A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip

Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Pierre Gaillard, Hadrien Hendrikx, Laurent Massoulié, Adrien Taylor

A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip

Mathieu Even[1,⋆], Raphaël Berthier[1,⋆], Francis Bach[1], Nicolas Flammarion[2], Pierre Gaillard[3], Hadrien Hendrikx[1], Laurent Massoulié[1,4] and Adrien Taylor[1]

⋆ Equal contributions

[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

[4]MSR-Inria Joint Centre

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; and 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. We provide continuized Nesterov acceleration under deterministic as well as stochastic gradients, with either additive or multiplicative noise. Finally, using our continuized framework and expressing the gossip averaging problem as the stochastic minimization of a certain energy function, we provide the first rigorous acceleration of asynchronous gossip algorithms.

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 [10]. They enjoy a low iteration cost 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. [14] and the many references therein), starting with the seminal work of Nesterov [42].

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.,

$$\forall x, y \in \mathbb{R}^d, \qquad f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 . \tag{1}$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

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

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

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 iteration cost, while achieving faster rates: it converges with rate $O(k^{-2})$ in the smooth convex case and with rate $O((1 - \sqrt{\mu/L})^k)$ in the smooth and strongly convex case [43]. These rates are then optimal among all black-box first-order methods that access gradients and linearly combine them [43, 41].

Nesterov acceleration relies on several sequences of iterates—two or three, depending on the formulation—and on a clever blend of gradient steps and mixing steps between the sequences. Different interpretations and motivations underlying the precise structure of accelerated schemes were approached in many works, including [12, 24, 3, 32, 2]. A large number of these works studied 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. Examples of studies relying on continuous time interpretation include [50, 33, 54, 53, 9, 18, 48, 49, 4, 5, 57, 40].

Continuized Nesterov acceleration. In this paper, we propose another continuous time equivalent to Nesterov acceleration, which we refer to as the *continuized* Nesterov acceleration, which avoids vanishing stepsizes. It is built by considering two sequences $x_t, z_t \in \mathbb{R}^d$, $t \in \mathbb{R}_{\geqslant 0}$, 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, and performs similarly to Nesterov original deterministic version both in theory and in simulations.

The continuized framework can be adapted to various settings and extensions of Nesterov acceleration. In what follows, we study how the continuized acceleration behaves in the presence of *additive* and *multiplicative* noise in the gradients. In the multiplicative noise setting, our acceleration satisfies a convergence rate similar to that of [30] and depends on the *statistical condition number* of the problem at hand. The two acceleration schemes are not directly comparable as we work in a continuized setting and only deal with pure multiplicative noise. Our analysis is nevertheless much simpler, as it closely mimics that of Nesterov acceleration.

Application to accelerated gossip algorithms. The continuized modeling is natural in asynchronous parallel computing where gradient steps arrive at random times. More importantly, there are situations where the continuized version of Nesterov acceleration can be practically implemented while the original acceleration can not. In distributed settings, for instance, the total number k of gradient steps taken in the network is typically not known to each particular node; an advantage of the continuized acceleration is that it requires to know only the time t and not k.

Gossip algorithms typically feature such asynchronous and distributed behaviors [11]. In gossip problems, nodes of a network aim at computing the global average of all their values by communicating only locally (with their neighbors), and without centralized coordination. In this set-up, pairs of adjacent nodes communicate at random times, asynchronously, and in parallel, so that the total number of past communications in the network at a given time is unknown to all nodes. In this paper, we formulate the gossip problem as a stochastic optimization problem. Thanks to the continuized formalism, we naturally obtain accelerated gossip algorithms that can be implemented in an asynchronous and distributed fashion.

Synchronous gossip algorithms rely on all nodes to communicate simultaneously [19]. Accelerating synchronous gossip algorithms have been studied in previous works, including *SSDA* [47], Chebyshev acceleration [39], Jacobi-Polynomial acceleration [7]. To that day, acceleration in the asynchronous setting has also been studied in a few works (see for instance geographic gossip [20], shift registers [37], *ESDAC* [25], and randomized Kaczmarz methods [38]). However, no algorithm in an asynchronous framework has been rigorously proven to achieve an accelerated rate for general graphs [21]. Other acceleration schemes [25, 38] relied on additional synchronizations between nodes, such as the knowledge of a global iteration counter. This departs from purely asynchronous operations, hence causing practical limitation. Our accelerated randomized gossip algorithm (Section 6) recovers the same accelerated rates, and only requires the knowledge of a common continuous-time $t \in \mathbb{R}_{\geqslant 0}$.

In this context, 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. In practice, the continuized framework is relevant for handling asynchrony in decentralized optimization, where agents of a network can not share a global iteration counter, preventing accelerated decentralized and asynchronous methods.

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

Structure of the paper. In Section 2, we recall standard results on gradient descent and Nesterov acceleration. In Section 3, we introduce a continuized variant of Nesterov acceleration. In Section 4, we show that discretizing the continuized acceleration yields an iterative method similar to that of Nesterov but with random parameters. In Section 5, we study continuized Nesterov acceleration under pure-multiplicative noise. We finally present accelerated asynchronous algorithms for the gossip problem in Section 6, as well as for decentralized optimization in Section 7.

2 Reminders on Nesterov acceleration

For the sake of comparison, let us first recall the classical Nesterov acceleration. To improve the convergence rate of gradient descent, Nesterov introduced iterations of three sequences, parametrized by $\tau_k, \tau_k', \gamma_k, \gamma_k', k \geqslant 0$, of the form

$$y_k = x_k + \tau_k(z_k - x_k)\,, \tag{3}$$
$$x_{k+1} = y_k - \gamma_k \nabla f(y_k)\,, \tag{4}$$
$$z_{k+1} = z_k + \tau_k'(y_k - z_k) - \gamma_k' \nabla f(y_k)\,. \tag{5}$$

Depending on whether the function f is known to be (1) convex, or (2) strongly convex with a known strong convexity parameter, Nesterov provided a set of parameter choices for achieving acceleration.

Theorem 1 (Convergence of accelerated gradient descent)**.** *Nesterov accelerated scheme satisfies:*

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*
$$A_0 = 0\,, \qquad\qquad A_{k+1} = A_k + \frac{1}{2}(1 + \sqrt{4A_k + 1})\,.$$
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 can be found as is in d'Aspremont et al. [14, Sections 4.4.1 and 4.5.3]. In the convex case, Nesterov acceleration achieves the rate $O(1/k^2)$, whereas gradient descent achieves a rate $O(1/k)$

(see [43, Corollary 2.1.2] for instance). In the strongly convex case, Nesterov acceleration achieves the rate $O((1-\sqrt{\mu/L})^k)$, where gradient descent achieves a rate $O((1-\mu/L)^k)$ (see [43, Theorem 2.1.15] for instance). In both cases, this results in a significant speedup in practice, see Figure 1.

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 1 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

This paper uses several mathematical notions related to random processes. The following sections expose the results from heuristic considerations of those notions, rigorously defined in Appendix C.

We argue that the accelerated iteration becomes more natural when considering 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 $T_1, T_2, T_3, \dots \geqslant 0$ be random times such that $T_1, T_2 - T_1, T_3 - T_2, \dots$ are independent identically distributed (i.i.d.), of law exponential with rate 1 (any constant rate would do, 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-} (Definition 6 in Appendix C). Our dynamics are parametrized by functions $\gamma_t, \gamma'_t, \tau_t, \tau'_t$, $t \geqslant 0$. At random times $T_1, T_2, \dots$, our sequences take gradient steps

$$x_{T_k} = x_{T_k-} - \gamma_{T_k} \nabla f(x_{T_k-}) \,, \tag{6}$$

$$z_{T_k} = z_{T_k-} - \gamma'_{T_k} \nabla f(x_{T_k-}) \,. \tag{7}$$

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, translation-invariant, ordinary differential equation (ODE)

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

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

Following the notation of stochastic calculus, we can write the process more compactly in terms of the Poisson point measure $\mathrm{d}N(t) = \sum_{k \geqslant 1} \delta_{T_k}(\mathrm{d}t)$, which has intensity the Lebesgue measure $\mathrm{d}t$,

$$\mathrm{d}x_t = \eta_t (z_t - x_t) \mathrm{d}t - \gamma_t \nabla f(x_t) \mathrm{d}N(t) \,, \tag{10}$$

$$\mathrm{d}z_t = \eta'_t (x_t - z_t) \mathrm{d}t - \gamma'_t \nabla f(x_t) \mathrm{d}N(t) \,. \tag{11}$$

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 [1]. Following this terminology, we refer to our acceleration (10)-(11) 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 can also be compared with coordinate gradient descent methods, that are easier to analyze when coordinates are selected randomly rather than in an ordered way [55]. Similarly, the continuized acceleration is simpler to analyze because the gradient steps (6)-(7) and the mixing steps (8)-(9) alternate randomly, due to the randomness of T_k, $k \geqslant 0$.

In analogy with Theorem 1, 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 1. In the theorem below, $\mathbb{E}$ denotes the expectation with respect to the Poisson point process $\mathrm{d}N(t)$, the only source of randomness.

Theorem 2 (Convergence of continuized Nesterov acceleration)**.** *The continuized Nesterov acceleration satisfies the following two points.*

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

We give an elementary sketch of proof in Appendix D.1 and a complete proof in Appendix D.2. Many authors have proposed continuous-time versions of Nesterov acceleration using differential calculus, see the numerous references in the introduction. For instance, in Su et al. [50], an ODE is obtained from Nesterov acceleration by taking the joint asymptotic where the stepsizes vanish and the number of iterates is rescaled. The resulting ODE must be discretized to be implemented; choosing the right discretization is not straightforward as it introduces stability and approximation errors that must be controlled [57, 49, 46].

On the contrary, our continuous time process (10)-(11) does not correspond to a limit where the stepsizes vanish. However, in Appendix F, we check that the random continuized acceleration has the same deterministic ODE scaling limit as Nesterov acceleration. This sanity check emphasizes that the continuized acceleration is fundamentally different from previous continuous-time equivalents.

Remark 1. *A similar Markovian structure can be obtained in a discrete setting by flipping* i.i.d. *coins to trigger gradient steps. By denoting* $p > 0$ *the probability to trigger a gradient step when flipping a coin,* (i) $p = 1$ *gives the classical setting, and* (ii) $p \to 0$ *while renormalizing time gives our continuized framework. In fact, this setting with updates triggered randomly is an interpolation between the classical and continuized frameworks, and consists in replacing exponential random variables by geometric random variables of parameter* p *for the waiting-time between updates. We thus believe the convergence guarantees described here and in the following can be adapted for this discrete scheme.*

4 Discrete implementation of the continuized acceleration with random parameters

In this section, we show that the continuized acceleration can be implemented exactly as a discrete algorithm. This contrasts with the discretization of ODEs that introduces discretization errors; here, we compute exactly

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

with the convention that $T_0 = 0$. 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. The resulting randomized discrete algorithm satisfies performance guarantees similar to those of Nesterov acceleration.

