A Lyapunov analysis for accelerated gradient methods: from deterministic to stochastic case

M. Laborde
McGill University

A. Oberman
McGill University

Abstract

Recent work by Su, Boyd and Candes made a connection between Nesterov's accelerated gradient descent method and an ordinary differential equation (ODE). We show that this connection can be extended to the case of stochastic gradients, and develop Lyapunov function based convergence rates proof for Nesterov's accelerated stochastic gradient descent. In the gradient case, we show Nesterov's method arises as a straightforward discretization of a modified ODE. Established Lyapunov analysis is used to recover the accelerated rates of convergence in both continuous and discrete time. Moreover, the Lyapunov analysis can be extended to the case of stochastic gradients. The result is a unified approach to acceleration in both continuous and discrete time, and in for both stochastic and full gradients.

1 Introduction

Recently Su et al. [2014] showed that Nesterov's method for accelerated gradient descent can be obtained as the discretization of an Ordinary Differential Equation (ODE). The work of Su et al. [2014] resulted in a renewed interest in the continuous time approach to first order optimization, for example Wibisono et al. [2016], Wilson et al. [2016], Wilson et al. [2019]. The goals of the approach are to: (i) develop new insights into accelerated algorithms, and (ii) obtain new optimization algorithms. However, so far there has limited progress made on the second goal.

There is more than one ODE which can be discretized to obtain Nesterov's method Shi et al. [2018]. By discretizing a specific choice of ODE, we show that: (i) with a constant learning rate, we obtain Nesterov's method, (ii) with a decreasing learning rate, we obtain an accelerated stochastic gradient descent algorithm. We prove, using a Lyapunov function approach, that for both the convex and strongly convex cases, the algorithms converge at the optimal rate for the last iterate of SGD, with rate constants which are better than previously available.

1.1 Discussion of results

We write a stochastic gradient, $\widehat{g}(x,\xi) = \tilde{\nabla} f(x,\xi)$ as

$$\widehat{g}(x,\xi) = \nabla f(x) + e(x,\xi), \tag{1}$$

and make the stochastic approximation assumption on the error,

$$\mathbb{E}[e] = 0 \text{ and } \operatorname{Var}(e) = \sigma^2. \tag{2}$$

Relaxing the assumption (2) to cover the finite sum case is the subject of future work. Making the assumption (2) allows our results to depend only on σ^2. In what follows, G^2, is a bound for $\mathbb{E}[\widehat{g}^2]$.

In the convex case, the optimal rate for the last iterate of SGD (see Shamir and Zhang [2013]) is order $\log(k)/\sqrt{k}$ with a rate constant that depends on G^2. (Jain et al. [2019] remove the log factor, but only assuming that the number of iterations is decided in advance.) We obtain the $\mathcal{O}(\log(k)/\sqrt{k})$ rate for the last iterate, with a constant which depends on σ, but is independent of the L-smoothness. See Table 1. Below we also present an algorithm with a learning rate parameterized by α. This algorithm has a convergence rate of $\mathcal{O}(1/k^{2-2\alpha})$ which goes to $1/\sqrt{k}$ as $\alpha \to 3/4$, but the rate constant is not controlled. See Proposition 4.5 for the full expression for the rate for both algorithms.

In the strongly convex case, we obtain the optimal $\mathcal{O}(1/k)$ rate for the last iterate, with constants independent of the L-smoothness bound of the gradient. (The constant, L, appears in the algorithm for initialization of the learning rate.) This improves on previous results, Nemirovski et al. [2009], Shamir and Zhang

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

Table 1: Convergence rate $\mathbb{E}[f(x_k) - f^*]$ after k steps. for f a convex, L-smooth function. G^2 is a bound on $\mathbb{E}[\tilde{\nabla} f(x)^2]$, and σ^2 given by (2). h_k is the learning rate. E_0 is the initial value of the Lyapunov function. Top: convex case, D is the diameter of the domain, c is a free parameter. Bottom: μ-strongly convex case, $C_f := \frac{L}{\mu}$ $h_k = \mathcal{O}(1/k)$.

	Shamir and Zhang [2013]	Acc. SGD (Proposition 4.5)
h_k	$\frac{c^2}{\sqrt{k}}$	$\frac{c}{k^{3/4}}$
Rate	$\left(\frac{D^2}{c^2} + c^2 G^2\right) \frac{(2+\log(k))}{\sqrt{k}}$	$\frac{\frac{E_0}{16c^2} + c^2\sigma^2(1+\log(k))}{(k^{1/4}-1)^2}$

Nemirovski et al. [2009]	Shamir and Zhang [2013]	Jain et al. [2019]	Acc. SGD (Proposition 5.6)
$\frac{2C_f G^2}{\mu k}$	$\frac{17G^2(1+\log(k))}{\mu k}$	$\frac{130G^2}{\mu k}$	$\frac{4\sigma^2}{\mu k + 4\sigma^2 E_0^{-1}}$

[2013], Jain et al. [2018], where the rate depends on G. See Table 1.

1.2 Outline of our approach

SGD can be interpreted as a time discretization of the gradient descent ODE, $\dot{x}(t) = -\nabla f(x(t))$, with perturbed gradients. Solutions of the ODE decrease a rate-generating Lyapunov function, (18) below. We extend this analysis to an ODE for accelerated gradient descent, which also has an associated Lyapunov function. The Lyapunov analysis in the accelerated case is substantially more complex than in the gradient descent case. So, rather than proceeding directly, we first perform an abstract Lyapunov function analysis which involves: (i) an ode, (ii) a rate-generating Lyapunov function, and (iii) passing from deterministic to stochastic gradients in the analysis.

