Katalyst: Boosting Convex Katayusha for Non-Convex Problems with a Large Condition Number

Zaiyi Chen, Yi Xu, Haoyuan Hu, Tianbao Yang

zaiyi.czy@alibaba-inc.com

2019-06-10
Overview

1. Introduction

2. Katalyst Algorithm and Theoretical Guarantee

3. Experiments
Problem Definition

Problem Definition

\[
\min_{x \in \mathbb{R}^d} \phi(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x)
\] (1)

we can obtain a better gradient complexity w.r.t. sample size \(n\) and accuracy \(\epsilon\) via variance reduced method (Johnson & Zhang, 2013) (SVRG-type).

We name the proposed algorithm Katalyst after Katyusha (Allen-Zhu, 2017) and Catalyst (Lin et al., 2015).
Problem Definition

\[
\min_{x \in \mathbb{R}^d} \phi(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x)
\]  

(1)

- we can obtain a better gradient complexity w.r.t. sample size \(n\) and accuracy \(\epsilon\) via variance reduced method (Johnson & Zhang, 2013) (SVRG-type).

- We name the proposed algorithm **Katalyst** after Katyusha (Allen-Zhu, 2017) and Catalyst (Lin et al., 2015).
Assumptions

- \{f_i\} are $L$-smooth.
- $\psi$ can be non-smooth but convex.
- $\phi$ is $\mu$-weakly convex.

**Definition 1**

*(L-smoothness)* A function $f$ is Lipschitz smooth with constant $L$ if its derivatives are Lipschitz continuous with constant $L$, that is

$$||\nabla f(x) - \nabla(y)|| \leq L||x - y||, \forall x, y \in \mathbb{R}^d$$

**Definition 2**

*(Weak convexity)* A function $\phi$ is $\mu$-weakly convex, if $\phi(x) + \frac{\mu}{2}||x||^2$ is convex.
Comparisons with Related Work

Table 1: Comparison of gradient complexities of variance reduction based algorithms for finding $\epsilon$-stationary point of (1). * marks the result is only valid when $L/\mu \leq \sqrt{n}$.

| Algorithms           | $L/\mu \geq \Omega(n)$ | $L/\mu \leq O(n)$ | Non-smooth $\psi$ |
|----------------------|-------------------------|-------------------|-------------------|
| SAGA (Reddi et al., 2016) | $O(n^{2/3}L/\epsilon^2)$ | $O(n^{2/3}L/\epsilon^2)$ | Yes |
| RapGrad (Lan & Yang, 2018)  | $\tilde{O}(\sqrt{nL\mu}/\epsilon^2)$ | $\tilde{O}((\mu n + \sqrt{nL\mu})/\epsilon^2)$ | Yes |
| SVRG (Reddi et al., 2016)   | $O(n^{2/3}L/\epsilon^2)$ | $O(n^{2/3}L/\epsilon^2)$ | Yes |
| Natasha1 (Allen-Zhu, 2017)  | NA                       | $O(nL/\epsilon^2)$ | No |
| RepeatSVRG (Allen-Zhu, 2017) | $\tilde{O}(n^{3/4}\sqrt{L\mu}/\epsilon^2)$ | $\tilde{O}((\mu n + n^{3/4}\sqrt{L\mu})/\epsilon^2)$ | Yes |
| 4WD-Catalyst (Paquette et al., 2018) | $O(nL/\epsilon^2)$ | $O(nL/\epsilon^2)$ | Yes |
| SPIDER (Fang et al., 2018)   | $O(\sqrt{nL}/\epsilon^2)$ | $O(\sqrt{nL}/\epsilon^2)$ | No |
| SNVRG (Zhou et al., 2018)    | $O(\sqrt{nL}/\epsilon^2)$ | $O(\sqrt{nL}/\epsilon^2)$ | No |
| Katalyst (this work)         | $\tilde{O}(\sqrt{nL\mu}/\epsilon^2)$ | $\tilde{O}((\mu n + L)/\epsilon^2)$ | Yes |

Our bound is proved optimal up to a logarithmic factor by a recent work (Zhou & Gu, 2019).
Overview

1. Introduction

2. Katalyst Algorithm and Theoretical Guarantee

3. Experiments
Interpretation - Our Basic Idea

Step 1
Interpretation - Our Basic Idea

Step > 1
Algorithm 1: Stagewise-SA($w_0$, $\{\eta_s\}$, $\mu$, $\{w_s\}$)