Theorem 3 (Discrete version of continuized acceleration)**.** *For any stochastic process of the form* (10)-(11)*, we have*

$$\tilde{y}_k = \tilde{x}_k + \tau_k(\tilde{z}_k - \tilde{x}_k)\,, \tag{12}$$
$$\tilde{x}_{k+1} = \tilde{y}_k - \tilde{\gamma}_k \nabla f(\tilde{y}_k)\,, \tag{13}$$
$$\tilde{z}_{k+1} = \tilde{z}_k + \tau'_k(\tilde{y}_k - \tilde{z}_k) - \tilde{\gamma}'_k \nabla f(\tilde{y}_k)\,, \tag{14}$$

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 2.(1),* $\tau_k = 1 - \left(\frac{T_k}{T_{k+1}}\right)^2$, $\tau'_k = 0$, $\tilde{\gamma}_k = \frac{1}{L}$*, and* $\tilde{\gamma}'_k = \frac{T_k}{2L}$*. Then*
$$\mathbb{E}\left[T_k^2\left(f(\tilde{x}_k) - f(x_*)\right)\right] \leqslant 2L\|z_0 - x_*\|^2\,.$$

2. *For the parameters of Theorem 2.(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* $\tilde{\gamma}'_k = \frac{1}{\sqrt{\mu L}}$*. Then*
$$\mathbb{E}\left[\exp\left(\sqrt{\frac{\mu}{L}}T_k\right)(f(\tilde{x}_k) - f(x_*))\right] \leqslant f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2 .$$

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 the bounds above. The performance guarantees are proved in Appendix D.2, and the formula for the discretization is studied in E. In Appendix A.1, we provide simulations confirming that this discrete random algorithm has a performance similar to Nesterov's original acceleration.

5 Continuized Nesterov acceleration of stochastic gradient descent

We now investigate the design of continuized accelerations of stochastic gradient descent. 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 $\mathcal{P}$. In the continuized framework, the randomness of the stochastic gradient and its time mix in a particularly convenient way. For similar reasons, Latz studied stochastic gradient descent as a gradient flow on a random function that is regenerated at a Poisson rate [35]. However, this approach has the same shortcomings as the other approaches based on gradient flows: the subsequent discretization introduces non-trivial errors. We avoid this problem here.

We keep the algorithms of the same form, 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 $T_1, T_2, \dots$,
$$x_{T_k} = x_{T_k-} - \gamma_{T_k}\nabla f(x_{T_k-}, \xi_k) ,$$
$$z_{T_k} = z_{T_k-} - \gamma'_{T_k}\nabla f(x_{T_k-}, \xi_k) .$$
Between these random times, the variables mix through the same ODE
$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t ,$$
$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t .$$
This can be written more compactly in terms of the Poisson point measure $\mathrm{d}N(t, \xi) = \sum_{k\geqslant 1}\delta_{(T_k,\xi_k)}(\mathrm{d}t, \mathrm{d}\xi)$ on $\mathbb{R}_{\geqslant 0} \times \Xi$, which has intensity $\mathrm{d}t \otimes \mathcal{P}$,
$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t\int_\Xi \nabla f(x_t, \xi)\mathrm{d}N(t, \xi) , \tag{15}$$
$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t - \gamma'_t\int_\Xi \nabla f(x_t, \xi)\mathrm{d}N(t, \xi) . \tag{16}$$
Here, the discussion depends on the properties satisfied by the stochastic gradients $\nabla f(x, \xi)$. In Appendix B, we study the so-called *additive noise* case. We show that the continuized acceleration satisfies perturbed convergence rates with the same choices of parameters as in Theorem 2. We thus show some robustness of the above acceleration to additive noise. Instead, in this section, we focus on the so-called *pure multiplicative noise* case, as it is crucial for the study of asynchronous gossip that follows. In this setting, parameters need to be chosen differently for our proof technique to work. A continuized acceleration is still possible, depending on the statistical condition number.

We now focus on functions f is of the following form, typical to least-squares supervised learning:
$$\forall x \in \mathbb{R}^d,\ f(x) = \mathbb{E}_{(a,b)\sim\mathcal{P}}\left[\frac{1}{2}(b - \langle x, a\rangle)^2\right] , \tag{17}$$
where $\xi = (a, b) \in \mathbb{R}^d \times \mathbb{R}$ is random of law $\mathcal{P}$. We assume that our *stochastic first order oracle* is the gradient of one realization of the expectation, namely,
$$\nabla f(x, \xi) = -(b - \langle x, a\rangle)a , \qquad \xi = (a, b) .$$
We investigate *noiseless*—or purely multiplicative—stochastic gradients, in the sense that almost surely, for $\xi = (a, b) \sim \mathcal{P}$:
$$b = \langle x_*, a\rangle,\ \text{so that}\ \nabla f(x_*, \xi) = 0 . \tag{18}$$

Noiseless stochastic gradients are relevant in several situations, such as coordinate gradient descent with randomly sampled coordinates [51, 44, 55] (where $\nabla f(x,\xi) = m\langle \nabla f(x), e_i\rangle e_i$ with i uniformly random in $\{1,\dots,d\}$), over-parameterized regime for least squares regression [52], function interpolation and gossip algorithms [8].

For a symmetric non-negative matrix A and a vector x, we denote $\|x\|_A^2 = x^\top A x$. Let $H = \mathbb{E}[aa^\top]$ be the Hessian of f. Let R^2 be the smallest positive real number such that:

$$\mathbb{E}\left[\|a\|^2 aa^\top\right] \preccurlyeq R^2 H\,. \tag{19}$$

Further, similarly to Jain et al. [30], we define the statistical condition number of the problem as the smallest $\tilde\kappa > 0$ such that:

$$\mathbb{E}\left[\|a\|_{H^{-1}}^2 aa^\top\right] \preccurlyeq \tilde\kappa H\,. \tag{20}$$

Theorem 4 (Continuized acceleration with pure multiplicative noise)**.** *Assume that* (18), (19) *and* (20) *hold true. Then the continuized acceleration satisfies the following.*

1. *Choose the parameters* $\eta_t = \frac{2}{t}, \eta'_t = 0, \gamma_t = \frac{1}{R^2}, \gamma'_t = \frac{t}{2R^2\tilde\kappa}$*. Then*

$$\frac{1}{2}\mathbb{E}\|x_t - x_*\|^2 \leqslant \frac{R^2\tilde\kappa\|z_0 - x_*\|_{H^{-1}}^2}{t^2}\,.$$

2. *Assume further that f is μ-strongly convex, i.e., all eigenvalues of H are greater or equal to μ, where $\mu > 0$. The condition number of f is then defined as $\kappa = R^2/\mu$. For the parameters* $\eta_t = \eta'_t = \frac{1}{\sqrt{\kappa\tilde\kappa}}$, $\gamma_t = \frac{1}{R^2}$ *and* $\gamma'_t = \frac{1}{R^2}\sqrt{\frac{\kappa}{\tilde\kappa}}$*, we have:*

$$\frac{1}{2}\mathbb{E}\|x_t - x_*\|^2 \leqslant \left(\frac{1}{2}\|x_0 - x_*\|^2 + \frac{\mu}{2}\|z_0 - x_*\|_{H^{-1}}^2\right)\exp\left(-\frac{t}{\sqrt{\kappa\tilde\kappa}}\right).$$

In the strongly convex case, the benefits of this acceleration are similar to those of Jain et al. [30] with classical discrete iterates: while stochastic gradient descent with stepsize $1/R^2$ is easily shown to achieve an exponential rate of convergence $1/\kappa$, the acceleration enjoys a rate of convergence of $1/\sqrt{\kappa\tilde\kappa}$. Note that from the definitions, $\tilde\kappa \leqslant \kappa$, thus the acceleration performs as least as well as the naive algorithm. However, depending on the distribution of a, the improvement might either be significant or null. We refer the reader to the rich discussion in Jain et al. [30] which provides insights on the interpretation of $\tilde\kappa$ and on the possibility to accelerate. Below, we provide a complementary perspective on the statistical condition number in the context of gossip algorithms, where it can be interpreted in terms of effective resistances of graphs.

Albeit more restrictive in terms of assumptions, our analysis is much simpler than that of Jain et al. [30], as it relies on a standard Lyapunov function, similar to that of the continuized acceleration (Theorem 2). In Appendix G, we use the same analysis framework to prove convergence of accelerated coordinate descent, which is another noiseless stochastic method.

6 Accelerating Randomized Gossip

The continuized framework allows designing accelerated decentralized algorithms requiring synchronized clocks, but no synchronization of the communications. In this section, we illustrate this statement in the simple case of gossip algorithms; the more general case of decentralized optimization is discussed in the next section.

Let $G = (\mathcal{V}, \mathcal{E})$ a connected graph representing a communication network of agents. Each agent $v \in \mathcal{V}$ is assigned a real number $x_0(v) \in \mathbb{R}$. The goal of the averaging (or gossip) problem is to design an iterative procedure allowing each agent of the network to know the average $\bar x = \frac{1}{m}\sum_{v\in\mathcal{V}} x_0(v)$ using only local communications, *i.e.*, communications between adjacent agents in the network.

We formalize the communication model of randomized gossip [11]. Time t is indexed continuously in $\mathbb{R}_{\geqslant 0}$. We generate a Poisson point measure $\mathrm{d}N(t,e) = \sum_{k\geqslant 1}\delta_{(T_k,\{v_k,w_k\})}$ with intensity measure $\mathrm{d}t\otimes\mathcal{P}$, where $\mathrm{d}t$ is the Lebesgue measure on $\mathbb{R}_{\geqslant 0}$ and $\mathcal{P} = (\mathcal{P}_{\{v,w\}})_{\{v,w\}\in\mathcal{E}}$ is a probability measure on the set $\mathcal{E}$ of edges. For $k \geqslant 0$, T_k is a time at which edge $\{v_k, w_k\}$ is *activated*: adjacent nodes v_k

and w_k can communicate and perform a pairwise update. The Poisson point measure assumption implies that edges are activated independently of one another and from the past: the activation times of edge $\{v, w\}$ form a Poisson point process of intensity $\mathcal{P}_{\{v,w\}}$.

