Journal of Machine Learning Research 21 (2020) 1-103 Submitted 3/19; Revised 7/20; Published 9/20

Asymptotic Analysis via Stochastic Differential Equations of Gradient Descent Algorithms in Statistical and Computational Paradigms

Yazhen Wang YZWANG@STAT.WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison WI 53706-1510, USA

Shang Wu SHANGWU@FUDAN.EDU.CN
Department of Statistics
School of Management
Fudan University
Shanghai 200433, China

Editor: Rina Foygel Barber

Abstract

This paper investigates the asymptotic behaviors of gradient descent algorithms (particularly accelerated gradient descent and stochastic gradient descent) in the context of stochastic optimization arising in statistics and machine learning, where objective functions are estimated from available data. We show that these algorithms can be computationally modeled by continuous-time ordinary or stochastic differential equations. We establish gradient flow central limit theorems to describe the limiting dynamic behaviors of these computational algorithms and the large-sample performances of the related statistical procedures, as the number of algorithm iterations and data size both go to infinity, where the gradient flow central limit theorems are governed by some linear ordinary or stochastic differential equations, like time-dependent Ornstein-Uhlenbeck processes. We illustrate that our study can provide a novel unified framework for a joint computational and statistical asymptotic analysis, where the computational asymptotic analysis studies the dynamic behaviors of these algorithms with time (or the number of iterations in the algorithms), the statistical asymptotic analysis investigates the large-sample behaviors of the statistical procedures (like estimators and classifiers) that are computed by applying the algorithms; in fact, the statistical procedures are equal to the limits of the random sequences generated from these iterative algorithms, as the number of iterations goes to infinity. The joint analysis results based on the obtained gradient flow central limit theorems lead to the identification of four factors—learning rate, batch size, gradient covariance, and Hessian—to derive new theories regarding the local minima found by stochastic gradient descent for solving non-convex optimization problems.

Keywords: acceleration, gradient descent, gradient flow central limit theorem, joint asymptotic analysis, joint computational and statistical analysis, Lagrangian flow central limit theorem, mini-batch, optimization, ordinary differential equation, stochastic differential equation, stochastic gradient descent, weak convergence

1. Introduction

Optimization plays an important role in scientific fields, ranging from machine learning to physical sciences and engineering and from statistics to social sciences and business. It lies at the core of data science as it provides a mathematical language for handling both computational algorithms and statistical inferences in data analysis. Numerous algorithms and methods have been proposed to solve optimization problems. Examples include Newton's method, gradient and subgradient descent, conjugate gradient methods, trust region methods, and interior point methods (Polyak, 1987; Boyd and Vandenberghe, 2004; Nemirovskii and Yudin, 1983; Nocedal and Wright, 2006; Ruszczynski, 2006; Boyd et al., 2011; Shor, 2012; Goodfellow et al., 2016). Practical problems arising in fields such as statistics and machine learning usually involve optimization settings where the objective functions are empirically estimated from available data in the form of a sum of differentiable functions. We refer to such optimization problems with random objective functions as stochastic optimization. As data sets grow rapidly in terms of scale and complexity, methods such as stochastic gradient descent can scale to the enormous size of big data and have been rather popular thus far. There has been recent surging interest in and great research work on the theory and practice of gradient descent and its extensions and variants. Further, there is extensive literature on stochastic approximation and recursive algorithms in machine learning, particularly stochastic gradient descent in deep learning (Ali et al., 2019; Chen et al., 2020; Dalalyan, 2017a, 2017b; Fan et al., 2018; Foster et al., 2019; Ge et al., 2015; Ghadimi and Lan, 2015; Jastrzębski et al., 2018; Jin et al., 2017; Kawaguchi, 2016; Keskar et al., 2017; Kushner and Yin, 2003; Lee et al., 2016; Li et al., 2016; Li et al., 2017a; Li et al., 2017b; Ma et al., 2019; Mandt et al., 2016, 2017; Nemirovski et al., 2009; Rakhlin et al., 2012; Ruppert, 1988; Shallue et al., 2019; Sirignano and Spiliopoulos, 2017; Su et al., 2016; Toulis et a., 2014; Toulis and Airoldi, 2015, 2016, 2017; Wibisono et al., 2016; Zhu, 2019). In spite of compelling theoretical and numerical evidence on the value of the concept of stochastic approximation and the acceleration phenomenon, there remains some conceptual and theoretical mystery in acceleration and stochastic approximation schemes.

1.1 Contributions

Both continuous-time and discrete-time means are adopted by computational and statistical (as well as machine learning) communities to study learning algorithms like stochastic gradient descent for solving optimization problems. The research on the computational aspect focuses more on the convergence and convergent dynamics of learning algorithms—in contrast the statistics research emphasizes statistical inferences of learning rules, where the learning rules are solutions of the optimization problems and the learning algorithms are designed to find the solutions. This paper adopts a new approach to combine both computational and statistical frameworks and develop a joint computational and statistical paradigm for analyzing gradient descent algorithms. Our joint study can handle computational convergence behaviors of the gradient descent algorithms as well as statistical large-sample performances of learning rules that are computed by the gradient descent algorithms. To be specific, in this paper, we derive continuous-time ordinary or stochastic differential equations to model the dynamic behaviors of these gradient descent algorithms

and investigate their limiting algorithmic dynamics and large-sample performances, as the number of algorithm iterations and data size both go to infinity.

For an optimization problem whose objective function is convex and deterministic, we consider a matched stochastic optimization problem whose random objective function is an empirical estimator of the deterministic objective function based on available data. The solution of the stochastic optimization specifies a decision rule like an estimator or a classifier based on the sampled data in statistics and machine learning, while its corresponding deterministic optimization problem characterizes—through its solution—the true value of the parameter in the population model. In other words, the two connected optimization problems associate with the data sample and its corresponding population model where the data are sampled from, and the stochastic optimization is considered to be a sample version of the deterministic optimization corresponding to the population. These two types of optimization problems refer to the deterministic population and stochastic sample optimization problems.

Consider random sequences that are generated from the gradient descent algorithms and their corresponding continuous-time ordinary or stochastic differential equations for the stochastic sample optimization setting. We show that the random sequences converge to solutions of the ordinary differential equations for the corresponding deterministic population optimization setup, and we derive their asymptotic distributions by some linear ordinary or stochastic differential equations like time-dependent Ornstein-Uhlenbeck processes. The asymptotic distributions are used to understand and quantify the limiting discrepancy between the random iterative sequences generated from each algorithm for solving the corresponding sample and population optimization problems. In particular, since the obtained asymptotic distributions characterize the limiting behavior of the normalized difference between the sample and population gradient (or Lagrangian) flows, the limiting distributions may be viewed as central limit theorems (CLT) for gradient (or Lagrangian) flows and are then called gradient (or Lagrangian) flow central limit theorems (GF-CLT or LF-CLT). Moreover, our analysis may offer a novel unified framework to conduct a joint asymptotic analysis for computational algorithms and statistical decision rules that are computed by applying the algorithms. As iterated computational methods, these gradient descent algorithms generate iterated sequences that converge to the exact decision rule or the true parameter value for the corresponding optimization problems, when the number of the iterations goes to infinity. Thus, as time (corresponding to the number of iterations) goes to infinity, the continuous-time differential equations may have distributional limits corresponding to the large-sample distributions of statistical decision rules as the sample size goes to infinity. In other words, the asymptotic analysis can be performed with both time and data size, where the time direction corresponds to the computational asymptotics on dynamic behaviors of the algorithms, and the data size direction associates with the statistical large-sample asymptotics on the statistical behaviors of decision rules—such as estimators and classifiers. The continuous-time modeling and the GF-CLT based joint asymptotic analysis may reveal new facts and shed some light on the phenomenon that stochastic gradient descent algorithms can escape from saddle points and converge to good local minimizers for solving non-convex optimization problems in deep learning.

In a nutshell, we highlight our main contributions in the following manner:

- We establish a new asymptotic theory for the discrepancy between the sample and population gradient (or Lagrangian) flows. In particular, the new limiting distributions for the normalized discrepancy are called the gradient (or Lagrangian) flow central limit theorems (GF-CLT or LF-CLT). See Sections 3.3 and 4.1-4.2.

- The obtained asymptotic theory provides a novel unified framework for a joint computational and statistical asymptotic analysis. Statistically, the joint analysis can facilitate inferential analysis of a learning rule computed by gradient descent algorithms. Computationally, the joint analysis enables us to understand and quantify a random fluctuation in and the related impact on the dynamic and convergence behavior of a gradient descent algorithm when it is applied to solve a stochastic optimization problem. In particular, the joint analysis can be employed to investigate the joint dynamic effect of data size and algorithm iterations on the computational and statistical errors for iterates generated by the algorithms, such as estimating the bias and variance of iterates and building tests and confidence sets for model parameters under the setting of a finite data sample and various algorithm iterations. See Sections 3.4 and 4.3.

- Computationally, we discover a novel theory that four factors—learning rate, batch size, gradient covariance, and Hessian—along with the associated identities are shown to influence the local minima found by stochastic gradient descent for solving a nonconvex optimization problem. It may also shed light on a certain intrinsic relationship among stochastic optimization, deterministic optimization, and statistical learning. See Section 4.4.

- Statistically, we illustrate implications of our results for statistical analysis of stochastic gradient descent and inference of outputs from stochastic gradient descent. See Section 4.5.

- The continuous-time approach is employed to demonstrate that it can provide a handy means and a beautiful framework for deriving elegant and deep results for stochastic dynamics of learning algorithms and statistical inference of learning rules.

1.2 Organization

The rest of the paper proceeds as follows. Section 2 introduces deterministic optimization and describes gradient descent, accelerated gradient descent, and their corresponding ordinary differential equations. Section 3 presents stochastic optimization and investigates asymptotic behaviors of the plain and accelerated gradient descent algorithms and their associated ordinary differential equations (with random coefficients) when the sample size goes to infinity. We illustrate the unified framework to conduct a joint analysis on computational and statistical asymptotics, where computational asymptotics deals with dynamic behaviors of the gradient descent algorithms with time (or iteration), and statistical asymptotics studies large-sample behaviors of statistical decision rules that are computed through the application of the algorithms. Section 4 considers stochastic gradient descent algorithms for large scale data and derives stochastic differential equations to model these algorithms. We establish asymptotic theory for these algorithms and their associated stochastic differential

equations and describe a joint analysis on computational and statistical asymptotics. All technical proofs are relegated in Section 5.

We adopt the following notations and conventions. For the stochastic sample optimization problem considered in Sections 3 and 4, we add a superscript n to notations for the associated processes and sequences in Section 3 and indices m and/or $*$ to notations for the corresponding processes and sequences affiliated with mini-batches in Section 4, while notations without such subscripts or superscripts are used for sequences and solutions (functions) corresponding to the deterministic population optimization problem given in Section 2. Our basic proof ideas can be described as follows. Each algorithm generates an iterated sequence for computing a learning rule, a step-wise empirical process is formed by the generated sequence, and a continuous process is obtained from the corresponding continuous-time differential equation. We derive asymptotic distributions by analyzing the differential equations, and we bound the differences between the empirical processes and their corresponding continuous processes by studying the optimization problems and utilizing the empirical process theory along with the related differential equations.

2. Ordinary Differential Equations for Gradient Descent Algorithms

This section establishes an optimization framework at the population-level that facilitates the corresponding finite-sample analysis in subsequent sections. Consider the following minimization problem,

$$\min_{\theta\in\Theta} g(\theta), \tag{2.1}$$

where the objective function $g(\theta)$ is defined on a parameter space $\Theta \subset I\!R^p$ and assumed to have L-Lipshitz continuous gradients. Iterative algorithms like gradient descent methods are often employed to numerically compute the solution of the minimization problem. Starting with some initial value x_0, the plain gradient descent algorithm is iteratively defined by

$$x_k = x_{k-1} - \delta\nabla g(x_{k-1}), \tag{2.2}$$

where ∇ denotes gradient operator, and δ is a positive constant that is often called a step size or learning rate.

It is easy to model $\{x_k,\ k = 0, 1, \cdots\}$ by a smooth curve $X(t)$ with the Ansatz $x_k \approx X(k\delta)$ as follows. Define a step function $x_\delta(t) = x_k$ for $k\delta \le t < (k+1)\delta$, and as $\delta \to 0$, $x_\delta(t)$ approaches $X(t)$ satisfying

$$\dot{X}(t) + \nabla g(X(t)) = 0, \tag{2.3}$$

where $\dot{X}(t)$ denotes the derivative of $X(t)$, and initial value $X(0) = x_0$. In fact, $X(t)$ is a gradient flow associated with the objective function $g(\cdot)$ in the optimization problem (2.1).

Nesterov's accelerated gradient descent scheme is a well-known algorithm that is much faster than the plain gradient descent algorithm. Starting with initial values x_0 and $y_0 = x_0$, Nesterov's accelerated gradient descent algorithm is iteratively defined by

$$x_k = y_{k-1} - \delta\nabla g(y_{k-1}), \qquad y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}), \tag{2.4}$$

where δ is a positive constant. Using (2.4), we derive a recursive relationship between consecutive increments

$$\frac{x_{k+1}-x_k}{\sqrt{\delta}} = \frac{k-1}{k+2}\frac{x_k-x_{k-1}}{\sqrt{\delta}} - \sqrt{\delta}\nabla g(y_k). \tag{2.5}$$

We model $\{x_k,\ k=0,1,\cdots\}$ by a smooth curve in a sense that x_k are its samples at discrete points—that is, we define a step function $x_\delta(t)=x_k$ for $k\sqrt{\delta}\le t<(k+1)\sqrt{\delta}$—and introduce the Ansatz $x_\delta(k\sqrt{\delta})=x_k\approx X(k\sqrt{\delta})$ for some smooth function $X(t)$ defined for $t\geqslant 0$. Let $\sqrt{\delta}$ be the step size. Taking $t=k\sqrt{\delta}$ and letting $\delta\to 0$ in equation (2.5), we obtain

$$\ddot{X}(t)+\frac{3}{t}\dot{X}(t)+\nabla g(X(t))=0, \tag{2.6}$$

with the initial conditions $X(0)=x_0$ and $\dot{X}(0)=0$. As the coefficient $3/t$ in the ordinary differential equation (2.6) is singular at $t=0$, the classical ordinary differential equation theory is not applicable to establish the existence or uniqueness of the solution to equation (2.6). The heuristic derivation of (2.6) is from Su et al. (2016), who established that equation (2.6) has a unique solution satisfying the initial conditions, and $x_\delta(t)$ converges to $X(t)$ uniformly on $[0,T]$ for any fixed $T>0$. Note the step size difference between the plain and accelerated cases, where the step size is $\delta^{1/2}$ for Nesterov's accelerated gradient descent algorithm and δ for the plain gradient descent algorithm. Su et al. (2016) showed that, because of the difference, the accelerated gradient descent algorithm moves much faster than the plain gradient descent algorithm along the curve $X(t)$. Wibisono et al. (2016) provided a more elaborate explanation on the acceleration phenomenon and developed a systematic continuous-time variational scheme to generate a large class of continuous-time differential equations and produce a family of accelerated gradient algorithms. The variational scheme utilizes a first-order rescaled gradient flow and a second-order Lagrangian flow, which are generalizations of gradient flow. We refer the solution $X(t)$ of the differential equation (2.3) to the gradient flow for the gradient descent algorithm (2.2), and the solution $X(t)$ to the differential equation (2.6) is called the Lagrangian flow for the accelerated gradient descent algorithm (2.4).

3. Gradient Descent for Stochastic Optimization

Let $\theta=(\theta_1,...,\theta_p)^{'}$ be the parameter that we are interested in, and U be a relevant random element on a probability space with a distribution Q. Consider an objective function $\ell(\theta;u)$ and its corresponding expectation $E[\ell(\theta;U)]=g(\theta)$. For example, in a statistical decision problem, we may take U to be a decision rule, $\ell(\theta;u)$ a loss function, and $g(\theta)=E[\ell(\theta;U)]$ its corresponding risk; in the M-estimation, we treat U as a sample observation and $\ell(\theta;u)$ as a ρ-function; in nonparametric function estimation and machine learning, we choose U as an observation and $\ell(\theta;u)$ equal to a loss function plus some penalty. For these problems, we consider the corresponding deterministic population minimization problem (2.1) for characterizing the true parameter value or its function as an estimand; however, practically, because $g(\theta)$ is usually unavailable, we have to employ its empirical version and consider a stochastic optimization problem, described as follows:

$$\min_{\theta\in\Theta}\mathcal{L}^n(\theta;\mathbf{U}_n), \tag{3.7}$$

where $\mathcal{L}^n(\theta; \mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; U_i)$, $\mathbf{U}_n = (U_1, \cdots, U_n)'$ is a sample, and we assume that $U_1, \cdots, U_n$ are i.i.d. and follow the distribution Q.

The minimization problem (2.1) characterizes the true value of the target estimand such as an estimation parameter in a statistical model and a classification parameter in a machine learning task. As the true objective function $g(\theta)$ is usually unknown in practice, we often solve the stochastic minimization problem (3.7) with observed data to obtain practically useful decision rules such as an M-estimator, a smoothing function estimator, and a machine learning classifier. The approach to obtaining practical procedures is based on the heuristic reasoning that as $n \to \infty$, the law of large number implies that $\mathcal{L}^n(\theta; \mathbf{U}_n)$ eventually converges to $g(\theta)$ in probability, and thus the solution of the stochastic sample minimization problem (3.7) approaches that of the deterministic population minimization problem (2.1).

3.1 Plain Gradient Descent Algorithm

Applying the plain gradient descent scheme to the stochastic sample minimization problem (3.7) with initial value x_0^n, we obtain the following iterative algorithm to compute the solution of the sample minimization problem (3.7),

$$x_k^n = x_{k-1}^n - \delta \nabla \mathcal{L}^n(x_{k-1}^n; \mathbf{U}_n), \tag{3.8}$$

where $\delta > 0$ is a step size or learning rate, and $\mathcal{L}^n$ is the objective function in the sample minimization problem (3.7).

Following the continuous curve approximation described in Section 2 we define a step function $x_\delta^n(t) = x_k^n$ for $k\delta \le t < (k+1)\delta$; for each n, as $\delta \to 0$, $x_\delta^n(t)$ approaches a smooth curve $X^n(t)$, $t \ge 0$, given by

$$\dot{X}^n(t) + \nabla \mathcal{L}^n(X^n(t); \mathbf{U}_n) = 0, \tag{3.9}$$

where $\nabla \mathcal{L}^n(X^n(t); \mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n \nabla \ell(X^n(t); U_i)$, gradient operator ∇ here is applied to $\mathcal{L}^n(\theta; \mathbf{U}_n)$ and $\ell(\theta; U_i)$ with respect to θ, and initial value $X^n(0) = x_0^n$. $X^n(t)$ is a gradient flow associated with $\mathcal{L}^n$ in the stochastic sample optimization problem (3.7).

As $\mathbf{U}_n$ and $X^n(t)$ are random, and our main interest is to study the distributional behaviors of the solution and algorithm, we may define a solution of equation (3.9) in the weak sense that there exist a process $X_\dagger^n(t)$ and a random vector $\mathbf{U}_n^\dagger = (U_1^\dagger, \cdots, U_n^\dagger)'$ defined on some probability space, such that $\mathbf{U}_n^\dagger$ is identically distributed as $\mathbf{U}_n$, $(\mathbf{U}_n^\dagger, X_\dagger^n(t))$ satisfies equation (3.9), and $X_\dagger^n(t)$ is called a (weak) solution of equation (3.9). Note that $X_\dagger^n(t)$ is not required to be defined on a fixed probability space with given random variables; instead, we define $X_\dagger^n(t)$ on some probability space with some associated random vector $\mathbf{U}_n^\dagger$ whose distribution is given by Q. The weak solution definition, which shares the same spirit as that for stochastic differential equations (Ikeda and Watanabe, 1981 and more in Section 4), will be rather handy in facilitating our asymptotic analysis in this paper. For simplicity, we exclude index † and "weak" when there is no confusion.

3.2 Accelerated Gradient Descent Algorithm

Nesterov's accelerated gradient descent scheme can be used to solve the sample minimization problem (3.7). Starting with initial values x_0^n and $y_0^n = x_0^n$, we obtain the following iterative

algorithm to compute the solution of the sample minimization problem (3.7),

$$x_k^n = y_{k-1}^n - \delta \nabla \mathcal{L}^n(y_{k-1}^n; \mathbf{U}_n), \qquad y_k^n = x_k^n + \frac{k-1}{k+2}(x_k^n - x_{k-1}^n). \tag{3.10}$$

Using the continuous curve approach described in Section 2, we can define a step function $x_\delta^n(t) = x_k$ for $k\sqrt{\delta} \le t < (k+1)\sqrt{\delta}$, and for every n, as $\delta \to 0$, we approximate $x_\delta^n(t)$ by a smooth curve $X^n(t)$, $t \ge 0$, governed by

$$\ddot{X}^n(t) + \frac{3}{t}\dot{X}^n(t) + \nabla \mathcal{L}^n(X^n(t); \mathbf{U}_n) = 0, \tag{3.11}$$

where initial values $X^n(0) = x_0^n$ and $\dot{X}^n(0) = 0$, $\nabla \mathcal{L}^n(X^n(t); \mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n \nabla \ell(X^n(t); U_i)$, and gradient operator ∇ here is applied to $\mathcal{L}^n(\theta; \mathbf{U}_n)$ and $\ell(\theta; U_i)$ with respect to θ. $X^n(t)$ is a Lagrangian flow associated with $\mathcal{L}^n$ in the sample optimization problem (3.7).

Again, we define a solution $X^n(t)$ of equation (3.11) in the weak sense—that is, that there exist a process $X^n(t)$ and a random vector $\mathbf{U}_n$ on some probability space so that the distribution of $\mathbf{U}_n$ is specified by Q, and $X^n(t)$ is a solution of equation (3.11).

3.3 Asymptotic Theory via Ordinary Differential Equations

In order to ensure that equations (2.3), (2.6), (3.9) and (3.11) and their solutions are well defined and study their asymptotics, we need to impose the following assumptions.

A0. Assume that initial values satisfy $x_0^n - x_0 = o_P(n^{-1/2})$.

A1. $\ell(\theta; u)$ is continuously twice differentiable in θ; $\forall\ u \in R^p$, $\exists\ h_1(u)$, such that $\forall\ \theta^1, \theta^2 \in \Theta$, $\|\nabla \ell(\theta^1; u) - \nabla \ell(\theta^2; u)\| \leqslant h_1(u)\|\theta^1 - \theta^2\|$, where $h_1(U)$ and $\nabla \ell(\theta_0; U)$ for some fixed θ_0 have finite fourth moments.

A2. $E[\ell(\theta; U)] = g(\theta)$, $E[\nabla \ell(\theta; U)] = \nabla g(\theta)$, $E[\mathbb{H}\ell(\theta; U)] = \mathbb{H}g(\theta)$, on the parameter space Θ, $g(\cdot)$ is continuously twice differentiable and strongly convex, and $\nabla g(\cdot)$ and $\mathbb{H}g(\cdot)$ are L-Lipschitz for some $L > 0$, where ∇ is the gradient operator (the first-order partial derivatives), and $\mathbb{H}$ is the Hessian operator (the second-order partial derivatives).

A3. Define cross auto-covariance $\boldsymbol{\varsigma}(\theta, \vartheta) = (\varsigma_{ij}(\theta, \vartheta))_{1 \le i,j \le p}$, $\theta, \vartheta \in \Theta$, where $\mathrm{Cov}[\frac{\partial}{\partial \theta_i}\ell(\theta; U), \frac{\partial}{\partial \vartheta_j}\ell(\vartheta; U)] = \varsigma_{ij}(\theta, \vartheta)$ are assumed to be continuously differentiable, and L-Lipschitz. Let $\sigma_{ij}(\theta) = \mathrm{Cov}[\frac{\partial}{\partial \theta_i}\ell(\theta; U), \frac{\partial}{\partial \theta_j}\ell(\theta; U)] = \varsigma_{ij}(\theta, \theta)$, and $\boldsymbol{\sigma}^2(\theta) = \mathrm{Var}[\nabla \ell(\theta; U)] = (\sigma_{ij}(\theta))_{1 \le i,j \le p} = \boldsymbol{\varsigma}(\theta, \theta)$ be positive definite.

A4. $\sqrt{n}[\nabla \mathcal{L}^n(\theta; \mathbf{U}_n) - \nabla g(\theta)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n [\nabla \ell(\theta; U_i) - \nabla g(\theta)]$ weakly converges to $\mathbf{Z}(\theta)$ uniformly over $\theta \in \Theta_X$, where $\mathbf{Z}(\theta)$ is a Gaussian process with mean zero and auto-covariance $\boldsymbol{\varsigma}(\theta, \vartheta)$ defined in A3, Θ_X is a bounded subset of Θ, and the interior of Θ_X contains the solutions $X(t)$ of the ordinary differential equations (2.3) and (2.6) connecting the initial value x_0 and the minimizer of $g(\theta)$.

Assumption A0 may relax the usual assumption of taking common initial values $x_0^n = x_0$. Assumptions A1 and A2 are often used to make optimization problems and differential

equations well defined and match the stochastic sample optimization problem (3.7) to the deterministic population optimization problem (2.1). Assumptions A3 and A4 guarantee that the solution of (3.7) and its associated differential equations provide large-sample approximations of those for (2.1). Assumption A4 can be easily justified by empirical processes with common conditions—like that $\nabla\ell(\theta; U)$, $\theta \in \Theta_X$, form a Donsker class (van der Vaart and Wellner, 2000)—since the solution curves $X(t)$ of the ordinary differential equations (2.3) and (2.6) are deterministic and bounded, and it is easy to select Θ_X. Examples that meet the assumptions include common statistical models and well-known loss and likelihood functions, such as usual exponential families and generalized linear models with squared-error and deviance loss functions (more specific cases will be provided later in the paper).

We remind readers of the notion convention specified at the end of Section 1 that adds a superscript n to sample-level notations for the processes and sequences associated with the stochastic sample optimization problem (3.7) in Section 3, while notations without such a superscript are for solutions and sequences corresponding to the deterministic population optimization problem (2.1) in Section 2. For a given $T > 0$, denote by $C([0, T])$ the space of all continuous functions on $[0, T]$ with the uniform metric $\max\{|b_1(t) - b_2(t)| : t \in [0, T]\}$ between functions $b_1(t)$ and $b_2(t)$. For solutions $X(t)$ and $X^n(t)$ of the ordinary differential equations (2.3) and (3.9) [or (2.6) and (3.11)], respectively, we define $V^n(t) = \sqrt{n}[X^n(t) - X(t)]$. Then $X(t)$, $X^n(t)$, and $V^n(t)$ live on $C([0, T])$. Treating them as random elements in $C([0, T])$, in the following theorem, we establish a weak convergence limit of $V^n(t)$.

