SIAM J. OPTIM.
Vol. 33, No. 3, pp. 1558–1586

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 finding and verifying Lyapunov functions for classes of ordinary and stochastic differential equations. More precisely, we extend the performance estimation framework, originally proposed by Drori and Teboulle [*Math. Program.*, 145 (2014), pp. 451–482], to continuous-time models. We retrieve convergence results comparable to those of discrete-time methods using fewer assumptions and inequalities and provide new results for a family of stochastic accelerated gradient flows.

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

MSC codes. 90C25, 90C30, 68Q25, 90C22

DOI. 10.1137/22M1498486

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. We consider optimization problems of the form

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

where f is convex and differentiable. First-order methods (that gather information about f by evaluating its gradient at past iterates) are very popular to solve these problems, due to their attractive low cost per iteration, and due to the fact that data science applications typically do not require very accurate solutions [9]. Gradient descent is a common first-order method, which starts from a point $x_0 \in \mathbf{R}^d$ and whose 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. Gradient descent with small step sizes is directly related to the so-called gradient flow:

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

with the notation $\dot{X}_t \triangleq \frac{d}{dt}X_t$ and where the solution X_t of the ordinary differential equation (ODE) verifies $X_{t_k} \approx x_k$ with the identification $t_k = \gamma k$. In numerical

* Received by the editors May 25, 2022; accepted for publication (in revised form) January 3, 2023; published electronically July 26, 2023.
https://doi.org/10.1137/22M1498486
Funding: 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).

† 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).

integration, gradient descent (1.2) is also known as the explicit Euler scheme for integrating gradient flows. Recently, Su, Boyd, and Candès [42] have interpreted Nesterov's accelerated gradient [29] in a similar fashion through its continuous-time version, paving the way to several continuous-time analyses of accelerated methods [39, 52, 51].

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 for some stochastic modeling. When ξ is drawn uniformly from a finite set of possible samples $(\xi_1,\dots,\xi_n)$, 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 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, and ξ_{i_k} is drawn uniformly at random in $(\xi_1,\dots,\xi_n)$. Thereby $\nabla\tilde{f}(x_k,\xi_{i_k})$ is an unbiased estimate of the full gradient: $\mathbf{E}_{i_k}[\nabla\tilde{f}(x_k,\xi_{i_k})] = \nabla f(x_k)$. Li, Tai, and E [25, 26] derived stochastic differential equations (SDEs) approximating SGD:

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

where $\sigma(X_t)$ is a noise parameter connected to parameters of the method, and these were further developed by Shi et al. [41, 53]. Relying on approximate theorems between SDEs and original stochastic gradient algorithms, SDEs have thus become a tool for analyzing convergence speeds of discrete-time methods. Usually, gradient flows (resp., first-order methods) are studied via worst-case convergence properties, which hold for any function of a given class, and any trajectory generated by the ODE (resp., optimization method). In many cases, continuous-time approaches seem to allow for shorter, simpler, and thereby more intuitive proofs. They also bring insights 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 proofs 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) originally coined by Drori and Teboulle [14] 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 analyzing convergence of continuous-time models, by constructing Lyapunov functions suited to a gradient-based ODE in a systematic way, using small-sized SDP reformulations. Furthermore, this procedure benefits from tightness properties, meaning that the feasibility of the SDP allows us to conclude that there exists a Lyapunov function within the prescribed family of Lyapunov functions under consideration. Reciprocally, infeasibility of the SDP allows us to conclude that there exists no such valid Lyapunov function within the family.

1.1. Lyapunov functions. Lyapunov functions are a standard tool for dealing with convergence properties of gradient flows. Such functions are also more generally used for studying stability properties of dynamical systems [21]. More precisely, consider a differentiable function f within a class $\mathcal{F}$ and an ODE aiming at minimizing f. For such a continuous dynamical system with a stationary point $x_\star \in \operatorname{argmin}_x f(x)$, we call $\mathcal{V} : \mathbf{R}^d \times \mathbf{R}^+ \to \mathbf{R}$ a Lyapunov function if it is differentiable and satisfies the following conditions for all trajectories X_t generated by the ODE:

- $\mathcal{V}(x,t) = 0 \iff x = x_\star$,
- $\mathcal{V}(X_t,t) \geqslant 0$,
- $\frac{d}{dt}\mathcal{V}(X_t,t) \leqslant 0$

for all $t \geqslant 0$.

Lyapunov functions are suited for deriving both linear (or exponential) and sublinear convergence rates. When looking for linear convergence rates (as we may expect for strongly convex functions), we typically use $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -\tau\mathcal{V}(X_t)$ instead of the third condition, where τ depends on the class of functions and on the ODE (for more details see Remark 2.3). In this ad hoc definition, we enforced nonnegativity along the trajectory X_t, but definitions often require nonnegativity on $\mathbf{R}^d$ [47]. There exist similar definitions of Lyapunov functions for discrete-time optimization methods [47, 37, 24].

With this approach, convergence guarantees highly depend on how rich the family of Lyapunov functions under consideration is. In this work, we use a family of quadratic Lyapunov functions that is popular and natural for studying both discrete-time [29, 15] and continuous-time [42] optimization schemes.

1.2. Prior works. Lyapunov functions are common for analyzing continuous-time and discrete-time models in convex optimization. For example, convergence proofs for Nesterov's accelerated gradient method typically rely on such Lyapunov analyses [29], [15, Theorem 4.8]. In the recent paper [4], 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 [42] and its high-resolution ODEs for strongly convex functions proposed in [39], or accelerated mirror descent, whose continuous-time dynamics were analyzed in [22].

Different techniques were developed to compute suitable Lyapunov functions. The authors of [51] put forward an approach based on the Bregman Lagrangian for accelerated methods in potentially non-Euclidean settings; this was further developed in [52]. The authors of [13] directly derived Lyapunov functions from Hamiltonian equations describing dynamics of ODEs. Using similar conservation laws in a dilated coordinate system, [43] 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 combinations of inequalities formulated as semidefinite programs was pioneered by Drori and Teboulle [14] and led to the notion of performance estimation problems. Their work was followed up in [50, 49] to provide worst-case bounds in a principled way, and was extended to the Lyapunov framework [47]. A competing strategy inspired by control theory was developed by [24, 19], 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 [16]. Connections between Lyapunov functions obtained via the IQC framework in continuous time and discrete time were highlighted by [37].

For stochastic differential equations (SDEs), convergence proofs can also be obtained through the Lyapunov approach, together with Ito's calculus. For some well-chosen Lyapunov functions, [30] analyzed SGD, SAGA [12], and SVRG [20]. The authors of [53, 54] extended the framework of Bregman Lagrangian Lyapunov functions [51] to the stochastic setting. To the best of our knowledge, a systematic way of verifying a Lyapunov function for SDEs has not been developed yet.

1.3. Contributions and organization. In this work, we are concerned with worst-case convergence analyses of ODEs and SDEs for modeling (stochastic) gradient based 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, which originate from a (possibly strongly) convex function. In short, we find Lyapunov functions as feasible points of certain linear matrix inequalities (LMIs). All codes for reproducing numerical results can be found at https://github.com/CMoucer/PEP_ODEs. Following this work, we have also added continuous-time models to the Python package PEPit [18]. After that, building on the first part of this work for gradient-based ODEs, we analyze continuous-time versions of stochastic optimization algorithms.

Section 3 studies properties of trajectories generated by SDEs that approximate 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 nonuniform version of averaging, allows us to reach an optimal trade-off between the two terms. Our results match those obtained for the stochastic gradient method, but with more compact analyses than those of the discrete-time setting.

In section 4, we prove that stochastic accelerated gradient flows require diminishing step sizes to converge in our setting. In contrast to first-order stochastic gradient flows, averaging does not preserve convergence for SDEs approximating accelerated gradient methods with constant 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.

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 $\lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y)$. Such functions (i.e., with full domain, $\mathrm{dom}(f) = \mathbf{R}^d$) are convex closed proper (CCP) (i.e., their epigraphs are nonempty closed convex sets). For simplicity, we assume in addition differentiability of f. Such a differentiable function f is convex if and only if for all $x, y \in \mathbf{R}^d$, $f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle$. A differentiable function f is L-smooth if its gradient is L-Lipschitz, that is if for any $x, y \ \in \mathbf{R}^d$,

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

Smoothness is a common assumption for analyzing optimization methods, which limits the growth rate of the function. A convex 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 convex functions from $\mathbf{R}^d$ to $\mathbf{R}$, with $0 \leqslant \mu \leqslant L \leqslant +\infty$. Weaker assumptions than strong convexity are also encountered in the literature for analyzing gradient algorithms and lead to similar convergence guarantees. Among them, Bolte et al. [7, Appendix 5.2] proved linear convergence of gradient descent under the Łojasiewicz inequality, first introduced by Łojasiewicz [27]. Other relaxed versions of strong convexity followed [28].

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. In this context, we show that verifying such a Lyapunov function can be formulated as verifying the feasibility of an LMI. This framework allows us to search for Lyapunov functions and to derive convergence bounds for nonautonomous gradient flows.

2.1. The gradient flow. Let us consider the gradient flow

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

Let $x_\star$ be a global minimizer of f. Without further assumptions, the function f is decreasing along the trajectory X_t solution to the gradient flow. The Lyapunov function $\mathcal{V}(X_t) = f(X_t) - f(x_\star)$ is indeed nonnegative and equal to zero at $x_\star$ and has a nonpositive derivative with respect to time $\frac{d}{dt}\mathcal{V}(X_t) = \dot{X}_t^\top \nabla f(X_t) = -\|\nabla f(X_t)\|^2$.

In the next section, we show how to obtain and verify such Lyapunov functions, and their corresponding convergence rates in the case of gradient flow originating from a strongly convex function.

