# 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]

## Abstract

An important class of non-convex objectives that has wide applications in machine learning consists of a sum of $n$ smooth functions and a non-smooth convex function. Tremendous studies have been devoted to conquering these problems by leveraging one of the two types of variance reduction techniques, i.e., SVRG-type that computes a full gradient occasionally and SAGA-type that maintains $n$ stochastic gradients at every iteration. An interesting question that has been largely ignored is how to improve the complexity of variance reduction methods for problems with a large condition number that measures the degree to which the objective is close to a convex function. In this paper, we present a simple but non-trivial boosting of a state-of-the-art SVRG-type method for convex problems (namely Katyusha) to enjoy an improved complexity for solving non-convex problems with a large condition number (that is close to a convex function). To the best of our knowledge, its complexity has the best dependence on $n$ and the degree of non-convexity, and also matches that of a recent SAGA-type accelerated stochastic algorithm for a constrained non-convex smooth optimization problem. Numerical experiments verify the effectiveness of the proposed algorithm in comparison with its competitors.

## 1. Introduction

The problem of interest in this paper belongs to the following class of non-convex optimization problems:

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

where each $f_i$ is a $L$-smooth function, and $\psi(\mathbf{x})$ is a "simple" closed convex function whose proximal mapping can be efficiently computed. The above problem covers constrained and non-constrained smooth optimization as special cases when $\psi(\mathbf{x})$ is the indicator function of a convex set and $\psi = 0$. This problem has broad applications in machine learning, and has been studied by numerous papers (Reddi et al., 2016a;b;c; Lan & Yang, 2018; Allen-Zhu, 2018; Allen-Zhu & Hazan, 2016). A number of stochastic algorithms were proposed by utilizing the finite-sum structure of the problem and smoothness of $f_i$ to derive faster convergence than stochastic gradient methods. These algorithms are based on two well-known variance-reduction techniques, namely the SVRG-type variance reduction (Johnson & Zhang, 2013) and the SAGA-type variance reduction (Defazio et al., 2014; Roux et al., 2012). The key difference between these two variance reduction techniques is that SVRG uses a full gradient that is computed periodically and SAGA uses a full gradient that is computed from its maintained historical gradients for each component $f_i$. Due to this difference, SAGA might require much higher memory than SVRG for many problems, which renders algorithms of SVRG-type more favorable than algorithms of SAGA-type.

Since the proposal of non-convex SVRG for solving non-convex problems in the form of (1) or its special case with $\psi = 0$ (Reddi et al., 2016a; Allen-Zhu & Hazan, 2016), several studies have tried to improve its complexity in terms of the number of components $n$ (Fang et al., 2018; Zhou et al., 2018). To the best of our knowledge, the state-of-the-art gradient complexity [1] of SVRG-type methods for finding a solution $\mathbf{x}$ such that $\mathrm{E}[\|\nabla\phi(\mathbf{x})\|] \leq \epsilon$ under the condition $\psi = 0$ and $\epsilon \leq 1/\sqrt{n}$ is given by $O(L\sqrt{n}/\epsilon^2)$. It was also shown in (Fang et al., 2018) that such a complexity is a lower bound for the problem (1), hence it cannot be improved in general.

However, most of previous studies have ignored the degree of non-convexity of each component function with few exceptions discussed later. A natural way to measure the degree of non-convexity is by considering a notion of $\mu$-weak convexity. In particular a function $f$ is said to be $\mu$-weakly convex if $f(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2$ is a convex function

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

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

[1] the number of stochastic gradient computations

*Table 1.* Comparison of gradient complexities of variance reduction based algorithms for finding $\epsilon$-stationary point of (1) with $\psi = 0$. The best complexity result for each setting is marked in red color. The top two algorithms, namely SAGA and RapGrad use the SAGA-type variance reduction technique, while others use the SVRG-type variance reduction technique. $\widetilde{O}(\cdot)$ hides some logarithmic factor. $^*$ 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., 2016c) | $O(n^{2/3}L/\epsilon^2)$ | $O(n^{2/3}L/\epsilon^2)$ | 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., 2016c) | $O(n^{2/3}L/\epsilon^2)$ | $O(n^{2/3}L/\epsilon^2)$ | Yes |
| Natasha1 (Allen-Zhu, 2017a) | NA | $O(n^{2/3}L^{1/3}\mu^{2/3}/\epsilon^2)^*$ | Yes |
| RepeatSVRG (Allen-Zhu, 2017a) | $\widetilde{O}(n^{3/4}\sqrt{L\mu}/\epsilon^2)$ | $\widetilde{O}((\mu n + n^{3/4}\sqrt{L\mu})/\epsilon^2)$ | 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 |

for $\mu > 0$, where $\|\cdot\|$ denotes the Euclidean norm. If $f$ is twice-differentiable, $\mu$-weak convexity is equivalent to that $\nabla^2 f(\mathbf{x}) \geq -\mu I$. Hence, the smaller the $\mu$, the closer the function to a convex function. For a smooth function with $L$-Lipchitz continuous gradient we define the condition number as $L/\mu$. Therefore, an interesting question is whether the gradient complexity of a SVRG-type method can be further improved for (1) with $\mu$-weakly convex functions $f_i$ when $\mu$ is very small. In another word, whether the gradient complexity can be made dependent on $\mu$ such that the closer $f_i$ is to a convex function the smaller is the complexity. In this paper, we provide an affirmative answer to this question. We show that when the condition number of each $f_i$ is large (i.e, $L/\mu \geq \Omega(n)$), we can improve the complexity to $\widetilde{O}(\sqrt{nL\mu}/\epsilon^2)$, which is better than that reported in (Fang et al., 2018; Zhou et al., 2018). To the best of our knowledge, this is the best result for a SVRG-type method for solving problem (1) under a large condition number, which also matches that of a recent work focusing on developing an accelerated SAGA-type method for solving constrained non-convex smooth optimization (Lan & Yang, 2018). We also establish a gradient complexity of $\widetilde{O}(\mu n/\epsilon^2)$ in the case of $L/\mu < O(n)$, which improves the complexity of (Fang et al., 2018; Zhou et al., 2018) when $\mu/L \leq 1/\sqrt{n}$, and is also slightly better than that of (Lan & Yang, 2018). The proposed algorithm is a simple but non-trivial boosting of convex Katyusha (Allen-Zhu, 2017b). The idea is by calling convex Katyusha for solving a sequence of regularized convex problems, which is similar to that used in the Catalyst technique for speeding up convex optimization (Lin et al., 2015). However, the key difference and novelty of the proposed algorithm is that we do not use any extrapolation step and the acceleration is simply achieved by carefully choosing the parameters (i.e., the number of epochs and the number of iterations for the inner loop) for convex Katyusha that are adaptive to the $\mu$-weak convexity of the problem. We refer to the proposed algorithm as Katalyst.

