A systematic approach to Lyapunov analyses of continuous-time models in convex optimization

Céline Moucer, Adrien Taylor, Francis Bach

A SYSTEMATIC APPROACH TO LYAPUNOV ANALYSES OF CONTINUOUS-TIME MODELS IN CONVEX OPTIMIZATION

CÉLINE MOUCER[†*], ADRIEN TAYLOR[†], AND FRANCIS BACH[†]

Abstract. First-order methods are often analyzed via their continuous-time models, where their worst-case convergence properties are usually approached via Lyapunov functions. In this work, we provide a systematic and principled approach to find and verify Lyapunov functions for classes of ordinary and stochastic differential equations. More precisely, we extend the performance estimation framework, originally proposed by Drori and Teboulle [9], to continuous-time models. We retrieve convergence results comparable to those of discrete methods using fewer assumptions and convexity inequalities, and provide new results for stochastic accelerated gradient flows.

Key words. Convex optimization, continuous-time models, first-order methods, worst-case analysis, performance estimation, stochastic differential equations, ordinary differential equations.

1. Introduction. Convex optimization is an important tool in the numerical analyst toolbox. It serves, among others, for framing modeling problems in data science and signal processing. A number of convex optimization problem take the form:

$$\min_{x\in\mathbf{R}^d} f(x), \tag{1.1}$$

where f is convex, differentiable, and $x \in \mathbf{R}^d$ contains the variables of the model. First-order methods are very popular to solve these problems, due to their attractive low cost per iteration, and to the fact that data science applications typically do not require very accurate solutions [6]. Gradient descent is a common first-order method, which starts from a point $x_0 \in \mathbf{R}^d$, its iterates are given by the simple recursion

$$x_{k+1} = x_k - \gamma\nabla f(x_k), \tag{1.2}$$

where $\gamma > 0$ is a step size. Denoting $\dot{X}_t = \frac{d}{dt}X_t$, gradient descent with small step sizes γ is directly related to the so-called gradient flow:

$$\dot{X}_t = -\nabla f(X_t),\ X_0 = x_0 \in \mathbf{R}^d, \tag{1.3}$$

where the solution X_t of the ODE verifies $X_{t_k} \approx x_k$ with the identification $t_k = \gamma k$. In numerical integration, gradient descent (1.2) is also known as the Euler explicit scheme for integrating the gradient flow. Recently, Su et al. [37] have interpreted Nesterov's accelerated gradient [25] in a similar fashion through its continuous version, paving the way to several continuous analyses of accelerated methods [34, 47, 46].

Many applications entail some randomness and require a stochastic modeling of the function f, which is often defined in terms of an expectation $f(x) = \mathbf{E}_\xi[\tilde{f}(x,\xi)]$. The function f is the expectation over some random variable ξ, and accounts some random modeling. When ξ is drawn uniformly from a finite set of possible samples $(\xi_1, ..., \xi_n)$, we have a finite sum $f(x) = \frac{1}{n}\sum_{k=1}^n \tilde{f}(x,\xi_k)$. As soon as the number of data points n is large, computing the gradient of a finite sum, as it is done in gradient-based methods, is possibly expensive (computing the gradient of each element of the

*Ecole Nationale des Ponts et Chaussées, Marne-la-Vallée, France.

†DI ENS, École normale supérieure, Université PSL, CNRS, INRIA, 75005 Paris, France (celine.moucer@inria.fr), (adrien.taylor@inria.fr), (francis.bach@inria.fr).

sum, which is possibly very large). Stochastic gradient descent (SGD) provides an alternative with lower computational burden per iteration, by evaluating only the gradient of a single $\tilde{f}(\cdot, \xi_{i_k})$ per iteration,

$$x_{k+1} = x_k - \gamma \nabla \tilde{f}(x_k, \xi_{i_k}),$$

where $\gamma > 0$ is the step size, ξ_{i_k} is drawn uniformly at random in $(\xi_1, ..., \xi_n)$, and thereby $\mathbf{E}_\xi[\nabla \tilde{f}(x_k, \xi_{i_k})] = f(x_k)$ is an unbiased estimate of the full gradient. Li et al. [21, 22] have proven a connection with stochastic differential equations (SDE):

$$dX_t = -\nabla f(X_t)dt + \sigma(X_t)dB_t,$$

where $\sigma(X_t)$ is a noise parameter connected to parameters of the method, that was further developed by Shi et al. [36, 48]. This connection has raised many questions regarding approximate equivalences between optimization methods and continuous-time models, and tools for analyzing convergence speeds of discrete methods. Usually, gradient flows and first-order methods are studied via worst-case convergence properties, that have to be verified for any function of a given class, and any trajectory generated by the ODE or the optimization method. In many cases, continuous approaches seem to allows for shorter and simpler proofs, together with intuitions on what can be expected from optimization methods.

The analysis of continuous-time models often relies on Lyapunov stability arguments, as in system theory and physics, where energy dissipation plays a crucial role. The existence of such Lyapunov functions provides direct convergence proof for ODEs under consideration. The main challenge in the Lyapunov approach is to find a suitable function that is decreasing along all trajectories generated by an ODE.

From an outsider point of view, these analyses are often seen as complicated and technical to reach. In this work, we remedy this problem by extending the systematic approach based on semidefinite programming (SDP) developed by Drori and Teboulle [9] for certifying convergence of optimization methods. This technique is referred as "performance estimation problems" (PEPs). The main contribution of this work is to provide a tool for simultaneously analyzing convergence of continuous-time models, and constructing Lyapunov functions suited to gradient flows in a systematic way, using small-sized SDPs reformulations. Furthermore, this procedure benefits from tightness properties, meaning that the feasibility of the SDP allows to conclude about the existence of some Lyapunov functions.

1.1. Lyapunov functions. When dealing with convergence rates of gradient flows, many proofs typically rely on a Lyapunov function. In control theory, such functions are common for studying stability of dynamical systems [17].

Definition 1.1. *Given a trajectory X_t , we call $\mathcal{V} : \mathbf{R}^d \to \mathbf{R}$ a Lyapunov function if it is differentiable and satisfies the conditions:*

1. $\mathcal{V}(x) = 0 \iff x = x_\star$,
2. $\mathcal{V}(X_t) \geqslant 0$,
3. $\frac{d}{dt}\mathcal{V}(X_t) \leqslant 0$.

Given an ODE starting from x_0 and a class of functions $\mathcal{F}$, a function $\mathcal{V}(\cdot)$ is a Lyapunov function if the inequality $\frac{d}{dt}\mathcal{V}(X_t) \leqslant 0$ is verified for all trajectories X_t generated by ODEs originating from functions $f \in \mathcal{F}$. There exist similar definitions of Lyapunov functions for discrete optimization methods [42, 32, 20].

Lyapunov functions are suited to deriving both linear and sublinear convergence rates. When looking for linear convergence rates (as we may expect for strongly convex functions), the third condition is typically replaced by $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -\tau\mathcal{V}(X_t)$, where τ depends on the class of functions and on the ODE.

With this approach, convergence guarantees highly depend on the choice of Lyapunov functions. For a specific ODE and class of functions $\mathcal{F}$, there are often multiple choices of valid Lyapunov functions. In this work, we look for Lyapunov functions in a class of quadratic functions that is popular both in the discrete [25, 10] and continuous time [37] literature.

1.2. Prior works. Lyapunov functions are common for analyzing continuous and discrete time models in convex optimization. Convergence proofs for Nesterov's accelerated gradient method typically rely on such Lyapunov functions [25], [10, Theorem 4.8]. In the recent work [3], the authors proposed Lyapunov-based analyses for many first-order methods, for linear and sublinear convergence rates. Continuous-time versions of optimization methods also often involve Lyapunov arguments, such as Nesterov's accelerated gradient flow introduced in [37], and its high-resolution ODEs for strongly convex functions proposed in [34], or accelerated mirror descent whose continuous dynamics was analyzed in [18].

Different techniques were developed to compute suitable Lyapunov functions. The authors of [46] proposed an approach based on Bregman Lagrangian for accelerated methods in potentially non-Euclidean settings, further developed in [47]. [8] directly derived Lyapunov functions from Hamiltonian equations describing dynamics of ODEs. Using similar conservation laws in a dilated coordinate system, [38] also generated Lyapunov functions in a principled way.

Given a class of functions and an optimization method, proving a convergence rate mostly consists in combining inequalities characterizing the class of functions at hand. Recently, the automated search for combination of inequalities formulated as semidefinite program was formalized by Drori and Teboulle [9], and led to the notion of performance estimation problems. Their work was followed up in [45, 44] to provide worst-case bounds in a principled way, and extended to the Lyapunov framework [42]. A competing strategy inspired by control theory was developed by [20, 15], where Lyapunov functions for discrete-time models are constructed using integral quadratic constraints (IQCs) and semidefinite programming; a similar approach was applied to continuous-time models in [11]. Connections between Lyapunov functions obtained via the IQC framework in continuous and discrete-time, were later highlighted by [32].

For stochastic differential equations (SDEs), convergence proofs can also be obtained through the Lyapunov approach, together with Ito's calculus. [26] analyzed both SGD, SAGA [7], and SVRG [16], for some well-chosen Lyapunov functions. [48, 49] extended the framework of [46] to the stochastic setting. To the best of our knowledge, a systematic way of verifying a Lyapunov functions in the stochastic setting has not been developed yet.

1.3. Contributions and organization. In this work, we are concerned with worst-case convergence analyses of ordinary and stochastic differential equations, for modeling optimization methods. We propose a principled approach to worst-case analyses based on Lyapunov functions, SDPs and Ito's calculus.

In Section 2, we extend the performance estimation approach developed for optimization methods to gradient flows, that originates from a (possibly strongly) convex function. In short, we find Lyapunov functions as feasible points of certain linear ma-

trix inequalities (LMIs). Building on the first part of this work for ODEs, we analyze continuous versions of stochastic optimization algorithms. All codes for numerical results are provided at https://github.com/CMoucer/PEP_ODEs.

Section 3 studies properties of trajectories generated by SDEs, as approximations to stochastic gradient methods. We obtain a simple version of the trade-off between forgetting the initial conditions and diminishing the noise, with and without averaging. It appears that decreasing step sizes, together with a non-uniform version of averaging, allows to reach an optimal trade-off. Our results match those obtained for the stochastic gradient method, in a compact way compared with discrete analyses.

In Section 4, we prove that accelerated gradient flows require diminishing step sizes to converge in the stochastic setting. In contrast to first-order stochastic gradient flow, averaging does not preserve convergence for fixed step size.

1.4. Assumptions. Throughout this work, functions to be minimized are convex (see Problem 1.1). Under this assumption, stationary points are global minimizers. We restrict ourselves to continuous-time versions of gradient descent, accelerated gradient descent and stochastic gradient descent. Such methods gather information about functions to be minimized by evaluating its (sub)gradient at past iterates.

Let us recall a few basic definitions and properties characterizing the classes of functions under consideration within the next sections. A function $f : \mathbf{R}^d \to \mathbf{R}$ is convex if for all $x, y \in \mathbf{R}^d$, and for all $\lambda \in [0,1]$, $f(\lambda x+(1-\lambda)y) \leqslant \lambda f(x)+(1-\lambda)f(y)$. We consider in particular the class of convex closed proper (CCP) functions (i.e., functions whose epigraphs are non-empty closed convex sets). For simplicity, we assume in addition differentiability of f, even if results do not require it for convex gradient differential inclusions [4, Section 3.2]. Then, f is convex if and only if for all $x, y \in \mathbf{R}^d$, $f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle$. Smoothness is a common assumption for analyzing optimization methods, that limits the growth rate of the function. A function f is L-smooth if the gradient is L-Lipschitz, that is if for any x, y,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

A differentiable function f is μ-strongly convex if for any $x, y \in \mathbf{R}^d$ it satisfies

$$\|\nabla f(x) - \nabla f(y)\| \geqslant \mu\|x - y\|.$$

Strong convexity ensures the function is not too flat, and the unicity of the minimizer $x_\star$. We denote by $\mathcal{F}_{\mu,L}$ the family of L-smooth μ-strongly closed convex proper functions, with $0 \leqslant \mu \leqslant L \leqslant +\infty$. Weaker assumptions than strong convexity are also encountered in the literature for analyzing gradient algorithms, and leads to similar convergence guarantees. Among them, Łojasiewicz [23] introduced the Łojasiewicz inequality, under which Polyak [28, Theorem 4] showed linear convergence of gradient descent. Other relaxed versions of strong convexity followed [5, 24].

2. A principled approach to Lyapunov functions for gradient flows. In this section, we study convergence properties of the gradient flow and its accelerated versions, via quadratic Lyapunov functions. We prove that verifying such a Lyapunov function can be formulated as a LMI. This framework allows to search for Lyapunov functions, and to derive convergence bounds for non-autonomous gradient flows.

2.1. Gradient flow. We consider the gradient flow

$$\dot{X}_t = -\nabla f(X_t),\ X_0 = x_0 \in \mathbf{R}^d.$$

Without further assumptions, the function f is decreasing along the trajectory X_t solution to the gradient flow. The Lyapunov function $\mathcal{V}(X_t) = f(X_t)$ has indeed a negative time-derivative $\frac{d}{dt}\mathcal{V}(X_t) = \dot{X}_t^\top \nabla f(X_t) = -\|\nabla f(X_t))\|^2$.

Under additional convexity assumptions, it is possible to derive Lyapunov functions in a principled way, and deduce worst-case convergence speeds of the gradient flow. As a first stage, let us consider gradient flows originating from strongly convex functions and establish linear (or exponential) convergence of the flows.

2.1.1. Minimizing strongly convex functions. Let f be μ-strongly convex ($\mu > 0$), admitting thus a unique minimizer $x_\star$ such that $f(x_\star) = f_\star$, and consider the Problem (1.1). Thanks to strong convexity, it is possible to prove linear convergence of the gradient flow to its stationary point. Scieur et al. proved in [33, Proposition 1.1] a convergence bound in function values for gradient flows originating from $f \in \mathcal{F}_{\mu,\infty}$,

$$f(X_t) - f_\star \leqslant e^{-2\mu t}(f(x_0) - f_\star). \tag{2.1}$$

This convergence guarantee follows directly from the time-derivative of the Lyapunov function $\mathcal{V}(X_t) = f(X_t) - f_\star$, together with strong convexity (or Łojasiewicz inequality): $\frac{d}{dt}\mathcal{V}(X_t) = \dot{X}_t^\top \nabla f(X_t) = -\|\nabla f(X_t)\|^2 \leqslant -2\mu(f(X_t) - f_\star) \leqslant -2\mu\mathcal{V}(X_t)$.

Given the specific gradient flows studied in this work, it is reasonable to search for Lyapunov functions made of linear combinations of function values, and a quadratic form in the trajectory X_t. We simply refer to them as quadratic Lyapunov functions:

$$\mathcal{V}_{a,c}(X_t) = a \cdot (f(X_t) - f_\star) + c \cdot \|X_t - x_\star\|^2, \tag{2.2}$$

where a, c are fixed nonnegative constants that do not depend on t. Such Lyapunov functions are common for proving convergence of gradient flows (and of optimization methods), and cover for instance the Lyapunov used to prove convergence of the gradient flow under strong convexity (2.1). Given a Lyapunov function $\mathcal{V}_{a,c}$, the idea is to find the smallest value of τ such that the condition

$$\frac{d}{dt}\mathcal{V}_{a,c}(X_t) \leqslant -\tau\mathcal{V}_{a,c}(X_t), \tag{2.3}$$

holds for any functions $f \in \mathcal{F}_{\mu,\infty}$ and any trajectory X_t generated by the gradient flow. After integrating, a convergence guarantee in function values is given by $\mathcal{V}_{a,c}(X_t) \leqslant e^{-\tau t}\mathcal{V}_{a,c}(X_0)$. Given a certain Lyapunov function $\mathcal{V}_{a,c}$ and a time t, we get that the largest acceptable τ is a solution to:

$$\begin{aligned} -\tau = \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N}, f \in \mathcal{F}_{\mu,\infty}} \quad & \frac{d}{dt}\mathcal{V}_{a,c}(X_t), \\ \text{subject to } & \mathcal{V}_{a,c}(X_t) = 1, \\ & \dot{X}_t = -\nabla f(X_t). \end{aligned} \tag{2.4}$$

This minimization problem is invariant in t. It is established in [45, 9] that these so-called performance estimation problems (PEP) can be formulated as SDPs (details are provided in Appendix A.1 for completeness). Because of the condition $f \in \mathcal{F}_{\mu,\infty}$, the maximization problem (2.4) is infinite-dimensional. Recall that a differentiable function $f \in \mathcal{F}_{\mu,\infty}$ verifies for all points $x, y \in \mathbf{R}^d$, $f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geqslant$

$\frac{\mu}{2}\|x-y\|^2$. Introducing alternate variables f_t, $f_\star$, g_t and $g_\star$ (informally: $f_t = f(X_t)$, $f_\star = f(x_\star)$, $g_t = \nabla f(X_t)$ and $g_\star = \nabla f(x_\star) = 0$), it holds that

$$\begin{aligned} -\tau = \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N}} \ & \frac{d}{dt}\mathcal{V}_{a,c}(X_t) \\ \text{subject to } & \mathcal{V}_{a,c}(X_t) = 1, \\ & \dot{X}_t = -g_t, \\ & f_i - f_j - \langle g_j, X_i - X_j\rangle \geqslant \frac{\mu}{2}\|X_i - X_j\|^2, \ \forall i,j = t,\star. \end{aligned} \tag{2.5}$$

The fact (2.5) produces an upper bound on $-\tau$ directly follows from the fact that any sampled strongly convex function satisfy these inequalities at the sampled points X_t and $x_\star$. Thereby, any feasible point to (2.4) corresponds to a feasible for (2.5) with the same objective value. In the other direction, [45, Corollary 2] (which provides a constructive way to obtain some $f \in \mathcal{F}_{\mu,\infty}$ that interpolates the triplets $(X_i, g_i, f_i)_{t,\star}$) ensures that any feasible point to (2.5) can be translated to a feasible point to (2.4) with also the same objective value, thereby reaching the equivalence between formulations (2.4) and (2.5).

