NESTEROV'S METHOD IS THE DISCRETIZATION OF A DIFFERENTIAL EQUATION WITH HESSIAN DAMPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Su et al. (2014) made a connection between Nesterov's method and an ordinary differential equation (ODE). We show if a Hessian damping term is added to the ODE from Su et al. (2014), then Nesterov's method arises as a straightforward discretization of the modified ODE. Analogously, in the strongly convex case, a Hessian damping term is added to Polyak's ODE, which is then discretized to yield Nesterov's method for strongly convex functions. Despite the Hessian term, both second order ODEs can be represented as first order systems.

Established Liapunov analysis is used to recover the accelerated rates of convergence in both continuous and discrete time. Moreover, the Liapunov 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

Su et al. (2014) made a connection between Nesterov's method for a convex, L-smooth function, f, and the second order, ordinary differential equation (ODE)

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

However Su et al. (2014) did not show that Nesterov's method arises as a discretization of (A-ODE). In order to obtain such a discretization, we consider the following ODE, which has an additional Hessian damping term with coefficient $1/\sqrt{L}$.

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

Notice that (H-ODE) is a perturbation of (A-ODE), and the perturbation goes to zero as $L \to \infty$. Similar ODEs have been studied by Alvarez et al. (2002), they have been shown to accelerate gradient descent in continuous time in (Attouch et al., 2016).

Next, we consider the case where f is also μ-strongly convex, and write $C_f := L/\mu$ for the condition number of f. Then Nesterov's method in the strongly convex case arises as discretization of the following second order ODE

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

(H-ODE-SC) is a perturbation of Polyak's ODE (Polyak, 1964)

$$\ddot{x} + 2\sqrt{\mu}\dot{x} + \nabla f(x) = 0$$

which is accelerates gradient when f is *quadratic* see (Scieur et al., 2017).

In each case, both continuous and discrete, as well and convex and strongly convex, it is possible to provide a proof of the rate using a Liapunov function. These proofs are already established in the literature: we give citations below, and also provide proof in the Appendix.

Moreover, the analysis for Nesterov's method in the full gradient can be extended to prove acceleration in the case of stochastic gradients. Acceleration of stochastic gradient descent has been established by Lin et al. (2015) and Frostig et al. (2015), see also Jain et al. (2018). A direct acceleration method with a connection to Nestero'v method was done by Allen-Zhu (2017). Our analysis unifies the continuous time ODE with the algorithm, and includes full gradient acceleration as a special case. The analysis proceeds by first rewriting (H-ODE) (and (H-ODE-SC)) as first order systems involving ∇f, and then replacing the ∇f with $g = \nabla f + e$. Both the continuous and discrete time methods achieve the accelerated rate of convergence, provided $|e|$ goes to zero quickly enough. The condition on $|e|$, is given below in (12) and (13) - it is faster than the corresponding rate for stochastic gradient descent. When $e = 0$ we recover the full gradient case.

The renewed interested in the continuous time approach began with the work of Su et al. (2014) and was followed Wibisono et al. (2016); Wilson et al. (2016). Continuous time analysis also appears in Flammarion & Bach (2015), Lessard et al. (2016), and Krichene et al. (2015). However, continuous time approaches to optimization have been around for a long time. Polyak's method Polyak (1964) is related to successive over relaxation for linear equations (Varga, 1957) which were initially used to accelerate solutions of linear partial differential equations (Young, 1954). A continuous time interpretation of Newton's method can be found in (Polyak, 1987) or Alvarez et al. (2002). The mirror descent algorithm of Nemirovskii et al. (1983) has a continuous time interpretation (Bubeck et al., 2015). The Liapunov approach for acceleration had already appeared in Beck & Teboulle (2009) for FISTA.

The question of when discretizations of dynamical systems also satisfy a Liapunov function has been studied in the context of stabilization in optimal control Levant (1993). More generally, Stuart & Humphries (1996) studies when a discretization of a dynamical system preserves a property such as energy dissipation.

2 An ODE representation for Nesterov's method

2.1 Convex case

Despite the Hessian term, (H-ODE-SC) can be represented as the following first order system.

Lemma 2.1. *The second order ODE* (H-ODE) *is equivalent to the first order system*

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

Proof. 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

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

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

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

Simplify to obtain (H-ODE). □

The system (1st-ODE) can be discretized using the forward Euler method with a constant time step, h, to obtain Nesterov's method.

Definition 2.2. *Define y_k as the following convex combination of x_k and v_k.*

$$y_k = \frac{kx_k + 2v_k}{k+2}. \tag{1}$$

Let $h > 0$ be a given small time step/learning rate and let $t_k = h(k+2)$. The forward Euler method for (1st-ODE) *with gradients evaluated at y_k is given by*

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

Remark 2.3. *The forward Euler method simply comes from replacing $\dot{x}$ with $(x_{k+1} - x_k)/h$ and similarly for v. Normally the velocity field is simply evaluated at x_k, v_k. The only thing different about* (FE-C) *from the standard forward Euler method is that ∇f is evaluated at y_k instead of x_k. However, this is still an explicit method. More general multistep methods and one leg methods in this context are discussed in Scieur et al. (2017).*

Recall the standard Nesterov's method from Nesterov (2013, Section 2.2)

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

Theorem 2.4. *The discretization of* (H-ODE) *given by* (FE-C)(1) *with $h = 1/\sqrt{L}$ and $t_k = h(k+2)$ is equivalent to the standard Nesterov's method* (Nest).

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

$$x_{k+1} - x_k = \frac{2}{k+2}(v_k - x_k) - \frac{1}{L}\nabla f(y_k)$$
$$v_{k+1} - v_k = -\frac{k+2}{2L}\nabla f(y_k)$$

Eliminate the variable v using (1) to obtain (Nest). □

2.2 Strongly Convex Case

Now we consider μ-strongly convex, and L-smooth functions, f, and write $C_f := \frac{L}{\mu}$ for the condition number. We first show that (H-ODE-SC) can be represented as a first order system.

