Stochastic FISTA Step Search Algorithm for Convex Optimization

Trang H. Tran[1] · Lam M. Nguyen[2] · Katya Scheinberg [1]

[1] Cornell University, School of Operations Research and Information Engineering
[2] IBM Research, Thomas J. Watson Research Center

Problem Statement

We consider the stochastic optimization problem

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where exact values of the function and its gradient may not be computed exactly. Instead we assume access to some stochastic oracles which return these estimations.
We make the following standard assumption.
Assumption 1. We assume that f is convex and L-smooth.

Now we define the oracles which our algorithm will call to estimate the values of our function and its gradient.
(i) **Stochastic first-order oracle (SFO(κ_g, δ_1)).** Given a point x, the oracle computes $g(x, \xi')$, a (random) estimate of the gradient $\nabla f(x)$, such that

$$\mathbb{P}_{\xi'}\left[\|g(x,\xi') - \nabla f(x)\| \le \kappa_g\|\nabla f(x)\|\right] \ge 1-\delta_1. \tag{2}$$

(ii) **Stochastic zeroth-order oracle (SZO$(\kappa_f, \alpha, \|g\|, \delta_2)$).** Given a point x and a constant $\alpha > 0$, the oracle computes $\tilde{f}(x, \xi)$, a (random) estimate of the function value $f(x)$ that satisfies

$$\mathbb{P}_{\xi}\left[|\tilde{f}(x,\xi) - f(x)| \le \kappa_f\alpha^2\|g\|^2\right] \ge 1-\delta_2. \tag{3}$$

Stochastic FISTA Step Search Algorithm

Our stochastic algorithm is based on the variant of FISTA algorithm [1] introduced in [7] for deterministic exact optimization. **Compared to [7], we keep track of prior successful updates in x_k^{prev} and use stochastic estimates.**

Algorithm 1: Stochastic FISTA Step Search Algorithm

1: Initialization: Choose an initial point $x_0 \in \mathbb{R}^d$. Set $t_1 = 1, t_0 = 0$, let $x_0^{prev} = x_0$. Set $0 < \gamma < 1$, $\theta_0 = \gamma$ with $\alpha_1 > 0$. Choose $\eta \in [1/2, 1]$.
2: **for** $k = 1, 2, \cdots$, **do**
3: Compute $(t_k^0, y_k) = \text{FISTAStep}(x_{k-1}, x_{k-1}^{prev}, t_{k-1}, \theta_{k-1})$
4: Compute the gradient estimate $g_k = g(y_k, \xi_k')$ using the stochastic first-order oracle and the reference point $p_{\alpha_k}(y_k) = y_k - \alpha_k g_k$.
5: Compute the function estimates $f_k^y = \tilde{f}(y_k, \xi_k)$ and $f_k^p = \tilde{f}(p_{\alpha_k}(y_k), \xi_k)$ using the stochastic zeroth-order oracle.
6: Check sufficient decrease condition $f_k^p \le f_k^y - \alpha_k\eta\|g_k\|^2$.
7: If unsuccessful: Set $(x_k, x_k^{prev}) = (x_{k-1}, x_{k-1}^{prev})$ and $t_k = t_{k-1}$, then set $\alpha_{k+1} = \gamma\alpha_k$ and $\theta_k = \frac{\theta_{k-1}}{\gamma}$.
8: If successful: Set $(x_k, x_k^{prev}) = (p_{\alpha_k}(y_k), x_{k-1})$ and $t_k = t_k^0$, then set $\alpha_{k+1} = \frac{\alpha_k}{\gamma}$, and $\theta_k = \frac{\alpha_k}{\alpha_{k+1}}$.
9: **end for**

Step $(t_{k+1}^0, y_{k+1}) = \text{FISTAStep}(x_k, x_k^{prev}, t_k, \theta_k)$ is defined as follows.

$$t_{k+1}^0 = \frac{1}{2}\left(1+\sqrt{1+4\theta_k t_k^2}\right), \text{ and } y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}^0}(x_k - x_k^{prev}). \tag{4}$$

