A Continuized View on Nesterov Acceleration

Raphaël Berthier, Francis Bach, Nicolas Flammarion, Pierre Gaillard, Adrien Taylor

A Continuized View on Nesterov Acceleration

Raphaël Berthier[1], Francis Bach[1], Nicolas Flammarion[2], Pierre Gaillard[3] and Adrien Taylor[1]

[1]Inria - Département d'informatique de l'ENS
PSL Research University, Paris, France
[2]School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
[3]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Abstract. We introduce the "continuized" Nesterov acceleration, a close variant of Nesterov acceleration whose variables are indexed by a continuous time parameter. The two variables continuously mix following a linear ordinary differential equation and take gradient steps at random times. This continuized variant benefits from the best of the continuous and the discrete frameworks: as a continuous process, one can use differential calculus to analyze convergence and obtain analytical expressions for the parameters; but a discretization of the continuized process can be computed exactly with convergence rates similar to those of Nesterov original acceleration. We show that the discretization has the same structure as Nesterov acceleration, but with random parameters.

1. Introduction

In the last decades, the emergence of numerous applications in statistics, machine learning and signal processing has led to a renewed interest in first-order optimization methods (Bottou et al., 2018). They enjoy a low computational complexity necessary to the analysis of large datasets. The performance of first-order methods was largely improved thanks to acceleration techniques (see the review by d'Aspremont et al., 2021, and the many references therein), starting with the seminal work of Nesterov (1983).

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, minimized at $x_* \in \mathbb{R}^d$. We assume throughout the paper that f is L-smooth, i.e.,

$$\forall x, y \in \mathbb{R}^d, \qquad f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 .$$

In addition, we sometimes assume that f is μ-strongly convex for some $\mu > 0$, i.e.,

$$\forall x, y \in \mathbb{R}^d, \qquad f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 .$$

For the problem of minimizing f, gradient descent is well-known to achieve a rate $f(x_k) - f(x_*) = O(k^{-1})$ in the smooth case, and a rate $f(x_k) - f(x_*) = O((1 - \mu/L)^k)$ in the smooth and strongly convex case. In both cases, Nesterov introduced an alternative method with essentially the same running-time complexity, that achieves faster rates; it converges at the rate $O(k^{-2})$ in the smooth convex case and at the rate $O((1 - \sqrt{\mu/L})^k)$ in the smooth and strongly convex case (Nesterov, 2003). These rates are then optimal among all methods that access gradients and linearly combine them (Nesterov, 2003; Nemirovskij and Yudin, 1983).

Nesterov acceleration introduces several sequences of iterates—two or three, depending on the formulations—and relies on a clever blend of gradient steps and mixing steps between the iterates. Many works contributed to interpret and motivate the precise structure of the iteration that lead to the success of the method, see for instance (Bubeck et al., 2015; Flammarion and Bach, 2015; Arjevani et al., 2016; Kim and Fessler, 2016; Allen-Zhu and Orecchia, 2017). A large number of these works found useful to study continuous time equivalents of Nesterov acceleration, obtained by taking the limit when stepsizes vanish, or from a variational framework. The continuous time index t of the limit allowed to use differential calculus to study the convergence of these equivalents. For examples of studies that use continuous time, see (Su et al., 2014; Krichene et al., 2015; Wilson et al., 2016; Wibisono et al., 2016; Betancourt et al., 2018; Diakonikolas and Orecchia, 2019; Shi et al., 2018, 2019; Attouch et al., 2018, 2019; Zhang et al., 2018; Siegel, 2019; Muehlebach and Jordan, 2019; Sanz-Serna and Zygalakis, 2020).

In this paper, we propose another way to obtain a continuous time equivalent of Nesterov acceleration, that we call the *continuized* version of Nesterov acceleration, that does not have

vanishing stepsizes. It is built by considering two variables $x_t, z_t \in \mathbb{R}^d$, $t \in \mathbb{R}_{\geqslant 0}$, that continuously mix following a linear ordinary differential equation (ODE), and that take gradient steps at random times $T_1, T_2, T_3, \dots$. Thus, in this modeling, mixing and gradient steps alternate randomly.

Thanks to the continuous index t and some stochastic calculus, one can differentiate averaged quantities (expectations) with respect to t. In particular, this leads to simple analytical expressions for the optimal parameters as functions of t, while the optimal parameters of Nesterov accelerations are defined by recurrence relations that are complicated to solve.

The discretization $\tilde{x}_k = x_{T_k}, \tilde{z}_k = z_{T_k}, k \in \mathbb{N}$, of the continuized process can be computed directly and exactly: the result is a recursion of the same form as Nesterov iteration, but with randomized parameters, that performs similarly to Nesterov original deterministic version both in theory and in simulations.

There are particular situations where Nesterov acceleration can not be implemented and the continuized acceleration can. First, a major advantage of the continuized acceleration over Nesterov acceleration is that the parameters of the algorithm depend only on time $t \in \mathbb{R}_{\geqslant 0}$, and not on the number of past gradient steps k. This is useful in distributed implementations, where the total number of gradient steps taken in the network may not be known to a particular node. Second, the continuized modeling can be relevant when gradient steps arrive at random times, as in asynchronous parallel computing for instance. Gossip algorithms represent another example where both features are present: the total number of past communications in the network at a given time is unknown to all nodes, and communication between nodes occur at random times. This motivated Even et al. (2020) to consider a similar continuized procedure, for communication steps instead of gradient steps, in order to accelerate gossip algorithms; their work is the source of inspiration for the present paper.

Beyond these particular situations, the continuized acceleration should be seen as a close approximation to Nesterov acceleration, that features both an insightful and convenient expression as a continuous time process and a direct implementation as a discrete iteration. We thus hope to contribute to the understanding of Nesterov acceleration. We believe that the continuized framework can be adapted to various settings and extensions of Nesterov acceleration; as an illustration of this statement, we study how the continuized acceleration behaves in the presence of additive noise on the gradients.

Notations. The index k always denotes a non-negative integer, while the indices t, s always denote non-negative reals.

Structure of the paper. In Section 2, we recall gradient descent and Nesterov acceleration, its choice of parameters, and its convergence rate as a function of the number of iterations k. In Section 3, we introduce our continuized variant of Nesterov acceleration, its choice of parameters and its convergence rate as functions of t. In Section 4, we show that the discretization of the continuized acceleration leads to an iteration of the same structure as Nesterov acceleration, with random parameters. We give the expressions for the parameters and the convergence rate in terms of the number of iterations k. Finally, in Section 5, we study the robustness of the continuized acceleration to additive noise.

2. Reminders on gradient descent and Nesterov acceleration

For the sake of comparison, let us first recall classical results of convex optimization. Consider the iterates of gradient descent with stepsize γ,

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

We have the following convergence of the function values $f(x_k)$, depending on whether the function f is (1) convex, or (2) strongly convex.

Theorem 1 (Convergence of gradient descent)**.** *Choose the stepsize $\gamma = 1/L$.*

(1) Then

$$f(x_k) - f(x_*) \leqslant \frac{2L\|x_0 - x_*\|^2}{k+4}\,.$$

(2) Assume further that f is μ-strongly convex, $\mu > 0$. Then

$$f(x_k) - f(x_*) \leqslant \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^k \|x_0 - x_*\|^2 .$$

These results (or similar bounds) can be found at many places in the literature; for instance the first bound is in (Nesterov, 2003, Corollary 2.1.2) and the second bound is a simple consequence of (Nesterov, 2003, Theorem 2.1.15). See also the recent book of Nesterov (2018).

To accelerate these rates of convergence, Nesterov introduced iterations of three sequences, parametrized by $\tau_k, \tau'_k, \gamma_k, \gamma'_k, k \geqslant 0$, of the form

$$y_k = x_k + \tau_k(z_k - x_k) , \tag{1}$$

$$x_{k+1} = y_k - \gamma_k \nabla f(y_k) , \tag{2}$$

$$z_{k+1} = z_k + \tau'_k(y_k - z_k) - \gamma'_k \nabla f(y_k) . \tag{3}$$

Depending on whether the function f is known to be (1) simply convex, or (2) strongly convex with a known strong convexity parameter, Nesterov gave choices of parameters leading to accelerated convergence rates.

Theorem 2 (Convergence of accelerated gradient descent)**.** *(1) Choose the parameters $\tau_k = 1 - \frac{A_k}{A_{k+1}}, \tau'_k = 0, \gamma_k = \frac{1}{L}, \gamma'_k = \frac{A_{k+1} - A_k}{L}, k \geqslant 0$, where the sequence $A_k, k \geqslant 0$, is defined by the recurrence relation*

$$A_0 = 0 , \qquad A_{k+1} = A_k + \frac{1}{2}(1 + \sqrt{4A_k + 1}) .$$

Then

$$f(x_k) - f(x_*) \leqslant \frac{2L\|x_0 - x_*\|^2}{k^2} .$$

(2) Assume further that f is μ-strongly convex, $\mu > 0$. Choose the constant parameters $\tau_k \equiv \frac{\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$, $\tau'_k \equiv \sqrt{\frac{\mu}{L}}$, $\gamma_k \equiv \frac{1}{L}$, $\gamma'_k \equiv \frac{1}{\sqrt{\mu L}}$, $k \geqslant 0$. Then

$$f(x_k) - f(x_*) \leqslant \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right)\left(1 - \sqrt{\frac{\mu}{L}}\right)^k .$$

This result, in this exact form, is proven by d'Aspremont et al. (2021, Sections 4.4.1 and 4.5.3).