Before ending this section, we present a motivating example of the considered easy non-convex problems with a large condition number. Let us consider least-squares regression with non-convex sparsity-promoting regularizers:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{a}_i^\top \mathbf{x}, b_i) + \lambda R(\mathbf{x}), \tag{2}$$

where $(\mathbf{a}_i, b_i), i = 1, \ldots, n$ denote a set of $n$ observed data with $\mathbf{a}_i \in \mathbb{R}^d$ representing the feature vector and $b_i \in \mathbb{R}$ representing the label of the $i$-th example, $\ell(\mathbf{a}_i^\top \mathbf{x}, b_i) = (\mathbf{a}_i^\top \mathbf{x} - b_i)^2$, $R(\mathbf{x})$ denotes a non-convex regularizer that enforces sparsity and $\lambda > 0$ is a regularization parameter. Commonly used non-convex sparsity-promoting regularizers include logarithmic sum penalty (Candès et al., 2008), transformed $\ell_1$ norm (Zhang & Xin, 2014), smoothly clipped absolute deviation (SCAD) regularization (Fan & Li, 2001), minimax concave penalty (MCP) regularization (Zhang, 2010). All of these regularizers can be written as a (scaled) $\ell_1$ norm minus a differentiable smooth function. Let us consider the logarithmic sum penalty $R(\mathbf{x}) = \sum_{i=1}^d \log(|\mathbf{x}_i| + \theta)$. It can be written as $R(\mathbf{x}) = 1/\theta\|\mathbf{x}\|_1 + R_2(\mathbf{x})$, where $R_2(\mathbf{x}) = \sum_{i=1}^d (\log(|x_i| + \theta) - |x_i|/\theta)$. It was shown that $R_2$ is a differentiable smooth non-convex function with a smoothness parameter $\mu = \frac{1}{\theta^2}$ (Wen et al., 2018). In order to formulate the problem as (1), we can defined $f_i(\mathbf{x}) = (\mathbf{a}_i^\top \mathbf{x} - b_i)^2/2 + \lambda R_2(\mathbf{x})$ and $\psi(\mathbf{x}) = \lambda\|\mathbf{x}\|_1/\theta$. Thus, we have $f_i$ is $\mu = \lambda/\theta^2$-weakly convex and $L = \max_i \|\mathbf{a}_i\|^2 + \lambda/\theta^2$-smooth. When the regularization parameter $\lambda$ is very small, then the condition number is very large. Similar discussions have been applied to other regularizers.

## 2. Related Work

Since the proposal of variance reduction techniques were proposed by Johnson & Zhang (2013); Roux et al. (2012); Zhang et al. (2013), they have received tremendous attention. In this paper we are mostly interested in non-convex problems. Hence, below we review some related works for non-convex optimization in the form of (1).

A SVRG-type method for solving non-convex smooth optimization - a special case of (1) with $\psi = 0$ were first proposed by two research groups independently (Reddi et al., 2016a; Allen-Zhu & Hazan, 2016). The gradient complexity of non-convex SVRG is given by $O(n^{2/3}L/\epsilon^2)$ for finding an $\epsilon$-stationary solution such that $\mathrm{E}[\|\nabla\phi(\mathbf{x})\|] \leq \epsilon$. It was later generalized to solving the general case (1) with $\psi$ being a non-smooth convex function by (Reddi et al., 2016b;c), which also includes a SAGA-type method. There are two basic variants of SVRG proposed in (Reddi et al., 2016a;b;c) one with a large mini-batch size ($n^{2/3}$) and one with a small step size $\Theta(1/n^{2/3}L)$. In the first variant, the step size can be set to a large value $\Theta(1/L)$. In the second variant, the mini-batch size can be set to 1. However, neither variant is practical, especially with a small step size $\Theta(1/n^{2/3}L)$, which usually leads to slow convergence in practice. In contrast, the proposed method uses a large step size $\Theta(1/L)$ and allows for using a mini-batch size of 1.

Recently, there are several improvements on the gradient complexity for SVRG-type methods in terms of dependence on $n$. In particular, two new SVRG-type algorithms were proposed in (Fang et al., 2018; Zhou et al., 2018), namely SPIDER and stochastic nested variance reduction for solving the problem (1) with $\psi = 0$. The gradient complexity of both algorithms is given by $O(\sqrt{n}L/\epsilon^2)$ for finding an $\epsilon$-stationary solution when $\epsilon \leq O(1/\sqrt{n})$.

Few works have taken the $\mu$-weak convexity of individual functions $f_i$ into account for the development of variance reduction methods (Allen-Zhu, 2017a; Lan & Yang, 2018). Under the weakly convex assumption, Allen-Zhu (2017a) proposed a novel acceleration of SVRG-style method, namely Natasha1, which established a state-of-the-art gradient complexity when condition number is small, i.e. $\sqrt{n} \geq L/\mu$. In the same paper, Allen-Zhu (2017a) also discussed another method, namely RepeatSVRG [2], which could converge faster than Natasha1 under a large condition number. The proposed method is more practical than RepeatSVRG in that it does not require setting $\epsilon$ aprior as in RepeatSVRG. In a more recent work, Lan & Yang (2018) proposed an SAGA-type method, which has the same gradient complexity of this work except for a worse memory cost. It is the first-work for deriving an $\mu$-dependent complexity of a variance-reduction method for solving smooth non-convex optimization problems. Our work is complementary by developing a SVRG-type method with the same complexity and for solving a broader family of problems with a non-smooth convex function $\psi$.

It is notable that accelerating the convergence for strongly convex and smooth optimization problems with a large condition number has received a lot of attention in the community (Lin et al., 2015; Frostig et al., 2015; Lan & Zhou, 2018; Allen-Zhu, 2017b). Recently, Paquette et al. (2018) also considered extending the Catalyst technique for speeding up convex optimization algorithms to solving non-convex problem (1). However, their gradient complexity for using SVRG is only $O(nL/\epsilon^2)$, which is worse than our result. Finally, we present a comparison between this work and previous works for solving (1) in Table 1.

[2] After the preliminary version of this manuscript was finished, it was brought to our attention that the updated arXiv manuscript (Allen-Zhu, 2018, V5) reported a new result for RepeatSVRG for our considered problem different from its proceedings version, which is in the same order as the result achieved in this work. It is less practical than our method.

