

Continuized Nesterov Momentum Achieves the $\mathcal{O}(\varepsilon^{-7/4})$ Complexity in Smooth Nonconvex Optimization

Julien Hermant^{*†} Jean-François Aujol[†] Charles Dossal[‡] Lorick Huang[‡]
 Aude Rondepierre[‡] Irène Waldspurger[§]

Abstract

For first-order optimization of non-convex functions with Lipschitz continuous gradient and Hessian, the best known complexity for reaching an ε -approximation of a stationary point is $\mathcal{O}(\varepsilon^{-7/4})$. The existing algorithms achieving this bound are based on momentum, but are always complemented with safeguard mechanisms that erase the accumulated momentum if a certain condition is violated. Whether such momentum-control mechanisms are fundamentally necessary has remained an open question. We show that randomizing the parameters enables to achieve this complexity in expectation when using momentum without any of such mechanisms, and we improve the numerical constant factor of the bound in the case of a large enough number of iterations. From an analysis perspective, we do so leveraging the continuized method, that interprets the algorithm as a realization of a continuous-time stochastic differential equation (SDE) involving a Poisson process. We show that this SDE converges in probability to the Heavy Ball ordinary differential equation when the stepsize goes to zero, paralleling the behavior of more classical instances of Nesterov momentum.

1 Introduction

Many real-world problems can be formulated as the minimization of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, that is, as $\min_{x \in \mathbb{R}^d} f(x)$. Yet, without structural assumptions - such as convexity of f - computing a global minimizer is generally intractable [22, Section 2.2]. In this work, assuming f is lower-bounded and differentiable, our goal is thus to find a stationary point of f , *i.e.* approximate $x^* \in \mathbb{R}^d$ such that

$$\nabla f(x^*) = 0.$$

The rising computational cost and expanding data scales in modern applications-ranging from machine learning to signal processing - have made *first-order optimization algorithms* widely used. These algorithms rely on gradient information to compute their iterates with gradient descent being a classic example. Among existing acceleration techniques designed to improve their convergence speed [38, 44], one of the most influential is *momentum*, whose idea dates back to the seminal work of [60].

Nesterov momentum The *Nesterov momentum* (NM) algorithm [53], sometimes called *Nesterov accelerated gradient*, is known to achieve optimal convergence bounds among first-order algorithms in various settings in convex optimization [52, 55], see Appendix A.1 for a more detailed discussion. There exist several ways to write Nesterov's momentum, see [24] for a discussion in the strongly-convex case, or see [36, Appendix B.2] in a more general setting. Among these different formulations, we consider the following

$$\begin{cases} \tilde{y}_k &= \tilde{x}_k + \alpha_k(\tilde{z}_k - \tilde{x}_k), \\ \tilde{x}_{k+1} &= \tilde{y}_k - \gamma \nabla f(\tilde{y}_k), \\ \tilde{z}_{k+1} &= \tilde{z}_k + \beta_k(\tilde{y}_k - \tilde{z}_k) - \gamma' \nabla f(\tilde{y}_k). \end{cases} \quad (\text{NM})$$

^{*}corresponding author: julien.hermant@math.u-bordeaux.fr

[†]Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400 Talence, France

[‡]IMT, Univ. Toulouse, INSA Toulouse, Toulouse, France

[§]CNRS, Univ. Paris Dauphine, Inria Mokaplan, France

We note that $\{\tilde{x}_k\}_{k \in \mathbb{N}}$ in (NM) reduces to gradient descent for the choice $\alpha_k = 0$, for all $k \in \mathbb{N}$. Because a wide range of applications is formulated as a non-convex optimization problem [32, 14, 23, 11, 27, 26, 45], an important question is whether the benefits of momentum extend beyond convex optimization. Empirically, it is often observed that the use of momentum improves the speed of convergence for non-convex minimization tasks such as neural network training [66, 33]. Theoretically, acceleration has been shown in some mild non-convex settings where finding a global minimizer remains tractable, see *e.g.* [37, 35, 31]. In this work, we only adopt *smoothness* assumptions, which fall outside regimes where guarantees of global minimization can be obtained.

Lipschitz Gradient When the objective f has a Lipschitz continuous gradient, finding an ε -stationary point, *i.e.* $x \in \mathbb{R}^d$ such that $\|\nabla f(x)\| \leq \varepsilon$, can be achieved by gradient descent in at most $\mathcal{O}(\varepsilon^{-2})$ gradient evaluations [55, (1.2.22)]. This bound is worst-case optimal among first-order algorithms [17]. The momentum mechanism thus does not provably accelerate gradient descent, achieving a similar rate [28]. This impossibility result suggests that the class of functions with merely Lipschitz-continuous gradients does not provide an adequate framework for understanding the acceleration mechanisms observed in practice. This motivates the study of higher-order smoothness assumptions.

Lipschitz Hessian Assuming in addition that the Hessian is Lipschitz continuous, the aforementioned bound can be improved to $\tilde{\mathcal{O}}(\varepsilon^{-7/4})$ [15, 41], up to some logarithmic factor further removed by [46] to achieve the bound $\mathcal{O}(\varepsilon^{-7/4})$. This class of functions therefore admits accelerated bounds, while still encompassing important non-convex problems. Examples include the symmetric low-rank matrix factorization problem, as long as the domain is a compact set [40], or squared loss regression of a neural network whose activation functions are with Lipschitz first and second derivatives, *e.g.* if using SoftPlus activations [62, Proposition 2].

The existing first-order algorithms achieving the $\mathcal{O}(\varepsilon^{-7/4})$ or $\tilde{\mathcal{O}}(\varepsilon^{-7/4})$ complexity rely on the following update as a core mechanism

$$\begin{cases} \tilde{y}_k = \tilde{x}_k + \alpha_k(\tilde{x}_k - \tilde{x}_{k-1}), \\ \tilde{x}_{k+1} = \tilde{x}_k - \gamma \nabla f(\tilde{y}_k). \end{cases} \quad (\text{NM}') \quad (1)$$

(NM') can be described as a Nesterov momentum algorithm, with fewer degrees of freedom compared with the 3-sequence version (NM). This can be seen as (NM) can be written as a 2-sequence form that reduces to (NM') for some parameter choices [35, Proposition 11], or by considering that the high-resolution limit-ODE associated to (NM) and (NM') can be written in the same form [34, Proposition 27]. However, these methods achieving the improved complexity do not rely on (NM') only; they also need to perform a *safety check* at each iteration. If it fails at iteration k , the **momentum built up so far is erased**, which is practically done by setting $\alpha_k = 0$. At a conceptual level, these procedures can be summarized as follows:

Algorithm 1: Structure of existing algorithms achieving the $\mathcal{O}(\varepsilon^{-7/4})$ complexity

- 1 **if** *Safety check succeeds* **then**
 - 2 | Run update rule (NM')
 - 3 **else**
 - 4 | Erase accumulated momentum (+ potentially trigger alternative mechanisms)
-

In [15, 41], the safety check monitors the level of non-convexity between consecutive iterates. When it exceeds a given threshold:

$$f(\tilde{x}_k) \leq f(\tilde{y}_k) + \langle \nabla f(\tilde{y}_k), \tilde{x}_k - \tilde{y}_k \rangle - \frac{\ell}{2} \|\tilde{x}_k - \tilde{y}_k\|^2, \quad \ell > 0,$$

accumulated momentum is erased, and a negative-curvature exploitation step is triggered as an alternative mechanism. An alternative is provided in [46], which employs a safety check based on the length of the trajectory residual. Exceeding a given threshold

$$k \sum_{t=0}^{k-1} \|\tilde{x}_{k+1} - \tilde{x}_k\|^2 > B^2, \quad B > 0,$$

it activates a restart mechanism, thus erasing the accumulated momentum. See detailed algorithmic descriptions and further discussion in Appendix A.3.

In both cases, failing the safety check triggers a **momentum reset**. Momentum resets, or momentum restarts, can be interesting. They have been shown to provide benefits in the convex setting, for example by enhancing practical behavior in some situations [58, 62], or by helping in the design of parameter-free algorithms [6]. However, such mechanisms are not mandatory to achieve theoretical acceleration in the convex setting. More generally, mechanisms that restart or control momentum do not seem to be necessary for momentum methods to perform well in many practical non-convex problems. It may be somewhat contradicted by the existing results, which suggest that these momentum control are needed to ensure accelerated convergence. This raises the following question:

Does momentum fundamentally requires control mechanisms to achieve the $\mathcal{O}(\varepsilon^{-7/4})$ complexity?

In a recent result, [57] shows that in the case of the vanilla Heavy Ball ordinary differential equation (ODE)

$$\ddot{x}_t + \alpha \dot{x}_t + \nabla f(x_t) = 0, \quad (\text{HB})$$

a suitable averaging of its solution $\bar{x}_t$ verifies $\min_{0 \leq t \leq T} \|\nabla f(\bar{x}_t)\| = \mathcal{O}(T^{-4/7})$, which in turn implies an ε -stationary solution is achieved in at most $\mathcal{O}(\varepsilon^{-7/4})$ execution time. Importantly, the Heavy Ball ODE can be viewed as a continuous-time analogue of momentum algorithms [65, 64]. It means that this continuous-time momentum dynamic does not require such mechanisms, and naturally raises the question of whether a similar result can be established in the discrete setting. To the best of our knowledge, no such discrete counterpart is currently known.¹

The continuized method Even, Berthier et al. [25] introduce the continuized Nesterov equations

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt - \gamma \nabla f(x_{t-})dN_t, \\ dz_t &= \eta'(x_t - z_t)dt - \gamma' \nabla f(x_{t-})dN_t, \end{cases} \quad (\text{CNE})$$

where $\eta, \eta', \gamma, \gamma'$ are real constants, $t \in \mathbb{R}_+$ and N_t is a Poisson process with intensity dt . The fundamental feature of this process is that it can be analyzed through continuous-time Lyapunov strategies using tools from stochastic calculus theory, while still yielding a computable algorithm that writes as (NM), with the specificity that the sequences $\{\alpha_k\}_{k \in \mathbb{N}}$ and $\{\beta_k\}_{k \in \mathbb{N}}$ are random. These stochastic sequences are simple, in the sense that they can be readily generated at the beginning of the algorithm. Later, [34] argued that, in non-convex settings, this continuized system helps narrow the gap between convergence results derived for momentum ODEs, such as (HB), and those obtained for their algorithmic counterparts. Our motivation combines this perspective with the absence of a discrete counterpart to the result of Okamura et al. [57].

1.1 Contributions

We draw inspiration from the ODE-based analysis of [57] to analyze the solution of (CNE), which yields our main result, stated as follows (see Theorem 5 for the formal version).

Theorem 1 (Informal version). *Let f be with Lipschitz gradient and Hessian, and let $\tilde{\varepsilon} \in (0, 1/2)$. For a number of iterations n large enough, (NM) with a suitable stochastic parameterization outputs a point $\tilde{x}$*

$$\mathbb{E} [\mathbf{1}_{\mathbb{A}_n} \|\nabla f(\tilde{x})\|] = \mathcal{O}(n^{-4/7}),$$

where $\mathbb{A}_n$ is a subset of the realizations that satisfies $\mathbb{P}(\mathbb{A}_n) \geq 1 - \tilde{\varepsilon}$.

It further implies that a stochastic parameterization of (NM) is able to achieve a point that satisfies $\mathbb{E} [\|\mathbf{1}_{\mathbb{A}_n} \nabla f(\tilde{x})\|] \leq \varepsilon$ in at most $\mathcal{O}(\varepsilon^{-7/4})$ gradient evaluations, without any momentum-reset mechanisms. We note that for n large enough, the hidden numerical factor of our bound improves over the best existing bound from [46]. Conceptually, our result indicates that (NM) by itself can achieve the complexity $\mathcal{O}(\varepsilon^{-7/4})$, at the cost of randomizing parameters, and thus, of a

¹In an independent work, [67] use the Performance Estimation Problem framework to show that a deterministic discretization of (HB) achieves this complexity without restarts. Yet, their algorithm still incorporates a momentum-control mechanism, that does not set its value to zero but reduces it when the velocity $\|\tilde{x}_{k+1} - \tilde{x}_k\|$ exceeds a threshold.

result in expectation. We also show that under Lipschitz gradient only, our algorithm recovers in expectation the best existing complexity $\mathcal{O}(\varepsilon^{-2})$.

The presence of stochastic parameters in the algorithm raises the question of the conceptual relation between the continuized method and more classical, deterministic parameterizations of (NM). We show that in the vanishing stepsize limit, a time-rescaled version of (CNE) converges in probability to (HB). This parallels the fact that classical deterministic parameterizations of (NM) converge to the same ODE in this vanishing stepsize limit [65, 63]. In this sense, (CNE) may be seen as an alternative, stochastic way to discretize the Heavy Ball (HB) ODE.

1.2 Related works

First order acceleration under Lipschitz gradient and Hessian The $\tilde{\mathcal{O}}(\varepsilon^{-7/4})$ complexity for first-order algorithms was first achieved by [15]. The proposed nested-loop algorithm combines (NM') with negative-curvature exploitation and the minimization of a regularized surrogate function. This approach is subsequently simplified in [41], where the algorithm reduces to (NM') combined with negative-curvature steps. Replacing the negative curvature exploitation with a restart mechanism, [46] achieves the $\mathcal{O}(\varepsilon^{-7/4})$ bound, improving by a logarithmic factor. This restart-based line of work has since been extended to a parameter-free algorithm [47], a forward-backward version [62] and a universal algorithm under Hölder continuous Hessian [48]. A different line of work, based on online-learning techniques rather than momentum, was initiated by [21], who achieved the complexity bound $\tilde{\mathcal{O}}(\varepsilon^{-7/4})$ under the supplementary assumption that f is Lipschitz. Their idea was later extended by [39], improving the dependence on ε to $\mathcal{O}(d^{1/4}\varepsilon^{-13/8})$ at the cost of an explicit dimension dependence, without this supplementary Lipschitzness assumption on f . These online-learning based methods are somewhat intricate, relying on nested algorithms and/or nested loops, but they show that in some settings, alternatives to momentum can achieve this accelerated bound. The first known lower-bound on the number of gradient evaluations needed to find an ε -stationary point in this setting was $\mathcal{O}(\varepsilon^{-12/7})$ [18], and was recently improved to $\mathcal{O}(\varepsilon^{-7/4})$ [70], closing the gap with the upper-bound.

Second-order stationary points In contrast to our setting, several works have focused on second-order stationary points, which under suitable assumptions correspond to local minimizers, see a formal definition in Appendix A.2. With probability $1 - \delta$, such a point is found in at most $\mathcal{O}(\varepsilon^{-7/4} \log(\frac{d}{\delta\varepsilon}))$ gradient and Hessian-vector product evaluations [1, 16], or with gradient evaluation only [41, 46]. Procedures based on first-order information have been designed that transform algorithms for finding first-order stationary points into ones that find second-order stationary points [69, 3]. Even though designing methods that achieve second-order stationary points efficiently is an important avenue of research, we do not consider this direction in this work.

Continuized Nesterov The continuized Nesterov equations are introduced in [25], motivated by the study of asynchronous algorithms. They show that the continuized Nesterov algorithm recovers, in expectation, the existing rate of the classic Nesterov Momentum for (strongly)-convex functions, further generalized to (strongly)-quasar convex functions [68, 34]. This method has known some extensions in the specific field of decentralized asynchronous algorithms [51, 50], but remains largely unexplored in optimization.

2 Background

In this work, we fix an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We denote by $\mathcal{E}(1)$ the exponential distribution with parameter 1, $\Gamma(k, 1)$ the Gamma distribution with shape parameter k and rate 1 and $\mathcal{P}(\lambda)$ the Poisson distribution with parameter λ . For $\phi : \mathbb{R} \rightarrow \mathbb{R}$, $t \in \mathbb{R}$, we note $\phi(t^-) := \lim_{s \rightarrow t^-} \phi(s)$ the left-limit. $\|\cdot\|$ denotes the Euclidean norm for vectors and $\|\cdot\|_2$ denotes the spectral norm for matrices. We use the notations $f^* := \min_{x \in \mathbb{R}^d} f(x)$ and $\Delta_f := f(x_0) - f^*$.

We consider the two following assumptions.

Assumption 1. (*Lipschitz gradient*) f is such that $\forall x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$.

Assumption 2. (*Lipschitz Hessian*) f is such that $\forall x, y \in \mathbb{R}^d$, $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L_2 \|x - y\|$.

For functions satisfying Assumption 1 and/or Assumption 2, a critical point could be a non-global minimizer, a saddle point, or a maximizer. As mentioned in Section 1, Assumption 1 alone is sufficient to find an ε -stationary point with first-order methods [55, 28], but the complexity of gradient descent cannot be improved [17]. Considering Assumptions 1-2 together allows for such improvement. In existing analysis [41, 46], Assumption 2 enables working with local quadratic approximations, on which momentum is known to be effective [59]. Our analysis does not rely on such approximations.

2.1 Continuized Nesterov Algorithm

In this section, we briefly describe the process studied in this work, and the resulting algorithm. We refer to [34, Section 3] for a more detailed presentation of the continuized framework. An intuitive way to present the process is as follows: consider a sequence of random times $\{T_k\}_{k \in \mathbb{N}}$ such that $T_0 = 0$ and $T_{k+1} - T_k \stackrel{i.i.d.}{\sim} \mathcal{E}(1)$ for $k \in \mathbb{N}$. For some constant parameters $\eta, \eta', \gamma, \gamma'$, we define a continuous-time process $(x_t, z_t)_{t \in \mathbb{R}_+}$ on each interval (T_k, T_{k+1}) as the solution of the following equation

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt, \\ dz_t &= \eta'(x_t - z_t)dt. \end{cases}$$

At $t = T_k$, the process jumps by performing gradient steps,

$$x_{T_k} = x_{T_k^-} - \gamma \nabla f(x_{T_k^-}), \quad z_{T_k} = z_{T_k^-} - \gamma' \nabla f(x_{T_k^-}).$$

This defines the *continuized Nesterov equation*, where we assume $x_0 = z_0$. It can be written in a more compact form as the following stochastic differential equation

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt - \gamma \nabla f(x_{t-})dN_t, \\ dz_t &= \eta'(x_t - z_t)dt - \gamma' \nabla f(x_{t-})dN_t, \end{cases} \quad (\text{CNE})$$

where $dN_t = \sum_{k \geq 0} \delta_{T_k}(dt)$ is a Poisson point measure with intensity dt . It mixes the continuous component, the dt factor, with the gradient component that acts at discrete random times, the dN_t factor. Replacing dN_t by dt in (CNE) would yield a momentum ODE, close to the classic Heavy-Ball ODE, see [34, Appendix E.2].

A remarkable property of (CNE) is that, upon defining the sequences $\tilde{y}_k := x_{T_{k+1}^-}$, $\tilde{x}_k := x_{T_k}$ and $\tilde{z}_k := z_{T_k}$, $k \in \mathbb{N}$, it satisfies a recursive relation that takes the form of a Nesterov momentum algorithm, with stochastic parameters depending on the random times $\{T_k\}_{k \in \mathbb{N}}$.

Proposition 2 ([34], Proposition 5). *Let $(x_t, z_t)_{t \in \mathbb{R}_+}$ follow (CNE) with $\eta + \eta' > 0$ and with underlying jumping times $\{T_k\}_{k \in \mathbb{N}}$. Define $\tilde{y}_k := x_{T_{k+1}^-}$, $\tilde{x}_{k+1} := x_{T_{k+1}}$ and $\tilde{z}_{k+1} := z_{T_{k+1}}$ as evaluations of this process. Then, $(\tilde{y}_k, \tilde{x}_k, \tilde{z}_k)$ writes as a specific parameterization of (NM), that we call the continuized Nesterov algorithm (CNA), namely*

$$\begin{cases} \tilde{y}_k &= \tilde{x}_k + \frac{\eta}{\eta + \eta'} (1 - e_k) (\tilde{z}_k - \tilde{x}_k) \\ \tilde{x}_{k+1} &= \tilde{y}_k - \gamma \nabla f(\tilde{y}_k) \\ \tilde{z}_{k+1} &= \tilde{z}_k + \eta' \frac{1 - e_k}{\eta + \eta e_k} (\tilde{y}_k - \tilde{z}_k) - \gamma' \nabla f(\tilde{y}_k) \end{cases} \quad (\text{CNA})$$

where $e_k = e^{-(\eta + \eta')(T_{k+1} - T_k)}$.

In words, (CNA) is an exact evaluation of a realization of (CNE). Intuitively, exact access to the continuous process (CNE) is possible because gradient information is only queried at discrete times. This direct correspondence between (CNA) and (CNE) is powerful: it enables convergence guarantees for a discrete algorithm to be derived via continuous-time analysis. This comes at the cost that even with deterministic gradients, the resulting algorithm is inherently stochastic, thus leading to non-deterministic convergence guarantees.

Practical implementation of (CNA) From a practical point of view, compared to a deterministic parameterization of (NM), the randomization procedure in (CNA) does not introduce any

significant additional difficulty. This is because, for any $k \in \mathbb{N}$, the increments $T_{k+1} - T_k$ are independent and identically distributed following an exponential law of parameter 1, and can therefore be easily and independently simulated. As a consequence, to run (CNA) with $n \in \mathbb{N}^*$ iterations, we can first generate at once n random variable $\Delta_k \stackrel{i.i.d.}{\sim} \mathcal{E}(1)$, and save the vector $\{\Delta_k\}_{k \in \{1, \dots, n\}}$. Then, for fixed parameters η, η', γ , and γ' , the algorithm (CNA) can be implemented in the same way as (NM), using the sequences

$$\alpha_k = \frac{\eta}{\eta + \eta'} \left(1 - e^{-(\eta + \eta')\Delta_k}\right), \quad \beta_k = \eta' \frac{1 - e^{-(\eta + \eta')\Delta_k}}{\eta' + \eta e^{-(\eta + \eta')\Delta_k}}, \quad k = 1, \dots, n.$$

Link between (CNA) and (NM') Former results that achieved the $\mathcal{O}(\varepsilon^{-7/4})$ complexity used a two-sequence formulation (NM') as their momentum component. Using [35, Proposition 11], we can rewrite (CNA) as a two-sequence algorithm:

$$\begin{cases} \tilde{y}_k = \tilde{x}_k + \frac{(1 - e_k)e_{k-1}}{1 - e_{k-1}}(\tilde{x}_k - \tilde{x}_{k-1}) + \frac{\eta}{\eta + \eta'}(1 - e_k) \left(\frac{\gamma'}{\gamma} - 1 - \frac{(\eta + \eta')e_{k-1}}{\eta(1 - e_{k-1})} \right) (\tilde{x}_k - \tilde{y}_{k-1}), \\ \tilde{x}_{k+1} = \tilde{y}_k - \gamma \nabla f(\tilde{y}_k), \end{cases}$$

with $e_k = e^{-(\eta + \eta')(T_{k+1} - T_k)}$. Compared with (NM'), there is a supplementary term, factor of $\tilde{x}_k - \tilde{y}_{k-1} = -\gamma \nabla f(\tilde{y}_{k-1})$. This supplementary term appears in algorithms that get an improved constant in the convergence bound in the smooth strongly convex setting, such as the Optimized Gradient Method (OGM) [43]. It also appears in results that achieves acceleration in the non-convex regime of strongly quasar-convex functions [35, Proposition 4]. With our choice of parameter in Theorem 5, this supplementary factor does not reduce to zero. It may indicate that algorithms of the form (NM') are unable to achieve the $\mathcal{O}(\varepsilon^{-7/4})$ complexity by their own, even if randomizing the parameter.

3 Convergence Result under Lipschitz Gradient Only

We derive a convergence result under Assumption 1 only.

Proposition 3. *Assume $T_1, \dots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d. of law $\mathcal{E}(1)$, with convention $T_0 = 0$. Under Assumption 1, iterations of (CNA) with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \sqrt{\frac{\gamma}{2}}$ and $\eta' > -\eta$ verify*

$$\mathbb{E} \left[\min_{0 \leq i \leq k-1} \|\nabla f(\tilde{y}_i)\|^2 \right] \leq \frac{4}{\gamma} \frac{\Delta_f}{k}.$$

Proposition 3 implies that we find, in expectation, an ε -stationary point in at most $\mathcal{O}(\varepsilon^{-2})$ gradient evaluations. Perhaps surprisingly, the parameter η' can be chosen freely in $(-\eta, +\infty)$, without affecting the result. This is a new result for the continuized algorithm (CNA), which shows it recovers the known bound of the classical Nesterov momentum algorithms in this setting, see [28]. Proposition 3 is a consequence of the following bound, established using Lyapunov analysis for the continuous-time system (CNE).

Lemma 4. *If $(x_t, z_t)_{t \in \mathbb{R}_+}$ is solution of (CNE) under Assumption 1 with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \gamma' - \gamma$ and $\eta' > -\eta$, we have*

$$\mathbb{E} \left[\int_0^t (\eta + \eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \right] \leq \mathbb{E} [f(x_0) - f^*]. \quad (1)$$

Sketch of Proof. We set $\bar{x}_t = (x_t, z_t)$ as a concatenation of $(x_t)_{t \geq 0}$ and $(z_t)_{t \geq 0}$, which satisfies $d\bar{x}_t = \zeta(\bar{x}_t)dt + G(\bar{x}_t)dN(t)$, where

$$\zeta(\bar{x}_t) = \begin{pmatrix} \eta(z_t - x_t) \\ \eta'(x_t - z_t) \end{pmatrix}, \quad G(\bar{x}_t) = \begin{pmatrix} -\gamma_t \nabla f(x_t) \\ -\gamma'_t \nabla f(x_t) \end{pmatrix}.$$

We set the Lyapunov function

$$\varphi(x, z) = f(x) + \frac{1}{2} \|x - z\|^2.$$

Intuitively, one can stochastically derivate $\varphi(\bar{x}_t)$ using an Itô formula, yielding the following relation

$$\varphi(\bar{x}_t) = \varphi(\bar{x}_0) + \int_0^t \langle \nabla \varphi(\bar{x}_s), \zeta(\bar{x}_s) \rangle ds + \int_0^t \varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s) ds + M_t,$$

where $(M_t)_{t \geq 0}$ is a martingale verifying $\mathbb{E}[M_t] = 0$ for all $t \geq 0$. Using Assumption 1 and the parameter choice, we compute that

$$\langle \nabla \varphi(\bar{x}_s), \xi(\bar{x}_s) \rangle + \varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s) \leq -(\eta + \eta') \|x_s - z_s\|^2 - \frac{\gamma}{4} \|\nabla f(x_s)\|^2,$$

which yields

$$\int_0^t (\eta + \eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \leq f(x_0) - f^* + M_t. \quad (2)$$

One can conclude using $\mathbb{E}[M_t] = 0$ for all $t \geq 0$. See the complete proof in Appendix B. $\square$

From Lemma 4, it follows that $\mathbb{E}[\min_{s \in [0, t]} \|\nabla f(x_s)\|^2] = \mathcal{O}(1/t)$. We note that a similar bound is obtained under Assumption 1 for a continuized system in [50, Proposition 6], but with different dynamics from (CNE), specifically adapted to asynchronous algorithms. With simple terms, because $\tilde{x}_k$ in (CNA) is exactly x_t in (CNE) evaluated at the time T_k , Proposition 3 follows from evaluating (2) at $t = T_k$, from which will follow a convergence result for (CNA), up to some technical considerations, see Appendix B. Lemma 4, in particular (2), is not only an intermediate result for Proposition 3. It also serves as a key building block for proving our main theorem in the Lipschitz Hessian setting (Theorem 5).

4 Main Results

We state in Section 4.1 our acceleration result for (CNA). In Section 4.2, we show that a time-rescaled version of the continuized system (CNE) converges in probability to the Heavy Ball ODE (HB) as the stepsize γ goes to zero.

4.1 Convergence Result

We show that (CNA) achieves the $\mathcal{O}(\varepsilon^{-7/4})$ complexity under Assumptions 1-2. It may be seen as a discrete-time version of a convergence result that considers the Heavy Ball ODE (HB) [57].

