Continuous-time Analysis of a Stochastic ADMM Method for Nonconvex Composite Optimization

Jiahong Guo* Xiao Wang† Xiantao Xiao‡

October 3, 2025

Abstract

In this paper, we focus on nonconvex composite optimization, whose objective is the sum of a smooth but possibly nonconvex function and a composition of a weakly convex function coupled with a linear operator. By leveraging a smoothing technique based on Moreau envelope, we propose a stochastic proximal linearized ADMM algorithm (SPLA). To understand its convergence behavior we consider a stochastic differential equation (SDE) that serves as a continuous-time stochastic model of the discrete scheme SPLA. Under mild conditions, we establish the almost-sure convergence for the smoothed objective function along the SDE's solution trajectory, and the associated in-expectation convergence rates in the context of Łojasiewicz inequality. We further establish the almost-sure global convergence and the in-expectation convergence rates of the SDE's solution. Building upon these convergence results, we derive the in-expectation convergence rates of a trajectory derived from the SDE's solution to some approximate critical point of the original nonsmooth problem. Finally, under certain conditions we obtain the convergence properties of the objective function values along the discrete iterates of SPLA.

1 Introduction

In this paper we consider the following nonconvex composite optimization problem

$$\min_{x\in\mathbb{R}^n} H(x) := f(x) + h(Ax), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable and possibly nonconvex function, $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator, and $h : \mathbb{R}^m \to (-\infty, +\infty]$ is a weakly convex function that is commonly referred to as the *regularizer*. The regularizer is used to guarantee certain desirable properties of the solution. Popular regularizers in the literature include the convex ℓ_1, ℓ_2 and total variation, as well as weakly convex regularizers such as the minimax concave penalty (MCP) and smoothly clipped absolute deviation (SCAD). Problem (1) arises in various applications including signal processing, image processing, machine learning, and statistics. In this paper, we assume that the exact gradient of f is unavailable, and only stochastic oracles can be accessed.

Regardless of the stochastic nature of f, the problem (1) in the convex setting has been extensively studied, and various primal-dual algorithms have been proposed, such as primal-dual hybrid gradient method, alternating direction method of multipliers (ADMM) and their variants. Related references include [23, 15, 48, 14, 35]. However, it is challenging to directly extend these algorithms to problem (1) in the fully nonconvex setting, due to the nonconvexity of f and h. To address this challenge, recent study

*Qilu Normal University, Jinan, China. (jhguo0722@163.com)
†Sun Yat-sen University, Guangzhou, China. (wangx936@mail.sysu.edu.cn)
‡Dalian University of Technology, Dalian, China. (xtxiao@dlut.edu.cn)

has focused on the Kurdyka-Łojasiewicz (KL) inequality, which provides a tool for analyzing the global convergence of optimization algorithms in the nonconvex setting. Relevant works include but are not limited to [1, 2, 3, 11, 51, 27, 29]. Li and Pong [31] applied ADMM to solve the nonconvex problem (1). They showed that under the assumptions that f and h are semialgebraic, and A is surjective, the iterates generated by ADMM converge to some critical point of the objective H. It is proved in [12] that the bounded iteration sequence generated by some Lagrangian-based methods, including proximal methods of multipliers and proximal ADMM, is globally convergent to a critical point in the semialgebraic setting. Under the assumption that the associated augmented Lagrangian satisfies the KL inequality, Boţ and Nguyen [14] proved that the iterates of proximal ADMM converge to a Karush-Kuhn-Tucker point, and established the corresponding convergence rates using the Łojasiewicz inequality. There has been a growing recognition of the significance of stochastic approximation techniques in the development of effective numerical algorithms for convex optimization problems, see, for instance, [19, 54, 5, 53]. For solving nonconvex problem (1), Bian et al. [7] proposed a general framework of stochastic ADMM with a class of variance-reduced gradient estimators. With the aid of the Łojasiewicz inequality and the properties of variance reduced gradient estimators, the global convergence and in-expectation convergence rates were also established.

Recently, there has been a lot of attention paid to the behavior of optimization algorithms from the viewpoint of continuous-time systems. For instance, the classic gradient descent method for minimizing a smooth function can be regarded as the Euler discretization of a first-order ordinary differential equation (ODE). Su et al. [50] showed that the exact limit of Nesterov's accelerated gradient method [43] is a second-order ODE. In [25], it was shown that the continuous limits of ADMM and its accelerated variant for solving (1) in the convex and smooth setting are first-order and second-order ODEs, respectively, and their convergence rates were also obtained. Reference [26] further extended the results in [25] to nonsmooth constrained problems under convex and strongly convex assumptions. An inertial system with damping, which can be viewed as the inertial continuous counterpart of ADMM-type methods, was studied in [4]. Li et al. [32] proposed the weak approximations via continuous-time stochastic differential equations (SDEs) to model the dynamics of stochastic gradient descent (SGD) algorithms. Subsequently, numerous studies have modeled SGD-type algorithms using SDEs, focusing on convex optimization with noisy gradient inputs [34, 20, 38, 39, 40, 41, 42]. Maulen-Soto et al. [38] established the almost-sure convergence of both the objective function value and SDE's solution in convex setting, as well as their corresponding in-expectation convergence rates. Dambrine et al. [20] demonstrated that SDEs offer superior approximation for stochastic gradient methods compared to classical ODEs. Moreover, inspired by [50], they proposed an inertial SDE of the Nesterov's accelerated method and derived more favorable convergence rates. Maulen-Soto et al. [40] further investigated the convergence for an inertial SDE with viscous time-varying damping and Hessian-driven damping. For nonsmooth and convex optimization, the stochastic differential inclusions with convergence results were established in [39] and [41]. Orvieto and Lucchi [45] proposed continuous-time models for stochastic optimization algorithms and derived convergence bounds for nonconvex optimization problems. Shi et al. [49] analyzed SGD using a learning-rate-dependent SDE, showing how the convergence rate of probability densities depends on the learning rate and explaining the effectiveness of learning rate decay in nonconvex optimization. The continuous-time perspective offers valuable insights into the long-term behavior of algorithms, without being tied to a specific time discretization scheme. However, despite some recent works on nonconvex problems, the use of SDE-based approaches for analyzing nonconvex optimization algorithms remains relatively limited.

Contributions. We introduce a stochastic proximal linearized ADMM (SPLA) for problem (1), based on the smoothing Moreau envelope of the objective function, and study the convergence properties of SPLA from a stochastic dynamical systems perspective by constructing a first-order SDE. We first establish the almost-sure global convergence of the smoothed function values along the SDE's solution trajectory. We also prove the in-expectation convergence rates of the smoothed function values under the Łojasiewicz

inequality. Furthermore, by exploiting the stability and regularity properties of the Moreau envelope and following the SDE's solution trajectory $\mathbf{x}(t)$, we construct a trajectory $\bar{\mathbf{x}}(t)$. For both trajectories we prove their almost-sure convergence and in-expectation convergence rates to an approximate critical point of the smoothed problem and the original problem, respectively. Finally, the convergence properties of the discrete iteration sequences generated by SPLA are established.

Compared with classical discrete-time analysis, the continuous-time analysis allows us to exploit tools from stochastic calculus, leading to deeper insights into the convergence behavior of algorithms. In particular, we establish in-expectation convergence rates for both the smoothed objective value and the SDE's solution, which are difficult to obtain using discrete techniques alone. This continuous-time viewpoint not only complements the discrete theory but also reveals fundamental structures underlying the stochastic dynamics of nonconvex optimization.

Organization. The rest of this paper is organized as follows. Section 2 introduces the necessary notations and preliminaries. In Section 3, we present a stochastic proximal linearized ADMM (SPLA) for problem (1), along with its associated first-order SDE. Section 4 is devoted to establishing the convergence properties of the objective function value by analyzing the SDE. We further derive the global convergence and convergence rates of solution trajectories in Section 5. Finally, we develop the convergence properties of the discrete algorithm SPLA in Section 6.

2 Notations and Preliminaries

Take $\mathbb{R}^n$ as an n-dimensional Euclidean space equipped with the standard inner product $\langle\cdot,\cdot\rangle$ and associated norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. Denote by $\mathbb{R}_+$ the set $[0,+\infty)$. Given a closed set $\mathcal{C}\subseteq\mathbb{R}^n$, the distance between a point $x\in\mathbb{R}^n$ and $\mathcal{C}$ by $\mathrm{dist}(x,\mathcal{C}) := \min_y\{\|x-y\| : y\in\mathcal{C}\}$. Let $\mathbb{R}^{n\times m}$ denote the space of real $n\times m$ matrices, and let I_n represent the $n\times n$ identity matrix. For a matrix $A\in\mathbb{R}^{n\times m}$, its Frobenius norm is defined as $\|A\|_{\mathrm{F}} := \sqrt{\mathrm{Tr}(AA^T)}$, where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

Consider a function $f:\mathbb{R}^n\to(-\infty,+\infty]$ and a point x with $f(x)$ being finite. The *Fréchet subdifferential* of f at x, written as $\partial f(x)$, is the set of all vectors $v\in\mathbb{R}^n$ such that

$$\liminf_{y\to x,\, y\neq x}\frac{f(y)-f(x)-\langle v,y-x\rangle}{\|y-x\|}\geq 0.$$

We denote the set of *critical points* of f by $\mathrm{crit}f := \{x\in\mathbb{R}^n : \mathrm{dist}(0,\partial f(x))=0\}$. Moreover, if f is differentiable, $\mathrm{crit}f = \{x\in\mathbb{R}^n : \nabla f(x)=0\}$. A function $f:\mathbb{R}^n\to\mathbb{R}$ is said to be *L-smooth*, if it is continuously differentiable and its gradient is L-Lipschitz continuous, i.e., $\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|$ for any $x,y\in\mathbb{R}^n$.

2.1 Moreau Envelope and Weak Convexity

Let $f:\mathbb{R}^n\to(-\infty,+\infty]$ be a proper and lower semicontinuous function. Given $\mu>0$, we define the *proximal mapping* of f as

$$\mathrm{prox}_{\mu f}(x) := \arg\min_{y\in\mathbb{R}^n}\left\{f(y)+\frac{1}{2\mu}\|y-x\|^2\right\}$$

and the *μ-Moreau envelope* of f as

$$f_\mu(x) := \inf_{y\in\mathbb{R}^n}\left\{f(y)+\frac{1}{2\mu}\|y-x\|^2\right\}.$$

It is shown in [46, Theorem 1.25] that if $\inf_{x\in\mathbb{R}^n} f(x)>-\infty$, then for any $x\in\mathbb{R}^n$, $f_\mu(x)$ increases to $f(x)$ as μ decreases to 0.

With $\varrho > 0$, a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is said to be *ϱ-weakly convex*, if the mapping $x \longmapsto f(x) + \frac{\varrho}{2}\|x\|^2$ is convex. For a proper, lower semicontinuous and ϱ-weakly convex function f, the Moreau envelope f_μ satisfies $\inf_{x\in\mathbb{R}^n} f(x) \leq f_\mu(x) \leq f(x)$ for any $x \in \mathbb{R}^n$. Moreover, as shown in [52, Proposition 3.1 and Corollary 3.4], the proximal mapping $\mathrm{prox}_{\mu f}$ is single-valued, and f_μ is convex and $\max\{\frac{1}{\mu}, \frac{\varrho}{1-\varrho\mu}\}$-smooth, provided that $\mu < 1/\varrho$.

2.2 Stochastic Differential Equation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_t\}_{t\geq 0}$ be a family of increasing sub-σ-fields of $\mathcal{F}$.

A stochastic process is a function $\mathbf{x} : \mathbb{R}_+ \times \Omega \to \mathbb{R}^n$, where for each $\omega \in \Omega$, the mapping $\omega \mapsto \mathbf{x}(t, \omega)$ defines the corresponding sample path. The stochastic process $\mathbf{x}$ is said to be almost surely (a.s.) continuous, if there exists an event $\mathcal{A} \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}) = 1$ such that for every $\omega \in \mathcal{A}$, $\mathbf{x}(\cdot, \omega)$ is continuous. For brevity, we often use $\mathbf{x}(t)$ to represent $\mathbf{x}(t, \cdot)$. We say that $\mathbf{x}$ is $\{\mathcal{F}_t\}$-adapted, if $\mathbf{x}(t)$ is $\mathcal{F}_t$-measurable for every $t \geq 0$.

Let W be a *Brownian motion* defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathbf{x} : \mathbb{R}_+ \to \mathbb{R}$ be a measurable $\{\mathcal{F}_t\}$-adapted stochastic process satisfying $\mathbb{E}[\int_S^T |\mathbf{x}(t)|^2 dt] < \infty$ for $0 \leq S < T < \infty$. Suppose $\{\mathbf{x}_n\}$ is a sequence of simple stochastic processes ([6, Definition 3.28]) such that $\lim_{n\to\infty} \mathbb{E}[\int_S^T |\mathbf{x}(t) - \mathbf{x}_n(t)|^2 dt] = 0$. Following [6, Definition 3.29], the *Itô's integral* of $\mathbf{x}$ with respect to the 1-dimensional Brownian motion $W(t)$ over $[S, T]$ is defined as

$$\int_S^T \mathbf{x}(t)\, dW(t) = \lim_{n\to\infty} \int_S^T \mathbf{x}_n(t)\, dW(t),$$

which owns the property $\mathbb{E}[\int_S^T \mathbf{x}(t)\, dW(t)] = 0$.

