

A LYAPUNOV ANALYSIS FOR ACCELERATED GRADIENT METHODS: FROM DETERMINISTIC TO STOCHASTIC CASE

MAXIME LABORDE AND ADAM M. OBERMAN

ABSTRACT. The article [SBC14] 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 if a Hessian damping term is added to the ODE from [SBC14], then Nesterov’s method arises as a straightforward discretization of the 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 which allows the full gradient case to be considered as a special case of the stochastic case. The result is a unified approach to convex acceleration in both continuous and discrete time and in both the stochastic and full gradient cases.

1. Introduction

In [SBC14], Su, Boyd and Candés made the connection between Nesterov’s accelerated gradient descent method and a second order differential equation. The goal of the approach appeared to be to develop insight into Nesterov’s algorithm, possibly leading to new optimization algorithms. This work resulted in a renewed interest in continuous time approach, for example in [WWJ16, WRJ16] and [WMW19] a Lyapunov analysis is done to recover the optimal rate of convergence. Continuous time analysis also appears in [FB15], [LRP16], and [KBB15], among many other recent works. Of course, the Lyapunov approach to proving convergence rates appears widely in optimization, for example, see [BT09] for FISTA.

So far there is less work on continuous time approaches to stochastic optimization. Stochastic Gradient Descent (SGD) [RM51] is a widely used optimization algorithm due to its ubiquitous use in machine learning [Bot91] [BCN16]. Convergence rates are available in a wide setting [LJSB12, BCN16, QRG+19]. When SGD is combined with momentum [Pol64, Nes13] empirical performance is improved, but this improvement is not always theoretically established [KNJK18]. The optimal convergence rate for SGD in the smooth, strongly convex case is order $1/k$. In the convex, nonsmooth case, the optimal rate goes down to $O(1/\sqrt{k})$ [NJLS09] or [Bub14]. Accelerated versions of stochastic gradient descent algorithms are comparatively more recent: they appear in [LMH15] as well as in [FGKS15] and [JKK+18]. A direct acceleration method with a connection to Nesterov’s method can be found in [AZ17]. For the continuous case, going from ODEs and Lyapunov analysis to a perturbed ODEs was done in [APR16] with results for accelerated gradient descent in continuous time.

Outline of approach. One goal of the continuous time approach is to take advantage of the Lyapunov function approach to obtaining convergence rates for differential equations, with the hope that an explicit discretization of the differential equation leads to an algorithm which also decreases the Lyapunov function. We present a general Lyapunov function approach which allows us to go from continuous time to discrete time, in both the full gradient and stochastic gradient setting. The discrete time result follows from the continuous time one by enforcing a restriction on the time step (learning rate), as in (CFL) below. Since we use a first order system to represent our ODE, the analysis can also be adapted to the non-smooth case.

The abstract analysis applies in particular to each of the cases: continuous time/algorithm, accelerated/standard gradient descent, convex/strongly convex, and full gradients/stochastic gradients. In each case, we combine a Lyapunov function and a differential equation/finite difference equation. There is a fairly systematic way to go from each case, which we make an effort to make clear. We use the same Lyapunov function $E(t, z)$ to go from continuous time to the algorithm. Going from full gradients to perturbed gradients requires adding a second term, $I$, to the Lyapunov function: this term satisfies an easily solved ODE in the continuous time case, and an easily solved recursion equation in the discrete time case.
We extend the analysis to the case of stochastic gradients, by first analyzing what we call the “perturbed” gradient case. We write

\[ \tilde{\nabla} f(x) = \nabla f(x) + \epsilon, \]

where \( \epsilon \) is an error term. We first perform the analysis in an abstract setting with time step/learning rate \( h_k \). The key step is to obtain the following inequality on a rate generating Lyapunov function

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

here \( r_E \geq 0 \) is a rate constant coming from the Lyapunov function, which is zero in convex case, and \( > 0 \) in the strongly convex case. The term \( \beta_k \) depends on the error \( \epsilon_k \) from (1), it is zero in the full gradient case. The algorithms involve no averaging of previous values.

When we consider the stochastic case, we take expectations in (1): \( E[\beta_k] \) is proportional to \( h_k \sigma^2 \), where \( \sigma^2 \) is the variance of the perturbation of the gradient. After taking expectations in (1), we establish convergence rates for standard and accelerated SGD in both the convex and strongly convex cases. In the strongly convex case, we obtain an algorithm which would correspond to an adaptive time step version of Nesterov’s method, with stochastic gradients, and time step of order \( 1/k \). The algorithm recovers the \( 1/k \) rate of convergence of other SGD algorithms, Propositions 4.8 and 6.16.

In the convex case, we obtain an algorithm which corresponds to Nesterov’s algorithm in the convex case, but with a scheduled learning rate. The learning rate schedule we obtain is different from existing ones: it corresponds to Nesterov’s algorithm with learning rate with \( h_k = 1/k^{3/4+\epsilon} \). The convergence rate is \( 1/k^{1/2-2\epsilon} \), for any \( \epsilon > 0 \), Proposition 5.19. A summary of our asymptotic results in the convex and strongly convex stochastic gradient cases can be found in Figure 1.

| Convex | time step \( h_k \) | rate of \( E[f(x_k)] - f^* \) |
|--------|-----------------|-------------------------------|
| SGD    | \( k^{-\alpha}, \alpha \in (2/3, 1] \) | \( \mathcal{O}(k^{\alpha-1}) \) if \( \alpha \in (2/3, 1) \), \( \mathcal{O}(\ln(k)^{-1}) \) if \( \alpha = 1 \). |
| Acc. SGD | \( k^{-\alpha}, \alpha \in (3/4, 1] \) | \( \mathcal{O}(k^{2(\alpha-1)}) \) if \( \alpha \in (3/4, 1) \), \( \mathcal{O}(\ln(k)^{-2}) \) if \( \alpha = 1 \). |
| \( \mu \)-Strongly Convex | SGD | \( \frac{2}{\mu k + 2(C_f+1)\sigma^2 E_0} \) | \( \frac{2(C_f+1)\sigma^2}{\mu k + 2(C_f+1)\sigma^2 E_0} \)
| Acc. SGD | \( \frac{2}{\mu k + 2\sigma^2 E_0} \) | \( \frac{2\sigma^2}{\mu k + 2\sigma^2 E_0} \)

**Figure 1.** Convergence rates in expectation of \( f(x_k) - f^* \): where \( h_k \) is a non constant learning rate and the error is such that \( E[\epsilon_k] = 0 \) and \( \text{Var}(\epsilon_k) = \sigma^2 \). \( E_0 \) represents the Lyapunov function at initial time and \( C_f := \frac{L}{\mu} \) denotes the condition number of the \( \mu \)-strongly convex, \( L \)-smooth function \( f \).

In addition, we obtain non-asymptotic error estimates in finite time, which measure the deviation from the rate in the full gradient case: the correction is a sum or integral of the error terms. By summing the errors, we get a result in the constant time step case with assuming decreasing error sizes. In the case of constant time steps, we need to assume that \( |\epsilon_i| \) is decreasing fast enough, however we do not take expectations of the error or make mean zero assumptions. In particular, we cover the biased gradient case.

**Remark 1.1** (interpretation of the results). One motivation/application for this work is the use of accelerated SGD in deep learning. The practical implementation uses Polyak’s momentum, and decrease time step after several epochs (a fixed number of time steps). In the first phase: see behaviour consistent with accelerated gradient descent, then the noise dominates, and the learning rate is decreased. Our analysis covers both phases: we show that after a finite number of iterations with constant learning rate, the decrease in \( f \) gap is consistent with the accelerated rate plus a correction due to the error in the gradients. Second, we show that we can obtain the asymptotic \( 1/k \) rate in the strongly convex case using accelerated gradient descent with decreasing time step.
Discussion of asymptotic error rates in the perturbed case. In the perturbed gradient case, we consider the case of decreasing errors $e_i$ but fixed learning rates. We show that if the size of the errors decreases quickly enough, we can recover the asymptotic convergence rates of the unperturbed gradient case. A sufficient condition is that

$$|e_k| \sim \frac{1}{k^\alpha}$$

with $\alpha > 1$ for perturbed gradient descent and $\alpha > 2$ for perturbed accelerated gradient descent. Thus, in order for the accelerated rates to be obtained, the perturbations of the gradients need to go to zero faster in the accelerated case.

Under these assumptions by introducing a perturbed Lyapunov function to compensate the effect of the error, inspired by the continuous time analysis of [APR16, ACPR18], we are able to obtain the same accelerated rate of convergence as in the deterministic case, Corollary 5.10 and Corollary 5.15, see Figure 2.

| Continuous | SGD | $\int_0^t |e(t)| < +\infty$ | $O(1/t)$ |
|------------|-----|-----------------------------|----------|
| Acc. SGD   | $\int_0^t |f(e(t))| < +\infty$ | $O(1/t^2)$ |
| Discrete   | SGD | $\sum_{k=0}^{\infty} |e_k| < +\infty$ | $O(1/k)$ |
| Acc. SGD   | $\sum_{k=0}^{\infty} k|e_k| < +\infty$ | $O(1/k^2)$ |

Figure 2. Convergence rates in the convex case: when $h$ is the learning rate such that $0 < h \leq 1/L$ in the gradient descent case and $0 < h \leq 1/\sqrt{L}$ in the accelerated case.

In addition, if we assume that $f$ is strongly convex, we can introduce another Lyapunov function in order to take advantage of the gap in the dissipation along the dynamics. This will imply an accelerated estimate on the decrease of the norm of $|\nabla f(y_k)|^2$, Corollary 5.18. Similarly, in the strongly convex case, the perturbation needs to go to zero faster in the accelerated case compared to the non-accelerated case. In the discrete case, $|e_k|$ should decrease as $\exp(-\mu k)$ in the gradient descent case, whereas, for the accelerated method, $|e_k|$ should decrease as $\exp(-\sqrt{\mu} k)$, which is faster in the relevant case $\mu < 1$. Then, under these assumptions, the same accelerated rate as in the deterministic case is achieved, see Corollary 6.10 in the continuous case and Corollary 6.13 in the discrete case. A summary of the results in the strongly convex case is given in Figure 3.

| Continuous | SGD | $\int_0^t \exp(\mu t)|e(t)| < +\infty$ | $O(\exp(-\mu t))$ |
|------------|-----|-------------------------------|----------------|
| Acc. SGD   | $\int_0^t \exp(\sqrt{\mu} t)|e(t)| < +\infty$ | $O(\exp(-\sqrt{\mu} t))$ |
| Discrete   | SGD | $\sum_{k=0}^{\infty} (1-h\mu)^{-k} |e_k| < +\infty$ | $O((1-h\mu)^k)$ |
| Acc. SGD   | $\sum_{k=0}^{\infty} (1-h\sqrt{\mu})^{-k} |e_k| < +\infty$ | $O((1-h\sqrt{\mu})^k)$ |

Figure 3. Convergence rates in the strongly convex case: when $h$ is the learning rate such that $0 < h \leq 2/(L + \mu)$ in the gradient descent case and $0 < h \leq 1/\sqrt{L}$ in the accelerated case.

Remark 1.2 (Applications of abstract perturbed gradient). The perturbation of the gradient can be abstract. In particular

1. Can have a stochastic gradient where the error is a mini-batch gradient. In order to convert from the mini-batch gradient to an abstract error, we require an estimate of the mean and variance of $e = \nabla f(x) - \nabla f(x)$ the minibatch error.

2. Can include the case where the error includes variance reduction [JZ13]. The correction by a snapshot of the full gradient at a snapshot location, which is updated every $m$ iterations, $e_k = \nabla f(y) - \nabla f(y) - (\nabla f(y_k) - \nabla f(y_k))$.

The combination of variance reduction and momentum was discussed in [AZ17].
(3) The error $e_k$ can also represent the difference between Nesterov’s method and Polyak’s momentum method, which comes from the error where $\nabla f(y_k)$ is replaced by $\nabla f(x_k)$. This difference can just be absorbed into the error in the gradient,
\[
\nabla f(y_k) + e_k = \nabla f(x_k) + \tilde{e}_k, \quad \tilde{e}_k = e_k + O(x_k - y_k).
\]

But in the early phase of the algorithm, where we compare to the accelerated gradient method rate, the finite time error estimate can simply include this term. So basically in Phase 1, Polyak SGD is not too different from Nesterov SGD.

1.1. Other related work. Our goal here is to obtain rates for optimization algorithms using a continuous time perspective. The idea put forward by many authors, notably [SBC14, WWJ16, WRJ16], is that the convergence proofs of accelerated algorithms do not give enough insight, and that building the connection with continuous time methods may bring the insight needed to more easily develop new algorithms.

However, continuous time approaches to optimization have been around for a long time. Polyak’s method [Pol64] is related to successive over relaxation for linear equations [Var57] which were initially used to accelerate solutions of linear partial differential equations [You54]. Continuous time interpretations of Newton’s method can be found in [Pol87] or [AABR02], and of mirror descent [NY83] can be found in [B+15, XWG18a, XWG18b].

Indeed, continuous time approach to solve first order convex optimization is a very well-developed theory and there exists a huge literature on the study of Nesterov’s method by continuous time and ODE arguments. The continuous time analysis can offer a very good framework for optimization and may lead to a better understanding of algorithms. However the project of continuous analysis to discrete is still not clearly defined, despite the recent work by [WWJ16, WRJ16, WMW19]. Related work studying discretizations of ordinary and partial differential equations which respect Lyapunov functions can be found in [SH96]: although in this case the discretizations are typically implicit, so they require further solution of equations to obtain an algorithm.

1.2. Notations and organization. Throughout the paper, denote $x_\ast = \arg\min_x f(x)$ and $-\infty < f^\ast := f(x_\ast) = \min_x f(x)$. We say that a function is $L$-smooth if $f : \mathbb{R}^d \to \mathbb{R}$ satisfies,
\[
f(y) - f(x) + \nabla f(x) \cdot (y - x) \leq \frac{L}{2} |x - y|^2,
\]
In addition, we consider also the class of $\mu$-strongly convex functions, i.e. $f - \frac{\mu}{2} |\cdot|^2$ is convex,
\[
f(x) + \nabla f(x) \cdot (y - x) \leq f(y) - \frac{\mu}{2} |y - x|^2,
\]
Combining these properties, we get, in particular, for $L$-smooth convex functions, we have
\[
\frac{1}{2L} |\nabla f(x)|^2 \leq f(x) - f^\ast.
\]
The condition number of a $\mu$-strongly convex, $L$-smooth function, $f$, is denoted $C_f$ and defined by $C_f := \frac{L}{\mu}$.

The paper is organized as follows. In section 2, we introduce second order ODEs with Hessian damping, (H-ODE) and (H-ODE-SC), and especially their associated first order systems, (1st-ODE) and (1st-ODE-SC). We show that Nesterov’s schemes derive from an explicit discretization of these systems in both convex and strongly convex cases. Section 3 is devoted to the presentation of an abstract Lyapunov analysis in order to obtain rates for optimization algorithms using a continuous time perspective. Then we extend this analysis to the perturbed case where the gradient is replaced by $\tilde{\nabla} f$, (1). In this case, providing that the error term decreases fast enough, we show an abstract convergence result with the same rate as in the unperturbed case. To conclude this section, we consider the case with a variable time step and error with fixed variance. Under these assumption, we provide abstract convergence results in expectation. Then, we apply this abstract analysis in Section 4 to the gradient descent case for convex and strongly convex functions. Finally, in sections 5 and 6, we extend this framework to the special case of accelerated gradient methods, applying it to the first order systems (1st-ODE) and (1st-ODE-SC). In particular, in the unperturbed case, we recover the usual optimal rates int the continuous and discrete setting. In addition, we give an accelerated rate for the gradient taking advantage of the gap in the dissipation of the Lyapunov functions. Concerning, the
perturbed case, we present a slightly extension of our abstract setting to recover the optimal rates from the unperturbed case. In addition, we give an accelerated rate in the stochastic case.

2. ODEs and derivation of Nesterov’s methods

In this section, we introduce, in both convex and strongly convex cases, a second order ODE which is a perturbation of the one introduced by Su, Boyd and Candés [SBC14], with a Hessian damping. However, the analysis of this ODE needs to assume $f$ to be twice differentiable. Nethertheless, we introduce a first order system, equivalent to the second order ODE when $f$ is smooth. We prove that Nesterov’s methods derive from an explicit discretization of our first order systems. One advantage of the first order system is that it allows to deal with non-smooth $f$. Moreover, to go to the stochastic gradient case, we really need a first order system for continuous time. Indeed, in [SBC14] and [SDJS18], a term in $\dot{x}$ appears in the Lyapunov function, and then it is not clear how to extend this to the stochastic case.

We start to expose it in the convex case and then we deal with the strongly convex case.