Lemma 2.5. *The second order ODE* (H-ODE-SC) *is equivalent to the first order system*

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

Proof. 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). □

System (1st-ODE-SC) can be discretized using a forward Euler method with a constant time step h to obtain Nesterov's method. Let $h > 0$ be a small time step, and apply the forward Euler method for (1st-ODE-SC) evaluated at y_k:

$$\begin{cases} x_{k+1} - x_k = \dfrac{h\sqrt{\mu}}{1+h\sqrt{\mu}}\lambda(v_k - x_k) - \dfrac{h}{\sqrt{L}}\nabla f(y_k), \\ v_{k+1} - v_k = \dfrac{h\sqrt{\mu}}{1+h\sqrt{\mu}}(x_k - v_k) - \dfrac{h}{\sqrt{\mu}}\nabla f(y_k) \end{cases} \tag{FE-SC}$$

where,

$$y_k = (1-\lambda_h)x_k + \lambda_h v_k, \qquad \lambda_h = \frac{h\sqrt{\mu}}{1+h\sqrt{\mu}}. \tag{2}$$

Now we recall the usual Nesterov's method for strongly convex functions from Nesterov (2013, Section 2.2)

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

Theorem 2.6. *The discretization of (H-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

$$\begin{aligned} x_{k+1} - x_k &= \frac{\sqrt{C_f{}^{-1}}}{1+\sqrt{C_f{}^{-1}}}(v_k - x_k) - \frac{1}{L}\nabla f(y_k) \\ v_{k+1} - v_k &= \frac{\sqrt{C_f{}^{-1}}}{1+\sqrt{C_f{}^{-1}}}(x_k - v_k) - \frac{1}{\sqrt{L\mu}}\nabla f(y_k) \end{aligned}$$

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

3 Liapunov Analysis

3.1 Convex case: continuous and discrete time

Definition 3.1. *Define the continuous time Liapunov function*

$$E(t, x, v) := t^2(f(x) - f^*) + 2|v - x^*|^2 \tag{3}$$

Define the discrete time Liapunov function E_k by

$$E_k = E(t_{k-1}, x_k, v_k) \tag{4}$$

Proposition 3.2. *Let f be a convex and L-smooth function. Let $(x(t), v(t))$ be a solution to (1st-ODE), then*

$$\frac{dE(t, x(t), v(t))}{dt} \leq -\frac{t^2}{\sqrt{L}}|\nabla f(x)|^2.$$

where $E(t, x, v)$ in given by (3). In particular, 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 $k \geq 0$,

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

In particular, if

$$h \leq \frac{1}{\sqrt{L}} \tag{5}$$

then E_k is decreasing. When equality holds in (5),

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

Most of the results stated above are already known, but for completeness we refer the proofs in Appendix A. Since (FE-C) is equivalent to Nesterov's method, the rate is known. The proof of the rate using a Liapunov function can be found in Beck & Teboulle (2009). Refer to **?** which shows that we can use the constant time step. The discrete Liapunov function (4) was used in Su et al. (2014); Attouch & Peypouquet (2016) to prove a rate.

3.2 Strongly convex case: continuous and discrete time

Definition 3.3. *Define the continuous time Liapunov function* $E(x, v)$

$$E(x, v) = f(x) - f^* + \frac{\mu}{2}|v - x^*|^2 \tag{6}$$

Define the discrete time Liapunov function by

$$E_k = E(x_k, v_v) = f(x_k) - f^* + \frac{\mu}{2}|v_k - x^*|^2. \tag{7}$$

Proposition 3.4. *Let* (x, v) *be the solution of* (1st-ODE-SC)*, then*

$$\frac{dE(x, v)}{dt} \leq -\sqrt{\mu}E(x, v) - \frac{1}{\sqrt{L}}|\nabla f(x)|^2 - \frac{\mu\sqrt{\mu}}{2}|v - x|^2. \tag{8}$$

In particular, for all $t > 0$*,*

$$E(x(t), v(t)) \leq \exp(-\sqrt{\mu}t)E(x_0, v_0).$$

Next, let x_k, v_k *be given by* (FE-SC) *with initial condition* (x_0, v_0)*. For* $h \leq \frac{1}{\sqrt{L}}$*, we have*

$$E_{k+1} - E_k \leq -h\sqrt{\mu}E_k. \tag{9}$$

In particular, for $h = \frac{1}{\sqrt{L}}$*,*

$$E_{k+1} \leq (1 - \sqrt{C_f{}^{-1}})E_k. \tag{10}$$

The discrete Liapunov function E_k was used to prove a rate in the strongly convex case by Wilson et al. (2016). The proof of (10) can be found in Wilson et al. (2016, Theorem 6). For completeness we also provide the proof in Appendix E.

4 Stochastic Accelerated method

In the appendix we present results in continuous and discrete time for (non-accelerated) stochastic gradient descent. We also present results in continuous time for the stochastic accelerated case in the Appendix.

We present the results in discrete time here.

4.1 Convex Stochastic Case: discrete time

In this section we consider stochastic gradients, which we write as a gradient plus an error term

$$\widetilde{\nabla} f(y_k) = \nabla f(y_k) + e_k \tag{11}$$

The stochastic gradient can be abstract, or it can error be a mini-batch gradient when f is a sum. Moreover, we can include the case where

$$e_k = \nabla f(\tilde{y}) - \nabla_I f(\tilde{y}) - (\nabla f(y_k) - \nabla_I f(y_k))$$

corresponding to a correction by a snapshot of the full gradient at a snapshot location, which is updated every m iterations, as inJohnson & Zhang (2013). The combination of gradient reduction and momentum was discussed in Allen-Zhu (2017).

In order to obtain the accelerated rate, our Liapuonov analysis requires that the $|e_i|$ be decreasing fast enough. This can also be accomplished in the minibatch setting by using larger minibatches. In this case, the rate of decrease of e_i required gives a schedule for minibatch sizes. A similar result was obtained in Attouch & Peypouquet (2016).

When we replace gradients with (11) the Forward Euler scheme (FE-C) becomes

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

where y_k is given by (1), h is a constant time step, and $t_k := h(k+2)$. In Appendix C, we study the continuous version of (Sto-FE-C) and obtain a rate of convergence using a Liapunov function.