A *stochastic differential equation* (SDE) is an equation of the form:

$$d\mathbf{x}(t) = F(t, \mathbf{x}(t))\, dt + G(t, \mathbf{x}(t))\, dW(t), \tag{2}$$

where W is an m-dimensional Brownian motion, $F : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$ and $G : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are measurable functions. The stochastic process $\mathbf{x} : \mathbb{R}_+ \to \mathbb{R}^n$ is called a solution of (2) for $t \in [0, T]$, if $\mathbf{x}$ is $\{\mathcal{F}_t\}$-adapted and a.s. continuous, $\int_0^T \|F(s, \mathbf{x}(s))\|\, ds < \infty$ a.s. and $\int_0^T \|G(s, \mathbf{x}(s))\|_{\mathrm{F}}^2\, ds < \infty$ a.s., and for every $t \in [0, T]$,

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t F(s, \mathbf{x}(s))\, ds + \int_0^t G(s, \mathbf{x}(s))\, dW(s) \quad \text{a.s.}$$

Let $\mathbf{x}(t)$ be a stochastic process governed by the SDE (2). The *Itô's formula* ([24]) states that for any twice continuously differentiable function $\phi : \mathbb{R}^n \to \mathbb{R}$, it holds that

$$\begin{aligned} d\phi(\mathbf{x}(t)) = &\left(\langle \nabla\phi(\mathbf{x}(t)), F(t, \mathbf{x}(t))\rangle + \frac{1}{2}\mathrm{Tr}\left(G(t, \mathbf{x}(t))G(t, \mathbf{x}(t))^T \nabla^2\phi(\mathbf{x}(t))\right)\right) dt \\ &+ \langle G(t, \mathbf{x}(t))^T \nabla\phi(\mathbf{x}(t)), dW(t)\rangle, \end{aligned} \tag{3}$$

where $\nabla^2\phi(\cdot)$ is the Hessian matrix of ϕ. The twice continuous differentiability of ϕ in Itô's formula can be weakened to L-smoothness. If ϕ is L-smooth, it follows from Rademacher's Theorem that $\nabla\phi$ is differentiable almost everywhere. Then adapting the methodology from [42, Proposition C.2], we can derive

$$\begin{aligned} d\phi(\mathbf{x}(t)) \leq &\left(\langle \nabla\phi(\mathbf{x}(t)), F(t, \mathbf{x}(t))\rangle + \frac{L}{2}\mathrm{Tr}\left(G(t, \mathbf{x}(t))G(t, \mathbf{x}(t))^T\right)\right) dt \\ &+ \langle G(t, \mathbf{x}(t))^T \nabla\phi(\mathbf{x}(t)), dW(t)\rangle. \end{aligned} \tag{4}$$

2.3 Assumption

Throughout the paper, we make some essential assumptions regarding the problem (1).

Assumption 1 *For problem (1), suppose that*

(i) $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$ *and* $\inf_{y\in\mathbb{R}^m} h(y) > -\infty$;

(ii) f *is* L_f*-smooth, and* h *is* ϱ*-weakly convex and* L_h*-Lipschitz continuous;*

(iii) A *is surjective.*

Remark 1 *Some comments on Assumption 1 are given as follows.*

(i) Under Assumption 1(i), as discussed in Subsection 2.1, the μ*-Moreau envelope* h_μ *is bounded from below. Moreover, Assumption 1(ii) ensures that* h_μ *is convex and* $\max\{\frac{1}{\mu}, \frac{\varrho}{1-\varrho\mu}\}$*-smooth, provided that the parameter* μ *satisfies* $0 < \mu < 1/\varrho$*. This condition is valid as long as* μ *is taken sufficiently small. Hence, we will always be working under the condition* $0 < \mu < 1/\varrho$*.*

(ii) Under Assumption 1(ii), it is straightforward to verify that the function $H_\mu(x) := f(x) + h_\mu(Ax)$ *is* L*-smooth, where*

$$L := L_f + \max\left\{\frac{1}{\mu}, \frac{\varrho}{1-\varrho\mu}\right\} \cdot \|A\|_{\mathrm{F}}^2.$$

Thus, H_μ *can be regarded as a smoothing approximation to* H*. Besides, Assumption 1(i) guarantees that* H_μ *is bounded from below, which is significant for the subsequent convergence analysis.*

(iii) It is worth noting that A *is surjective if and only if* AA^T *is positive definite. Additionally, according to [16, Subsection 3.2], Assumption 1(ii)-(iii) is necessary to establish the relationship between* $\mathrm{crit}H$ *and* $\mathrm{crit}H_\mu$*, as stated in Lemma 1.*

Given $\epsilon > 0$, we call a point $x \in \mathbb{R}^n$ an *ϵ-critical point* of H if $\mathrm{dist}(0, \partial H(x)) \le \epsilon$. The set of all ϵ-critical points of H is denoted by $\mathrm{crit}_\epsilon H$. The following lemma shows that a critical point of H_μ can be an ϵ-critical point of H, provided that μ is sufficiently small. The proof of this lemma follows [16, Subsection 3.2].

Lemma 1 *Suppose that Assumption 1(ii)-(iii) holds, and let the parameter* μ *satisfy* $0 < \mu < \min\left\{\frac{1}{\varrho}, \frac{\epsilon\sqrt{\lambda_{\min}(AA^T)}}{L_f L_h}\right\}$*, where* $\lambda_{\min}(AA^T)$ *denotes the smallest eigenvalue of* AA^T*. Then, for any given* $x \in \mathrm{crit}H_\mu$*, the point*

$$\bar{x} := x - A^T(AA^T)^{-1}(Ax - \mathrm{prox}_{\mu h}(Ax)) \tag{5}$$

is an ϵ*-critical point of* H*, i.e.,* $\bar{x} \in \mathrm{crit}_\epsilon H$*.*

3 SPLA and Continuous-Time System

In this section, we present the details of a stochastic proximal linearized ADMM algorithm, SPLA, for solving problem (1), and then derive the corresponding first-order SDE of SPLA.

Consider an equivalent form of (1):

$$\min_{x\in\mathbb{R}^n,\, z\in\mathbb{R}^m} f(x) + h(z) \quad \text{s.t. } z = Ax.$$

The classical ADMM, as described in [15], runs the update:

$$x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \left\{ f(x) + \langle u^k, Ax - z^k\rangle + \frac{\rho}{2}\|Ax - z^k\|^2 \right\}, \tag{6a}$$

$$z^{k+1} \in \arg\min_{z\in\mathbb{R}^m} \left\{ h(z) + \langle u^k, Ax^{k+1} - z\rangle + \frac{\rho}{2}\|Ax^{k+1} - z\|^2 \right\}, \tag{6b}$$

$$u^{k+1} = u^k + Ax^{k+1} - z^{k+1}, \tag{6c}$$

where $\rho > 0$ refers to a penalty parameter. It is worth noting that an inherent challenge arises in updating x^{k+1} due to the presence of the coupled term Ax. Meanwhile, the potential nonlinearity and nonconvexity of f can make it challenging to compute the proximal mapping of f. To address these issues, a linearization approach with respect to x^k in (6a) is used to obtain a more tractable point by solving subproblem

$$\min_{x\in\mathbb{R}^n} \quad \langle \nabla f(x^k), x - x^k\rangle + \langle u^k, Ax - z^k\rangle + \frac{\rho}{2}\|Ax - z^k\|^2 + \frac{1}{2\eta}\|x - x^k\|_M^2. \tag{7}$$

Here, M is a symmetric positive definite matrix and $\|x\|_M^2 := x^T Mx$. Then (7) admits a unique solution. However, $\rho A^T A + \frac{1}{\eta}M$ is a dense matrix in general. To further ease the potential computational burden, we set $M = \eta(\tau I - \rho A^T A)$ in (7), leading to the update of x^k:

$$x^{k+1} = x^k - \frac{1}{\tau}\left(\nabla f(x^k) + \rho A^T(Ax^k - z^k + \frac{1}{\rho}u^k)\right).$$

In many scenarios, it is challenging to compute the gradient ∇f at a given point. To cope with this issue, a stochastic gradient estimator $\widetilde{\nabla} f$ can be used to approximate the full gradient ∇f. Notice that h is a weakly convex function in the setting of problem (1). Since the Moreau envelope of a weakly convex function is smooth under certain conditions (shown in Subsection 2.1) and the critical points of H are closely related to those of H_μ (see Lemma 1), we consider the following Algorithm 1, where $\widetilde{\nabla} f(x^k)$ is a stochastic approximation to $\nabla f(x^k)$ and h_μ is the μ-Moreau envelope of h. Let $\xi^k := \widetilde{\nabla} f(x^k) - \nabla f(x^k)$ denote the gradient noise at x^k. We assume that the random vector ξ^k has zero mean conditioned on x^k, i.e., $\mathbb{E}[\xi^k|x^k] = 0$, and its covariance matrix is given by $\Sigma(x^k) := \mathbb{E}[\xi^k(\xi^k)^T|x^k]$. Clearly, there exists a function $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ such that $\Sigma(x^k) = \sigma(x^k)\sigma(x^k)^T$.

Algorithm 1 SPLA

Initialization: Choose initial point $(x^0, z^0, u^0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$, parameters $\eta, \rho > 0$ and $\tau > \rho\|A\|_{\mathrm{F}}^2 + 1/\eta$.

for $k = 0, 1, 2, \ldots$ **do**

Update x^k, z^k, u^k as follows:

$$x^{k+1} = x^k - \frac{1}{\tau}\left(\widetilde{\nabla} f(x^k) + \rho A^T(Ax^k - z^k + \frac{1}{\rho}u^k)\right), \tag{8a}$$

$$z^{k+1} = \arg\min_{z\in\mathbb{R}^m} \left\{ h_\mu(z) + \langle u^k, Ax^{k+1} - z\rangle + \frac{\rho}{2}\|Ax^{k+1} - z\|^2 \right\}, \tag{8b}$$

$$u^{k+1} = u^k + Ax^{k+1} - z^{k+1}. \tag{8c}$$

end for

Under Assumption 1, we consider the following SDE:

$$d\mathbf{x}(t) = -\frac{1}{\lambda}\nabla H_\mu(\mathbf{x}(t))\,dt + \frac{1}{\lambda\sqrt{\rho}}\sigma(\mathbf{x}(t))\,dW(t), \quad \text{with } \mathbf{x}(0) = x^0, \tag{9}$$

where $\lambda > \|A\|_{\mathrm{F}}^2$ and $W(t)$ is an m-dimensional Brownian motion. In Appendix B, we provide a derivation to obtain (9) with a sufficiently large ρ in an informal yet insightful way. We now state an assumption on the diffusion term σ. This mild assumption, also adopted in [38], guarantees the well-posedness of the SDE (9) and will be maintained throughout the remainder of the paper.

Assumption 2 *The function $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ is L_σ-Lipschitz continuous, and there exists a constant $\bar{\sigma} > 0$ such that $\|\sigma(x)\|_{\mathrm{F}} \le \bar{\sigma}$ for any $x \in \mathbb{R}^n$.*

Remark 2 *Assumptions 1 and 2 ensure the Lipschitz continuity of both $\nabla H_\mu(x)$ and $\sigma(x)$. Hence, following Theorem 7 in Appendix A, we can guarantee the existence and uniqueness of the solution $\mathbf{x}(t)$ of (9), as well as $\mathbb{E}\left[\sup_{t\in[0,T]} \|\mathbf{x}(t)\|^2\right] < +\infty$ for every $T > 0$.*

The continuous-time SDE (9) is motivated by the structure of the proposed SPLA. Specifically, it can be viewed as a stochastic perturbation of the continuous-time dynamical system associated with the deterministic proximal linearized ADMM. Its drift term captures the linearized proximal structure of SPLA, while diffusion term models the stochastic gradient noise. Under certain smoothness assumptions, the solution of (9) provides a first-order weak approximation of the discrete iterates of SPLA. The formal description is shown rigorously in Proposition 1.

In the sequel, we analyze the convergence behavior of SPLA through its associated SDE (9). Throughout the remainder of the paper, we will assume that Assumptions 1 and 2 hold and let $\mathbf{x}$ be a solution of the SDE (9).

4 Convergence Properties of $H_\mu(\mathbf{x}(t))$

In this section we will study the convergence properties of the smoothed objective function H_μ along the solution trajectory $\mathbf{x}(t)$.

4.1 Almost-Sure Convergence of $H_\mu(\mathbf{x}(t))$

To proceed, we first define

$$U_t := \int_t^\infty \|\sigma(\mathbf{x}(s))\|_{\mathrm{F}}^2 \, ds, \quad M_t := \int_0^t \langle \sigma(\mathbf{x}(s))^T \nabla H_\mu(\mathbf{x}(s)), dW(s)\rangle,$$

and introduce

$$\mathcal{L}(\mathbf{x}(t)) := H_\mu(\mathbf{x}(t)) + \frac{L}{2\lambda^2\rho} U_t - \frac{1}{\lambda\sqrt{\rho}} M_t.$$

The following lemma characterizes the behavior of $\mathcal{L}(\mathbf{x}(t))$.

