Stochastic FISTA adaptive step search algorithm for convex composite optimization

July 24, 2024

Trang H. Tran
joint work with Katya Scheinberg (Georgia Tech)
and Lam M. Nguyen (IBM Research)

Cornell University

Outline

1. Problem Setting
2. Prior Literature in FISTA Algorithm
3. Adaptive Stochastic Method Framework
4. Stochastic FISTA Step Search Algorithm

Convex Composite Optimization

$$\min_{x \in \mathbb{R}^n} \phi(x) = f(x) + h(x),$$

- f is convex and L-smooth,
- h is proper, closed and convex.
- It is easy to compute

$$\mathsf{prox}_h(y) = \arg\min_{x \in \mathbb{R}^n} \left\{ h(x) + \frac{1}{2} \|x - y\|^2 \right\},$$

Basic operations and notations

- Proximal model

$$Q_\alpha(x, y) = f(y) + \nabla f(y)^\top (x - y) + \frac{1}{2\alpha}\|x - y\|^2 + h(x).$$

- Proximal step

$$p_\alpha(y) := \arg\min_x Q_\alpha(x, y) = \mathsf{prox}_{\alpha h}(y - \alpha \nabla f(y)).$$

- Gradient mapping

$$D_\alpha(y) := \frac{1}{\alpha}\Big(y - \mathsf{prox}_{\alpha h}(y - \alpha \nabla f(y))\Big) = \frac{1}{\alpha}\Big(y - p_\alpha(y)\Big), \text{ for } \alpha > 0.$$

When $h \equiv 0$, then $p_\alpha(y) = y - \alpha \nabla f(x)$ and $D_\alpha(y) \equiv \nabla f(y)$.

Inexact/ stochastic gradient case

For g - an estimate of $\nabla f(y)$

- $\tilde{Q}_\alpha(x, y) = f(y) + g^\top (x - y) + \frac{1}{2\alpha}\|x - y\|^2 + h(x)$.
- $\tilde{p}_\alpha(y) := \arg\min_x \tilde{Q}_\alpha(x, y) = \mathsf{prox}_{\alpha h}(y - \alpha g)$.

FISTA Algorithm [Beck and Teboulle, 2009]

Algorithm FISTA, fixed step size

1: Initialization: Choose $x_0 \in \mathbb{R}^n$ and $\alpha > 0$. Set $t_1 = 1, t_0 = 0$, $y_1 = x_0$,
2: **for** $k = 1, 2, \cdots,$ **do**
3: Compute $\nabla f(y_k)$ and $x_k = p_\alpha(y_k) := \mathrm{prox}_{\alpha h}(y_k - \alpha \nabla f(y_k))$
4: $(t_{k+1}, y_{k+1}) = \mathrm{FISTAStep}(x_k, x_{k-1}, t_k)$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \text{ and } y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}).$$

5: **end for**

FISTA Algorithm [Beck and Teboulle, 2009]

Algorithm FISTA, fixed step size

1: Initialization: Choose $x_0 \in \mathbb{R}^n$ and $\alpha > 0$. Set $t_1 = 1, t_0 = 0$, $y_1 = x_0$,
2: **for** $k = 1, 2, \cdots,$ **do**
3: Compute $\nabla f(y_k)$ and $x_k = p_\alpha(y_k) := \mathrm{prox}_{\alpha h}(y_k - \alpha \nabla f(y_k))$
4: $(t_{k+1}, y_{k+1}) = \mathrm{FISTAStep}(x_k, x_{k-1}, t_k)$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \text{ and } y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}).$$

5: **end for**

Maintain a key property:

$$t_{k+1}(t_{k+1} - 1) = t_k^2, \quad t_{k+1} \approx t_k + \frac{1}{2} \approx \frac{k}{2}$$

$$\phi(x_k) - \phi(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\alpha t_k^2} = \mathcal{O}(\frac{1}{k^2})$$

FISTA with adaptive step size

- Check sufficient decrease condition $\phi(p_{\alpha_k}(y_k)) \leq Q_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$, if not satisfied then decrease step size
- In this (deterministic) version the step size is monotone
- Maintain a key property:

$$\alpha_{k+1} t_{k+1}(t_{k+1} - 1) \leq \alpha_k t_k^2, \quad t_{k+1} \approx t_k + \frac{1}{2} \approx \frac{k}{2}$$

$$\phi(x_k) - \phi(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\alpha_k t_k^2} = \mathcal{O}(\frac{1}{k^2})$$

Inexact FISTA [Schmidt et al., 2011]