Definition 4.1. *Define the discrete stochastic Liapunov function $\tilde{E}_k := E_k + I_k$, for $k \geq 0$, where E_k is given by (4) and and, $e_{-1} := 0$ and for $k \geq 0$,*

$$I_k := h \sum_{i=0}^{k} 2t_i \langle v_i - x^*, e_{i-1} \rangle .$$

Theorem 4.2. *Assume that the sequence e_k satisfies*

$$\sum_{i=1}^{+\infty} i|e_i| < +\infty \tag{12}$$

and set $h = \frac{1}{\sqrt{L}}$. Then, $\sup_{i\geq 1} |v_i - x^*| < +\infty$ and

$$\tilde{E}_{k+1} \leq \tilde{E}_k, \qquad k \geq 0$$

We immediately have the following result.

Corollary 4.3. *Suppose e_k satisfies (12) and $h = \frac{1}{\sqrt{L}}$. Then, for $k \geq 0$,*

$$f(x_k) - f^* \leq \frac{C}{(k+1)^2},$$

with

$$C = 2L((f(x_0) - f^*) + |v_0 - x^*|^2) + 2 \sup_{i\geq 1} |v_i - x^*| \sum_{i=0}^{+\infty} (i+3)|e_i|.$$

Remark 4.4. *The assumption on e_k is satisfied, for example, by a sequence of the form $|e_k| = 1/k^\alpha$ for any $\alpha > 2$. By comparison for SGD, the corresponding condition is satisfied by such sequences with $\alpha > 1$. Thus the norm of the noise needs to go to zero faster for accelerated SGD compared to regular SGD (see Appendix B) in order to obtain the rate.*

Remark 4.5. *In Theorem 4.2, we focus on the maximum possible time step $h = 1/\sqrt{L}$. The result is still true if we shrink the time step. In this case, I_k can be defined using the tails $h \sum_{i=k+1}^{\infty} 2t_i \langle v_i - x^*, e_{i-1} \rangle$, see Attouch & Peypouquet (2016).*

4.2 Strongly convex stochastic case: discrete time

In this section, we consider that stochastic gradient, which we write as a gradient plus an error, as in section 4.1. In Appendix B.2, we study the Stochastic gradient descent and Appendix C.2 is devoted to the analysis of the continuous framework of Stochastic Accelerated method. The Forward Euler scheme (FE-SC) becomes

$$\begin{cases} x_{k+1} - x_k = \lambda_h(v_k - x_k) - \dfrac{h}{\sqrt{L}}(\nabla f(y_k) + e_k), \\ v_{k+1} - v_k = \lambda_h(x_k - v_k) - \dfrac{h}{\sqrt{\mu}}(\nabla f(y_k) + e_k), \end{cases} \tag{Sto-FE-SC}$$

where e_k is a given error and

$$y_k = (1 - \lambda_h)x_k + \lambda_h v_k, \qquad \lambda_h = \frac{h\sqrt{\mu}}{1 + h\sqrt{\mu}}.$$

Inspired by the continuous framework (Appendix C.2), we define a discrete Lyapunov function.

Definition 4.6. *Define $\tilde{E}_k := E_k + I_k$, where E_k is given by (7) and*

$$I_k := h\sqrt{\mu}\,(1 - h\sqrt{\mu})^k \sum_{i=0}^{k} (1 - h\sqrt{\mu})^{-i} \langle v_i - x^*, e_{i-1} \rangle,$$

with the convention $e_{-1} = 0$.

Then we obtain the following convergence result for sequences generated by (Sto-FE-SC).

Theorem 4.7. *Let x_k, v_k be two sequences generated by the scheme (Sto-FE-SC) with initial condition (x_0, v_0). Suppose that $h = \frac{1}{\sqrt{L}}$ and the sequence $(e_k)_k$ satisfies*

$$\sum_{i=0}^{+\infty} (1 - \sqrt{{C_f}^{-1}})^{-i} e_i < +\infty. \tag{13}$$

Then,

$$\tilde{E}_{k+1} \leqslant (1 - \sqrt{{C_f}^{-1}})^k \tilde{E}_k.$$

In addition, $\sup_{i \geq 0} |v_i - x^*| \leq M$ *for a positive constant* M *and*

$$f(x_k) - f^* + \frac{\mu}{2}|v_k - x^*|^2 \leq A(1 - \sqrt{{C_f}^{-1}})^k,$$

with

$$A = f(x_0) - f^* + \frac{\mu}{2}|v_0 - x^*|^2 + M \sum_{i=0}^{+\infty} (1 - \sqrt{{C_f}^{-1}})^{-i} e_{i-1}$$

We include the proof of Theorem 4.7 since this result is new.

Proof of Theorem 4.7. First we prove that

$$\begin{aligned} E_{k+1} - E_k \quad \leq \quad & -\sqrt{{C_f}^{-1}} E_k \\ & -\sqrt{{C_f}^{-1}} \langle \lambda_h (x_k - v_k) - \frac{1}{\sqrt{\mu}} (\nabla f(y_k) + e_k), e_k \rangle \\ & -\sqrt{{C_f}^{-1}} \langle v_k - x^*, e_k \rangle \end{aligned}$$

For the term I_k, we obtain

$$\begin{aligned} I_{k+1} - I_k \quad \leq \quad & \sqrt{{C_f}^{-1}} (1 - \sqrt{{C_f}^{-1}})^k \Big((1 - \sqrt{{C_f}^{-1}}) \sum_{i=0}^{k+1} (1 - \sqrt{{C_f}^{-1}})^{-i} \langle v_i - x^*, e_{i-1} \rangle \\ & - \sum_{i=0}^{k} (1 - \sqrt{{C_f}^{-1}})^{-i} \langle v_i - x^*, e_{i-1} \rangle \Big) \\ = \quad & -\sqrt{{C_f}^{-1}} I_k + \sqrt{{C_f}^{-1}} \langle v_{k+1} - x^*, e_k \rangle. \end{aligned}$$