Lemma 2 *For any $t \ge t_0 \ge 0$, it holds that*

$$\mathcal{L}(\mathbf{x}(t)) \le \mathcal{L}(\mathbf{x}(t_0)) - \frac{1}{\lambda}\int_{t_0}^t \|\nabla H_\mu(\mathbf{x}(s))\|^2 \, ds \quad a.s. \tag{10}$$

Proof. Proof. Note that (9) corresponds to (2) with $F(t, \mathbf{x}(t)) = -\frac{1}{\lambda}\nabla H_\mu(\mathbf{x}(t))$ and $G(t, \mathbf{x}(t)) = \frac{1}{\lambda\sqrt{\rho}}\sigma(\mathbf{x}(t))$. Thus following (4) with $\phi = H_\mu$ and by the L-smoothness of H_μ, we have

$$dH_\mu(\mathbf{x}(t)) \le -\frac{1}{\lambda}\|\nabla H_\mu(\mathbf{x}(t))\|^2 \, dt + \frac{L}{2\lambda^2\rho}\|\sigma(\mathbf{x}(t))\|_{\mathrm{F}}^2 \, dt$$

$$+\frac{1}{\lambda\sqrt{\rho}}\langle\sigma(\mathbf{x}(t))^T\nabla H_\mu(\mathbf{x}(t)),dW(t)\rangle. \tag{11}$$

By integrating the above inequality, it follows that for any $t\geq t_0\geq 0$,

$$\begin{aligned}H_\mu(\mathbf{x}(t))\leq H_\mu(\mathbf{x}(t_0))&-\frac{1}{\lambda}\int_{t_0}^{t}\|\nabla H_\mu(\mathbf{x}(s))\|^2\,ds\\&+\frac{L}{2\lambda^2\rho}\int_{t_0}^{t}\|\sigma(\mathbf{x}(s))\|_{\mathrm{F}}^2\,ds+\frac{1}{\lambda\sqrt{\rho}}\int_{t_0}^{t}\langle\sigma(\mathbf{x}(s))^T\nabla H_\mu(\mathbf{x}(s)),dW(s)\rangle\quad\text{a.s.}\end{aligned} \tag{12}$$

Rearranging it yields (10). □

Building on the descent property of $\mathcal{L}(\mathbf{x}(t))$, we can prove the in-expectation convergence of both $\mathcal{L}(\mathbf{x}(t))$ and $H_\mu(\mathbf{x}(t))$.

Theorem 1 *Suppose that $\mathbb{E}[U_0]<+\infty$, then*

(i) $\mathbb{E}[\mathcal{L}(\mathbf{x}(t))]$ *converges to a finite value, denoted by $\bar{\mathcal{L}}$, as $t\to\infty$;*

(ii) $\mathbb{E}[H_\mu(\mathbf{x}(t))]$ *converges to $\bar{\mathcal{L}}$, as $t\to\infty$;*

(iii) $\mathbb{E}[\int_0^\infty\|\nabla H_\mu(\mathbf{x}(s))\|^2\,ds]<+\infty$.

Proof. Proof. By Assumption 2, it follows that for any $0<T<\infty$,

$$\mathbb{E}\left[\int_0^T\|\sigma(\mathbf{x}(s))^T\nabla H_\mu(\mathbf{x}(s))\|^2\,ds\right]\leq\bar{\sigma}^2\mathbb{E}\left[\int_0^T\|\nabla H_\mu(\mathbf{x}(s))\|^2\,ds\right].$$

According to the L-Lipschitz continuity of ∇H_μ, we have $\|\nabla H_\mu(\mathbf{x}(t))\|^2\leq 2L^2\|\mathbf{x}(t)-x\|^2+2\|\nabla H_\mu(x)\|^2$ for any finite $x\in\mathbb{R}^n$. It has been shown in Remark 2 that the solution $\mathbf{x}(t)$ satisfies $\mathbb{E}[\sup_{t\in[0,T]}\|\mathbf{x}(t)\|^2]<+\infty$ for every $T>0$, therefore, we derive

$$\mathbb{E}\left[\int_0^T\|\sigma(\mathbf{x}(s))^T\nabla H_\mu(\mathbf{x}(s))\|^2\,ds\right]<+\infty,$$

which indicates that M_t is a Itô's integral. Then by the property of Itô's integral presented in Subsection 2.2, it holds that $\mathbb{E}[M_t]=0$ for any $t\in[0,T]$, hence,

$$\mathbb{E}[\mathcal{L}(\mathbf{x}(t))]=\mathbb{E}[H_\mu(\mathbf{x}(t))]+\frac{L}{2\lambda^2\rho}\mathbb{E}[U_t]. \tag{13}$$

Then it follows from Lemma 2 that for any $t\geq t_0\geq 0$,

$$\mathbb{E}[\mathcal{L}(\mathbf{x}(t))]\leq\mathbb{E}[\mathcal{L}(\mathbf{x}(t_0))]-\frac{1}{\lambda}\mathbb{E}\left[\int_{t_0}^{t}\|\nabla H_\mu(\mathbf{x}(s))\|^2\,ds\right], \tag{14}$$

which, together with Assumption 1(i) and $\mathbb{E}[U_0]\geq 0$, indicates that $\mathbb{E}[\mathcal{L}(\mathbf{x}(t))]$ is nonincreasing and bounded from below. Thus, it converges to some finite value, defined as $\bar{\mathcal{L}}$. This, together with $\lim_{t\to\infty}\mathbb{E}[U_t]=0$, gives $\lim_{t\to\infty}\mathbb{E}[H_\mu(\mathbf{x}(t))]=\bar{\mathcal{L}}$. By letting $t\to\infty$ in (14), we further derive

$$\mathbb{E}\left[\int_0^\infty\|\nabla H_\mu(\mathbf{x}(s))\|^2\,ds\right]<+\infty. \tag{15}$$

The proof is completed. □

Let $N_t := \int_0^t \sigma(\mathbf{x}(s))\, dW(s)$. It follows from [6, Propositions 3.19 and 3.20], together with $\mathbb{E}[U_0] < +\infty$ and (15), that N_t and M_t are continuous martingales and satisfy

$$\mathbb{E}[\|N_t\|^2] = \mathbb{E}\left[\int_0^t \|\sigma(\mathbf{x}(s))\|_{\mathrm{F}}^2\, ds\right] \le \mathbb{E}\left[\int_0^\infty \|\sigma(\mathbf{x}(s))\|_{\mathrm{F}}^2\, ds\right] < +\infty, \tag{16}$$

$$\mathbb{E}[|M_t|^2] = \mathbb{E}\left[\int_0^t \|\sigma(\mathbf{x}(s))^T \nabla H_\mu(\mathbf{x}(s))\|^2\, ds\right] \le \bar{\sigma}^2 \mathbb{E}\left[\int_0^\infty \|\nabla H_\mu(\mathbf{x}(s))\|^2\, ds\right] < +\infty. \tag{17}$$

Thus, $\sup_{t\ge 0} \mathbb{E}[\|N_t\|^2] < +\infty$ and $\sup_{t\ge 0} \mathbb{E}[|M_t|^2] < +\infty$. Then by Theorem 8 in Appendix A, there exist random variables N_∞ and M_∞ satisfying $\mathbb{E}[\|N_\infty\|^2] < +\infty$, $\mathbb{E}[|M_\infty|^2] < +\infty$ such that $N_t \to N_\infty$ a.s. and $M_t \to M_\infty$ a.s. Moreover, by the almost-sure convergence of N_t and M_t, [6, Proposition 3.20(ii)] guarantees that

$$\mathbb{E}\left[\sup_{t\ge 0} \|N_t\|^2\right] < +\infty \quad \text{and} \quad \mathbb{E}\left[\sup_{t\ge 0} |M_t|^2\right] < +\infty. \tag{18}$$

Building on Theorem 1 and the aforementioned properties of N_t and M_t, we now establish the almost-sure convergence of $H_\mu(\mathbf{x}(t))$, as well as the asymptotic vanishing of its gradient.

Theorem 2 *Suppose that $\mathbb{E}[U_0] < +\infty$. Then, $\|\nabla H_\mu(\mathbf{x}(t))\|$ converges to zero almost surely, and $H_\mu(\mathbf{x}(t))$ converges almost surely.*

Proof. Proof. From (15), we derive

$$\int_0^\infty \|\nabla H_\mu(\mathbf{x}(s))\|^2\, ds < +\infty \quad \text{a.s.} \tag{19}$$

This implies that there exists an event $\mathcal{A}_1 \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}_1) = 1$ such that for every $\omega \in \mathcal{A}_1$, $\int_0^\infty \|\nabla H_\mu(\mathbf{x}(s,\omega))\|^2\, ds < +\infty$. Hence, we have $\liminf_{t\to\infty} \|\nabla H_\mu(\mathbf{x}(t,\omega))\| = 0$. To complete the proof, it suffices to prove $\limsup_{t\to\infty} \|\nabla H_\mu(\mathbf{x}(t,\omega))\| = 0$. On the contrary, we assume that $\limsup_{t\to\infty} \|\nabla H_\mu(\mathbf{x}(t,\omega))\| > 0$, then there exists a constant $\delta > 0$ such that

$$\liminf_{t\to\infty} \|\nabla H_\mu(\mathbf{x}(t,\omega))\| < \delta < \limsup_{t\to\infty} \|\nabla H_\mu(\mathbf{x}(t,\omega))\|.$$

By the definition of lim sup, there exists an increasing sequence $\{t_k\}$ with $\lim_{k\to\infty} t_k = \infty$ such that $\|\nabla H_\mu(\mathbf{x}(t_k,\omega))\| > \delta$ for all k. Moreover, for any $\epsilon_0 > 0$, we can extract a subsequence $\{t_{k_j}\}$ of $\{t_k\}$ such that $t_k - t_{k_{j-1}} > \epsilon_0$ for all $j \ge 1$. Without loss of generality, we fix $\epsilon_0 = 1$ and relabel this subsequence as $\{t_k\}$. Thus, there exists an increasing sequence $\{t_k\}$ such that $\|\nabla H_\mu(\mathbf{x}(t_k,\omega))\| > \delta$ and $t_k - t_{k-1} > 1$ for all k.

Since $N_t \to N_\infty$ a.s., there exists an event $\mathcal{A}_2 \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}_2) = 1$ such that $N_t(\omega) \to N_\infty(\omega)$ for every $\omega \in \mathcal{A}_2$, which, together with $\int_0^\infty \|\nabla H_\mu(\mathbf{x}(s,\omega))\|^2\, ds < +\infty$ for every $\omega \in \mathcal{A}_1$, implies that for every $\omega \in \mathcal{A}_1 \cap \mathcal{A}_2$ and any $\epsilon_1, \epsilon_2 > 0$, there exists $K > 0$ such that for any $k > K$ and $t > t_k$,

$$\|N_t(\omega) - N_{t_k}(\omega)\|^2 < \frac{\epsilon_1 \lambda^2 \rho}{4} \quad \text{and} \quad \int_{t_k}^\infty \|\nabla H_\mu(\mathbf{x}(s,\omega))\|^2\, ds < \frac{\epsilon_2 \lambda^2}{2}.$$

Fix ϵ_2 and choose ϵ_1 such that $\epsilon_1 < \min\{\frac{\delta^2}{4L^2}, \epsilon_2\}$. Note that the intervals $[t_k, t_k + \frac{\epsilon_1}{2\epsilon_2}], k \ge 0$ are disjoint. From (9), there exists an event $\mathcal{A}_3 \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}_3) = 1$ such that for every $\omega \in \mathcal{A}_3$,

$$\|\mathbf{x}(t,\omega) - \mathbf{x}(t_k,\omega)\| \le \frac{1}{\lambda} \int_{t_k}^t \|\nabla H_\mu(\mathbf{x}(s,\omega))\|\, ds + \frac{1}{\lambda\sqrt{\rho}} \|N_t(\omega) - N_{t_k}(\omega)\|.$$

Therefore, for every $\omega \in \mathcal{A} := \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$ with $\mathbb{P}(\mathcal{A}) = 1$, for any $k > K$ and $t \in [t_k, t_k + \frac{\epsilon_1}{2\epsilon_2}]$, we have

$$\|\mathbf{x}(t,\omega) - \mathbf{x}(t_k,\omega)\|^2 \leq \frac{2}{\lambda^2}(t - t_k)\int_{t_k}^{t} \|\nabla H_\mu(\mathbf{x}(s,\omega))\|^2\, ds + \frac{2}{\lambda^2\rho}\|N_t(\omega) - N_{t_k}(\omega)\|^2 < \epsilon_1,$$

which, together with the L-smoothness of H_μ and $\epsilon_1 < \min\{\frac{\delta^2}{4L^2}, \epsilon_2\}$, ensures that

$$\|\nabla H_\mu(\mathbf{x}(t,\omega)) - \nabla H_\mu(\mathbf{x}(t_k,\omega))\|^2 \leq L^2\|\mathbf{x}(t,\omega) - \mathbf{x}(t_k,\omega)\|^2 < L^2\epsilon_1 < \frac{\delta^2}{4}.$$

Combining the above inequality and $\|\nabla H_\mu(\mathbf{x}(t_k,\omega))\| > \delta$ follows that

$$\|\nabla H_\mu(\mathbf{x}(t,\omega))\| \geq \|\nabla H_\mu(\mathbf{x}(t_k,\omega))\| - \|\nabla H_\mu(\mathbf{x}(t,\omega)) - \nabla H_\mu(\mathbf{x}(t_k,\omega))\| > \frac{\delta}{2}.$$

Then, for every $\omega \in \mathcal{A}$,

$$\int_0^\infty \|\nabla H_\mu(\mathbf{x}(s,\omega))\|^2\, ds \geq \sum_{k>K}\int_{t_k}^{t_k+\frac{\epsilon_1}{2\epsilon_2}} \|\nabla H_\mu(\mathbf{x}(s,\omega))\|^2\, ds > \sum_{k>K}\frac{\delta^2\epsilon_1}{8\epsilon_2} = +\infty,$$

which contradicts with (19). Hence, for every $\omega \in \mathcal{A}$,

$$\limsup_{t\to\infty} \|\nabla H_\mu(\mathbf{x}(t,\omega))\| = \liminf_{t\to\infty} \|\nabla H_\mu(\mathbf{x}(t,\omega))\| = 0,$$

which indicates that $\lim_{t\to\infty} \|\nabla H_\mu(\mathbf{x}(t))\| = 0$ holds almost surely.