The key step is to obtain the following inequality on a rate-generating Lyapunov function.

$$E(t_{k+1}, z_{k+1}) \leq (1 - r_E h_k) E(t_k, z_k) + h_k \beta_k. \quad (3)$$

Here, h_k is the learning rate, and $r_E \geq 0$ is a rate constant. The term β_k depends on the error e_k from (1), and reduces to zero in the full gradient case $e = 0$. When we take expectations in (3), $\mathbb{E}[\beta_k]$ is proportional to $h_k\sigma^2$. The inequality (3) can be seen as a generalization of the fundamental lemma in Bottou et al. [2016], where the objective is replaced by a Liapunov function.

Schmidt et al. [2011] used a related approach, but they considered the case where the magnitude of the errors decrease quickly enough to obtain the rates in the deterministic case. A continuous time version of the analysis from Schmidt et al. [2011] was performed in Attouch et al. [2016]. Serhat Aybat et al. [2019] recently performed a similar Lyapunov analysis however they do not obtain a rate.

1.3 Other related work

Continuous time analysis also appears in Flammarion and Bach [2015], Lessard et al. [2016], and Krichene et al. [2015], among many other recent works. The Lyapunov approach to proving convergence rates appears widely in optimization, for example, see Beck and Teboulle [2009] for FISTA.

Convergence rates for averaged SGD are available in a wide setting, (Bottou et al. [2016], Lacoste-Julien et al. [2012], Rakhlin et al. [2012], Qian et al. [2019]). In the non-asymptotic regime, the last iterate of SGD is often preferred in current applications. The optimal convergence rate for SGD is $O(1/k)$ in the smooth, strongly convex case, and $O(1/\sqrt{k})$ in the convex, nonsmooth case (Nemirovski et al. [2009], Bubeck [2014]).

When SGD is combined with momentum (Polyak [1964], Nesterov [2013]) empirical performance is improved, but this improvement is not always theoretically established (Kidambi et al. [2018]). Accelerated versions of stochastic gradient descent algorithms are comparatively more recent: they appear in Lin et al. [2015] as well as in Frostig et al. [2015] and Jain et al. [2018]. A direct acceleration method with a connection to Nesterov's method can be found in Allen-Zhu [2017].

Organization We organize the paper as follows. Section 2 is devoted to the presentation of an abstract Lyapunov analysis. We apply this analysis in the full gradient, and the stochastic gradient case. In Section 3 we apply the abstract analysis to the gradient descent ODE, in the case of convex and strongly convex functions.

Finally, in Sections 4 and 5 we study the accelerated ODEs, and the related Lyapunov functions in the convex and strongly convex cases, respectively. In the convex case we show that the discretization of (1st-ODE) below leads to Nesterov's method with a constant learning rate. However setting the learning rate to $h_k = O(k^{-3/4})$ in (FE-C), (which is different from the usual order $k^{-1/2}$ rate) leads to an accelerated stochastic gradient descent algorithm. In the strongly convex case, we begin with the ODE, (1st-ODE-SC), which is discretized with a learning rate h_k to result in (FE-SC). In the full gradient case, setting the constant learning rate $h_k = 1/\sqrt{L}$ leads to Nesterov's method. In the stochastic gradient case, setting $h_k = O(1/k)$ as described in Proposition 5.6, leads to an accelerated SGD algorithm.

Notation Let f be proper convex function, and write $x_* = \operatorname{argmin}_x f(x)$ and $f^* = f(x_*)$. We say f is L-smooth if

$$f(y) - f(x) + \nabla f(x) \cdot (x - y) \leq \frac{L}{2}|x - y|^2, \qquad (4)$$

and f is μ-strongly convex if

$$f(x) + \nabla f(x) \cdot (y - x) \leq f(y) - \frac{\mu}{2}|x - y|^2, \qquad (5)$$

Write $C_f := \frac{L}{\mu}$ for the condition number of f.

2 Abstract Lyapunov Analysis: going from continuous to discrete time

In this section we present an abstract analysis showing how to go from deterministic continuous time to stochastic discrete time. We make a definition of the continuous problem and start exposing the problem in an abstract setting. Then, we show how this framework can be extended to study the stochastic case. We introduce abstract assumptions (e.g inequalities (8), (14) and (15)) which are easily verified in the gradient case. These assumptions are later proven in the accelerated case.

2.1 ODEs, Perturbed ODEs and discretizations

Definition 2.1. *Let $g(t, z, p)$ be L_g-Lipschitz continuous, and affine in the variable p,*

$$g(t, z, p) = g_1(t, z) + g_2(t, z)p. \qquad (6)$$

Consider the Ordinary Differential Equation

$$\dot{z}(t) = g(t, z(t), \nabla f(z(t))) \qquad \text{(ODE)}$$

Referring to (1)*, Consider also the perturbed ODE,*

$$\dot{z}(t) = g(t, z(t), \tilde{\nabla} f(z(t))). \qquad \text{(PODE)}$$

(ODE) has unique solutions in all time for every initial condition $z(0) = z_0 \in \mathbb{R}^n$. Moreover, if we assume that $e(t)$ is Lipschitz continuous in time, then (PODE) has unique solutions in all time for every initial condition $z(0) = z_0 \in \mathbb{R}^n$. On the other hand, if we wish to consider a model of $e(t)$ which is more consistent with random, mean zero errors, then (PODE) is no longer well-posed as an ODE. However, we can consider a Stochastic Differential Equation (Oksendal [2013], Pavliotis [2016]), which would lead to similar results to the discrete case where we take expectations of the mean zero error term. We do not pursue the SDE approach here to simplify the exposition.