Putting all together, we obtain

$$\begin{aligned} \tilde{E}_{k+1} - \tilde{E}_k \quad = \quad & E_{k+1} - E_k + I_{k+1} - I_k \\ \leq \quad & -\sqrt{{C_f}^{-1}} \tilde{E}_k \\ & + \frac{1}{L}|e_k|^2 + \frac{\sqrt{{C_f}^{-1}}}{\sqrt{L}} \langle \lambda (v_k - x_k) + \frac{1}{\sqrt{\mu}} \nabla f(y_k), e_k \rangle \\ & + \sqrt{{C_f}^{-1}} \langle v_{k+1} - v_k, e_k \rangle \end{aligned}$$

And by definition of $v_{k+1} - v_k$, we have

$$\begin{aligned}\tilde{E}_{k+1} - \tilde{E}_k &\leq -\sqrt{C_f}^{-1}\langle \lambda_h(x_k - v_k) - \frac{1}{\sqrt{L\mu}}(\nabla f(y_k) + e_k), e_k\rangle \\ &\quad +\sqrt{C_f}^{-1}\langle \lambda_h(x_k - v_k) - \frac{1}{\sqrt{L\mu}}(\nabla f(y_k) + e_k), e_k\rangle \\ &\leq -\sqrt{C_f}^{-1}\tilde{E}_k.\end{aligned}$$

We conclude, as in the convex case, applying discrete Gronwall Lemma and (13). □

References

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

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.

Hedy Attouch and Juan Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM J. Optim.*, 26(3):1824–1834, 2016. ISSN 1052-6234. doi: 10.1137/15M1046095. URL https://doi.org/10.1137/15M1046095.

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.

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

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

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

Roy Frostig, Rong Ge, Sham Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *International Conference on Machine Learning*, pp. 2540–2548, 2015.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pp. 545–604, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2845–2853. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5843-accelerated-mirror-descent-in-continuous-and-discrete-time.pdf.

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

Arie Levant. Sliding order and sliding accuracy in sliding mode control. *International journal of control*, 58(6):1247–1263, 1993.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pp. 3384–3392, 2015.

Arkadii Nemirovskii, David Borisovich Yudin, and Edgar Ronald Dawson. Problem complexity and method efficiency in optimization. 1983.

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

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

Boris T Polyak. Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*, 1987.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. ISSN 00034851. URL http://www.jstor.org/stable/2236626.

Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. Integration methods and accelerated optimization algorithms. *arXiv preprint arXiv:1702.06751*, 2017.

AM Stuart and AR Humphries. *Dynamical systems and numerical analysis, volume 2 of Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 1996.

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

Richard S Varga. A comparison of the successive overrelaxation method and semi-iterative methods using chebyshev polynomials. *Journal of the Society for Industrial and Applied Mathematics*, 5 (2):39–46, 1957.

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

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

David Young. Iterative methods for solving partial difference equations of elliptic type. *Transactions of the American Mathematical Society*, 76(1):92–111, 1954.

A Continuous framework: ODE and rate

Proof of Prof 3.2. By definition of E, we have

$$\begin{aligned} \frac{dE}{dt} &\leq 2t(f(x)-f^*) + t^2\langle \nabla f(x), \dot{x}\rangle \\ &\quad +4\langle v - x^*, \dot{v}\rangle \\ &\leq 2t(f(x)-f^*) + 2t\langle \nabla f(x), v - x\rangle - \frac{t^2}{\sqrt{L}}|\nabla f(x)|^2 \\ &\quad -2t\langle v - x^*, \nabla f(x)\rangle \\ &\leq 2t(f(x)-f^* - \langle x - x^*, \nabla f(x)\rangle) - \frac{t^2}{\sqrt{L}}|\nabla f(x)|^2. \end{aligned}$$

The proof is concluded by convexity,

$$f(x) - f^* - \langle x - x^*, \nabla f(x)\rangle \leq 0. \qquad \square$$

Proof of Proposition 3.4. Using (1st-ODE-SC), we obtain

$$\begin{aligned}\frac{dE(x,v)}{dt} &= \langle \nabla f(x), \dot{x}\rangle + \lambda\sqrt{\mu}\langle v - x^*, \dot{v}\rangle \\ &= \sqrt{\mu}\langle \nabla f(x), v - x\rangle - \frac{1}{\sqrt{L}}|\nabla f(x)|^2 - \mu\sqrt{\mu}\langle v - x^*, v - x\rangle - \sqrt{\mu}\langle \nabla f(x), v - x^*\rangle \\ &= -\sqrt{\mu}\langle \nabla f(x), x - x^*\rangle - \frac{1}{\sqrt{L}}|\nabla f(x)|^2 - \frac{\mu\sqrt{\mu}}{2}\left[|v - x^*|^2 + |v - x|^2 - |x - x^*|^2\right]\end{aligned}$$

By strong convexity, we have

$$\begin{aligned}\frac{dE(x,v)}{dt} &\leq -\sqrt{\mu}\left(f(x) - f^* + \frac{\mu}{2}|x - x^*|^2\right) - \frac{1}{\sqrt{L}}|\nabla f(x)|^2 \\ &\quad -\frac{\mu\sqrt{\mu}}{2}\left[|v - x^*|^2 + |v - x|^2 - |x - x^*|^2\right] \\ &\leq -\sqrt{\mu}E(x,v) - \frac{1}{\sqrt{L}}|\nabla f(x)|^2 - \frac{\mu\sqrt{\mu}}{2}|v - x|^2.\end{aligned}$$

□

B Stochastic Gradient Descent

B.1 Convex case: continuous and discrete time

Let $e : [0, +\infty) \to \mathbb{R}^d$ be a integrable function. Consider the gradient descent

$$\dot{x} = -(\nabla f(x) + e(t)). \tag{14}$$

Then define the Lyapunov function, $\tilde{E}$, by

$$\tilde{E}(t,x) = E(t,x) + I(t),$$

where,

$$E(t,x) = t(f(x) - f^*) + \frac{1}{2}|x - x^*|^2,$$

and,

$$I(t) = \int_0^t \langle x(s) - x^* + s\nabla f(x(s)), e(s)\rangle\, ds.$$

Then the following result holds.