2.1.1. Minimizing strongly convex functions. Let f be μ-strongly convex (i.e., $f \in \mathcal{F}_{\mu,\infty}$) with $\mu > 0$, and let $x_\star$ be its unique minimizer such that $f(x_\star) = f_\star$. In this context, it is possible to prove linear convergence of the gradient flow to its stationary point. Scieur et al. proved in [38, Proposition 1.1] the following convergence bound in function values for the gradient flow:

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

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

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 $a, c \geqslant 0$ are constants that do not depend on t, such that the function $\mathcal{V}_{a,c}$ is nonnegative and nonincreasing along the flow. Quadratic Lyapunov functions are common for proving convergence of gradient flows and cover for instance the Lyapunov function 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 largest nonnegative value $\tau_\star$ such that the condition

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

holds for any dimension $d \in \mathbf{N}$, any function $f \in \mathcal{F}_{\mu,\infty}$, and any trajectory X_t generated by the gradient flow. After integration, (2.3) allows us to obtain a convergence guarantee of the form $\mathcal{V}_{a,c}(X_t) \leqslant e^{-\tau_\star t}\mathcal{V}_{a,c}(X_0)$. Given a certain Lyapunov function $\mathcal{V}_{a,c}$ and a time t, we get that the largest acceptable $\tau_\star$ is a solution to

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

This minimization problem is invariant with respect to t. It is established in [50, 14] that these so-called performance estimation problems (PEPs) can be formulated as SDPs. Because of the variable $f \in \mathcal{F}_{\mu,\infty}$, the maximization problem (2.4) is infinite-dimensional. Recall that a differentiable function $f \in \mathcal{F}_{\mu,\infty}$ verifies for all points $x, y \in \mathbf{R}^d$, $f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geqslant \frac{\mu}{2}\|x - y\|^2$. Using 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

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

The fact that (2.5) produces an upper bound on $-\tau_\star$ directly follows from the fact that any sampled strongly convex function satisfies these inequalities at the sampled points (here X_t and $x_\star$). Thereby, any feasible point to (2.4) corresponds to a feasible point for (2.5) with the same objective value. In the other direction, [50, Corollary 2] (which provides a constructive way to obtain some $f \in \mathcal{F}_{\mu,\infty}$ that interpolates the triplets $(X_i, g_i, f_i)_{i=t,\star}$) ensures that any feasible point to (2.5) can be translated to a feasible point to (2.4) with the same objective value, thereby reaching the equivalence between formulations (2.4) and (2.5).

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

$$\begin{aligned} -\tau_\star = \max_{G\succcurlyeq 0, F\in\mathbf{R}^2} \ & b_0^\top F + \operatorname{Tr}(A_0 G), \\ \text{subject to } & b_1^\top F + \operatorname{Tr}(A_1 G) \geqslant 0, \\ & b_2^\top F + \operatorname{Tr}(A_2 G) \geqslant 0, \\ & b_3^\top F + \operatorname{Tr}(A_3 G) = 1. \end{aligned} \tag{2.6}$$

where $A_0 = \left(\begin{smallmatrix} 0 & -c \\ -c & -a \end{smallmatrix}\right)$, $A_1 = \left(\begin{smallmatrix} -\mu/2 & 1/2 \\ 1/2 & 0 \end{smallmatrix}\right)$, $A_2 = \left(\begin{smallmatrix} -\mu/2 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $A_3 = \left(\begin{smallmatrix} c & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $b_0 = [0, 0]^\top$, $b_1 = [-1,\ 1]^\top$, $b_2 = [1,\ -1]^\top$, and $b_3 = a[1,\ -1]^\top$. Consider the corresponding Lagrangian, for $F \in \mathbf{R}^2$, $G \succcurlyeq 0$, $\tau \in \mathbf{R}$, and $\lambda_1, \lambda_2 \geqslant 0$:

$$\begin{aligned} \mathcal{L}(F, G, \tau, \lambda_1, \lambda_2) = {} & b_0^\top F + \operatorname{Tr}(A_0 G) + \tau \cdot (b_3^\top F + \operatorname{Tr}(A_3 G) - 1) \\ & + \lambda_1 \cdot (b_1^\top F + \operatorname{Tr}(A_1 G)) + \lambda_2 \cdot (b_2^\top F + \operatorname{Tr}(A_2 G)). \end{aligned}$$

The saddle point of the Lagrangian is given by

$$\tau_\star = \min_{\tau, \lambda_1 \geqslant 0, \lambda_2 \geqslant 0} \max_{F, G \succcurlyeq 0} \mathcal{L}(F, G, \tau, \lambda_1, \lambda_2).$$

The Lagrangian dual of the SDP (2.6) is obtained by maximizing over $F \in \mathbf{R}^2$, $G \succcurlyeq 0$:

$$\begin{aligned} -\tau_\star = \min_{\tau,\lambda_1,\lambda_2} \; & -\tau, \\ \text{subject to } & S = A_0 + \lambda_1 A_1 + \lambda_2 A_2 + \tau A_3 \preccurlyeq 0, \\ & b_0 + \lambda_1 b_1 + \lambda_2 b_2 + \tau b_3 = 0, \\ & \tau \in \mathbf{R},\ \lambda_1, \lambda_2 \geqslant 0. \end{aligned}$$

The equality with $\tau_\star$ comes from strong duality, which holds via Slater's conditions (it is relatively easy to show that there exists a Slater point using the same construction as in [50, Theorem 6]). Finally, since any feasible τ is a lower bound for $\tau_\star$, those developments allow arriving to the equivalence between verifying a quadratic Lyapunov function and verifying the feasibility of an LMI.

THEOREM 2.1. *Let $a, c, \tau \geqslant 0$ and $\mu > 0$. 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 dimensions $d \in \mathbf{N}$, for all $f \in \mathcal{F}_{\mu,\infty}$, and all trajectory X_t solutions to the gradient flow* (1.3)*, where $\mathcal{V}_{a,c}$ is a quadratic Lyapunov function* (2.2)*.*
- *There exist $\lambda_1, \lambda_2 \geqslant 0$ such that*

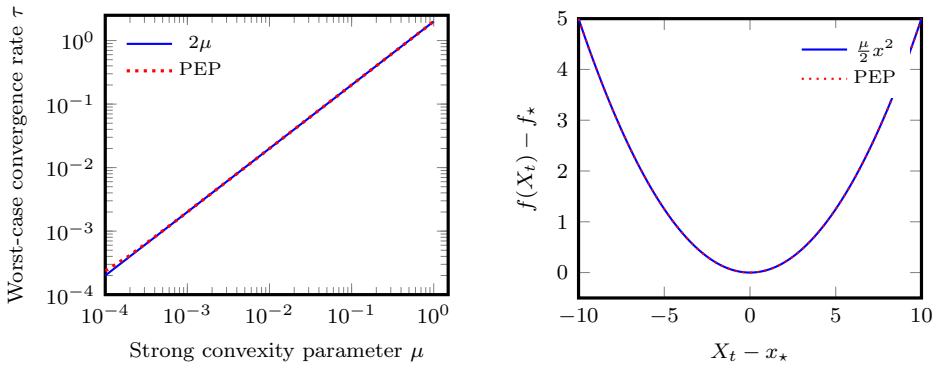
$$S = \begin{pmatrix} \tau c - \frac{\mu}{2}(\lambda_1 + \lambda_2) & -c + \frac{\lambda_1}{2} \\ -c + \frac{\lambda_1}{2} & -a \end{pmatrix} \preccurlyeq 0,\ \tau a = \lambda_1 - \lambda_2. \tag{2.7}$$

Remark 2.2. As a corollary of our result and for the class of quadratic Lyapunov functions (2.2) (and later (2.8), (2.11)), it turns out that only two interpolation inequalities in $(X_t, x_\star)$ and $(x_\star, X_t)$ are involved in convergence proofs for continuous-time models (see Theorem 2.7 for second-order gradient flows). In other words, the framework reveals shorter proofs in continuous time.

A few conclusions can be drawn from the LMI equivalence from Theorem 2.1. For fixed value of a, c, τ, the LMI 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 in the class $\mathcal{F}_{\mu,\infty}$. Second, we can simultaneously optimize over the class of quadratic Lyapunov functions and over the convergence rate. Indeed, given a rate τ, the LMI is jointly convex in $\lambda_1, \lambda_2, a, c$. Therefore, thanks to linearity of the feasibility problem in τ, a bisection search allows us to optimize over it and to find the worst-case guarantee $\tau_\star$.

In Figure 1a, we obtain the fastest linear convergence rate that can be achieved using quadratic Lyapunov functions (2.2) (and even for all Lyapunov functions, since the rate is tight on a function $f(x) = \frac{\mu}{2}\|x\|_2^2$ for any time t). Together with Theorem 2.1 we retrieve the known linear worst-case convergence speed in $e^{-2\mu}$ from Scieur et al. [38, Proposition 1.1] without improvement. The numerical approach ensures tightness of the procedure by construction, as it guarantees the existence of a numerical function f that exactly achieves this convergence guarantee (see Figure 1b and the method in [48, Chapter 3]). The next section builds on the same technique to analyze the gradient flow originating from a (possibly nonstrongly) convex function. In this scenario, the difficulty comes from the time-dependence of Lyapunov functions.

2.1.2. Minimizing convex functions. Let $f \in \mathcal{F}_{0,\infty}$ and $x_\star$ be a minimizer of f. In this case, worst-case convergence rates are often sublinear. Again, as in discrete time, it is possible to obtain convergence guarantees using time-dependent quadratic Lyapunov functions.



(a) Worst-case rate $\tau_\star$ for the class of quadratic Lyapunov functions (2.2)

(b) Reconstruction of a function $f \in \mathcal{F}_{\mu,\infty}$ that interpolates $x_\star$ and X_t, while matching the convergence rate $\tau_\star = 2\mu$, 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 [*38*, Proposition* 1.1*], for trajectories X_t generated by gradient flow* (1.3) *originating from a μ-strongly convex function.*

The Lyapunov function $\mathcal{V}(X_t,t) = t(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ from [42, p. 7] verifies $\frac{d}{dt}\mathcal{V}(X_t,t) \leqslant 0$ for any dimension $d \in \mathbf{N}$, any function $f \in \mathcal{F}_{0,\infty}$, and any trajectory 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 integration between 0 and t, we recover a convergence bound in function values from the literature [42, p. 7], [16, section 6.3.1]:

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

Let us adapt the techniques from section 2.1.1 by considering 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 $a_t, c_t \geqslant 0$ are functions differentiable with respect to time such that the function $\mathcal{V}_{a_t,c_t}$ is nonnegative and nonincreasing along the flow. When a quadratic Lyapunov function decreases along the trajectory X_t, that is, $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t) \leqslant 0$, a convergence guarantee in function values is given by

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

Verifying a quadratic Lyapunov function can be cast as verifying that the following maximization problem is nonpositive:

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

Remark 2.3. 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 2.4. *Let $a_t, c_t \geqslant 0$ be continuously differentiable with respect to time. The following assertions are equivalent:*

- *The inequality $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t,t) \leqslant 0$ is satisfied for all dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{0,\infty}$, and all trajectories X_t generated by the gradient flow* (1.3)*, where $\mathcal{V}_{a_t,c_t}$ is a quadratic Lyapunov function defined in* (2.8)*.*
- *There exist $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that*

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

Proof. The LMI is obtained following the previous methodology (see Appendix A.1). ∎

Choosing $\lambda_t^{(1)} = 1$, $\lambda_t^{(2)} = 0$, together with $c_t = \frac{1}{2}$ and $a_t = t$, the conditions in the LMI from Theorem 2.4 are satisfied. We retrieve the Lyapunov function $\mathcal{V}(x,t) = t(f(x) - f_\star) + \frac{1}{2}\|x - x_\star\|^2$ from [42, p. 7].