To solve the gossip problem, Boyd et al. [11] proposed the following naive strategy: each agent $v \in \mathcal{V}$ keeps a local estimate $x_t(v)$ of the average and, upon activation of edge $\{v_k, w_k\}$ at time $T_k \in \mathbb{R}_{\geqslant 0}$, the activated nodes v_k, w_k average their current estimates

$$x_{T_k}(v_k),\ x_{T_k}(w_k) \quad \longleftarrow \quad \frac{x_{T_k-}(v_k) + x_{T_k-}(w_k)}{2}\,.$$

In this section, we accelerate this naive procedure. Our strategy is to apply Section 5 as follows. Consider the energy function

$$f(x) = \sum_{\{v,w\}\in\mathcal{E}} \frac{\mathcal{P}_{\{v,w\}}}{2}(x(v) - x(w))^2\,, \qquad x = (x(v))_{v\in\mathcal{V}}\,. \tag{21}$$

This function is convex, smooth, and writes in the form (17):

$$f(x) = \mathbb{E}_{\{v,w\}\sim\mathcal{P}}\left[\frac{1}{2}\left\langle x, a_{\{v,w\}}\right\rangle^2\right], \tag{22}$$

where $a_{\{v,w\}} = e_v - e_w$ and $(e_v)_{v\in\mathcal{V}}$ forms the canonical basis of $\mathbb{R}^{\mathcal{V}}$. As in Section 5, a stochastic gradient of f is obtained by taking the gradient of one realization of the expectation, namely:

$$\nabla f(x, \{v, w\}) = \langle x, a_{\{v,w\}}\rangle a_{\{v,w\}} = \begin{cases} x(v) - x(w) & \text{at coordinate } v, \\ x(w) - x(v) & \text{at coordinate } w, \\ 0 & \text{at all other coordinates.} \end{cases} \tag{23}$$

As a consequence, a stochastic gradient step with stepsize $1/2$ corresponds to a local averaging alongside edge $\{v, w\}$, where $\{v, w\} \sim \mathcal{P}$. More generally, the randomized gossip algorithm as described by Boyd et al. [11] is the stochastic gradient descent:

$$\mathrm{d}x_t = -\frac{1}{2}\int_{\mathbb{R}_{\geqslant 0}\times\mathcal{E}} \nabla f(x_t, \{v, w\})\mathrm{d}N(t, \{v, w\})\,. \tag{24}$$

Using Section 5, we can accelerate this algorithm if we know the strong convexity parameter of f and the constants R^2 and $\tilde{\kappa}$ as defined in (19) and (20) respectively. These constants can be intepreted as graph-related quantities here.

Definition 1 (Graph-related quantities)**.** *The Laplacian matrix $\mathcal{L} \in \mathbb{R}^{\mathcal{V}\times\mathcal{V}}$ of graph G with weights $(\mathcal{P}_{\{v,w\}})_{\{v,w\}\in\mathcal{E}}$ on the edges is the matrix with entries $\mathcal{L}_{v,w} = -\mathcal{P}_{\{v,w\}}$ if $\{v, w\} \in \mathcal{E}$, $\mathcal{L}_{v,v} = \sum_{w\sim v} \mathcal{P}_{\{v,w\}}$, and $\mathcal{L}_{v,w} = 0$ if $\{v, w\} \notin \mathcal{E}$. We denote μ_{gossip} the second smallest eigenvalue of $\mathcal{L}$, corresponding to its smallest positive eigenvalue. For $\{v, w\} \in \mathcal{E}$, let $R_{\mathrm{eff}}(v, w) = (e(v) - e(w))^\top \mathcal{L}^{-1}(e(v) - e(w))$ be the effective resistance of edge $\{v, w\}$, and $R_{\max} = \max_{\{v,w\}\in\mathcal{E}} R_{\mathrm{eff}}(v, w)$ be the maximal resistance in the graph.*

The function f is quadratic with Hessian $\mathcal{L}$, and strongly convex with parameter μ_{gossip} on the hyperplane $F = \{x \in \mathbb{R}^{\mathcal{V}} : \sum_{v\in\mathcal{V}} x(v) = \bar{x}\}$; hence we use the (perhaps abusive) notation μ_{gossip} throughout. Moreover, the conditions (19) and (20) are satisfied with $R^2 = 2$, $\tilde{\kappa} = R_{\max}$.

These parameters being given, the accelerated stochastic gradient descent updates (15)-(16) can be instantiated as follows. Each agent $v \in \mathcal{V}$ keeps two local estimates $x_t(v), z_t(v)$ of $\bar{x}$, initialized at $x_0(v)$. Upon activation of edge $\{v_k, w_k\}$ at time T_k,

$$\begin{aligned} x_{T_k}(v_k) = x_{T_k}(w_k) &= \frac{x_{T_k-}(v_k) + x_{T_k-}(w_k)}{2}\,, \\ z_{T_k}(v_k) = z_{T_k-}(v_k) &+ \frac{1}{\sqrt{2\mu_{\mathrm{gossip}} R_{\max}}}(x_{T_k-}(w_k) - x_{T_k-}(v_k))\,, \\ z_{T_k}(w_k) = z_{T_k-}(w_k) &+ \frac{1}{\sqrt{2\mu_{\mathrm{gossip}} R_{\max}}}(x_{T_k-}(v_k) - x_{T_k-}(w_k))\,. \end{aligned}$$

Between these updates, $x_t(v)$ and $z_t(v)$ locally mix at all nodes $v \in \mathcal{V}$, according to the coupled ODE:

$$\mathrm{d}x_t(v) = \sqrt{\frac{2\mu_{\text{gossip}}}{R_{\max}}}(z_t(v) - x_t(v))\mathrm{d}t,$$
$$\mathrm{d}z_t(v) = \sqrt{\frac{2\mu_{\text{gossip}}}{R_{\max}}}(x_t(v) - z_t(v))\mathrm{d}t.$$

This algorithm is *asynchronous* in the sense that it does not require global synchronous operations: the mixing of local variables does not require any synchronization since parameter $t \in \mathbb{R}_{\geqslant 0}$ is available at all nodes independently from the number of past updates, while a local pairwise update between adjacent nodes v and w only requires a local synchronization.

Theorem 5 (Accelerated randomized gossip)**.** *Let $(x_t(v))_{v\in\mathcal{V},t\geqslant 0}$ be generated with accelerated randomized gossip. For any $t \in \mathbb{R}_{\geqslant 0}$:*

$$\sum_{v\in\mathcal{V}} \frac{1}{2}\mathbb{E}\left[\left(x_t(v) - \bar{x}\right)^2\right] \leqslant 2\left(\sum_{v\in\mathcal{V}} \frac{1}{2}\left(x_0(v) - \bar{x}\right)^2\right) \exp\left(-\sqrt{\frac{\mu_{\text{gossip}}}{2R_{\max}}}t\right).$$

Let $\theta_{\text{ARG}} = \sqrt{\frac{\mu_{\text{gossip}}}{2R_{\max}}}$ be the rate of convergence of accelerated randomized gossip, and $\theta_{\text{RG}} = \mu_{\text{gossip}}$ be the rate of convergence of randomized gossip [11]. We have $\theta_{\text{ARG}} \geqslant \theta_{\text{RG}}/\sqrt{2}$. Let us exhibit scenarios over which accelerated randomized gossip gains several orders of magnitude. Denoting $\mathcal{P}_{\min} = \min_{\{v,w\}\in\mathcal{E}} \mathcal{P}_{\{v,w\}}$, Ellens et al. [22] ensures that for $\{v, w\} \in \mathcal{E}$, $\mathcal{P}_{\min}R_{\text{eff}}(v, w) \leqslant 1$, so that $R_{\max} \leqslant \mathcal{P}_{\min}^{-1}$.

Corollary 1 (Comparison with randomized gossip)**.** *Accelerated randomized gossip achieves a rate satisfying:*

$$\sqrt{\frac{\theta_{\text{RG}}\mathcal{P}_{\min}}{2}} \leqslant \theta_{\text{ARG}}.$$

Assume furthermore that there exist some constants $c > 0$ such that for all $\{v, w\} \in \mathcal{E}$, $\mathcal{P}_{\{v,w\}} \leqslant c\mathcal{P}_{\min}$ and $d_v + d_w \leqslant 2d$. Then, with $C = 1/\sqrt{2cd}$:

$$C\sqrt{\frac{\theta_{\text{RG}}}{|\mathcal{V}|}} \leqslant \theta_{\text{ARG}}.$$

Assume now for simplicity that the Poisson intensities $\mathcal{P}_{\{v,w\}}$ are all equal to $1/|\mathcal{E}|$. Denoting $|\mathcal{V}| = m$, on the cyclic and the line graph, this gives us $\theta_{\text{ARG}} = \Omega(1/m^2)$ while $\theta_{\text{RG}} \asymp 1/m^3$. On a d-dimensional grid, we have $\theta_{\text{ARG}} = \Omega(1/m^{1+1/d})$ and $\theta_{\text{RG}} \asymp 1/m^{1+2/d}$. However, on graphs with unbounded degrees, no improvements are observed, as illustrated in Figure 2, Appendix A.2. In the case of the complete graph, this is expected since at least $\theta_{\text{RG}}^{-1} \asymp m$ communications are needed to compute the average. We thus recover the same rates as Dimakis et al. [20] for the graphs they study, but generalized to any network.

7 Accelerating Asynchronous Decentralized Optimization

Our continuized framework for accelerating randomized gossip can be extended to the more general problem of decentralized optimization: each node v in the network G previously defined holds a function $f_v : \mathbb{R}^d \to \mathbb{R}$, μ-strongly convex and L-smooth. Nodes of the network collaborate to solve:

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} f_v(x) \right\}. \tag{25}$$

As in gossip averaging, only local communications are allowed. Quantities related to f_v can only be computed at node v. In the case of empirical risk minimization, f_v represents the empirical risk related to node v's local data. Setting $f_v(x) = \frac{1}{2}\|x - x_0(v)\|^2$ leads to the averaging problem previously described. Similarly to Section 6, time is indexed continuously by t in $\mathbb{R}_{\geqslant 0}$, and communications are ruled by the same Poisson point measure $\mathrm{d}N(t,e) = \sum_{k\geqslant 1} \delta_{(T_k,\{v_k,w_k\})}$ on $\mathbb{R}_{\geqslant 0} \times \mathcal{E}$. Yet, we no longer assume (as in Theorem 4) that the function f is quadratic. Instead, we write a dual formulation of Problem (25) and minimize it using a continuized version of accelerated coordinate descent [45] that we present in Appendix G. This leads to an accelerated decentralized algorithm to solve (25). Our algorithm mimics the behavior of accelerated randomized gossip: a node possesses two local parameters that mix continuously through a time-independent ODE. At time T_k, adjacent nodes v_k and w_k use their local function in order to compute gradient conjugates $\nabla f_v^*(x(v)), \nabla f_w^*(x(w))$. Since the local functions are not simple quadratics anymore, the stochastic gradients $\nabla f(x, \{v,w\})$ from Equation (26) are replaced by terms proportional to:

$$G(y, \{v,w\}) = \begin{cases} \nabla f_v^*(y(v)) - \nabla f_w^*(y(w)) & \text{at coordinate } v, \\ -\nabla f_v^*(y(v)) - \nabla f_w^*(y(w)) & \text{at coordinate } w, \\ 0 & \text{at all other coordinates.} \end{cases} \tag{26}$$

Due to lack of space, we describe the iterations more in details in Appendix H, together with a relevant choice of parameters. The crucial point is that, similarly to the gossip averaging case, we do not require nodes to be aware of a global iteration counter. Yet, we still obtain the same convergence rate as [25], as provided by the following theorem. The same approach can be used to "continuize" other accelerated randomized gossip algorithms for decentralized optimization, such as ADFS [26].