Theorem 5. *Let $\{T_k\}_{k \in \mathbb{N}}$ such that $T_0 = 0$, $T_{k+1} - T_k \stackrel{i.i.d}{\sim} \mathcal{E}(1)$. Let $\tilde{\varepsilon} \in (0, 1/2]$, and $c_{upp} \geq 1$ such that it satisfies $1 - \tilde{\varepsilon} - \log(1 - \tilde{\varepsilon}) = c_{upp} - \log(c_{upp})$, which implies $c_{upp} \leq 1 + 2\tilde{\varepsilon}$. Under Assumptions 1-2, consider the iterations of (CNA) with $\gamma = \frac{1}{L}$, $\gamma' = \frac{1}{L} + \sqrt{\frac{1}{2L}}$, $\eta = \sqrt{\frac{1}{2L}}$, $\eta' = \alpha - \sqrt{\frac{1}{2L}}$, where*

$$\alpha = \left(\frac{\sqrt{3} \cdot 64 c_{upp} e}{(1 - \tilde{\varepsilon})^2} \right)^{2/7} \left(\frac{L_2^2 \Delta_f}{L^3 n} \right)^{1/7}.$$

Then, if n is large enough to ensure $2((1 + 2\tilde{\varepsilon})n)^{-1} < \alpha \leq 1$, we have

$$\mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} \min_{k \in \{1, \dots, n\}} \|\nabla f(\bar{x}_k)\| \right] \leq \mathcal{B}_n L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{-4/7},$$

where $\bar{x}_k = \sum_{i=0}^{k-1} \lambda_{i,k} \tilde{y}_i$ with $\lambda_{i,k} = \frac{e^{\alpha T_i + 1}}{\sum_{j=1}^k e^{\alpha T_j}}$, $\mathbb{P}(\mathbb{A}_n) \geq 1 - \tilde{\varepsilon}$, and $\mathcal{B}_n$ is such that it is uniformly bounded with n and $\lim_{n \rightarrow +\infty} \mathcal{B}_n \leq 9.2(1 + 2\tilde{\varepsilon})^2 \left(\frac{1 + 2\tilde{\varepsilon}}{(1 - \tilde{\varepsilon})^2} \right)^{1/7}$.

A more general statement can be found in Theorem 19, in which we provide a result while relaxing the constraint of the value of n , to the cost of a less clean bound. From Theorem 5, we deduce that (CNA) outputs a point $\tilde{x} := \operatorname{argmin}_{k \in \{1, \dots, n\}} \|\nabla f(\bar{x}_k)\|$ that verifies $\mathbb{E}[\mathbb{1}_{\mathbb{A}_n} \|\nabla f(\tilde{x})\|] \leq \varepsilon$ in at most $\mathcal{O}\left(\frac{L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}}\right)$ gradient evaluations, as long as ε is small enough. We may remove the expectation and obtain a result with a given probability, the Markov property being a direct way to do so. The stochasticity induced by the process created significant challenges in the analysis, that we solved by deriving new concentration inequalities. One of these establishes that for any $\tilde{\varepsilon} \in (0, 1)$ and $n \in \mathbb{N}^*$, there exists $\mathcal{C}_n$ such that with probability $1 - \tilde{\varepsilon}$

$$\forall i \in \{1, \dots, n\}, \quad \sum_{j=1}^i e^{-\alpha(T_i - T_j)} \leq \mathcal{C}_n \mathbb{E} \left[\sum_{j=1}^i e^{-\alpha(T_i - T_j)} \right],$$

where $\mathcal{C}_n$ is uniformly upper-bounded with n and $T_i - T_j \stackrel{i.i.d}{\sim} \Gamma(i - j + 1)$. The set $\mathbb{A}_n$ arises from these inequalities, see details in Theorem 12.

Numerical factor and asymptotic improvement The factor $\mathcal{B}_n$, explicitly stated in (21), has a somewhat intricate finite-time expression. Choosing a $\tilde{\varepsilon}$ value close to zero increases the size of the set $\mathbb{A}_n$, on which our result holds, at the cost of a poorer constant $\mathcal{B}_n$ in *finite time*. However, asymptotically with n , the factors that deteriorate the bound when $\tilde{\varepsilon}$ is small vanish. So, in this asymptotic regime, we can choose $\tilde{\varepsilon}$ small enough to ensure for instance $\lim_{n \rightarrow +\infty} \mathcal{B}_n \leq 10$. Then for ε small enough, from Theorem 5 we deduce that we reach in expectation a ε -solution in at most $\frac{10^{7/4} L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}} \approx \frac{56 L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}}$ gradient evaluations. The best existing bound [46] is $\frac{82^{7/4} L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}} \approx \frac{2234 L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}}$, such that we considerably improve the numerical constant factor in the asymptotic regime.

Remark 6. We note that as the analysis of [57] requires knowledge of the terminal time T to tune parameters in the continuous setting, our result uses the final iteration n to tune η' in Theorem 5. There is a possibility that this dependence can be removed by using an iteration-dependent schedule of the form $\eta'_k = \mathcal{O}(k^{-1/7}) - \sqrt{\gamma/2}$, at the cost of significant additional technical complexity. In a similar way, prior works achieving the $\mathcal{O}(\varepsilon^{-7/4})$ complexity involve ε -depending parameters [15, 41, 46], with the notable exception of the parameter-free restart method of [47], or the independent work [67].

4.2 Convergence of the continuized system to the Heavy Ball equation

Compared with more classical momentum algorithms, the continuized method may seem somewhat atypical. In this section, we draw a link with a more classical object of the momentum literature: the Heavy Ball equation (HB). In the L -smooth, convex/strongly convex case, [25] uses the heuristic $dN(\gamma^{-1/2}s) \approx \gamma^{-1/2}s$ in the limit $\gamma \rightarrow 0$ to argue that (CNE), parameterized with values ensuring acceleration, converges to the Heavy Ball equation. We provide a rigorous statement in our setting. To do so, we want to consider (CNE) with $0 < \gamma \leq 1/L$, $\gamma' = \gamma + \sqrt{\gamma/2}$, $\eta = \sqrt{\gamma/2}$, $\eta' = \alpha_T - \sqrt{\gamma/2}$, where $\alpha = \left(\frac{\sqrt{3} \cdot 64 c_{upp} e}{(1-\tilde{\varepsilon})^2}\right)^{2/7} \left(\frac{\gamma^3 L_2^2 \Delta_f}{n}\right)^{1/7}$, and study the limit of this system as γ goes to zero after a suitable time-rescaling. We recall that Theorem 5 is obtained using these parameters, with the special choice $\gamma = \frac{1}{L}$. However, we will provide a modification on these parameters. Because α is tuned to optimize discrete quantities, it involves $n \in \mathbb{N}$ that corresponds to a number of iterations. We did not derive a convergence result for (CNE) in continuous time, but it is reasonable to make the correspondence between the final iteration n and a terminal time $T \in \mathbb{R}_{\geq 0}$, such that we replace n by T in the definition of α .

Theorem 7. Let $(x_t, z_t)_{t \in \mathbb{R}_+}$ be solution of (CNE), which we recall to be

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt - \gamma \nabla f(x_t) dN_t, \\ dz_t &= \eta'(x_t - z_t)dt - \gamma' \nabla f(x_t) dN_t, \end{cases} \quad (\text{CNE})$$

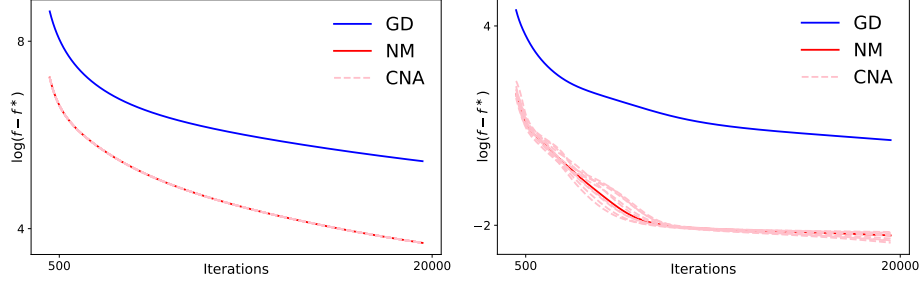


Figure 1: We plot one run of gradient descent, one run of a deterministic instance of (NM), and ten runs of (CNA) on a matrix factorization problem. On the left, we choose γ according to the theoretical result of [40, Lemma 6]. On the right, we chose a more aggressive stepsize and momentum parameter α , i.e. smaller. We note that because we pushed the parameters closer to the stability limit, one of the 10 runs of (CNA) diverged and is not plotted.

where $\gamma' = \gamma + (\gamma/2)^{\frac{1}{2}}$, $\eta = (\gamma/2)^{\frac{1}{2}}$, $\eta' = \alpha_{(\gamma/2)^{-1/2}T} - (\gamma/2)^{\frac{1}{2}}$, with

$$\alpha_{(\gamma/2)^{-1/2}T} = \left(\frac{\sqrt{3} \cdot 64c_{upp}e}{(1-\tilde{\varepsilon})^2} \right)^{2/7} \left(\frac{\gamma^3 L_2^2 \Delta_f}{T(\gamma/2)^{-1/2}} \right)^{\frac{1}{7}} = K_{\tilde{\varepsilon}, c_{upp}} \left(\frac{L_2^2 \Delta_f}{\sqrt{2}T} \right)^{\frac{1}{7}} \gamma^{1/2},$$

for some $T > 0$, $\tilde{\varepsilon} \in (0, 1/2]$ and $K_{\tilde{\varepsilon}, c_{upp}} = \left(\frac{\sqrt{3} \cdot 64c_{upp}e}{(1-\tilde{\varepsilon})^2} \right)^{2/7}$. Then, the time-rescaled process $x_s^\gamma := x_{(\gamma/2)^{-1/2}s}$ converges uniformly in probability on $[0, T]$ to the Heavy Ball ODE as $\gamma \rightarrow 0$, namely

$$\sup_{s \in [0, T]} \|X_s - x_s^\gamma\| \xrightarrow{\mathbb{P}_{\gamma \rightarrow 0}} 0,$$

where

$$\ddot{X}_s + 2^{3/7} K_{\tilde{\varepsilon}, c_{upp}} \left(\frac{L_2^2 \Delta_f}{T} \right)^{1/7} \dot{X}_s + \nabla f(X_s) = 0,$$

as long as $x_0 = X_0$.

See the proof in Appendix C. This result indicates that we may view (CNE) as a specific, stochastic discretization of (HB). It provides a strong theoretical connection with more classical, deterministic parameterization of Nesterov momentum, which converge to the same kind of ODE [65, 63]. Empirically, we observe that in the small stepsize regime, the behavior of (CNA) is indeed very close to a deterministic version. On Figure 1, we run (CNA) on a matrix factorization problem, that satisfies Assumptions 1 and 2 on a subspace of the domain, see [40, Lemma 6] for admissible values of L and L_2 . We also run a deterministic instance of (NM) by approximating the parameters of (CNA), designed as follows. We have $\frac{\eta}{\eta+\eta'} (1 - e^{-(\eta+\eta')\Delta_k}) \approx \eta\Delta_k$ if $\eta + \eta'$ is small, while $\eta' \frac{1 - e^{-(\eta+\eta')\Delta_k}}{\eta' + \eta e^{-(\eta+\eta')\Delta_k}} \approx \frac{\eta'(\eta+\eta)\Delta_k}{\eta' + \eta(1 - (\eta+\eta')\Delta_k)}$. Replacing roughly Δ_k by its expectation 1, we then chose $\alpha_k = \eta$ and $\beta_k = \frac{\eta'(\eta+\eta)}{\eta' + \eta(1 - (\eta+\eta')\Delta_k)}$ as our deterministic sequences of parameters for (NM). We observe on Figure 1 that in the small stepsize regime, the continuized version behaves very similarly to the deterministic version, with function values curves being superposed. Behaviors start differentiating when increasing stepsize and choosing a more aggressive momentum parameter.

5 Overview of the proof of Theorem 5

This section provides a proof outline of Theorem 5, with most technical details deferred to the appendix. We recall that $(x_t, z_t)_{t \in \mathbb{R}_+}$ are defined as follows

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt - \gamma \nabla f(x_t) dN_t, \\ dz_t &= \eta'(x_t - z_t)dt - \gamma' \nabla f(x_t) dN_t, \end{cases} \quad (\text{CNE})$$

with $dN_t = \sum_{k \geq 1} \delta_{T_k}(dt)$ for $\{T_k\}_{k \in \mathbb{N}}$ a sequence of random variables satisfying $T_{k+1} - T_k \stackrel{i.i.d}{\sim} \mathcal{E}(1)$, with convention $T_0 = 0$. Throughout the proof, we note $\alpha := \eta + \eta'$ and use the shorthand notation $\int_0^t := \int_{[0,t]}$ to save space. Inspired from the averaging [57, Equations (4)-(5)], we consider a Poisson-average of the trajectory x_t

$$\bar{x}_t := \int_0^t w_t(s) x_s - dN_s, \quad (3)$$

where we set $w_t(s) = \frac{\alpha e^{\alpha s}}{\int_0^t \alpha e^{\alpha s'} dN(s')}$. To readers unfamiliar with Poisson processes, we emphasize that Poisson integrals are stochastic sums. Namely, for a measurable function ϕ ,

$$\int_0^t \phi(s) dN_s = \sum_{i=1}^{N_t} \phi(T_i)$$

where $N_t = \sum_{k \geq 1} \mathbb{1}_{T_k \leq t} \sim \mathcal{P}(t)$. So, we can also write $\bar{x}_t = \sum_{i=1}^{N_t} w_t(T_i^-) x_{T_i^-}$. In particular, at $t = T_n$, by definition $N_{T_n} = n$ and the number of terms in the sum is fixed, becoming $\sum_{i=1}^n w_t(T_i^-) x_{T_i^-}$. To ensure $\int_0^t w_t(s) dN_s = 1$, we assume $N_t \geq 1$ (otherwise the integral is zero). This is not a restriction, because $N_t < 1$ would imply no jumps occurred, *i.e.* the underlying algorithm has not started yet.

Plan of the Proof Our proof strategy is the following, divided into seven steps.

- I. We establish some upper-bounds on trajectory-depending quantities, that follow from the Lyapunov control of Section 3, which uses the Lipschitz gradient property. These controls are used in step III and IV.
- II. We upper-bound by three terms an expected weighed average $\mathbb{E} \left[\int_0^{T_n} \Lambda_t \|\nabla f(\bar{x}_t)\| dN_t \right]$, for some weighting parameter Λ_t . It uses the Hessian Lipschitz property, and a stochastic integration by part formula.
- III. We bound the first of these three terms, by a quantity of the order $\mathcal{O}(\sqrt{\alpha n})$. The main argument is the Cauchy-Schwartz inequality and our Lyapunov control from I.
- IV. We bound with high probability the two remaining terms, by a quantity of the order $\mathcal{O}(\alpha^{-3})$. This part is the most tedious of the proof. Apart from our Lyapunov control from I, it requires to use a concentration inequality on the underlying jump times $\{T_k\}_{k \in \mathbb{N}}$, thus a result holding on a subset of realization $\mathbb{A}_n$.
- V. We combine the bounds derived in III and IV, and show that with the choice $\alpha = \mathcal{O}(n^{-1/7})$, we obtain a result of the form

$$\mathbb{E} \left[\min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] = \mathcal{O}(n^{3/7}).$$

- VI. We obtain a lower-bound of the form $\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \geq Cn$ for some constant $C < 1$, that holds with high probability.
- VII. We state our final convergence results. The final key argument is that although it involves the continuous-time process $\bar{x}_t$, $\arg \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\|$, can in fact be computed by (CNA).

Difference and similarities in the analysis with Okamura et al. [57] Some of our arguments are directly inspired from [57]. This is the case for our definition of averaging (3), that is essentially [57, Equations (4)-(5)] with ds replaced by dN_s . Our Part II is also directly inspired: Lemma 8 is a continuized-adapted version of Lemma 1 in [57], the difference in our analysis mainly coming from the Poisson process and the fact that (CNE) is a two-trajectory system, in contrast with the Heavy Ball equation (HB). Similarly, our Lemma 9 provides a stochastic integration by

part, inspired by the integration by parts [57, Equation 18], but in our case the stochastic nature of the process requires some care when using such arguments. Beyond these directly-adaptable points, the rest of the analysis differs because the stochasticity induced by the process create non trivial difficulties. In particular, Theorem 12 provides new concentrations inequalities that specifically address such difficulties, and is an essential argument of the proof.

5.1 Part I - Lyapunov Controls from Section 3

In this first part, we establish bounds on trajectory-depending quantities that we can deduce from the analysis carried in Section 3. Under Assumption 1, we proved in Lemma 4 inequality (2), namely

$$\int_0^t (\eta + \eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \leq f(x_0) - f^* + M_t,$$

with M_t a martingale such that $\mathbb{E}[M_t] = 0$. From this, we deduce

$$\begin{aligned} \int_0^t (\eta + \eta') \|x_s - z_s\|^2 ds &\leq f(x_0) - f^* + M_t, \\ \int_0^t \frac{\gamma}{4} \|\nabla f(x_{s-})\|^2 dN_s &\leq f(x_0) - f^* + U_t, \end{aligned} \quad (4)$$

$$\int_0^t (\eta + \eta') \|x_{s-} - z_{s-}\|^2 dN_s \leq f(x_0) - f^* + V_t, \quad (5)$$

where for some martingales U_t, V_t with mean zero, (4) and (5) hold because $\int_0^t (\eta + \eta') \|x_s - z_s\|^2 ds = \int_0^t (\eta + \eta') \|x_{s-} - z_{s-}\|^2 dN_s$ and $\int_0^t \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds = \int_0^t \frac{\gamma}{4} \|\nabla f(x_{s-})\|^2 dN_s$, up to martingales additive terms whose expectations are zero. Now, using Theorem 23, evaluating these bounds at $t = T_n$, they become

$$\mathbb{E} \left[\int_0^{T_n} (\eta + \eta') \|x_s - z_s\|^2 ds \right] \leq \mathbb{E} [f(x_0) - f^*], \quad (6)$$

$$\mathbb{E} \left[\int_0^{T_n} \frac{\gamma}{4} \|\nabla f(x_{s-})\|^2 dN_s \right] \leq \mathbb{E} [f(x_0) - f^*], \quad (7)$$

$$\mathbb{E} \left[\int_0^{T_n} (\eta + \eta') \|x_{s-} - z_{s-}\|^2 dN_s \right] \leq \mathbb{E} [f(x_0) - f^*]. \quad (8)$$

The three inequalities (6), (7), (8) will play a central role. A large part of the remainder of the proof is devoted to preparing the ground for the application of these inequalities.

5.2 Part II - Using the Hessian Lipschitz Property to Obtain a Three Terms Upper-Bound

We adapt the analysis of (HB) carried by [57] to derive an analysis of (CNE). Ultimately, our goal is to obtain a control of the norm of the gradient of the average (3), namely of $\|\nabla f(\bar{x}_t)\|$. The Lipschitz Hessian property (Assumption 2) allows us to control the distance of $\nabla f(\bar{x}_t)$ to a quantity of interest, namely the average of the gradient.

Lemma 8. *Let $(x_t, z_t)_{t \in \mathbb{R}_+}$ be solution of (CNE). Under Assumption 2, we have*

$$\begin{aligned} &\left\| \int_0^t w_t(s) \nabla f(x_{s-}) dN_s - \nabla f(\bar{x}_t) \right\| \\ &\leq L_2 \int_0^t \eta \|z_s - x_s\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds \end{aligned} \quad (9)$$

$$+ L_2 \int_0^t \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau \right) dN_s. \quad (10)$$

See the proof in Appendix D.1. The quantities $\|z_s - x_s\|^2$ and $\|\nabla f(x_{s-})\|^2$ appearing in (9) and (10) are interesting. Indeed, as showed in Section 5.1, we can bound the expectation of integrals of these quantities.

In contrast, we do not have a direct control of $\int_0^t w_t(s) \nabla f(x_{s-}) dN_s$. However, because of the choice of averaging $w_s(\cdot)$ and the definition of (CNE), a stochastic integration by parts formula reduces it to a more tractable quantity.

Lemma 9. *If $(x_t, z_t)_{t \in \mathbb{R}_+}$ is solution of (CNE), for $t > 0$ we have*

$$\int_0^t w_t(s) \nabla f(x_{s-}) dN_s = -\frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma}. \quad (11)$$

See the proof in Appendix D.2. Combining Lemmas 8 and 9 and using a triangular inequality yields the following upper-bound

$$\begin{aligned} \|\nabla f(\bar{x}_t)\| &\leq \frac{w_t(t) \|z_t - x_t\|}{\gamma' - \gamma} \\ &\quad + L_2 \int_0^t \eta \|z_s - x_s\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds \\ &\quad + L_2 \int_0^t \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau \right) dN_s. \end{aligned}$$

This bound is nonetheless not sufficient to use our controls of Section 5.1; in particular, we do not have a bound involving the non-integrated quantity $\|z_t - x_t\|$. So, we integrate between $t = 0$ and $t = T_n$ and take the expectation on the above inequality, after multiplication of each side by a suitable quantity.

Lemma 10. *Noting $\Lambda_t := \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2$, for $(x_t, z_t)_{t \in \mathbb{R}_+}$ solution of (CNE) under Assumption 2, we have*

$$\begin{aligned} \mathbb{E} \left[\int_0^{T_n} \Lambda_t \|\nabla f(\bar{x}_t)\| dN_t \right] &\leq \underbrace{\frac{1}{\gamma' - \gamma} \mathbb{E} \left[\int_0^{T_n} \Lambda_t w_t(t) \|z_t - x_t\| dN_t \right]}_{\text{Term 1}} \\ &\quad + \underbrace{\alpha^2 \eta^2 L_2 \mathbb{E} \left[\int_0^{T_n} \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds \right]}_{\text{Term 2}} \\ &\quad + \underbrace{\alpha^2 \gamma^2 L_2 \mathbb{E} \left[\int_0^{T_n} \|\nabla f(x_{s-})\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau dN_t \right) dN_s \right]}_{\text{Term 3}}. \end{aligned} \quad (12)$$

See the proof in Appendix D.3. Section 5.3 is devoted to the control of Term 1. Section 5.4 is devoted to the control of Term 2 and Term 3, these two terms sharing a similar analysis.

5.3 Part III: Control of Term 1

We obtain the following control of Term 1 in (12).

Lemma 11. *Under Assumptions 1 and 2, let $(x_t, z_t)_{t \in \mathbb{R}_+}$ be solution of (CNE) with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \sqrt{\frac{\gamma}{2}}$ and $\eta' = \alpha - \eta$ for $\alpha > 0$. Then, we have*

$$\mathbb{E} \left[\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| dN_t \right] \leq \sqrt{\alpha n} \sqrt{\frac{A_n}{\gamma}} \sqrt{\Delta_f},$$

with $A_n = 12 \left(1 + \frac{3}{2}\alpha\right) (1 + 2\alpha)$.

Sketch of Proof. The main idea is to use Cauchy-Schwartz inequality and our controls from Section 5.1. Indeed, using the notation $\Lambda_t := \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2$ and the definition $w_t(t) = \frac{\alpha e^{\alpha t}}{\int_0^t \alpha e^{\alpha s} dN_s}$, Term 1 becomes

$$\begin{aligned} \mathbb{E} \left[\int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\|^2 dN_t \right] &= \mathbb{E} \left[\int_0^{T_n} \frac{\alpha}{\gamma' - \gamma} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right) \|z_t - x_t\| dN_t \right] \\ &\leq \frac{\alpha}{\gamma' - \gamma} \sqrt{\mathbb{E} \left[\int_0^{T_n} \Lambda_t dN_t \right]} \sqrt{\mathbb{E} \left[\int_0^{T_n} \|z_t - x_t\|^2 dN_t \right]}, \end{aligned} \quad (13)$$

where we used the Cauchy-Schwarz inequality. $\sqrt{\mathbb{E} \left[\int_0^{T_n} \Lambda_t dN_t \right]}$ can be computed explicitly. For the other term, we still cannot use our controls from Section 5.1. Before, we use that

$$\begin{aligned} \int_0^{T_n} \|z_t - x_t\|^2 dN_t &= \int_0^{T_n} \|z_{t-} - x_{t-} - (\gamma' - \gamma) \nabla f(x_{t-})\|^2 dN_t \\ &\leq 2 \int_0^{T_n} \|z_{t-} - x_{t-}\|^2 dN_t + 2(\gamma' - \gamma)^2 \int_0^{T_n} \|\nabla f(x_{t-})\|^2 dN_t \end{aligned}$$

We note here that this step is necessary because our control (5) holds when integrating $\|z_{t-} - x_{t-}\|^2$, not $\|z_t - x_t\|^2$. The reason is that the relation $\int_0^t \phi(s) dN_s = \int_0^t \phi(s) ds + M_t$ for some centered martingale $(M_t)_{t \geq 0}$ holds as long as ϕ is a *previsible* process with respect to N , namely if $\phi(t)$ is measurable with respect to the filtration $\{N_s\}_{0 \leq s < t}$. While $\|z_{t-} - x_{t-}\|^2$ is previsible, $\|z_t - x_t\|^2$ is not. $\square$

See Appendix D.4 for the full proof.

5.4 Part IV: Control of Term 2 and Term 3

Compared with Term 1, the remaining Terms 2 and 3 from (12) are more challenging. When using the classic deterministic (HB) equation, the analogue of the triple integrals inside the parentheses are deterministic, and can be directly computed [57]. In our case, if the expectation of these triple integrals can be computed, we have to deal with the expectation of the full expression. To do so, we are faced with the lack of independence between the trajectory-dependent factors and these triple integrals.

To illustrate this, we consider Term 3, namely

$$\int_0^{T_n} \|\nabla f(x_{s-})\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau dN_t \right) dN_s.$$

It can be rearranged as

$$\int_0^{T_n} \left(\int_s^{T_n} \int_s^t e^{\alpha(\tau-t)} e^{\alpha(s-t)} \|\nabla f(x_{s-})\|^2 \int_0^s e^{\alpha(\sigma-s)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau dN_t \right) dN_s.$$

Writing it as a sum, it becomes

$$\sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \left\| \nabla f(x_{T_i-}) \right\|^2 \sum_{j=1}^i e^{\alpha(T_j - T_i)} (k - j). \quad (14)$$

Taking the expectation, we have

$$\begin{aligned} \mathbb{E}[(14)] &= \mathbb{E} \left[\sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \left\| \nabla f(x_{T_i-}) \right\|^2 \sum_{j=1}^i e^{\alpha(T_j - T_i)} (k - j) \right] \\ &= \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \left\| \nabla f(x_{T_i-}) \right\|^2 \sum_{j=1}^i e^{\alpha(T_j - T_i)} (k - j) \right] \end{aligned}$$

Importantly, the increments of $\{T_k\}_{k \in \mathbb{N}}$ are independent, namely for any integers $1 \leq \ell_0 < \ell_1$, we have $T_{\ell_1} - T_{\ell_0} \perp \phi(T_1, \dots, T_{\ell_0})$, for any measurable function ϕ . In our case, as $l, k \geq i$, we have $T_k - T_l \perp \|\nabla f(x_{T_i^-})\|$ and $T_i - T_l \perp \|\nabla f(x_{T_i^-})\|$. So, we deduce

$$\mathbb{E}[(14)] = \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \right] \mathbb{E} \left[\left\| \nabla f(x_{T_i^-}) \right\|^2 \sum_{j=1}^i e^{\alpha(T_j - T_i)} (k - j) \right]$$

The challenge we face here is that $\nabla f(x_{T_i^-})$ and $\sum_{j=1}^i e^{\alpha(T_j - T_i)} (k - j)$ are not independent. We address this difficulty in the next section.

5.4.1 Splitting the expectations

In the previous section, we showed that we arrive at a point where trajectory-dependent factors are multiplied with some trajectory-independent terms, from which they are no independent, which prevents to separate the expectations. Precisely, the problematic trajectory-independent factors we obtain are

$$\sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)}, \quad \sum_{j=1}^i e^{-\alpha(T_i - T_j)}, \quad \sum_{j=1}^i (i - j) e^{-\alpha(T_i - T_j)}, \quad i \in \{1, \dots, n\}.$$

Our solution is to show that with high probability, these terms can be upper-bounded by their expectation multiplied by a constant factor. The trickiest part is to find a probability and a constant that do not become trivial as n grows. In particular, it cannot work by directly using the Markov property.

