|^2. \end{aligned} \tag{18}$$

Since

$$\begin{aligned} \|z_{n+1} - x_\star\|^2 =& \|z_n - x_\star\|^2 \\ &+ 2\langle z_{n+1} - z_n, z_{n+1} - x_\star \rangle - \|z_{n+1} - z_n\|^2, \end{aligned} \tag{19}$$

the linear combination of inequalities $2\gamma_{n+1}(n+r-1)^2/(r-1) \times$(18)+$(r-1) \times$(19) leads to

$$\begin{aligned} &\frac{2\gamma_{n+1}(n+r-1)^2}{r-1}(H(x_{n+1}) - H_\star) + (r-1)\|z_{n+1} - x_\star\|^2 \\ &\leq \frac{2\gamma_{n+1} n(n+r-1)}{r-1}(H(x_n) - H_\star) + (r-1)\|z_n - x_\star\|^2 \\ &- 2\gamma_{n+1}(n+r-1)\langle \nabla f(\xi_{n+1}, y_n) - \nabla F(y_n), z_{n+1} - x_\star \rangle \\ &- (1 - L\gamma_{n+1}) \frac{(n+r-1)^2}{r-1} \|\gamma_{n+1} T_{\gamma_{n+1}}(\xi_{n+1}, y_n)\|^2. \end{aligned}$$

Denoting χ_{n+1} the sum of the last two terms, we finally have

$$\begin{aligned} &V_{n+1} \leq V_n \\ &\quad + \frac{2}{r-1}(\gamma_{n+1} n(n+r-1) - \gamma_n(n+r-2)^2)(H(x_n) - H_\star) \\ &\quad + \chi_{n+1}. \end{aligned}$$

To control χ_{n+1}, first consider the case where $G \equiv 0$. Recalling (17), and that $\mathbb{E}_n \langle \nabla f(\xi_{n+1}, y_n) - \nabla F(y_n), z_n - x_\star \rangle = 0$,

$$\begin{aligned} \mathbb{E}_n \chi_{n+1} = (1 + L\gamma_{n+1}) \gamma_{n+1}^2 \frac{(n+r-1)^2}{r-1} \sigma^2 \\ - (1 - L\gamma_{n+1}) \gamma_{n+1}^2 \frac{(n+r-1)^2}{r-1} \|\nabla F(y_n)\|^2. \end{aligned}$$

In the general case, note that, if $L\gamma_{n+1} < 1$,

$$\begin{aligned} &|2\langle \nabla f(\xi_{n+1}, y_n) - \nabla F(y_n), \gamma_{n+1} T_{\gamma_{n+1}}(\xi_{n+1}, y_n) \rangle| \\ &\leq \frac{1 - L\gamma_{n+1}}{\gamma_{n+1}} \|\gamma_{n+1} T_{\gamma_{n+1}}(\xi_{n+1}, y_n)\|^2 \\ &+ \frac{\gamma_{n+1}}{1 - L\gamma_{n+1}} \|\nabla f(\xi_{n+1}, y_n) - \nabla F(y_n)\|^2. \end{aligned}$$

Using (17) and $\mathbb{E}_n \langle \nabla f(\xi_{n+1}, y_n) - \nabla F(y_n), z_n - x_\star \rangle = 0$, we have

$$\mathbb{E}_n \chi_{n+1} \leq \frac{{\gamma_{n+1}}^2}{1 - L\gamma_{n+1}} \frac{(n+r-1)^2}{r-1} \sigma^2,$$

and lemma 4 is proven.

5.2. Proof of theorem 5

6. REFERENCES

[1] Y. F. Atchade, G. Fort, and E. Moulines, “On stochastic proximal gradient algorithms,” *ArXiv e-prints, 1402.2365*, Feb. 2014.

[2] Y. F Atchadé, G. Fort, and E. Moulines, “On perturbed proximal gradient algorithms,” *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 310–342, 2017.

[3] G. Fort, L. Risser, Y. F. Atchadé, and E. Moulines, “Stochastic fista algorithms: So fast?,” in *SSP 2018*. IEEE, 2018, pp. 796–800.

[4] Y. E Nesterov, “A method for solving the convex programming problem with convergence rate $1/k^2$,” in *Dokl. Akad. Nauk SSSR*, 1983, vol. 269, pp. 543–547.

[5] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[6] A. S. Nemirovsky and D. B. Yudin, “Problem complexity and method efficiency in optimization,” 1983.

[7] Y. E Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2013.

[8] W. Su, S. Boyd, and E. Candes, “A differential equation for modeling nesterov’s accelerated gradient method: Theory and insights,” in *NIPS*, 2014, pp. 2510–2518.

[9] S. Bubeck, Y. T. Lee, and M. Singh, “A geometric alternative to nesterov’s accelerated gradient descent,” *arXiv preprint arXiv:1506.08187*, 2015.

[10] A. Wibisono, A. C. Wilson, and M. I Jordan, “A variational perspective on accelerated methods in optimization,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 47, pp. E7351–E7358, 2016.

[11] H. Attouch and J. Peypouquet, “The rate of convergence of nesterov’s accelerated forward-backward method is actually faster than $1/k^2$,” *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1824–1834, 2016.

[12] D. Scieur, V. Roulet, F. Bach, and A. d’Aspremont, “Integration methods and accelerated optimization algorithms,” *arXiv preprint arXiv:1702.06751*, 2017.

[13] H. Attouch, Z. Chbani, and H. Riahi, “Combining fast inertial dynamics for convex optimization with tikhonov regularization,” *Journal of Mathematical Analysis and Applications*, vol. 457, no. 2, pp. 1065–1094, 2018.

[14] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont, “Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity,” *Mathematical Programming*, vol. 168, no. 1-2, pp. 123–175, 2018.

[15] L. Bottou and O. Bousquet, “The tradeoffs of large scale learning,” in *NIPS*, 2008, pp. 161–168.

[16] L. Bottou, F. E Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[17] A. Agarwal, M. J Wainwright, P. L Bartlett, and P. K Ravikumar, “Information-theoretic lower bounds on the oracle complexity of convex optimization,” in *NIPS*, 2009, pp. 1–9.

[18] E. Moulines and F. Bach, “Non-asymptotic analysis of stochastic approximation algorithms for machine learning,” in *NIPS*, 2011, pp. 451–459.

[19] H. Robbins and D. Siegmund, “A convergence theorem for non negative almost supermartingales and some applications,” in *Optimizing Methods in Statistics*, pp. 233–257. Academic Press, New York, 1971.