Inexact FISTA with fixed step size

Algorithm

1: Initialization: Choose $x_0 \in \mathbb{R}^n$ and $\alpha > 0$. Set $t_1 = 1, t_0 = 0$, $y_1 = x_0$,
2: **for** $k = 1, 2, \cdots,$ **do**
3: Compute $g_k \approx \nabla f(y_k)$ and $x_k = \tilde{p}_\alpha(y_k) := \text{prox}_{\alpha h}(y_k - \alpha g_k)$
4: $(t_{k+1}, y_{k+1}) = \text{FISTAStep}(x_k, x_{k-1}, t_k)$
5: **end for**

By controlling the sum of errors - allowing a ‘budget' of inexactness:

$$\|\nabla f(y_k) - g_k\|^2 \leq \mathcal{O}(\frac{1}{k^{4+\beta}}), \quad \beta > 0$$

$$\phi(x_k) - \phi(x^*) \leq \mathcal{O}(\frac{1}{k^2})$$

How to extend this to stochastic setting?

Our first order oracle

$$g_k = g(y_k, \xi_k) \approx \nabla f(y_k),\ \xi_k \text{ is a r.v.}$$

$$\mathbb{E}_{\xi_k}[\|\nabla f(y_k) - g_k\|^2 |\, y_k] \leq \mathcal{O}(\frac{1}{k^{4+\beta}}), \quad \beta > 0$$

- Past literature on stochastic accelerated methods assumes unbiased gradients.
- In Vaswani, Bach, Schmidt 2019 [Vaswani et al., 2019], strong growth condition is assumed $(h(x) \equiv 0)$

$$\mathbb{E}_{\xi_k}[\|\nabla f(y_k) - g_k\|^2 |\, y_k] \leq \rho \|\nabla f(y_k)\|^2$$

to get $\mathcal{O}(\frac{1}{k^2})$ rate for fixed step size.

Question: Can we show convergence rate of $\mathcal{O}(\frac{1}{k^2})$ for biased estimates, composite functions, and adaptive step sizes?

Difficulty of stochastic, fully adaptive settings

- (Deterministic) sufficient decrease $\phi(p_{\alpha_k}(y_k)) \leq Q_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$ is not satisfied when the step size is large
- (Stochastic) sufficient decrease $\phi(\tilde{p}_{\alpha_k}(y_k)) \leq \tilde{Q}_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$ may be not satisfied (additionally) because of stochastic gradient

Difficulty of stochastic, fully adaptive settings

- (Deterministic) sufficient decrease $\phi(p_{\alpha_k}(y_k)) \leq Q_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$ is not satisfied when the step size is large
- (Stochastic) sufficient decrease $\phi(\tilde{p}_{\alpha_k}(y_k)) \leq \tilde{Q}_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$ may be not satisfied (additionally) because of stochastic gradient

- If condition is not satisfied $\rightarrow$ decrease step size
- If condition is satisfied $\rightarrow$ should increase step size to "offset" the decrease

Difficulty of stochastic, fully adaptive settings

- (Deterministic) sufficient decrease $\phi(p_{\alpha_k}(y_k)) \leq Q_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$ is not satisfied when the step size is large
- (Stochastic) sufficient decrease $\phi(\tilde{p}_{\alpha_k}(y_k)) \leq \tilde{Q}_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$ may be not satisfied (additionally) because of stochastic gradient

- If condition is not satisfied $\rightarrow$ decrease step size
- If condition is satisfied $\rightarrow$ should increase step size to "offset" the decrease

- A prior stochastic framework relies on the fact that if step size of the algorithm is small enough, it increases.

Adaptive Stochastic Method Framework

- This framework is originated in [Cartis and Scheinberg, 2018], then utilized in [Blanchet et al., 2019, Paquette and Scheinberg, 2020, Berahas et al., 2021, Jin et al., 2023]...
- Step size parameter can both increase and decrease. The framework does not require unbiased estimator.

Adaptive Stochastic Method Framework

- This framework is originated in [Cartis and Scheinberg, 2018], then utilized in [Blanchet et al., 2019, Paquette and Scheinberg, 2020, Berahas et al., 2021, Jin et al., 2023]...
- Step size parameter can both increase and decrease. The framework does not require unbiased estimator.
- Two main ingredients:
  - If α_k is below a threshold, and the model is "good enough", then sufficient decrease is guaranteed.
  - The model is "good enough" with some fixed probability $p > \frac{1}{2}$.

Adaptive Stochastic Method Framework

- This framework is originated in [Cartis and Scheinberg, 2018], then utilized in [Blanchet et al., 2019, Paquette and Scheinberg, 2020, Berahas et al., 2021, Jin et al., 2023]...
- Step size parameter can both increase and decrease. The framework does not require unbiased estimator.
- Two main ingredients:
  - If α_k is below a threshold, and the model is "good enough", then sufficient decrease is guaranteed.
  - The model is "good enough" with some fixed probability $p > \frac{1}{2}$.
- But original FISTA can only decrease step size.
- We need a variant of FISTA with full backtracking.