From a high-level perspective, Nesterov acceleration iterates over several variables, alternating between gradient steps (always with respect to the gradient at y_k) and mixing steps, where the running value of a variable is replaced by a linear combination of the other variables. However, the precise way gradient and mixing steps are coupled is rather mysterious, and the success of the proof of Theorem 2 relies heavily on the detailed structure of the iterations. In the next section, we try to gain perspective on this structure by developing a continuized version of the acceleration.

3. Continuized version of Nesterov acceleration

In this section and the following ones, we use several mathematical notions related to random processes. It should be possible to understand the paper with only a heuristic understanding of these notions. The rigorous definitions are provided in Appendix A.

We argue that the accelerated iteration becomes more natural if we consider two variables x_t, z_t indexed by a continuous time $t \geqslant 0$, that are continuously mixing and that take gradient steps at random times. More precisely, let $T_1, T_2, T_3, \ldots \geqslant 0$ be random times such that $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are independent identically distributed (i.i.d.), of law exponential with rate 1 (any constant rate would do, but we choose 1 to make the comparison with discrete time k straightforward). By convention, we choose that our stochastic processes $t \mapsto x_t, t \mapsto z_t$ are càdlàg almost surely, i.e., right continuous with well-defined left-limits x_{t-}, z_{t-} (see Definition 5 in Appendix A). Our dynamics are parametrized by functions $\gamma_t, \gamma'_t, \tau_t, \tau'_t$, $t \geqslant 0$. At the random times $T_1, T_2, \ldots$, our sequences take gradient steps

$$x_{T_k} = x_{T_k-} - \gamma_{T_k} \nabla f(x_{T_k-}) , \tag{4}$$

$$z_{T_k} = z_{T_k-} - \gamma'_{T_k} \nabla f(x_{T_k-}) . \tag{5}$$

Because of the memoryless property of the exponential distribution, in a infinitesimal time interval $[t, t + \mathrm{d}t]$, the variables take gradients steps with probability $\mathrm{d}t$, independently of the past.

Between these random times, the variables mix through a linear ordinary differential equation (ODE)

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t\,, \tag{6}$$

$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t\,. \tag{7}$$

Following the notation of stochastic calculus, we can write the process more compactly in terms of the Poisson point measure $\mathrm{d}N(t) = \sum_{k\geqslant 0} \delta_{T_k}(\mathrm{d}t)$, which has intensity the Lebesgue measure $\mathrm{d}t$,

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t \nabla f(x_t)\mathrm{d}N(t)\,, \tag{8}$$

$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t - \gamma'_t \nabla f(x_t)\mathrm{d}N(t)\,. \tag{9}$$

Before giving convergence guarantees for such processes, let us digress quickly on why we can expect an iteration of this form to be mathematically appealing.

First, from a Markov chain indexed by a discrete time index k, one can associate the so-called *continuized* Markov chain, indexed by a continuous time t, that makes transition with the same Markov kernel, but at random times, with independent exponential time intervals (Aldous and Fill, 2002). Following this terminology, we refer to our acceleration (8)-(9) as the continuized acceleration. The continuized Markov chain is appreciated for its continuous time parameter t, while keeping many properties of the original Markov chain; similarly the continuized acceleration is arguably simpler to analyze, while performing similarly to Nesterov acceleration.

Second, it is also interesting to compare with coordinate gradient descent methods, that are easier to analyze when coordinates are selected randomly rather than in an ordered way (Wright, 2015). Similarly, the continuized acceleration is simpler to analyze because the gradient steps (4)-(5) and the mixing steps (6)-(7) alternate randomly, due to the randomness of $T_1, T_2, \ldots$

In analogy with Theorem 2, we give choices of parameters that lead to accelerated convergence rates, in the convex case (1) and in the strongly convex case (2). Convergence is analyzed as a function of t. As $\mathrm{d}N(t)$ is a Poisson point process with rate 1, t is the expected number of gradient steps done by the algorithm. Thus t is analoguous to k in Theorem 2.

Theorem 3 (Convergence of continuized Nesterov acceleration). *(1) Choose the parameters* $\eta_t = \frac{2}{t}, \eta'_t = 0, \gamma_t = \frac{1}{L}, \gamma'_t = \frac{t}{2L}$. *Then*

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{2L\|z_0 - x_*\|^2}{t^2}\,.$$

(2) Assume further that f is μ-strongly convex, $\mu > 0$. Choose the constant parameters $\eta_t = \eta'_t \equiv \sqrt{\frac{\mu}{L}}$, $\gamma_t \equiv \frac{1}{L}$, $\gamma'_t \equiv \frac{1}{\sqrt{\mu L}}$. *Then*

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right)\exp\left(-\sqrt{\frac{\mu}{L}}t\right)\,.$$

Sketch. A complete and rigorous proof is given in Appendix B.1. Here, we only provide the heuristic of the main lines of the proof.

The proof is similar to the one of Nesterov acceleration: we prove that for some choices of parameters $\eta_t, \eta'_t, \gamma_t, \gamma'_t$, $t \geqslant 0$, and for some functions A_t, B_t, $t \geqslant 0$,

$$\phi_t = A_t\left(f(x_t) - f(x_*)\right) + B_t\|z_t - x_*\|^2$$

is a supermartingale. In particular, this implies that $\mathbb{E}\phi_t$ is a Lyapunov function, i.e., a non-increasing function of t.

To prove that ϕ_t is a supermartingale, it is sufficient to prove that for all infinitesimal time intervals $[t, t + \mathrm{d}t]$, $\mathbb{E}_t\phi_{t+\mathrm{d}t} \leqslant \phi_t$, where $\mathbb{E}_t$ denotes the conditional expectation knowing all the past of the Poisson process up to time t. Thus we would like to compute the first order variation of $\mathbb{E}_t\phi_{t+\mathrm{d}t}$. This implies computing the first order variation of $\mathbb{E}_t f(x_{t+\mathrm{d}t})$.

From (8), we see that $f(x_t)$ evolves for two reasons between t and $t + \mathrm{d}t$:

- x_t follows the linear ODE (6), which results in the infinitesimal variation $f(x_t) \to f(x_t) + \eta_t\langle\nabla f(x_t), z_t - x_t\rangle\mathrm{d}t$, and

- with probability $\mathrm{d}t$, x_t takes a gradient step, which results in a macroscopic variation $f(x_t) \to f\left(x_t - \gamma_t \nabla f(x_t)\right)$.

Combining both variations, we obtain that

$$\mathbb{E}_t f(x_{t+\mathrm{d}t}) \approx f(x_t) + \eta_t \langle \nabla f(x_t), z_t - x_t \rangle \mathrm{d}t + \mathrm{d}t \left(f\left(x_t - \gamma_t \nabla f(x_t)\right) - f(x_t) \right) ,$$

where the $\mathrm{d}t$ in the second term corresponds to the probability that a gradient step happens; note that the latter event is independent of the past up to time t.

A similar computation can be done for $\mathbb{E}_t \|z_t - x_*\|^2$. Putting things together, we obtain

$$\begin{aligned}
\mathbb{E}_t \phi_{t+\mathrm{d}t} - \phi_t \approx \mathrm{d}t \Bigg(& \frac{\mathrm{d}A_t}{\mathrm{d}t} (f(x_t) - f(x_*)) + A_t \eta_t \langle \nabla f(x_t), z_t - x_t \rangle \\
& - A_t \left(f\left(x_t - \gamma_t \nabla f(x_t)\right) - f(x_t) \right) + \frac{\mathrm{d}B_t}{\mathrm{d}t} \|z_t - x_*\|^2 \\
& + 2 B_t \eta_t' \langle z_t - x_*, x_t - z_t \rangle + B_t \big(\|z_t - \gamma_t' \nabla f(x_t) - x_*\|^2 - \|z_t - x_*\|^2 \big) \Bigg) .
\end{aligned}$$

Using convexity and strong convexity inequalities, and a few computations, we obtain the following upper bound:

$$\begin{aligned}
\mathbb{E}_t \phi_{t+\mathrm{d}t} - \phi_t \lesssim \mathrm{d}t \Bigg(& \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t \eta_t \right) \langle \nabla f(x_t), x_t - x_* \rangle + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t \eta_t' \right) \|z_t - x_*\|^2 \\
& + (A_t \eta_t - 2 B_t \gamma_t') \langle \nabla f(x_t), z_t - x_* \rangle + \left(B_t \eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t} \frac{\mu}{2} \right) \|x_t - x_*\|^2 \\
& + \left(B_t \gamma_t'^2 - A_t \gamma_t \left(1 - \frac{L \gamma_t}{2} \right) \right) \|\nabla f(x_t)\|^2 \Bigg) .
\end{aligned}$$

We want this infinitesimal variation to be non-positive. Here, we choose the parameters so that $\gamma_t = 1/L$, and all prefactors in the above expression are zero. This gives some constraints on the choices of parameters. We show that only one degree of freedom is left: the choice of the function A_t, that must satisfy the ODE

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \left(\sqrt{A_t} \right) = \frac{\mu}{4L} \sqrt{A_t} ,$$

but whose initialization remains free. Once the initialization of the function A_t is chosen, this determines the full function A_t and, through the constraints, all parameters of the algorithm. As ϕ_t is a supermartingale (by design), a bound on the performance of the algorithm is given by

$$\mathbb{E} f(x_t) - f(x_*) \leqslant \frac{\mathbb{E}\phi_t}{A_t} \leqslant \frac{\phi_0}{A_t} .$$

The results presented in Theorem 3 correspond to one special choice of initialization for the function A_t.

