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

Problem Definition

Problem Definition

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

- we can obtain a better gradient complexity w.r.t. sample size n and accuracy ϵ 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

Problem Definition

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

- we can obtain a better gradient complexity w.r.t. sample size n and accuracy ϵ 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.
- ψ can be non-smooth but convex.
- ϕ is μ-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

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

Definition 2

(Weak convexity) A function ϕ is μ-weakly convex, if $\phi(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex.

Comparisons with Related Work

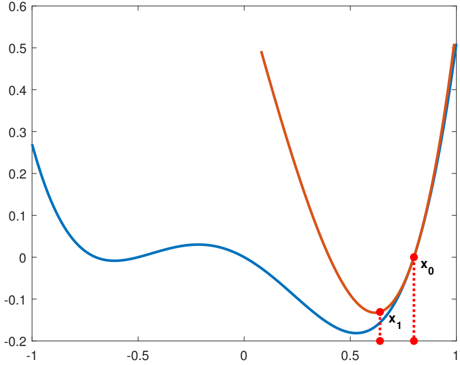
Table 1: Comparison of gradient complexities of variance reduction based algorithms for finding ϵ-stationary point of (1). * marks the result is only valid when $L/\mu \leq \sqrt{n}$.

Algorithms	$L/\mu \geq \Omega(n)$	$L/\mu \leq O(n)$	Non-smooth ψ
SAGA (Reddi et al., 2016)	$O(n^{2/3}L/\epsilon^2)$	$O(n^{2/3}L/\epsilon^2)$	Yes
RapGrad (Lan & Yang, 2018)	$\widetilde{O}(\sqrt{nL\mu}/\epsilon^2)$	$\widetilde{O}((\mu n + \sqrt{nL\mu})/\epsilon^2)$	indicator function
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	$O(n^{2/3}L^{2/3}\mu^{1/3}/\epsilon^2)^*$	Yes
RepeatSVRG (Allen-Zhu, 2017)	$\widetilde{O}(n^{3/4}\sqrt{L\mu}/\epsilon^2)$	$\widetilde{O}((\mu n + n^{3/4}\sqrt{L\mu})/\epsilon^2)$	Yes
4WD-Catalyst (Paquette et al., 2018)	$O(nL/\epsilon^2)$	$O(nL/\epsilon^2)$	Yes
SPIDER (Fang et al., 2018)	$O(\sqrt{n}L/\epsilon^2)$	$O(\sqrt{n}L/\epsilon^2)$	No
SNVRG (Zhou et al., 2018)	$O(\sqrt{n}L/\epsilon^2)$	$O(\sqrt{n}L/\epsilon^2)$	No
Katalyst (this work)	$\widetilde{O}(\sqrt{nL\mu}/\epsilon^2)$	$\widetilde{O}((\mu n + L)/\epsilon^2)$	Yes

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

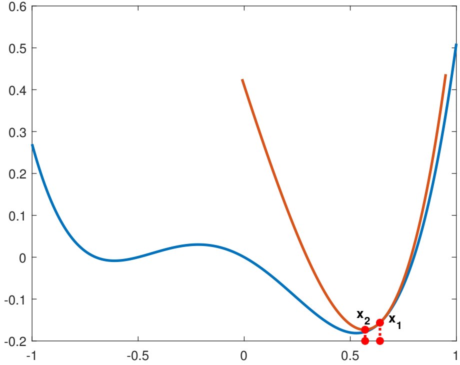
Overview

Interpretation - Our Basic Idea



Step 1

Interpretation - Our Basic Idea



Step > 1

A Unified Framework

Meta Algorithm

Algorithm 1: Stagewise-SA($\mathbf{w}_0$, $\{\eta_s\}$, μ, $\{\mathbf{w}_s\}$)

Input : a non-increasing sequence $\{w_s\}$, $\mathbf{x}_0 \in \mathrm{dom}(\psi)$, $\gamma = (2\mu)^{-1}$;

for $s = 1, \ldots, S$ **do**

  $f_s(\cdot) = \phi(\cdot) + \frac{1}{2\gamma}\| \cdot - \mathbf{x}_{s-1}\|^2$;