Proposition B.1. *Let x be a solution of (14) with initial condition x_0. Then,*

$$\frac{d\tilde{E}(t,x)}{dt} \leq -t|\nabla f(x)|^2.$$

In addition, if f is L-smooth, $\sup_{s\geq 0}|x(s) - x^*| < +\infty$, $\sup_{s\geq 0} s|\nabla f(x(s))| < +\infty$ and

$$f(x(t)) - f^* \leq \frac{1}{t}\left(f(x_0) - f^* + \frac{1}{2}|x_0 - x^*|^2 + \sup_{s\geq 0}|x(s) - x^* + s\nabla f(x(s))|\|e\|_{L^1(0,+\infty)}\right).$$

Proof. For all $t > 0$, we have

- $\frac{dE(t,x)}{dt} = (f(x) - f^* - \langle \nabla f(x), x - x^*\rangle) - t|\nabla f(x)| - \langle x - x^* + t\nabla f(x), e\rangle$,
- $\frac{dI(t)}{dt} = \langle x - x^* + t\nabla f(x), e\rangle$.

Then, since f is convex, we obtain the first result. We deduce that $\tilde{E}$ is decreasing. Arguing as Attouch et al. (2016) along with the co-coercivity inequality, we prove that $\sup_{s\geq 0}|x(s) - x^*| < +\infty$, $\sup_{s\geq 0} s|\nabla f(x(s))| < +\infty$ which concludes the proof. □

The discretization of (14) is

$$x_{k+1} - x_k = -h(\nabla f(x_k) + e_k), \tag{15}$$

where $e_k = e(hk)$.

Define $\tilde{E}_k$ by

$$\tilde{E}_k = E_k + I_k,$$

where, for $t_k := hk$,

$$E_k = t_k(f(x_k) - f^*) + \frac{1}{2}|x_k - x^*|^2,$$

and,

$$I_k = h\sum_{i=0}^{k-1}\langle x_i - x^* - t_{i+1}(\nabla f(x_i) + e_i), e_i\rangle.$$

Proposition B.2. *Let x_k be the sequence generated by (15) with initial condition x_0. Assume that h satisfies, for all $k \geq 0$,*

$$h(Lt_{k+1} + 1) - 2t_{k+1} \leq 0 \equiv h \leq \frac{1}{L}. \tag{16}$$

Then the $\tilde{E}_k$ is decreasing. In addition if $(e_k)_k$ and $(t_{k+1}|e_k|^2)_k$ are summable, $\sup_{i\geq 0}|x_i - x^*| < +\infty$, $\sup_{i\geq 0}|t_{i+1}\nabla f(x_i)| < +\infty$ and

$$f(x_k) - f^* \leq \frac{1}{t_k}\left[\frac{1}{2}|x_0 - x^*|^2 + \sup_{i\geq 0}|x_i - x^* + t_{i+1}\nabla f(x_i)|\sum_{i=0}^{+\infty}(|e_i| + t_{i+1}|e_i|^2)\right].$$

Proof. By L-smoothness and convexity of f, we have

$$\begin{aligned}
E_{k+1} - E_k \quad &\leq \quad -ht_{k+1}\langle\nabla f(x_k), \nabla f(x_k) + e_k\rangle \\
&\quad\quad +(Lt_{k+1} + 1)\frac{h^2}{2}|\nabla f(x_k) + e_k|^2 \\
&\quad\quad -\langle x_k - x^*, e_k\rangle \\
&\quad\quad +h(f(x_k) - f^* - \langle\nabla f(x_k), x_k - x^*\rangle) \\
&\leq \quad ((Lt_{k+1} + 1)h - 2t_{k+1})\frac{h}{2}|\nabla f(x_k) + e_k|^2 \\
&\quad\quad -h\langle x_k - x^*, e_k\rangle + t_{k+1}h\langle\nabla f(x_k) + e_k, e_k\rangle.
\end{aligned}$$

In addition,

$$I_{k+1} - I_k = h\langle x_k - x^* - t_{k+1}(\nabla f(x_k) + e_k), e_k\rangle,$$

therefore,

$$\tilde{E}_{k+1} - \tilde{E}_k \leq ((Lt_{k+1} + 1)h - 2t_{k+1})\frac{h}{2}|\nabla f(x_k) + e_k|^2 \leq 0,$$

when h satisfies (16). We conclude the proof with the same argument as Proposition B.1. □

B.2 STRONGLY CONVEX CASE: CONTINUOUS AND DISCRETE TIME

Let us study the equation

$$\dot{x} = -(\nabla f(x) + e(t)), \tag{17}$$

for an error function, e satisfying

$$\int_0^{+\infty} e^{\mu s}|e(s)|\, ds < +\infty. \tag{18}$$

This condition on the error function is classical Robbins & Monro (1951). The case $e = 0$ is satisfied trivially and corresponds to the gradient descent ODE.

We define the function $E : [0, +\infty) \times \mathbb{R}^d \to [0, +\infty)$ by

$$E(t, x) = \frac{1}{2}|x - x^*|^2 + I(t),$$

where,

$$I(t) = e^{-\mu t} \int_0^t e^{\mu s} \langle x(s) - x^*, e(s) \rangle \, ds.$$

Then we have the following result.

Proposition B.3. *Let x be a solution of (17) with initial data x_0 and suppose that e satisfies (18). Then,*

$$\frac{dE(t, x)}{dt} \leq -\mu E(t, x).$$

In addition, $\sup_{t \geq 0} |x - x^*| < +\infty$ *and*

$$\frac{1}{2}|x - x^*|^2 \leq e^{-\mu t} \left(\frac{1}{2}|x_0 - x^*|^2 + \sup_{s \geq 0} |x(s) - x^*| \int_0^{+\infty} e^{\mu s} |e(s)| \, ds \right).$$

Proof. For all $t > 0$,

$$\begin{aligned} \frac{dE(t, x)}{dt} &= -\langle x - x^*, \nabla f(x) \rangle - \langle x - x^*, e \rangle - \mu I(t) + \langle x - x^*, e \rangle \\ &\leq -\frac{\mu}{2}|x - x^*|^2 - \mu I(t) = -\mu E(t, x). \end{aligned}$$

Therefore $E(t, x(t))$ is decreasing and then for all $t > 0$,

$$\frac{1}{2}|x(t) - x^*|^2 \leq \frac{1}{2}|x_0 - x^*| + \int_0^t |x(s) - x^*| e^{\mu s} |e(s)| \, ds.$$

By Gronwall Lemma and (18), we deduce that $\sup_{t \geq 0} |x - x^*| < +\infty$ and the proof is concluded. $\square$

The discretization of (17) is

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

where $e_k = e(hk)$. We define E_k, for $k \geq 1$, by

$$E_k = \frac{1}{2}|x_k - x^*|^2 + I_k,$$

where,

$$I_k = (1 - h\mu)^k h \sum_{i=0}^{k} (1 - h\mu)^{-i} \langle x_i - x^*, e_{i-1} \rangle,$$

with the notation $e_{-1} = 0$.