Theorem 1 *Under Assumptions A0–A4, as $n \to \infty$, $V^n(t)$ weakly converges to a Gaussian process $V(t)$, where $V(t)$ is the unique solution of the following linear differential equations,*

$$\dot{V}(t) + [\mathbb{H}g(X(t))]V(t) + \mathbf{Z}(X(t)) = 0, \;\; V(0) = 0 \tag{3.12}$$

for the plain gradient descent case, and

$$\ddot{V}(t) + \frac{3}{t}\dot{V}(t) + [\mathbb{H}g(X(t))]V(t) + \mathbf{Z}(X(t)) = 0, \;\; V(0) = \dot{V}(0) = 0 \tag{3.13}$$

for the accelerated gradient descent case, where the deterministic functions $X(t)$ in (3.12) and (3.13) are the solutions of the ordinary differential equations (2.3) and (2.6), respectively, $\mathbb{H}$ is the Hessian operator, random coefficient $\mathbf{Z}(\cdot)$ is the Gaussian process given by Assumption A4.

In particular, if Gaussian process $\mathbf{Z}(\theta) = \sigma(\theta)\mathbf{Z}$, where random variable $\mathbf{Z} \sim N_p(0, \mathbf{I}_p)$, and $\boldsymbol{\sigma}(\theta)$ is defined in Assumption A3, then $V(t) = \Pi(t)\mathbf{Z}$ on $C([0, T])$, and the deterministic matrix $\Pi(t)$ is the unique solution of the following linear differential equations,

$$\dot{\Pi}(t) + [\mathbb{H}g(X(t))]\Pi(t) + \boldsymbol{\sigma}(X(t)) = 0, \;\; \Pi(0) = 0 \tag{3.14}$$

for the plain gradient descent case, and

$$\ddot{\Pi}(t) + \frac{3}{t}\dot{\Pi}(t) + [\mathbb{H}g(X(t))]\Pi(t) + \boldsymbol{\sigma}(X(t)) = 0, \;\; \Pi(0) = \dot{\Pi}(0) = 0 \tag{3.15}$$

for the accelerated gradient descent case, where $X(t)$ in (3.14) and (3.15) are the solutions of the ordinary differential equations (2.3) and (2.6), respectively, $\mathbb{H}$ is the Hessian operator, and $\boldsymbol{\sigma}(\cdot)$ is defined in Assumption A3.

Remark 1 *As discussed in Sections 2 and 3.1, for the gradient descent case $X(t)$ and $X^n(t)$ are gradient flows associated with the population optimization* (2.1) *and the sample optimization* (3.7)*, respectively, and thus refer to the corresponding population and sample gradient flows. Consequently, the Gaussian limiting distribution of $V^n(t)$ describes the asymptotic distribution of the difference between the sample and population gradient flows, with a normalization factor $\sqrt{n}$. Hence, it is natural to view the Gaussian limiting distribution as the central limit theorem for the gradient flows, and we call it the gradient flow central limit theorem (GF-CLT). Similarly, for the accelerated case $X(t)$ and $X^n(t)$ are Lagrangian flows associated with the population optimization* (2.1) *and the sample optimization* (3.7)*, respectively, and thus refer to the corresponding population and sample Lagrangian flows. The Gaussian limiting distribution for the normalized discrepancy between the sample and population Lagrangian flows can be naturally viewed as the central limit theorem for the Lagrangian flows, and we call it the Lagrangian flow central limit theorem (LF-CLT).*

Remark 2 *As we discussed earlier in Section 3, as $n \to \infty$, $\mathcal{L}^n(\theta; \mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; U_i)$ converges to $g(\theta)$ in probability, and the solutions of the population minimization (2.1) and the sample optimization (3.7) must be very close to each other. We may heuristically illustrate the derivation of Theorem 1 as follows. The central limit theorem may lead us to that as $n \to \infty$, $\nabla\mathcal{L}^n(\theta; \mathbf{U}_n)$ is asymptotically distributed as $\nabla g(\theta) + n^{-1/2}\mathbf{Z}(\theta)$. Then, asymptotically, differential equations (3.9) and (3.11) are, respectively, equivalent to*

$$\dot{X}^n(t) + \nabla g(X^n(t)) + n^{-1/2}\mathbf{Z}(X^n(t)) = 0, \tag{3.16}$$

$$\ddot{X}^n(t) + \frac{3}{t}\dot{X}^n(t) + \nabla g(X^n(t)) + n^{-1/2}\mathbf{Z}(X^n(t)) = 0. \tag{3.17}$$

Applying the perturbation method for solving ordinary differential equations, we write approximation solutions of equations (3.16) and (3.17) as $X^n(t) \approx X(t) + n^{-1/2}V(t)$ and substitute it into (3.16) and (3.17). With $X(t)$ satisfying the ordinary differential equations (2.3) or (2.6), using the Taylor expansion and ignoring higher order terms, we can easily obtain equations (3.12) and (3.13) for the weak convergence limit $V(t)$ of $V^n(t)$ in the two cases, respectively.

The step process $x_\delta^n(t)$ is used to model iterates x_k^n generated from the gradient descent algorithms (3.8) and (3.10). To study their weak convergence, we need to introduce the Skorokhod space, denoted by $D([0,T])$, of all cádlág functions on $[0,T]$, equipped with the Skorokhod metric (Billingsely, 1999). Then, $x_\delta^n(t)$ lives on $D([0,T])$, and treating it as a random element in $D([0,T])$, we derive its weak convergence limit in the following theorem.

Theorem 2 *Under Assumptions A0–A4, as $\delta \to 0$ and $n \to \infty$, we have*

$$\sup_{t\in[0,T]} |x_\delta^n(t) - X^n(t)| = O_P(\delta^{1/2}|\log\delta|),$$

where $x_\delta^n(t)$ are the continuous-time step processes for discrete x_k^n generated from algorithms (3.8) and (3.10), with continuous curves $X^n(t)$ defined by the ordinary differential equations (3.9) and (3.11), for the cases of plain and accelerated gradient descent algorithms, respectively. In particular, we may select (n, δ), such that $n\delta|\log\delta|^2 \to 0$ as $\delta \to 0$ and

$n \to \infty$, and then for the selected (n, δ), $n^{1/2}[x_\delta^n(t) - X(t)]$ weakly converges to $V(t)$ on $D([0,T])$, where $X(t)$ is the solution of the ordinary differential equations (2.3) or (2.6), and $V(t)$ is given by Theorem 1. That is, $\sqrt{n}[x_\delta^n(t) - X(t)]$ and $\sqrt{n}[X^n(t) - X(t)]$ share the same weak convergence limit.

Remark 3 *There are two types of asymptotic analyses in the set up. One type is to employ continuous differential equations to model discrete iterate sequences generated from the gradient descent algorithms, which is associated with δ treated as the step size between consecutive sequence points. Another type involves the use of random objective functions in stochastic optimization, which are estimated from the sample data of size n. We refer the first and second types as computational and statistical asymptotics, respectively. The computational asymptotic analysis is that for each n, the ordinary differential equations (3.9) and (3.11)[or (3.16) and (3.17)] provide continuous solutions as the limits of discrete iterate sequences generated from algorithms (3.8) and (3.10), respectively, when δ is allowed to go to zero. Theorem 1 provides the statistical asymptotic analysis to describe the behavior difference between the sample gradient flow $X^n(t)$ and the population gradient flow $X(t)$, as the sample size n goes to infinity. Theorem 2 involves both types of asymptotics and indicates that as $\delta \to 0$ and $n \to \infty$, $x_\delta^n(t) - X^n(t)$ is of order $\delta^{1/2}|\log \delta|$. It is easy to select (n, δ) so that $x_{\delta_n}^n(t) - X^n(t)$ is of order smaller than $n^{-1/2}$. Then, $x_{\delta_n}^n(t)$ has the same asymptotic distribution $V(t)$ as $X^n(t)$.*

3.4 Unified Framework for Joint Computational and Statistical Analysis

The two types of asymptotics associated with δ and n appear to be rather different, with one for computational algorithms and one for statistical procedures. This section further elaborates regarding these analyses and provides a joint framework to unify both viewpoints. Denote the solutions of the deterministic population optimization problem (2.1) and the stochastic sample optimization problem (3.7) by $\check{\theta}$ and $\hat{\theta}_n$, respectively. In the statistical setup, $\check{\theta}$ and $\hat{\theta}_n$ represent the true estimand and its associated estimator, respectively. Using the definitions of $\check{\theta}$ and $\hat{\theta}_n$ and the Taylor expansion, we have $\nabla g(\check{\theta}) = 0$,

$$0 = \nabla \mathcal{L}^n(\hat{\theta}_n; \mathbf{U}_n) = \nabla \mathcal{L}^n(\check{\theta}; \mathbf{U}_n) + \mathbb{H}\mathcal{L}^n(\check{\theta}; \mathbf{U}_n)(\hat{\theta}_n - \check{\theta}) + \text{remainder},$$

the law of large number implies that $\mathbb{H}\mathcal{L}^n(\check{\theta}; \mathbf{U}_n)$ converges in probability to $\mathbb{H}g(\check{\theta})$ as $n \to \infty$, and Assumption A4 indicates that

$$\nabla \mathcal{L}^n(\check{\theta}; \mathbf{U}_n) = \nabla g(\check{\theta}) + n^{-1/2}\mathbf{Z}(\check{\theta}) + \text{remainder} = n^{-1/2}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z} + \text{remainder},$$

where $\mathbf{Z}$ stands for a standard normal random vector. Thus, $n^{1/2}(\hat{\theta}_n - \check{\theta})$ is asymptotically distributed as $[\mathbb{H}g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$. On the other hand, the gradient descent algorithms generate iterate sequences corresponding to $X(t)$ and $X^n(t)$, which are expected to approach the solutions of the population optimization (2.1) and the sample optimization (3.7), respectively. Hence, $X(t)$ and $X^n(t)$ must move toward $\check{\theta}$ and $\hat{\theta}_n$, respectively, and $V_n(t)$ and $V(t)$ must reach their corresponding targets $n^{1/2}(\hat{\theta}_n - \check{\theta})$ and $[\mathbb{H}g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$. Below, we provide a framework to connect $(X^n(t), X(t))$ with $(\hat{\theta}_n, \check{\theta})$ and $(V^n(t), V(t))$ with $\Big(n^{1/2}(\hat{\theta}_n - \check{\theta}), [\mathbb{H}g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}\Big)$.

Since the time interval considered thus far is $[0,T]$ for any arbitrary $T>0$, we may extend the finite time interval to $\mathbb{R}_+=[0,+\infty)$ and consider $C(\mathbb{R}_+)$, the space of all continuous functions on $\mathbb{R}_+$, to be equipped with a metric d for the topology of uniform convergence on compacta, where

$$d(b_1,b_2)=\sum_{r=1}^{\infty}2^{-r}min\left\{1,\max_{0\le s\le r}|b_1(s)-b_2(s)|\right\}.$$

The solutions $X(t)$, $X^n(t)$, $V(t)$ and $V^n(t)$ of the ordinary differential equations (2.3), (2.6), (3.9), (3.11)–(3.17) all live on $C(\mathbb{R}_+)$ and we can study their weak convergence on $C(\mathbb{R}_+)$. Similarly, we adopt the Skorokhod space $D(\mathbb{R}_+)$ equipped with the Skorokhod metric for the weak convergence study of $x_\delta^n(t)$ (Billingsely, 1999). The following theorem establishes the weak convergence of these processes on $D(\mathbb{R}_+)$ and indicates their asymptotic behaviors as $t\to\infty$.

Theorem 3 *Suppose that Assumptions A0–A4 are met, $\mathbb{H}g(\check{\theta})$ is positive definite, all eigenvalues of $\int_0^t\mathbb{H}g(X(s))ds$ diverge as $t\to\infty$, $\mathbb{H}g(\theta_1)$ and $\mathbb{H}g(\theta_2)$ commute for any $\theta_1\neq\theta_2$, and $n\delta|\log\delta|^2\to 0$ as $\delta\to 0$ and $n\to\infty$. Then, on $D(\mathbb{R}_+)$, as $\delta\to 0$ and $n\to\infty$, $V^n(t)=\sqrt{n}[X^n(t)-X(t)]$ and $\sqrt{n}[x_\delta^n(t)-X(t)]$ weakly converge to $V(t)$, $t\in[0,+\infty)$.*

Furthermore, for the plain gradient descent case, as $t\to\infty$ and $k\to\infty$, we have

(1) x_k, $x_\delta(t)$, and $X(t)$ converge to $\check{\theta}$, where x_k, $x_\delta(t)$, and $X(t)$ are defined in Section 2 (see the gradient descent algorithms and ordinary differential equations 2.2–2.6).

(2) x_k^n, $x_\delta^n(t)$, and $X^n(t)$ converge to $\hat{\theta}_n$ in probability and, thus, $V^n(t)$ converges to $\sqrt{n}(\hat{\theta}_n-\check{\theta})$ in probability, where x_k^n, $x_\delta^n(t)$, and $X^n(t)$ are defined in algorithms and equations (3.8)-(3.11).

(3) The limiting distributions of $V(t)$ as $t\to\infty$ and $\sqrt{n}(\hat{\theta}_n-\check{\theta})$ as $n\to\infty$ are identical and given by a normal distribution with mean zero and variance $[\mathbb{H}g(\check{\theta})]^{-1}\boldsymbol{\sigma}^2(\check{\theta})[\mathbb{H}g(\check{\theta})]^{-1}$, where $V(t)$, defined in the ordinary differential equations (3.12) and (3.13), is the weak convergence limit of $V^n(t)$ as $n\to\infty$.

Remark 4 *Denote the limits of the processes in Theorem 3 as $t,k\to\infty$ by the corresponding processes with t and k replaced by ∞. Then, Theorem 3 shows that for the plain gradient descent case, $x_\infty=x_\delta(\infty)=X(\infty)=\check{\theta}$, $x_\infty^n=x_\delta^n(\infty)=X^n(\infty)=\hat{\theta}_n$, $V^n(\infty)=\sqrt{n}[X^n(\infty)-X(\infty)]=\sqrt{n}[x_\delta^n(\infty)-X(\infty)]=\sqrt{n}(\hat{\theta}_n-\check{\theta})$, $V(\infty)=[\mathbb{H}g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\mathbf{Z}=[\mathbb{H}g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$; $V(t)$ weakly converges to $V(\infty)$ as $t\to\infty$, and $V^n(\infty)$ weakly converges to $V(\infty)$ as $n\to\infty$. In particular, as the process $V^n(t)$ is indexed by n and t, its limits are the same regardless the order of $n\to\infty$ and $t\to\infty$. Moreover, as $\check{\theta}=X(\infty)$ is the minimizer of the convex function $g(\cdot)$, the positive definite assumption $\mathbb{H}g(\check{\theta})=\mathbb{H}g(X(\infty))>0$ is rather reasonable; since the limit $\mathbb{H}g(X(\infty))$ of $\mathbb{H}g(X(t))$ as $t\to\infty$ has all positive eigenvalues, it is natural to expect that $\int_0^\infty\mathbb{H}g(X(s))ds$ has diverging eigenvalues. We conjecture that for the accelerated gradient descent case, similar asymptotic results might hold, as $k,t\to\infty$.*

With the augmentation of $t = \infty$, we extend $[0, +\infty)$ further to $[0, +\infty]$, consider $X(t)$, $x_\delta(t)$, $X^n(t)$, $x^n_\delta(t)$, $V(t)$, and $V^n(t)$ on $t \in [0, \infty]$ and derive the limits of $V^n(t)$ and $\sqrt{n}[x^n_\delta(t) - X(t)]$ on $[0, \infty]$ by Theorem 3. As $\delta \to 0$ and $n \to \infty$, the limiting distributions of $V^n(t) = \sqrt{n}[X^n(t) - X(t)]$ and $\sqrt{n}[x^n_\delta(t) - X(t)]$ are $V(t)$ for $t \in [0, \infty]$, where $(V^n(t), V(t))$ describe the dynamic evolution of the gradient descent algorithms for $t \in [0, \infty)$ and the statistical distribution of $\sqrt{n}(\hat{\theta}_n - \check{\theta})$ for $t = \infty$.

In a unified framework, the joint asymptotic analysis describes distribution limits of $X^n(t)$ and $x^n_\delta(t)$ from both computation and statistical viewpoints in the following manner. For $t \in [0, \infty)$, $X(t)$ and $V(t)$ represent the limiting behaviors of $X^n(t)$ and $x^n_\delta(t)$ corresponding to the computational algorithms, and $X(\infty)$ and $V(\infty)$ illustrate the limiting behaviors of the corresponding statistical decision rule $\hat{\theta}_n$. We use the following simple example to explicitly illustrate the joint asymptotic analysis.

Example 1. Suppose that $U_i = (U_{1i}, U_{2i})'$, $i = 1, \cdots, n$, are iid random vectors, where U_{1i} and U_{2i} are independent and follow a normal distribution $N(\theta_1, \tau^2)$ with mean θ_1 and known variance τ^2 and an exponential distribution with mean θ_2, respectively, and $\theta = (\theta_1, \theta_2)'$. Define $\ell(\theta; U_i) = (U_i - \theta)'(U_i - \theta)/2$, and denote by $\check{\theta}$ the true value of the parameter θ in the model. Then, $\mathcal{L}(\theta; \mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n (U_i - \theta)'(U_i - \theta)/2$, $g(\theta) = E[\ell(\theta; U_i)] = [(\theta - \check{\theta})'(\theta - \check{\theta}) + \tau^2 + \check{\theta}_2^2]/2$, $\nabla g(\theta) = \theta - \check{\theta}$, $\nabla \ell(\theta; U_i) = \theta - U_i$, $\nabla \mathcal{L}(\theta; \mathbf{U}_n) = \theta - \bar{U}_n$, and $\boldsymbol{\sigma}^2(\theta) = Var(U_1 - \theta) = \text{diag}(\tau^2, \check{\theta}_2^2)$, where $\bar{U}_n = (\bar{U}_{1n}, \bar{U}_{2n})'$ is the sample mean. It is evident that the corresponding population minimization problem (2.1) and sample minimization problem (3.7) have explicit solutions: $g(\theta)$ has the minimizer $\check{\theta}$, and $\mathcal{L}(\theta; \mathbf{U}_n)$ has the minimizer $\hat{\theta}_n = \bar{U}_n$. For this example, algorithms (2.2), (3.8), (2.4), and (3.10) yield recursive formulas $x_k = x_{k-1} + \delta(\check{\theta} - x_{k-1})$, and $x^n_k = x^n_{k-1} + \delta(\bar{U}_n - x^n_{k-1})$ for the plain gradient descent case; moreover, $x_k = x_{k-1} + \delta(\check{\theta} - y_{k-1})$, $y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$, $x^n_k = x^n_{k-1} + \delta(\bar{U}_n - y^n_{k-1})$, $y^n_k = x^n_k + \frac{k-1}{k+2}(x^n_k - x^n_{k-1})$ for the accelerated gradient descent case. While it may not be so obvious to explicitly describe the dynamic behaviors of these algorithms for the accelerated case, below we clearly illustrate the behaviors of their corresponding ordinary differential equations through closed-form expressions. First, we consider the plain gradient descent case where closed form expressions are very simple. The ordinary differential equations (2.3) and (3.9) admit the following simple solutions,

$$X(t) = (X_1(t), X_2(t))' = \check{\theta} + (x_0 - \check{\theta})e^{-t}, \;\; X^n(t) = (X^n_1(t), X^n_2(t))' = \bar{U}_n + (x^n_0 - \bar{U}_n)e^{-t},$$

$$V^n(t) = (V^n_1(t), V^n_2(t))' = \sqrt{n}(\bar{U}_n - \check{\theta})(1 - e^{-t}) + \sqrt{n}(x^n_0 - x_0)e^{-t}.$$

Note that $Z_1 = \sqrt{n}(\bar{U}_{1n} - \check{\theta}_1)/\tau \sim N(0, 1)$, $\sqrt{n}(\bar{U}_{2n}/\check{\theta}_2 - 1)$ converges in distribution to a standard normal random variable Z_2, and Z_1 and Z_2 are independent. As in Theorem 1, let $\mathbf{Z} = (Z_1, Z_2)'$, $V(t) = \Pi(t)\mathbf{Z}$, where $\Pi(t) = -(1 - e^{-t})\text{diag}(\tau, \check{\theta}_2)$ is the matrix solution of the linear differential equation (3.14) in this case. Then, for $t \in [0, \infty)$, we have

$$V^n(t) = \begin{pmatrix} \tau Z_1 \\ \check{\theta}_2 Z_2 \end{pmatrix}(1 - e^{-t}) + o_P(1) = V(t) + o_P(1),$$

which confirms that $V^n(t)$ converges to $V(t)$, as shown in Theorem 1. Further, as $t \to \infty$, $X(t) \to \check{\theta} = X(\infty)$, $X^n(t) \to \hat{\theta}_n = \bar{U}_n = X^n(\infty)$, and $V^n(t) \to V^n(\infty) = \sqrt{n}(\bar{U}_n - \check{\theta})$; as $n \to \infty$, $V^n(\infty) \to V(\infty) = \Pi(\infty)\mathbf{Z} = -(\tau Z_1, \check{\theta}_2 Z_2)'$, which provides the asymptotic

distribution of the estimator $\hat{\theta}_n = X^n(\infty)$. In summary, the behaviors of $X(t)$, $X^n(t)$, $V^n(t)$, and $V(t)$ over $[0, \infty]$ provide a complete description on the dynamic evolution of the gradient descent algorithms when applied to solve the stochastic sample optimization problem. For example, as functions of t, $X(t)$ and $X^n(t)$ can be used to describe how the sequences generated from the algorithms evolve along iterations; we may use the convergence of $V^n(t)$ to $V^n(\infty)$ and $V(t)$ to $V(\infty)$, as $t \to \infty$, to illustrate how the generated sequences converge to the target optimization solutions (estimators); the convergence of $V^n(\infty)$ to $V(\infty)$ as $n \to \infty$ may be employed to characterize the asymptotic distributions of the target optimization solutions; moreover, their relationship with n and t can be used to investigate the joint dynamic effect of data size and algorithm iterations on the computational and statistical errors in the sequences generated by the algorithms. The key signature in this case is the exponential decay factor e^{-t} that appears in all relationships. The joint asymptotic analysis with both n and t provides a unified picture for the statistical asymptotic analysis with $n \to \infty$ and the computational asymptotic analysis with $t \to \infty$.

For the accelerated case, solution $X(t)$ of the ordinary differential equation (2.6) admits an expression via the Bessel function (Watson, 1995),

$$X(t) = \check{\theta} + \frac{2(x_0 - \check{\theta})}{t} J_1(t),$$

where $x_0 = (x_{0,1}, x_{0,2})'$ is an initial value of $X(t) = (X_1(t), X_2(t))'$, and $J_1(u)$ is the Bessel function of the first kind of order one,

$$J_1(u) = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!!(2j+2)!!} u^{2j+2},$$

with the following symptotic behaviors as $u \to 0$ and $u \to \infty$,

$$J_1(u) \sim \frac{u}{2} \text{ as } u \to 0, \text{ and } J_1(u) \sim \sqrt{\frac{2}{\pi u}} cos\left(u - \frac{3\pi}{4}\right) \text{ as } u \to \infty.$$

The ordinary differential equation (3.11) has the following solution,

$$X^n(t) = \bar{U}_n + \frac{2(x_0^n - \bar{U}_n)}{t} J_1(t), \;\; V^n(t) = \sqrt{n}(\bar{U}_n - \check{\theta}) \left[1 - \frac{2}{t} J_1(t)\right] + \sqrt{n}(x_0^n - x_0) \frac{2}{t} J_1(t).$$

As in Theorem 1, let $V(t) = \Pi(t)\mathbf{Z}$, where it is relatively simple to use the properties of the Bessel function $J_1(u)$ to verify that $\Pi(t) = -[1 - 2J_1(t)/t]\text{diag}(\tau, \check{\theta}_2)$ is the matrix solution of the linear differential equation (3.15) in this case. Then, for $t \in [0, \infty)$, we have

$$V^n(t) = \begin{pmatrix} \tau Z_1 \\ \check{\theta}_2 Z_2 \end{pmatrix} \left[1 - \frac{2}{t} J_1(t)\right] + o_P(1) = V(t) + o_P(1).$$

The result matches the weak convergence of $V^n(t)$ to $V(t)$ shown in Theorem 1, and as $t \to \infty$, $X(t) \to \check{\theta} = X(\infty)$, $X^n(t) \to \hat{\theta}_n = \bar{U}_n = X^n(\infty)$, and $V^n(t) \to V^n(\infty) = \sqrt{n}(\bar{U}_n - \check{\theta})$; as $n \to \infty$, $V^n(\infty) \to V(\infty) = \Pi(\infty)\mathbf{Z} = -(\tau Z_1, \check{\theta}_2 Z_2)'$, which indicates the asymptotic distribution of the estimator $\hat{\theta}_n = X^n(\infty)$. Again, the behaviors of $X(t)$, $X^n(t)$, $V^n(t)$,

and $V(t)$ over $[0, \infty]$ describe the dynamic evolution of the accelerated gradient descent algorithm, such as how the sequences generated from the algorithm evolve along iterations (via $X(t)$ and $X^n(t)$ as functions of t) and converge to the target optimization solutions (via the convergence of $V^n(t)$ to $V^n(\infty)$ and $V(t)$ to $V(\infty)$ as $t \to \infty$), as well as connect to the asymptotic distributions of the target optimization solutions (via the convergence of $V^n(\infty)$ to $V(\infty)$ as $n \to \infty$). We find that the polynomial decay factor $\frac{2}{t}J_1(t)$ appears in all relationships for the accelerated case, and the major difference in the two cases is exponential decay $1 - e^{-t}$ for the plain case vs polynomial decay $1 - \frac{2}{t}J_1(t)$ for the accelerated case.

Remark 5 *Solving problems with large-scale data often requires some tradeoffs between statistical efficiency and computational efficiency; thus, we must account for both statistical errors and computational errors. We illustrate the potential of the joint asymptotic analysis framework for the study of the two types of errors. Note that*

$$x_\delta^n(t) - \check{\theta} = x_\delta^n(t) - \hat{\theta}_n + \hat{\theta}_n - \check{\theta},$$

where $x_\delta^n(t)$ (or x_k^n) are the values computed by the gradient descent algorithms for solving the stochastic sample optimization problem (3.7) based on sampled data, and $\check{\theta}$ is the exact solution of the deterministic population optimization problem (2.1) corresponding to the true value of θ, with $\hat{\theta}_n$ the exact solution of the sample optimization problem (3.7) corresponding to the estimator of θ. The total error $x_\delta^n(t) - \check{\theta}$ consists of computational error $x_\delta^n(t) - \hat{\theta}_n$ (of order t^{-1} or t^{-2}) and statistical error $\hat{\theta}_n - \check{\theta}$ (of order usually $n^{-1/2}$). Since $X(t)$ approaches the solution $\check{\theta} = X(\infty)$ of the population optimization problem (2.1), and $\sqrt{n}[x_\delta^n(t) - X(t)]$ weakly converges to $V(t)$, we may utilize $X(t) - X(\infty)$ and $Var(V(t))/n$ to approximate the bias and variance of $x_\delta^n(t)$ (or x_k^n), respectively. Moreover, the theory of Gaussian processes allows us to find an interval, such that with high probability $V(t)$ falls into the interval for all $t \in [0, T]$, and rescaling the interval by $n^{-1/2}$ yields an approximate simultaneous interval for $x_\delta^n(t) - X(t)$, $t \in [0, T]$. The simultaneous interval enables us to derive tests and confidence sets based on $x_\delta^n(t)$ (or x_k^n) for model parameters, and assess the closeness between iterates x_k^n and $X(t)$ and, thus, the convergence of x_k^n toward the target value $X(\infty)$.