Problem Setting

Filtration. To formalize the conditioning on the past, let $\mathcal{F}_{k-1}^{G\cdot F}$ denote the σ-algebra generated by the random variables $G_0, \ldots, G_{k-1}$ (gradient estimates) and $F_0^y, F_0^p, \ldots, F_{k-1}^y, F_{k-1}^p$ (function estimates). Therefore $\{\mathcal{F}_{k-1}^{G\cdot F}\}_{k\ge1}$ is a filtration.
Epsilon accurate solution and the hitting time N_ϵ to reach ϵ accuracy. Let x^* be a global minimizer of f and $f_* > -\infty$ is the minimum value of f over its domain. We say that x is an ϵ-accurate solution of problem $\min_{x\in\mathbb{R}^n} f(x)$ if it satisfies $f(x) - f_* \le \epsilon$.
For a level of accuracy ϵ, we aim to derive a bound on the expected number of iterations the algorithm needs to reach the given accuracy level. We define N_ϵ as the hitting time for the event $\{f(X_k) - f_* \le \epsilon\}$, where k is the iteration of the algorithm.
True Iteration. We say iteration k is true if $\|g_k - \nabla f(y_k)\| \le \kappa_g\|\nabla f(y_k)\|$ and $|f_k^y - f(y_k)| \le \kappa_f\alpha_k^2\|g_k\|^2$, $|f_k^p - f(p_{\alpha_k}(y_k))| \le \kappa_f\alpha_k^2\|g_k\|^2$, respectively. We use a variable I_k^{true} to indicate that iteration k is true.
The condition on the gradient is known as the norm condition [3, 2, 6].
With these definitions, we present our key Assumption 2.
Assumption 2.
(i) For every iteration k, the following bound holds:

$$\mathbb{E}\left[\|(\nabla f(y_k) - g_k)\|^2|\mathcal{F}_{k-1}^{G\cdot F}\right] \le \frac{1}{\alpha_k^2 t_k^2\cdot k^{2+\beta}} \tag{5}$$

$$\mathbb{E}[(f_k^y - f(y_k))|\mathcal{F}_{k-1}^{G\cdot F}] \le \frac{1}{\alpha_k t_k^2\cdot k^{1+\beta'}}; \quad \mathbb{E}[(f(x_k) - f_k^p)|\mathcal{F}_{k-1}^{G\cdot F}] \le \frac{1}{\alpha_k t_k^2\cdot k^{1+\beta'}} \tag{6}$$

for some $\beta > 0$ and $\beta' > 0$.
(ii) Let $\eta \in [1/2, 1]$ and $\kappa_g \le \frac{1-\eta}{2-\eta} \le \frac{1}{3}$. There exists a probability $p \in (0.5, 1]$ such that:

$$\mathbb{P}(I_k^{\text{true}}|\mathcal{F}_{k-1}^{G\cdot F}) \ge p \text{ for all } k \ge 0 \tag{7}$$

Assumption 2(i) implies that variance of gradient noise decrease with a particular rate, where α_k has constant order and t_k scales with order k. Therefore this assumption is the same as the rate of deterministic gradient error [8], however, our condition is in expectation. Assumption 2(ii) is used in prior work [4, 2, 5] and is needed to ensure that successful iterations happen sufficiently often when α_k is small enough, which is in the next Lemma.

Bound on the Expected Hitting Time

We start with a property for the step sizes $\{\mathcal{A}_k\}_{k\ge1}$.

Lemma 1

Suppose that iteration k is true. Let $\kappa_g \le \frac{1-\eta}{2-\eta} \le \frac{1}{3}$. If $\alpha_k \le \bar{\alpha} = \frac{1-\eta-(2-\eta)\kappa_g}{(0.5L+2\kappa_f)(1-\kappa_g)}$ then $f_k^p \le f_k^y - \eta\alpha_k\|g_k\|^2$, i.e. the step is successful.

Without loss of generality, we assume that $\alpha_1 \ge \bar{\alpha}$. **Lemma 1 states that good model estimates and sufficiently small step size leads to successful step.**

Theorem 1: Bounding $\mathbb{E}(N_\epsilon)$ based on $\mathbb{E}(N_{TS})$

We assume Assumption 2(ii). We consider N_{TS} the number of true successful iterations with $\alpha_k \ge \bar{\alpha} : \sum_{k=1}^{N_\epsilon-1}\mathbb{1}\{\mathcal{A}_k \ge \bar{\alpha}\}\cdot\mathbb{1}\{\text{Iteration } k \text{ is true and successful}\}$. We have the **bound on $\mathbb{E}(N_\epsilon)$ based on $\mathbb{E}(N_{TS})$:**

$$\mathbb{E}(N_\epsilon) \le \frac{2p}{(2p-1)^2}\left(2\mathbb{E}(N_{TS}) + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right). \tag{8}$$

The main proof of Theorem 1 is heavily based on a prior framework for line search methods [4], where the same principle is further applied in [2, 5]. The next question focuses on the expectation of N_{TS}, which we present in the next section.