Similar to Theorem 2.1, the LMI from Theorem 2.4, and hence the problem of looking for a Lyapunov function, is jointly convex in $\lambda_t^{(1)}$, $\lambda_t^{(2)}$, c_t, a_t, $\dot{a}_t$, $\dot{c}_t$. This Lyapunov analysis can also be validated numerically, as for the gradient flow originating from strongly convex functions.

2.2. Accelerated gradient flows. A major improvement to gradient descent dates back to Nesterov [29], who introduced 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 $\gamma, \alpha_k \geqslant 0$ depend on the class of functions to minimize. The combination of past iterates allows more control over the accumulated error. The idea of incorporating momentum was first introduced by Polyak [32] with the heavy-ball method, starting from $x_0, x_1 \in \mathbf{R}^d$, and for a momentum $\alpha_k > 0$:

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

Yet, compared to Nesterov's accelerated gradient method, the heavy-ball method lacks global acceleration beyond quadratics.

When the step size γ goes to zero, these schemes happen to be closely related to second-order differential equations, where $\beta_t \geqslant 0$ is a continuous function depending 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 [39, 37, 52, 19]. 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 [53, 41, 52]. Taking integration theory's point of view, Scieur et al. [38] proved that these multistep methods may even be seen as discretization schemes of the gradient flow (for quadratics).

Again, ODEs and multistep 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 a quadratic

Lyapunov function for second-order gradient flows, taken in the class of quadratic functions

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

where $P = \left(\begin{smallmatrix} p_t^{(11)} & p_t^{(12)} \\ p_t^{(12)} & p_t^{(22)} \end{smallmatrix}\right), a_t \geqslant 0$ are continuously differentiable with respect to time and such that the Lyapunov function is nonnegative and nonincreasing along the flow. After integration of a quadratic Lyapunov function $\mathcal{V}_{a_t,c_t}$ between 0 and t, this approach leads to convergence bounds, for instance, in function values $f(X_t) - f_\star \leqslant \frac{\mathcal{V}(x_0)}{a_t}$.

2.2.1. Minimizing strongly convex functions. Let $f \in \mathcal{F}_{\mu,L}$, with strong convexity parameter $\mu > 0$, and let Nesterov's accelerated gradient method's parameters by defined by $\gamma = \frac{1}{L}$ and $\alpha = \frac{1-\sqrt{\mu\gamma}}{1+\sqrt{\mu\gamma}}$ in (2.9). When the step size γ goes to zero, the continuous-time limit of y_k in (2.9) is exactly the Polyak damped oscillator [32],

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

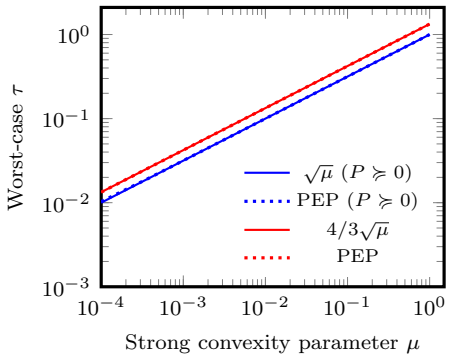
as it has already been highlighted in previous works [16, 37, 39]. The heavy-ball method (2.10) reaches the same continuous-time limit when reducing the step size. Shi et al. [39] derived a convergence guarantee in $f(X_t) - f_\star = \mathcal{O}(e^{-\frac{\sqrt{\mu}t}{4}})$ using a Lyapunov-based approach. This bound was improved to $f(X_t) - f_\star = \mathcal{O}(e^{-\sqrt{\mu}t})$ by Wilson, Recht, and Jordan [52, Appendix B] using the Bregman–Lagrangian approach, and by Sanz-Serna and Zygalakis using the IQC framework [37, 16]. Using the methodology from Theorem 2.1, we show that verifying linear convergence guarantees using quadratic Lyapunov functions with constant parameters (2.11) can be cast as an LMI.

THEOREM 2.5. *Let $\mu > 0$ and $\tau \geqslant 0$. Let $a \geqslant 0$, and let P be a symmetric matrix. The following assertions are equivalent:*

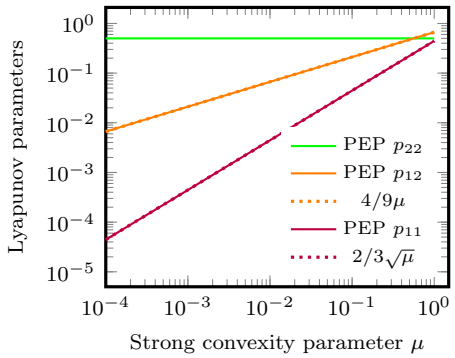
- *The inequality $\frac{d}{dt}\mathcal{V}_{a,P}(X_t) \leqslant -\tau\mathcal{V}_{a,P}(X_t)$, is satisfied for all dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{\mu,\infty}$, and all trajectories X_t generated by the Polyak damped oscillator* (2.12)*, where $\mathcal{V}_{a,P}$ is a quadratic Lyapunov function* (2.11)*.*
- *There exist $\lambda_1, \lambda_2, \nu_1, \nu_2 \geqslant 0$, such that*

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

Proof. The equivalence is obtained using the methodology developed in section 2.1.1 and introducing the Gram matrix $G = P^\top P$, where $P = (\dot{X}_t, X_t - x_\star, g_t)$, where g_t holds for $\nabla f(X_t)$. The first LMI refers to the nonincreasing condition for the Lyapunov function. The second one refers to the positivity constraint $\mathcal{V}_{a,P}(X_t) \geqslant 0$ for all dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{0,\infty}$, and all trajectories X_t generated by Polyak damped oscillator (2.12), as done for discrete-time methods in [47, Th. 7]. ∎



(a) Best guarantees found within the class of quadratic Lyapunov functions (2.11).



(b) Lyapunov parameters P in (2.11) for $\tau = 4/3\sqrt{\mu}$ and $a = 1$, as a function of the condition number μ.

FIG. 2. *Comparison between the worst-case guarantee obtained numerically with PEP, and its references, for the Polyak damped oscillator* (2.12) *originating from μ-strongly convex functions, and for quadratic Lyapunov functions* (2.11).

As for the gradient flow, the LMI is jointly convex in $\lambda_1, \lambda_2, \nu_1, \nu_2 \geqslant 0$, in the Lyapunov parameters a, P, and is linear in τ. Hence, we can perform a bisection search over τ to find the fastest linear convergence rate that can be verified using quadratic Lyapunov functions, as done in Figure 2a. The framework provides in addition a numerical tool for choosing Lyapunov parameters for which the worst-case linear convergence rate is achieved. Figure 2b helped in providing an intuition for parameters in Corollary 2.6.

COROLLARY 2.6. *Let $\mu \geqslant 0$. The function*

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

verifies $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -4/3\sqrt{\mu}\mathcal{V}(X_t)$ for all dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{\mu,\infty}$, and all trajectories X_t generated by the Polyak damped oscillator (2.12)*. A tight rate is achieved for $f(x) = \frac{1}{2}\mu x^2$.*

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 [47] in discrete time, where a stricter positivity condition on $P \succcurlyeq 0$ hindered proving tight convergence of Nesterov's accelerated gradient. Similarly in our context, the Lyapunov function from Corollary 2.12 is defined by $P = \left(\begin{smallmatrix} 4/9\mu & 2/3\sqrt{\mu} \\ 2/3\sqrt{\mu} & 1/2 \end{smallmatrix}\right)$, which is not positive semidefinite. Usually in the continuous-time models literature, we only consider matrices P that are positive semidefinite, such as in the Lyapunov function from [37, 39, Theorem 4.3],

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

which verifies $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -\sqrt{\mu}\mathcal{V}(X_t)$ for all dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{\mu,\infty}$ and all trajectories X_t generated by the Polyak damped oscillator. This Lyapunov is a feasible point of the LMI (2.13) from Theorem 2.5, with $\tau = \sqrt{\mu}$, $\lambda_1 = \sqrt{\mu}$, $\lambda_2 = 0$,

$\nu_1 = 0$, $\nu_2 = a = 1$. By relaxing the condition $P \succcurlyeq 0$, we thus improve results from Sanz-Serna and Zygalakis and Wilson et al. by a factor 4/3 in Corollary 2.6.

Figure 2a together with Corollary 2.6 allows us 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 the gradient flow, rates are sublinear when the accelerated gradient flow originates from convex functions. Let $f \in \mathcal{F}_{0,L}$, $\gamma \leqslant \frac{1}{L}$ be the step size and $\alpha_k = \frac{k-1}{k+2}$ be the scheme parameter in Nesterov's accelerated method. Su, Boyd, and Candès [42, section 2] proved the connection between the first-order scheme and a second-order ODE known as the accelerated gradient flow (AGF):

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

Su, Boyd, and Candès [42, Theorem 3] proved that the following inequality is verified for all functions $f \in \mathcal{F}_{0,\infty}$ and all trajectories X_t generated by the accelerated gradient flow (2.14):

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

Their proof exhibits a Lyapunov function $\mathcal{V}(X_t, t) = t^2(f(X_t) - f_\star) + 2\|(X_t - x_\star) + \frac{t}{2}\dot{X}_t\|^2$, which 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). In this context, we obtain again an LMI equivalence for verifying a quadratic Lyapunov function (2.11).