  $\mathbf{x}_s = \text{Katyusha}(f_s, \mathbf{x}_{s-1}, K_s, \mu, L+\mu)$ // $\mathbf{x}_s$ is usually an averaged solution;

end

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

$$f_s(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(f_i(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_{s-1}\|^2\right)}_{\hat{f}_i(\mathbf{x})} + \underbrace{\frac{\gamma^{-1} - \mu}{2}\|\mathbf{x} - \mathbf{x}_{s-1}\|^2 + \psi(\mathbf{x})}_{\hat{\psi}(\mathbf{x})}$$

Algorithm

Algorithm 2: Katyusha$(f, x_0, K, \sigma, \widehat{L})$

Initialize: $\tau_2 = \frac{1}{2},\ \tau_1 = \min\{\sqrt{\frac{n\sigma}{3L}}, \frac{1}{2}\},\ \eta = \frac{1}{3\tau_1 L},\ \theta = 1 + \eta\sigma,\ m = \lceil \frac{\log(2\tau_1 + 2/\theta - 1)}{\log \theta} \rceil + 1,\ \mathbf{y}_0 = \zeta_0 = \widetilde{\mathbf{x}}^0 \leftarrow \mathbf{x}_0;$

for $k = 0, \ldots, K-1$ **do**

- $\mathbf{u}^k = \nabla \hat{f}(\widetilde{x}^k);$
- **for** $t = 0, \ldots, m-1$ **do**
  - $j = km + t;$
  - $\mathbf{x}_j = \tau_1 \zeta_j + \tau_2 \widetilde{\mathbf{x}}^k + (1 - \tau_1 - \tau_2)\mathbf{y}_j;$
  - $\widetilde{\nabla}_{j+1} = \mathbf{u}^k + \nabla \hat{f}_i(\mathbf{x}_{j+1}) - \nabla \hat{f}_i(\widetilde{\mathbf{x}}^k);$
  - $\zeta_{j+1} = \arg\min_{\zeta} \frac{1}{2\eta}\|\zeta - \zeta_j\|^2 + \langle \widetilde{\nabla}_{j+1}, \zeta \rangle + \hat{\psi}(\zeta);$
  - $\mathbf{y}_{j+1} = \arg\min_{\mathbf{y}} \frac{3\widehat{L}}{2}\|\mathbf{y} - \mathbf{x}_{j+1}\|^2 + \langle \widetilde{\nabla}_{j+1}, \mathbf{y} \rangle + \hat{\psi}(\zeta);$
- **end**
- $\widetilde{\mathbf{x}}^{k+1} = \frac{\sum_{t=0}^{m-1} \theta^t \mathbf{y}_{sm+t+1}}{\sum_{j=0}^{m-1} \theta^t};$

end

Output : $\widetilde{x}^K;$

Theorem 3

Let $w_s = s^\alpha, \alpha > 0, \gamma = \frac{1}{2\mu}$, $\widehat{L} = L + \mu$, $\sigma = \mu$, *and in each call of Katyusha let* $\tau_1 = \min\{\sqrt{\frac{N\sigma}{3\widehat{L}}}, \frac{1}{2}\}$, *step size* $\eta = \frac{1}{3\tau_1 \widehat{L}}$, $\tau_2 = 1/2$, $\theta = 1 + \eta\sigma$, *and* $K_s = \left\lceil \frac{\log(D_s)}{m \log(\theta)} \right\rceil$, $m = \left\lfloor \frac{\log(2\tau_1 + 2/\theta - 1)}{\log \theta} \right\rfloor + 1$, *where* $D_s = \max\{4\hat{L}/\mu, \hat{L}^3/\mu^3, L^2 s/\mu^2\}$. *Then we have that*

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

where $\mathbf{z} = prox_{\gamma\phi}(\mathbf{x})$, τ *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* $\mathbf{x}_{\tau+1}$ *such that*

$$\max(\mathrm{E}[\|\nabla\phi_\gamma(\mathbf{x}_{\tau+1})\|^2], L^2\mathrm{E}[\|\mathbf{x}_{\tau+1} - \mathbf{z}_{\tau+1}\|^2]) \leq \epsilon^2$$

is

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

Theory

Theorem 4

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* $\mathbf{x}_{\tau+1}$ *such that* $\mathrm{E}[\|\nabla\phi(\mathbf{x}_{\tau+1})\|^2] \leq \epsilon^2$ *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

Experiments I

Squared hinge loss + (log-sum penalty (LSP) / transformed ℓ_1 penalty (TL1)).

