Novel stepsize for some accelerated and stochastic optimization methods

Nguyen Phung Hai Chung, Hoang Van Chung and Pham Thi Hoai

Abstract

Abstract. New first-order methods now need to be improved to keep up with the constant developments in machine learning and mathematics. They are commonly used methods to solve optimization problems. Among them, the algorithm branch based on gradient descent has developed rapidly with good results achieved. Not out of that trend, in this article, we research a new method combined with acceleration methods to provide updated results for optimization problems commonly found in machine learning. Besides, realizing the remarkable increase in parameters in recent deep learning models, we also research stochastic methods to increase speed in optimizing model parameters. Also in this article, theories to prove the convergence of the proposed algorithms are also given and there are experiments to prove the effectiveness of those.

1 Introduction

Along with the strong development of machine learning in recent years, first-order methods have also been rapidly improved and play a core role in optimizing models. The most popular is gradient descent and its variations as they are always the choice for problems in machine learning. Previously, gradient descent was also a key method in solving nonlinear problems and other problems in mathematics. With such wide application, a lot of research revolves around gradient descent. Most of them are methods to improve acceleration and step size to achieve new results in convergence theory and experiment. With the motivation from those studies, we propose some methods that combine several acceleration methods with new step sizes. We present them and their stochastic versions in detail along with theorems proving convergence and experiments. Also in this study, some new result for machine learning and mathematics problems are presented.

Throughout the study, we use some notations for convenience in presenting algorithms and theory. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the objective function that we want to optimize, the problem formulation that we consider is the basic unconstrained optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

where $d \in \mathbb{N}$ be the dimension or the number of parameters of the function, $[d] = \{1, 2, 3, ..., d\}$. We assume that (1) has a solution and denote its optimal value by f^*. We denote ∇f be the gradient of f. Assume that at each iteration we get sample ξ^k to make a stochastic gradient of f denoted as $\nabla_{\xi^k} f(x^k)$, which in this paper is briefly written as $\nabla_\xi f(x^k)$. We also assume that $\mathbb{E}\left[\nabla_\xi f(x)\right] = \nabla f(x)$ for all $x \in \mathbb{R}^d$ and note $\mathbb{E}_{n-1}\left[\cdot\right]$ the conditional expectation knowing $f_1, ..., f_{n-1}$. Finally, we note $\{\varsigma_k\}$ is a sequence and ς_k is its k-th component.

Gradient Descent, the mehod originally proposed by Augustin-Louis Cauchy[1], is a first-order optimization method with the idea of updating the variable $x^k \in \mathbb{R}^d$ at each iteration $k > 0$ with the formula

$$x^{k+1} = x^k - \lambda_k \nabla f(x^k)$$

where $\lambda_k > 0$ is the stepsize at iteration k. In efforts to improve Gradient Descent, the accelerated methods of Boris T. Polyak, Gradient Descent with Momentum or Heavy Ball method[15] and Nesterov's Accelerated Gradient of Yurii Nesterov[12] became the most prominent. By improving the

variable x^k update method, Gradient Descent with Momentum

$$v^{k+1} = \gamma v^k + \lambda_k \nabla f\left(x^k\right)$$
$$x^{k+1} = x^k - v^{k+1}$$

and Nesterov's Accelerated Gradient

$$v^{k+1} = \gamma v^k + \nabla f\left(x^k\right)$$
$$x^{k+1} = x^k - \lambda_k\left(\gamma v^{k+1} + \nabla f\left(x^k\right)\right)$$

where $0 \leq \gamma < 1$ is momentum factor, have achieved many achievements and significantly improved results on optimization problems. The above methods are also opportunities to open up new research directions based on Gradient Descent to improve results and solve optimization problems. Inheriting such motivation along with learning about related research in section 2, in this study we combine acceleration methods based on Gradient Descent with new step sizes to create new methods in section 4. Besides, we also propose stochastic methods in section 5. Along with the proposed methods, we present proof of their convergence in sections and conduct experiments to measure effectiveness when applied to problems in the fields of mathematics, machine learning and deep learning in section 6.

2 Related work

Since their appearance, Gradient Descent with momentum[15] and Nesterov's Accelerated Gradient[12] have achieved many achievements in the field of optimization. They open up new research directions to improve the above algorithms and prove them with new methods. The study [11] proposed an adaptive step size for Gradient Descent and achieved various good results in optimization problems when combined with acceleration. The study [5] provides proof of global convergence and global intercepts of the convergence rate of Gradient Descent with momentum and Nesterov's Accelerated Gradient. In addition, [10] provides an improved analysis that shows Stochastic Gradient Descent with momentum converges as quickly as Stochastic Gradient Descent on a smooth objective function, with both strongly convex and non-convex settings. The study [16] showed that Stochastic Gradient Descent with momentum converges with a convergence rate $O((1-\gamma)^-2)$ assuming that the gradients are bounded. In [3], they improve this ratio to $O((1-\gamma)^-1)$. The study [10] obtained similar results but with weaker assumption. And in [2], they provided an improved analysis of Stochastic Gradient Descent with momentum with a tight Liapunov analysis.

3 Step Size

Algorithm 1 Novel step size

Inittialization. Select $\lambda_0 > 0, 0 < \eta_1 < \eta_0$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Choose $x^0 \in \mathbb{R}^d$.
$x^1 = x^0 - \lambda_0 \nabla f(x^0)$
for $k = 1, 2, \ldots$ **do**
  if $\lambda_{k-1} > \eta_0 \frac{\left\|x^k - x^{k-1}\right\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$ **then**
    $\lambda_k = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$
  else
    $\lambda_k = (1 + \varepsilon_{k-1})\, \lambda_{k-1}$
  end if
  $x^{k+1} = x^k - \lambda_k \nabla f(x^k)$
end for

We provide the step size lemmas that are used in all the algorithms proposed in this study. To facilitate proving the convergence of the algorithms, we show that the step size has lower bounded and converges.

Lemma 3.1. *[9] Let $\{\lambda_k\}$ be the sequence generated by Algorithm 1 where f is smooth and its gradient is L-Lipschitz continuous, then*

$$\lambda_k \geq \min(\lambda_0, \frac{\eta_1}{L}) \quad \forall k \geq 0$$

and if f is smooth, strongly convex and its gradient is L-Lipschitz continuous, then

$$\lambda_k < \frac{(1+\varepsilon_{k-1})\eta_0}{\mu} \quad \forall k \geq 0$$

Proof. It's obviously true with $k = 0$. With $k \geq 1$, we consider two case:

- If $\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\| > \frac{\eta_0}{\lambda_{k-1}} \left\|x^k - x^{k-1}\right\|$ then $\lambda_k = \frac{\eta_1 \|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$. Because of L-smooth assumption on f, we have $\|\nabla f(x^k) - \nabla f(x^{k-1}\| \leq L\|x^k - x^{k-1}\|$, so $\lambda_k \geq \frac{\eta_1}{L}$.
- If $\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\| < \frac{\eta_0}{\lambda_{k-1}} \left\|x^k - x^{k-1}\right\|$ then $\lambda_k = (1+\varepsilon_{k-1})\lambda_{k-1} \geq \lambda_{k-1}$.

By induction we get that $\forall k \geq 0, \lambda_k \geq \min(\lambda_0, \frac{\eta_1}{L})$. Moreover, if f is smooth, strongly convex and its gradient is L-Lipschitz continuous, we can deduce that

$$\forall k \geq 0, \lambda_k < \frac{(1+\varepsilon_{k-1})\eta_0 \left\|x^k - x^{k-1}\right\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|} = \frac{(1+\varepsilon_{k-1})\eta_0}{\mu}$$

□

Lemma 3.2. *[9] Let $\{\lambda_k\}$ be the sequence generated by Algorithm (1) where f is smooth and its gradient is L-Lipschitz continuous, then $\{\lambda_k\}$ converges to $\overline{\lambda} < \infty$.*

Proof. Firstly we will proof that $\ln(\frac{\lambda_{k+1}}{\lambda_k}) \leq \ln(1+\varepsilon_k), \forall k \geq 0$. Let's consider two case:

- If $\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\| > \frac{\eta_0}{\lambda_{k-1}} \left\|x^k - x^{k-1}\right\| \Leftrightarrow \frac{\eta_0 \|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|} < \lambda_{k-1}$ then

  $$\lambda_k = \frac{\eta_1 \|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|} \overset{\eta_1 \leq \eta_0}{<} \frac{\eta_0 \|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|} < \lambda_{k-1},$$

  Then $\frac{\lambda_k}{\lambda_{k-1}} < 1 \Rightarrow \frac{\lambda_k}{\lambda_{k-1}} < 1 + \varepsilon_{k-1}$ (because $\{\varepsilon_k\}$ is a positive sequence).
- If $\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\| < \frac{\eta_0}{\lambda_{k-1}} \left\|x^k - x^{k-1}\right\|$ then $\lambda_k = (1+\varepsilon_{k-1})\lambda_{k-1}$. Then $\frac{\lambda_k}{\lambda_{k-1}} = 1 + \varepsilon_{k-1}$.

From two case, we have

$$\frac{\lambda_k}{\lambda_{k-1}} \leq 1 + \varepsilon_{k-1}, \quad \forall k \geq 1$$

$$\Leftrightarrow \frac{\lambda_{k+1}}{\lambda_k} = 1 + \varepsilon_k, \quad \forall k \geq 0$$

Secondly we will show the main result of the Lemma. Let $a_k = \ln(\lambda_{k+1}) - \ln(\lambda_k)$. We have $a_k = a_k^+ - a_k^-$, where $a_k^+ = \max(0, a_k), a_k^- = -\min(0, a_k)$. So $a_k^+ \geq 0, a_k^- \geq 0, \forall k \geq 0$. We have

$$a_k = \ln(\frac{\lambda_{k+1}}{\lambda_k}) \leq \ln(1+\varepsilon_k) \leq \varepsilon_k, \forall k \geq 0,$$

so $a_k^+ \leq \varepsilon_k$. Because $\sum_{k=0}^{\infty} \varepsilon_k$ is convergent, we have $\sum_{k=0}^{\infty} a_k^+$ is convergent.
Consider

$$\ln(\lambda_{k+1}) - \lambda_0 = \sum_{i=0}^{k} a_i = \sum_{i=0}^{k} (a_i^+ - a_i^-) = \sum_{i=0}^{k} a_i^+ - \sum_{i=0}^{k} a_i^-$$

Assert $\lim_{k \to +\infty} \sum_{i=0}^{k} a_i^- = +\infty$ then $\lim_{k \to +\infty} \ln(\lambda_k) = -\infty \Leftrightarrow \lim_{k \to +\infty} \lambda_k = 0$. But in Lemma (3.1) we showed that $\lambda_k \geq \min(\lambda_0, \frac{\eta_1}{L}) > 0, \forall k \geq 0$. So $\sum_{k=0}^{+\infty} a_k^-$ is convergent. Because of that, we have $\lim_{k \to +\infty} ln(\lambda_k) < +\infty \Rightarrow \lim_{k \to +\infty} \lambda_k < +\infty$ □