In a second stage, we introduce $G = \begin{pmatrix} \|X_t - x_\star\|^2 & \langle X_t - x_\star, g_t\rangle \\ \langle X_t - x_\star, g_t\rangle & \|g_t\|^2 \end{pmatrix} \succeq 0$ a Gram matrix and a vector $F = [f_t, f_\star]$, thereby obtaining a semidefinite reformulation:

$$\begin{aligned} -\tau = \max_{G \succeq 0, F \in \mathbf{R}^2, d \in \mathbf{N}} \ & b_0^\top F + \operatorname{Tr}(A_0 G) \\ \text{subject to } & b_1^\top F + \operatorname{Tr}(A_1 G) \geqslant 0, \\ & b_2^\top F + \operatorname{Tr}(A_2 G) \geqslant 0. \end{aligned} \tag{2.6}$$

where $A_0 = \begin{pmatrix} c\cdot\tau & -c \\ -c & -a \end{pmatrix}$, $A_1 = \begin{pmatrix} -\mu/2 & 1/2 \\ 1/2 & 0 \end{pmatrix}$, $A_2 = \begin{pmatrix} -\mu/2 & 0 \\ 0 & 0 \end{pmatrix}$, $b_0 = a\cdot\tau[1,\ -1]^\top$ $b_1 = [-1,\ 1]^\top$ and $b_2 = [1,\ -1]^\top$. Those developments allow arriving to the following worst-case results.

THEOREM 2.1. *Let $\mathcal{V}_{a,c}$ be a quadratic Lyapunov function (2.2) with $a, c \geqslant 0$, $\tau \geqslant 0$, and functions $f \in \mathcal{F}_{\mu,\infty}$ be strongly convex with parameter $\mu > 0$. We consider gradient flows (1.3) originating from functions $f \in \mathcal{F}_{\mu,\infty}$ and starting from an initial point $x_0 \in \mathbf{R}^d$, where $d \in \mathbf{N}$ is the dimension. The following assertions are equivalent:*

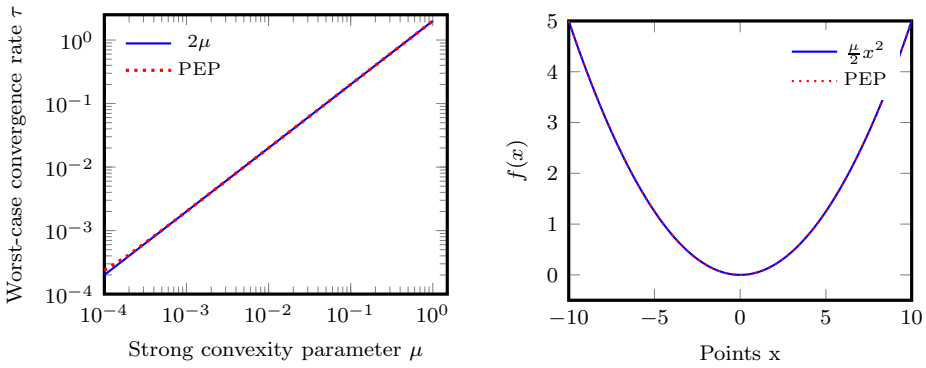
- *The inequality $\frac{d}{dt}\mathcal{V}_{a,c}(X_t) \leqslant -\tau\mathcal{V}_{a,c}(X_t)$ is satisfied for all $f \in \mathcal{F}_{\mu,\infty}$, all trajectories X_t solutions to the gradient flow and all dimensions $d \in \mathbf{N}$.*
- *There exist $\lambda_1, \lambda_2 \geqslant 0$ such that*

$$S = \begin{pmatrix} \tau c - \frac{\mu}{2}(\lambda_1 + \lambda_2) & -c + \frac{\lambda_1}{2} \\ -c + \frac{\lambda_1}{2} & -a \end{pmatrix} \preceq 0, \ \tau a = \lambda_1 - \lambda_2. \tag{2.7}$$

Proof. The procedure to obtain the LMI is fully detailed in Appendix A.1, and consists in taking the standard Lagrangian dual of the SDP (2.6). ∎

A few conclusions can be drawn from the LMI equivalence from Theorem 2.1. First, it provides a necessary and sufficient condition for a quadratic Lyapunov function $\mathcal{V}_{a,c}$ to decrease at a specific rate $\tau \geqslant 0$ for all functions $f \in \mathcal{F}_{\mu,\infty}$. The infeasibility of the LMI allows to conclude about the existence of a Lyapunov function such that the rate τ is achieved for all functions in the class. Second, it is possible to

search over the class of quadratic Lyapunov functions that certify linear convergence. Given a rate τ, the LMI is indeed jointly convex in $\lambda_1, \lambda_2, a, c$. Finally, thanks to linearity of the feasibility problem in τ, a bisection search allows to optimize over the convergence rate. In other words, we can optimize jointly over the Lyapunov function and the worst-case guarantee.



(a) Worst-case rate τ on the class of quadratic Lyapunov functions for the gradient flow.

(b) Reconstruction of a function $f \in \mathcal{F}_\mu$ that interpolates x_0 and X_t, while matching the convergence rate $\tau = -2\mu$, with $\mu = 0.1$

Fig. 1: Comparison between numerical values for τ obtained by solving the LMI (2.7) and the reference established in the literature [33, Proposition 1.1], for trajectories X_t generated by gradient flow (1.3) on μ-strongly convex functions.

In Figure 1a, we obtain the fastest linear convergence rate that can be achieved using quadratic Lyapunov functions. Together with Theorem 2.1 we retrieve the known linear worst-case convergence speed in $e^{-2\mu}$ from Scieur et al. [33, Proposition 1.1], without improvement. The numerical approach though allows to ensure tightness with a numerical function f that matches this convergence guarantee (see Figure 1b and the method in [43, Chapter 3]). Let us build on these results to analyze the gradient flow originating from a (possibly non strongly) convex function, where the difficulty comes from the time-dependence of Lyapunov functions. Following theorems are obtained using the same methodology.

2.1.2. Minimizing convex functions. In the case where $f \in \mathcal{F}_{0,\infty}$, worst-case convergence rates often are sublinear. Again, as in discrete time, it is possible to obtain convergence guarantees using time-dependent quadratic Lyapunov functions.

The Lyapunov function $\mathcal{V}(X_t, t) = t(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ from [37, p. 7] verifies $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$ for any functions $f \in \mathcal{F}_{0,\infty}$, and any trajectories X_t generated by the gradient flow (1.3) (*proof*: $\frac{d}{dt}\mathcal{V}(X_t) = t\langle \nabla f(X_t), \dot{X}_t\rangle + f(X_t) - f_\star + \langle \dot{X}_t, X_t - x_\star\rangle = -t\|\nabla f(X_t)\|^2 + f(X_t) - f_\star - \langle \nabla f(X_t), X_t - x_\star\rangle \leqslant -t\|\nabla f(X_t)\|^2$ using convexity). After integrating, we recover a convergence bound in function values from the literature [37, p.7], [11, Section 6.3.1],

$$f(X_t) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2t}.$$

We consider the family of quadratic Lyapunov functions:

$$\mathcal{V}_{a_t, c_t}(X_t, t) = a_t(f(X_t) - f_\star) + c_t\|X_t - x_\star\|^2, \tag{2.8}$$

where a_t, c_t are differentiable functions from $\mathbf{R}^+$ to $\mathbf{R}^+$. When the Lyapunov function is decreasing along the trajectory X_t, that is $\frac{d}{dt}\mathcal{V}(X_t) \leqslant 0$, a convergence guarantee in function values is given by

$$f(X_t) - f_\star \leqslant \frac{\mathcal{V}(x_0, 0)}{a_t} = \frac{a_0(f(x_0) - f_\star) + c_0\|x_0 - x_\star\|^2}{a_t}.$$

Looking for a worst-case guarantee with a Lyapunov approach can be cast as:

$$\begin{aligned} 0 \geqslant \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N}, f \in \mathcal{F}_{0,\infty}} \quad & \frac{d}{dt}\mathcal{V}_{a_t,c_t} \\ \text{subject to} \quad & \dot{X}_t = -\nabla f(X_t). \end{aligned}$$

Remark 2.2. The strongly convex case as defined above is a particular case of the convex one, using a specific Lyapunov function $\Phi(\cdot)$, such that $\mathcal{V}(X_t, t) = e^{\tau t}\Phi(X_t)$ where $\Phi(X_t) = a \cdot (f(X_t) - f_\star) + c \cdot \|X_t - x_\star\|^2$. Then, $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$ is equivalent to $\frac{d}{dt}\Phi(X_t) \leqslant -\tau\Phi(X_t)$.

THEOREM 2.3. *Let $\mathcal{V}_{a_t,c_t}$ defined in (2.8) be a quadratic Lyapunov function for a_t, c_t nonnegative differentiable functions and functions $f \in \mathcal{F}_{0,\infty}$ be convex. Given the gradient flow (1.3) originating from convex functions $f \in \mathcal{F}_{0,\infty}$ and starting from $x_0 \in \mathbf{R}^d$, where $d \in \mathbf{N}$ is the dimension. The following assertions are equivalent:*

- *The inequality $\dfrac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t, t) \leqslant 0$ is satisfied for all functions $f \in \mathcal{F}_{0,\infty}$, for all trajectories X_t generated by gradient flows, all functions $f \in \mathcal{F}_{\mu,\infty}$ and for all dimensions $d \in \mathbf{N}$.*
- *There exist $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that*

$$S = \begin{pmatrix} \dot{c}_t & -c_t + \frac{\lambda_t^{(1)}}{2} \\ -c_t + \frac{\lambda_t^{(1)}}{2} & -a_t \end{pmatrix} \preceq 0, \ \dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

If these assertions are satisfied, c_t is decreasing, and $a_t \geqslant tc_t$.

Proof. The LMI is obtained following the previous methodology (Appendix A.1), and the inequality $a_t \geqslant tc_t$ directly from the LMI. ∎

Choosing $c_t = \frac{1}{2}$ and $a_t = t$ for the Lyapunov function parameters, and $\lambda_t^{(1)} = 1$, $\lambda_t^{(2)} = 0$, we retrieve the Lyapunov function $\mathcal{V}(x, t) = t(f(x) - f_\star) + \frac{1}{2}\|x - x_\star\|^2$ from [37, p. 7], that verifies $f(X_t) - f_\star \leqslant \frac{1}{2t}\|x_0 - x_\star\|^2$ for all convex functions $f \in \mathcal{F}_{0,\infty}$, for all trajectories X_t generated by the gradient flow, and all dimensions $d \in \mathbf{N}$.

The differential LMI equivalence from Theorem 2.3 is jointly convex in $\lambda_t^{(1)}$, $\lambda_t^{(2)}$, c_t, a_t, $\dot{a}_t$, $\dot{c}_t$, introducing derivatives (to be compared with iterates for discrete optimization methods). In contrast with the minimization of strongly convex functions, numerically solving this LMI is not available because of the dependence in t. Other quadratic Lyapunov functions could be derived. However, the condition $a_t \geqslant c_t t$ together with $\dot{c}_t \leqslant 0$ implies that the convergence cannot be faster than $\frac{1}{t}$.

2.2. Accelerated gradient flows. A major improvement to gradient descent dates back to Nesterov in 1983 [25], with an accelerated gradient method (AGM),

$$\begin{aligned} x_{k+1} &= y_k - \gamma\nabla f(y_k), \\ y_{k+1} &= x_{k+1} + \alpha_k(x_{k+1} - x_k), \end{aligned} \tag{2.9}$$

where γ, α_k are nonnegative parameters depending on the class of functions to minimize. The current iterate is computed thanks to a so-called momentum. This combination of past iterates allows more control over the accumulated error. This idea was first introduced by Polyak [27] with the heavy-ball method, starting from $x_0, x_1 \in \mathbf{R}^d$,

$$x_{k+2} = x_{k+1} + \alpha_k(x_{k+1} - x_k) - \gamma \nabla f(x_{k+1}), \tag{2.10}$$

where $\alpha_k > 0$ is the momentum method. Yet, compared with Nesterov's accelerated gradient method, the heavy-ball method lacks global acceleration beyond quadratics.

When reducing the step size γ, these schemes happen to be closely related to second-order differential equations, for $\beta_t \geqslant 0$ a function that depends on α_k,

$$\ddot{X}_t + \beta_t \dot{X} + \nabla f(X_t) = 0.$$

Recently, accelerated gradient methods have been analyzed using second-order differential equations [34, 32, 47, 15]. Reversely, the accelerated gradient method and the heavy-ball method may be seen as discretization schemes of these second order ODEs, as many other schemes. Discretization techniques are, among others, discussed by [48, 36, 47]. Taking integration theory's point of view, Scieur et al. [33] proved that these multi-step methods may even be seen as discretization schemes of the gradient flow (for quadratics).

Again, ODEs and multi-step first-order methods as defined above are often handled using quadratic Lyapunov functions. We extend the systematic Lyapunov approach developed previously to accelerated gradient flows. Let $\mathcal{V}_{a_t,P_t}$ be the family of quadratic Lyapunov functions for second-order gradient flows,

$$\mathcal{V}_{a_t,P_t}(X_t, t) = a_t(f(X_t) - f_\star) + \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top P_t \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}, \tag{2.11}$$

where $P = \begin{pmatrix} p_t^{(11)} & p_t^{(12)} \\ p_t^{(12)} & p_t^{(22)} \end{pmatrix}$ is a symmetric matrix with differentiable parameters, and a_t is a differentiable function, such that the Lyapunov is nonnegative when evaluated on the gradient flow. After integrating, this approach leads to convergence bounds for instance in function values, such that $f(X_t) - f_\star \leqslant \frac{\mathcal{V}(x_0)}{a_t}$.

2.2.1. Minimizing strongly convex functions. Let $f \in \mathcal{F}_{\mu,L}$, with strong convexity parameter $\mu > 0$, and scheme parameters be defined by $\gamma = \frac{1}{L}$ and $\alpha = \frac{1-\sqrt{\mu\gamma}}{1+\sqrt{\mu\gamma}}$ in Nesterov's accelerated gradient methods (2.9). When reducing the step size γ, the continuous-time limit of y_k in (2.9) is exactly the Polyak damped oscillator [27],

$$\ddot{X}_t + 2\sqrt{\mu}\dot{X}_t + \nabla f(X_t) = 0, \tag{2.12}$$

where $t = \gamma k$, as it has already been highlighted in previous works [11, 32, 34]. This ODE is also the limit of the heavy-ball method (2.10). Shi et al. [34] proved a convergence guarantee in $f(X_t) - f_\star = O(e^{-\frac{\sqrt{\mu}t}{4}})$ using a Lyapunov-based approach. This bound was improved to $f(X_t) - f_\star = O(e^{-\sqrt{\mu}t})$, by Wilson et al. [47, Appendix B] using the Bregman-Lagrangian approach, and by Sanz-Serna and Zygalakis using the IQC framework [32, 11]. We compute linear convergence guarantees using quadratic Lyapunov functions with constant parameters (2.11).