THEOREM 2.7. *Let $a_t \geqslant 0$, $P_t \succcurlyeq 0$ be functions continuously differentiable with respect to time. 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 dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{0,\infty}$, and all trajectories X_t generated by the accelerated gradient flow* (2.14)*, where $\mathcal{V}_{a_t,P_t}$ is a quadratic Lyapunov function* (2.11)*.*
- *There exist $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that*

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

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

Proof. The proof follows those from Theorems 2.5 and 2.1. ∎

The Lyapunov function $\mathcal{V}(X_t,t) = t^2(f(X_t)-f_\star) + 2\left\|(X_t - x_\star) + \frac{t}{2}\dot{X}\right\|^2$ exhibited by Su, Boyd, and Candès [42, Theorem 3] is a feasible point of the LMI, with $a_t = t^2$ and $P_t = 2\left(\begin{smallmatrix} 1 & t/2 \\ t/2 & t^2/4 \end{smallmatrix}\right)$ for the Lyapunov parameters, and $\lambda_t^{(1)} = t$, $\lambda_t^{(2)} = 0$. It is possible to retrieve these results numerically, as was done for the accelerated gradient flow originating from strongly convex functions.

2.3. Higher-order convergence and time dilation. In this section, we analyze convergence of the first and second-order nonautonomous gradient flows and provide convergence guarantees depending on their parametrization. It appears that higher-order convergence of nonautonomous gradient flows is highly connected to time dilation.

2.3.1. A nonautonomous first-order gradient flow. Let $f \in \mathcal{F}_{0,\infty}$. Let the nonautonomous first-order gradient flow be defined by

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

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 by 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 rescaling. 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 nonautonomous gradient flow. The following corollaries can be obtained by performing this change of variable in Theorem 2.4.

COROLLARY 2.8. *Let $\mu \geqslant 0$.*

- *If $\mu > 0$, the 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 dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{\mu,\infty}$, and all trajectories X_t generated by the nonautonomous gradient flow* (2.15). *A convergence guarantee is given by $f(X_t) - f_\star \leqslant e^{-2\mu \int_0^t \alpha_s ds}(f(x_0) - f_\star)$.*
- *If $\mu = 0$, the function $\mathcal{V}(X_t, t) = (\int_0^t \alpha_s ds)(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ verifies $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$ for all dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{0,\infty}$, and all trajectories X_t generated by the nonautonomous gradient flow* (2.15). *A convergence guarantee is given by $f(X_t) - f_\star \leqslant \frac{1}{2\int_0^t \alpha_s ds}\|x_0 - x_\star\|^2$.*

Remark 2.9. When $\alpha_t = 1$ above, that is, $\tau_t = t$, we recover exactly the Lyapunov functions $\mathcal{V}(X_t, t) = t(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ from Theorem 2.4 for convex functions, and $\mathcal{V}(X_t) = e^{2\mu t}(f(X_t) - f_\star)$ from Theorem 2.1 for strongly convex functions.

As mentioned by Orvieto and Lucchi [30] in the stochastic setting, and for accelerated methods by Wibisono, Wilson, and Jordan [51], one can thus work either with X_t generated by the nonautonomous gradient flow (2.15) or with Y_t generated by the gradient flow. However, acceleration on X_t is not preserved after discretizing the flow. For example, applying an explicit Euler scheme to the nonautonomous gradient flow (2.15) with $\alpha_s = \alpha > 0$ and originating from a function $f \in \mathcal{F}_{0,L}$, a condition on step sizes h arises, $0 \leqslant h \leqslant \frac{2}{L\alpha}$.

When focusing on continuous-time models for analyzing explicit first-order methods, we usually prefer working with the gradient flow (1.3). However, the nonautonomous gradient flow (2.15) may be useful for analyzing other methods such as proximal methods. More generally, in the next section, we analyze a family of second-order gradient flows without adjusting the time scale (taking $\alpha_t = 1$).

2.3.2. A nonautonomous second-order gradient flow. Nesterov's accelerated gradient flow reaches an $\mathcal{O}(\frac{1}{t^2})$ convergence in function values (see Theorem 2.7). We study convergence properties of a nonautonomous second-order gradient flow and compare them with those of the accelerated gradient flow for the family of quadratic Lyapunov functions (2.11). Let $\beta_t \geqslant 0$ be a continuous function, and a second-order nonautonomous gradient flow

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

Remark 2.10. Wibisono, Wilson, and Jordan [51, Theorem 2.2] proved that this ODE is related to the family of ODEs defined by $\dot{Y}_t + \tilde{\beta}_t Y_t + \tilde{\alpha}_t \nabla f(Y_t) = 0$. Let $\alpha_t > 0$ be a continuously differentiable function with respect to time, $\tau_t = \int_0^t \alpha_s ds$ a time rescaling and X_t a solution to (2.16). The trajectory $Y_t = X_{\tau_t}$ is solution to $\ddot{Y}_t + (\beta_{\tau_t}\sqrt{\alpha_t} - \frac{\dot{\alpha}_t}{2\alpha_t})\dot{Y}_t + \alpha_t \nabla f(Y_t) = 0$. In contrast with α_t in (2.15), note that changing β_t in (2.16) does not correspond to a time rescaling.

Theorem 2.11 provides a systematic condition for the function $\mathcal{V}$ taken among the class of quadratic Lyapunov functions (2.11) to be a Lyapunov function for the second-order gradient flow (2.16).

THEOREM 2.11. *Let $a_t \geqslant 0$ and $P_t \succcurlyeq 0$ be continuously differentiable with respect to time, and let $\mu \geqslant 0$. 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 dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{\mu,\infty}$, and all trajectories X_t generated by the nonautonomous second-order gradient flow* (2.16)*, where $\mathcal{V}_{a_t,P_t}$ is a quadratic Lyapunov function of the form* (2.11)*.*
- *There exist $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that*

$$\begin{pmatrix} -\frac{\mu}{2}(\lambda_t^{(1)} + \lambda_t^{(2)}) + \dot{p}_t^{(11)} & p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} & -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} \\ p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} & 2(p_t^{(12)} - \beta_t p_t^{(22)}) + \dot{p}_t^{(22)} & -p_t^{(22)} + \frac{a_t}{2} \\ -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} & -p_t^{(22)} + \frac{a_t}{2} & 0 \end{pmatrix} \preccurlyeq 0, \tag{2.17}$$

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

The LMI (2.17) from Theorem 2.11 is parametrized by β_t. When $\beta_t = \frac{3}{t}$ and $\mu = 0$, we recover Theorem 2.7 for Nesterov's accelerated gradient flow.

COROLLARY 2.12. *Let $\mu \geqslant 0$. The 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 $\mu > 0$, $a_t = e^{\tau t}$, with $\tau = \min(\sqrt{\mu}, \frac{2}{3}\beta_t)$,*
- *if $\mu = 0$, $a_t = \min\big((\sqrt{a_0} + (\sqrt{p_0^{(11)}}/2)t)^2, \lim_{\epsilon\to 0,\epsilon>0} a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds}\big)$,*

verifies $\frac{d}{dt}\mathcal{V}(X_t,t) \leqslant 0$ for all dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{\mu,\infty}$, and all trajectories X_t generated by the second-order gradient flow (2.16)*.*

Proof. The proof follows from Theorem 2.11 and is detailed in Appendix A.2. ∎

Given a convex function $f \in \mathcal{F}_{0,\infty}$ and a quadratic Lyapunov function, Corollary 2.12 allows us to conclude the following about the convergence of the second-order gradient flow (2.16): given the class of quadratic Lyapunov functions (2.11), it cannot converge faster in function values than Nesterov's accelerated gradient flow, i.e., not faster than $\mathcal{O}(\frac{1}{t^2})$.

To analyze Nesterov's accelerated gradient methods using ODEs, Su, Boyd, and Candès [42] introduced a parametrized second-order gradient flow that fits the model (2.16):

$$\ddot{X}_t + \frac{r}{t}\dot{X}_t + \nabla f(X_t) = 0, \tag{2.18}$$

where $r \geqslant 0$. 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 any function $f \in \mathcal{F}_{0,\infty}$ and any trajectory X_t generated by the accelerated gradient flow (2.18) [42, Theorem 5]. When $r < 3$, Attouch, Chbani, and Riahi [1, Theorem 2.1] proved a convergence bound in function values $f(X_t) - f_\star = \mathcal{O}(\frac{1}{t^{2r/3}})$. Using Corollary 2.12, 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 2.13. Polynomial convergence can be achieved up to a time rescaling, as was shown by Wisobono, Wilson, and Jordan [51]. Given $\tau_t = t^{p/2}$ ($\alpha_t = \frac{p}{2}t^{p/2-1}$), Nesterov's accelerated gradient flow ($r = 3$) transforms into $\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + \frac{p^2}{4}t^{p-2}\nabla f(X_t) = 0$ for $p \geqslant 2$. Corollary 2.12 ensures convergence in function values $f(X_t) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2t^p}$ for all dimensions $d \in \mathbf{N}$, all functions $f \in \mathcal{F}_{0,\infty}$, and all trajectories X_t generated by the second-order gradient flow (2.16).

We have extended the performance estimation approach to continuous-time models using Lyapunov functions. Given a (possibly accelerated) gradient flow and a class of functions, we presented a semidefinite formulation equivalent with the existence of a certain type of quadratic Lyapunov function. The next section is devoted to the analysis of continuous-time models approximating SGD.

3. SDEs for modeling SGD. Convergence results for stochastic convex optimization often require additional assumptions on problem classes, refined choices of step sizes, and averaged iterates. Their analyses raise more complex proofs in contrast with deterministic methods. Prior works have been concerned with connections between stochastic methods and stochastic differential equations (SDEs) [25, 30, 53, 40, 39]. This section is devoted to convergence analyses of SDEs approximating stochastic methods using a systematic Lyapunov approach. Verifying a small-sized LMI will be sufficient (but not necessary) for verifying a quadratic Lyapunov function.