In this sketch of proof, our derivation of the infinitesimal variation is intuitive and elementary; however it can be made more rigorous and concise—albeit more technical—using classical results from stochastic calculus, namely Proposition 2. This is our approach in Appendix B.1. □

Many authors have proposed continuous-time equivalents in order to understand better Nesterov acceleration using differential calculus, see the numerous references in the introduction. For instance, in the seminal work of Su et al. (2014), the equivalence is obtained from Nesterov acceleration by taking the joint asymptotic where the stepsizes vanish and the number of iterates is rescaled. The resulting limit is an ODE that must be discretized to be implemented; choosing the right discretization is not straightforward as it introduces stability and approximation errors that must be controlled, see (Zhang et al., 2018; Shi et al., 2019; Sanz-Serna and Zygalakis, 2020).

On the contrary, our continuous time equivalent (8)-(9) does not correspond to a limit where the stepsizes vanish. However, in Appendix D, we check that the continuized acceleration has the same ODE scaling limit as Nesterov acceleration. This sanity check emphasizes that the continuized acceleration is fundamentally different from previous continuous-time equivalents.

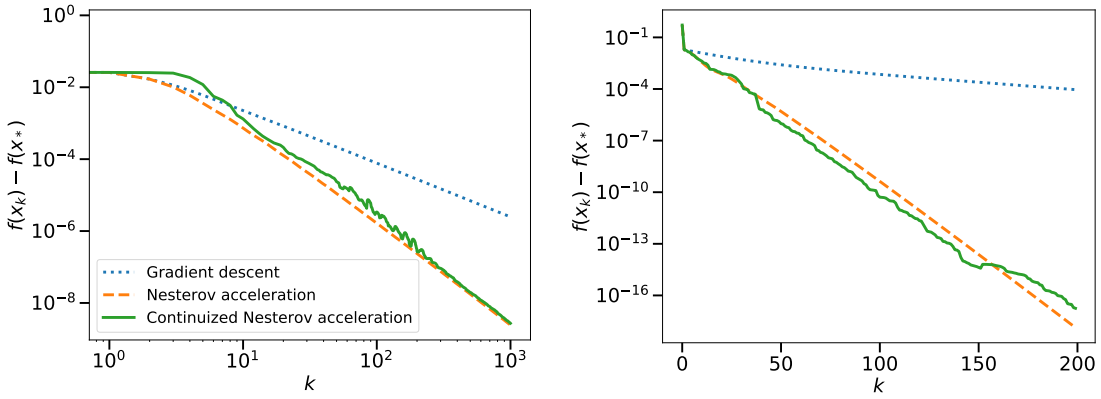

FIGURE 1. Comparison between gradient descent, Nesterov acceleration, and the continuized version of Nesterov acceleration, on a convex function (left) and a strongly convex function (right). For the continuized acceleration, which is randomized, the results shown corresponds to a single run. (Results were stable across runs.)

4. Discrete implementation of the continuized implementation with random parameters

In this section, we show that the continuized acceleration can be implemented exactly as a discrete algorithm. Denote

$$\tilde{x}_k = x_{T_k}\,, \qquad \tilde{y}_k = x_{T_{k+1}-}\,, \qquad \tilde{z}_k = z_{T_k}\,.$$

The three sequences $\tilde{x}_k, \tilde{y}_k, \tilde{z}_k$, $k \geqslant 0$, satisfy a recurrence relation of the same structure as Nesterov acceleration, but with random weights.

Theorem 4 (Discrete version of continuized acceleration)**.** *For any stochastic process of the form* (8)-(9)*, we have*

$$\tilde{y}_k = \tilde{x}_k + \tau_k(\tilde{z}_k - \tilde{x}_k)\,, \tag{10}$$
$$\tilde{x}_{k+1} = \tilde{y}_k - \tilde{\gamma}_k \nabla f(\tilde{y}_k)\,, \tag{11}$$
$$\tilde{z}_{k+1} = \tilde{z}_k + \tau'_k(\tilde{y}_k - \tilde{z}_k) - \tilde{\gamma}'_k \nabla f(\tilde{y}_k)\,, \tag{12}$$

for some random parameters $\tau_k, \tau'_k, \tilde{\gamma}_k, \tilde{\gamma}'_k$ (that are functions of $T_k, T_{k+1}, \eta_t, \eta'_t, \gamma_t, \gamma'_t$).

(1) For the parameters of Theorem 3.1, $\tau_k = 1 - \left(\frac{T_k}{T_{k+1}}\right)^2$, $\tau'_k = 0$, $\tilde{\gamma}_k = \frac{1}{L}$, and $\tilde{\gamma}'_k = \frac{T_k}{2L}$.

(2) For the parameters of Theorem 3.2, $\tau_k = \frac{1}{2}\left(1 - \exp\left(-2\sqrt{\frac{\mu}{L}}(T_{k+1} - T_k)\right)\right)$, $\tau'_k = \tanh\left(\sqrt{\frac{\mu}{L}}(T_{k+1} - T_k)\right)$, $\tilde{\gamma}_k = \frac{1}{L}$, and $\tilde{\gamma}'_k = \frac{1}{\sqrt{\mu L}}$.

This theorem is proved in Appendix C.

In Figure 1, we compare this continuized Nesterov acceleration (10)-(12) with the classical Nesterov acceleration (1)-(3) and gradient descent. In the strongly convex case (right), we run the algorithms with the parameters of Theorem 2.2 and 4.2 on the function

$$f(x_1, x_2, x_3) = \frac{\mu}{2}(x_1 - 1)^2 + \frac{3\mu}{2}(x_2 - 1)^2 + \frac{L}{2}(x_3 - 1)^2\,,$$

with $\mu = 10^{-2}$ and $L = 1$. In the convex case, we run the algorithms with the parameters of Theorem 2.1 and 4.1 on the function

$$f(x_1, \ldots, x_{100}) = \frac{1}{2}\sum_{i=1}^{100} \frac{1}{i^2}\left(x_i - \frac{1}{i}\right)^2\,,$$

which has negligible strong convexity parameter. All iterations were initialized from $x_0 = z_0 = 0$.

In order to have a straightforward theoretical comparison with Nesterov acceleration, we describe the performance $f(\tilde{x}_k) - f(x_*) = f(x_{T_k}) - f(x_*)$ of the continuized acceleration in terms of the number k of gradient operations.

Theorem 5 (Convergence of the discretized version)**.** *The discrete implementation* (10)-(12)*, with random weights, of the continuized acceleration, satisfies:*

(1) For the parameters of Theorem 4.1,

$$\mathbb{E}\left[T_k^2\left(f(\tilde{x}_k) - f(x_*)\right)\right] \leqslant 2L\|z_0 - x_*\|^2 .$$

(2) Assume further that f is μ-strongly convex, $\mu > 0$. For the parameters of Theorem 4.2,

$$\mathbb{E}\left[\exp\left(\sqrt{\frac{\mu}{L}}T_k\right)\left(f(\tilde{x}_k) - f(x_*)\right)\right] \leqslant f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2 .$$

This theorem is proved in Appendix B.1. The law of T_k is well known: it is the sum of k i.i.d. random variables of law exponential with rate 1; this is called an Erlang or Gamma distribution with shape parameter k and rate 1. One can use well-known properties of this law, such as its concentration around its expectation $\mathbb{E}T_k = k$, to derive corollaries of Theorem 5.

5. Robustness of the continuized Nesterov acceleration to additive noise

We now investigate how the continuized version of Nesterov acceleration performs under stochastic noise. We should emphasize that a similar study has been done on Nesterov acceleration directly (Lan, 2012; Hu et al., 2009; Xiao, 2010; Devolder, 2011; Cohen et al., 2018; Aybat et al., 2020). However, in the continuized framework, the randomness of the stochastic gradient and its time mix in a particularly convenient way.

We assume that we do not have direct access to the gradient $\nabla f(x)$ but to a random estimate $\nabla f(x,\xi)$, where $\xi \in \Xi$ is random of law $\mathcal{P}$. We assume that our estimate is unbiased, i.e.,

$$\forall x \in \mathbb{R}^d , \qquad \mathbb{E}_\xi \nabla f(x,\xi) = \nabla f(x) , \tag{13}$$

and has a uniformly bounded variance, i.e., there exists $\sigma^2 \geqslant 0$ such that

$$\forall x \in \mathbb{R}^d , \qquad \mathbb{E}_\xi \left\|\nabla f(x,\xi) - \nabla f(x)\right\|^2 \leqslant \sigma^2 . \tag{14}$$

These assumptions typically hold in the additive noise model, where $\nabla f(x,\xi) = \nabla f(x) + \xi$, where $\xi \in \mathbb{R}^d$ is satisfies $\mathbb{E}\xi = 0$, $\mathbb{E}\|\xi\|^2 \leqslant \sigma^2$. By an abuse of terminology, we say that our stochastic gradients have "additive noise" when (13) and (14) hold.