4. Stochastic Gradient Descent via Stochastic Differential Equations for Stochastic Optimization

Solving the stochastic sample optimization problem (3.7) by the associated algorithms (3.8) and (3.10) requires evaluating the sum-gradient for all data—that is, it demands expensive evaluations of the gradients $\nabla \ell(\theta; U_i)$ from summand functions $\ell(\theta; U_i)$ with all data U_i, $i = 1, \cdots, n$. For big data problems, there is an enormous amount of data available and such evaluation of the sums of gradients for all data becomes prohibitively expensive. In order to overcome the computational burden, stochastic gradient descent uses a so-called mini-batch of data to evaluate the corresponding subset of summand functions at each iteration. Each mini-batch is a relatively small data set that is sampled from (i) the large training data set $\mathbf{U}_n$ or (ii) the underlying population distribution Q. For the case of subsampling from the original data set $\mathbf{U}_n$, it turns out that mini-batch subsampling in the stochastic gradient descent scheme is similar to the m out of n (with or without replacement) bootstraps for gradients (Bickel et al., 1997). While bootstrap resampling is widely used to draw inferences

in statistics, resampling used here and in the learning community is motivated purely from the computational purpose. Specifically, assume integer m is much smaller than n, and denote by $\mathbf{U}_m^* = (U_1^*, \cdots, U_m^*)'$ a mini-batch. For the case (ii), $\mathbf{U}_m^* = (U_1^*, \cdots, U_m^*)'$ is an i.i.d. sample taken from the distribution Q. For case (i), $\mathbf{U}_m^* = (U_1^*, \cdots, U_m^*)'$ is a subsample taken from $\mathbf{U} = (U_1, \cdots, U_n)'$, where $U_1^*, \cdots, U_m^*$ are randomly drawn with or without replacement from $U_1, \cdots, U_n$. For the case with replacement, $U_1^*, \cdots, U_m^*$ represent an i.i.d. sample taken from $\hat{Q}_n$, where $\hat{Q}_n$ is the empirical distribution of $U_1, \cdots, U_n$. In this paper, we consider the case in which mini-batches are sampled from the underlying distribution Q. Since mini-batch size m is negligible in comparison with data size n, the bootstrap sampling case (ii) can be handled via strong approximation (Csörgö and Mason, 1989; Csörgö et al., 1999; Massart, 1989; Rio, 1993a, 1993b) by converting case (ii) into the essentially proven scenario of case (i), where mini-bataches are sampled from the underlying distribution Q.

The main computational idea in the stochastic gradient descent algorithm is to replace $\mathcal{L}^n(\theta; \mathbf{U}_n)$ in algorithms (3.8) and (3.10) for solving the sample optimization problem (3.7) by a smaller sample version $\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*)$ at each iteration, where

$$\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*) = \frac{1}{m} \sum_{i=1}^{m} \ell(\theta; U_i^*).$$

We remind readers of the notion convention specified at the end of Section 1 that adds indices m and/or $*$ to notations for the corresponding processes and sequences affiliated with mini-batches in Section 4, while notations with a superscript n and without such subscripts or superscripts correspond to the stochastic sample optimization problem (3.7) and the deterministic population optimization problem (2.1), respectively.

4.1 Stochastic Gradient Descent

The stochastic gradient descent scheme replaces $\nabla\mathcal{L}^n(x_{k-1}^n; \mathbf{U}_n)$ in algorithm (3.8) by a smaller sample version at each iteration to obtain the following recursive algorithm,

$$x_k^m = x_{k-1}^m - \delta \nabla \hat{\mathcal{L}}^m(x_{k-1}^m; \mathbf{U}_{mk}^*), \tag{4.18}$$

where $\mathbf{U}_{mk}^* = (U_{1k}^*, \cdots, U_{mk}^*)'$, $k = 1, 2, \cdots$, are independent mini-batches.

We may naively follow the continuous curve approach described in Section 2 to approximate $\{x_k^m, k = 0, 1, \cdots\}$ by a smooth curve similar to the case in Section 3. However, unlike the scenario in Section 3, algorithms (4.18) [and (4.26) for the accelerated case in Section 4.2 later] are designed for the computational purpose, and they do not correspond to any optimization problem with a well-defined objective function, such as $g(\theta)$ in the population optimization problem (2.1) or $\mathcal{L}^n(\theta; \mathbf{U}_n)$ in the sample optimization problem (3.7), since samples $\mathbf{U}_{mk}^*$ used in $\hat{\mathcal{L}}^m(x_{k-1}^m; \mathbf{U}_{mk}^*)$ change with iteration k. The analysis for stochastic gradient descent will be rather different from that studied in Section 3. Below, we define a 'pseudo objective function' for the stochastic gradient descent case.

Define a mini-batch process $\mathbf{U}_m^*(t) = (U_1^*(t), \cdots, U_m^*(t))'$ and a step process $x_\delta^m(t)$, $t \geq 0$, for x_k^m in (4.18) as follows:

$$\mathbf{U}_m^*(t) = \mathbf{U}_{mk}^* \text{ and } x_\delta^m(t) = x_k^m \text{ for } k\delta \leq t < (k+1)\delta. \tag{4.19}$$

To facilitate the analysis, we adopt a convention $x_\delta^m(t) = x_0^m$ for $t < 0$. Then, $\hat{\mathcal{L}}^m(x_\delta^m(t-\delta); \mathbf{U}_m^*(t)) = \hat{\mathcal{L}}^m(x_{k-1}^m; \mathbf{U}_{mk})$ for $k\delta \le t < (k+1)\delta$. $\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*(t))$ may be treated as a counterpart of $\mathcal{L}^n(\theta; \mathbf{U}_n)$. As $m \to \infty$, $\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*(t))$ approaches $g(\theta)$ for each δ, and the stochastic gradient descent algorithm (4.18) can still solve the sample optimization problem (3.7) numerically. However, as t evolves, $\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*(t))$ changes from iteration to iteration, and depends on δ as well as m, since mini-batches change as the algorithm iterates, and the number of the mini-batches involved is determined by the time t and the step size δ. There is no single bona fide objective function here, and the 'pseudo objective function' $\mathcal{L}^m(\theta; \mathbf{U}_m^*(t))$ cannot serve the role of genuine objective functions such as $g(\theta)$ and $\mathcal{L}^n(\theta; \mathbf{U}_n)$. The approach in Sections 2 and 3 cannot be directly applied to obtain an ordinary differential equation like equation (3.9). In fact, as evident below, for this case there exists no such analog ordinary differential equation. Instead, we will derive asymptotic stochastic differential equations for algorithm (4.18). The new asymptotic stochastic differential equations may be considered as counterparts of the ordinary differential equation (3.16), which is an asymptotic version of the ordinary differential equation (3.9), but the key difference is that the asymptotic stochastic differential equations must depend on the step size δ as well as m to account for the mini-batch effect (see more details later after the stochastic differential equations (4.21) and (4.22) regarding the associated random variability). Our derivation and stochastic differential equations rely on the asymptotic behavior of $\nabla\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*(t)) - \nabla g(\theta)$ as $\delta \to 0$ and $m \to \infty$.

We need the following initial condition to guarantee the validity of our asymptotic analysis.

A5. Assume that initial values satisfy $x_0^m - x_0 = o_P((\delta/m)^{1/2})$.

We describe the asymptotic behavior of $\nabla\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*(t))$ in the following theorem.

Theorem 4 *Define a partial sum process*

$$H_\delta^m(t) = (m\delta)^{1/2} \sum_{t_k \le t} \left[\nabla\hat{\mathcal{L}}^m(x_\delta^m(t_{k-1}); \mathbf{U}_m^*(t_k)) - \nabla g(x_\delta^m(t_{k-1}))\right], \ t \ge 0, \tag{4.20}$$

where $t_k = k\delta$, $k = 0, 1, 2, \cdots$. Under Assumptions A1–A5, as $\delta \to 0$ and $m \to \infty$, we have that on $D([0,T])$, $H_\delta^m(t)$ weakly converges to $H(t) = \int_0^t \boldsymbol{\sigma}((X(u))d\mathbf{B}(u)$, $t \in [0,T]$, where $\mathbf{B}$ is a p-dimensional standard Brownian motion, $\boldsymbol{\sigma}(\theta)$ is defined in Assumption A3, and $X(t)$ is the solution of the ordinary differential equation (2.3).

Remark 6 *As discussed earlier, due to mini-batches used in algorithm (4.18), there is no corresponding optimization problem with a well-defined objective function. Consequently, we do not have any δ-free differential equation analog to the ordinary differential equation (3.16). In other words, here, there is no analog continuous modeling to derive differential equations free of δ, obtained by letting $\delta \to 0$. This may be explained from Theorem 4 as follows. It is easy to see that $H_\delta^m(t)$ is a normalized partial sum process for $[T/\delta]$ random variables $\nabla\hat{\mathcal{L}}^m(x_\delta^m(t_{k-1}); \mathbf{U}_m^*(t_k))$ whose variances are of order m^{-1}, and the weak convergence theory for partial sum processes indicates that a normalized factor $(m\delta)^{1/2}$ in the definition (4.20) is required to obtain a weak convergence limit for $H_\delta^m(t)$. On the other*

hand, to obtain an analog to the $\mathbf{Z}$ *term in equation (3.16), we need to find some kind of continuous-time limit for* $\nabla\hat{\mathcal{L}}^m(\theta;\mathbf{U}_m^*(t)) - \nabla g(\theta)$. *As* $\mathbf{U}_m^*(t)$ *is an empirical process for independent subsamples* $\mathbf{U}_{mk}^*$, $\nabla\mathcal{L}^m(\theta;\mathbf{U}_m^*(t)) - \nabla g(\theta)$ *may behave like a sort of discrete-time weighted white noise (in fact, a martingale difference sequence). Therefore, a possible continuous-time limit for* $\nabla\mathcal{L}^m(\theta;\mathbf{U}_m^*(t)) - \nabla g(\theta)$ *is related to a continuous-time white noise, which is defined as the derivative* $\dot{\mathbf{B}}(t)$ *of Brownian motion* $\mathbf{B}(t)$ *in the sense of the Dirac delta function (a generalized function). In the notation of Theorem 4, we may informally write* $H(t) = \int_0^t \boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du$ *in terms of white noise* $\dot{\mathbf{B}}(t)$, *and* $\nabla\hat{\mathcal{L}}^m(x_\delta^m(t-\delta);\mathbf{U}_m^*(t)) - \nabla g(X(t))$ *corresponds to the derivative* $\dot{H}(t) = \boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t)$ *of* $H(t)$. *While the factor* $\delta^{1/2}$ *on the right-hand side of the definition (4.20) is required to normalize a partial sum process with* $[T/\delta]$ *random variables for obtaining a weak convergence limit, from the white noise perspctive, here, we require a normalized factor* $\delta^{1/2}$ *to move from a discrete-time white noise to a continuous-time white noise. As a matter of fact, the weak convergence is very natural from the viewpoint of weak convergence for stochastic processes (Jacod and Shiryaev, 2003; He et al., 1992). Because of the white noise type stochastic variation due to different mini-batches used from iteration to iteration in algorithm (4.18), the continuous modeling for stochastic gradient descent will be* δ*-dependent, which will be given below.*

Using the definitions of $x_\delta^m(t)$ in (4.19) and $H_\delta^m(t)$ in (4.20), we recast algorithm (4.18) as

$$x_\delta^m(t+\delta) - x_\delta^m(t) = -\nabla g(x_\delta^m(t))\delta - (\delta/m)^{1/2}\,[H_\delta^m(t+\delta) - H_\delta^m(t)].$$

Theorem 4 suggests an approximation of the step process $H_\delta^m(t)$ by the continuous process $H(t)$, and we approximate the step process $x_\delta^m(t)$ by a continuous process $X_\delta^m(t)$. Taking the step size δ as dt, $H_\delta^m(t+\delta) - H_\delta^m(t)$ as $dH(t) = \boldsymbol{\sigma}(X(t))d\mathbf{B}(t)$, and $x_\delta^m(t+\delta) - x_\delta^m(t)$ as $dX_\delta^m(t)$, we transform the above difference equation into the following stochastic differential equation,

$$dX_\delta^m(t) = -\nabla g(X_\delta^m(t))dt - (\delta/m)^{1/2}\boldsymbol{\sigma}(X(t))d\mathbf{B}(t), \tag{4.21}$$

where $X(t)$ is the solution of the ordinary differential equation (2.3), and $\mathbf{B}(t)$ is a p-dimensional standard Brownian motion. The solution $X_\delta^m(t)$ of the stochastic differential equation (4.21) may be considered as a continuous approximation of x_k^m [or $x_\delta^m(t)$] generated from the stochastic gradient descent algorithm (4.18) [or (4.19)]. Since $X_\delta^m(t)$ is expected to be close to $X(t)$, and the Brownian term in (4.21) is of higher order, we may replace $X(t)$ in (4.21) by $X_\delta^m(t)$ to better mimic the recursive relationship in (4.18). In other words, we also consider the following stochastic differential equation,

$$d\check{X}_\delta^m(t) = -\nabla g(\check{X}_\delta^m(t))dt - (\delta/m)^{1/2}\boldsymbol{\sigma}(\check{X}_\delta^m(t))d\mathbf{B}(t). \tag{4.22}$$

Since we are interested in distributional behaviors, we consider solutions of the stochastic differential equations (4.21) and (4.22) in the weak sense—that is, for each fixed δ and m, there exist versions of the continuous processes $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ along with Brownian motion $\mathbf{B}(t)$ on some probability space to satisfy equations (4.21) and (4.22) (Ikeda and Watanabe, 1981).

The stochastic Brownian terms in (4.21) and (4.22) are employed to account for the random fluctuations due to the use of mini-batches for gradient estimation from iteration

to iteration in the stochastic gradient descent algorithm (4.18), where $m^{-1/2}$ and $\delta^{1/2}$ are statistical normalization factors with m for mini-batch size and $[T/\delta]$ for the total number of iterations considered in $[0,T]$ (as δ for the step size). At each iteration, we resort to a mini-batch for gradient estimation; thus, the number of iterations in $[0,T]$ is equal to the number of mini-batches used in $[0,T]$, and the factor $\delta^{1/2}$ accounts for the effect due to the total number of mini-batches used in $[0,T]$, while $m^{-1/2}$ accounts for the effect of m observations in each mini-batch.

The theorem below derives the asymptotic distribution of $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$. Let $V_\delta^m(t) = (m/\delta)^{1/2}[X_\delta^m(t) - X(t)]$ and $\check{V}_\delta^m(t) = (m/\delta)^{1/2}[\check{X}_\delta^m(t) - X(t)]$. Treating them as random elements in $C([0,T])$, we derive their weak convergence limit in the following theorem.

Theorem 5 *Under Assumptions A1–A5, as $\delta \to 0$ and $m \to \infty$, we have*

$$\sup_{0\le t\le T} |X_\delta^m(t) - \check{X}_\delta^m(t)| = O_P(m^{-1}\delta), \tag{4.23}$$

and both $V_\delta^m(t)$ and $\check{V}_\delta^m(t)$, $t \in [0,T]$, weakly converge to $V(t)$ which is a time-dependent Ornstein-Uhlenbeck process satisfying

$$dV(t) = -[\mathbb{H}g(X(t))]V(t)dt - \boldsymbol{\sigma}(X(t))d\mathbf{B}(t), \;\; V(0) = 0, \tag{4.24}$$

where $\mathbb{H}$ is the Hessian operator, $\mathbf{B}$ is a p-dimensional standard Brownian motion, $\boldsymbol{\sigma}(\theta)$ is defined in Assumption A3, and $X(t)$ is the solution of the ordinary differential equation (2.3).

Remark 7 *As $X(t)$ and $X^n(t)$ in the gradient descent case are viewed as the population and sample gradient flows, respectively, in Remark 1, we may treat $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ as stochastic gradient flows in the stochastic gradient descent case, and regard the Gaussian limiting distribution of $V_\delta^m(t)$ and $\check{V}_\delta^m(t)$ as the central limit theorem for the stochastic gradient flows, which simply refers to the gradient flow central limit theorem (GF-CLT).*

Remark 8 *Theorem 5 reveals that while $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ have the same weak convergence limit, they are an order of magnitude closer to each other than to $X(t)$. This may also be evident from the fact that the difference between the stochastic differential equations (4.21) and (4.22) is at the high order Brownian term with $X_\delta^m(t)$ replaced by its limit $X(t)$. The linear stochastic differential equation (4.24) indicates that $V(t)$ has the following explicit expression for $t \in [0,T]$ under the condition that $\mathbb{H}g(X(u))$ and $\mathbb{H}g(X(v))$ commute for all $u \neq v$,*

$$V(t) = -\int_0^t \exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right]\boldsymbol{\sigma}(X(u))d\mathbf{B}(u). \tag{4.25}$$

The step process $x_\delta^m(t)$ defined in (4.19) is the empirical process for iterates x_k^m generated from the stochastic gradient descent algorithm (4.18). Treating $x_\delta^m(t)$ as a random element in $D([0,T])$, we consider its asymptotic distribution in the following theorem.

Theorem 6 *Under Assumptions A1–A5, as $\delta \to 0$ and $m \to \infty$, we have*

$$\sup_{t \le T} |x_\delta^m(t) - X_\delta^m(t)| = o_P(m^{-1/2}\delta^{1/2}) + O_P(\delta|\log\delta|^{1/2}),$$

where $x_\delta^m(t)$ and $X_\delta^m(t)$ are defined by algorithm (4.19) and the stochastic differential equation (4.21), respectively. In particular, if we choose (δ, m), such that $m\delta|\log\delta| \to 0$ as $\delta \to 0$ and $m \to \infty$, then for the chosen (δ, m), $(m/\delta)^{1/2}[x_\delta^m(t) - X(t)]$ weakly converges to $V(t)$, where $V(t)$ is governed by the stochastic differential equation (4.24).

Remark 9 *Theorem 6 indicates that iterate sequences x_k^m generated from the stochastic gradient descent algorithm (4.18) can be very close to the continuous curves $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$, which are governed by the stochastic differential equations (4.21) and (4.22), respectively. With the appropriate choices of (δ, m), we can make the empirical process $x_\delta^m(t)$ for x_k^m to share the same weak convergence limit as the continuous curves $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$. The results enable us to study discrete algorithms by analyzing their corresponding continuous stochastic differential equations and their relatively simple weak limit.*

Remark 10 *We may consider stochastic gradient descent with momentum and/or diminishing learning rate and obtain the corresponding stochastic differential equations. For example, δ in (4.18) can be replaced by diminishing learning rate $\delta_k = \eta k^{-\alpha}$ for some $\alpha \in (0, 1)$ and constant $\eta > 0$, and the same arguments lead us to stochastic differential equations such as (4.21) and (4.22) with extra factor $(t+1)^{-\alpha}$. For the momentum case, we need to add an extra linear term in the drifts of $X_\delta^m(t)$ (or $\check{X}_\delta^m(t)$). For example, we consider stochastic gradient descent with momentum*

$$x_k^m = \gamma x_{k-1}^m - \delta\nabla\hat{\mathcal{L}}^m(x_{k-1}^m; \mathbf{U}_{mk}^*), \qquad \delta = \eta k^{-\alpha}, \;\; \gamma = 1 - \beta\eta,$$

and obtain the following stochastic differential equation,

$$dX_\delta^m(t) = -[\nabla g(X_\delta^m(t))(t+1)^{-\alpha} + \beta X_\delta^m(t)]dt - (\eta/m)^{1/2}\boldsymbol{\sigma}(X(t))(t+1)^{-\alpha}d\mathbf{B}(t).$$

4.2 Accelerated Stochastic Gradient Descent

We apply Nesterov's acceleration scheme to stochastic gradient descent by replacing $\nabla\mathcal{L}^n(y_{k-1}^n; \mathbf{U}_n)$ in algorithm (3.10) with a subsampled version at each iteration as follows:

$$x_k^m = y_{k-1}^m - \delta\nabla\hat{\mathcal{L}}^m(y_{k-1}^m; \mathbf{U}_{mk}^*), \qquad y_k^m = x_k^m + \frac{k-1}{k+2}(x_k^m - x_{k-1}^m), \tag{4.26}$$

where we use initial values x_0^m and $y_0^m = x_0^m$, and $\mathbf{U}_{mk}^* = (U_{1k}^*, \cdots, U_{mk}^*)'$, $k = 1, 2, \cdots$, are independent mini-batches.

The continuous modeling for algorithm (4.26) is conceptually in parallel with the case for the stochastic gradient descent algorithm (4.18) in Section 4.1, but the tricky part is that we face numerous mathematical difficulties in multiple steps related to singularity in the second-order stochastic differential equations involved.

As we have illustrated the continuous modeling of x_k^m generated from algorithm (4.18) in Section 4.1, it is evident that our derivation of stochastic differential equations relies on

the asymptotic behavior of $\nabla\hat{\mathcal{L}}^m(\theta;\mathbf{U}^*_m(t)) - \nabla g(\theta)$ as $\delta \to 0$ and $m \to \infty$. Similar to the cases in Sections 2 and 3.2, we define step processes

$$x^m_\delta(t) = x^m_k,\ \ y^m_\delta(t) = y^m_k,\ \ \mathbf{U}^*_m(t) = \mathbf{U}^*_{mk},\ \ \text{for}\ \ k\sqrt{\delta} \le t < (k+1)\sqrt{\delta}, \tag{4.27}$$

and approximate $x^m_\delta(t)$ by a smooth curve $X^m_\delta(t)$, which is given by (4.35) below. Note the difference between the step sizes δ and $\delta^{1/2}$ for the plain and accelerated cases, respectively, as indicated at the end of Section 2.

Theorem 7 *Define a partial sum process*

$$H^m_\delta(t) = (m^2\delta)^{1/4} \sum_{t_k \le t} \left[\nabla \hat{L}^m(y^m_\delta(t_{k-1}); \mathbf{U}^*_m(t_k)) - \nabla g(y^m_\delta(t_{k-1}))\right],\ t \ge 0, \tag{4.28}$$

where $t_k = k\delta^{1/2}$, $k = 0, 1, 2, \cdots$. Under Assumptions A1–A5, as $\delta \to 0$ and $m \to \infty$, we have that on $D([0,T])$, $H^m_\delta(t)$ weakly converges to $H(t) = \int_0^t \boldsymbol{\sigma}((X(u))d\mathbf{B}(u)$, $t \in [0,T]$, where $\mathbf{B}$ is a p-dimensional standard Brownian motion, $\boldsymbol{\sigma}(\theta)$ is defined in Assumption A3, and $X(t)$ is the solution of the ordinary differential equation (2.6).

Now, we are ready to derive the second-order stochastic differential equation corresponding to algorithm (4.26). First, note that in the population-level, the second-order ordinary differential equation (2.6) can be equivalently written as

$$\begin{cases} dX(t) = Z(t)dt, \\ dZ(t) = -\left[\frac{3}{t} Z(t) + \nabla g(X(t))\right] dt, \end{cases} \tag{4.29}$$

where $Z(t) = \dot{X}(t)$; algorithm (2.4) is equivalent to

$$\begin{cases} x_{k+1} = x_k + \sqrt{\delta}\, z_k, \\ z_{k+1} = \left[1 - \frac{3}{k+3}\right] z_k - \sqrt{\delta}\, \nabla g\left(x_k + \frac{2k+3}{k+3}\sqrt{\delta}\, z_k\right), \end{cases} \tag{4.30}$$

where $z_k = (x_{k+1} - x_k)/\sqrt{\delta}$, which can be recasted as

$$\begin{cases} \frac{x_{k+1}-x_k}{\sqrt{\delta}} = z_k, \\ \frac{z_{k+1}-z_k}{\sqrt{\delta}} = -\frac{3}{t_k + 3\sqrt{\delta}} z_k - \nabla g\left(x_k + \frac{2k+3}{k+3}\sqrt{\delta}\, z_k\right), \end{cases} \tag{4.31}$$

where we take $t_k = k\sqrt{\delta}$. We approximate (x_k, z_k) by continuous curves $(X(t), Z(t))$. Note that as $\delta \to 0$, $3\sqrt{\delta} \to 0$ and $\frac{2k+3}{k+3}\sqrt{\delta}\, z_k \to 0$ in (4.31), which are negligible relative to t_k and x_k. We take step size $\sqrt{\delta}$ as dt and turn the discrete difference equation system (4.31) into the continuous differential equation system (4.29).

Second, we replace (x_k, z_k) in (4.30) by (x^m_k, z^m_k), where $z^m_k = (x^m_{k+1} - x^m_k)/\sqrt{\delta}$, and write the sample-level algorithm (4.26) in the following equivalent forms,

$$\begin{cases} x^m_{k+1} = x^m_k + \sqrt{\delta}\, z^m_k, \\ z^m_{k+1} = \left[1 - \frac{3}{k+3}\right] z^m_k - \sqrt{\delta}\, \nabla g\left(x^m_k + \frac{2k+3}{k+3}\sqrt{\delta}\, z^m_k\right) - \frac{\delta^{1/4}}{\sqrt{m}}[H^m_\delta(t_{k+1}) - H^m_\delta(t_k)], \end{cases} \tag{4.32}$$

or equivalently,

$$\begin{cases} \frac{x_{k+1}^m - x_k^m}{\sqrt{\delta}} = z_k^m, \\ \frac{z_{k+1}^m - z_k^m}{\sqrt{\delta}} = -\frac{3}{t_k+3\sqrt{\delta}} z_k^m - \nabla g\left(x_k^m + \frac{2k+3}{k+3}\sqrt{\delta}\, z_k^m\right) - \frac{\delta^{1/4}}{\sqrt{m}} \frac{H_\delta^m(t_{k+1}) - H_\delta^m(t_k)}{\sqrt{\delta}}, \end{cases} \tag{4.33}$$

where again $t_k = k\sqrt{\delta}$. Third, we approximate (x_k^m, z_k^m) by some continuous process $(X_\delta^m(t), Z_\delta^m(t))$. As Theorem 7 suggests, we substitute $H_\delta^m(t)$ by $H(t)$, with $dH(t) = \boldsymbol{\sigma}(X(t))d\mathbf{B}(t)$; dropping the negligible terms $3\sqrt{\delta}$ and $\frac{2k+3}{k+3}\sqrt{\delta}\, z_k^m$ and taking the step size $\sqrt{\delta}$ as dt, we move from the discrete difference equation system (4.33) to the following stochastic differential equation system,

$$\begin{cases} dX_\delta^m(t) = Z_\delta^m(t)dt, \\ dZ_\delta^m(t) = -\left[\frac{3}{t} Z_\delta^m(t) + \nabla g(X_\delta^m(t))\right] dt - \frac{\delta^{1/4}}{\sqrt{m}} \boldsymbol{\sigma}(X(t))d\mathbf{B}(t), \end{cases} \tag{4.34}$$

which—together with $\dot{X}_\delta^m(t) = Z_\delta^m(t)$—is equivalent to the following second-order stochastic differential equation,

$$\ddot{X}_\delta^m(t) + \frac{3}{t}\dot{X}_\delta^m(t) + \nabla g(X_\delta^m(t)) + (\delta/m^2)^{1/4}\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0, \tag{4.35}$$

where initial conditions $X_\delta^m(0) = x_0^m$ and $\dot{X}_\delta^m(0) = 0$, $X(t)$ is defined by the ordinary differential equation (2.6), $\mathbf{B}(t)$ is a p-dimensional Brownian motion, and $\dot{\mathbf{B}}(t)$ is white noise defined as the time derivative of $\mathbf{B}(t)$ in the sense of generalized functions (Hida and Si, 2008).