Input: a non-increasing sequence $\{w_s\}$, $x_0 \in \text{dom}(\psi)$, $\gamma = (2\mu)^{-1}$;

1 for $s = 1, \ldots, S$ do
2 $f_s(\cdot) = \phi(\cdot) + \frac{1}{2\gamma} \| \cdot - x_{s-1} \|^2$;
3 $x_s = \text{Katyusha}(f_s, x_{s-1}, K_s, \mu, L + \mu)$ // $x_s$ is usually an averaged solution;
4 end

Output: $x_\tau$, $\tau$ is randomly chosen from $\{0, \ldots, S\}$ according to the probabilities $p_\tau = \frac{w_{\tau+1}}{\sum_{k=0}^{S} w_{k+1}}$, $\tau = 0, \ldots, S$;

\[
f_s(x) = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i(x) + \frac{\mu}{2} \|x - x_{s-1}\|^2) + \frac{\gamma^{-1} - \mu}{2} \|x - x_{s-1}\|^2 + \hat{\psi}(x)
\]
Algorithm 2: Katyusha($f, x_0, K, \sigma, \hat{L}$)

Initialize: $\tau_2 = \frac{1}{2}$, $\tau_1 = \min\{\sqrt{\frac{n\sigma}{3L}}, \frac{1}{2}\}$, $\eta = \frac{1}{3\tau_1 L}$, $\theta = 1 + \eta\sigma$, $m = \left\lceil \frac{\log(2\tau_1 + 2/\theta - 1)}{\log \theta} \right\rceil + 1$, $y_0 = \zeta_0 = \tilde{x}^0 \leftarrow x_0$;

1. for $k = 0, \ldots, K - 1$ do
   2. $u^k = \nabla \hat{f}(\tilde{x}^k)$;
   3. for $t = 0, \ldots, m - 1$ do
      4. $j = km + t$;
      5. $x_j = \tau_1 \zeta_j + \tau_2 \tilde{x}^k + (1 - \tau_1 - \tau_2) y_j$;
      6. $\tilde{\nabla}_{j+1} = u^k + \nabla \hat{f}_i(x_{j+1}) - \nabla \hat{f}_i(\tilde{x}^k)$;
      7. $\zeta_{j+1} = \arg \min_{\zeta} \frac{1}{2\eta} \| \zeta - \zeta_j \|^2 + \langle \tilde{\nabla}_{j+1}, \zeta \rangle + \hat{\psi}(\zeta)$;
      8. $y_{j+1} = \arg \min_y \frac{3\hat{L}}{2} \| y - x_{j+1} \|^2 + \langle \tilde{\nabla}_{j+1}, y \rangle + \hat{\psi}(\zeta)$;
   9. end
10. $\tilde{x}^{k+1} = \frac{\sum_{t=0}^{m-1} \theta^t y_{sm+t+1}}{\sum_{j=0}^{m-1} \theta^t}$;
11. end

Output: $\tilde{x}^K$;
Theory

**Theorem 3**

Let $w_s = s^\alpha$, $\alpha > 0$, $\gamma = \frac{1}{2\mu}$, $\hat{L} = L + \mu$, $\sigma = \mu$, and in each call of Katyusha let 

$$\tau_1 = \min\{\sqrt{\frac{N \sigma}{3L}}, \frac{1}{2}\},$$

step size $\eta = \frac{1}{3\tau_1 L}$, $\tau_2 = 1/2$, $\theta = 1 + \eta \sigma$, and $K_s = \left\lfloor \frac{\log(D_s)}{m \log(\theta)} \right\rfloor$,

$$m = \left\lceil \frac{\log(2\tau_1 + 2/\theta - 1)}{\log \theta} \right\rceil + 1,$$

where $D_s = \max\{4\hat{L}/\mu, \hat{L}^3/\mu^3, L^2 s/\mu^2\}$. Then we have that

$$\max\{E[\|\nabla \phi_\gamma(x_{\tau+1})\|^2], E[L^2\|x_{\tau+1} - z_{\tau+1}\|^2]\} \leq \frac{34\mu\Delta_\phi(\alpha + 1)}{S + 1} + \frac{98\mu\Delta_\phi(\alpha + 1)}{(S + 1)\alpha^I \alpha < 1},$$