Complexity Analysis

To bound $\mathbb{E}(N_{TS})$, we consider the stochastic process $\mathcal{N}_k = 2\alpha_k^{succ}t_k^2 v_k + \|u_k\|^2$, where $v_k = f(x_k) - f_*$ and $u_k = (t_k - 1)x_k^{prev} + x^* - t_k x_k$. We also denote the step size $\alpha_k^{succ} = \alpha_{k'}$, where $k' = \max\{k'' \text{ successful }, k'' \le k\}$, the step size at the previous successful iteration prior to k. **The next Lemma describes the changes of $\mathcal{N}_k$ during successful and unsuccessful iterations.**

Lemma 2: Changes in $\mathcal{N}_k$

Let $\{x_k\}_{k\ge1}$ be the iterates of Algorithm 1 and the stochastic process $\mathcal{N}_k$ be defined above.
(i) If iteration k is unsuccessful, then we have $\mathcal{N}_k = \mathcal{N}_{k-1}$.
(ii) We assume Assumption 1 and let $\eta \ge \frac{1}{2}$. If k is a successful iteration then we have $\mathcal{N}_k - \mathcal{N}_{k-1} \le B_k$, where $B_k = 2\alpha_k t_k(\nabla f(y_k) - g_k)^\top u_{k-1} + 2\alpha_k t_k^2(f_k^y - f(y_k) + f(x_k) - f_k^p)$.

In Lemma 2(ii), B_k dictates how the noise of the model at iteration k affects the bound.

Theorem 2: The number of large true successful steps

Let $\{x_k\}_{k\ge1}$ be the iterates of Algorithm 1. We assume Assumption 1 and Assumption 2(i). Then we have

$$\mathbb{E}[N_{TS}] \le \frac{\sqrt{2D_1}}{\sqrt{\bar{\alpha}\epsilon}}. \tag{9}$$

where $D_1 = 3\alpha_0^{succ}t_0^2 v_0 + 2\|u_0\|^2 + \frac{3(\beta+2)^2}{\beta^2} + \frac{2(\beta+2)}{\beta+1} + \frac{6(\beta'+1)}{\beta'}$.

The (expected) iteration complexity of Algorithm 1 is $\mathcal{O}(\epsilon^{-1/2})$, which matches with the complexity of the deterministic accelerated momentum.

Theorem 3: Expected iteration complexity

Let $\{x_k\}_{k\ge1}$ be the iterates of Algorithm 1. We assume Assumption 1 and Assumption 2. The number of iterations N_ϵ to reach an ϵ accuracy solution satisfies the following bound:

$$\mathbb{E}(N_\epsilon) \le \frac{2p}{(2p-1)^2}\left(\frac{2\sqrt{2D_1}}{\sqrt{\bar{\alpha}\epsilon}} + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right). \tag{10}$$

where $\bar{\alpha}$ is defined in Lemma 1 and D_1 is defined in Theorem 2.

Key References

[1] Amir Beck and Marc Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: **SIAM Journal on Imaging Sciences** (2009).
[2] A. S. Berahas, L. Cao, and K. Scheinberg. "Global Convergence Rate Analysis of a Generic Line Search Algorithm with Noise". In: **SIAM Journal on Optimization** (2021).
[3] Richard G. Carter. "On the Global Convergence of Trust Region Algorithms Using Inexact Gradient Information". In: **SIAM Journal on Numerical Analysis** (1991).
[4] C. Cartis and K. Scheinberg. "Global convergence rate analysis of unconstrained optimization methods based on probabilistic models". In: **Mathematical Programming** (2018).
[5] Courtney Paquette and Katya Scheinberg. "A Stochastic Line Search Method with Expected Complexity Analysis". In: **SIAM Journal on Optimization** (2020).
[6] Boris Polyak. "Introduction to Optimization". In: **Translations Series in Mathematics and Engineering** (2020).
[7] Katya Scheinberg, Donald Goldfarb, and Xi Bai. "Fast First-Order Methods for Composite Convex Optimization with Backtracking". In: **Foundations of Computational Mathematics** (2014).
[8] Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. "Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization". In: **24th International Conference on Neural Information Processing Systems** (2011).