As we have discussed and demonstrated for the stochastic gradient descent case in Section 4.1, similar to the stochastic differential equations (4.21) and (4.22) for the stochastic gradient descent algorithm, the second-order stochastic differential equations (4.34) and (4.35) depend on δ and m through the stochastic Brownian terms. They are used to account for the random fluctuation in mini-batches used for gradient estimation from iteration to iteration in algorithm (4.26), where $m^{-1/2}$ and $\delta^{1/4}$ are statistical normalization factors with m for the mini-batch size and $[T/\delta^{1/2}]$ for the total number of iterations considered in $[0, T]$ (as $\delta^{1/2}$ for the step size), or equivalently, the total number of mini-batches used in $[0, T]$.

The theorem below indicates that the second-order stochastic differential equation (4.35) has a unique solution. Here, again, we consider the solution in the weak sense that for each fixed δ and m, there exist continuous process $X_\delta^m(t)$ and Brownian motion $\mathbf{B}(t)$ on some probability space to satisfy equation (4.35). As in Section 4.1, process $X_\delta^m(t)$ provides a continuous approximation of iterate x_k^m given by algorithm (4.26). As $\delta \to 0$ and $m \to \infty$, the Brownian term in equation (4.35) disappears, and $X_\delta^m(t)$ approaches $X(t)$ defined by the ordinary differential equation (2.6). Define $V_\delta^m(t) = (m^2/\delta)^{1/4}[X_\delta^m(t) - X(t)]$. Then, $X(t)$, $X_\delta^m(t)$, and $V_\delta^m(t)$ live on $C([0, T])$. Treating them as random elements in $C([0, T])$, we derive a weak convergence limit of $V_\delta^m(t)$ in the following theorem.

Theorem 8 *Under Assumptions A1–A5, the second-order stochastic differential equation (4.35) has a unique solution in the weak sense, and as $\delta \to 0$ and $m \to \infty$, $V_\delta^m(t)$ weakly*

converges to a Gaussian process $V(t)$ on $C([0,T])$, where $V(t)$ is the unique solution of the following linear second-order stochastic differential equation,

$$\ddot{V}(t) + \frac{3}{t}\dot{V}(t) + [\mathbb{H}g(X(t))]V(t) + \boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0, \tag{4.36}$$

where $\mathbb{H}$ is the Hessian operator, $X(t)$ is the solution of the ordinary differential equation (2.6), $\boldsymbol{\sigma}(\theta)$ is defined in Assumption A3, $\mathbf{B}(t)$ is a p-dimensional standard Brownian motion, and initial conditions $V(0) = \dot{V}(0) = 0$.

Remark 11 *As $X(t)$ and $X^n(t)$ in the accelerated case are viewed as the population and sample Lagrangian flows in Remark 1, respectively, we treat $X_\delta^m(t)$ as a stochastic Lagrangian flow in the accelerated stochastic gradient descent case and regard the Gaussian limiting distribution of $V_\delta^m(t)$ as the central limit theorem for the stochastic Lagrangian flows, which we simply call the Lagrangian flow central limit theorem (LF-CLT).*

The step process $x_\delta^m(t)$ in definition (4.27) is the empirical process for iterate x_k^m generated from algorithm (4.26). Treating $x_\delta^m(t)$ as a random element in $D([0,T])$ we consider its asymptotic distribution in the follow theorem.

Theorem 9 *Assume that there exists $a \in (0, 1/2)$, such that $\delta m^{2/(1-2a)}$ is bounded below from zero. Then, under Assumptions A1–A5, as $\delta \to 0$ and $m \to \infty$, we have*

$$\sup_{t \le T} |x_\delta^m(t) - X_\delta^m(t)| = o_p(m^{-1/2}\delta^{1/4}) + O_p(\delta^{1/2}|\log\delta|),$$

where $x_\delta^m(t)$ and $X_\delta^m(t)$ are given by the definition (4.27) and the stochastic differential equation (4.35), respectively. In particular if we choose (δ, m) to further satisfy that $m^{1/2}\delta^{1/4}|\log\delta| \to 0$ as $\delta \to 0$ and $m \to \infty$, then for the chosen (δ, m), $(m^2/\delta)^{1/4}[x_\delta^m(t) - X(t)]$ weakly converges to $V(t)$, $t \in [0,T]$, where $V(t)$ is governed by the stochastic differential equation (4.36).

Remark 12 *As mentioned earlier, similar to the stochastic gradient descent case, the continuous modeling depends on both δ and m, and Theorems 7-9 are in parallel with Theorems 4-6. However, for the accelerated case, the challenges are largely with regard to the technical proofs. For example, we must handle second-order stochastic differential equations like (4.35) with singularity (similar to the singularity case for the ordinary differential equations (2.6) and (3.15)); it is difficult to analyze the complex recursive relationship in the accelerated stochastic gradient descent algorithm (4.26). Theorems 8 and 9 show that iterate sequences x_k^m generated from the accelerated stochastic gradient descent algorithm (4.26) may be very close to the continuous curve $X_\delta^m(t)$ governed by the stochastic differential equation (4.35), and appropriate choices of (δ, m) enable the empirical process $x_\delta^m(t)$ for iterates x_k^m to have the same weak convergence limit as the continuous curve $X_\delta^m(t)$.*

Remark 13 *The two conditions on (δ, m) are compatible. The bound condition indicates that for some generic constant C, $\delta^{a/2-1/4}m^{-1/2} < C$ or $\delta > Cm^{-2/(1-2a)}$, and the condition $m^{1/2}\delta^{1/4}|\log\delta| \to 0$ requires that δ should decrease faster than m^{-2}. For example, if we take $\delta = m^{-b}$ for any $b > 2$, then $\delta^{a/2-1/4}m^{-1/2} \le 1$ holds for $1/2 > a > 1/2 - 1/b$, and $m^{1/2}\delta^{1/4}|\log\delta| = bm^{1/2-p/4}\log m \to 0$.*

Below, we continue to study Example 1 considered in Section 3.4 under the stochastic gradient descent case.

Example 2. In Example 1, we have already calculated $\nabla g(\theta) = \theta - \check{\theta}$, $\ddot{g}(\theta) = I$, $\boldsymbol{\sigma}(\theta) = \text{diag}(\tau, \check{\theta}_2)$, and $X(t) = \check{\theta} + (x_0 - \check{\theta})e^{-t}$. For the stochastic gradient descent case, solving the stochastic differential equation (4.21), we obtain

$$\begin{aligned} X_\delta^m(t) &= x_0^m e^{-t} + \check{\theta}(1 - e^{-t}) - \sqrt{\frac{\delta}{m}} \int_0^t e^{u-t} \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u) \\ &= \check{\theta} + (x_0^m - \check{\theta})e^{-t} - \sqrt{\frac{\delta}{m}} \left(\tau \int_0^t e^{u-t} dB_1(u),\ \check{\theta}_2 \int_0^t e^{u-t} dB_2(u) \right)' \\ &= X(t) + (x_0^m - x_0)e^{-t} + \sqrt{\frac{\delta}{m}} \text{diag}(\tau, \check{\theta}_2) \Lambda(t) \\ &= X(t) + (x_0^m - x_0)e^{-t} + \sqrt{\frac{\delta}{m}} V(t), \end{aligned} \tag{4.37}$$

where $\Lambda(t) = -(\int_0^t e^{u-t} dB_1(u), \int_0^t e^{u-t} dB_2(u))$ is an Ornstein-Uhlenbeck process whose stationary distribution is a bivariate normal distribution with mean zero and variance equal to half of the identity matrix, and $V(t) = \text{diag}(\tau, \check{\theta}_2)\Lambda(t)$ is the solution of the stochastic differential equation (4.24). It is evident that the weak convergence of $V_\delta^m(t) = (m/\delta)^{1/2}[X_\delta^m(t) - X(t)]$ to $V(t)$. For the accelerated case, as we have seen, the solution of the ordinary differential equation (2.6) has the following form,

$$X(t) = \check{\theta} + \frac{2(x_0 - \check{\theta})}{t} J_1(t).$$

Below we provide solutions of the stochastic differential equations (4.35) and (4.36) in this case. First, we consider the solution $V(t)$ of the stochastic differential equation (4.36). It is easy to verify that $tV(t)$ satisfies the inhomogeneous Bessel equation of the first-order with constant term $t^3 \text{diag}(\tau, \check{\theta}_2)\dot{\mathbf{B}}(t)$, and its solution can be expressed as follows:

$$V(t) = \frac{\pi}{2} \frac{J_1(t)}{t} \int_0^t \check{J}_1(u) u^2 \text{diag}(\tau, \check{\theta}_2) d\mathbf{B}(u) - \frac{\pi}{2} \frac{\check{J}_1(t)}{t} \int_0^t J_1(u) u^2 \text{diag}(\tau, \check{\theta}_2) d\mathbf{B}(u),$$

where $J_1(t)$ and $\check{J}_1(t)$ are the Bessel functions (Gatson, 1995) of the first and second kind of order one, respectively. Since in this case, ∇g is linear, $\ddot{g} = 1$, and the stochastic differential equations (4.35) and (4.36) differ by a shift $\check{\theta}$ and a scale $m^{-1/2}\delta^{1/4}$, we can easily find

$$\begin{aligned} X_\delta^m(t) &= \check{\theta} + \frac{2(x_0^m - \check{\theta})}{t} J_1(t) + m^{-1/2}\delta^{1/4} V(t) \\ &= X(t) + \frac{2(x_0^m - x_0)}{t} J_1(t) + m^{-1/2}\delta^{1/4} V(t). \end{aligned}$$

With the initial value given in A5, it is evident that $V_\delta^m(t) = (m^2/\delta)^{1/4}[X_\delta^m(t) - X(t)]$ weakly converges to $V(t)$.

4.3 Joint Computational and Statistical Asymptotic Analysis for Stochastic Gradient Descent

As we advocate a joint asymptotic analysis framework in Section 3.4, here $X(t)$, $X_\delta^m(t)$, $V_\delta^m(t)$, and $V(t)$ provide a joint asymptotic analysis for the dynamic behaviors of the stochastic gradient descent algorithms (4.18) and (4.26). The weak convergence results established in Theorems 4-9 can be used to demonstrate the corresponding weak convergence results in $C(\mathbb{R}_+)$ and $D(\mathbb{R}_+)$. It is more complicated to consider the asymptotic analysis with $t \to \infty$ for the stochastic gradient descent case and extend the convergence results further from $[0, \infty)$ to $[0, \infty]$. As $t \to \infty$, the Brownian motion $\mathbf{B}(t)$ behaves like $(2t \log\log t)^{1/2}$, and process $H(t)$ often diverges; however, there may exist meaningful distributional limits for processes $X_\delta^m(t)$, $x_\delta^m(t)$, $V_\delta^m(t)$, and $V(t)$. For the stochastic gradient descent case, we establish the weak convergence of $V_\delta^m(t)$ to $V(t)$ on $D(\mathbb{R}_+)$ and study their asymptotic behaviors as $t \to \infty$ in the following theorem.

Theorem 10 *Suppose that Assumptions A1–A5 are met, $\mathbb{H}g(\check{\theta})$ is positive definite, all eigenvalues of $\int_0^t \mathbb{H}g(X(s))ds$ diverge as $t \to \infty$, $\mathbb{H}g(\theta_1)$ and $\mathbb{H}g(\theta_2)$ commute for any $\theta_1 \neq \theta_2$, and assume $m^{1/2}\delta|\log\delta|^{1/2} \to 0$, as $\delta \to 0$ and $m \to \infty$. Then, we obtain the following results.*

(i) As $\delta \to 0$ and $m \to \infty$, $V_\delta^m(t) = (m/\delta)^{1/2}[X_\delta^m(t) - X(t)]$ and $(m/\delta)^{1/2}[x_\delta^m(t) - X(t)]$ weakly converge to $V(t)$ on $D(\mathbb{R}_+)$.

(ii) The stochastic differential equation (4.24) admits a unique stationary distribution denoted by $V(\infty)$, where $V(\infty)$ follows a normal distribution with mean zero and covariance matrix $\Gamma(\infty)$ satisfying the following algebraic Ricatti equation,

$$\Gamma(\infty)\mathbb{H}g(X(\infty)) + \mathbb{H}g(X(\infty))\Gamma(\infty) = \boldsymbol{\sigma}^2(X(\infty)). \tag{4.38}$$

(iii) Further assume that there exists a unique stationary distribution, denoted by $X_\delta^m(\infty)$, for the stochastic differential equation (4.21). Then, as $\delta \to 0$ and $m \to \infty$, $V_\delta^m(\infty) = (m/\delta)^{1/2}[X_\delta^m(\infty) - X(\infty)]$ converges in distribution to $V(\infty)$.

Remark 14 *Similar to Theorem 3 and Remark 4, Theorem 10 indicates that for the stochastic gradient descent case, as $\delta \to 0$ and $m \to \infty$, $X_\delta^m(\infty)$ approaches $X(\infty) = \check{\theta}$, $V_\delta^m(t) = \sqrt{m/\delta}[X_\delta^m(t) - X(t)]$ converges to $V(t)$, $t \in [0, \infty]$, and $V(t)$ weakly converges to $V(\infty)$ as $t \to \infty$. Intuitively, $V(t)$ is a time-dependent Ornstein-Uhlenbeck process with stationary distribution $V(\infty)$ as its limit when $t \to \infty$, and similarly the solution $X_\delta^m(t)$ of the stochastic differential equation (4.21) may admit a stationary distribution $X_\delta^m(\infty)$ as the limiting distribution of $X_\delta^m(t)$ when $t \to \infty$ (Da Prato and Zabczyk, 1996; Gardiner, 2009). Naturally, $X_\delta^m(\infty)$ corresponds to $V(\infty)$. Mandt et al. (2017) essentially take these results as their major model assumptions to establish that stochastic gradient descent can be treated as a statistical estimation procedure in the Bayesian framework. Kushner and Yin (2003) mainly investigated the convergence of stochastic approximation algorithms, such as x_k^n in (3.8) and x_k^m in (4.18), by the so-called mean ordinary differential equations. The main ideas are described in the following manner. Random effects in the algorithms asymptotically average out, their convergence dynamics are determined effectively by the tail behaviors of the iterates from the algorithms, and the mean ordinary differential equations can asymptotically approximate the tail iterates (which are the iterates with iteration index*

k shifted toward infinity). Kushner and Yin (2003, chapter 10) also studied tail iterates centered at the true target $X(\infty) = \check{\theta}$ (instead of $X(t)$ in our case) to obtain a stationary Ornstein-Uhlenbeck process, instead of the time-dependent Ornstein-Uhlenbeck process (4.24) in this paper. The stationary Ornstein-Uhlenbeck process corresponds to $V(\infty)$ in our case, which is employed to describe the convergent behaviors of the algorithms around the actual target $X(\infty) = \check{\theta}$. In the convergence study of the Langevin Monte Carlo algorithm, Dalalyan (2017a, 2017b) and Dalalyan and Karagulyan (2019) derived explicit error bounds on the algorithm's sampling distribution with respect to the target invariant distribution of the Langevin diffusion.

Note that stochastic gradient descent is designed for the pure computational purpose, and there is no corresponding objective function nor analog of minimizer $\hat{\theta}_n$ for the stochastic gradient descent algorithm, as mini-batches (and their corresponding gradient estimators) change along iterations. It is not clear whether there are known statistical estimation methods that correspond to the limits of $x_\delta^m(t)$ and $X_\delta^m(t)$ as $t \to \infty$. Below, we provide an explicit illustration of the point through Examples 1 and 2 considered in Sections 3.4 and 4.2.

Example 3. First, from Examples 1 and 2 we evaluate

$$H(t) = \int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u) = \big(\tau B_1(t), \check{\theta}_2 B_2(t)\big)',$$

where $\boldsymbol{\sigma}(X(u)) = \mathrm{diag}(\tau, \check{\theta}_2)$, and $X(u) = \check{\theta} + (x_0 - \check{\theta})e^{-u}$. By the law of the iterated logarithm for Brownian motion, $H(t)$ diverges like $(t \log\log t)^{1/2}$ as $t \to \infty$. From (4.37), we have

$$\begin{aligned} X_\delta^m(t) &= X(t) + (x_0^m - x_0)e^{-t} + \sqrt{\frac{\delta}{m}}\mathrm{diag}(\tau, \check{\theta}_2)\Lambda(t) \\ &= X(t) + (x_0^m - x_0)e^{-t} + \sqrt{\frac{\delta}{m}}V(t), \end{aligned} \tag{4.39}$$

where $\Lambda(t) = -(\int_0^t e^{u-t}dB_1(u), \int_0^t e^{u-t}dB_2(u))$ is an Ornstein-Uhlenbeck process whose stationary distribution is a bivariate normal distribution with mean zero and variance equal to half of the identity matrix, $V(t) = \mathrm{diag}(\tau, \check{\theta}_2)\Lambda(t)$, and $V_\delta^m(t) = (m/\delta)^{1/2}(x_0^m - x_0)e^{-t} + V(t)$. As $t \to \infty$, $\Lambda(t)$ approaches its stationary distribution given by $\mathbf{Z}/\sqrt{2}$, where $\mathbf{Z} = (Z_1, Z_2)'$, and Z_1 and Z_2 are independent standard normal random variables. Using expression (4.39), we conclude that as $t \to \infty$, $X_\delta^m(t)$ converges in distribution to $X_\delta^m(\infty) = \check{\theta} + (\delta/m)^{1/2}\mathrm{diag}(\tau, \check{\theta}_2)\mathbf{Z}/\sqrt{2}$. For the initial values satisfying Assumption A5, $x_0^m - x_0 = o((\delta/m)^{1/2})$, $V_\delta^m(t)$ weakly converges to $V(t)$, and $V_\delta^m(\infty)$ weakly converges to $V(\infty) = (\tau Z_1, \check{\theta}_2 Z_2)/\sqrt{2}$.

On the other hand, the stochastic gradient descent algorithm (4.18) yields

$$x_k^m = x_{k-1}^m + \delta(\bar{U}_{mk}^* - x_{k-1}^m), \;\; k = 1, 2, \cdots,$$

where $\mathbf{U}_{mk}^* = (U_{1k}^*, \cdots, U_{mk}^*)$, $k \geq 1$, are mini-batches, and $\bar{U}_{mk}^*$ is the sample mean of $U_{1k}^*, \cdots, U_{mk}^*$. In comparison with the recursive relationship $x_k^n = x_{k-1}^n + \delta(\bar{U}_n - x_{k-1}^n)$ for the stochastic sample optimization (3.7) based on all data and $x_k = x_{k-1} + \delta(\check{\theta} - x_{k-1})$

for the deterministic population optimization (2.1), the differences are $\delta(\bar{U}^*_{mk} - \bar{U}_n)$ and $\delta(\bar{U}^*_{mk} - \check{\theta})$, respectively. In fact, for the stochastic gradient descent case, we rewrite the recursive relationship as $x^m_k = (1-\delta)x^m_{k-1} + \delta\bar{U}^*_{mk}$ and obtain

$$x^m_\delta(t) = x^m_0(1-\delta)^{[t/\delta]} + \delta\sum_{k\delta\le t}(1-\delta)^{[t/\delta]-k}\bar{U}^*_{mk}. \tag{4.40}$$

Similarly, we have

$$x^n_\delta(t) = x^n_0(1-\delta)^{[t/\delta]} + \bar{U}_n\delta\sum_{k\delta\le t}(1-\delta)^{[t/\delta]-k},\ \ x_\delta(t) = x_0(1-\delta)^{[t/\delta]} + \check{\theta}\delta\sum_{k\delta\le t}(1-\delta)^{[t\delta]-k}.$$

Letting $t\to\infty$, we obtain

$$x^n_\delta(\infty) = \bar{U}_n\delta\sum_{k=1}^{\infty}(1-\delta)^{k-1} = \bar{U}_n,\ \ x_\delta(\infty) = \check{\theta},$$

$$x^m_\delta(\infty) = \delta\lim_{\ell\to\infty}\sum_{j=0}^{\ell}(1-\delta)^j\bar{U}^*_{m,\ell-j} = \delta\sum_{k=1}^{\infty}(1-\delta)^{k-1}\bar{U}^{**}_{mk},$$

where sequence $\{\bar{U}^{**}_{mk}\}_k$ is defined as the reverse sequence of $\{\bar{U}^*_{mk}\}_k$. It is evident that $X(\infty) = x_\delta(\infty) = \check{\theta}$, $X^n(\infty) = x_\delta(\infty) = \hat{\theta}_n$, and $X^m_\delta(\infty)$ and $x^m_\delta(\infty)$ approach $\check{\theta}$ but do not correspond to any statistical estimation procedures like $\hat{\theta}_n$. For $t\in[0,\infty)$, when δ is small and m is relatively large, $x^m_\delta(t)$ can be naturally approximated by its 'limit' $x^m_0e^{-t} + \check{\theta}(1-e^{-t}) - (\delta/m)^{1/2}\int_0^t e^{u-t}\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)$, which is equal to $X^m_\delta(t)$, where the last term on the right-hand side of (4.40)—after being centered at $\check{\theta}$ and normalized by $\delta^{1/2}$—weakly converges to $\int_0^t e^{u-t}\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)$. To compare these processes, we assume initial values $x^m_0 = x^n_0 = x_0$ for simplicity. Then, we have

$$x^n_\delta(t) = x_\delta(t) + (\bar{U}_n - \check{\theta})\left[1-(1-\delta)^{[t/\delta]}\right], \tag{4.41}$$

$$\begin{aligned} x^m_\delta(t) &= x^n_\delta(t) + \delta\sum_{k\delta\le t}(1-\delta)^{[t/\delta]-k}(\bar{U}^*_{mk} - \bar{U}_n)\\ &= x_\delta(t) + (\bar{U}_n - \check{\theta})\left[1-(1-\delta)^{[t/\delta]}\right] + \delta\sum_{k\delta\le t}(1-\delta)^{[t/\delta]-k}(\bar{U}^*_{mk} - \bar{U}_n). \end{aligned} \tag{4.42}$$

The second and third terms on the right-hand side of (4.42) account for, respectively, the variability due to statistical estimation and the random fluctuation due to the use of mini-batches for gradient estimation from iteration to iteration in the stochastic gradient descent algorithm. Note that $\bar{U}_n - \check{\theta}$ and $\bar{U}^*_{mk} - \bar{U}_n$ are of orders $n^{-1/2}$ and $m^{-1/2}$, respectively. This is true even for the case that mini-batches $\mathbf{U}^*_{mk} = (U^*_{1k},\cdots,U^*_{mk})$, $k\ge 1$, are sampled from the large data set $\mathbf{U}_n$ for the bootstrap resampling case. In fact, we may resort to the strong approximation (Komlós et al., 1975, 1976; Csörgö et al., 1999; Csörgö and Mason, 1989) to obtain

$$\bar{U}^*_{mk} - \bar{U}_n = m^{-1/2}A_{mk} + O_P(m^{-1}\log m),\ \ \bar{U}_n - \check{\theta} = n^{-1/2}D_n + O_P(n^{-1}\log n), \tag{4.43}$$

where A_{mk}, $k = 1, 2, \cdots$, are almost i.i.d. random variables defined by a sequence of independent Brownian bridges on some probability spaces, with random variables D_n defined by another sequence of independent Brownian bridges on the probability spaces. As $m/n \to 0$, we easily conclude that the second term on the right-hand side of (4.42) is of higher order than the third term, where the the third term represents the cumulative mini-batch-subsampling effect up to the $k = [t/\delta]$-th iteration, with the second term for the statistical estimation error. Equations (4.41) and (4.42) show that as $m, n \to \infty$, $x_\delta^n(t)$ and $x_\delta^m(t)$ approach $\breve{\theta}$; moreover, on average both gradient descent and stochastic gradient descent algorithms remain on target, with the only difference being their random variabilities. Theorems 1 and 2 establish an order of $n^{-1/2}\mathbf{Z}$ for the random variability of the gradient descent algorithm using all data, while Theorems 5 and 6 indicate that for the stochastic gradient descent algorithm, the cumulative random fluctuation up to the $[t/\delta]$-iteration can be modeled by process $(\delta/m)^{1/2}V(t)$, where $V(t)$ given by the stochastic differential equation (4.24) (or its expression 4.25) is a time-dependent Ornstein-Uhlenbeck process that may admit a stationary distribution with mean zero and variance $\boldsymbol{\sigma}^2(X(\infty))/[2\mathbb{H}g(X(\infty))]$, factor $m^{-1/2}$ accounts for the effect of each mini-batch of size m, and factor $\delta^{1/2}$ represents the effect of the total number of mini-batches that is proportional to $1/\delta$. The normalized factor $(\delta/m)^{1/2}$ implies that while each mini-batch of size m is not as efficient as the full data sample of size n, the repetitive use of mini-batch subsampling in stochastic gradient descent utilizes more data and improves its efficiency, with the improvement represented by $\delta^{1/2}$, where $1/\delta$ is proportional to the total number of mini-batches up to the time t (or the t/δ-th iteration). In other words, repeatedly subsampling compensates the efficiency loss due to a mini-batch of small size at each iteration. Intuitively, this implies that the stochastic gradient descent algorithm invokes different mini-batches to cause some random fluctuation when moving from one iteration to another; as the number of iterations increases, subsampling improves efficiency with factor $(\delta/m)^{1/2}$ instead of $m^{-1/2}$ in order to make up loss from $n^{-1/2}$ to $m^{-1/2}$—that is, updating with the use of a large number of mini-batches can improve accuracy for the stochastic gradient descent algorithm.