Definition 2.2. *For a given learning rate schedule $h_k \geq 0$ and $t_k = \sum_{i=0}^{k} h_i$, the forward Euler discretization of* (ODE) *corresponds to the sequence*

$$z_{k+1} = z_k + h_k g(t_k, z_k, \nabla f(z_k)), \qquad \text{(FE)}$$

given an initial value z_0. Similarly, the forward Euler discretization of (PODE) *is given by*

$$z_{k+1} = z_k + h_k g(t_k, z_k, \nabla f(z_k) + e_k) \qquad \text{(FEP)}$$

The solution of (FE) or of (FEP) can be interpolated to be a function of time $z^h : [0, T) \to \mathbb{R}^n$ by simply setting $z^h(t_k) = z_k$ along with piecewise constant or piecewise linear interpolation between time steps. It is a standard result from numerical analysis of ODE theory (Iserles [2009]) that functions z^h converge to $z(t)$ with error of order h_k, provided $h_k \leq 1/L_g$.

2.2 Lyapunov analysis for the unperturbed ODE

First, we give the definition of a rate-generating Lyapunov function for (ODE).

Definition 2.3. *We say $E(t, z)$ is a rate-generating Lyapunov function for* (ODE) *if, for all $t > 0$, $E(t, z^*) = 0$ and $\nabla E(t, z^*) = 0$ where z^* is a stationary solution of* (ODE)*, i.e. $g(t, z^*, \nabla f(z^*)) = 0$, and if there are constants $r_E, a_E \geq 0$ such that*

$$\begin{aligned} \partial_t E(t, z) + \nabla E(t, z) \cdot g(t, z, \nabla f(z)) \\ \leq -r_E E(t, z) - a_E |g(t, z, \nabla f(z))|^2 \end{aligned} \qquad (7)$$

Remark 2.4. *We use a similar definition which is used in the convex case, where now $r_E = 0$ and the constant a_E is extended to be a function of time $a_E(t)$. The analysis below does not change.*

Then, we can deduce the following rate in the continuous case.

Lemma 2.5. *Let E be a rate generating Lyapunov function for* (ODE)*. Then*

$$E(t, z(t)) \leq E(0, z(0)) \exp(-r_E t)$$

Proof.

$$\begin{aligned}&\frac{d}{dt}E(t,z(t))\\&=\partial_t E(t,z(t))+\nabla E(t,z(t))\cdot g(t,z(t),\nabla f(z(t)))\\&\qquad\le -r_E E(t,z(t))-a_E|g(t,z(t),\nabla f(z(t)))|^2\end{aligned}$$

by assumption (7). Gronwall's inequality gives the result. □

2.3 Dissipation of E along (FEP)

Now we consider the discrete case where ∇f has been replaced by a perturbed gradient $\tilde{\nabla} f=\nabla f+e$ and then we compute the dissipation of E along (FEP).

First, we assume, in addition, there exists $L_E>0$ such that E satisfies,

$$\begin{aligned}E(t_{k+1},z_{k+1})-E(t_k,z_k)&\le \partial_t E(t_k,z_k)(t_{k+1}-t_k)\\&+\langle\nabla E(t_k,z_k),z_{k+1}-z_k\rangle+\frac{L_E}{2}|z_{k+1}-z_k|^2.\end{aligned}\tag{8}$$

Lemma 2.6. *Let z_k be the solution of* (FEP). *Suppose E is a rate-generating Lyapunov function for* (ODE) *which satisfies* (8). *Then if h_k satisfies*

$$h_k\le\frac{2a_E}{L_E},\tag{9}$$

we have

$$E(t_{k+1},z_{k+1})-E(t_k,z_k)\le -h_k r_E E(t_k,z_k)+h_k\beta_k\tag{10}$$

where β_k is defined by

$$\begin{aligned}&\beta_k:=\langle\nabla E(t_k,z_k),g_2(t_k,z_k)e_k\rangle\\&+L_E h_k\left\langle g(t_k,z_k,\nabla f(z_k))+\frac{1}{2}g_2(t_k,z_k)e_k,g_2(t_k,z_k)e_k\right\rangle.\end{aligned}\tag{11}$$

2.4 Convergence in expectation: $r_E>0$

Now we consider the case where the error e_k satisfies (2). We require that h_k satisfies (9). (This assumption is easy to enforce, since, in practice, we can perform a few steps of the algorithm, and reset, so that until h_k small enough). We consider the case $r_E>0$, which applies to the strongly convex case.

By definition of β_k,

$$\mathbb{E}[\beta_k]=\frac{h_k L_E g_2(t_k,z_k)^2\sigma^2}{2},$$

and then, taking the expectation in (10), we obtain

$$\begin{aligned}&\mathbb{E}[E(t_{k+1},z_{k+1})]\\&\le(1-h_k r_E)E(t_k,z_k)+\frac{h_k^2 L_E g_2(t_k,z_k)^2\sigma^2}{2}.\end{aligned}\tag{12}$$

Then, we deduce the following result

Proposition 2.7 (Case $r_E>0$)**.** *Assume that $r_E>0$ and $\overline{g_2}=\max_{(t,z)}g_2(t,z)<+\infty$. If*

$$h_k:=\frac{2}{r_E(k+\alpha^{-1}E_0^{-1})}\qquad\textit{where}\qquad\alpha=\frac{r_E^2}{2L_E\overline{g_2}^2\sigma^2},$$

then,

$$\mathbb{E}[E(t_k,z_k)]\le\frac{1}{\alpha(k+\alpha^{-1}E_0^{-1})}.\tag{13}$$

2.5 Convergence in expectation: $r_E=0$

Since the convex case corresponds to the $r_E=0$ case, the rate of convergence has to be incorporated into the Lyapunov function. Then we assume that there exists five constants $a_1,a_2,a_3,b_1,b_2\ge 0$ such that

$$\mathbb{E}[E(t_{k+1},z_{k+1})]-E(t_k,z_k)\le\frac{(a_1+a_2t_k+a_3t_k^2)h_k^2\sigma^2}{2}\tag{14}$$

and in addition, that

$$\mathbb{E}[E(t_k,z_k)]\ge(b_1t_k+b_2t_k^2)(\mathbb{E}[f(x_k)]-f^*).\tag{15}$$

We will see below that these conditions are easily verified in the non-accelerated case, and that it also holds for the accelerated case. Then we obtain