2.1. Convex case. In [SBC14] Su, Boyd and Candés made a connection between Nesterov’s method for a convex, $L$-smooth function, $f$, and the second order ordinary differential equation (ODE)

\begin{equation}
\ddot{x} + \frac{3}{t} \dot{x} + \nabla f(x) = 0
\end{equation}

which can be written as the first order system

\begin{equation}
\begin{cases}
\dot{x} = \frac{2}{t}(v - x) \\
\dot{v} = -\frac{t}{2} \nabla f(x)
\end{cases}
\end{equation}

Our starting point is the following system of first order ODEs, which is a slight modification of (2.1)

\begin{equation}
\begin{cases}
\dot{x} = \frac{2}{t}(v - x) - \frac{1}{\sqrt{L}} \nabla f(x) \\
\dot{v} = -\frac{t}{2} \nabla f(x)
\end{cases}
\end{equation}

The system (1st-ODE) is equivalent to the following second order differential equation with a Hessian damping

\begin{equation}
\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)
\end{equation}

**Derivation of (H-ODE).** Solve for $v$ in the first line of (1st-ODE)

$$v = \frac{t}{2} (\dot{x} + \frac{1}{\sqrt{L}} \nabla f(x)) + x$$

differentiate to obtain

$$\dot{v} = \frac{1}{2} (\ddot{x} + \frac{1}{\sqrt{L}} \nabla f(x)) + \frac{t}{2} (\ddot{x} + \frac{1}{\sqrt{L}} D^2 f(x) \cdot \dot{x}) + \dot{x}.$$  

Insert into the second line of (1st-ODE)

$$\frac{1}{2} (\ddot{x} + \frac{1}{\sqrt{L}} \nabla f(x)) + \frac{t}{2} (\ddot{x} + \frac{1}{\sqrt{L}} D^2 f(x) \cdot \dot{x}) + \dot{x} = -\frac{t}{2} \nabla f(x).$$

Simplify to obtain (H-ODE). □

We will show below that solutions of (1st-ODE) decrease the same Lyapunov function faster than solutions of (A-ODE). Interestingly, it leads to the second order ODE (H-ODE), which has an additional Hessian damping term with coefficient $1/\sqrt{L}$. This Hessian damping term combines continuous time Newton method and accelerated dynamic (A-ODE). Notice that (H-ODE) is a perturbation of (A-ODE) of order $1/\sqrt{L}$, and the perturbation goes to zero as $L \to \infty$. Similar ODEs have been studied by [AABR02], they have been shown to accelerate gradient descent in continuous time in [APR16].

In the ODE (H-ODE) the coefficient of $\dot{x}$ is damped by the Hessian and then depends on $x$. In addition the coefficient of $\nabla f(x)$ is perturbed by $\frac{1}{\sqrt{L}}$ which goes to zero asymptotically. This equation corresponds at a
first order perturbation $O \left( h = \frac{1}{\sqrt{L}} \right)$ of (A-ODE). Recently in [SDJS18], Shi, Du, Jordan and Su introduced a family of second order differential equations called high-resolution differential equation. This equation is derived from Nesterov’s method using terms of order $O(1)$ and $O(h)$ instead of only terms of order $O(1)$ to derive (A-ODE). In this context, (H-ODE) corresponds to the high-resolution equation with the parameter $\frac{1}{\sqrt{L}}$.

However, we demonstrate below that the first order system (1st-ODE) is more amenable to analysis, allowing for short clean proofs which generalize to the perturbed gradient case. The system (1st-ODE) can be discretized to recover Nesterov’s method using an explicit discretization with a time step $h = \frac{1}{\sqrt{L}}$, Proposition 2.2. By a Lyapunov analysis, we recover the usual optimal rates, in both continuous and discrete cases, and, in addition, we obtain an extra gap in the dissipation of the Lyapunov function, Proposition 5.2, which gives us an estimate on the decrease of $|\nabla f(x)|^2$, Corollary 5.5.

Nesterov’s method for a convex, $L$-smooth function, $f$, can be written as [Nes13, Section 2.2]

(C-Nest)

\[
\begin{align*}
\begin{cases}
x_{k+1} &= y_k - \frac{1}{L} \nabla f(y_k) \\
y_{k+1} &= x_{k+1} + \frac{k}{k+3} (x_{k+1} - x_k)
\end{cases}
\]

Definition 2.1. Let $h > 0$ be a given small time step/learning rate and let $t_k = h(k + 2)$. 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,

(FE-C)

\[
\begin{align*}
\begin{cases}
x_{k+1} - x_k &= \frac{2h}{L} (v_k - x_k) - \frac{h}{\sqrt{L}} \nabla f(y_k), \\
v_{k+1} - v_k &= -\frac{ht_k}{2} \nabla f(y_k), \\
y_k &= \left(1 - \frac{3}{k+2}\right)x_k + \frac{2}{k+2} v_k.
\end{cases}
\]

Then the following result holds.

Proposition 2.2. The discretization of (1st-ODE) given by (FE-C) with $h = 1/\sqrt{L}$ is equivalent to the standard Nesterov’s method (C-Nest).

Proof. The system (FE-C) with $h = 1/\sqrt{L}$ and $t_k = h(k + 2)$ becomes

\[
\begin{align*}
\begin{cases}
x_{k+1} - x_k &= \frac{2h}{L} (v_k - x_k) - \frac{h}{\sqrt{L}} \nabla f(y_k) \\
v_{k+1} - v_k &= -\frac{ht_k}{2} \nabla f(y_k)
\end{cases}
\]

Eliminate the variable $v_k$ using the definition of $y_k$ in (FE-C) to obtain (C-Nest). \hfill \Box

2.2. Strongly convex case. In the case of a $\mu$-strongly convex function, we are interested to another second order differential equation with a Hessian damping. For a $\mu$-strongly, convex function $f$, consider the first order system

(1st-ODE-SC)

\[
\begin{align*}
\begin{cases}
\dot{x} &= \sqrt{\mu}(v - x) - \frac{1}{\sqrt{L}} \nabla f(x), \\
\dot{v} &= \sqrt{\mu}(x - v) - \frac{1}{\sqrt{L}} \nabla f(x).
\end{cases}
\]

which is equivalent to the second order equation with Hessian damping for a smooth $f$

(H-ODE)

\[
\ddot{x} + 2 \sqrt{\mu} \dot{x} + \nabla f(x) = -\frac{1}{\sqrt{L}} \left(D^2 f(x) \cdot \dot{x} + \sqrt{\mu} \nabla f(x) \right).
\]

Equivalence between (1st-ODE-SC) and (H-ODE-SC). Solve for $v$ in the first line of (1st-ODE-SC)

\[
v = \frac{1}{\sqrt{\mu}} (\dot{x} + \frac{1}{\sqrt{L}} \nabla f(x)) + x
\]

differentiate to obtain

\[
\dot{v} = \frac{1}{\sqrt{\mu}} (\ddot{x} + \frac{1}{\sqrt{L}} D^2 f(x) \cdot \dot{x} + \dot{x}).
\]

Insert into the second line of (1st-ODE-SC)

\[
\frac{1}{\sqrt{\mu}} (\ddot{x} + \frac{1}{\sqrt{L}} D^2 f(x) \cdot \dot{x} + \dot{x}) = -\dot{x} - \left(\frac{1}{\sqrt{L}} + \frac{1}{\sqrt{\mu}}\right) \nabla f(x).
\]
Simplify to obtain (H-ODE-SC).

The equation (H-ODE-SC) can be seen as a combination between Polyak’s ODE (A-ODE-SC)

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

which is an accelerates gradient method when \( f \) is quadratic see [SRBd17], 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 [SDJS18] with the highest parameter value \( \frac{\sqrt{\mu}}{\mu} \). Using a Lyapunov analysis, we will show in Section 6 that the same Lyapunov function of (A-ODE-SC) decreases faster along (1st-ODE-SC) and allows an acceleration in the decrease of the gradient. The asymptotic exponential rates are retrieved in the continuous and discrete setting, Proposition 6.2. In addition, rewriting (H-ODE-SC) as a first order a system (SC-Nest) permits to derive Nesterov’s method using an explicit discretization with a time step \( h = \frac{1}{\sqrt{\lambda}} \), Proposition 2.5, and to extend the Lyapunov analysis in the perturbed case.

Nesterov’s method in the strongly convex case can be written as follows.

(SC-Nest) \[
\begin{align*}
    x_{k+1} &= y_k - \frac{1}{L} \nabla f(y_k) \\
    y_{k+1} &= x_{k+1} + \frac{1}{1+\sqrt{C_f^{-1}}} (x_{k+1} - x_k)
\end{align*}
\]

**Definition 2.3.** Let \( h > 0 \) be a small time step, and take an explicit Euler method for (1st-ODE-SC) evaluated at \( y_k \), defined below, and with \( h\sqrt{\mu} \) replaced by \( \lambda_h = \frac{h\sqrt{\mu}}{1+h\sqrt{\mu}} \),

(FE-SC) \[
\begin{align*}
    x_{k+1} - x_k &= \lambda_h (v_k - x_k) - \frac{h}{\sqrt{L}} \nabla f(y_k), \\
    v_{k+1} - v_k &= \lambda_h (x_k - v_k) - \frac{h}{\sqrt{\mu}} \nabla f(y_k) \\
    y_k &= (1 - \lambda_h)x_k + \lambda_h v_k, \quad \lambda_h = \frac{h\sqrt{\mu}}{1+h\sqrt{\mu}}
\end{align*}
\]

**Remark 2.4.** 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 \( \sqrt{\mu} \).

**Proposition 2.5.** 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).

**Proof.** (FE-SC) with \( h = 1/\sqrt{L} \) becomes \[
\begin{align*}
    x_{k+1} - x_k &= \frac{\sqrt{C_f^{-1}}}{1+\sqrt{C_f^{-1}}} (v_k - x_k) - \frac{1}{L} \nabla f(y_k), \\
    v_{k+1} - v_k &= \frac{\sqrt{C_f^{-1}}}{1+\sqrt{C_f^{-1}}} (x_k - v_k) - \frac{1}{\sqrt{\mu}} \nabla f(y_k)
\end{align*}
\]

Eliminate the variable \( v_k \) using the definition of \( y_k \) to obtain (SC-Nest). \( \square \)

### 3. Abstract Lyapunov Analysis: Going from Continuous to Discrete Time

The advantage of continuous time Lyapunov analysis to obtain convergence rates is that there is no time step, so the number of terms we need to control is simpler. By presenting our ODEs as first order systems, we use explicit gradient calculations, and avoid the tedious substitution of terms which results from used the second order ODE. In addition, the Lyapunov functions are cleaner, since they involve only variables without time derivatives. The first order system approach becomes even more important in the stochastic case, where derivatives of error terms might otherwise appear. Finally, the first order system formulation allows the analysis to go through for nonsmooth objectives, although we do not pursue that here.

In this section we present theorems showing how to go from continuous time Lyapunov functions to discrete time, in both the full gradient and perturbed cases in the abstract setting.

We make a definition of the continuous to discrete problem and start exposing the problem in an abstract setting, then, we show how this framework can be extended to study a perturbed gradient descent (i.e. an error is made in the evaluation of the gradient). We discuss also how this framework can be adapt to deal
with accelerated gradient methods. To conclude this section, we present an abstract convergence rate for variable learning rates.

In the first subsection, we define the class of ODEs we consider and the associated Forward Euler methods. We define a generic Lyapunov function, which may depend on time, and provide conditions for the Lyapunov function to give a rate in continuous time, and then show how this rate extends to the forward Euler method, provided a restriction on the time step is satisfied.

In the subsequent subsection, we show how the same analysis can be extended to the case of perturbed gradients. Again we are in the abstract Lyapunov function setting. Our analysis shows that we can recover the rates corresponding to the full gradient case, provided that the error in the gradient decreases fast enough. Although this an unusual way to present the rates, presenting the results in this way gives a unified approach to the perturbed and full gradient cases. However, to conclude this section, we present an abstract convergence rate in the case where the learning rate is not constant anymore and the error has zero-mean and a fixed variance.

3.1. ODEs, Perturbed ODEs and discretizations. Consider an abstract ordinary differential equation, generated by the velocity field $g$ as follows.

**Definition 3.1.** Let $g(t, z, p)$ be $L_g$-Lipschitz continuous, and affine in the variable $p$,

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

Consider the ODE

\begin{equation}
(\text{ODE}) \quad \dot{z}(t) = g(t, z(t), \nabla f(z(t)))
\end{equation}

Let $e(t)$ be a perturbation of the gradient $\nabla f(z(t))$ as in (1). Consider the Perturbed ODE

\begin{equation}
(\text{PODE}) \quad \dot{z}(t) = g(t, z(t), \nabla f(z(t)) + e(t)),
\end{equation}

(PODE) 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 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 (SDE) [Oks13, Pav16], 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 3.2.** For a given time step (learning rate) $h \geq 0$, the forward Euler discretization of (ODE) corresponds to the sequence

\begin{equation}
(z_k)_{k \geq 0}, \quad z_{k+1} = z_k + hg(t_k, z_k, \nabla f(z_k)), \quad t_k = kh
\end{equation}

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

\begin{equation}
(z_k)_{k \geq 0}, \quad z_{k+1} = z_k + hg(t_k, z_k, \nabla f(z_k) + e_k), \quad t_k = kh
\end{equation}

The solution of (FE) or of (FEP) can be interpolated to be a function of time $z^h : [0, T) \rightarrow \mathbb{R}^n$ 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 [Ise09] that functions $z^h$ converge to $z(t)$ with error of order $h$, provided $h \leq 1/L_g$.

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

**Definition 3.3.** 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{equation}
(5) \quad \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{equation}

**Remark 3.4.** Definition 3.3 can be extended to consider nonegative time depending gap $a_E = a_E(t)$. This may appear especially in the convex case and the analysis below does not change.

Then, we can deduce the following rate in the continuous case.
Lemma 3.5. 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. \[
\frac{d}{dt} E(t, z(t)) = \partial_t E(t, z(t)) + \nabla E(t, z(t)) \cdot g(t, z(t), \nabla f(z(t))) \leq -r_E E(t, z(t)) - a_E |g(t, z(t), \nabla f(z(t)))|^2
\]
by assumption (3.3). Gronwall’s inequality gives the result. \( \square \)

In the discrete setting, we obtain

Lemma 3.6. Let \( z_k \) be the solution of the forward Euler method (FE) for (ODE). Let \( E \) be a rate generating Lyapunov function for (ODE) so that (3.3) holds. Suppose in addition that there exists \( L_E > 0 \) such that \( E \) satisfies,
\[
E(t_{k+1}, z_{k+1}) - E(t_k, z_k) \leq \partial_t E(t_k, z_k)(t_{k+1} - t_k) + \langle \nabla E(t_k, z_k), z_{k+1} - z_k \rangle + \frac{L_E}{2}|z_{k+1} - z_k|^2,
\]
Then
\[
E(t_{k+1}, z_{k+1}) \leq (1 - hr_E)E(t_k, z_k)
\]
provided
\[
(CFL) \quad h \leq \frac{2a_E}{L_E}
\]
In particular, choosing equality in (CFL) we have
\[
E(t_k, z_k) \leq \left(1 - \frac{2a_E r_E}{L_E}\right)^k E(0, z_0)
\]

Remark 3.7. Note that condition (3.6) is a generalization of the \( L \)-smoothness condition in space and is automatically satisfied in the case where the Lyapunov function does not depend on time by \( L_E \)-smoothness. Below in Lemma 4.1, we will see that this assumption is satisfied in the gradient descent case for convex and strongly convex functions.

Proof. Estimate \( E(t_{k+1}, z_{k+1}) - E(t_k, z_k) \) using (3.6) to get
\[
E(t_{k+1}, z_{k+1}) - E(t_k, z_k) \leq \partial_t E(t_k, z_k)(t_{k+1} - t_k) + \nabla E(t_k, z_k)(z_{k+1} - z_k) + \frac{L_E}{2}|z_{k+1} - z_k|^2
\]
\[
\leq -hr_E E(t_k, z_k) - ha_E |g(t_k, z_k, \nabla f(z_k))|^2 + \frac{h^2 L_E}{2}|g(t_k, z_k, \nabla f(z_k))|^2
\]
\[
\leq -hr_E E(t_k, z_k) - \left(ha_E - \frac{h^2 L_E}{2}\right)|g(t_k, z_k, \nabla f(z_k))|^2
\]
So apply (CFL) to get
\[
E(t_{k+1}, z_{k+1}) - E(t_k, z_k) \leq -hr_E E(t_k, z_k)
\]
which also gives (3.6). \( \square \)

3.3. Lyapunov analysis for the perturbed ODE. First, we compute the dissipation of a rate-generating Lyapunov function in the unperturbed case along (PODE).

Lemma 3.8. Let \( z \) be a solution of (PODE) and suppose \( E(t, z) \) is a rate-generating Lyapunov function for (ODE) which satisfies (3.6). Then
\[
\frac{d}{dt} E(t, z(t)) \leq -r_E E(t, z(t)) - a_E |g(t, z(t), \nabla f(z(t)))|^2 + \langle \nabla E(t, z), g_2(t, z(t))e(t) \rangle.
\]

Proof. Since \( z \) is solution of (PODE) and \( E \) satisfies (3.3),
\[
\frac{d}{dt} E(t, z(t)) = \partial_t E(t, z(t)) + \langle \nabla E(t, z(t)), g(t, z(t))\nabla f(z(t)) + e(t)\rangle
\]
\[
= \partial_t E(t, z(t)) + \langle \nabla E(t, z(t)), g(t, z(t))\nabla f(z(t)) \rangle + \langle \nabla E(t, z), g_2(t, z(t))e(t) \rangle
\]
\[
\leq -r_E E(t, z(t)) - a_E |g(t, z(t), \nabla f(z(t)))|^2 + \langle \nabla E(t, z), g_2(t, z(t))e(t) \rangle. \quad \square
\]
Observe that when we go from the unperturbed ODE (ODE) to the perturbed ODE (PODE), the additional term \( \langle \nabla E(s, z), g_2(s, z(t)) e(t) \rangle \) appears in the time derivative of \( E \) along (PODE). In this section we show how to add a term to the original Lyapunov function to obtain a Lyapunov function in the perturbed case.