## 3. Katalyst

In this section we present the proposed Katalyst algorithm and its analysis. We first present some notations. For simplicity of presentation, we let $f = \sum_{i=1}^n f_i(\mathbf{x})/n$, and let $\mathrm{prox}_{\lambda\psi}(\mathbf{x}) = \arg\min_{\mathbf{z}} \psi(\mathbf{z}) + \frac{1}{2\lambda}\|\mathbf{z}-\mathbf{x}\|^2$ denote the proximal mapping of a function $\psi$. For problem (1), a point $\mathbf{x} \in \mathrm{dom}(\psi)$ is a first-order stationary point if $0 \in \partial\phi(\mathbf{x})$, where $\partial\phi$ denotes the partial gradient of $\phi$. However, it is hard for an iterative algorithm to find an exact stationary point with a finite number of iterations. Therefore, some notion of $\epsilon$-stationary is usually considered.

In the literature, several notions of $\epsilon$-stationarity were considered by accommodating the non-smooth term $\psi$ in different way. The first measure is simply using the subdifferentiable of the objective function $\phi$. Under this measure, a point $\mathbf{x}$ is said to be $\epsilon$-stationary if $\mathrm{dist}(0, \partial\phi(\mathbf{x})) \leq \epsilon$, where dist denotes the Euclidean distance from a point to a set and $\partial\phi(\mathbf{x}) = \nabla f(\mathbf{x}) + \partial\psi(\mathbf{x})$. The second measure is using the proximal gradient defined as:

$$\mathcal{G}_\eta(\mathbf{x}) = \frac{1}{\eta}(\mathbf{x} - \mathrm{prox}_{\eta\psi}(\mathbf{x} - \nabla f(\mathbf{x}))). \tag{3}$$

Under this measure, a point $\mathbf{x}$ is said to be $\epsilon$-stationary if $\|\mathcal{G}_\eta(\mathbf{x})\|^2 \leq \epsilon$. This convergence measure has been used in (Reddi et al., 2016c; Allen-Zhu, 2017a). The third stationarity measure that is more general is defined by using a notion of nearly stationary. In particular, a point $\mathbf{x}$ is called $(\epsilon, \delta)$-nearly stationary if there exists a point $\widehat{\mathbf{x}}$ such that

$$\|\mathbf{x} - \widehat{\mathbf{x}}\| \leq \delta, \quad \mathrm{dist}(0, \partial\phi(\widehat{\mathbf{x}})) \leq \epsilon. \tag{4}$$

This convergence measure has been used in (Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018a; Lan & Yang,

**Algorithm 1** Katalyst for Non-Convex Optimization

**Initialize:** non-decreasing positive weights $\{w_s\}$, $\mathbf{x}_0 \in \text{dom}(\psi)$, $\gamma = (2\mu)^{-1}$
**for** $s = 1, \ldots, S+1$ **do**
  Let $f_s(\cdot) = \phi(\cdot) + \frac{1}{2\gamma}\|\cdot - \mathbf{x}_{s-1}\|^2$
  $\mathbf{x}_s = \text{Katyusha}(f_s, \mathbf{x}_{s-1}, K_s, \mu, L+\mu)$
**end for**
**Return:** $\mathbf{x}_{\tau+1}$, $\tau$ is randomly chosen from $\{0, \ldots, S\}$ according to probabilities $p_\tau = \frac{w_{\tau+1}}{\sum_{k=0}^{S} w_{k+1}}, \tau = 0, \ldots, S$.

2018; Chen et al., 2018). The third convergence measure is more general that covers the first two measures as special cases. This can be easily seen for the first convergence measure with $\widehat{\mathbf{x}} = \mathbf{x}$ and $\delta = 0$. For the second convergence measure, we can show that when $\|\mathcal{G}_\eta(\mathbf{x})\| \leq \epsilon$ holds with $\eta = 1/L$, we have $\|\mathbf{x} - \mathbf{z}\| \leq \epsilon/L$ and $\text{dist}(0, \partial\phi(\mathbf{z})) \leq \|\mathcal{G}_\eta(\mathbf{x})\| + L\|\mathbf{x} - \mathbf{z}\| \leq 2\epsilon$, where $\mathbf{z} = \text{prox}_{\eta\psi}(\mathbf{x} - \nabla f(\mathbf{x}))$.

In this paper, we use the third stationarity measure that is same as that used in (Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018a; Lan & Yang, 2018; Chen et al., 2018), which is more suitable for our algorithm than other measures. To this end, we introduce the Moreau envelope of $\phi$

$$\phi_\lambda(\mathbf{x}) = \min_{\mathbf{z}} \phi(\mathbf{z}) + \frac{1}{2\lambda}\|\mathbf{z} - \mathbf{x}\|^2.$$

Further, the optimal solution to the above problem is $\text{prox}_{\lambda\phi}(\mathbf{x})$. It is known that if $\phi(\mathbf{x})$ is $\mu$-weakly convex and $\lambda < \mu^{-1}$, then its Moreau envelope $\phi_\lambda(\mathbf{x})$ is $C^1$-smooth with the gradient given by $\nabla\phi_\lambda(\mathbf{x}) = \lambda^{-1}(\mathbf{x} - \text{prox}_{\lambda\phi}(\mathbf{x}))$ (see e.g. (Davis & Drusvyatskiy, 2018b)). A small norm of $\nabla\phi_\lambda(\mathbf{x})$ has an interpretation that $\mathbf{x}$ is close to $\widehat{\mathbf{x}} = \text{prox}_{\lambda\phi}(\mathbf{x})$ that is $\epsilon$-stationary. In particular for any $\mathbf{x} \in \mathbb{R}^d$, let $\widehat{\mathbf{x}} = \text{prox}_{\lambda\phi}(\mathbf{x})$, then we have

$$\begin{cases} \phi(\widehat{\mathbf{x}}) \leq \phi(\mathbf{x}), \\ \|\mathbf{x} - \widehat{\mathbf{x}}\| = \lambda\|\nabla\phi_\lambda(\mathbf{x})\|, \\ \text{dist}(0, \partial\phi(\widehat{\mathbf{x}})) \leq \|\nabla\phi_\lambda(\mathbf{x})\|. \end{cases} \tag{5}$$

This means that a point $\mathbf{x}$ satisfying $\|\nabla\phi_\lambda(\mathbf{x})\| \leq \epsilon$ is close to a point in distance of $O(\epsilon)$ that is $\epsilon$-stationary. Below, we will prove the convergence in terms of $\|\nabla\phi_\lambda(\mathbf{x})\|$ for some $\lambda > 0$ and $\|\mathbf{x} - \text{prox}_{\lambda\phi}(\mathbf{x})\|$ as well, which is consistent with that in (Lan & Yang, 2018).

