Supplement for ”Katalyst: Boosting Convex Katyusha for Non-Convex Problems with a Large Condition Number”

Zaiyi Chen [1 2] **Yi Xu** [3] **Haoyuan Hu** [1] **Tianbao Yang** [3]

A. Proof of Theorem 3

We need the following lemma for proving Theorem 3.

Lemma 2. *(Allen-Zhu, 2017) Regarding the modified-Katyusha algorithm (Algorithm 2), suppose that $\tau_1 \leq \frac{1}{3\eta\hat{L}}$, $\tau_2 = 1/2$. Defining $D_t := f(\mathbf{y}_t) - f(\mathbf{x})$, $\widetilde{D}^k := f(\widetilde{\mathbf{x}}^k) - f(\mathbf{x})$ for any $\mathbf{x}$, conditioned on iterations $\{0, \ldots, t-1\}$ in k-th epoch and all iterations before k-th epoch, we have that*

$$\begin{aligned} 0 \leq & \frac{(1-\tau_1-\tau_2)}{\tau_1} D_t - \frac{1}{\tau_1}\mathrm{E}[D_{t+1}] + \frac{\tau_2}{\tau_1}\widetilde{D}^k \\ & + \frac{1}{2\eta}\|\zeta_t - \mathbf{x}\|^2 - \frac{1+\eta\sigma}{2\eta}\mathrm{E}[\|\zeta_{t+1} - \mathbf{x}\|^2] \end{aligned} \tag{1}$$

Proof. [of Theorem 3] Define $\theta = 1 + \eta\sigma$ and multiply (1) by θ^t on both side. By summing up the inequalities in (1) in the k-th epoch, we have that

$$\begin{aligned} 0 \leq & \mathrm{E}_k\left[\frac{1-\tau_1-\tau_2}{\tau_1}\sum_{t=0}^{m-1} D_{km+t}\theta^t - \frac{1}{\tau_1}\sum_{t=0}^{m-1} D_{km+t+1}\theta^t\right] \\ & + \frac{\tau_2}{\tau_1}\widetilde{D}^k \sum_{t=1}^{m-1}\theta^t + \frac{1}{2\eta}\|\zeta_{km} - \mathbf{x}\|^2 \\ & - \frac{\theta^m}{2\eta}\mathrm{E}_{k+1}[\|\zeta_{(k+1)m} - \mathbf{x}\|^2] \end{aligned}$$

where $\mathrm{E}_k[\cdot]$ denotes expectation in k-th epoch conditional on $0, \ldots, k-1$ epochs. Using the convexity of $f(\cdot)$, we have that

$$\begin{aligned} & \frac{\tau_1+\tau_2-1+1/\theta}{\tau_1}\theta\mathrm{E}_k[\widetilde{D}^{k+1}]\sum_{t=0}^{m-1}\theta^t + \\ & \frac{1-\tau_1-\tau_2}{\tau_1}\theta^m\mathrm{E}[D_{(k+1)m}] + \frac{\theta^m}{2\eta}\mathrm{E}_k[\|\zeta_{(k+1)m} - \mathbf{x}\|^2] \\ & \leq \frac{\tau_2}{\tau_1}\widetilde{D}^k\sum_{t=0}^{m-1}\theta^t + \frac{1-\tau_1-\tau_2}{\tau_1}D_{km} + \frac{1}{2\eta}\|\zeta_{km} - \mathbf{x}\|^2 \end{aligned} \tag{2}$$

Substituting $\tau_2 = 1/2$ and $m \leq \lceil\frac{\log(2\tau_1+2/\theta-1)}{\log\theta}\rceil + 1$, we have that

$$\begin{aligned} & \theta^m\frac{1}{2\theta\tau_1}\mathrm{E}_k[\widetilde{D}^{k+1}]\sum_{t=0}^{m-1}\theta^t + \frac{1/2-\tau_1}{\tau_1}\theta^m\mathrm{E}[D_{(k+1)m}] \\ & + \frac{\theta^m}{2\eta}\mathrm{E}_k[\|\zeta_{(k+1)m} - \mathbf{x}\|^2] \\ & \leq \frac{1}{2\tau_1}\widetilde{D}^k\sum_{t=0}^{m-1}\theta^t + \frac{1/2-\tau_1}{\tau_1}D_{km} + \frac{1}{2\eta}\|\zeta_{km} - \mathbf{x}\|^2 \end{aligned}$$