From (10), there exists an event $\mathcal{A}_4 \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}_4) = 1$ such that for every $\omega \in \mathcal{A}_4$ and for any $t \geq t_0 \geq 0$,

$$\mathcal{L}(\mathbf{x}(t,\omega)) \leq \mathcal{L}(\mathbf{x}(t_0,\omega)) - \frac{1}{\lambda}\int_{t_0}^{t} \|\nabla H_\mu(\mathbf{x}(s,\omega))\|^2\, ds. \tag{20}$$

This implies that $\mathcal{L}(\mathbf{x}(t,\omega))$ is nonincreasing. From (18), it holds that $\sup_{t\geq 0}|M_t|^2 < +\infty$ a.s. Thus, there exists an event $\mathcal{A}_5 \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}_5) = 1$ such that for every $\omega \in \mathcal{A}_5$, $M_t(\omega)$ is bounded. Take $\bar{\mathcal{A}} := \mathcal{A}_4 \cap \mathcal{A}_5$. Then, we derive that for every $\omega \in \bar{\mathcal{A}}$ with $\mathbb{P}(\bar{\mathcal{A}}) = 1$, $\mathcal{L}(\mathbf{x}(t,\omega))$ is nonincreasing and bounded from below. Thus, $\mathcal{L}(\mathbf{x}(t,\omega))$ converges for every $\omega \in \bar{\mathcal{A}}$, which further indicates that $\mathcal{L}(\mathbf{x}(t))$ converges almost surely. Since $M_t \to M_\infty$ a.s. and $U_t \to 0$ a.s., it follows that $H_\mu(\mathbf{x}(t))$ converges almost surely. □

Remark 3 *Following the proof of Theorem 2, $\mathcal{L}(\mathbf{x}(t))$ and $H_\mu(\mathbf{x}(t))$ converge almost surely, i.e., there exist random variables $\mathcal{L}_\infty$ and $H_{\mu,\infty}$ such that $\mathcal{L}(\mathbf{x}(t)) \to \mathcal{L}_\infty$ a.s. and $H_\mu(\mathbf{x}(t)) \to H_{\mu,\infty}$ a.s. and $\mathcal{L}_\infty = H_{\mu,\infty} - \frac{1}{\lambda\sqrt{\rho}}M_\infty$.*

We have established the in-expectation convergence and the almost-sure convergence of $H_\mu(\mathbf{x}(t))$ in Theorems 1 and 2, respectively. As a direct consequence, we conclude that

$$\lim_{t\to\infty} \mathbb{E}[H_\mu(\mathbf{x}(t))] = \mathbb{E}[\lim_{t\to\infty} H_\mu(\mathbf{x}(t))]. \tag{21}$$

This follows from the Dominated Convergence Theorem (see [18, 47]), for which we only need to verify the integrability condition. From (12), it holds that

$$H_\mu(\mathbf{x}(t)) \leq H_\mu(\mathbf{x}(0)) + \frac{L}{2\lambda^2\rho}U_0 + \frac{1}{\lambda\sqrt{\rho}}\sup_{t\geq 0}|M_t - M_0| \quad \text{a.s.} \tag{22}$$

Moreover, by $\mathbb{E}[U_0] < +\infty$ and

$$\mathbb{E}[\sup_{t\geq 0}|M_t - M_0|] \leq \left(\mathbb{E}[\sup_{t\geq 0}|M_t - M_0|^2]\right)^{1/2} < +\infty$$

deduced from (18), the right-hand side of (22) is integrable. Consequently, by the almost-sure convergence of $H_\mu(\mathbf{x}(t))$ to $H_{\mu,\infty}$, the Dominated Convergence Theorem ensures that $\mathbb{E}[H_\mu(\mathbf{x}(t))] \to \mathbb{E}[H_{\mu,\infty}] = \bar{\mathcal{L}}$, and thus (21) follows.

4.2 Convergence Rate of $H_\mu(\mathbf{x}(t))$ under Łojasiewicz Inequality

In this subsection, we assume that $\mathbf{x}$ is an almost surely bounded solution of (9), i.e., $\sup_{t\geq 0}\|\mathbf{x}(t)\| < +\infty$ a.s., which is standard in the convergence analysis of nonconvex optimization algorithms (see [7, 22, 17, 33] for instance).

For any fixed sample $\omega \in \Omega$, we now define the set consisting of cluster points of the $\mathbf{x}(t,\omega)$ as follows:

$$\mathcal{C}_\omega := \{x_\infty(\omega) \in \mathbb{R}^n : \exists \text{ an increasing sequence } \{t_k\} \text{ such that } \mathbf{x}(t_k,\omega) \to x_\infty(\omega) \text{ as } k \to \infty\}.$$

Lemma 3 *Suppose that $\mathbf{x}$ is almost surely bounded and $\mathbb{E}[U_0] < +\infty$. Then there exists an event $\mathcal{A} \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}) = 1$ such that for all $\omega \in \mathcal{A}$, the following statements hold:*

(i) The set $\mathcal{C}_\omega$ is nonempty, compact and

$$\lim_{t\to\infty} \mathrm{dist}((\mathbf{x}(t,\omega)), \mathcal{C}_\omega) = 0.$$

(ii) H_μ is finite and constant on $\mathcal{C}_\omega$.

(iii) $\mathcal{C}_\omega \subseteq \mathrm{crit}H_\mu$.

Proof. Proof. Since $\mathbf{x}$ is almost surely bounded, there exists an event $\mathcal{A} \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}) = 1$, such that for every $\omega \in \mathcal{A}$, $\mathbf{x}(\cdot,\omega)$ is bounded. This implies that $\mathcal{C}_\omega$ is nonempty and bounded. Using a similar proof as in [13, Lemma 3.3(vii)], it follows that $\mathcal{C}_\omega$ is compact. From the definition of cluster points, we obtain $\mathrm{dist}((\mathbf{x}(t,\omega)), \mathcal{C}_\omega) \to 0$ as $t \to \infty$, which completes the proof of item (i).

For any $x_\infty(\omega) \in \mathcal{C}_\omega$, by the definition of cluster points, there exists an increasing sequence $\{t_k\}$ such that $\mathbf{x}(t_k,\omega) \to x_\infty(\omega)$. According to the continuity of H_μ, it follows that $H_\mu(\mathbf{x}(t_k,\omega)) \to H_\mu(x_\infty(\omega))$ as $t_k \to \infty$. It has been shown in Theorem 2 that $H_\mu(\mathbf{x}(t)) \to H_{\mu,\infty}$ a.s., then $H_\mu(x_\infty(\omega)) = H_{\mu,\infty}(\omega)$ for every $\omega \in \mathcal{A}$. Hence H_μ is finite and constant on $\mathcal{C}_\omega$.

We next show that for any $x_\infty(\omega) \in \mathcal{C}_\omega$, it holds that $x_\infty(\omega) \in \mathrm{crit}H_\mu$, i.e., $\nabla H_\mu(x_\infty(\omega)) = 0$. Since ∇H_μ is continuous, then $\|\nabla H_\mu(\mathbf{x}(t_k,\omega))\| \to \nabla H_\mu(x_\infty(\omega))$ due to $\mathbf{x}(t_k,\omega) \to x_\infty(\omega)$. Recall that in Theorem 2 we have shown $\|\nabla H_\mu(\mathbf{x}(t))\| \to 0$ a.s., which implies that, for all $\omega \in \mathcal{A}$, $\|\nabla H_\mu(\mathbf{x}(t,\omega))\| \to 0$. Together with $\|\nabla H_\mu(\mathbf{x}(t_k,\omega))\| \to \nabla H_\mu(x_\infty(\omega))$, it further yields $\nabla H_\mu(x_\infty(\omega)) = 0$. The proof is completed. □

To analyze the convergence rate of $H_\mu(\mathbf{x}(t))$, we first provide the definition of the Łojasiewicz inequality. The concept of Łojasiewicz inequality was originally introduced in seminal works such as [36, 28], and has been further developed and extended in the context of nonconvex optimization, including [10, 1, 2]. Below, we present a slightly stronger variant of the classical Łojasiewicz inequality, as stated in [33, Definition 2.1].

Definition 1 *A smooth function $f:\mathbb{R}^n\to\mathbb{R}$ is said to satisfy Łojasiewicz inequality at a point $\bar{x}\in\mathbb{R}^n$, if there exist constants $\eta\in(0,+\infty]$, $\theta\in(0,1)$, $\varsigma>0$, and a neighborhood U of $\bar{x}$ such that for all $x\in U\cap\{x:0<|f(x)-f(\bar{x})|<\eta\}$, the following inequality holds:*

$$\varsigma|f(x)-f(\bar{x})|^{-\theta}\cdot\|\nabla f(x)\|\geq 1. \tag{23}$$

Remark 4 *The Łojasiewicz inequality stated in Definition 1 holds for semialgebraic functions and real subanalytic functions (see [8, 9, 10]). In the context of the problem considered in this paper, the function H_μ satisfies the Łojasiewicz inequality at all point $x\in\mathbb{R}^n$ provided that both f and h are semialgebraic, see [2, Section 4].*

Prior to applying the Łojasiewicz inequality in our convergence analysis, we observe from Lemma 3 that the Łojasiewicz inequality can be reformulated in a uniformized version as follows. This result has been established in [33, Lemma 3.5].

Lemma 4 *Let $\mathcal{C}$ be a compact set and let $f:\mathbb{R}^n\to\mathbb{R}$ be a smooth function. Assume that f is constant on $\mathcal{C}$ and satisfies the Łojasiewicz inequality at each point of $\mathcal{C}$. Then, there exist $\epsilon,\eta>0$, $\theta\in(0,1)$, $\varsigma>0$, such that for all $\bar{x}\in\mathcal{C}$ and $x\in\{x\in\mathbb{R}^n:\operatorname{dist}(x,\mathcal{C})<\epsilon\}\cap\{x\in\mathbb{R}^n:0<|f(x)-f(\bar{x})|<\eta\}$, one has $\varsigma|f(x)-f(\bar{x})|^{-\theta}\cdot\|\nabla f(x)\|\geq 1$.*

Now, we prove the in-expectation convergence rates of $H_\mu(\mathbf{x}(t))$ to $H_{\mu,\infty}$ based on Lemma 4 and Theorem 2.

Theorem 3 *Suppose that $\mathbf{x}$ is almost surely bounded and $\mathbb{E}[U_0]<+\infty$. Denote by $H_{\mu,\infty}$ the almost-sure limit of $H_\mu(\mathbf{x}(t))$. If there exists an event $\mathcal{A}\in\mathcal{F}$ with $\mathbb{P}(\mathcal{A})=1$ such that H_μ satisfies the Łojasiewicz inequality at every $x\in\bigcup_{\omega\in\mathcal{A}}\mathcal{C}_\omega$, then the following statements hold true:*

(i) If $\theta\in(0,1/2]$, there exist constants $a_1,b_1,c_1>0$ and a sufficiently large $T>0$ such that

$$\mathbb{E}[|H_\mu(\mathbf{x}(t))-H_{\mu,\infty}|]\leq a_1\exp(-b_1(t-T))+\frac{c_1}{\rho},\quad\forall t>T. \tag{24}$$

(ii) If $\theta\in(1/2,1)$, there exist constants $a_2,c_2>0$ and a sufficiently large $T>0$ such that

$$\mathbb{E}[|H_\mu(\mathbf{x}(t))-H_{\mu,\infty}|]\leq a_2(t-T)^{\frac{1}{1-2\theta}}+\frac{c_2}{\rho},\quad\forall t>T. \tag{25}$$

Proof. Proof. We have shown in Theorem 2 that $H_\mu(\mathbf{x}(t))\to H_{\mu,\infty}$ a.s., which implies that there exists an event $\mathcal{A}_0\in\mathcal{F}$ with $\mathbb{P}(\mathcal{A}_0)=1$ such that for every $\omega\in\mathcal{A}_0$, $H_\mu(\mathbf{x}(t,\omega))\to H_{\mu,\infty}(\omega)$. Hence, for any $\eta>0$, there exists a time $T_0>0$ such that $|H_\mu(\mathbf{x}(t,\omega))-H_{\mu,\infty}(\omega)|<\eta$ for any $t\geq T_0$. Moreover, by Lemma 3(i), for any $\epsilon>0$, there exists $T_1>0$ such that $\operatorname{dist}(\mathbf{x}(t,\omega),\mathcal{C}_\omega)<\epsilon$ for any $t\geq T_1$ and every $\omega\in\mathcal{A}$. By the above discussion, for every $\omega\in\mathcal{A}_0\cap\mathcal{A}$, we have $\mathbf{x}(t,\omega)\in\{x\in\mathbb{R}^n:\operatorname{dist}(x,\mathcal{C}_\omega)<\epsilon\}\cap\{x:0<|H_\mu(x)-H_{\mu,\infty}(\omega)|<\eta\}$ for all $t\geq T_2:=\max\{T_0,T_1\}$. Since H_μ is constant on $\mathcal{C}_\omega$ with value $H_{\mu,\infty}(\omega)$ and satisfies the Łojasiewicz inequality with $\theta\in(0,1)$, by Lemma 4 we obtain

$$\|\nabla H_\mu(\mathbf{x}(t,\omega))\|^2\geq\frac{1}{\varsigma^2}|H_\mu(\mathbf{x}(t,\omega))-H_{\mu,\infty}(\omega)|^{2\theta},\quad\forall t\geq T_2, \tag{26}$$

which indicates

$$\mathbb{E}[\|\nabla H_\mu(\mathbf{x}(t))\|^2]\geq\frac{1}{\varsigma^2}\mathbb{E}[|H_\mu(\mathbf{x}(t))-H_{\mu,\infty}|^{2\theta}],\quad\forall t\geq T_2. \tag{27}$$