Theorem 12. *Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0, 1)$, $\alpha \in (0, 1)$. Let $K = \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2 \frac{n}{\varepsilon}) \rceil$, and $c_{upp} \geq 1$ be chosen such that $c_{upp} - \log(c_{upp}) = c_{inf} - \log(c_{inf})$. Let*

$$C_n = 32(4 + 4C_n + C_n^2) \frac{c_{upp}^e}{c_{inf}^2}, \quad \text{where } C_n = c_{inf} \alpha K.$$

There exists a set $\mathbb{A}_n \subset \Omega$ of realizations satisfying $\mathbb{P}(\mathbb{A}_n) \geq 1 - \varepsilon$ such that the three following assertions hold

$$\begin{aligned} (1) \quad & \forall i \in \{K + 1, \dots, n\}, \quad \sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \leq C_n \mathbb{E} \left[\sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \right], \\ (2) \quad & \forall i \in \{1, \dots, n\}, \quad \sum_{j=1}^i e^{-\alpha(T_i - T_j)} \leq C_n \mathbb{E} \left[\sum_{j=1}^i e^{-\alpha(T_i - T_j)} \right], \\ (3) \quad & \forall i \in \{1, \dots, n\}, \quad \sum_{j=1}^i (i - j) e^{-\alpha(T_i - T_j)} \leq C_n \mathbb{E} \left[\sum_{j=1}^i (i - j) e^{-\alpha(T_i - T_j)} \right]. \end{aligned}$$

Section E is devoted to the complete proof of Theorem 12. One of its key step is to use concentration inequalities to obtain that with a high probability that does not depend on n , we ensure both an upper and lower-bound on the increments $T_i - T_j$, of the form

$$c_{inf} \mathbb{E}[T_i - T_j] - c_{inf} K \leq T_i - T_j \leq c_{upp} \mathbb{E}[T_i - T_j] + c_{upp} K,$$

where $K = \mathcal{O}(\log(n/\varepsilon))$. Precisely, the set $\mathbb{A}_n$ is the set on which the above holds. These bounds then allow to relate the sums with their expectations.

Thanks to Theorem 12, with high probability one can fully separate the expectation of the factor that depend on the trajectory from those who do not. This is formalized in the following result.

Lemma 13. *We assume the same setting as Theorem 12.*

1. *Noting $A_k := \int_{[T_k, T_n]} \int_{[T_k, t]} \int_{[0, T_k]} e^{\alpha(T_k - t)} e^{\alpha(\tau - t)} e^{\alpha(\sigma - T_k)} (\tau - \sigma) dN_\sigma dN_\tau dN_t$, we have*

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \int_0^{T_n} \eta^2 \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds \right] \\ & \leq C_n \max_{k \in \{1, \dots, n\}} \mathbb{E}[A_k] \mathbb{E} \left[\int_0^{T_n} \eta^2 \|z_s - x_s\|^2 ds \right] \end{aligned} \tag{15}$$

2. Noting $B_k := \int_{T_k}^{T_n} \int_{T_k}^t \int_0^{T_k} e^{\alpha(\tau-t)} e^{\alpha(T_k-t)} e^{\alpha(\sigma-T_k)} (N_{\tau-} - N_{\sigma-}) dN_{\sigma} dN_{\tau} dN_t$, we have

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \int_0^{T_n} \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_{\sigma} dN_{\tau} dN_t \right) dN_s \right] \\ & \leq \mathcal{C}_n \max_{k \in \{1, \dots, n\}} \mathbb{E}[B_k] \mathbb{E} \left[\int_0^{T_n} \gamma^2 \|\nabla f(x_{s-})\|^2 dN_s \right] \end{aligned} \quad (16)$$

The proof of Lemma 13 uses the ideas introduced in the beginning of Section 5.4, such as writing the integrals as sums (see (14)). See Appendix D.5 for the complete proof.

5.4.2 Final Control

Thanks to Lemma 13, we see that we can control separately the expectations of trajectory-dependent factors, and the factors $\max_{k \in \{1, \dots, n\}} \mathbb{E}[A_k]$ and $\max_{k \in \{1, \dots, n\}} \mathbb{E}[B_k]$. The first can be dealt with using the Lyapunov controls of Section 5.1. For the two other terms, we note that we can write the quantities A_k and B_k as sums, for instance in the case of B_k

$$\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_{\sigma} dN_{\tau} dN_t = \sum_{i=k}^n \sum_{j=1}^k \sum_{l=k}^i e^{2\alpha(T_l-T_i)} e^{\alpha(T_j-T_l)} (l-j).$$

The expectation of the terms in this sum can be computed using basic properties of Gamma laws. Indeed, as $T_l - T_i \sim \Gamma(l-i, 1)$, we recognize that $\mathbb{E}[e^{2\alpha(T_l-T_i)}]$ is the characteristic function of a gamma law evaluated at 2α , whose value is $(1+2\alpha)^{-(l-i)}$. The terms appearing in $\mathbb{E}[A_k]$ are not significantly more difficult. Then, $\max_{k \in \{2, \dots, n\}} \mathbb{E}[A_k]$ and $\max_{k \in \{2, \dots, n\}} \mathbb{E}[B_k]$ can be upper-bounded with basic algebraic manipulations, yielding the following bounds.

Lemma 14. *Consider A_k and B_k as defined in Lemma 13. Then*

$$\begin{aligned} 1. \max_{k \in \{2, \dots, n\}} \mathbb{E}[A_k] & \leq \frac{(1+\alpha)(1+2\alpha)}{\alpha^4}. \\ 2. \max_{k \in \{1, \dots, n\}} \mathbb{E}[B_k] & \leq \frac{(1+\alpha)^2(1+2\alpha)}{\alpha^4}. \end{aligned}$$

See the proof in Appendix D.6. Combining Lemma 14 with our controls from Section 5.1, we finally obtain the following bound on Terms 2 and 3.

Lemma 15. *Assume the same setting as in Theorem 12. Under Assumptions 1 and 2, let $(x_t, z_t)_{t \in \mathbb{R}_+}$ be solution of (CNE) with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \sqrt{\frac{\gamma}{2}}$ and $\eta' = \alpha - \eta$ for $\alpha \in (0, 1)$. Then, we have*

$$\begin{aligned} & \alpha^2 \eta^2 L_2 \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \int_0^{T_n} \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_{\sigma} dN_{\tau} dN_t \right) ds \right] \\ & + \alpha^2 \gamma^2 L_2 \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \int_0^{T_n} \|\nabla f(x_{s-})\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_{\sigma} dN_{\tau} dN_t \right) dN_s \right] \\ & \leq \frac{\mathcal{C}_n L_2 \gamma}{\alpha^3} B_n \Delta_f, \end{aligned}$$

$$\text{with } B_n := \frac{1}{2}(1+\alpha)(1+2\alpha)(1+\alpha 8(1+\alpha)).$$

See the proof in Appendix D.7.

5.5 Part V: Tuning α

Combining Lemma 10, and the controls provided by Lemma 11 (Term 1) and Lemma 15 (Terms 2 and 3), we obtain that under the statement of Theorem 5, we have

$$\begin{aligned} \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 \|\nabla f(\bar{x}_t)\|^2 dN_t \right] & \leq \sqrt{\alpha n} \sqrt{\frac{A_n}{\gamma}} \sqrt{\Delta_f} \\ & + \frac{\mathcal{C}_n L_2 \gamma}{\alpha^3} B_n \Delta_f \end{aligned} \quad (17)$$

with, $A_n := 12(1 + \frac{3}{2}\alpha)(1 + 2\alpha)$ and $B_n := \frac{1}{2}(1 + \alpha)(1 + 2\alpha)(1 + \alpha 8(1 + \alpha))$.

Because we integrate with respect to dN_t on $[0, T_n]$, we have

$$\min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \leq \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 \|\nabla f(\bar{x}_t)\| dN_t.$$

Therefore, from (17), it follows that

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \\ & \leq \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 \|\nabla f(\bar{x}_t)\| dN_t \right] \\ & \stackrel{(17)}{\leq} \sqrt{\alpha n} \sqrt{\frac{A_n}{\gamma}} \sqrt{\Delta_f} + \frac{C_n L_2 \gamma}{\alpha^3} B_n \Delta_f \\ & = n^{3/7} \left(\sqrt{A_n C_\alpha L^{1+a_1} L_2^{a_2} \Delta_f^{1+a_3}} + B_n C_\alpha^{-3} C_n L^{-1-3a_1} L_2^{1-3a_2} \Delta_f^{1-3a_3} \right), \end{aligned} \tag{18}$$

where in the last inequality, we chose $\gamma = 1/L$ and defined

$$\alpha = C_\alpha n^{-1/7} L^{a_1} L_2^{a_2} \Delta_f^{a_3},$$

where $C_\alpha > 0$, $a_1, a_2, a_3 \in \mathbb{R}$ are defined such that $\alpha \leq 1$. The choice $\alpha = \mathcal{O}(n^{-1/7})$ is derived from a trade-off between $\sqrt{\alpha}$ and α^{-3} . It is consistent with the choice made in [57], in the case of studying the Heavy Ball (HB) ODE.

We exhibit two admissible choices of α , subject to the constrain $\alpha \leq 1$. We note that in general, this forces to set a condition on the value of the number of iteration n , which may depend on L , L_2 or Δ_f . If one wants to avoid such conditions, one can choose $C_\alpha = 1$, $a_1 = a_2 = a_3 = 0$.

Corollary 16 (Choice I). *Let $\alpha = n^{-1/7} \leq 1$. In this case, we have*

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \\ & \leq 16(1 + 2n^{-1/7})^4 \left(1 + 8(1 + 2n^{-1/7}) \right) (4 + 4C_n + C_n^2) \frac{c_{upp} e}{c_{inf}^2} \left(\sqrt{L \Delta_f} + L^{-1} L_2 \Delta_f \right) n^{3/7}, \end{aligned} \tag{19}$$

with $C_n = c_{inf} n^{-1/7} \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2 \frac{n}{\varepsilon}) \rceil$.

Otherwise, if we set a condition of n , we can obtain a cleaner result.

Corollary 17 (Choice II). *Let*

$$\alpha = \left(\frac{\sqrt{3} \cdot 64 c_{upp} e}{c_{inf}^2} \right)^{2/7} \left(\frac{L_2^2 \Delta_f}{L n} \right)^{\frac{1}{7}}.$$

If $\alpha \leq 1$, we have

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \\ & \leq (1 + C_n + C_n^2/4) H(\alpha) 7 \cdot 6^{-6/7} 12^{3/7} 64^{1/7} \left(\frac{c_{upp} e}{c_{inf}^2} \right)^{1/7} \left(\frac{c_{upp} e}{c_{inf}^2} \right)^{1/7} L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{3/7}, \end{aligned}$$

with $C_n = c_{inf} \alpha \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2 \frac{n}{\varepsilon}) \rceil$ and $H(\alpha) = (1 + \alpha)(1 + 2\alpha)(1 + \alpha 8(1 + \alpha))$.

The proofs of Corollary 16 and 17 are in Appendix D.8.

5.6 Part VI: Control of the double Poisson integral factor

Up to now, we managed to get an upper-bound of the form

$$\mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] = \mathcal{O}(n^{3/7}), \quad (20)$$

where we recall that by Theorem 12, $\mathbb{A}_n$ is such that $\mathbb{P}(\mathbb{A}_n) \geq 1 - \varepsilon$. As we chose $\alpha = \mathcal{O}(n^{-1/7})$, we have (see Lemma 40-(ii))

$$\mathbb{E} \left[\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] = \mathcal{O}(n),$$

which heuristically suggests that $\mathbb{E} [\mathbb{1}_{\mathbb{A}_n} \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\|] = \mathcal{O}(n^{-4/7})$. However, the two factors inside the expectations in (20) are not independent, as they both depend on the n first jump times. We therefore cannot separate the expectation. To circumvent this difficulty, we use the following result.

Lemma 18. *Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0, 1)$, $\alpha \in (0, 1]$ and $K = \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2\frac{n}{\varepsilon}) \rceil$. Let $c_{upp} \geq 1$ be chosen such that $c_{inf} - \log(c_{inf}) = c_{upp} - \log(c_{upp})$ and $C_n = c_{inf} \alpha K$. Then, with the same set $\mathbb{A}_n$ as in Theorem 12, we have*

$$\mathbb{1}_{\mathbb{A}_n} \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \geq \mathbb{1}_{\mathbb{A}_n} \frac{n}{c_{upp}^2} e^{-2\frac{c_{upp}}{c_{inf}} C_n} \left(1 - \frac{2}{c_{upp} \alpha n} \right).$$

The proof of Lemma 18 borrows arguments used to prove Theorem 12, see full proof in Appendix D.9.

5.7 Part VII: Final Results

We now state our final result. There is a result with our first choice $\alpha = n^{-1/7}$, that (almost) does not require any restriction on n , at the cost of a less clean bound. A second result holds with a more careful choice of α , but it induces a constraint on n that has to be large enough.

Theorem 19 (Complete Version). *Let $\{T_k\}_{k \in \mathbb{N}}$ such that $T_0 = 0$, $T_{k+1} - T_k \stackrel{i.i.d}{\sim} \mathcal{E}(1)$. Let $c_{inf}, \varepsilon \in (0, 1)$, and $c_{upp} \geq 1$ such that it satisfies $c_{inf} - \log(c_{inf}) = c_{upp} - \log(c_{upp})$.*

Under Assumptions 1-2, consider the iterations of (CNA) with $\gamma = \frac{1}{L}$, $\gamma' = \frac{1}{L} + \sqrt{\frac{1}{2L}}$, $\eta = \sqrt{\frac{1}{2L}}$, $\eta' = \alpha - \sqrt{\frac{1}{2L}}$.

(i) *Let $\alpha = n^{-1/7}$. If $n \geq 3$, we have*

$$\mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \min_{k \in \{1, \dots, n\}} \|\nabla f(\bar{x}_k)\| \right] \leq \mathcal{A}_n \left(\sqrt{L \Delta_f} + L^{-1} L_2 \Delta_f \right) n^{-4/7},$$

where

$$\mathcal{A}_n := 16(1+2n^{-1/7})^4 \left(1 + 8(1+2n^{-1/7}) \right) (4+4C_n+C_n^2) \left(1 - \frac{2}{c_{upp} n^{6/7}} \right)^{-1} \frac{c_{upp}^3}{c_{inf}^2} e^{1+2\frac{c_{upp}}{c_{inf}} C_n},$$

with $C_n = c_{inf} \alpha \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2\frac{n}{\varepsilon}) \rceil$.

(ii) *Let $\alpha = \left(\frac{\sqrt{3} \cdot 64 c_{upp} e}{c_{inf}^2} \right)^{2/7} \left(\frac{L_2^2 \Delta_f}{L^3 n} \right)^{\frac{1}{7}}$. If n is large enough to ensure $2(c_{upp} n)^{-1} < \alpha \leq 1$, then we have*

$$\mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \min_{k \in \{1, \dots, n\}} \|\nabla f(\bar{x}_k)\| \right] \leq \mathcal{B}_n L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{-4/7},$$

with

$$\mathcal{B}_n := (1 + C_n + C_n^2/4)H(\alpha)76^{-6/7}12^{3/7}(64e)^{1/7} \left(1 - \frac{2}{c_{upp}\alpha n}\right)^{-1} c_{upp}^2 \left(\frac{c_{upp}}{c_{inf}^2}\right)^{1/7} e^{2\frac{c_{upp}}{c_{inf}}C_n},$$

with $C_n = c_{inf}\alpha \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2\frac{n}{\varepsilon}) \rceil$ and $H(\alpha) = (1+\alpha)(1+2\alpha)(1+\alpha 8(1+\alpha))$.

In both statements, $\bar{x}_k = \sum_{i=0}^{k-1} \lambda_{i,k} \tilde{y}_i$ with $\lambda_{i,k} = \frac{e^{\alpha T_{i+1}}}{\sum_{j=1}^k e^{\alpha T_j}}$, and $\mathbb{P}(\mathbb{A}_n) \geq 1 - \varepsilon$.

While the constant ε determines the size of the set $\mathbb{A}_n$, the role of c_{inf} and c_{upp} might be more difficult to grasp. These two constants appear during the proof of Theorem 12. Choosing c_{inf} close to one, which makes c_{upp} close to 1, makes C_n increase. Choosing $c_{inf} \ll 1$, inducing $c_{upp} \gg 1$, reduces C_n but still hurts the bound because of factors such as c_{upp}^2 .

Proof. Corollary 16 and 17 provide bounds that involve $\min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\|$, which is about the continuous-time process $\bar{x}_t$. We have to make sure that the bound will be achievable by our algorithm, precisely, we have to show that the point $\arg \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\|$ can be outputted by (CNA). We have

$$\bar{x}_k := \bar{x}_{T_k} = \int_{[0, T_k]} w_{T_k}(s) x_{s-} dN_s = \sum_{i=1}^k w_{T_k}(T_i) x_{T_i-} = \sum_{i=0}^{k-1} w_{T_k}(T_{i+1}) \tilde{y}_i,$$

using $x_{T_i-} = \tilde{y}_{i-1}$ by Proposition 2. Also, recall that $w_t(s) = \frac{e^{\alpha s}}{\int_0^t e^{\alpha s} dN_s}$, such that we fix

$$\lambda_{i,k} := w_{T_k}(T_{i+1}) = \frac{e^{\alpha T_{i+1}}}{\sum_{j=1}^k e^{\alpha T_j}}.$$

So, for $\bar{x}_k = \sum_{i=0}^{k-1} \lambda_{i,k} \tilde{y}_i$, $\lambda_{i,k} = \frac{e^{\alpha T_{i+1}}}{\sum_{j=1}^k e^{\alpha T_j}}$, we have

$$\min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| = \min_{k \in \{1, \dots, n\}} \|\nabla f(\bar{x}_k)\|.$$

Then, the result is a direct combination of the bounds provided Corollary 16 and 17 with Lemma 18. We just have to make sure that $\left(1 - \frac{2}{c_{upp}\alpha n}\right) > 0$. In the case $\alpha = n^{-1/7}$, it holds as long as $n \geq 3$. More generally it is verified as long as $\alpha > \frac{2}{c_{upp}n}$. In the case $\alpha = \left(\frac{\sqrt{3} \cdot 64 c_{upp} e}{c_{inf}^2}\right)^{2/7} \left(\frac{L_2^2 \Delta_f}{L n}\right)^{1/7}$, it induces the following constraint on n

$$n > \left(\frac{\sqrt{3} \cdot 64 c_{upp} e}{c_{inf}^2}\right)^{-1/3} \left(\frac{L_2^2 \Delta_f}{L}\right)^{-1/6} 2^{7/6} c_{upp}^{-7/6}.$$

□

We can conclude with the proof of Theorem 5, which can be seen as a corollary of Theorem 19.

Proof of Theorem 5. It is a direct application of (ii) in Theorem 19, fixing $c_{inf} = 1 - \tilde{\varepsilon}$. Then, recalling the expression of $\mathcal{B}_n$

$$\mathcal{B}_n := (1 + C_n + C_n^2/4)H(\alpha)76^{-6/7}12^{3/7}(64e)^{1/7} \left(1 - \frac{2}{c_{upp}\alpha n}\right)^{-1} c_{upp}^2 \left(\frac{c_{upp}}{(1-\tilde{\varepsilon})^2}\right)^{1/7} e^{2\frac{c_{upp}}{1-\tilde{\varepsilon}}C_n}, \quad (21)$$

with $C_n = (1 - \tilde{\varepsilon})\alpha \lceil (-\tilde{\varepsilon} - \log(1 - \tilde{\varepsilon}))^{-1} \log(2\frac{n}{\tilde{\varepsilon}}) \rceil$ and $H(\alpha) = (1 + \alpha)(1 + 2\alpha)(1 + \alpha 8(1 + \alpha))$.

As $\alpha = \mathcal{O}(1/n^{1/7})$, we have that $C_n \rightarrow_{n \rightarrow +\infty} 0$, $H(\alpha) \rightarrow_{n \rightarrow +\infty} 1$, and $\left(1 - \frac{2}{c_{upp}\alpha n}\right)^{-1} \rightarrow_{n \rightarrow +\infty} 1$. So,

$$\lim_{n \rightarrow +\infty} \mathcal{B}_n = 76^{-6/7}12^{3/7}(64e)^{1/7} c_{upp}^2 \left(\frac{c_{upp}}{(1-\tilde{\varepsilon})^2}\right)^{1/7} \leq 9.2 c_{upp}^2 \left(\frac{c_{upp}}{(1-\tilde{\varepsilon})^2}\right)^{1/7}.$$

It remains to show the bound $c_{upp} \leq 1 + 2\tilde{\varepsilon}$, which is done in Lemma 41.

□

6 Conclusion

In this work, we show that (NM) with randomized parameter achieves the $\mathcal{O}(\varepsilon^{-7/4})$ complexity with high probability, demonstrating that it is not necessary to safeguard momentum with reset mechanisms. Prior to this work, the lack of necessity of such mechanisms was known in the continuous case [57], but not in the discrete case. It was argued in [34] that in general, the continuized method can tighten the gap between result derived for ODEs—such as (HB)—and for algorithms. In this work, we provide a new concrete example for the class of functions with Lipschitz gradient and Hessian.

In another line of work, we note that [15] showed that for functions with Lipschitz gradient, Hessian and third derivative, one can further improve the complexity to $\mathcal{O}(\varepsilon^{-5/3})$. The underlying algorithm is arguably complicated, but still based on (NM') as a core component. It is of interest whether simpler algorithms—such as (CNA)—can achieve this complexity, even with a certain probability.

Acknowledgment

This work was supported by PEPR PDE-AI and ANR SOS2ID (grant ANR-24-CE40-3786). We thank Raphaël Berthier for helpful discussions on the limit of the continuized system in the vanishing stepsize regime.

References

- [1] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, page 1195–1199, 2017.
- [2] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.
- [3] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.
- [4] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pages 3:1–3:22, 2017.
- [5] Hedy Attouch and Alexandre Cabot. Convergence rates of inertial forward-backward algorithms. *SIAM Journal on Optimization*, 28(1):849–874, 2018.
- [6] Jean-François Aujol, Luca Calatroni, Charles Dossal, Hippolyte Labarrière, and Aude Rondepierre. Parameter-free fista by adaptive restart and backtracking. *SIAM Journal on Optimization*, 34(4):3259–3285, 2024.
- [7] Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Optimal convergence rates for Nesterov acceleration. *SIAM Journal on Optimization*, 29(4):3131–3153, 2019.
- [8] Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of the Heavy-Ball method under the Łojasiewicz property. *Mathematical Programming*, 198(1):195–254, 2023.
- [9] Afonso S Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Conference on learning theory*, pages 361–382. PMLR, 2016.
- [10] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

- [11] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.
- [12] Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.
- [13] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *Advances in Neural Information Processing Systems*, 29, 2016.
- [14] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.
- [15] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. “convex until proven guilty”: Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pages 654–663. PMLR, 2017.
- [16] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.
- [17] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.
- [18] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021.
- [19] Antonin Chambolle and Charles H Dossal. On the convergence of the iterates of’ fista”. *Journal of Optimization Theory and Applications*, 166(3):25, 2015.
- [20] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.
- [21] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pages 6643–6670. PMLR, 2023.
- [22] Marina Danilova, Pavel Dvurechensky, Alexander Gasnikov, Eduard Gorbunov, Sergey Guminov, Dmitry Kamzolov, and Innokentiy Shibaev. Recent theoretical advances in non-convex optimization. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*, pages 79–163. Springer, 2022.
- [23] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.
- [24] Aaron Defazio. On the curved geometry of accelerated optimization. *Advances in Neural Information Processing Systems*, 32, 2019.
- [25] Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Hadrien Hendrikx, Pierre Gaillard, Laurent Massoulié, and Adrien Taylor. Continuized accelerations of deterministic and stochastic gradient descents, and of gossip algorithms. *Advances in Neural Information Processing Systems*, 34:28054–28066, 2021.
- [26] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International conference on machine learning*, pages 1233–1242. PMLR, 2017.
- [27] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

- [28] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.
- [29] Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Optimal first-order methods for convex functions with a quadratic upper bound. *arXiv preprint arXiv:2205.15033*, 2022.
- [30] Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method. *arXiv preprint arXiv:2307.11291*, 2023.
- [31] Kanan Gupta, Jonathan W Siegel, and Stephan Wojtowytsch. Nesterov acceleration despite very noisy gradients. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [32] Moritz Hardt. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 651–660. IEEE, 2014.
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [34] Julien Hermant, Jean-François Aujol, Charles Dossal, Lorick Huang, and Aude Rondepierre. Continuized nesterov acceleration for non-convex optimization. *arXiv preprint arXiv:2512.16533*, 2025.
- [35] Julien Hermant, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Study of the behaviour of nesterov accelerated gradient in a non convex setting: the strongly quasar convex case. *arXiv preprint arXiv:2405.19809*, 2024.
- [36] Julien Hermant, Marien Renaud, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Gradient correlation is a key ingredient to accelerate sgd with momentum. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [37] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, pages 1894–1938. PMLR, 2020.
- [38] Geoffrey Hinton. Neural networks for machine learning. Coursera Lecture 6e, 2012. Available at <https://www.coursera.org/learn/neural-networks>.
- [39] Ruichen Jiang, Aryan Mokhtari, and Francisco Patitucci. Improved complexity for smooth nonconvex optimization: a two-level online learning approach with quasi-newton methods. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 2225–2236, 2025.
- [40] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.
- [41] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.
- [42] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.
- [43] Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1):81–107, 2016.
- [44] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

- [45] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.
- [46] Huan Li and Zhouchen Lin. Restarted nonconvex accelerated gradient descent: No more poly-logarithmic factor in the in the $o(\epsilon^{-7/4})$ complexity. *Journal of Machine Learning Research*, 24(157):1–37, 2023.
- [47] Naoki Marumo and Akiko Takeda. Parameter-free accelerated gradient descent for nonconvex minimization. *SIAM J. Optim.*, 34:2093–2120, 2022.
- [48] Naoki Marumo and Akiko Takeda. Universal heavy-ball method for nonconvex optimization under hölder continuous Hessians. *Mathematical Programming*, 212(1):147–175, 2025.
- [49] Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto Imbuzeiro Oliveira. Solving SDPs for synchronization and maxcut problems via the Grothendieck inequality. In *Conference on learning theory*, pages 1476–1515. PMLR, 2017.
- [50] Adel Nabli, Eugene Belilovsky, and Edouard Oyallon. A²CiD²: Accelerating Asynchronous Communication in Decentralized Deep Learning. *Advances in Neural Information Processing Systems*, 36:47451–47474, 2023.
- [51] Adel Nabli and Edouard Oyallon. Dadao: Decoupled accelerated decentralized asynchronous optimization. In *International Conference on Machine Learning*, pages 25604–25626. PMLR, 2023.
- [52] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience*, 1983.
- [53] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.
- [54] Yurii Nesterov. Introductory lectures on convex optimization. applied optimization, 2004.
- [55] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.
- [56] Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006.
- [57] Kaito Okamura, Naoki Marumo, and Akiko Takeda. Heavy-ball differential equation achieves $\mathcal{O}(\epsilon^{-7/4})$ convergence for nonconvex functions. *arXiv preprint arXiv:2406.06100*, 2024.
- [58] Brendan O’Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.
- [59] Michael O’Neill and Stephen J Wright. Behavior of accelerated gradient methods near critical points of nonconvex functions. *Mathematical Programming*, 176(1):403–427, 2019.
- [60] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- [61] Philip E Protter. Stochastic differential equations. In *Stochastic integration and differential equations*, pages 249–361. Springer, 2012.
- [62] Marien Renaud, Julien Hermant, Deliang Wei, and Yu Sun. Provably accelerated imaging with restarted inertia and score-based image priors. *arXiv preprint arXiv:2510.07470*, 2025.
- [63] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.
- [64] Jonathan W Siegel. Accelerated first-order methods: Differential equations and Lyapunov functions. *arXiv preprint arXiv:1903.05671*, 2019.

- [65] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov’s accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.
- [66] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, 2013.
- [67] Kansei Ushiyama and Shun Sato. A restart-free accelerated algorithm for non-convex minimization: Continuous and discrete analysis. *arXiv preprint arXiv:2606.30050*, 2026.
- [68] Jun-Kun Wang and Andre Wibisono. Continuized acceleration for quasar convex functions in non-convex optimization, 2023.
- [69] Yi Xu, Rong Jin, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. *Advances in neural information processing systems*, 31, 2018.
- [70] Dongruo Zhou. Sharp first-order lower bounds for higher-order smooth nonconvex optimization. *arXiv preprint arXiv:2606.05438*, 2026.