THEOREM 2.4. *Let $\mathcal{V}_{a,P}$ be a quadratic Lyapunov function (2.11), where $a \geqslant 0$, and P a symmetric matrix. Let $\tau \geqslant 0$, $\mu \geqslant 0$, $d \in \mathbf{N}$ be the dimension, and the Polyak damped oscillator (2.12) be starting from $x_0 \in \mathbf{R}^d$, and originating from strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$. The following assertions are equivalent:*

- *The inequality $\frac{d}{dt}\mathcal{V}_{a,P}(X_t) \leqslant -\tau \mathcal{V}_{a,P}(X_t)$ is satisfied for all functions $f \in \mathcal{F}_{\mu,\infty}$, all trajectories X_t generated by Polyak damped oscillator, and all dimensions $d \in \mathbf{R}^d$.*
- *There exist $\lambda_1, \lambda_2, \nu_1, \nu_2 \geqslant 0$, such that*

$$
\begin{aligned}
&\begin{pmatrix} -\frac{\mu}{2}(\lambda_1+\lambda_2) + \tau p_{11} & p_{11} - 2\sqrt{\mu}p_{12} + \tau p_{12} & -p_{12} + \frac{\lambda_1}{2} \\ * & 2(p_{12} - 2\sqrt{\mu}p_{22}) + \tau p_{22} & -p_{22} + \frac{a}{2} \\ * & * & 0 \end{pmatrix} \preceq 0, \\
&\tau a = \lambda_1 - \lambda_2, \\
&\begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \frac{\mu}{2}(\nu_1+\nu_2) & 0 & \frac{-\nu_1}{2} \\ 0 & 0 & 0 \\ \frac{-\nu_1}{2} & 0 & 0 \end{pmatrix} \succeq 0, \\
&a = \nu_2 - \nu_1.
\end{aligned} \tag{2.13}
$$

Proof. The LMI equivalence is obtained introducing the Gram matrix $G = P^\top P$, where $P = (\dot{X}_t, X_t - x_\star, g_t)$ and using the PEP methodology. The first LMI refers to the non-increasing condition for the Lyapunov function, and the second LMI to the positivity of the Lyapunov for all trajectories and convex functions. ∎

The LMI is a feasibility problem, jointly convex in $\lambda_1, \lambda_2, \nu_1, \nu_2 \geqslant 0$ and in Lyapunov parameters a, P. Given parameters a, P and a rate τ, it is a verification tool for $\mathcal{V}_{a,P}$ to be a Lyapunov function decreasing at rate τ. As for gradient flows, we can perform a bisection search over τ to find the fastest linear convergence rate that can be verified using quadratic Lyapunov functions (see Figure 2a). It provides a numerical tool for choosing Lyapunov parameters (Figure 2b) for which the worst-case linear convergence rate is achieved.

COROLLARY 2.5. *Let us consider the Polyak damped oscillator (2.12) starting from $x_0 \in \mathbf{R}^d$ where $d \in \mathbf{N}$ is the dimension, and originating from strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$ where $\mu > 0$. The Lyapunov function*

$$
\mathcal{V}(X_t) = f(X_t) - f_\star + \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top \begin{pmatrix} 4/9\mu & 2/3\sqrt{\mu} \\ 2/3\sqrt{\mu} & 1/2 \end{pmatrix} \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix},
$$

verifies $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -4/3\sqrt{\mu}\mathcal{V}(X_t)$ for all functions $f \in \mathcal{F}_{\mu,\infty}$, all trajectories X_t generated by Polyak damped oscillator, and all dimensions $d \in \mathbf{N}$.

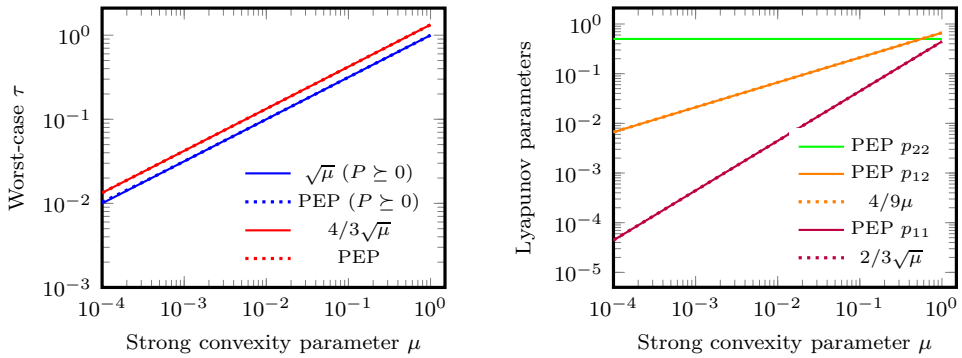
Proof. Taking $\lambda_1 = 4/3\sqrt{\mu}$, $\lambda_2 = 0$, $\nu_1 = 0$ and $\nu_2 = 1$, we verify the LMI for this Lyapunov function $\mathcal{V}$, with $\tau = 4/3\sqrt{\mu}$. ∎

This class of quadratic Lyapunov functions is inspired by [42] in discrete time, where a stricter positivity condition on $P \succeq 0$ hindered proving tight convergence of Nesterov's accelerated gradient. Similarly, the Lyapunov function from Corollary 2.12 is defined by $P = \begin{pmatrix} 4/9\mu & 2/3\sqrt{\mu} \\ 2/3\sqrt{\mu} & 1/2 \end{pmatrix}$, which is not positive semidefinite. In the continuous-time models' literature, we usually choose matrices P positive semidefinite,

such as in the Lyapunov function from [32, 34, Theorem 4.3],

$$\mathcal{V}(X_t) = f(X_t) - f_\star + \frac{1}{2}\begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top \begin{pmatrix} \mu & \sqrt{\mu} \\ \sqrt{\mu} & 1 \end{pmatrix} \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}$$

that verifies $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -\sqrt{\mu}\mathcal{V}(X_t)$ for all functions $f \in \mathcal{F}_{\mu,\infty}$, all trajectories X_t generated by Polyak damped oscillator, and all dimensions $d \in \mathbf{N}$. This Lyapunov is a feasible point of the LMI (2.13) from Theorem 2.4, with $\tau = \sqrt{\mu}$, $\lambda_1 = \sqrt{\mu}$, $\lambda_2 = 0$, $\nu_1 = 0$, $\nu_2 = a = 1$. Relaxing the condition $P \succeq 0$ improves results from Sanz-Serna and Zygalakis and Wilson et al. by a factor $4/3$ in Corollary 2.5.



(a) Worst-case guarantee optimized over the class of quadratic Lyapunov functions.

(b) Lyapunov parameters P for $\tau = 4/3\sqrt{\mu}$ and $a = 1$, as a function of the condition number μ.

Fig. 2: Comparison between numerical worst-case rate guarantee obtained numerically with PEP, and with references. Trajectories X_t are generated by the damped gradient flow (2.12) for μ-strongly convex functions.

Figure 2b provides a numerical help for computing the Lyapunov function from Corollary 2.5. Figure 2a together with Corollary 2.5 allows to conclude that this bound cannot be improved when changing the Lyapunov function among the class of quadratic functions (2.11).

2.2.2. Minimizing convex functions. As for gradient flow, rates are sublinear when the accelerated gradient flow originates from convex functions. Let $f \in \mathcal{F}_{0,L}$, the step size be $\gamma \leqslant \frac{1}{L}$, and $\alpha = \frac{k-1}{k+2}$ be the scheme parameter in Nesterov's accelerated method. Su et al. [37, Section 2] proved the connection between the first-order scheme and a second order ODE known as accelerated gradient flow (AGF):

$$\ddot{X}_t + \frac{3}{t}\dot{X}_t + \nabla f(X_t) = 0. \tag{2.14}$$

Su et al. [37, Theorem 3] proved the following inequality is verified for all functions f and all trajectories X_t generated by AGF,

$$f(X_t) - f_\star \leq 2\frac{\|x_0 - x_\star\|^2}{t^2}.$$

Their proof exhibits a Lyapunov function $\mathcal{V}(X_t, t) = t^2(f(X_t) - f_\star) + \|(X_t - x_\star) + \frac{t}{2}\dot{X}_t\|^2$, whose derivative is decreasing along trajectories X_t (*proof:* $\frac{d}{dt}\mathcal{V}(X_t, t) =$

$2t(f(X_t)-f_\star)+t^2\langle\dot{X}_t,\nabla f(X_t)\rangle+2\langle X_t-x_\star+\frac{t}{2}\dot{X}_t,3\dot{X}_t+t\ddot{X}_t\rangle=2t(f(X_t)-f_\star)-\langle\nabla f(X_t),X_t-x_\star\rangle\leqslant 0$ by convexity of f). The following theorem provides a systematic condition for a quadratic function $\mathcal{V}$ (2.11) to be a Lyapunov function for AGF (2.14).

THEOREM 2.6. *Let $\mathcal{V}_{a_t,P_t}$ be a quadratic Lyapunov function (2.11), where $a_t\geqslant 0$ is a differentiable function, and $P_t\succeq 0$ with differentiable parameters. Given the accelerated gradient flow (2.14) starting from $x_0\in\mathbf{R}^d$ where $d\in\mathbf{N}$ is the dimension, and originating from convex functions $f\in\mathcal{F}_{0,\infty}$, the following assertions are equivalent:*

- *The inequality $\frac{d}{dt}\mathcal{V}_{a_t,P_t}(X_t,t)\leqslant 0$ is satisfied for all functions $f\in\mathcal{F}_{0,\infty}$, all trajectories X_t generated by the accelerated gradient flow and all dimensions $d\in\mathbf{N}$.*
- *There exist $\lambda_t^{(1)},\lambda_t^{(2)}\geqslant 0$ such that*

$$S=\begin{pmatrix}\dot{p}_t^{(11)} & p_t^{(11)}-\frac{3}{t}p_t^{(12)}+\dot{p}_t^{(12)} & -p_t^{(12)}+\frac{\lambda_t^{(1)}}{2}\\ * & 2(p_t^{(12)}-\frac{3}{t}p_t^{(22)})+\dot{p}_t^{(22)} & -p_t^{(22)}+\frac{a_t}{2}\\ * & * & 0\end{pmatrix}\preceq 0,\ \dot{a}_t=\lambda_t^{(1)}-\lambda_t^{(2)}.$$

Proof. The proof follows those from Theorem 2.4 and 2.1. ∎

The LMI from Theorem 2.6 is differentiable and convex in $\lambda_t^{(1)}$, $\lambda_t^{(2)}$. Even when studying second-order gradient flows, the number of interpolation inequalities (or of dual values $\lambda_t^{(i)}$) is bounded by two, enforcing short proofs.

Theorem 2.6 allows to retrieve the Lyapunov function exhibited by Su et al. [37, Theorem 3], and its associated convergence guarantee. Choosing $a_t=t^2$ and $P_t=2\begin{pmatrix}1 & t/2\\ t/2 & t^2/4\end{pmatrix}$ for the Lyapunov parameters, and $\lambda_t^{(1)}=t$, $\lambda_t^{(2)}=0$, we prove that that the Lyapunov function $\mathcal{V}(X_t,t)=t^2(f(X_t)-f_\star)+2\left\|(X_t-x_\star)+\frac{t}{2}\dot{X}\right\|^2$ verifies $\frac{d}{dt}\mathcal{V}(X_t,t)\leqslant 0$, for all functions $f\in\mathcal{F}_{0,\infty}$, for all trajectories X_t generated accelerated gradient flows, and all dimensions $d\in\mathbf{N}$. As for gradient flows originating from convex functions, we did not compute numerically worst-case scenarios because of the time-dependence of the Lyapunov function.

2.3. Higher-order convergence and time dilation. In this section, we analyze convergence of non-autonomous gradient flows, and provide convergence rates depending on their parameters. It appears that higher order convergence of gradient flows is highly connected to time dilation.

2.3.1. Non-autonomous first-order gradient flow. Let f be a convex function, and X_t be the solution to the non-autonomous first-order gradient flow,

$$\dot{X}_t=-\alpha_t\nabla f(X_t),\ X_0=x_0\in\mathbf{R}^d, \tag{2.15}$$

where $\alpha_t\geqslant 0$ is a continuous function (such that the flow is converging). It is natural to wonder if it is possible to accelerate such gradient flows when changing α_t. A change of variable connects this ODE to the gradient flow (1.3), for which $\alpha_t=1$. Let Y_t be the solution to the gradient flow, and $\tau_t=\int_0^t\alpha_s ds$ be a time change variable. Then, the variable $X_t=Y_{\tau_t}$ verifies $\dot{X}_t=\frac{d}{dt}Y_{\tau_t}=\alpha_t\dot{Y}_{\tau_t}=-\alpha_t\nabla f(Y_{\tau_t})=-\alpha_t\nabla f(X_t)$, which is exactly the non-autonomous gradient flow. The following corollaries can be obtained by performing the appropriate change of variable in LMI from Theorem 2.3.

COROLLARY 2.7. *Let us consider non-autonomous gradient flows* (2.15) *starting from* $x_0 \in \mathbf{R}^d$ *where* $d \in \mathbf{N}$ *is the dimension, and originating from possibly strongly convex functions* $f \in \mathcal{F}_{\mu,\infty}$, *where* $\mu \geqslant 0$.

- *If* $\mu = 0$, *the Lyapunov function* $\mathcal{V}(X_t, t) = \left(\int_0^t \alpha_s ds\right)(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ *verifies* $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$ *for all functions* $f \in \mathcal{F}_{0,\infty}$, *for all trajectories* X_t *generated by non-autonomous gradient flows and all dimensions* $d \in \mathbf{N}$.
  A convergence guarantee is given by $f(X_t) - f_\star \leqslant \dfrac{1}{2\int_0^t \alpha_s ds}\|x_0 - x_\star\|^2$.
- *If* $\mu > 0$, *the Lyapunov function* $\mathcal{V}(X_t, t) = e^{2\mu\int_0^t \alpha_s ds}(f(X_t) - f_\star)$ *verifies* $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant -2\mu\alpha_t\mathcal{V}(X_t, t)$ *for all strongly convex functions* $f \in \mathcal{F}_{\mu,\infty}$ *(*$\mu > 0$*), all trajectories* X_t *generated by the non-autonomous gradient flow and all dimensions* $d \in \mathbf{N}$.
  A convergence guarantee is given by $f(X_t) - f_\star \leqslant e^{-2\mu\int_0^t \alpha_s ds}(f(x_0) - f_\star)$.

Remark 2.8. When considering $\alpha_t = 1$ above, that is $\tau_t = t$, we recover exactly the results from Theorem 2.1 and Theorem 2.3.

As mentioned by Orvieto and Lucchi [26] in the stochastic setting, and for accelerated methods by Wibisono et al. [46], one can thus either work with Y_t generated by the non-autonomous gradient flow (2.15), or with X_t generated by the gradient flow. However, the acceleration on Y_t is not preserved after discretizing the flow. Applying explicit Euler scheme to a non-autonomous gradient flow (2.15) with $\alpha_s = \alpha > 0$ and originating from $f \in \mathcal{F}_{0,L}$, a condition on step sizes $h > 0$ arises $0 \leqslant h \leqslant \frac{2}{L\alpha}$.

When focusing on continuous-time models for analyzing explicit optimization methods, that only calls for past gradient iterates, we prefer working with the gradient flow (1.3). However, non-autonomous gradient flows (2.15) may be useful for analyzing other methods such as proximal methods. More generally, and in the next section, we analyze gradient flows without adjusting the time scale (taking $\alpha_t = 1$).

2.3.2. A non-autonomous second-order gradient flows. Nesterov's accelerated gradient flow reaches an $\frac{1}{t^2}$ convergence (see Theorem 2.6) in function values. Considering the family of quadratic Lyapunov functions (2.11), we study convergence of non-autonomous second-order gradient flows and draw comparison with the accelerated gradient flow. Let $\beta_t \geqslant 0$ be a continuous function, and a second-order non-autonomous ODE,

$$\ddot{X}_t + \beta_t \dot{X}_t + \nabla f(X_t) = 0. \tag{2.16}$$

Remark 2.9. Wibisono et al. [46, Theorem 2.2] proved this ODE is related to the family of ODEs defined by $\dot{Y}_t + \tilde{\beta}_t Y_t + \tilde{\alpha}_t \nabla(Y_t) = 0$. For α_t a strictly nonnegative differentiable function, a time change formula $\tau_t = \int_0^t \sqrt{\alpha_s} ds$, and X_t a solution to (2.16), the trajectory $Y_t = X_{\tau_t}$ is solution to $\ddot{Y}_t + (\beta_{\tau_t}\sqrt{\alpha_t} - \frac{\dot{\alpha}_t}{2\alpha_t})\dot{Y}_t + \alpha_t \nabla f(Y_t) = 0$.

For any two functions $\beta_t^{(1)}, \beta_t^{(2)} \geqslant 0$, there is no time change formula that connects their associated ODE. In other words, they are in the same time-scale. Theorem 2.10 provides an LMI equivalence for analyzing convergence of trajectories X_t generated by second-order gradient flows (2.16) using quadratic Lyapunov functions (2.11).

THEOREM 2.10. *Let us consider* $\mathcal{V}_{a_t,P_t}$ *quadratic Lyapunov functions* (2.11), *and non-autonomous second-order gradient flows* (2.16) *starting from* $x_0 \in \mathbf{R}^d$ *where* $d \in$

$\mathbf{N}$ *is the dimension, and originating from possibly non-strongly convex functions* $f \in \mathcal{F}_{\mu,\infty}$ *where* $\mu \geqslant 0$. *The following assertions are equivalent:*

- *The inequality* $\frac{d}{dt}\mathcal{V}_{a_t,P_t}(X_t,t) \leqslant 0$ *is satisfied for all functions* $f \in \mathcal{F}_{\mu,\infty}$, *for* X_t *generated by non-autonomous second-order gradient flows, and all dimensions* $d \in \mathbf{N}$.
- *There exist* $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ *such that,*

$$\begin{pmatrix} -\frac{\mu}{2}(\lambda_t^{(1)}+\lambda_t^{(2)}) + \dot{p}_t^{(11)} & p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} & -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} \\ * & 2(p_t^{(12)} - \beta_t p_t^{(22)}) + \dot{p}_t^{(22)} & -p_t^{(22)} + \frac{a_t}{2} \\ * & * & 0 \end{pmatrix} \preceq 0,$$

$$\dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

For a given function $\beta_t \geqslant 0$, the LMI remains convex in $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$. Compared with the LMI derived from Nesterov's accelerated gradient method in Theorem 2.6, the LMI is parametrized by β_t.