### 3.1. Algorithm

The Katalyst algorithm is presented in Algorithm 1, which falls into the same framework presented in (Chen et al., 2018). The idea is to construct a strongly convex function $f_s$ at each stage and then call a stochastic algorithm

**Algorithm 2** Katyusha($f, x_0, K, \sigma, \hat{L}$)

**Initialize:** $\tau_2 = \frac{1}{2}$, $\tau_1 = \min\{\sqrt{\frac{n\sigma}{3\hat{L}}}, \frac{1}{2}\}$, $\eta = \frac{1}{3\tau_1\hat{L}}$, $\theta = 1 + \eta\sigma$, $m = \lceil \frac{\log(2\tau_1 + 2/\theta - 1)}{\log\theta} \rceil + 1$
$y_0 = \zeta_0 = \widetilde{x}^0 \leftarrow x_0$
**for** $k = 0, \ldots, K-1$ **do**
  $u^k = \nabla\hat{f}(\widetilde{x}^k)$
  **for** $t = 0, \ldots, m-1$ **do**
    $j = km + t$
    $x_{j+1} = \tau_1\zeta_j + \tau_2\widetilde{x}^k + (1 - \tau_1 - \tau_2)y_j$
    $\widetilde{\nabla}_{j+1} = u^k + \nabla\hat{f}_i(x_{j+1}) - \nabla\hat{f}_i(\widetilde{x}^k)$
    $\zeta_{j+1} = \arg\min_\zeta \frac{1}{2\eta}\|\zeta - \zeta_j\|^2 + \langle\widetilde{\nabla}_{j+1}, \zeta\rangle + \hat{\psi}(\zeta)$
    Option I: $y_{j+1} = \arg\min_y \frac{3\hat{L}}{2}\|y - x_{j+1}\|^2 + \langle\widetilde{\nabla}_{j+1}, y\rangle + \hat{\psi}(\zeta)$
    Option II: $y_{j+1} = x_{j+1} + \tau_1(\zeta_{j+1} - \zeta_j)$
  **end for**
  compute $\widetilde{x}^{k+1} = \frac{\sum_{t=0}^{m-1}\theta^t y_{sm+t+1}}{\sum_{j=0}^{m-1}\theta^t}$
**end for**
**Output** $\widetilde{x}^K$

(Katyusha here) for approximately solving the constructed function. One may consider directly applying their Theorem 1 to prove the convergence. However, their analysis only concerns the convergence of $\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|$ without explicit considering the convergence of $\|\mathbf{x} - \text{prox}_{\gamma\phi}(\mathbf{x})\|$, which is important for proving the convergence of $\|\nabla\phi(\mathbf{x})\|$ when $\psi = 0$. By using the second inequality in (5), one can bound $2\mu\|\mathbf{x} - \text{prox}_{\gamma\phi}(\mathbf{x})\|$ by $\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|$. Nevertheless, in the case of $\mu \ll 1$, such analysis will yield much worse gradient complexity than that is achieved below. Hence, we need a more refined analysis of the proposed algorithm with a careful setting of Katyusha for solving each subproblem.

A modified Katyusha is employed at each stage for solving the regularized subproblem $f_s(\mathbf{x})$, which is assumed to be $\sigma$-strongly convex and have $\hat{L}$-Lipschitz continuous gradients for the smooth components. The modified Katyusha is presented in Algorithm 2. Given the way that $f_s$ is constructed, we can write it as

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

It is easy to see that $\hat{f}_i(\mathbf{x})$ is convex and $\hat{L} = (L + \mu)$-smooth, and $\hat{\psi}(\mathbf{x})$ is $\sigma = (\gamma^{-1} - \mu)$-strongly convex, which satisfy the conditions made in (Allen-Zhu, 2017b). In each call of the modified Katyusha, $\hat{f}_i$ is considered as the smooth component, and $\hat{\psi}$ is considered as the non-smooth regular-

izer. The key difference between our modified Katyusha and the original Katyusha algorithm for solving smooth and strongly convex problems in (Allen-Zhu, 2017b) lies at the setting of $\tau_1$, $m$ and $K$. For example in (Allen-Zhu, 2017b), the value of $\tau_1$ is set to $\tau_1 = \min(\sqrt{m\sigma/3\hat{L}}, 1/2)$. However, in our modified Katyusha the value of $\tau_1$ is independent of $m$. The value of $m$ is also different from that suggested in (Allen-Zhu, 2017b), which is suggested to $2n$. The value of $K$ (the number of epochs) in the original Katyusha is chosen such that the objective gap is less than $\epsilon$. In our modified Katyusha, it is set to make sure that the objective function $f_s(\mathbf{x})$ is decreased by a sufficient amount. Actually, we do not solve $\min_{\mathbf{x}} f_s(\mathbf{x})$ to an $\epsilon$-accuracy level in terms of the objective value. Below, we present the gradient complexity of Katalyst (i.e., the order of number of evaluations of $\nabla\phi_i(\mathbf{x})$) based on the following basic assumptions.

**Assumption 1.** *For problem (1), we assume that (i) $f_i(\cdot)$ is $L$-smooth and $\mu$-weakly convex, (ii) $\psi$ is a non-smooth convex function, and (iii) there exists $\Delta_\phi > 0$ such that $\phi(\mathbf{x}_0) - \min_{\mathbf{x}} \phi(\mathbf{x}) \leq \Delta_\phi$.*

**Theorem 1.** *Suppose Assumption 1 holds. 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 $\tau_1 = \min\{\sqrt{\frac{n\sigma}{3\hat{L}}}, \frac{1}{2}\}$, step size $\eta = \frac{1}{3\tau_1\hat{L}}$, $\tau_2 = 1/2$, $\theta = 1 + \eta\sigma$, and*

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

*where $D_s = \max\{24\hat{L}/\mu, 2\hat{L}^3/\mu^3, 8L^2 s/\mu^2\}$. Then we have that*

$$\max\{\mathrm{E}[\|\nabla\phi_\gamma(\mathbf{x}_{\tau+1})\|^2], \mathrm{E}[L^2\|\mathbf{x}_{\tau+1} - \mathbf{z}_{\tau+1}\|^2]\} \leq \frac{34\mu\Delta_\phi(\alpha+1)}{S+1} + \frac{48\mu\Delta_\phi(\alpha+1)}{(S+1)\alpha^{\mathbb{I}_{\alpha<1}}},$$

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

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

Indeed, when $\psi = 0$ we can derive a slightly stronger result stated in the following theorem.

**Theorem 2.** *Suppose Assumption 1 holds and $\psi = 0$. With the same parameter values as in Theorem 1 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*

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

**Remark:** Our results in the above two theorems match that in (Lan & Yang, 2018). Indeed, our result in Theorem 1 is slightly more general than that in (Lan & Yang, 2018), which only considers the constrained smooth optimization with $\psi$ being the indicator function of a convex set.

### 3.2. Analysis

We first state the convergence property of modified Katyusha (Algorithm 2) for solving following problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \hat{f}(\mathbf{x}) + \hat{\psi}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_i(\mathbf{x}) + \hat{\psi}(\mathbf{x}), \quad (6)$$

where each $\hat{f}_i$ is $\hat{L}$-smooth and convex, $\hat{\psi}(\mathbf{x})$ is $\sigma$-strongly convex.