Proposition B.4. *Assume that $h \leq \frac{1}{L}$. Then,*

$$E_{k+1} - E_k \leq -h\mu E_k.$$

In addition, if the sequence $(1 - h\mu)^{-i}|e_i|$ is summable, $\sup_{i \geq 1} |x_i - x^*| < +\infty$ *and we deduce,*

$$\frac{1}{2}|x_k - x^*|^2 \leq (1 - h\mu)^k \left(\frac{1}{2}|x_0 - x^*|^2 + h \sup_{i \geq 1} |x_i - x^*| \sum_{i=0}^{+\infty} (1 - h\mu)^{-i-1} |e_i| \right).$$

Proof. First, as usual, we have

$$\begin{aligned} \frac{1}{2}|x_{k+1} - x^*|^2 - \frac{1}{2}|x_k - x^*|^2 &= -h\langle \nabla f(x_k), x_k - x^* \rangle - h\langle e_k, x_k - x^* \rangle + \frac{h^2}{2}|\nabla f(x_k) + e_k|^2 \\ &\leq -\frac{h\mu}{2}|x_k - x^*|^2 + h(f^* - f(x_k)) + \frac{h^2}{2}|\nabla f(x_k) + e_k|^2 \\ &\leq -\frac{h\mu}{2}|x_k - x^*|^2 - \frac{h}{2L}|\nabla f(x_k)|^2 + \frac{h^2}{2}|\nabla f(x_k) + e_k|^2. \end{aligned}$$

In addition,

$$\begin{aligned} I_{k+1} - I_k &= h(1-h\mu)^k \left((1-h\mu) \sum_{i=0}^{k+1} (1-h\mu)^{-i} \langle x_i - x^*, e_{i-1} \rangle - \sum_{i=0}^{k} (1-h\mu)^{-i} \langle x_i - x^*, e_{i-1} \rangle \right) \\ &= -h\mu I_k + h\langle x_{k+1} - x^*, e_k \rangle. \end{aligned}$$

Combining these two inequalities,

$$\begin{aligned} E_{k+1} - E_k &\leq -h\mu E_k + h\langle x_{k+1} - x_k, e_k \rangle - \frac{h}{2L}|\nabla f(x_k)|^2 + \frac{h^2}{2}|\nabla f(x_k) + e_k|^2 \\ &\leq -h\mu E_k + \frac{h}{2}\left(h - \frac{1}{L}\right)|\nabla f(x_k)|^2 - \frac{h^2}{2}|e_k|^2 \\ &\leq -h\mu E_k, \end{aligned}$$

when $h \leq \frac{1}{L}$.

In order to conclude, we also need to establish that E_k is bounded below. That follows from discrete Gronwall's inequality, as was already done in the continuous case in Proposition B.3. □

C Stochastic accelerated continuous time

In this section, we consider that an error $e(t)$ is made in the calculation of the gradient.

C.1 Convex case

We study the following perturbation of system (1st-ODE),

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

where e is a function satisfying

$$\int_0^{+\infty} s|e(s)| < +\infty. \tag{19}$$

The corresponding ODE is

$$\ddot{x} + \frac{3}{t}\dot{x} + \frac{1}{\sqrt{L}} D^2 f(x) \cdot \dot{x} + \left(\frac{1}{t\sqrt{L}} + 1\right)\nabla f(x) = -\left(\frac{1}{t\sqrt{L}} + 1\right) e(t) - \frac{1}{\sqrt{L}} e'(t).$$

We follow the argument from Attouch et al. (2016, section 5) to define a Lyapunov function for this system. Let $\tilde{E}$ be defined by

$$\tilde{E}(t, x, v) = E(t, x, v) + I(t, x, v),$$

where,

$$E(t, x, v) = t^2(f(x) - f^*) + 2|v - x^*|^2,$$

and

$$I(t, x, v) = \int_0^t s\langle 2(v - x^*) + \frac{s}{\sqrt{L}} \nabla f(x), e(s) \rangle \, ds.$$

Lemma C.1. *Let (x, v) be a solution of (Sto-1st-ODE) with initial condition $(x(0), v(0)) = (x_0, v_0)$ and suppose that e satisfies (19). Then*

$$\frac{d\tilde{E}}{dt}(t, x, v) \leq -\frac{t^2}{\sqrt{L}} |\nabla f(x)|^2.$$

In addition, $\sup_{t \geq 0} |v(t) - x^*| < +\infty$ *and* $\sup_{t \geq 0} |t\nabla f(x)| < +\infty$.

Proof. Following the proof of Proposition 3.2, we have

$$\frac{dE}{dt}(t,x,v) \leq -\frac{t^2}{\sqrt{L}}|\nabla f(x)|^2 - \frac{t^2}{\sqrt{L}}\langle \nabla f(x), e(t)\rangle - 2t\langle v - x^*, e(t)\rangle.$$

In addition,

$$\frac{dI}{dt}(t,x,v) = \frac{t^2}{\sqrt{L}}\langle \nabla f(x), e(t)\rangle + 2t\langle v - x^*, e(t)\rangle.$$

Then,

$$\frac{d\tilde{E}}{dt} \leq -\frac{t^2}{\sqrt{L}}|\nabla f(x)|^2.$$

In particular, $\tilde{E}$ is decreasing and

$$t^2(f(x) - f^*) + 2|v - x^*|^2 \leq 2|x_0 - x^*|^2 - \int_0^t s\langle 2(v - x^*) + Cs\nabla f(x), e(s)\rangle\, ds.$$

Using the inequality of co-coercitivity, we obtain

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

Using (19), we conclude applying Gronwall Lemma. □

Then we deduce

Proposition C.2. *Let (x, v) be a solution of (Sto-1st-ODE) with initial condition $(x(0), v(0)) = (x_0, v_0)$ and suppose that e satisfies (19). Then,*

$$f(x(t)) - f^* \leq \frac{1}{t^2}\left(2|v_0 - x^*|^2 + \sup_{s\geq 0}\left|2(v(s) - x^*) + \frac{s}{\sqrt{L}}\nabla f(x(s))\right| \int_0^{+\infty} s|e(s)|\, ds\right).$$

C.2 Strongly convex stochastic case: continuous time

Define the perturbed system of (1st-ODE-SC) by

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

where e is a locally integrable function.