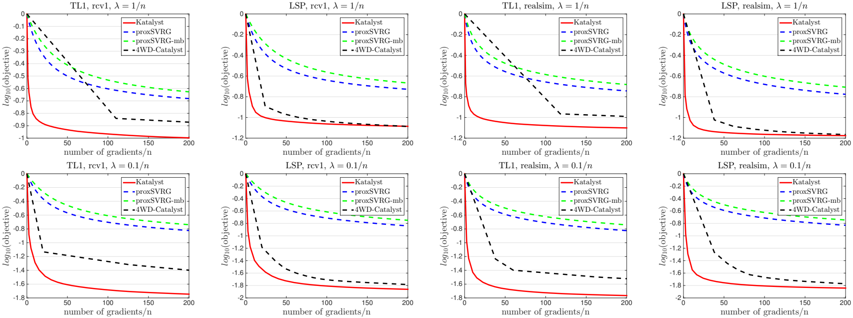


Figure 1: Comparison of different algorithms for two tasks on different datasets

Experiments II

We use Smoothed SCAD given in (Lan & Yang, 2018),

$$R_{\lambda,\gamma,\epsilon}(x) = \begin{cases} \lambda(x^2+\epsilon)^{\frac{1}{2}}, & \text{if} (x^2+\epsilon)^{\frac{1}{2}} \le \lambda, \\ \dfrac{2\gamma\lambda(x^2+\epsilon)^{\frac{1}{2}} - (x^2+\epsilon) - \lambda^2}{2(\gamma-1)}, & \\ & \text{if } \lambda < (x^2+\epsilon)^{\frac{1}{2}} < \gamma\lambda, \\ \dfrac{\lambda^2(\gamma+1)}{2}, & \text{otherwise,} \end{cases}$$

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

$$\min_{\mathbf{x}\in\mathbb{R}^d} \phi(\mathbf{x}) := \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{a}_i^\top \mathbf{x} - b_i)^2 + \frac{\rho}{2}\sum_{i=1}^{d} R_{\lambda,\gamma,\epsilon}(x_i)$$

Experiments II.1

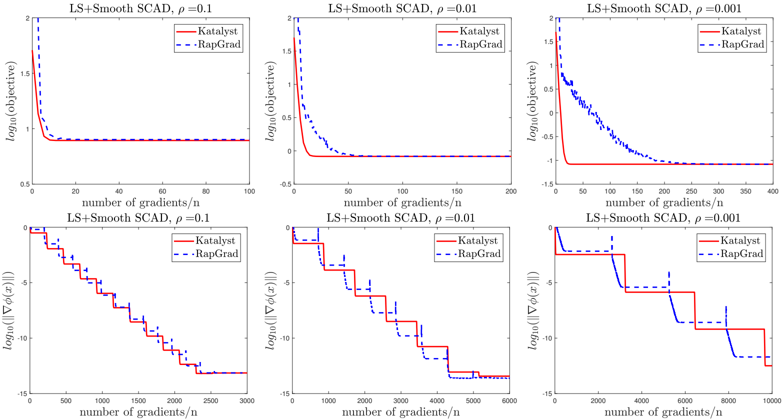


Figure 2: Theoretical performances of RapGrad and Katalyst.

Experiments II.2

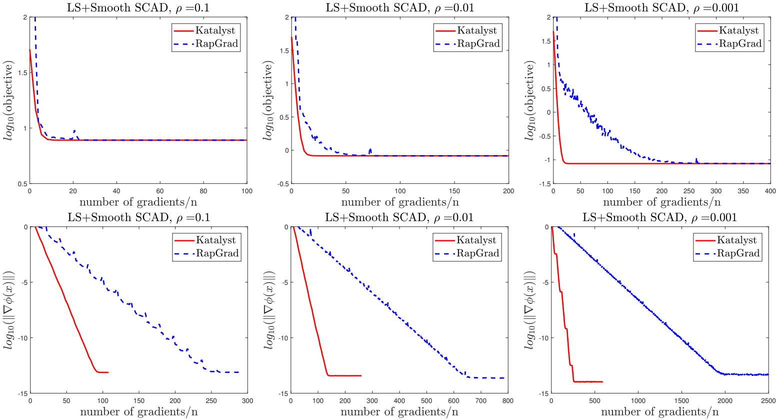


Figure 3: Empirical performances of RapGrad and Katalyst with early termination.

The End

References I

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.

References II

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.