In order to compensate for the additional term coming from the error \( e(t) \) we are motivated to define the perturbed Lyapunov function by
\[
\dot{E}(t, z) = E(t, z) + I(t, z(\cdot))
\]
where \( I(t, z(\cdot)) \) satisfies
\[
I'(t, z(\cdot)) = -\langle \nabla E(s, z(s)), g_2(s, z(s)) e(s) \rangle - r_E I(t, z(\cdot)), \quad I(0, z(\cdot)) = 0
\]
Note that unlike \( E \), \( I \) depends on the history of \( z \) and on \( e \), so we emphasize this with the notation. The preceding time dependent ODE is easily solved by standard methods. The solution is given by
\[
I(t, z(\cdot)) = -\exp(-r_E t) \int_0^t \exp(r_E s) \langle \nabla E(s, z(s)), g_2(s, z(s)) e(s) \rangle ds.
\]

**Definition 3.9.** Write
\[
J(t, z(\cdot)) = \int_0^t \exp(r_E s) \langle \nabla E(s, z(s)), g_2(s, z(s)) e(s) \rangle ds.
\]
so that \( J(t, z(\cdot)) = -\exp(r_E t) I(t, z(\cdot)) \). Define the perturbed Lyapunov function
\[
\tilde{E}(t, z(\cdot)) = E(t, z(t)) - \exp(-r_E t) J(t, z(\cdot))
\]

**Proposition 3.10.** Let \( z(t) \) be a solution of the perturbed ODE (PODE) and let \( E \) be a rate-generating Lyapunov function for the unperturbed ODE (ODE).

Then
\[
E(t, z(t)) \leq \exp(-r_E t)(E(0, z(0)) + J(t, z(\cdot))).
\]

**Proof.** We first establish
\[
\tilde{E}(t, z(t)) \leq \tilde{E}(0, z(0)) \exp(-r_E t)
\]
By assumption (3.3) and the calculation at the beginning of this section,
\[
\frac{d}{dt} \tilde{E}(t, z(t)) = \partial_t E(t, z) + \nabla E(t, z) \cdot g(t, z, \nabla f(z(t))) + I'(t, z(\cdot)) \leq -r_E E(t, z(t)) - a_E |g(t, z(t), \nabla f(z(t)))|^2 + \langle \nabla E(t, z), g_2(t, z(t)) e(t) \rangle + I'(t) \leq -r_E E(t, z(t)) - a_E |g(t, z(t), \nabla f(z(t)))|^2 - r_E I(t, z(\cdot)) \leq -r_E \tilde{E}(t, z(t)) - a_E |g(t, z(t), \nabla f(z(t)))|^2.
\]
Gronwall’s inequality completes the proof of (3.3).

From (3.3), using the definition of \( E \), and the fact that \( \tilde{E}(0, z(0)) = E(0, z(0)) \), we have
\[
\tilde{E}(t, z(\cdot)) = E(t, z(t)) - \exp(-r_E t) J(t, z(\cdot)) \leq \exp(-r_E t) E(0, z(0))
\]
which gives the second result.

**Corollary 3.11.** Under the assumptions of the previous proposition, for all \( t > 0 \), \( |\nabla E(t, z(t))| \) satisfies
\[
\sup_{0 \leq s \leq t} |\nabla E(s, z(s))| \leq M(t, z(\cdot)),
\]
where
\[
M(t) := 1 + 2L_E E(0, z(0)) \exp(-r_E t) + \exp(-r_E t) 4L_E^2 E(0, z(0)) \int_0^t |g_2(s, z(s)) e(s)| \exp \left( 2L_E \int_s^t |g_2(u, z(u)) e(u)| du \right) ds + 2L_E \exp(-r_E t) \int_0^t \exp(r_E s) |g_2(s, z(s)) e(s)| \exp \left( 2L_E \int_s^t |g_2(u, z(u)) e(u)| du \right) ds.
\]
and if \( e \) satisfies

\[
\int_0^{+\infty} \exp(r_E s) |g_2(s, z(s)) e(s)| \, ds < +\infty,
\]

then \( M(t, z(\cdot)) \) is bounded in \( L^\infty(\mathbb{R}_+) \) and

\[
E(t, z(t)) \leq \exp(-r_E t) \left( E(0, z(0)) + M(t, z(\cdot)) \int_0^t \exp(r_E s) |g_2(s, z(s)) e(s)| \, ds \right) = O(\exp(-r_E t)).
\]

**Proof.** From Proposition 3.10, we have

\[
\exp(r_E t) E(t, z(t)) \leq (E(0, z(0)) + J(t, z(\cdot))).
\]

Since for all \( t \), by (1.2),

\[
E(t, z(t)) \geq \frac{1}{2L_E} |\nabla E(t, z(t))|^2 \geq \frac{1}{2L_E} (|\nabla E(t, z(t))| - 1),
\]

and

\[
J(t, z(\cdot)) \leq \int_0^t \exp(r_E s) |\nabla E(s, z(s))||g_2(s, z(s)) e(s)| \, ds.
\]

Then,

\[
\exp(r_E t) |\nabla E(t, z(t))| \leq (\exp(r_E t) + 2L_E E(0, z(0))) + 2L_E \int_0^t \exp(r_E s) |\nabla E(s, z(s))||g_2(s, z(s)) e(s)| \, ds.
\]

Use Gronwall’s Lemma \( t \mapsto \exp(r_E t) |\nabla E(t, z(t))| \) to obtain the first part of the proof. We conclude noticing that all term in the right hand side are bounded when \( t \not> +\infty \) under the assumption (3.11). \( \square \)

### 3.4. Lyapunov analysis for the perturbed algorithm.

Now we consider the discrete case. As in the continuous case, the strategy is to see which extra terms arise in the dissipation of the original Lyapunov function along the perturbed equation, and then build an additional term into the perturbed Lyapunov function to cancel them out. The following lemma computes the excess term.

**Lemma 3.12.** Let \( z_k \) be the solution of (FEP). Suppose \( E \) is a rate-generating Lyapunov function for (ODE) which satisfies (3.6). Then

\[
E(t_{k+1}, z_{k+1}) - E(t_k, z_k) \leq -hr_E E(t_k, z_k) - h \left( a_E - \frac{L_E h}{2} \right) |g(t_k, z_k, \nabla f(z_k))|^2 + h \beta_k
\]

where \( \beta_k \) is defined by

\[
\beta_k := \langle \nabla E(t_k, z_k), g_2(t_k, z_k) e_k \rangle + L_E h \left( g(t_k, z_k, \nabla f(z_k)) + \frac{1}{2} g_2(t_k, z_k) e_k, g_2(t_k, z_k) e_k \right).
\]

**Remark 3.13.** Note (3.12) is a perturbation of the analogous result in the continuous case: the first term in \( \beta_k \) is a discretization of the corresponding term in the continuous case - the remaining terms are perturbations of order \( h \).

**Remark 3.14.** Note also that if under the standard assumptions on the error, \( E[e_i] = 0 \) and \( \text{Var}(e_i) = \sigma^2 \), along with independence, then

\[
E[\beta_k] = h \frac{L_E g_2^2}{2} \sigma^2
\]
Proof.

\[ E(t_{k+1}, z_{k+1}) - E(t_k, z_k) \leq \partial_t E(t_k, z_k)(t_{k+1} - t_k) + \langle \nabla E(t_k, z_k), z_{k+1} - z_k \rangle + \frac{L_E}{2}|z_{k+1} - z_k|^2 \]

by (3.6)

\[ \leq h \left( \partial_t E(t_k, z_k) + \langle \nabla E(t_k, z_k), g(t_k, z_k, \nabla f(z_k)) \rangle \right) \]

+ \( h \langle \nabla E(t_k, z_k), g_2(t_k, z_k)e_k \rangle + \frac{L_E h^2}{2}|g(t_k, z_k, \nabla f(z_k) + e_k)|^2 \)

\[ \leq -hr_E E(t_k, z_k) - ha_E |g(t_k, z_k, \nabla f(z_k))|^2 \]

by (FEP)

\[ \leq -hr_E E(t_k, z_k) - h a_E \frac{L_E}{2}|g(t_k, z_k, \nabla f(z_k))|^2 \]

+ \( h \langle \nabla E(t_k, z_k), g_2(t_k, z_k)e_k \rangle + \frac{L_E h^2}{2}|g(t_k, z_k, \nabla f(z_k) + e_k)|^2 \)

which concludes the proof.

Following the argument in the continuous, and using the previous calculation, we see that the problem is to define \( I_k \) such that \( I_0 = 0 \) and

\[ I_{k+1} - I_k = -hr_E I_k - h \beta_k, \]

Lemma 3.15. The solution of (3.4) is given by

\[ I_k = -(1 - hr_E)^k \sum_{i=0}^{k-1} (1 - hr_E)^{-i-1} \beta_i, \]

Proof. Applying the definition (3.15) we obtain

\[ I_{k+1} - I_k = -(1 - hr_E)^k \sum_{i=0}^{k} (1 - hr_E)^{-i-1} \beta_i - \sum_{i=0}^{k-1} (1 - hr_E)^{-i-1} \beta_i \]

\[ = -(1 - hr_E)^k \beta_k - hr_E(1 - hr_E)^{-k-1} \beta_k \]

\[ = -hr_E I_k - h \beta_k \]

as required.

The arguments above lead us to the following definition.

Definition 3.16. Define the perturbed Lyapunov function \( \tilde{E}_k \) by

\[ \tilde{E}_k := E(t_k, z_k) - (1 - hr_E)^k J_k = E(t_k, z_k) + I_k \]

where

\[ J_k = h \sum_{i=0}^{k-1} (1 - hr_E)^{-i-1} \beta_i, \]

and \( \beta_i \) given by (3.12).

Lemma 3.17. Let \( z_k \) be the solution of the forward Euler method (FEP) for (PODE) with time step \( h > 0 \). Let \( \tilde{E} \) be the perturbed Lyapunov function defined in Definition 3.16. Suppose \( E \) is a rate-generating Lyapunov function for (ODE) which satisfies (3.6). Choose

\[ 0 < h \leq \frac{2a_E}{L_E} \]

Then

\[ \tilde{E}(t_{k+1}, z_{k+1}) \leq (1 - hr_E) \tilde{E}(t_k, z_k) \]

Proof. From Lemma 3.12, (3.12) combined with (CFL) we have

\[ E(t_{k+1}, z_{k+1}) - E(t_k, z_k) \leq -hr_E E(t_k, z_k) + h\beta_k. \]

Also, from Lemma 3.15, (3.4) holds for \( I_k \). Combining these two estimates use the definition of \( \hat{E}_k \) to obtain the result

\[ \hat{E}(t_{k+1}, z_{k+1}) - \hat{E}(t_k, z_k) \leq -hr_E \hat{E}(t_k, z_k) \]

which gives (3.17).

\[ \square \]

Proposition 3.18. Under the assumptions of Lemma 3.17 and suppose \( h \) satisfies (CFL), then

\[ E(t_k, z_k) \leq (1 - hr_E)^k (E(t_0, z_0) + J_k) \]

Note the similar pattern of the result (3.18) to corresponding result (3.10) in the continuous case.

Proof. First, using the definitions in (3.17) and by induction, we have

\[ E(t_k, z_k) \leq (1 - hr_E)^k E(t_0, z_0) - I_k \]

\[ = (1 - hr_E)^k E(t_0, z_0) + (1 - hr_E)^k J_k \]

\[ = (1 - hr_E)^k (E(t_0, z_0) + J_k). \]

\[ \square \]

Corollary 3.19. Under the assumptions of Lemma 3.17, and assume in addition that \( E \) satisfies

\[ E(t_k, z_k) \geq C_1 |\hat{\beta}_k| - C_2, \quad \hat{\beta}_k = \nabla E(t_k, z_k) + L_E h g(t_k, z_k, \nabla f(z_k)), \]

for some constants \( C_1 > 0 \) and \( C_2 \geq 0 \). Then, for all \( k > 0 \), \(|\hat{\beta}_k| \) satisfies

\[ \sup_{0 \leq i \leq k} |\hat{\beta}_i| \leq M_k, \]

where \( M_k \) depends on

\[ \sum_{i=0}^{k-1} (1 - hr_E)^{-i} |g_2(t_i, z_i)e_i|. \]

and, in particular, if (3.19) is finite when \( k \to +\infty \), then \( M_k \) is bounded in \( l^\infty \) and, for all \( k \geq 0 \),

\[ E(t_k, z_k) \leq (1 - hr_E)^k \left( E(t_0, z_0) + M_k \sum_{i=0}^{k-1} (1 - hr_E)^{-i-1} |g_2(t_i, z_i)e_i| \right) = O((1 - hr_E)^k). \]

Proof. The proof is similar to the continuous case, Corollary 3.11, using condition (3.19) to apply Gronwall’s Lemma on \( (1 - hr_E)^{-i} |\hat{\beta}_i| \).

\[ \square \]

Remark 3.20. Note that assumption (3.19) is not satisfied in general in the accelerated setting as we will see it in the following. Indeed, in Nesterov’s method, the gradient of \( f \) is evaluated at a different point than the Lyapunov function.

3.5. Discussion of accelerated methods. Contrary to the gradient descent case, the situation is not quite so simple in the accelerated case. First note that there is a gap between the discrete and continuous setting: more than one ODE can be consistent with a discrete algorithm. This means, in principle, that there may be whole parameterized families of ODEs which satisfy the condition (3.21) and which can be discretized to obtain the algorithm. In the accelerated case, we need to use such a parameterized family. It may also be too restrictive to require that the Forward Euler discretization of the ODE satisfies the same Lyapunov function. However, we can consider a parameterized ODE system and we may need to consider a parameterized Lyapunov function.

Consider the parameterized velocity field \( g(t, z, \nabla f(y^\epsilon); \epsilon) \) where, \( z = (x, v) \), \( y^\epsilon = x + \theta_\epsilon (v - x) \), with \( \theta_\epsilon \in (0, 1) \), for \( \epsilon > 0 \), and \( \theta_0 = 0 \), and

\[ g(t, z, \nabla f(y^\epsilon); \epsilon) = g_1(t, z; \epsilon) + g_2(t, z; \epsilon) \nabla f(y^\epsilon), \]
which is $L_g$-Lipschitz continuous, uniformly in $\epsilon$. In continuous time, set $\epsilon = 0$, and consider

\[(\text{ODE-h}) \quad \dot{z}(t) = g(t, z(t), \nabla f(x(t)); 0)\]

However, for a given learning rate (time step) $h \geq 0$, we allow a perturbation which depends on $h$, i.e. $\epsilon = h$, and instead consider the algorithm, $z_k = (x_k, v_k)$,

\[(\text{FE-h}) \quad z_{k+1} = z_k + h g(t_k, z_k, \nabla f(y^h_k); h), \quad y^h_k = x_k + \theta_h (v_k - x_k),\]

provided an initial value $z_0$.

**Definition 3.21.** The (continuous) function $E(t, z; h)$ is a rate-generating Lyapunov function for (ODE-h) if there exists $r_E, a_E \geq 0$, such that

\[d\frac{dt}{dt} E(t, z(t); 0) \leq -r_E E(t, z(t); 0) - a_E |\nabla f(x)|^2.\]

for every solution $z(t) = (x(t), v(t)) : [0, \infty) \rightarrow \mathbb{R}^n \times \mathbb{R}^n$ of (ODE-h). If, in addition, there is $L_E > 0$, so that

\[E(t_k+1, z_k+1; h) \leq (1 - r_E h) E(t_k, z_k; h) - (L_E h - a_E) h |\nabla f(y^h_k)|^2,\]

for all solutions $z_k$ of (FE-h) with $0 < h \leq \frac{a_E}{L_E}$, then we call $E$ a rate-generating Lyapunov function for the sequence (FE-h).

In Sections 5 and 6, we will adapt the previous analysis to the accelerated gradient method in the convex and strongly convex case. In particular, we will extend the accelerated gradient method to the perturbed gradient case. The main difficulty resides in the fact that, in the discrete case, the gradient and the Lyapunov function are not evaluated at the same point anymore and we will use different methods to overcome this issue.