Definition C.3. *Define the continuous time Liapunov function $E(x, v)$*

$$E(x, v) = f(x) - f^* + \frac{\mu}{2}|v - x^*|^2 \tag{20}$$

Define the perturbed Liapunov function $\tilde{E}$, by

$$\tilde{E}(t, x, v) := E(x, v) + I(t, x, v),$$

$$I(t, x) := e^{-\sqrt{\mu}t}\int_0^t e^{\sqrt{\mu}s}\left\langle \sqrt{\mu}(v(s) - x^*) + \frac{1}{\sqrt{L}}\nabla f(x), e(s)\right\rangle ds.$$

Proposition C.4. *We have,*

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

Proof. Using (8), we obtain

$$\begin{aligned}
\frac{d}{dt}\tilde{E}(t,x,v) &\leq \frac{d}{dt}E(x,v) - \sqrt{\mu}I(t,x) + \langle \sqrt{\mu}(v - x^*) + \frac{1}{\sqrt{L}}\nabla f(x), e\rangle \\
&\leq -\sqrt{\mu}E(x,v) - \langle \sqrt{\mu}(v - x^*) + \frac{1}{\sqrt{L}}\nabla f(x), e\rangle - \frac{1}{\sqrt{L}}|\nabla f(x)|^2 - \frac{\sqrt{\mu}\mu}{2}|v - x|^2 \\
&\quad -\sqrt{\mu}I(t,x) + \langle \sqrt{\mu}(v - x^*) + \frac{1}{\sqrt{L}}\nabla f(x), e\rangle \\
&\leq -\sqrt{\mu}\tilde{E}(t,x,v) - \frac{1}{\sqrt{L}}|\nabla f(x)|^2 - \frac{\sqrt{\mu}\mu}{2}|v - x|^2
\end{aligned}$$

□

Lemma C.5. *Suppose f is bounded from below and $s \mapsto e^{\sqrt{\mu}s}e(s) \in L^1$. Let (x, v) be a solution of* (Sto-1st-ODE-SC)*, then $\sup_{t\geq 0}|v(t)-x^*| < +\infty$ and $\sup_{t\geq 0}|\nabla f(x)| < +\infty$.*

Proof. Same as Attouch et al. (2016, Lemma 5.2), using the fact that $\frac{1}{2L}|\nabla f(x)|^2 \leq f(x) - f^*$, $t \mapsto \tilde{E}(x(t), v(t))$ is decreasing and Gronwall's inequality. □

Then, combining the two previous result, we obtain:

Corollary C.6. *Suppose that $s \mapsto e^{\lambda s}e(s)$ is a $L^1(0, +\infty)$ function. Let (x, v) be a solution of* (Sto-1st-ODE-SC) *with initial condition $(x(0), v(0)) = (x_0, v_0)$. Then,*

$$f(x(t)) - f^* + \frac{\mu}{2}|v(t)-x^*|^2 \leq Ce^{-\lambda t},$$

where,

$$C = f(x_0) - f^* + \frac{\mu}{2}|v_0 - x^*|^2 + \|e^{\lambda s}e(s)\|_{L^1(0,+\infty)} \sup_{s\geq 0}\left|\sqrt{\mu}(v(s)-x^*) + \frac{1}{\sqrt{L}}\nabla f(x)\right|.$$

Proof. By Proposition C.4 and Gronwall's Lemma, we have

$$\tilde{E}(t, x(t), v(t)) \leq e^{-\sqrt{\mu}t}\tilde{E}(0, x_0, v_0).$$

This is equivalent to

$$\begin{aligned} f(x(t)) - f^* + \frac{\mu}{2}|v(t)-x^*|^2 \quad &\leq \quad e^{-\sqrt{\mu}t}\Big[f(x_0) - f^* + \frac{\mu}{2}|v_0 - x^*|^2 \\ &\quad + \int_0^t \left|\sqrt{\mu}(v(s)-x^*) + \frac{1}{\sqrt{L}}\nabla f(x)\right| |e^{\sqrt{\mu}s}g(s)|ds\Big] \\ &\leq \quad Ce^{-\sqrt{\mu}t}, \end{aligned}$$

which concludes the proof with Lemma C.5. □

D Proof Theorem 4.2

First, using the convexity and the L-smoothness of f, we obtain the following classical inequality (see Attouch & Peypouquet (2016) or Su et al. (2014) in the case $e_k = 0$),

$$\begin{aligned} f(x_{k+1}) - f^* \quad &\leq \quad \frac{k}{k+2}(f(x_k) - f^*) + \frac{2}{k+2}\langle \nabla f(y_k), v_k - x^*\rangle \\ &\quad + \frac{h}{\sqrt{L}}\langle e_k, \nabla f(y_k) + e_k\rangle + \left(\frac{h^2}{2} - \frac{h}{\sqrt{L}}\right)|\nabla f(y_k) + e_k|^2. \end{aligned}$$

Then, we have

$$\begin{aligned} t_k^2(f(x_{k+1}) - f^*) \quad &- \quad t_{k-1}^2(f(x_k) - f^*) \\ &\leq \quad \left(\frac{kt_k^2}{k+2} - t_{k-1}^2\right)(f(x_k) - f^*) + \frac{2t_k^2}{k+2}\langle \nabla f(y_k), v_k - x^*\rangle \\ &\quad + \frac{ht_k^2}{\sqrt{L}}\langle e_k, \nabla f(y_k) + e_k\rangle + \left(\frac{h}{2} - \frac{1}{\sqrt{L}}\right)ht_k^2|\nabla f(y_k) + e_k|^2. \end{aligned}$$