Substituting it into (14) gives that

$$\mathbb{E}[\mathcal{L}(\mathbf{x}(t))] \leq \mathbb{E}[\mathcal{L}(\mathbf{x}(T_2))] - \frac{1}{\varsigma^2\lambda}\int_{T_2}^{t}\mathbb{E}[|H_\mu(\mathbf{x}(s)) - H_{\mu,\infty}|^{2\theta}]\,ds. \tag{28}$$

Following Theorem 1 and (21), it holds that $\bar{\mathcal{L}} = \mathbb{E}[H_{\mu,\infty}]$, which, together with (13), yields

$$\mathbb{E}[\mathcal{L}(\mathbf{x}(t))] - \bar{\mathcal{L}} \leq \mathbb{E}[|H_\mu(\mathbf{x}(t)) - H_{\mu,\infty}|] + \frac{L}{2\lambda^2\rho}\mathbb{E}[U_t]. \tag{29}$$

For any $\theta \in [1/2, 1)$, we have

$$\begin{aligned}\big(\mathbb{E}[\mathcal{L}(\mathbf{x}(t))] - \bar{\mathcal{L}}\big)^{2\theta} &\leq \left(\mathbb{E}[|H_\mu(\mathbf{x}(t)) - H_{\mu,\infty}|] + \frac{L}{2\lambda^2\rho}\mathbb{E}[U_t]\right)^{2\theta}\\ &\leq 2\mathbb{E}[|H_\mu(\mathbf{x}(t)) - H_{\mu,\infty}|]^{2\theta} + \frac{2L^{2\theta}}{(2\lambda^2\rho)^{2\theta}}\mathbb{E}[U_t]^{2\theta}\\ &\leq 2\mathbb{E}[|H_\mu(\mathbf{x}(t)) - H_{\mu,\infty}|^{2\theta}] + \frac{2L^{2\theta}}{(2\lambda^2\rho)^{2\theta}}\mathbb{E}[U_t]^{2\theta},\end{aligned} \tag{30}$$

where the second inequality is obtained by the fact that $|a+b|^\theta \leq |a|^\theta + |b|^\theta$ and $(a+b)^2 \leq 2a^2 + 2b^2$, the last inequality is deduced from Jensen's inequality $\mathbb{E}[|\xi|] \leq \mathbb{E}[|\xi|^p]^{1/p}$ for $p \geq 1$. Combining inequalities (28) and (30), we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\mathbf{x}(t))] - \bar{\mathcal{L}} \leq \mathbb{E}[\mathcal{L}(\mathbf{x}(T_2))] - \bar{\mathcal{L}} &- \frac{1}{2\varsigma^2\lambda}\int_{T_2}^{t}(\mathbb{E}[\mathcal{L}(\mathbf{x}(s))] - \bar{\mathcal{L}})^{2\theta}\,ds\\ &+ \frac{L^{2\theta}}{2^{2\theta}\varsigma^2\lambda^{4\theta+1}\rho^{2\theta}}\int_{T_2}^{t}\mathbb{E}[U_s]^{2\theta}\,ds.\end{aligned} \tag{31}$$

Since the condition $\mathbb{E}[U_0] < +\infty$ holds, there exists a sufficiently large $T_3 \geq T_2$ such that $\mathbb{E}[U_s] < 1$ for any $s \geq T_3$. Therefore, we obtain $\mathbb{E}[U_s]^{2\theta} \leq \mathbb{E}[U_s]$ for $s \geq T_3$ from $\theta \in [1/2, 1)$, and

$$\mathbb{E}[\mathcal{L}(\mathbf{x}(t))] - \bar{\mathcal{L}} \geq \mathbb{E}[|H_\mu(\mathbf{x}(t)) - H_{\mu,\infty}|] - \frac{L}{2\lambda^2\rho}, \quad \forall t \geq T_3 \tag{32}$$

from (13). By letting $\mathscr{L}(t) := \mathbb{E}[\mathcal{L}(\mathbf{x}(t))] - \bar{\mathcal{L}}$, it follows from (31) that for any $t \geq T_3 > 0$,

$$\mathscr{L}(t) \leq \mathscr{L}(T_3) - \frac{1}{2\varsigma^2\lambda}\int_{T_3}^{t}\mathscr{L}(s)^{2\theta}\,ds + \frac{L^{2\theta}}{2^{2\theta}\varsigma^2\lambda^{4\theta+1}\rho^{2\theta}}(t - T_3). \tag{33}$$

Case (i): $\theta \in (0, 1/2]$. In this case, the proof can be reduced to the one corresponding to $\theta = 1/2$. Indeed, since $H_\mu(\mathbf{x}(t,\omega)) \to H_{\mu,\infty}(\omega)$ as $t \to +\infty$, there exists a sufficiently large $T > 0$ such that $|H_\mu(\mathbf{x}(t,\omega)) - H_{\mu,\infty}(\omega)| < 1$ for all $t \geq T$. Consequently, if (26) holds for $\theta \in (0, 1/2)$, it also remains valid for $\theta = 1/2$. Therefore, to prove (i) it suffices to consider the specific case with $\theta = 1/2$. When $\theta = 1/2$, from (33), we take

$$b = \frac{1}{2\varsigma^2\lambda}, \quad c = \frac{L}{2\varsigma^2\lambda^3\rho}.$$

According to Lemma 5 in Appendix A with $g(t) = f(t) = \mathscr{L}(t)$ and $t_0 = T_3$, it holds that

$$\mathscr{L}(t) \leq \mathscr{L}(T_3)\exp(-b(t - T_3)) + \frac{c}{b}. \tag{34}$$

Furthermore, by (32) and the definitions of b, c, we have

$$\mathbb{E}[|H_\mu(\mathbf{x}(t)) - H_{\mu,\infty}|] \leq \mathscr{L}(T_3)\exp(-\frac{1}{2\varsigma^2\lambda}(t-T_3)) + \frac{3L}{2\lambda^2\rho}.$$

Thus there exist a_1, b_1 and c_1 such that (24) holds.

Case (ii): $\theta \in (1/2, 1)$. From (33), we set

$$b = \frac{1}{2\varsigma^2\lambda}, \quad c = \frac{L^{2\theta}}{2^{2\theta}\varsigma^2\lambda^{4\theta+1}\rho^{2\theta}}.$$

By $\mathbb{E}[\mathcal{L}(\mathbf{x}(t))] \to \bar{\mathcal{L}}$ (shown in Theorem 1), for sufficiently large T_3, it holds that $\mathscr{L}(T_3) = \mathbb{E}[\mathcal{L}(\mathbf{x}(T_3))] - \bar{\mathcal{L}} < (\frac{c}{b})^{\frac{1}{2\theta}}$. Then, by applying Lemma 5 in Appendix A with $g(t) = f(t) = \mathscr{L}(t)$ and $t_0 = T_3$, we have

$$\mathscr{L}(t) \leq \left[\mathscr{L}(T_3)^{1-2\theta} + (2\theta-1)b(t-T_3)\right]^{\frac{1}{1-2\theta}} + \left(\frac{c}{b}\right)^{\frac{1}{2\theta}}, \tag{35}$$

which, together with (32) and the definitions of b and c, gives that

$$\mathbb{E}[|H_\mu(\mathbf{x}(t)) - H_{\mu,\infty}|] \leq \left[\mathscr{L}(T_3)^{1-2\theta} + \frac{2\theta-1}{2\varsigma^2\lambda}(t-T_3)\right]^{\frac{1}{1-2\theta}} + \frac{3L}{2\lambda^2\rho}.$$

Therefore, (25) holds for sufficiently large T. □

Corollary 1 *Under the conditions of Theorem 3, we have:*

(i) For $\theta \in (0, 1/2]$, there exist constants $a_1, b_1, c_1 > 0$ and a sufficiently large $T > 0$ such that

$$\mathbb{E}[|H(\mathbf{x}(t)) - H_{\mu,\infty}|] \leq a_1\exp(-b_1(t-T)) + \frac{c_1}{\rho} + \frac{\mu L_h^2}{2(1-\varrho\mu)}, \quad \forall t > T.$$

(ii) For $\theta \in (1/2, 1)$, there exist constants $a_2, c_2 > 0$ and a sufficiently large $T > 0$ such that

$$\mathbb{E}[|H(\mathbf{x}(t)) - H_{\mu,\infty}|] \leq a_2(t-T)^{\frac{1}{1-2\theta}} + \frac{c_2}{\rho} + \frac{\mu L_h^2}{2(1-\varrho\mu)}, \quad \forall t > T.$$

Proof. Proof. Applying the triangle inequality yields

$$\mathbb{E}[|H(\mathbf{x}(t)) - H_{\mu,\infty}|] \leq \mathbb{E}[|H_\mu(\mathbf{x}(t)) - H_{\mu,\infty}|] + \mathbb{E}[|H(\mathbf{x}(t)) - H_\mu(\mathbf{x}(t))|]. \tag{36}$$

The first term of the right-hand side of the above inequality admits the upper bounds as in Theorem 3. For the second term, from Lemma 6 in Appendix A, we have

$$0 \leq H(\mathbf{x}(t)) - H_\mu(\mathbf{x}(t)) = h(\mathbf{x}(t)) - h_\mu(A\mathbf{x}(t)) \leq \frac{\mu L_h^2}{2(1-\varrho\mu)},$$

which further derives

$$\mathbb{E}[H(\mathbf{x}(t)) - H_\mu(\mathbf{x}(t))] \leq \frac{\mu L_h^2}{2(1-\varrho\mu)}.$$

Combining the two upper bounds for the right-hand side of (36) yields the desired result. □

Note that the bias terms $\frac{c_i}{\rho}(i = 1, 2)$ and $\frac{\mu L_h^2}{2(1-\varrho\mu)}$ can be made arbitrarily small, i.e., reduced to any prescribed accuracy, by taking the penalty parameter ρ sufficiently large and the Moreau envelope parameter μ (with $0 < \mu < 1/\varrho$) sufficiently small.

5 Convergence Properties of $\mathbf{x}(t)$ and $\bar{\mathbf{x}}(t)$

In this section we will establish the convergence properties of the solution $\mathbf{x}$ to (9) and of $\bar{\mathbf{x}}$ as defined below in (37), including their almost-sure global convergence as well as in-expectation convergence rates.

Motivated by Lemma 1, to approach an ϵ-critical point of H, it is sufficient to get close to a critical point of H_μ. More specifically, if we can prove the convergence of $\mathbf{x}(t)$ towards a critical point of H_μ under certain conditions, then by applying Lemma 1 we can show that $\bar{\mathbf{x}}(t)$, defined by

$$\bar{\mathbf{x}}(t) := \mathbf{x}(t) - A^T(AA^T)^{-1}(A\mathbf{x}(t) - \mathrm{prox}_{\mu h}(A\mathbf{x}(t))), \tag{37}$$

is convergent to an ϵ-critical point of H, when $\mu \le \epsilon\sqrt{\lambda_{\min}(AA^T)}/(L_f L_h)$.

Theorem 4 *Suppose that $\mathbf{x}$ is almost surely bounded and there exist $C > 0$ and $a > 1$ such that*

$$\mathbb{E}[U_t] \le \frac{C}{(t+1)^a}, \quad \forall t \ge 0.$$

If there exists an event $\mathcal{A} \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}) = 1$ such that H_μ satisfies the Łojasiewicz inequality at every $x \in \bigcup_{\omega\in\mathcal{A}} \mathcal{C}_\omega$, then $\mathbf{x}(t)$ converges almost surely to a critical point of H_μ, and $\bar{\mathbf{x}}(t)$ converges almost surely to an ϵ-critical point of H, provided that $\mu \le \epsilon\sqrt{\lambda_{\min}(AA^T)}/(L_f L_h)$.

Proof. Proof. Note the fact that $\mathbb{E}[U_t] \le \frac{C}{(t+1)^a}$ implies $\mathbb{E}[U_0] < +\infty$. Therefore, the convergence results in Section 4 also hold true.