**Theorem 3.** *(One call of Katyusha) Suppose that $\tau_1 = \min\{\sqrt{\frac{n\sigma}{3\hat{L}}}, \frac{1}{2}\}$, $\tau_2 = 1/2$, $\eta = \frac{1}{3\tau_1\hat{L}}$, and $m = \lceil\frac{\log(2\tau_1+2/\theta-1)}{\log\theta}\rceil + 1$. Defining $\theta := 1 + \eta\sigma$, $D_t := f(\mathbf{y}_t) - f(\mathbf{x})$, $\widetilde{D}^k := f(\widetilde{\mathbf{x}}^k) - f(\mathbf{x})$ for any $\mathbf{x}$, Algorithm 2 outputs a solution $\widetilde{\mathbf{x}}^K$ of problem (6) such that*

$$\mathrm{E}[\widetilde{D}^K] \leq 4\tau_1\theta^{-mK}\left(\frac{1-\tau_1}{\tau_1}\widetilde{D}^0 + \frac{1}{2\eta}\|\zeta_0 - \mathbf{x}\|^2\right). \quad (7)$$

The proof of above theorem is deferred to Appendix A.

*Proof.* [of Theorem 1] Given Thoerem 3, our analysis is divided into several parts. First, we verify the value of $K$ is a valid one. Then, we apply the above theorem to show the convergence for solving each constructed function $f_s$. Then, we prove the convergence of $\|\nabla\phi_\gamma(\mathbf{x}_{\tau+1})\|$, followed by the convergence analysis of $L\|\mathbf{x}_{\tau+1} - \mathbf{z}_{\tau+1}\|$. Then, we briefly prove Theorem 2. Finally, we derive the gradient complexity.

**Validation of $K$:** Overall, we need

$$\theta^{-mK} \leq \min\left\{\frac{\mu}{24\hat{L}}, \frac{\mu^3}{2\hat{L}^3}, \frac{\mu^2}{8L^2 s}\right\}.$$

Define $D_s = \max\{24\hat{L}/\mu, 2\hat{L}^3/\mu^3, 8L^2 s/\mu^2\} \geq 16$. We can set $K = \lceil\frac{\log(D_{\max})}{m\log\theta}\rceil$. Then,

$$K \geq \left\lceil \frac{4}{m \log \theta} \right\rceil \geq 1,$$

where the last inequality follows that $2/\theta \geq 2\tau_1 + 2/\theta - 1 \geq 1$ always hold according to the setting of $\tau_1 = \min\{\sqrt{\frac{n\mu}{3\hat{L}}}, \frac{1}{2}\}$ and $\eta = \frac{1}{3\tau_1 \hat{L}}$.

**Convergence of $\|\nabla\phi_\gamma(\cdot)\|$.** Let $\mathbf{z}_s = \arg\min_{\mathbf{x}} f_s(\mathbf{x})$ and $\mathbf{x}_*$ denote the global minimum of $\min_{\mathbf{x}} \phi(\mathbf{x})$. It is notable that $\|\mathbf{x}_{s-1} - \mathbf{z}_s\|/\gamma = \nabla\phi_\gamma(\mathbf{x}_{s-1})$. Below, we will use $K$ to denote $K_s$. $\mathrm{E}_s$ denotes the expectation over randomness in the $s$-th stage conditioned on all previous stages. Applying Theorem 3 to the $s$-th call of Katyusha, we have

$$\begin{aligned}\mathrm{E}_s[f_s(\mathbf{x}_s) - f_s(\mathbf{z}_s)] \leq & 4\theta^{-mK}(f_s(\mathbf{x}_{s-1}) - f_s(\mathbf{z}_s)) \\ & + \frac{2\tau_1\theta^{-mK}}{\eta}\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2. \quad (8)\end{aligned}$$

It is easy to see that

$$\begin{aligned}& f_s(\mathbf{x}_{s-1}) - f_s(\mathbf{z}_s)) \\ =& \phi(\mathbf{x}_{s-1}) - \phi(\mathbf{z}_s) - \frac{1}{2\gamma}\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2 \\ \leq& \phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*) - \frac{1}{2\gamma}\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2.\end{aligned}$$

Thus, we have

$$\begin{aligned}& \mathrm{E}_s[f_s(\mathbf{x}_s) - f_s(\mathbf{z}_s)] \\ \leq& 4\theta^{-mK}(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*)) + \frac{2\tau_1\theta^{-mK}}{\eta}\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2 \\ \leq& \underbrace{4\theta^{-mK}(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*)) + 2\theta^{-mK}\hat{L}\|\mathbf{x}_{s-1} - \mathbf{z}_s\|^2}_{\mathcal{E}_s}.\end{aligned}$$

Based on the above result and by utilizing the strong convexity of $f_s$ and simple algebra, we have the following result whose proof is in Appendix B.

**Lemma 1.** *Let $\Delta_s = \phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_s)$ and $\theta^{-mK} \leq \mu/(24\hat{L})$. Then we have that*

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

It implies that

$$\begin{aligned}& \|\nabla\phi_\gamma(\mathbf{x}_{s-1})\|^2 \\ \leq& \mathrm{E}_s[8\Delta_s/\gamma] + 96\theta^{-mK}(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*))/\gamma.\end{aligned}$$

Multiplying both sides by $w_s$, we have that

$$\begin{aligned}& w_s\mathrm{E}_s[\|\nabla\phi_\gamma(\mathbf{x}_{s-1})\|^2] \\ \leq& \mathrm{E}_s\Big[8w_s\Delta_s/\gamma + 96\theta^{-mK}w_s(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*))/\gamma\Big].\end{aligned}$$

By summing over $s = 1, \ldots, S+1$, we have

$$\begin{aligned}& \mathrm{E}[\sum_{s=1}^{S+1} w_s\|\nabla\phi_\gamma(\mathbf{x}_{s-1})\|^2] \\ \leq& \mathrm{E}\Big[\frac{8}{\gamma}\sum_{s=1}^{S+1} w_s\Delta_s + \frac{96}{\gamma}\sum_{s=1}^{S+1} w_s\theta^{-mK}(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*))\Big].\end{aligned}$$

Taking the expectation w.r.t. $\tau \in \{0, \ldots, S\}$, we have that

$$\begin{aligned}\mathrm{E}[\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|^2] \leq \mathrm{E}\Big[& \frac{8\sum_{s=1}^{S+1} w_s\Delta_s}{\gamma\sum_{s=1}^{S+1} w_s} \\ & + \frac{96\sum_{s=1}^{S+1} w_s\theta^{-mK}(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*))}{\gamma\sum_{s=1}^{S+1} w_s}\Big].\end{aligned}$$