Telescoping above inequality over all epochs $k = 0, \ldots, K-1$ we have that

$$\begin{aligned} \mathrm{E}[\widetilde{D}^K] \leq 2\theta\tau_1\theta^{-mK}\Bigg(&\frac{1}{2\tau_1}\widetilde{D}^0 + \frac{1/2-\tau_1}{\tau_1\sum_{t=0}^{m-1}\theta^t}D_0 \\ & + \frac{1}{2\eta\sum_{t=0}^{m-1}\theta^t}\|\zeta_0 - \mathbf{x}\|^2\Bigg) \end{aligned}$$

Since $\sum_{t=0}^{m-1}\theta^t \geq 1$, $\tau_1 \leq \frac{1}{2}$ and $\theta \leq 2$, we have

$$\mathrm{E}[\widetilde{D}^K] \leq 4\tau_1\theta^{-mK}(\frac{1-\tau_1}{\tau_1}\widetilde{D}^0 + \frac{1}{2\eta}\|\zeta_0 - \mathbf{x}\|^2)$$

We can use the same analysis by plugging $\mathbf{x} = \zeta_0$ in (1) to prove that $\mathrm{E}[f(\widetilde{\mathbf{x}}^K) - f(\widetilde{\mathbf{x}}^0)] \leq 0$ - an objective value decreasing property that will be used later. □

[1] Cainiao AI, China [2] University of Science and Technology of China, China [3] University of Iowa, USA. Correspondence to: <zaiyi.czy@alibaba-inc.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

B. Proof of Lemma 1

Proof. First we have hat

$$\begin{aligned}\mathrm{E}[f_s(\mathbf{x}_s)] =& \mathrm{E}\left[\phi(\mathbf{x}_s)) + \frac{1}{2\gamma}\|\mathbf{x}_s - \mathbf{x}_{s-1}\|^2\right] \leq f_s(\mathbf{z}_s) + \mathcal{E}_s \\ \leq& f_s(\mathbf{x}_{s-1}) + \mathcal{E}_s = \phi(\mathbf{x}_{s-1}) + \mathcal{E}_s\end{aligned}$$

Besides, we also have that

$$\begin{aligned}&\|\mathbf{x}_s - \mathbf{x}_{s-1}\|^2 \\ =&\|\mathbf{x}_s - \mathbf{z}_s + \mathbf{z}_s - \mathbf{x}_{s-1}\|^2 \\ =&\|\mathbf{x}_s - \mathbf{z}_s\|^2 + \|\mathbf{z}_s - \mathbf{x}_{s-1}\|^2 + 2\langle \mathbf{x}_s - \mathbf{z}_s, \mathbf{z}_s - \mathbf{x}_{s-1}\rangle \\ \geq&(1-\alpha_s^{-1})\|\mathbf{x}_s - \mathbf{z}_s\|^2 + (1-\alpha_s)\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2\end{aligned}$$

where the inequality follows from the Young's inequality with $0 < \alpha_s < 1$. Combining above inequalities, then we have