Theorem 6 (Accelerated asynchronous decentralized optimization)**.** *For* $(x_t(v))_{v\in\mathcal{V}} = (\nabla f_v^*(z_t(v)))_{v\in\mathcal{V}}$ *generated by the accelerated coordinate descent on the dual of Problem* (25)*:*

$$\sum_{v\in\mathcal{V}} \frac{1}{2}\mathbb{E}\left[\|x_t(v) - x_*\|^2\right] \leqslant C \left(\sum_{v\in\mathcal{V}} \frac{1}{2}\|x_0(v) - x_*\|^2\right) \exp\left(-\frac{\theta'_{\mathrm{ARG}}}{\sqrt{\kappa}} t\right),$$

where $\kappa = \mu/L$ is an upper bound on the condition number of f, C is a constant that depends on the graph and κ, and θ'_{ARG} is the rate of convergence of accelerated randomized gossip on the graph G as defined in Theorem 5 but with graph resistances are defined in a different way (see Theorem 10).

8 Conclusion

In this work, we introduced a continuized version of Nesterov's accelerated gradients. In a nutshell, the method has two sequences of iterates which take gradient steps at random times. In between gradient steps, the two sequences mix following a simple ordinary differential equation, whose parameters are picked to ensure 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, as the differential equation governing the mixing procedure has a simple analytical solution. A discretization of the continuized method corresponds to an accelerated gradient method with random parameters.

Continuization strategies were introduced in the context of Markov chains [1]. Here, they allow using acceleration mechanisms in asynchronous distributed optimization, where usually agents are not aware of the total number of iterations taken so far. This is showcased in the context of asynchronous gossip algorithms.

Acknowledgements: The authors thank Sam Power for pointing out the class of piecewise deterministic Markov processes and related references, and an anonymous reviewer for suggesting Remark 1. 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), from the DGA, and from the MSR-INRIA joint centre.

References

[1] David Aldous and James Allen Fill. Reversible markov chains and random walks on graphs. 2002. Unfinished monograph, recompiled 2014, available at http://www.stat.berkeley.edu/$\sim$aldous/RWG/book.html.

[2] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS '17, 2017.

[3] Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On lower and upper bounds in smooth and strongly convex optimization. *Journal of Machine Learning Research*, 17(126):1–51, 2016.

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

[5] Hedy Attouch, Zaki Chbani, and Hassan Riahi. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leqslant 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019.

[6] Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. Robust accelerated gradient methods for smooth strongly convex functions. *SIAM Journal on Optimization*, 30(1):717–751, 2020.

[7] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Accelerated gossip in networks of given dimension using jacobi polynomial iterations. *SIAM Journal on Mathematics of Data Science*, 2(1):24–47, 2020.

[8] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. In *Advances in Neural Information Processing Systems*, volume 33, pages 2576–2586, 2020.

[9] Michael Betancourt, Michael Jordan, and Ashia Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

[10] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[11] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

[12] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[13] Michael Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. 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, 2018.

[14] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. 2021.

[15] Mark HA Davis. Piecewise-deterministic markov processes: a general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3): 353–376, 1984.

[16] Mark HA Davis. *Markov models & optimization*. Routledge, 2018.

[17] Olivier Devolder. Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.

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

[19] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

[20] Alexandros D. G. Dimakis, Anand D. Sarwate, and Martin J. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Transactions on Signal Processing*, 56(3): 1205–1216, 2008. ISSN 1053-587X. doi: 10.1109/tsp.2007.908946. URL http://dx.doi.org/10.1109/TSP.2007.908946.

[21] Alexandros DG Dimakis, Anand D Sarwate, and Martin J Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Transactions on Signal Processing*, 56(3): 1205–1216, 2008.

[22] W. Ellens, F.M. Spieksma, P. Van Mieghem, A. Jamakovic, and R.E. Kooij. Effective graph resistance. *Linear Algebra and its Applications*, 435(10):2491–2506, 2011. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2011.02.024. URL https://www.sciencedirect.com/science/article/pii/S0024379511001443. Special Issue in Honor of Dragos Cvetkovic.

[23] Mathieu Even, Hadrien Hendrikx, and Laurent Massoulié. Decentralized optimization with heterogeneous delays: a continuous-time approach. *arXiv preprint arXiv:2106.03585*, 2021.

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

[25] Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. Accelerated decentralized optimization with local updates for smooth and strongly convex objectives, 2018.

[26] Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. An accelerated decentralized stochastic proximal algorithm for finite sums. *arXiv preprint arXiv:1905.11394*, 2019.

[27] Chonghai Hu, Weike Pan, and James Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 781–789, 2009.

[28] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014.

[29] Jean Jacod and Albert Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media, 2013.

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

[31] Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic.uchicago.edu/shai/papers/KakadeShalevTewari09.pdf*, 2009.

[32] Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1):81–107, 2016.

[33] Walid Krichene, Alexandre Bayen, and Peter Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems*, 28:2845–2853, 2015.

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

[35] Jonas Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31(4):1–25, 2021.

[36] Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274. Springer, 2016.

[37] Ji Liu, Brian D.O. Anderson, Ming Cao, and A. Stephen Morse. Analysis of accelerated gossip algorithms. *Automatica*, 49(4):873 – 883, 2013. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2013.01.001. URL http://www.sciencedirect.com/science/article/pii/S0005109813000022.

[38] Nicolas Loizou, Michael Rabbat, and Peter Richtárik. Provably accelerated randomized gossip algorithms. In *ICASSP 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 7505–7509. IEEE, 2019.

[39] Eduardo Montijano, Juan Montijano, and C. Sagues. Chebyshev polynomials in distributed consensus applications. *IEEE Transactions on Signal Processing*, 61, 11 2011. doi: 10.1109/TSP.2012.2226173.

[40] Michael Muehlebach and Michael Jordan. A dynamical systems perspective on Nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR, 2019.

[41] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

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

[43] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

[44] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[45] Yurii Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017. doi: 10.1137/16M1060182. URL https://doi.org/10.1137/16M1060182.

[46] Jesús María Sanz-Serna and Konstantinos Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *arXiv preprint arXiv:2009.00673*, 2020.

[47] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning*, 2017.

[48] Bin Shi, Simon Du, Michael Jordan, and Weijie Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

[49] Bin Shi, Simon Du, Weijie Su, and Michael Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, volume 32, pages 5744–5752, 2019.

[50] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27:2510–2518, 2014.

[51] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

[52] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

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

[54] Ashia Wilson, Benjamin Recht, and Michael I Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

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

[56] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.

[57] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems*, volume 31, pages 3900–3909, 2018.

A Numerical Simulations

A.1 Simulations of the discretized continuized acceleration

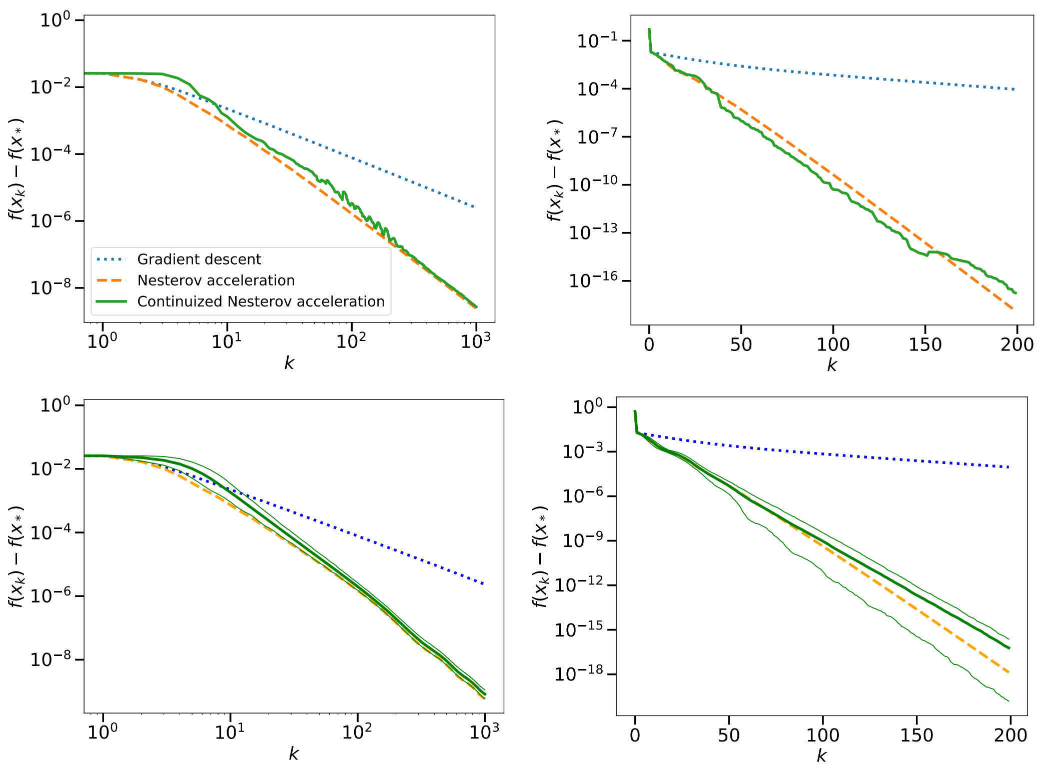


Figure 1: Comparison between gradient descent, Nesterov acceleration, and the continuized version of Nesterov acceleration, on a convex function (left plots) and a strongly convex function (right plots). For the continuized acceleration, which is randomized, the results shown in the above plots correspond to a single run. In the plots below, the thick line represents the average performance over $N = 1000$ runs of the continuized acceleration, while the thin lines represent the 5% and 95% quantiles.