Corollary 2.11. *Let us consider non-autonomous second-order gradient flows starting from* $x_0 \in \mathbf{R}^d$ (2.16) *where* $d \in \mathbf{N}$ *is the dimension, and originating from possibly non-strongly convex functions* $f \in \mathcal{F}_{\mu,\infty}$ *where* $\mu \geqslant 0$. *The Lyapunov function*

$$\mathcal{V}(X_t,t) = a_t(f(X_t) - f_\star) + \frac{1}{2a_t}\|a_t\dot{X}_t + \dot{a}_t(X_t - x_\star)\|^2,$$

with a_t *defined by:*

- *If* $\mu > 0$, $a_t = \min(\sqrt{\mu}, \frac{2}{3}\beta_t)$,
- *If* $\mu = 0$, $a_t = \min((\sqrt{a_0} + (\sqrt{p_0^{(11)}}/2)t)^2, \lim_{\epsilon\to 0,\epsilon>0} a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds})$,

verifies $\frac{d}{dt}\mathcal{V}(X_t,t) \leqslant 0$ *for all functions* $f \in \mathcal{F}_{\mu,\infty}$, *all* X_t *generated the second order gradient flows and all dimensions* $d \in \mathbf{N}$.

Proof. The proof follows from Theorem 2.10, and is detailed in Appendix A.2. ∎

Non-autonomous second-order gradient flows (2.16) cannot converge faster than Nesterov's accelerated gradient flow in function values, that is to say not faster than $\frac{1}{t^2}$, when using quadratic Lyapunov functions parametrized by $\beta_t \geqslant 0$ from Corollary 2.11. To analyze Nesterov's accelerated gradient methods using ODEs, Su et al. [37] introduced parametrized second-order gradient flows, that fit the model (2.16),

$$\ddot{X}_t + \frac{r}{t}\dot{X}_t + \nabla f(X_t) = 0. \tag{2.17}$$

When $r \geqslant 3$, the guarantee $f(X_t) - f_\star \leqslant \frac{(r-1)^2\|x_0 - x_\star\|^2}{2t^2}$ holds for all convex functions f and all trajectories X_t generated by accelerated gradient flows [37, Theorem 5]. When $r < 3$, Attouch et al. [1, Theorem 2.1] proved a convergence bound in $f(X_t) - f_\star = \mathcal{O}(\frac{1}{t^{2r/3}})$, improving the results from [37, Theorem 7] that required additional assumptions on f. Using Corollary 2.11, we retrieve a similar bound in function values $f(X_t) - f_\star \leqslant \frac{2\|x_0 - x_\star\|^2 r^2}{9t^{\frac{2r}{3}}}$.

Remark 2.12. Polynomial convergence can be achieved up to a change of variable, as it was shown by Wisobono et al. [46] with $\tau_t = t^{p/2}$ ($\alpha_t = \frac{p}{2}t^{p/2-1}$). Nesterov's accelerated gradient flow ($r = 3$) has an ODE $\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + \frac{p^2}{4}t^{p-2}\nabla f(X_t) = 0$ for

$p \geqslant 2$. For all convex functions and all trajectories X_t generated by the ODE starting from x_0, a convergence bound in function value is given by $f(X_t) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2t^p}$.

We have extended the performance estimation approach to continuous-time models, using Lyapunov functions. Given an ODE and a class of functions, we presented a semidefinite formulation equivalent with the existence of a quadratic Lyapunov function. This LMI provides a principled way to generate Lyapunov functions, even when dealing with non-autonomous parametrized ODEs. It turns out only two convex inequalities (in $(X_t, x_\star)$ and $(x_\star, X_t)$, see Theorem 2.6) are involved in convergence proofs for continuous-time models. For strongly convex functions, we proved numerically worst-case guarantees from Corollary 2.1 and cannot be improved using a specific family of quadratic Lyapunov functions. Even if the connection between discrete and continuous-time does not allow to directly transfer results to the analysis of first-order methods, their convergence analyses are be closely related.

3. SDEs for modeling SGD. Convergence results for stochastic convex optimization often require additional assumptions on function classes, refined choices of step sizes and averaged iterates. Their analyses raise challenges and more complex proofs in contrast with deterministic methods. Prior works have been concerned with the connection between stochastic methods and stochastic differential equations (SDEs) [21, 26, 48]. This section is devoted to convergence analyses of SDEs using the Lyapunov framework. As for ODEs, verifying a Lyapunov function will be cast as verifying the feasibility of a small-sized LMI.

Let us recall the stochastic gradient descent method (SGD)

$$x_{k+1} = x_k - \gamma \nabla \tilde{f}(x_k, \xi_{i_k}), \tag{3.1}$$

where $\gamma > 0$ is the step size, ξ_{i_k} are uniformly drawn in $(\xi_1, ..., \xi_n)$, and where $\nabla \tilde{f}(x_k, \xi_{i_k})$ is an unbiased estimate of full gradient $\nabla f(x_k)$. Li et al. [21] introduced stochastic modified equations (SMEs) to model SGD, rewriting it as

$$x_{k+1} = x_k - \gamma \tilde{f}(x_k, \xi_{i_k}) + \sqrt{h} V_k(x_k),$$

where $V_k(x) = \sqrt{\gamma}(\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{i_k})$ has zero mean and a covariance matrix equal to $\gamma \Sigma(x_k) = \gamma(\sum_{i=1}^n (\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{i_k}))(\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{i_k}))^\top)$.

The corresponding SDE is given by

$$dX_t = -\nabla f(X_t)dt + (\gamma \Sigma(X_t))^{1/2} dB_t, \tag{3.2}$$

where B_t is a standard Brownian motion, is an (order-1 weak) approximation of SGD ([21, Theorem 1], [22]). The SDE approximation of SGD allows to take into account the role of fixed step size in the dynamics of SGD (while keeping them small). Under mild assumptions on f, Li et al. proved the weak approximation of SGD by this SDE on a finite interval $[0, T]$: there exists $C > 0$ such that $\|\mathbf{E}[x_k] - \mathbf{E}[X(k\gamma)]\| \leqslant C\gamma$ for $k \in [0, \frac{T}{\gamma}]$. However, this approach is limited since C depends exponentially on T. In the literature, matching rates are often obtained extending Lyapunov functions from continuous to discrete time [35, 26, 34].

Remark 3.1. The SDE (3.2) is an approximation of SGD for small step sizes γ. When taking the step size to zero, the noise term actually disappears, and the limiting ODE of SGD is exactly the gradient flow (1.3). Similarly, the stochastic Langevin dynamics $x_{k+1} = x_k - \gamma \nabla f(x_k) - \sqrt{\gamma}\xi_k$ has the limiting ODE $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$, where the step size is not taken into account.

Compared with the gradient flow, SGD does not converge to a stationary point under fixed step sizes [2, 41]. Convergence to a stationary point requires diminishing step sizes such as $\gamma_k = \frac{1}{\sqrt{k}}$. Li et al. [21, Section 4.1] (and later Orvieto and Lucchi [26, Section 2.1]) proposed to include this varying learning rate in the dynamics:

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

where γ is the maximum allowed learning rate and $h_k \in [0, 1]$ is the varying part. For $h_t \geqslant 0$ a continuous function corresponding to discrete h_k, the SDE is given by

$$dX_t = -h_t \nabla f(X_t) dt + h_t (\gamma \Sigma(X_t))^{1/2} dB_t. \tag{3.3}$$

We treat the covariance matrix $\Sigma(X_t)$ as symmetric, already implied by the notation $\Sigma(X_t)^{1/2}$, but unstructured with bounded variance $\Sigma(X_t) \preceq \Sigma$ along any trajectory X_t generated by the approximating SDE (3.3). Compared to ODEs, functions $f \in \mathcal{F}_{0,L}$ to be optimized using SDEs are in addition assumed to be possibly L-smooth with $L \in (0, \infty]$, and to be twice differentiable.

We propose to analyze approximating SDEs with averaging techniques, and to include later varying step sizes. Verifying Lyapunov functions thanks to small-sizes LMI, we retrieve convergence results from discrete optimization methods, using appropriate choices of step sizes.

3.1. Lyapunov functions do not always extend to the stochastic setting. The analysis of the gradient flow in the deterministic case provides Lyapunov functions that are decreasing along trajectories generated by ODEs. The direct transfer of these Lyapunov functions to the stochastic setting is not always suited to the variance term, as detailed below. Under constant step sizes $\gamma > 0$, an approximating SDE of SGD is

$$dX_t = -\nabla f(X_t) dt + (\gamma \Sigma(X_t))^{1/2} dB_t.$$

In SDE theory, a time-differential of a function of a solution to a stochastic process is given by Ito's Lemma [39, Theorem 4.2].

LEMMA 3.2 (Ito's Lemma). *For g a twice differentiable function, and X_t a stochastic process solution to the SDE (3.2),*

$$dg(X_t, t) = \frac{\partial}{\partial t} g(X_t, t) dt + \frac{\partial}{\partial x} g(X_t, t) dX_t + \frac{1}{2} \gamma \mathrm{Tr}\left(\frac{\partial^2}{\partial x^2} g(X_t, t) \Sigma(X_t)\right) dt.$$

3.1.1. Minimizing strongly convex functions. When the SDE originates from (possibly non-smooth) strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$, the Lyapunov function from the deterministic setting extends well to SDEs. In the deterministic setting, we have shown in Theorem 2.1 that the function $\mathcal{V}(x, t) = e^{2\mu t}(f(x) - f_\star)$ is a Lyapunov function in the worst-case. Applying Ito's formula to this Lyapunov function with X_t a solution to the SDE (3.2), we obtain $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \frac{1}{2} e^{2\mu t} \gamma \mathbf{E}\mathrm{Tr}(\nabla^2_{xx} f(X_t) \Sigma(X_t))$. After integrating between 0 and t, we have

$$\mathbf{E}(f(X_t) - f_\star) \leqslant e^{-2\mu t} \left(f(x_0) - f_\star + \frac{1}{2} \gamma \int_0^t e^{2\mu s} \mathbf{E}\mathrm{Tr}(\nabla^2_{xx} f(X_s) \Sigma(X_s)) ds \right).$$

Additional requirements on f, such as smoothness and twice differentiability, are needed for convergence. For example, if f is in addition L-smooth with $L < \infty$ and using the bounded covariance assumption $\Sigma(X_t) \preceq \Sigma$, the variance term is bounded

by $\frac{1}{2}L\gamma\mathrm{Tr}(\Sigma)$. Under constant step sizes, the SDE approximating SGD converges to a diffusion, and cannot get to a stationary point $x_\star$ in the worst-case, but the extra term is linear in the step size γ. As in the deterministic case, the forgetting of initial conditions remains of order $O(e^{-2\mu t})$.

3.1.2. Minimizing convex functions. When $f \in \mathcal{F}_{0,\infty}$, Lyapunov functions induce convergence bounds with a possibly diverging variance term. When considering deterministic gradient flows (1.3) originating from convex functions, a Lyapunov function followed from Theorem 2.3: $\mathcal{V}(x,t) = t(f(x) - f_\star) + \frac{1}{2}\|x - x_\star\|^2$. Assuming X_t are solutions X_t to SDEs (3.2) originating from convex functions $f \in \mathcal{F}_{0,\infty}$, we compute the derivative to this Lyapunov function thanks to Ito's formula $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t) \leqslant -t\mathbf{E}\|\nabla f(X_t)\|^2 + \mathbf{E}\frac{1}{2}\mathrm{Tr}((t\nabla_x^2 f(X_t) + I)\Sigma(X_t))$. Thanks to twice differentiability, bounded covariance $\Sigma(X_t) \preceq \Sigma$, and assuming in addition L-smoothness of f, a convergence bound is given by

$$\mathbf{E}(f(X_t) - f_\star) \leqslant \frac{\|x_0 - x_\star\|^2}{t} + \frac{1}{2}(L\frac{t}{2} + 1)\mathrm{Tr}(\Sigma).$$

Taylor and Bach proved a comparable convergence bound for SGD [41, Theorem 5], applying the Lyapunov performance estimation approach under similar assumptions (bounded variance, smoothness of f). Tough, we cannot conclude about convergence of SGD in the worst-case without further assumptions. Optimization methods have been developed to ensure the global convergence of SGD to the optimum, among them averaging [30] and diminishing step sizes.

3.2. Diminishing the step size is a key to success. We study convergence of SDEs with varying step sizes (3.3). In contrast to the deterministic setting, in which varying step sizes correspond to a time rescaling, the time change plays a direct role on the variance term (explicit formula by Orvieto and Lucchi in [26, Theorem 5]). Our problem can be formulated as follows, where $\mathcal{V}$ are Lyapunov functions,

$$\max_{X_t \in \mathbf{R}^d, d\in\mathbf{N},\ f\in\mathcal{F}_{0,\infty}} \ \frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t),$$
$$\text{subject to } \ dX_t = -h_t\nabla f(X_t)dt + h_t(\gamma\Sigma(X_t))^{1/2}dB_t,$$

where $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t) = \mathbf{E}[\frac{\partial}{\partial t}\mathcal{V}(X_t,t) + \frac{\partial}{\partial x}\mathcal{V}(X_t,t)\frac{dX_t}{dt}] + \frac{\gamma}{2}\mathbf{E}\mathrm{Tr}(\frac{\partial^2}{\partial x^2}\mathcal{V}(X_t,t)\Sigma(X_t))$ is computed with Ito's formula. The first two terms corresponds exactly to taking the derivative in trajectories generated by ODEs (2.15) (or the SDEs (3.3) with $\gamma = 0$), and the last term to a variance term. Because of the trace in a second-order derivative and in the covariance matrix $\Sigma(X_t)$, we do not take this term into account in an LMI reformulation. Instead, we propose to first derive a family of Lyapunov functions using LMIs from the deterministic setting, and then to optimize their parameters so that the variance term converges conveniently.

Corollary 3.3. *Let $f \in \mathcal{F}_{0,\infty}$ be twice differentiable functions, and let us consider SDEs with varying step sizes (3.3) starting from $X_0 \in \mathbf{R}^d$, where $d \in \mathbf{N}$ is the dimension. The quadratic Lyapunov function from the family (2.8)*

$$\mathcal{V}(X_t,t) = a_t^{(1)}(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2,$$

with $\dot{a}_t^{(1)} = 2h_t$ verifies $\frac{d}{dt}\mathbf{E}(\mathcal{V}(X_t,t) \leqslant h_t^2\, Tr((\nabla_{xx}^2 f(X_t)a_t^{(1)} + \frac{1}{2}I_d)\Sigma(X_t))$ for all twice differentiable functions $f \in \mathcal{F}_{0,\infty}$, all trajectories X_t generated by SDEs (3.3),

and all dimensions $d \in \mathbf{N}$. *Then, it holds that* $\mathbf{E}[f(X_t) - f_\star] \leqslant \dfrac{\|X_0 - X_\star\|^2}{a_t^{(1)}} + \dfrac{\gamma}{2a_t^{(1)}} \displaystyle\int_0^t h_s^2 \, Tr((\nabla^2_{xx} f(X_s) a_s^{(1)} + \frac{1}{2} I_d)\Sigma(X_s))ds.$

Proof. The Lyapunov function $\mathcal{V}$ directly comes from Corollary 2.7, for non-autonomous first-order gradient flows. The bound on $\mathbf{E}[f(X_t) - f_\star]$ is obtained using Ito's formula on $\mathcal{V}$ along trajectories X_t generated by approximating SDEs (3.3). ∎

The convergence bound from Corollary 3.3 is divided into two terms: a term that forgets the initial conditions and a variance term due to noise. Convergence is mostly controlled by the step size ($\dot{a}_t^{(1)} = 2h_t$). Bach and Moulines [2, Theorem 5] provided a comparable but much more complex analysis for stochastic gradient descent, for a specific family of step sizes. Let us compare our results, assuming $h_t = \frac{1}{(t+1)^\alpha}$, where $\alpha \geqslant 0$. The Lyapunov function from Corollary 3.3 is given by $\mathcal{V}(X_t, t) = a_t^{(1)}(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ with $a_t^{(1)} = (t+1)^{1-\alpha}$. The forgetting of the initial condition is thus bounded by $\frac{\|x_0 - x_\star\|^2}{(t+1)^{1-\alpha}}$. Provided $\Sigma(X_t) \preceq \Sigma$, the variance term is bounded by:

$$\begin{cases} \frac{\gamma \mathrm{Tr}(\Sigma)}{(t+1)^{1-\alpha}} \left(L \frac{(t+1)^{2-3\alpha} - 1}{2-3\alpha} + \frac{1}{2} \frac{(t+1)^{1-2\alpha} - 1}{1-2\alpha} \right) & \text{if } 0 \leqslant \alpha < 1, \\ \frac{\gamma \mathrm{Tr}(\Sigma)}{\log(t)} \left(L \int_0^t \frac{\log(s+1)}{(s+1)^2} ds + \frac{1}{2}(1 - \frac{1}{t+1}) \right) & \text{if } \alpha = 1. \end{cases}$$

Because of the variance term, the Lyapunov converges if and only if $\alpha \geqslant \frac{1}{2}$. In other words, convergence is not guaranteed for constant step sizes ($\alpha = 0$). If $\alpha \in (1/2, 2/3)$, the convergence in function value is bounded by $O(\frac{1}{t^{2\alpha-1}})$. If $\alpha \in (2/3, 1)$, the convergence in function value is bounded by $O(\frac{1}{t^{1-\alpha}})$. As for SGD [2, Theorem 5], the convergence regime changes at $\alpha = \frac{2}{3}$ with a global convergence rate in $\frac{1}{t^{1/3}}$, for which the variance term and the term that forgets the initial conditions converges at the same rate (up to $\log(t)$). It is therefore possible to reach convergence with diminishing step sizes. Other techniques have been developed to improve the trade-off between faster convergence and larger step sizes.