Stochastic gradient descent (SGD) is given by

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

where $\gamma > 0$ is the step size, ξ_{i_k} are uniformly drawn in $(\xi_1, \dots, \xi_n)$, and where $\nabla \tilde{f}(x_k, \xi_{i_k})$ is an unbiased estimate of full gradient $\nabla f(x_k)$. Li, Tai, and E [25] introduced stochastic modified equations (SMEs) to model SGD, rewriting it as

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

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. The SDE (3.2) is an (order-1 weak) approximation of SGD [25, Theorem 1], [26], which allows us to take into account the role of constant step size in the dynamics of SGD (while keeping them small). Under mild assumptions on f, Li, Tai, and E [26] 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, the approximation point of view from this approach is relatively limited since C depends exponentially on T.

Remark 3.1. The SDE (3.2) is an approximation of SGD for small step sizes $\gamma \geqslant 0$. When the step size goes to zero, the noise term actually disappears, and the limiting ODE of SGD is exactly the gradient flow (1.3). In contrast, the stochastic Langevin dynamics [31] $x_{k+1} = x_k - \gamma \nabla f(x_k) - \sqrt{\gamma}\xi_k$ have the limiting ODE $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$, where the step size is not taken into account.

Compared to gradient descent, SGD does not converge to a stationary point under constant step sizes [3, 46]. Convergence to a stationary point requires diminishing step sizes such as $\gamma_k = \frac{1}{\sqrt{k}}$. Li, Tai, and E [25, section 4.1] (and later Orvieto and Lucchi [30, 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 time-varying part. For $h_t \geqslant 0$ a continuous function playing the role of step sizes h_k, the SDE is

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

We treat the covariance matrix $\Sigma(X_t)$ as symmetric, as already implied by the notation $\Sigma(X_t)^{1/2}$, but unstructured with bounded variance $\Sigma(X_t) \preccurlyeq \Sigma$ 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 continuously differentiable.

In this section, we analyze approximating SDEs with averaging techniques and include time-varying step sizes later. Verifying Lyapunov functions thanks to small-sized LMIs, we retrieve convergence results from discrete-time 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 any trajectory generated by (1.3) (see section 2). 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 $\gamma > 0$ (and $h_t = 1$), 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 differential with respect to time of a function of a solution to a stochastic process is given by Ito's lemma [44, Theorem 4.2].

LEMMA 3.2 (Ito's lemma). *Let g be a twice continuously differentiable function and X_t be a stochastic process solution to the SDE* (3.2)*; then*

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

3.1.1. Minimizing strongly convex functions. When the SDE originates from (possibly nonsmooth) 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 along the gradient flow. Given X_t a solution to the SDE (3.2), $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \frac{1}{2}e^{2\mu t}\gamma \mathbf{E}\mathrm{Tr}(\nabla^2_{xx} f(X_t)\Sigma(X_t))$ follows from Ito's formula. After integration between 0 and t, we have

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

We cannot conclude about the convergence of the trajectory X_t without additional requirements on the problem class. For example, let the smoothness parameter of f be finite $L < \infty$ and recalling the bounded covariance assumption $\Sigma(X_t) \preccurlyeq \Sigma$, the variance term is bounded by $\frac{1}{2}L\gamma\mathrm{Tr}(\Sigma)$. Therefore, under constant step sizes, the SDE approximating SGD converges to a diffusion. In addition, it cannot get to a stationary point $x_\star$, because of the extra term, which depends linearly in the step size γ. As in the deterministic case, the forgetting of initial conditions remains of order $\mathcal{O}(e^{-2\mu t})$.

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 the deterministic gradient flow (1.3) originating from functions $f \in \mathcal{F}_{0,\infty}$, the Lyapunov function $\mathcal{V}(x,t) = t(f(x) - f_\star) + \frac{1}{2}\|x - x_\star\|^2$ followed from Theorem 2.4. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function, and let X_t be any trajectory solution to the SDE (3.2). Thanks to Ito's formula, we compute the derivative of $\mathcal{V}(\cdot)$ with respect to time as follows: $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t) \leqslant -t\mathbf{E}\|\nabla f(X_t)\|^2 + \mathbf{E}\frac{1}{2}\mathrm{Tr}((t\nabla_x^2 f(X_t) + I)\Sigma(X_t))$. Using the bounded covariance assumption $\Sigma(X_t) \preccurlyeq \Sigma$ and assuming in addition f to be L-smooth with $L < +\infty$, a convergence bound in function values is given by

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

Such an inequality does not allow us to conclude about convergence of the SDE (3.3) without further assumptions.

In discrete-time, Taylor and Bach proved a comparable convergence bound for SGD [46, Theorem 5], applying the Lyapunov performance estimation approach under similar assumptions (bounded variance, smoothness of f). Optimization methods and alternative techniques have been developed to ensure the global convergence of SGD to the optimum, among them averaging [34] and diminishing step sizes.

3.2. Diminishing the step size is a key to success. We study convergence of SDEs with time-varying step sizes (3.3). In contrast to the deterministic setting, in which time-varying step sizes correspond to a time rescaling whose benefit disappears after discretization, such a time rescaling plays a direct role in the variance term (explicit formula by Orvieto and Lucchi in [30, Theorem 5]).

Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function, X_t be generated by (3.3), and $\mathcal{V}$ be a function. Our goal is to control the maximization problem

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

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 using Ito's formula. The first two terms correspond exactly to taking the derivative in trajectories generated by ODEs (2.15) (or the SDEs (3.3) with $\gamma = 0$), and the last term corresponds to a variance term. Because of the trace in a second-order derivative of f and in the covariance matrix $\Sigma(X_t)$, we do not take this term into account in an LMI formulation. Instead, we propose to first derive a family of Lyapunov functions for associated ODEs using LMIs (deterministic setting), and then to optimize their parameters so that the variance term converges conveniently. In other words, the following functions $\mathcal{V}$ are Lyapunov functions for induced ODEs ($\gamma = 0$),

but are not Lyapunov functions for the SDEs under consideration. This approach allows a systematic computation of such quadratic functions and leads to convergence guarantees.

COROLLARY 3.3. *Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function, and let $X_t \in \mathbf{R}^d$ be generated by the SDE* (3.3)*. The quadratic function*

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

with $\dot{a}_t^{(1)} = 2h_t$, verifies $\frac{d}{dt}\mathbf{E}(\mathcal{V}(X_t,t) \leqslant h_t^2 \mathbf{E}\mathrm{Tr}((\nabla^2_{xx} f(X_t) a_t^{(1)} + \frac{1}{2} I_d)\Sigma(X_t))$.

Furthermore, it holds that

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{a_t^{(1)}} + \frac{\gamma}{2a_t^{(1)}} \int_0^t h_s^2 \mathbf{E}\mathrm{Tr}((\nabla^2_{xx} f(X_s) a_s^{(1)} + \tfrac{1}{2} I_d)\Sigma(X_s))ds.$$

Proof. The function $\mathcal{V}$ is obtained using Corollary 2.8 and is a Lyapunov function for a nonautonomous first-order gradient flow. The bound for $\mathbf{E}[f(X_t) - f_\star]$ is derived using Ito's formula on $\mathcal{V}$ along trajectories X_t generated by the SDE (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 [3, Theorem 5] provided a comparable but much more complex analysis for stochastic gradient descent, for a specific family of step sizes. To compare to our results, let us consider step sizes defined by $h_t = \frac{1}{(t+1)^\alpha}$, where $\alpha \geqslant 0$. A possible convergence guarantee arises from Corollary 3.3 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 $\Sigma(X_t) \preccurlyeq \Sigma$, the variance term is bounded by

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

The variance term converges to zero if and only if $\alpha \geqslant \frac{1}{2}$. In other words, convergence is not guaranteed for constant step sizes ($\alpha = 0$). For $\alpha \in (1/2, 2/3)$, the convergence in function value is bounded by $\mathcal{O}(\frac{1}{t^{2\alpha - 1}})$. For $\alpha \in (2/3, 1)$, the convergence in function value is bounded by $\mathcal{O}(\frac{1}{t^{1-\alpha}})$. As for SGD [3, 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 converge at the same rate (up to $\log(t)$). It is therefore possible to reach convergence with diminishing step sizes. Other techniques, such as averaging, 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 [34, 33] is a standard way to improve convergence of SGD. In the discrete-time setting, convergence guarantees are considered at an averaged sequence, where i_k is drawn uniformly in $[1, \ldots, n]$ and n is the sample size:

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

Other averaging techniques were developed later, such as primal averaging [45] and averaging with respect to some nonnegative function [23].

3.3.1. Polyak–Ruppert averaging. In this section, we analyze convergence properties of an SDE (3.3) approximating SGD with time-varying step sizes under Polyak–Ruppert averaging. Taylor and Bach [46, Theorem 6] provided a systematic design of Lyapunov functions for (3.4) and a condition on the step size for convergence to the optimum. An approximating SDE for Polyak–Ruppert averaging is given by

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

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 functions taking the averaged sequence into account:

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

where $a_t^{(1)}, a_t^{(2)} \geqslant 0$ and $P_t = \left(\begin{smallmatrix} p_t^{(11)} & p_t^{(12)} \\ p_t^{(12)} & p_t^{(22)} \end{smallmatrix}\right) \succcurlyeq 0$ are continuously differentiable with respect to time. Given a quadratic function $\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}$ and an SDE (3.5) under Polyak–Ruppert averaging, we present in Theorem 3.4 a way to control the quantity:

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

THEOREM 3.4. *Let $h_t \geqslant 0$, and let $a_t \geqslant 0$, $P_t \succcurlyeq 0$ be continuously differentiable with respect to time. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and $(X_t, \bar{X}_t) \in \mathbf{R}^d \times \mathbf{R}^d$ be a trajectory generated by the SDE under Polyak–Ruppert averaging* (3.5)*. Let $\mathcal{V}$ be a quadratic function* (3.6)*, such that there exist $\lambda_t^{(1)}, \dots, \lambda_t^{(6)} \geqslant 0$ verifying*

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

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

Then, the following inequality is satisfied:

$$\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,\bar{X}_t,t) \leqslant \frac{1}{2}\mathbf{E}\mathrm{Tr}\left((a_t^{(1)}\nabla_{xx}f(X_t) + 2p_t^{(11)}I_d)\Sigma(X_t)\right)h_t^2\gamma.$$

Proof. The proof follows the method from section 2.1.1 (see Appendix B.1). ∎

The variance term $\frac{1}{2}\mathbf{E}\mathrm{Tr}((a_t^{(1)}\nabla_{xx}f(X_t) + 2p_t^{(11)}I_d)\Sigma(X_t))h_t^2\gamma$ from Theorem 3.4 increases with $a_t^{(1)}$, and its convergence requires additional assumptions on f, for example smoothness. For this reason, we propose to analyze convergence guarantees based on functions $\mathcal{V}_{0,a_t^{(2)},P_t}$, on the averaged sequence only.