$$\begin{aligned}&\frac{(1-\alpha_s)}{2\gamma}\mathrm{E}_s\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2 \\ \leq&\mathrm{E}_s\left[\Delta_s + \frac{(\alpha_s^{-1}-1)}{2\gamma}\|\mathbf{x}_s - \mathbf{z}_s\|^2 + \mathcal{E}_s\right] \\ \leq&\mathrm{E}_s[\Delta_s] + \frac{(\alpha_s^{-1}-1)}{2\gamma}\mathrm{E}_s[\|\mathbf{x}_s - \mathbf{z}_s\|^2] + \mathcal{E}_s \\ \leq&\mathrm{E}_s[\Delta_s] + \frac{(\alpha_s^{-1}-1)+\gamma\sigma}{\gamma\sigma}\mathcal{E}_s \\ \leq&\mathrm{E}_s[\Delta_s] + \frac{(\alpha_s^{-1}-1)+\gamma\sigma}{\gamma\sigma}\Big[4\theta^{-mK}(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*)) \\ &+2\theta^{-mK}\hat{L}\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2\Big]\end{aligned}$$

where the first inequality follows from the definition $\Delta_s := \phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_s)$, and the third inequality uses the strong convexity of $f_s(\mathbf{x})$, whose strong convexity parameter is $\sigma = \gamma^{-1} - \mu$. Substituting $\alpha_s = 1/2$, $\gamma = 1/(2\mu)$, and $\sigma = \mu$, $\hat{L} \leq 2L$ and $\theta^{-mK} \leq \mu/(24\hat{L})$, we have that

$$\frac{1}{8\gamma}\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2 \leq \mathrm{E}_s[\Delta_s] + 12\theta^{-mK}(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*))$$

□

C. A Technical Lemma

Lemma 3. *For a non-decreasing sequence $w_s, s = 0, \dots, S+1$, we have*

$$\mathrm{E}\left[\sum_{s=1}^{S+1} w_s\Delta_s\right] \leq \Delta_\phi w_{S+1}$$

Proof.

$$\begin{aligned}&\sum_{s=1}^{S+1} w_s\Delta_s = \sum_{s=1}^{S+1} w_s(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_s)) \\ =&\sum_{s=1}^{S+1}(w_{s-1}\phi(\mathbf{x}_{s-1}) - w_s\phi(\mathbf{x}_s)) \\ &+\sum_{s=1}^{S+1}(w_s - w_{s-1})\phi(\mathbf{x}_{s-1}) \\ =&w_0\phi(\mathbf{x}_0) - w_{S+1}\phi(\mathbf{x}_{S+1}) + \sum_{s=1}^{S+1}(w_s - w_{s-1})\phi(\mathbf{x}_{s-1}) \\ =&\sum_{s=1}^{S+1}(w_s - w_{s-1})(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_{S+1}))\end{aligned}$$

where the third equality follows from the extension that $w_0 = 0$. Taking expectation on both sides, we have

$$\begin{aligned}&\mathrm{E}\left[\sum_{s=1}^{S+1} w_s\Delta_s\right] \\ =&\sum_{s=1}^{S+1}(w_s - w_{s-1})\mathrm{E}[(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_{S+1}))] \\ \leq&\sum_{s=1}^{S+1}(w_s - w_{s-1})[\phi(\mathbf{x}_0) - \phi(\mathbf{x}_*)] \\ \leq&\Delta_\phi w_{S+1}\end{aligned}$$

where we use the fact that $\mathrm{E}[f_s(\mathbf{x}_s) - f_s(\mathbf{x}_{s-1})] \leq 0$ (this is the objective value decreasing property of Katyusha) implying $\mathrm{E}[\phi(\mathbf{x}_s) - \phi(\mathbf{x}_{s-1})] \leq 0$ and hence $\mathrm{E}[\phi(\mathbf{x}_s)] \leq \phi(\mathbf{x}_0)$ for $s \geq 0$. □

D. Decomposition of LSP and TL1

It is easy to verify that for LSP, $r_1(\mathbf{x}) = \frac{\lambda}{\beta}\|\mathbf{x}\|_1$ and $r_2(x) = \lambda\sum_{i=1}^d(|x|/\beta - \log(\beta + |x|))$. For TL1, $r_1(\mathbf{x}) = \lambda\frac{\beta+1}{\beta}\|\mathbf{x}\|_1$ and $r_2(\mathbf{x}) = \lambda\sum_{i=1}^d \frac{(\beta+1)|x_i|^2}{\beta(\beta+|x_i|)}$. For smoothness of r_2 for both regularizers, we refer readers to (Wen et al., 2018).