Lemma 3.3. *There exists a fixed number $\bar{k}$ such that*

$$\left\|\nabla f\left(x^k\right)-\nabla f\left(x^{k-1}\right)\right\| \leq \frac{\eta_0}{\lambda_{k-1}}\left\|x^k-x^{k-1}\right\| \quad \forall k \geq \bar{k}$$

and therefore $\lambda_k > \lambda_{k-1}\, \forall k \geq \bar{k}$

Proof. Suppose by contradiction that there exists $\{k_j\}, k_j \to +\infty$ such that

$$\left\|\nabla f\left(x^{k_j}\right)-\nabla f\left(x^{k_j-1}\right)\right\| > \frac{\eta_0}{\lambda_{k_j-1}}\left\|x^{k_j}-x^{k_j-1}\right\|.$$

For this case

$$\lambda_{k_j}=\eta_1 \frac{\left\|x^{k_j}-x^{k_j-1}\right\|}{\left\|\nabla f\left(x^{k_j}\right)-\nabla f\left(x^{k_j-1}\right)\right\|}$$

Consequently,

$$\frac{\eta_1\left\|x^{k_j}-x^{k_j-1}\right\|}{\lambda_{k_j}}=\left\|\nabla f\left(x^{k_j}\right)-\nabla f\left(x^{k_j-1}\right)\right\| > \frac{\eta_0}{\lambda_{k_j-1}}\left\|x^{k_j}-x^{k_j-1}\right\|$$

i.e.,

$$\frac{\lambda_{k_j}}{\lambda_{k_j-1}} < \frac{\eta_1}{\eta_0} \quad \forall k_j$$

On the other hand, from Lemma 3.2 we have

$$\lim_{k_j \to +\infty} \lambda_{k_j} = \lim_{k_j \to +\infty} \lambda_{k_j-1} = \lim_{k \to +\infty} \lambda_k = \lambda^*. \tag{2}$$

hence we deduce that

$$\frac{\lambda^*}{\lambda^*} \leq \frac{\eta_1}{\eta_0} < 1$$

It is a contradiction and we finish the proof. □

4 Accelerated NGD

In this section, we propose two accelerated method. The algorithm 2 use our new step size for Gradient Descent with momentum (or Heavy Ball method) and called Novel Gradient Descent with momentum (NGDm). Beside, the algorithm 3 also use our new step size for Nesterov's Accelerated Gradient and called Novel Gradient Descent with Nesverov's accelerated (NGD Nesterov). After present the algorithms, the proves of convergence are also presented.

4.1 Assumptions

Assumption 4.1. *We assume that f is lower bounded by f^**

$$\forall x \in \mathbb{R}^d, f(x) \geq f^*.$$

Assumption 4.2. *We assume f is smooth and μ-strongly convex, i.e its gradient is L-Lipschitz continuous:*

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

With above assumptions, we can prove that the algorithm 2 and algorithm 3 converges under certain parameter conditions.

Algorithm 2 Novel Gradient Descent with momentum (NGDm)

Inittialization. Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0, 0 < \gamma < 1$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Choose $x^0 \in \mathbb{R}^n, \lambda_{-1} = \lambda_0$.
$v^1 = \nabla f\left(x^0\right)$
$x^1 = x^0 - \lambda_0 v^1$
for $k = 1, 2, \ldots$ **do**
  if $\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\| > \frac{\eta_0}{\lambda_{k-1}} \left\|x^k - x^{k-1}\right\|$ **then**
    $\lambda_k = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$
  else
    $\lambda_k = (1 + \varepsilon_{k-1})\, \lambda_{k-1}$
  end if
  $v^{k+1} = \gamma v^k + \lambda_k \nabla f\left(x^k\right)$
  $x^{k+1} = x^k - v^{k+1}$
end for

4.2 Heavy Ball Method

Theorem 4.1. *Under assumptions in 4.1, if* $\frac{(1+\varepsilon_{k-1})\eta_0}{1-\gamma} \leq \frac{\mu}{L}\, \forall k \in \mathrm{N}$ *and* $0 \leq \gamma < 1$*, the sequence* $\{x^k\}$ *generated by algorithm 2 satisfies*

$$f(\overline{x}^K) - f^* \leq \frac{1}{K+1}\left(\frac{\gamma}{1-\gamma}(f(x^0) - f^*) + \frac{1}{2\lambda_0(1-\gamma)}\|x^0 - (1-\gamma)x^*\|^2 + \frac{C_1}{1-\gamma}\right)$$

where

$$C_1 = \sum_{k=1}^{\bar{k}-1}\left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k-1}}\right)\|x^k - \gamma x^{k-1} - (1-\gamma)x^*\|^2$$

and $\bar{k}$ *satisfy lemma 3.3 and*

$$\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^{K} x^k$$

Proof. Rearranging the algorithm 2, at each iteration k, we have

$$x^{k+1} - \gamma x^k = x^k - \gamma x^{k-1} - \lambda_k \nabla f(x^k) \tag{3}$$

Apply norm to (3), we get

$$\begin{aligned}\|x^{k+1} - \gamma x^k - (1-\gamma)x^*\|^2 &= \|x^k - \gamma x^{k-1} - (1-\gamma)x^*\|^2 + \lambda_k^2\|\nabla f(x^k)\|^2 \\ &\quad - 2\langle x^k - \gamma x^{k-1} - (1-\gamma)x^*, \lambda_k \nabla f(x^k)\rangle \\ &= \|x^k - \gamma x^{k-1} - (1-\gamma)x^*\|^2 + \lambda_k^2\|\nabla f(x^k)\|^2 \\ &\quad - 2(1-\gamma)\lambda_k\langle x^k - x^*, \nabla f(x^k)\rangle - 2\gamma\lambda_k\langle x^k - x^{k-1}, \nabla f(x^k)\rangle \end{aligned} \tag{4}$$

Because f is a smooth convex function and its gradient is Lipschitz continuous with constant L, apply to (4) Theorem 2.1.5 in [13]

$$f(x) - f(y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle x - y, \nabla f(x)\rangle$$

we get

$$\begin{aligned}\|x^{k+1}-\gamma x^k-(1-\gamma)x^*\|^2 \le \|x^k-\gamma x^{k-1}-(1-\gamma)x^*\|^2+\lambda_k^2\|\nabla f(x^k)\|^2 \\ -2(1-\gamma)\lambda_k\left(f(x^k)-f^*+\frac{1}{2L}\|\nabla f(x^k)\|^2\right) \\ -2\gamma\lambda_k\left(f(x^k)-f(x^{k-1})+\frac{1}{2L}\|\nabla f(x^k)-\nabla f(x^{k-1})\|^2\right) \\ \le \|x^k-\gamma x^{k-1}-(1-\gamma)x^*\|^2-2(1-\gamma)\lambda_k\left(f(x^k)-f^*\right) \\ -2\gamma\lambda_k\left(f(x^k)-f(x^{k-1})\right)+\left(\lambda_k^2-\frac{(1-\gamma)\lambda_k}{L}\right)\|\nabla f(x^k)\|^2 \qquad (5)\end{aligned}$$

From lemma 3.1, we note $\lambda_k < \frac{(1+\varepsilon_{k-1})\eta_0}{\mu}\,\forall k \in \mathrm{N}$. Combining with $\frac{(1+\varepsilon_{k-1})\eta_0}{1-\gamma} \le \frac{\mu}{L}\,\forall k \in \mathrm{N}$ and divide the two side of 5 in $2\lambda_k$, we could deduce that

$$\begin{aligned}\frac{1}{2\lambda_k}\|x^{k+1}-\gamma x^k-(1-\gamma)x^*\|^2 \le \frac{1}{2\lambda_k}\|x^k-\gamma x^{k-1}-(1-\gamma)x^*\|^2-(1-\gamma)\left(f(x^k)-f^*\right) \\ -\gamma\left(f(x^k)-f(x^{k-1})\right) \qquad (6)\end{aligned}$$

Summing over $k=0,...,K$ gives

$$\begin{aligned}(1-\gamma)\sum_{k=0}^{K}(f(x^k)-f^*)+\sum_{k=0}^{K}\left(\gamma(f(x^k)-f^*)+\frac{1}{2\lambda_k}\|x^{k+1}-\gamma x^k-(1-\gamma)x^*\|^2\right) \\ \le \sum_{k=0}^{K}\left(\gamma(f(x^{k-1})-f^*)+\frac{1}{2\lambda_k}\|x^k-\gamma x^{k-1}-(1-\gamma)x^*\|^2\right)\end{aligned}$$

With lemma 3.3, there exists a fixed number $\bar{k}$ such that

$$\begin{aligned}(1-\gamma)\sum_{k=0}^{K}(f(x^k)-f^*) \le \gamma(f(x^0)-f^*)+\frac{1}{2\lambda_0}\|x^0-(1-\gamma)x^*\|^2-\frac{1}{2\lambda_K}\|x^{K+1}-(1-\gamma)x^*\|^2 \\ \underbrace{+\sum_{k=1}^{\bar{k}-1}\left(\frac{1}{2\lambda_k}-\frac{1}{2\lambda_{k-1}}\right)\|x^k-\gamma x^{k-1}-(1-\gamma)x^*\|^2}_{C_1<\infty} \\ \underbrace{+\sum_{k=\bar{k}}^{K}\left(\frac{1}{2\lambda_k}-\frac{1}{2\lambda_{k-1}}\right)\|x^k-\gamma x^{k-1}-(1-\gamma)x^*\|^2}_{C_2<0}\end{aligned}$$

And because f is convex, we have

$$f(\overline{x}^K)-f^* \le \frac{1}{K+1}\left(\frac{\gamma}{1-\gamma}(f(x^0)-f^*)+\frac{1}{2\lambda_0(1-\gamma)}\|x^0-(1-\gamma)x^*\|^2+\frac{C_1}{1-\gamma}\right)$$

□

4.3 Nesterov's Accelerated Method

Theorem 4.2. *Under assumptions in 4.1, if* $(1+\varepsilon_{k-1})\eta_0 \le \frac{\mu}{L}\,\forall k \in \mathrm{N}$ *and* $0 \le \gamma < 1$*, the sequence* $\{x^k\}$ *generated by algorithm 3 satisfies*

$$f(\overline{x}^K)-f^* \le \frac{1}{K+1}\left(\frac{\gamma}{1-\gamma}f(x^0-f^*)+\frac{(1-\gamma)}{2\lambda_0}\|x^0-x^*\|^2+(1-\gamma)C_3\right)$$

Algorithm 3 Novel Gradient Descent with Nesterov's accelerated (NGD Nesterov)