COROLLARY 3.5 (averaging and diminishing step sizes). *Let $h_t \geqslant 0$ and $a_t^{(2)} \geqslant 0$ be continuously differentiable. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and $(X_t, \bar{X}_t)$ be a trajectory generated by the SDE under Polyak–Ruppert averaging* (3.5). *Assuming that $\dot{a}_t^{(2)} \leqslant \frac{a_t^{(2)}}{t}$ and $t \to \frac{a_t^{(2)}}{th_t}$ is a nonincreasing function, the 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 $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, \bar{X}_t, t) \leqslant \frac{a_t^{(2)}}{t} h_t \mathbf{E}\mathrm{Tr}(\Sigma(X_t))$. Furthermore, it holds that $\mathbf{E}[f(\bar{X}_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{2a_t^{(2)}} + \frac{\gamma}{2a_t^{(2)}} \int_0^t \frac{a_s^{(2)}}{s} h_s \mathbf{E}\mathrm{Tr}(\Sigma(X_s))ds$.

Proof. The proof follows from the LMI in Theorem 3.4 with the choices $\lambda_t^{(3)} = \lambda_t^{(6)} = \lambda_t^{(2)} = 0$, $\lambda_t^{(5)} = \frac{a_t^{(2)}}{t}$, $\lambda_t^{(4)} = \lambda_t^{(1)} = h_t p_t^{(11)}$, $\dot{a}_t^{(2)} = \frac{a_t^{(2)}}{t}$, $\dot{p}_t^{(11)} = (\frac{a_t^{(2)}}{th_t})^{\cdot}$, $\dot{p}_t^{(12)} = \dot{p}_t^{(22)} = 0$. ∎

When $a_t^{(2)} = t$ (its maximal possible value), the step size verifies $\dot{h}_t \leqslant 0$. The variance term does not diverge if and only if h_t is constant. Recalling the unbounded covariance $\Sigma_t \preccurlyeq \Sigma$, 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 convergence and diminishing step size, as obtained without averaging.

Under the assumptions of Corollary 3.5, recalling that $\Sigma(X_t) \preccurlyeq \Sigma$ is a bounded covariance matrix, let $h_t = \frac{1}{(t+1)^\alpha}$ be the step size and $a_t^{(2)} = t^\beta$, where $\alpha \geqslant 0$ and $0 \leqslant \beta \leqslant 1$. The decreasing condition imposes $\alpha + \beta \leqslant 1$. In comparison with step size requirements drawn from Corollary 3.3, where convergence required $\alpha \geqslant \frac{1}{2}$, averaging allows larger step sizes.

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 behaves as $\mathcal{O}(\frac{1}{t^\beta})$, and the variance term behaves as $\mathcal{O}(\frac{1}{t^\alpha})$ if $\beta \neq \alpha$. When $\alpha = \beta$, the variance term behaves as $\mathcal{O}(\frac{\log(t)}{2t^\beta})$. Hence, a natural choice is $\alpha = \beta = \frac{1}{2}$, retrieving results from [3, Theorem 4], [46, Table 2] in discrete-time optimization.

3.3.2. Weighted averaging. Polyak–Ruppert averaging performs uniform averaging of any trajectory X_t over the time step. We introduce weighted averaging to analyze SGD that is defined with respect to a function $u_t \geqslant 0$ [23]:

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

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

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

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

THEOREM 3.6. *Let $\mu \geqslant 0$. Let $d \in \mathbf{N}$ be the dimension, $f \in \mathcal{F}_{\mu,\infty}$ be a twice continuously differentiable (possibly strongly convex) function, and $(X_t, \bar{X}_t)$ be trajectories*

generated by an SDE under generalized averaging (3.7). *Assuming* $\big(\frac{\int_0^t \dot{u}_s ds}{2h_t}\big) 2\mu u_t$, *the 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 $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, \bar{X}_t^u, t) \leqslant \frac{1}{2} u_t h_t \gamma \mathbf{E}\mathrm{Tr}(\Sigma(X_t))$.

Then, it holds that

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

Proof. The proof follows those of Theorem 3.4 and Corollary 3.5 with $\frac{1}{t} \to C_t^u$. ∎

Under the convexity assumption ($\mu = 0$), u_t verifies $(\frac{\int_0^t \dot{u}_s ds}{2h_t}) \leqslant 0$. Assuming step sizes of the form $h_t = \frac{1}{(t+1)^\alpha}$ and an averaging function $u_t = \frac{1}{(t+1)^\beta}$, with $\alpha, \beta \geqslant 0$, it follows that $\beta \geqslant \alpha$. Recall that the covariance matrix $\Sigma(X_t) \preccurlyeq \Sigma$ is bounded. From Theorem 3.6, the term that forgets the initial conditions behaves as $\mathcal{O}(\frac{1}{(t+1)^{1-\beta}})$, and the variance term behaves as $\mathcal{O}(\frac{1}{(t+1)^\alpha})$. Both terms converge at the same rate for $\alpha = \beta = \frac{1}{2}$ (up to $\log(t+1)$). In discrete time, similar convergence results have been derived for SGD under Polyak–Ruppert averaging [3, Theorem 6].

Under the strong convexity assumption ($\mu > 0$), polynomial convergence can be reached for the term that contains the initial conditions. However, the variance term cannot converge faster than the step size. Recalling that the covariance matrix $\Sigma(X_t) \preccurlyeq \Sigma$ is bounded, if $u_t = (t+1)^\beta$ and $h_t = (t+1)^{-\alpha}$ with $\alpha, \beta \geqslant 0$, the condition $\big(\frac{\int_0^t \dot{u}_s ds}{2h_t}\big) \leqslant 2\mu u_t$ implies that $\alpha = 0$. The variance term is then exactly equal to the step size $h_t = h_0$. Hence, weighted 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 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 using approximating SDEs (3.3). Continuous-time analyses lead to similar convergence results, while benefiting from simpler formulations and fewer assumptions especially on step sizes. Using this approach, we analyzed the trade-off between nonuniform averaging and step sizes, paving the way to a better understanding of averaging techniques. In the next section, 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 a family of second-order gradient flows, including in particular 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 a twice continuously differentiable function and $\gamma > 0$ be constant step sizes. An approximating SDE (for order-1 weak approximations) for Nesterov's accelerated gradient is given by [26, Theorem 16, section 4.4]

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

As for SGD, the function $\mathcal{V}(x,\dot{x},t) = t^2(f(x)-f_\star) + 2\|(x-x_\star)+\frac{t}{2}\dot{x}\|^2$ obtained from Theorem 2.7 does not allow us to conclude about convergence to a stationary point of trajectories X_t generated by the stochastic accelerated gradient flows (4.1). Using the bounded covariance assumption $\Sigma(X_t) \preccurlyeq \Sigma$ and applying Ito's formula to $\mathcal{V}(\cdot)$ along X_t, we have

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

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 [42]. Let us define the stochastic accelerated gradient flow under Polyak–Ruppert averaging:

$$\begin{aligned} &d^2X_t + \frac{3}{t}dX_t + \nabla f(X_t)dt + \sqrt{\gamma\Sigma}dB_t = 0, \\ &d\bar{X}_t = \frac{X_t - \bar{X}_t}{t}dt. \end{aligned} \tag{4.2}$$

The family of quadratic functions in consideration is given by

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

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

THEOREM 4.1. *Let $a_t^{(1)}, a_t^{(2)} \geqslant 0$ and $P_t \succcurlyeq 0$ be continuously differentiable with respect to time. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and $(X_t, \bar{X}_t)$ be a trajectory generated by the stochastic accelerated gradient flow under Polyak–Ruppert averaging* (4.2) *with constant step sizes ($h_t = 1$). Let, in addition, $\mathcal{V} = \mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(\cdot)$ be a quadratic function* (4.3)*. Then it holds that*

- *When $a_t^{(1)} = 0$, if the function $\mathcal{V}$ verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,\bar{X}_t,t) \leqslant \mathrm{Tr}(2p_t^{(11)}\gamma\Sigma(X_t))$, then $\mathcal{V} = 0$.*
- *When $a_t^{(2)} = 0$, the 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 $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t) \leqslant \mathbf{E}\mathrm{Tr}(2a_t^{(1)}\gamma\Sigma(X_t))$, with $a_t^{(1)} \leqslant t^2$. Furthermore, it holds that $\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{2\|x_0 - x_\star\|^2}{a_t^{(1)}} + \frac{1}{2a_t^{(1)}}\gamma\int_0^t a_s^{(1)}\mathbf{E}\mathrm{Tr}(\Sigma(X_s))ds$.

Proof. The proof follows from Theorem 2.11 (see Appendix B.2). The first statement holds when considering a function $\beta_t \geqslant 0$ instead of $\frac{3}{t}$. ∎

Without further assumptions on the covariance Σ_t, the accelerated gradient flow under Polyak–Ruppert averaging (4.2) with constant step sizes admits no quadratic function that allows forgetting of the initial conditions while reducing the variance term. Therefore, Polyak–Ruppert averaging plays a different role in the stochastic accelerated gradient flow compared to the stochastic gradient flow approximating SGD (3.5).

4.2. Diminishing step sizes do not help preserve acceleration. Averaging under constant step sizes was not conclusive for finding a convergence guarantee for Nesterov's accelerated gradient flow with a diffusion term. We consider a second-order nonautonomous stochastic gradient flow with time-varying step sizes $h_t \geqslant 0$,

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

To study the convergence of (4.4), we generate a quadratic function within the class (2.11).

THEOREM 4.2. *Let $h_t \geqslant 0$, $\gamma > 0$, and $a_t \geqslant 0$ be continuously differentiable with respect to time. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and X_t be a trajectory generated by the stochastic second-order gradient flow* (4.4)*. 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 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 $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \frac{1}{4}\mathbf{E}\mathrm{Tr}(a_t h_t \Sigma(X_t))\gamma$.