Proposition 2.8. *Assume that $r_E=0$ and* (14)-(15) *hold. If $h_k=\frac{c}{k^\alpha}$, $t_k=\sum_{i=1}^k h_i$, then the following holds:*

- *Case $a_1,a_2,b_1>0$, $a_3=b_2=0$: If $\alpha\in\left[\frac{2}{3},1\right)$, then*

$$\mathbb{E}[f(x_k)]-f^*\le\begin{cases}\dfrac{\frac{1-\alpha}{c}E_0+\left(\frac{a_1c(1-\alpha)\alpha}{2\alpha-1}+\frac{a_2c^2(3\alpha-1)}{2(3\alpha-2)}\right)\sigma^2}{b_1(k^{1-\alpha}-1)}, & \alpha\in\left(\frac{2}{3},1\right)\\[2ex]\dfrac{\frac{1}{3c}E_0+\left(\frac{2a_1c}{3}+\frac{a_2c^2}{2}(1+\log(k))\right)\sigma^2}{b_1(k^{1/3}-1)}, & \alpha=\frac{2}{3}.\end{cases}$$

- *Case $a_1=a_2=b_1=0$, $a_3>0$ and $b_2>0$: If $\alpha\in\left[\frac{3}{4},1\right)$, then*

$$\mathbb{E}[f(x_k)]-f^*\le\begin{cases}\dfrac{\frac{(1-\alpha)^2}{c^2}E_0+\frac{a_3c^2(4\alpha-2)\sigma^2}{2(4\alpha-3)}}{b_2(k^{1-\alpha}-1)^2}, & \alpha\in\left(\frac{3}{4},1\right)\\[2ex]\dfrac{\frac{1}{16c^2}E_0+\frac{a_3c^2\sigma^2}{2}(1+\log(k))}{b_2(k^{1/4}-1)^2}, & \alpha=\frac{3}{4}.\end{cases}$$

3 Application to gradient descent

In this section, we apply our previous abstract analysis to gradient descent for convex and strongly convex functions. In this case, g from (6) is simply given by

$$g(t,z,p)=-p$$

Let f be a μ-strongly convex, L-smooth function. Consider

$$\dot{x}(t) = -\nabla f(x(t)) \tag{16}$$

and its associated perturbed forward Euler scheme, where the gradient is replaced by $\tilde{\nabla} f = \nabla f + e$

$$x_{k+1} - x_k = -h_k(\nabla f(x_k) + e_k), \tag{17}$$

with initial condition x_0. Define the convex and strongly convex Lyapunov functions

$$\begin{aligned} E^c(t,x) &:= t(f(x) - f^*) + \frac{1}{2}|x - x^*|^2, \\ E^{sc}(x) &:= f(x) - f^* + \frac{\mu}{2}|x - x^*|^2. \end{aligned} \tag{18}$$

3.1 Continuous analysis

First, we show that the Lyapunov functions satisfy (8).

Proposition 3.1.
- *E^c is a Lyapunov function in the sense of Definition 2.3 with $r_{E^c} = 0$ and $a_{E^c} = t$. In addition, E^c satisfies (8) with $L_{E^c} = Lt_{k+1} + 1$.*
- *E^{sc} is a Lyapunov function in the sense of Definition 2.3 with $r_{E^{sc}} = \mu$ and $a_{E^{sc}} = 1$. In addition, E^{sc} satisfies (8) with $L_{E^{sc}} = L + \mu$.*

Applying Lemma 2.5 to E^c and E^{sc}, we recover the usual rates.

Corollary 3.2. *Let x be the solution of (16). Then, for all $t \geq 0$,*

$$f(x(t)) - f^* \leq \frac{1}{2t}|x_0 - x_*|^2 \qquad (cvx)$$

$$f(x(t)) - f^* + \frac{\mu}{2}|x(t) - x_*|^2 \leq e^{-\mu t} E^{sc}(x_0) \qquad (st.\ cvx)$$

3.2 Perturbed gradient descent

In the case where the gradient is replaced by a perturbed gradient $\tilde{\nabla} f = \nabla f + e$, where e is an error term, by Lemma 2.5, the dissipation of E^c and E^{sc} along (17) is given in the next proposition.

Proposition 3.3. *Let x_k be the sequence generated by perturbed gradient descent (17).*

- *Convex case: If h_k satisfies, $h_k \leq \frac{1}{L}$, and $t_k = \sum_{i=0}^k h_i$, then, E_k^c satisfies*

$$E_{k+1}^c \leq E_k^c + h_k \beta_k^c, \tag{19}$$

where $\beta_k^c := -\langle x_k - x^* - t_{k+1}\nabla f(x_k), e_k\rangle + h_k(Lt_{k+1} + 1)\left\langle \nabla f(x_k) + \frac{1}{2}e_k, e_k\right\rangle$.

- *Strongly convex case: If h_k satisfies $h_k \leq \frac{2}{L+\mu}$, then, E_k^{sc} satisfies*

$$E_{k+1}^{sc} \leq (1 - h_k\mu)E_k^{sc} + h_k\beta_k^{sc}, \tag{20}$$

where $\beta_k^{sc} = \langle \mu(x_k - x^*) + \nabla f(x_k), e_k\rangle + h_k(L+\mu)\left\langle \nabla f(x_k) + \frac{1}{2}e_k, e_k\right\rangle$.