**3.6. Variable time step and convergence in expectation.** Now we consider the case where the error $e_k$ satisfies

\[E[e_k] = 0 \quad \text{and} \quad \text{Var}(e_k) = \mathbb{E}[|e_k|^2] = \sigma^2 > 0,\]

and where the time step/learning rate $h$ is not constant anymore i.e. $h = h_k$, we require that $h_0$ satisfies CFL, which from induction, need to be small enough, so that means: take warm up where $h$ small enough, and run a few steps until $E_0$ small enough. Then, (3.12) becomes

\[E(t_{k+1}, z_{k+1}) - E(t_k, z_k) \leq -h_k r_E E(t_k, z_k) + h_k \beta_k,\]

where $\beta_k$ also depends on $h_k$

\[\beta_k := \langle \nabla E(t_k, z_k), g_2(t_k, z_k) e_k \rangle + h_k \left( \frac{L_E}{2} g_2(t_k, z_k) e_k + L_E g(t_k, z_k, \nabla f(z_k)), g_2(t_k, z_k) e_k \right).\]

so

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

Taking the expectation in (3.6), we obtain

\[\mathbb{E}[E(t_{k+1}, z_{k+1})] \leq (1 - h_k r_E) E(t_k, z_k) + \frac{h_k^2 L_E g_2(t_k, z_k)^2 \sigma^2}{2}.\]

Then, we deduce the following result

**Proposition 3.22 (Case $r_E > 0$).** Assume that $r_E > 0$ and $\bar{g}_2 = \max_{(t, z)} g_2(t, z) < +\infty$. If

\[h_k := \frac{2}{r_E (k + \alpha^{-1} E_0^{-1})} \quad \text{where} \quad \alpha = \frac{\bar{g}_2^2}{2 L_E \sigma^2} \]

then,

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

Note, the assumption that $g_2$ is bounded and $r_E > 0$ apply to strongly convex gradient descent as well as strongly convex accelerated gradient descent.
Proposition 3.23 (Case \( r_E = 0 \)). Assume that \( r_E = 0 \) and \( E \) satisfies (3.6)-(3.6). If \( h_k = k^{-\alpha} \), \( t_k = \sum_{i=0}^{k} h_i \), 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
  \[
  \mathbb{E}[f(x_k)] - f^* = \mathcal{O}\left( \frac{1}{k^{1-\alpha}} \right)
  \]
  and if \( \alpha = 1 \), then
  \[
  \mathbb{E}[f(x_k)] - f^* = \mathcal{O}\left( \frac{1}{\ln(k)} \right)
  \]

- **Case** \( a_3 > 0 \) and \( b_2 > 0 \): If \( \alpha \in \left( \frac{4}{3}, 1 \right) \), then
  \[
  \mathbb{E}[f(x_k)] - f^* = \mathcal{O}\left( \frac{b_1}{k^{1-\alpha}} + \frac{b_2}{k^{2-2\alpha}} \right)
  \]
  and if \( \alpha = 1 \), then
  \[
  \mathbb{E}[f(x_k)] - f^* = \mathcal{O}\left( \frac{b_1}{\ln(k)} + \frac{b_2}{\ln(k)^2} \right)
  \]

We will see that this result can be applied for convex gradient descent and convex accelerated gradient method.

**Proof.** First, since \( t_k = \sum_{i=0}^{k} \frac{1}{r_i} \), we need \( \alpha < 1 \). Summing (3.6) over from 0 to \( k - 1 \), we obtain

\[
\mathbb{E}[E(t_k, x_k)] \leq \frac{\sigma^2}{2} \sum_{i=0}^{k-1} h_i^2 (a_1 + a_2 t_i + a_3 t_i^2).
\]

Now we want to prove that \( \mathbb{E}[E(t_k, x_k)] \) is bounded.

- **Case** \( a_1, a_2, b_1 > 0, a_3 = b_2 = 0 \): In that case, we need to prove that
  \[
  a_1 \sum_{i=0}^{\infty} \frac{1}{i^{2\alpha}} + a_2 \sum_{i=0}^{\infty} \frac{1}{i^{2\alpha}} \sum_{j=0}^{i} \frac{1}{j^{\alpha}} < +\infty.
  \]
For the first series converges if and only if $\alpha > \frac{1}{2}$. Before studying the second series, remark that
\[
\sum_{j=0}^{i} \frac{1}{j^{\alpha}} \sim \begin{cases} \frac{1}{\alpha} \ln(i) & \text{if } \alpha \neq 1, \\ \frac{1}{\alpha - 1} & \text{if } \alpha = 1. \end{cases}
\]
and then
\[
\frac{1}{i^{2\alpha}} \sum_{j=0}^{i} \frac{1}{j^{\alpha}} \sim \begin{cases} \frac{1 - 3\alpha}{\alpha - 1} \ln(i) & \text{if } \alpha \neq 1, \\ \frac{1}{\ln(i)} & \text{if } \alpha = 1. \end{cases}
\]
which implies that the series converges if and only if $\alpha = 1$ or $\frac{2}{3} < \alpha < 1$. So we have proved that $E[E(t_k, x_k)]$ is bounded if $\alpha = 1$ or $\frac{2}{3} < \alpha < 1$. By (3.6), we obtain
\[
b_1 t_k (E[f(x_k)] - f^*) \leq C.
\]
In addition,
\[
t_k = \sum_{i=0}^{k} \frac{1}{i^{2\alpha}} \geq \begin{cases} \frac{k^{1 - \alpha}}{1 - \alpha} & \text{if } \alpha \neq 1, \\ \ln(k) & \text{if } \alpha = 1. \end{cases}
\]
which concludes the proof of the first item.

- **Case $a_3 > 0$ and $b_2 > 0$:** In that case, we need to prove that
  \[
a_1 \sum_{i=0}^{\infty} \frac{1}{i^{2\alpha}} + a_2 \sum_{i=0}^{\infty} \frac{1}{i^{2\alpha}} \sum_{j=0}^{i} \frac{1}{j^{\alpha}} + a_3 \sum_{i=0}^{\infty} \frac{1}{i^{2\alpha}} \left( \sum_{j=0}^{i} \frac{1}{j^{\alpha}} \right)^2 < +\infty
\]
  Arguing as in the first part of the proof the two first series converge if $\alpha = 1$ or $\frac{2}{3} < \alpha < 1$. Expanding the square,
  \[
  \frac{1}{i^{2\alpha}} \left( \sum_{j=0}^{i} \frac{1}{j^{\alpha}} \right)^2 = \frac{1}{i^{2\alpha}} \sum_{j=0}^{i} \frac{1}{j^{2\alpha}} + \frac{1}{i^{2\alpha}} \sum_{j,d=0,j \neq d}^{i} \frac{1}{j^{\alpha} l^{\alpha}}.
  \]
  Then in our case, the series converge if $\frac{1}{2} < \alpha < 1$ or $\alpha = 1$. Combining the fact that
  \[
b_1 t_k + b_2 t_k^2 \geq \begin{cases} b_1 \frac{k^{1 - \alpha}}{1 - \alpha} + b_2 \frac{k^{2 - 2\alpha}}{2 - \alpha}, & \text{if } \alpha \neq 1, \\ b_1 \ln(k) + b_2 \ln(k)^2 & \text{if } \alpha = 1. \end{cases}
\]
with (3.6) concludes the proof.

4. Applications to Gradient descent

In this section, we apply our previous abstract analysis to gradient descent for convex and strongly convex functions. In this case, $g$ is given by
\[
g(t, z, p) = -p \quad \text{i.e.} \quad g_1(t, z) = 0 \text{ and } g_2(t, z) = -1.
\]

Let $f$ be a $\mu$-strongly convex, $L$-smooth function. Consider
\[
(28) \quad \dot{x}(t) = -\nabla f(x(t))
\]
and its associated forward Euler scheme
\[
(29) \quad x_{k+1} - x_k = -h\nabla f(x_k)
\]
as well as their perturbed version, where the gradient is replaced by $\tilde{\nabla} f = \nabla f + e$,
\[
(30) \quad \dot{x}(t) = -(\nabla f(x(t)) + e(t)).
\]
and the forward Euler scheme
\[
(31) \quad x_{k+1} - x_k = -h(\nabla f(x_k) + e_k),
\]
with initial condition $x_0$. 
Then, define the functions
\[ E^c(t, x) := t(f(x) - f^*) + \frac{1}{2}|x - x^*|^2, \]
\[ E^{sc}(x) := f(x) - f^* + \mu_0|x - x^*|^2. \]
for the convex and strongly convex cases, respectively.

In the following, we will apply the previous abstract analysis to \( E^c \) and \( E^{sc} \).

4.1. **Unperturbed case.** First, we show that the Lyapunov functions satisfy (3.6).

**Proposition 4.1.**

- \( E^c \) is a Lyapunov function in the sense of Definition 3.3 with \( r_{E^c} = 0 \) and \( a_{E^c} = t \). In addition, \( E^c \) satisfies (3.6) with \( L_{E^c} = Lt_{k+1} + 1 \).
- \( E^{sc} \) is a Lyapunov function in the sense of Definition 3.3 with \( r_{E^{sc}} = \mu \) and \( a_{E^{sc}} = 1 \). In addition, \( E^{sc} \) satisfies (3.6) with \( L_{E^{sc}} = L + \mu \).

**Proof.**

- In the convex case, we first start to look for (3.3):
\[
\partial_t E^c(t, z) - \nabla E^c(t, z) \nabla f(z) = f(z) - f^* - \langle \nabla f(z) + z - x^*, \nabla f(z) \rangle \\
= f(z) - f^* - \langle z - x^*, \nabla f(z) \rangle - t|\nabla f(z)|^2 \\
\leq -t|\nabla f(z)|^2,
\]
by convexity, which gives \( r_{E^c} = 0 \) and \( a_{E^c} = t \). Now, by 1-convexity of the quadratic term and \( L \)-smoothness of \( f \),
\[
E^c(t_{k+1}, z_{k+1}) - E^c(t_k, z_k) \leq t_k(f(z_k) - f^*) + \langle \nabla f(z_k), z_{k+1} - z_k \rangle + \frac{L}{2}|z_{k+1} - z_k|^2 - t_k(f(z_k) - f^*) \\
+ \frac{1}{2}|z_{k+1} - x^*|^2 - \frac{1}{2}|z_k - x^*|^2 \\
\leq \langle t_k \nabla f(z_k) + z_k - x^*, z_{k+1} - z_k \rangle + \frac{Lt_{k+1} + 1}{2}|z_{k+1} - z_k|^2 + (t_{k+1} - t_k)(f(z_k) - f^*) \\
+ (t_{k+1} - t_k)(\nabla f(z_k), z_{k+1} - z_k) \\
\leq (t_{k+1} - t_k)\partial_t E^c(t_k, z_k) + \langle \nabla E^c(t_k, z_k), z_{k+1} - z_k \rangle + \frac{Lt_{k+1} + 1}{2}|z_{k+1} - z_k|^2,
\]
since
\[
(t_{k+1} - t_k)(\nabla f(z_k), z_{k+1} - z_k) = -h^2|\nabla f(z_k)|^2 \leq 0.
\]
Then \( L_{E^c} = Lt_{k+1} + 1 \).

- In the strongly convex case,
\[
\partial_t E^{sc}(z) - \nabla E^{sc}(z) \nabla f(z) = -\langle \nabla f(z) + \mu(z - x^*), \nabla f(z) \rangle \\
= -\mu\langle z - x^*, \nabla f(z) \rangle - |\nabla f(z)|^2 \\
\leq -\mu\left(f(z) - f^* - \frac{\mu}{2}|z - x^*|^2\right) - |\nabla f(z)|^2,
\]
by strong convexity and then \( r_{E^{sc}} = \mu \) and \( a_{E^{sc}} = 1 \). Concerning (3.6), since \( E^{sc} \) is time independent, (3.6) is equivalent to \( L \)-smoothness condition which gives \( L_{E^{sc}} = L + \mu \).

□

Then, applying Lemma 3.5 and Lemma 3.6 to \( E^c \) and \( E^{sc} \), we obtain the usual rates in the convex and strongly convex case.

**Corollary 4.2.** Let \( x \) be the solution of (4) and \( x_k \) the sequence generated by (4). Then,

- **Convex case:** for all \( h \leq \frac{1}{L} \),
\[
f(x(t)) - f^* \leq \frac{1}{2t}|x_0 - x_*|^2 \text{ and } f(x_k) - f^* \leq \frac{1}{2hk}|x_0 - x_*|^2,
\]

- **Strongly convex case:** for all \( h \leq \frac{1}{L + \mu} \),
\[
f(x(t)) - f^* \leq \frac{1}{2t}|x_0 - x_*|^2 \text{ and } f(x_k) - f^* \leq \frac{1}{2hk}|x_0 - x_*|^2,
\]
• Strongly convex case: for all $h \leq \frac{2}{\bar{L}}$,
\[ f(x(t)) - f^* + \frac{\mu}{2} \| x(t) - x^* \|^2 \leq e^{-\mu t} E^{sc}(x_0) \] and \[ f(x_k) - f^* + \frac{\mu}{2} \| x_k - x^* \|^2 \leq (1 - h \mu)^k E^{sc}(x_0), \]

4.2. Perturbed gradient descent: convex case. In this section, we replace gradients in (4) and (4) by perturbed gradient $\nabla f = \nabla f + e$, where $e$ is an error term, and then we consider (4) and (4).

Following (3.9), define the perturbed Lyapunov function, $E^c$, by
\[ E^c(t, x) = E^c(t, x) + I^c(t, x(\cdot)), \]
where $E^c$ is defined as previously and,
\[ I^c(t, x(\cdot)) = \int_0^t (\nabla E^c(s, x(z)), e(s)) ds \]
\[ = \int_0^t (x(s) - x^* + s \nabla f(x(s)), e(s)) ds. \]

Then applying Proposition 3.10 and Corollary 3.11, we have

**Proposition 4.3.** Let $x$ be a solution of (4). Then,
\[ E^c(t, x) \leq E^c(0, x_0) - I^c(t, x(\cdot)). \]
In addition, if $e \in L^1(\mathbb{R}_+)$, then $\sup_{t \geq 0} |x(s) - x^* + s \nabla f(x(s))| \leq M_\infty < +\infty$, and
\[ f(x(t)) - f^* \leq \frac{1}{t} \left( \frac{1}{2} \| x_0 - x^* \|^2 + M_\infty \| e \|_{L^1} \right). \]

Following the abstract case, define the discrete perturbed Lyapunov function, $E^c_k$, by
\[ E^c_k = E^c(t_k, x_k) + I^c_k \]
where $t_k := hk$, and
\[ I^c_k = -h \sum_{i=0}^{k-1} \beta_i, \]
with,
\[ \beta_i := -\langle x_i - x^* - t_{i+1} \nabla f(x_i), e_i \rangle + h(L_{i+1} + 1) \left( \nabla f(x_i) + \frac{1}{2} e_i, e_i \right). \]

In the next proposition we sum up results coming from Lemma 3.12, Lemma 3.17 and Corollary 3.19.

**Proposition 4.4.** Let $x_k$ be the sequence generated by perturbed gradient descent (4). Assume that $h$ satisfies
\[ h \leq \frac{1}{L}. \]

Then, $E^c_k$ satisfies
\[ E^c_{k+1} \leq E^c_k + h \beta_k, \]
where $\beta_k$ is defined by (4.2). Then
\[ f(x_k) - f^* \leq \frac{1}{t_k} \left( \frac{1}{2} \| x_0 - x^* \|^2 - I_k \right). \]

In addition, $\sup_{0 \leq i \leq k} |x_i - x^*| + |t_{i+1} \nabla f(x_i)| \leq M_k$, where $M_k$ depends on
\[ \sum_{i=0}^{k-1} (|e_i| + i|e_i|^2). \]
Assuming that (4.4) is finite then $M_k$ is bounded in $l^\infty$ and
\[ f(x_k) - f^* \leq \frac{1}{t_k} \left( \frac{1}{2} \| x_0 - x^* \|^2 + M_k h \sum_{i=0}^{k-1} (|e_i| + i|e_i|^2) \right) = O \left( \frac{1}{k} \right). \]
4.3. **Variable time step and convergence in expectation: convex case.** In this section we 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 (3.6).

Note that (4.4) still holds for an adaptative time step and then, since \( \mathbb{E}[\tilde{\beta}_k] = h_k^2 (L t_{k+1} + 1) \sigma^2 \),
\[
\mathbb{E} [E_{k+1}^c] \leq E_k^c + \frac{h_k^2 (L t_{k+1} + 1) \sigma^2}{2},
\]
and
\[
\mathbb{E} [E_k^c] \geq t_k (\mathbb{E} [f(x_k)] - f^*),
\]
which correspond to (3.6) and (3.6) with
\[
a_1 = \frac{L (h_k + 1)}{2}, \quad a_2 = \frac{L}{2}, \quad a_3 = 0, \quad b_1 = 1 \text{ and } b_2 = 0.
\]
So Proposition 3.23 gives

**Proposition 4.5.** Assume \( h_k := k^{-\alpha} \) and \( t_k = \sum_{i=0}^{k} h_i \), then the following holds:

- If \( \alpha \in \left( \frac{2}{3}, 1 \right) \), then
  \[
  \mathbb{E} [f(x_k)] - f^* = O \left( \frac{1}{k^{1-\alpha}} \right),
  \]
- if \( \alpha = 1 \), then
  \[
  \mathbb{E} [f(x_k)] - f^* = O \left( \frac{1}{\ln(k)} \right).
  \]

4.4. **Perturbed gradient descent: strongly convex case.** Now consider \( \mu \)-strongly convex function \( f \). Define the perturbed Lyapunov function \( \tilde{E}^{sc} : [0, +\infty) \times \mathbb{R}^d \rightarrow [0, +\infty) \) by
\[
\tilde{E}^{sc}(t, x) = E^{sc}(x) + I^{sc}(t, x(\cdot)),
\]
where,
\[
I^{sc}(t, x(\cdot)) = e^{-\mu t} \int_0^t e^{\mu s} \langle \nabla E^{sc}(x(s)), e(s) \rangle \, ds
= e^{-\mu t} \int_0^t e^{\mu s} \langle \mu (x(s) - x^*) + \nabla f(x(s)), e(s) \rangle \, ds
\]
\[= -e^{-\mu t} J^{sc}(t, x(\cdot)).\]

Then the following result holds.