Next, we bound the numerators of the two terms in the above bound. For the first term in the above bound, we use Lemma 3 in the Appendix C and have $\mathrm{E}\left[\sum_{s=1}^{S+1} w_s\Delta_s\right] \leq \Delta_\phi w_{S+1}$. We can bound the second term as following:

$$\begin{aligned}& \mathrm{E}\Big[\sum_{s=1}^{S+1} w_s\theta^{-mK}(\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*))\Big] \\ \leq& \sum_{s=1}^{S+1} w_s\theta^{-mK}\mathrm{E}[\phi(\mathbf{x}_{s-1}) - \phi(\mathbf{x}_*)] \leq \Delta_\phi\sum_{s=1}^{S+1} w_s\theta^{-mK},\end{aligned}$$

where we use the fact $\mathrm{E}[\phi(\mathbf{x}_s)] \leq \mathrm{E}[\phi(\mathbf{x}_{s-1})]$ as shown in the proof of Lemma 3 in the Appendix C. As a result,

$$\begin{aligned}& \mathrm{E}[\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|^2]] \\ \leq& \Big[\frac{8\Delta_\phi w_{S+1}}{\gamma\sum_{s=1}^{S+1} w_s} + \frac{96\Delta_\phi\sum_{s=1}^{S+1} w_s\theta^{-mK}}{\gamma\sum_{s=1}^{S+1} w_s}\Big] \\ \leq& \Big[\frac{8\Delta_\phi w_{S+1}}{\gamma\sum_{s=1}^{S+1} w_s} + \frac{12\Delta_\phi\sum_{s=1}^{S+1} w_s s^{-1}}{\gamma\sum_{s=1}^{S+1} w_s}\Big],\end{aligned}$$

where we use the fact $\theta^{-mK} \leq 1/(8s)$. Then by simple algebra (cf. (Chen et al., 2018)), we have

$$\mathrm{E}[\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|^2] \leq \frac{16\mu\Delta_\phi(\alpha+1)}{S+1} + \frac{24\mu\Delta_\phi(\alpha+1)}{(S+1)\alpha^{\mathbb{I}_{\alpha<1}}}.$$

Due to the objective decreasing property, we have

$$\mathrm{E}[\phi(\mathbf{x}_s) + \frac{1}{2\gamma}\|\mathbf{x}_s - \mathbf{x}_{s-1}\|^2 - \phi(\mathbf{x}_{s-1})] \leq 0,$$

which implies by a similar analysis

$$\frac{1}{2\gamma}\mathrm{E}[\|\mathbf{x}_{\tau+1} - \mathbf{x}_\tau\|^2] \leq \frac{\Delta_\phi(\alpha+1)}{S+1}.$$

Since $\phi_\gamma(\mathbf{x})$ has $(\gamma^{-1} - \mu)$-Lipschitz continuous gradient (cf. Lemma 2.1 in (Drusvyatskiy & Paquette, 2018)), then

we have

$$\begin{aligned}
&\mathrm{E}[\|\nabla\phi_\gamma(\mathbf{x}_{\tau+1})\|^2]\\
\leq&2\mathrm{E}[\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|^2]+2(\gamma^{-1}-\mu)^2\mathrm{E}[\|\mathbf{x}_{\tau+1}-\mathbf{x}_\tau\|^2]\\
\leq&2\mathrm{E}[\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|^2]+\frac{2\mu\Delta_\phi(\alpha+1)}{S+1}\\
\leq&\frac{34\mu\Delta_\phi(\alpha+1)}{S+1}+\frac{48\mu\Delta_\phi(\alpha+1)}{(S+1)\alpha^{\mathbb{I}_{\alpha<1}}}.
\end{aligned}$$

**Convergence of $L\|\mathbf{x}_{\tau+1}-\mathbf{z}_{\tau+1}\|$.** By the strong convexity of $f_s$, we have $\mathrm{E}[\|\mathbf{x}_s-\mathbf{z}_s\|^2]\leq\frac{2}{\sigma}\mathcal{E}_s$. To proceed, we have

$$\begin{aligned}
&L^2\|\mathbf{x}_s-\mathbf{z}_s\|^2\leq\frac{4L^2\theta^{-mK}}{\sigma}(2(\phi(\mathbf{x}_{s-1})-\phi(\mathbf{x}_*))\\
&+\hat{L}\|\mathbf{x}_{s-1}-\mathbf{z}_s\|^2)\\
\leq&2\theta^{-mK}\left[\frac{4L^2}{\sigma}(\phi(\mathbf{x}_{s-1})-\phi(\mathbf{x}_*))+\frac{\hat{L}^3}{\mu^3}\|\nabla\phi_\gamma(\mathbf{x}_{s-1})\|^2\right]\\
\leq&\frac{8L^2\theta^{-mK}}{\sigma}(\phi(\mathbf{x}_{s-1})-\phi(\mathbf{x}_*))+\|\nabla\phi_\gamma(\mathbf{x}_{s-1})\|^2,
\end{aligned}$$

where we use the fact $\|\mathbf{x}_{s-1}-\mathbf{z}_s\|/\gamma=\|\nabla\phi_\gamma(\mathbf{x}_{s-1})\|$ and $\theta^{-mK}\leq\mu^3/(2\hat{L}^3)$. Then following the same analysis as above,

$$\begin{aligned}
&\mathrm{E}[L^2\|\mathbf{x}_{\tau+1}-\mathbf{z}_{\tau+1}\|^2]\\
\leq&\frac{8L^2\Delta_\phi\sum_{s=1}^{S+1}w_s\theta^{-mK}}{\sigma\sum_{s=1}^{S+1}w_s}+\mathrm{E}[|\nabla\phi_\gamma(\mathbf{x}_\tau)|^2].
\end{aligned}$$

Since $\theta^{-mK}\leq\mu^2/(8L^2s)$, then

$$\mathrm{E}[L^2\|\mathbf{x}_{\tau+1}-\mathbf{z}_{\tau+1}\|^2]\leq\frac{\mu\Delta_\phi(16+25\alpha^{-\mathbb{I}_{\alpha<1}})(\alpha+1)}{S+1}$$

When $\psi(\cdot)=0$ and considering $\alpha$ as a constant, we have

$$\begin{aligned}
&\mathrm{E}[\|\nabla\phi(\mathbf{x}_{\tau+1})\|^2]\\
\leq&\mathrm{E}[\|\nabla\phi(\mathbf{x}_{\tau+1})-\nabla\phi(\mathbf{z}_{\tau+1})+\nabla\phi(\mathbf{z}_{\tau+1})\|^2]\\
\leq&2\mathrm{E}[L^2\|\mathbf{x}_{\tau+1}-\mathbf{z}_{\tau+1}\|^2+\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|^2]\leq O\left(\frac{\mu\Delta_\phi}{S+1}\right).
\end{aligned}$$

Indeed, for $\psi(\cdot)=0$, we can do slightly better by bounding $f_s(\mathbf{x}_{s-1})-f_s(\mathbf{z}_s)\leq\frac{\hat{L}}{2}\|\mathbf{x}_{s-1}-\mathbf{z}_s\|^2$. Then $\mathcal{E}_s$ becomes $4\theta^{-mK}\hat{L}\|\mathbf{x}_{s-1}-\mathbf{z}_s\|^2$ and $\theta^{-mK}(\phi(\mathbf{x}_{s-1})-\phi(\mathbf{z}_s))$ in the proceeding analysis is gone, which removes the requirement $\theta^{-mK}\leq\mu^2/(16L^2s)$. As a result, we can set $K=\lceil\log(D)/(m\log\theta)\rceil$, where $D=\max(48L/\mu,2\hat{L}^3/\mu^3)$.