4.4 Convergence Analysis of Stochastic Gradient Descent for Non-Convex Optimization

Our asymptotic results may have implications for stochastic gradient descent used in non-convex optimization particularly in deep learning. Recent studies often suggest that stochastic gradient descent algorithms can escape from saddle points and find good local minima (Jastrzębski et al., 2018; Jin et al., 2017; Keskar et al., 2017; Lee et al., 2016; Shallue et al., 2019). We provide new rigorous analysis and heuristic intuition to shed some light on the phenomenon. First, note that we can relax the convexity assumption on the objective function $g(\theta)$ for the deterministic population optimization (2.1) in Theorems 4–6 and, thus, Theorem 10 can be easily adopted to non-convex optimization with $\breve{\theta}$ being a critical point of $g(\theta)$. Suppose that stochastic gradient descent processes converge to the critical point $\breve{\theta}$. Applying the large deviation theory to the stochastic differential equations (4.21) and (4.22) corresponding to the gradient descent algorithm, we find that as δ/m goes to zero, if the critical point is a saddle point of $g(\theta)$, the continuous processes generated from the stochastic differential equations can escape from the saddle point in a polynomial time (pro-

portional to $(m/\delta)^{1/2}\log(m/\delta))$ (Kifer, 1981, Theorems 2.1-2.3; Li et al., 2017b, Theorem 3.3); in contrst, if the critical point is a local minimizer of $g(\theta)$, the continuous processes take an exponential time (proportional to $\exp\{c(m/\delta)^{1/2}\}$ for some generic constant c) to leave any given neighborhood of the local minimizer (Dembo and Zeitouni, 2010, Chapter 5; Li et al., 2017b, Theorem 3.2). We may also explain the phenomenon from the limiting distribution perspective. Theorem 5 indicates that the continuous processes $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ generated from the stochastic differential equations (4.21) and (4.22) are asymptotically the same as the deterministic solution $X(t)$ of the ordinary differential equation (2.3) plus $(\delta/m)^{1/2}V(t)$, where $V(t)$ is the solution of the stochastic differential equation (4.24). The limiting process $V(t)$ is a time-dependent Ornstein-Uhlenbeck process given by the expression (4.25). Then, we propose the following theorem for the behaviors of $g(X_\delta^m(t))$ and $g(\check{X}_\delta^m(t))$ around the critical point $\check{\theta}$.

Theorem 11 *Suppose that Assumptions A1–A5 (except for the convexity of $g(\cdot)$) are met, and the gradient descent process $X(t)$ given by the ordinary differential equation (2.3) converges to a critical point $\check{\theta}$ of $g(\cdot)$. Then, we have the following results,*

$$\begin{aligned} g(X_\delta^m(t)) &= g(X(t)) + (\delta/m)^{1/2}\nabla g(X(t))V_\delta^m(t) + \frac{\delta}{2m}[V_\delta^m(t)]'\mathbb{H}g(X(t))V_\delta^m(t) + o_P(\delta/m) \\ &= g(\check{\theta}) + \frac{1}{2}[X(t)-\check{\theta}+(\delta/m)^{1/2}V_\delta^m(t)]'\mathbb{H}g(\check{\theta})[X(t)-\check{\theta}+(\delta/m)^{1/2}V_\delta^m(t)] \\ &\quad + o_P\left(\delta/m + [X(t)-\check{\theta}]^2\right), \end{aligned} \tag{4.44}$$

$$\begin{aligned} \nabla g(X_\delta^m(t)) &= \nabla g(X(t)) + (\delta/m)^{1/2}\mathbb{H}g(X(t))V_\delta^m(t) + o_P((\delta/m)^{1/2}) \\ &= \mathbb{H}g(\check{\theta})[X(t)-\check{\theta}+(\delta/m)^{1/2}V_\delta^m(t)] + o_P\left((\delta/m)^{1/2} + |X(t)-\check{\theta}|\right), \end{aligned} \tag{4.45}$$

and the same equalities hold with X_δ^m replaced by $\check{X}_\delta^m$, where $X(t)$, $X_\delta^m(t)$, and $\check{X}_\delta^m(t)$ are the solutions of the differential equations (2.3), (4.21), and (4.22), respectively; $V_\delta^m(t) = (m/\delta)^{1/2}[X_\delta^m(t) - X(t)]$, and the equalities hold in the weak sense that we may consider $X_\delta^m(t)$, $V_\delta^m(t)$, and $V(t)$ on some common probability spaces through Skorokhod's representation.

If $\check{\theta} = X(\infty)$ is a local minimizer with positive definite $\mathbb{H}g(\check{\theta})$, then as $t \to \infty$, $V(t)$ has a limiting stationary distribution with mean zero, its covariance matrix $\Gamma(\infty)$ satisfies the algebraic Ricatti equation (4.38), and

$$E[g(X_\delta^m(t))] = g(X(t)) + \frac{\delta}{4m}tr[\boldsymbol{\sigma}^2(X(\infty))] + o(\delta/m), \tag{4.46}$$

$$E[|\nabla g(X_\delta^m(t))|^2] = |\nabla g(X(t))|^2 + \frac{\delta}{2m}tr[\boldsymbol{\sigma}^2(X(\infty))\mathbb{H}g(X(\infty))] + o(\delta/m). \tag{4.47}$$

If $\check{\theta}$ is a saddle point, $V(t)$ diverges and, thus, does not have any limiting distribution.

Theorem 11 shows that as $X(t)$ gets close to the critical point $\check{\theta}$ within the range of order $(\delta/m)^{1/2}$, $g(X_\delta^m(t))$ and $g(\check{X}_\delta^m(t))$ are approximately quadratic. As Theorem 5 indicates that $V(t)$ is the limit of $V_\delta^m(t)$, we may replace $V_\delta^m(t)$ by $V(t)$ in the expansions of $g(X_\delta^m(t))$ and $\nabla g(X_\delta^m(t))$ and find that $V(t)$ plays a key role in determining the behavior of the stochastic gradient descent algorithm. If the critical point $\check{\theta}$ is a saddle point of $g(\theta)$,

$\mathbb{H}g(\cdot)$ is non-positive definite around the saddle point; then, the time-dependent Ornstein-Uhlenbeck process $V(t)$ does not have any stationary distribution—in fact, it diverges. Thus, processes $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ have unstable behaviors around the saddle point and can make big moves, which leads them to escape from the saddle point. On the other hand, if the critical point $\check{\theta}$ is a local minimizer of $g(\theta)$, then $g(\cdot)$ may be approximately quadratic with positive definite $\mathbb{H}g(\cdot)$ around the local minimizer. Moreover, $V(t)$ has a stationary distribution, and all the processes maintain stable stochastic behaviors. In addition, the stochastic component $(\delta/m)^{1/2}V(t)$ plays a dominant role in determining the behaviors of $g(X_\delta^m(t))$ and $g(\check{X}_\delta^m(t))$ around the local minimizer. In fact, equations (4.44)-(4.47) imply that $g(X_\delta^m(t))$ and $g(\check{X}_\delta^m(t))$ behave as

$$g(X(t)) + (\delta/m)^{1/2}\nabla g(X(t))V(t) + \frac{\delta}{2m}[V(t)]'\mathbb{H}g(X(t))V(t),$$

whose mean is asymptotically equal to

$$g(X(t)) + \frac{\delta}{4m}tr[\boldsymbol{\sigma}^2(X(\infty))],$$

and $\nabla g(X_\delta^m(t))$ and $\nabla g(\check{X}_\delta^m(t))$ function in a similar manner as

$$\nabla g(X(t)) + (\delta/m)^{1/2}\mathbb{H}g(X(t))V(t),$$

which has mean $\nabla g(X(t))$ and variance asymptotically equal to

$$\frac{\delta}{2m}tr[\boldsymbol{\sigma}^2(X(\infty))\mathbb{H}g(X(\infty))].$$

First, it must be noted that the stochastic components in equations (4.44)–(4.47) depend on the learning rate δ and the batch size m only through their ratio δ/m. Second, they are characterized by the local geometry of the objective function around the local minimizer, where the local geometric characteristics include the Hessian $\mathbb{H}g(X(t))$ and the gradient covariance $\boldsymbol{\sigma}^2(X(t))$. In particular, utilizing the joint analysis along with the algebraic Ricatti equation (4.38) for the stationary covariance of the Ornstein-Uhlenbeck process, we establish equations (4.46) and (4.47) to specify how the minima found by stochastic gradient descent are influenced by four factors: the learning rate δ, batch size m, gradient covariance $\boldsymbol{\sigma}^2(\check{\theta})$, and Hessian $\mathbb{H}g(\check{\theta})$. These may have implications regarding the behavior of stochastic gradient descent for non-convex optimization. For example, equations (4.46) and (4.47) indicate that the ratio δ/m of learning rate to batch size is inversely proportional to $tr[\boldsymbol{\sigma}^2(\check{\theta})]$ for a given level of the expected loss at $\check{\theta}$ and $tr[\boldsymbol{\sigma}^2(\check{\theta})\mathbb{H}g(\check{\theta})]$ for a specific level of the expected loss gradient at $\check{\theta}$. In other words, for a larger δ/m, stochastic gradient descent tends to find a local minimum with smaller $tr[\boldsymbol{\sigma}^2(\check{\theta})]$ and $tr[\boldsymbol{\sigma}^2(\check{\theta})\mathbb{H}g(\check{\theta})]$. For a more sharp (or wide) local minimizer $\check{\theta}$, we have larger (or smaller) $\mathbb{H}g(\check{\theta})$ as well as faster (or slower) changing gradient around $\check{\theta}$, which points to a tendency of larger (or smaller) $tr[\boldsymbol{\sigma}^2(\check{\theta})]$ and $tr[\boldsymbol{\sigma}^2(\check{\theta})\mathbb{H}g(\check{\theta})]$. However, $tr[\boldsymbol{\sigma}^2(\check{\theta})]$ and $tr[\boldsymbol{\sigma}^2(\check{\theta})\mathbb{H}g(\check{\theta})]$ together do not characterize sharpness or flatness of local minimizers, and batch size alone does not determine the ratio of learning rate to batch size. Hence, our results do not directly support or contradict the claims in Keskar et al. (2017) on large/small batch methods for finding

sharp/flat local minima regarding generalization errors, which requires further theoretical and numerical studies (Shallue et al., 2019).

A case in point is a special case studied in Jastrzębski et al. (2018) that identified three factors that influence the minimum found by stochastic gradient descent. We describe the special case in the following manner. Suppose that U has a pdf $f(u;\theta)$, and the loss function $\ell(\theta;u) = -\log f(u;\theta)$. Since we take the loss as a negative log likelihood, this is the MLE case, and the gradient covariance $\boldsymbol{\sigma}^2(\theta)$ corresponds to the negative Fisher information, which in turn is equal to $E[\mathbb{H}\ell(\theta;U)] = \mathbb{H}g(\theta)$. In this case, because the stochastic differential equation (4.24) has the commutable diffusion coefficient $\boldsymbol{\sigma}(X(t))$ and drift $\mathbb{H}g(X(t)) = \boldsymbol{\sigma}^2(X(t))$, we have an explicit expression (4.25) for the time-dependent Ornstein-Uhlenbeck process $V(t)$, with the simple stationary distribution $N(0,\Gamma(\infty)) = N(0,\mathbf{I})/2$. With these explicit forms and $\mathbb{H}g(X(t)) = \boldsymbol{\sigma}^2(X(t))$, Jastrzębski et al. (2018, Equation 9) employed direct calculations for this specific example to essentially establish a special form of (4.46) with only three of the four factors regarding a relationship between the ratio of learning rate to batch size and the width of the minimum found by stochastic gradient descent. However, their corresponding formula no longer holds for the general case. In fact, for this case, given $\mathbb{H}g(X(t)) = \boldsymbol{\sigma}^2(X(t))$ and the explicit expressions of $V(t)$ and its stationary distribution, our general results can easily recover the relation in Jastrzębski et al. (2018, Equation 9). Moreover, our results are supported by additional numerical studies (Luo and Wang, 2020; Wang, 2019).

Foster et al. (2019) revealed that the complexity of stochastic optimization can be decomposed into the complexity of its corresponding deterministic population optimization and the sample complexity, where the optimization complexity represents the minimal amount of effort required to find near-stationary points, and the sample complexity of an algorithm refers to the number of training-samples required to learn a target function sufficiently well. Equation (4.45) indicates that finding near-stationary points of $g(X_\delta^m(t))$ can be converted into making $|\nabla g(X(t))|$ small and controlling $(\delta/m)^{1/2}\mathbb{H}g(X(t))V(t)$. Making $|\nabla g(X(t))|$ small means finding a near-stationary point for the corresponding deterministic population optimization. Equation (4.47) implies that the control of $(\delta/m)^{1/2}$ $\mathbb{H}g(X(t))V(t)$ can be achieved through bounding its variance—namely, imposing a bound on $tr[\boldsymbol{\sigma}^2(X(\infty))\ \mathbb{H}g(X(\infty))]$ along with selecting a sufficiently small ratio δ/m of learning rate to batch size—where the variance $tr[\boldsymbol{\sigma}^2(X(\infty))\mathbb{H}g(X(\infty))]$ is used to describe the sample complexity of the associated statistical learning problem for the time-dependent Ornstein-Uhlenbeck process. This indicates that our results are in line with Foster et al. (2019), and future study may reveal further intrinsic connection between our results and those of Foster et al. (2019).

Example 4. Consider the problem of orthogonal tensor decomposition (Ge et al., 2015; Li et al., 2016). A fourth-order tensor $\Upsilon \in \mathbb{R}^{d^4}$ has an orthogonal tenor decomposition if it can be written as

$$\Upsilon = \sum_{j=1}^{d} \boldsymbol{\alpha}_j^{\otimes 4},$$

where $\boldsymbol{\alpha}_j$'s are orthonormal vectors in $\mathbb{R}^d$ satisfying $\|\boldsymbol{\alpha}_j\| = 1$ and $\boldsymbol{\alpha}_j^\dagger\boldsymbol{\alpha}_k = 0$ for $j \neq k$, and the problem is to find tensor components $\boldsymbol{\alpha}_j$'s given such a tensor Υ. Since the tensor decomposition problem has inherent symmetry—that is, a tensor decomposition is

unique only up to component permutation and sign-flips—the symmetry property makes the corresponding optimization problem multiple local minima and, thus, non-convex.

A formulation of orthogonal tensor decomposition as an optimization problem to find one component was proposed in Frieze et al. (1996) as follows:

$$\max_{\|\boldsymbol{\beta}\|_2=1} \Upsilon(\boldsymbol{\beta},\boldsymbol{\beta},\boldsymbol{\beta},\boldsymbol{\beta}).$$

Take $\Upsilon = E[U^{\otimes 4}]$ to be the fourth-order tensor whose (i_1,i_2,i_3,i_4)-th entry is $E(U_{i_1}U_{i_2}U_{i_3}U_{i_4})$, where U is a d-dimensional random vector with distribution Q. Assume that $U = \mathbf{A}W$, where W is bounded, and has symmetric and i.i.d. components with unit variance, and $\mathbf{A}$ is an orthonormal matrix whose column vectors $\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_d$ form an orthonormal basis. Let ψ_k be the k-th moment of i.i.d. components of W, with $\psi_1 = 0$, $\psi_2 = 1$, and ψ_4 equal to its kurtosis. The optimization problem can be equivalently casted as the problem of finding components $\boldsymbol{\alpha}_j$'s into the solution to the following population optimization problem,

$$\min -sign(\psi_4-3)E[(\boldsymbol{\beta}^\dagger U)^4] = \min \sum_{j=1}^{d} -(\boldsymbol{\alpha}_j^\dagger\boldsymbol{\beta})^4 \quad \text{subject to } \|\boldsymbol{\beta}\| = 1.$$

It is well known that there is an unidentifiable tensor structure for $\psi_4 = 3$. For $\psi_4 \neq 3$, we may consider the empirical objective function $\sum_{i=1}^n -sign(\psi_4-3)(\boldsymbol{\beta}^\dagger U_i)^4$ based on available data $U_1,\cdots,U_n$, and study the corresponding stochastic optimization. The objective function of the population optimization has the gradient and Hessian in the tangent space as follows:

$$sign(\psi_4-3)\nabla\Upsilon(\boldsymbol{\beta},\boldsymbol{\beta},\boldsymbol{\beta},\boldsymbol{\beta}) = 4\left([\beta_1^2 - \|\boldsymbol{\beta}\|_4^4]\beta_1,\cdots,[\beta_d^2-\|\boldsymbol{\beta}\|_4^4]\beta_d\right),$$

$$sign(\psi_4-3)\mathbb{H}\Upsilon(\boldsymbol{\beta},\boldsymbol{\beta},\boldsymbol{\beta},\boldsymbol{\beta}) = -12\mathrm{diag}(\beta_1^2,\cdots,\beta_d^2) + 4\|\boldsymbol{\beta}\|_4^4\mathbf{I}_d.$$

Applying both gradient descent and stochastic gradient descent algorithms for solving the population and sample optimization problems, we obtain iterates x_k at the population-level and iterates x_k^m at the sample-level. As learning rate $\delta \to 0$, $x_{[t/\delta]}$ converges in probability to the population gradient flow $X(t)$ that satisfies

$$\frac{dX_i}{dt} = 4X_i\left(X_i^2 - \sum_{\ell=1}^{d} X_\ell^4\right), \quad i = 1,\cdots,d,$$

and $(m/\delta)^{1/2}[x^m_{[t/\delta]} - X(t)]$ has a weak convergence limit $V(t)$ that satisfies

$$dV(t) = -\boldsymbol{\mu}(X(t))V(t)dt - \boldsymbol{\sigma}(X(t))d\mathbf{B}(t),$$

where

$$\boldsymbol{\mu}(\boldsymbol{\beta}) = -12\mathrm{diag}(\beta_1^2,\cdots,\beta_d^2) + 4\|\boldsymbol{\beta}\|_4^4\mathbf{I}_d, \quad \boldsymbol{\sigma}^2(\boldsymbol{\beta}) = 16Cov([\boldsymbol{\beta}^\dagger\mathbf{W}]^3\mathbf{W}).$$

In order to better understand the complex gradient flow system and time-dependent Ornstein-Uhlenbeck process limit, we derive explicit expressions for the case of $d = 2$, where $X(t) = (X_1(t), X_2(t))'$ has the following closed-form solution,

$$X_1^2(t) = 0.5 + 0.5[1 + c\exp(-4t)]^{-0.5}, \quad X_2^2(t) = 1 - X_1^2(t),$$

with constant c depending on the initial value. In particular, if the initial vector$([X_1(0)]^2 < [X_2(0)]^2$ (resp. $([X_1(0)]^2 > [X_2(0)]^2)$, then $X_1(t)$ approaches 1 (resp. 0) as $t \to \infty$. Direct calculations yield

$$\boldsymbol{\sigma}^2(u)/16 = E([u_1W_1 + u_2W_2]^3\mathbf{W}\mathbf{W}^\dagger) - E([u_1W_1 + u_2W_2]^3\mathbf{W})[E([u_1W_1 + u_2W_2]^3\mathbf{W})]^\dagger,$$

where $E([u_1W_1 + u_2W_2]^3\mathbf{W}) = (u_1^3\psi_4 + 3u_1u_2^2, u_2^3\psi_4 + 3u_1^2u_2)$,

$$E([u_1W_1 + u_2W_2]^6W_1^2) = \sum_{\ell=0}^{6} C_\ell^6 u_1^\ell u_2^{6-\ell}\psi_{\ell+2}\psi_{6-\ell},$$

$$E([u_1W_1 + u_2W_2]^6W_2^2) = \sum_{\ell=0}^{6} C_\ell^6 u_2^\ell u_1^{6-\ell}\psi_{\ell+2}\psi_{6-\ell}, \text{ and}$$

$$E([u_1W_1 + u_2W_2]^6W_1W_2) = \sum_{\ell=0}^{6} C_\ell^6 u_1^\ell u_2^{6-\ell}\psi_{\ell+1}\psi_{6-\ell+1}.$$

We may simplify $\boldsymbol{\mu}(X(t))$ and $\boldsymbol{\sigma}(X(t))$ further by approximating $X(t)$ with its limit w_*(some critical point). For example, if $X(t)$ approaches critical point $w_* = (1, 0)$ (saddle point), we may approximate $\boldsymbol{\mu}(X(t))$ and $\boldsymbol{\sigma}^2(X(t))$ by $\boldsymbol{\mu}(w_*)$ and $\boldsymbol{\sigma}^2(w_*)$, where

$$\boldsymbol{\mu}(w_*) = -12\text{diag}(w_{*1}^2, w_{*2}^2) + 4\mathbf{I} = \text{diag}(-8, 4), \quad \boldsymbol{\sigma}^2(w_*) = \text{diag}(\psi_8 - \psi_4^2, \psi_6),$$

and obtain an approximate stochastic differential equation for the weak convergence limit $V(t)$ that satisfies

$$dV(t) = 4\left[-\text{diag}(-2, 1)V(t)dt - [\text{diag}(\psi_8 - \psi_4^2, \psi_6)]^{1/2}d\mathbf{B}_t\right].$$

On the other hand, if $X(t)$ approaches critical point $w_* = 2^{-1/2}(1, -1)$ (local minimizer), we have $\boldsymbol{\mu}(w_*) = -12\text{diag}(w_{*1}^2, w_{*2}^2) + 4\mathbf{I}/d = -8I/d = -4\mathbf{I}$, and $\boldsymbol{\sigma}^2(w_*)$ is equal to

$$\frac{1}{8}\begin{pmatrix} \psi_8 + 16\psi_6 + 15\psi_4^2 - 26\psi_3\psi_5 - (\psi_4 + 3)^2 & 30\psi_3\psi_5 - 12\psi_6 - 20\psi_4^2 + (\psi_4 + 3)^2 \\ 30\psi_3\psi_5 - 12\psi_6 - 20\psi_4^2 + (\psi_4 + 3)^2 & \psi_8 + 16\psi_6 + 15\psi_4^2 - 26\psi_3\psi_5 - (\psi_4 + 3)^2 \end{pmatrix},$$

$$dV(t) = -4V(t)dt - \boldsymbol{\sigma}(u_*)d\mathbf{B}_t.$$

It is evident from the stochastic differential equations that $V(t)$ has a stationary distribution for the local minimizer case, while $V(t)$ diverges for the saddle point case (in fact, the first component of $V(t)$ has a variance with exponential growth in t). Moreover, algorithms are available to compute numerical solutions to the ordinary or stochastic differential equations (Butcher, 2008; Kloeden and Platen, 1992).

4.5 Statistical Analysis of Stochastic Gradient Descent for Output Inference

There is a great current interest in the statistical analysis of stochastic gradient descent. Examples include statistical variability analysis and Bayesian inference (Chen et al., 2020; Li et al., 2018; Mandt et al., 2017; Toulis and Airoldi, 2017). Our results may have important implications on the statistical analysis of stochastic gradient descent. For the case

of stochastic gradient descent, Theorems 5 and 6 reveal that output sequence x_k^m generated from the stochastic gradient descent algorithm (4.18) is asymptotically equivalent to the continuous processes $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ generated from the stochastic differential equations (4.21) and (4.22), respectively; in turn, they are both asymptotically the same as $(\delta/m)^{1/2}V(t)$ plus the deterministic solution $X(t)$ of the ordinary differential equation (2.3), where $V(t)$ is the solution of the stochastic differential equation (4.24). The limiting process $V(t)$ is a time-dependent Ornstein-Uhlenbeck process, and its stationary distribution is a normal distribution with mean zero and covariance $\mathbf{\Gamma}(\infty)$ specified by the algebraic Ricatti equation (4.38), which is given by Theorem 10. This suggests that the statistical inference based on x_k^m can be asymptotically equivalent to the statistical inference based on discrete samples from $X(t) + (\delta/m)^{1/2}V(t)$. As $t \to \infty$, $X(t)$ converges to the true minimizer $X(\infty) = \check{\theta}$, $V(t)$ converges in distribution to $V(\infty)$, which follows its stationary distribution $N(0, \mathbf{\Gamma}(\infty))$. Thus, inferences based on x_k^m can be asymptotically equivalent to inferences based on discrete samples from the Ornstein-Uhlenbeck process with stationary distribution $N(\check{\theta}, \delta\mathbf{\Gamma}(\infty)/m)$. Below, we discuss two specific cases.

Consider the Bayesian treatment of stochastic gradient descent in Mandt et al. (2017). As described above, Theorems 5, 6, and 10 imply that outputs from the stochastic gradient descent algorithm (4.18) are asymptotically equivalent to discrete samples from the Ornstein-Uhlenbeck process with stationary distribution $N(\check{\theta}, \delta\mathbf{\Gamma}(\infty)/m)$, where $\mathbf{\Gamma}(\infty)$ is given by the algebraic Ricatti equation (4.38). Applying the Bernstein-von Mises theorem to discrete samples from the Ornstein-Uhlenbeck process, we find that the posterior distribution is asymptotically equal to a normal distribution with mean and covariance equal to the MLE of $\check{\theta}$ and the Fisher information evaluated at the MLE, respectively. Since the stochastic gradient descent outputs are asymptotically equivalent to discrete samples from the Ornstein-Uhlenbeck process, the posterior distribution based on outputs from stochastic gradient descent is asymptotically the same as the posterior distribution for the Ornstein-Uhlenbeck model; thus, it is asymptotically equal to the normal distribution. The obtained results can be employed to justify the essential inference assumptions in Mandt et al. (2017) and Li et al. (2018) that stochastic gradient descent is a stationary Ornstein-Uhlenbeck process, and the corresponding posterior distribution is Gaussian.

Another case is the average output from stochastic gradient descent. Denote by $\bar{x}_\delta^m$ the average of N outputs $x_{k_i}^m = x_\delta^m(t_{k_i})$, $i = 1, \cdots, N$, from the stochastic gradient descent algorithm (4.18), where N may depend on m and δ, and $N \to \infty$ as $\delta \to 0$ and $m \to \infty$. By Skorokhod's representation theorem and Theorems 5 and 6, we have that as $\delta \to 0$ and $m \to \infty$, the average of $x_\delta^m(t)$ has the same asymptotic distribution as the average of $X_\delta^m(t)$ and the difference between

$$(m/\delta)^{1/2}\left[N^{-1}\sum_{i=1}^N X_\delta^m(t_{k_i}) - N^{-1}\sum_{i=1}^N X(t_{k_i})\right] \text{ and } N^{-1}\sum_{i=1}^N V(t_{k_i})$$

is negligible. Note that deterministic $N^{-1}\sum_{i=1}^N X(t_{k_i})$ converges to $X(\infty) = \check{\theta}$ and that for large N, the distribution of $N^{-1/2}\sum_{i=1}^N V(t_{k_i})$ can be approximated by a normal distribution with mean zero and covariance $A^{-1}SA^{-1}$, where with notations in Theorem 10 we set

$$A = \boldsymbol{H}g(\check{\theta}) = \boldsymbol{H}g(X(\infty)), \quad S = \boldsymbol{\sigma}^2(\check{\theta}) = \boldsymbol{\sigma}^2(X(\infty)). \tag{4.48}$$

This suggests that $(m/\delta)^{1/2}(\bar{x}_\delta^m - \check{\theta})$ has an asymptotic normal distribution with mean zero and covariance $A^{-1}SA^{-1}$, and we may use outputs from stochastic gradient descent to estimate $A^{-1}SA^{-1}$ and employ the associated Ornstein-Uhlenbeck process to justify the estimation approaches. See Chen et al. (2020) and Li et al. (2018) for the covariance estimation study of stochastic gradient descent.