**Proposition 4.6.** Let \( x \) be a solution of (4) with initial data \( x_0 \), then
\[
f(x(t)) - f^* + \frac{\mu}{2} \| x(t) - x^* \|^2 \leq e^{-\mu t} \left( E^{sc}(x_0) + J^{sc}(t, x(\cdot)) \right),
\]
In addition, if \( \exp(\mu \cdot)e \in L^1(\mathbb{R}^+) \), then sup\( t \geq 0 \| x(s) - x^* + \nabla f(x(s)) \| \leq M_{\infty} < +\infty \) and
\[
f(x(t)) - f^* + \frac{\mu}{2} \| x(t) - x^* \|^2 \leq e^{-\mu t} \left( E^{sc}(x_0) + M_{\infty} \| \exp(\mu \cdot)e \|_{L^1} \right) = O(e^{-\mu t}).
\]

Now in the discrete case, define the discrete perturbed Lyapunov function \( \tilde{E}_k^{sc} \), for \( k \geq 0 \), by
\[
\tilde{E}_k^{sc} = E^{sc}(x_k) + I_k^{sc},
\]
where \( x_k \) is generated by the forward Euler discretization of (4), (4), and
\[
I_k^{sc} = (1 - h\mu)^k \sum_{i=0}^{k-1} (1 - h\mu)^{-i-1} \beta_i = -(1 - h\mu)^k J_k^{sc},
\]
where
\[
\beta_i = \langle \mu (x_i - x^*) + \nabla f(x_i), e_i \rangle + h (L + \mu) \left( \nabla f(x_i) + \frac{1}{2} e_i, e_i \right).
\]

As in the convex case, in the next proposition we sum up results coming from Lemma 3.12, Lemma 3.17 and Corollary 3.19.
Proposition 4.7. Let \( x_k \) be the sequence generated by perturbed gradient descent (4). Assume that \( h \) satisfies
\[
h \leq \frac{2}{L + \mu}.
\]
\( E_k^{sc} \) satisfies
\[
E_{k+1}^{sc} \leq E_k^{sc} + h \beta_k,
\]
where \( \beta_k \) is defined by (4.4). Then
\[
f(x_k) - f^* + \frac{\mu}{2} |x_k - x^*|^2 \leq (1 - h \mu)^k (E_0^{sc} + J_k^{sc}).
\]
In addition, \( \sup_{0 \leq i \leq k} (|x_i - x^*| + |\nabla f(x_i)|) \leq M_k \), where \( M_k \) depends on
\[
\sum_{i=0}^{k-1} (1 - h \mu)^{-i} |e_i|.
\]
Assuming that (4.7) is finite when \( k \nearrow +\infty \), then \( M_k \) is bounded in \( t^\infty \) and
\[
f(x_k) - f^* + \frac{\mu}{2} |x_k - x^*|^2 \leq (1 - h \mu)^k \left( E_0^{sc} + M_k \sum_{i=0}^{k-1} (1 - h \mu)^{-i} |e_i| \right) = \mathcal{O}((1 - h \mu)^k).
\]

4.5. Variable time step and convergence in expectation: strongly convex case. Now consider the case of a variable time step \( h_k \) and a zero-mean and constant Variance error \( e_k \) i.e. \( e_k \) satisfies (3.6).

As in the convex case, (4.7) is still true for an adaptive time step \( h_k \) and then, since \( \mathbb{E}[\beta_k] = h_k^2 (L + \mu)^2 \),
\[
\mathbb{E}[E_{k+1}^{sc}] \leq (1 - \mu h_k) E_k^{sc} + \frac{h_k^2 (L + \mu)^2}{2}.
\]
We are in the case where \( r_{E^{sc}} = \mu > 0 \) and \( g_2^2 = 1 \), then Proposition 3.22 gives

Proposition 4.8. If
\[
h_k^{sc} := \frac{2}{\mu (k + \alpha^{-1} E_0^{-1})}, \quad \alpha^{sc} := \frac{\mu}{2 (C_f + 1) \sigma^2}
\]
and \( t_k = \sum_{i=0}^{k} h_i \), then the following holds:
\[
\mathbb{E}[E(x_k)] \leq \frac{1}{\alpha^{sc} k + E_0^{-1}} = \frac{2 (C_f + 1) \sigma^2}{\mu k + 2 (C_f + 1) \sigma^2 E_0^{-1}}.
\]

5. Accelerated method: convex case

In the remainder of the paper, we will extend the analysis developed in section 3 to the accelerated gradient method in both continuous and discrete time. We study the perturbed case as in the full gradient descent case. However, we will see, in the perturbed case, that the definition of \( I \) requires slight modifications. Indeed, the Lyapunov function and the gradient of \( f \) are not evaluated at the same point. We will also consider the case where the time step is varying and the error has zero-mean and a constant variance. Section 5 considers the convex case while Section 6 focuses on the strongly convex case.

In this section, we study system (1st-ODE) as well as its discretization (FE-C). Then we extend the Lyapunov analysis to the perturbed case. We present also a convergence rate for the expectation of \( f \) in the case of the time step varies, \( h_k \), and the error has a constant variance.

5.1. Unperturbed gradient case. Define the perturbed ODE by setting \( z = (x, v) \) and \( g(t, x, v, \nabla f(y^*); \epsilon) \) as follows
\[
g(t, x, v, \nabla f(y^*); \epsilon) = g_1(t, x, v) + g_2(t, x, v) \nabla f(y^*)
\]
where,
\[
g_1(t, x, v) = \begin{pmatrix} -\frac{2}{t} & \frac{2}{t} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix} \quad \text{and} \quad g_2(t, x, v) = \begin{pmatrix} -\frac{1}{\sqrt{L}} \\ -\frac{2}{t} \end{pmatrix}.
\]
and

\[ y' = x + \frac{2\epsilon}{t}(v - x). \]

Let \( h > 0 \) be a given small time step/learning rate and let \( t_k = h(k + 2) \). Consider the perturbation with \( \epsilon = h \) of the forward Euler method for \( g \) given by

\[
\begin{align*}
x_{k+1} - x_k &= \frac{h}{t_k}(v_k - x_k) - \frac{h}{\sqrt{L}}\nabla f(y_k), \\
v_{k+1} - v_k &= -\frac{h}{t_k}\nabla f(y_k), \\
y_k &= \frac{h}{t_k + 2} = x_k + \frac{2}{t_k + 2}(v_k - x_k).
\end{align*}
\]

**Definition 5.1.** Define the continuous time parameterized Lyapunov function \( E^{ac,c}(t, x, v; \epsilon) := (t - \epsilon)^2(f(x) - f^*) + 2|v - x^*|^2 \)

Define the discrete time Lyapunov function \( E_k^{ac,c} \) by

\[ E_k^{ac,c} = E^{ac,c}(t_k, x_k, v_k; h) = E^{ac,c}(t_{k-1}, x_k, v_k; 0) \]

In the next proposition, we show that \( E_k^{ac,c} \) is a rate-generating Lyapunov function, in the sense of Definition 3.3. For system (1st-ODE) and its explicit discretization (FE-C).

**Proposition 5.2.** Suppose \( f \) is convex and \( L \)-smooth. Define the velocity field by (5.1) and let \( E_k^{ac,c} \) be given by (5.1). Then \( E_k^{ac,c}(\cdot; 0) \) is a continuous Lyapunov function with \( r_{E_k^{ac,c}} = 0 \) and \( a_{E_k^{ac,c}} = \frac{t^2}{\sqrt{L}} \) (with gap \( t^2|\nabla f(x)|^2 \)) i.e.

\[
\frac{d}{dt}E^{ac,c}(t, x(t), v(t); 0) \leq \frac{t^2}{\sqrt{L}}|\nabla f(x)|^2,
\]

and \( E_k^{ac,c} \) is a discrete Lyapunov function, with \( L_{E_k^{ac,c}} = t^2_k \), for the sequence generated by (FE-C), for all \( k \geq 0 \),

\[
E_{k+1}^{ac,c} \leq E_k^{ac,c} - h^2(f(x_k) - f^*) + \left(h - \frac{1}{\sqrt{L}}\right)t_k^2h|\nabla f(y_k)|^2,
\]

for \( h \leq \frac{1}{\sqrt{L}} \).

Since (FE-C) is equivalent to Nesterov’s method, the rate is known. The proof of the rate using a Lyapunov function can be found in [BT09]. A proof which shows that we can use the constant time step can be found in [Bec17]. The discrete Lyapunov function (5.1) was used in [SBC14, APR16] to prove a rate.

**Remark 5.3.** Note, compared to Su-Boyd-Can-dès’ ODE (A-ODE), there is a gap in the dissipation of the Lyapunov function \( E_k^{ac,c} \), which will not be there if the extra term, \( -\frac{1}{t_k}\nabla f(x) \), was missing. In particular, if \( \bar{z} \) solution of (A-ODE), and \( z \) solution of (1st-ODE), then we can prove faster convergence due to the gap. Indeed, this gap permits to improve the asymptotic rate of the descent of the gradient, see Corollary 5.5.

**Corollary 5.4.** Let \( f \) be a convex and \( L \)-smooth function. Let \( (x(t), v(t)) \) be a solution to (1st-ODE), then for all \( t > 0 \),

\[ f(x(t)) - f^* \leq \frac{2}{t^2}|v_0 - x^*|^2. \]

Furthermore, let \( x_k, v_k \) be given by (FE-C). Then for all \( k \geq 0 \) and \( h \leq \frac{1}{\sqrt{L}} \),

\[ f(x_k) - f^* \leq \frac{1}{(k + 1)^2} \left( f(x_0) - f^* + 2L|v_0 - x^*|^2 \right). \]

**Proof of Proposition 5.2.** First, in the continuous case, by definition of \( E \), we have

\[
\frac{d}{dt}E^{ac,c}(t, x(t), v(t); 0) \leq 2t(f(x) - f^*) + t^2|\nabla f(x)|, \dot{x} \\
+ 4\langle v - x^*, \dot{v} \rangle \\
\leq 2t(f(x) - f^*) + 2t\langle \nabla f(x), v - x \rangle - \frac{t^2}{\sqrt{L}}|\nabla f(x)|^2 \\
- 2t\langle v - x^*, \nabla f(x) \rangle \\
\leq 2t(f(x) - f^* - \langle x - x^*, \nabla f(x) \rangle) - \frac{t^2}{\sqrt{L}}|\nabla f(x)|^2.
\]
The proof is concluded by convexity,
\[ f(x) - f^* - \langle x - x^*, \nabla f(x) \rangle \leq 0. \]

In the discrete case, using the convexity and the $L$-smoothness of $f$, we obtain the following classical inequality, see [SBC14, APR16],
\[
t_i^2 (f(x_{k+1}) - f^*) - t_{k-1}^2 (f(x_k) - f^*) \\
\leq \left( \frac{k t_k^2}{k+2} - t_{k-1}^2 \right) (f(x_k) - f^*) + \frac{2t_k^2}{k+2} (\nabla f(y_k), v_k - x^*) \\
+ \left( \frac{h}{2} - \frac{1}{\sqrt{L}} \right) h t_k^2 |\nabla f(y_k)|^2.
\]

By definition of $v_{k+1}$, we have
\[
2|v_{k+1} - x^*|^2 - 2|v_k - x^*|^2 = -2h t_k \langle v_k - x^*, \nabla f(y_k) \rangle + \frac{h^2 t_k^2}{2} |\nabla f(y_k)|^2.
\]

Combining these two previous inequalities, we obtain
\[
E_{k+1}^{ac,c} - E_k^{ac,c} \leq -h^2 (f(x_k) - f^*) + \left( h - \frac{1}{\sqrt{L}} \right) t_k^2 h |\nabla f(y_k)|^2.
\]

Notice that due to the extra gap we obtain an improvement in the rate of convergence of $|\nabla f(y_k)|^2$. It is well-known that by (1.2) and Corollary 5.4, we have
\[
\frac{1}{2L} |\nabla f(x_k)|^2 \leq \frac{1}{(k+1)^2} \left( f(x_0) - f^* + 2L|v_0 - x^*|^2 \right).
\]

The gap obtained in the dissipation of $E^{ac,c}$ gives a faster convergence rate.

**Corollary 5.5.** For all $h < \frac{1}{\sqrt{L}}$, $k \geq 1$,
\[
\min_{0 \leq i \leq k} |\nabla f(y_i)|^2 \leq \frac{3\sqrt{L}}{h^3(1 - \sqrt{L} h)} \frac{1}{k^3} E^{ac,c}(x_0, v_0).
\]

**Proof.** By Proposition 5.2, for all $k \geq 0$,
\[
E_k^{ac,sc} - E_0^{ac,sc} \leq \sum_{i=0}^{k-1} (E_{i+1}^{ac,sc} - E_i^{ac,sc}) \\
\leq h \left( h - \frac{1}{\sqrt{L}} \right) \sum_{i=0}^{k} t_i^2 |\nabla f(y_i)|^2 \\
\leq h^3 \left( h - \frac{1}{\sqrt{L}} \right) \min_{0 \leq i \leq k} |\nabla f(y_i)|^2 \sum_{i=0}^{k} (i + 2)^2.
\]

To simplify the bound, notice that
\[
\sum_{i=0}^{k} (i + 2)^2 = \frac{2k^3 + 6k^2 + 13k}{6} \geq \frac{1}{3} k^3,
\]
which concludes the proof. \(\square\)

**Remark 5.6.** The optimal choice for $h$ is $h = \frac{3}{4\sqrt{L}}$ and then, we have
\[
\min_{0 \leq i \leq k} |\nabla f(y_i)|^2 \leq \frac{256L^2}{9k^3} E^{ac,c}(x_0, v_0).
\]
5.2. Perturbed gradient: continuous time. In this section, we consider that an error $e(t)$ is made in the evaluation of the gradient at time $t$. We study the following perturbation of system (1st-ODE),  

\[
(\text{Per-1st-ODE}) \quad \begin{cases}
\dot{x} = \frac{3}{t}(v - x) - \frac{1}{t\sqrt{L}}(\nabla f(x) + e(t)), \\
\dot{v} = -\frac{1}{t}(\nabla f(x) + e(t)).
\end{cases}
\]

where $e$ is a function which represents an error in the calculation of the gradient $\nabla f(x)$.

If we assume that $e$ and $f$ are smooth, the corresponding second order ODE would be

\[
\ddot{x} + \frac{2}{t}\dot{x} + \frac{1}{t\sqrt{L}}D^2f(x) \cdot \dot{x} + \left(\frac{1}{t\sqrt{L}} + 1\right)\nabla f(x) = -\left(\frac{1}{t\sqrt{L}} + 1\right)e(t) - \frac{1}{\sqrt{L}}e'(t),
\]

which corresponds to (H-ODE) perturbed by the term $-\left(\frac{1}{t\sqrt{L}} + 1\right)e(t) - \frac{1}{\sqrt{L}}e'(t)$.

**Definition 5.7.** Define the perturbed Lyapunov function for this system, $\tilde{E}^{ac,c}$, by

\[
\tilde{E}^{ac,c}(t, x, v; \epsilon) = E^{ac,c}(t, x, v; \epsilon) + I^{ac,c}(t, x(\cdot), v(\cdot); \epsilon),
\]

where $E^{ac,c}$ is defined as in (5.1) and

\[
I^{ac,c}(t, x(\cdot), v(\cdot); \epsilon) = \int_{0}^{t} s(2(v - x^*)) + \frac{s}{2\sqrt{L}}\nabla f(y'(s)), e(s)) + 2e(t)(\nabla f(y'(s)) + \frac{e(s)}{2})ds,
\]

where $y^*$ is defined by (5.1).

**Remark 5.8.** We will see in Proposition 5.11 that $I^{ac,c}(\cdot; \epsilon)$ is the good definition in the sense of Definition 3.16. However, in the discrete case ($\epsilon = h$), the gradient in $I^{ac,c}(\cdot; h)$ is evaluated in $y_k$ while $f$ in $E^{ac,c}(\cdot; h)$ is evaluated at $x_k$. We will overcome this difficulty defining another Lyapunov function which does not depend on $\nabla f(y_k)$ Nevertheless, $\tilde{E}^{ac,c}$ will be still useful assuming that $f$ is in addition $\mu$-strongly convex to obtain an accelerated rate for $\min_{i \leq k} |\nabla f(y_i)|^2$. Moreover $\tilde{E}^{ac,c}$ will also allow us to apply the abstract analysis with a variable time step and an error with zero mean and fixed Variance, developed in Section 3.6.

**Proposition 5.9.** Let $(x, v)$ be a solution of (Per-1st-ODE) with initial condition $(x(0), v(0)) = (x_0, v_0)$. Then

\[
\frac{d}{dt} \tilde{E}^{ac,c}(t, x, v; 0) \leq -\frac{t^2}{\sqrt{L}}|\nabla f(x)|^2.
\]

and

\[
f(x) - f^* \leq \frac{1}{t^2} \left(2|v_0 - x^*|^2 - I^{ac,c}(t, x(\cdot), v(\cdot))\right).
\]

**Proof.** Following the proof of Proposition 5.2, we have

\[
\frac{d}{dt} E^{ac,c}(t, x, v; 0) \leq -\frac{t^2}{\sqrt{L}}|\nabla f(x)|^2 - \frac{t^2}{\sqrt{L}}(\nabla f(x), e(t)) - 2t(v - x^*, e(t)).
\]

In addition,

\[
\frac{d}{dt} I^{ac,c}(t) = \frac{t^2}{\sqrt{L}}(\nabla f(x), e(t)) + 2t(v - x^*, e(t)).
\]

Then,

\[
\frac{d}{dt} \tilde{E}^{ac,c}(t, x, v; 0) \leq -\frac{t^2}{\sqrt{L}}|\nabla f(x)|^2;
\]

and the rest of the proof follows directly.