We keep the same algorithms, replacing gradients by stochastic gradients. Let $\xi_1, \xi_2, \dots$ be i.i.d. random variables of law $\mathcal{P}$. We take stochastic gradient steps at the random times $T_1, T_2, \dots$,

$$\begin{aligned} x_{T_k} &= x_{T_k-} - \gamma_{T_k}\nabla f(x_{T_k-}, \xi_k) , \\ z_{T_k} &= z_{T_k-} - \gamma'_{T_k}\nabla f(x_{T_k-}, \xi_k) . \end{aligned}$$

Between these random times, the variables mix through the same ODE

$$\begin{aligned} \mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t , \\ \mathrm{d}z_t &= \eta'_t(x_t - z_t)\mathrm{d}t . \end{aligned}$$

This can be written more compactly in terms of the Poisson point measure $\mathrm{d}N(t,\xi) = \sum_{k\geqslant 0}\delta_{(T_k,\xi_k)}(\mathrm{d}t, \mathrm{d}\xi)$ on $\mathbb{R}_{\geqslant 0}\times\Xi$, which has intensity $\mathrm{d}t\otimes\mathcal{P}$,

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t\int_\Xi \nabla f(x_t,\xi)\mathrm{d}N(t,\xi) , \tag{15}$$

$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t - \gamma'_t\int_\Xi \nabla f(x_t,\xi)\mathrm{d}N(t,\xi) . \tag{16}$$

Theorem 6 (Continuized acceleration with noise)**.** *Assume that the stochastic gradients are unbiased* (13) *and have a variance uniformly bounded by σ^2* (14)*. Then the continuized acceleration* (15)-(16) *satisfies the following.*

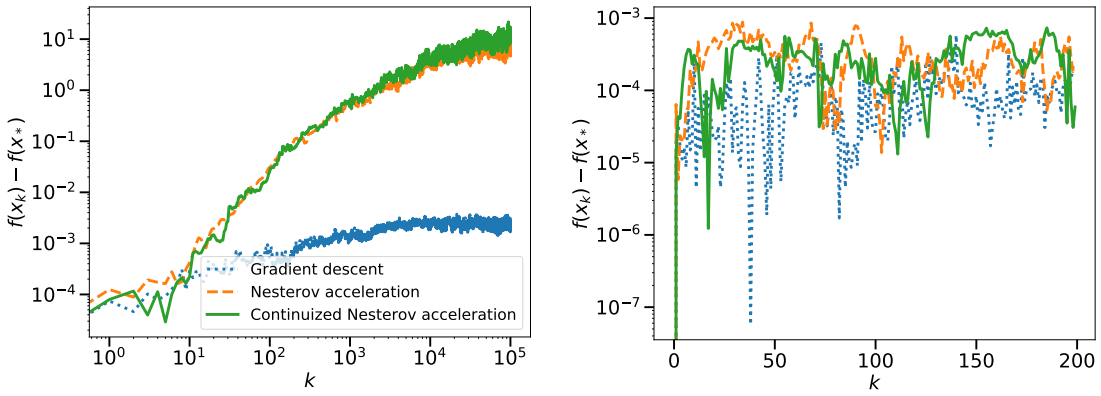

FIGURE 2. Effect of additive noise on gradient descent, Nesterov acceleration, and the continuized version of Nesterov acceleration, on a convex function (left) and a strongly convex function (right). The results shown corresponds to a single run. (Results were stable across runs.)

(1) For the parameters of Theorem 3.1,

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{2L\|z_0 - x_*\|^2}{t^2} + \sigma^2 \frac{t}{3L}\,.$$

(2) Assume further that f is μ-strongly convex, $\mu > 0$. For the parameters of Theorem 3.2,

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right) \exp\left(-\sqrt{\frac{\mu}{L}}t\right) + \sigma^2 \frac{1}{\sqrt{\mu L}}\,.$$

This theorem is proved in Appendix B.2.

In the above bounds, L is a parameter of the algorithm, that can be taken greater than the best known smoothness constant of the function f. Increasing L reduces the stepsizes of the algorithm and performs some variance reduction. If the bound σ^2 on the variance is known, one can choose L optimizing the above bounds in order to obtain algorithms that adapt to additive noise.

In Figure 2, we run the same simulations as in Figure 1, with two differences: (1) we add isotropic Gaussian noise on the gradients, with covariance $10^{-4}\,\mathrm{Id}$, and (2) we initialized algorithms at the optimum, i.e., $x_0 = z_0 = x_*$. Initializing at the optimum enables to isolate the effect of the additive noise only. These simulations confirm Theorem 6: the noise term is (sub-)linearly increasing in the convex case and constant in the strongly convex case.

Note that similarly to Theorem 5, one could obtain convergence bounds for the discrete implementation under the presence of additive noise.

6. Conclusion

In this work, we introduced a continuized version of Nesterov's accelerated gradients. In a nutshell, the method has two sequences of iterates from which gradient steps are taken at random times. In between gradient steps, the two sequences mix following a simple ordinary differential equation, whose parameters are picked for ensuring good convergence properties of the method.

As compared to other continuous time models of Nesterov acceleration, a key feature of this approach is that the method can be implemented without any approximation step, as the differential equation governing the mixing procedure has a simple analytical solution. When discretized, the continuized method corresponds to an accelerated gradient method with random parameters.

Continuization strategies were introduced in the context of Markov chains (Aldous and Fill, 2002). Here, they allow using acceleration mechanisms in asynchronous distributed optimization, where agents are usually not aware of total the number of iterations taken so far, as showcased in the context of asynchronous gossip algorithms by Even et al. (2020). Possible future research directions include extending to constrained and non-Euclidean settings.

Acknowledgements

This work was funded in part by the French government under management of 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) and from the DGA.

References

Aldous, D. and Fill, J. A. (2002). Reversible markov chains and random walks on graphs. Unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/$\sim$aldous/RWG/book.html`.

Allen-Zhu, Z. and Orecchia, L. (2017). Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS '17.

Arjevani, Y., Shalev-Shwartz, S., and Shamir, O. (2016). On lower and upper bounds in smooth and strongly convex optimization. *Journal of Machine Learning Research*, 17(126):1–51.

Attouch, H., Chbani, Z., Peypouquet, J., and Redont, P. (2018). Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1):123–175.

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

Aybat, N. S., Fallah, A., Gurbuzbalaban, M., and Ozdaglar, A. (2020). Robust accelerated gradient methods for smooth strongly convex functions. *SIAM Journal on Optimization*, 30(1):717–751.

Betancourt, M., Jordan, M., and Wilson, A. (2018). On symplectic optimization. *arXiv preprint arXiv:1802.03653*.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Bubeck, S., Lee, Y. T., and Singh, M. (2015). A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*.

Cohen, M., Diakonikolas, J., and Orecchia, L. (2018). On acceleration with noise-corrupted gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR.

d'Aspremont, A., Scieur, D., and Taylor, A. (2021). Acceleration methods.

Devolder, O. (2011). Stochastic first order methods in smooth convex optimization. Technical report, CORE.

Diakonikolas, J. and Orecchia, L. (2019). The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689.

Even, M., Hendrikx, H., and Massoulié, L. (2020). Asynchrony and acceleration in gossip algorithms. *arXiv preprint arXiv:2011.02379*.

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

Hu, C., Pan, W., and Kwok, J. (2009). Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 781–789.

Ikeda, N. and Watanabe, S. (2014). *Stochastic differential equations and diffusion processes*. Elsevier.

Jacod, J. and Shiryaev, A. (2013). *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media.

Kim, D. and Fessler, J. A. (2016). Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1):81–107.

Krichene, W., Bayen, A., and Bartlett, P. (2015). Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems*, 28:2845–2853.

Lan, G. (2012). An optimal method for stochastic composite optimization. *Math. Program.*, 133(1-2, Ser. A):365–397.

Le Gall, J.-F. (2016). *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274. Springer.

Muehlebach, M. and Jordan, M. (2019). A dynamical systems perspective on Nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR.

Nemirovskij, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 27(2):372–376.

Nesterov, Y. (2003). *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media.

Nesterov, Y. (2018). *Lectures on Convex Optimization*, volume 137. Springer.

Sanz-Serna, J. M. and Zygalakis, K. (2020). The connections between Lyapunov functions for some optimization algorithms and differential equations. *arXiv preprint arXiv:2009.00673*.

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

Shi, B., Du, S., Su, W., and Jordan, M. (2019). Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, volume 32, pages 5744–5752.

Siegel, J. W. (2019). Accelerated first-order methods: Differential equations and Lyapunov functions. *arXiv preprint arXiv:1903.05671*.

Su, W., Boyd, S., and Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27:2510–2518.

Wibisono, A., Wilson, A. C., and Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358.

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

Wright, S. (2015). Coordinate descent algorithms. *Math. Program.*, 151(1, Ser. B):3–34.

Xiao, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596.

Zhang, J., Mokhtari, A., Sra, S., and Jadbabaie, A. (2018). Direct Runge-Kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems*, volume 31, pages 3900–3909.

Appendix A. Stochastic calculus toolbox

In this appendix, we give a short introduction to the mathematical tools that we use in this paper. For more details, the reader can consult the more rigorous monographs of Jacod and Shiryaev (2013); Ikeda and Watanabe (2014); Le Gall (2016).

A.1. Poisson point measures. We fix $\mathcal{P}$ a probability law on some space Ξ.