3.3. Averaging for larger step sizes. Polyak-Ruppert averaging [30, 29] is a standard way to improve convergence of SGD. In the discrete setting, convergence guarantees are considered at an averaged sequence defined by:

$$\begin{aligned} x_{k+1} &= x_k - \gamma \nabla f_{i_k}(x_k), \\ \bar{x}_k &= \frac{1}{k} \sum_{i=1}^k x_i. \end{aligned} \tag{3.4}$$

Other averaging techniques were later developed, such as primal averaging [40] and averaging with respect to some nonnegative function [19]. Primal averaging is detailed for comparison in Appendix C.

3.3.1. Polyak-Ruppert averaging. In this section, we analyze convergence properties of SDEs (3.3) with varying step sizes under Polyak-Ruppert averaging. Taylor and Bach [41, Theorem 6] provided a decreasing Lyapunov function along $(\bar{x}_k, x_k)$ (3.4), and a condition on the step size for convergence to the optimum. An

approximating SDE for Polyak-Ruppert averaging is:

$$\begin{aligned} dX_t &= -h_t \nabla f(X_t)dt + h_t(\gamma\Sigma(X_t))^{1/2}dB_t, \\ d\bar{X}_t &= \frac{X_t - \bar{X}_t}{t}dt, \end{aligned} \tag{3.5}$$

with step size $\gamma > 0$ that is taken close to zero, and a variable term $h_t \in [0,1]$. We introduce the family of quadratic Lyapunov functions:

$$\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t) = \begin{pmatrix} a_t^{(1)} \\ a_t^{(2)} \end{pmatrix}^\top \begin{pmatrix} f(X_t) - f_\star \\ f(\bar{X}_t) - f_\star \end{pmatrix} + \begin{pmatrix} X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}^\top P_t \begin{pmatrix} X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}, \tag{3.6}$$

where $a_t^{(1)}, a_t^{(2)} \geqslant 0$ are differentiable real functions, and $P_t = \begin{pmatrix} p_t^{(11)} & p_t^{(12)} \\ p_t^{(12)} & p_t^{(22)} \end{pmatrix} \succeq 0$ has differentiable parameters. Given a Lyapunov function $\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}$, and SDEs (3.5), the performance estimation problem is formulated by:

$$\begin{aligned} \max_{f\in\mathcal{F}_{0,\infty}, d\in\mathbf{N}, X_t\in\mathbf{R}^d} \quad & \frac{d}{dt}\mathbf{E}\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t), \\ \text{subject to } & dX_t = -h_t\nabla f(X_t)dt + h_t(\gamma\Sigma(X_t))^{1/2}dB_t, \\ & d\bar{X}_t = \frac{X_t - \bar{X}_t}{t}dt. \end{aligned}$$

A possible control of this quantity is presented in Theorem 3.4.

THEOREM 3.4. *Let $f \in \mathcal{F}_{0,\infty}$ be twice differentiable functions, and SDEs be taken under Polyak-Ruppert averaging (3.5) starting from $x_0 \in \mathbf{R}^d$, where $d \in \mathbf{N}$ is the dimension, and $\mathcal{V}$ be quadratic Lyapunov functions defined by (3.6). The following assertions are equivalent,*

- *The inequality $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,\bar{X}_t,t) \leqslant \frac{1}{2}Tr((a_t^{(1)}\nabla_{xx}f(X_t) + 2p_t^{(11)}I_d)\Sigma(X_t))h_t^2\gamma$ is satisfied for all functions $f \in \mathcal{F}_{0,\infty}$, all trajectories $(X_t,\bar{X}_t)$ generated by SDEs under Polyak-Ruppert averaging (3.5), and all dimensions $d \in \mathbf{N}$.*
- *There exist $\lambda_t^{(1)},...,\lambda_t^{(6)} \geqslant 0$ such that,*

$$\begin{pmatrix} \dot{p}_t^{11} + \frac{2p_t^{(12)}}{t} & \dot{p}_t^{12} - \frac{p_t^{(12)}}{t} + \frac{p_t^{(22)}}{t} & \frac{\lambda_t^{(6)}+\lambda_t^{(4)}}{2} - h_t p_t^{(11)} & \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} \\ * & \dot{p}_t^{22} - \frac{2p_t^{(22)}}{t} & -\frac{\lambda_t^{(6)}}{2} - h_t p_t^{(12)} & -\frac{a_t^{(2)}}{2t} + \frac{\lambda_t^{(5)}+\lambda_t^{(3)}}{2} \\ * & * & -a_t^{(1)} & 0 \\ * & * & * & 0 \end{pmatrix} \preceq 0,$$

$$\dot{a}_t^{(1)} + \lambda_t^{(1)} + \lambda_t^{(5)} = \lambda_t^{(4)} + \lambda_t^{(6)},$$
$$\dot{a}_t^{(2)} + \lambda_t^{(2)} + \lambda_t^{(6)} = \lambda_t^{(3)} + \lambda_t^{(5)} + \dot{a}_t^{(1)}.$$

Proof. The proof follows the methodology from Section 2.1.1, and using the Gram matrix $G = P^\top P$, where $P = (X_t - x_\star, \bar{X}_t - x_\star, g_t, \bar{g}_t)$ (see Appendix B.1). ∎

The variance term increases with $a_t^{(1)}$, and its convergence requires an additional smoothness assumption on f. For this reason, we propose to analyze the convergence based on Lyapunov functions on the averaged sequence only ($\mathcal{V}_{0,a_t^{(2)},P_t}$).

COROLLARY 3.5 (Averaging and diminishing step sizes). *Let $f \in \mathcal{F}_{0,\infty}$ be twice differentiable functions, SDEs be taken under Polyak-Ruppert averaging (3.5) starting from $x_0 \in \mathbf{R}^d$ where $d \in \mathbf{N}$ is the dimension. Assuming $\dot{a}_t^{(2)} \leqslant \frac{a_t^{(2)}}{t}$ and $t \to \frac{a_t^{(2)}}{th_t}$ is a non-increasing function, the Lyapunov function*

$$\mathcal{V}(X_t, \bar{X}_t, t) = a_t^{(2)}(f(\bar{X}_t) - f_\star) + \frac{a_t^{(2)}}{2th_t}\|X_t - X_\star\|^2$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, \bar{X}_t, t) \leqslant \frac{a_t^{(2)}}{t} h_t \, Tr(\Sigma(X_t))$ for all twice differentiable functions $f \in \mathcal{F}_{0,\infty}$, all trajectories X_t generated by SDEs (3.5) and all dimensions $d \in \mathbf{N}$. Then, it holds that $\mathbf{E}[f(\bar{X}_t) - f_\star] \leqslant \dfrac{\|x_0 - x_\star\|^2}{2a_t^{(2)}} + \dfrac{\gamma}{2a_t^{(2)}} \displaystyle\int_0^t \frac{a_s^{(2)}}{s} h_s \, Tr(\Sigma(X_s))ds$.

Proof. The proof follows from Theorem 3.4 (see Appendix B.1). ∎

When $a_t^{(2)} = t$ (its maximal possible value), the step size verifies $\dot{h}_t \geqslant 0$. The variance term does not diverge if and only if h_t is constant. Then, a convergence bound is given by $\mathbf{E}[f(\bar{X}_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{2t} + \frac{1}{2}\mathrm{Tr}(\sigma)\gamma h$. The decreasing condition on $t \to \frac{a_t^{(2)}}{th_t}$ suggests a trade-off between converging and diminishing step size, as obtained without averaging.

Under the assumptions of Corollary 3.5, let us consider a bounded covariance matrix $\Sigma(X_t) \preceq \Sigma$, a step size $h_t = \frac{1}{(t+1)^\alpha}$, and $a_t^{(2)} = t^\beta$ for $\alpha \geqslant 0$ and $0 \leqslant \beta \leqslant 1$ some parameters. The decreasing condition imposes $\alpha + \beta \leqslant 1$. A different behavior is expected from α and β: on the one hand, an ideal step size should be large (α small), and on the other hand, we aim at converging as fast as possible (β large). The term that forgets the initial conditions is bounded by $O(\frac{1}{t^\beta})$, and the variance term by $O(\frac{1}{t^\alpha})$ if $\beta \neq \alpha$. When $\alpha = \beta$, the variance term is in $\frac{\log(t)}{2t^\beta}$. Hence, a natural choice is $\alpha = \beta = \frac{1}{2}$, retrieving results from [2, Theorem 4] [41, Table 2] in discrete optimization.

3.3.2. Weighted averaging. Polyak-Ruppert performs uniform averaging of trajectories X_t over the time step. We introduce weighted averaging to analyze SGD, that is defined with respect to a function $u_t \geqslant 0$ [19],

$$\bar{x}_t^u = \frac{1}{\int_0^t u_s ds} \int_0^t u_s x_s ds.$$

Under weighted averaging, and with $C_t^u = \frac{u_t}{\int_0^t u_s ds}$, the SDE is

$$\begin{aligned} dX_t &= -h_t \nabla f(X_t)dt + h_t(\gamma\Sigma(X_t))^{1/2} dB_t, \\ d\bar{X}_t^u &= (X_t - \bar{X}_t^u) C_t^u dt. \end{aligned} \tag{3.7}$$

We study convergence of this generalized version of Polyak-Ruppert averaging, and compare it to traditional averaging techniques.

THEOREM 3.6. *Let $f \in \mathcal{F}_{\mu,\infty}$ be twice differentiable functions, possibly strongly convex $\mu \geqslant 0$, and SDEs be given by (3.7), starting from $x_0 \in \mathbf{R}^d$ where $d \in \mathbf{N}$ is the dimension. Assuming $\left(\frac{u_t}{2h_t}\right)^{\cdot} \leqslant 2\mu u_t$, the Lyapunov function*

$$\mathcal{V}(X_t, \bar{X}_t^u, t) = \frac{u_t}{C_t^u}(f(\bar{X}_t^u) - f_\star) + \frac{u_t}{2h_t}\|X_t - x_\star\|^2,$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, \bar{X}_t^u, t) \leqslant \frac{1}{2}u_t h_t \gamma\, Tr(\Sigma(X_t))$ *for all twice differentiable functions* $f \in \mathcal{F}_{0,\infty}$*, all trajectories* $(X_t, \bar{X}_t)$ *generated by SDEs with averaging* (3.7)*, and all dimensions* $d \in \mathbf{N}$.

Then, it holds that $\mathbf{E}[f(\bar{X}_t^u) - f_\star] \leqslant \dfrac{\|x_0 - x_\star\|^2 u_0}{2h_0 \int_0^t u_s ds} + \dfrac{\gamma}{4\int_0^t u_s ds}\displaystyle\int_0^t u_s h_s\, Tr(\Sigma(X_s))ds.$

Proof. These results are obtained by replacing $\frac{1}{t} \to C_t^u$ (see Appendix B.2). ∎

Let us consider convex functions $f \in \mathcal{F}_{0,\infty}$, a bounded covariance matrix $\Sigma(X_t) \preceq \Sigma$, a step size $h_t = \frac{1}{(t+1)^\alpha}$, and an averaging function $u_t = \frac{1}{(t+1)^\beta}$, with $\alpha, \beta \geqslant 0$. From Theorem 3.6, the terms that forgets the initial conditions is in $\frac{1}{(t+1)^{1-\beta}}$, and the variance term in $\frac{1}{(t+1)^\alpha}$. Both terms converge at same rate for $\alpha = \beta = \frac{1}{2}$ (up to $\log(t+1)$). We retrieve convergence results for SGD under Polyak-Ruppert averaging [2, Theorem 6].

For strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$, polynomial convergence can be reached for the term that contains the initial conditions. However, the variance term cannot converge faster than the step size. Given a bounded covariance matrix $\Sigma(X_t) \preceq \Sigma$, the variance term behaves after integrating by part $\frac{\int_0^t u_s h_s ds}{\int_0^t u_s ds} = h_t - \frac{\int_0^t (\int u)\dot{h}_s ds}{\int_0^t u_s ds}$. For diminishing step sizes, $\frac{\int_0^t u_s h_s ds}{\int_0^t u_s ds} \geqslant h_t$. Hence, averaging allows a better convergence for the terms containing initial conditions, but does not play a role in the variance term. To conclude, weighted averaging does not improve convergence results obtained under Polyak-Ruppert or primal averaging. The trade-off between the forgetting of the initial conditions and the noise term mostly relies on step sizes.

We have analyzed convergence of SGD together with averaging techniques in continuous-time, using approximating SDEs (3.3). Compared with discrete time, the continuous-time analysis leads to similar convergence results, while benefiting from simpler formulations, fewer assumptions especially on step sizes. Using this approach, we analyzed the trade-off between non-uniform averaging and step sizes, paving the way to a better understanding of averaging techniques. In the next session, we explore new convergence analyses for stochastic accelerated methods.

4. Accelerating the gradient flow. For both stochastic and deterministic models, we have retrieved known convergence results for continuous-time models approximating optimization methods. In this section, we provide convergence guarantees for second order gradient flows, and in particular for AGF (2.14).

In the deterministic setting, convergence of gradient descent was improved using a momentum. In this section, let $f \in \mathcal{F}_{0,\infty}$ be twice differentiable function and $\gamma > 0$ be constant step sizes. Li [22, Theorem 16, Section 4.4] proved that Nesterov accelerated gradient has the approximating SDE (for order-1 weak approximations),

$$d^2X_t + \frac{3}{t}dX_t + \nabla f(X_t)dt + \sqrt{\gamma\Sigma(X_t)}dB_t = 0. \tag{4.1}$$

As for SGD, the Lyapunov function $\mathcal{V}(x,t) = t^2(f(x) - f_\star) + 2\|(x - x_\star) + \frac{t}{2}\dot{x}\|^2$ obtained from Theorem 2.6 does not allow to conclude about convergence to a stationary point of trajectories X_t generated by stochastic accelerated gradient flows (4.1). Assuming bounded covariance $\Sigma(X_t) \preceq \Sigma$ and applying Ito's formula to $\mathcal{V}(\cdot)$ along X_t,

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{t^2} + \gamma \mathrm{Tr}(\Sigma) 3t.$$

In the following, we explore Polyak-Ruppert averaging together with diminishing step sizes, to analyze convergence of second-order SDEs.

4.1. Averaging does not preserve convergence rates. Averaging was a key to success for improving convergence of SGD (see Section 3). It is natural to wonder if averaging preserves the acceleration of Nesterov's gradient flow [37]. Let us define stochastic sedond-order gradient flows with Polyak-Ruppert averaging:

$$\begin{aligned} &d^2X_t + \beta_t dX_t + \nabla f(X_t)dt + \sqrt{\gamma\Sigma}dB_t = 0, \\ &d\bar{X}_t = \frac{X_t - \bar{X}_t}{t}dt, \end{aligned} \tag{4.2}$$

where $\beta_t \geqslant 0$ is a function, and a family of quadratic Lyapunov functions,

$$\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t) = \begin{pmatrix} a_t^{(1)} \\ a_t^{(2)} \end{pmatrix}^\top \begin{pmatrix} f(X_t) - f_\star \\ f(\bar{X}_t) - f_\star \end{pmatrix} + \begin{pmatrix} \dot{X}_t \\ X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}^\top P_t \begin{pmatrix} \dot{X}_t \\ X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}, \tag{4.3}$$

where $P_t \succeq 0$ and $a_t^{(1)}, a_t^{(2)} \geqslant 0$ are differentiable functions.