Inittialization. Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0, 0 < \gamma < 1$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Choose $x^0 \in \mathbb{R}^n, \lambda_{-1} = \lambda_0$.
$v^1 = \nabla f\left(x^0\right)$
$x^1 = x^0 - \lambda_0 v^1$
for $k = 1, 2, \ldots$ **do**
    if $\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\| > \frac{\eta_0}{\lambda_{k-1}} \left\|x^k - x^{k-1}\right\|$ **then**
        $\lambda_k = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$
    else
        $\lambda_k = (1 + \varepsilon_{k-1})\, \lambda_{k-1}$
    end if
    $v^{k+1} = \gamma v^k + \lambda_k \nabla f\left(x^k\right)$
    $x^{k+1} = x^k - \left(\gamma v^{k+1} + \lambda_k \nabla f\left(x^k\right)\right)$
end for

where

$$C_3 = \sum_{k=1}^{\bar{k}-1} \left(\frac{1}{2\lambda_k} - \frac{1}{2\lambda_{k-1}} \right) \|x^k + a^k - x^*\|^2$$
$$a^{k+1} = \frac{\gamma}{1-\gamma}(x^{k+1} - x^k + \lambda_k \nabla f(x^k))$$

and $\bar{k}$ *satisfy lemma 3.3 and*

$$\overline{x}^K = \frac{1}{K+1} \sum_{k=0}^{K} x^k$$

Proof. Modify the algorithm 3, we have

$$y^{k+1} = x^k - \lambda_k \nabla f(x^k) \tag{7}$$
$$x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \tag{8}$$

Assume that $0 \leq \gamma < 1$ we have

$$a^{k+1} = \frac{\gamma}{1-\gamma}(x^{k+1} - x^k + \lambda_k \nabla f(x^k)) \tag{9}$$

From (7) and (9), we get

$$\begin{aligned} x^{k+1} + a^{k+1} &= \frac{x^{k+1}}{1-\gamma} + \frac{\gamma}{1-\gamma}(\lambda_k \nabla f(x^k) - x^k) \\ &= x^k + a^k - \frac{\lambda_k \nabla f(x^k)}{1-\gamma} \end{aligned}$$

Consider that,

$$\begin{aligned} \|x^{k+1} + a^{k+1} - x^*\|^2 &= \|x^k + a^k - x^*\|^2 + \frac{{\lambda_k}^2}{(1-\gamma)^2} \|\nabla f(x^k)\|^2 - \frac{2\lambda_k}{1-\gamma} \langle x^k + a^k - x^*, \nabla f(x^k) \rangle \\ &= \|x^k + a^k - x^*\|^2 + \frac{{\lambda_k}^2}{(1-\gamma)^2} \|\nabla f(x^k)\|^2 - \frac{2\lambda_k}{1-\gamma} \langle x^k - x^*, \nabla f(x^k) \rangle \\ &\quad - \frac{2\gamma\lambda_k}{(1-\gamma)^2} \langle x^k - x^{k-1}, \nabla f(x^k) \rangle - \frac{2\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2} \langle \nabla f(x^{k-1}), \nabla f(x^k) \rangle \end{aligned}$$

Because f is a smooth convex function and its gradient is Lipschitz continuous with constant L, with Theorem 2.1.5 in [13]

$$f(x) - f(y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle x - y, \nabla f(x)\rangle$$

We have,

$$\begin{aligned}
\|x^{k+1} + a^{k+1} - x^*\|^2 \leq{}& \|x^k + a^k - x^*\|^2 - \frac{2\lambda_k}{1-\gamma}(f(x^k) - f(x^*)) - \frac{\lambda_k}{(1-\gamma)L}\|\nabla f(x^k)\|^2 \\
&- \frac{2\gamma\lambda_k}{(1-\gamma)^2}(f(x^k) - f(x^{k-1})) - \frac{\gamma\lambda_k}{L(1-\gamma)^2}\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2 \\
&- \frac{2\gamma\lambda_k\lambda^{k-1}}{(1-\gamma)^2}\langle\nabla f(x^{k-1}), \nabla f(x^k)\rangle + \frac{\lambda_k^2\|\nabla f(x^k)\|^2}{(1-\gamma)^2}\|x^{k+1} + a^{k+1} - x^*\|^2 \\
\leq{}& \|x^k + a^k - x^*\|^2 - \frac{2\lambda_k}{1-\gamma}(f(x^k) - f(x^*)) - \frac{2\gamma\lambda_k}{(1-\gamma)^2}(f(x^k) - f(x^{k-1})) \\
&\underbrace{- \frac{\gamma\lambda_k}{L(1-\gamma)^2}\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2 + \frac{\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2}\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2}_{A} \\
&\underbrace{- \left(\frac{\lambda_k}{(1-\gamma)L} + \frac{\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2} - \frac{\lambda_k^2}{(1-\gamma)^2}\right)\|\nabla f(x^k)\|^2}_{B} - \frac{\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2}\|\nabla f(x^{k-1})\|^2
\end{aligned} \tag{10}$$

From lemma 3.1, we note $\lambda_k < \frac{(1+\varepsilon_{k-1})\eta_0}{\mu}\,\forall k \in \mathrm{N}$. Combining with $(1+\varepsilon_{k-1})\eta_0 \leq \frac{\mu}{L}\,\forall k \in \mathrm{N}$ and divide the two side of 10 in $2\lambda_k$, we could deduce that

$$\frac{1}{2\lambda_k}\|x^{k+1} + a^{k+1} - x^*\|^2 \leq \frac{1}{2\lambda_k}\|x^k + a^k - x^*\|^2 - \frac{f(x^k) - f(x^*)}{(1-\gamma)} - \frac{\gamma(f(x^k) - f(x^{k-1}))}{(1-\gamma)^2}$$

Summing over $k = 0, ..., K$ gives

$$\begin{aligned}
\frac{1}{1-\gamma}\sum_{k=0}^{K}(f(x^k) - f^*) + \sum_{k=0}^{K}\left(\frac{\gamma}{(1-\gamma)^2}(f(x^k) - f^*) + \frac{1}{2\lambda_k}\|x^{k+1} + a^{k+1} - x^*\|^2\right) \\
\leq \sum_{k=0}^{K}\left(\frac{\gamma}{(1-\gamma)^2}(f(x^{k-1}) - f^*) + \frac{1}{2\lambda_k}\|x^k + a^k - x^*\|^2\right)
\end{aligned}$$

With lemma 3.3, there exists a fixed number $\bar{k}$ such that

$$\begin{aligned}
\frac{1}{1-\gamma}\sum_{k=0}^{K}(f(x^k) - f^*) \leq{}& \frac{\gamma}{(1-\gamma)^2}(f(x^0) - f^*) + \frac{1}{2\lambda_0}\|x^0 - x^*\|^2 - \frac{1}{2\lambda_K}\|x^{K+1} + a^{K+1} - x^*\|^2 \\
&\underbrace{+ \sum_{k=1}^{\bar{k}-1}\left(\frac{1}{2\lambda_k} - \frac{1}{2\lambda_{k-1}}\right)\|x^k + a^k - x^*\|^2}_{C_3<\infty} \\
&\underbrace{+ \sum_{k=\bar{k}}^{K}\left(\frac{1}{2\lambda_k} - \frac{1}{2\lambda_{k-1}}\right)\|x^k + a^k - x^*\|^2}_{C_4<0}
\end{aligned}$$

And because f is convex, we have

$$f(\overline{x}^K) - f^* \leq \frac{1}{K+1}\left(\frac{\gamma}{1-\gamma}f(x^0 - f^*) + \frac{(1-\gamma)}{2\lambda_0}\|x^0 - x^*\|^2 + (1-\gamma)C_3\right)$$

□

5 Stochastic Algorithms

Algorithm 4 Novel step size for Stochastic version

1: **Inittialization.** Select $\lambda_0 > 0, 0 < \eta_1 < \eta_0$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Choose $\lambda_{-1} = \lambda_0$.
2: **for** $k = 1, 2, \ldots$ **do**
3: **if** $\left\|\nabla f_\xi(x^k) - \nabla f_\xi(x^{k-1})\right\| > \frac{\eta_0}{\lambda_{k-1}} \left\|x^k - x^{k-1}\right\|$ **then**
4: $\lambda_k = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{\left\|\nabla f_\xi(x^k) - \nabla f_\xi(x^{k-1})\right\|}$
5: **else**
6: $\lambda_k = (1 + \varepsilon_{k-1})\, \lambda_{k-1}$
7: **end if**
8: **end for**

Consider the problem

$$\min\{f(x) : x \in \mathbb{R}^n\} \tag{11}$$

where $f(x) \overset{\text{def}}{=} \mathbb{E}\left[f_\xi(x)\right]$ and $f_\xi : \mathbb{R}^n \to \mathbb{R}$ and assume $f(x)$ has optimal value f^*. From lemmas in section 3, it can be easily deduced the bellowing lemmas for stochastic algorithms. The proof is completely similar.

Lemma 5.1. *Let $\{\lambda_k\}$ be the sequence generated by Algorithm 4 where f_ξ is smooth and its gradient is L-Lipschitz continuous, then $\lambda_k \geq \min(\lambda_0, \frac{\eta_1}{L})\ \forall k \geq 0$.*

Lemma 5.2. *Let $\{\lambda_k\}$ be the sequence generated by Algorithm (4) where f_ξ is smooth and its gradient is L-Lipschitz continuous, then $\{\lambda_k\}$ converges to $\overline{\lambda} < \infty$.*

Besides deterministic methods, stochastic methods are also proposed with the main purpose of applying to deep learning models.

5.1 Stochastic Method

Assumption 5.1. *We assume f_ξ is smooth, i.e its gradient is L-Lipschitz continuous:*

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

This assumption also implies that f is smooth

Assumption 5.2. *Each f_ξ, is μ-strongly convex. This assumption implies that f is also μ-strongly convex.*

Lemma 5.3. *Suppose Assumption (5.1), (5.2) hold, λ_k is the stepsize sequence and η_0, η_1 is defined in Algorithm (5). Then exists k_1 such that*

$$\min\{\frac{\eta_1}{L}, \lambda_0\} = \gamma \leq \lambda_k \leq \frac{\eta_0^2}{\eta_1 \mu}, \qquad \forall k \geq k_1 \tag{12}$$

Proof. Indeed, consider the two possible cases:

- Case 1: If $L_k = \left\|\nabla f_{\zeta^k}\left(x^k\right) - \nabla f_{\zeta^k}\left(x^{k-1}\right)\right\| > \frac{\eta_0}{\lambda_{k-1}} \left\|x^k - x^{k-1}\right\|$ then

$$\lambda_k = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{L_k} = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{\left\|\nabla f_{\zeta^k}\left(x^k\right) - \nabla f_{\zeta^k}\left(x^{k-1}\right)\right\|} \leq \frac{\eta_1}{\mu} \leq \frac{\eta_0}{\mu} \leq \frac{\eta_0^2}{\eta_1 \mu}.$$

(Because f is μ-strongly convex $\|x - y\| \leq \frac{1}{\mu} \left\|\nabla f_{\zeta^k}(x) - \nabla f_{\zeta^k}(y)\right\|, \quad \forall x, y.$)

Algorithm 5 Novel stochastic gradient descent (SNGD)