A Complementary details to Section 1

A.1 Accelerated gradient algorithms under convexity

A seminal work [60] shows that the heavy ball momentum algorithm

$$x_{n+1} = x_n + \alpha(x_n - x_{n-1}) - \beta \nabla f(x_n)$$

can significantly improve over the convergence speed of gradient descent when minimizing strongly convex quadratic functions. [53, 54] further shows (NM) ensures an acceleration result for L -smooth (strongly)-convex functions. As mentioned in Section 1, the corresponding bounds are optimal among all first-order algorithms [52, 55]. There exists a rich literature building upon these contributions. Far from being exhaustive, it includes new interpretations of the algorithms [65, 4], extensions to the non-smooth case *e.g.* via composite proximal optimization [10, 19, 5], extensions to the stochastic gradient case [28, 2, 36], and acceleration results under relaxed geometrical assumptions [37, 8, 7]. To add some nuance, it is worth noting that even in the convex setting, inertial methods do not necessarily ensure accelerated convergence [29, 30].

A.2 Second-order stationary points

We provide a rigorous statement of the notion of second-order ε -stationary point, mentioned in our introductory section. We then briefly comment on it.

Definition 20 (Second-order ε -stationary point, [56]). *For a L_2 -Lipschitz Hessian function f , x is an ε -second-order stationary point if*

$$\|\nabla f(x)\| \leq \varepsilon, \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\varepsilon L_2}.$$

A second-order ε -stationary point approximates a second-order stationary point, namely a point x^* such that

$$\|\nabla f(x^*)\| = 0, \quad \lambda_{\min}(\nabla^2 f(x^*)) \geq 0.$$

In particular, such points are ensured to avoid *strict saddle-points*, that is, stationary points such that $\lambda_{\min}(\nabla^2 f(x^*)) < 0$. The 1d function $x \mapsto x^3$ is a simple example of a function that has a second-order stationary point (in $x = 0$) that is not a local minimizer, thus called *non-strict saddle-points*. Fortunately, several works indicate that many applications are prevented from such pathological stationary points [20, 42, 9, 49, 13, 11, 27, 26].

Algorithm 2: Negative curvature exploitation and Nesterov momentum - Detailed version
[41]

Input: $x_0, \eta, \theta, \gamma, s$
1 $v_0 \leftarrow 0$
2 **for** $t = 0, 1, \dots$ **do**
3 $y_t \leftarrow x_t + (1 - \theta)v_t$;
4 $x_{t+1} \leftarrow y_t - \eta \nabla f(y_t)$;
5 $v_{t+1} \leftarrow x_{t+1} - x_t$;
6 **if** $f(x_t) \leq f(y_t) + \langle \nabla f(y_t), x_t - y_t \rangle - \frac{\gamma}{2} \|x_t - y_t\|^2$ **then**
7 $(x_{t+1}, v_{t+1}) \leftarrow \text{Negative-Curvature-Exploitation}(x_t, v_t, s)$
8 **Negative-Curvature-Exploitation** (x_t, v_t, s)
9 **if** $\|v_t\| \geq s$ **then**
10 $x_{t+1} \leftarrow x_t$
11 **else**
12 $\delta = s \cdot v_t / \|v_t\|$
13 $x_{t+1} \leftarrow \operatorname{argmin}_{x \in \{x_t + \delta, x_t - \delta\}} f(x)$
14 **return** $(x_{t+1}, 0)$;

A.3 Detailed formulations of existing safety check and alternative mechanisms

For completeness, we give detailed versions of existing algorithms that achieve the complexity $\mathcal{O}(e^{-7/4})$. Algorithm 2 is a detailed version of an algorithm that performs a negative curvature step [41]. Algorithm 3 is a detailed version of an algorithm that uses a restart mechanism [46]. We make a few remarks:

- We chose to state Algorithm 2 as the version from [41] rather than the earlier one from [15], because of the significant length and trickiness of the latter.
- In its original statement, Algorithm 2 also involves a "stochastic perturbation step", used to find an ε -second-order stationary point (Definition 20). Because we do not study this concept, we avoided this step in our presentation.
- In the convergence result associated with Algorithm 3 [46, Theorem 1], the restart threshold is set to $B = \sqrt{\varepsilon/L_2}$, where ε denotes the target accuracy and L_2 is the Lipschitz constant of the Hessian. As noted by the authors, this value of B can be very small in practice, causing the algorithm to restart at nearly every iteration. In such cases, the method effectively reduces to standard gradient descent, which eliminates the practical benefit of momentum. This behavior is confirmed by numerical experiments, showing little to no empirical acceleration due to the near absence of momentum steps.

To address this issue, the authors propose an alternative restart scheme [46, Algorithm 2], in which restarts are allowed to occur less frequently. While keeping $B = \sqrt{\varepsilon/L_2}$, the user is permitted to initialize a larger value B_0 , potentially satisfying $B_0 \gg B$. The restart condition is then replaced by

$$k \sum_{t=0}^{k-1} \|x^{t+1} - x^t\|^2 > \max\{B^2, B_0^2\},$$

and B_0 is gradually decreased until $B_0 \leq B$, at which point the algorithm reduces to Algorithm 3. It is shown [46, Theorem 2] that this modification introduces an additional $\mathcal{O}\left(L^{1/2} \varepsilon^{-1/4} L_2^{1/4} \log\left(\frac{L_2 B_0}{\varepsilon}\right)\right)$ term in the gradient complexity, which does not significantly degrade the overall rate. This modification shows great empirical behaviour, enabling acceleration over gradient descent. However, this alternative scheme requires additional function evaluations, leading to a total complexity of $\mathcal{O}(\varepsilon^{-3/2})$ function queries.

Algorithm 3: Restarted Nesterov Momentum - Detailed version [46]

Input: $x^{-1} = x^0$, $B > 0$, K , η , θ .
1 $k \leftarrow 0$
2 **while** $k < K$ **do**
3 $y^k \leftarrow x^k + (1 - \theta)(x^k - x^{k-1})$
4 $x^{k+1} \leftarrow y^k - \eta \nabla f(y^k)$
5 $k \leftarrow k + 1$
6 **if** $k \sum_{t=0}^{k-1} \|x^{t+1} - x^t\|^2 > B^2$ **then**
7 $x^{-1} \leftarrow x^k$,
8 $x^0 \leftarrow x^k$,
9 $k \leftarrow 0$
10 $K_0 \leftarrow \arg \min_{\lfloor K/2 \rfloor \leq k \leq K-1} \|x^{k+1} - x^k\|$
11 $\hat{y} \leftarrow \frac{1}{K_0 + 1} \sum_{k=0}^{K_0} y^k$
12 **return** $\hat{y}$

B Proof of Lemma 4 and Proposition 3

Under Assumption 1, we have this classical descent lemma [54].

Lemma 21. *Under Assumption 1, if $\gamma \leq \frac{1}{L}$, we have*

$$f(x - \gamma \nabla f(x)) - f(x) \leq \gamma \left(\frac{L}{2} \gamma - 1 \right) \|\nabla f(x)\|^2.$$

To perform Lyapunov analysis, our tool is the following Itô formula, which intuitively allows stochastic derivation.

Proposition 22 ([25], Proposition 2). *Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a smooth function. Let $x_t \in \mathbb{R}^d$ be a solution of*

$$dx_t = \zeta(x_t)dt + G(x_t-)dN_t$$

where $\zeta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a locally Lipschitz function and $G : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is measurable. Then,

$$\varphi(x_t) = \varphi(x_0) + \int_0^t \langle \nabla \varphi(x_s), \zeta(x_s) \rangle ds + \int_{[0,t]} \varphi(x_s + G(x_s)) - \varphi(x_s) ds + M_t, \quad (22)$$

where M_t is a martingale such that $\mathbb{E}[M_t] = 0$, $\forall t \geq 0$.

We first restate Lemma 4.

Lemma 4. *If $(x_t, z_t)_{t \in \mathbb{R}_+}$ is solution of (CNE) under Assumption 1 with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \gamma' - \gamma$ and $\eta' > -\eta$, we have*

$$\mathbb{E} \left[\int_0^t (\eta + \eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \right] \leq \mathbb{E}[f(x_0) - f^*]. \quad (1)$$

Proof. Let $\bar{x}_t = (x_t, z_t)$ where $(x_t, z_t)_{t \in \mathbb{R}_+}$ satisfies (CNE). It satisfies $d\bar{x}_t = \zeta(\bar{x}_t)dt + G(\bar{x}_t)dN(t)$, where

$$\zeta(\bar{x}_t) = \begin{pmatrix} \eta(z_t - x_t) \\ \eta'(x_t - z_t) \end{pmatrix}, \quad G(\bar{x}_t) = \begin{pmatrix} -\gamma_t \nabla f(x_t) \\ -\gamma'_t \nabla f(x_t) \end{pmatrix}$$

We apply Proposition 22 to $\varphi(\bar{x}_t)$, where

$$\varphi(x, z) = f(x) + \frac{1}{2} \|x - z\|^2,$$

inducing

$$\varphi(\bar{x}_t) = \varphi(\bar{x}_0) + \int_0^t \langle \nabla \varphi(\bar{x}_s), \zeta(\bar{x}_s) \rangle ds + \int_0^t \varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s) ds + M_t, \quad (23)$$

where M_t is a martingale.

Computations We compute $\langle \nabla \varphi(\bar{x}_s), \zeta(\bar{x}_s) \rangle$ and $\varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s)$. We have

$$\frac{\partial \varphi}{\partial x} = \nabla f(x) + x - z, \quad \frac{\partial \varphi}{\partial z} = z - x.$$

So,

$$\begin{aligned} \langle \nabla \varphi(\bar{x}_s), \zeta(\bar{x}_s) \rangle &= \langle \nabla f(x_s) + x_s - z_s, \eta(z_s - x_s) \rangle + \langle z_s - x_s, \eta'(x_s - z_s) \rangle \\ &= \eta \langle \nabla f(x_s), z_s - x_s \rangle - (\eta + \eta') \|x_s - z_s\|^2. \end{aligned} \quad (24)$$

Also,

$$\begin{aligned} \varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s) &= f(x_s - \gamma \nabla f(x_s)) - f(x_s) \\ &\quad + \frac{1}{2} \left(\|x_s - \gamma \nabla f(x_s) - (z_s - \gamma' \nabla f(x_s))\|^2 - \|x_s - z_s\|^2 \right) \\ &= f(x_s - \gamma \nabla f(x_s)) - f(x_s) + \frac{(\gamma' - \gamma)^2}{2} \|\nabla f(x_s)\|^2 \\ &\quad + (\gamma' - \gamma) \langle \nabla f(x_s), x_s - z_s \rangle \\ &\leq ((\gamma' - \gamma)^2 - \gamma(2 - L\gamma)) \frac{1}{2} \|\nabla f(x_s)\|^2 \\ &\quad + (\gamma' - \gamma) \langle \nabla f(x_s), x_s - z_s \rangle. \end{aligned} \quad (25)$$

The last inequality uses Lemma 21, assuming $\gamma \leq \frac{1}{L}$. We combine (24) and (25)

$$\begin{aligned} &\langle \nabla \varphi(\bar{x}_s), \zeta(\bar{x}_s) \rangle + \varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s) \\ &\leq (\eta - (\gamma' - \gamma)) \langle \nabla f(x_s), z_s - x_s \rangle - (\eta + \eta') \|x_s - z_s\|^2 \\ &\quad + ((\gamma' - \gamma)^2 - \gamma(2 - L\gamma)) \frac{1}{2} \|\nabla f(x_s)\|^2 \end{aligned} \quad (26)$$

Parameter tuning We now choose the parameters driving (CNE) such that we obtain the desired result. We set $\eta = \gamma' - \gamma$ to cancel scalar product. Then fixing $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$ ensures

$$(\gamma' - \gamma)^2 - \gamma(2 - L\gamma) \leq \frac{\gamma}{2} - \gamma = -\frac{\gamma}{2}.$$

The latter holds because $\gamma \leq \frac{1}{L} \Rightarrow -\gamma(2 - L\gamma) \leq -\gamma$. We thus have

$$\langle \nabla \varphi(\bar{x}_s), \zeta(\bar{x}_s) \rangle + \varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s) \leq -(\eta + \eta') \|x_s - z_s\|^2 - \frac{\gamma}{4} \|\nabla f(x_s)\|^2. \quad (27)$$

Conclusion Combining (27), (26) and (23), we get

$$\begin{aligned} \varphi(\bar{x}_t) &\leq \varphi(\bar{x}_0) + \int_0^t -(\eta + \eta') \|x_s - z_s\|^2 - \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds + M_t \\ &\Rightarrow \int_0^t (\eta + \eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \leq f(x_0) - f^* + M_t, \end{aligned} \quad (28)$$

using $\varphi(\bar{x}_0) - \varphi(\bar{x}_t) \leq f(x_0) - \min_x f(x)$ and that $x_0 = z_0$. It remains to take expectation on (28)

$$\mathbb{E} \left[\int_0^t (\eta + \eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \right] \leq \mathbb{E} [f(x_0) - f^*],$$

where we used $\mathbb{E}[M_t] = \mathbb{E}[M_0] = 0$. □

We now prove the discrete version of the result, which we first restate.

Proposition 3. *Assume $T_1, \dots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d. of law $\mathcal{E}(1)$, with convention $T_0 = 0$. Under Assumption 1, iterations of (CNA) with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \sqrt{\frac{\gamma}{2}}$ and $\eta' > -\eta$ verify*

$$\mathbb{E} \left[\min_{0 \leq i \leq k-1} \|\nabla f(\tilde{y}_i)\|^2 \right] \leq \frac{4}{\gamma} \frac{\Delta_f}{k}.$$

Proof. From (28), we have

$$\int_0^t \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \leq f(x_0) - f^* + M_t.$$

Also, we have

$$\int_0^t \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds = \int_0^t \frac{\gamma}{4} \|\nabla f(x_{s-})\|^2 dN_s + \int_0^t \frac{\gamma}{4} \|\nabla f(x_{s-})\|^2 (ds - dN_s), \quad (29)$$

where $\int_0^t \frac{\gamma}{4} \|\nabla f(x_{s-})\|^2 (ds - dN_s)$ is a centered martingale. Then, there exists a centered martingale U_t such that

$$\int_0^t \frac{\gamma}{4} \|\nabla f(x_{s-})\|^2 dN_s \leq \Delta_f + U_t.$$

We use the following stopping theorem.

Theorem 23 ([34], Theorem 6). *Let $(\varphi_t)_{t \in \mathbb{R}_+}$ be a nonnegative process with cadlag trajectories, such that it verifies*

$$\varphi_t \leq K_0 + M_t,$$

for some positive random variable K_0 , some martingale $(M_t)_{t \geq 0}$ with $M_0 = 0$. Then, for an almost surely finite stopping time τ , one has

$$\mathbb{E}[\varphi_\tau] \leq \mathbb{E}[K_0].$$

Applying Theorem 23 with $\tau = T_k$, $\varphi_t = \int_0^t \frac{\gamma}{4} \|\nabla f(x_{s-})\|^2 dN_s$, $K_0 = \Delta_f$ and $M_t = U_t$, we obtain

$$\mathbb{E} \left[\int_0^{T_k} \|\nabla f(x_{s-})\|^2 dN_s \right] \leq \frac{4}{\gamma} \Delta_f.$$

Finally, we have

$$\int_0^{T_k} \|\nabla f(x_{s-})\|^2 dN_s = \sum_{i=1}^k \left\| \nabla f(x_{T_i^-}) \right\|^2 = \sum_{i=1}^k \|\nabla f(\tilde{y}_{i-1})\|^2 \geq k \min_{0 \leq i \leq k-1} \|\nabla f(\tilde{y}_i)\|^2,$$

thus the result. $\square$

C Proof of Theorem 7

Recall $\gamma' = \gamma + (\gamma/2)^{\frac{1}{2}}$, $\eta = (\gamma/2)^{\frac{1}{2}}$, $\eta'_T = \alpha_{(\gamma/2)^{-1/2}T} - (\gamma/2)^{\frac{1}{2}}$. From the proof of Lemma 4, we have

$$\mathbb{E}[f(x_t) - f^*] \leq \mathbb{E}[f(x_0) - f^*].$$

Also, by L -smoothness we have $\|\nabla f(x)\|^2 \leq 2L(f(x) - f^*)$ for any $x \in \mathbb{R}^d$ [54, Section 1.2.3], such that for any $t \geq 0$,

$$\mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \leq 2L(f(x_0) - f^*).$$

Taking the sup on t , we get

$$\sup_{t>0} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \leq 2L(f(x_0) - f^*). \quad (30)$$

We define the rescaled process $(x_s^\gamma, z_s^\gamma) := (x_{(\gamma/2)^{-1/2}s}, z_{(\gamma/2)^{-1/2}s})$ and $N_s^\gamma := N_{(\gamma/2)^{-1/2}s}$ on the interval $s \in [0, T]$. This means that the process $(x_t)_{t \geq 0}$ has to be define on the interval of time $s \in [0, (\gamma/2)^{-1/2}T]$, which justifies the choice of terminal time in $\alpha_{(\gamma/2)^{-1/2}T}$. We also have that $M_s^\gamma := N_s^\gamma - (\gamma/2)^{-1/2}s$ is a martingale, whose predictable quadratic variation is

$$\langle M_s^\gamma \rangle = (\gamma/2)^{-1/2}s. \quad (31)$$

The rescaled process satisfies

$$\begin{cases} dx_s^\gamma &= \eta(\gamma/2)^{-1/2}(z_s^\gamma - x_s^\gamma)ds - (\gamma/2)^{1/2}\nabla f(x_{s-}^\gamma)ds - \gamma\nabla f(x_{s-}^\gamma)dM_s^\gamma, \\ dz_s^\gamma &= \eta'(\gamma/2)^{-1/2}(x_s^\gamma - z_s^\gamma)ds - \gamma'(\gamma/2)^{-1/2}\nabla f(x_{s-}^\gamma)ds - \gamma'\nabla f(x_{s-}^\gamma)dM_s^\gamma, \end{cases}$$

which becomes, once the parameters are replaced by their dependence in γ

$$\begin{cases} dx_s^\gamma &= (z_s^\gamma - x_s^\gamma)ds - (\gamma/2)^{1/2}\nabla f(x_{s-}^\gamma)ds - \gamma\nabla f(x_{s-}^\gamma)dM_s^\gamma, \\ dz_s^\gamma &= (2^{\frac{3}{7}}K_{\tilde{\varepsilon}, c_{upp}} \left(\frac{L_2^2 \Delta_f}{T} \right)^{1/7} - 1)(x_s^\gamma - z_s^\gamma)ds - ((2\gamma)^{1/2} + 1)\nabla f(x_{s-}^\gamma)ds - \gamma'\nabla f(x_{s-}^\gamma)dM_s^\gamma, \end{cases}$$

By (30), we deduce

$$\begin{aligned} \mathbb{E} \left[\int_0^t \|\nabla f(x_s^\gamma)\|^2 ds \right] &= \int_0^t \mathbb{E} \left[\|\nabla f(x_s^\gamma)\|^2 \right] ds \\ &\leq t\Delta_f. \end{aligned} \quad (32)$$

In the following step, we show that some of the component converges uniformly to zero in probability as γ goes to zero.

Step 1. We have

$$\mathbb{E} \left[\sup_{s \in [0, T]} \left\| \gamma^{1/2} \int_0^s \nabla f(x_r^\gamma) dr \right\|^2 \right] \leq \mathbb{E} \left[\sup_{s \in [0, T]} \gamma s \int_0^s \|\nabla f(x_r^\gamma)\|^2 dr \right] \leq \mathbb{E} \left[\gamma T \int_0^T \|\nabla f(x_r^\gamma)\|^2 dr \right],$$

where we used Cauchy-Schwartz. Using (32), we obtain

$$\mathbb{E} \left[\sup_{s \in [0, T]} \left\| \gamma^{1/2} \int_0^s \nabla f(x_r^\gamma) dr \right\|^2 \right] \leq \gamma T^2 \Delta_f \xrightarrow{\gamma \rightarrow 0} 0,$$

which implies that $\sup_{s \in [0, T]} \|\gamma^{1/2} \int_0^s \nabla f(x_r^\gamma) dr\|$ converges to zero in probability.

Then, by Doob's inequality, we get

$$\mathbb{E} \left[\sup_{s \in [0, T]} \left\| \gamma \int_0^s \nabla f(x_r^\gamma) dM_r^\gamma \right\|^2 \right] \leq 4\mathbb{E} \left[\left\| \gamma \int_0^T \nabla f(x_r^\gamma) dM_r^\gamma \right\|^2 \right].$$

By an isometry theorem and (31), we deduce

$$4\mathbb{E} \left[\left\| \gamma \int_0^T \nabla f(x_r^\gamma) dM_r^\gamma \right\|^2 \right] = 4\mathbb{E} \left[\gamma^2 \int_0^T \|\nabla f(x_r^\gamma)\|^2 d\langle M_r^\gamma \rangle \right] = 4\mathbb{E} \left[\gamma^{3/2} \int_0^T \|\nabla f(x_r^\gamma)\|^2 dr \right].$$

Using again (32), we obtain

$$\mathbb{E} \left[\sup_{s \in [0, T]} \left\| \gamma \int_0^s \nabla f(x_r^\gamma) dM_r^\gamma \right\|^2 \right] \leq \gamma^{3/2} T \Delta_f \xrightarrow{\gamma \rightarrow 0} 0,$$

which implies $\sup_{s \in [0, T]} \|\gamma \int_0^s \nabla f(x_r^\gamma) dM_r^\gamma\| \xrightarrow{\gamma \rightarrow 0} 0$ in probability.

There remains one term, for which we use the same arguments.

$$\mathbb{E} \left[\sup_{s \in [0, T]} \left\| \gamma' \int_0^s \nabla f(x_r^\gamma) dM_r^\gamma \right\|^2 \right] \leq 4\mathbb{E} \left[\gamma'^2 \int_0^s \|\nabla f(x_r^\gamma)\|^2 d\langle M_r^\gamma \rangle \right] = 4\mathbb{E} \left[\gamma'^2 \gamma^{-1/2} \int_0^T \|\nabla f(x_r^\gamma)\|^2 dr \right].$$

Using (32), it becomes

$$\mathbb{E} \left[\sup_{s \in [0, T]} \left\| \gamma' \int_0^s \nabla f(x_r^\gamma) dM_r^\gamma \right\|^2 \right] \leq 4\gamma'^2 \gamma^{-1/2} T \Delta_f \leq 4\gamma^{1/2} \Delta_f \rightarrow_{\gamma \rightarrow 0} 0.$$

Which implies $\sup_{s \in [0, T]} \|\gamma' \int_0^s \nabla f(x_s^\gamma) dM_s^\gamma\| \rightarrow 0$ in probability.

Step 2. Now, we define

$$\begin{cases} dX_s &= (Z_s - X_s)ds, \\ dZ_s &= (2^{\frac{3}{7}} K_{\varepsilon, c_{upp}} \left(\frac{L_2^2 \Delta_f}{T} \right)^{1/7} - 1)(X_s - Z_s)ds - \nabla f(X_s)ds, \end{cases} \quad (33)$$

We want to show that the rescaled process converges in probability to this system with $\gamma \rightarrow 0$. We note $\Delta X_s^\gamma := X_s - x_s^\gamma$ and $\Delta Z_s^\gamma := Z_s - z_s^\gamma$. We have

$$\Delta X_s^\gamma = \int_0^s (\Delta Z_r^\gamma - \Delta X_r^\gamma) dr + R_{X, \gamma}(s),$$

where $R_{X, \gamma}(s) := -(\gamma/2)^{1/2} \int_0^s \nabla f(x_s^\gamma) ds - \gamma \int_0^s \nabla f(x_s^\gamma) dM_s^\gamma$, and

$$\Delta Z_s^\gamma = A \int_0^s (\Delta X_r^\gamma - \Delta Z_r^\gamma) dr - \int_0^s (\nabla f(X_r) - \nabla f(x_r^\gamma)) dr + R_{Z, \gamma}(s),$$

denoting $A := (2^{\frac{3}{7}} K_{\varepsilon, c_{upp}} \left(\frac{L_2^2 \Delta_f}{T} \right)^{1/7} - 1)$ and $R_{(Z, \gamma)}(s) = -(2\gamma)^{1/2} \int_0^s \nabla f(x_s^\gamma) ds - \gamma' \int_0^s \nabla f(x_s^\gamma) dM_s^\gamma$.

Using several time the triangular inequality, and $\|\int \cdot ds\| \leq \int \|\cdot\| ds$, we get

$$\|\Delta X_s^\gamma\| \leq \int_0^s (\|\Delta Z_r^\gamma\| + \|\Delta X_r^\gamma\|) dr + \|R_{X, \gamma}(s)\|.$$

Similarly,

$$\|\Delta Z_s^\gamma\| \leq A \int_0^s (\|\Delta Z_r^\gamma\| + \|\Delta X_r^\gamma\|) dr + \int_0^s \|\nabla f(X_r) - \nabla f(x_r^\gamma)\| dr + \|R_{Z, \gamma}(s)\|,$$

and using also that ∇f is L -Lipschitz

$$\|\Delta Z_s^\gamma\| \leq A \int_0^s (\|\Delta Z_r^\gamma\| + \|\Delta X_r^\gamma\|) dr + L \int_0^s \|\Delta X_r^\gamma\| dr + \|R_{Z, \gamma}(s)\|.$$

Using the rough inequality $\|\Delta X_r^\gamma\| \leq \|\Delta Z_r^\gamma\| + \|\Delta X_r^\gamma\|$, we get

$$\|\Delta X_s^\gamma\| + \|\Delta Z_s^\gamma\| \leq (A + L) \int_0^s (\|\Delta Z_r^\gamma\| + \|\Delta X_r^\gamma\|) dr + \|R_{X, \gamma}(s)\| + \|R_{Z, \gamma}(s)\|. \quad (34)$$

We take the sup on $[0, s]$, denoting $D_\gamma(r) := \sup_{s \in [0, r]} (\|\Delta X_s^\gamma\| + \|\Delta Z_s^\gamma\|)$, and use Fatou's Lemma to obtain

$$D_\gamma(s) \leq (A + L) \int_0^s D_\gamma(r) dr + \sup_{s \in [0, T]} \|R_{X, \gamma}(s)\| + \sup_{s \in [0, T]} \|R_{Z, \gamma}(s)\|.$$

Using the Grönwall Lemma, we deduce

$$D_\gamma(T) \leq e^{(A+L)T} \left(\sup_{s \in [0, T]} \|R_{X, \gamma}(s)\| + \sup_{s \in [0, T]} \|R_{Z, \gamma}(s)\| \right).$$

Yet, from Step 1 we can deduce

$$\sup_{s \in [0, T]} \|R_{X, \gamma}(s)\| + \sup_{s \in [0, T]} \|R_{Z, \gamma}(s)\| \xrightarrow{\gamma \rightarrow 0} 0$$

in probability. We can conclude that in probability, $D_\gamma(T) \xrightarrow{\gamma \rightarrow 0} 0$, or

$$\sup_{s \in [0, T]} \|X_s - x_s^\gamma\| \xrightarrow{\mathbb{P}}_{\gamma \rightarrow 0} 0, \quad \sup_{s \in [0, T]} \|Z_s - z_s^\gamma\| \xrightarrow{\mathbb{P}}_{\gamma \rightarrow 0} 0.$$

To conclude, we use [34, Proposition 27], to deduce that X_s can be rewritten as the solution of the following ODE

$$\ddot{X}_s + 2^{\frac{3}{7}} K_{\tilde{\varepsilon}, \text{cupp}} \left(\frac{L_2^2 \Delta f}{T} \right)^{1/7} \dot{X}_s + \nabla f(X_s) = 0.$$

D Proof of the Lemmas stated in Section 5

We prove formally all the Lemmas used in Section 5 to prove Theorem 5.