Then we deduce

**Corollary 5.10.** Let $(x, v)$ be a solution of (Per-1st-ODE) with initial condition $(x(0), v(0)) = (x_0, v_0)$. Then,

\[
\sup_{0 \leq s \leq t} |v(s) - x^*| + |s\nabla f(s)| < M(t),
\]
where $M(t)$ depends on

$$\int_0^t s\|e(s)\|. \tag{43}$$

Assume that (5.10) is bounded, then $M(t)$ is bounded in $L^\infty(\mathbb{R}_+) \text{ and,}$

$$f(x(t)) - f^* \leq \frac{1}{t^2} \left( 2|v_0 - x^*|^2 + M(t) \int_0^t s\|e(s)\| \right) = O\left( \frac{1}{t^2} \right).$$

Proof. From the previous proposition, $E^{ac,c}$ is decreasing and

$$t^2(f(x) - f^*) + 2|v - x^*|^2 \leq 2|v_0 - x^*|^2 - \int_0^t s\left( 2(v - x^*) + \frac{s}{\sqrt{L}} \nabla f(x), e(s) \right) ds.$$ 

Using (1.2), we obtain

$$\frac{1}{2L} \|\nabla f(x)\| + 2|v - x^*| \leq 2|x_0 - x^*|^2 + \frac{1}{2L} + 2 + \int_0^t \left( \frac{1}{\sqrt{L}} \|s\nabla f(x)\| + 2|v - x^*| \right) |se(s)| ds.$$

And since $e$ satisfies (5.10), we conclude applying Gronwall’s Lemma. \qed

5.3. Perturbed gradient: discrete time. Replacing gradients with $\tilde{\nabla} f$, the $\epsilon = h$-perturbation of the Forward Euler scheme (FE-C) becomes

$$\begin{align*}
(\text{Per-FE-C}) \quad \begin{cases} 
 x_{k+1} - x_k = \frac{2h}{t_k} (v_k - x_k) - \frac{h}{\sqrt{L}} (\nabla f(y_k) + e_k), \\
 v_{k+1} - v_k = -h \frac{t_k}{2} (\nabla f(y_k) + e_k),
\end{cases}
\end{align*}$$

where $y_k$ is as in (FE-C), $h$ is a constant time step, and $t_k := h(k + 2)$.

In this setting, we first give two estimates of the dissipation of $E^{ac,c}_k$ along (Per-FE-C).

**Proposition 5.11.** Let $x_k, v_k, y_k$ be sequences generated by (Per-FE-C)-(FE-C). Then,

$$E^{ac,c}_{k+1} - E^{ac,c}_k \leq -h^2 (f(x_k) - f^*) + \left( h - \frac{1}{\sqrt{L}} \right) t_k^2 h |\nabla f(y_k)|^2 + h\beta_k, \tag{44}$$

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

$$E^{ac,c}_{k+1} - E^{ac,c}_k \leq \left( \frac{h}{1 - \frac{1}{\sqrt{L}}} \right) t_k^2 h |\nabla f(y_k)|^2 + |e_k|^2 + h\overline{\beta}_k, \tag{45}$$

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

Proof. First, using the convexity and the $L$-smoothness of $f$, we obtain the following classical inequality (see [APR16] or [SBC14] in the case $e_k = 0$ and Appendix A for $e_k \neq 0$),

$$t_k^2 (f(x_{k+1} - f^*) - f_{k-1} (f(x_k - f^*) \leq -h^2 (f(x_k) - f^*) + 2ht_k \langle \nabla f(y_k), v_k - x^* \rangle - \left( \frac{1}{\sqrt{L}} - \frac{h}{2} \right) ht_k^2 |\nabla f(y_k)|^2$$

$$- \eta \frac{t_k^2}{\sqrt{L}} \langle \nabla f(y_k), e_k \rangle + h^2 t_k^2 \langle \nabla f(y_k), \frac{e_k}{2} \rangle.$$

By definition of $v_{k+1}$, we have

$$2|v_{k+1} - x^*|^2 - 2|v_k - x^*|^2 = -2ht_k \langle v_k - x^*, \nabla f(y_k) + e_k \rangle + \frac{h^2 t_k^2}{2} |\nabla f(y_k) + e_k|^2.$$

$$= -2ht_k \langle v_k - x^*, \nabla f(y_k) + \frac{h^2 t_k^2}{2} |\nabla f(y_k)|^2$$

$$- 2ht_k \langle v_k - x^*, e_k \rangle + h^2 t_k^2 \langle \nabla f(y_k) + \frac{e_k}{2}, e_k \rangle.$$
Therefore,
\[ E_{k+1}^{ac,c} - E_k^{ac,c} \leq -h^2(f(x_k) - f^*) - \left( \frac{1}{\sqrt{L}} - h \right) h t_k^2 |\nabla f(y_k)|^2 - 2 h t_k \langle v_k - x^* - \frac{t_k}{\sqrt{L}} \nabla f(y_k), e_k \rangle + 2 h^2 t_k^2 \left\langle \nabla f(y_k) + \frac{e_k}{2}, e_k \right\rangle, \]
and (5.11) is proved. For the second inequality, arguing as in Appendix A, we obtain
\[ t_k^2 (f(x_{k+1} - f^*) - t_k^2 (f(x_k - f^*)) \leq -h^2(f(x_k) - f^*) + 2 h t_k \langle \nabla f(y_k), v_k - x^* \rangle - \left( \frac{1}{\sqrt{L}} - \frac{h}{2} \right) h t_k^2 |\nabla f(y_k)|^2 + \frac{h t_k^2 |\nabla f(y_k)|^2}{\sqrt{L}}. \]
So, we conclude the proof of (5.11).

**Remark 5.12.** Inequalities (5.11) and (5.11) still hold with a variable time step \( h_k \) with \( t_k = \sum_{i=0}^{k} h_i \). Then, in Section 5.5, we will use (5.11) to apply Proposition 3.23.

We cannot use the same method as in Section 3 to define the Lyapunov function because the perturbed part of Lyapunov function would be
\[ I_k^{ac,c} := -h \sum_{i=0}^{k-1} \beta_i \]
which implies we need to control \( |\nabla f(y_k)| \). However, \( E_k^{ac,c} \), controls only \( |\nabla f(x_k)| \). In Section 5.4, we will show that we can still use this perturbed Lyapunov function assuming that \( f \) is \( \mu \)-strongly convex. Indeed, in this case, we have,
\[ t_i^2 \frac{\mu}{2} |x_k - x|^2 + \frac{\mu}{2} |v_k - x|^2 \leq E_k^{ac,c} \text{ and } t_i |\nabla f(y_k)| \leq L |y_k - x^*| \leq (1 - \lambda_h) L t_i |x_k - x^*| + \lambda_h L t_i |v_k - x^*|. \]
However in this section, we focus on the optimal case \( (h = \frac{1}{\sqrt{L}}) \) and then, we can define a simpler discrete perturbed Lyapunov function as follow.

**Definition 5.13.** Define the discrete perturbed Lyapunov function \( E_k^{ac,c} := E_k^{ac,c} + T_k^{ac,c} \), for \( k \geq 0 \), where \( E_k^{ac,c} \) is given by (5.1) and, for \( k \geq 0 \),
\[ T_k^{ac,c} := h \sum_{i=0}^{k-1} 2 t_i \langle v_{i+1} - x^*, e_i \rangle. \]

**Proposition 5.14.** Let \( x_k, v_k, y_k \) be sequences generated by (Per-FE-C)-(FE-C). Then, for \( h = \frac{1}{\sqrt{L}} \), \( E_k^{ac,c} \) is decreasing and
\[ f(x_k) - f^* \leq \frac{L}{(k+2)^2} (E_0^{ac,c} - T_k^{ac,c}). \]

**Proof.**
\[ E_{k+1}^{ac,c} - E_k^{ac,c} = 2 h t_k \langle v_{k+1} - x^*, e_k \rangle = 2 h t_k \langle v_k - x^*, e_k \rangle - h^2 t_k^2 \langle \nabla f(y_k) + e_k, e_k \rangle. \]
Combine this inequality and (5.11) to obtain
\[ E_{k+1}^{ac,c} - E_k^{ac,c} \leq -h^2(f(x_k) - f^*) + \left( h - \frac{1}{\sqrt{L}} \right) t_k^2 h |\nabla f(y_k)|^2 + \frac{h t_k^2 |\nabla f(y_k)|^2}{\sqrt{L}} (e_k, \nabla f(y_k) + e_k). \]
Since \( h = \frac{1}{\sqrt{L}} \), we deduce that \( E_k^{ac,c} \) is decreasing and the rest of the proof follows.

We immediately have the following result.
Corollary 5.15. Under the assumption of Proposition 5.14, then \( \max_{0 \leq i \leq k} |v_i - x^*| \leq M_k \) where \( M_k \) depends on 
\[
(47) \quad \sum_{i=1}^{k-1} |e_i|.
\]
Suppose (5.15) is finite, then \( M_k \) is bounded in \( l^\infty \) and 
\[
 f(x_k) - f^* \leq \frac{L}{(k+1)^2} \left( E_0^{ac,c} + M_k \sum_{i=1}^{k-1} |e_i| \right) = O \left( \frac{1}{(k+1)^2} \right).
\]
Proof. Since \( E_k^{ac,c} \) is decreasing from Proposition 5.14, 
\[
2|v_k - x^*|^2 \leq 2|v_0 - x^*|^2 + \frac{1}{L} \sum_{i=0}^{k-1} |v_i - x^*|(i+3)|e_i|.
\]
and the discrete version of Gronwall’s Lemma gives the result. \( \square \)

5.4. Decrease of \( \|\nabla f(y_k)\| \) in the strongly convex case. If, in addition, we assume that \( f \) is \( \mu \)-strongly convex, it is possible to improve the decrease of \( \|\nabla f(y_k)\| \). In order to do that, we introduce a different discrete perturbed Lyapunov function than (5.13).

Definition 5.16. Define the perturbed Lyapunov function \( \tilde{E}_k^{ac,c} := \tilde{E}^{ac,c}(t_k, x_k, v_k; h) \) where \( \tilde{E}^{ac,c} \) is defined in Definition 5.7 i.e. \( \tilde{E}^{ac,c}(t_k, x_k, v_k; h) = E_k^{ac,c} + I_k^{ac,c} \), as in for \( k \geq 0 \), where \( E_k^{ac,c} \) is given by (5.1) and and for \( k \geq 0 \), 
\[
 I_k^{ac,c} := I^{ac,c}(t_k, (x_k), (v_k); h) = -h \sum_{i=0}^{k-1} \beta_i,
\]
where 
\[
 \beta_i = -t_k \left( 2(v_k - x^*) + \frac{1}{2\sqrt{L}} \nabla f(y_k), e_k \right) + 2ht_k^2 \left( \nabla f(y_k) + \frac{e_k}{2} , e_k \right).
\]
First, we show that

Proposition 5.17. The perturbed Lyapunov function satisfies 
\[
 \tilde{E}_{k+1}^{ac,c} - \tilde{E}_k^{ac,c} \leq \left( h - \frac{1}{\sqrt{L}} \right) t_k^2 h |\nabla f(y_k)|^2, \quad k \geq 0
\]
Proof. From (5.11), we have 
\[
 E_{k+1}^{ac,c} - E_k^{ac,c} \leq \left( h - \frac{1}{\sqrt{L}} \right) ht_k^2 |\nabla f(y_k)|^2 + h\beta_k
\]
In addition, 
\[
 I_{k+1}^{ac,c} - I_k^{ac,c} = -h\beta_k.
\]
Therefore we get 
\[
 \tilde{E}_{k+1}^{ac,c} - \tilde{E}_k^{ac,c} \leq \left( h - \frac{1}{\sqrt{L}} \right) t_k^2 h |\nabla f(y_k)|^2
\]
Then, we obtain the following result.

Corollary 5.18. For all \( h < \frac{1}{\sqrt{L}} \), we have 
\[
 \sup_{0 \leq i \leq k} |v_i - x^*| + k|x_i - x^*| \leq M_k,
\]
where \( M_k \) depends on 
\[
(48) \quad \sum_{i=0}^{k-1} i^2 |e_i|.
\]
Moreover, if $e_k$ has a finite second moment i.e. (5.18) is bounded, then, $M_k$ is bounded and

$$
\min_{0 \leq i \leq k} |\nabla f(y_i)|^2 \leq \frac{3\sqrt{L}}{h^3(1 - \sqrt{L}h)} \frac{1}{k^3} \left( E^{ac,c}(x_0, v_0) + CM_k \sum_{i=0}^{k-1} \overline{e_i} \right) = O\left( \frac{1}{k^3} \right).
$$

Proof. From Proposition 5.17, $E^{ac,c}_k$ is decreasing and by $\mu$-convexity and $L$-smoothness, we have

$$
\frac{t_f^2 \mu}{2} |x_k - x^*|^2 + 2|v_k - x^*|^2 \leq E^{ac,sc}_k
$$

$$
\leq E^{ac,sc}_0 + h \sum_{i=0}^{k-1} t_i^2 |e_i|^2 + h \sum_{i=0}^{k-1} \left( 2|v_i - x^*| + \frac{t_i}{\sqrt{L}}|\nabla f(y_i)| \right) t_i |e_i|
$$

$$
\leq E^{ac,sc}_0 + h \sum_{i=0}^{k-1} t_i^2 |e_i|^2 + h \sum_{i=0}^{k-1} \left( 2|v_i - x^*| + \sqrt{L}t_i|y_i - x^*| \right) t_i |e_i|
$$

$$
\leq E^{ac,sc}_0 + h \sum_{i=0}^{k-1} t_i^2 |e_i|^2 + h \sum_{i=0}^{k-1} \left( (2 + \sqrt{L}t_i\lambda_h)|v_i - x^*| + \sqrt{L}t_i(1 - \lambda_h)|x_i - x^*| \right) t_i |e_i|
$$

Then, Gronwall’s Lemma gives a bound on $\sup_{k \geq 1} |v_k - x^*|$ and $\sup_{k \geq 1} k|x_k - x^*|$ which depends on (5.18) and is bounded as soon as (5.18) is. Then, the rest of the proof is similar to the unperturbed case, Corollary 5.5, taking advantage of the gap in the dissipation of $E^{ac,c}_k$, Proposition 5.17.

5.5. Variable time step: convergence in expectation. To conclude this section, we 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 (3.6). We do not assume strong convexity anymore, only convexity.

From (5.11) and Remark 5.12, we can deduce that if $x_k, v_k$ are generated by (Per-FE-C) with $h_k, t_k = \sum_{i=0}^{k} h_i$, and

$$
y_k = \left( 1 - \frac{2h_k}{t_k} \right) x_k + \frac{2h_k}{t_k} v_k,
$$

then (5.11) becomes

$$
E^{ac,c}_{k+1} - E^{ac,c}_k \leq -h_k^2 (f(x_k) - f^*) + \left( h_k - \frac{1}{\sqrt{L}} \right) t_k^2 h_k |\nabla f(y_k)|^2 + h_k \beta_k.
$$

Since $E[\beta_k] = h_k t_k^2 \sigma^2$, then

$$
E[E^{ac,c}_{k+1} - E^{ac,c}_k] \leq h_k t_k^2 \sigma^2,
$$

and

$$
E[E^{ac,c}_k] \leq E[f(x_k)] - f^*.
$$

which corresponds to (3.6) and (3.6) with

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

Then we can apply Proposition 3.23 to obtain the following convergence results.

**Proposition 5.19.** Assume $h_k = k^{-\alpha}$ and $t_k = \sum_{i=0}^{k} h_i$, then the following holds:

- If $\alpha \in \left( \frac{3}{4}, 1 \right)$, then
  $$
  E[f(x_k)] - f^* = O\left( \frac{1}{k^{2-2\alpha}} \right)
  $$

- If $\alpha = 1$, then
  $$
  E[f(x_k)] - f^* = O\left( \frac{1}{(\ln(k))^2} \right).
  $$

**Remark 5.20.** Compared to gradient descent, Proposition 4.5, the range of power $\alpha$ is smaller but for a fixed $h_k = k^{-\alpha}$, with $\alpha \in \left( \frac{3}{4}, 1 \right]$, then the rate in Proposition 5.19 is accelerated. Indeed,

accelerated rate = (GD rate)$^2$.  

6. 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 extend the obtained results to the perturbed case providing that the error decreases fast enough. At the end, we present a convergence rate for the expectation of $f$ in the case of the time step varies, $h_k$, and the error has a fixed variance.