Inittialization. Select $\lambda_0 > 0, 0 < \eta_1 < \eta_0$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Choose $x^0 \in \mathbb{R}^n, \lambda_{-1} = \lambda_0$, ξ^0.
$x^1 = x^0 - \lambda_0 \nabla f_{\xi^0}\left(x^0\right)$
for $k = 1, 2, \ldots$ **do**
  Sample ξ^k and optionally ζ^k
  Option I: $L_k = \left\|\nabla f_{\xi^k}\left(x^k\right) - \nabla f_{\xi^k}\left(x^{k-1}\right)\right\|$
  Option II: $L_k = \left\|\nabla f_{\zeta^k}\left(x^k\right) - \nabla f_{\zeta^k}\left(x^{k-1}\right)\right\|$
  if $L_k > \frac{\eta_0}{\lambda_{k-1}}\left\|x^k - x^{k-1}\right\|$ **then**
    $\lambda_k = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{L_k}$
  else
    $\lambda_k = (1 + \varepsilon_{k-1})\,\lambda_{k-1}$
  end if
  $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}\left(x^k\right)$
end for

- Otherwise, we have $\left\|\nabla f_{\zeta^k}\left(x^k\right) - \nabla f_{\zeta^k}\left(x^{k-1}\right)\right\| \leq \frac{\eta_0}{\lambda_{k-1}}\left\|x^k - x^{k-1}\right\|$ meaning that $\lambda_{k-1} \leq \frac{\eta_0}{\mu}$. Therefore,

$$\lambda_k = (1 + \varepsilon_{k-1})\lambda_{k-1} \leq (1 + \varepsilon_{k-1})\frac{\eta_0}{\mu}.$$

$\lim_{n\to\infty} \varepsilon_k = 0$ so exists k_1 satisfy $\varepsilon_{k-1} \leq \frac{\eta_0}{\eta_1} - 1, \quad \forall k \geq k_1$. So $\lambda_k \leq \frac{\eta_0^2}{\eta_1 \mu}, \quad \forall k \geq k_1$

Because $f(x)$ is L-smooth so $\left\|\nabla f_{\zeta^k}(x) - \nabla f_{\zeta^k}(y)\right\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$.
For $k = 1$ if $L_1 > \frac{\eta_0}{\lambda_0}\left\|x^1 - x^0\right\|$ then $\lambda_1 = \eta_1 \frac{\left\|x^1 - x^0\right\|}{L_1} \geq \frac{\eta_1}{L}$, otherwise $\lambda_1 = (1 + \varepsilon_0)\lambda_0 \geq \lambda_0$. By induction, we get that $\lambda_k \geq \min\{\frac{\eta_1}{L}, \lambda_0\} = \gamma \quad \forall k \geq 0$. □

Proposition 5.1. *([14] Lemma 1) Denote* $\sigma^2 \overset{def}{=} \mathbb{E}\left[\left\|\nabla f_\xi\left(x^*\right)\right\|^2\right]$ *and assume* f *to be* L*-smooth and convex. Then it holds for any* x

$$\mathbb{E}\left[\|\nabla f_\xi(x)\|^2\right] \leq 4L\left(f(x) - f_*\right) + 2\sigma^2. \tag{13}$$

Another fact that we will use is a strong convexity bound, which states for any x, y

$$\langle \nabla f(x), x - y\rangle \geq \frac{\mu}{2}\|x - y\|^2 + f(x) - f(y). \tag{14}$$

Theorem 5.1. *Suppose Assumption (5.1), (5.2) hold,* $\{\lambda_k\}$ *is the stepsize sequence and* η_0, η_1 *is defined in Algorithm (5). If we choose some* $\frac{\eta_0^2}{\eta_1} \leq \frac{\mu}{2L}$*, then exists* k_1 *such that*

$$\mathbb{E}\left[\left\|x^{k+1} - x^*\right\|^2\right] \leq \exp\left(-\mu(k - k_1 + 1)\gamma\right) C_0 + 2\frac{\eta_0^2 \sigma^2}{\eta_1 \mu^2}$$

where $C_0 \overset{def}{=} \mathbb{E}\left[\left\|x^{k_1} - x^*\right\|^2\right]$ *and* $\sigma^2 \overset{def}{=} \mathbb{E}\left[\left\|\nabla f_\xi\left(x^*\right)\right\|^2\right]$.

Proof. Under assumptions on $\frac{\eta_0^2}{\eta_1} \leq \frac{\mu}{2L}$, we have $\lambda_k \leq \frac{\eta_0^2}{\eta_1 \mu} \leq \frac{1}{2L}$. Since λ_k is independent of ξ^k, we

have $\mathbb{E}\left[\lambda_k \nabla f_{\xi^k}\left(x^k\right)\right] = \mathbb{E}\left[\lambda_k\right]\mathbb{E}\left[\nabla f\left(x^k\right)\right]$ and

$$\begin{aligned}
\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] &= \mathbb{E}\left[\left\|x^k-x^*\right\|^2\right] - 2\mathbb{E}\left[\lambda_k\left\langle\nabla f\left(x^k\right), x^k-x^*\right\rangle\right] + \mathbb{E}\left[\lambda_k^2\right]\mathbb{E}\left[\left\|\nabla f_{\xi^k}\left(x^k\right)\right\|^2\right] \\
&\overset{(14)}{\leq} \mathbb{E}\left[\left(1-\lambda_k\mu\right)\left\|x^k-x^*\right\|^2\right] - 2\mathbb{E}\left[\lambda_k\left(f\left(x^k\right)-f_*\right)\right] + \mathbb{E}\left[\lambda_k^2\right]\mathbb{E}\left[\left\|\nabla f_{\xi^k}\left(x^k\right)\right\|^2\right] \\
&\overset{(13)}{\leq} \mathbb{E}\left[\left(1-\lambda_k\mu\right)\left\|x^k-x^*\right\|^2\right] - 2\mathbb{E}[\lambda_k \underbrace{\left(1-2\lambda_k L\right)}_{\geq 0}\left(f\left(x^k\right)-f_*\right)] + 2\mathbb{E}\left[\lambda_k^2\right]\sigma^2 \\
&\overset{(12)}{\leq} \mathbb{E}\left[1-\lambda_k\mu\right]\mathbb{E}\left[\left\|x^k-x^*\right\|^2\right] + 2\eta_0^2\frac{\mathbb{E}\left[\lambda_k\right]\sigma^2}{\eta_1\mu}.
\end{aligned}$$

Therefore, if we subtract $2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}$ from both sides, we obtain

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2 - 2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}\right] \leq \mathbb{E}\left[1-\lambda_k\mu\right]\mathbb{E}\left[\left\|x^k-x^*\right\|^2 - 2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}\right].$$

If $\mathbb{E}\left[\left\|x^k-x^*\right\|^2\right] \leq 2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}$ for some k, it follows that $\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right] \leq 2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}$ for any $t \geq k$. Otherwise, we can reuse the produced bound to obtain

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] \leq \left(\prod_{t=k_1}^{k}\mathbb{E}\left[1-\lambda_t\mu\right]\right)\mathbb{E}\left[\left\|x^{k_1}-x^*\right\|^2\right] + 2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}.$$

By inequality $1-x \leq e^{-x}$, we have $\prod_{t=k_1}^{k}\mathbb{E}\left[1-\lambda_t\mu\right] \leq \exp\left(-\mu\sum_{t=k_1}^{k}\lambda_t\right)$. In addition, recall that in accordance with (12) we have $\lambda_k \geq \gamma$. Thus,

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] \leq \exp\left[-\mu(k-k_1+1)\gamma\right]\mathbb{E}\left[\left\|x^{k_1}-x^*\right\|^2\right] + 2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}.$$

□

So if we can choose $\{\varepsilon_k\}$ such that $k_1 = 1$ then

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] \leq \exp\left[-\mu k\gamma\right]\mathbb{E}\left[\left\|x^1-x^*\right\|^2\right] + 2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}.$$

We also have

$$\mathbb{E}\left[\left\|x^1-x^*\right\|^2\right] \leq 2\left\|x^0-x^*\right\| + 2\lambda_0^2\mathbb{E}\left[\left\|\nabla f_{\xi^0}\left(x^0\right)\right\|^2\right] \leq 2\left\|x^0-x^*\right\| + 2\lambda_0^2\left(2L^2\left\|x^0-x^*\right\|^2 + 2\sigma^2\right) = M.$$

So

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] \leq \exp\left(-\mu k\gamma\right)M + 2\frac{\eta_0^2\sigma^2}{\eta_1\mu^2}.$$

In this case we can choose ε_k such that $\varepsilon_0 \leq \frac{\eta_0}{\eta_1} - 1$.

5.2 Stochastic Momentum Method

Before given the prove of convergence for the algorithm 6, we need some assumptions as follows.

5.2.1 Assumptions

Assumption 5.3. *We assume that f is lower bounded by f^**

$$\forall x \in \mathbb{R}^d, f(x) \geq f^*.$$

Assumption 5.4. *We assume l_2-norm of the stochastic gradients of f are bounded, i.e, there exist σ so that*

$$\forall x \in \mathbb{R}^d, \mathbb{E}\left[\|\nabla f_\xi(x)\|^2\right] \leq \sigma^2.$$

Assumption 5.5. *We assume f_ξ is smooth, i.e its gradient is L-Lipschitz continuous:*

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

This assumption also implies that f is smooth

Algorithm 6 Stochastic novel gradient descent momentum (SNGDm)

Inittialization. Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0, 0 < \gamma < 1$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Choose $x^0 \in \mathbb{R}^n, \lambda_{-1} = \lambda_0$, ξ^0.
$v^1 = \nabla f_{\xi^0}\left(x^0\right)$
$x^1 = x^0 - \lambda_0 v^1$
for $k = 1, 2, \ldots$ **do**
  if $\left\|\nabla f_{\xi^k}\left(x^k\right) - \nabla f_{\xi^k}\left(x^{k-1}\right)\right\| > \frac{\eta_0}{\lambda_{k-1}}\left\|x^k - x^{k-1}\right\|$ **then**
    $\lambda_k = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{\left\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^{k-1})\right\|}$
  else
    $\lambda_k = (1 + \varepsilon_{k-1})\,\lambda_{k-1}$
  end if
  $v^{k+1} = \gamma v^k + \nabla f_{\xi^k}\left(x^k\right)$
  $x^{k+1} = x^k - \lambda_k v^{k+1}$
end for

5.2.2 Proof of Convergent

For all $n \in \mathbb{N}^*$, we note $G^k = \nabla f(x^{k-1})$ and $g^k = \nabla f_\xi(x^{k-1})$.