E. Comparisons with RapGrad

In this section, we conduct some experiments for solving least square (LS) regression problem with the smoothly clipped absolute deviation (Smoothed SCAD) penalty by RapGrad (Lan & Yang, 2018) and Katalyst. To handle the non-smoothness of SCAD (Fan & Li, 2001) at $\mathbf{x} = 0$, we add a small positive number ϵ to obtain a smooth

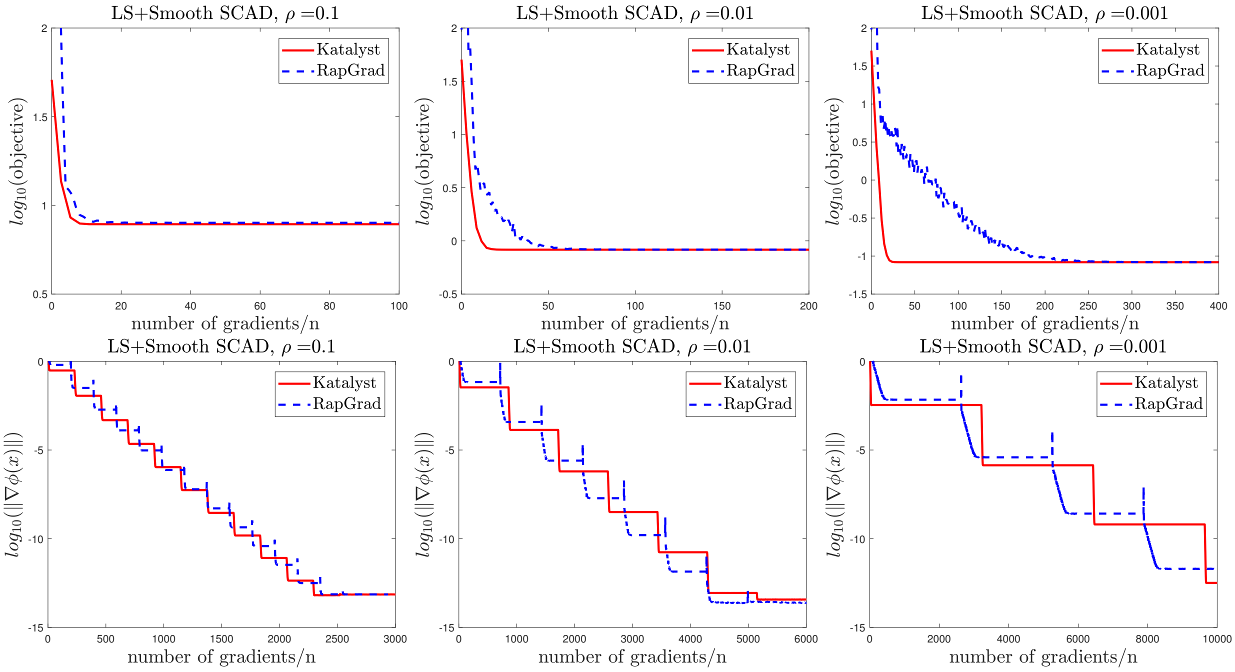

Figure 2. Theoretical performances of RapGrad and Katalyst.