3.3 Variable time step and convergence in expectation

Now, consider the case of a variable time step h_k and a zero-mean and fixed Variance error e_k i.e. e_k satisfies (2). From Proposition 3.3, since $\mathbb{E}[\beta_k^c] = \frac{h_k^2(Lt_{k+1}+1)\sigma^2}{2}$ and $\mathbb{E}[\beta_k^{sc}] = \frac{h_k^2(L+\mu)\sigma^2}{2}$,

$$\begin{aligned} \mathbb{E}[E_{k+1}^c] &\leq E_k^c + \frac{h_k^2(Lt_{k+1}+1)\sigma^2}{2} \\ \mathbb{E}[E_{k+1}^{sc}] &\leq (1-\mu h_k)E_k^{sc} + \frac{h_k^2(L+\mu)\sigma^2}{2}. \end{aligned}$$

In addition,

$$\mathbb{E}[E_k^c] \geq t_k(\mathbb{E}[f(x_k)] - f^*),$$

which correpsond, for E_k^c, to (14) and (15) with

$$a_1 = \frac{L(h_k+1)}{2},\ a_2 = \frac{L}{2},\ a_3 = 0,\ b_1 = 1 \text{ and } b_2 = 0.$$

And, for E_k^{sc}, to the case where $r_{E^{sc}} = \mu > 0$ and $g_2^2 = 1$, then Proposition 2.7 and Proposition 2.8 give:

Proposition 3.4.
- *Convex case ($r_{E^c} = 0$): If $h_k^c := \frac{c}{k^\alpha}$ satisfies $h_k^c \leq \frac{1}{L}$, and $t_k = \sum_{i=1}^k h_i^c$, then the following holds:*

$$\mathbb{E}[f(x_k)] - f^* \leq \begin{cases} \dfrac{\frac{1-\alpha}{c}E_0 + \frac{Lc^2(3\alpha-1)\sigma^2}{2(3\alpha-2)}}{k^{1-\alpha}-1}, & \alpha \in \left(\frac{2}{3}, 1\right) \\ \dfrac{\frac{1}{3c}E_0 + \left(3c + Lc^2(1+\ln(k))\right)\frac{\sigma^2}{2}}{k^{1/3}-1}, & \alpha = \frac{2}{3}. \end{cases}$$

- *Strongly convex case ($r_{E^{sc}} = \mu > 0$): If*

$$h_k^{sc} := \frac{2}{\mu(k + (\alpha^{sc}E_0^{sc})^{-1})},\quad \alpha^{sc} := \frac{\mu}{2(C_f+1)\sigma^2}$$

and $t_k = \sum_{i=0}^k h_i$, then, if h_k^{sc} satisfies $h_k^{sc} \leq \frac{2}{L+\mu}$, the following holds:

$$\mathbb{E}[E(x_k)] \leq \frac{2(C_f+1)\sigma^2}{\mu k + 2(C_f+1)\sigma^2 E_0^{sc-1}}.$$

4 Accelerated method: convex case

In the remainder of the paper, we will extend the analysis developed in Section 3 to the accelerated gradient methods. In this section, we consider the case of convex functions.

4.1 ODE and derivation of Nesterov's method

Nesterov's method for a convex, L-smooth function, f, can be written as [Nesterov, 2013, Section 2.2]

$$\begin{cases} x_{k+1} = y_k - \dfrac{1}{L}\nabla f(y_k) \\ y_{k+1} = x_{k+1} + \dfrac{k}{k+3}(x_{k+1} - x_k) \end{cases} \qquad \text{(C-Nest)}$$

Su et al. [2014] made a connection between (C-Nest) and the second order ODE

$$\ddot{x} + \frac{3}{t}\dot{x} + \nabla f(x) = 0 \tag{A-ODE}$$

(A-ODE) can be written as the first order system

$$\begin{cases} \dot{x} = \dfrac{2}{t}(v - x) \\ \dot{v} = -\dfrac{t}{2}\nabla f(x). \end{cases} \tag{21}$$

Our starting point is the following system of first order ODEs, which is a perturbation of (21)

$$\begin{cases} \dot{x} = \dfrac{2}{t}(v - x) - \dfrac{1}{\sqrt{L}}\nabla f(x) \\ \dot{v} = -\dfrac{t}{2}\nabla f(x), \end{cases} \tag{1st-ODE}$$

Solutions of (1st-ODE) decrease the same Lyapunov function faster than solutions of (A-ODE). Write $z = (x, v)$ to write (1st-ODE) in the form (6)

$$g(t, z, \nabla f(x)) = \frac{2}{t}\begin{pmatrix} v - x \\ 0 \end{pmatrix} + \begin{pmatrix} -\frac{1}{\sqrt{L}} \\ -\frac{t}{2} \end{pmatrix}\nabla f(x).$$

The system (1st-ODE) is equivalent to the following ODE

$$\ddot{x} + \frac{3}{t}\dot{x} + \nabla f(x) = -\frac{1}{\sqrt{L}}\left(D^2 f(x)\cdot\dot{x} + \frac{1}{t}\nabla f(x)\right) \tag{H-ODE}$$

which has an additional Hessian damping term with coefficient $1/\sqrt{L}$. Notice that (H-ODE) is a perturbation of (A-ODE) of order $\frac{1}{\sqrt{L}}$, and the perturbation goes to zero as $L \to \infty$. Similar ODEs have been studied by Alvarez et al. [2002], they have been shown to accelerate gradient descent in continuous time in Attouch et al. [2016].

There is more than one ODE which can be discretized to obtain Nesterov's method. Shi et al. [2018] introduced a family of high resolution second order ODEs which also lead to Nesterov's method. In this context, (H-ODE) corresponds to the high-resolution equation with the parameter $\frac{1}{\sqrt{L}}$. Making the specific choice of first order system (1st-ODE) considerably simplifies the analysis, allowing for shorter, clearer proofs which generalize to the stochastic gradient case (which was not treated in Shi et al. [2018]).