Lemma 5.4. *[3] Given $0 \le \gamma < 1$, $\{x^k\}$ and $\{v^k\}$ defined by Algorithm (6) and with assumption in (5.3), we have*

$$\mathbb{E}\left[\|v^k\|^2\right] \le \frac{\sigma^2}{(1-\gamma)^2} \quad \forall k \in \mathbb{N}^*.$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\|v^k\|^2\right] &= \mathbb{E}\left[\|\sum_{i=0}^{k-1}\gamma^i g^{k-i}\|^2\right] \\ &= \mathbb{E}\left[\langle\sum_{i=0}^{k-1}\gamma^i g^{k-i}, \sum_{j=0}^{k-1}\gamma^j g^{k-j}\rangle\right].\end{aligned}$$

Using Cauchy-Schwarz,

$$\begin{aligned}\mathbb{E}\left[\|v^k\|^2\right] &\le \sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\gamma^i\gamma^j\left(\frac{\mathbb{E}\left[\|g^{k-i}\|^2\right]}{2} + \frac{\mathbb{E}\left[\|g^{k-j}\|^2\right]}{2}\right) \\ &\le \sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\gamma^i\gamma^j\sigma^2 \\ &\le \frac{\sigma^2}{(1-\gamma)^2}.\end{aligned}$$

□

Lemma 5.5. *(Sum of a geometric term times index [3]). Given $0 < a < 1, i \in \mathbb{N}$ and $Q \in \mathbb{N}$ with $Q \ge i$,*

$$\sum_{q=i}^{Q} a^q q = \frac{a^i}{1-a}\left(i - a^{Q-i+1}Q + \frac{a - a^{Q+1-i}}{1-a}\right) \le \frac{a}{(1-a)^2}.$$

Lemma 5.6. *(Descent lemma [3]) Given $0 \le \gamma < 1$, $\{x^k\}, \{v^k\}, \{\lambda_k\}$ defined by Algorithm (6) and $\lambda_k \le \bar{\lambda} \quad \forall k \in \mathbb{N}$, we have*

$$\mathbb{E}\left[\nabla f(x^{k-1})^T v^k\right] \ge \sum_{i=0}^{k-1}\gamma^i \mathbb{E}\left[\|\nabla f(x^{k-i-1})\|^2\right] - \frac{\bar{\lambda}L\gamma\sigma^2}{(1-\gamma)^3}.$$

Proof. We note $G^k = \nabla f(x^{k-1})$ is the expected gradient and $g^k = \nabla f_\xi(x^{k-1})$ is the stochastic gradient at iteration k. Consider

$$\begin{aligned} {G^k}^T v^k &= \sum_{i=0}^{k-1} \gamma^i {G^k}^T g^{k-i} \\ &= \sum_{i=0}^{k-1} \gamma^i {G^{k-i}}^T g^{k-i} + \sum_{i=0}^{k-1} \gamma^i (G^k - G^{k-i})^T g^{k-i}. \end{aligned} \tag{15}$$

We apply

$$\forall r > 0,\, x,\, y \in \mathbb{R},\, ||xy|| \leq \frac{r}{2}||x||^2 + \frac{||y||^2}{2r}$$

with $x = G^k - G^{k-i}$, $y = g^{k-i}$ to (15), we have

$${G^k}^T v^k \geq \sum_{i=0}^{k-1} \gamma^i {G^{k-i}}^T g^{k-i} - \sum_{i=0}^{k-1} \frac{\gamma^i}{2} \left(r||G^k - G^{k-i}||^2 + \frac{||g^{k-i}||}{r} \right). \tag{16}$$

Because f is L-smooth, so we have

$$\begin{aligned} ||G^k - G^{k-i}||^2 &\leq L^2 ||x^k - x^{k-i}||^2 \\ &\leq L^2 || \sum_{l=1}^{i} x^{k-l+1} - x^{k-l} ||^2 \\ &\leq L^2 || \sum_{l=1}^{i} \lambda_{k-l-1} v^{k-l} ||^2. \end{aligned}$$

Note that $\{\lambda_k\}$ is convergent (lemma 3.2), let $\overline{\lambda} \geq \lambda_k \, \forall k \in \mathbb{N}$, we have

$$\begin{aligned} ||G^k - G^{k-i}||^2 &\leq L^2 \left|\left| \sum_{l=1}^{i} \overline{\lambda} v^{k-l} \right|\right|^2 \\ &\leq \overline{\lambda}^2 L^2 i \sum_{l=1}^{i} ||v^{k-l}||^2. \end{aligned} \tag{17}$$

From (15) and (17), we get

$${G^k}^T v^k \geq \sum_{i=0}^{k-1} \gamma^i {G^{k-i}}^T g^{k-i} - \sum_{i=0}^{k-1} \frac{\gamma^i}{2} \left(r(\overline{\lambda} L)^2 i \sum_{l=1}^{i} ||v^{k-l}||^2 + \frac{||g^{k-i}||^2}{r} \right).$$

Taking expectation of two side, we have

$$\mathbb{E}\left[{G^k}^T v^k\right] \geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[{G^{k-i}}^T g^{k-i}\right] - \sum_{i=0}^{k-1} \frac{\gamma^i}{2} \left(r(\overline{\lambda} L)^2 i \sum_{l=1}^{i} \mathbb{E}\left[||v^{k-l}||^2\right] + \frac{\mathbb{E}\left[||g^{k-i}||^2\right]}{r} \right). \tag{18}$$

Beside that

$$\begin{aligned} \mathbb{E}\left[{G^{k-i}}^T g^{k-i}\right] &= \mathbb{E}\left[\mathbb{E}_{n-k-1}\left[\nabla F(x^{n-k-1})^T \nabla f(x^{n-k-1})\right]\right] \\ &= \mathbb{E}\left[\nabla F(x^{n-k-1})^T \nabla F(x^{n-k-1})\right] \\ &= \mathbb{E}\left[||G^{k-i}||^2\right]. \end{aligned} \tag{19}$$

and from Assumption 5.4 we have

$$\mathbb{E}\left[||g^{k-i}||^2\right] \leq \sigma^2 \tag{20}$$

and from Lemma 5.4

$$\mathbb{E}\left[||v^{k-l}||^2\right] \leq \frac{\sigma^2}{(1-\gamma)^2} \tag{21}$$

Injecting (19), (20), (21) to (18), we get

$$\mathbb{E}\left[{G^k}^T v^k\right] \geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2\right] - \sum_{i=0}^{k-i} \frac{\gamma^i}{2}\left(r\frac{(\overline{\lambda}Li)^2\sigma^2}{(1-\gamma)^2} + \frac{\sigma^2}{r}\right) \tag{22}$$

Replace $r = \frac{1-\gamma}{\overline{\lambda}Li}$ to (22), we obtain

$$\begin{aligned}\mathbb{E}\left[{G^k}^T v^k\right] &\geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2\right] - \sum_{i=0}^{k-1} \frac{\gamma^i}{2}\left(\frac{2\overline{\lambda}Li\sigma^2}{1-\gamma}\right) \\ &\geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2\right] - \frac{\overline{\lambda}L}{1-\gamma}\sigma^2 \sum_{i=0}^{k-1} \gamma^i i\end{aligned}$$

Using Lemma 5.5, deduce

$$\mathbb{E}\left[{G^k}^T v^k\right] \geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2\right] - \frac{\overline{\lambda}L\gamma\sigma^2}{(1-\gamma)^3}$$

□

Before introducing convergent theorem of NSGDm, we remind about iteration standard for non convex stochastic optimization ([6]). Our results bound the expected squared norm of the gradient at iteration τ, which is a random index with value in $\{0, 1..., N-1\}$, so that for a number of iterations $N \in \mathbb{N}^*$

$$\forall i \in \mathbb{N},\, j < N, \mathbb{P}\left[\tau = j\right] \propto 1 - \gamma^{N-j}. \tag{23}$$

Theorem 5.2. *(Convergent of SNGDm [3]) Given assumptions from section 5.2.1, τ defined above, for a iteration $N > \frac{1}{1-\gamma}$, $x^0 \in \mathbb{R}^d$, $\lambda_k \leq \overline{\lambda}\, \forall k \in \mathbb{N}$, $0 \geq \gamma < 1$, $\{x^k\}$ defined as by Algorithm (6) then*

$$\mathbb{E}\left[||\nabla f(x^\tau)||^2\right] \leq \frac{1-\gamma}{\overline{\lambda}\tilde{N}}(f(x^0) - f^*) + \frac{N}{\tilde{N}}\frac{\overline{\lambda}L\sigma^2(1+\gamma)}{2(1-\gamma)^2}$$

where $\tilde{N} = N - \frac{\gamma}{1-\gamma}$.

Proof. Because f is smooth and take a specifie iteration $k \in \mathbb{N}^*$, we have

$$f(x^k) \leq f(x^{k-1}) - \lambda_{k-1}{G^k}^T v^k + \frac{{\lambda_{k-1}}^2 L||v^k||^2}{2}$$

Let $\overline{\lambda} \geq \lambda_k$, $\forall k \in \mathbb{N}$, with Lemma 5.4 and Lemma 5.6, take expectation of two side, we get

$$\begin{aligned}\mathbb{E}\left[f(x^k)\right] &\leq \mathbb{E}\left[f(x^{k-1})\right] - \overline{\lambda}\left(\sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2\right]\right) + \frac{\overline{\lambda}^2 L\gamma\sigma^2}{(1-\gamma)^3} + \frac{\overline{\lambda}L\sigma^2}{2(1-\gamma)^2} \\ &\leq \mathbb{E}\left[f(x^{k-1})\right] - \overline{\lambda}\left(\sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2\right]\right) + \frac{\overline{\lambda}L(1+\gamma)\sigma^2}{2(1-\gamma)^3}\end{aligned}$$

rearranging and summing over $k \in \{1...N\}$ we get

$$\overline{\lambda}\sum_{k=1}^{N}\sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2\right] \leq f(x^0) - \mathbb{E}\left[f(x^N)\right] + N\frac{\overline{\lambda}^2 L\sigma^2}{2(1-\gamma)^3}. \tag{24}$$

Zoom on the right hand side, replace $j = k - i$, we get

$$\begin{aligned} RHS &= \overline{\lambda}\sum_{k=1}^{N}\sum_{j=1}^{k}\gamma^{k-j}\mathbb{E}\left[||G^j||^2\right] \\ &= \overline{\lambda}\sum_{j=1}^{N}\mathbb{E}\left[||G^j||^2\right]\sum_{k=j}^{N}\gamma^{k-j} \\ &= \frac{\overline{\lambda}}{1-\gamma}\sum_{j=1}^{N}\mathbb{E}\left[||\nabla f(x^{j-1})||^2\right](1-\gamma^{N-j+1}) \\ &= \frac{\overline{\lambda}}{1-\gamma}\sum_{j=0}^{N-1}\mathbb{E}\left[||\nabla f(x^{j})||^2\right](1-\gamma^{N-j}). \end{aligned}$$

As the definition of τ in (23)

$$\sum_{j=0}^{N-1}(1-\gamma^{N-j}) = N - \gamma\frac{1-\gamma^N}{1-\gamma} \geq N - \frac{\gamma}{1-\gamma}$$

and let $\tilde{N} = N - \frac{\gamma}{1-\gamma}$, we obtain

$$RHS \geq \frac{\overline{\lambda}\tilde{N}}{1-\gamma}\mathbb{E}\left[||\nabla f(x^{\tau})||^2\right] \tag{25}$$

From (24) and (25), with Assumption 5.3 that f is lower bounded by f^*, we obtain

$$\mathbb{E}\left[||\nabla f(x^{\tau})||^2\right] \leq \frac{1-\gamma}{\overline{\lambda}\tilde{N}}(f(x^0) - f^*) + \frac{N}{\tilde{N}}\frac{\overline{\lambda}L\sigma^2(1+\gamma)}{2(1-\gamma)^2}.$$

□

5.3 Stochastic Nesterov's Accelerated Method

Algorithm 7 Stochastic Novel Nesterov's Accelerated Gradient Descent (SNAGD)