**Gradient Complexity:** Finally, we analyze the gradient complexity. Let us consider the gradient complexity at the $s$-th stage, which is

$$\begin{aligned}
(n+m)K\leq&\frac{2\log(D_s)}{\log(2\tau_1+2\tau_1\eta\mu+1-\eta\mu)}n\\
&+\frac{2\log(D_s)}{\log(1+\eta\mu)}.
\end{aligned}$$

Let $\tau_1=\frac{c}{\eta\mu}$, where $0\leq c=\frac{\mu}{3\hat{L}}\leq\frac{1}{3}$. We have that

$$\begin{aligned}
&(n+m)K=nK+mK\\
\leq&\frac{2\log(D_s)}{\log(2\tau_1+2\tau_1\eta\mu+1-\eta\mu)}n+\frac{2\log(D_s)}{\log(1+\eta\mu)}\\
\leq&\frac{2\log(D_s)}{\log(2\tau_1+2c+1-\frac{c}{\tau_1})}n+\frac{2\log(D_s)}{\log(1+\frac{c}{\tau_1})}.
\end{aligned}$$

We analyze two cases.

**Case 1:** If $n\geq\frac{3\hat{L}}{4\mu}$, then $\tau_1=\frac{1}{2}$, we have that

$$(n+m)K\leq O\bigg(\log(D_s)n+\frac{\log D_s}{\log(1+2c)}\bigg).$$

Since $0\leq c\leq1/3$ so $\log(1+2c)\geq c$, then

$$(n+m)K\leq O\bigg((n+\frac{\hat{L}}{\mu})\log D_s\bigg).$$

Then the total gradient complexity for finding $\mathrm{E}\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|^2\leq\epsilon^2$ is $O((\mu n+L)\log(L/(\mu\epsilon))$.

**Case 2:** If $n\leq\frac{3\hat{L}}{4\mu}$, then $\tau_1=\sqrt{\frac{n\mu}{3\hat{L}}}\in(0,\frac{1}{2}]$. We have that following inequalities hold

$$\frac{\tau_1}{c}=\sqrt{\frac{3n\hat{L}}{\mu}}\leq\frac{3\hat{L}}{2\mu},\quad\log(\frac{\tau_1+c}{c})\leq\log\frac{3\hat{L}}{\mu},$$
$$\log(1+c/\tau_1)\geq c/(2\tau_1)$$

and due to (1) if $2\tau_1+2c-\frac{c}{\tau_1}\leq1/2$, then $\log(2\tau_1+2c+1-\frac{c}{\tau_1})\geq\tau_1+c-c/(2\tau_1)$; (2) if $2\tau_1+2c-\frac{c}{\tau_1}\geq1/2$, then $\log(2\tau_1+2c+1-\frac{c}{\tau_1})\geq\log(1.5)$, we have

$$\begin{aligned}
&\frac{1}{\log(2\tau_1+2c+1-\frac{c}{\tau_1})}\\
\leq&\max\left\{\frac{1}{\log(1.5)},\frac{1}{\sqrt{\frac{n\mu}{3\hat{L}}}+\frac{\mu}{3\hat{L}}-\sqrt{\frac{\mu}{12n\hat{L}}}}\right\}\leq O\bigg(\sqrt{\frac{\hat{L}}{n\mu}}\bigg).
\end{aligned}$$

Thus we have $(n+m)K\leq O\bigg(\sqrt{\frac{n\hat{L}}{\mu}}\log\frac{\hat{L}}{\mu}\bigg)$, and the total gradient complexity for finding $\mathrm{E}\|\nabla\phi_\gamma(\mathbf{x}_\tau)\|^2\leq\epsilon^2$ is $O(\sqrt{\mu nL}\log(L/(\mu\epsilon))$.

□

## 4. Experiments

In this section, we conduct some experiments for solving regularized classification problem in the form of (2) with $\ell(\mathbf{x};\mathbf{a}_i,b_i)=\frac{1}{2}(\max(0,1-b_i\mathbf{a}_i^\top\mathbf{x}))^2$ being a squared hinge loss that is more suitable for classification.

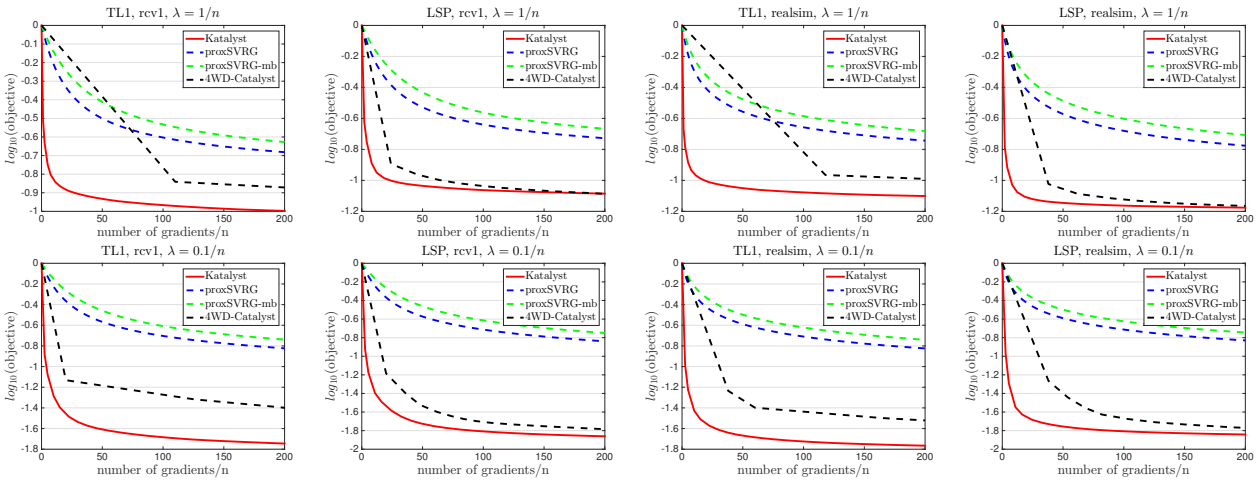


*Figure 1.* Comparison of different algorithms for two tasks on different datasets