By defintion of v_{k+1}, we have

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

In addition,

$$I_{k+1} - I_k = 2ht_k\langle v_{k+1} - x^*, e_k\rangle.$$

Combining these three previous inequalities, we obtain

$$\begin{aligned}\tilde{E}_{k+1}-\tilde{E}_k &\leq -h^2(f(x_k)-f^*)+\left(h-\frac{1}{\sqrt{L}}\right)t_k^2h|\nabla f(y_k)+e_k|^2\\ &\quad +2ht_k\langle e_k,v_{k+1}-v_k\rangle+\frac{ht_k^2}{\sqrt{L}}\langle e_k,\nabla f(y_k)+e_k\rangle\\ &\leq -h^2(f(x_k)-f^*)+\left(h-\frac{1}{\sqrt{L}}\right)t_k^2h|\nabla f(y_k)+e_k|^2\\ &\quad \left(\frac{ht_k^2}{\sqrt{L}}-h^2t_k^2\right)\langle e_k,\nabla f(y_k)+e_k\rangle.\end{aligned}$$

Since $h=\frac{1}{\sqrt{L}}$, we deduce that $\tilde{E}_k$ is decreasing. In particular,

$$2|v_k-x^*|^2\leq 2|v_0-x^*|^2+\frac{1}{L}\sum_{i=0}^{k-1}|v_i-x^*|(i+3)|e_i|.$$

and the discrete version of Gronwall Lemma gives the result since $(i+3)|e_i|$ is a summable sequence due to (12).

E Proof of Proposition 3.4

To simplify, we denote $\lambda_h=\frac{h\sqrt{\mu}}{1+h\sqrt{\mu}}$. Note, however, 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 in terms of gradients evaluated at y_k.

Lemma E.1. *Suppose that f is a μ-stgrongly convex and L-smooth function, then*

$$f(x_{k+1})-f(x_k)\leq\langle\nabla f(y_k),y_k-x_k\rangle-\frac{\mu}{2}|y_k-x_k|^2+\frac{h}{2}\left(h-\frac{2}{\sqrt{L}}\right)|\nabla f(y_k)|^2. \tag{21}$$

Proof. First, we remark that

$$\begin{aligned}f(x_{k+1})-f(x_k) &= f(x_{k+1})-f(y_k)+f(y_k)-f(x_k)\\ &\leq \langle\nabla f(y_k),x_{k+1}-y_k\rangle+\frac{L}{2}|x_{k+1}-y_k|^2\\ &\quad +\langle\nabla f(y_k),y_k-x_k\rangle-\frac{\mu}{2}|y_k-x_k|^2.\end{aligned}$$

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 (21). □

Proof of Proposition 3.4. Once (9) is established, since the expression on the right hand side is monotone in h, the largest choice of h is given by $h=\frac{1}{\sqrt{L}}$, which leads immediately to (10).

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 (21) and the quadratic term in E.

The second term in the Liapunov function gives, using 1-smoothness of the quadratic term in E.

$$\begin{aligned}\frac{\mu}{2}\left(|v_{k+1}-x^*|^2-|v_k-x^*|^2\right) &= \mu\langle v_k-x^*,v_{k+1}-v_k\rangle+\frac{\mu}{2}|v_{k+1}-v_k|^2\\ &= -\mu\lambda_h\langle v_k-x^*,v_k-x_k\rangle\\ &\quad -h\sqrt{\mu}\langle v_k-x^*,\nabla f(y_k)\rangle\\ &\quad +\frac{\mu}{2}|v_{k+1}-v_k|^2.\end{aligned}$$

Before going on, using the fact from (2), that y_k is a convex combination of x_k and v_k, we have

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

which gives

$$\begin{aligned} \frac{\mu}{2}\left(|v_{k+1}-x^*|^2 - |v_k - x^*|^2\right) &= -h\mu\sqrt{\mu}\langle v_k - x^*, v_k - y_k\rangle \\ &\quad -h\sqrt{\mu}\langle v_k - y_k, \nabla f(y_k)\rangle - h\sqrt{\mu}\langle y_k - x^*, \nabla f(y_k)\rangle \\ &\quad \frac{\mu}{2}|v_{k+1} - v_k|^2 \\ &\le -\frac{h\mu\sqrt{\mu}}{2}\left(|v_k - x^*|^2 + |v_k - y_k|^2 - |y_k - x^*|^2\right) \\ &\quad -\langle y_k - x_k, \nabla f(y_k)\rangle - h\sqrt{\mu}\left(f(y_k) - f^* + \frac{\mu}{2}|y_k - x^*|^2\right) \\ &\quad \frac{\mu}{2}\left(|y_k - x_k|^2 + \frac{2h}{\sqrt{\mu}}\langle y_k - x_k, \nabla f(y_k)\rangle + \frac{h^2}{\mu}|\nabla f(y_k)|^2\right), \end{aligned}$$

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

$$\begin{aligned} \tfrac{\mu}{2}\left(|v_{k+1}-x^*|^2 - |v_k - x^*|^2\right) &\le -h\sqrt{\mu}E_k - \langle y_k - x_k, \nabla f(y_k)\rangle \\ &\quad +\left(\tfrac{\mu}{2} + \tfrac{h\sqrt{\mu}L}{2} - \tfrac{\sqrt{\mu}}{2h}\right)|y_k - x_k|^2 + \tfrac{h^2}{2}|\nabla f(y_k)|^2. \end{aligned} \tag{22}$$

Combining (21) and (22), we have

$$E_{k+1} - E_k \quad \le \quad -h\sqrt{\mu}E_k + \left(h^2 - \frac{h}{\sqrt{L}}\right)|\nabla f(y_k)|^2 + \left(\frac{h\sqrt{\mu}L}{2} - \frac{\sqrt{\mu}}{2h}\right)|x_k - y_k|^2$$

which concludes the proof of (9). □