Inittialization. Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0, 0 < \gamma < 1$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{\infty}\varepsilon_k < \infty$. Choose $x^0 \in \mathbb{R}^n, \lambda_{-1} = \lambda_0$, ξ^0.
$v^1 = \nabla f_{\xi^0}\left(x^0\right)$
$x^1 = x^0 - \lambda_0 v^1$
for $k = 1, 2, \ldots$ **do**
  if $\left\|\nabla f_{\xi^k}\left(x^k\right) - \nabla f_{\xi^k}\left(x^{k-1}\right)\right\| > \frac{\eta_0}{\lambda_{k-1}}\left\|x^k - x^{k-1}\right\|$ **then**
    $\lambda_k = \eta_1 \frac{\left\|x^k - x^{k-1}\right\|}{\left\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^{k-1})\right\|}$
  else
    $\lambda_k = (1 + \varepsilon_{k-1})\,\lambda_{k-1}$
  end if
  $v^{k+1} = \gamma v^k + \nabla f_{\xi^k}\left(x^k\right)$
  $x^{k+1} = x^k - \lambda_k\left(\gamma v^{k+1} + \nabla f_{\xi^k}\left(x^k\right)\right)$
end for

5.3.1 Assumptions

Assumption 5.6. *We assume that f is lower bounded by f^**

$$\forall x \in \mathbb{R}^d, f(x) \geq f^*.$$

Assumption 5.7. *We assume the stochastic gradients have bounded variance and the gradients of f are uniformly bounded, i.e, there exist σ and δ so that*

$$\forall x \in \mathbb{R}^d, \|\nabla f(x)\|^2 \leq \delta^2, \quad \mathbb{E}\left[\|\nabla f_\xi(x)\|^2\right] - \|\nabla f(x)\|^2 \leq \sigma^2$$

Assumption 5.8. *We assume f_ξ is smooth, i.e its gradient is L-Lipschitz continuous:*

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

This assumption also implies that f is smooth

5.3.2 Proof of Convergent

Lemma 5.7. *[3] Given $0 \leq \gamma < 1$, $\{x^k\}$ and $\{v^k\}$ defined by Algorithm (7) and with assumption in (5.7), we have*

$$\mathbb{E}\left[\|v^k\|^2\right] \leq \frac{\sigma^2 + \delta^2}{(1-\gamma)^2} \quad \forall k \in \mathbb{N}^*.$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\|v^k\|^2\right] &= \mathbb{E}\left[\|\sum_{i=0}^{k-1} \gamma^i g^{k-i}\|^2\right] \\ &= \mathbb{E}\left[\langle \sum_{i=0}^{k-1} \gamma^i g^{k-i}, \sum_{j=0}^{k-1} \gamma^j g^{k-j} \rangle\right].\end{aligned}$$

Using Cauchy-Schwarz,

$$\begin{aligned}\mathbb{E}\left[\|v^k\|^2\right] &\leq \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \gamma^i\gamma^j \left(\frac{\mathbb{E}\left[\|g^{k-i}\|^2\right]}{2} + \frac{\mathbb{E}\left[\|g^{k-j}\|^2\right]}{2}\right) \\ &\leq \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \gamma^i\gamma^j(\sigma^2 + \delta^2) \\ &\leq \frac{(\sigma^2 + \delta^2)}{(1-\gamma)^2}.\end{aligned}$$

□

Lemma 5.8. *(Descent lemma [3]) Given $0 \leq \gamma < 1$, $\{x^k\}, \{v^k\}, \{\lambda_k\}$ defined by Algorithm (7) and $\lambda_k \leq \bar{\lambda} \quad \forall k \in \mathbb{N}$, we have*

$$\mathbb{E}\left[\nabla f(x^{k-1})^T v^k\right] \geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[\|\nabla f(x^{k-i-1})\|^2\right] - \frac{\bar{\lambda} L \gamma(\sigma^2 + \delta^2)}{(1-\gamma)^3}.$$

Proof. We note $G^k = \nabla f(x^{k-1})$ is the expected gradient and $g^k = \nabla f_\xi(x^{k-1})$ is the stochastic gradient at iteration k. Consider

$$\begin{aligned}{G^k}^T v^k &= \sum_{i=0}^{k-1} \gamma^i {G^k}^T g^{k-i} \\ &= \sum_{i=0}^{k-1} \gamma^i {G^{k-i}}^T g^{k-i} + \sum_{i=0}^{k-1} \gamma^i (G^k - G^{k-i})^T g^{k-i}. \qquad (26)\end{aligned}$$

We apply

$$\forall r > 0,\ x,\ y \in \mathbb{R},\ ||xy|| \leq \frac{r}{2}||x||^2 + \frac{||y||^2}{2r}$$

with $x = G^k - G^{k-i}$, $y = g^{k-i}$ to (26), we have

$$G^{k^T} v^k \geq \sum_{i=0}^{k-1} \gamma^i G^{k-i^T} g^{k-i} - \sum_{i=0}^{k-1} \frac{\gamma^i}{2} \left(r||G^k - G^{k-i}||^2 + \frac{||g^{k-i}||}{r} \right).$$

Because f is L-smooth, so we have

$$\begin{aligned} ||G^k - G^{k-i}||^2 &\leq L^2 ||x^k - x^{k-i}||^2 \\ &\leq L^2 || \sum_{l=1}^{i} x^{k-l+1} - x^{k-l}||^2 \\ &\leq L^2 || \sum_{l=1}^{i} \lambda_{k-l-1} v^{k-l}||^2. \end{aligned}$$

Note that $\{\lambda_k\}$ is convergent (lemma 3.2), let $\overline{\lambda} \geq \lambda_k \; \forall k \in \mathbb{N}$, we have

$$\begin{aligned} ||G^k - G^{k-i}||^2 &\leq L^2 \left|\left| \sum_{l=1}^{i} \overline{\lambda} v^{k-l} \right|\right|^2 \\ &\leq \overline{\lambda}^2 L^2 i \sum_{l=1}^{i} ||v^{k-l}||^2. \end{aligned} \tag{27}$$

From (26) and (27), we get

$$G^{k^T} v^k \geq \sum_{i=0}^{k-1} \gamma^i G^{k-i^T} g^{k-i} - \sum_{i=0}^{k-1} \frac{\gamma^i}{2} \left(r(\overline{\lambda} L)^2 i \sum_{l=1}^{i} ||v^{k-l}||^2 + \frac{||g^{k-i}||^2}{r} \right).$$

Taking expectation of two side, we have

$$\mathbb{E}\left[G^{k^T} v^k \right] \geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[G^{k-i^T} g^{k-i} \right] - \sum_{i=0}^{k-1} \frac{\gamma^i}{2} \left(r(\overline{\lambda} L)^2 i \sum_{l=1}^{i} \mathbb{E}\left[||v^{k-l}||^2 \right] + \frac{\mathbb{E}\left[||g^{k-i}||^2 \right]}{r} \right). \tag{28}$$

Beside that

$$\begin{aligned} \mathbb{E}\left[G^{k-i^T} g^{k-i} \right] &= \mathbb{E}\left[\mathbb{E}_{n-k-1}\left[\nabla F(x^{n-k-1})^T \nabla f(x^{n-k-1}) \right] \right] \\ &= \mathbb{E}\left[\nabla F(x^{n-k-1})^T \nabla F(x^{n-k-1}) \right] \\ &= \mathbb{E}\left[||G^{k-i}||^2 \right]. \end{aligned} \tag{29}$$

and from Assumption 5.7 we have

$$\mathbb{E}\left[||g^{k-i}||^2 \right] \leq \sigma^2 + \delta^2 \tag{30}$$

and from Lemma 5.7

$$\mathbb{E}\left[||v^{k-l}||^2 \right] \leq \frac{\sigma^2 + \delta^2}{(1-\gamma)^2} \tag{31}$$

Injecting (29), (30), (31) to (28), we get

$$\mathbb{E}\left[G^{k^T} v^k \right] \geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2 \right] - \sum_{i=0}^{k-i} \frac{\gamma^i}{2} \left(r \frac{(\overline{\lambda} L i)^2 (\sigma^2 + \delta^2)}{(1-\gamma)^2} + \frac{\sigma^2 + \delta^2}{r} \right) \tag{32}$$

Replace $r = \frac{1-\gamma}{\overline{\lambda} L i}$ to (32), we obtain

$$\begin{aligned} \mathbb{E}\left[G^{k^T} v^k \right] &\geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2 \right] - \sum_{i=0}^{k-1} \frac{\gamma^i}{2} \left(\frac{2\overline{\lambda} L i (\sigma^2 + \delta^2)}{1-\gamma} \right) \\ &\geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2 \right] - \frac{\overline{\lambda} L}{1-\gamma} (\sigma^2 + \delta^2) \sum_{i=0}^{k-1} \gamma^i i \end{aligned}$$

Using Lemma 5.5, deduce

$$\mathbb{E}\left[{G^k}^T v^k\right] \geq \sum_{i=0}^{k-1} \gamma^i \mathbb{E}\left[||G^{k-i}||^2\right] - \frac{\overline{\lambda} L \gamma (\sigma^2 + \delta^2)}{(1-\gamma)^3}$$

□

Theorem 5.3. *(Convergent of SNAGD [3]) Given assumptions from section 5.3.1, τ defined above, for a iteration $N > \frac{1}{1-\gamma}$, $x^0 \in \mathbb{R}^d$, $\lambda_k \leq \overline{\lambda} \forall k \in \mathbb{N}$, $0 \geq \gamma < 1$ and $\{x^k\}$ defined as by Algorithm (7) then*

$$\mathbb{E}\left[||\nabla f(x^\tau)||^2\right] \leq \frac{1-\gamma}{\overline{\lambda}\gamma\tilde{N}}(f(x^0) - f^*) + \frac{N}{\tilde{N}}\left(\frac{\overline{\lambda}^2 L \gamma^2 (2-\gamma)(\sigma^2+\delta^2)}{(1-\gamma)^3} + \frac{(\overline{\lambda} + 2\overline{\lambda}^2 L)(\sigma^2+\delta^2)}{2} + \frac{\overline{\lambda}\delta^2}{2}\right)$$

where $\tilde{N} = N - \frac{\gamma}{1-\gamma}$.