The system (1st-ODE) can be discretized to recover Nesterov's method using an explicit discretization with a constant time step $h = \frac{1}{\sqrt{L}}$, as demonstrated in Proposition 4.2.

Definition 4.1. *Let $h_k > 0$ be a decreasing sequence of small time step/learning rate and t_k an increasing sequence of discrete time. The discretization of* (1st-ODE) *corresponds to an explicit time discretization with gradients evaluated at y_k, the convex combination of x_k and v_k, defined below,*

$$\begin{cases} x_{k+1} - x_k = \frac{2h_k}{t_k}(v_k - x_k) - \frac{h_k}{\sqrt{L}}\nabla f(y_k), \\ v_{k+1} - v_k = -\frac{h_k t_k}{2}\nabla f(y_k), \\ y_k = \left(1 - \frac{2h_k}{t_k}\right)x_k + \frac{2h_k}{t_k}v_k. \end{cases} \tag{FE-C}$$

Then the following result holds.

Proposition 4.2. *The discretization of* (1st-ODE) *given by* (FE-C) *with $h_k = h = 1/\sqrt{L}$ and $t_k = h(k+2)$ is equivalent to the standard Nesterov's method* (C-Nest).

4.2 Perturbed gradient: discrete time

Replacing gradients with $\tilde{\nabla} f = \nabla f + e$, the discretization of (1st-ODE), (FE-C), with a time step h_k, becomes

$$\begin{cases} x_{k+1} - x_k = \dfrac{2h_k}{t_k}(v_k - x_k) - \dfrac{h_k}{\sqrt{L}}(\nabla f(y_k) + e_k), \\ v_{k+1} - v_k = -h_k\dfrac{t_k}{2}(\nabla f(y_k) + e_k), \end{cases} \tag{Per-FE-C}$$

where y_k is as in (FE-C), $t_k = \sum_{i=0}^k h_i$.

Definition 4.3. *Define the continuous time parametrized Lyapunov function*

$$E^{ac,c}(t, x, v; \epsilon) := (t-\epsilon)^2(f(x) - f^*) + 2|v - x^*|^2 \tag{22}$$

Define the discrete time Lyapunov function E_k^c by

$$E_k^{ac,c} = E^{ac,c}(t_k, x_k, v_k; h_k) = E^{ac,c}(t_{k-1}, x_k, v_k; 0) \tag{23}$$

It is well-known that $E^{ac,c}$ is Lyapunov function for (1st-ODE), see Su et al. [2014]. But note, compared to Su-Boyd-Candés' ODE (A-ODE), there is a gap in the dissipation of the Lyapunov function $E^{ac,c}$, which will not be there if the extra term, $-\frac{1}{\sqrt{L}}\nabla f(x)$, was missing, see Appendix C.2. In particular, if z and $\tilde{z}$ are solutions of (A-ODE) and (1st-ODE) respectively, then we can prove faster convergence due to the gap.

Proposition 4.4. *Let x_k, v_k, y_k be sequences generated by* (Per-FE-C)-(FE-C)*. Then, for $h_k \leq \frac{1}{\sqrt{L}}$,*

$$E_{k+1}^{ac,c} - E_k^{ac,c} \leq h_k\beta_k, \tag{24}$$

where $\beta_k := -t_k\langle 2(v_k - x^*) - \frac{t_k}{\sqrt{L}}\nabla f(y_k), e_k\rangle + 2h_k t_k^2\left\langle \nabla f(y_k) + \frac{e_k}{2}, e_k\right\rangle$.

4.3 Convergence in expectation

Assume e_k satisfies (2). We deduce that $\mathbb{E}[\beta_k] = h_k t_k^2 \sigma^2$. Then, Proposition 4.4 gives

$$\mathbb{E}[E_{k+1}^{ac,c}] - E_k^{ac,c} \leq h_k^2 t_k^2 \sigma^2.$$

In addition,

$$\mathbb{E}[E_k^{ac,c}] \geq t_{k-1}^2 \mathbb{E}[f(x_k)] - f^*.$$

That corresponds to (14) and (15) with

$$a_1 = a_2 = 0,\ a_3 = 2,\ b_1 = 0 \text{ and } b_2 = 1. \tag{25}$$

Then we can apply Proposition 2.8 to obtain the following convergence result.

Proposition 4.5. *Assume $h_k := \frac{c}{k^\alpha} \leq \frac{1}{\sqrt{L}}$ and $t_k = \sum_{i=1}^k h_i$, then we have the following bound on $\mathbb{E}[f(x_k)] - f^*$*

$$\begin{aligned} &\frac{\frac{(1-\alpha)^2}{c^2}E_0 + c^2\sigma^2\frac{\alpha-1/2}{\alpha-3/4}}{(k^{1-\alpha}-1)^2}, \quad \alpha \in \left(\frac{3}{4}, 1\right) \\ &\frac{\frac{1}{16c^2}E_0 + c^2\sigma^2(1+\log(k))}{(k^{1/4}-1)^2}, \quad \alpha = \frac{3}{4}. \end{aligned}$$

Remark 4.6. *Since in (25), the parameters do not depend on L. We observe that the rates of convergence of $\mathbb{E}[f(x_k)] - f^*$ in the previous proposition do not depend on the smoothness of f and are accelerated compared to SGD (Proposition 3.4).*

5 Accelerated method: Strongly Convex case

This section is devoted to the analysis of (H-ODE-SC) and in particular to its first order system (1st-ODE-SC) in continuous and discrete time using a Lyapunov analysis in the strongly convex case. Then, we present a convergence rate for the expectation of f.