For $\theta \in [1/2, 1)$, note that $\mathbb{E}[U_t] \le \frac{C}{(t+1)^a}$ and $\mathbb{E}[U_t]^{2\theta} \le \mathbb{E}[U_t]$ for any $t \ge T_3$, where T_3 is defined in the proof of Theorem 4. It then follows that

$$\int_{T_3}^t \mathbb{E}[U_s]^{2\theta}\, ds \le \int_{T_3}^t \mathbb{E}[U_s]\, ds \le \int_{T_3}^t \frac{C}{(s+1)^a}\, ds \le \frac{C}{a-1}.$$

Substituting the above inequality into (31), with $\mathscr{L}(t) = \mathbb{E}[\mathcal{L}(\mathbf{x}(t))] - \bar{\mathcal{L}}$ we have

$$\mathscr{L}(t) \le \mathscr{L}(T_3) - \frac{1}{2\varsigma^2\lambda}\int_{T_3}^t \mathscr{L}(s)^{2\theta}\, ds + \frac{CL^{2\theta}}{2^{2\theta}\varsigma^2\lambda^{4\theta+1}\rho^{2\theta}(a-1)}.$$

By Lemma 5 in Appendix A, we derive that for all $t \ge T_3 > 0$,

$$\mathscr{L}(t) \le Q(t) := \begin{cases} a_1 \exp(-b_1(t - T_3)), & \text{if } \theta \in (0, 1/2], \\ a_2(t - T_3)^{\frac{1}{1-2\theta}}, & \text{if } \theta \in (1/2, 1), \end{cases} \tag{38}$$

where a_1, a_2, b_1 are positive constants.

Let

$$V_t := \int_t^\infty \mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|^2]\, ds \quad \text{and} \quad R_t := (-Q'(t))^{-1/2}.$$

It follows from (14) and (38) that

$$\frac{1}{\lambda}V_t \le \mathbb{E}[\mathcal{L}(\mathbf{x}(t))] - \bar{\mathcal{L}} = \mathscr{L}(t) \le Q(t), \quad \forall t \ge T_3 > 0. \tag{39}$$

Applying Cauchy-Schwarz inequality, we have

$$\int_t^\infty \mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|]\, ds \le \left(\int_t^\infty R_s\mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|^2]\, ds\right)^{1/2}\left(\int_t^\infty R_s^{-1}\, ds\right)^{1/2}.$$

According to integration by parts, we obtain that for all $t \geq T_3$,

$$\begin{aligned}
&\int_t^\infty R_s \mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|^2]\, ds = -\int_t^\infty R_s\, dV_s = -(R_s V_s)|_t^\infty + \int_t^\infty V_s\, dR_s \\
&\leq -(R_s V_s)|_t^\infty + \lambda \int_t^\infty Q(s)\, dR_s = -(R_s V_s)|_t^\infty + \lambda (R_s Q(s))|_t^\infty - \lambda \int_t^\infty R_s\, dQ(s) \\
&= -R_t(\lambda Q(t) - V_t) + \lim_{t\to\infty} R_t(\lambda Q(t) - V_t) - \lambda \int_t^\infty R_s\, dQ(s) \leq -\lambda \int_t^\infty R_s\, dQ(s),
\end{aligned}$$

where the first inequality follows from $V_s \leq \lambda Q(s)$ for all $s \geq t \geq T_3$, the last inequality is derived using $0 \leq V_t \leq \lambda Q(t)$ and $\lim_{t\to\infty} Q(t) = 0$. Combining the above two inequalities yields that

$$\begin{aligned}
\int_t^\infty \mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|]\, ds &\leq \left(-\lambda \int_t^\infty R_s\, dQ(s)\right)^{1/2} \left(\int_t^\infty R_s^{-1}\, ds\right)^{1/2} \\
&= \left(\lambda \int_t^\infty (-Q'(s))^{1/2}\, ds\right)^{1/2} \left(\int_t^\infty (-Q'(s))^{1/2}\, ds\right)^{1/2} \\
&= \sqrt{\lambda} \int_t^\infty (-Q'(s))^{1/2}\, ds.
\end{aligned}$$

Hence, following (38), for $\theta \in [1/2, 1)$ and all $t \geq T_3$, it holds that

$$\int_t^\infty \mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|]\, ds \leq \sqrt{\lambda} \int_t^\infty (-Q'(s))^{1/2}\, ds < +\infty, \tag{40}$$

which implies that $\int_0^\infty \mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|\, ds] < +\infty$. Further, we have

$$\int_0^\infty \|\nabla H_\mu(\mathbf{x}(s))\|\, ds < +\infty \quad \text{a.s.} \tag{41}$$

From (9), there exists an event $\mathcal{A}_1 \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}_1) = 1$ such that for every $\omega \in \mathcal{A}_1$ and for any $t_2 \geq t_1 \geq 0$,

$$\|\mathbf{x}(t_2,\omega) - \mathbf{x}(t_1,\omega)\| \leq \frac{1}{\lambda} \int_{t_1}^{t_2} \|\nabla H_\mu(\mathbf{x}(s,\omega))\|\, ds + \frac{1}{\lambda\sqrt{\rho}} \|N_{t_2}(\omega) - N_{t_1}(\omega)\|,$$

where $N_t = \int_0^t \sigma(\mathbf{x}(s))\, dW(s)$. Since N_t converges almost surely to N_∞, there exists an event $\mathcal{A}_2 \in \mathcal{F}$ with $\mathbb{P}(\mathcal{A}_2) = 1$ such that $N_t(\omega) \to N_\infty(\omega)$ for every $\omega \in \mathcal{A}_2$. Thus, for any $\epsilon > 0$, there exists $T_4 > 0$ such that $\|N_{t_2}(\omega) - N_{t_1}(\omega)\| < \frac{\epsilon\lambda\sqrt{\rho}}{2}$ for all $t_2 \geq t_1 \geq T_4$ and $\omega \in \mathcal{A}_2$. Moreover, by (41) there exists an event $\mathcal{A}_3$ with $\mathbb{P}(\mathcal{A}_3) = 1$ such that for every $\omega \in \mathcal{A}_3$, $\int_0^\infty \|\nabla H_\mu(\mathbf{x}(s,\omega))\|\, ds < +\infty$. Consequently, for any $\epsilon > 0$, there exists $T_5 > 0$ such that $\int_{t_1}^{t_2} \|\nabla H_\mu(\mathbf{x}(s,\omega))\|\, ds < \frac{\epsilon\lambda}{2}$ holds for all $t_2 \geq t_1 \geq T_5$ and $\omega \in \mathcal{A}_3$. Take $\bar{\mathcal{A}} := \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}$ and $T := \max\{T_3, T_4, T_5\}$, then for every $\omega \in \bar{\mathcal{A}}$ with $\mathbb{P}(\bar{\mathcal{A}}) = 1$, for any $t_2 \geq t_1 \geq T$, $\|\mathbf{x}(t_2,\omega) - \mathbf{x}(t_1,\omega)\| \leq \epsilon$. Hence, by the arbitrariness of ϵ and Cauchy's criterion for convergence, $\mathbf{x}(t,\omega)$ is convergent for every $\omega \in \bar{\mathcal{A}}$. It, together with Lemma 3(iii), implies that $\mathbf{x}(t,\omega)$ converges to a critical point of H_μ. Thus, $\mathbf{x}(t)$ converges almost surely to some critical point of H_μ, named as x_∞.

By the continuity of $\mathrm{prox}_{\mu h}$, it is clear that $\bar{\mathbf{x}}(t)$ converges almost surely to

$$\bar{x}_\infty := x_\infty - A^T(AA^T)^{-1}(Ax_\infty - \mathrm{prox}_{\mu h}(Ax_\infty)). \tag{42}$$

Then it follows from Lemma 1 and $\mu \leq \epsilon\sqrt{\lambda_{\min}(AA^T)}/(L_f L_h)$ that $\bar{x}_\infty \in \mathrm{crit}_\epsilon H$. □

Remark 5 *A straightforward calculation shows that the condition* $\mathbb{E}[U_t] \le C/(t+1)^a$ *with* $a > 1$ *is satisfied if* $\|\sigma(\mathbf{x}(t))\|_{\mathrm{F}} = o(1/(t+1))$. *This corresponds to the discrete-time condition* $\|\sigma(x^k)\|_{\mathrm{F}} = o(1/(k+1))$ *which implies that the gradient noise* $\xi^k = \widetilde{\nabla} f(x^k) - \nabla f(x^k)$ *decays faster than* $1/(k+1)$. *Such decay can be effectively achieved by gradually increasing the mini-batch size used to estimate the stochastic gradient* $\widetilde{\nabla} f(x^k)$ *for the large-scale optimization problem* $\min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^N f_i(x) + h(Ax)$, *where* N *is very large. For example, if the mini-batch size* S_k *satisfies* $|S_k| = O((k+1)^{2a})$ *for some* $a > 1$, *then the variance of the stochastic gradient satisfies* $\mathbb{E}[\|\xi^k\|^2|x^k] = O(1/(k+1)^{2a})$ *where* $\xi^k = \frac{1}{|S_k|}\sum_{i\in S_k} \nabla f_i(x^k) - \nabla f(x^k)$, *which in turn ensures* $\|\sigma(x^k)\|_{\mathrm{F}} = O(1/(k+1)^a)$, $a > 1$.

Theorem 5 *Under the conditions of Theorem 4, let* x_∞ *be the almost-sure limit of* $\mathbf{x}(t)$, *and define* $\bar{\mathbf{x}}(t)$ *and* $\bar{x}_\infty$ *as in (37) and (42), respectively. Then, there exist* $v_1, v_2 > 0$ *such that:*

(i) If $\theta \in (0, 1/2]$, *there exist constants* $u_1, u_2 > 0$ *and* $T > 0$ *such that for all* $t > T$,

$$\mathbb{E}[\|\mathbf{x}(t) - x_\infty\|] \le u_1 \exp(-u_2(t-T)) + \frac{v_2}{(t+1)^{a/2}},$$
$$\mathbb{E}[\|\bar{\mathbf{x}}(t) - \bar{x}_\infty\|] \le v_1 u_1 \exp(-u_2(t-T)) + \frac{v_1 v_2}{(t+1)^{a/2}}.$$

(ii) If $\theta \in (1/2, 1)$, *there exist a constant* $w_1 > 0$ *and* $T > 0$ *such that for all* $t > T$,

$$\mathbb{E}[\|\mathbf{x}(t) - x_\infty\|] \le w_1 (t-T)^{\frac{1-\theta}{1-2\theta}} + \frac{v_2}{(t+1)^{a/2}},$$
$$\mathbb{E}[\|\bar{\mathbf{x}}(t) - \bar{x}_\infty\|] \le v_1 w_1 (t-T)^{\frac{1-\theta}{1-2\theta}} + \frac{v_1 v_2}{(t+1)^{a/2}}.$$

Proof. Proof. It has been shown in Theorem 4 that $\mathbf{x}(t) \to x_\infty$ a.s. and $N_t \to N_\infty$ a.s. Then, from (9), we have for any $t \ge 0$,

$$\|\mathbf{x}(t) - x_\infty\| \le \frac{1}{\lambda}\int_t^\infty \|\nabla H_\mu(\mathbf{x}(s))\|\, ds + \frac{1}{\lambda\sqrt{\rho}}\|N_t - N_\infty\| \quad \text{a.s.}$$

By taking expectation on both sides of the above inequality, it holds that

$$\mathbb{E}[\|\mathbf{x}(t) - x_\infty\|] \le \frac{1}{\lambda}\int_t^\infty \mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|]\, ds + \frac{1}{\lambda\sqrt{\rho}}\mathbb{E}[\|N_t - N_\infty\|]. \tag{43}$$

Under the assumption that $\mathbb{E}[U_t] \le \frac{C}{(t+1)^a}$ for $t \ge 0$, we have

$$\mathbb{E}[\|N_t - N_\infty\|] \le \mathbb{E}[\|N_t - N_\infty\|^2]^{1/2} = \mathbb{E}\left[\int_t^\infty \|\sigma(\mathbf{x}(s))\|_{\mathrm{F}}^2\, ds\right]^{1/2} \le \frac{\sqrt{C}}{(t+1)^{a/2}}. \tag{44}$$

Following (38) and (40), there exists a sufficiently large $T > 0$ such that for all $t \ge T$,

$$\int_t^\infty \mathbb{E}[\|\nabla H_\mu(\mathbf{x}(s))\|]\, ds \le \begin{cases} 2\sqrt{\frac{a_1\lambda}{b_1}}\exp(-\frac{b_1}{2}(t-T)), & \text{if } \theta \in (0, 1/2], \\ \frac{\sqrt{a_2\lambda(2\theta-1)}}{1-\theta}(t-T)^{\frac{1-\theta}{1-2\theta}}, & \text{if } \theta \in (1/2, 1), \end{cases} \tag{45}$$

where a_1, a_2, and b_1 are the same as those in (38). Substituting (44) and (45) into (43) yields that for $t \geq T$,

$$\mathbb{E}[\|\mathbf{x}(t) - x_\infty\|] \leq \begin{cases} 2\sqrt{\frac{a_1}{b_1\lambda}}\exp(-\frac{b_1}{2}(t-T)) + \frac{\sqrt{C}}{\lambda\sqrt{\rho}(t+1)^{a/2}}, & \text{if } \theta \in (0, 1/2], \\ \frac{\sqrt{a_2(2\theta-1)}}{(1-\theta)\sqrt{\lambda}}(t-T)^{\frac{1-\theta}{1-2\theta}} + \frac{\sqrt{C}}{\lambda\sqrt{\rho}(t+1)^{a/2}}, & \text{if } \theta \in (1/2, 1). \end{cases}$$