6.1. Unperturbed gradient: continuous and discrete time. Define the perturbed ODE by setting $z = (x, v)$ and $g(t, x, v, \nabla f(y^*); \epsilon) = g_1(t, x, v; \epsilon) + g_2(t, x, v)\nabla f(y^*)$ as follows

$$g_1(t, x, v; \epsilon) = \left( \frac{-\epsilon \sqrt{\mu}}{1 + \epsilon \sqrt{\mu}}, \frac{-\epsilon \sqrt{\mu}}{1 + \epsilon \sqrt{\mu}} \right) (x) \quad \text{and} \quad g_2(t, x, v) = \left( \frac{-\sqrt{\mu}}{\sqrt{\mu}} \right).$$

with,

$$y^* = x + \frac{\epsilon \sqrt{\mu}}{1 + \epsilon \sqrt{\mu}} (v - x).$$

Then, we recall that the second order equation with Hessian damping (H-ODE-SC) can be rewritten as

$$\begin{align*}
\text{(1st-ODE-SC)} & \quad \begin{cases}
\dot{x} = \sqrt{\mu}(v - x) - \frac{h}{\sqrt{\mu}} \nabla f(x), \\
\dot{v} = \sqrt{\mu}(x - v) - \frac{h}{\sqrt{\mu}} \nabla f(x),
\end{cases}
\end{align*}$$

which is the case where $\epsilon = 0$, and given a small time step $h > 0$, the $\epsilon = h$ perturbation of the forward Euler method for (1st-ODE-SC) gives

$$\begin{align*}
(x_{k+1}, v_{k+1}) &= (x_k - \lambda_h v_k, v_k) - \frac{h}{\sqrt{\mu}} \nabla f(y_k), \\
\text{FE-C}\quad y_k &= (1 - \lambda_h) x_k + \lambda_h v_k, \quad \lambda_h = \frac{h \sqrt{\mu}}{1 + h \sqrt{\mu}}.
\end{align*}$$

Now define the Lyapunov function associated to this problem.

**Definition 6.1.** Define the continuous time Lyapunov function, $E_{ac,sc}^{ac,sc}$, by

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

and the discrete in time Lyapunov function by

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

In the next proposition, we show that $E_{ac,sc}^{ac,sc}$ is a rate-generating Lyapunov function, in the sense of Definition 3.21, for system (1st-ODE-SC) and its explicit discretization (FE-SC).

**Proposition 6.2.** Suppose $f$ is $\mu$-strongly convex and $L$-smooth. Let $(x, v)$ be a solution of (1st-ODE-SC) and $(x_k, v_k)$ be sequences generated by (FE-SC). Let $E_{ac,sc}^{ac,sc}$ be given by (6.1). Then $E_{ac,sc}^{ac,sc}$ is a continuous Lyapunov function with $r_{E_{ac,sc}^{ac,sc}} = \sqrt{\mu}$ and $a_{E_{ac,sc}^{ac,sc}} = \frac{1}{\sqrt{\mu}}$, i.e.

$$\frac{d}{dt} E_{ac,sc}^{ac,sc}(x, v) \leq -\sqrt{\mu} E_{ac,sc}^{ac,sc}(x, v) - \frac{1}{\sqrt{L}} |\nabla f(x)|^2 - \frac{\mu \sqrt{\mu}}{2} |v - x|^2.$$

and $E_{k+1}^{ac,sc}$ is a discrete Lyapunov function, with $L_{E_{ac,sc}^{ac,sc}} = 1$, for the sequence generated by (FE-C), for all $k \geq 0$,

$$E_{k+1}^{ac,sc} \leq (1 - h \sqrt{\mu}) E_k^{ac,sc} + \left( h^2 - \frac{h}{\sqrt{\mu}} \right) |\nabla f(y_k)|^2 + \left( \frac{h \sqrt{\mu} L}{2} - \frac{\sqrt{\mu}}{2h} \right) |x_k - y_k|^2.$$

for $h \leq \frac{1}{\sqrt{\mu}}$.

Then we retrieve the usual optimal rates in the continuous and discrete cases.
Corollary 6.3. Let $f$ be a $\mu$-strongly convex and $L$-smooth function. Let $(x(t), v(t))$ be a solution to (1st-ODE), then for all $t > 0$,
\[
    f(x(t)) - f^* + \frac{\mu}{2} |v(t) - x^*|^2 \leq \exp(-\sqrt{\mu}t) E^{ac,sc}(x_0, v_0).
\]
Furthermore, let $x_k, v_k$ be given by (FE-C). Then for all $k \geq 0$ and $h \leq \frac{1}{\sqrt{L}}$,
\[
    f(x_k) - f^* + \frac{\mu}{2} |v_k - x^*|^2 \leq (1 - h \sqrt{\mu})^k \left( f(x_0) - f^* + \frac{\mu}{2} |v_0 - x^*|^2 \right).
\]

The proof of Corollary 6.3 results immediately from Proposition 6.2 and then we focus on the proof of (6.2) and (6.2) in the following.

The discrete Lyapunov function $E^{ac,sc}$ was used to prove a rate in the strongly convex case by [WRJ16]. The proof of (6.2) can be found in [WRJ16, Theorem 6]. For completeness we also provide the proof. We split the proof in two parts: first we prove the gap inequality in the continuous case and then we consider the discrete case.

Proof of (6.2). Using (1st-ODE-SC), we obtain
\[
    \frac{d}{dt} E^{ac,sc}(x, v) = \langle \nabla f(x), \dot{x} \rangle + \sqrt{\mu} \langle v - x^*, \dot{v} \rangle
\]
\[
    = \sqrt{\mu} \langle \nabla f(x), v - x \rangle - \frac{1}{\sqrt{L}} \| \nabla f(x) \|^2 - \mu \sqrt{\mu} \langle v - x^*, v - x \rangle - \sqrt{\mu} \langle \nabla f(x), v - x^* \rangle
\]
\[
    = -\sqrt{\mu} \langle \nabla f(x), x - x^* \rangle - \frac{1}{\sqrt{L}} \| \nabla f(x) \|^2 - \frac{\mu \sqrt{\mu}}{2} \left[ |v - x^*|^2 + |v - x|^2 - |x - x^*|^2 \right]
\]

By strong convexity, we have
\[
    \frac{d}{dt} E^{ac,sc}(x, v) \leq -\sqrt{\mu} \left( f(x) - f^* + \frac{\mu}{2} |x - x^*|^2 \right) - \frac{1}{\sqrt{L}} \| \nabla f(x) \|^2
\]
\[
    - \frac{\mu \sqrt{\mu}}{2} \left[ |v - x^*|^2 + |v - x|^2 - |x - x^*|^2 \right]
\]
\[
    \leq -\sqrt{\mu} E^{ac,sc}(x, v) - \frac{1}{\sqrt{L}} \| \nabla f(x) \|^2 - \frac{\mu \sqrt{\mu}}{2} |v - x|^2.
\]

which establishes (6.2). \qed

In order to prove the gap inequality in the discrete case, note, since the gradients are evaluated at $y_k$, not $x_k$, the first step is to use strong convexity and $L$-smoothness to estimate the differences of $E^k_{ac,sc}$ in terms of gradients evaluated at $y_k$.

Lemma 6.4. Suppose that $f$ is a $\mu$-strongly convex and $L$-smooth function, then
\[
    f(x_{k+1}) - f(x_k) \leq \langle \nabla f(y_k), y_k - x_k \rangle - \frac{\mu}{2} |y_k - x_k|^2 + \frac{h}{2} \left( \frac{h}{\sqrt{L}} \right) \| \nabla f(y_k) \|^2.
\]

Proof. First, we remark that
\[
    f(x_{k+1}) - f(x_k) = f(x_{k+1}) - f(y_k) + f(y_k) - f(x_k)
\]
\[
    \leq \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L}{2} |x_{k+1} - y_k|^2
\]
\[
    + \langle \nabla f(y_k), y_k - x_k \rangle - \frac{\mu}{2} |y_k - x_k|^2.
\]

Since the first line of (1st-ODE-SC) can be rewritten as
\[
    x_{k+1} = y_k - \frac{h}{\sqrt{L}} \nabla f(y_k),
\]
we obtain (6.4). \qed

Now we can prove (6.2).
Proof of (6.2). In the proof we will estimate the linear term \( \langle y_k - x_k, \nabla f(y_k) \rangle \) in terms of \( \langle y_k - x^*, \nabla f(y_k) \rangle \) plus a correction which is controlled by the gap (the negative quadratic) in (6.4) and the quadratic term in \( E_{k}^{ac,sc} \).

The second term in the Lyapunov function gives, using 1-smoothness of the quadratic term in \( E_{k}^{ac,sc} \),

\[
\frac{\mu}{2} (|v_{k+1} - x^*|^2 - |v_k - x^*|^2) = \mu (v_k - x^*, v_{k+1} - v_k) + \frac{\mu}{2} |v_{k+1} - v_k|^2
\]

\[
= -\mu \lambda_h (v_k - x^*, v_k - x_k) + h \sqrt{\mu} (v_k - x^*, \nabla f(y_k)) + \frac{\mu}{2} |v_{k+1} - v_k|^2.
\]

Before going on, by definition of \( y_k \) in (FE-SC) as a convex combination of \( x_k \) and \( v_k \), we have

\[
\lambda_h (v_k - x_k) = \frac{\lambda_h}{1 - \lambda_h} (v_k - y_k) = h \sqrt{\mu} (v_k - y_k) \quad \text{and} \quad v_k - y_k = \frac{1 - \lambda_h}{\lambda_h} (y_k - x_k) = \frac{1}{h \sqrt{\mu}} (y_k - x_k)
\]

which gives

\[
\frac{\mu}{2} (|v_{k+1} - x^*|^2 - |v_k - x^*|^2) = -h \mu \sqrt{\mu} (v_k - x^*, v_k - y_k)
\]

\[
- h \sqrt{\mu} (v_k - y_k, \nabla f(y_k)) - h \sqrt{\mu} (y_k - x^*, \nabla f(y_k)) + \frac{\mu}{2} |v_{k+1} - v_k|^2
\]

\[
\leq -\frac{h \mu \sqrt{\mu}}{2} (|v_k - x^*|^2 + |v_k - y_k|^2 - |y_k - x^*|^2)
\]

\[
- (y_k - x_k, \nabla f(y_k)) - h \sqrt{\mu} f(y_k) - f^* + \frac{\mu}{2} (y_k - x^*)^2
\]

\[
+ \frac{\mu}{2} \left( |y_k - x_k|^2 + \frac{2h}{\sqrt{\mu}} (y_k - x_k, \nabla f(y_k)) + \frac{h^2}{\mu} |\nabla f(y_k)|^2 \right),
\]

by strong convexity. Then using the \( L \)-smoothness of \( f \), we obtain

\[
\frac{\mu}{2} (|v_{k+1} - x^*|^2 - |v_k - x^*|^2) \leq -h \sqrt{\mu} E_{k}^{ac,sc} - (y_k - x_k, \nabla f(y_k)) + \left( \frac{\mu}{2} + \frac{h \sqrt{\mu}}{2} - \frac{L}{2 \mu} \right) |y_k - x_k|^2 + \frac{h^2}{2 \mu} |\nabla f(y_k)|^2.
\]

Combining (6.4) and (6.1), we obtain (6.2). \( \square \)

Then, we can deduce from the gap in (6.2) an accelerated convergence for \( \min_{0 \leq i \leq k} |\nabla f(y_i)|^2 \) and \( \min_{0 \leq i \leq k} |x_i - y_i|^2 \) for \( h < \frac{1}{\sqrt{L}} \).

Corollary 6.5. For all \( h < \frac{1}{\sqrt{L}} \), \( k \geq 1 \),

\[
\min_{0 \leq i \leq k} |\nabla f(y_i)|^2 \leq \frac{\sqrt{L}}{h(1 - h \sqrt{L})} \left( \frac{h \sqrt{\mu}}{(1 - h \sqrt{\mu})^{-k} - 1} \right) E_{0}^{ac,cs},
\]

and

\[
\min_{0 \leq i \leq k} |x_i - y_i|^2 \leq \frac{2h}{\sqrt{\mu}(1 - h^2 L)} \left( \frac{h \sqrt{\mu}}{(1 - h \sqrt{\mu})^{-k} - 1} \right) E_{0}^{ac,cs}.
\]

Remark 6.6. The rate in Corollary (6.5) is better than the usual rate \( (1 - h \sqrt{\mu})^k \). Indeed, for \( k \geq 1 \),

\[
(1 - h \sqrt{\mu})^k (1 - h \sqrt{\mu})^{-k} - 1 = 1 - (1 - h \sqrt{\mu})^k \geq h \sqrt{\mu},
\]

and then,

\[
(1 - h \sqrt{\mu})^k \geq \frac{h \sqrt{\mu}}{(1 - h \sqrt{\mu})^{-k} - 1}.
\]
Proof. From (6.4), we have

\[
E^{ac,sc}_{k+1} - (1 - h\sqrt{\mu})^k E^{ac,sc}_0 = \sum_{i=0}^{k-1} (1 - h\sqrt{\mu})^{k-1-i} (E^{ac,sc}_{i+1} - (1 - h\sqrt{\mu})E^{ac,sc}_i)
\]

\[
\leq \sum_{i=0}^{k-1} (1 - h\sqrt{\mu})^{k-1-i} \left( (h^2 - \frac{h}{\sqrt{L}}) |\nabla f(y_i)|^2 + \left( \frac{h\sqrt{\mu}L}{2} - \frac{\sqrt{\mu}}{2h} \right) |x_i - y_i|^2 \right)
\]

\[
\leq \min_{0 \leq i \leq k} \left( \left( \frac{h^2 - \frac{h}{\sqrt{L}}}{\sqrt{\mu}} \right) |\nabla f(y_i)|^2 + \left( \frac{h\sqrt{\mu}L}{2} - \frac{\sqrt{\mu}}{2h} \right) |x_i - y_i|^2 \right) \sum_{i=0}^{k-1} (1 - h\sqrt{\mu})^{k-1-i}.
\]

Therefore,

\[
\min_{0 \leq i \leq k} \left( \left( \frac{h^2 - \frac{h}{\sqrt{L}}}{\sqrt{\mu}} \right) |\nabla f(y_i)|^2 + \left( \frac{h\sqrt{\mu}L}{2} - \frac{\sqrt{\mu}}{2h} \right) |x_i - y_i|^2 \right) \leq \frac{h\sqrt{\mu}}{(1 - h\sqrt{\mu})^{k-1}} E^{ac,cs}_0,
\]

which concludes the proof. \(\square\)

Remark 6.7. In the proof we also show that for all \(h < \frac{1}{\sqrt{L}}\),

\[
\sum_{i=0}^{k-1} (1 - h\sqrt{\mu})^{i-1} |\nabla f(y_i)|^2 \leq \frac{\sqrt{L}}{h(1 - h\sqrt{L})} E^{ac,cs}_0,
\]

and

\[
\sum_{i=0}^{k-1} (1 - h\sqrt{\mu})^{i-1} |x_i - y_i|^2 \leq \frac{2h}{\sqrt{\mu}(1 - h^2 L)} E^{ac,cs}_0.
\]

6.2. Perturbed gradient: continuous time. Now, define the perturbed system of (1st-ODE-SC) by

(Per-1st-ODE-SC)

\[
\begin{cases}
\dot{x} = \sqrt{\mu}(v - x) - \frac{1}{\sqrt{L}}(\nabla f(x) + e(t)), \\
\dot{v} = \sqrt{\mu}(x - v) - \frac{1}{\sqrt{\mu}}(\nabla f(x) + e(t)).
\end{cases}
\]

where \(e\) is a locally integrable function.

Definition 6.8. Define the continuous in time perturbed Lyapunov function \(\tilde{E}^{ac,sc}\) by

\[
\tilde{E}^{ac,sc}(t, x, v) := E^{ac,sc}(x, v) + I^{ac,sc}(t, x(\cdot), v(\cdot)),
\]

where \(E^{ac,sc}\) is given by (6.1) and

\[
I^{ac,sc}(t, x(\cdot), v(\cdot)) := \exp\left(\frac{-\sqrt{\mu}t}{2}\right) |\nabla f(x)|^2 - \frac{\sqrt{\mu}}{2} |v - x|^2.
\]

The next proposition gives the dissipation of \(\tilde{E}^{ac,sc}\).

Proposition 6.9. Let \((x, v)\) be a solution of (Per-1st-ODE-SC) with initial condition \((x(0), v(0)) = (x_0, v_0)\), then

\[
\frac{d}{dt} \tilde{E}^{ac,sc}(t, x, v) \leq -\sqrt{\mu} \tilde{E}^{ac,sc} - \frac{1}{\sqrt{L}} |\nabla f(x)|^2 - \frac{\sqrt{\mu}L}{2} |v - x|^2,
\]

and

\[
f(x(t)) - f^* + \frac{\mu}{2} |v_k - x|^2 \leq \exp(-\sqrt{\mu}t) \left( E^{ac,sc}(x_0, v_0) + J(t, x(\cdot), v(\cdot)) \right).
\]
Proof. Using (6.2), we obtain
\[
\frac{d}{dt} \tilde{E}^{ac,sc}(t, x, v) \leq \frac{d}{dt} E^{ac,sc}(x, v) - \sqrt{\mu} E^{ac,sc}(t, v) + \langle \sqrt{\mu}(v - x^*) + \frac{1}{\sqrt{L}} \nabla f(x), e \rangle
\]
\[
 \leq - \sqrt{\mu} E^{ac,sc}(x, v) - \langle \sqrt{\mu}(v - x^*) + \frac{1}{\sqrt{L}} \nabla f(x), e \rangle - \frac{1}{\sqrt{L}} |\nabla f(x)|^2 - \frac{\sqrt{\mu}}{2} |v - x|^2
\]
\[
- \sqrt{\mu} E^{ac,sc}(t, v) + \langle \sqrt{\mu}(v - x^*) + \frac{1}{\sqrt{L}} \nabla f(x), e \rangle
\]
\[
 \leq - \sqrt{\mu} E^{ac,sc}(t, x, v) - \frac{1}{\sqrt{L}} |\nabla f(x)|^2 - \frac{\sqrt{\mu}}{2} |v - x|^2
\]

Then, \( t \mapsto \tilde{E}^{ac,sc}(t, x(t), v(t)) \) is decreasing and the last inequality follows directly.