In Figure 1, we compare this continuized Nesterov acceleration (12)-(14) with the classical Nesterov acceleration (3)-(5) and gradient descent. In the strongly convex case (right), we run the algorithms with the parameters of Theorem 1.(2) and 3.(2) on the function

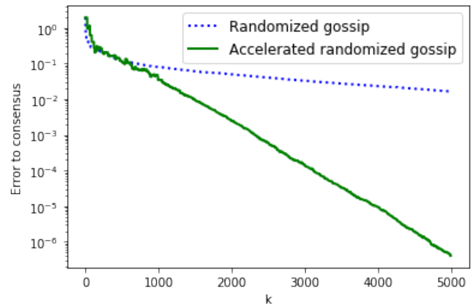
$$f(x_1, x_2, x_3) = \frac{\mu}{2}(x_1 - 1)^2 + \frac{3\mu}{2}(x_2 - 1)^2 + \frac{L}{2}(x_3 - 1)^2 \,,$$

with $\mu = 10^{-2}$ and $L = 1$. In the convex case, we run the algorithms with the parameters of Theorem 1.(1) and 3.(1) on the function

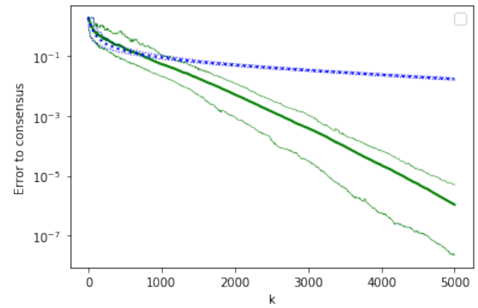
$$f(x_1, \dots, x_{100}) = \frac{1}{2} \sum_{i=1}^{100} \frac{1}{i^2} \left(x_i - \frac{1}{i} \right)^2 \,,$$

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

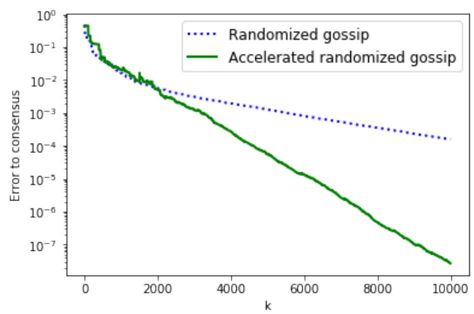
A.2 Simulation of Accelerated Randomized Gossip



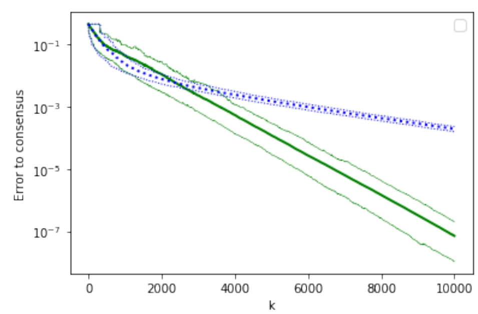
(a) Line graph, 30 nodes



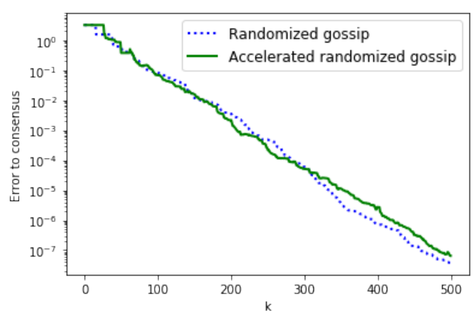
(b) Line graph, 30 nodes



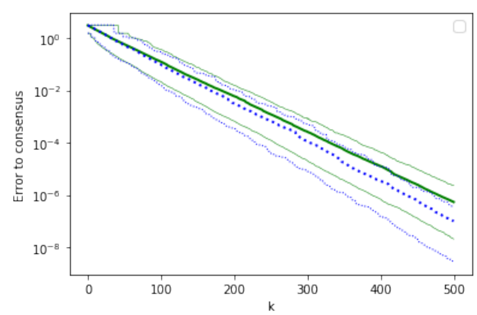
(c) 2D-Grid, 225 nodes



(d) 2D-Grid, 225 nodes



(e) Complete graph, 10 nodes



(f) Complete graph, 10 nodes

Figure 2: Comparison between randomized gossip [11] and accelerated randomized gossip from Section 6, on 3 different graphs: line with 30 nodes, 2D-Grid with 225 nodes and complete graph with 30 nodes. The probability $\mathcal{P}$ on the set of edges that determines at every activation which edge is activated is uniform in all cases. Parameters of the algorithm are taken as in Theorem 5. In all simulations, initialization was taken with a vector x_0 such that $x_0(v) = 0$ at all nodes, except one where $x_0(v) = 1$. Figures on the left represent one run of the algorithms. Figures on the right represent the average performance (thick line) for $N = 1000$ runes with the same settings, and the 5% and 95% quantiles (thin lines). As expected, we observe acceleration one the line and the grid, but no such phenomenon on the complete graph.

B Robustness of the continuized Nesterov acceleration to additive noise

In this section, we study the continuized acceleration (15)-(16) under stochastic gradients. We assume that our gradient estimates are unbiased, i.e.,

$$\forall x \in \mathbb{R}^d\,, \qquad \mathbb{E}_\xi \nabla f(x, \xi) = \nabla f(x)\,, \tag{27}$$

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

$$\forall x \in \mathbb{R}^d , \qquad \mathbb{E}_\xi \left\|\nabla f(x,\xi) - \nabla f(x)\right\|^2 \leqslant \sigma^2 . \tag{28}$$

These assumptions typically hold in the additive noise model, where $\nabla f(x, \xi) = \nabla f(x) + \xi$, and $\xi \in \mathbb{R}^d$ 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 (27) and (28) hold.

We should emphasize that similar studies of Nesterov acceleration under additive noise has been done [34, 27, 56, 17, 13, 6].

Theorem 7 (Continuized acceleration with additive noise)**.** *Assume that the stochastic gradients are unbiased* (27) *and have a variance uniformly bounded by* σ^2 (28)*. Then the continuized acceleration* (15)*-*(16) *satisfies the following.*

1. *For the parameters of Theorem 2.(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 2.(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 D.3.

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 3, 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} 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 7: the noise term is (sub-)linearly increasing in the convex case and constant in the strongly convex case.

Note that similarly to Theorem 3, one could obtain convergence bounds for the discrete implementation under the presence of additive noise.

C 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 [29], Ikeda and Watanabe [28], Le Gall [36].

C.1 Poisson point measures

We fix $\mathcal{P}$ a probability law on some space Ξ.

Definition 2. *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 $\mathrm{d}N(t, \xi) = \sum_{k \geqslant 1} \delta_{(T_k, \xi_k)}(\mathrm{d}t, \mathrm{d}\xi)$ 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,*

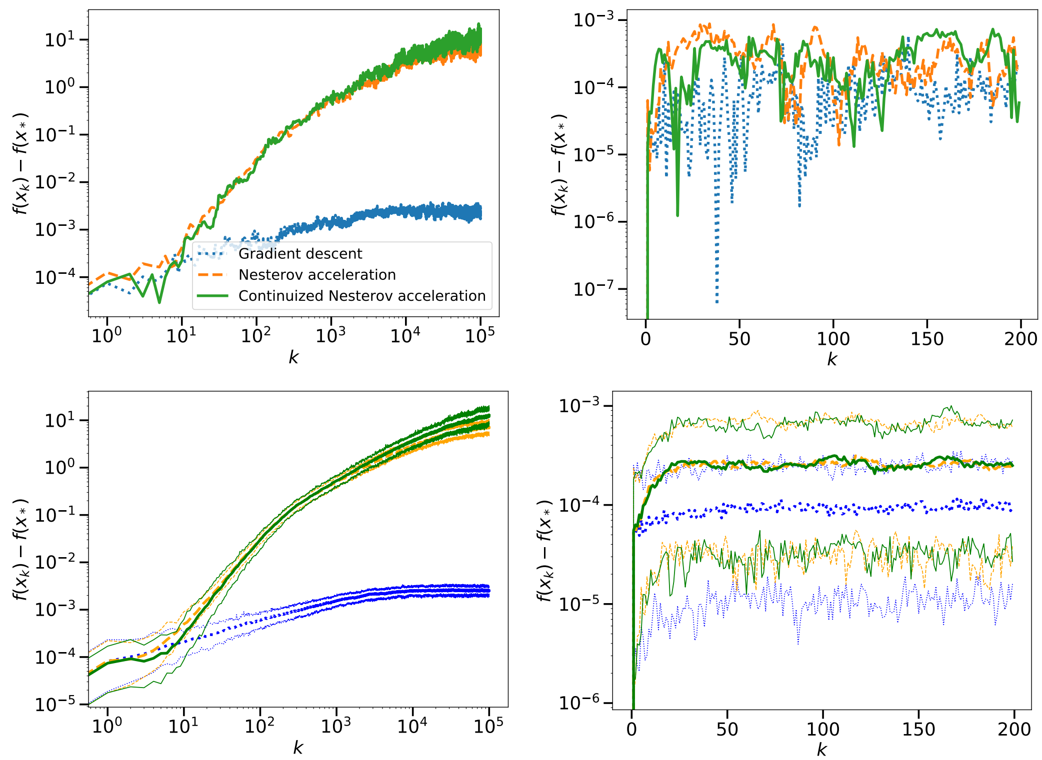


Figure 3: Effect of additive noise on gradient descent, Nesterov acceleration, and the continuized version of Nesterov acceleration, on a convex function (left) and a strongly convex function (right). All algorithms are started from the optimum x_*. The results shown in the above plots correspond to a single run. In the plots below, the thick line represents the average performance over $N = 100$ runs of each algorithm, while the thin lines represent the 5% and 95% quantiles.

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

Definition 3. *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*

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

C.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 4. *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 5. *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 6. *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 8 (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. $\mathbb{E}X_T \leqslant X_0$).*

C.3 Stochastic ordinary differential equation with Poisson jumps

The continuized processes are the composition of an ordinary differential equation and stochastic Poisson jumps. It is thus a piecewise-deterministic Markov process [15, 16], a special case of stochastic models that do not include any diffusion term. The stochastic calculus of this class of processes is particularly intuitive: there is no Ito correction term as with diffusive processes.

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 7. *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 $\mathrm{d}N(t, \xi) = \sum_{k\geqslant 1} \delta_{(T_k,\xi_k)}(\mathrm{d}t, \mathrm{d}\xi)$ given by Proposition 1, then

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

Here, we consider only autonomous equations as b and G are a function of x_t, but not of t. However, there is no loss of generality, one can study time-dependent systems by studying the equation in the variable (t, x_t). This trick is used in Appendix D.

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

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

Moreover, we have the decomposition

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

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

D 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.

D.1 Sketch of proof for Theorem 2

A complete and rigorous proof is given in Appendix D.2. 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, $t \geqslant 0$,

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

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

- x_t follows the linear ODE (8), 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

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

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

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

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 2 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 D.2.