Definition 1. *A* (homogenous) Poisson point measure *on $\mathbb{R}_{\geqslant 0} \times \Xi$, with intensity $\nu(\mathrm{d}t, \mathrm{d}\xi) = \mathrm{d}t \otimes \mathrm{d}\mathcal{P}(\xi)$, is a random measure N on $\mathbb{R}_{\geqslant 0} \times \Xi$ such that*

- *For any disjoint measurable subsets A and B of $\mathbb{R}_{\geqslant 0} \times \Xi$, $N(A)$ and $N(B)$ are independent.*
- *For any measurable subset A of $\mathbb{R}_{\geqslant 0} \times \Xi$, $N(A)$ is a Poisson random variable with parameter $\nu(A)$. (If $\nu(A) = \infty$, $N(A)$ is equal to ∞ almost surely.)*

Proposition 1. *Let N be a Poisson point measure on $\mathbb{R}_{\geqslant 0} \times \Xi$ with intensity $\mathrm{d}t \otimes \mathrm{d}\mathcal{P}(\xi)$.*

There exists a decomposition $\mathrm{d}N(t, \xi) = \sum_{k \geqslant 0} \delta_{(T_k, \xi_k)}(\mathrm{d}t, \mathrm{d}\xi)$ on $\mathbb{R}_{\geqslant 0} \times \Xi$ where $0 < T_1 < T_2 < T_3 < \dots$ and $\xi_1, \xi_2, \xi_3, \dots \in \Xi$ satisfy:

- *$T_1, T_2 - T_1, T_3 - T_2, \dots$ are i.i.d. of law exponential with rate 1,*
- *$\xi_1, \xi_2, \xi_3, \dots$ are i.i.d. of law $\mathcal{P}$ and independent of the $T_1, T_2, T_3, \dots$.*

Definition 2. *Let N be a Poisson point measure on $\mathbb{R}_{\geqslant 0} \times \Xi$ with intensity $\mathrm{d}t \otimes \mathrm{d}\mathcal{P}(\xi)$. The* filtration *$\mathcal{F}_t$, $t \geqslant 0$,* generated by *N is defined by the formula*

$$\mathcal{F}_t = \sigma\left(N([0, s] \times A)\,,\ s \leqslant t, A \subset \Xi \text{ measurable}\right)\,.$$

A.2. Martingales and supermartingales. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_t$, $t \geqslant 0$, a filtration on this probability space.

Definition 3. *A random process $x_t \in \mathbb{R}^d$, $t \geqslant 0$, is* adapted *if for all $t \geqslant 0$, x_t is $\mathcal{F}_t$-measurable. An adapted process $x_t \in \mathbb{R}$, $t \geqslant 0$ is a* martingale *(resp.* supermartingale*) if for all $0 \leqslant s \leqslant t$, $\mathbb{E}[x_t |\ \mathcal{F}_s] = x_s$ (resp. $\mathbb{E}[x_t |\ \mathcal{F}_s] \leqslant x_s$).*

Definition 4. *A random variable $T \in [0, \infty]$ is a* stopping time *if for all $t \geqslant 0$, $\{T \leqslant t\} \in \mathcal{F}_t$.*

Definition 5. *A function $x_t, t \geqslant 0$, is said to be* càdlàg *if it is right continuous and for every $t > 0$, the limit $x_{t-} := \lim_{s \to t, s < t} x_s$ exists and is finite.*

Theorem 7 (Martingale stopping theorem)**.** *Let x_t, $t \geqslant 0$, be a martingale (resp. supermartingale) with càdlàg trajectories and uniformly integrable. Let T be a stopping time. Then $\mathbb{E}X_T = X_0$ (resp. $\mathbb{E}X_T \leqslant X_0$).*

A.3. Stochastic ordinary differential equation with Poisson jumps. We fix $\mathcal{P}$ a probability law on some space Ξ, N a Poisson point measure on $\mathbb{R}_{\geqslant 0} \times \Xi$ with intensity $\mathrm{d}t \otimes \mathrm{d}\mathcal{P}(\xi)$, and denote $\mathcal{F}_t$, $t \geqslant 0$, the filtration generated by N.

Definition 6. *Let $b : \mathbb{R}^d \to \mathbb{R}^d$ and $G : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ be two functions. An random process $x_t \in \mathbb{R}^d$, $t \geqslant 0$, is said to be a solution of the equation*

$$\mathrm{d}x_t = b(x_t)\mathrm{d}t + \int_\Xi G(x_t, \xi)\mathrm{d}N(t, \xi)$$

if it is adapted, càdlàg, and for all $t \geqslant 0$,

$$x_t = x_0 + \int_0^t b(x_s)\mathrm{d}s + \int_{[0,t] \times \Xi} G(x_{s-}, \xi)\mathrm{d}N(s, \xi)\,.$$

If we consider the decomposition $\mathrm{d}N(t, \xi) = \sum_{k \geqslant 0} \delta_{(T_k, \xi_k)}(\mathrm{d}t, \mathrm{d}\xi)$ given by Proposition 1, then

$$\int_{[0,t] \times \Xi} G(x_{s-}, \xi)\mathrm{d}N(s, \xi) = \sum_{k \geqslant 0} \mathbf{1}_{\{T_k \leqslant t\}} G(x_{T_k -}, \xi_k)\,.$$

Proposition 2. *Let $x_t \in \mathbb{R}^d$ be a solution of*

$$\mathrm{d}x_t = b(x_t)\mathrm{d}t + \int_\Xi G(x_t, \xi)\mathrm{d}N(t, \xi)$$

and $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Then

$$\varphi(x_t) = \varphi(x_0) + \int_0^t \langle \nabla\varphi(x_s), b(x_s)\rangle \mathrm{d}s + \int_{[0,t]\times\Xi} (\varphi(x_{s-} + G(x_{s-},\xi)) - \varphi(x_{s-}))\, \mathrm{d}N(s,\xi)\,.$$

Moreover, we have the decomposition

$$\begin{aligned}\int_{[0,t]\times\Xi} &(\varphi(x_{s-} + G(x_{s-},\xi)) - \varphi(x_{s-}))\, \mathrm{d}N(s,\xi) \\ &= \int_0^t \int_\Xi (\varphi(x_s + G(x_s,\xi)) - \varphi(x_s))\, \mathrm{d}t\mathrm{d}\mathcal{P}(\xi) + M_t\,,\end{aligned}$$

where $M_t = \int_{[0,t]\times\Xi} (\varphi(x_{s-} + G(x_{s-},\xi)) - \varphi(x_{s-}))\, (\mathrm{d}N(s,\xi) - \mathrm{d}t\mathrm{d}\mathcal{P}(\xi))$ is a martingale.

This proposition is an elementary calculus of variations formula: to compute the value of the observable $\varphi(x_t)$, one must sum the effects of the continuous part and of the Poisson jumps. Moreover, the integral with respect to the Poisson measure N becomes a martingale if the same integral with respect to its intensity measure $\mathrm{d}t \otimes \mathrm{d}\mathcal{P}(\xi)$ is removed.

Appendix B. Analysis of the continuized Nesterov acceleration

To encompass the proofs in the convex and in the strongly convex cases in a unified way, we assume f is μ-strongly convex, $\mu \geqslant 0$. If $\mu > 0$, this corresponds to assuming the μ-strong convexity in the usual sense; if $\mu = 0$, it means that we only assume the function to be convex. In other words, the proofs in the convex case can be obtained by taking $\mu = 0$ below.

In this section, $\mathcal{F}_t$, $t \geqslant 0$, is the filtration associated to the Poisson point measure N.

B.1. Noiseless case: proofs of Theorems 3 and 5. In this section, we analyze the convergence of the continuized iteration (8)-(9), that we recall for the reader's convenience:

$$\begin{aligned}\mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t \nabla f(x_t)\mathrm{d}N(t)\,,\\ \mathrm{d}z_t &= \eta'_t(x_t - z_t)\mathrm{d}t - \gamma'_t \nabla f(x_t)\mathrm{d}N(t)\,.\end{aligned}$$

The choices of parameters $\eta_t, \eta'_t, \gamma_t, \gamma'_t$, $t \geqslant 0$, and the corresponding convergence bounds follow naturally from the analysis. We seek sufficient conditions under which the function

$$\phi_t = A_t\,(f(x_t) - f_*) + B_t \|z_t - x_*\|^2$$

is a supermartingale.

The process $\bar{x}_t = (t, x_t, z_t)$ satisfies the equation

$$\mathrm{d}\bar{x}_t = b(\bar{x}_t)\mathrm{d}t + G(\bar{x}_t)\mathrm{d}N(t)\,, \qquad b(\bar{x}_t) = \begin{pmatrix} 1 \\ \eta_t(z_t - x_t) \\ \eta'_t(x_t - z_t)\end{pmatrix}\,, \qquad G(\bar{x}_t) = \begin{pmatrix} 0 \\ -\gamma_t\nabla f(x_t) \\ -\gamma'_t \nabla f(x_t)\end{pmatrix}\,.$$

We thus apply Proposition 2 to $\phi_t = \varphi(\bar{x}_t) = \varphi(t, x_t, z_t)$ where

$$\varphi(t,x,z) = A_t\,(f(x) - f(x_*)) + B_t\|z - x_*\|^2\,,$$

we obtain:

$$\phi_t = \phi_0 + \int_0^t \langle \nabla\varphi(\bar{x}_s), b(\bar{x}_s)\rangle \mathrm{d}s + \int_0^t (\varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s))\, \mathrm{d}s + M_t\,,$$

where M_t is a martingale. Thus, to show that φ_t is a supermartingale, it is sufficient to show that the map $t \mapsto \int_0^t \langle \nabla\varphi(\bar{x}_s), b(\bar{x}_s)\rangle \mathrm{d}s + \int_0^t (\varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s)))\, \mathrm{d}s$ is non-increasing almost surely, i.e.,

$$I_t := \langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle + \varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) \leqslant 0\,.$$