THEOREM 4.1. *Let $f \in \mathcal{F}_{0,\infty}$ be twice differentiable functions, $\mathcal{V} = \mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}$ be quadratic Lyapunov functions, and stochastic accelerated gradient flows under Polyak-Ruppert averaging (4.2) with constant step sizes $h_t = 1$, starting from $x_0 \in \mathbf{R}^d$, where $d \in \mathbf{N}$ is the dimension. Then it holds that:*

- *When $a_t^{(1)} = 0$, if the function $\mathcal{V}$ verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,\bar{X}_t,t) \leqslant Tr(2p_t^{(11)}\gamma\Sigma(X_t))$ for all twice-differentiable functions $f \in \mathcal{F}_{0,\infty}$, all trajectories $(X_t,\bar{X}_t)$ generated by SDEs (4.2), and all dimensions $d \in \mathbf{N}$, then $\mathcal{V} = 0$.*
- *When $a_t^{(2)} = 0$, the Lyapunov function*

$$\mathcal{V}(X_t,t) = a_t^{(1)}(f(X_t) - f_\star) + \frac{1}{2a_t^{(1)}}\|a_t^{(1)}\dot{X}_t + \dot{a}_t^{(1)}(X_t - x_\star)\|^2,$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t) \leqslant Tr(2a_t^{(1)}\gamma\Sigma(X_t))$, with $a_t^{(1)} \leqslant t^2$ for all functions $f \in \mathcal{F}_{0,\infty}$, all trajectories X_t generated by stochastic second-order gradient flows (4.2) and all dimensions $d \in \mathbf{N}$. Then, it holds that $\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{2\|x_0 - x_\star\|^2}{a_t^{(1)}} + \frac{1}{2a_t^{(1)}}\gamma\int_0^t a_s^{(1)}\, Tr(\Sigma(X_s))ds$.

Proof. The proof follow from Theorem 2.10 (see Appendix B.3). ∎

To conclude, with constant step sizes, there is no Lyapunov functions that allows to forget the initial conditions while reducing the variance term. Primal averaging leads to similar results (see Appendix C). Therefore, averaging plays a different role in second-order SDEs than in first-order SDEs under Polyak-Ruppert averaging.

4.2. Diminishing step sizes. Averaging was not conclusive for finding a convergence guarantee of Nesterov's accelerated gradient flow with a diffusion term. Let us consider second-order stochastic gradient flows with varying step sizes $h_t \geqslant 0$,

$$d^2X_t + \beta_t dX_t + h_t\nabla f(X_t)dt + h_t\sqrt{\gamma\Sigma(X_t)}dB_t = 0. \tag{4.4}$$

We propose a Lyapunov function among the class of quadratic functions (2.11).

THEOREM 4.2. *Let $f \in \mathcal{F}_{0,\infty}$ be twice differentiable functions, X_t be a trajectory generated by stochastic second-order gradient flows (4.4) with varying step sizes $h_t \geqslant 0$ and $\gamma > 0$, starting from $x_0 \in \mathbf{R}^d$, where $d \in \mathbf{N}$ is the dimension. Assuming $\dot{a}_t \leqslant a_t \frac{2}{3}(\beta_t + \frac{1}{2}\frac{\dot{h}_t}{h_t})$ and $t \to \frac{(\dot{a}_t)^2}{2h_t a_t}$ a decreasing function, the Lyapunov function*

$$\mathcal{V}(X_t) = a_t(f(X_t) - f_\star) + \frac{1}{2h_t a_t}\|a_t\dot{X}_t + \dot{a}_t(X_t - x_\star)\|^2,$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \frac{1}{4}Tr(a_t h_t \Sigma(X_t))\gamma$ for all functions $f \in \mathcal{F}_{0,\infty}$, all trajectories X_t generated by stochastic second-order gradient flows (4.4) and all dimensions $d \in \mathbf{N}$.

Proof. This result is obtained extending the LMI for ODEs from Theorem 2.10 and Corollary 2.11 to varying step sizes. ∎

Let us now assume that $h_t = \frac{1}{(t+1)^\alpha}$ and $\beta_t = \frac{b}{t}$, where $\alpha, b > 0$. It follows from Theorem 4.2, that $\dot{a}_t \leqslant \beta\frac{a_t}{t}$, where $\beta \leqslant \frac{2b-\alpha}{3}$ and $\alpha + \beta \leqslant 2$, and

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{\beta^2}{t^\beta}\|x_0 - x_\star\|^2 + \frac{\gamma}{4t^\beta}\int_0^t \frac{s^\beta}{(s+1)^\alpha}\mathrm{Tr}(\Sigma(X_s))ds.$$

We assume bounded covariance of $\Sigma(X_t) \preceq \Sigma$, then $\beta \leqslant \min(\frac{2b-\alpha}{3}, 2-\alpha)$. On the one hand, the smaller the step sizes, the better the convergence for the term that contains the initial conditions. On the other hand, the variance term behaves as $\frac{1}{t^\beta}$ if $\beta \leqslant \alpha - 1$, and as $\frac{1}{t^{\alpha-1}}$ otherwise (convergence requiring then $\alpha \leqslant 1$ and $\beta \leqslant 1$). For Nesterov's accelerated gradient flow with $b = 3$, we have $\beta = 2 - \alpha \leqslant \alpha - 1$, and therefore $\alpha \geqslant \frac{3}{2}$. Taking $\alpha = \frac{3}{2}$, a convergence bound is given by:

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{9}{4\sqrt{t}}\|x_0 - x_\star\|^2 + \frac{\log t}{\sqrt{t}}\gamma\mathrm{Tr}(\Sigma).$$

We retrieve result from Corollary 3.5 for the SDE approximating SGD with Polyak-Ruppert averaging, with smaller step sizes. It does not seem possible to accelerate SGD when diminishing the step size. Ghadimi and Lan [31, Corollary 3] proved a convergence bound for a randomized stochastic accelerated gradient method with $\beta = 2$ and $\alpha = \frac{1}{2}$, that we do not retrieved. However, in their approach, the function is minimized over a compact, convex domain, whereas our approach focuses on an unbounded domain.

5. Conclusion and future work. We have developed a systematic approach for generating Lyapunov functions for families of ODEs and SDEs. Verifying such a Lyapunov function can be formulated as an LMI. From this formulation, it is possible to derive Lyapunov functions and the associated convergence bounds. We retrieve results from discrete optimization methods with shorter proofs, and fewer assumptions.

While obtaining guarantees for stochastic optimization methods might be tedious, the SDE approach allows for simpler analyses of the trade-off between the variance term and term that forgets the initial conditions. A shortcoming of this approach is that this analysis does not include approximation guarantees between optimization methods and their continuous-time counterparts. In the deterministic setting, stability techniques are often developed to quantify this approximation efficiency [13]. In stochastic analysis, stochastic modified equation have been introduced by Li et al. [21, Theorem 1] to better approximate SGD, and stochastic methods with momentum. For non-convex functions, some analysis have also been done by Shi et al [36]. However,

these approximation theorems often require many assumptions on the class of functions, which we believe could be further simplified using computer-assisted proofs.

Acknowledgments. This work was funded by MTE and the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19- P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063).

REFERENCES

[1] H. Attouch, Z. Chbani, and H. Riahi, *Rate of convergence of the nesterov accelerated gradient method in the subcritical case $\alpha \leqslant 3$.*, ESAIM: Control, Optimisation and Calculus of Variations, 25:2 (2019).

[2] F. Bach and E. Moulines, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, in Neural Information Processing Systems (NIPS), 2011.

[3] N. Bansal and A. Gupta, *Potential-function proofs for gradient methods*, Theory of Computing, 15 (2019), pp. 1–32.

[4] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, *Characterizations of Lojasiewicz inequalities: Subgradient flows, talweg, convexity*, Transactions of the American Mathematical Society (AMS), 362 (2010), pp. 3319–3363.

[5] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Mathematical Programming, 165 (2017), pp. 471–507.

[6] L. Bottou and O. Bousquet, *The tradeoffs of large scale learning*, in Advances in Neural Information Processing Systems (NIPS), 2007.

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

[8] J. Diakonikolas and M. Jordan, *Generalized momentumbased methods: A hamiltonian perspective*, SIAM Journal on Optimization, 31(1) (2021), p. 915–944.

[9] Y. Drori and M. Teboulle, *Performance of first-order methods for smooth convex minimization: a novel approach*, Mathematical Programming, 145 (2014), pp. 451–482.

[10] A. d'Aspremont, D. Scieur, and A. Taylor, *Acceleration methods*, Foundations and Trends in Optimization, 5 (2021), pp. 1–245.

[11] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, *Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems*, SIAM Journal on Optimization, 28 (2018), pp. 2654–2689.

[12] N. Flammarion and F. Bach, *From averaging to acceleration, there is only a step-size*, in Conference on Learning Theory (COLT), 2015.

[13] W. Gautschi, *Numerical Analysis*, Springer Science and Business Media, (2011).

[14] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, *Global convergence of the heavy-ball method for convex optimization*, in European control conference (ECC), 2015.

[15] B. Hu and L. Lessard, *Dissipativity theory for Nesterov's accelerated method*, in International Conference on Machine Learning (ICML), 2017.

[16] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems (NIPS), 2013.

[17] R. E. Kalman and J. E. Bertram, *Control system analysis and design via the "second method" of lyapunov: I-continuous-time systems.*, Journal of Basic Engineering, 82(2) (1960), pp. 371–393.

[18] W. Krichene, A. Bayen, and P. L. Bartlett, *Accelerated mirror descent in continuous and discrete time*, in Advances in Neural Information Processing Systems (NIPS), 2015.

[19] N. Le Roux, *Anytime tail averaging*, preprint arXiv:1902.05083, (2019).

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

[21] Q. Li, C. Tai, and W. E, *Stochastic modified equations and adaptive stochastic gradient algorithms*, in International Conference on Machine Learning (ICML), 2017.

[22] Q. Li, C. Tai, and W. E, *Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations*, The Journal of Machine Learning Research (JMLR), 20 (2019), pp. 1–47.

[23] S. Lojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels, in: Les equations aux dérivées partielles*, Editions du centre National de la Recherche Scientifique,

(1963), pp. 87–89.
[24] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, 175 (2019), pp. 69–107.
[25] Y. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Dokl. akad. nauk Sssr, 269 (1983), pp. 543–547.
[26] A. Orvieto and A. Lucchi, *Continuous-time models for stochastic optimization algorithms*, in Advances in Neural Information Processing Systems (NeurIPS), 2020.
[27] B. Polyak, *Some methods of speeding up the convergence of iteration methods*, 1964.
[28] B. T. Polyak, *Gradient methods for minimizing functionals*, USSR Computational Mathematics and Mathematical Physics, 4 (1963), pp. 864–878.
[29] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization, 30 (1992), pp. 838–855.
[30] D. Ruppert, *Efficient estimations from a slowly convergent Robbins Monroe process. technical report.*, 1988.
[31] G. L. Saeed Ghadimi, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Mathematical Programming Series A, 156 (2015), pp. 59–99.
[32] J. M. Sanz Serna and K. C. Zygalakis, *The connections between lyapunov functions for some optimization algorithms and differential equations*, SIAM Journal on Numerical Analysis, 59 (2021), pp. 1542–1565.
[33] D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont, *Integration methods and accelerated optimization algorithms*, in Advances in Neural Information Processing Systems (NIPS), 2017.
[34] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Mathematical Programming, (2018), pp. 1–70.
[35] B. Shi, S. S. Du, W. J. Su, and M. I. Jordan, *Acceleration via symplectic discretization of high-resolution differential equations*, in Advances in Neural Information Processing Systems (NeurIPS), 2019.
[36] B. Shi, W. J. Su, and M. I. Jordan, *On learning rates and schr\" odinger operators*, preprint arXiv:2004.06977, (2020).
[37] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, The Journal of Machine Learning Research (JMLR), 17 (2016), pp. 1–43.
[38] J. J. Suh, G. Roh, and E. K. Ryu, *Continuous-time analysis of agm via conservation laws in dilated coordinate systems*, 2022, https://arxiv.org/abs/2202.05501.
[39] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*, Institute of Mathematical Statistics Textbooks, Cambridge University Press, 2019.
[40] W. Tao, Z. Pan, G. Wu, and Q. Tao, *Primal averaging: A new gradient evaluation step to attain the optimal individual convergence*, IEEE transactions on cybernetics, 50 (2018), pp. 835–845.
[41] A. Taylor and F. Bach, *Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions*, in Conference on Learning Theory (COLT), 2021.
[42] A. Taylor, B. V. Scoy, and L. Lessard, *Lyapunov functions for first-order methods: Tight automated convergence guarantees*, in International Conference on Machine Learning (ICML), 2018.
[43] A. B. Taylor, *Interpolation and performance estimation of first-order methods for convex optimization, PhD Thesis, Chapter 3 : Convex interpolation*, 2017.
[44] A. B. Taylor, J. M. Hendrickx, and F. Glineur, *Exact worst-case performance of first-order methods for composite convex optimization*, SIAM Journal on Optimization, 27 (2017), p. 1283–1313.
[45] A. B. Taylor, J. M. Hendrickx, and F. Glineur, *Smooth strongly convex interpolation and exact worst-case performance of first-order methods*, Mathematical Programming, 161 (2017), pp. 307–345.
[46] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, in Proceedings of the National Academy of Sciences, 2016.
[47] A. C. Wilson, B. Recht, and M. I. Jordan, *A Lyapunov analysis of accelerated methods in optimization*, The Journal of Machine Learning Research (JMLR), 22 (2021), pp. 1–34.
[48] P. Xu, T. Wang, and Q. Gu, *Accelerated stochastic mirror descent: From continuous-time dynamics to discrete-time algorithms*, in International Conference on Artificial Intelligence and Statistics (AISTATS), 2018.
[49] P. Xu, T. Wang, and Q. Gu, *Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions*, in International Conference on Machine Learning (ICML), 2018.

Appendix A. Proof for ODEs.

A.1. Proofs for Theorems 2.3 and 2.1.

Proof. Let us consider X_t a solution to the gradient flow (1.3) starting from $x_0 \in \mathbf{R}^d$, and a quadratic Lyapunov function of the form $\mathcal{V}_{a_t,c_t}(X_t,t) = a_t \cdot (f(X_t) - f_\star) + c_t \cdot \|X_t - x_\star\|^2$ such that a_t, c_t are differentiable and $\mathcal{V}_{a_t,c_t}$ is nonnegative over the trajectory X_t. The family of function f is supposed to be μ-strongly convex, with $\mu \geqslant 0$. We are going to prove the LMI equivalence of Theorem 2.3 and Theorem 2.1.

Let us first rephrase our objective, that is computing a Lyapunov function that is decreasing along X_t.

$$\begin{aligned} 0 \geqslant \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N},\ f \in \mathcal{F}_{\mu,\infty}} \quad & \frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t,t) \\ \text{subject to} \quad & \dot{X}_t = -\nabla f(X_t). \end{aligned}$$

Since $f \in \mathcal{F}_{\mu,\infty}$, this problem is at first sight infinite-dimensional, and thus not easily solvable. Let us reformulate it as an feasibility condition over the class of functions,

$$\begin{aligned} 0 \geqslant \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N}} \quad & \frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t,t), \\ \text{subject to } & \dot{X}_t = -\nabla f(X_t), \\ & \exists f \in \mathcal{F}_{\mu,\infty} : \begin{cases} g_t = \nabla f(X_t) & f_t = f(X_t), \\ g_\star = \nabla f(x_\star) = 0 & f_\star = f(x_\star). \end{cases} \end{aligned}$$

This problem remains infinite-dimensional, but we can reformulate it thanks to an interpolation theorem of the class of functions $\mathcal{F}_{\mu,L}$ on (X_i, g_i, f_i), with $0 \leqslant \mu \leqslant L \leqslant \infty$. A set $\{(X_i, g_i, f_i)_{i\in I}\}$ is said to be $\mathcal{F}_{\mu,L}$-*interpolable* if there exist $f \in \mathcal{F}_{\mu,L}$ such that $f_i = f(X_i)$ and $g_i = \nabla f(X_i)$ for all $i \in I$. It turns out that the formulation for this class of functions is quite simple.

THEOREM A.1. *[45, Theorem 4]: Set $\{(X_i, g_i, f_i)\}_{i\in I}$ if $\mathcal{F}_{\mu,L}$ interpolable if and only if the following set of conditions holds for every pair of indices $i \in I$ and $j \in J$*

$$f_i - f_j - \langle g_j, X_i - X_j\rangle \geqslant \frac{\left(\frac{1}{L}\|g_i - g_j\|^2 + \mu\|X_i - X_j\|^2 - 2\frac{\mu}{L}\langle g_i - g_j, X_i - g_j\rangle\right)}{2(1-\frac{\mu}{L})}.$$

This theorem allows to replace the problem above by a quadratic program, linear in $f_t, f_\star$) and quadratic in $(x_t, g_t, x_\star)$.

$$\begin{aligned} 0 \geqslant \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N}} \quad & \frac{d}{dt}\mathcal{V}_{a_t,c_t}, \\ \text{subject to } & \dot{X}_t = -g_t, \\ & f_i - f_j - \langle g_j, X_i - X_j\rangle \geqslant \frac{\mu}{2}\|X_i - X_j\|^2,\ \forall i,j = t,\star. \end{aligned}$$

Let G be a Gram matrix defined by $G = \begin{pmatrix} \|X_t - x_\star\|^2 & \langle X_t - x_\star, g_t\rangle \\ \langle X_t - x_\star, g_t\rangle & \|g_t\|^2 \end{pmatrix} \succeq 0$ and the vector $F = [f_t, f_\star]$. The quadratic program can thus be formulated into a semidefinite

program,

$$\begin{aligned} 0 \geqslant \max_{G\succeq 0, F\in\mathbf{R}^2} & b_0^\top F + \mathrm{Tr}(A_0 G), \\ \text{subject to } & b_1^\top F + \mathrm{Tr}(A_1 G) \geqslant 0, \\ & b_2^\top F + \mathrm{Tr}(A_2 G) \geqslant 0, \end{aligned}$$

where $A_0 = \begin{pmatrix} \dot{c}_t & -c_t \\ -c_t & -a_t \end{pmatrix}$, $A_1 = \begin{pmatrix} -\mu/2 & 1/2 \\ 1/2 & 0 \end{pmatrix}$, $A_2 = \begin{pmatrix} -\mu/2 & 0 \\ 0 & 0 \end{pmatrix}$, $b_0 = \dot{a}_t[1,\ -1]^\top$ $b_1 = [-1,\ 1]^\top$ and $b_2 = [1,\ -1]^\top$.