5.1 ODE and derivation of Nesterov's method

In the case of a μ-strongly convex function, f, Nesterov's method can be written as follows [Nesterov, 2013, Section 2.2],

$$\begin{cases} x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \\ y_{k+1} = x_{k+1} + \frac{1-\sqrt{C_f^{-1}}}{1+\sqrt{C_f^{-1}}}(x_{k+1} - x_k). \end{cases} \tag{SC-Nest}$$

This method can be seen as the discretization of Polyak's ODE

$$\ddot{x} + 2\sqrt{\mu}\dot{x} + \nabla f(x) = 0, \tag{A-ODE-SC}$$

which is an accelerated gradient method when f is *quadratic* see Scieur et al. [2017], and can be rewritten as the following first order system

$$\begin{cases} \dot{x} = \sqrt{\mu}(v - x), \\ \dot{v} = \sqrt{\mu}(x - v) - \frac{1}{\sqrt{\mu}}\nabla f(x). \end{cases} \tag{26}$$

Here, as in the convex case, we are interested in a perturbed version of (26),

$$\begin{cases} \dot{x} = \sqrt{\mu}(v - x) - \frac{1}{\sqrt{L}}\nabla f(x), \\ \dot{v} = \sqrt{\mu}(x - v) - \frac{1}{\sqrt{\mu}}\nabla f(x), \end{cases} \tag{1st-ODE-SC}$$

which corresponds to (6) where $z = (x, v)$ and

$$g(t, z, \nabla f(x)) = \sqrt{\mu}\begin{pmatrix} v - x \\ x - v \end{pmatrix} + \begin{pmatrix} -\frac{1}{\sqrt{L}} \\ -\frac{1}{\sqrt{\mu}} \end{pmatrix}\nabla f(x).$$

This system is equivalent to the second order equation with Hessian damping for a smooth f

$$\ddot{x} + 2\sqrt{\mu}\dot{x} + \nabla f(x) = -\frac{1}{\sqrt{L}}\left(D^2 f(x)\cdot\dot{x} + \sqrt{\mu}\nabla f(x)\right), \tag{H-ODE-SC}$$

see Appendix D.1. The equation (H-ODE-SC) can be seen as a combination between Polyak's ODE, (A-ODE-SC) and the ODE for Newton's method.

Similarly to the convex case, notice that (H-ODE-SC) can be seen as the high-resolution equation from Shi et al. [2018] with the highest parameter value $\frac{1}{\sqrt{L}}$. Using a Lyapunov analysis, we will show that the same Lyapunov function of (A-ODE-SC) decreases faster along (1st-ODE-SC). The asymptotic exponential rates are retrieved in the continuous and discrete setting, Proposition D.1. In addition, rewriting (H-ODE-SC) as a first order system (SC-Nest) allows us to derive Nesterov's method using an explicit discretization with a time step $h = \frac{1}{\sqrt{L}}$, Proposition 5.3, and to extend the Lyapunov analysis in the perturbed case.

Definition 5.1. *Let $h_k > 0$ be a decreasing sequence of small time step/learning rate. Take an explicit Euler method for* (1st-ODE-SC) *evaluated at y_k, defined below, and with $h_k\sqrt{\mu}$ replaced by $\lambda_h = \frac{h_k\sqrt{\mu}}{1+h_k\sqrt{\mu}}$*

$$\begin{cases} x_{k+1} - x_k = \lambda_h(v_k - x_k) - \frac{h_k}{\sqrt{L}}\nabla f(y_k), \\ v_{k+1} - v_k = \lambda_h(x_k - v_k) - \frac{h_k}{\sqrt{\mu}}\nabla f(y_k) \\ y_k = (1-\lambda_h)x_k + \lambda_h v_k, \qquad \lambda_h = \frac{h_k\sqrt{\mu}}{1+h_k\sqrt{\mu}}. \end{cases} \tag{FE-SC}$$

Remark 5.2. *As in the convex case, to obtain Nesterov's method, we need to evaluate the gradient at y_k, which is a perturbation of x_k. In addition, in the strongly convex case, we perturb by $\sqrt{\mu}$.*

Proposition 5.3. *The discretization of* (1st-ODE-SC) *given by* (FE-SC) *with* $h = 1/\sqrt{L}$ *is equivalent to the standard Nesterov's method* (SC-Nest).

5.2 Perturbed gradient: discrete time

Replacing gradients with $\tilde{\nabla} f = \nabla f + e$, the discretization of (1st-ODE-SC), (FE-SC), with a time step h_k, becomes

$$\begin{cases} x_{k+1} - x_k = \lambda_h(v_k - x_k) - \frac{h_k}{\sqrt{L}}(\nabla f(y_k) + e_k), \\ v_{k+1} - v_k = \lambda_h(x_k - v_k) - \frac{h_k}{\sqrt{\mu}}(\nabla f(y_k) + e_k), \\ y_k = (1-\lambda_h)x_k + \lambda_h v_k, \qquad \lambda_h = \frac{h_k\sqrt{\mu}}{1+h_k\sqrt{\mu}}. \end{cases} \tag{Per-FE-SC}$$

Definition 5.4. *Define the continuous time Lyapunov function,* $E^{ac,sc}$, *by*

$$E^{ac,sc}(x, v) = f(x) - f^* + \frac{\mu}{2}|v - x^*|^2, \tag{27}$$

and the discrete in time Lyapunov function by

$$E_k^{ac,sc} = E^{ac,sc}(x_k, v_k) = f(x_k) - f^* + \frac{\mu}{2}|v_k - x^*|^2. \tag{28}$$

In Appendix D.2, we show that $E^{ac,sc}$ is, indeed, a Lyapunov function for (1st-ODE-SC). Now, in the next proposition, we give the dissipation of $E^{ac,sc}$ along (Per-FE-SC).