We now compute

$$\begin{aligned}\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle &= \partial_t\varphi(\bar{x}_t) + \langle \partial_x\varphi(\bar{x}_t), \eta_t(z_t - x_t)\rangle + \langle \partial_z\varphi(\bar{x}_t), \eta'_t(x_t - z_t)\rangle \\ &= \frac{\mathrm{d}A_t}{\mathrm{d}t}\,(f(x_t) - f(x_*)) + \frac{\mathrm{d}B_t}{\mathrm{d}t}\|z_t - x_*\|^2 + A_t\eta_t\langle \nabla f(x_t), z_t - x_t\rangle \\ &\quad + 2B_t\eta'_t\langle z_t - x_*, x_t - z_t\rangle\,.\end{aligned}$$

Here, we use that as f is μ-strongly convex,

$$f(x_t) - f(x_*) \leqslant \langle \nabla f(x_t), x_t - x_* \rangle - \frac{\mu}{2}\|x_t - x_*\|^2 \,,$$

and the simple bound

$$\begin{aligned}\langle z_t - x_*, x_t - z_t\rangle &= \langle z_t - x_*, x_t - x_*\rangle - \|z_t - x_*\|^2 \leqslant \|z_t - x_*\|\|x_t - x_*\| - \|z_t - x_*\|^2 \\ &\leqslant \frac{1}{2}\left(\|z_t - x_*\|^2 + \|x_t - x_*\|^2\right) - \|z_t - x_*\|^2 = \frac{1}{2}\left(\|x_t - x_*\|^2 - \|z_t - x_*\|^2\right) .\end{aligned}$$

This gives

$$\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle \leqslant \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\langle \nabla f(x_t), x_t - x_*\rangle + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\frac{\mu}{2}\right)\|x_t - x_*\|^2 \tag{17}$$

$$+ \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\|z_t - x_*\|^2 + A_t\eta_t\langle \nabla f(x_t), z_t - x_*\rangle \,. \tag{18}$$

Further,

$$\begin{aligned}\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) &= A_t\left(f(x_t - \gamma_t\nabla f(x_t)) - f(x_t)\right) \\ &\quad + B_t\left(\|(z_t - x_*) - \gamma_t'\nabla f(x_t)\|^2 - \|z_t - x_*\|^2\right) .\end{aligned}$$

As f is L-smooth,

$$\begin{aligned}f(x_t - \gamma_t\nabla f(x_t)) - f(x_t) &\leqslant \langle \nabla f(x_t), -\gamma_t\nabla f(x_t)\rangle + \frac{L}{2}\|\gamma_t\nabla f(x_t)\|^2 \\ &= -\gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\|\nabla f(x_t)\|^2 \,.\end{aligned}$$

This gives

$$\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) \leqslant \left(B_t\gamma_t'^2 - A_t\gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\right)\|\nabla f(x_t)\|^2 - 2B_t\gamma_t'\langle \nabla f(x_t), z_t - x_*\rangle \,. \tag{19}$$

Finally, combining (17)-(18) with (19), we obtain

$$I_t \leqslant \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\langle \nabla f(x_t), x_t - x_*\rangle + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\|z_t - x_*\|^2 \tag{20}$$

$$+ (A_t\eta_t - 2B_t\gamma_t')\langle \nabla f(x_t), z_t - x_*\rangle + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\frac{\mu}{2}\right)\|x_t - x_*\|^2 \tag{21}$$

$$+ \left(B_t\gamma_t'^2 - A_t\gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\right)\|\nabla f(x_t)\|^2 \,. \tag{22}$$

Remember that $I_t \leqslant 0$ is a sufficient condition for ϕ_t to be a supermartingale. Here, we choose the parameters $\eta_t, \eta_t', \gamma_t, \gamma_t', t \geqslant 0$, so that all prefactors are 0. We start by taking $\gamma_t \equiv \frac{1}{L}$ (other choices $\gamma_t < \frac{2}{L}$ could be possible but would give similar results) and we want to satisfy

$$\frac{\mathrm{d}A_t}{\mathrm{d}t} = A_t\eta_t \,, \qquad \frac{\mathrm{d}B_t}{\mathrm{d}t} = B_t\eta_t' \qquad A_t\eta_t = 2B_t\gamma_t' \,, \qquad B_t\eta_t' = \frac{\mathrm{d}A_t}{\mathrm{d}t}\frac{\mu}{2} \,, \qquad B_t\gamma_t'^2 = \frac{A_t}{2L} \,.$$

To satisfy the last equation, we choose

$$\gamma_t' = \sqrt{\frac{A_t}{2LB_t}} \,. \tag{23}$$

To satisfy the third equation, we choose

$$\eta_t = \frac{2B_t\gamma_t'}{A_t} = \sqrt{\frac{2B_t}{LA_t}} \,. \tag{24}$$

To satisfy the fourth equation, we choose

$$\eta_t' = \frac{\mathrm{d}A_t}{\mathrm{d}t}\frac{\mu}{2B_t} = \frac{A_t\eta_t\mu}{2B_t} = \mu\sqrt{\frac{A_t}{2LB_t}} \,. \tag{25}$$

Having now all parameters $\eta_t, \eta_t', \gamma_t, \gamma_t'$ constrained, we now have that ϕ_t is Lyapunov if

$$\frac{\mathrm{d}A_t}{\mathrm{d}t} = A_t\eta_t = \sqrt{\frac{2A_tB_t}{L}} \,, \qquad\qquad \frac{\mathrm{d}B_t}{\mathrm{d}t} = B_t\eta_t' = \mu\sqrt{\frac{A_tB_t}{2L}} \,.$$

This only leaves the choice of the initialization (A_0, B_0) as free: both the algorithm and the Lyapunov depend on it. (Actually, only the relative value A_0/B_0 matters.) Instead of solving the above system of two coupled non-linear ODEs, it is convenient to turn them into a single second-order linear ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right) = \frac{1}{2\sqrt{A_t}}\frac{\mathrm{d}A_t}{\mathrm{d}t} = \sqrt{\frac{B_t}{2L}}\,, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{B_t}\right) = \frac{1}{2\sqrt{B_t}}\frac{\mathrm{d}B_t}{\mathrm{d}t} = \mu\sqrt{\frac{A_t}{8L}}\,. \tag{26}$$

This can also be restated as

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\sqrt{A_t}\right) = \frac{\mu}{4L}\sqrt{A_t}\,, \qquad \sqrt{B_t} = \sqrt{2L}\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right)\,. \tag{27}$$

B.1.1. *Proof of the first part (convex case).* We now assume $\mu = 0$, and we choose the solution such that $A_0 = 0$ and $B_0 = 1$. From (26), we have $\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{B_t}\right) = 0$, thus $B_t \equiv 1$, and $\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right) = \frac{1}{\sqrt{2L}}$, thus $\sqrt{A_t} = \frac{t}{\sqrt{2L}}$. The parameters of the algorithm are given by (23)-(25): $\eta_t = \frac{2}{t}$, $\eta_t' = 0$, $\gamma_t' = \frac{t}{\sqrt{2L}}$ (and we had chosen $\gamma_t = \frac{1}{L}$).

From the fact that ϕ_t is a supermartingale, we obtain that the associated algorithm satisfies

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{\mathbb{E}\phi_t}{A_t} \leqslant \frac{\phi_0}{A_t} = \frac{2L\|z_0 - x_*\|^2}{t^2}\,.$$

This proves the first part of Theorem 3.

Further, one can apply martingale stopping Theorem 7 to the supermartingale ϕ_t with the stopping time T_k to obtain

$$\mathbb{E}\left[A_{T_k}\left(f(\tilde{x}_k) - f(x_*)\right)\right] = \mathbb{E}\left[A_{T_k}\left(f(x_{T_k}) - f(x_*)\right)\right] \leqslant \mathbb{E}\phi_{T_k} \leqslant \phi_0 = \|z_0 - x_*\|^2\,.$$

This proves the first part of Theorem 5.

B.1.2. *Proof of the second part (strongly convex case).* We now assume $\mu > 0$. We consider the solution of (27) that is exponential:

$$\sqrt{A_t} = \sqrt{A_0}\exp\left(\frac{1}{2}\sqrt{\frac{\mu}{L}}t\right)\,, \qquad \sqrt{B_t} = \sqrt{A_0}\sqrt{\frac{\mu}{2}}\exp\left(\frac{1}{2}\sqrt{\frac{\mu}{L}}t\right)\,.$$

The parameters of the algorithm are given by (23)-(25): $\eta_t = \eta_t' = \sqrt{\frac{\mu}{L}}$, $\gamma_t' = \frac{1}{\sqrt{\mu L}}$ (and we had chosen $\gamma_t = \frac{1}{L}$).

From the fact that ϕ_t is a supermartingale, we obtain that the associated algorithm satisfies

$$\begin{aligned}\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{\mathbb{E}\phi_t}{A_t} \leqslant \frac{\phi_0}{A_t} &= \frac{A_0(f(x_0) - f(x_*)) + A_0\frac{\mu}{2}\|z_0 - x_*\|^2}{A_t} \\ &= \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right)\exp\left(-\sqrt{\frac{\mu}{L}}t\right)\,.\end{aligned}$$

This proves the second part of Theorem 3. Similarly to above, one can also apply the martingale stopping theorem to prove the second part of Theorem 5.

Remark 1. *In the above derivation, in both the convex and strongly convex cases, we choose a particular solution of* (27)*, while several solutions are possible. In the convex case, we make the choice $A_0 = 0$ to have a succinct bound that does not depend on $f(x_0) - f(x_*)$. More importantly, in the strongly convex case, we choose the solution that satisfies the relation $\sqrt{\frac{\mu}{2}}\sqrt{A_t} = \sqrt{B_t}$, which implies that $\eta_t, \eta_t', \gamma_t'$, are constant functions of t, and $\eta_t = \eta_t'$. These conditions help solving in closed form the continuous part of the process*

$$\begin{aligned}\mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t\,, \\ \mathrm{d}z_t &= \eta_t'(x_t - z_t)\mathrm{d}t\,,\end{aligned}$$

which is crucial if we want to have a discrete implementation of our method (for more details, see Theorem 4 and its proof). However, in the strongly convex case, considering other solutions would be interesting, for instance to have an algorithm converging to the convex one as $\mu \to 0$.