Proof. We note $G^k = \nabla f(x^{k-1})$ is the expected gradient and $g^k = \nabla f_\xi(x^{k-1})$ is the stochastic gradient at iteration k. Because f is smooth and take a specifie iteration $k \in \mathbb{N}^*$, we have

$$f(x^k) \leq f(x^{k-1}) - \lambda_{k-1}\gamma {G^k}^T v^k - \lambda_{k-1} {G^k}^T g^k + \frac{{\lambda_{k-1}}^2 L ||\gamma v^k + g^k||^2}{2} \tag{33}$$

Apply Cauchy-Schwarz inequality and assumption 5.7, we have

$$\begin{aligned}
\|\gamma v^k + g^k\|^2 &= \|\gamma v^k\|^2 + \|g^k\|^2 + 2\langle \gamma v^k, g^k \rangle \\
&\leq \|\gamma v^k\|^2 + \|g^k\|^2 + 2\|\gamma v^k\|\|g^k\| \\
&\leq 2\|\gamma v^k\|^2 + 2\|g^k\|^2
\end{aligned} \tag{34}$$

and

$$-\langle {G^k}^T g^k \rangle \leq \|{G^k}^T g^k\| \leq \frac{\|G^k\|^2}{2} + \frac{\|g^k\|^2}{2} \leq \frac{\delta^2}{2} + \frac{\|g^k\|^2}{2} \tag{35}$$

Let $\overline{\lambda} \geq \lambda_k$, $\forall k \in \mathbb{N}$, injecting (35) and (34) to (33) and taking expectation of two side with Lemma 5.7, Lemma 5.8, apply assumption 5.7, we get

$$\begin{aligned}
\mathbb{E}\left[f(x^k)\right] &\leq \mathbb{E}\left[f(x^{k-1})\right] - \overline{\lambda}\gamma\left(\sum_{i=0}^{k-1}\gamma^i \mathbb{E}\left[||G^{k-i}||^2\right]\right) + \frac{\overline{\lambda}^2 L\gamma^2(\sigma^2+\delta^2)}{(1-\gamma)^3} + \frac{\overline{\lambda}^2 L\gamma^2(\sigma^2+\delta^2)}{2(1-\gamma)^2} \\
&\quad + \frac{\overline{\lambda}(\sigma^2+\delta^2)}{2} + \frac{\overline{\lambda}\delta^2}{2} + \overline{\lambda}^2 L(\sigma^2+\delta^2) \\
&\leq \mathbb{E}\left[f(x^{k-1})\right] - \overline{\lambda}\gamma\left(\sum_{i=0}^{k-1}\gamma^i \mathbb{E}\left[||G^{k-i}||^2\right]\right) + \frac{\overline{\lambda}^2 L\gamma^2(2-\gamma)(\sigma^2+\delta^2)}{(1-\gamma)^3} + \frac{(\overline{\lambda} + 2\overline{\lambda}^2 L)(\sigma^2+\delta^2)}{2} + \frac{\overline{\lambda}\delta^2}{2}
\end{aligned}$$

rearranging and summing over $k \in \{1...N\}$ we get

$$\overline{\lambda}\gamma\sum_{k=1}^{N}\sum_{i=0}^{k-1}\gamma^i \mathbb{E}\left[||G^{k-i}||^2\right] \leq f(x^0) - \mathbb{E}\left[f(x^N)\right] + N\left(\frac{\overline{\lambda}^2 L\gamma^2(2-\gamma)(\sigma^2+\delta^2)}{(1-\gamma)^3} + \frac{(\overline{\lambda} + 2\overline{\lambda}^2 L)(\sigma^2+\delta^2)}{2} + \frac{\overline{\lambda}\delta^2}{2}\right) \tag{36}$$

Zoom on the right hand side, replace $j = k - i$, we get

$$\begin{aligned} RHS &= \overline{\lambda}\gamma \sum_{k=1}^{N} \sum_{j=1}^{k} \gamma^{k-j} \mathbb{E}\left[||G^j||^2\right] \\ &= \overline{\lambda}\gamma \sum_{j=1}^{N} \mathbb{E}\left[||G^j||^2\right] \sum_{k=j}^{N} \gamma^{k-j} \\ &= \frac{\overline{\lambda}\gamma}{1-\gamma} \sum_{j=1}^{N} \mathbb{E}\left[||\nabla f(x^{j-1})||^2\right] (1 - \gamma^{N-j+1}) \\ &= \frac{\overline{\lambda}\gamma}{1-\gamma} \sum_{j=0}^{N-1} \mathbb{E}\left[||\nabla f(x^{j})||^2\right] (1 - \gamma^{N-j}). \end{aligned}$$

As the definition of τ in (23)

$$\sum_{j=0}^{N-1} (1 - \gamma^{N-j}) = N - \gamma \frac{1-\gamma^N}{1-\gamma} \geq N - \frac{\gamma}{1-\gamma}$$

and let $\tilde{N} = N - \frac{\gamma}{1-\gamma}$, we obtain

$$RHS \geq \frac{\overline{\lambda}\gamma\tilde{N}}{1-\gamma} \mathbb{E}\left[||\nabla f(x^{\tau})||^2\right] \tag{37}$$

From (36) and (37), with Assumption 5.6 that f is lower bounded by f^*, we obtain

$$\mathbb{E}\left[||\nabla f(x^{\tau})||^2\right] \leq \frac{1-\gamma}{\overline{\lambda}\gamma\tilde{N}}(f(x^0) - f^*) + \frac{N}{\tilde{N}} \left(\frac{\overline{\lambda}^2 L\gamma^2(2-\gamma)(\sigma^2+\delta^2)}{(1-\gamma)^3} + \frac{(\overline{\lambda} + 2\overline{\lambda}^2 L)(\sigma^2+\delta^2)}{2} + \frac{\overline{\lambda}\delta^2}{2} \right)$$

□

6 Experiments

In this section, experiments about the proposed algorithms are presented consist of four main problems: logistic regression, matrix factorization, cubic regularization and neural network. Through experiments, we choose the convergence sequences as $\varepsilon = \frac{1}{k^\alpha}$ for SNGDm and SNAGD and $\varepsilon = \frac{\alpha \log(k)^\beta}{k^{1.1}}$ for NGD, NGDm, NGD Nesterov and SNGD. Beside that, we note lr as the learning rate, also known as the step size or λ in our proposed algorithms, γ as the momentum hyper parameter and η_0 and η_1 as the hyper parameters which have been mentioned in our algorithms.

6.1 Accelerated NGD

In this part, we run the experiments with three problems.

1. **Logistic Regression.** The objective function $\frac{1}{n}\sum_{i=1}^{n} log(1 + exp(-b_i a_i^T x)) + \frac{\gamma}{2}\|x\|^2$ is the logistic loss with l_2-regularization. In this objective function, n is the number of observations, $\gamma > 0$ which is often chosen as $\frac{1}{n}$, is the regularization parameter and $(a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}, i \in \mathbb{N}^*$ are the observations. We apply the proposed algorithms NGDm and NGD Nesterov on the popular benchmark dataset such as Covtype, Mushroom, W8a.

2. **Matrix Factorization.** The objective function $\min_{X=[U,V]} f(X) = f(U,V) = \frac{1}{2}\|UV^T - A\|_F^2$ with $A \in \mathbb{R}^{m\times n}$, $r < \min\{m,n\}$, $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$, is nonconvex problem. We used Movielens 100K dataset[7] which is polular dataset for benchmark in Matrix Factorization problems. Like in [11], r is chosen as 10, 20 and 30. All algorithms start with the same initial point and its values are random.

3. **Cubic Regularization.** The objective function $f(x) = g^T x + \frac{1}{2} x^T H x + \frac{M}{6}\|x\|^3$, where $g \in \mathbb{R}^d$, $H \in \mathbb{R}^{d\times d}$ and $M > 0$ is only smooth locally. In this objective function, g and H are the gradient and the Hessian respectively. All algorithms start with the same initial point as $x = 0 \in \mathbb{R}^d$, with $M = 10, 20, 100$.

6.1.1 Logistic Regression

We compare proposed algorithms as NGDm, NGD Nesterov to related algorithms such as GD[1], AdGD, AdGD-accel[11]. In the dataset detail,

- **Covtype.** has 581,012 samples total and 54 dimensions.
- **Mushrooms.** has 8,124 samples total and 112 dimensions.
- **W8a.** has 49,749 samples total and 300 dimensions.

As in Figure 1, NGDm and NGD Nesterov have good results on all problems, especially NGDm has the most outstanding results. In the experiment, the algorithms use hyperparameters as follows:

- **GD.** $\frac{1}{L}$ lr where L is smoothness constant.
- **AdGD and AdGD-accel.** default hyperparameters as in public source code of [11][1].
- **NGD.** 1e-3 lr, 0.2 η_0, 0.15 η_1, 4.5 β and 2.0 α
- **NGD Nesterov and NGDm.** 1e-3 lr, 0.2 η_0, 0.19 η_1, 0.0 β and 3.0 α and 0.9 γ.

6.1.2 Matrix Factorization

Matrix factorization is commonly found in recommender system. This problem's popularity makes optimization for it very useful. Using Movielens 100K, a popular dataset in matrix factorization problem, we want to evaluate algorithms in practical application in the most objective way. As result in Figure 2, NGDm achieves a new result, much better than the algorithms considered together. In detail, the algorithms use hyperparameters as follows:

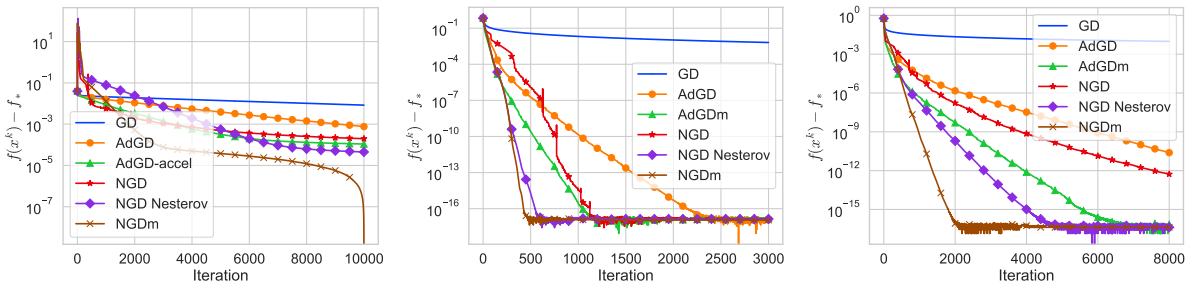
- **GD.** $\frac{1}{L}$ lr where L is smoothness constant.
- **AdGD and AdGD-accel.** default hyperparameters as in public source code of [11]
- **NGD.** 1e-5 lr, 0.49 η_0, 0.48 η_1, 0 β and 75 α for three cases of r
- **NGDm.** 1e-5 lr, 0.49 η_0, 0.48 η_1, 1 β, 2 α and 0.9 γ for three cases of r
- textbfNGD Nesterov. there are three different hyper parameters for three cases of r. In detail,
  1. **r=10.** 0.001 lr, 0.49 η_0, 0.44 η_1, 3 β and 5 α
  2. **r=20.** 0.001 lr, 0.49 η_0, 0.44 η_1, 4 β and 2 α
  3. **r=30.** 0.001 lr, 0.49 η_0, 0.44 η_1, 2 β and 1 α

6.1.3 Cubic Regularization

This problem is a subproblem of Newton's method. In the experiment, the algorithms use hyperparameters as follows:

- **GD.** line search lr in numbers spaced evenly on a log scale from -1 to 1
- **AdGD and AdGD-accel.** default hyperparameters as in public source code of [11].
- **NGD.** 1e-5 lr, 0.49 η_0, 0.45 η_1, 3.5 β and 4.0 α
- **NGD Nesterov and NGDm.** 1e-3 lr, 0.2 η_0, 0.19 η_1, 0.0 β and 3.0 α and 0.9 γ.