D.2 Noiseless case: proofs of Theorem 2 and of the bounds of Theorem 3

In this section, we analyze the convergence of the continuized iteration (10)-(11), 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

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

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

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

we obtain:

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

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

$$\begin{aligned}
\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle &= \partial_t\varphi(\bar{x}_t) + \langle \partial_x\varphi(\bar{x}_t), \eta_t(z_t - x_t)\rangle + \langle \partial_z\varphi(\bar{x}_t), \eta_t'(x_t - z_t)\rangle \\
&= \frac{\mathrm{d}A_t}{\mathrm{d}t}\left(f(x_t) - f(x_*)\right) + \frac{\mathrm{d}B_t}{\mathrm{d}t}\frac{1}{2}\|z_t - x_*\|^2 + A_t\eta_t\langle \nabla f(x_t), z_t - x_t\rangle \\
&\quad + B_t\eta_t'\langle z_t - x_*, x_t - z_t\rangle\,.
\end{aligned}$$

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

$$\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle \leqslant \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\langle \nabla f(x_t), x_t - x_*\rangle + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2 \tag{29}$$

$$+ \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\frac{1}{2}\|z_t - x_*\|^2 + A_t\eta_t\langle \nabla f(x_t), z_t - x_*\rangle\,. \tag{30}$$

Further,

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

As f is L-smooth,

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

This gives

$$\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) \leqslant \left(B_t \gamma_t'^2 - A_t \gamma_t \left(2 - L\gamma_t\right)\right) \frac{1}{2}\|\nabla f(x_t)\|^2 - B_t \gamma_t' \langle \nabla f(x_t), z_t - x_*\rangle . \tag{31}$$

Finally, combining (29)-(30) with (31), we obtain

$$I_t \leqslant \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t \eta_t\right) \langle \nabla f(x_t), x_t - x_*\rangle + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t \eta_t'\right) \frac{1}{2}\|z_t - x_*\|^2 \tag{32}$$

$$+ (A_t \eta_t - B_t \gamma_t')\langle \nabla f(x_t), z_t - x_*\rangle + \left(B_t \eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right) \frac{1}{2}\|x_t - x_*\|^2 \tag{33}$$

$$+ \left(B_t \gamma_t'^2 - A_t \gamma_t \left(2 - L\gamma_t\right)\right) \frac{1}{2}\|\nabla f(x_t)\|^2 . \tag{34}$$

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

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

To satisfy the last equation, we choose

$$\gamma_t' = \sqrt{\frac{A_t}{L B_t}} . \tag{35}$$

To satisfy the third equation, we choose

$$\eta_t = \frac{B_t \gamma_t'}{A_t} = \sqrt{\frac{2 B_t}{L A_t}} . \tag{36}$$

To satisfy the fourth equation, we choose

$$\eta_t' = \frac{\mathrm{d}A_t}{\mathrm{d}t} \frac{\mu}{B_t} = \frac{A_t \eta_t \mu}{B_t} = \mu \sqrt{\frac{A_t}{L B_t}} . \tag{37}$$

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

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

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:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right) = \frac{1}{2\sqrt{A_t}} \frac{\mathrm{d}A_t}{\mathrm{d}t} = \frac{1}{2}\sqrt{\frac{B_t}{L}} , \qquad \frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{B_t}\right) = \frac{1}{2\sqrt{B_t}} \frac{\mathrm{d}B_t}{\mathrm{d}t} = \frac{\mu}{2}\sqrt{\frac{A_t}{L}} . \tag{38}$$

This can also be restated as

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\sqrt{A_t}\right) = \frac{\mu}{4L}\sqrt{A_t} , \qquad \sqrt{B_t} = 2\sqrt{L}\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right) . \tag{39}$$

D.2.1 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 (38), we have $\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{B_t}\right) = 0$, thus $B_t \equiv 1$, and $\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right) = \frac{1}{2\sqrt{L}}$, thus $\sqrt{A_t} = \frac{t}{2\sqrt{L}}$. The parameters of the algorithm are given by (35)-(37): $\eta_t = \frac{2}{t}$, $\eta'_t = 0$, $\gamma'_t = \frac{t}{2L}$ (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 2.

Further, one can apply martingale stopping Theorem 8 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 formula of Theorem 3.1.

D.2.2 Proof of the second part (strongly convex case)

We now assume $\mu > 0$. We consider the solution of (39) that is exponential:

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

The parameters of the algorithm are given by (35)-(37): $\eta_t = \eta'_t = \sqrt{\frac{\mu}{L}}$, $\gamma'_t = \frac{1}{\sqrt{\mu L}}$ (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 2. Similarly to above, one can also apply the martingale stopping theorem to prove the formula of Theorem 3.2.

Remark 2. *In the above derivation, in both the convex and strongly convex cases, we choose a particular solution of* (39)*, 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 $\sqrt{\mu}\sqrt{A_t} = \sqrt{B_t}$, 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 3 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 $\mu \to 0$.

D.3 With additive noise: proof of Theorem 7

The proof of this theorem is along the same lines as the proof of Theorem 2 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

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

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

$$\phi_t = \phi_0 + \int_0^t I_s \mathrm{d}s + M_t\,, \tag{40}$$

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 (29)-(30) holds. The computation of the second term becomes

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

As f is L-smooth,

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

By assumptions (27) and (28), 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

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

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

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

Combining the bounds, we obtain

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

which is an additive perturbation of the bound (32)-(34) in the noiseless case, with a perturbation proportional to σ^2. The choices of parameters of Theorem 2 cancel all first five prefactors, and satisfy $\gamma_t = \frac{1}{L}$, $A_t L\gamma_t^2 = B_t\gamma_t'^2$. We thus obtain

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

This bound controls the increase of ϕ_t. Using the decomposition (50), we obtain

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

D.3.1 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

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

D.3.2 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

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

D.4 With Pure Multiplicative Noise: Proof of Theorem 4

The proof of this theorem mimics the proof of Theorem 2, with a slightly different Lyapunov function. We recall that in Section 5, the function f is of the form:

$$\forall x \in \mathbb{R}^d, f(x) = \mathbb{E}\left[\frac{1}{2}(\langle a, x\rangle - b)^2\right],$$

where $\xi = (a, b) \in \mathbb{R}^d \times \mathbb{R}$ is of law $\mathcal{P}$. Thanks to the *noiseless assumption*, for $H = \mathbb{E}\left[aa^\top\right]$, we also have:

$$\forall x \in \mathbb{R}^d, f(x) = \frac{1}{2}\|x - x_*\|_H^2.$$

The Lyapunov function studied in the proof of Theorem 2 would then write as, for $t \in \mathbb{R}_{\geqslant 0}$:

$$\phi_t = \frac{A_t}{2}\|x_t - x_*\|_H^2 + \frac{B_t}{2}\|z_t - x_*\|^2.$$

An acceleration of stochastic gradient descent using this Lyapunov function has been done by Vaswani et al. [52]. In order to have an analysis similar to Nesterov acceleration, the authors make a strong growth condition, which is too strong for many stochastic gradient problems and for our application to gossip algorithms. Instead, our analysis requires a bounded statistical condition number $\tilde{\kappa}$, and performs a shift in terms of dependency over H: $\|x - x_*\|_H^2$ becomes $\|x - x_*\|^2$, and $\|z_t - x_*\|^2$ becomes $\|z_t - x_*\|_{H^{-1}}^2$. The new Lyapunov function writes:

$$\phi_t = \frac{A_t}{2}\|x_t - x_*\|^2 + \frac{B_t}{2}\|z_t - x_*\|_{H^{-1}}^2.$$

As in Theorem 2, the proof consists in proving that for carefully chosen parameters, ϕ_t is a supermatingale. 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:

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

where M_t is a martingale and

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

Since the Lyapunov function is not the same, we need to explicit here each term. The first term writes:

$$\begin{aligned}\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle =& \frac{1}{2}\frac{\mathrm{d}A_t}{\mathrm{d}t}\|x_t - x_*\|^2 + \frac{1}{2}\frac{\mathrm{d}B_t}{\mathrm{d}t}\|z_t - x_*\|^2_{H^{-1}} \\ &+ A_t\eta_t\langle x_t - x_*, z_t - x_t\rangle + B_t\eta'_t\langle H^{-1}(z_t - x_*), x_t - z_t\rangle.\end{aligned}$$

Mimicking the proof of Theorem 2, we write

$$\frac{1}{2}\|x_t - x_*\|^2 \leqslant \|x_t - x_*\|^2 - \frac{\mu}{2}\|x_t - x_*\|^2_{H^{-1}},$$

and

$$\begin{aligned}\langle H^{-1}(z_t - x_*), x_t - z_t\rangle &= \langle z_t - x_*, x_t - x_*\rangle_{H^{-1}} - \|z_t - x_*\|^2_{H^{-1}} \\ &\leqslant \frac{1}{2}\big(\|x_t - x_*\|^2_{H^{-1}} - \|z_t - x_*\|^2_{H^{-1}}\big).\end{aligned}$$

Hence:

$$\begin{aligned}\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle \leqslant& \frac{\mathrm{d}A_t}{\mathrm{d}t}\|x_t - x_*\|^2 + \left(B_t\eta'_t - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2_{H^{-1}} \\ &+ \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta'_t\right)\frac{1}{2}\|z_t - x_*\|^2_{H^{-1}} + A_t\eta_t\langle x_t - x_*, z_t - x_t\rangle\,.\end{aligned}$$

Further,

$$\begin{aligned}\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) =& \frac{A_t}{2}\left(\|x_t - \gamma_t\nabla f(x_t,\xi) - x_*\|^2 - \|x_t - x_*\|^2\right) \\ &+ \frac{B_t}{2}\left(\|(z_t - x_*) - \gamma'_t\nabla f(x_t,\xi)\|^2_{H^{-1}} - \|z_t - x_*\|^2_{H^{-1}}\right)\,.\end{aligned}$$

Then, expanding and taking expectation over ξ of the first term:

$$\begin{aligned}\mathbb{E}_\xi\left[\frac{1}{2}\|x_t - \gamma_t\nabla f(x_t,\xi) - x_*\|^2 - \frac{1}{2}\|x_t - x_*\|^2\right] &= \frac{\gamma_t^2}{2}\mathbb{E}_\xi\left[\|\nabla f(x_t,\xi)\|^2\right] - \gamma_t\langle H(x_t - x_*), x_t - x_*\rangle \\ &\leqslant \left(\frac{R^2\gamma_t^2}{2} - \gamma_t\right)\|x_t - x_*\|^2_H,\end{aligned}$$

where we used the definition of R^2 in Equation (19):

$$\begin{aligned}\mathbb{E}_\xi\left[\|\nabla f(x_t,\xi)\|^2\right] &= (x_t - x_*)^\top\mathbb{E}\left[aa^\top aa^\top\right](x_t - x_*) \\ &= (x_t - x_*)^\top\mathbb{E}\left[\|a\|^2 aa^\top\right](x_t - x_*) \\ &\leqslant R^2(x_t - x_*)^\top H(x_t - x_*).\end{aligned}$$

The second term writes:

$$\begin{aligned}\frac{1}{2}\mathbb{E}_\xi\left[\|(z_t - x_*) - \gamma'_t\nabla f(x_t,\xi)\|^2_{H^{-1}} - \|z_t - x_*\|^2_{H^{-1}}\right] &= \frac{\gamma_t'^2}{2}\mathbb{E}_\xi\left[\|\nabla f(x_t,\xi)\|^2_{H^{-1}}\right] \\ &\quad - \gamma'_t\langle x_t - x_*, z_t - x_*\rangle \\ &\leqslant \frac{\tilde{\kappa}\gamma_t'^2}{2}\|x_t - x_*\|^2_H \\ &\quad - \gamma'_t\langle x_t - x_*, z_t - x_*\rangle,\end{aligned}$$

where we used the definition of $\tilde{\kappa}$ in Equation (20):

$$\begin{aligned}\mathbb{E}_\xi\left[\|\nabla f(x_t,\xi)\|^2_{H^{-1}}\right] &= (x_t - x_*)^\top\mathbb{E}\left[aa^\top H^{-1}aa^\top\right](x_t - x_*) \\ &= (x_t - x_*)^\top\mathbb{E}\left[a\|a\|^2_{H^{-1}}a^\top\right](x_t - x_*) \\ &\leqslant \tilde{\kappa}(x_t - x_*)^\top H(x_t - x_*).\end{aligned}$$

Combining these inequalities gives the following upper-bound on I_t:

$$I_t \leqslant \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\|x_t - x_*\|^2 + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta'_t\right)\frac{1}{2}\|z_t - x_*\|^2_{H^{-1}}$$
$$+ (A_t\eta_t - B_t\gamma'_t)\langle x_t - x_*, z_t - x_*\rangle + \left(B_t\eta'_t - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2_{H^{-1}}$$
$$+ \left(\tilde{\kappa}B_t\gamma'^2_t - A_t\gamma_t\left(2 - R^2\gamma_t\right)\right)\frac{1}{2}\|x_t - x_*\|^2_H$$

Since $I_t \leqslant 0$ is still a sufficient condition for ϕ_t to be a supermartingale, we choose parameters such that all prefactors are equal to 0. We first take $\gamma_t = \frac{1}{R^2}$, and we want to satisfy:

$$\frac{\mathrm{d}A_t}{\mathrm{d}t} = A_t\eta_t\,, \qquad \frac{\mathrm{d}B_t}{\mathrm{d}t} = B_t\eta'_t \qquad A_t\eta_t = B_t\gamma'_t\,, \qquad B_t\eta'_t = \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\,, \qquad B_t\gamma'^2_t = \frac{A_t}{\tilde{\kappa}R^2}\,.$$

To satisfy that last equality, we choose:

$$\gamma'_t = \sqrt{\frac{A_t}{B_t\tilde{\kappa}R^2}}.$$

The rest of the proof then follows just as in the proof of Theorem D.2.