Note that there is a difference between the asymptotic covariances $\mathbf{\Gamma}(\infty)$ and $A^{-1}SA^{-1}$ for stochastic gradient descent described above and in the literature. For example, in Chen et al. (2020), Kushner and Yin (2003), Li et al. (2018), and Polyak and Juditsky (1992), the average output from stochastic gradient descent has asymptotic covariance $A^{-1}SA^{-1}$, while Mandt et al. (2017) and Theorem 10 indicate that the asymptotic covariance of the corresponding outputs is equal to the stationary covariance $\mathbf{\Gamma}(\infty)$ [defined in (4.38)] of the associated Ornstein-Uhlenbeck process. We explain and reconcile the difference between the covariances $A^{-1}SA^{-1}$ and $\mathbf{\Gamma}(\infty)$ in the following manner. On the one hand, although the Ornstein-Uhlenbeck process $V(t)$ approaches its normal stationary distribution with mean zero and covariance $\mathbf{\Gamma}(\infty)$, its re-scaled average $\frac{1}{\sqrt{N}}\sum_{i=1}^N V(t_{k_i}) \approx \frac{1}{\sqrt{N}}\int_0^N V(s)ds$ has asymptotic covariance $A^{-1}SA^{-1}$. Indeed, without confusion, we denote by $V(t)$ the stationary solution of the Ornstein-Uhlenbeck model as $dV(t) = -AV(t)dt + S^{1/2}d\mathbf{B}_t$ and define its auto-covariance function as $\zeta(s_1 - s_2) = E[V(s_1)(V(s_2))']$. Then, the variance of $\frac{1}{\sqrt{N}}\int_0^N V(s)ds$ is equal to

$$\begin{aligned}
&\frac{1}{N}\int_0^N \int_0^N E[V(s_1)(V(s_2))']ds_1 ds_2 = \frac{1}{N}\int_0^N \int_0^N \zeta(s_1 - s_2)ds_1 ds_2 \\
&= \int_{-\infty}^{\infty} \zeta(u)du + O(N^{-1}) = A^{-1}SA^{-1} + O(N^{-1}), \text{ as } N \to \infty,
\end{aligned}$$

where A and S are given by the expressions (4.48) and the last equality is due to the fact that

$$\begin{aligned}
&\zeta(0) = Var(V(s)) = \int_0^\infty e^{-As}Se^{-As}ds = \Gamma(\infty) \text{ satisfying } \zeta(0)A + A\zeta(0) = S, \\
&\zeta(s) = e^{-As}\zeta(0), \;\; \zeta(-s) = \zeta(0)e^{-As}, \;\; s \geq 0, \text{ and} \\
&\int_{-\infty}^{\infty} \zeta(u)du = A^{-1}\zeta(0) + \zeta(0)A^{-1} = A^{-1}[\zeta(0)A + A\zeta(0)]A^{-1} = A^{-1}SA^{-1}.
\end{aligned}$$

On the other hand, the stationary covariance $\mathbf{\Gamma}(\infty)$ is derived by treating the stochastic gradient descent recursive equation as an approximate VAR(1) model (Polyak and Juditsky, 1992), where the VAR(1) model can be expressed as $V_k = \Psi V_{k-1} + e_k$, with $\Psi = I - \delta A$, and random errors e_k has covariance $Var(e_k) = \delta S$. The VAR(1) model can be approximated by an Ornstein-Uhlenbeck model $dV(t) = -AV(t)dt + S^{1/2}d\mathbf{B}_t$. From the VAR(1) equation and stationarity, we obtain

$$Var(V_k) = \Psi Var(V_{k-1})\Psi + \delta S = Var(V_k) - \delta A Var(V_k) - \delta Var(V_k)A + \delta^2 A Var(V_k)A + \delta S.$$

Canceling out $Var(V_k)$, dividing by δ on both sides, and then letting $\delta \to 0$, we have

$$A\,Var(V_k) + Var(V_k)\,A = S,$$

which recovers the algebraic Ricatti equation (4.38) for the stationary covariance $\mathbf{\Gamma}(\infty)$ of the Ornstein-Uhlenbeck process $V(t)$. In particular, for the one-dimensional case, the AR(1) variance has an expression $Var(e_k)/(1-\Psi^2)$. Plugging $\Psi = I - \delta A$ and $Var(e_k) = \delta S$ into the variance formula, we obtain

$$\frac{Var(e_k)}{1-\Psi^2} = \frac{\delta S}{1-(1-\delta A)^2} = \frac{S}{2A} + O(\delta),$$

where the leading term $\frac{S}{2A}$ is the exact stationary variance of the Ornstein-Uhlenbeck process.

5. Proofs of Theorems

Denote by C generic constant free of (δ, m, n) whose value may change from appearance to appearance. For simplicity, we assume initial values $x_0^n = x_0^m = x_0$. In this proof section, lemmas are established under the conditions and assumptions in corresponding theorems, and we often do not repeatedly list these conditions and assumptions in the lemmas. To track processes under different circumstances and facilitate long technical arguments, we adopt the following notations and conventions.

It is often necessary to place processes and random variables on some common probability spaces. At such occasions, we often automatically change probability spaces and consider versions of the processes and the random variables on new probability spaces, without altering notations. Because of this convention and Skorokhod's representation theorem, we often switch between "convergence in probability" and "convergence in distribution." Moreover, because of the convention, when no confusion occurs, we attempt to use the same notation for random variables or processes with identical distribution.

Convention 1. We reserve x's and y's for sequences generated from gradient descent algorithms and the corresponding empirical processes, respectively, and X's for solutions of ordinary differential equations (ODEs) and stochastic differential equations (SDEs).

Convention 2. As described at the end of Section 1, to solve optimization (3.7) using gradient descent algorithms, we add superscripts n and m to notations for the associated processes and sequences based on all data in Section 3 and based on mini-batches in Section 4, respectively, while notations without any superscript are for sequences and functions corresponding to optimization (2.1).

Convention 3. We reserve V's for normalized solutions difference between differential equations associated with optimization (2.1) and optimization (3.7) under the cases for all data and mini-batches, while we reserve V without any superscript as their corresponding weak convergence limits.

Convention 4. As described at the end of Section 1, we add a superscript $*$ to notations U's associated with mini-batches, and as in Convention 2, their corresponding process notations have a superscript m.

Convention 5. We denote by $|\Psi|$ the absolute value of scalar Ψ, the Euclidean norm of vector Ψ, or the spectral norm of matrix Ψ.

5.1 Proof of Theorem 1

We show that the solution of the linear differential equations (3.12) and (3.13) are Gaussian, assuming existence and uniqueness. For equation (3.12), its solution has an expression $V(t) = \Pi_0(t)\int_0^t [\Pi_0(s)]^{-1}\mathbf{Z}(X(s))ds$, where $\Pi_0(t)$ is a p by p deterministic matrix constructed by the Magnus expansion for solving the homogeneous linear differential equation $\dot{V}(t) + [\mathbb{H}g(X(t))]V(t) = 0$ (Blanes et al., 2009). Thus, the limiting distribution of $V^n(t)$ is Gaussian. For equation (3.13) in the accelerated case, we may convert the second-order homogeneous linear differential equation $\ddot{V}(t) + \frac{3}{t}\dot{V}(t) + [\mathbb{H}g(X(t))]V(t) = 0$ into an equivalent first-order homogeneous linear differential equation system

$$\begin{pmatrix} \dot{V}(t) \\ \dot{\Xi}(t) \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ -\mathbb{H}g(X(t)) & -\frac{3}{t} \end{bmatrix} \begin{pmatrix} V(t) \\ \Xi(t) \end{pmatrix},$$

where $\Xi(t) = \dot{V}(t)$. Similar to the first-order case, we apply the Magnus expansion to solve the first-order homogeneous linear differential equation system and then show that the solution of the differential equation (3.13) linearly depends on $Z(\cdot)$. Therefore, the limiting distribution of $V^n(t)$ is also Gaussian. As a matter of fact, the theorem shows that in the special case of $\mathbf{Z}(\theta) = \sigma(\theta)\mathbf{Z}$, we have $V(t) = \Pi(t)\mathbf{Z}$ to clearly indicate the Gaussian limiting distribution.

Now, we are ready to provide detailed arguments for the accelerated case, as results for the plain case are relatively easier to show and will be established subsequently. Henceforth, for simplicity, we provide proof arguments only for the case of $\mathbf{Z}(\theta) = \boldsymbol{\sigma}(\theta)\mathbf{Z}$, as the proof for general $\mathbf{Z}(\theta)$ is essentially the same.

5.1.1 Differential Equation Derivation

With $\mathbf{U}_n = (U_1, \cdots, U_n)^\tau$, let $R^n(\theta; \mathbf{U}_n) = (R_1^n(\theta; \mathbf{U}), \cdots, R_p^n(\theta; \mathbf{U}_n))^\tau$, where

$$R_j^n(\theta; \mathbf{U}_n) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta_j}\ell(\theta; U_i) - \frac{\partial}{\partial\theta_j}g(\theta)\right], \qquad j = 1, \cdots, p.$$

Then, we obtain

$$R^n(\theta; \mathbf{U}_n) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta; U_i) - \nabla g(\theta)\right].$$

For the accelerated case, we can re-express ODE (3.11) as

$$\ddot{X}^n(t) + \frac{3}{t}\dot{X}^n(t) + \nabla g(X^n(t)) + \frac{1}{\sqrt{n}}R^n(X^n(t); \mathbf{U}_n) = 0. \tag{5.49}$$

By Lemma 5 below we obtain that $X^n(t)$ converges in probability to $X(t)$ uniformly over any finite interval. Thus, for large n, $X^n(t)$ falls into Θ_X, and Assumption A4 implies that as $n \to \infty$, $R^n(X^n(t); \mathbf{U}_n) = O_P(1)$, and $n^{-1/2}R^n(X^n(t); \mathbf{U}_n)| \to 0$. Hence, ODEs (3.11) and (5.49) both converge to ODE (2.6).

From Assumption A4 we have that $R^n(\theta; \mathbf{U}_n)$ converges in distribution to $\boldsymbol{\sigma}(\theta)\mathbf{Z}$ uniformly over $\theta \in \Theta_X$, and the generalization of Skorokhod's representation theorem in Lemma 1 below shows that there exist $\mathbf{U}_n^\dagger$ and $\mathbf{Z}_\dagger$ defined on some common probability

spaces with $\mathbf{Z}_\dagger \sim N_p(0, \mathbf{I}_p)$ and $\mathbf{U}_n^\dagger$ identically distributed as $\mathbf{U}_n$, such that as $n \to \infty$, $R^n(\theta; \mathbf{U}_n^\dagger) - \boldsymbol{\sigma}(\theta)\mathbf{Z}_\dagger = o(1)$ uniformly over $\theta \in \Theta_X$. Thus, we hold that the solution $X^n(t)$ of equations (3.11) is identically distributed as the solution $X_\dagger^n(t)$ of

$$\ddot{X}_\dagger^n(t) + \frac{3}{t}\dot{X}_\dagger^n(t) + \nabla g(X_\dagger^n(t)) + \frac{1}{\sqrt{n}} R^n(X_\dagger^n(t); \mathbf{U}_n^\dagger) = 0,$$

which in turn may be written as

$$\ddot{X}_\dagger^n(t) + \frac{3}{t}\dot{X}_\dagger^n(t) + \nabla g(X_\dagger^n(t)) + \frac{1}{\sqrt{n}} \boldsymbol{\sigma}(X_\dagger^n(t))\mathbf{Z}_\dagger + o\left(n^{-1/2}\right) = 0. \tag{5.50}$$

In particular, (5.50) is equivalent to (3.17) up to the order of $n^{-1/2}$, which implies that as $n \to \infty$, ODEs (3.11), (3.17), and (5.50) all converge to ODE (2.6), and $X_\dagger^n(t)$ almost surely converges to $X(t)$. Since the solutions of equations (3.11), (3.17), and (5.50) are defined in the distribution sense, when there is no confusion, with a little abuse of notations, we exclude index † and write equation (5.50) as

$$\ddot{X}^n(t) + \frac{3}{t}\dot{X}^n(t) + \nabla g(X^n(t)) + \frac{1}{\sqrt{n}} \boldsymbol{\sigma}(X^n(t))\mathbf{Z} + o\left(n^{-1/2}\right) = 0, \tag{5.51}$$

where $\mathbf{Z}$ is a Gaussian random vector with distribution $N_p(0, \mathbf{I}_p)$, and initial conditions $X^n(0) = x_0$ and $\dot{X}^n(0) = 0$.

The arguments for establishing Theorem 1 in Su et al. (2016) can be directly applied to establish the existence and uniqueness of the solution $X^n(t)$ to (5.49) for each n. We can employ the same arguments with $\nabla g(\cdot)$ replaced by $\mathbb{H}g(X(t))\,\Pi(t) + \boldsymbol{\sigma}(X(t))$ or $\mathbb{H}g(X(t))\,V(t) + \boldsymbol{\sigma}(X(t))\mathbf{Z}$ to show that linear differential equations (3.15) and (3.13) have unique solutions.

For the plain gradient descent case, Lemma 2 below shows that $X^n(t)$ converges to $X(t)$ uniformly over any finite interval. Similarly, we can establish that ODE (3.9) is asymptotically equivalent to ODE (3.16), and the standard ODE theory reveals that they have unique solutions.

Now, we provide a generalization of Skorokhod's representation theorem. Assumption A4 indicates that $R^n(\theta; \mathbf{U}_n)$ converges in distribution to $\mathbf{Z}(\theta)$, and Skorokhod's representation theorem enables the realizations of $R^n(\theta; \mathbf{U}_n)$ and $\mathbf{Z}(\theta)$ on some common probability spaces with almost sure convergence. The following lemma generalizes Skorokhod's representation theorem for a joint representation of $\mathbf{U}_n$ and $R^n(\theta; \mathbf{U}_n)$ along with almost sure convergence for $R^n(\theta; \mathbf{U}_n)$ and $\mathbf{Z}(\theta)$.

Lemma 1 *There exist* $\mathbf{U}_n^\dagger$ *and* $\mathbf{Z}_\dagger(\theta)$ *defined on some common probability spaces with* $\mathbf{Z}_\dagger(\theta)$ *and* $\mathbf{U}_n^\dagger$ *identically distributed as* $\mathbf{Z}(\theta)$ *and* $\mathbf{U}_n$*, respectively, such that as* $n \to \infty$*,* $R^n(\theta; \mathbf{U}_n^\dagger) - \mathbf{Z}_\dagger(\theta) = o(1)$ *uniformly over* $\theta \in \Theta_X$*.*

Proof. Our proof argument follows the construction proof of Skorokhod's representation theorem in Billingsley (1999, Theorem 6.7), with some delicate modifications involving the joint distribution of $\mathbf{U}_n$ and $R^n(\theta; \mathbf{U}_n)$ as well as its associated conditional distributions.

Let Ψ_θ be the normal distribution of $\mathbf{Z}(\theta)$. Assume random variables $\mathbf{U}_n$ are defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Denote by $\Psi_{\theta,n}$ the joint distribution of $\mathbf{U}_n$ and $R^n(\theta; \mathbf{U}_n)$,

and by $\Psi_{\theta,n,U}$ and $\Psi_{\theta,n,R}$ the marginal distributions of $\mathbf{U}_n$ and $R^n(\theta;\mathbf{U}_n)$, respectively. Let $\Xi_0(\theta),\cdots,\Xi_k(\theta)$ be the partition of $\mathbb{R}^p$ (the support of normal distribution Ψ_θ), such that (i) $\Psi_\theta(\Xi_0(\theta))<\epsilon$, (ii) the boundaries of $\Xi_0(\theta),\cdots,\Xi_k(\theta)$ have probability zero under Ψ_θ, and (iii) the diameters of $\Xi_1(\theta),\cdots,\Xi_k(\theta)$ are bounded by ϵ. Here, we use notation $\Xi_i(\theta)$ to indicate the possible dependence of the partitions on θ. For $r=1,2,\cdots$, we take $\epsilon_r=2^{-r}$ and obtain partition $\Xi_0^r(\theta),\cdots,\Xi_{k_r}^r(\theta)$. Assumption A4 indicates that $R^n(\theta;\mathbf{U}_n)$ converges in distribution to $\mathbf{Z}(\theta)$ uniformly over $\theta\in\Theta_X$, which implies that for each r there exists an integer n_r^* (free of θ), such that for $n\geq n_r^*$,

$$\Psi_{\theta,n,R}\left(\Xi_i^r(\theta)\right)\geq(1-\epsilon_r)\Psi_\theta\left(\Xi_i^r(\theta)\right),\qquad i=1,\cdots,k_r,\ \theta\in\Theta_X.$$

As in Billingsley (1999, Theorem 6.7), we can always find a probability space to support a random element with any given distribution; moreover, by passing to the appropriate large or infinite product space, we can show that there exists a probability space $(\Omega_\dagger,\mathcal{F}_\dagger,I\!P_\dagger)$ to support random variables ξ, $\mathbf{Z}_\dagger(\theta)$, $\check{\mathbf{U}}_n$, and $\mathbf{U}_{n,i}^\dagger$, and $\mathbf{\Lambda}_n$, $n,i\geq1$—all independent of each other—with the following four properties.

(i) ξ follows a uniform distribution on $[0,1]$.

(ii) $\mathbf{Z}_\dagger(\theta)$ follows a normal distribution Ψ_θ, and $\check{\mathbf{U}}_n$ has distribution $\Psi_{\theta,n,U}$.

(iii) For each $n_r^*\leq n<n_{r+1}^*$ and for each $\Xi_i^r(\theta)$ with non-zero probability under Ψ_θ, we take $\mathbf{U}_{n,i}^\dagger$ to be an independent random variable on $(\Omega_\dagger,\mathcal{F}_\dagger,I\!P_\dagger)$, such that $\mathbf{U}_{n,i}^\dagger$ and $R^n(\theta;\mathbf{U}_{n,i}^\dagger)$ jointly follow distribution $\Psi_{\theta,n}(\cdot|\Xi_i^r(\theta))$, which denotes the joint conditional distribution of $\mathbf{U}_n(\omega)$ and $R^n\left(\theta;\mathbf{U}_n(\omega)\right)$ given $R^n\left(\theta;\mathbf{U}_n(\omega)\right)\in\Xi_i^r(\theta)$ (the restriction of the joint distribution $\Psi_{\theta,n}$ on the set $\{\mathbf{u},R^n\left(\theta;\mathbf{u}\right)\in\Xi_i^r(\theta)\}$)—that is, for any Borel sets $A_1\subset\mathbb{R}^m$ and $A_2\subset\mathbb{R}^p$,

$$\begin{aligned}
&I\!P_\dagger\left[\mathbf{U}_{n,i}^\dagger(\omega_\dagger)\in A_1,R^n\left(\theta;\mathbf{U}_{n,i}^\dagger(\omega_\dagger)\right)\in A_2\right]\\
&=I\!P_\dagger\left[\check{\mathbf{U}}_n(\omega_\dagger)\in A_1,R^n\left(\theta;\check{\mathbf{U}}_n(\omega_\dagger)\right)\in A_2\middle|\,R^n\left(\theta;\check{\mathbf{U}}_n(\omega_\dagger)\right)\in\Xi_i^r(\theta)\right]\\
&=I\!P\left[\mathbf{U}_n(\omega)\in A_1,R^n\left(\theta;\mathbf{U}_n(\omega)\right)\in A_2\middle|\,R^n\left(\theta;\mathbf{U}_n(\omega)\right)\in\Xi_i^r(\theta)\right]\\
&=\Psi_{\theta,n}(A_1\times A_2|\Xi_i^r(\theta)).
\end{aligned}$$

Taking $A_1=\mathbb{R}^m$ in the above equality, we obtain the marginal result for $R^n(\theta;\mathbf{U}_{n,i}^\dagger)$,

$$\begin{aligned}
&I\!P_\dagger\left[R^n\left(\theta;\mathbf{U}_{n,i}^\dagger(\omega_\dagger)\right)\in A_2\right]=I\!P_\dagger\left[R^n\left(\theta;\check{\mathbf{U}}_n(\omega_\dagger)\right)\in A_2\middle|\,R^n\left(\theta;\check{\mathbf{U}}_n(\omega_\dagger)\right)\in\Xi_i^r(\theta)\right]\\
&=I\!P\left[R^n\left(\theta;\mathbf{U}_n(\omega)\right)\in A_2\middle|\,R^n\left(\theta;\mathbf{U}_n(\omega)\right)\in\Xi_i^r(\theta)\right]=\Psi_{\theta,n,R}\left(A_2|\Xi_i^r(\theta)\right).
\end{aligned}$$

(iv) For each $n_r^*\leq n<n_{r+1}^*$, the distribution of $\mathbf{\Lambda}_n$ is given by

$$\nu_n(A)=\epsilon_r^{-1}\sum_{i=0}^{k_r}\Psi_{\theta,n}(A\times\mathbb{R}^p|\Xi_i^r(\theta))\left[\Psi_{\theta,n,R}(\Xi_i^r(\theta))-(1-\epsilon_r)\Psi_\theta(\Xi_i^r(\theta))\right].$$

Now, we define $\mathbf{U}_n^\dagger$ on $(\Omega_\dagger,\mathcal{F}_\dagger,I\!P_\dagger)$ in the following manner. For $n<n_1^*$, take $\mathbf{U}_{n,\dagger}=\check{\mathbf{U}}_n$. For each $n_r^*\leq n<n_{r+1}^*$, define

$$\mathbf{U}_n^\dagger=\sum_{i=0}^{k_r}\mathbf{U}_{n,i}^\dagger1\{\xi\leq1-\epsilon_r,\mathbf{Z}_\dagger(\theta)\in\Xi_i^r(\theta)\}+\mathbf{\Lambda}_n1\{\xi>1-\epsilon_r\}.$$

We derive the distribution of $\mathbf{U}_n^\dagger$. For any Borel set $A_1 \subset I\!R^m$,

$$I\!P_\dagger(\mathbf{U}_n^\dagger \in A_1) = \sum_{i=0}^{k_r} I\!P_\dagger \left[\mathbf{U}_{n,i}^\dagger \in A_1, \mathbf{Z}_\dagger(\theta) \in \Xi_i^r(\theta), \xi \le 1-\epsilon_r\right]$$
$$+I\!P_\dagger(\mathbf{\Lambda}_n \in A_1, \xi > 1-\epsilon_r)$$

$$= (1-\epsilon_r)\sum_{i=0}^{k_r} I\!P_\dagger \left[\mathbf{U}_{n,i}^\dagger \in A_1\right] I\!P_\dagger \left[\mathbf{Z}_\dagger(\theta) \in \Xi_i^r(\theta)\right] + \epsilon_r \nu_n(A_1)$$
$$= (1-\epsilon_r)\sum_{i=0}^{k_r} \Psi_{\theta,n}(A_1 \times I\!R^p | \Xi_i^r(\theta)) \Psi_\theta(\Xi_i^r(\theta)) + \epsilon_r \nu_n(A_1)$$
$$= \sum_{i=0}^{k_r} \Psi_{\theta,n}(A \times I\!R^p | \Xi_i^r(\theta)) \Psi_{\theta,n,R}(\Xi_i^r(\theta))$$
$$= \sum_{i=0}^{k_r} I\!P\left[\mathbf{U}_n \in A_1, R^n\left(\theta; \mathbf{U}_n(\omega)\right) \in \Xi_i^r(\theta)\right]$$
$$= I\!P[\mathbf{U}_n \in A_1] = \Psi_{\theta,n,U}(A_1),$$

where the fifth equality is on account of the definition of distribution ν_n (which is a reverse construction). In other words, we have shown that $\mathbf{U}_n^\dagger$ is identically distributed as $\mathbf{U}_n$ for all n. Let $\Omega_{r,\dagger} = \{\xi \le 1-\epsilon_r, \mathbf{Z}_\dagger(\theta) \notin \Xi_0^r(\theta)\}$ and $\Omega_\dagger^* = \liminf_{r\to\infty} \Omega_{r,\dagger}$. Then, $I\!P_\dagger(\Omega_{r,\dagger}^*) > 1-2\epsilon_r$, and an application of the Borel-Cantelli lemma leads to $I\!P_\dagger(\Omega_\dagger^*) = 1$. For $n_r^* \le n < n_{r+1}^*$, on set $\Omega_{r,\dagger}$, $R^n(\theta; \mathbf{U}_n^\dagger)$ and $\mathbf{Z}_\dagger(\theta)$ fall into the same set $\Xi_i^r(\theta)$, whose diameter is less than ϵ_r. Thus, on $\Omega_\dagger^*$, $R^n(\theta; \mathbf{U}_n^\dagger)$ a.s. converges to $\mathbf{Z}_\dagger(\theta)$ uniformly over $\theta \in \Theta_X$. ■

5.1.2 Weak Convergence and Tightness

To prove the weak convergence of $V_n(t)$ to $V(t)$, we need to establish the usual finite-dimensional convergence as well as uniform tightness (or stochastic equicontinuity) (Kim and Pollard, 1990, Theorem 2.3; Pollard, 1988; Van der Vaart and Wellner, 2000). We establish finite-dimensional convergence below.