Recall that we have proven that $\bar{\mathbf{x}}(t) \to \bar{x}_\infty$ in Theorem 4. Notice that $A^T(AA^T)^{-1}$ is a Moore-Penrose pseudoinverse of A, then by the definition of Frobenius condition number $\mathrm{cond}(A)$, it follows that $\|A^T(AA^T)^{-1}\|_{\mathrm{F}} \cdot \|A\|_{\mathrm{F}} = \mathrm{cond}(A)$. Then, we derive

$$\begin{aligned} \mathbb{E}[\|\bar{\mathbf{x}}(t) - \bar{x}_\infty\|] &\leq \mathbb{E}[\|\mathbf{x}(t) - x_\infty\|] + \mathbb{E}[\|A^T(AA^T)^{-1}(A\mathbf{x}(t) - Ax_\infty)\|] \\ &\quad + \mathbb{E}[\|A^T(AA^T)^{-1}(\mathrm{prox}_{\mu h}(A\mathbf{x}(t)) - \mathrm{prox}_{\mu h}(Ax_\infty))\|] \\ &\leq \left(1 + \frac{2-\mu\varrho}{1-\mu\varrho}\mathrm{cond}(A)\right)\mathbb{E}[\|\mathbf{x}(t) - x_\infty\|], \end{aligned}$$

where the last inequality is due to the $\frac{1}{1-\mu\varrho}$-Lipschitz continuity of $\mathrm{prox}_{\mu h}$ (see [16, Proposition 3.3]). Combining the above two inequalities, we derive the in-expectation convergence rates of $\|\bar{\mathbf{x}}(t) - \bar{x}_\infty\|$ as follows:

$$\mathbb{E}[\|\bar{\mathbf{x}}(t) - \bar{x}_\infty\|] \leq \begin{cases} 2v_1\sqrt{\frac{a_1}{b_1\lambda}}\exp(-\frac{b_1}{2}(t-T)) + \frac{v_1\sqrt{C}}{\lambda\sqrt{\rho}(t+1)^{a/2}}, & \text{if } \theta \in (0, 1/2], \\ \frac{v_1\sqrt{a_2(2\theta-1)}}{(1-\theta)\sqrt{\lambda}}(t-T)^{\frac{1-\theta}{1-2\theta}} + \frac{v_1\sqrt{C}}{\lambda\sqrt{\rho}(t+1)^{a/2}}, & \text{if } \theta \in (1/2, 1), \end{cases}$$

where $v_1 = 1 + \frac{2-\mu\varrho}{1-\mu\varrho}\mathrm{cond}(A)$. □

Remark 6 *For all $\theta \in (0, 1)$, the convergence of $\mathbb{E}[\|\mathbf{x}(t) - x_\infty\|]$ and $\mathbb{E}[\|\bar{\mathbf{x}}(t) - \bar{x}_\infty\|]$ is sublinear with the rate depending on the value of θ:*

(i) For $\theta \in (0, 1/2]$. Both $\mathbb{E}[\|\mathbf{x}(t) - x_\infty\|]$ and $\mathbb{E}[\|\bar{\mathbf{x}}(t) - \bar{x}_\infty\|]$ converge at a sublinear rate, since the exponential term $\exp(-u_2(t-T))$ is eventually dominated by $\frac{1}{(t+1)^{a/2}}$.

(ii) For $\theta \in (1/2, 1)$. The convergence remains sublinear, but the leading term depends on the value of θ. The decay is governed by $\frac{1}{(t+1)^{a/2}}$ if $\theta \in (\frac{1}{2}, \frac{2+a}{2a+2})$, the term $(t-T)^{\frac{1-\theta}{1-2\theta}}$ becomes dominant if $\theta \in (\frac{2+a}{2a+2}, 1)$.

6 Convergence of Discrete Sequences $H_\mu(x^k)$ and $H(x^k)$

In this section, we apply the continuous-time results to the discrete scheme and characterize the convergence properties of $H_\mu(x^k)$ and $H(x^k)$ during the iteration process of SPLA.

Denote by $\mathcal{G}$ the set of continuous functions with most polynomial growth, i.e., $g \in \mathcal{G}$ if there exist integers $\kappa_1, \kappa_2 > 0$ such that

$$|g(x)| \leq \kappa_1(1 + \|x\|^{2\kappa_2}) \quad \text{for all } x \in \mathbb{R}^n.$$

For each integer $\alpha \geq 1$, we denote by $\mathcal{G}^\alpha$ the set of α-times continuously differentiable functions, which, together with their partial derivatives up to order α, belong to $\mathcal{G}$. We write $\mathcal{G}_w^\alpha$ for the set of functions g

possessing weak derivatives up to order α such that, for each multi-index β with size $|\beta| \leq \alpha$, there exist positive integers $\kappa_1, \kappa_2 > 0$ satisfying

$$\|D^\beta g(x)\| \leq \kappa_1(1 + \|x\|^{2\kappa_2}) \quad \text{for almost every } x \in \mathbb{R}^n,$$

where $D^\beta g$ denotes the order β weak derivative of g. Clearly, $\mathcal{G}^\alpha \subset \mathcal{G}_w^\alpha$. We now state a standard weak-approximation result adapted from [32, Corollary 10] and [30, Theorem 3], which is a key theoretical tool for linking the continuous-time model to its discrete counterpart.

Proposition 1 *Given $T > 0$, let $\{(x^k, z^k, u^k) : k \geq 0\}$ be the sequence generated by Algorithm 1 with $\rho > \max\{1, 1/T\}$, $\{\mathbf{x}(t) : t \in [0, T]\}$ be a solution trajectory defined by the SDE (9). Under Assumptions 1 and 2, and further assuming f, $h_\mu \in \mathcal{G}_w^3$, then for every $g \in \mathcal{G}^2$, there exists a constant $C > 0$ (independent of ρ) such that*

$$\max_{k=0,1,\cdots,\lfloor \rho T \rfloor} \left| \mathbb{E}[g(x^k)] - \mathbb{E}[g(\mathbf{x}(k/\rho))] \right| \leq \frac{C}{\rho}.$$

In order to obtain the convergence of $H_\mu(x^k)$ and $H(x^k)$, we require additional smoothness assumptions on f and h_μ. These assumptions are not part of our standing conditions for the nonsmooth problem (1) and are imposed here exclusively for this purpose.

Theorem 6 *Under the conditions of Theorem 3, we further assume $f, h_\mu \in \mathcal{G}^2 \cap \mathcal{G}_w^3$. Then, for the sequence $\{x^k\}$ generated by SPLA, the following statements hold for any fixed $T > 0$ and all $k \leq \rho T$:*

(i) If $\theta \in (0, 1/2]$, there exist constants $a_1, b_1, c_1 > 0$ such that

$$|\mathbb{E}[H_\mu(x^k) - H_{\mu,\infty}]| \leq a_1 \exp(-b_1 k) + \frac{c_1}{\rho}.$$

(ii) If $\theta \in (1/2, 1)$, there exist constants $a_2, c_2 > 0$ such that

$$|\mathbb{E}[H_\mu(x^k) - H_{\mu,\infty}]| \leq a_2 k^{\frac{1}{1-2\theta}} + \frac{c_2}{\rho}.$$

Proof. Proof. Under the assumption that f, $h_\mu \in \mathcal{G}^2 \cap \mathcal{G}_w^3$, the result follows directly from Theorem 3 and Proposition 1 by setting $g = H_\mu$. □

Similar to Corollary 1, we obtain the following corollary which follows directly from Lemma 6 in Appendix A.

Corollary 2 *Under the conditions of Theorem 6, for any fixed $T > 0$ and all $k \leq \rho T$, we have:*

(i) For $\theta \in (0, 1/2]$, there exist constants $a_1, b_1, c_1 > 0$ such that

$$|\mathbb{E}[H(x^k) - H_{\mu,\infty}]| \leq a_1 \exp(-b_1 k) + \frac{c_1}{\rho} + \frac{\mu L_h^2}{2(1 - \varrho\mu)}.$$

(ii) For $\theta \in (1/2, 1)$, there exist constants $a_2, c_2 > 0$ such that

$$|\mathbb{E}[H(x^k) - H_{\mu,\infty}]| \leq a_2 k^{\frac{1}{1-2\theta}} + \frac{c_2}{\rho} + \frac{\mu L_h^2}{2(1 - \varrho\mu)}.$$

7 Conclusion

In this paper, we analyzed the convergence properties of a stochastic proximal linearized ADMM (SPLA) for nonconvex optimization from a dynamic perspective. We first established a fundamental connection between SPLA and a first-order SDE, which enabled us to investigate the algorithm from the viewpoint of stochastic dynamical system. Under mild conditions, we proved the almost-sure convergence of the smoothed objective, as well as in-expectation convergence rates by leveraging the Łojasiewicz inequality. We further showed the SDE's solution converges almost surely to some critical point of the smoothed problem, and derived corresponding in-expectation convergence rates. Based on the SDE's solution, we constructed a stochastic process that converges almost surely to an approximate critical point of the original objective function, along with its in-expectation convergence rates. Finally, we presented convergence properties of the discrete sequences generated by the algorithm SPLA.

A Some Technical Results

Theorem 7 (Existence and Uniqueness Theorem) *[44, Theorem 5.2.1] For the SDE (2), if there exists some constant $C \geq 0$ such that for every $T > 0$,*

$$\|F(t,x) - F(t,y)\| + \|G(t,x) - G(t,y)\|_{\mathrm{F}} \leq C\|x - y\|, \quad \forall x, y \in \mathbb{R}^n, \forall t \in [0,T],$$

then (2) has a unique solution $\mathbf{x}$ and $\mathbb{E}[\sup_{t\in[0,T]} \|\mathbf{x}(t)\|^2] < +\infty$ for every $T > 0$.

Theorem 8 (Martingale Convergence Theorem) *[21, Theorem 4.1] Let $\{M_t, t \geq 0\} : \Omega \to \mathbb{R}$ be a continuous martingale such that $\sup_{t\geq 0} \mathbb{E}[|M_t|^p] < +\infty$ for some $p > 1$. Then there exists a random variable M_∞ satisfying $\mathbb{E}[|M_\infty|^p] < +\infty$ such that $\lim_{t\to\infty} M_t = M_\infty$ a.s.*

Lemma 5 (Comparison Lemma with Explicit Solution) *Let $t_0 \geq 0$ and $T > t_0$. Given constants $a \geq 1$, $b > 0$ and $c \geq 0$, consider the Cauchy problem*

$$\begin{cases} \dfrac{dg(t)}{dt} = -bg(t)^a + c \quad \text{for almost all } t > t_0, \\ g(t_0) = g_0 > 0, \end{cases} \tag{46}$$

which admits an absolutely continuous solution $g : [t_0, T] \to \mathbb{R}_+$. Then the following statements hold:

(i) If $g_0 < (\frac{c}{b})^{\frac{1}{a}}$, then for all $t \geq t_0$, the solution g satisfies

$$g(t) \leq \begin{cases} g_0 \exp(-b(t-t_0)) + \dfrac{c}{b}, & \text{if } a = 1, \\ \left[g_0^{1-a} + (a-1)b(t-t_0)\right]^{\frac{1}{1-a}} + \left(\dfrac{c}{b}\right)^{\frac{1}{a}}, & \text{if } a > 1. \end{cases} \tag{47}$$

(ii) If a bounded from below and lower semicontinuous function $f : [t_0, T] \to \mathbb{R}_+$ satisfies

$$f(t) \leq f(s) - b\int_s^t f(u)^a \, du + c(t-s)$$

for $T \geq t > s \geq t_0$ and $f(t_0) = g_0$, then $f(t) \leq g(t)$ holds for all $t \in [t_0, T]$.

Proof. Proof. For $a = 1$, the conclusion follows directly from the explicit integration of (46).

For $a > 1$, from (46) we observe that $\frac{dg(t)}{dt}$ is 0 if $g(t) = (\frac{c}{b})^{\frac{1}{a}}$, and negative if $g(t) > (\frac{c}{b})^{\frac{1}{a}}$, and positive if $g(t) < (\frac{c}{b})^{\frac{1}{a}}$. Since $g_0 < (\frac{c}{b})^{\frac{1}{a}}$, we have $\max_{t\geq t_0} g(t) = (\frac{c}{b})^{\frac{1}{a}}$. To determine the solution of (46), we first solve the simplified equation:

$$\begin{cases} \dfrac{dh(t)}{dt} = -bh(t)^a \quad \text{for } t > t_0, \\ h(t_0) = g_0. \end{cases} \tag{48}$$

It is easy to obtain the solution of (48):

$$h(t) = \left[g_0^{1-a} + (a-1)b(t-t_0)\right]^{\frac{1}{1-a}},$$

which is decreasing in t and converges to 0 as $t \to +\infty$. By $g(t) \leq \left(\frac{c}{b}\right)^{\frac{1}{a}}$ and $h(t) \geq 0$ for $t \geq t_0$, it follows that $\epsilon(t) := g(t) - h(t) \leq \left(\frac{c}{b}\right)^{\frac{1}{a}}$. Therefore,

$$g(t) = h(t) + \epsilon(t) \leq \left[g_0^{1-a} + (a-1)b(t-t_0)\right]^{\frac{1}{1-a}} + \left(\frac{c}{b}\right)^{\frac{1}{a}}, \quad \forall t \geq t_0,$$

Then, item (i) holds.