Luckily it exists! (Scheinberg, Goldfarb, Xi, 2011)

$(t_{k+1}, y_{k+1}) = \mathsf{FISTAStep}(x_k, x_{k-1}, t_k, \theta_k)$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k t_k^2}}{2},$$
$$y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}).$$

Maintain a key property:

$$\alpha_{k+1} t_{k+1}(t_{k+1} - 1) \leq \alpha_k t_k^2, \quad \theta_k = \frac{\alpha_k}{\alpha_{k+1}}, \quad \alpha_k t_k^2 \geq \left(\sum_{i<k} \sqrt{\alpha_i}/2 \right)^2$$

$$\phi(x_k) - \phi(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\alpha_k t_k^2} = \mathcal{O}(\frac{1}{k^2})$$

FISTA with fully adaptive step size

Algorithm FISTA-BKTR [Scheinberg et al., 2014]

1: Initialization: Choose $x_0 \in \mathbb{R}^n$, $\gamma \in (0,1)$ and $\alpha_1 > 0$. Set $t_1 = 1, t_0 = 0$, $y_1 = x_0 = x_{-1}$, $\theta_0 = 1$
2: **for** $k = 1, 2, \cdots,$ **do**
3: Compute $\nabla f(y_k)$
4: Compute $p_{\alpha_k}(y_k) := \text{prox}_{\alpha_k h}(y_k - \alpha_k \nabla f(y_k))$
5: If $\phi(p_{\alpha_k}(y_k)) > Q_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$
6: Compute $(t_k, y_k) = \text{FISTAStep}(x_{k-1}, x_{k-2}, t_{k-1}, \theta_{k-1})$. Set $\alpha_{k+1} = \gamma\alpha_k$ and $\theta_k = \frac{\theta_{k-1}}{\gamma}$.
7: Return to Step 5.
8: Otherwise
9: Set $x_k = p_{\alpha_k}(y_k)$ $\alpha_{k+1} = \frac{\alpha_k}{\gamma}$ and $\theta_k = \gamma$.
10: $(t_{k+1}, y_{k+1}) = \text{FISTAStep}(x_k, x_{k-1}, t_k, \theta_k)$
11: **end for**

Stochastic version

Algorithm Stochastic FISTA Step Search Algorithm

Initialization:
Choose $x_0 \in \mathbb{R}^n$, $\gamma \in (0,1)$, $\alpha_1 > 0$. Set $t_1 = 1, t_0 = 0$, $x_0^{prev} = x_0$ and $\theta_0 = \gamma$.
for $k = 1, 2, \cdots,$ **do**
    Compute $(t_k^0, y_k) = \text{FISTAStep}(x_{k-1}, x_{k-1}^{prev}, t_{k-1}, \theta_{k-1})$
    Compute $g_k = g(y_k, \xi_k)$ and $\tilde{p}_{\alpha_k}(y_k) = y_k - \alpha_k g_k$.
        If $\phi(\tilde{p}_{\alpha_k}(y_k)) > \tilde{Q}_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$
            Set $\left(x_k, x_k^{prev}\right) = \left(x_{k-1}, x_{k-1}^{prev}\right)$ and $t_k = t_{k-1}$,
             Then set $\alpha_{k+1} = \gamma \alpha_k$ and $\theta_k = \frac{\theta_{k-1}}{\gamma}$.
        Otherwise
            Set $\left(x_k, x_k^{prev}\right) = (p_{\alpha_k}(y_k), x_{k-1})$ and $t_k = t_k^0$,
            then set $\alpha_{k+1} = \frac{\alpha_k}{\gamma}$, and $\theta_k = \gamma$.
end for

We need another condition on the oracle

$\exists \bar{\alpha}$: $\quad \alpha \leq \bar{\alpha} \quad \Rightarrow \mathbb{P}\left\{\alpha_{k+1} = \frac{\alpha_k}{\gamma} \middle| \, past\right\} \geq p > \frac{1}{2}$

- We need $\phi(\tilde{p}_{\alpha_k}(y_k)) \leq \tilde{Q}_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$ with probability $p > \frac{1}{2}$.
- When $h(x) \equiv 0$ (no prox), then sufficient to assume:

$$\mathbb{P}_{\xi_k}\left\{\|g_k(y_k, \xi_k) - \nabla f(y_k)\| \leq \rho \|\nabla f(y_k)\| \,|\, y_k\right\} \geq p > \frac{1}{2}$$

(ρ sufficiently small).