In this convex case, strong duality holds via Slater's conditions [45, Theorem 6]. Consider the Lagrangian dual of the SDP,

$$\begin{aligned} \min_{\lambda_t^{(1)},\lambda_t^{(2)}\geqslant 0} & \ 0 \\ \text{subject to } & A_0 + \lambda_t^{(1)} A_1 + \lambda_t^{(2)} A_2 \preceq 0 \\ & \lambda_t^{(2)} b_2 + \lambda_t^{(1)} b_1 + b_0 = 0. \end{aligned}$$

This problem is a feasibility problem,

$$\begin{aligned} \min_{\lambda_t^{(1)},\lambda_t^{(2)}\geqslant 0} & \ 0 \\ \text{subject to } & \ S = \begin{pmatrix} \dot{c}_t - \frac{\mu}{2}(\lambda_t^{(1)} + \lambda_t^{(2)}) & -c_t + \frac{\lambda_t^{(1)}}{2} \\ -c_t + \frac{\lambda_t^{(1)}}{2} & -a_t \end{pmatrix} \preceq 0, \\ & \dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}. \end{aligned}$$

When $\mu = 0$, this formulation is exactly the LMI formulation obtained in Theorem 2.3. When $\mu > 0$, the LMI corresponds to Theorem 2.1 for $\dot{a}_t = e^{2\mu t} a$ and $\dot{c}_t = e^{2\mu t} c$. ∎

A.2. Proof for Corollary 2.11.

Proof. From Theorem 2.3, we get a LMI reformulation to finding a suitable Lyapunov function: There exist $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that,

$$\begin{aligned} & \begin{pmatrix} -\frac{\mu}{2}(\lambda_t^{(1)} + \lambda_t^{(2)}) + p_{11}^{\cdot} & p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} & -p_t^{(12)} + \frac{\lambda_1}{2} \\ * & 2(p_t^{(12)} - \beta_t p_t^{(22)}) + \dot{p}_t^{(22)} & -p_t^{(22)} + \frac{a_t}{2} \\ * & * & 0 \end{pmatrix} \preceq 0, \\ & \dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}. \end{aligned}$$

Convex functions. Let us first assume the functions f to minimize are convex ($\mu = 0$). While determining a suitable Lyapunov function $\mathcal{V}_{a_t,P_t}$, we try to find a large parameter a_t at time t. To this end, we saturate $\dot{a}_t = \lambda_t^{(1)}$ ($\lambda_t^{(2)} = 0$). The LMI can then be simplified into:

$$\begin{aligned} & S = \begin{pmatrix} \dot{p}_t^{(11)} & p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} \\ * & 2(p_t^{(12)} - \beta_t p_t^{(22)}) + \dot{p}_t^{(22)} \end{pmatrix} \preceq 0, \\ & p_t^{(22)} = \frac{a_t}{2}, \\ & p_t^{(12)} = \frac{\dot{a}_t}{2}. \end{aligned}$$

From linear algebra, the matrix S is semi-definite negative if and only if the diagonal terms are negative. We obtain two conditions on a_t:

$$\dot{p}_t^{(11)} \leqslant 0,$$
$$\dot{a}_t - \beta_t a_t + \frac{1}{2}\dot{a}_t \leqslant 0.$$

In addition, we assumed $P_t \preceq 0$ to enforce positivity of the Lyapunov function $\mathcal{V}_{a_t,P_t}$, which is equivalent to $p_t^{(11)} \geqslant \frac{(p_t^{(12)})^2}{p_t^{(22)}} = \frac{(\dot{a}_t)^2}{2a_t}$.

On the one hand, we conclude that the function $\frac{\dot{a}_t}{2\sqrt{a_t}} \leqslant p_0^{(11)}$, and after integrating between 0 and t, that $\sqrt{a_t} \leqslant \sqrt{a_0} + \frac{\sqrt{p_0^{(11)}}}{2}t$ ($a_t = O(t^2)$). On the other hand, we conclude that $\dot{a}_t \leqslant \frac{2}{3}\beta_t a_t$. For all $\epsilon > 0$, after integrating between ϵ, and t, that $a_t \leqslant a_\epsilon e^{\int_\epsilon^\top \frac{2}{3}\beta_s ds}$. Therefore, $a_t \leqslant \lim_{\epsilon\to 0} a_\epsilon e^{\int_\epsilon^\top \frac{2}{3}\beta_s ds}$, and a bound on a_t is given by:

$$a_t \leqslant \min((\sqrt{a_0} + (\sqrt{p_0^{(11)}}/2)t)^2, \lim_{\epsilon\to 0} a_\epsilon e^{\int_\epsilon^\top \frac{2}{3}\beta_s ds}).$$

Assuming for instance $a_0 = 0$ and $p_0^{(11)} = 2$, we obtain:

$$a_t \leqslant \min(t^2, \lim_{\epsilon\to 0} a_\epsilon e^{\int_\epsilon^\top \frac{2}{3}\beta_s ds}).$$

Strongly convex functions. Similar result can be obtain assuming an exponential form for the parameters $a_t = ae^{t\tau}$, and $P_t = Pe^{t\tau}$, where $\tau > 0$ is a convergence parameter to determine. ∎

Taking a generic form for $\beta_t = \frac{r}{t}$), the condition $\lim_{\epsilon\to 0} a_\epsilon e^{\int_\epsilon^\top \frac{2}{3}\beta_s ds}$ takes more sense. Indeed, $a_t = t^{2r/3}$ is a solution to $a_t = \lim_{\epsilon\to 0} a_\epsilon e^{\int_\epsilon^\top \frac{2}{3}\beta_s ds}$ (which requires $a_0 = 0$).

Appendix B. Proofs for SDEs.

B.1. Proof for Theorem 3.4 and Corollary 3.5.

Proof for Theorem 3.4. We consider the stochastic gradient flow under Polyak-Ruppert averaging (3.5), that is recalled here:

$$dX_t = -h_t \nabla f(X_t)dt + h_t(\gamma\Sigma(X_t))^{1/2}dB_t,$$
$$d\bar{X}_t = \frac{X_t - \bar{X}_t}{t}dt,$$

where h_t is the varying part of the step size, and $\gamma > 0$ the maximum allowed step size.

Let us consider the family of quadratic Lyapunov functions defined by

$$\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t) = \begin{pmatrix} a_t^{(1)} \\ a_t^{(2)} \end{pmatrix}^\top \begin{pmatrix} f(X_t) - f_\star \\ f(\bar{X}_t) - f_\star \end{pmatrix} + \begin{pmatrix} X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}^\top P_t \begin{pmatrix} X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}, \tag{B.1}$$

where $a_t^{(1)}, a_t^{(2)}$ are differentiable real functions, and $P_t = \begin{pmatrix} p_t^{(11)} & p_t^{(12)} \\ p_t^{(12)} & p_t^{(22)} \end{pmatrix} \succeq 0$ has differentiable parameters. We look for a worst-case guarantee in the Lyapunov approach,

that can be cast as a minimization problem at time t,

$$\begin{aligned}\max_{X_t\in\mathbf{R}^d, d\in\mathbf{N},\ f\in\mathcal{F}_{0,\infty}} \quad & \frac{d}{dt}\mathbf{E}\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t),\\ \text{subject to} \quad & dX_t = -h_t\nabla f(X_t)dt + h_t(\gamma\Sigma(X_t))^{1/2}dB_t,\\ & d\bar{X}_t = \frac{X_t-\bar{X}_t}{t}dt.\end{aligned}$$

Let us rewrite the SDE,

$$\begin{pmatrix} dX_t \\ d\bar{X}_t\end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \frac{1}{t} & -\frac{1}{t}\end{pmatrix}\begin{pmatrix} X_t \\ \bar{X}_t\end{pmatrix}dt + \begin{pmatrix} -h_t & 0 \\ 0 & 0\end{pmatrix}\begin{pmatrix}\nabla f(X_t)\\ \nabla f(\bar{X}_t)\end{pmatrix}dt + \begin{pmatrix} h_t(\gamma\Sigma(X_t))^{1/2}\\ 0\end{pmatrix}dB_t.$$

Let us denote $Y_t = \begin{pmatrix} X_t \\ \bar{X}_t\end{pmatrix}$, and $\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t) = \mathcal{V}(Y_t,t)$, then $\frac{d}{dt}\mathcal{V}(Y_t,t) = \frac{\partial}{\partial t}\mathcal{V}(Y_t,t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt} + \frac{1}{2}\gamma h_t^2\mathrm{Tr}[\frac{\partial^2}{\partial Y_t^2}\mathcal{V}(Y_t,t)^\top\begin{pmatrix}\Sigma(X_t)\\ 0\end{pmatrix}]$. Then, we have,

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t) &= \frac{d}{dt}\mathbf{E}\mathcal{V}(Y_t,t),\\ &= \mathbf{E}[\frac{\partial}{\partial t}\mathcal{V}(Y_t,t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt}]\\ &\quad + \frac{1}{2}\gamma h_t^2\mathrm{Tr}(\frac{\partial^2}{\partial Y_t^2}\mathcal{V}(Y_t,t)^\top\begin{pmatrix}\Sigma(X_t)\\ 0\end{pmatrix}).\end{aligned}$$

The noise term only depends on the second derivative of f in X_t and contains a noise term $\Sigma(X_t)$. We propose to not take the noise term into account in the SDP formulation. The inequality

$$\frac{d}{dt}\mathbf{E}\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t) \leqslant \frac{1}{2}\gamma h_t^2\mathrm{Tr}(\frac{\partial^2}{\partial X_t^2}\mathcal{V}(X_t,\bar{X}_t,t)^\top\Sigma(X_t)), \tag{B.2}$$

is satisfied for any trajectory $(X_t,\bar{X}_t)$ generated by the SDE under Polyak-Ruppert averaging (3.5) and any function $f\in\mathcal{F}_{0,\infty}$. This is directly equivalent to $\mathbf{E}[\frac{\partial}{\partial t}\mathcal{V}(Y_t,t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt}] \leqslant 0$, where $Y_t = \begin{pmatrix} X_t & \bar{X}_t\end{pmatrix}^\top$. We notice that this inequality is exactly equivalent to $\frac{d}{dt}\mathcal{V}(\tilde{Y}_t,t)\leqslant 0$ where $\tilde{Y}_t = \begin{pmatrix}\tilde{X}_t & \tilde{\bar{X}}_t\end{pmatrix}^\top$ are trajectories generated by the deterministic gradient flow obtained from the SDE (3.5) with $\gamma = 0$. Thus, an equivalent reformulation to Equation (B.2) is given by:

$$\begin{aligned}0 \geqslant \max_{X_t\in\mathbf{R}^d, d\in\mathbf{N},\ f\in\mathcal{F}_{0,\infty}} \quad & \frac{d}{dt}\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t)\\ \text{subject to} \quad & dX_t = -h_t\nabla f(\bar{X}_t)dt,\\ & d\bar{X}_t = \frac{X_t-\bar{X}_t}{t}dt.\end{aligned}$$

Using convex interpolation on triplets $\{(X_t,g_t,f_t),\ (\bar{X}_t,\bar{g}_t,\bar{f}_t),\ (X_\star,0,f(X_\star))\}$ such that there exists $f\in\mathcal{F}_{0,\infty}$ verifying $f_t = f(X_t)$, $g_t = \nabla f(X_t)$, $f(\bar{X}_t) = \bar{f}_t$ and $\bar{f}_t = \nabla f(\bar{X}_t)$, the infinite dimensional problem has a quadratic program formulation. Introducing the Gram matrix $G = Y^\top Y \succeq 0$, $Y = [X_t - x_\star, \bar{X}_t - x_\star, g_t, \bar{g}_t]$, the

maximization problem is reformulated as an SDP program, whose dual is the following LMI,

$$\begin{pmatrix} \dot{p}_t^{(11)} + \frac{2p_t^{(12)}}{t} & \dot{p}_t^{(12)} - \frac{p_t^{(12)}}{t} + \frac{p_t^{(22)}}{t} & \frac{\lambda_t^{(6)}+\lambda_t^{(4)}}{2} - h_t p_t^{(11)} & \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} \\ * & \dot{p}_t^{(22)} - \frac{2p_t^{(22)}}{t} & -h_t p_t^{(12)} - \frac{\lambda_t^{(6)}}{2} & \frac{\lambda_t^{(3)}+\lambda_t^{(5)}}{2} - \frac{a_t^{(2)}}{2t} \\ * & * & -a_t^{(11)} & 0 \\ * & * & * & 0 \end{pmatrix} \preceq 0,$$

$$\lambda_t^{(6)} + \lambda_t^{(4)} = \lambda_t^{(5)} + \lambda_t^{(1)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(2)} + \lambda_t^{(6)} + \dot{a}_t^{(2)} = \lambda_t^{(3)} + \lambda_t^{(5)},$$
$$\lambda_t^{(1)} + \lambda_t^{(2)} + \dot{a}_t^{(2)} = \lambda_t^{(4)} + \lambda_t^{(5)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(i)} \geqslant 0,\ i \in \{1, ..., 6\},$$

where $\lambda_t^{(i)},\ i \in \{1, ..., 6\}$ are dual values associated with interpolation inequalities. These three inequalities can be reduced to two independent equalities. ∎

Proof for Corollary 3.5. Using the LMI formulation of Theorem 3.4, the Lyapunov function $\mathcal{V}(X_t, \bar{X}_t, t) = a_t^{(2)}(f(\bar{X}_t) - f_\star) + p_t^{(11)}\|X_t - X_\star\|^2$ satisfies the LMI:

$$\begin{pmatrix} \dot{p}_t^{(11)} & 0 & \frac{\lambda_t^{(6)}+\lambda_t^{(4)}}{2} - h_t p_t^{(11)} & \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} \\ * & 0 & -\frac{\lambda_t^{(6)}}{2} & \frac{\lambda_t^{(3)}+\lambda_t^{(5)}}{2} - \frac{a_t^{(2)}}{2t} \\ * & * & 0 & 0 \\ * & * & * & 0 \end{pmatrix} \preceq 0,$$

$$\lambda_t^{(6)} + \lambda_t^{(4)} = \lambda_t^{(5)} + \lambda_t^{(1)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(2)} + \lambda_t^{(6)} + \dot{a}_t^{(2)} = \lambda_t^{(3)} + \lambda_t^{(5)},$$
$$\lambda_t^{(i)} \geqslant 0,\ i \in \{1, ..., 6\}.$$

From the LMI, we conclude that $\lambda_t^{(3)} = \lambda_t^{(6)} = 0$, $\lambda_t^{(5)} = \frac{a_t^{(2)}}{t}$, $\lambda_t^{(4)} = h_t p_t^{(11)}$. The LMI is verified if, and only if,

$$\dot{a}_t^{(2)} \leqslant \lambda_t^{(2)} + \dot{a}_t^{(2)} = \lambda_t^{(5)} = \frac{a_t^{(2)}}{t}$$
$$\dot{p}_t^{(11)} = \left(\frac{\dot{a}_t^{(2)}}{t h_t}\right) \leqslant 0.$$

The bound in function values follows from this Lyapunov function integrated between 0 and t. ∎

B.2. Proof for Theorem 3.6.