[1] https://github.com/ymalitsky/adaptive_gd



(a) Covtype Dataset (b) Mushroom Dataset (c) W8a Dataset

Fig. 1: The logistic regression objective results

6.2 Stochastic Algorithm

The experiments were run to evaluate the stochastic variants of NGD such as SNGDm and SNAGD on Cifar10 and Mnist dataset. About the datasets,

- **Mnist**[4] (Modified National Institute of Standards and Technology database, 10 classes) is a large collection of handwritten digits. It has a training set of 60,000 examples, and a test set of 10,000 examples. Each sample is a 28×28 pixels image.
- **Cifar10**[8] (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60000 32×32 color images. There are 6000 images per class with 5000 training and 1000 testing images per class.

Although having the same classes and number of samples, but two datasets are different about sample type. This leads to different optimizations for these datasets. One often use both of them for optimization algorithm benchmark. As result show in Figure 4 and Figure 5, NGD Nesterov and NGDm also have an outstanding result for both datasets. Beside that, we present the learning rate of algorithms. Unlike AdSGD, the learning rate of NGDm and NGD Nesterov increases with each cycle and tends to converge. This has also been proven with the theory above. In detail, the algorithms use hyperparameters as follows:

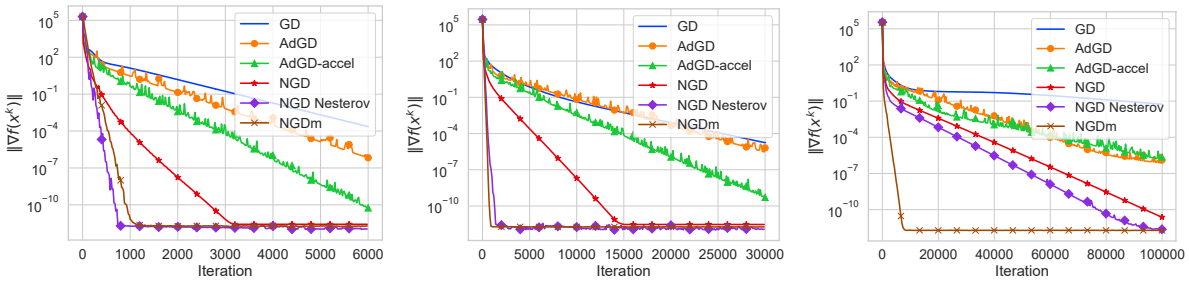
- **SGD and SGDm.** 0.01 lr and 0.9 momentum.
- **AdSGD.** default hyper parameters as in public source code of [11]
- **SNGD.** 1e-3 lr, 0.4 η_0, 0.35 η_1, 4.0 β and 3.0 α
- **SNAGD and SNGDm.** 1e-5 lr, 0.2 η_0, 0.15 η_1 and 0.9 α and 0.9 γ.

6.3 About Line Search Strategy

We implement a line search strategy to find the best possible hyper parameter in a set of hyper parameters, which can make it easier to apply our proposed algorithms to problems. We consider some hyper parameters to achieve good result: lr, η_0, η_1, β, α, where

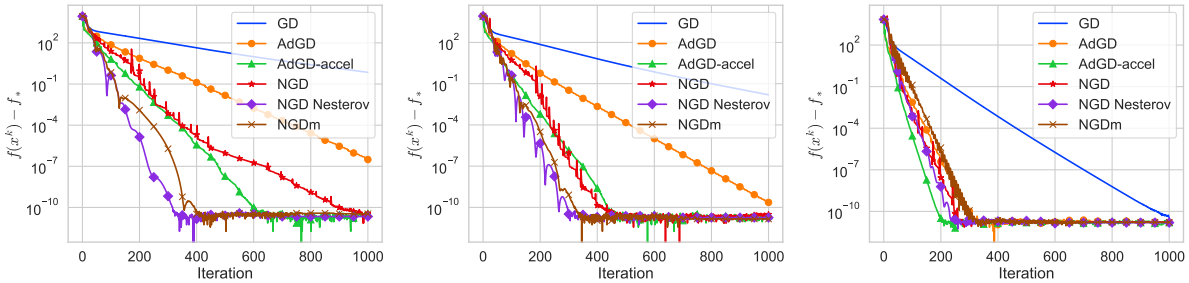
- lr in $[1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3]$.
- η_0 in $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6]$.
- η_1 in $[\eta_0 - 0.01, \eta_0 - 0.05, \eta_0 - 0.1]$.
- α and β in $[0, 1, 2, 3, 4, 5]$.

We run line search in a small number of iterations to find best hyper parameters. There are other good results but we only present the best results in this study.



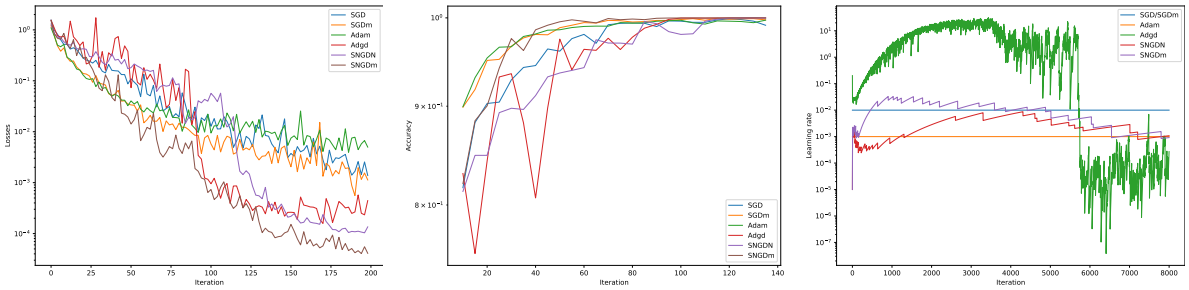
(a) r = 10 (b) r = 20 (c) r = 30

Fig. 2: Results for Matrix Factorization.



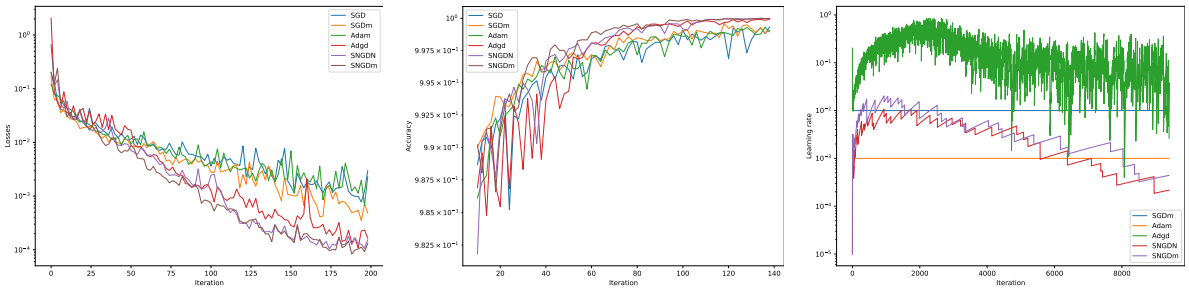
(a) M = 10 (b) M = 20 (c) M = 100

Fig. 3: Results for Cubic Regularization.



(a) Cifar10 Dataset, Objective (b) Cifar10 Accuracy (c) Cifar10 Learning Rate

Fig. 4: Results for Cifar10 Dataset.



(a) Mnist Dataset, Objective (b) Mnist Accuracy (c) Mnist Learning Rate

Fig. 5: Results for Mnist Dataset.

7 Conclusion

In this study, we combined a new step size with acceleration and stochastic methods, to propose new algorithms. We have also provided proofs of convergence for the proposed algorithms. In addition, experiments also show that the new algorithms work very well on some optimization problems and achieve new results.

Reference

[1] Augustin-Louis Cauchy. "ANALYSE MATHÉMATIQUE. – Méthodc générale pour la résolution des systèmes d'équations simultanées". In: 1847. URL: https://api.semanticscholar.org/CorpusID:123755271.

[2] Aaron Defazio. "Momentum via Primal Averaging: Theoretical Insights and Learning Rate Schedules for Non-Convex Optimization". In: (2021). arXiv: 2010.00406 [cs.LG].

[3] Alexandre Défossez et al. "A Simple Convergence Proof of Adam and Adagrad". In: 2022. arXiv: 2003.02395 [stat.ML].

[4] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: vol. 29. 6. IEEE, 2012, pp. 141–142.

[5] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. "Global convergence of the Heavy-ball method for convex optimization". In: (2014). arXiv: 1412.7457 [math.OC].

[6] Saeed Ghadimi and Guanghui Lan. "Stochastic First- and Zeroth-order Methods for Nonconvex Stochastic Programming". In: 2013. arXiv: 1309.5549 [math.OC].

[7] F. Maxwell Harper, Joseph A. Konstan, and Joseph A. "The MovieLens Datasets: History and Context". In: ACM Trans. Interact. Intell. Syst. 5 (2016), 19:1–19:19.

[8] Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: 2009. URL: https://api.semanticscholar.org/CorpusID:18268744.

[9] Hongwei Liu, Ting Wang, and Zexian Liu. "Some Modified Fast Iterative Shrinkage Thresholding Algorithms with a New Adaptive Non-Monotone Stepsize Strategy for Nonsmooth and Convex Minimization Problems". In: Comput. Optim. Appl. 83.2 (Nov. 2022), pp. 651–691. ISSN: 0926-6003. DOI: 10.1007/s10589-022-00396-6. URL: https://doi.org/10.1007/s10589-022-00396-6.

[10] Yanli Liu, Yuan Gao, and Wotao Yin. "An Improved Analysis of Stochastic Gradient Descent with Momentum". In: (2020). arXiv: 2007.07989 [math.OC].

[11] Yura Malitsky and Konstantin Mishchenko. "Adaptive Gradient Descent without Descent". In: Proceedings of the 37th International Conference on Machine Learning. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6702–6712. URL: https://proceedings.mlr.press/v119/malitsky20a.html.

[12] Yurii Nesterov. "A method for solving the convex programming problem with convergence rate $O(1/k^2)$". In: Proceedings of the USSR Academy of Sciences 269 (1983), pp. 543–547. URL: https://api.semanticscholar.org/CorpusID:145918791.

[13] Yurii Nesterov. "Introductory Lectures on Convex Optimization - A Basic Course". In: (2014). URL: https://api.semanticscholar.org/CorpusID:62288331.

[14] Lam Nguyen et al. "SGD and Hogwild! Convergence Without the Bounded Gradients Assumption". In: Proceedings of the 35th International Conference on Machine Learning. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 3750–3758. URL: https://proceedings.mlr.press/v80/nguyen18c.html.

[15] B.T. Polyak. "Some methods of speeding up the convergence of iteration methods". In: vol. 4. 5. 1964, pp. 1–17. DOI: https://doi.org/10.1016/0041-5553(64)90137-5. URL: https://www.sciencedirect.com/science/article/pii/0041555364901375.

[16] Tianbao Yang, Qihang Lin, and Zhe Li. "Unified Convergence Analysis of Stochastic Momentum Methods for Convex and Non-convex Optimization". In: (2016). arXiv: 1604.03257 [math.OC].