Proposition 5.5. *Let* x_k, v_k *be two sequences generated by the scheme* (Per-FE-SC) *with initial condition* (x_0, v_0). *Suppose that* $h_k \le \frac{1}{\sqrt{L}}$, *then*

$$E_{k+1}^{ac,sc} \le (1 - h_k\sqrt{\mu})E_k^{ac,sc} + h_k\beta_k, \tag{29}$$

where $\beta_k = 2h_k\left\langle \nabla f(y_k) + \frac{e_k}{2}, e_k \right\rangle - \left\langle \sqrt{\mu}(x_k - y_k + v_k - x^*) - \frac{1}{\sqrt{L}}\nabla f(y_k), e_k \right\rangle$.

5.3 Variable time step and convergence in expectation

Now we consider the case where e_k satisfies (2). From Proposition 5.5, since $\mathbb{E}[\beta_k] = h_k^2\sigma^2$, we have

$$\mathbb{E}[E_{k+1}^{ac,sc}] \le (1 - \sqrt{\mu}h_k)E_k^{ac,sc} + h_k^2\sigma^2.$$

We are in the case where $r_{E^{ac,sc}} = \sqrt{\mu} > 0$, $L_{E^{ac,sc}} g_2^2 = 2$, then Proposition 2.7 gives

Proposition 5.6. *If*

$$h_k^{ac,sc} := \frac{2}{\sqrt{\mu}(k + (\alpha^{ac,sc}E_0^{ac,sc})^{-1})}, \qquad \alpha^{ac,sc} := \frac{\mu}{4\sigma^2},$$

then, assuming that $h_0^{ac,sc} \le \frac{1}{\sqrt{L}}$, *the following holds:*

$$\mathbb{E}[E^{ac,sc}(x_k)] \le \frac{4\sigma^2}{\mu k + 4\sigma^2 E_0^{ac,sc-1}}.$$

Remark 5.7. *Observe that, as in the accelerated convex case, the rate is independent of the smoothness parameter* L.

6 Conclusions

In this work, we obtained a modified algorithm for stochastic gradient descent, based on a continuous time approach. In both the convex and strongly convex cases, we began with an ODE, which, when discretized with a constant learning rate, led to the Nesterov's method in the full gradient case. However, in the stochastic gradient case, allows for a variable time step, led to an accelerated algorithm for SGD. We proved a convergence rate for the last iterate of the algorithm with an improved rate constant compared to the prior results. The rate constant depends on the variance of the stochastic gradient, but not on the L-smoothness of the gradient.

The proof technique relied heavily on continuous time analysis and abstract Lyapunov function analysis. We chose a first order system of ODEs which could be discretized with constant learning rate to recover Nesterov's method. Then, switching to the stochastic gradient case, we performed the abstract Lyapunov analysis, which gave convergence rates dependent on an abstract Lyapunov function. We then applied these abstract results to the Lyapunov functions and ODEs corresponding to Nesterov's method, to obtain a rate. The resulting proofs were more organized and concise that a direct analysis would have been. In addition, the abstract Lyapunov function analysis could be used to obtain rates for other stochastic algorithms, for example saddle point problems.

The convergence rate depends on the initial value of the Lyapunov function. This means that any time the value of the Lyapunov function is below the estimate, we can reinitialize to obtain a stronger estimate of the convergence rate. In the worst case, it may be that the estimate is not improved, but in practice, we can bootstrap the estimate.

This work can be extended in the following ways. Since we use a first order system to represent our ODE, the analysis can be adapted without much difficulty to the non-smooth case. Finally, the assumptions on the stochastic gradients (2) could be relaxed: for example, to cover the x dependent variance case for finite sums.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-18-1-0167 (A.O.).

References

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

F. Alvarez, H. Attouch, J. Bolte, and P. Redont. A second-order gradient-like dissipative dynamical system with hessian-driven damping.-application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–780, 2002.

H. Attouch, J. Peypouquet, and P. Redont. Fast convex optimization via inertial dynamics with hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

S. Bubeck. Convex Optimization: Algorithms and Complexity. *ArXiv e-prints*, May 2014.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695, 2015.

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

A. Iserles. *A first course in the numerical analysis of differential equations.* Number 44. Cambridge university press, 2009.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 545–604. PMLR, 06–09 Jul 2018.

P. Jain, D. Nagaraj, and P. Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Proceedings of the Annual Conference On Learning Theory (COLT)*, pages 1752–1755, 2019.

R. Kidambi, P. Netrapalli, P. Jain, and S. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems 28*, pages 2845–2853. Curran Associates, Inc., 2015.

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems 28*, pages 3384–3392. Curran Associates, Inc., 2015.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

A. M. Oberman and M. Prazeres. Stochastic Gradient Descent with Polyak's Learning Rate. *arXiv e-prints*, art. arXiv:1903.08688, Mar 2019.

B. Oksendal. *Stochastic differential equations: an introduction with applications.* Springer Science & Business Media, 2013.

G. A. Pavliotis. *Stochastic processes and applications.* Springer, 2016.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

X. Qian, P. Richtarik, R. Gower, A. Sailanbayev, N. Loizou, and E. Shulgin. Sgd with arbitrary sampling: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209, 2019.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1571–1578, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

M. Schmidt, N. L. Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 1458–1466, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont. Integration methods and accelerated optimization algorithms. *arXiv preprint arXiv:1702.06751*, 2017.

N. Serhat Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar. Robust Accelerated Gradient Methods for Smooth Strongly Convex Functions. *arXiv e-prints*, art. arXiv:1805.10579, May 2019.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, page 201614734, 2016.

A. Wilson, L. Mackey, and A. Wibisono. Accelerating Rescaled Gradient Descent: Fast Optimization of Smooth Functions. *arXiv e-prints*, art. arXiv:1902.08825, Feb 2019.

A. C. Wilson, B. Recht, and M. I. Jordan. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.