Then, we obtain:

**Corollary 6.10.** Let \((x, v)\) be a solution of (Per-1st-ODE-SC) with initial condition \((x(0), v(0)) = (x_0, v_0)\). Then,
\[
\sup_{0 \leq s \leq t} 2|v(s) - x^*| + \frac{1}{\sqrt{L}} |\nabla f(x(s))| \leq M(t),
\]
where \(M(t)\) depends on
\[
\int_0^t \exp(\sqrt{\mu}s)|e(s)| \, ds.
\]
Assume that (6.10) is finite, then \(M \in L^\infty(\mathbb{R}_+)\) and, \(f(x(t)) - f^* + \frac{\mu}{2} |v(t) - x^*|^2 \leq \exp(-\sqrt{\mu}t) \left(E^{ac,sc}(x_0, v_0) + M(t) \int_0^t \exp(\sqrt{\mu}s)|e(s)| \, ds\right) = O(\exp(-\sqrt{\mu}t)).
\]

**Proof.** By Proposition 6.9, we have
\[
f(x(t)) - f^* + \frac{\mu}{2} |v(t) - x^*|^2 \leq \exp(-\sqrt{\mu}t) \left[f(x_0) - f^* + \frac{\mu}{2} |v_0 - x^*|^2
\]
\[
+ \int_0^t \left|\sqrt{\mu}(v(s) - x^*) + \frac{1}{\sqrt{L}} \nabla f(x) \right| \exp(\sqrt{\mu}s)|e(s)| \, ds\right].
\]
The first part of the proof is due to Gronwall’s Lemma and then, we conclude the proof. \(\square\)

### 6.3. Perturbed gradient: discrete time

Consider the \(h\)-perturbation of the forward Euler discretization of (Per-1st-ODE-SC) given by
\[
\begin{cases}
  x_{k+1} - x_k = \lambda_h (v_k - x_k) - \frac{h}{\sqrt{L}} (\nabla f(y_k) + e_k), \\
  v_{k+1} - v_k = \lambda_h (x_k - v_k) - \frac{h}{\sqrt{\mu}} (\nabla f(y_k) + e_k), \\
  y_k = (1 - \lambda_h)x_k + \lambda_h v_k,
\end{cases}
\]
(Per-FE-SC)

where \(e_k\) is a given error.

Inspired by the continuous framework, define the discrete perturbed Lyapunov function, \(\tilde{E}^{ac,sc}_k\).

**Definition 6.11.** Define \(\tilde{E}^{ac,sc}_k := E^{ac,sc}_k + I^{ac,sc}_k\), where \(E^{ac,sc}_k\) is given by (6.1) and
\[
I^{ac,sc}_k := -h (1 - h \sqrt{\mu})^k \sum_{i=0}^{k-1} (1 - h \sqrt{\mu})^{-i-1} \beta_i
\]
\[
= -h (1 - h \sqrt{\mu})^k J^{ac,sc}_k,
\]
where,
\[
\beta_i := -\left(\sqrt{\mu}(x_k - y_k + v_k - x^*) - \frac{1}{\sqrt{L}} \nabla f(y_k)\right) + 2h \left(\nabla f(y_i) + \frac{e_i}{2}, e_i\right).
\]

Then we obtain the following convergence result for sequences generated by (Per-FE-SC).
Proposition 6.12. Let \( x_k, v_k \) be two sequences generated by the scheme (Per-FE-SC) with initial condition \((x_0, v_0)\). Suppose that \( h \leq \frac{1}{\sqrt{L}} \), then

\[
E_{k+1}^{ac,sc} \leq (1 - h \sqrt{\mu}) E_k^{ac,sc} + \left( h^2 - \frac{h}{\sqrt{L}} \right) |\nabla f(y_k)|^2 + \left( \frac{h^2}{2} - \frac{h}{\sqrt{L}} \right) \frac{h \sqrt{\mu}}{2} |x_k - y_k|^2 + h \beta_k.
\]

Then, \( \tilde{E}_k^{ac,sc} \) is decreasing and,

\[
f(x_k) - f^* + \frac{\mu}{2} |v_k - x^*|^2 \leq (1 - h \sqrt{\mu})^k (E_0^{ac,sc} + J_k^{ac,sc}),
\]

Proof of Proposition 6.12. First, arguing as in Proposition 6.2,

\[
f(x_{k+1}) - f(x_k) \leq \langle \nabla f(y_k), y_k - x_k \rangle - \frac{\mu}{2} |y_k - x_k|^2 + \left( h^2 - \frac{h}{\sqrt{L}} \right) |\nabla f(y_k)|^2
\]

and,

\[
\frac{\mu}{2} |v_{k+1} - x^*|^2 - \frac{\mu}{2} |v_k - x^*|^2 \leq -h \sqrt{\mu} E_k^{ac,sc} + \left( \sqrt{\mu} + Lh \right) \frac{h \sqrt{\mu}}{2} |x_k - y_k|^2 + \frac{h^2}{2} |\nabla f(y_k)|^2
\]

\[-h \sqrt{\mu} \langle v_k - x^* + x_k - y_k, e_k \rangle + h^2 \langle \nabla f(y_k) + \frac{e_k}{2}, e_k \rangle.
\]

Summing these two inequalities,

\[
E_{k+1}^{ac,sc} - E_k^{ac,sc} \leq -h \sqrt{\mu} E_k^{ac,sc} + \left( \frac{h^2}{2} - \frac{h}{\sqrt{L}} \right) |\nabla f(y_k)|^2 + \left( \frac{h \sqrt{\mu} L}{2} - \frac{h \sqrt{\mu}}{2h} \right) |x_k - y_k|^2
\]

\[-h \sqrt{\mu} \langle x_k - y_k + v_k - x^*, e_k \rangle + \frac{h}{\sqrt{L}} \langle \nabla f(y_k) + e_k, e_k \rangle.
\]

For the term \( I_k \), we obtain

\[
I_{k+1}^{ac,sc} - I_k^{ac,sc} \leq -h \sqrt{\mu} I_k^{ac,sc} - h \beta_k.
\]

Putting all together, we obtain (6.12) and then \( \tilde{E}_k^{ac,sc} \) is decreasing which concludes the proof. \( \Box \)

Remark that \( x_k - y_k + v_k - x^* = \lambda h (x_k - x^*) + (1 - \lambda h) (v_k - x^*) \) and \( |\nabla f(y_k)| \leq L |y_k - x^*| \leq L (1 - \lambda h) |x_k - x^*| + L \lambda h |v_k - x^*| \), then

\[
|\sqrt{\mu} (x_k - y_k + v_k - x^*) - \frac{1}{\sqrt{L}} \nabla f(y_i)| \leq (\lambda h (1 - L) + L) |x_k - x^*| + (1 + \lambda h (L - 1)) |v_k - x^*|.
\]

Therefore, we achieve the same rate as in the unperturbed case.

Corollary 6.13. Under the assumptions of Proposition 6.12, then \( \sup_{0 \leq i \leq k} |v_i - x^*| + \sup_{i \geq 0} |x_i - x^*| \leq M_k \), where \( M_k \) depends on

\[
\sum_{i=0}^{k-1} (1 - h \sqrt{\mu})^{-i} |e_i|.
\]

If (6.13) is bounded when \( k \searrow +\infty \), then \( M_k \in L^\infty \) and

\[
f(x_k) - f^* + \frac{\mu}{2} |v_k - x^*|^2 \leq (1 - h \sqrt{\mu})^k \left( E_0^{ac,sc} + CM_k \sum_{i=0}^{k-1} (1 - h \sqrt{\mu})^{-i} |e_i| \right) = O ((1 - h \sqrt{\mu})^k) \).
\]

Proof. Since \( J_k^{ac,sc} \) satisfies

\[
J_{k+1}^{ac,sc} \leq h \sum_{i=0}^{k-1} |\sqrt{\mu} (x_k - y_k + v_k - x^*) - \frac{1}{\sqrt{L}} \nabla f(y_i)| (1 - h \sqrt{\mu})^{-i-1} |e_i| \]

\[
\leq h \sum_{i=0}^{k-1} ((\lambda h (1 - L) + L) |x_k - x^*| + (1 + \lambda h (L - 1)) |v_k - x^*|) (1 - h \sqrt{\mu})^{-i-1} |e_i|.
\]
Combine with \( E_k^{ac,sc} \geq \frac{h^2}{2} (|x_k - x^*|^2 + |v_k - x^*|^2) \), we conclude, as in the convex case, applying discrete Gronwall’s lemma and (6.13).

Moreover, taking advantage of the gap in the dissipation of \( \tilde{E}_k^{ac,sc} \), we obtain

**Corollary 6.14.** For all \( h < \frac{1}{\sqrt{L}} \), \( k \geq 1 \),

\[
\min_{0 \leq i \leq k} |\nabla f(y_i)|^2 \leq \frac{\sqrt{L}}{h(1-h\sqrt{L})} \left( \frac{h\sqrt{\mu}}{(1-h\sqrt{\mu})^{-k}-1} \right) (E_0^{ac,cs} + I_\infty),
\]

and

\[
\min_{0 \leq i \leq k} |x_i - y_i|^2 \leq \frac{2h}{\sqrt{\mu}(1-h^2L)} \left( \frac{h\sqrt{\mu}}{(1-h\sqrt{\mu})^{-k}-1} \right) (E_0^{ac,cs} + I_\infty).
\]

**Proof.** The proof is similar to the one of Corollary 6.5 using (6.12). \( \square \)

**Remark 6.15.** As in the deterministic case, we have

\[
\sum_{i=0}^{k-1} (1-h\sqrt{\mu})^{-i-1} |\nabla f(y_i)|^2 \leq \frac{\sqrt{L}}{h(1-h\sqrt{L})} (E_0^{ac,cs} + I_\infty),
\]

and

\[
\sum_{i=0}^{k-1} (1-h\sqrt{\mu})^{-i-1} |x_i - y_i|^2 \leq \frac{2h}{\sqrt{\mu}(1-h^2L)} (E_0^{ac,cs} + I_\infty).
\]

6.4. **Variable time step and convergence in expectation.** Now we consider the case of a variable time step \( h_k \) and a zero-mean and a constant variance error \( e_k \), (3.6).

As in the convex case, the proof of (6.12) is still true for an adaptative time step \( h_k \) and then, since \( \mathbb{E}[\beta_k] = \frac{h^2 \sigma^2}{2} \),

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

We are in the case where \( r_{E^{ac,sc}} = \sqrt{\mu} > 0 \), \( L_{E^{ac,sc}} \) and \( g_2^2 = 1 \) (Proposition 6.2), then Proposition 3.22 gives

**Proposition 6.16.** If

\[
h_k^{ac,sc} := \frac{2}{\sqrt{\mu} (k + \alpha^{-1} E_0^{ac,sc})}, \quad \alpha^{ac,sc} := \frac{\mu}{2 \sigma^2}
\]

and \( t_k = \sum_{i=0}^{k} h_i \), then the following holds:

\[
\mathbb{E}[E^{ac,sc}(x_k)] \leq \frac{1}{\alpha^{ac,sc} + E_0^{ac,sc}} = \frac{2 \sigma^2}{\mu k + 2 \sigma^2 E_0^{ac,sc}}.
\]

**Remark 6.17.** If we compare this result to the gradient descent case, Proposition 4.8, starting with \( E_0^{ac} = E_0^{ac,sc} = 0 \), we remark that since \( \alpha^{ac,sc} \geq \alpha^{sc} \), even if the time step has to be smaller in the accelerated case \( h_k^{ac,sc} \leq h_k^{sc} \) we obtain an accelerated rate:

\[
\frac{2 \sigma^2}{\mu k + 2 \sigma^2 E_0^{-1}} \leq \frac{1}{\alpha^{ac,sc} + E_0^{-1}} = \frac{1}{\alpha^{sc} + E_0^{-1}} = \frac{2(C_f + 1) \sigma^2}{\mu k + 2 (C_f + 1) \sigma^2 E_0^{-1}}.
\]

**Acknowledgments.** Research supported by AFOSR FA9550-18-1-0167 (A.O.).

**References**

[AABR02] Felipe Alvarez, Hedy Attouch, Jérôme 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.

[ACPR18] Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1-2):123–175, 2018.

[APR16] Hedy Attouch, Juan Peypouquet, and Patrick Redont. Fast convex optimization via inertial dynamics with hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016.
Appendix A. Proof of inequality (5.3)

The proof is a perturbed version of the one in [BT09, SBC14, APR16]. First we prove the following inequality:

Lemma A.1. Assume $f$ is a convex, L-smoothness function. For all $x, y, z$, $f$ satisfies

$$f(z) \leq f(x) + \langle \nabla f(y), z - x \rangle + \frac{L}{2} |z - y|^2. \tag{68}$$

Proof. By $L$-smoothness,

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

and since $f$ is convex,

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle.$$

We conclude the proof combining these two inequalities.

Now apply inequality (A.1) at $(x, y, z) = (x_k, y_k, x_{k+1})$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(y_k), x_{k+1} - x_k \rangle + \frac{L}{2} |x_{k+1} - y_k|^2$$

$$\leq f(x_k) + \frac{2h}{tk} \langle \nabla f(y_k), v_k - x_k \rangle - \frac{h}{\sqrt{L}} |\nabla f(y_k)|^2$$

$$- \frac{h}{\sqrt{L}} (\nabla f(y_k), e_k) + \frac{h^2}{2} |\nabla f(y_k) + e_k|^2,$$

and then

$$f(x_{k+1}) \leq f(x_k) + \frac{2h}{tk} \langle \nabla f(y_k), v_k - x_k \rangle - \left( \frac{1}{\sqrt{L}} - \frac{h}{2} \right) h |\nabla f(y_k)|^2 - \frac{h}{\sqrt{L}} (\nabla f(y_k), e_k) + h^2 \langle \nabla f(y_k) + e_k, e_k \rangle. \tag{69}$$

If we apply (A.1) also at $(x, y, z) = (x^*, y_k, x_{k+1})$ we obtain

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

$$\leq f^* + \langle \nabla f(y_k), y_k - x^* \rangle - \frac{h}{\sqrt{L}} |\nabla f(y_k)|^2$$

$$- \frac{h}{\sqrt{L}} (\nabla f(y_k), e_k) + \frac{h^2}{2} |\nabla f(y_k) + e_k|^2.$$
then,

\[(70) \quad f(x_{k+1}) \leq f^* + \langle \nabla f(y_k), y_k - x^* \rangle - \left( \frac{1}{\sqrt{L}} - \frac{h}{2} \right) h|\nabla f(y_k)|^2 - \frac{h}{\sqrt{L}} \langle \nabla f(y_k), e_k \rangle + h^2 \left\langle \nabla f(y_k) + \frac{e_k}{2}, e_k \right\rangle. \]

Summing \((1 - \frac{2h}{t_k})(A)\) and \(\frac{2h}{t_k}(A)\), we have

\[f(x_{k+1}) - f^* \leq \left(1 - \frac{2h}{t_k}\right) (f(x_k) - f^*) + \frac{2h}{t_k} \langle \nabla f(y_k), v_k - x^* \rangle - \left( \frac{1}{\sqrt{L}} - \frac{h}{2} \right) h|\nabla f(y_k)|^2
\]

\[-\frac{h}{\sqrt{L}} \langle \nabla f(y_k), e_k \rangle + h^2 \left\langle \nabla f(y_k) + \frac{e_k}{2}, e_k \right\rangle.\]

Then,

\[t_k^2 (f(x_{k+1}) - f^*) \leq (t_k - 2h) t_k (f(x_k) - f^*) + 2ht_k \langle \nabla f(y_k), v_k - x^* \rangle - \left( \frac{1}{\sqrt{L}} - \frac{h}{2} \right) ht_k^2 |\nabla f(y_k)|^2
\]

\[-\frac{ht_k^2}{\sqrt{L}} \langle \nabla f(y_k), e_k \rangle + h^2 t_k^2 \left\langle \nabla f(y_k) + \frac{e_k}{2}, e_k \right\rangle
\]

\[\leq (t_{k-1}^2 - h^2) (f(x_k) - f^*) + 2ht_k \langle \nabla f(y_k), v_k - x^* \rangle - \left( \frac{1}{\sqrt{L}} - \frac{h}{2} \right) ht_k^2 |\nabla f(y_k)|^2
\]

\[-\frac{ht_k^2}{\sqrt{L}} \langle \nabla f(y_k), e_k \rangle + h^2 t_k^2 \left\langle \nabla f(y_k) + \frac{e_k}{2}, e_k \right\rangle,\]

which concludes the proof.