We need another condition on the oracle

$\exists \bar{\alpha}$: $\quad \alpha \leq \bar{\alpha} \quad \Rightarrow \; \mathbb{P}\left\{\alpha_{k+1} = \frac{\alpha_k}{\gamma} \middle| \; past\right\} \geq p > \frac{1}{2}$

- We need $\phi(\tilde{p}_{\alpha_k}(y_k)) \leq \tilde{Q}_{\alpha_k}(p_{\alpha_k}(y_k), y_k)$ with probability $p > \frac{1}{2}$.
- When $h(x) \equiv 0$ (no prox), then sufficient to assume:

$$\mathbb{P}_{\xi_k}\left\{\|g_k(y_k, \xi_k) - \nabla f(y_k)\| \leq \rho \|\nabla f(y_k)\| \,|\, y_k\right\} \geq p > \frac{1}{2}$$

(ρ sufficiently small).

- In the prox case assume

$$\mathbb{P}_{\xi_k}\left\{\|g_k(y_k, \xi_k) - \nabla f(y_k)\| \leq \rho \|D_\alpha(y_k)\| \,|\, y_k\right\} \geq p > \frac{1}{2}$$

(Used in prior work, e.g. [Baraldi and Kouri, 2023], but for deterministic error.)

Main Result

Assume:

$$\mathbb{P}_{\xi_k}\left\{\|g_k(y_k,\xi_k)-\nabla f(y_k)\| \le \rho\|D_\alpha(y_k)\| \,|\, y_k\right\} \ge p > \frac{1}{2}$$

and

$$\mathbb{E}_{\xi_k}[\alpha_k\|\nabla f(y_k)-g_k\|^2 \,|\, y_k] \le \mathcal{O}(\frac{1}{t_k^2 k^{2+\beta}}), \quad \beta > 0$$

Expected complexity

Let T_ϵ be a hitting time for $\mathbb{I}\{\phi(X_k)-\phi(x^*)\le\epsilon\}$ then, there exists constant $\bar{\alpha}$ such that

$$\mathbb{E}(T_\epsilon) \le \frac{2p}{(2p-1)^2}\left(\left(\frac{Const(\alpha_0^{succ},x_0,\beta)}{\bar{\alpha}\epsilon}\right)^{\frac{1}{2}} + \log_\gamma\left(\frac{\bar{\alpha}}{\alpha_1}\right)\right) + 1,$$

Overview of main results and future work

- We extended analysis of adaptive stochastic methods based on random gradient estimates to accelerated methods.
- We extend this framework to composite optimization. Analysis for ISTA follow smoothly.
- We relax analysis in [Schmidt et al., 2011], from inexact to stochastic setting.
- Future work: assumptions on the noise and sufficient decrease condition may be improvable.
- Future work: inexact function oracles and high probability complexity bound.

References I

Baraldi, R. J. and Kouri, D. P. (2023).
A proximal trust-region method for nonsmooth optimization with inexact function and gradient evaluations.
Mathematical Programming, 201(1-2):559–598.

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

Berahas, A. S., Cao, L., and Scheinberg, K. (2021).
Global convergence rate analysis of a generic line search algorithm with noise.
SIAM Journal on Optimization, 31(2):1489–1518.

Blanchet, J., Cartis, C., Menickelly, M., and Scheinberg, K. (2019).
Convergence rate analysis of a stochastic trust-region method via supermartingales.
INFORMS Journal on Optimization, 1:92–119.

References II

Cartis, C. and Scheinberg, K. (2018).
Global convergence rate analysis of unconstrained optimization methods based on probabilistic models.
Mathematical Programming, 169(2):337–375.

Jin, B., Scheinberg, K., and Xie, M. (2023).
Sample complexity analysis for adaptive optimization algorithms with stochastic oracles.

Paquette, C. and Scheinberg, K. (2020).
A stochastic line search method with expected complexity analysis.
SIAM Journal on Optimization, 30(1):349–376.

Scheinberg, K., Goldfarb, D., and Bai, X. (2014).
Fast First-Order Methods for Composite Convex Optimization with Backtracking.
Foundations of Computational Mathematics, 14(3):389–417.

Schmidt, M., Roux, N. L., and Bach, F. R. (2011).
Convergence rates of inexact proximal-gradient methods for convex optimization.
CoRR, abs/1109.2415.

References III

Vaswani, S., Bach, F., and Schmidt, M. (2019).

Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron.

In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR.

THANK YOU!!!

Trang H. Tran - htt27@cornell.edu

https://htt-trangtran.github.io/