where $z = \text{prox}\gamma\phi(x)$, $\tau$ is randomly chosen from $\{0, \ldots, S\}$ according to probabilities

$$p_\tau = \frac{w_{\tau+1}}{\sum_{k=0}^S w_{k+1}}, \tau = 0, \ldots, S.$$  

Furthermore, the total gradient complexity for finding $x_{\tau+1}$ such that

$$\max(\max(\mathbb{E}[\|\nabla \phi_\gamma(x_{\tau+1})\|^2], L^2\mathbb{E}[\|x_{\tau+1} - z_{\tau+1}\|^2]) \leq \epsilon^2$$

is

$$N(\epsilon) = \begin{cases} 
O\left((\mu n + \sqrt{n\mu L}) \log \left(\frac{L}{\mu \epsilon} \frac{1}{\epsilon^2}\right)\right), & n \geq \frac{3L}{4\mu}, \\
O\left(\sqrt{nL\mu} \log \left(\frac{L}{\mu \epsilon} \frac{1}{\epsilon^2}\right)\right), & n \leq \frac{3L}{4\mu}.
\end{cases}$$
Suppose $\psi = 0$. With the same parameter values as in Theorem 3 except that $K = \left\lceil \frac{\log(D)}{m \log(\theta)} \right\rceil$, where $D = \max(48\hat{L}/\mu, 2\hat{L}^3/\mu^3)$. The total gradient complexity for finding $x_{\tau+1}$ such that $E[\|\nabla \phi(x_{\tau+1})\|^2] \leq \epsilon^2$ is

$$N(\epsilon) = \begin{cases} O\left((\mu n + \sqrt{n\mu L}) \log \left( \frac{L}{\mu} \right) \frac{1}{\epsilon^2} \right), & n \geq \frac{3L}{4\mu}, \\
O\left(\sqrt{nL\mu} \log \left( \frac{L}{\mu} \right) \frac{1}{\epsilon^2} \right), & n \leq \frac{3L}{4\mu}. \end{cases}$$
Overview

1 Introduction

2 Katalyst Algorithm and Theoretical Guarantee

3 Experiments
Experiments I

Squared hinge loss + (log-sum penalty (LSP) / transformed $\ell_1$ penalty (TL1)).

Figure 1: Comparison of different algorithms for two tasks on different datasets
We use Smoothed SCAD given in (Lan & Yang, 2018),

\[
R_{\lambda, \gamma, \epsilon}(x) = \begin{cases} 
  \frac{2\gamma}{2(\gamma-1)} \left( x^2 + \epsilon \right)^{\frac{1}{2}} - (x^2 + \epsilon) - \lambda^2, & \text{if } \lambda < (x^2 + \epsilon)^{\frac{1}{2}} < \gamma \lambda, \\
  \frac{\lambda^2(\gamma + 1)}{2}, & \text{otherwise,}
\end{cases}
\]

where \( \gamma > 2 \), \( \lambda > 0 \), and \( \epsilon > 0 \). Then the problem is

\[
\min_{x \in \mathbb{R}^d} \phi(x) := \frac{1}{2n} \sum_{i=1}^{n} (a_i^\top x - b_i)^2 + \frac{\rho}{2} \sum_{i=1}^{d} R_{\lambda, \gamma, \epsilon}(x_i)
\]
Figure 2: Theoretical performances of RapGrad and Katalyst.
Figure 3: Empirical performances of RapGrad and Katalyst with early termination.
The End
Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 89–97, 2017.

Allen-Zhu, Z. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 1200–1205, 2017.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *NeurIPS*, pp. 687–697, 2018.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.

Lan, G. and Yang, Y. Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization. *CoRR*, abs/1805.05411, 2018.
Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pp. 3384–3392, 2015.

Paquette, C., Lin, H., Drusvyatskiy, D., Mairal, J., and Harchaoui, Z. Catalyst for gradient-based nonconvex optimization. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pp. 613–622, 2018.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pp. 1145–1153, 2016.

Zhou, D. and Gu, Q. Lower bounds for smooth nonconvex finite-sum optimization. *arXiv preprint arXiv:1901.11224*, 2019.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *NeurIPS*, pp. 3925–3936, 2018.