E Proof of Theorem 3

By integrating the ODE

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

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

$$\tilde{y}_k = x_{T_{k+1}-} = x_{T_k} + \tau_k(z_{T_k} - x_{T_k}) = \tilde{x}_k + \tau_k(\tilde{z}_k - \tilde{x}_k)\,, \tag{41}$$
$$z_{T_{k+1}-} = z_{T_k} + \tau''_k(x_{T_k} - z_{T_k}) = \tilde{z}_k + \tau''_k(\tilde{x}_k - \tilde{z}_k)\,.$$

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,

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

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

Further, from (6)-(7), we obtain the equations

$$\tilde{x}_{k+1} = x_{T_{k+1}} = x_{T_{k+1}-} - \gamma_{T_{k+1}}\nabla f(x_{T_{k+1}-}) = \tilde{y}_k - \gamma_{T_{k+1}}\nabla f(\tilde{y}_k)\,, \tag{42}$$
$$\tilde{z}_{k+1} = z_{T_{k+1}} = z_{T_{k+1}-} - \gamma'_{T_{k+1}}\nabla f(x_{T_{k+1}-}) = \tilde{z}_k + \tau'_k(\tilde{y}_k - \tilde{z}_k) - \gamma'_{T_{k+1}}\nabla f(\tilde{y}_k)\,. \tag{43}$$

The stated equation (12)-(14) are the combination of (41), (42) and (43).

1. The parameters of Theorem 2.(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

   $$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t = \frac{2}{t}(z_t - x_t)\mathrm{d}t\,,$$
   $$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t = 0\,,$$

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

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

   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 2.(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 $\tilde{\gamma}'_k = \gamma'_{T_k} = \frac{1}{\sqrt{\mu L}}$.

F Heuristic ODE scaling limit of the continuized acceleration

F.1 Convex case

With the choices of parameters of Theorem 2.(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

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

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. [50] for Nesterov acceleration.

F.2 Strongly-convex case

With the choices of parameters of Theorem 2.(2), the continuized acceleration is

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

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

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

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

$$\frac{\mathrm{d}x_s}{\mathrm{d}s} = \sqrt{\mu}(z_s - x_s)\,, \tag{44}$$

$$\frac{\mathrm{d}z_s}{\mathrm{d}s} = \sqrt{\mu}(x_s - z_s) - \frac{1}{\sqrt{\mu}}\nabla f(x_s)\,. \tag{45}$$

From (44), we have $z_s = x_s + \frac{1}{\sqrt{\mu}}\frac{\mathrm{d}x_s}{\mathrm{d}s}$, and by substitution in (45), we obtain

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

This is the so-called "low-resolution" ODE for Nesterov acceleration of Shi et al. [48].

G Continuized Accelerated Coordinate Descent with arbitrary sampling

In this section, we focus on the following problem:

$$\min_{x\in\mathbb{R}^d} f(x), \tag{46}$$

where f is of the form $f : x \mapsto g(Rx)$ for some function g and projector $R \in \mathbb{R}^{d\times d}$ (such that $R^2 = R$). We further assume that f is smooth with respect to some matrix $M \in \mathbb{R}^{d\times d}$ and μ-strongly convex with respect to R, *i.e.*:

$$\frac{\mu}{2}\|x-y\|_R^2 \leqslant f(x) - f(y) - \nabla f(x)^\top(x-y) \leqslant \frac{1}{2}\|x-y\|_M^2.$$

Note that μ can be equal to zero, but convergence will be slower in this case. We analyze the convergence of the following continuized coordinate descent iteration:

$$\begin{aligned} \mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t\int_\Xi \frac{R_{\xi\xi}}{\mathcal{P}_\xi}\nabla f(x_t,\xi)\mathrm{d}N(t,\xi)\,, \\ \mathrm{d}z_t &= \eta_t'(x_t - z_t)\mathrm{d}t - \gamma_t'\int_\Xi \nabla f(x_t,\xi)\mathrm{d}N(t,\xi)\,, \end{aligned} \tag{47}$$

where

$$\nabla f(x_t,\xi) = \frac{1}{\mathcal{P}_\xi}\nabla_\xi f(x_t), \tag{48}$$

with the coordinate gradient $\nabla_\xi f(x_t) = e_\xi e_\xi^\top \nabla f(x_t)$, with $e_\xi \in \mathbb{R}^d$ the unit vector associated with coordinate $\xi \in \{1,\dots,d\}$ and $\mathcal{P}_\xi$ and $\mathrm{d}N$ are defined as in Section 6. Note that these iterations are slightly different from the previous stochastic gradient iteration since the stochastic gradient is not the same for x_t and z_t (same direction but different magnitudes). The following theorem is a continuized version of Hendrikx et al. [25], which is itself largely based on Nesterov and Stich [45].

Theorem 9 (Continuized acceleration of coordinate descent)**.** *Assume that the stochastic gradients are of the coordinate descent form* (48)*. Besides, choose parameter L such that:*

$$L \geqslant \max_{\xi \in \Xi} \frac{M_{\xi\xi} R_{\xi\xi}}{\mathcal{P}_\xi^2} \,. \tag{49}$$

Then the continuized acceleration (60) *satisfies the following:*

1. *For $\eta_t = \frac{2}{t}, \eta'_t = 0, \gamma_t = \frac{1}{L}, \gamma'_t = \frac{t}{2L}$,*

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

2. *Assume further that $\mu > 0$ and 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 ,*

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

Proof. Similarly to the proof in Appendix D.3, the proof of this theorem is along the same lines as the proof of Theorem 2, and we only highlight the major differences. The process $\bar{x}_t = (t, x_t, z_t)$ satisfies the equation

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

We also consider a slightly different Lyapunov function ϕ_t that takes into account the projector R:

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

This change of norm is essential to take into account the fact that f is not strongly convex with respect to the euclidean norm, but only with respect to $\|\cdot\|_R$. We apply Proposition 2 to $\phi_t = \varphi(\bar{x}_t) = \varphi(t, x_t, z_t)$ and obtain

$$\phi_t = \phi_0 + \int_0^t I_s \mathrm{d}s + M_t \,, \tag{50}$$

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 (29)-(30) holds. The computation of the second term becomes

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

As f is M-smooth,

$$f\left(x_t - \gamma_t \frac{R_{\xi\xi}}{\mathcal{P}_\xi} \nabla f(x_t, \xi) \right) - f(x_t) \leqslant \langle \nabla f(x_t), -\gamma_t \frac{R_{\xi\xi}}{\mathcal{P}_\xi} \nabla f(x_t, \xi) \rangle + \frac{1}{2}\|\gamma_t \frac{R_{\xi\xi}}{\mathcal{P}_\xi} \nabla f(x_t, \xi)\|_M^2 \,.$$

In the additive case, the variance is bounded by σ^2. In this case, we have that:

$$\|\frac{R_{\xi\xi}}{\mathcal{P}_\xi} \nabla f(x_t, \xi)\|_M^2 = \frac{M_{\xi\xi} R_{\xi\xi}}{\mathcal{P}_\xi^2} \|\nabla f(x_t, \xi)\|_R^2 \leqslant L\|\nabla f(x_t, \xi)\|_R^2, \tag{51}$$

and similarly:

$$\langle \nabla f(x_t), -\gamma_t \frac{R_{\xi\xi}}{\mathcal{P}_\xi} \nabla f(x_t, \xi) \rangle = -\gamma_t \frac{R_{\xi\xi}}{\mathcal{P}_\xi^2} \|\nabla_\xi f(x_t)\|^2 = \gamma_t \|\nabla f(x_t, \xi)\|_R^2. \tag{52}$$

Thus:

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

Similarly, thanks to the unbiasedness of $\nabla f(x_t,\xi)$,

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

This gives

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

Combining the bounds, we obtain

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

We see that we obtain a result that is very similar to that of the deterministic case. The choices of parameters of Theorem 9 cancel all first five prefactors, and satisfy $\gamma_t = \frac{1}{L}$, $A_t L\gamma_t^2 = B_t\gamma_t'^2$. We thus obtain $I_t \leqslant 0$ and so ϕ_t is a supermartingale, and the rest follows as in Appendix D.2. □

H Accelerated Decentralized Optimization with Randomized Gossip Communications.