Proof. Following the same reasoning as in the proof for Theorem 2.3, the worst-case guarantee can be formulated as the maximization problem:

$$0 \geqslant \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N},\ f \in \mathcal{F}_{0,\infty}} \frac{d}{dt}\mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}(X_t, \bar{X}_t, t)$$
$$\text{subject to } \ dX_t = -h_t \nabla f(X_t)dt + h_t(\gamma \Sigma(X_t))^{1/2} dB_t,$$
$$d\bar{X}_t^u = (X_t - \bar{X}_t^u) C_t^u dt,$$

where $\bar{x}_t^u = \frac{1}{\int_0^t u_s ds} \int_0^\top u_s x_s ds$ and $C_t^u = \frac{u_t}{\int_0^t u_s ds}$. Its LMI equivalent reformulation is given by:

$$\begin{pmatrix} \dot{p}_t^{(11)} & \dot{p}_t^{(12)} & \frac{\lambda_t^{(6)}+\lambda_t^{(4)}}{2} - h_t p_t^{(11)} & a_t^{(2)} C_t^u - \frac{\lambda_t^{(5)}}{2} \\ +2p_t^{(12)} C_t^u & +(p_t^{(22)} - p_t^{(12)}) C_t^u & & \\ * & \dot{p}_t^{(22)} - 2p_t^{(22)} C_t^u & -h_t p_t^{(12)} - \frac{\lambda_t^{(6)}}{2} & \frac{\lambda_t^{(3)}+\lambda_t^{(5)}}{2} - a_t^{(2)} C_t^u \\ * & * & -a_t^{(11)} & 0 \\ * & * & * & 0 \end{pmatrix} \preceq 0,$$

$$\lambda_t^{(6)} + \lambda_t^{(4)} = \lambda_t^{(5)} + \lambda_t^{(1)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(2)} + \lambda_t^{(6)} + \dot{a}_t^{(2)} = \lambda_t^{(3)} + \lambda_t^{(5)},$$
$$\lambda_t^{(1)} + \lambda_t^{(2)} + \dot{a}_t^{(2)} = \lambda_t^{(4)} + \lambda_t^{(5)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(i)} \geqslant 0, \ i \in \{1, ..., 6\}.$$

Considering the Lyapunov function $\mathcal{V}(X_t, \bar{X}_t^u, t) = a_t^{(2)}(f(\bar{X}_t^u) - f_\star) + p_t^{(11)} \|X_t - x_\star\|^2$, it follows that $\dot{a}_t^{(2)} \leqslant C_t^u a_t^{(2)}$, and $p_t^{(11)} = \frac{a_t^{(2)} C_t^u}{2h_t}$ with $\dot{p}_t^{(11)} \leqslant 0$. After integrating, we have $a_t^{(2)} \leqslant a_0^{(2)} u_t$. When saturating this inequality, we get the Lyapunov function from the Theorem $\mathcal{V}(X_t, \bar{X}_t^u, t) = \frac{u_t}{C_t^u}(f(\bar{X}_t^u) - f_\star) + \frac{u_t}{2h_t}\|X_t - x_\star\|^2$ with $\left(\dot{\frac{u_t}{2h_t}}\right) \leqslant 0$. When considering interpolation inequalities under strong convexity of f, the condition is $\left(\dot{\frac{u_t}{2h_t}}\right) \leqslant 2\mu u_t$. ∎

B.3. Proof for Theorem 4.1.

Proof. We follow the same reasoning as in Theorem 2.3 and Theorem 3.4, the worst-case guarantee can be formulated as the maximization problem:

$$\begin{aligned} 0 \geqslant \max_{f \in \mathcal{F}_{0,\infty}, d \in \mathbf{N}, X_t \in \mathrm{R}^d} & \ \frac{d}{dt} \mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}(X_t, \bar{X}_t, t), \\ & \text{subject to } d^2 X_t + \beta_t dX_t + \nabla f(X_t) dt + \sqrt{\gamma \Sigma} dB_t = 0, \\ & d\bar{X}_t = \frac{X_t - \bar{X}_t}{t} dt, \end{aligned}$$

where $\mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}(X_t, \bar{X}_t, t) = \begin{pmatrix} a_t^{(1)} \\ a_t^{(2)} \end{pmatrix}^\top \begin{pmatrix} f(X_t) - f_\star \\ f(\bar{X}_t) - f_\star \end{pmatrix} + \begin{pmatrix} \dot{X}_t \\ X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}^\top P_t \begin{pmatrix} \dot{X}_t \\ X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}$ with $P_t \succeq 0$, with differentiable coefficients, and $a_t^{(1)}, a_t^{(2)}$ are nonnegative differentiable functions. Introducing the Gram matrix $G = P^\top P$ with $P = [\dot{X}_t, \ X_t -$

$X_\star,\ \bar{X}_t - x_\star,\ \nabla f(X_t),\ \nabla f(\bar{X}_t)]$, the LMI equivalent reformulation is given by:

$$\begin{pmatrix} \dot{p}_t^{(11)} & \dot{p}_t^{(12)} - \beta_t p_t^{(12)} & \dot{p}_t^{(13)} - \frac{p_t^{(13)}}{t} - & -p_t^{(11)} & 0 \\ -2\beta_t p_t^{(11)} & + \frac{p_t^{(13)}}{t} + p_t^{(22)} & p_t^{(12)}\beta_t + p_t^{(23)} & & \\ * & \dot{p}_t^{(22)} + 2\frac{p_t^{(22)}}{t} & p_t^{(23)} - 2\frac{2p_t^{(22)}}{t} & a_t^{(1)} - p_t^{(12)} & \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} \\ & & + \frac{p_t^{(33)}}{t} & + \frac{\lambda_t^{(4)} + \lambda_t^{(6)}}{2} & \\ * & * & \dot{p}_t^{(33)} - \frac{2p_t^{(33)}}{t} & -\frac{\lambda_t^{(6)}}{2} - p_t^{(13)} & \frac{\lambda_t^{(5)} + \lambda_t^{(3)} - a_t^{(2)}/t}{2} \\ * & * & * & 0 & 0 \\ * & * & * & * & 0 \end{pmatrix} \preceq 0,$$

$$\lambda_t^{(6)} + \lambda_t^{(4)} = \lambda_t^{(5)} + \lambda_t^{(1)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(2)} + \lambda_t^{(6)} + \dot{a}_t^{(2)} = \lambda_t^{(3)} + \lambda_t^{(5)},$$
$$\lambda_t^{(1)} + \lambda_t^{(2)} + \dot{a}_t^{(2)} = \lambda_t^{(4)} + \lambda_t^{(5)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(i)} \geqslant 0,\ i \in \{1, ..., 6\}.$$

Thus, $p_t^{(11)} = 0$, and because P_t is positive semidefinite, $p_t^{(12)} = p_t^{(13)} = 0$. Then, $a_t^{(1)} = 0$. If in addition $a_t^{(2)} = 0$, the unique feasible Lyapunov function is $\mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t} = 0$. Otherwise, the LMI can be simplified to the LMI of Theorem 2.10. ∎

Appendix C. Primal averaging. In this appendix, we analyze primal averaging, that is to compare with Polyak-Ruppert averaging. Both continuous-time models happens to have very similar behavior.

C.1. First-order gradient flow. Let us consider *primal averaging*, introduced by Tao et al. [40], that led to a simpler analysis of averaging in SGD [41, Theorem 7]

$$\begin{aligned} dX_t &= -h_t \nabla f(\bar{X}_t)dt + h_t(\gamma \Sigma(X_t))^{1/2} dB_t, \\ d\bar{X}_t &= \frac{X_t - \bar{X}_t}{t} dt. \end{aligned} \tag{C.1}$$

Unlike Polyak-Ruppert averaging, the gradient is evaluated on the averaged sequence $\bar{X}_t$, that drives the dynamics on X_t. The family of Lyapunov functions $\mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}$ causes a noise $\frac{1}{2}\mathrm{Tr}((a_t^{(1)} \nabla_{xx} f(\bar{X}_t) + 2p_t^{(11)} I_d)\Sigma) h_t^2 \gamma$. As for Polyak-Ruppert averaging, we consider the family of Lyapunov functions $\mathcal{V}_{0, a_t^{(2)}, P_t}$. It is possible to obtain an LMI condition on the Lyapunov, as done in Theorem 3.4.

THEOREM C.1. *Let $f \in \mathcal{F}_{0,\infty}$ be twice differentiable functions, and SDEs (C.1) starting from $x_0 \in \mathbf{R}^d$, where $d \in \mathbf{N}$ is the dimension and originating from convex functions f. Assuming $\dot{a}_t^{(2)} \leqslant \frac{a_t^{(2)}}{t}$ and $t \to \frac{a_t^{(2)}}{t h_t}$ is a non-increasing function, the following Lyapunov function*

$$\mathcal{V}(X_t, t) = a_t^{(2)}(f(\bar{X}_t) - f_\star) + \frac{a_t^{(2)}}{2th_t}\|X_t - X_\star\|^2,$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \frac{\gamma}{2} Tr(\Sigma(X_t)) h_t \frac{a_t^{(2)}}{t}$ for all twice-differentiable functions $f \in \mathcal{F}_{0,\infty}$, for all trajectories $(X_t, \bar{X}_t)$ generated by the SDEs (C.1), and all dimensions $d \in \mathbf{N}$.

Then, it holds that $\mathbf{E}[f(\bar{X}_t) - f_\star] \leqslant \dfrac{\|x_0 - x_\star\|^2}{2a_t^{(2)}h_0} + \dfrac{\gamma}{2a_t^{(2)}} \displaystyle\int_0^t Tr(\Sigma(X_s))h_s \frac{a_s^{(2)}}{s} ds.$

Proof. This results is obtained following the same reasoning as for Polyak-Ruppert averaging in Theorem 2.3. The equivalent LMI to inequality

$$\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \mathrm{Tr}(\Sigma)\gamma h_t,$$

for any twice-differentiable function $f \in \mathcal{F}_{0,\infty}$ and any trajectory $(X_t, \bar{X}_t)$ generated by the SDE (C.1), is given by

$$\begin{pmatrix} \dot{p}_t^{(11)} + \frac{2p_t^{(12)}}{t} & \dot{p}_t^{(12)} - \frac{p_t^{(12)}}{t} + \frac{p_t^{(22)}}{t} & \frac{\lambda_t^{(6)}+\lambda_t^{(4)}}{2} & \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} - h_t p_t^{(11)} \\ * & \dot{p}_t^{(22)} - \frac{2p_t^{(22)}}{t} & -\frac{\lambda_t^{(6)}}{2} & \frac{\lambda_t^{(3)}+\lambda_t^{(5)}}{2} - \frac{a_t^{(2)}}{2t} - h_t p_t^{(12)} \\ * & * & -a_t^{(11)} & 0 \\ * & * & * & 0 \end{pmatrix} \preceq 0,$$

$$\lambda_t^{(6)} + \lambda_t^{(4)} = \lambda_t^{(5)} + \lambda_t^{(1)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(2)} + \lambda_t^{(6)} + \dot{a}_t^{(2)} = \lambda_t^{(3)} + \lambda_t^{(5)},$$
$$\lambda_t^{(1)} + \lambda_t^{(2)} + \dot{a}_t^{(2)} = \lambda_t^{(4)} + \lambda_t^{(5)} + \dot{a}_t^{(1)},$$
$$\lambda_t^{(i)} \geqslant 0, \ i \in \{1, ..., 6\}.$$

The Lyapunov function defined by $\mathcal{V}(X_t, \bar{X}_t, t) = a_t^{(2)}(f(\bar{X}t) - f_\star) + p_t^{(11)}\|X_t - x_\star\|^2$ verifies the following LMI

$$\begin{pmatrix} \dot{p}_t^{(11)} + \frac{2p_t^{(12)}}{t} & \dot{p}_t^{(12)} - \frac{p_t^{(12)}}{t} + \frac{p_t^{(22)}}{t} & \frac{\lambda_t^{(6)}+\lambda_t^{(4)}}{2} & \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} - h_t p_t^{(11)} \\ * & \dot{p}_t^{(22)} - \frac{2p_t^{(22)}}{t} & -\frac{\lambda_t^{(6)}}{2} & \frac{\lambda_t^{(3)}+\lambda_t^{(5)}}{2} - \frac{a_t^{(2)}}{2t} - h_t p_t^{(12)} \\ * & * & 0 & 0 \\ * & * & * & 0 \end{pmatrix} \preceq 0,$$

$$\lambda_t^{(6)} + \lambda_t^{(4)} = \lambda_t^{(5)} + \lambda_t^{(1)},$$
$$\lambda_t^{(2)} + \lambda_t^{(6)} + \dot{a}_t^{(2)} = \lambda_t^{(3)} + \lambda_t^{(5)},$$
$$\lambda_t^{(i)} \geqslant 0, \ i \in \{1, ..., 6\}.$$

Thus, $\lambda_t^{(6)} = \lambda_t^{(4)} = \lambda_t^{(5)} = \lambda_t^{(1)} = 0$, and

$$\dot{a}_t^{(2)} \leqslant \lambda_t^{(2)} + \dot{a}_t^{(2)} = \lambda_t^{(3)} = \frac{a_t^{(2)}}{2t},$$
$$\dot{p}_t^{(11)} = \left(\frac{\dot{a}_t^{(2)}}{2th_t}\right) \leqslant 0.$$

□

Primal averaging involves at most two interpolation inequalities whereas Polyak-Ruppert averaging may involve up to four inequalities to obtain the same convergence guarantee. Compared with Polyak- Ruppert averaging, primal averaging has similar convergence guarantees but requires less interpolation inequalities.

Remark C.2. In discrete time, convergence of primal averaging is often proven using its connection to the heavy ball method [12, 14]. In continuous time, primal

averaging also allows to describe the dynamics of the averaged sequence by a second-order SDE $\ddot{\bar{X}}_t + \frac{2}{t}\dot{\bar{X}}_t + \frac{1}{t}\nabla f(\bar{X}_t) + \frac{\sqrt{\gamma\Sigma(X_t)}}{t}\frac{dB_t}{dt} = 0$. The convergence bound above can be obtained directly from this SDE. This analysis allows to understand the relationship between averaging and accelerated gradient flows.

Under primal averaging, the convergence in function values is exactly the same as under Polyak-Ruppert averaging, but involves only one interpolation inequality. When considering optimization methods, [41, Theorem 7] also pointed out that primal averaging leads to shorter proofs.

C.2. Second order-gradient flow. We have seen primal averaging for stochastic first-order gradient flows (3.3) leads to a second-order SDE. For the accelerated gradient flow under primal averaging, it is possible to obtain a third order SDE,

$$d^3\bar{X}_t + (\frac{3}{t} + \beta_t)d^2\bar{X}_t + \frac{\beta_t}{t}d\bar{X}_t + \frac{1}{t}\nabla f(\bar{X}_t)dt + \sqrt{\frac{\gamma\Sigma(\bar{X}_t)}{t^2}}dB_t = 0.$$

A natural quadratic Lyapunov is given by $\mathcal{V}_{a_t,P_t}(X_t) = a_t(f(X_t) - f_\star) + U_t^\top P_t U_t$, where $U_t = \begin{pmatrix} X_t - X_\star, & \dot{X}_t, & \ddot{X}_t \end{pmatrix}^\top$, and a_t is a nonnegative differentiable function, and $P_t \succeq 0$ has differentiable parameters.

THEOREM C.3. *Let $f \in \mathcal{F}_{0,\infty}$ be twice differentiable functions, $\beta_t \geqslant 0$ be differentiable functions, and consider the following ODE starting from $x_0 \in \mathbf{R}^d$ where $d \in \mathbf{N}$ is the dimension,*

$$d^3X_t + \alpha_t d^2X_t + \beta_t dX_t + \gamma_t\nabla f(X_t)dt = 0.$$

If a quadratic Lyapunov function $\mathcal{V}_{a_t,P_t}$ as defined above verifies $\frac{d}{dt}\mathcal{V}_{a_t,P_t}(X_t,t) \leqslant 0$ for all twice-differentiable functions $f \in \mathcal{F}_{0,\infty}$, all solutions X_t to ODEs defined above, and all dimensions $d \in \mathbf{N}$, then $\mathcal{V}_{a_t,P_t} = 0$.

Proof. Following the reasoning from Appendix A.1, the problem of finding a suitable Lyapunov function can be formulated as a maximization problem

$$\begin{aligned} 0 \geqslant \max_{f\in\mathcal{F}_{0,\infty}, d\in\mathbf{N}, X_t\in\mathbf{R}^d} \ & \frac{d}{dt}\mathcal{V}_{a_t,P_t}, \\ \text{subject to } \ & d^3X_t + \alpha_t d^2X_t + \beta_t dX_t + \gamma_t\nabla f(X_t)dt = 0, \end{aligned}$$

has the following LMI equivalence,

$$\begin{pmatrix} \dot{p}_t^{(11)} - 2\alpha_t p_t^{(11)} & \dot{p}_t^{(12)} - \beta_t p_t^{(11)} - \alpha_t p_t^{(12)} & \dot{p}_t^{(13)} - \alpha_t p_t^{(13)} & -h_t p_t^{(11)} \\ +2p_t^{(12)} & +p_t^{(22)} + p_t^{(13)} & +p_t^{(23)} & \\ * & \dot{p}_t^{(22)} - 2\beta_t p_t^{(12)} + 2p_t^{(23)} & \dot{p}_t^{(23)} - \beta_t p_t^{(13)} & \frac{a_t}{2} - h_t p_t^{(12)} \\ & & +p_t^{(33)} & \\ * & * & \dot{p}_t^{(33)} & \frac{\lambda_t^{(1)}}{2} - h_t p_t^{(13)} \\ * & * & * & 0 \end{pmatrix} \preceq 0,$$

$$\dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

It follows directly from the LMI that $p_t^{(11)} = 0$. Since $P_t \succeq 0$, $p_t^{(11)} = p_t^{(12)} = p_t^{(13)} = 0$. Thus $\lambda_t^{(1)} = 0 = \lambda_t^{(2)}$ and $a_t = 0$, and the Lyapunov $\mathcal{V}_{a_t,P_t} = 0$ is the unique solution to this LMI. ∎

Theorem C.3 concludes in an other way that it is not possible to obtain a convergence guarantee on an averaged sequence of an accelerated method. However, we can wonder if there exists other transformations that ensures convergence of both the averaged sequence and the noise.