D.1 Proof of Lemma 8

The proof follows similar steps as for Lemma 1 in [57]. We show the following inequalities

$$\begin{aligned} & \left\| \int_0^t w_t(s) \nabla f(x_{s-}) dN_s - \nabla f(\bar{x}_t) \right\| \\ & \leq \frac{L_2}{2} \int_0^t w_t(s) \|x_{s-} - \bar{x}_t\|^2 dN_s \end{aligned} \quad (35)$$

$$= \frac{L_2}{2} \int \int_{0 \leq \sigma \leq \tau \leq t} w_t(\sigma) w_t(\tau) \|x_{\tau-} - x_{\sigma-}\|^2 dN_\sigma dN_\tau \quad (36)$$

$$\begin{aligned} & \leq L_2 \int_0^t \eta^2 \|z_s - x_s\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds \\ & + L_2 \int_0^t \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau \right) dN_s. \end{aligned} \quad (37)$$

Inequality (35). Using Taylor expansion of the gradient, we have

$$\begin{aligned} & \nabla f(x_{s-}) - \nabla f(\bar{x}_t) \\ & = \int_0^1 \nabla^2 f((1 - \sigma)\bar{x}_t + \sigma x_{s-})(x_{s-} - \bar{x}_t) d\sigma \\ & = \nabla^2 f(\bar{x}_t)(x_{s-} - \bar{x}_t) + \int_0^1 (\nabla^2 f((1 - \sigma)\bar{x}_t + \sigma x_{s-}) - \nabla^2 f(\bar{x}_t))(x_{s-} - \bar{x}_t) d\sigma. \end{aligned}$$

Multiply both sides by $w_t(s)$ and integrate with respect to $dN(\cdot)$ on $s \in [0, t]$

$$\begin{aligned} & \int_0^t w_t(s) (\nabla f(x_{s-}) - \nabla f(\bar{x}_t)) dN_s \\ & = \int_0^t w_t(s) \nabla^2 f(\bar{x}_t)(x_{s-} - \bar{x}_t) dN_s \\ & + \int_0^t w_t(s) \int_0^1 (\nabla^2 f((1 - \sigma)\bar{x}_t + \sigma x_{s-}) - \nabla^2 f(\bar{x}_t))(x_{s-} - \bar{x}_t) d\sigma dN_s. \end{aligned} \quad (38)$$

Because $\int_0^t w_t(s) dN_s = 1$, we have

$$\int_0^t w_t(s) (\nabla f(x_{s-}) - \nabla f(\bar{x}_t)) dN_s = \int_0^t w_t(s) \nabla f(x_{s-}) dN_s - \nabla f(\bar{x}_t),$$

and because $\bar{x}_t = \int_0^t w_t(s) x_{s-} dN_s$, we get

$$\int_0^t w_t(s) \nabla^2 f(\bar{x}_t) (x_{s-} - \bar{x}_t) dN_s = 0.$$

Also,

$$\begin{aligned} & \left\| \int_0^t w_t(s) \int_0^1 (\nabla^2 f((1-\sigma)\bar{x}_t + \sigma x_{s-}) - \nabla^2 f(\bar{x}_t)) (x_{s-} - \bar{x}_t) d\sigma dN_s \right\| \\ & \leq \int_0^t w_t(s) \int_0^1 \left\| \nabla^2 f((1-\sigma)\bar{x}_t + \sigma x_{s-}) - \nabla^2 f(\bar{x}_t) \right\|_2 \|x_{s-} - \bar{x}_t\| d\sigma dN_s \\ & \leq \int_0^t w_t(s) \int_0^1 \sigma L_2 \|x_{s-} - \bar{x}_t\|^2 d\sigma dN_s \quad (\text{Assumption 2}) \\ & = \frac{L_2}{2} \int_0^t w_t(s) \|x_{s-} - \bar{x}_t\|^2 dN_s. \end{aligned}$$

Finally, taking the norm, (38) becomes

$$\left\| \int_0^t w_t(s) \nabla f(x_{s-}) dN_s - \nabla f(\bar{x}_t) \right\| \leq \frac{L_2}{2} \int_0^t w_t(s) \|x_{s-} - \bar{x}_t\|^2 dN_s.$$

Equality (36). We have

$$\begin{aligned} \left\| \int_0^t w_t(\sigma) x_{\sigma-} dN_\sigma \right\|^2 &= \left\langle \int_0^t w_t(\sigma) x_{\sigma-} dN_\sigma, \int_0^t w_t(\tau) x_{\tau-} dN_\tau \right\rangle \\ &= \int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \langle x_{\sigma-}, x_{\tau-} \rangle dN_\sigma dN_\tau. \end{aligned} \quad (39)$$

Also

$$\begin{aligned} & \int \int_{0 \leq \sigma \leq \tau \leq t} w_t(\sigma) w_t(\tau) \|x_{\tau-} - x_{\sigma-}\|^2 dN_\sigma dN_\tau \\ &= \frac{1}{2} \int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \|x_{\tau-} - x_{\sigma-}\|^2 dN_\sigma dN_\tau \\ &= \int_0^t w_t(\sigma) \|x_{\sigma-}\|^2 dN_\sigma - \int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \langle x_{\sigma-}, x_{\tau-} \rangle dN_\sigma dN_\tau \end{aligned}$$

where the last equality is obtained by developing the squared norm, and because

$$\int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \|x_{\tau-}\|^2 dN_\sigma dN_\tau = \int_0^t w_t(\tau) \|x_{\tau-}\|^2 dN_\tau \int_0^t w_t(\sigma) dN_\sigma = \int_0^t w_t(\tau) \|x_{\tau-}\|^2 dN_\tau.$$

Then,

$$\begin{aligned}
& \int_0^t w_t(\sigma) \|x_{\sigma-}\|^2 dN_\sigma - \int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \langle x_{\sigma-}, x_{\tau-} \rangle dN_\sigma dN_\tau \\
&= \int_0^t w_t(\sigma) \|x_{\sigma-}\|^2 dN_\sigma - \left\| \int_0^t w_t(\sigma) x_{\sigma-} dN_\sigma \right\|^2, \quad \text{using (39)} \\
&= \int_0^t w_t(\sigma) \|x_{\sigma-}\|^2 dN_\sigma - \|\bar{x}_t\|^2 \\
&= \int_0^t w_t(\sigma) \|x_{\sigma-}\|^2 dN_\sigma - 2\|\bar{x}_t\|^2 + \|\bar{x}_t\|^2 \\
&= \int_0^t w_t(\sigma) \|x_{\sigma-}\|^2 dN_\sigma - 2 \left\langle \int_0^t w_t(\sigma) x_{\sigma-} dN_\sigma, \underbrace{\int_0^t w_t(\sigma) x_{\sigma-} dN_\sigma}_{=\bar{x}_t} \right\rangle \\
&\quad + \|\bar{x}_t\|^2 \underbrace{\int_0^t w_t(\sigma) dN_\sigma}_{=1} \\
&= \int_0^t w_t(\sigma) \|x_{\sigma-}\|^2 dN_\sigma - 2 \int_0^t w_t(\sigma) \langle x_{\sigma-}, \bar{x}_t \rangle dN_\sigma + \int_0^t w_t(\sigma) \|\bar{x}_t\|^2 dN_\sigma \\
&= \int_0^t w_t(\sigma) \|x_{\sigma-} - \bar{x}_t\|^2 dN_\sigma.
\end{aligned}$$

Inequality (37). By definition of $(x_t)_{t \geq 0}$, for all $\sigma \in [0, t]$ we have

$$x_t = x_{\sigma-} + \int_\sigma^t \eta(z_s - x_s) ds - \int_{[\sigma, t]} \gamma \nabla f(x_{s-}) dN_s$$

Taking the left-limit at t , it becomes

$$x_{t-} = x_{\sigma-} + \int_\sigma^t \eta(z_s - x_s) ds - \int_{[\sigma, t)} \gamma \nabla f(x_{s-}) dN_s$$

For $0 \leq \sigma \leq \tau \leq t$, using $\|a - b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for $a, b \in \mathbb{R}^d$ and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\|x_{\tau-} - x_{\sigma-}\|^2 &= \left\| \int_\sigma^\tau \eta(z_s - x_s) ds - \int_{[\sigma, \tau)} \gamma \nabla f(x_{s-}) dN_s \right\|^2 \\
&\leq 2 \left\| \int_\sigma^\tau \eta(z_s - x_s) ds \right\|^2 + 2 \left\| \int_{[\sigma, \tau)} \gamma \nabla f(x_{s-}) dN_s \right\|^2 \\
&\leq 2 \left(\int_\sigma^\tau \|\eta(z_s - x_s)\| ds \right)^2 + 2 \left(\int_{[\sigma, \tau)} \|\gamma \nabla f(x_{s-})\| dN_s \right)^2 \\
&\leq 2 \left(\int_\sigma^\tau \|\eta(z_s - x_s)\|^2 ds \right) \left(\int_\sigma^\tau ds \right) + 2 \left(\int_{[\sigma, \tau)} dN_s \right) \left(\int_{[\sigma, \tau)} \|\gamma \nabla f(x_{s-})\|^2 dN_s \right) \\
&\leq 2(\tau - \sigma) \left(\int_\sigma^\tau \eta^2 \|(z_s - x_s)\|^2 ds \right) + 2(N_{\tau-} - N_{\sigma-}) \int_\sigma^\tau \gamma^2 \|\nabla f(x_{s-})\|^2 dN_s.
\end{aligned} \tag{40}$$

In the last inequality, we used $\int_{[\sigma, \tau]} \|\gamma \nabla f(x_{s-})\|^2 dN_s \leq \int_{[\sigma, \tau]} \|\gamma \nabla f(x_{s-})\|^2 dN_s$, recalling our notation $\int_{\sigma}^{\tau} dN_s = \int_{[\sigma, \tau]}$. Then, we have

$$\begin{aligned}
& \int \int_{0 \leq \sigma \leq \tau \leq t} w_t(\sigma) w_t(\tau) \|x_{\tau-} - x_{\sigma-}\|^2 dN_{\sigma} dN_{\tau} \\
& \stackrel{(40)}{\leq} \int \int_{0 \leq \sigma \leq \tau \leq t} w_t(\sigma) w_t(\tau) 2(\tau - \sigma) \left(\int_{\sigma}^{\tau} \eta^2 \|z_s - x_s\|^2 ds \right) dN_{\sigma} dN_{\tau} \\
& + \int \int_{0 \leq \sigma \leq \tau \leq t} w_t(\sigma) w_t(\tau) 2(N_{\tau-} - N_{\sigma-}) \left(\int_{\sigma}^{\tau} \gamma^2 \|\nabla f(x_{s-})\|^2 dN_s \right) dN_{\sigma} dN_{\tau} \\
& = \int \int_{0 \leq \sigma \leq s \leq \tau \leq t} w_t(\sigma) w_t(\tau) 2(\tau - \sigma) \eta^2 \|z_s - x_s\|^2 ds dN_{\sigma} dN_{\tau} \\
& + \int \int_{0 \leq \sigma \leq s \leq \tau \leq t} w_t(\sigma) w_t(\tau) 2(N_{\tau-} - N_{\sigma-}) \gamma^2 \|\nabla f(x_{s-})\|^2 dN_s dN_{\sigma} dN_{\tau} \\
& = 2 \int_0^t \eta^2 \|z_s - x_s\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_{\sigma} dN_{\tau} \right) ds \\
& + 2 \int_0^t \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau-} - N_{\sigma-}) dN_{\sigma} dN_{\tau} \right) dN_s.
\end{aligned}$$

D.2 Proof of Lemma 9

We set $\tilde{w}(s) = \alpha e^{\alpha s}$, such that $s \mapsto \tilde{w}(s)$ is differentiable, satisfying

$$\frac{d\tilde{w}(s)}{ds} = \alpha \tilde{w}(s).$$

Summing the two lines of (CNE) and using $\eta + \eta' = \alpha$, we get

$$dx_s - dz_s = \alpha(z_s - x_s)ds - (\gamma - \gamma')\nabla f(x_{s-})dN_s. \quad (41)$$

We multiply each side of (41) by $\tilde{w}(s)$ and use $d\tilde{w}(s) = \alpha \tilde{w}(s)ds$ to get

$$\begin{aligned}
\tilde{w}(s)(dx_s - dz_s) &= \tilde{w}(s)\alpha(z_s - x_s)ds - \tilde{w}(s)(\gamma - \gamma')\nabla f(x_{s-})dN_s \\
&= (z_s - x_s)d\tilde{w}(s) - \tilde{w}(s)(\gamma - \gamma')\nabla f(x_{s-})dN_s.
\end{aligned}$$

By rearranging, we obtain

$$\tilde{w}(s)(dx_s - dz_s) + (x_s - z_s)d\tilde{w}(s) = -\tilde{w}(s)(\gamma - \gamma')\nabla f(x_{s-})dN_s. \quad (42)$$

We write (42) in an integral form, such that

$$\int_0^t \tilde{w}(s)(dx_s - dz_s) + \int_0^t (x_s - z_s)d\tilde{w}(s) = -(\gamma - \gamma') \int_0^t \tilde{w}(s)\nabla f(x_{s-})dN_s \quad (43)$$

We use the following integration by part formula for stochastic integrals, see a precise statement in [61, p.83].

Corollary 24 ([61]). *Let X, Y be semimartingales with at least one of X or Y continuous. Then*

$$X_t Y_t = X_0 Y_0 + \int_0^t X_{s-} dY_s + \int_0^t Y_{s-} dX_s.$$

We apply Corollary 24 to $X_t = x_t - z_t$ and $Y_t = \tilde{w}(t)$, such that $X_0 Y_0 = 0$, and so

$$\int_0^t \tilde{w}(s)(dx_s - dz_s) + \int_0^t (x_s - z_s)d\tilde{w}(s) = \tilde{w}(t)(x_t - z_t). \quad (44)$$

We combine (43) with (44), and multiply both sides by $\left(\int_0^t \alpha e^{\alpha s} dN_s\right)^{-1}$ to conclude. Note that we can use Corollary 24 as $t \mapsto \tilde{w}(t)$ is deterministic and continuous, and $x_t - z_t$ is a semimartingale as a difference of semimartingales. To see that $(x_t)_{t \geq 0}$ is a semi martingale, remark that it writes

$$\begin{aligned} x_t &= x_0 + \eta \int_0^t (z_s - x_s) ds - \gamma \int_0^t \nabla f(x_{s-}) dN_s \\ &= x_0 + \underbrace{\int_0^t (\eta(z_s - x_s) - \gamma \nabla f(x_s)) ds}_{\text{adapted} + \text{continuous} + \text{bounded variation}} - \underbrace{\gamma \int_0^t \nabla f(x_{s-}) (dN_s - ds)}_{\text{martingale}}, \end{aligned}$$

which is a semimartingale by definition. The same argument holds for $(z_t)_{t \geq 0}$.

D.3 Proof of Lemma 10

From Lemma 8, we have

$$\begin{aligned} \|\nabla f(\bar{x}_t)\| &\leq \left\| \int_0^t w_t(s) \nabla f(x_{s-}) dN_s \right\| \\ &\quad + L_2 \int_0^t \eta^2 \|z_s - x_s\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds \\ &\quad + L_2 \int_0^t \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau \right) dN_s \\ &= \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| + L_2 \int_0^t \eta^2 \|z_s - x_s\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds \\ &\quad + L_2 \int_0^t \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau \right) dN_s, \end{aligned}$$

where the equality is because of Lemma 9. Multiplying both sides by $\Lambda_t := \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2$ and integrating on $[0, T_n]$ with respect to dN_t , we get

$$\begin{aligned} \int_0^{T_n} \Lambda_t \|\nabla f(\bar{x}_t)\| dN_t &\leq \int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| dN_t \\ &\quad + L_2 \int_0^{T_n} \Lambda_t \int_0^t \eta^2 \|z_s - x_s\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds dN_t \\ &\quad + L_2 \int_0^{T_n} \Lambda_t \int_0^t \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau \right) dN_s dN_t \\ &\stackrel{(i)}{=} \int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| dN_t \\ &\quad + \alpha^2 L_2 \int_0^{T_n} \int_0^t \eta^2 \|z_s - x_s\|^2 \left(\int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau \right) ds dN_t \\ &\quad + \alpha^2 L_2 \int_0^{T_n} \int_0^t \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau \right) dN_s dN_t \\ &\stackrel{(ii)}{=} \int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| dN_t \\ &\quad + \alpha^2 L_2 \int_0^{T_n} \eta^2 \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds \\ &\quad + \alpha^2 L_2 \int_0^{T_n} \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau dN_t \right) dN_s. \end{aligned}$$

Equality (i) holds as $\Lambda_t = \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 = \left(\int_0^t \alpha e^{\alpha s} dN_s \right)^2 e^{-2\alpha t}$, such that

$$\Lambda_t w_t(\sigma) w_t(\tau) = \left(\int_0^t \alpha e^{\alpha s} dN_s \right)^2 e^{-2\alpha t} \frac{\alpha e^{\alpha \sigma}}{\int_0^t \alpha e^{\alpha s} dN_s} \frac{\alpha e^{\alpha \tau}}{\int_0^t \alpha e^{\alpha s} dN_s} = \alpha^2 e^{-2\alpha t} e^{\alpha \sigma} e^{\alpha \tau}.$$

Equality (ii) holds by Tonelli, enabling to swap the integration order of s and t . We conclude by taking the expectation.

D.4 Proof of Lemma 11

Recalling $w_t(t) = \frac{\alpha e^{\alpha t}}{\int_0^t \alpha e^{\alpha s} dN_s}$, one has

$$\begin{aligned} \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| &= \frac{\alpha e^{\alpha t}}{(\gamma' - \gamma) \int_0^t \alpha e^{\alpha s} dN_s} \|z_t - x_t\| \\ &= \frac{\alpha}{(\gamma' - \gamma) \int_0^t \alpha e^{\alpha(s-t)} dN_s} \|z_t - x_t\| \end{aligned}$$

So, using the notation $\Lambda_t := \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2$, we have

$$\begin{aligned} \mathbb{E} \left[\int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\|^2 dN_t \right] &= \mathbb{E} \left[\int_0^{T_n} \frac{\alpha}{\gamma' - \gamma} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right) \|z_t - x_t\| dN_t \right] \\ &\leq \frac{\alpha}{\gamma' - \gamma} \sqrt{\mathbb{E} \left[\int_0^{T_n} \Lambda_t dN_t \right]} \sqrt{\mathbb{E} \left[\int_0^{T_n} \|z_t - x_t\|^2 dN_t \right]}, \end{aligned} \quad (45)$$

where we used the Cauchy-Schwarz inequality. We have

$$\begin{aligned} \int_0^{T_n} \|z_t - x_t\|^2 dN_t &= \int_0^{T_n} \|z_{t-} - x_{t-} - (\gamma' - \gamma) \nabla f(x_{t-})\|^2 dN_t \\ &\leq 2 \int_0^{T_n} \|z_{t-} - x_{t-}\|^2 dN_t + 2(\gamma' - \gamma)^2 \int_0^{T_n} \|\nabla f(x_{t-})\|^2 dN_t \end{aligned}$$

Then, from (7), we have

$$\mathbb{E} \left[\int_0^{T_n} \|\nabla f(x_{t-})\|^2 dN_t \right] \leq \frac{4}{\gamma} \Delta_f, \quad (46)$$

and from (8), we have

$$\mathbb{E} \left[\int_0^{T_n} \|z_{t-} - x_{t-}\|^2 dN_t \right] \leq \frac{1}{\eta + \eta'} \Delta_f = \frac{1}{\alpha} \Delta_f. \quad (47)$$

Combining (46) and (47), we obtain

$$\begin{aligned} \sqrt{\mathbb{E} \left[\int_0^{T_n} \|z_t - x_t\|^2 dN_t \right]} &\leq \sqrt{2\Delta_f} \sqrt{\frac{1}{\alpha} + \frac{4(\gamma' - \gamma)^2}{\gamma}} \\ &\leq \sqrt{2\alpha^{-1}\Delta_f} \sqrt{1 + \frac{4(\gamma' - \gamma)^2}{\gamma}}, \end{aligned} \quad (48)$$

the last inequality using $\alpha^{-1} \geq 1$. Note that we assume $\gamma' = \gamma + \sqrt{\gamma/2}$, such that $\frac{(\gamma' - \gamma)^2}{\gamma} = \frac{1}{2}$, and $\sqrt{1 + \frac{4(\gamma' - \gamma)^2}{\gamma}} = \sqrt{1 + 2} = \sqrt{3}$. Combining this with (45), Lemma 40-(i) and (48), we deduce

$$\begin{aligned}
\mathbb{E} \left[\int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_{t-} - x_{t-})}{\gamma' - \gamma} \right\| dN_t \right] &\leq \frac{\alpha}{\gamma' - \gamma} \sqrt{3 \left(1 + \frac{3}{2}\alpha\right) (1 + 2\alpha)n} \sqrt{2 \frac{\Delta_f}{\alpha}} \\
&= \sqrt{2\alpha n} \sqrt{\frac{6 \left(1 + \frac{3}{2}\alpha\right) (1 + 2\alpha)}{\gamma}} \sqrt{\Delta_f} \\
&= \sqrt{\alpha n} \sqrt{\frac{A_n}{\gamma}} \sqrt{\Delta_f},
\end{aligned} \tag{49}$$

where we used $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$ and defined $A_n = 12 \left(1 + \frac{3}{2}\alpha\right) (1 + 2\alpha)$.

D.5 Proof of Lemma 13

We prove each statement separately.

Equation (15) We have

$$\begin{aligned}
&\int_0^{T_n} \eta^2 \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds \\
&= \sum_{i=0}^{n-1} \int_{(T_i, T_{i+1})} \eta^2 \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds,
\end{aligned}$$

where we used that integration with respect to ds on a singleton is zero.

Then, noting $U_i := \int_{(T_i, T_{i+1})} \eta^2 \|z_s - x_s\|^2 ds$, we have

$$\begin{aligned}
&\sum_{i=0}^{n-1} \int_{(T_i, T_{i+1})} \eta^2 \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds \\
&\stackrel{(i)}{=} \sum_{i=0}^{n-1} \left(\int_{(T_i, T_{i+1})} \eta^2 \|z_s - x_s\|^2 ds \right) \left(\int_{T_{i+1}}^{T_n} \int_{T_{i+1}}^t \int_0^{T_i} e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) \\
&\stackrel{(ii)}{=} \sum_{i=1}^{n-1} \left(\int_{(T_i, T_{i+1})} \eta^2 \|z_s - x_s\|^2 ds \right) \left(\int_{T_{i+1}}^{T_n} \int_{T_{i+1}}^t \int_0^{T_i} e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) \\
&= \sum_{i=1}^{n-1} U_i \left(\int_{T_{i+1}}^{T_n} \int_{T_{i+1}}^t \int_0^{T_i} e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) \\
&= \sum_{i=1}^{n-1} \left(\int_{T_{i+1}}^{T_n} \int_{T_{i+1}}^t \int_0^{T_i} e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) U_i dN_\sigma dN_\tau dN_t \right),
\end{aligned} \tag{50}$$

where (i) uses the fact that $\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t$ is constant for $s \in (T_i, T_{i+1})$, as no new jump occurs on this interval. (ii) uses that $\int_0^{T_0} \cdot dN_t = \int_0^0 \cdot dN_t = 0$, as almost surely no jumps occur at $t = 0$. Because we integrate over $[0, T_n]$ with Poisson integrals, we can write (50) as a sum

$$(50) = \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \sum_{j=1}^i e^{\alpha(T_k - T_l)} e^{\alpha(T_j - T_i)} (T_k - T_j) U_i. \tag{51}$$

We multiply each side of (51) by $\mathbb{1}_{\mathbb{A}_n}$, and take the expectation

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{\mathbb{A}_n} \cdot (51)] &= \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \sum_{j=1}^i e^{\alpha(T_k - T_l)} e^{\alpha(T_j - T_l)} (T_k - T_j) U_i \right] \\
&= \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} e^{\alpha(T_k - T_l)} e^{\alpha(T_{i+1} - T_l)} \sum_{j=1}^i e^{\alpha(T_j - T_{i+1})} (T_k - T_j) U_i \right] \\
&\stackrel{(i)}{\leq} \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} e^{\alpha(T_k - T_l)} e^{\alpha(T_{i+1} - T_l)} \sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_k - T_j) U_i \right] \\
&= \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_{i+1} - T_l)} U_i \mathbb{1}_{\mathbb{A}_n} \sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_k - T_{i+1}) \right] \\
&\quad + \sum_{i=K}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_{i+1} - T_l)} U_i \mathbb{1}_{\mathbb{A}_n} \sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right] \\
&\quad + \sum_{i=1}^{K-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_{i+1} - T_l)} U_i \sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right].
\end{aligned}$$

(i) uses that $\sum_{j=1}^i a_j \leq \sum_{j=1}^{i+1} a_j$, as long as $a_j \geq 0$ for any $j \in \{1, \dots, i+1\}$. The last line uses $\mathbb{1}_{\mathbb{A}_n} \leq 1$. Now, we use Proposition 32 to deal with the third term, and Theorem 12 to deal with the first and second term, such that

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{\mathbb{A}_n} \cdot (51)] &\leq \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} (T_k - T_{i+1}) U_i \mathcal{C}_n \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} \right] \right] \\
&\quad + \sum_{i=K}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} U_i \mathcal{C}_n \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right] \right] \\
&\quad + \sum_{i=1}^{K-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_{i+1} - T_l)} U_i c_{upp} K^2 \right] \\
&\leq \mathcal{C}_n \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} (T_k - T_{i+1}) U_i \right] \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} \right] \\
&\quad + \mathcal{C}_n \sum_{i=K}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} U_i \right] \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right] \\
&\quad + \sum_{i=1}^{K-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_{i+1} - T_l)} U_i \right] c_{upp} K^2 \\
&\stackrel{(i)}{\leq} \mathcal{C}_n \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} (T_k - T_{i+1}) U_i \right] \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} \right] \\
&\quad + \mathcal{C}_n \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} U_i \right] \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right].
\end{aligned} \tag{52}$$

(i) uses that for $i \geq K$ and by Lemma 35, we have

$$\mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right] \geq \frac{e^{-1}}{32} \min \left\{ (i+1)^2, \frac{1}{\alpha^2} \right\} \geq \frac{e^{-1}}{32} \min \left\{ K^2, \frac{1}{\alpha^2} \right\} \geq \frac{c_{inf}^2 e^{-1}}{32} K^2,$$

the last inequality using the assumption $\alpha^{-1} \geq c_{inf}K$. So, we deduce

$$\mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right] \geq \frac{c_{inf}^2 e^{-1}}{32 c_{upp}} c_{upp} K^2 \geq C_n^{-1} c_{upp} K^2,$$

which implies

$$c_{upp} K^2 \leq C_n \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right].$$

Then, U_i depends on $\{T_k\}_{k \in \{1, \dots, i\}}$, as it depends on x_s and z_s for $s \in (T_i, T_{i+1})$, which depend on $\{T_k\}_{k \in \{1, \dots, i\}}$. Also, U_i depends on the length of the interval (T_i, T_{i+1}) , such that U_i also depends on T_{i+1} . So, we have that U_i only depends on $\{T_k\}_{k \in \{1, \dots, i+1\}}$. Because the increments of $\{T_k\}_{k \in \mathbb{N}}$ are independent, because $k, l \geq i+1$, we have $T_{i+1} - T_l \perp U_i$ and $T_k - T_l \perp U_i$. So, (52) becomes

$$\begin{aligned} \mathbb{E} [\mathbb{1}_{\mathbb{A}_n} \cdot (51)] &\leq C_n \sum_{i=1}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} (T_k - T_{i+1}) \right] \mathbb{E} [U_i] \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} \right] \\ &\quad + C_n \sum_{i=K}^{n-1} \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} \right] \mathbb{E} [U_i] \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right] \\ &= C_n \sum_{i=1}^{n-1} \mathbb{E} [U_i] \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} (T_k - T_{i+1}) \right] \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} \right] \\ &\quad + C_n \sum_{i=K}^{n-1} \mathbb{E} [U_i] \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} \right] \mathbb{E} \left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right] \\ &\stackrel{(i)}{=} C_n \sum_{i=1}^{n-1} \mathbb{E} [U_i] \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} (T_k - T_{i+1}) \sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} \right] \\ &\quad + C_n \sum_{i=1}^{n-1} \mathbb{E} [U_i] \sum_{l=i+1}^n \sum_{k=i+1}^l \mathbb{E} \left[e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} \sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_{i+1} - T_j) \right] \\ &= C_n \sum_{i=1}^{n-1} \mathbb{E} [U_i] \mathbb{E} \left[\sum_{l=i+1}^n \sum_{k=i+1}^l e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} \sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_k - T_j) \right] \end{aligned} \tag{53}$$

where in (i) we used the independence of the increments of the $\{T_k\}_{k \in \mathbb{N}}$. Finally, writing the sums as Poisson integrals, noting

$$\begin{aligned} A_{i+1} &:= \sum_{l=i+1}^n \sum_{k=i+1}^l e^{\alpha(T_{i+1} - T_l)} e^{\alpha(T_k - T_l)} \sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})} (T_k - T_j) \\ &= \int_{[T_{i+1}, T_n]} \int_{[T_{i+1}, t]} e^{\alpha(T_{i+1} - t)} e^{\alpha(\tau - t)} \int_{[0, T_{i+1}]} e^{\alpha(\sigma - T_{i+1})} (\tau - \sigma) dN_\sigma dN_\tau dN_t, \end{aligned}$$

and recalling $U_i := \int_{(T_i, T_{i+1})} \eta^2 \|z_s - x_s\|^2 ds$, (53) becomes

$$\begin{aligned} \mathbb{E} [\mathbb{1}_{\mathbb{A}_n} \cdot (51)] &\leq C_n \sum_{i=1}^{n-1} \mathbb{E} \left[\int_{(T_i, T_{i+1})} \eta^2 \|z_s - x_s\|^2 ds \right] \mathbb{E} [A_{i+1}] \\ &\leq C_n \max_{i \in \{1, \dots, n-1\}} \mathbb{E} [A_{i+1}] \sum_{i=1}^{n-1} \mathbb{E} \left[\int_{(T_i, T_{i+1})} \eta^2 \|z_s - x_s\|^2 ds \right] \\ &\leq C_n \max_{i \in \{1, \dots, n\}} \mathbb{E} [A_i] \mathbb{E} \left[\int_0^{T_n} \eta^2 \|z_s - x_s\|^2 ds \right]. \end{aligned}$$