We now consider the setting of decentralized optimization considered in Section 7. More specifically, recall that we wish to solve:

$$\min_{x\in\mathbb{R}^d}\left\{f(x) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} f_v(x)\right\}, \tag{53}$$

where the function f_v is privately held by node $v \in \mathcal{V}$. To solve this problem, a classical approach is to use a dual formulation [47, 25]. We first rewrite Problem (53) as:

$$\min_{X\in\mathbb{R}^{|\mathcal{V}|\times d},\ X_u = X_v\ \forall\{u,v\}\in\mathcal{E}}\left\{F(X) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} f_v(X_v)\right\}, \tag{54}$$

where $X_v \in \mathbb{R}^d$ corresponds to the local parameter of node v, and the equality constraints ensures equivalence between (53) and (54). The constraints are linear and can be expressed in matrix form as:

$$A^\top X = 0, \tag{55}$$

with $A \in \mathbb{R}^{\mathcal{E}\times\mathcal{V}}$ such that $\ker(A^\top) = \mathrm{Span}(1, ..., 1)$ the constant vector. The natural choice for matrix A is to choose a square root of the Laplacian matrix of graph G. For $(e_v)_{v\in\mathcal{V}}$ and $(e_{\{v,w\}})_{\{v,w\}\in\mathbb{E}}$ the canonical bases of $\mathbb{R}^{\mathcal{V}}$ and $\mathbb{R}^{\mathcal{E}}$, A is thus that for any $\{v,w\} \in \mathcal{E}$:

$$Ae_{\{v,w\}} = \sqrt{\mathcal{P}_{\{v,w\}}}(e_v - e_w).$$

Matrix A then satisfies $AA^\top = \mathcal{L}$ the Laplacian matrix of graph G with weights $\mathcal{P}_{\{v,w\}}$. Indeed, if $W_{\{v,w\}} = \mathcal{P}_{\{v,w\}}(e_v - e_w)(e_v - e_w)^\top$ corresponds to the gossip matrix for edge $\{v,w\}$, A is such that:

$$AA^\top = \sum_{\{v,w\}\in\mathcal{E}} W_{\{v,w\}} = \mathcal{L}. \tag{56}$$

Then, introducing Lagrange multipliers λ, we obtain through Lagrangian duality that Problem (53) is equivalent to:

$$\max_{\lambda\in\mathbb{R}^{\mathcal{E}\times d}} -F^*(A\lambda), \tag{57}$$

with F^* the convex conjugate of F. Following the approach of Hendrikx et al. [25], we then apply Accelerated Coordinate Descent to this dual problem. Yet, we use the *continuized* version of Theorem 9, which allows us to remove the global iterations counter on which previous approaches rely. We see that Problem (57) has exactly the right form to apply Theorem 9, leading to the following dual iterations:

$$\begin{aligned}
\mathrm{d}\lambda_t^{(y)} &= \eta_t(\lambda_t^{(z)} - \lambda_t^{(y)})\mathrm{d}t - \gamma_t \int_{\mathbb{R}_{\geqslant 0}\times\mathcal{E}} \frac{R_{\{v,w\}}}{\mathcal{P}_{\{v,w\}}^2} e_{\{v,w\}} e_{\{v,w\}}^\top A^\top \nabla F^*(A\lambda_t^{(y)})\mathrm{d}N(t,\{v,w\})\,, \\
\mathrm{d}\lambda_t^{(z)} &= \eta_t'(\lambda_t^{(y)} - \lambda_t^{(z)})\mathrm{d}t - \gamma_t' \int_{\mathbb{R}_{\geqslant 0}\times\mathcal{E}} \frac{1}{\mathcal{P}_{\{v,w\}}} e_{\{v,w\}} e_{\{v,w\}}^\top A^\top \nabla F^*(A\lambda_t^{(y)})\mathrm{d}N(t,\{v,w\})\,,
\end{aligned} \tag{58}$$

where $P = A^\dagger A$ with $A^\dagger$ is the pseudo-inverse of A, $R_{\{v,w\}} = e_{\{v,w\}}^\top A^\dagger A e_{\{v,w\}}$. Now, we multiply these iterations by A on the left (which is standard), and we rewrite them with the following iterates:

$$y_t = A\lambda_t^{(y)}, \qquad z_t = A\lambda_t^{(z)}. \tag{59}$$

Note that $y_t, z_t \in \mathbb{R}^{|\mathcal{V}|\times d}$, and are thus variables associated with *nodes* of the graph.

$$\begin{aligned}
\mathrm{d}y_t &= \eta_t(z_t - y_t)\mathrm{d}t - \gamma_t \int_{\mathbb{R}_{\geqslant 0}\times\mathcal{E}} \frac{R_{\{v,w\}}}{\mathcal{P}_{\{v,w\}}^2} W_{\{v,w\}} \nabla F^*(y_t)\mathrm{d}N(t,\{v,w\})\,, \\
\mathrm{d}z_t &= \eta_t'(y_t - z_t)\mathrm{d}t - \gamma_t' \int_{\mathbb{R}_{\geqslant 0}\times\mathcal{E}} \frac{1}{\mathcal{P}_{\{v,w\}}} W_{\{v,w\}} \nabla F^*(y_t)\mathrm{d}N(t,\{v,w\})\,,
\end{aligned} \tag{60}$$

where we recall that $W_{\{v,w\}} = \mathcal{P}_{\{v,w\}}(e_v - e_w)(e_v - e_w)^\top$ corresponds to the gossip matrix for edge $\{v,w\}$. Besides, the dual gradients $\nabla F^*(y_t)$ are such that $e_v^\top \nabla F^*(y_t) = \nabla f_v^*(e_v^\top y_t)$, and so each component can be computed locally at node v.

In summary, the distributed decentralized algorithm writes as follows. Upon activation of edge $\{v_k, w_k\}$ at time T_k,

$$\begin{aligned}
G_{\{v_k,w_k\}}(T_k) &= \omega_{\{v_k,w_k\}}\left[\nabla f^*((y_{T_k^-})_{v_k}) - \nabla f^*((y_{T_k^-})_{w_k})\right] \\
y_{T_k}(v_k) &= y_{T_k^-}(v_k) - \gamma_t \frac{R_{\{v_k,w_k\}}}{\mathcal{P}_{\{v_k,w_k\}}^2} G_{\{v_k,w_k\}}(T_k)\,, \\
y_{T_k}(w_k) &= y_{T_k^-}(w_k) + \gamma_t \frac{R_{\{v_k,w_k\}}}{\mathcal{P}_{\{v_k,w_k\}}^2} G_{\{v_k,w_k\}}(T_k)\,, \\
z_{T_k}(v_k) &= z_{T_k^-}(v_k) - \gamma_t' G_{\{v_k,w_k\}}(T_k)\,, \\
z_{T_k}(w_k) &= z_{T_k^-}(w_k) + \gamma_t' G_{\{v_k,w_k\}}(T_k)\,.
\end{aligned} \tag{61}$$

Between these updates, $y_t(v)$ and $z_t(v)$ locally mix at all nodes $v \in \mathcal{V}$, according to the coupled ODE:

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

This algorithm can be implemented with local computations and pairwise communications only, since an update along edge $\{v,w\}$ only involves the parameters and functions of nodes v and w. In order to fully describe this algorithm, we need to specify the various parameters. We do so, with the corresponding rate of convergence, in the following theorem.

Theorem 10 (Asynchronous Accelerated Decentralized Optimization)**.** *Assume that each f_v is μ-strongly-convex with $\mu > 0$ and L-smooth. Let $L_{\mathrm{dual}} = \frac{1}{\mu}\max_{\{v,w\}} \frac{R_{\{v,w\}}}{\mathcal{P}_{\{v,w\}}}$, where we recall* *that* $R_{\{v,w\}} = (A^\dagger A)_{\{v,w\},\{v,w\}}$*. Then, let* $\theta'_{\rm ARG} = \sqrt{\mu_{\rm gossip}/\max_{\{v,w\}}\frac{R_{\{v,w\}}}{\mathcal{P}_{\{v,w\}}}}$ *where* $\mu_{\rm gossip}$ *is the smallest non-zero eigenvalue of the Laplacian of the graph* $\mathcal{G}$*, and* $\kappa = L/\mu$ *is a bound on the condition number of* f*. We choose the constant parameters* $\eta_t = \eta'_t \equiv \frac{\theta'_{\rm ARG}}{\sqrt{\kappa}}$*,* $\gamma_t \equiv \frac{1}{L_{\rm dual}}$*,* $\gamma'_t \equiv \sqrt{\frac{L}{\mu_{\rm gossip}L_{\rm dual}}}$*. The iterates produced by the algorithm described in* (61) *verify:*

$$\mathbb{E}\sum_{v\in V}\frac{1}{2}\|\nabla f_v^*(z_t(v)) - x_\star\|^2 \leqslant C_0^{\rm dual}\exp\left(-\frac{\theta'_{\rm ARG}}{\sqrt{\kappa}}t\right),$$

with $C_0^{\rm dual} = \frac{\lambda_{\max}(AA^\top)}{\mu}\left(F^*(A\lambda_0^{(y)}) - F^*(A\lambda_\star) + \frac{\mu_{\rm gossip}}{2L}\|\lambda_0^{(z)} - \lambda_\star\|^2_{A^\dagger A}\right)$*, with* $\lambda_\star$ *a solution to the dual problem.*

Note that $\theta'_{\rm ARG}$ is slightly different from $\theta_{\rm ARG}$. Yet, following Hendrikx et al. [25], an equivalent of Corollary 1 can be obtained for $\theta'_{\rm ARG}$. To obtain Theorem 6, we simply choose $\lambda_0^{(y)} = \lambda_0^{(z)}$ and bound the dual function suboptimality by the distance to optim using the smoothness and strong convexity of F^*.

We stress the fact that the accelerated algorithm described in this section, as well as accelerated randomized gossip in Section 6, are decentralized and asynchronous: operations are local and do not require any global synchronization, provided that a continuous time clock can be shared. This is possible only thanks to the continuized framework. However, there are some limitations: even if these algorithms are the first to achieve these rates without any global synchronization, computations and communications are here assumed to happen instantly, or to take a negligible time. Handling communication and computation physical capacity constraints such as delays or node/edge overloads in our algorithms as in [23] combined with accelerated schemes is left for future works.

Proof. First note that the Hessian of the dual objective writes for some $\lambda \in \mathbb{R}^{|\mathcal{E}|\times d}$:

$$A^\top\nabla^2 F^*(A\lambda)A \succcurlyeq \frac{1}{L}A^\top A, \tag{62}$$

since F^* is L^{-1} strongly-convex when F is L-smooth [31]. Thus, the dual objective is $\mu_{\rm gossip}/L$ strongly convex on the orthogonal of the kernel of A. Similarly, the smoothness of the dual objective in direction $\{v,w\}$ is equal to:

$$M_{\{v,w\}\{v,w\}} = e_{\{v,w\}}^\top A^\top\nabla^2 F^*(A\lambda)Ae_{\{v,w\}} \preccurlyeq \frac{1}{\mu}e_{\{v,w\}}^\top A^\top Ae_{\{v,w\}} = \frac{\mathcal{P}_{\{v,w\}}}{2\mu}. \tag{63}$$

Thus, we have that:

$$L_{\rm dual} = \max_{\{v,w\}}\frac{M_{\{v,w\}\{v,w\}}R_{\{v,w\}}}{\mathcal{P}^2_{\{v,w\}}} = \frac{1}{\mu}\max_{\{v,w\}}\frac{R_{\{v,w\}}}{\mathcal{P}_{\{v,w\}}}. \tag{64}$$

Then, the result follows directly from applying Theorem (9), together with the smoothness of the dual gradients, since:

$$\mathbb{E}\sum_{v\in V}\frac{1}{2}\|\nabla f_v^*(z_t(v)) - x_\star\|^2 \leqslant \mathbb{E}\frac{1}{2\mu}\|A\lambda_t^{(z)} - A\lambda_\star\|^2 \leqslant \frac{\lambda_{\max}(AA^\top)}{2\mu}\mathbb{E}\|\lambda_t^{(z)} - \lambda_\star\|^2_R. \tag{65}$$

□

Note that the primal parameter that we are interested in is $x_t = \nabla f^*(z_t)$, and not y_t or z_t which are dual parameters.