B.2. With additive noise: proof of Theorem 6. The proof of this theorem is along the same lines as the proof of Theorem 3 above. Here, we only give the major differences.

We analyze the convergence of the continuized stochastic iteration (15)-(16), that we recall for the reader's convenience:

$$\begin{aligned}\mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t \int_\Xi \nabla f(x_t,\xi)\mathrm{d}N(t,\xi)\,,\\ \mathrm{d}z_t &= \eta'_t(x_t - z_t)\mathrm{d}t - \gamma'_t \int_\Xi \nabla f(x_t,\xi)\mathrm{d}N(t,\xi)\,.\end{aligned}$$

In this setting, we loose the property that

$$\phi_t = A_t\left(f(x_t) - f_*\right) + B_t\|z_t - x_*\|^2$$

is a supermartingale. However, we bound the increase of ϕ_t.

The process $\bar{x}_t = (t, x_t, z_t)$ satisfies the equation

$$\mathrm{d}\bar{x}_t = b(\bar{x}_t)\mathrm{d}t + \int_\Xi G(\bar{x}_t,\xi)\mathrm{d}N(t,\xi), \quad b(\bar{x}_t) = \begin{pmatrix}1\\ \eta_t(z_t - x_t)\\ \eta'_t(x_t - z_t)\end{pmatrix}, \quad G(\bar{x}_t,\xi) = \begin{pmatrix}0\\ -\gamma_t\nabla f(x_t,\xi)\\ -\gamma'_t\nabla f(x_t,\xi)\end{pmatrix}.$$

We apply Proposition 2 to $\phi_t = \varphi(\bar{x}_t) = \varphi(t, x_t, z_t)$ and obtain

$$\phi_t = \phi_0 + \int_0^t I_s\mathrm{d}s + M_t\,, \tag{28}$$

where M_t is a martingale and

$$I_t = \langle\nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle + \mathbb{E}_\xi\varphi(\bar{x}_t + G(\bar{x}_t,\xi)) - \varphi(\bar{x}_t)\,.$$

The computation of the first term remains the same: the inequality (17)-(18) holds. The computation of the second term becomes

$$\begin{aligned}\mathbb{E}_\xi\varphi(\bar{x}_t + G(\bar{x}_t,\xi)) - \varphi(\bar{x}_t) = A_t\left(\mathbb{E}_\xi f(x_t - \gamma_t\nabla f(x_t,\xi)) - f(x_t)\right)\\ + B_t\left(\mathbb{E}_\xi\|(z_t - x_*) - \gamma'_t\nabla f(x_t,\xi)\|^2 - \|z_t - x_*\|^2\right)\,.\end{aligned}$$

As f is L-smooth,

$$\begin{aligned}f(x_t - \gamma_t\nabla f(x_t,\xi)) - f(x_t) &\leqslant \langle\nabla f(x_t), -\gamma_t\nabla f(x_t,\xi)\rangle + \frac{L}{2}\|\gamma_t\nabla f(x_t,\xi)\|^2\,,\\ \mathbb{E}_\xi f(x_t - \gamma_t\nabla f(x_t,\xi)) - f(x_t) &\leqslant \langle\nabla f(x_t), -\gamma_t\mathbb{E}_\xi\nabla f(x_t,\xi)\rangle + \frac{L}{2}\mathbb{E}_\xi\|\gamma_t\nabla f(x_t,\xi)\|^2\,.\end{aligned}$$

Bu assumptions (13) and (14), the stochastic gradient $\nabla f(x,\xi)$ is unbiased and has a variance bounded by σ^2, which implies $\mathbb{E}_\xi\|\nabla f(x_t,\xi)\|^2 \leqslant \|\nabla f(x_t)\|^2 + \sigma^2$. Thus

$$\mathbb{E}_\xi f(x_t - \gamma_t\nabla f(x_t,\xi)) - f(x_t) \leqslant -\gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\|\nabla f(x_t)\|^2 + \sigma^2\frac{L\gamma_t^2}{2}\,.$$

Similarly,

$$\begin{aligned}\mathbb{E}_\xi\|(z_t - x_*) - \gamma'_t\nabla f(x_t,\xi)\|^2 - \|z_t - x_*\|^2 &= -2\gamma'_t\langle\mathbb{E}_\xi\nabla f(x_t,\xi), z_t - x_*\rangle + \gamma_t'^2\mathbb{E}_\xi\|\nabla f(x_t,\xi)\|^2\\ &\leqslant -2\gamma'_t\langle\nabla f(x_t), z_t - x_*\rangle + \gamma_t'^2\|\nabla f(x_t)\|^2 + \sigma^2\gamma_t'^2\,.\end{aligned}$$

This gives

$$\begin{aligned}\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) \leqslant \left(B_t\gamma_t'^2 - A_t\gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\right)\|\nabla f(x_t)\|^2 - 2B_t\gamma'_t\langle\nabla f(x_t), z_t - x_*\rangle\\ + \sigma^2\left(A_t\frac{L\gamma_t^2}{2} + B_t\gamma_t'^2\right)\,.\end{aligned}$$

Combining the bounds, we obtain

$$\begin{aligned} I_t \leqslant & \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\langle \nabla f(x_t), x_t - x_*\rangle + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\|z_t - x_*\|^2 \\ & + (A_t\eta_t - 2B_t\gamma_t')\langle \nabla f(x_t), z_t - x_*\rangle + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\frac{\mu}{2}\right)\|x_t - x_*\|^2 \\ & + \left(B_t\gamma_t'^2 - A_t\gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\right)\|\nabla f(x_t)\|^2 + \sigma^2\left(A_t\frac{L\gamma_t^2}{2} + B_t\gamma_t'^2\right), \end{aligned}$$

which is an additive perturbation of the bound (20)-(22) in the noiseless case, with a perturbation proportional to σ^2. The choices of parameters of Theorem 3 cancel all first five prefactors, and satisfy $\gamma_t = \frac{1}{L}$, $A_t\frac{L\gamma_t^2}{2} = B_t\gamma_t'^2$. We thus obtain

$$I_t \leqslant \sigma^2\frac{A_t}{L}\,.$$

This bound controls the increase of ϕ_t. Using the decomposition (28), we obtain

$$\begin{aligned} \mathbb{E}f(x_t) - f(x_*) \leqslant \frac{\mathbb{E}\phi_t}{A_t} & \leqslant \frac{\phi_0}{A_t} + \frac{\int_0^t \mathbb{E}I_s\mathrm{d}s}{A_t} \\ & \leqslant \frac{A_0(f(x_0) - f(x_*)) + B_0\|z_0 - x_*\|^2}{A_t} + \frac{\sigma^2}{L}\frac{\int_0^t A_s\mathrm{d}s}{A_t}\,. \end{aligned}$$

B.2.1. *Proof of the first part (convex case).* In this case, $A_t = \frac{t^2}{2L}$ and $B_0 = 1$. Thus $\int_0^t A_s\mathrm{d}s = \frac{1}{2L}\frac{t^3}{3}$. Thus

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{2L\|z_0 - x_*\|^2}{t^2} + \sigma^2\frac{t}{3L}\,.$$

B.2.2. *Proof of the second part (strongly convex case).* In this case, $A_t = A_0\exp\left(\sqrt{\frac{\mu}{L}}t\right)$ and $B_0 = A_0\frac{\mu}{2}$. Thus $\int_0^t A_s\mathrm{d}s \leqslant A_0\sqrt{\frac{\mu}{L}}^{-1}\exp\left(\sqrt{\frac{\mu}{L}}t\right) = \sqrt{\frac{L}{\mu}}A_t$. Thus

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right)\exp\left(-\sqrt{\frac{\mu}{L}}t\right) + \sigma^2\frac{1}{\sqrt{\mu L}}\,.$$

Appendix C. Proof of Theorem 4

By integrating the ODE

$$\begin{aligned} \mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t\,, \\ \mathrm{d}z_t &= \eta_t'(x_t - z_t)\mathrm{d}t\,, \end{aligned}$$

between T_k and $T_{k+1}-$, we obtain that there exists τ_k, τ_k'', such that

$$\begin{aligned} \tilde{y}_k = x_{T_{k+1}-} &= x_{T_k} + \tau_k(z_{T_k} - x_{T_k}) = \tilde{x}_k + \tau_k(\tilde{z}_k - \tilde{x}_k)\,, \\ z_{T_{k+1}-} &= z_{T_k} + \tau_k''(x_{T_k} - z_{T_k}) = \tilde{z}_k + \tau_k''(\tilde{x}_k - \tilde{z}_k)\,. \end{aligned} \tag{29}$$

From the first equation, we have $\tilde{x}_k = \frac{1}{1-\tau_k}(\tilde{y}_k - \tau_k\tilde{z}_k)$, which gives by substitution in the second equation,

$$\begin{aligned} z_{T_{k+1}-} &= \tilde{z}_k + \tau_k''\left(\frac{1}{1-\tau_k}(\tilde{y}_k - \tau_k\tilde{z}_k) - \tilde{z}_k\right) \\ &= \tilde{z}_k + \tau_k'(\tilde{y}_k - \tilde{z}_k)\,, \end{aligned}$$

where $\tau_k' = \frac{\tau_k''}{1-\tau_k}$.