Proof. This result is obtained by extending the LMI for ODEs from Theorem 2.11 and Corollary 2.12 to time-varying step sizes. ∎

To compare to previous results, we consider parametrized step sizes $h_t = \frac{1}{(t+1)^\alpha}$ and SDEs with $\beta_t = \frac{b}{t}$, where $\alpha, b > 0$. We derive a convergence bound using Theorem 4.2 together with Theorem 2.11, that $\dot{a}_t \leqslant \beta\frac{a_t}{t}$, where $\beta \leqslant \min(\frac{2b-\alpha}{3}, 2-\alpha)$:

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

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, under bounded covariance $\Sigma(X_t) \preccurlyeq \Sigma$, the variance term behaves as $\mathcal{O}(\frac{1}{t^\beta})$ if $\beta \leqslant \alpha - 1$, and as $\mathcal{O}(\frac{1}{t^{\alpha-1}})$ 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 the result from Corollary 3.5 for the SDE approximating SGD under Polyak–Ruppert averaging. Yet this result requires smaller step sizes ($\alpha \geqslant \frac{3}{2}$). It does not seem possible to accelerate SGD with diminishing step sizes. Ghadimi and Lan [36, Corollary 3] proved a convergence bound for a stochastic accelerated gradient method with $\beta = 2$ and $\alpha = \frac{1}{2}$ that we do not retrieve. Yet, in their approach, functions are 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 finding quadratic Lyapunov functions for families of ODEs and SDEs approximating SGD. Verifying such a Lyapunov function is cast as verifying the feasibility of a small-sized LMI. From this formulation, it is possible to efficiently search for quadratic Lyapunov functions for arriving to convergence bounds. While we retrieve convergence guarantees similar to those of discrete-time systems, continuous-time models require fewer assumptions on the problem classes and can be analyzed through shorter proofs.

While obtaining guarantees for stochastic optimization methods might be tedious, the SDE approach allows for simpler analysis of the trade-off between the variance term and the 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 efficiently [17]. In stochastic analysis, stochastic modified equations have been introduced by Li, Tai, and E [25, Theorem 1] to better approximate SGD, and stochastic methods with momentum. For nonconvex functions, some analyses have also been done by Shi, Su, and Jordan [41]. However, these approximation theorems often require many assumptions on the class of functions, which we believe could be further simplified using computer-assisted proofs.

Concerning possible extensions, this work relies on a specific family of quadratic Lyapunov functions provided by (2.11) and (2.8). Whereas it was sufficient for our purposes, it turns out that it is possible to extend it by taking into account terms of the form $\int_0^t f(X_s)ds$ or integrals of the quadratic terms that are used in our Lyapunov functions. While we did not consider those terms here, they could be useful for studying other methods or in other settings. The attentive reader might also have realized that the PEP technique for continuous-time systems can be applied without difficulty to differential [7, section 3.2] and monotone inclusion [5, 6] problems. For monotone inclusions, interpolation results for casting the PEPs as semidefinite programs can be found in [35]. Finally, we note that it is still not clear how to use PEP-related techniques for directly studying higher-order methods and assumptions (already appearing in the variance term in the stochastic setting), which are also common for continuous-time systems [2], or equivalents in the monotone inclusion setting [10, 8]. The problem here is the lack of a clean performance estimation reformulation, beyond the somewhat indirect approach by [11]. We leave those investigations for future work.

Appendix A. Proof for ODEs.

A.1. Proofs for Theorem 2.4.

Proof. Following the same procedure as for Theorem 2.1 under the strong-convexity assumption, and given a quadratic Lyapunov function $\mathcal{V}_{a_t,c_t}(X_t,t) = a_t(f(x_t) - f_\star) + c_t\|X_t - x_\star\|_2^2$ with $a_t, c_t \geqslant 0$, the maximization problem can be formulated into a semidefinite program

$$\begin{aligned} 0 \geqslant \max_{G\succcurlyeq 0, F\in\mathbf{R}^2} & b_0^\top F + \operatorname{Tr}(A_0 G), \\ \text{subject to } & b_i^\top F + \operatorname{Tr}(A_i G) \geqslant 0,\ i \in \{1,2\}, \end{aligned}$$

where $A_0 = \left(\begin{smallmatrix} \dot{c}_t & -c_t \\ -c_t & -a_t \end{smallmatrix}\right)$, $A_1 = \left(\begin{smallmatrix} 0 & 1/2 \\ 1/2 & 0 \end{smallmatrix}\right)$, $A_2 = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $b_0 = \dot{a}_t[1,\ -1]^\top$, $b_1 = [-1,\ 1]^\top$, and $b_2 = [1,\ -1]^\top$, whose Lagrangian dual is given by the feasibility problem

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

This formulation is exactly the LMI feasibility problem from the statement of Theorem 2.4. ∎

A.2. Proof for Corollary 2.12.

Proof. We prove the claim in the convex case; the strongly convex setting is deduced by assuming an exponential form for the parameters $a_t = ae^{t\tau}$ and $P_t = Pe^{t\tau}$, where $\tau > 0$ is a (linear) convergence rate to be determined.

From Theorem 2.11, verifying the Lyapunov function can be framed as verifying the feasibility of an LMI. That is, the desired Lyapunov function can be deduced from the existence of $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that

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

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

Because of the zero diagonal term in $S = (s_{ij})_{1 \geqslant ij \geqslant 3} \preccurlyeq 0$, it follows that $p_t^{(22)} = \frac{a_t}{2}$ and $p_t^{(12)} = \frac{\dot{a}_t}{2}$ (otherwise, the submatrix $\left(\begin{smallmatrix} s_{22} & s_{23} \\ s_{23} & 0 \end{smallmatrix}\right) \preccurlyeq 0$ has a strictly nonpositive determinant, which is a contradiction).

Let us now assume that $\lambda_t^{(2)} = 0$, which leads to $\dot{a}_t = \lambda_t^{(1)}$. Because of the null entries, the matrix S can be reduced to a 2×2 semidefinite negative matrix. Such a matrix has nonpositive diagonal terms (since $s_{11}s_{22} \geqslant (s_{12})^2$ and $s_{11} + s_{22} \geqslant 0$), which are $\dot{p}_t^{(11)} \leqslant 0$ and $2(p_t^{(12)} - \beta_t p_t^{(22)}) + \dot{p}_t^{(22)}$, which simplifies into $\dot{a}_t \leqslant \frac{2}{3}\beta_t a_t$.

For all $\epsilon > 0$, after integration of $\dot{a}_t \leqslant \frac{2}{3}\beta_t a_t$ between ϵ and t, we have that $a_t \leqslant a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds}$. Therefore, $a_t \leqslant \lim_{\epsilon \to 0} a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds}$. Recalling that P_t is positive semidefinite, its determinant is nonnegative $p_t^{(11)} p_t^{(22)} - (p_t^{(12)})^2 \geqslant 0$, that is, $p_t^{(11)} \geqslant \frac{(p_t^{(12)})^2}{p_t^{(22)}} = \frac{(\dot{a}_t)^2}{2a_t}$. After integration between 0 and t, we obtain the following bound on a_t: $\sqrt{a_t} \leqslant \sqrt{a_0} + \frac{\sqrt{p_0^{(11)}}}{2} t$.

In other words, $a_t \leqslant \min\left((\sqrt{a_0} + \frac{\sqrt{p_0^{(11)}}}{2} t)^2, \lim_{\epsilon \to 0} a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds}\right)$. ∎

Appendix B. Proofs for SDEs.

B.1. Proof for Theorem 3.4.

Proof. We rewrite the SDE into

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

and denote $Y_t = \begin{pmatrix} X_t \\ \bar{X}_t \end{pmatrix}$.

We consider the quadratic function (3.6) $\mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}(X_t, \bar{X}_t, t) = \mathcal{V}(Y_t, t)$. Applying Ito's formula, its derivative with respect to time is $\frac{d}{dt}\mathcal{V}(Y_t, t) = \frac{\partial}{\partial t}\mathcal{V}(Y_t, t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt} + \frac{1}{2}\gamma h_t^2 \mathrm{Tr}[\frac{\partial^2}{\partial Y_t^2}\mathcal{V}(Y_t, t)^\top \left(\begin{smallmatrix} \Sigma(X_t) \\ 0 \end{smallmatrix}\right)]$. By taking the expectation of Ito's formula, we have $\frac{d}{dt}\mathbf{E}\mathcal{V}(Y_t, t) = \mathbf{E}[\frac{\partial}{\partial t}\mathcal{V}(Y_t, t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt}] + \frac{1}{2}\gamma h_t^2 \mathbf{E}\mathrm{Tr}(\frac{\partial^2}{\partial Y_t^2}\mathcal{V}(Y_t, t)^\top \left(\begin{smallmatrix} \Sigma(X_t) \\ 0 \end{smallmatrix}\right))$.

The variance term $\frac{1}{2}\gamma h_t^2 \mathbf{E}\mathrm{Tr}(\frac{\partial^2}{\partial Y_t^2}\mathcal{V}(Y_t, t)^\top \left(\begin{smallmatrix} \Sigma(X_t) \\ 0 \end{smallmatrix}\right))$ depends on the second derivative of f in the space variable (X_t), and on the covariance Σ_t. We do not take this term into account in the performance estimation framework.

We formulate the problem of verifying a quadratic function as verifying the inequality $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,\bar{X}_t,t) \leqslant \frac{1}{2}\gamma h_t^2 \mathbf{E}\mathrm{Tr}(\frac{\partial^2}{\partial X_t^2}\mathcal{V}(X_t,\bar{X}_t,t)^\top \Sigma(X_t))$ (that is, $\mathbf{E}[\frac{\partial}{\partial t}\mathcal{V}(Y_t,t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt}] \leqslant 0$) holds for any function twice continuously differentiable $f \in \mathcal{F}_{0,\infty}$ and any trajectory $(X_t,\bar{X}_t)$ generated by the SDE under Polyak–Ruppert averaging (3.5).

This problem is equivalent to verifying that $\frac{d}{dt}\mathcal{V}(\tilde{Y}_t,t) \leqslant 0$ holds for any twice continuously differentiable function $f \in \mathcal{F}_{0,\infty}$ and any trajectory $\tilde{Y}_t = (\tilde{X}_t \quad \dot{\tilde{X}}_t)^\top$ generated by deterministic gradient flow from the SDE (3.5) with $\gamma = 0$.