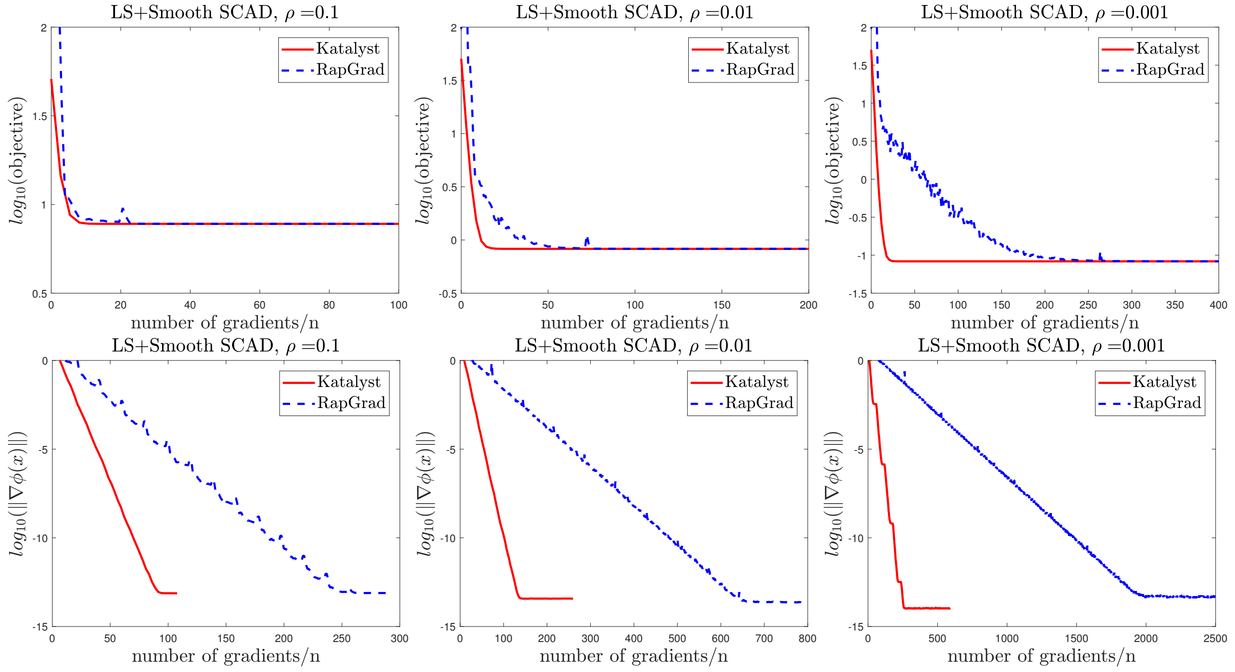

Figure 3. Empirical performances of RapGrad and Katalyst with early termination.

approximation $R_{\lambda,\gamma,\epsilon}$ (Lan & Yang, 2018):

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

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

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

It is easy to show that the weakly convexity parameter and smoothness parameter are given by $\mu = \frac{\rho}{2(\gamma-1)}$ and $L = \frac{\rho\lambda\epsilon^{-1/2}}{2} + \max_{1\leq i\leq n} \|\mathbf{a}_i\|^2$.

We test RapGrad and Katalyst on a randomly generated data set of size $n = 1000$ and $d = 100$. The parameters of $R_{\lambda,\gamma,\epsilon}$ are setting as follows, $\epsilon = 10^{-3}$, $\lambda = 2$, $\gamma = 4$ and $\rho = 0.01$.

We first set all parameters in RapGrad and Katalyst to their theoretical values. The results are reported in Figure 2. We can see that two algorithms perform similarly in the view of gradient norm $\|\nabla\phi\|$, while Katalyst reduces the objective value faster than RapGrad even when ρ is small.

We can observe from Figure 2 that both RapGrad and Katalyst are conservative on estimating the required iterations for solving subproblems, which waste a large number of gradient computations. Next, we try early termination strategy for solving subproblems to see whether it could help to get better performance following (Lan & Yang, 2018). The number of the inner loops of RapGrad and Katalyst are tested from s to $s/50$ and from K to $K/50$ respectively, and the best results are reported in Figure 3. We can see that both algorithms have little improvements in reducing objective values. Meanwhile, the plateaus of gradient norm are disappeared and two algorithms obtain about 10 times faster convergence speed in $\|\nabla\phi\|$.

References

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.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Lan, G. and Yang, Y. Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization. *CoRR*, abs/1805.05411, 2018.

Wen, B., Chen, X., and Pong, T. K. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 69(2):297–324, Mar 2018.