Item (ii) follows directly from [37, Proposition 2.3]. □

Lemma 6 *Suppose that $f : \mathbb{R}^n \to (-\infty, +\infty]$ is ϱ-weakly convex and L-Lipschitz continuous. Let f_μ be the μ-Moreau envelope of f. Then for any $x \in \mathbb{R}^n$, it holds that*

$$0 \leq f(x) - f_\mu(x) \leq \frac{\mu L^2}{2(1-\varrho\mu)}.$$

Proof. Proof. Since f is ϱ-weakly convex, for any $x, y \in \mathbb{R}^n$ and any $v \in \partial f$, we have

$$f(y) \geq f(x) + \langle v, y - x\rangle - \frac{\varrho}{2}\|y - x\|^2.$$

This, together with the definition of Moreau envelope, yields that

$$f_\mu(x) = \inf_{y\in\mathbb{R}^n}\left\{f(y) + \frac{1}{2\mu}\|y-x\|^2\right\} \geq f(x) + \min_{y\in\mathbb{R}^n}\left\{\langle v, y-x\rangle + (\frac{1}{2\mu} - \frac{\varrho}{2})\|x-y\|^2\right\},$$

which implies

$$f(x) - f_\mu(x) \leq \frac{\mu}{2(1-\varrho\mu)}\|v\|^2.$$

The L-Lipschitz continuity of f ensures $\|v\| \leq L$, which leads to the conclusion. □

B An Informal Derivation of (9)

Let $\{x^k\}, \{z^k\}, \{u^k\}$ be generated by Algorithm 1. Substituting $\xi^k = \widetilde{\nabla} f(x^k) - \nabla f(x^k)$ into (8a), and using the optimality condition of (8b)

$$0 = \nabla h_\mu(z^{k+1}) - \rho(Ax^{k+1} - z^{k+1} + \frac{1}{\rho}u^k),$$

we obtain

$$0 = \tau(x^{k+1} - x^k) + \nabla f(x^k) + A^T\nabla h_\mu(z^{k+1}) + \xi^k - \rho A^T A(x^{k+1} - x^k) + \rho A^T(z^{k+1} - z^k). \tag{49}$$

Following the similar continuous-time modeling approach as in [50], we assume that the discrete sequences $\{x^k\}, \{z^k\}, \{u^k\}$ are interpolations of smooth stochastic processes $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{u}$, i.e., $x^k \approx \mathbf{x}(k/\rho)$, $z^k \approx \mathbf{z}(k/\rho)$ and $u^k \approx \mathbf{u}(k/\rho)$. Put $k = t/s$ with $s = 1/\rho$. Then, for sufficiently small s, we approximate $\mathbf{x}(t) \approx x^{t/s} = x^k, \mathbf{x}(t+s) \approx x^{(t+s)/s} = x^{k+1}, \mathbf{z}(t) \approx z^{t/s} = z^k, \mathbf{z}(t+s) \approx z^{(t+s)/s} = z^{k+1}, \mathbf{u}(t) \approx u^{t/s} = u^k, \mathbf{u}(t+s) \approx u^{(t+s)/s} = u^{k+1}$. By applying Taylor expansion we derive the relations

$$u^{k+1} = \mathbf{u}(t) + s\dot{\mathbf{u}}(t) + O(s^2), \tag{50}$$

$$z^{k+1} = \mathbf{z}(t) + s\dot{\mathbf{z}}(t) + O(s^2), \tag{51}$$

$$x^{k+1} = \mathbf{x}(t) + s\dot{\mathbf{x}}(t) + O(s^2). \tag{52}$$

Then it follows from (8c) that

$$s\dot{\mathbf{u}}(t) = A\mathbf{x}(t) - \mathbf{z}(t) + s(A\dot{\mathbf{x}}(t) - \dot{\mathbf{z}}(t)) + O(s^2).$$

Note that as $s \to 0$, we have

$$A\mathbf{x}(t) = \mathbf{z}(t) \tag{53}$$

for any $t \in \mathbb{R}_+$, which further indicates from the arbitrariness of t that

$$A\dot{\mathbf{x}}(t) = \dot{\mathbf{z}}(t). \tag{54}$$

Assume that ξ^k is a zero mean random vector with covariance $\Sigma(x^k)$. Then ξ^k can be approximated by $\xi^k \approx -\rho^{1/2}\sigma(\mathbf{x}(t))(W(t+\frac{1}{\rho}) - W(t))$, where $W(t)$ is a standard Brownian motion and $\Sigma(x) = \sigma(x)\sigma(x)^T$. Hence, (49) becomes

$$\begin{aligned} 0 = {} & \tau(s\dot{\mathbf{x}}(t) + O(s^2)) + \nabla f(\mathbf{x}(t)) + A^T\nabla h_\mu(\mathbf{z}(t) + O(s)) \\ & - s^{1/2}\sigma(x(t))\frac{dW(t)}{dt} - A^TA\dot{\mathbf{x}}(t) + A^T\dot{\mathbf{z}}(t) + O(s). \end{aligned} \tag{55}$$

Although $\frac{dW(t)}{dt}$ is not defined in the classical sense due to the almost-sure nowhere differentiability of Brownian motion, we use this notation formally to represent a white noise process. Since $\tau s > \|A\|_{\mathrm{F}}^2 + s/\eta$, there exists a constant $\lambda > \|A\|_{\mathrm{F}}^2$ such that

$$\begin{aligned} 0 = {} & \lambda\dot{\mathbf{x}}(t) + \nabla f(\mathbf{x}(t)) + A^T\nabla h_\mu(\mathbf{z}(t) + O(s)) - s^{1/2}\sigma(x(t))\frac{dW(t)}{dt} \\ & - A^TA\dot{\mathbf{x}}(t) + A^T\dot{\mathbf{z}}(t) + O(s). \end{aligned} \tag{56}$$

The continuity of ∇h_μ and (53) imply $\nabla h_\mu(\mathbf{z}(t) + O(s)) \to \nabla h_\mu(A\mathbf{x}(t))$ as $s \to 0$. Moreover, the stochastic term $s^{1/2}\sigma(x(t))\frac{dW(t)}{dt}$ remains dominant as $s \to 0$, since its Euler discretization is the stochastic gradient noise ξ^k, therefore, it cannot be neglected as $s \to 0$. Then, taking the limit $s \to 0$ in (56), we finally obtain the SDE (9).

References

[1] Hédy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116:5–16, 2009.

[2] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Math. Program.*, 35(2):5–16, 2010.

[3] Hédy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, 137:91–129, 2013.

[4] Hédy Attouch, Zaki Chbani, Fadili Jalal, and Hassan Riahi. Fast convergence of dynamical ADMM via time scaling of damped inertial dynamics. *J. Optim. Theory Appl.*, 193:704–736, 2022.

[5] Samaneh Azadi and Suvrit Sra. Towards an optimal stochastic alternating direction method of multipliers. In Eric P. Xing and Tony Jebara, editors, *Proc. 31st ICML*, pages 620–628, Bejing, China, 2014.

[6] Paul H. Bezandry and Toka Diagana. An introduction to stochastic differential equations. In *Almost Periodic Stochastic Processes*, pages 61–115. Springer New York, New York, NY, 2011.

[7] Fengmiao Bian, Jiangwei Liang, and Xiaoqun Zhang. A stochastic alternating direction method of multipliers for non-smooth and non-convex optimization. *Inverse Probl.*, 37(7):075009, 2021.

[8] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. A nonsmooth Morse–Sard theorem for subanalytic functions. *J. Math. Anal. Appl.*, 321(2):729–740, 2006.

[9] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4):1205–1223, 2007.

[10] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.*, 18(2):556–572, 2007.

[11] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146:459–494, 2014.

[12] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Nonconvex Lagrangian-based optimization: Monitoring schemes and global convergence. *Math. Oper. Res.*, 43:1210–1232, 2018.

[13] Radu Ioan Boţ, Ernö Robert Csetnek, and Szilárd Csaba László. A second-order dynamical approach with variable damping to nonconvex smooth minimization. *Appl. Anal.*, 99(3):361–378, 2020.

[14] Radu Ioan Boţ and Dang-Khoa Nguyen. The proximal alternating direction method of multipliers in the nonconvex setting: Convergence analysis and rates. *Math. Oper. Res.*, 45(2):682–712, 2020.

[15] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

[16] Axel Böhm and Stephen J. Wright. Variable smoothing for weakly convex composite functions. *J. Optim. Theory Appl.*, 188:628–649, 2021.

[17] Camille Castera, Jérôme Bolte, Cédric Févotte, and Edouard Pauwels. An inertial newton algorithm for deep learning. *J. Mach. Learn. Res.*, 22:5977–6007, 2021.

[18] Han-Fu Chen. *Stochastic Approximation and Its Applications*. Springer New York, NY, 2010.

[19] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM J. Optim.*, 24(4):1779–1814, 2014.

[20] Marc Dambrine, Ch Dossal, Bénédicte Puig, and Aude Rondepierre. Stochastic differential equations for modeling first order optimization methods. *SIAM J. Optim.*, 34(2):1402–1426, 2024.

[21] J.L. Doob. *Stochastic Processes*. Probability and Statistics Series. Wiley, 1953.

[22] Derek Driggs, Junqi Tang, Jingwei Liang, Mike Davies, and Carola-Bibiane Schönlieb. A stochastic proximal alternating minimization for nonsmooth and nonconvex optimization. *SIAM J. Imaging Sci.*, 14(4):1932–1970, 2021.

[23] Ernie Esser, Xiaoqun Zhang, and Tony F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.*, 3(4):1015–1046, 2010.

[24] Lawrence C. Evans. *An Introduction to Stochastic Differential Equations*. American Mathematical Society, Providence, Rhode Island, 2012.

[25] Guilherme Franç a, Daniel P. Robinson, and René Vidal. ADMM and accelerated ADMM as continuous dynamical systems. In Jennifer Dy and Andreas Krause, editors, *Proc. 35th ICML*, volume 80, pages 1559–1567, Red Hook, NY, 10–15 Jul 2018. PMLR.

[26] Guilherme França, Daniel P. Robinson, and René Vidal. A nonsmooth dynamical systems perspective on accelerated extensions of ADMM. *IEEE Trans. Automat. Control*, 68(5):2966–2978, 2023.

[27] Jonas Geiping and Michael Moeller. Composite optimization by nonconvex majorization-minimization. *SIAM J. Imaging Sci.*, 11(4):2494–2528, 2018.

[28] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier*, 48(3):769–783, 1998.

[29] Szilárd Csaba László. Convergence rates for an inertial algorithm of gradient type associated to a smooth non-convex minimization. *Math. Program.*, 190(1):285–329, 2021.

[30] Chris Junchi Li. A general continuous-time formulation of stochastic ADMM and its variants, 2024. arXiv: 2404.14358.

[31] Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM J. Optim.*, 25(4):2434–2460, 2015.

[32] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *J. Mach. Learn. Res.*, 20(40):1–47, 2019.

[33] Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the Kurdyka-Łojasiewicz inequality. *SIAM J. Optim.*, 33(2):1092–1120, 2023.

[34] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In *Adv. Neural Inf. Process. Syst.*, volume 34, pages 12712–12725, 2021.

[35] Yanli Liu, Yunbei Xu, and Wotao Yin. Acceleration of primal–dual methods by preconditioning and simple subproblem procedures. *J. Sci. Comput.*, 86(2):1–34, 2021.

[36] Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les Équ. Dériv. Partielles*, pages 8–89, 1963.

[37] Radosław Matusik, Andrzej Nowakowski, Sławomir Plaskacz, and Andrzej Rogowski. Finite-time stability for differential inclusions with applications to neural networks. *SIAM J. Control Optim.*, 58(5):2854–2870, 2020.

[38] Rodrigo Maulen-Soto, Jalal Fadili, and Hedy Attouch. An SDE perspective on stochastic convex optimization. *Math. Oper. Res.*, In press, 2024.

[39] Rodrigo Maulen-Soto, Jalal Fadili, and Hedy Attouch. Tikhonov regularization for stochastic convex optimization in Hilbert spaces, 2024. arXiv:2403.06708v4.

[40] Rodrigo Maulen-Soto, Jalal Fadili, Hedy Attouch, and Peter Ochs. An SDE perspective on stochastic inertial gradient dynamics with time-dependent viscosity and geometric damping, 2024. arXiv:2407.04562.

[41] Rodrigo Maulen-Soto, Jalal Fadili, Hedy Attouch, and Peter Ochs. Stochastic inertial dynamics via time scaling and averaging, 2025. arXiv:2403.16775.

[42] Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM J. Optim.*, 28(1):163–197, 2018.

[43] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[44] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, Heidelberg, 2003.

[45] Antonio Orvieto and Aurelien Lucchi. Continuous-time models for stochastic optimization algorithms. In *Adv. Neural Inf. Process. Syst.*, volume 32, 2019.

[46] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*. Springer-Verlag, Berlin, New York, NY, 1998.

[47] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*. Soc. Ind. Appl. Math., Philadelphia, PA, 2021.

[48] Ron Shefi and Marc Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM J. Optim.*, 24(1):269–297, 2014.

[49] Bin Shi, Weijie Su, and Michael I. Jordan. On learning rates and schrödinger operators. *J. Mach. Learn. Res.*, 24(379):1–53, 2023.

[50] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.*, 17(153):1–43, 2016.

[51] Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: further properties and nonmonotone linesearch algorithms. *SIAM J. Optim.*, 28(3):2274–2303, 2018.

[52] Hoheisel Tim, Laborde Maxime, and Oberman Adam. A regularization interpretation of the proximal point method for weakly convex functions. *J. Dyn. Games*, 7(1):79–96, 2020.

[53] Yi Xu, Mingrui Liu, Qihang Lin, and Tianbao Yang. ADMM without a fixed penalty parameter: Faster convergence with new adaptive penalization. In *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, pages 1267–1277, Red Hook, NY, USA, 2017.

[54] Wenliang Zhong and James Kwok. Fast stochastic alternating direction method of multipliers. In Eric P. Xing and Tony Jebara, editors, *Proc. 31st ICML*, pages 46–54, Bejing, China, 2014.