Equation (16)

$$\begin{aligned} & \int_0^{T_n} \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau dN_t \right) dN_s \\ &= \int_0^{T_n} \left(\int_s^{T_n} \int_s^t e^{\alpha(\tau-t)} e^{\alpha(s-t)} \gamma^2 \|\nabla f(x_{s-})\|^2 \int_0^s e^{\alpha(\sigma-s)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau dN_t \right) dN_s \end{aligned} \quad (54)$$

Because we integrate over $[0, T_n]$, (54) can be written as a sum

$$(54) = \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 \sum_{j=1}^i e^{\alpha(T_j-T_i)} (k-j). \quad (55)$$

We used that $N_{T_k^-} - N_{T_j^-} = (k-1) - (j-1)$. Multiplying by $\mathbf{1}_{\mathbb{A}_n}$ and taking expectation, we have

$$\begin{aligned} \mathbb{E} [\mathbf{1}_{\mathbb{A}_n} \cdot (54)] &= \mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 \sum_{j=1}^i e^{\alpha(T_j-T_i)} (k-j) \right] \\ &= \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 \sum_{j=1}^i e^{\alpha(T_j-T_i)} (k-j) \right] \\ &= \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 (k-i) \mathbf{1}_{\mathbb{A}_n} \sum_{j=1}^i e^{\alpha(T_j-T_i)} \right] \\ &\quad + \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 \mathbf{1}_{\mathbb{A}_n} \sum_{j=1}^i e^{\alpha(T_j-T_i)} (i-j) \right] \end{aligned}$$

Now, Theorem 12 ensures that there exists a constant $\mathcal{C}_n$, such that we have

$$\begin{aligned} \mathbb{E} [\mathbf{1}_{\mathbb{A}_n} \cdot (54)] &\leq \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 \mathcal{C}_n (k-i) \mathbb{E} \left[\sum_{j=1}^i e^{\alpha(T_j-T_i)} \right] \right] \\ &\quad + \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 \mathcal{C}_n \mathbb{E} \left[\sum_{j=1}^i e^{\alpha(T_j-T_i)} (i-j) \right] \right] \\ &\leq \mathcal{C}_n \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 \right] (k-i) \mathbb{E} \left[\sum_{j=1}^i e^{\alpha(T_j-T_i)} \right] \\ &\quad + \mathcal{C}_n \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k-T_l)} e^{\alpha(T_i-T_l)} \gamma^2 \left\| \nabla f(x_{T_i^-}) \right\|^2 \right] \mathbb{E} \left[\sum_{j=1}^i e^{\alpha(T_j-T_i)} (i-j) \right]. \end{aligned} \quad (56)$$

Now, because the increments of $\{T_k\}_{k \in \mathbb{N}}$ are independent, as $l, k \geq i$, we have $T_k - T_l \perp$

$\|\nabla f(x_{T_i^-})\|$ and $T_i - T_l \perp \|\nabla f(x_{T_i^-})\|$. So, (56) becomes

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{\mathbb{A}_n} \cdot (54)] &\leq C_n \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \right] \gamma^2 \mathbb{E} \left[\|\nabla f(x_{T_i^-})\|^2 \right] (k - i) \mathbb{E} \left[\sum_{j=1}^i e^{\alpha(T_j - T_i)} \right] \\
&\quad + C_n \sum_{i=1}^n \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \gamma^2 \right] \mathbb{E} \left[\|\nabla f(x_{T_i^-})\|^2 \right] \mathbb{E} \left[\sum_{j=1}^i e^{\alpha(T_j - T_i)} (i - j) \right] \\
&= C_n \sum_{i=1}^n \gamma^2 \mathbb{E} \left[\|\nabla f(x_{T_i^-})\|^2 \right] \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \right] (k - i) \mathbb{E} \left[\sum_{j=1}^i e^{\alpha(T_j - T_i)} \right] \\
&\quad + C_n \sum_{i=1}^n \gamma^2 \mathbb{E} \left[\|\nabla f(x_{T_i^-})\|^2 \right] \sum_{l=i}^n \sum_{k=i}^l \mathbb{E} \left[e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \right] \mathbb{E} \left[\sum_{j=1}^i e^{\alpha(T_j - T_i)} (i - j) \right] \\
&\stackrel{(i)}{=} C_n \sum_{i=1}^n \gamma^2 \mathbb{E} \left[\|\nabla f(x_{T_i^-})\|^2 \right] \mathbb{E} \left[\sum_{l=i}^n \sum_{k=i}^l e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \sum_{j=1}^i e^{\alpha(T_j - T_i)} (k - j) \right], \tag{57}
\end{aligned}$$

where (i) used the independence property of the increments of $\{T_k\}_{k \in \mathbb{N}}$. Rewriting (57) as Poisson integrals, we note

$$\begin{aligned}
B_i &:= \sum_{l=i}^n \sum_{k=i}^l e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \sum_{j=1}^i e^{\alpha(T_j - T_i)} (k - j) \\
&= \int_{T_i}^{T_n} \int_{T_i}^t e^{\alpha(\tau - t)} e^{\alpha(T_i - t)} \int_0^{T_i} (N_{\tau^-} - N_{\sigma^-}) e^{\alpha(\sigma - T_i)} dN_{\sigma} dN_{\tau} dN_t,
\end{aligned}$$

such that

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{\mathbb{A}_n} \cdot (54)] &\leq C_n \sum_{i=1}^n \gamma^2 \mathbb{E} \left[\|\nabla f(x_{T_i^-})\|^2 \right] \mathbb{E}[B_i] \\
&\leq C_n \max_{i \in \{1, \dots, n\}} \mathbb{E}[B_i] \sum_{i=1}^n \gamma^2 \mathbb{E} \left[\|\nabla f(x_{T_i^-})\|^2 \right] \\
&= C_n \max_{i \in \{1, \dots, n\}} \mathbb{E}[B_i] \mathbb{E} \left[\int_0^{T_n} \gamma^2 \|\nabla f(x_{s^-})\|^2 dN_s \right]
\end{aligned}$$

D.6 Proof of Lemma 14

We recall

$$\begin{aligned}
A_k &:= \int_{[T_k, T_n]} \int_{[T_k, t]} \int_{[0, T_k]} e^{\alpha(T_k - t)} e^{\alpha(\tau - t)} e^{\alpha(\sigma - T_k)} (\tau - \sigma) dN_{\sigma} dN_{\tau} dN_t, \\
B_k &:= \int_{T_k}^{T_n} \int_{T_k}^t \int_0^{T_k} e^{\alpha(\tau - t)} e^{\alpha(T_k - t)} e^{\alpha(\sigma - T_k)} (N_{\tau^-} - N_{\sigma^-}) dN_{\sigma} dN_{\tau} dN_t.
\end{aligned}$$

We will split the proof in several sub lemmas. We first bound $\max_{k \in \{1, \dots, n\}} \mathbb{E}[B_k]$.

Lemma 25. *We have*

$$\max_{k \in \{1, \dots, n\}} \mathbb{E}[B_k] \leq \frac{(1 + \alpha)^2 (1 + 2\alpha)}{\alpha^4}$$

Proof. Let $k \leq n$.

$$\begin{aligned}
B_k &= \int_{T_k}^{T_n} \int_{T_k}^t \int_0^{T_k} e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_{\sigma} dN_{\tau} dN_t \\
&= \sum_{i=k}^n \sum_{j=1}^k \sum_{l=k}^i e^{\alpha(T_l-T_i)} e^{\alpha(T_j-T_l)} (l-j) \\
&= \sum_{i=k}^n \sum_{j=1}^k \sum_{l=k}^i e^{2\alpha(T_l-T_i)} e^{\alpha(T_j-T_l)} (l-j).
\end{aligned}$$

We take expectation, use $T_l - T_i \perp T_j - T_l$ and Lemma 39

$$\mathbb{E}[B_k] = \mathbb{E} \left[\sum_{i=k}^n \sum_{j=1}^k \sum_{l=k}^i e^{2\alpha(T_l-T_i)} e^{\alpha(T_j-T_l)} (l-j) \right] \quad (58)$$

$$= \sum_{i=k}^n \sum_{j=1}^k \sum_{l=k}^i \mathbb{E} \left[e^{2\alpha(T_l-T_i)} \right] \mathbb{E} \left[e^{\alpha(T_j-T_l)} \right] (l-j) \quad (59)$$

$$\begin{aligned}
&= \sum_{i=k}^n \sum_{j=1}^k \sum_{l=k}^i (1+2\alpha)^{l-i} (1+\alpha)^{j-l} (l-j) \\
&= \sum_{\substack{i,j,l \text{ s.t.} \\ 1 \leq j \leq k \leq l \leq i \leq n}} (1+2\alpha)^{l-i} (1+\alpha)^{j-l} (l-j) \\
&= \sum_{\substack{j,l \text{ s.t.} \\ 1 \leq j \leq k \leq l \leq n}} (1+\alpha)^{j-l} (l-j) \sum_{i=l}^n (1+2\alpha)^{l-i}
\end{aligned}$$

We have $\sum_{i=l}^n (1+2\alpha)^{l-i} \leq \sum_{i=l}^{+\infty} (1+2\alpha)^{l-i} = 1 + \frac{1}{2\alpha}$. We plug this inequality in the above, such that

$$\mathbb{E}[B_k] \leq \left(1 + \frac{1}{2\alpha}\right) \sum_{\substack{j,l \text{ s.t.} \\ 1 \leq j \leq k \leq l \leq n}} (1+\alpha)^{j-l} (l-j) \quad (60)$$

Now, we do the change of variable $(j, l) \rightarrow (j, l-j) := (j, d)$, such that

$$\mathbb{E}[B_k] \leq \left(1 + \frac{1}{2\alpha}\right) \sum_{d=0}^{n-1} \sum_{j=\max\{1, k-d\}}^{\min\{k, n-d\}} (1+\alpha)^{-d} d. \quad (61)$$

Indeed, for a fixed k , we both have $1 \leq j \leq k$, and $k \leq l \leq n \Rightarrow k-d \leq l-d \stackrel{(i)}{=} j \leq n-d$, where (i) is due to the change of variable. Then, as $\sum_{j=\max\{1, k-d\}}^{\min\{k, n-d\}} (1+\alpha)^{-d} d \leq \sum_{j=k-d}^k (1+\alpha)^{-d} d$, we deduce

$$\begin{aligned}
\mathbb{E}[B_k] &\leq \left(1 + \frac{1}{2\alpha}\right) \sum_{d=0}^{n-1} \sum_{j=k-d}^k (1+\alpha)^{-d} d \\
&\leq \left(1 + \frac{1}{2\alpha}\right) \sum_{d=0}^{n-1} (1+\alpha)^{-d} d(d+1) \\
&\leq \left(1 + \frac{1}{2\alpha}\right) \sum_{d=0}^{+\infty} (1+\alpha)^{-d} d(d+1) \\
&= \left(1 + \frac{1}{2\alpha}\right) \frac{2}{\alpha^3} (1+\alpha)^2 \\
&= \frac{(1+\alpha)^2 (1+2\alpha)}{\alpha^4}.
\end{aligned} \quad (62)$$

□

We now turn to $\max_{k \in \{1, \dots, n\}} \mathbb{E}[A_k]$. The computations are similar, thanks to the following elementary result.

Lemma 26. *If $T_n \sim \Gamma(n, 1)$, then*

$$\mathbb{E}[T_n e^{-\alpha T_n}] = \frac{n}{(1 + \alpha)^{n+1}}$$

Proof. Let $\varphi(\alpha) = \mathbb{E}[e^{-\alpha T_n}]$ the moment-generating function of T_n . As $T_n \sim \Gamma(n, 1)$, we have $\phi(\alpha) = (1 + \alpha)^{-n} \Rightarrow \phi'(\alpha) = -n(1 + \alpha)^{-n-1}$. So

$$\phi(\alpha) = -\mathbb{E}[T_n e^{-\alpha T_n}] \Rightarrow \mathbb{E}[T_n e^{-\alpha T_n}] = \frac{n}{(1 + \alpha)^{n+1}}.$$

□

Lemma 27. *We have*

$$\max_{k \in \{1, \dots, n\}} \mathbb{E}[A_k] \leq \frac{(1 + \alpha)(1 + 2\alpha)}{\alpha^4}$$

Proof. The proof is almost identical to the one of Lemma 25. The only difference lies in (59), where $l - j$ is replaced by $T_l - T_j$. Then, a direct application of Lemma 26 gives $\mathbb{E}[T_n e^{-\alpha T_n}] = (1 + \alpha)^{-1} \mathbb{E}[n e^{-\alpha T_n}]$ for some $T_n \sim \Gamma(n, 1)$. □

D.7 Proof of Lemma 15

1. Bounding Term 2. Using Lemma 13 and Lemma 14, we have

$$\begin{aligned} & \mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} \alpha^2 L_2 \int_0^{T_n} \eta^2 \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds \right] \\ & \leq C_n L_2 \frac{(1 + \alpha)(1 + 2\alpha)}{\alpha^2} \mathbb{E} \left[\int_0^{T_n} \eta^2 \|z_s - x_s\|^2 ds \right]. \end{aligned} \quad (63)$$

From (6), we have

$$\mathbb{E} \left[\int_0^{T_n} \|z_s - x_s\|^2 ds \right] \leq \frac{1}{\eta + \eta'} \Delta_f = \frac{1}{\alpha} \Delta_f. \quad (64)$$

So, combining (63) and (64), we deduce

$$\begin{aligned} & \mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} \alpha^2 L_2 \int_0^{T_n} \eta^2 \|z_s - x_s\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds \right] \\ & \leq C_n L_2 \eta^2 \frac{(1 + \alpha)(1 + 2\alpha)}{\alpha^3} \Delta_f \\ & = C_n L_2 \gamma \frac{(1 + \alpha)(1 + 2\alpha)}{2\alpha^3} \Delta_f, \end{aligned} \quad (65)$$

where the last equality uses $\eta = \sqrt{\frac{\gamma}{2}}$.

2. Bounding Term 3. Using Lemma 13 and Lemma 14, we have

$$\begin{aligned} & \mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} \alpha^2 L_2 \int_0^{T_n} \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau dN_t \right) dN_s \right] \\ & \leq C_n L_2 \frac{(1 + \alpha)^2 (1 + 2\alpha)}{\alpha^2} \mathbb{E} \left[\int_0^{T_n} \gamma^2 \|\nabla f(x_{s-})\|^2 dN_s \right] \end{aligned} \quad (66)$$

From (7), we have

$$\mathbb{E} \left[\int_0^{T_n} \|\nabla f(x_{s-})\|^2 dN_s \right] \leq \frac{4}{\gamma} \Delta_f. \quad (67)$$

So, combining (66) and (67), we deduce

$$\begin{aligned} & \mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} \alpha^2 L_2 \int_0^{T_n} \gamma^2 \|\nabla f(x_{s-})\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau-} - N_{\sigma-}) dN_\sigma dN_\tau dN_t \right) dN_s \right] \\ & \leq 4C_n L_2 \gamma \frac{(1+\alpha)^2(1+2\alpha)}{\alpha^2} \Delta_f. \end{aligned} \quad (68)$$

Combining (65) and (68), we obtain the desired result.

D.8 Proofs of Section 5.5

Proof of Corollary 16 From (18)

$$\begin{aligned} & \mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \\ & = n^{3/7} \left(\sqrt{A_n L \Delta_f} + B_n C_n L^{-1} L_2 \Delta_f \right) \\ & \leq 4\sqrt{3} B_n C_n \left(\sqrt{L \Delta_f} + L^{-1} L_2 \Delta_f \right) n^{3/7} \end{aligned} \quad (69)$$

where we use $C_n \geq 1$ and $\sqrt{A_n} = 2\sqrt{3} \sqrt{(1 + \frac{3}{2}\alpha)(1+2\alpha)} \leq 2\sqrt{3}(1+2\alpha) \leq 2\sqrt{3}(1+2\alpha)(1+\alpha)(1+\alpha 8(1+\alpha)) = 4\sqrt{3} B_n$. Using $\alpha = n^{-1/7}$, we have

$$B_n C_n = 2\sqrt{3}(1+2n^{-1/7})(1+n^{-1/7}) \left(1 + n^{-1/7} 8(1+n^{-1/7}) \right) 32(4+4C_n+C_n^2) \frac{c_{upp}^e}{c_{inf}^2},$$

with $C_n = c_{inf} n^{-1/7} K$.

Proof of Corollary 17 From (18)

$$\begin{aligned} & \mathbb{E} \left[\mathbf{1}_{\mathbb{A}_n} \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \\ & \leq n^{3/7} \left(\sqrt{A_n C_\alpha L^{1+a_1} L_2^{a_2} \Delta_f^{1+a_3}} + B_n C_\alpha^{-3} C_n L^{-1-3a_1} L_2^{1-3a_2} \Delta_f^{1-3a_3} \right), \end{aligned} \quad (70)$$

with, $A_n := 12(1 + \frac{3}{2}\alpha)(1+2\alpha)$ and $B_n := \frac{1}{2}(1+\alpha)(1+2\alpha)(1+\alpha 8(1+\alpha))$. We want to find a_1, a_2, a_3 such that the exponents in the left and right term are the same. Therefore, we solve

$$\begin{cases} \frac{1+a_1}{2} = -1-3a_1, \\ \frac{a_2}{2} = 1-3a_2, \\ \frac{1+a_3}{2} = 1-3a_3, \end{cases} \implies (a_1, a_2, a_3) = \left(-\frac{3}{7}, \frac{2}{7}, \frac{1}{7} \right).$$

Then, (70) becomes

$$\begin{aligned}
& \mathbb{E} \left[\mathbb{1}_{\mathbb{A}_n} \min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \\
& \leq n^{3/7} \left(\sqrt{A_n \left(1 + \frac{3}{2}\right) C_\alpha + B_n C_\alpha^{-3} \mathcal{C}_n} \right) L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} \\
& \stackrel{(i)}{=} \left(\sqrt{A_n C_\alpha + B_n C_\alpha^{-3} \left(32 \frac{c_{upp} e}{c_{inf}^2} (4 + 4C_n + C_n^2) \right)} \right) L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{3/7} \tag{71} \\
& \stackrel{(ii)}{\leq} (1 + C_n + C_n^2/4) \left(\sqrt{A_n C_\alpha + 128 B_n \frac{c_{upp} e}{c_{inf}^2} C_\alpha^{-3}} \right) L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{3/7} \\
& \stackrel{(iii)}{\leq} (1 + C_n + C_n^2/4) H(\alpha) \left(\sqrt{12 C_\alpha + 64 \frac{c_{upp} e}{c_{inf}^2} C_\alpha^{-3}} \right) L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{3/7}.
\end{aligned}$$

In (i) we use the definition of $\mathcal{C}_n$, namely $\mathcal{C}_n = 32(4 + 4C_n + C_n^2) \frac{c_{upp} e}{c_{inf}^2}$. In (ii) we use $(4 + 4C_n + C_n^2) = 4(1 + C_n + C_n^2/4)$ and $(1 + C_n + C_n^2/4) \geq 1$. In (iii), replace A_n and B_n by their values and we set $H(\alpha) := \max\{\sqrt{(1 + \frac{3}{2}\alpha)(1 + 2\alpha)}, (1 + \alpha)(1 + 2\alpha)(1 + \alpha 8(1 + \alpha))\} = (1 + \alpha)(1 + 2\alpha)(1 + \alpha 8(1 + \alpha))$. We then optimize on C_α . Note that we keep implicit the dependence on C_α in $H(\alpha)$, and do not consider it in the optimization of C_α . Denoting $a = 12$, $b = 64 \frac{c_{upp} e}{c_{inf}^2}$, we want to minimize

$$\sqrt{a C_\alpha} + b C_\alpha^{-3}.$$

The function $\phi : t \mapsto \sqrt{at} + bt^{-3}$ is derivable on $\mathbb{R}_+^*$, and $\phi'(t) = \frac{\sqrt{a}}{2\sqrt{t}} - \frac{3b}{t^4}$. It follows that ϕ reaches its minimum on $\mathbb{R}_+^*$ at the point

$$\phi'(t) = 0 \Leftrightarrow t = \left(\frac{6b}{\sqrt{a}} \right)^{2/7}.$$

Then, the minimum of ϕ , denoted $\phi_{\min}$, is

$$\begin{aligned}
\phi_{\min} &= \phi \left(\left(\frac{6b}{\sqrt{a}} \right)^{2/7} \right) \\
&= \sqrt{a \left(\frac{6b}{\sqrt{a}} \right)^{2/7}} + \frac{b}{\left(\frac{6b}{\sqrt{a}} \right)^{6/7}} \\
&= a^{1/2} \cdot 6^{1/7} b^{1/7} a^{-1/14} + b \cdot a^{3/7} 6^{-6/7} b^{-6/7} \\
&= 6^{1/7} a^{3/7} b^{1/7} + 6^{-6/7} a^{3/7} b^{1/7} \\
&= \left(6^{1/7} + 6^{-6/7} \right) a^{3/7} b^{1/7} \\
&= 7 \cdot 6^{-6/7} a^{3/7} b^{1/7}.
\end{aligned}$$

Then, choosing $C_\alpha := \left(\frac{6b}{\sqrt{a}} \right)^{2/7}$, and replacing a and b by their values, we obtain

$$C_\alpha = \left(\frac{6b}{\sqrt{a}} \right)^{2/7} = \left(\frac{6 \cdot 64 \frac{c_{upp} e}{c_{inf}^2}}{\sqrt{12}} \right)^{2/7} = \left(\frac{\sqrt{3} \cdot 64 c_{upp} e}{c_{inf}^2} \right)^{2/7},$$

which in turn implies

$$\sqrt{A_n C_\alpha} + 64 \frac{c_{upp} e}{c_{inf}^2} C_\alpha^{-3} = 7 \cdot 6^{-6/7} 12^{3/7} 64^{1/7} \left(\frac{c_{upp} e}{c_{inf}^2} \right)^{1/7}.$$

Finally, we deduce that with the choice

$$\alpha = \left(\frac{\sqrt{3} \cdot 64 c_{upp} e}{c_{inf}^2} \right)^{2/7} \left(\frac{L_2^2 \Delta_f}{L^3 n} \right)^{1/7},$$

we obtain

$$\begin{aligned} & \mathbb{E} \left[\min_{t \in \{T_1, \dots, T_n\}} \|\nabla f(\bar{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \\ & \leq (1 + C_n + C_n^2/4) H(\alpha) 7 \cdot 6^{-6/7} 12^{3/7} 64^{1/7} \left(\frac{c_{upp} e}{c_{inf}^2} \right)^{1/7} \left(\frac{c_{upp} e}{c_{inf}^2} \right)^{1/7} L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{3/7} \quad (72) \\ & \approx 9.2(1 + C_n + C_n^2/4) H(\alpha) \left(\frac{c_{upp}}{c_{inf}^2} \right)^{1/7} L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{3/7}. \end{aligned}$$

D.9 Proof of Lemma 18

We note

$$\Delta_n := \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t = \sum_{i=1}^n \left(\sum_{j=1}^i \alpha e^{-\alpha(T_i - T_j)} \right)^2.$$

By Proposition 29, on the set $\mathbb{A}_n$, we have that for all $j \leq i \leq n$,

$$T_i - T_j \leq c_{upp}(i - j) + c_{upp}K.$$

Hence,

$$\mathbb{1}_{\mathbb{A}_n} \Delta_n \geq \mathbb{1}_{\mathbb{A}_n} \alpha^2 \sum_{i=1}^n \left(\sum_{j=1}^i e^{-c_{upp}\alpha(i-j+K)} \right)^2 = \mathbb{1}_{\mathbb{A}_n} \alpha^2 e^{-2c_{upp}\alpha K} \sum_{i=1}^n \left(\sum_{j=1}^i e^{-c_{upp}\alpha(i-j)} \right)^2.$$

Set $r := e^{-c_{upp}\alpha}$. Then

$$\sum_{j=1}^i e^{-c_{upp}\alpha(i-j)} = \sum_{\ell=0}^{i-1} r^\ell = \frac{1 - r^i}{1 - r}.$$

Therefore,

$$\mathbb{1}_{\mathbb{A}_n} \Delta_n \geq \mathbb{1}_{\mathbb{A}_n} \frac{\alpha^2}{(1 - r)^2} e^{-2c_{upp}\alpha K} \sum_{i=1}^n (1 - r^i)^2.$$

Using $1 - e^{-x} \leq x$ for all $x \geq 0$, we get

$$\frac{\alpha^2}{(1 - r)^2} = \frac{\alpha^2}{(1 - e^{-c_{upp}\alpha})^2} \geq \frac{\alpha^2}{(c_{upp}\alpha)^2} = \frac{1}{c_{upp}^2}.$$

Moreover,

$$e^{-2c_{upp}\alpha K} = e^{-2 \frac{c_{upp}}{c_{inf}} C_n}.$$

Hence,

$$\mathbb{1}_{\mathbb{A}_n} \Delta_n \geq \mathbb{1}_{\mathbb{A}_n} \frac{1}{c_{upp}^2} e^{-2 \frac{c_{upp}}{c_{inf}} C_n} \sum_{i=1}^n (1 - r^i)^2.$$

Now,

$$\sum_{i=1}^n (1 - r^i)^2 = n - 2 \sum_{i=1}^n r^i + \sum_{i=1}^n r^{2i} \geq n - 2 \sum_{i=1}^{\infty} r^i = n - \frac{2r}{1 - r} = n - \frac{2}{e^{c_{upp}\alpha} - 1}.$$

Therefore,

$$\mathbb{1}_{\mathbb{A}_n} \Delta_n \geq \mathbb{1}_{\mathbb{A}_n} \frac{1}{c_{upp}^2} e^{-2 \frac{c_{upp}}{c_{inf}} C_n} \left(n - \frac{2}{e^{c_{upp}\alpha} - 1} \right).$$

Finally, since $e^x - 1 \geq x$ for all $x \geq 0$,

$$\frac{2}{e^{c_{upp}\alpha} - 1} \leq \frac{2}{c_{upp}\alpha},$$

such that

$$\mathbb{1}_{\mathbb{A}_n} \Delta_n \geq \mathbb{1}_{\mathbb{A}_n} \frac{1}{c_{upp}^2} e^{-2 \frac{c_{upp}}{c_{inf}} C_n} \left(n - \frac{2}{c_{upp}\alpha} \right) = \mathbb{1}_{\mathbb{A}_n} \frac{n}{c_{upp}^2} e^{-2 \frac{c_{upp}}{c_{inf}} C_n} \left(1 - \frac{2}{c_{upp}\alpha n} \right).$$

E Proof of Theorem 12

We prove Theorem 12 in this section, which we restate now.