Further, from (4)-(5), we obtain the equations

$$\tilde{x}_{k+1} = x_{T_{k+1}} = x_{T_{k+1}-} - \gamma_{T_{k+1}}\nabla f(x_{T_{k+1}-}) = \tilde{y}_k - \gamma_{T_{k+1}}\nabla f(\tilde{y}_k)\,, \tag{30}$$

$$\tilde{z}_{k+1} = z_{T_{k+1}} = z_{T_{k+1}-} - \gamma_{T_{k+1}}'\nabla f(x_{T_{k+1}-}) = \tilde{z}_k + \tau_k'(\tilde{y}_k - \tilde{z}_k) - \gamma_{T_{k+1}}'\nabla f(\tilde{y}_k)\,. \tag{31}$$

The stated equation (10)-(12) are the combination of (29), (30) and (31).

(1) The parameters of Theorem 3.1 are $\eta_t = \frac{2}{t}, \eta_t' = 0, \gamma_t = \frac{1}{L}$ and $\gamma_t' = \frac{t}{2L}$. In this case, the ODE

$$\begin{aligned} \mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t = \frac{2}{t}(z_t - x_t)\mathrm{d}t\,, \\ \mathrm{d}z_t &= \eta_t'(x_t - z_t)\mathrm{d}t = 0\,, \end{aligned}$$

can be integrated in closed form: for $t \geqslant t_0$,

$$\begin{aligned} x_t &= z_{t_0} + \left(\frac{t_0}{t}\right)^2 (x_{t_0} - z_{t_0}) = x_{t_0} + \left(1 - \left(\frac{t_0}{t}\right)^2\right)(z_{t_0} - x_{t_0})\,, \\ z_t &= z_{t_0}\,. \end{aligned}$$

In particular, taking $t_0 = T_k$, $t = T_{k+1}-$, we obtain $\tau_k = 1 - \left(\frac{T_k}{T_{k+1}}\right)^2$, $\tau_k'' = 0$ and thus $\tau_k' = \frac{\tau_k''}{1-\tau_k} = 0$. Finally, $\tilde{\gamma}_k = \gamma_{T_k} = \frac{1}{L}$ and $\tilde{\gamma}_k' = \gamma_{T_k}' = \frac{T_k}{2L}$.

(2) The parameters of Theorem 3.2 are $\eta_t = \eta_t' \equiv \sqrt{\frac{\mu}{L}}, \gamma_t \equiv \frac{1}{L}$ and $\gamma_t' \equiv \frac{1}{\sqrt{\mu L}}$. In this case, the ODE

$$\begin{aligned} \mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t = \sqrt{\frac{\mu}{L}}(z_t - x_t)\mathrm{d}t\,, \\ \mathrm{d}z_t &= \eta_t'(x_t - z_t)\mathrm{d}t = \sqrt{\frac{\mu}{L}}(x_t - z_t)\mathrm{d}t\,, \end{aligned}$$

can also be integrated in closed form: for $t \geqslant t_0$,

$$\begin{aligned} x_t &= \frac{x_{t_0} + z_{t_0}}{2} + \frac{x_{t_0} - z_{t_0}}{2}\exp\left(-2\sqrt{\frac{\mu}{L}}(t - t_0)\right) \\ &= x_{t_0} + \frac{1}{2}\left(1 - \exp\left(-2\sqrt{\frac{\mu}{L}}(t - t_0)\right)\right)(z_{t_0} - x_{t_0})\,, \\ z_t &= \frac{x_{t_0} + z_{t_0}}{2} + \frac{z_{t_0} - x_{t_0}}{2}\exp\left(-2\sqrt{\frac{\mu}{L}}(t - t_0)\right) \\ &= z_{t_0} + \frac{1}{2}\left(1 - \exp\left(-2\sqrt{\frac{\mu}{L}}(t - t_0)\right)\right)(x_{t_0} - z_{t_0})\,. \end{aligned}$$

In particular, taking $t_0 = T_k$, $t = T_{k+1}-$, we obtain $\tau_k = \tau_k'' = \frac{1}{2}\left(1 - \exp\left(-2\sqrt{\frac{\mu}{L}}(T_{k+1} - T_k)\right)\right)$ and thus $\tau_k' = \frac{\tau_k''}{1-\tau_k} = \tanh\left(\sqrt{\frac{\mu}{L}}(T_{k+1} - T_k)\right)$. Finally, $\tilde{\gamma}_k = \gamma_{T_k} = \frac{1}{L}$ and $\tilde{\gamma}_k' = \gamma_{T_k}' = \frac{1}{\sqrt{\mu L}}$.

Appendix D. Heuristic ODE scaling limit of the continuized acceleration

D.1. Convex case. With the choices of parameters of Theorem 3.1, the continuized acceleration is

$$\begin{aligned} \mathrm{d}x_t &= \frac{2}{t}(z_t - x_t)\mathrm{d}t - \frac{1}{L}\nabla f(x_t)\mathrm{d}N(t)\,, \\ \mathrm{d}z_t &= -\frac{t}{2L}\nabla f(x_t)\mathrm{d}N(t)\,. \end{aligned}$$

The ODE scaling limit is obtained by taking the limit $L \to \infty$ (so that the stepsize $1/L$ vanishes) and rescaling the time $s = t/\sqrt{L}$. Some law of large number argument heuristically gives us that, as $L \to \infty$, $\mathrm{d}N(t) = \mathrm{d}N(\sqrt{L}s) \approx \sqrt{L}\mathrm{d}s$. Thus in the limit, we obtain

$$\begin{aligned} \mathrm{d}x_s &= \frac{2}{\sqrt{L}s}(z_s - x_s)\sqrt{L}\mathrm{d}s - \frac{1}{L}\nabla f(x_s)\sqrt{L}\mathrm{d}s\,, \\ \mathrm{d}z_s &= -\frac{\sqrt{L}s}{2L}\nabla f(x_s)\sqrt{L}\mathrm{d}s\,. \end{aligned}$$

The second term of the first equation becomes negligible in the limit. Thus the equations simplify to

$$\frac{\mathrm{d}x_s}{\mathrm{d}s} = \frac{2}{s}(z_s - x_s)\,,$$
$$\frac{\mathrm{d}z_s}{\mathrm{d}s} = -\frac{s}{2}\nabla f(x_s)\,.$$

Thus

$$-\frac{s}{2}\nabla f(x_s) = \frac{\mathrm{d}z_s}{\mathrm{d}s} = \frac{\mathrm{d}}{\mathrm{d}s}\left(x_s + \frac{s}{2}\frac{\mathrm{d}x_s}{\mathrm{d}s}\right) = \frac{\mathrm{d}x_s}{\mathrm{d}s} + \frac{1}{2}\frac{\mathrm{d}x_s}{\mathrm{d}s} + \frac{s}{2}\frac{\mathrm{d}^2x_s}{\mathrm{d}s^2}\,,$$

and thus

$$\frac{\mathrm{d}^2x_s}{\mathrm{d}s^2} + \frac{3}{s}\frac{\mathrm{d}x_s}{\mathrm{d}s} + \nabla f(x_s) = 0\,.$$

This is the same limiting ODE as the one found by Su et al. (2014) for Nesterov acceleration.

D.2. Strongly-convex case. With the choices of parameters of Theorem 3.2, the continuized acceleration is

$$\mathrm{d}x_t = \sqrt{\frac{\mu}{L}}(z_t - x_t)\mathrm{d}t - \frac{1}{L}\nabla f(x_t)\mathrm{d}N(t)\,,$$
$$\mathrm{d}z_t = \sqrt{\frac{\mu}{L}}(x_t - z_t)\mathrm{d}t - \frac{1}{\sqrt{\mu L}}\nabla f(x_t)\mathrm{d}N(t)\,.$$

Again, we take joint scaling $L \to \infty$, $s = t/\sqrt{L}$, with the approximation $\mathrm{d}N(t) \approx \sqrt{L}\mathrm{d}s$. We obtain

$$\mathrm{d}x_s = \sqrt{\frac{\mu}{L}}(z_s - x_s)\sqrt{L}\mathrm{d}s - \frac{1}{L}\nabla f(x_s)\sqrt{L}\mathrm{d}s\,,$$
$$\mathrm{d}z_s = \sqrt{\frac{\mu}{L}}(x_s - z_s)\sqrt{L}\mathrm{d}s - \frac{1}{\sqrt{\mu L}}\nabla f(x_s)\sqrt{L}\mathrm{d}s\,.$$

As before, the second term of the first equation becomes negligible in the limit. Thus the equations simplify to

$$\frac{\mathrm{d}x_s}{\mathrm{d}s} = \sqrt{\mu}(z_s - x_s)\,, \tag{32}$$
$$\frac{\mathrm{d}z_s}{\mathrm{d}s} = \sqrt{\mu}(x_s - z_s) - \frac{1}{\sqrt{\mu}}\nabla f(x_s)\,. \tag{33}$$

From (32), we have $z_s = x_s + \frac{1}{\sqrt{\mu}}\frac{\mathrm{d}x_s}{\mathrm{d}s}$, and by substitution in (33), we obtain

$$\frac{\mathrm{d}^2x_s}{\mathrm{d}s^2} + 2\sqrt{\mu}\frac{\mathrm{d}x_s}{\mathrm{d}s} + \nabla f(x_s) = 0\,.$$

This is the so-called "low-resolution" ODE for Nesterov acceleration of Shi et al. (2018).