For the accelerated case, taking a difference between ODEs (2.6) and (5.50), we have

$$[\ddot{X}_\dagger^n(t) - \ddot{X}(t)] + \frac{3}{t}[\dot{X}_\dagger^n(t) - \dot{X}(t)] + \nabla[g(X_\dagger^n(t)) - g(X(t))] + \frac{1}{\sqrt{n}}\boldsymbol{\sigma}(X_\dagger^n(t))\mathbf{Z}_\dagger = o\left(n^{-1/2}\right).$$

Let $V_\dagger^n(t) = \sqrt{n}[X_\dagger^n(t) - X(t)]$. As $n \to \infty$, $X_\dagger^n(t) \to_{a.s.} X(t)$, $\boldsymbol{\sigma}(X_\dagger^n(t)) = \boldsymbol{\sigma}(X(t)) + o(1)$, and $\nabla[g(X_\dagger^n(t)) - g(X(t))] = \nabla^2 g(X(t))[X_\dagger^n(t) - X(t)] + o(X_\dagger^n(t) - X(t))$; thus, $V_\dagger^n(t)$ satisfies

$$\ddot{V}_\dagger^n(t) + \frac{3}{t}\dot{V}_\dagger^n(t) + \boldsymbol{I\!H} g(X(t)) V_\dagger^n(t) + \boldsymbol{\sigma}(X(t))\mathbf{Z}_\dagger = o(1).$$

As $n \to \infty$, $V_\dagger^n(t)$ almost surely converge to the unique solution $V_\dagger(t)$ of the following linear differential equation,

$$\ddot{V}_\dagger(t) + \frac{3}{t}\dot{V}_\dagger(t) + [\mathbb{H}g(X(t))]V_\dagger(t) + \boldsymbol{\sigma}(X(t))\mathbf{Z}_\dagger = 0,$$

where $X(t)$ is the solution of equation (2.6), random variable $\mathbf{Z}_\dagger \sim N_p(0, \mathbf{I}_p)$, and initial conditions $V_\dagger(0) = \dot{V}_\dagger(0) = 0$. As $V(t)$ and $V_\dagger(t)$ are governed by the equations with the same form but identically distributed random coefficients $\mathbf{Z}$ and $\mathbf{Z}_\dagger$, we easily see that $V(t)$ and $V_\dagger(t)$ are identically distributed.

The almost sure convergence of $V_\dagger^n(t)$ to $V_\dagger(t)$ implies the joint convergence of $(V_\dagger^n(t_1), \cdots, V_\dagger^n(t_k))$ to $(V_\dagger(t_1), \cdots, V_\dagger(t_k))$ for any integer k and any $t_1, \cdots, t_k \in \mathbb{R}_+$. From the identical distributions of $X^n(t)$ with $X_\dagger^n(t)$, $V^n(t)$ with $V_\dagger^n(t)$, and $V(t)$ with $V_\dagger(t)$ we immediately conclude that $(V^n(t_1), \cdots, V^n(t_k))$ converges in distribution to $(V(t_1), \cdots, V(t_k))$. This establishes the finite-dimensional distribution convergence of $V^n(t)$ to $V(t)$.

For the plain gradient descent case, an application of the similar argument to ODEs (2.3) and (3.16) can establish the finite-dimensional convergence.

Now, we show the tightness of $V_n(t)$. To establish the tightness of $V_n(t)$ on $[0, T]$, we need to show that for any $\varepsilon > 0$ and $\eta > 0$, there exists a positive constant δ, such that

$$\limsup_{n\to\infty} P\left[\sup_{(t_1,t_2)\in\mathcal{T}(T,\delta)} |V_n(t_1) - V_n(t_2)| > \eta\right] < \varepsilon, \tag{5.52}$$

where $\mathcal{T}(T, \delta) = \{(t_1, t_2), t_1, t_2 \in \mathbb{R}_+, \max(t_1, t_2) \le T, |t_1 - t_2| < \delta\}$. The tightness of $V_n(t)$ on $\mathbb{R}_+$ requires the above result for any $T < \infty$.

Note that as (5.52) requires only some probability evaluation, with the abuse of notations, we drop index † and work on equation (5.51).

5.1.3 Weak Convergence Proof for the Plain Gradient Descent Case

Lemma 2 *For any given $T > 0$, we have*

$$\max_{t\in[0,T]} |X^n(t) - X(t)| = O_P(n^{-1/2}).$$

Proof. From ODEs (2.3) and (3.9), we obtain

$$\dot{X}^n(t) - \dot{X}(t) = -[\nabla g(X^n(t)) - \nabla g(X(t))] - n^{-1/2}R^n(X^n(t); \mathbf{U}_n),$$

and using Assumptions A1 and A2, we arrive at

$$|\nabla g(X^n(t)) - \nabla g(X(t))| \le L|X^n(t) - X(t)|,$$
$$n^{-1/2}|R^n(X^n(t); \mathbf{U}_n) - R^n(X(t); \mathbf{U}_n)| \le \left(n^{-1}\sum_{i=1}^n h_1(U_i) + L\right)|X^n(t) - X(t)|.$$

Combining them we have

$$\begin{aligned} |X^n(t) - X(t)| \;\le\; & n^{-1/2}\int_0^t |R^n(X(s); \mathbf{U}_n)|ds \\ & + \left(n^{-1}\sum_{i=1}^n h_1(U_i) + 2L\right)\int_0^t |X^n(s) - X(s)|ds, \end{aligned}$$

and an application of Gronwall's inequality leads to

$$|X^n(t) - X(t)| \le n^{-1/2} \int_0^t |R^n(X(s); \mathbf{U}_n)| ds$$
$$+ n^{-1/2} \left(n^{-1} \sum_{i=1}^n h_1(U_i) + 2L \right) \int_0^t e^{\left(n^{-1}\sum_{i=1}^n h_1(U_i)+2L\right)u} du \int_0^u |R^n(X(s); \mathbf{U}_n)| ds,$$

which implies that

$$\max_{t\in[0,T]} |X^n(t) - X(t)| \le n^{-1/2} \int_0^T |R^n(X(s); \mathbf{U}_n)| ds$$
$$+ n^{-1/2} \left(n^{-1} \sum_{i=1}^n h_1(U_i) + 2L \right) \int_0^T e^{\left(n^{-1}\sum_{i=1}^n h_1(U_i)+2L\right)u} du \int_0^u |R^n(X(s); \mathbf{U}_n)| ds.$$

Since Assumptions A3 and A4 indicate that $\sup_t |R^n(X(t); \mathbf{U}_n)| \sim \sup_t |\boldsymbol{\sigma}(X(t))\mathbf{Z}| = O_P(1)$, and $n^{-1}\sum_{i=1}^n h_1(U_i)$ converges in probability to $E[h_1(U)] < \infty$, the above inequality shows that $\max_{t\in[0,T]} |X^n(t) - X(t)| = O_P(n^{-1/2})$. ■

Lemma 3 *For any given $T > 0$, $V^n(t)$ is stochastically equicontinuous on [0,T].*

Proof. Lemma 2 has shown $\max_{t\in[0,T]} |V^n(t)| = O_P(1)$. From ODEs (2.3) and (3.9), we have

$$\dot{V}^n(t) = \sqrt{n}[\dot{X}^n(t) - \dot{X}(t)] = -\sqrt{n}[\nabla g(X^n(t)) - \nabla g(X(t))] - R^n(X^n(t); \mathbf{U}_n),$$
$$|\dot{V}^n(t)| \le \sqrt{n}|\nabla g(X^n(t)) - \nabla g(X(t))| + |R^n(X^n(t); \mathbf{U}_n)|$$
$$\le L\sqrt{n}|X^n(t)) - X(t)| + |R^n(X^n(t); \mathbf{U}_n)|.$$

Lemma 2 shows that $\sqrt{n}|X^n(t)) - X(t)| = O_P(1)$, which indicates that for large n, $X^n(t)$ falls into Θ_X and assumption A4 in turn implies $|\sup_t R^n(X^n(t); \mathbf{U}_n)| \sim \sup_t |\boldsymbol{\sigma}(X^n(t))\mathbf{Z}| = O_P(1)$. Substituting these into the upper bound of $|\dot{V}^n(t)|$, we prove that $\max_{t\in[0,T]} |\dot{V}^n(t)| = O_P(1)$. Combining this with $\max_{t\in[0,T]} |V^n(t)| = O_P(1)$ shown in Lemma 2, we immediately establish the lemma. ■

Proof of Theorem 1 for the plain gradient descent case. The same perturbation argument in Section 5.1.2 can be used to show finite-dimensional distribution convergence of $V^n(t)$ to $V(t)$ for simple ODE (3.9) in the plain gradient descent case. With the tightness of $V^n(t)$ shown in Lemma 3 together with the finite distribution convergence, we immediately prove the weak convergence of $V^n(t)$ to $V(t)$ in the plain gradient descent case. ■

5.1.4 Weak Convergence Proof for the Accelerated Case

We can use the same proof as that given in Su et al. (2016, Theorem 1) to show that ODE (3.11) has a unique solution for each n and $\mathbf{U}_n$. While the proof arguments in Su et al. (2016, Theorem 1) mainly require local ODE properties, like those near a neighbor of zero, our weak convergence analysis needs to investigate global behaviors of processes generated from SDEs and ODEs with random coefficients. First, we extend and refine a

few local results for the global case and establish several preparatory lemmas for proving weak convergence in the theorem.

Given an interval $\mathcal{I} = [s,t]$ and a process $Y(t)$, define for $a \in (0,1]$,

$$M_a(s,t;Y) = M_a(\mathcal{I};Y) = \sup_{u\in[s,t]} \left| \frac{\dot{Y}(u) - \dot{Y}(s)}{(u-s)^a} \right|. \tag{5.53}$$

In the proof of Theorem 1, we take $a = 1$ and use $M_1(s,t;Y)$. We need $M_a(s,t;Y)$ with $a < 1$ subsequently in the proof of Theorem 8.

Lemma 4 *For $X(t)$ and $X^n(t)$, we have*

$$M_1(s,t;X) \le \frac{1}{1 - L(t-s)^2/6} \left[\left(\frac{3}{s} + \frac{L(t-s)}{2} \right) |\dot{X}(s)| + |\nabla g(X(s))| \right],$$

$$M_1(s,t;X^n) \le \frac{1}{1 - [\zeta(\mathbf{U}_n) + 2L](t-s)^2/6}$$

$$\left[\left(\frac{3}{s} + \frac{[\zeta(\mathbf{U}_n) + 2L](t-s)}{2} \right) |\dot{X}^n(s)| + |\nabla g(X^n(s))| + n^{-1/2} |R^n(X^n(s);\mathbf{U}_n)| \right],$$

$$M_1(s,t;X^n - X) \le \frac{1}{1 - [\zeta(\mathbf{U}_n) + 2L](t-s)^2/6} \Big\{ (3/s + (t-s)[\zeta(\mathbf{U}_n) + 2L]) |\dot{X}^n(s) - \dot{X}(s)|$$

$$+ [2\zeta(\mathbf{U}_n) + 5L] |X^n(s) - X(s)| + n^{-1/2} |R^n(X^n(s);\mathbf{U}_n)|$$

$$+ n^{-1/2} \sup_{u\in[s,t]} |R^n(X(u);\mathbf{U}_n) - R^n(X(s);\mathbf{U}_n)| \Big\},$$

when $s > 0$ and $t - s < \sqrt{6/[\zeta(\mathbf{U}_n) + 2L]}$, $\zeta(\mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n h_1(U_i)$, and $h_1(\cdot)$ is given in Assumption A1. In particular, for $s = 0$,

$$M_1(0,t;X) \le \frac{|\nabla g(x_0)|}{1 - Lt^2/6}, \quad M_1(0,t;X^n) \le \frac{|\nabla g(x_0)| + n^{-1/2}|R^n(x_0;\mathbf{U}_n)|}{1 - [\zeta(\mathbf{U}_n) + 2L]t^2/6},$$

$$M_1(0,t;X^n - X) \le \frac{n^{-1/2}}{1 - [\zeta(\mathbf{U}_n) + 2L]t^2/6}$$

$$\left[|R^n(x_0;\mathbf{U}_n)| + \sup_{u\in[0,t]} |R^n(X(u);\mathbf{U}_n) - R^n(x_0;\mathbf{U}_n)| \right].$$

Proof. Because of similarity, we provide proof arguments only for $M_1(s,t;X^n - X)$. As $V^n(t) = \sqrt{n}[X^n(t) - X(t)]$, we have $M_1(s,t;V^n) = \sqrt{n} M_1(s,t;X^n - X)$ and establish the inequality for $M_1(s,t;V^n)$. $V^n(t)$ satisfies the differential equation

$$\ddot{V}^n(t) + \frac{3}{t}\dot{V}^n(t) + \sqrt{n}\nabla[g(X^n(t)) - g(X(t))] + R^n(X^n(t);\mathbf{U}_n) = 0. \tag{5.54}$$

Let

$$H(t;V^n) = \sqrt{n}\nabla[g(X^n(t)) - g(X(t))] + R^n(X^n(t);\mathbf{U}_n),$$

and $J(s,t;H,V^n)=\int_s^t u^3[H(u;V^n)-H(s;V^n)]du$. Then, we have

$$|H(t;V^n)-H(s;V^n)|\le\sqrt{n}|\nabla[g(X^n(t))-g(X^n(s))-g(X(t))+g(X(s))]|$$
$$+|R^n(X^n(t);\mathbf{U}_n)-R^n(X^n(s);\mathbf{U}_n)|.$$

As in the proof of Lemma 2, using Assumptions A1 and A2, we obtain

$$\sqrt{n}\,|\nabla[g(X^n(t))-g(X^n(s))-g(X(t))+g(X(s))]|$$
$$\le L\sqrt{n}|X^n(t)-X(t)|+L\sqrt{n}|X^n(s)-X(s)|,$$
$$|R^n(X^n(t);\mathbf{U}_n)-R^n(X^n(s);\mathbf{U}_n)|\le|R^n(X^n(t);\mathbf{U}_n)-R^n(X(t);\mathbf{U}_n)|$$
$$+|R^n(X^n(s);\mathbf{U}_n)-R^n(X(s);\mathbf{U}_n)|+|R^n(X(t);\mathbf{U}_n)-R^n(X(s);\mathbf{U}_n)|,$$
$$|R^n(X^n(u);\mathbf{U}_n)-R^n(X(u);\mathbf{U}_n)|\le[\zeta(\mathbf{U}_n)+L]\sqrt{n}|X^n(u)-X(u)|,$$
$$\sqrt{n}[X^n(t)-X(t)]=V^n(t)=\int_s^t[\dot V^n(u)-\dot V^n(s)]du+V^n(s)+(t-s)\dot V^n(s).$$

Putting together these results, we arrive at

$$|H(t;V^n)-H(s;V^n)|\le[\zeta(\mathbf{U}_n)+2L]$$
$$\left[\int_s^t|\dot V^n(u)-\dot V^n(s)|du+2|V^n(s)|+(t-s)|\dot V^n(s)|\right]+|R^n(X(t);\mathbf{U}_n)-R^n(X(s);\mathbf{U}_n)|.$$

On the other hand, we have

$$\int_s^t|\dot V^n(u)-\dot V^n(s)|du\le\int_s^t(u-s)\frac{|\dot V^n(u)-\dot V^n(s)|}{u-s}du\le\int_s^t(u-s)M_1(s,t;V^n)du$$
$$=\frac{M_1(s,t;V^n)(t-s)^2}{2},$$
$$\int_s^t M_1(s,u;V^n)u^3(u-s)^2du/2\le M_1(s,t;V^n)t^3(t-s)^3/6.$$

Substituting the above inequalities into the upper bound for $|H(u;V^n)-H(s;V^n)|$ and the definition of $J(s,t;H,V^n)$, we conclude that

$$|J(s,t;H,V^n)|\le t^3(t-s)\sup_{u\in[s,t]}|R^n(X(u);\mathbf{U}_n)-R^n(X(s);\mathbf{U}_n)|$$
$$+[\zeta(\mathbf{U}_n)+2L]\left\{M_1(s,t;V^n)t^3(t-s)^3/6+[2|V^n(s)|+(t-s)|\dot V^n(s)|]t^3(t-s)\right\}.$$

ODE (5.54) is equivalent to

$$\frac{t^3\dot V^n(t)}{dt}=-t^3H(t;V^n),\text{which implies that}$$
$$t^3\dot V^n(t)-s^3\dot V^n(s)=-\int_s^t u^3H(u;V^n)du=-\frac{t^4-s^4}{4}H(s;V^n)-J(s,t;H,V^n),$$
$$\frac{\dot V^n(t)-\dot V^n(s)}{t-s}=-\frac{t^3-s^3}{t^3(t-s)}\dot V^n(s)-\frac{t^4-s^4}{4t^3(t-s)}H(s;V^n)-\frac{J(s,t;H,V^n)}{t^3(t-s)},$$

and using the upper bound of $|J(s,t;H,V^n)|$ and algebraic manipulation, we obtain

$$\begin{aligned}
&\frac{|\dot{V}^n(t)-\dot{V}^n(s)|}{t-s} \le \frac{t^3-s^3}{t^3(t-s)}|\dot{V}^n(s)| + \frac{t^4-s^4}{4t^3(t-s)}|H(s;V^n)| + \frac{|J(s,t;H,V^n)|}{t^3(t-s)} \\
&\le \frac{t^2+st+s^2}{t^3}|\dot{V}^n(s)| + \frac{(t^2+s^2)(t+s)}{4t^3}|H(s;V^n)| \\
&+ [\zeta(\mathbf{U}_n)+2L]\left[M_1(s,t;V^n)\frac{(t-s)^2}{6} + 2|V^n(s)| + (t-s)|\dot{V}^n(s)|\right] \\
&+ \sup_{u\in[s,t]}|R^n(X(u);\mathbf{U}_n) - R^n(X(s);\mathbf{U}_n)|.
\end{aligned}$$

As the above inequality holds for any $t > s$, we replace t by v, take the maximum over $v \in [s,t]$, and use the definition of $M_1(s,t;V^n)$ (which is increasing in t) to obtain

$$\begin{aligned}
&M_1(s,t;V^n) \le \frac{3}{s}|\dot{V}^n(s)| + |H(s;V^n)| + [\zeta(\mathbf{U}_n)+2L]M_1(s,t;V^n)\frac{(t-s)^2}{6} \\
&+ [\zeta(\mathbf{U}_n)+2L][2|V^n(s)| + (t-s)|\dot{V}^n(s)|] + \sup_{u\in[s,t]}|R^n(X(u);\mathbf{U}_n) - R^n(X(s);\mathbf{U}_n)|, \\
&\le \frac{3}{s}|\dot{V}^n(s)| + L|V^n(s)| + |R^n(X^n(s);\mathbf{U}_n)| + [\zeta(\mathbf{U}_n)+2L]M_1(t,s;V^n)\frac{(t-s)^2}{6} \\
&+ [\zeta(\mathbf{U}_n)+2L][2|V^n(s)| + (t-s)|\dot{V}^n(s)|] + \sup_{u\in[s,t]}|R^n(X(u);\mathbf{U}_n) - R^n(X(s);\mathbf{U}_n)|.
\end{aligned}$$

Further, solving for $M_1(s,t;V^n)$ yields

$$\begin{aligned}
&M_1(s,t;V^n) \le \frac{1}{1-[\zeta(\mathbf{U}_n)+2L](t-s)^2/6}\Big\{(3/s + (t-s)[\zeta(\mathbf{U}_n)+2L])|\dot{V}^n(s)| \\
&+[2\zeta(\mathbf{U}_n)+5L]|V^n(s)| + |R^n(X^n(s);\mathbf{U}_n)| + \sup_{u\in[s,t]}|R^n(X(u);\mathbf{U}_n) - R^n(X(s);\mathbf{U}_n)|\Big\},
\end{aligned}$$

when $s > 0$ and $t - s < \sqrt{6/[\zeta(\mathbf{U}_n)+2L]}$. If $s = 0$, we replace the coefficient $3/s$ by $1/t$ in the above inequality, and $V^n(0) = \dot{V}^n(0) = 0$, $X^n(0) = X(0) = x_0$. Then, we obtain

$$M_1(0,t;V^n) \le \frac{1}{1-[\zeta(\mathbf{U}_n)+2L]t^2/6}\left[|R^n(x_0;\mathbf{U}_n)| + \sup_{u\in[0,t]}|R^n(X(u);\mathbf{U}_n) - R^n(x_0;\mathbf{U}_n)|\right],$$

which specifically implies that

$$\sup_{t\le\sqrt{3/[\zeta(\mathbf{U}_n)+2L]}}\frac{|\dot{X}^n(t)-\dot{X}(t)|}{t} \le 2n^{-1/2}\left[2|R^n(x_0;\mathbf{U}_n)| + \sup_{u\in[0,t]}|R^n(X(u);\mathbf{U}_n)|\right] \to 0,$$

that is, $\dot{X}^n(t) \to \dot{X}(t)$ uniformly over $\left[0, \sqrt{3/[\zeta(\mathbf{U}_n)+2L]}\right]$. ■

Lemma 5 *For any given $T > 0$, we have*

$$\max_{t\in[0,T]}|X^n(t) - X(t)| = O_P(n^{-1/2}), \quad \max_{t\in[0,T]}|V^n(t)| = O_P(1),$$

$$\max_{t\in[0,T]}|\dot{X}^n(t) - \dot{X}(t)| = O_P(n^{-1/2}), \quad \max_{t\in[0,T]}|\dot{V}^n(t)| = O_P(1).$$

Proof. As $V^n(t) = \sqrt{n}[X^n(t) - X(t)]$, we need to establish the results for $X^n(t) - X(t)$ only. Since, as $n \to \infty$, $\zeta(\mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n h_1(U_i) \to E(h_1(U))$. Divide the interval $[0,T]$ into $N = \left\lfloor T\sqrt{[E(h_1(U)) + 2L]/3} \right\rfloor + 1$ number of subintervals with length close to $\sqrt{3/[E(h_1(U)) + 2L]}$ (except for the last one), and denote them by $\mathcal{I}_i = [s_{i-1}, s_i]$, $i = 1, \cdots, N$ (with $s_0 = 0$, $s_N = T$, $\mathcal{I}_1 = [0, s_1]$, $\mathcal{I}_N = [s_{N-1}, T]$). First, for $t \in \mathcal{I}_1$, from Lemma 4 we have

$$|\dot{X}^n(t) - \dot{X}(t)| \le |\mathcal{I}_1| M_1(\mathcal{I}_1; X^n - X) \le Cn^{-1/2}\left[|R^n(x_0; \mathbf{U}_n)| + |R^n(X(s_1); \mathbf{U}_n)|\right],$$

$$|X^n(t) - X(t)| \le \int_{\mathcal{I}_1} |\dot{X}^n(u) - \dot{X}(u)| du \le Cn^{-1/2}\left[|R^n(x_0; \mathbf{U}_n)| + |R^n(X(s_1); \mathbf{U}_n)|\right].$$

Assumption A4 implies that $R^n(x_0; \mathbf{U}_n) = O_P(1)$, and $R^n(X(s_1); \mathbf{U}_n) = O_P(1)$; thus, the upper bounds of $\dot{X}^n(t) - \dot{X}(t)$ and $X^n(t) - X(t)$ over $t \in \mathcal{I}_1$ are $O_P(n^{-1/2})$.

For $t \in \mathcal{I}_i$, $i = 2, \cdots, N$, from Lemma 4 we have

$$\begin{aligned}
&|\dot{X}^n(t) - \dot{X}(t) - \dot{X}^n(s_{i-1}) + \dot{X}(s_{i-1})| \le |\mathcal{I}_i| M_1(\mathcal{I}_i; X^n - X) \\
&\le C\left[[\zeta(\mathbf{U}_n) + C_1]|\dot{X}^n(s_{i-1}) - \dot{X}(s_{i-1})| + [\zeta(\mathbf{U}_n) + C_2]|X^n(s_{i-1}) - X(s_{i-1})|\right] \\
&+ Cn^{-1/2}\left\{|R^n(X^n(s_{i-1}); \mathbf{U}_n)| + 2\sup_{u \ge 0}|R^n(X(u); \mathbf{U}_n)|\right\},
\end{aligned}$$

and

$$\begin{aligned}
&|X^n(t) - X(t)| \le |X^n(s_{i-1}) - X(s_{i-1})| + |\mathcal{I}_i||\dot{X}^n(s_{i-1}) - \dot{X}(s_{i-1})| \\
&+ \int_{\mathcal{I}_i} |\dot{X}^n(u) - \dot{X}(u) - \dot{X}^n(s_{i-1}) + \dot{X}(s_{i-1})| du \\
&\le C\left[[\zeta(\mathbf{U}_n) + C_1]|\dot{X}^n(s_{i-1}) - \dot{X}(s_{i-1})| + [\zeta(\mathbf{U}_n) + C_2]|X^n(s_{i-1}) - X(s_{i-1})|\right] \\
&+ Cn^{-1/2}\left\{|R^n(X^n(s_{i-1}); \mathbf{U}_n)| + 2\sup_{u \ge 0}|R^n(X(u); \mathbf{U}_n)|\right\}.
\end{aligned}$$

We use the above two inequalities to prove by induction that the upper bounds of $X^n(t) - X(t)$ and $\dot{X}^n(t) - \dot{X}(t)$ on $[0,T]$ are $O_P(n^{-1/2})$, and the upper bounds of $X^n(t) - X(t)$ and $\dot{X}^n(t) - \dot{X}(t)$ on $[0,T]$ are $O_P(n^{-1/2})$. Assume that the upper bounds of $X^n(t) - X(t)$ and $\dot{X}^n(t) - \dot{X}(t)$ on $\cup_{j=1}^{i-1}\mathcal{I}_j$ are $O_P(n^{-1/2})$. Note that N is free of n, and by induction we show that the upper bounds of $X^n(t) - X(t)$ and $\dot{X}^n(t) - \dot{X}(t)$ over $t \le s_{i-1}$ are $O_P(n^{-1/2})$—in particular $X^n(s_{i-1}) \to X(s_{i-1})$ in probability—and, thus, assumption A4 indicates that $R^n(X^n(s_{i-1}); \mathbf{U}_n) = O_P(1)$, and $\sup_{u \ge 0}|R^n(X(u); \mathbf{U}_n)| = O_P(1)$. The above-mentioned two inequalities immediately reveal that their upper bounds on $\mathcal{I}_i$ are also $O_P(n^{-1/2})$. Hence, we establish that the bounds of $X^n(t) - X(t)$ and $\dot{X}^n(t) - \dot{X}(t)$ on $\cup_{j=1}^N \mathcal{I}_j = [0,T]$ are $O_P(n^{-1/2})$. ■

Lemma 6 *$V^n(t)$ is stochastically equicontinuous on $[0,T]$.*

Proof. Lemma 5 indicates that $\max_{t\in[0,T]}|V^n(t)| = O_P(1)$ and $\max_{t\in[0,T]}|\dot{V}^n(t)| = O_P(1)$, which implies that $V^n(t)$ is stochastically equicontinuous on $[0,T]$. ■

Proof of Theorem 1. Lemma 6 along with the finite distribution convergence immediately lead to that as $n \to \infty$, $V^n(t)$ weakly converges to $V(t)$. ■

5.2 Proof of Theorem 2

We prove Theorem 2 in two subsections for the plain and accelerated cases.

5.2.1 Proof for the Plain Gradient Descent Case

Lemma 7 *For the case of the plain gradient descent algorithm, we have*

$$\max_{t\in[0,T]} |x_\delta^n(t) - X^n(t)| = O_P(\delta), \quad \max_{k\le T\delta^{-1}} |x_k^n - X^n(k\delta)| = O_P(\delta),$$

where $\{x_k^n\}$ is generated from algorithm (3.8), with $x_\delta^n(t)$ its continuous-time step process, and $X^n(t)$ the solution of ODE (3.9).