Theorem 12. *Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0, 1)$, $\alpha \in (0, 1)$. Let $K = \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2 \frac{n}{\varepsilon}) \rceil$, and $c_{upp} \geq 1$ be chosen such that $c_{upp} - \log(c_{upp}) = c_{inf} - \log(c_{inf})$. Let*

$$C_n = 32(4 + 4C_n + C_n^2) \frac{c_{upp}^e}{c_{inf}^2}, \text{ where } C_n = c_{inf}\alpha K.$$

There exists a set $\mathbb{A}_n \subset \Omega$ of realizations satisfying $\mathbb{P}(\mathbb{A}_n) \geq 1 - \varepsilon$ such that the three following assertions hold

- (1) $\forall i \in \{K+1, \dots, n\}, \quad \sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \leq C_n \mathbb{E} \left[\sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \right],$
- (2) $\forall i \in \{1, \dots, n\}, \quad \sum_{j=1}^i e^{-\alpha(T_i - T_j)} \leq C_n \mathbb{E} \left[\sum_{j=1}^i e^{-\alpha(T_i - T_j)} \right],$
- (3) $\forall i \in \{1, \dots, n\}, \quad \sum_{j=1}^i (i - j) e^{-\alpha(T_i - T_j)} \leq C_n \mathbb{E} \left[\sum_{j=1}^i (i - j) e^{-\alpha(T_i - T_j)} \right].$

Sketch of Proof. Define

$$H_0^i := \sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)}, \quad H_1^i := \sum_{j=1}^i e^{-\alpha(T_i - T_j)}, \quad H_2^i := \sum_{j=1}^i (i - j) e^{-\alpha(T_i - T_j)}.$$

The proof follows the same five-step strategy for each of the three sums H_0^i, H_1^i, H_2^i .

Step 1 — Choice of block size K . Since $T_i - T_j = \sum_{k=j}^{i-1} \Delta_k$ with $\Delta_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{E}(1)$, the Chernov inequality controls deviations of a sum of *fixed* length K from its expectation, whose value is K . However, we need a bound valid for all gap lengths $i - j$ simultaneously. The key device is to fix a block size K (whose value will be determined in Step 2) and decompose each sum into consecutive blocks of size K plus a residual block.

Step 2 — Concentration of $T_i - T_j$ (Proposition 29). Applying Chernov to each block of size K , combined with a union bound over the $O(n)$ possible block starting points, gives with probability $1 - \varepsilon$, for all $1 \leq j \leq i \leq n$:

$$c_{inf}(i - j) - c_{inf}K \leq T_i - T_j \leq c_{upp}(i - j) + c_{upp}K.$$

The value $K = \lceil (c_{inf} - 1 - \log c_{inf})^{-1} \log(2n/\varepsilon) \rceil$ is chosen to make the failure probability of each block at most $\varepsilon/(2n)$, so that using a union bound implies a probability over all the blocks of $\varepsilon/2$, that does not grows with n .

Step 3 — High-probability upper-bounds on H_k^i (Propositions 30–31–32). Substituting the bounds of Step 2 into each sum yields upper-bounds with high probability on $\{H_k^i\}_{k \in \{0,1,2\}}$. These upper-bound involve terms that depend on K .

Step 4 — lower-bounds on expectations (Lemmas 35–36–34). Computing the characteristic function of the Gamma distribution, $\mathbb{E}[H_k^i]$ is computed exactly in closed form, then lower-bounded by quantity whose form matches those of the upper-bound in Step 3.

Step 5 — Conclusion (Section E.3). The upper-bounds from Step 3 involve terms that depend on K , while the lower-bounds from Step 4 do not. We define C_n such that $K = \mathcal{O}(C_n/\alpha)$, allowing to combine these bounds. We obtain that the sums H_k^i are upper-bounded by their expectations, up to constants involving C_n . As long as $\alpha = n^{-\beta}$ for some $\beta > 0$, the constant C_n will be uniformly upper-bounded with n , and will verify $C_n \rightarrow_{n \rightarrow +\infty} 0$. □

E.1 Upper-bounding the sums

A key result for our proof is based on a rather classical concentration inequality called the Chernov inequality. We use it in the case of gamma laws, see *e.g.* [12, Section 2.4].

Lemma 28 (Chernov inequality). *For $k \in \mathbb{N}^*$, let $T_k \sim \Gamma(k, 1)$, $0 < c_{inf} \leq 1 \leq c_{upp}$. Then,*

$$(i) \quad \mathbb{P}(T_k \leq c_{inf}k) \leq e^{-(c_{inf}-1-\log(c_{inf}))k}.$$

$$(ii) \quad \mathbb{P}(T_k \geq c_{upp}k) \leq e^{-(c_{upp}-1-\log(c_{upp}))k}.$$

Proposition 29. *Let $\varepsilon, c_{inf} \in (0, 1)$, $n > 0$, and $K := \lceil (c_{inf} - 1 - \log c_{inf})^{-1} \log(2n/\varepsilon) \rceil$. Let $c_{upp} \geq 1$ be chosen such that $c_{upp} - \log(c_{upp}) = c_{inf} - \log(c_{inf})$. Let the set $\mathbb{A}_n$ such that we have for any $1 \leq j \leq i \leq n$*

$$c_{inf}(i - j) - c_{inf}K \leq T_i - T_j \leq c_{upp}(i - j) + c_{upp}K.$$

Then, we have $\mathbb{P}(\mathbb{A}_n) \geq 1 - \varepsilon$.

Proof. We note $\Delta_k := T_{k+1} - T_k$ such that

$$T_i - T_j = \sum_{k=j}^{i-1} \Delta_k,$$

where $\Delta_k \stackrel{i.i.d}{\sim} \mathcal{E}(1)$. Let an integer $K > 0$, to be further fixed.

Lower-bound. The idea is to decompose $\sum_{k=j}^{i-1} \Delta_k$ into $\lfloor (i - j)/K \rfloor$ blocks of size K starting at $j, j + K, j + 2K, \dots$, plus a residual block of size $(i - j) \bmod K < K$. Chernov is applied to each complete block; the residual block is non-negative ($\Delta_k \geq 0$ a.s.) and is simply dropped, at the cost of the term $-c_{inf}K$ in the final bound.

We have $\sum_{k=j}^{j-1+K} \Delta_k = T_{j+K} - T_j$ which follows the same law as $T_K \sim \Gamma(K, 1)$. Therefore, for any $c_{inf} \in (0, 1)$, applying Lemma 28 (i),

$$\sum_{k=j}^{j-1+K} \Delta_k \leq c_{inf}K, \tag{73}$$

with probability at most $e^{-K(c_{inf}-1-\log c_{inf})}$. Choosing $K = \lceil (c_{inf} - 1 - \log c_{inf})^{-1} \log(2n/\varepsilon) \rceil$, this probability is at most $\varepsilon/(2n)$. By a union bound over $j \in \{1, \dots, n - K\}$, with probability $1 - \varepsilon/2$, (73) fails for all such j , namely

$$\sum_{k=j}^{j-1+K} \Delta_k > c_{inf}K. \tag{74}$$

Now, consider the following decomposition

$$\sum_{k=j}^{i-1} \Delta_k = \underbrace{\sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k}_{\lfloor (i-j)/K \rfloor \text{ complete blocks}} + \underbrace{\sum_{k=j+\lfloor \frac{i-j}{K} \rfloor K}^{i-1} \Delta_k}_{\text{residual block}}.$$

For $s \in \{0, \dots, \lfloor \frac{i-j}{K} \rfloor - 1\}$, we have $j + sK \in \{j, \dots, i - K\} \subset \{1, \dots, n - K\}$. So, (74) gives that with probability $1 - \varepsilon/2$, for any such s we have $\sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k > c_{inf}K$. The residual block can be dropped, as it is nonnegative. Then,

$$\begin{aligned}
\sum_{k=j}^{i-1} \Delta_k &= \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k + \sum_{k=j+\lfloor \frac{i-j}{K} \rfloor K}^{i-1} \Delta_k \\
&\geq \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor - 1} c_{inf}K \\
&= c_{inf}K \left\lfloor \frac{i-j}{K} \right\rfloor \\
&\geq c_{inf}K \left(\frac{i-j}{K} - 1 \right) \quad (\lfloor x \rfloor \geq x - 1) \\
&= c_{inf}(i-j) - c_{inf}K.
\end{aligned} \tag{75}$$

Upper-bound. As for the lower-bound, the strategy for the upper-bound exploits the following decomposition

$$\sum_{k=j}^{i-1} \Delta_k = \underbrace{\sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k}_{\lfloor (i-j)/K \rfloor \text{ complete blocks}} + \underbrace{\sum_{k=j+\lfloor \frac{i-j}{K} \rfloor K}^{i-1} \Delta_k}_{\text{residual block}}.$$

However, a difference is that the residual block cannot be dropped by a non-negativity argument, because here we want to obtain an upper-bound. A simple strategy is to bound a residual block by a complete block of size K , namely

$$\sum_{k=j+\lfloor \frac{i-j}{K} \rfloor K}^{i-1} \Delta_k \leq \sum_{k=j+\lfloor \frac{i-j}{K} \rfloor K}^{j+(\lfloor \frac{i-j}{K} \rfloor + 1)K-1} \Delta_k.$$

However, some terms in the sum of the right-hand side are not defined if $j + (\lfloor \frac{i-j}{K} \rfloor + 1)K - 1 \geq i$. So, for the purpose of the analysis, we introduce $\{\Delta_k\}_{n \leq k \leq n+K-1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{E}(1)$, independent of $\{\Delta_k\}_{k \leq n-1}$, such that we can write

$$\begin{aligned}
\sum_{k=j}^{i-1} \Delta_k &= \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k + \sum_{k=j+\lfloor \frac{i-j}{K} \rfloor K}^{i-1} \Delta_k \\
&\leq \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k + \sum_{k=j+\lfloor \frac{i-j}{K} \rfloor K}^{j+(\lfloor \frac{i-j}{K} \rfloor + 1)K-1} \Delta_k \\
&= \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k
\end{aligned} \tag{76}$$

Then, applying Lemma 28 (ii), we have that for any $c_{upp} > 1$ and for $j \in \{1, \dots, n-1\}$

$$\sum_{k=j}^{j-1+K} \Delta_k \geq c_{upp}K, \tag{77}$$

with probability at most

$$e^{-K(c_{upp}-1-\log(c_{upp}))}.$$

Choosing c_{upp} such that $c_{upp} - \log(c_{upp}) = c_{inf} - \log(c_{inf})$, we obtain

$$\left\lceil (c_{upp} - 1 - \log(c_{upp}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \right\rceil = \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \right\rceil =: K,$$

and it follows that this probability is at most $\varepsilon/(2n)$. By a union bound over $j \in \{1, \dots, n-1\}$, with probability $1 - \frac{\varepsilon}{2}$, we have

$$\sum_{k=j}^{j-1+K} \Delta_k \leq c_{upp}K \quad (78)$$

So, from (76), we deduce that with probability $1 - \varepsilon/2$, we have

$$\sum_{k=j}^{i-1} \Delta_k \leq \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k \leq \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor} c_{upp}K = c_{upp}K \left(\left\lfloor \frac{i-j}{K} \right\rfloor + 1 \right) \leq c_{upp}(i-j) + c_{upp}K. \quad (79)$$

Conclusion. The claim follows by combining (75) and (79), each holding under probability $1 - \frac{\varepsilon}{2}$. $\square$

Proposition 30. Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0, 1)$, $\alpha \in (0, 1]$ and $K = \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2\frac{n}{\varepsilon}) \rceil$. With probability $1 - \varepsilon$, we have

$$\forall i \in \{1, \dots, n\}, \sum_{j=1}^i e^{-\alpha(T_i - T_j)} \leq \min\left\{i, \frac{2}{\alpha c_{inf}} + K\right\}.$$

Proof. We bound

$$H_1^i := \sum_{j=1}^i e^{-\alpha(T_i - T_j)}.$$

We note that as long as $i > K$, the $i - K$ first terms of H_1^i can be controlled using the lower-bound on $T_i - T_j$ from Proposition 29. It remains the last terms, whose amount is at most K ; these terms can be bounded by 1.

Let $i \leq K$. We have

$$\sum_{j=1}^i e^{-\alpha(T_i - T_j)} \leq i \leq \min\left\{i, \frac{2}{\alpha c_{inf}} + K\right\}. \quad (80)$$

Let $i > K$. With probability $1 - \varepsilon$, we have

$$\begin{aligned} \sum_{j=1}^i e^{-\alpha(T_i - T_j)} &= \sum_{j=1}^{i-K} e^{-\alpha(T_i - T_j)} + \sum_{j=i-K+1}^i e^{-\alpha(T_i - T_j)} \\ &\leq \sum_{j=1}^{i-K} e^{-\alpha(T_i - T_j)} + K \\ &\stackrel{(i)}{\leq} \sum_{j=1}^{i-K} e^{-\alpha(c_{inf}(i-j) - c_{inf}K)} + K \\ &\stackrel{(ii)}{=} \sum_{\ell=0}^{i-K-1} e^{-\alpha c_{inf} \ell} + K \\ &\stackrel{(iii)}{\leq} \min\left\{i - K, \frac{1}{1 - e^{-\alpha c_{inf}}}\right\} + K \\ &\stackrel{(iv)}{\leq} \min\left\{i, \frac{2}{\alpha c_{inf}} + K\right\}. \end{aligned} \quad (81)$$

For step (i), we apply $T_i - T_j \geq c_{inf}(i - j) - c_{inf}K$. This is due to Proposition 29, valid with probability $1 - \varepsilon$, for any $j \leq i - K$ since $i - j \geq K$. For step (ii) we set the change of variable

$\ell = i - K - j$, and simplify $e^{\alpha c_{inf} K} e^{-\alpha c_{inf} K} = 1$. (iii) uses $\sum_{\ell=0}^{i-K-1} e^{-\alpha c_{inf} \ell} = \frac{1 - e^{-\alpha c_{inf} (i-K)}}{1 - e^{-\alpha c_{inf}}} \leq \frac{1}{1 - e^{-\alpha c_{inf}}}$. (iv) uses $e^{-t} \leq 1 - t/2$ for $t = \alpha c_{inf} \in (0, 1]$, giving $\frac{1}{1 - e^{-\alpha c_{inf}}} \leq \frac{2}{\alpha c_{inf}}$. $\square$

Proposition 31. Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0, 1)$, $\alpha \in (0, 1]$ and $K = \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log(2\frac{n}{\varepsilon}) \rceil$. With probability $1 - \varepsilon$, we have

$$\forall i \in \{1, \dots, n\}, \sum_{j=1}^i (i-j) e^{-\alpha(T_i - T_j)} \leq \min \left\{ \frac{i^2}{2}, \frac{4}{(\alpha c_{inf})^2} + \frac{2K}{\alpha c_{inf}} + \frac{K^2}{2} \right\}.$$

Proof. We bound

$$H_2^i := \sum_{j=1}^i (i-j) e^{-\alpha(T_i - T_j)}.$$

Compared with H_1^i , the weight $(i-j)$ adds a linear factor. When treating H_2^i , the main difference is that the sum $\sum_{\ell} \ell q^\ell$ is handled via $\sum_{\ell=0}^{\infty} \ell q^\ell = q/(1-q)^2$.

If $i \leq K$,

$$H_2^i \leq \sum_{j=1}^i (i-j) = \frac{i(i-1)}{2} \leq \frac{i^2}{2} \leq \min \left\{ \frac{i^2}{2}, \frac{4}{(\alpha c_{inf})^2} + \frac{2K}{\alpha c_{inf}} + \frac{K^2}{2} \right\}.$$

If $i > K$, with probability $1 - \varepsilon$, we have

$$\begin{aligned} H_2^i &= \sum_{j=1}^{i-K} (i-j) e^{-\alpha(T_i - T_j)} + \sum_{j=i-K+1}^i (i-j) e^{-\alpha(T_i - T_j)} \\ &\stackrel{(i)}{\leq} \sum_{j=1}^{i-K} (i-j) e^{-\alpha(c_{inf}(i-j) - c_{inf}K)} + \sum_{j=i-K+1}^i (i-j) \\ &\stackrel{(ii)}{\leq} \sum_{\ell=0}^{i-K-1} (\ell + K) q^\ell + \frac{K^2}{2}. \end{aligned} \tag{82}$$

Step (i) uses $T_i - T_j \geq c_{inf}(i-j) - c_{inf}K$, holding with probability $1 - \varepsilon$, thanks to Proposition 29. It also uses $e^{-\alpha(T_i - T_j)} \leq 1$. In step (ii) we set $q := e^{-\alpha c_{inf}}$ and the change of variable $\ell = i - j - K$.

Thus,

$$H_2^i \leq \sum_{\ell=0}^{i-K-1} \ell q^\ell + K \sum_{\ell=0}^{i-K-1} q^\ell + \frac{K^2}{2}. \tag{83}$$

Now,

$$\sum_{\ell=0}^{m-1} \ell q^\ell \leq \min \left\{ \sum_{\ell=0}^{m-1} \ell, \sum_{\ell=0}^{\infty} \ell q^\ell \right\} = \min \left\{ \frac{m(m-1)}{2}, \frac{q}{(1-q)^2} \right\},$$

and

$$\sum_{\ell=0}^{m-1} q^\ell \leq \min \left\{ m, \frac{1}{1-q} \right\}.$$

Applying this with $m = i - K$, we obtain

$$\begin{aligned} H_2^i &\leq \min \left\{ \frac{(i-K)(i-K-1)}{2}, \frac{q}{(1-q)^2} \right\} \\ &\quad + K \min \left\{ i-K, \frac{1}{1-q} \right\} + \frac{K^2}{2}. \end{aligned} \tag{84}$$

Using $\min(a, b) + \min(c, d) \leq \min(a+c, b+d)$, we get

$$\begin{aligned} H_2^i &\leq \min \left\{ \frac{(i-K)(i-K-1)}{2} + K(i-K) + \frac{K^2}{2}, \right. \\ &\quad \left. \frac{q}{(1-q)^2} + \frac{K}{1-q} + \frac{K^2}{2} \right\}. \end{aligned} \tag{85}$$

The first term is equal to $\frac{i^2-i+K}{2}$. Since $i > K$, it follows that $\frac{i^2-i+K}{2} \leq \frac{i^2}{2}$. Then,

$$H_2^i \leq \min \left\{ \frac{i^2}{2}, \frac{q}{(1-q)^2} + \frac{K}{1-q} + \frac{K^2}{2} \right\}. \quad (86)$$

As moreover $\alpha c_{inf} \leq 1$, then

$$1 - q = 1 - e^{-\alpha c_{inf}} \geq \frac{\alpha c_{inf}}{2},$$

hence

$$\frac{1}{1-q} \leq \frac{2}{\alpha c_{inf}}, \quad \frac{q}{(1-q)^2} \leq \frac{1}{(1-q)^2} \leq \frac{4}{(\alpha c_{inf})^2}.$$

Therefore,

$$H_2^i \leq \min \left\{ \frac{i^2}{2}, \frac{4}{(\alpha c_{inf})^2} + \frac{2K}{\alpha c_{inf}} + \frac{K^2}{2} \right\}. \quad (87)$$

□

The following proposition differs from Propositions 30 and 31 for two reasons. First, because the weights $(T_i - T_j)$ are random, we need both the upper bound $(T_i - T_j \leq c_{upp}(i-j) + c_{upp}K)$ and the lower-bound $(T_i - T_j \geq c_{inf}(i-j) - c_{inf}K)$ from Proposition 29. The upper-bound controls the weight, the lower bound gives exponential decay. Then, a difficulty is that for small i , there is no direct upper-bound of $(T_i - T_j)$ that depends on i , which will lead to a crude upper-bound in this case that does not depend on i . Fortunately, it will be sufficient to prove our main result (Theorem 5).

Proposition 32. *Let $c_{inf}, \varepsilon \in (0, 1)$, c_{upp} and K defined as in Proposition 29. With probability $1 - \varepsilon$, we have*

$$\forall i \in \{1, \dots, K\}, \quad \sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \leq c_{upp} K^2,$$

and

$$\forall i \in \{K+1, \dots, n\}, \quad \sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \leq \min \left\{ c_{upp} i^2, \frac{4c_{upp}}{(\alpha c_{inf})^2} + \frac{4c_{upp}K}{\alpha c_{inf}} + c_{upp} K^2 \right\}$$

Proof. We bound

$$H_0^i := \sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)}.$$

We recall $c_{inf} < 1 < c_{upp}$ are chosen such that

$$K = \left\lceil (c_{upp} - 1 - \log(c_{upp}))^{-1} \log\left(\frac{2n}{\varepsilon}\right) \right\rceil = \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(\frac{2n}{\varepsilon}\right) \right\rceil.$$

If $i \leq K$, with probability $1 - \varepsilon$ we have

$$H_0^i \leq \sum_{j=1}^i (T_i - T_j) \stackrel{(i)}{\leq} \sum_{j=1}^i c_{upp} K \leq c_{upp} i K \leq c_{upp} K^2.$$

(i) uses that as $i - j \leq K$, we have $T_i - T_j \leq T_{j+K} - T_j \leq c_{upp} K$, the last inequality being due to the Chernov inequality (see (78)).

If $i > K$, we write

$$\sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} = \sum_{j=1}^{i-K} (T_i - T_j) e^{-\alpha(T_i - T_j)} + \sum_{j=i-K+1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)}. \quad (88)$$

For the second sum, with probability $1 - \varepsilon$,

$$\sum_{j=i-K+1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \leq \sum_{j=i-K+1}^i (T_i - T_j) \stackrel{(i)}{\leq} c_{upp} K^2.$$

Again, (i) uses that as $i - j \leq K$, we have $T_i - T_j \leq T_{j+K} - T_j \leq c_{upp} K$.

Then, with probability $1 - \varepsilon$, we have by Proposition 29,

$$T_i - T_j \geq c_{inf}(i - j) - c_{inf} K,$$

and

$$T_i - T_j \leq c_{upp}(i - j) + c_{upp} K.$$

We use both inequalities simultaneously, the upper-bound controls the factor $(T_i - T_j)$, the lower-bound gives exponential decay in $e^{-\alpha(T_i - T_j)}$. Let $q := e^{-\alpha c_{inf}}$. Therefore,

$$\begin{aligned} H_0^i &\leq \sum_{j=1}^{i-K} (c_{upp}(i - j) + c_{upp} K) e^{-\alpha(c_{inf}(i-j) - c_{inf} K)} + c_{upp} K^2 \\ &\stackrel{(i)}{\leq} \sum_{\ell=0}^{i-K-1} (c_{upp} \ell + 2c_{upp} K) q^\ell + c_{upp} K^2. \end{aligned} \tag{89}$$

In (i) we set the change $\ell = i - j - K$. Then,

$$H_0^i \leq c_{upp} \sum_{\ell=0}^{i-K-1} \ell q^\ell + 2c_{upp} K \sum_{\ell=0}^{i-K-1} q^\ell + c_{upp} K^2. \tag{90}$$

Now,

$$\sum_{\ell=0}^{m-1} \ell q^\ell \leq \min \left\{ \sum_{\ell=0}^{m-1} \ell, \sum_{\ell=0}^{\infty} \ell q^\ell \right\} = \min \left\{ \frac{m(m-1)}{2}, \frac{q}{(1-q)^2} \right\},$$

and

$$\sum_{\ell=0}^{m-1} q^\ell \leq \min \left\{ m, \frac{1}{1-q} \right\}.$$

Applying this with $m = i - K$, we obtain

$$\begin{aligned} H_0^i &\leq c_{upp} \min \left\{ \frac{(i-K)(i-K-1)}{2}, \frac{q}{(1-q)^2} \right\} \\ &\quad + 2c_{upp} K \min \left\{ i-K, \frac{1}{1-q} \right\} + c_{upp} K^2. \end{aligned} \tag{91}$$

Using $\min(a, b) + \min(c, d) \leq \min(a + c, b + d)$, we get

$$\begin{aligned} H_0^i &\leq \min \left\{ c_{upp} \frac{(i-K)(i-K-1)}{2} + 2c_{upp} K(i-K) + c_{upp} K^2, \right. \\ &\quad \left. c_{upp} \frac{q}{(1-q)^2} + \frac{2c_{upp} K}{1-q} + c_{upp} K^2 \right\}. \end{aligned} \tag{92}$$

The first component in the min is equal to $c_{upp} \frac{i^2 - i + 2iK + K - K^2}{2}$. Then,

$$\begin{aligned} c_{upp} \frac{i^2 - i + 2iK + K - K^2}{2} &= c_{upp} \frac{i^2 - i + (i + 1/2)^2 - (K - i - 1/2)^2}{2} \\ &\leq c_{upp} \frac{i^2 - i + (i + 1/2)^2 - (1/2)^2}{2}, \text{ as } K < i \\ &= c_{upp}. \end{aligned}$$

So, we deduce

$$H_0^i \leq \min \left\{ c_{upp} i^2, c_{upp} \frac{q}{(1-q)^2} + \frac{2c_{upp}K}{1-q} + c_{upp}K^2 \right\}. \quad (93)$$

As moreover $\alpha c_{inf} \leq 1$, then

$$1 - q = 1 - e^{-\alpha c_{inf}} \geq \frac{\alpha c_{inf}}{2},$$

hence

$$\frac{1}{1-q} \leq \frac{2}{\alpha c_{inf}}, \quad \frac{q}{(1-q)^2} \leq \frac{1}{(1-q)^2} \leq \frac{4}{(\alpha c_{inf})^2}.$$

Therefore,

$$H_0^i \leq \min \left\{ c_{upp} i^2, \frac{4c_{upp}}{(\alpha c_{inf})^2} + \frac{4c_{upp}K}{\alpha c_{inf}} + c_{upp}K^2 \right\}. \quad (94)$$

□

E.2 lower-bounding expectations

The three lemmas below compute or lower-bound the terms $\mathbb{E}[H_k^i]$, recalling

$$H_0^i := \sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)}, \quad H_1^i := \sum_{j=1}^i e^{-\alpha(T_i - T_j)}, \quad H_2^i := \sum_{j=1}^i (i - j) e^{-\alpha(T_i - T_j)}.$$

Since $T_i - T_j \sim \Gamma(i - j, 1)$, the Laplace transform of the Gamma distribution gives exact closed-form expressions, which are then lower-bounded using elementary inequalities.

Lemma 33. *We have*

- $\mathbb{E}[H_0^i] = \frac{1 - i(1+\alpha)^{-(i-1)} + (i-1)(1+\alpha)^{-i}}{\alpha^2}$
- $\mathbb{E}[H_2^i] = (1 + \alpha) \frac{1 - i(1+\alpha)^{-(i-1)} + (i-1)(1+\alpha)^{-i}}{\alpha^2}$

Proof. Applying Lemma 39, we have

$$\mathbb{E}[H_2^i] = \sum_{j=1}^i (i - j) \mathbb{E} \left[e^{-\alpha(T_i - T_j)} \right] = \sum_{j=1}^i (i - j) (1 + \alpha)^{-(i-j)} = \sum_{k=1}^{i-1} k (1 + \alpha)^{-k}.$$

By Lemma 38 applied to $\gamma = (1 + \alpha)^{-1}$, we have

$$\begin{aligned} \mathbb{E}[H_2^i] &= (1 + \alpha)^{-1} \frac{1 - i(1 + \alpha)^{-(i-1)} + (i-1)(1 + \alpha)^{-i}}{(1 - (1 + \alpha)^{-1})^2} \\ &= (1 + \alpha) \frac{1 - i(1 + \alpha)^{-(i-1)} + (i-1)(1 + \alpha)^{-i}}{\alpha^2}. \end{aligned}$$

By Lemma 26, since $T_i - T_j \sim \Gamma(i - j, 1)$ and setting $k = i - j$:

$$\begin{aligned} \mathbb{E}[H_0^i] &= \sum_{j=1}^i \mathbb{E} \left[(T_i - T_j) e^{-\alpha(T_i - T_j)} \right] = \sum_{k=1}^{i-1} k (1 + \alpha)^{-(k+1)} \\ &= (1 + \alpha)^{-1} \sum_{k=1}^{i-1} k (1 + \alpha)^{-k} = (1 + \alpha)^{-1} \mathbb{E}[H_2^i] \\ &= \frac{1 - i(1 + \alpha)^{-(i-1)} + (i-1)(1 + \alpha)^{-i}}{\alpha^2}. \end{aligned}$$

□

Lemma 34.

$$\mathbb{E}[H_1^i] \geq (1 - e^{-1}) \min\left\{i, \frac{1}{\alpha}\right\}.$$

Proof. Applying Lemma 39 and bounding the discrete sum by an integral, we have:

$$\mathbb{E}[H_1^i] = \sum_{j=0}^{i-1} (1 + \alpha)^{-j} \geq \sum_{j=0}^{i-1} e^{-\alpha j} \geq \int_0^i e^{-\alpha t} dt = \frac{1 - e^{-\alpha i}}{\alpha}.$$

We also applied the standard inequality $\log(1 + x) \leq x$ to $x = \alpha$, which implies $(1 + \alpha)^{-1} \geq e^{-\alpha}$, then $(1 + \alpha)^{-j} \geq e^{-\alpha j}$. Now, we split the proof between the case $i \leq \alpha^{-1}$ and $i > \alpha^{-1}$.