**Penalties and Parameters.** We choose two different non-convex and non-smooth penalty functions as the regularizers, namely log-sum penalty (LSP) $R(\mathbf{x}) = \sum_{i=1}^{d} \log(\beta + |x_i|), (\beta > 0)$ and transformed $\ell_1$ (TL1) penalty $R(\mathbf{x}) = \sum_{i=1}^{d} \frac{(\beta+1)|x_i|}{\beta+|x_i|}, (\beta > 0)$ where $\beta > 0$ is a parameter. Both LSP and TL1 can be written as a difference of convex functions: $R(\mathbf{x}) = r_1(\mathbf{x}) - r_2(\mathbf{x})$, where $r_1(\mathbf{x})$ is a scaled $\ell_1$ norm and $r_2(x)$ is smooth and convex (cf. details provided in Appendix D). Then the problem becomes

$$\min_{\mathbf{x}\in\mathbb{R}^d} \phi(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{(\ell(\mathbf{x}; \mathbf{a}_i, b_i) - r_2(\mathbf{x}))}_{f_i(\mathbf{x})} + \underbrace{r_1(\mathbf{x})}_{\Psi(\mathbf{x})},$$

For LSP, it is easy to show that the weakly convexity parameter and smoothness parameter of $f_i(\mathbf{x})$ are given by $\mu = \frac{\lambda}{\beta^2}$ and $L = \frac{\lambda}{\beta^2} + \max_{1\le i\le d} \|\mathbf{a}_i\|^2$. For TL1, it is easy to show that the weakly convexity parameter and smoothness parameter of $f_i(\mathbf{x})$ are given by $\mu = \frac{2(\beta+1)\lambda}{\beta^2}$ and $L = \frac{2(\beta+1)\lambda}{\beta^2} + \max_{1\le i\le d} \|\mathbf{a}_i\|^2$. We fix $\beta = 1$ but set two different values of $\lambda \in \{\frac{1}{n}, \frac{0.1}{n}\}$. The experiments are performed on two data sets from libsvm website (Chang & Lin, 2011), namely rcv1 ($n = 20,242$ and $d = 47,236$) and real-sim ($n = 72,309$ and $d = 20,958$).

**Baselines and Settings.** We compare the proposed Katalyst with proxSVRG, its mini-batch variant (named proxSVRG-mb in experiments) (Reddi et al., 2016c) and 4WD-Catalyst (Paquette et al., 2018). Other algorithms like RapGrad, SPIDER, SNVRG are not applicable to the considered problem. Since smoothness parameter $L$ and weak convexity parameter $\mu$ are given as discussed above, we implement Algorithm 1 in (Paquette et al., 2018) for 4WD-Catalyst. All parameters in three baselines including step size and the number of iterations for the inner loop are set to their theoretical values suggested in the original papers. We also compared Katalyst with RapGrad (Lan & Yang, 2018) for solving least square regression problem with smoothed SCAD penalty (Lan & Yang, 2018). The detailed description of experiments and the results are included in the Appendix.

**Results.** We report the results in Figure 1, where the x-axis is (number of gradients)$/n$ and the y-axis is log-scale of the objective value. For 4WD-Catalyst, we only plot the result at the end of each stage since it selects the better solution of two sub-problems. It is worth noting that we do not include the complexity of computing $f_\kappa(\bar{\mathbf{x}}_s; \mathbf{x}_{s-1})$ in solving sub-problem for 4WD-Catalyst, i.e. (Paquette et al., 2018, eqn. (7) of Algorithm 1), which would introduce more CPU time in practice.

We can observe that when using a smaller $\lambda$ that gives a smaller value of $\mu$-convexity parameter, Katalyst has relatively larger speed-up compared with the two variants of proxSVRG, which supports the presented complexity of Katalyst that is adaptive to the weakly convex property. Katalyst is also more efficient than 4WD-Catalyst, which needs to solve two sub-problems at each stage to satisfy a certain criterion that requires many iterations in practice.

## 5. Conclusion

In this paper, we have developed a SVRG-type accelerated stochastic algorithm for solving a family of non-convex optimization problems whose objective consists of a finite-sum of smooth functions and a non-smooth convex function. We proved that the gradient complexity can be improved when the condition number is very large compared to the number of smooth components, which achieves the best complexity among all SVRG-type methods and also matches that of an existing SAGA-type stochastic algorithm.

## Acknowledgement

The authors thank the anonymous reviewers for their helpful comments. We thank Prof. Enhong Chen for his assistance, who is supposed to be in the author list. Some work of Zaiyi Chen was done while he was a Ph.D. candidate in the University of Science and Technology of China. Yi Xu and Tianbao Yang are partially supported by National Science Foundation (IIS-1545995).

## References

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, 2017a.

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, 2017b.

Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non- convex parameter. *CoRR*, /abs/1708.08694/V5, 2018.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pp. 699–707, 2016.

Candès, E. J., Wakin, M. B., and Boyd, S. P. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, Dec 2008.

Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, Z., Yang, T., Yi, J., Zhou, B., and Chen, E. Universal stagewise learning for non-convex problems with convergence on averaged solutions. *CoRR*, /abs/1808.06296, 2018.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *CoRR*, abs/1803.06523, 2018a.

Davis, D. and Drusvyatskiy, D. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *CoRR*, /abs/1802.02988, 2018b.

Davis, D. and Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *arXiv preprint arXiv:1707.03505*, 2017.

Defazio, A., Bach, F. R., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1646–1654, 2014.

Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, Jul 2018.

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.

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.

Frostig, R., Ge, R., Kakade, S., and Sidford, A. Unregularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2540–2548. JMLR.org, 2015.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (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.

Lan, G. and Zhou, Y. An optimal randomized incremental gradient method. *Math. Program.*, 171(1-2):167–215, 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., Hefny, A., Sra, S., Póczós, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, pp. 314–323. JMLR.org, 2016a.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. Fast incremental method for smooth nonconvex optimization. In *55th IEEE Conference on Decision and Control (CDC)*, pp. 1971–1977, 2016b.

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, 2016c.

Roux, N. L., Schmidt, M. W., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2672–2680, 2012.

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.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894 – 942, 2010.

Zhang, L., Mahdavi, M., and Jin, R. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 980–988. 2013.

Zhang, S. and Xin, J. Minimization of transformed l_1 penalty: Theory, difference of convex function algorithm, and robust application in compressed sensing. *CoRR*, abs/1411.5735, 2014.

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