We follow the methodology developed in section 2 for ODEs. We formulate the verification of such a Lyapunov function as the maximization problem

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

This maximization problem can be formulated as the following SDP program:

$$\min_{\lambda_t^{(i)}\geqslant 0,\ i\in\{1,\dots,6\}} 0$$
$$\begin{pmatrix} \dot{p}_t^{(11)} + \frac{2p_t^{(12)}}{t} & \dot{p}_t^{(12)} + \frac{p_t^{(22)}-p_t^{(12)}}{t} & \frac{\lambda_t^{(6)}+\lambda_t^{(4)}-2h_tp_t^{(11)}}{2} & \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} \\ \dot{p}_t^{(12)} + \frac{p_t^{(22)}-p_t^{(12)}}{t} & \dot{p}_t^{(22)} - \frac{2p_t^{(22)}}{t} & -\frac{\lambda_t^{(6)}-2h_tp_t^{(12)}}{2} & -\frac{a_t^{(2)}}{2t} + \frac{\lambda_t^{(5)}+\lambda_t^{(3)}}{2} \\ \frac{\lambda_t^{(6)}+\lambda_t^{(4)}-2h_tp_t^{(11)}}{2} & -\frac{\lambda_t^{(6)}-2h_tp_t^{(12)}}{2} & -a_t^{(1)} & 0 \\ \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} & -\frac{a_t^{(2)}}{2t} + \frac{\lambda_t^{(5)}+\lambda_t^{(3)}}{2} & 0 & 0 \end{pmatrix} \preccurlyeq 0,$$
$$\dot{a}_t^{(1)} + \lambda_t^{(1)} + \lambda_t^{(5)} = \lambda_t^{(4)} + \lambda_t^{(6)},$$
$$\dot{a}_t^{(2)} + \lambda_t^{(2)} + \lambda_t^{(6)} = \lambda_t^{(3)} + \lambda_t^{(5)} + \dot{a}_t^{(1)}.$$

Note that the Gram matrix G and function value vector F to be introduced are different from those in the setting explored in section 2; more precisely, $G = Q^\top Q \succcurlyeq 0$, $Q = [X_t - x_\star, \bar{X}_t - x_\star, g_t, \bar{g}_t]$, and $F = (f(X_t), f(\bar{X}_t), f_\star)$.

The final inequality follows from Ito's formula. ∎

B.2. Proof for Theorem 4.1.

Proof. From the reasoning from Theorems 2.4 and 3.4, the worst-case guarantee can be formulated into a maximization problem:

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

A control of this quantity can be achieved with the following LMI, for $\lambda_t^{(i)} \geqslant 0,\ i \in \{1,\dots,6\}$, and where $*$ corresponds to the symmetric entries of the matrix, and where $\beta_t = \frac{3}{t}$:

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

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

Thus, $p_t^{(11)} = 0$, and because P_t is positive semidefinite, $p_t^{(12)} = p_t^{(13)} = 0$. If in addition $a_t^{(1)} = 0$, 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 2.11, and the Lyapunov function follows from Corollary 2.12. ∎

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 Optim. Calc. Var., 25 (2019), p. 2.

[2] H. Attouch, J. Peypouquet, and P. Redont, *Fast convex optimization via inertial dynamics with Hessian driven damping*, J. Differential Equations, 261 (2016), pp. 5734–5783.

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

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

[5] R. I. Boţ and E. R. Csetnek, *Second order forward-backward dynamical systems for monotone inclusion problems*, SIAM J. Control Optim., 54 (2016), pp. 1423–1443, https://doi.org/10.1137/15M1012657.

[6] R. I. Boţ and E. R. Csetnek, *A dynamical system associated with the fixed points set of a nonexpansive operator*, J. Dynam. Differential Equations, 29 (2017), pp. 155–168.

[7] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, *Characterizations of Lojasiewicz inequalities: Subgradient flows, talweg, convexity*, Trans. Amer. Math. Soc., 362 (2010), pp. 3319–3363.

[8] R. I. Bot, E. R. Csetnek, and D.-K. Nguyen, *Fast OGDA in Continuous and Discrete Time*, preprint, https://arxiv.org/abs/2203.10947, 2022.

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

[10] T. Chavdarova, M. I. Jordan, and M. Zampetakis, *Last-Iterate Convergence of Saddle Point Optimizers via High-Resolution Differential Equations*, preprint, https://arxiv.org/abs/2112.13826, 2021.

[11] E. De Klerk, F. Glineur, and A. B. Taylor, *Worst-case convergence analysis of inexact gradient and Newton methods through semidefinite programming performance estimation*, SIAM J. Optim., 30 (2020), pp. 2053–2082, https://doi.org/10.1137/19M1281368.

[12] 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.

[13] J. Diakonikolas and M. Jordan, *Generalized momentum-based methods: A Hamiltonian perspective*, SIAM J. Optim., 31 (2021), pp. 915–944, https://doi.org/10.1137/20M1322716.

[14] Y. Drori and M. Teboulle, *Performance of first-order methods for smooth convex minimization: A novel approach*, Math. Program., 145 (2014), pp. 451–482.

[15] A. d'Aspremont, D. Scieur, and A. Taylor, *Acceleration methods*, Found. Trends Optim., 5 (2021), pp. 1–245.

[16] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, *Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems*, SIAM J. Optim., 28 (2018), pp. 2654–2689, https://doi.org/10.1137/17M1136845.

[17] W. Gautschi, *Numerical Analysis*, 2nd ed., Springer, New York, 2012.
[18] B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, and A. Dieuleveut, *PEPit: Computer-Assisted Worst-Case Analyses of First-Order Optimization Methods in Python*, preprint, https://arxiv.org/abs/2201.04040, 2022.
[19] B. Hu and L. Lessard, *Dissipativity theory for Nesterov's accelerated method*, in International Conference on Machine Learning (ICML), 2017.
[20] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems (NIPS), 2013.
[21] R. E. Kalman and J. E. Bertram, *Control system analysis and design via the "second method" of Lyapunov: I-continuous-time systems*, J. Basic Eng., 82 (1960), pp. 371–393.
[22] 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.
[23] N. Le Roux, *Anytime Tail Averaging*, preprint, https://arxiv.org/abs/1902.05083, 2019.
[24] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM J. Optim., 26 (2016), pp. 57–95, https://doi.org/10.1137/15M1009597.
[25] Q. Li, C. Tai, and W. E, *Stochastic modified equations and adaptive stochastic gradient algorithms*, in International Conference on Machine Learning (ICML), 2017.
[26] Q. Li, C. Tai, and W. E, *Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations*, J. Mach. Learn. Res. (JMLR), 20 (2019), pp. 1–47.
[27] S. Lojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in Les Équations aux Dérivées Partielles (Paris, 1961), Éditions du Centre National de la Recherche Scientifique (CNRS), Paris, France, 1963, pp. 87–89.
[28] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Math. Program., 175 (2019), pp. 69–107.
[29] 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 (in Russian).
[30] A. Orvieto and A. Lucchi, *Continuous-time models for stochastic optimization algorithms*, in Advances in Neural Information Processing Systems (NeurIPS), 2020.
[31] P. Langevin, *Sur la théorie du mouvement brownien [on the theory of brownian motion]*, C. R. Acad. Sci. Paris, 146 (1908), pp. 530–533.
[32] B. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phys., 4 (1964), pp. 1–17.
[33] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855, https://doi.org/10.1137/0330046.
[34] D. Ruppert, *Efficient Estimations from a Slowly Convergent Robbins Monroe Process*, Technical report, 1988.
[35] E. K. Ryu, A. B. Taylor, C. Bergeling, and P. Giselsson, *Operator splitting performance estimation: Tight contraction factors and optimal parameter selection*, SIAM J. Optim., 30 (2020), pp. 2251–2271, https://doi.org/10.1137/19M1304854.
[36] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99.
[37] J. M. Sanz Serna and K. C. Zygalakis, *The connections between Lyapunov functions for some optimization algorithms and differential equations*, SIAM J. Numer. Anal., 59 (2021), pp. 1542–1565, https://doi.org/10.1137/20M1364138.
[38] 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.
[39] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Math. Program. Ser. A, 195 (2022), pp. 79–148.
[40] 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.
[41] B. Shi, W. J. Su, and M. I. Jordan, *On Learning Rates and Schrödinger Operators*, preprint, arXiv:2004.06977, 2020.
[42] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, J. Mach. Learn. Res. (JMLR), 17 (2016), pp. 1–43.
[43] J. J. Suh, G. Roh, and E. K. Ryu, *Continuous-time analysis of accelerated gradient methods via conservation laws in dilated coordinate systems*, in Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, PMLR 162, 2022, pp. 20640–20667.
[44] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*, Institute of Mathematical Statistics Textbooks, Cambridge University Press, Cambridge, UK, 2019.

[45] W. Tao, Z. Pan, G. Wu, and Q. Tao, *Primal averaging: A new gradient evaluation step to attain the optimal individual convergence*, IEEE Trans. Cybernet., 50 (2018), pp. 835–845.
[46] A. Taylor and F. Bach, *Stochastic first-order methods: Non-asymptotic and computer-aided analyses via potential functions*, in Conference on Learning Theory (COLT), 2019.
[47] 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.
[48] A. B. Taylor, *Interpolation and Performance Estimation of First-Order Methods for Convex Optimization*, Ph.D. thesis, Chapter 3: Convex interpolation, 2017.
[49] A. B. Taylor, J. M. Hendrickx, and F. Glineur, *Exact worst-case performance of first-order methods for composite convex optimization*, SIAM J. Optim., 27 (2017), pp. 1283–1313, https://doi.org/10.1137/16M108104X.
[50] A. B. Taylor, J. M. Hendrickx, and F. Glineur, *Smooth strongly convex interpolation and exact worst-case performance of first-order methods*, Math. Program., 161 (2017), pp. 307–345.
[51] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. E7351–E7358.
[52] A. C. Wilson, B. Recht, and M. I. Jordan, *A Lyapunov analysis of accelerated methods in optimization*, J. Mach. Learn. Res. (JMLR), 22 (2021), pp. 1–34.
[53] 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.
[54] 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.