Case $i \leq \alpha^{-1}$. Then $\alpha i \in [0, 1]$. By concavity of $t \mapsto 1 - e^{-t}$ on $[0, 1]$, we have $1 - e^{-t} \geq (1 - e^{-1})t$. Applying it to $t = \alpha i$ we deduce

$$\frac{1 - e^{-\alpha i}}{\alpha} \geq (1 - e^{-1})i.$$

Case $i > \alpha^{-1}$. We have $\alpha i > 1$, which implies $1 - e^{-\alpha i} > 1 - e^{-1}$. Hence

$$\frac{1 - e^{-\alpha i}}{\alpha} > \frac{1 - e^{-1}}{\alpha}$$

From the two cases, we deduce

$$\mathbb{E}[H_1^i] \geq (1 - e^{-1}) \min\left\{i, \frac{1}{\alpha}\right\}.$$

□

Lemma 35. *We have*

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{32} \min\left\{i^2, \frac{1}{\alpha^2}\right\}.$$

Proof. Writing $r = (1 + \alpha)^{-1}$, Lemma 33 gives

$$\mathbb{E}[H_0^i] = \frac{1 - i r^{i-1} + (i-1)r^i}{\alpha^2} = \frac{1 - r^{i-1}(i - r(i-1))}{\alpha^2}.$$

Since

$$i - (i-1)r = 1 + (i-1)(1-r) = 1 + \frac{\alpha(i-1)}{1+\alpha},$$

noting $y := \frac{\alpha(i-1)}{1+\alpha}$, it follows that

$$\mathbb{E}[H_0^i] = \frac{1 - r^{i-1}(1+y)}{\alpha^2}.$$

We use $\log(1 + x) \geq x/(1 + x)$ at $x = \alpha$, which induces

$$r^{i-1} = e^{-(i-1)\log(1+\alpha)} \leq e^{-(i-1)\frac{\alpha}{1+\alpha}} = e^{-y},$$

and hence

$$\mathbb{E}[H_0^i] \geq \frac{1 - (1+y)e^{-y}}{\alpha^2}.$$

Now, we note the following integration by part property: $\int_0^y t e^{-t} dt = [-t e^{-t}]_0^y + \int_0^y e^{-t} dt = -y e^{-y} + (1 - e^{-y}) = 1 - (1+y)e^{-y}$. We lower-bound this integral form, depending on the value of y .

- $y \leq 1$: $\int_0^y t e^{-t} dt \geq e^{-1} \int_0^y t dt = \frac{e^{-1}}{2} y^2$.
- $y \geq 1$: $\int_0^y t e^{-t} dt \geq \int_0^1 t e^{-t} dt \geq \frac{e^{-1}}{2}$.

From this we deduce

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{2\alpha^2} \min\{y^2, 1\}.$$

As $0 < \alpha \leq 1$, then $y \geq \alpha(i-1)/2$, so $\min\{y^2, 1\} \geq \min\{\alpha^2(i-1)^2/4, 1\}$. This implies

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{8} \min\left\{(i-1)^2, \frac{1}{\alpha^2}\right\}.$$

For $i \geq 2$, we have $(i-1) \geq i/2$ so $(i-1)^2 \geq i^2/4$. We then conclude

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{32} \min\left\{i^2, \frac{1}{\alpha^2}\right\}.$$

□

Lemma 36. *We have*

$$\mathbb{E}[H_2^i] \geq \frac{e^{-1}}{32} \min\left\{i^2, \frac{1}{\alpha^2}\right\}.$$

Proof. The claim can be deduced from Lemma 35. Indeed, recall

$$\mathbb{E}[H_0^i] = \sum_{k=1}^{i-1} \frac{k}{(1+\alpha)^{k+1}} \leq \sum_{k=1}^{i-1} \frac{k}{(1+\alpha)^k} = \mathbb{E}[H_2^i],$$

where we used $\alpha > 0$.

□

E.3 Combining lower and upper-bounds

We conclude the proof of Theorem 12 in this section.

We define

$$C_n := \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \right\rceil c_{inf} \alpha = K c_{inf} \alpha. \quad (95)$$

In the proof, we will replace K with $C_n/(\alpha c_{inf})$. Therefore, C_n measures the cost of replacing K by $(\alpha c_{inf})^{-1}$ in the upper-bounds of Section E.1. Note that for our convergence result, our final choice of α will be of order $1/n^{1/7}$, such that C_n will be of order $\log(\frac{n}{\varepsilon})n^{-1/7}$. Therefore, we will have $C_n \rightarrow 0$ as $n \rightarrow +\infty$, and that C_n is uniformly bounded on $n \in \mathbb{N}^*$.

The proof for the 3 sums are similar, it combines the upper-bounds from Section E.1 and the lower-bounds from Section E.2.

(1st step.) From Proposition 32, with probability $1 - \varepsilon$, we have

$$\forall i \in \{K+1, \dots, n\}, \quad H_0^i \leq \min\left\{c_{upp} i^2, \frac{4c_{upp}}{(\alpha c_{inf})^2} + \frac{4c_{upp}K}{\alpha c_{inf}} + c_{upp}K^2\right\}.$$

Using the definition (95) of C_n , this becomes

$$\forall i \in \{K+1, \dots, n\}, \quad H_0^i \leq \min\left\{c_{upp} i^2, (4 + 4C_n + C_n^2) \frac{c_{upp}}{(\alpha c_{inf})^2}\right\}.$$

As $(4 + 4C_n + C_n^2)/c_{inf}^2 \geq 1$, it implies that with probability $1 - \varepsilon$, $\forall i \in \{K+1, \dots, n\}$, we have

$$H_0^i \leq \frac{c_{upp}}{c_{inf}^2} (4 + 4C_n + C_n^2) \min\left\{i^2, \frac{1}{\alpha^2}\right\}, \quad (96)$$

Now, from Lemma 35, we have

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{32} \min\left\{i^2, \frac{1}{\alpha^2}\right\}. \quad (97)$$

Combining (96) and (97), we deduce that with probability $1 - \varepsilon$, $\forall i \in \{K+1, \dots, n\}$, we have

$$\sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \leq 32(4 + 4C_n + C_n^2) \frac{c_{upp} e}{c_{inf}^2} \mathbb{E} \left[\sum_{j=1}^i (T_i - T_j) e^{-\alpha(T_i - T_j)} \right].$$

(2nd step.) From Proposition 30, with probability $1 - \varepsilon$

$$\forall i \in \{1, \dots, n\}, H_1^i \leq \min \left\{ i, \frac{2}{\alpha c_{inf}} + K \right\}.$$

Using the definition (95) of C_n , we get that with probability $1 - \varepsilon$, $\forall i \in \{1, \dots, n\}$,

$$\forall i \in \{1, \dots, n\}, H_1^i \leq \min \left\{ i, \frac{2 + C_n}{\alpha c_{inf}} \right\}. \quad (98)$$

Now, from Lemma 34, for any $i \in \{1, \dots, n\}$,

$$(1 - e^{-1})^{-1} \mathbb{E} [H_1^i] \geq \min \left\{ i, \frac{1}{\alpha} \right\}. \quad (99)$$

Combining (98) and (99), we deduce that with probability $1 - \varepsilon$,

$$\forall i \in \{1, \dots, n\}, H_1^i \leq \frac{2 + C_n}{c_{inf}} (1 - e^{-1})^{-1} \mathbb{E} [H_1^i].$$

(3d step.) By Proposition 31, with probability $1 - \varepsilon$,

$$\forall i \in \{1, \dots, n\}, H_2^i \leq \min \left\{ \frac{i^2}{2}, \frac{4}{(\alpha c_{inf})^2} + \frac{2K}{\alpha c_{inf}} + K^2 \right\}.$$

Using the definition (95) of C_n , we get that with probability $1 - \varepsilon$, $\forall i \in \{1, \dots, n\}$,

$$H_2^i \leq \min \left\{ \frac{i^2}{2}, \frac{4 + 2C_n + C_n^2}{(\alpha c_{inf})^2} \right\}.$$

Because $1/2 < 4 + 2C_n + C_n^2$, we deduce that with probability $1 - \varepsilon$, $\forall i \in \{1, \dots, n\}$,

$$H_2^i \leq \frac{(4 + 2C_n + C_n^2)}{c_{inf}^2} \min \left\{ i^2, \frac{1}{\alpha^2} \right\}. \quad (100)$$

Now, from Lemma 36, for any $i \in \{1, \dots, n\}$,

$$\mathbb{E}[H_2^i] \geq \frac{e^{-1}}{32} \min \left\{ i^2, \frac{1}{\alpha^2} \right\}. \quad (101)$$

Combining (100) and (101), we deduce that with probability $1 - \varepsilon$,

$$\forall i \in \{1, \dots, n\}, H_2^i \leq 32(4 + 2C_n + C_n^2) \frac{e}{c_{inf}^2} \mathbb{E} [H_2^i].$$

Conclusion. We conclude by defining

$$C_n := \max \left\{ 32(4 + 4C_n + C_n^2) \frac{c_{upp}e}{c_{inf}^2}, \frac{2 + C_n}{c_{inf}} (1 - e^{-1})^{-1}, 32(4 + 2C_n + C_n^2) \frac{e}{c_{inf}^2} \right\}.$$

It follows that

$$C_n = 32(4 + 4C_n + C_n^2) \frac{c_{upp}e}{c_{inf}^2}.$$

F Additional technical Lemmas used to prove Theorem 5

In this section, we state and prove lemmas used to prove Theorem 5, but not explicitly stated in Section 5. We start with some elementary results.

Lemma 37. Let $\alpha > 0$, $i \geq 1$. We have for $c > 0$

$$\sum_{j=1}^i (1 + c\alpha)^{-(i-j)} = \frac{1 + c\alpha - (1 + c\alpha)^{-i+1}}{c\alpha},$$

Proof.

$$\begin{aligned} \sum_{j=1}^i (1 + c\alpha)^{-(i-j)} &= \sum_{k=0}^{i-1} (1 + c\alpha)^{-k} \\ &= \frac{1 - (1 + c\alpha)^{-i}}{1 - (1 + c\alpha)^{-1}} \\ &= \frac{1 + c\alpha - (1 + c\alpha)^{-i+1}}{c\alpha} \end{aligned}$$

□

Lemma 38. Let $\gamma \neq 1$, $n \in \mathbb{N}^*$. We have

$$\sum_{k=1}^n k\gamma^k = \gamma \frac{1 - (n+1)\gamma^n + n\gamma^{n+1}}{(1 - \gamma)^2}. \quad (102)$$

Proof. Define $\varphi(\gamma) := \sum_{k=1}^n \gamma^k = \frac{\gamma - \gamma^{n+1}}{1 - \gamma}$. We have

$$\varphi'(y) = \sum_{k=1}^n k\gamma^{k-1} = \frac{(1 - (n+1)\gamma^n)(1 - \gamma) + \gamma - \gamma^{n+1}}{(1 - \gamma)^2} = \frac{1 - (n+1)\gamma^n + n\gamma^{n+1}}{(1 - \gamma)^2}.$$

Multiply by γ allows to get (102).

□

Lemma 39. If $i \leq j$, we have

$$\mathbb{E} \left[e^{-\alpha(T_j - T_i)} \right] = (1 + \alpha)^{-j+i}.$$

Proof. It is trivially true if $i = j$. If $i < j$, we have

$$e^{-\alpha(T_j - T_i)} = \prod_{k=0}^{j-i-1} e^{-\alpha(T_{k+i+1} - T_{k+i})}$$

By independence of the $T_{k+1} - T_k$ for all k , we have

$$\mathbb{E} \left[\prod_{k=0}^{j-i-1} e^{-\alpha(T_{k+i+1} - T_{k+i})} \right] = \prod_{k=0}^{j-i-1} \mathbb{E} \left[e^{-\alpha(T_{i+j+1} - T_{i+j})} \right].$$

Recall $T_{k+1} - T_k$ for all k follows an exponential law with parameter 1, such that we have $\mathbb{E} \left[e^{-\alpha(T_{i+j+1} - T_{i+j})} \right] = (1 + \alpha)^{-1}$. So,

$$\begin{aligned} \prod_{k=0}^{j-i-1} \mathbb{E} \left[e^{-\alpha(T_{i+j+1} - T_{i+j})} \right] &= \prod_{k=0}^{j-i-1} (1 + \alpha)^{-1} \\ &= (1 + \alpha)^{-j+i}. \end{aligned}$$

□

Lemma 40. For $n \geq 1$, $\alpha > 0$, we have

(i) The following upper-bound

$$\mathbb{E} \left[\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \leq \left(1 + \frac{3}{2}\alpha \right) (1 + 2\alpha)n.$$

(ii) The following equality

$$\begin{aligned} & \mathbb{E} \left[\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] \\ &= \frac{(2 + \alpha)(1 + 2\alpha)}{2} n + \frac{(1 + 2\alpha)(1 + \frac{3}{2}\alpha)}{2\alpha} (1 - (1 + 2\alpha)^{-n}) \\ & \quad - \frac{2(1 + \alpha)(1 + 2\alpha)}{\alpha} (1 - (1 + \alpha)^{-n}) \end{aligned}$$

Proof. **Point (i)** We have

$$\begin{aligned} \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t &= \sum_{i=1}^n \left(\sum_{j=1}^i \alpha e^{\alpha(T_j - T_i)} \right)^2 \\ &= \alpha^2 \sum_{i=1}^n \sum_{j=1}^i e^{-2\alpha(T_i - T_j)} + 2\alpha^2 \sum_{i=1}^n \sum_{1 \leq j < l \leq i} e^{-\alpha(T_i - T_j)} e^{-\alpha(T_i - T_l)}. \end{aligned} \tag{103}$$

We focus on the first term in (103). We have by using Lemma 39 and Lemma 37 ($c = 2$)

$$\begin{aligned} \mathbb{E} \left[\alpha^2 \sum_{i=1}^n \sum_{j=1}^i e^{-2\alpha(T_i - T_j)} \right] &= \alpha^2 \sum_{i=1}^n \sum_{j=1}^i \mathbb{E} \left[e^{-2\alpha(T_i - T_j)} \right] \\ &= \alpha^2 \sum_{i=1}^n \sum_{j=1}^i (1 + 2\alpha)^{-(i-j)} \\ &= \alpha^2 \sum_{i=1}^n \frac{1 + 2\alpha - (1 + 2\alpha)^{-i+1}}{2\alpha} \end{aligned}$$

Omitting the non positive term, we have

$$\mathbb{E} \left[\alpha^2 \sum_{i=1}^n \sum_{j=1}^i e^{-2\alpha(T_i - T_j)} \right] \leq \alpha n \frac{1 + 2\alpha}{2}.$$

We focus on the second term of (103). We have

$$2\alpha^2 \sum_{i=1}^n \sum_{1 \leq j < l \leq i} e^{-\alpha(T_i - T_j)} e^{-\alpha(T_i - T_l)} = 2\alpha^2 \sum_{i=1}^n \sum_{1 \leq j < l \leq i} e^{-2\alpha(T_i - T_l)} e^{-\alpha(T_l - T_j)}$$

We take expectation and we use Lemma 39

$$\begin{aligned} 2\alpha^2 \mathbb{E} \left[\sum_{i=1}^n \sum_{1 \leq j < l \leq i} e^{-\alpha(T_i - T_j)} e^{-\alpha(T_i - T_l)} \right] &= 2\alpha^2 \sum_{i=1}^n \sum_{1 \leq j < l \leq i} \mathbb{E} \left[e^{-2\alpha(T_i - T_l)} e^{-\alpha(T_l - T_j)} \right] \\ &= 2\alpha^2 \sum_{i=1}^n \sum_{1 \leq j < l \leq i} \mathbb{E} \left[e^{-2\alpha(T_i - T_l)} \right] \mathbb{E} \left[e^{-\alpha(T_l - T_j)} \right] \\ &= 2\alpha^2 \sum_{i=1}^n \sum_{1 \leq j < l \leq i} (1 + 2\alpha)^{-(i-l)} (1 + \alpha)^{-(l-j)} \\ &= 2\alpha^2 \sum_{i=1}^n \sum_{l=1}^i \sum_{j=1}^{l-1} (1 + 2\alpha)^{-(i-j)} (1 + \alpha)^{-(i-j)}. \end{aligned}$$

Now, using twice Lemma 37 with $c = 1$ and $c = 2$, and omitting non negative terms

$$\begin{aligned}
2\alpha^2 \sum_{i=1}^n \sum_{l=1}^i \sum_{j=1}^{l-1} (1+2\alpha)^{-(i-l)} (1+\alpha)^{-(i-j)} &= 2\alpha^2 \sum_{i=1}^n \sum_{l=1}^i (1+2\alpha)^{-(i-l)} \sum_{j=1}^{l-1} (1+\alpha)^{-(l-j)} \\
&= 2\alpha^2 \sum_{i=1}^n \sum_{l=1}^i (1+2\alpha)^{-(i-l)} \left(\frac{1+\alpha - (1+\alpha)^{-i+1}}{\alpha} \right) \\
&\leq 2\alpha(1+\alpha) \sum_{i=1}^n \sum_{l=1}^i (1+2\alpha)^{-(i-l)} \\
&= 2\alpha(1+\alpha) \sum_{i=1}^n \left(\frac{1+2\alpha - (1+2\alpha)^{-i+1}}{2\alpha} \right) \\
&\leq (1+\alpha)(1+2\alpha)n
\end{aligned}$$

We then deduce that

$$2\alpha^2 \mathbb{E} \left[\sum_i^n \sum_{1 \leq j < l \leq i} e^{-\alpha(T_i - T_j)} e^{-\alpha(T_i - T_l)} \right] \leq (1+\alpha)(1+2\alpha)n.$$

Finally, we have the bound

$$\begin{aligned}
\mathbb{E} \left[\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] &\leq \alpha \frac{1+2\alpha}{2} n + (1+\alpha)(1+2\alpha)n \\
&= \left(1 + \frac{3}{2}\alpha \right) (1+2\alpha)n.
\end{aligned}$$

Point (ii) Let $Y_k = \left(\int_{[0, T_k]} \alpha e^{\alpha(s-T_k)} dN_s \right)^2$. We have

$$\begin{aligned}
Y_{k+1} &= \left(\int_{[0, T_{k+1}]} \alpha e^{\alpha(s-T_{k+1})} dN_s \right)^2 \\
&= \left(e^{\alpha(T_k - T_{k+1})} \int_{[0, T_k]} \alpha e^{\alpha(s-T_k)} dN_s + \int_{[T_k, T_{k+1}]} \alpha e^{\alpha(s-T_{k+1})} dN_s \right)^2.
\end{aligned}$$

We have $\int_{[T_k, T_{k+1}]} \alpha e^{\alpha(s-T_{k+1})} dN_s = \alpha$. So

$$Y_{k+1} = e^{2\alpha(T_k - T_{k+1})} Y_k + 2\alpha e^{\alpha(T_k - T_{k+1})} \int_{[0, T_k]} \alpha e^{\alpha(s-T_k)} dN_s + \alpha^2.$$

We take the expectation

$$\begin{aligned}
\mathbb{E}[Y_{k+1}] &= \mathbb{E} \left[e^{2\alpha(T_k - T_{k+1})} Y_k + 2\alpha e^{\alpha(T_k - T_{k+1})} \int_{[0, T_k]} \alpha e^{\alpha(s-T_k)} dN_s + \alpha^2 \right] \\
&= \mathbb{E} \left[e^{2\alpha(T_k - T_{k+1})} Y_k \right] + 2\alpha \mathbb{E} \left[e^{\alpha(T_k - T_{k+1})} \int_{[0, T_k]} \alpha e^{\alpha(s-T_k)} dN_s \right] + \alpha^2 \\
&= \mathbb{E} \left[e^{2\alpha(T_k - T_{k+1})} \right] \mathbb{E}[Y_k] + 2\alpha \mathbb{E} \left[e^{\alpha(T_k - T_{k+1})} \right] \mathbb{E} \left[\int_{[0, T_k]} \alpha e^{\alpha(s-T_k)} dN_s \right] + \alpha^2 \\
&= (1+2\alpha)^{-1} \mathbb{E}[Y_k] + 2\alpha(1+\alpha)^{-1} (1+\alpha - (1+\alpha)^{-k+1}) + \alpha^2 \\
&= (1+2\alpha)^{-1} \mathbb{E}[Y_k] + 2\alpha(1 - (1+\alpha)^{-k}) + \alpha^2 \\
&= (1+2\alpha)^{-1} \mathbb{E}[Y_k] - 2\alpha(1+\alpha)^{-k} + \alpha(2+\alpha).
\end{aligned}$$

Noting $u_k := \mathbb{E}[Y_k]$, we get a recursive formula of the form

$$u_{k+1} = au_k + b_k + c,$$

whose solution is given by

$$u_k = a^k \sum_{i=0}^{k-1} \frac{b_i + c}{a^{i+1}} = \sum_{i=0}^{k-1} (b_i + c) a^{k-i-1}. \quad (104)$$

We compute (104) in two steps. We start with

$$\begin{aligned} \sum_{i=0}^{k-1} ca^{k-i-1} &= c \sum_{j=0}^{k-1} a^j \\ &= c \frac{1 - a^k}{1 - a} \\ &= \alpha(2 + \alpha) \frac{1 - (1 + 2\alpha)^{-k}}{1 - (1 + 2\alpha)^{-1}} \\ &= \frac{(2 + \alpha)(1 + 2\alpha)}{2} - \frac{2 + \alpha}{2} (1 + 2\alpha)^{-k+1}. \end{aligned} \quad (105)$$

Then, we have

$$\begin{aligned} a^{k-1} \sum_{i=0}^{k-1} \frac{b_i}{a^i} &= -2\alpha a^{k-1} \sum_{i=0}^{k-1} \left(\frac{1 + 2\alpha}{1 + \alpha} \right)^i \\ &= -2\alpha a^{k-1} \left(1 - \left(\frac{1 + 2\alpha}{1 + \alpha} \right)^k \right) \left(1 - \left(\frac{1 + 2\alpha}{1 + \alpha} \right) \right)^{-1} \\ &= 2(1 + \alpha) (a^{k-1} - a^{-1}(1 + \alpha)^{-k}) \\ &= 2(1 + \alpha)(1 + 2\alpha)((1 + 2\alpha)^{-k} - (1 + \alpha)^{-k}) \\ &= 2(1 + \alpha)(1 + 2\alpha)(1 + 2\alpha)^{-k} - 2(1 + \alpha)(1 + 2\alpha)(1 + \alpha)^{-k}. \end{aligned} \quad (106)$$

Injecting (105) and (106) into (104), we get the following expression for u_k

$$\begin{aligned} u_k &= \frac{(2 + \alpha)(1 + 2\alpha)}{2} - \frac{2 + \alpha}{2} (1 + 2\alpha)^{-k+1} + 2(1 + \alpha)(1 + 2\alpha)(1 + 2\alpha)^{-k} \\ &\quad - 2(1 + \alpha)(1 + 2\alpha)(1 + \alpha)^{-k} \\ &= \frac{(2 + \alpha)(1 + 2\alpha)}{2} + \left(-\frac{(2 + \alpha)(1 + 2\alpha)}{2} + 2(1 + \alpha)(1 + 2\alpha) \right) (1 + 2\alpha)^{-k} \\ &\quad - 2(1 + \alpha)(1 + 2\alpha)(1 + \alpha)^{-k} \\ &= \frac{(2 + \alpha)(1 + 2\alpha)}{2} + (1 + 2\alpha)(1 + \frac{3}{2}\alpha)(1 + 2\alpha)^{-k} - 2(1 + \alpha)(1 + 2\alpha)(1 + \alpha)^{-k}. \end{aligned}$$

It remains to sum over $k = 0, \dots, n$, to get

$$\begin{aligned} \sum_{k=1}^n u_k &= \frac{(2 + \alpha)(1 + 2\alpha)}{2} n + \frac{(1 + 2\alpha)(1 + \frac{3}{2}\alpha)}{2\alpha} (1 - (1 + 2\alpha)^{-n}) \\ &\quad - \frac{2(1 + \alpha)(1 + 2\alpha)}{\alpha} (1 - (1 + \alpha)^{-n}). \end{aligned}$$

To conclude, remark that

$$\mathbb{E} \left[\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \right] = \sum_{k=1}^n \mathbb{E} \left[\left(\int_{[0, T_k]} \alpha e^{\alpha(s-T_k)} dN_s \right)^2 \right] = \sum_{k=1}^n \mathbb{E}[Y_k] = \sum_{k=1}^n u_k.$$

□

Lemma 41. *Let $\tilde{\varepsilon} \in (0, 1/2]$, and $c_{upp} \geq 1$ such that it satisfies $1 - \tilde{\varepsilon} - \log(1 - \tilde{\varepsilon}) = c_{upp} - \log(c_{upp})$. Then, we have $c_{upp} \leq 1 + 2\tilde{\varepsilon}$.*

Proof. We denote $h(t) = t - \log(t)$. By assumption, we have $h(c_{upp}) = h(1 - \tilde{\varepsilon})$. We have that h is non-decreasing on $[1, +\infty)$. Then, if we show $h(1 + c_0\tilde{\varepsilon}) \geq h(1 - \tilde{\varepsilon})$ for some $c_0 \geq 1$, it will induce $c_{upp} \leq 1 + c_0\tilde{\varepsilon}$. First,

$$h(1 + c_0\tilde{\varepsilon}) - 1 = c_0\tilde{\varepsilon} - \log(1 + c_0\tilde{\varepsilon}) = \int_0^{c_0\tilde{\varepsilon}} \frac{t}{1+t} dt.$$

Since $t \mapsto (1+t)^{-1}$ is decreasing, for $t \in [0, c_0\tilde{\varepsilon}]$, $\frac{1}{1+t} \geq \frac{1}{1+c_0\tilde{\varepsilon}}$, therefore,

$$h(1 + c_0\tilde{\varepsilon}) - 1 \geq \frac{1}{1+c_0\tilde{\varepsilon}} \int_0^{c_0\tilde{\varepsilon}} t dt = \frac{c_0^2\tilde{\varepsilon}^2}{2(1+c_0\tilde{\varepsilon})}.$$

On the other hand,

$$h(1 - \tilde{\varepsilon}) - 1 = -\tilde{\varepsilon} - \log(1 - \tilde{\varepsilon}) = \int_0^{\tilde{\varepsilon}} \frac{t}{1-t} dt.$$

Since $t \mapsto (1-t)^{-1}$ is increasing on $[0, \tilde{\varepsilon}]$, $\frac{1}{1-t} \leq \frac{1}{1-\tilde{\varepsilon}}$, thus

$$h(1 - \tilde{\varepsilon}) - 1 \leq \frac{1}{1-\tilde{\varepsilon}} \int_0^{\tilde{\varepsilon}} t dt = \frac{\tilde{\varepsilon}^2}{2(1-\tilde{\varepsilon})}.$$

Consequently, it is enough that $\frac{c_0^2\tilde{\varepsilon}^2}{2(1+c_0\tilde{\varepsilon})} \geq \frac{\tilde{\varepsilon}^2}{2(1-\tilde{\varepsilon})}$, or equivalently $c_0^2(1 - \tilde{\varepsilon}) \geq 1 + c_0\tilde{\varepsilon}$. This is equivalent to

$$\tilde{\varepsilon} \leq \frac{c_0^2 - 1}{c_0^2 + c_0} = 1 - \frac{1}{c_0}.$$

Therefore, for any $c_0 > 1$, choosing $a := 1 - \frac{1}{c_0}$ gives, for all $\tilde{\varepsilon} \in (0, a]$, $h(1 + c_0\tilde{\varepsilon}) \geq h(1 - \tilde{\varepsilon}) = h(c_{upp})$, from which we deduce

$$c_{upp} \leq 1 + c_0\tilde{\varepsilon}.$$

The result is the particular case $c_0 = 2 \Rightarrow \tilde{\varepsilon} \in \left(0, \frac{1}{2}\right]$. □