Proof. Algorithm (3.8) is the Euler scheme for solving ODE (3.9), and we apply the standard ODE theory to obtain the global approximation error for the Euler scheme. First, by Assumption A1, we have that $\nabla\mathcal{L}^n(\theta;\mathbf{U}_n)$ is Lipschtiz in θ with Lipschitz constant $\frac{1}{n}\sum_{i=1}^n h_1(U_i)$, which converges in probability to $E[h_1(U)] < \infty$. On the other hand, taking derivatives on both sides of ODE (3.9), we obtain

$$\ddot{X}^n(t) = -\nabla^2\mathcal{L}^n(X^n(t);\mathbf{U}_n)\dot{X}^n(t) = \nabla^2\mathcal{L}^n(X^n(t);\mathbf{U}_n)\nabla\mathcal{L}^n(X^n(t);\mathbf{U}_n).$$

Using Lemma 2, we conclude that for large n, $X^n(t)$ falls into Θ_X; thus, Assumption A4 indicates that $\sup_t|\nabla^\kappa\mathcal{L}^n(X^n(t);\mathbf{U}_n)| \sim \sup_t|\nabla^\kappa g(X^n(t)) + n^{-1/2}\boldsymbol{\sigma}_k(X^n(t))\mathbf{Z}_\kappa| = O_P(1)$, where $\mathbf{Z}_\kappa$ are standard normal random variables. Combining these results, we obtain $\sup_{t\in[0,T]}|\ddot{X}^n(t)| = O_P(1)$. An application of the standard ODE theory for the global approximation error of the Euler scheme (Butcher, 2008) leads to

$$\max_{t\in[0,T]} |x_\delta^n(t) - X^n(t)| \le \delta\left(\frac{2}{n}\sum_{i=1}^n h_1(U_i)\right)^{-1} \sup_{t\in[0,T]}|\ddot{X}^n(t)|\left[\exp\left(\frac{T}{n}\sum_{i=1}^n h_1(U_i)\right) - 1\right]$$
$$= O_P(\delta). \blacksquare$$

Proof of Theorem 2. Lemma 7 establishes the first order result for $x_\delta^n(t) - X^n(t)$, and the weak convergence result is the consequence of the order result and Theorem 1. ■

5.2.2 Proof for the Accelerated Gradient Descent Case

Note that (x_k, y_k) and (x_k^n, y_k^n) are generated from accelerated gradient descent algorithms (2.4) and (3.10), respectively, and $X(t)$ and $X^n(t)$ are the respective solutions of ODEs (2.6) and (3.11).

Lemma 8 *For fixed $T > 0$, as $\delta \to 0$, we have*

$$\max_{k\le T\delta^{-1/2}} \left|x_k - X(k\delta^{1/2})\right| = O(\delta^{1/2}|\log\delta|), \tag{5.55}$$

$$\max_{k\le T\delta^{-1/2}} \left|z_k - \dot{X}(k\delta^{1/2})\right| = O(\delta^{1/2}|\log\delta|), \tag{5.56}$$

where sequence x_k is generated from algorithm (2.4), $X(t)$ is the solution of the corresponding ODE (2.6), and $z_k = (x_{k+1} - x_k)/\delta^{\frac{1}{2}}$ is given in (4.31).

Proof. We rewrite (2.4) as

$$x_{k+2} = y_{k+1} - \delta \nabla g(y_{k+1}), \;\; y_{k+1} = x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k) = x_k + \frac{2k+3}{k+3}\delta^{\frac{1}{2}} z_k,$$

and obtain

$$z_{k+1} = \left(1 - \frac{3}{k+3}\right) z_k - \delta^{\frac{1}{2}} \nabla g \left(x_k + \frac{2k+3}{k+3}\delta^{\frac{1}{2}} z_k\right). \tag{5.57}$$

Denote by y^* the critical point of $g(\cdot)$. Then, we have

$$|\nabla g(y_k)| = |\nabla g(y_k) - \nabla g(y^*)| \leq L|y_k - y^*| \leq C_1,$$

where C_1 is some constant, and

$$|z_0| = |x_1 - x_0|/\delta^{\frac{1}{2}} = \delta^{\frac{1}{2}} |\nabla g(x_0)| \leq C_1 \delta^{\frac{1}{2}}, \tag{5.58}$$

$$|z_k| \leq \frac{k-1}{k+2}|z_{k-1}| + C_1\delta^{\frac{1}{2}} \leq (k+1)C_1\delta^{\frac{1}{2}}. \tag{5.59}$$

To compare x_k and $X(k\delta^{\frac{1}{2}})$ and derive the difference between them, we first need to identify the relationship between $X(k\delta^{\frac{1}{2}})$ and $X((k+1)\delta^{\frac{1}{2}})$ and between $\dot{X}(k\delta^{\frac{1}{2}})$ and $\dot{X}((k+1)\delta^{\frac{1}{2}})$. As in (4.29), we let $Z = \dot{X}$, and ODE (2.6) is equivalent to

$$\dot{X} = Z, \;\; \dot{Z} = -\frac{3}{t} Z - \nabla g(X).$$

Then, with convention $t_k = k\delta^{\frac{1}{2}}$, we have for $k \geq 1$,

$$X(t_{k+1}) = X(t_k) + \int_{t_k}^{t_{k+1}} Z(u)du = X(t_k) + \delta^{\frac{1}{2}} Z(t_k) + \int_{t_k}^{t_{k+1}} [Z(u) - Z(t_k)]du, \tag{5.60}$$

$$\begin{aligned} Z(t_{k+1}) &= Z(t_k) - \int_{t_k}^{t_{k+1}} \frac{3}{u} Z(u)du - \int_{t_k}^{t_{k+1}} \nabla g(X(u))du \\ &= \left(1 - \frac{3}{k}\right) Z(t_k) - \int_{t_k}^{t_{k+1}} \left[\frac{3}{u} Z(u) - \frac{3}{t_k} Z(t_k)\right] du - \\ &\quad \delta^{\frac{1}{2}} \nabla g(X(t_k)) - \int_{t_k}^{t_{k+1}} [\nabla g(X(u)) - \nabla g(X(t_k))]du. \end{aligned} \tag{5.61}$$

Lemma 4 shows that on $(0, T]$, $|\dot{X}(t)|/t$ is bounded, $|Z(t)| \leq Ct$, and $|\dot{Z}(t)| = |\ddot{X}(t)| \leq C$ for some constant C. Then, we easily derive bounds for the following integrals that appear on the right-hand sides of (5.60) and (5.61);

$$\left| \int_{t_k}^{t_{k+1}} [Z(u) - Z(t_k)]du \right| = O(\delta),$$

$$\begin{aligned}\left|\int_{t_k}^{t_{k+1}}\left[\frac{3}{u}Z(u)-\frac{3}{t_k}Z(t_k)\right]du\right| &\leq \int_{t_k}^{t_{k+1}}\left|\frac{3}{u}[Z(u)-Z(t_k)]\right|du \\ &\quad+\int_{t_k}^{t_{k+1}}\left|\left(\frac{3}{u}-\frac{3}{t_k}\right)Z(t_k)\right|du \\ &\leq \frac{C\delta}{t_k}+\frac{3(t_{k+1}-t_k)^2}{t_kt_{k+1}}Ct_k=O(\delta^{\frac{1}{2}}k^{-1}),\end{aligned}$$

$$\left|\int_{t_k}^{t_{k+1}}[\nabla g(X(u))-\nabla g(X(t_k))]du\right|\leq L\int_{t_k}^{t_{k+1}}|X(u)-X(t_k)|du=O(\delta).$$

Plugging these integrals bounds into (5.60) and (5.61), we conclude

$$X(t_{k+1})=X(t_k)+\delta^{\frac{1}{2}}Z(t_k)+O(\delta),$$

$$Z(t_{k+1})=\left(1-\frac{3}{k}\right)Z(t_k)-\delta^{\frac{1}{2}}\nabla g(X(t_k))+O(\delta^{\frac{1}{2}}k^{-1})+O(\delta).$$

Let $a_k=|x_k-X(t_k)|, b_k=|z_k-Z(t_k)|$, and $S_k=b_0+b_1+...+b_k$. Using the definition of z_k and (5.57)-(5.59), we have

$$a_0=0,\ \ a_{k+1}\leq a_k+\delta^{\frac{1}{2}}b_k+O(\delta),$$

$$a_k\leq\delta^{\frac{1}{2}}S_{k-1}+O(k\delta), \tag{5.62}$$

$$b_0=|z_0|\leq C_1\delta^{\frac{1}{2}},\ \ b_1=|z_1-Z(t_1)|=O(\delta^{\frac{1}{2}}),$$

$$\begin{aligned}b_{k+1} &\leq \left(1-\frac{3}{k+3}\right)b_k+\frac{9}{k(k+3)}|Z(t_k)|+ \\ &\quad L\delta^{\frac{1}{2}}\left|x_k+\frac{2k+3}{k+3}\delta^{\frac{1}{2}}z_k-X(t_k)\right|+O(\delta^{\frac{1}{2}}k^{-1})+O(\delta) \\ &\leq b_k+O(\delta^{\frac{1}{2}}k^{-1})+L\delta^{\frac{1}{2}}a_k+2L\delta(k+1)C_1\delta^{\frac{1}{2}}+O(\delta^{\frac{1}{2}}k^{-1})+O(\delta) \\ &\leq b_k+L\delta S_{k-1}+L\delta^{\frac{1}{2}}O(k\delta)+O(\delta)+O(\delta^{\frac{1}{2}}k^{-1}) \\ &\leq b_k+L\delta S_{k-1}+O(\delta^{\frac{1}{2}}(k+1)^{-1}).\end{aligned} \tag{5.63}$$

Since for $1\leq k\leq T\delta^{-\frac{1}{2}}$, $k\delta^{\frac{1}{2}}=O(1), O(\delta)=O(\delta^{\frac{1}{2}}k^{-1})$, $k^{-1}\leq 2(k+1)^{-1}$. Moreover, with $b_1=O(\delta^{\frac{1}{2}})$, it is evident that can see that (5.63) holds for $k=0$. Therefore, there exists some constant $C_2>0$, such that

$$b_{k+1}\leq b_k+L\delta S_{k-1}+C_2\delta^{\frac{1}{2}}(k+1)^{-1}.$$

Define a new sequence b'_k from b_k in the following manner. Let $b'_0=b_0$, $b'_{k+1}=b'_k+L\delta S'_{k-1}+C_2\delta^{\frac{1}{2}}(k+1)^{-1}$, where $S'_k=b'_0+b'_1+...+b'_k$. Then, we can easily prove by induction that $b_k\leq b'_k$. Indeed, if $b_j\leq b'_j$ for $j=0,1,...,k$, then $S_{k-1}\leq S'_{k-1}$,

$$b_{k+1}\leq b_k+L\delta S_{k-1}+C_2\delta^{\frac{1}{2}}(k+1)^{-1}\leq b'_k+L\delta S'_{k-1}+C_2\delta^{\frac{1}{2}}(k+1)^{-1}=b'_{k+1}.$$

On the other hand, as $L\delta S'_{k-1} + C_2\delta^{\frac{1}{2}}(k+1)^{-1} > 0$, $\{b'_k\}$ is an increasing sequence. Thus, $S'_{k-1} \le kb'_k$, and

$$b'_{k+1} \le b'_k + L\delta k b'_k + C_2\delta^{\frac{1}{2}}(k+1)^{-1}.$$

Again, we define another sequence b^*_k from b'_k in the following manner. Let $b^*_0 = b'_0$, $b^*_{k+1} = b^*_k + L\delta k b^*_k + C_2\delta^{\frac{1}{2}}(k+1)^{-1}$. The same induction argument can prove that $b'_k \le b^*_k$. The recursive definition of b^*_k easily leads to the following expression,

$$b^*_k = \delta^{\frac{1}{2}}\left(C_1\prod_{j=1}^{k-1}(1+L\delta j) + C_2\sum_{i=1}^{k} i^{-1}\prod_{j=i}^{k-1}(1+L\delta j)\right),$$

and, hence, we obtain

$$\begin{aligned}
S_{\lfloor T\delta^{-\frac{1}{2}}\rfloor-1} &\le T\delta^{-\frac{1}{2}} b_{\lfloor T\delta^{-\frac{1}{2}}\rfloor} \le T\delta^{-\frac{1}{2}} b'_{\lfloor T\delta^{-\frac{1}{2}}\rfloor} \le T\delta^{-\frac{1}{2}} b^*_{\lfloor T\delta^{-\frac{1}{2}}\rfloor} \\
&\le C\left(\prod_{j=1}^{\lfloor T\delta^{-\frac{1}{2}}\rfloor-1}(1+L\delta j) + \sum_{i=1}^{\lfloor T\delta^{-\frac{1}{2}}\rfloor} i^{-1}\prod_{j=i}^{\lfloor T\delta^{-\frac{1}{2}}\rfloor-1}(1+L\delta j)\right) \\
&\le C\left(\prod_{j=1}^{\lfloor T\delta^{-\frac{1}{2}}\rfloor-1}(1+L\delta T\delta^{-\frac{1}{2}}) + \sum_{i=1}^{\lfloor T\delta^{-\frac{1}{2}}\rfloor} i^{-1}\prod_{j=i}^{\lfloor T\delta^{-\frac{1}{2}}\rfloor-1}(1+L\delta T\delta^{-\frac{1}{2}})\right) \\
&\le Ce^{LT^2}\left(1+\sum_{i=1}^{\lfloor T\delta^{-\frac{1}{2}}\rfloor} i^{-1}\right) \le C\log(T\delta^{-\frac{1}{2}}) = O(|\log\delta|).
\end{aligned}$$

Finally, using the above inequality and (5.62), we arrive at

$$\max_{k\le T\delta^{-\frac{1}{2}}}\left|x_k - X(k\delta^{\frac{1}{2}})\right| \le \delta^{\frac{1}{2}} S_{\lfloor T\delta^{-\frac{1}{2}}\rfloor-1} + O(T\delta^{\frac{1}{2}}) = O(\delta^{\frac{1}{2}}|\log\delta|),$$

which proves (5.55). It is easy to see that the left-hand side of (5.56) is bounded by $b^*_{[T\delta^{-1/2}]}$, which is of order $\delta^{\frac{1}{2}}|\log\delta|$. ■

Lemma 9

$$\max_{t\in[0,T]}|x^n_\delta(t) - X^n(t)| = O_p(\delta^{\frac{1}{2}}|\log\delta|),$$

where $x^n_\delta(t)$ is the continuous-time step processes for discrete sequence x^n_k generated from algorithm (3.10), and $X^n(t)$ is the continuous-time solution of the corresponding ODE (3.11).

Proof. The objective function associated with (10) and (11) is $\nabla\mathcal{L}^n(\theta;\mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta;U_i)$, which has Lipschitz constant $\frac{1}{n}\sum_{i=1}^n h_1(U_i) = O_p(1)$. Then, for any $\epsilon > 0$, there exists some constant $L_0 > 0$, such that for all n, $P\left(\frac{1}{n}\sum_{i=1}^n h_1(U_i) > L_0\right) < \epsilon$. For each

n, on the event $\{\frac{1}{n}\sum_{i=1}^n h_1(U_i) \le L_0\}$, Lemma 8 indicates that there exists constant M (which depends on L_0 only and is free of n), such that

$$\max_{k\le T\delta^{-\frac{1}{2}}} \left|x_k^n - X^n(k\delta^{\frac{1}{2}})\right| \le M\delta^{\frac{1}{2}}|\log\delta|.$$

Consequently, we have

$$P\left(\max_{k\le T\delta^{-\frac{1}{2}}} \left|x_k^n - X^n(k\delta^{\frac{1}{2}})\right| > M\delta^{\frac{1}{2}}|\log\delta|\right) \le P\left(\frac{1}{n}\sum_{i=1}^n h_1(U_i) > L\right) < \epsilon$$

holds for each n, that is

$$\max_{k\le T\delta^{-\frac{1}{2}}} \left|x_k^n - X^n(k\delta^{\frac{1}{2}})\right| = O_p(\delta^{\frac{1}{2}}|\log\delta|).$$

Lemma 4 indicates that $\sup_{t\in[0,T]}|\dot{X}^n(t)| = O_p(1)$ and, hence, we obtain

$$\sup_{s,t\in[0,T],t-s\le\delta^{\frac{1}{2}}} |X^n(t) - X^n(s)| \le \delta^{\frac{1}{2}} \sup_{t\in[0,T]} |\dot{X}^n(t)| = O_p(\delta^{\frac{1}{2}}).$$

Finally, for any t we can find k, such that $t_k \le t < t_{k+1}$, and show that

$$\max_{t\in[0,T]} |x_\delta^n(t) - X^n(t)| \le \max_{t\in[0,T]} \{|x_k^n - X^n(t_k)| + |X^n(t_k) - X^n(t)|\} = O_p(\delta^{\frac{1}{2}}|\log\delta|). \blacksquare$$

Proof of Theorem 2. Lemma 9 establishes the order result for $x_\delta^n(t) - X^n(t)$, and the weak convergence result is the consequence of the order result and Theorem 1. ■

5.3 Proof of Theorem 3

Using Assumption A4 and the standard empirical process argument (van der Vaart and Wellner, 2000), we can show that $\hat{\theta}_n$ is $\sqrt{n}$-consistent. Define $\vartheta = n^{1/2}(\theta - \check{\theta})$. We apply Taylor expansion to obtain

$$\begin{aligned}
&\mathcal{L}^n(\theta, \mathbf{U}_n) = \mathcal{L}^n(\check{\theta}, \mathbf{U}_n) + \nabla\mathcal{L}^n(\check{\theta}, \mathbf{U}_n)(\theta - \check{\theta}) + (\theta - \check{\theta})'\mathbb{H}\mathcal{L}^n(\check{\theta}, \mathbf{U}_n)(\theta - \check{\theta})/2 \\
&+ o_P(n^{-1/2}) \\
&= \mathcal{L}^n(\check{\theta}, \mathbf{U}_n) + n^{-1/2}[\nabla g(\check{\theta}) + n^{-1/2}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}]\vartheta + n^{-1}\vartheta'\mathbb{H}g(\check{\theta})\vartheta/2 + o_P(n^{-1}) \\
&= \mathcal{L}^n(\check{\theta}, \mathbf{U}_n) + n^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}\vartheta + n^{-1}\vartheta'\mathbb{H}g(\check{\theta})\vartheta/2 + o_P(n^{-1}),
\end{aligned}$$

where $\mathbf{Z}$ stands for the standard normal random vector, the second equality is due to Assumptions 2 and 4, Skorokhod's representation theorem, and the law of large numbers, and the third equality is from $\nabla g(\check{\theta}) = 0$. As $\hat{\theta}_n$ is the minimizer of $\mathcal{L}^n(\theta, \mathbf{U}_n)$, $\hat{\vartheta}_n = n^{1/2}(\hat{\theta}_n - \check{\theta})$ asymptotically minimizes $\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}\vartheta + \vartheta'\mathbb{H}g(\check{\theta})\vartheta/2$ over ϑ and thus, has an asymptotic distribution $[\mathbb{H}g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$. Note that $C(\mathbb{R}_+)$ is a subspace of $D(\mathbb{R}_+)$, and because of the metrics used in $C(\mathbb{R}_+)$ and $D(\mathbb{R}_+)$, the weak convergence of these process on $D(\mathbb{R}_+)$ is determined by their weak convergence on $D([0,T])$ for all integers T only (Billingsely, 1999; Jacod and Shiryaev, 2002). Treating $X(t)$, $X^n(t)$, $V(t)$, $V^n(t)$, and $x_\delta^n(t)$ as random

elements in $D(\mathbb{R}_+)$, since the weak convergence results established in Theorems 1 and 2 hold for $X^n(t)$ and $x^n_\delta(t)$ on $D([0,T])$ for any $T>0$, we conclude from these established weak convergence results that $V^n(t)=\sqrt{n}[X^n(t)-X(t)]$ and $\sqrt{n}[x^n_\delta(t)-X(t)]$ weakly converge to $V(t)$ on $D(\mathbb{R}_+)$.

On the other hand, it is known that as $k\to\infty$, x_k generated from algorithms (2.2) and (2.4) converge to the solution $\check{\theta}$ of (2.1) with speeds of orders $(\delta k)^{-1}$ and $(\sqrt{\delta}k)^{-2}$, respectively, while as $t\to\infty$, their corresponding continuous curves $X(t)$ as the solutions of ODEs (2.3) and (2.6) approach $\check{\theta}$ with speeds of orders t^{-1} and t^{-2}, respectively (Nesterov, 1983, 2004; Su et al., 2016). Similarly, for fixed n, as $k,t\to\infty$, x^n_k and $x^n_\delta(t)$ from algorithms (3.8) and (3.10) and $X^n(t)$ from ODEs (3.9) and (3.11) approach the solution $\hat{\theta}_n$ of (3.7). For the weak limit $V(t)$ governed by (3.14) or (3.12), as $t\to\infty$, both ODEs lead to $[\mathbb{H}g(X(\infty))]V(\infty)+\boldsymbol{\sigma}(X(\infty))\mathbf{Z}=0$, or equivalently, $V(\infty)=[\mathbb{H}g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\mathbf{Z}$. In fact, the solutions of (3.14) and (3.12) admit simple explicit expressions, for example,

$$V(t)=\int_0^t \exp\left[-\int_s^t \mathbb{H}g(X(u))du\right]\boldsymbol{\sigma}(X(s))ds\mathbf{Z}, \tag{5.64}$$

$$\forall\epsilon>0,\ \exists t_0>0,\text{ such that }\forall s>t_0,\left|[\mathbb{H}g(X(s))]^{-1}\boldsymbol{\sigma}(X(s))-[\mathbb{H}g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\right|<\epsilon,$$

$$\int_{t_0}^t \exp\left[-\int_s^t \mathbb{H}g(X(u))du\right]\boldsymbol{\sigma}(X(s))ds=\int_{t_0}^t \exp\left[-\int_s^t \mathbb{H}g(X(u))du\right]\mathbb{H}g(X(s))$$
$$\left\{[\mathbb{H}g(X(s))]^{-1}\boldsymbol{\sigma}(X(s))-[\mathbb{H}g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\right\}ds$$
$$+\int_{t_0}^t \exp\left[-\int_s^t \mathbb{H}g(X(u))du\right]\mathbb{H}g(X(s))ds[\mathbb{H}g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty)). \tag{5.65}$$

Since the assumptions indicate that $\boldsymbol{\sigma}(X(s))$ and $\mathbb{H}g(X(s))$ are bounded continuous on $[0,t_0]$, $\int_0^{t_0}|\boldsymbol{\sigma}(X(s))|ds$ is finite, and $\int_{t_0}^t \mathbb{H}g(X(u))du$ has finite eigenvalues. We immediately conclude that the eigenvalues of $\int_{t_0}^t \mathbb{H}g(X(s))ds$—which are no less than the eigenvalues of $\int_0^t \mathbb{H}g(X(s))ds$ minus the maximum eigenvalue of $\int_0^{t_0}\mathbb{H}g(X(s))ds$—diverge as $t\to\infty$. Therefore, we obtain

$$\left|\int_0^{t_0}\exp\left[-\int_s^t \mathbb{H}g(X(u))du\right]\boldsymbol{\sigma}(X(s))ds\right|\le\left|\exp\left[-\int_{t_0}^t \mathbb{H}g(X(u))du\right]\right|\int_0^{t_0}|\boldsymbol{\sigma}(X(s))|ds$$
$$\to 0,$$

$$\int_{t_0}^t \exp\left[-\int_s^t \mathbb{H}g(X(u))du\right]\mathbb{H}g(X(s))ds=1-\exp\left[-\int_{t_0}^t \mathbb{H}g(X(s))ds\right]\to 1,$$

$$\int_{t_0}^t\left|\exp\left[-\int_s^t \mathbb{H}g(X(u))du\right]\mathbb{H}g(X(s))\right|\left|[\mathbb{H}g(X(s))]^{-1}\boldsymbol{\sigma}(X(s))\right.$$
$$\left.-[\mathbb{H}g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\right|ds$$
$$\le\epsilon-\epsilon\left|\exp\left[-\int_{t_0}^t \mathbb{H}g(X(s))ds\right]\right|\le\epsilon,$$

which goes to zero, as we let $\epsilon\to 0$. Combining these results with (5.64) and (5.65), we conclude that as $t\to\infty$,

$$\int_0^t \exp\left[-\int_s^t \mathbb{H}g(X(u))du\right]\boldsymbol{\sigma}(X(s))ds\to[\mathbb{H}g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty)),$$

and $V(t)$ converges in distribution to $[\boldsymbol{I}\boldsymbol{H}g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\mathbf{Z}$. ■

5.4 Proofs of Theorems 4-6

Theorem 4 is proved by Lemma 10, with Theorems 5 and 6 shown in Lemma 18, where both lemmas are established in this subsection.

Denote by $\hat{Q}^*_{mk}$ the empirical distribution of mini-batch $U^*_{1k}, \cdots, U^*_{mk}$. Then, we have

$$\nabla\hat{\mathcal{L}}^m(\theta; \mathbf{U}^*_{mk}) = \int \nabla\ell(\theta; u)\hat{Q}^*_{mk}(du),$$

$$\int \nabla\ell(\theta; u)Q(du) = E[\nabla\ell(\theta; U)] = \nabla g(\theta).$$

Let $R^m(\theta; \mathbf{U}^*_m(t)) = (R^m_1(\theta; \mathbf{U}^*_m(t)), \cdots, R^m_p(\theta; \mathbf{U}^*_m(t)))'$, where

$$R^m_j(\theta; \mathbf{U}^*_m(t)) = \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial\theta_j}\ell(\theta; U^*_i(t)) - \frac{\partial}{\partial\theta_j}g(\theta)\right], \quad j = 1, \cdots, p.$$

We have

$$m^{-1/2}R^m(\theta; \mathbf{U}^*_m(t)) = \int \nabla\ell(\theta; u)\hat{Q}^*_{mk}(du) - \int \nabla\ell(\theta; u)Q(du),$$
$$\nabla\hat{\mathcal{L}}^m(x^m_{k-1}; \mathbf{U}^*_{mk}) = \nabla g(x^m_{k-1}) + m^{-1/2}R^m(x^m_{k-1}; \mathbf{U}^*_{mk}).$$

It is evident that $R^m(x^m_{k-1}; \mathbf{U}^*_{mk})$, $k = 1, \cdots, T/\delta$, are martingale differences and that $H^m_\delta(t)$ is a martingale. We may use the martingale theory (He et al., 1992; Jacod and Shiryaev, 2003) to establish weak convergence of $H^m_\delta(t)$ to the stochastic integral $H(t)$. Below, we employ a more direct approach to prove the weak convergence and obtain further convergence rate results.

Lemma 10 *As $\delta \to 0$ and $m \to \infty$, $H^m_\delta(t)$ weakly converges to $H(t) = \int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)$, $t \in [0, T]$.*

Proof. Let

$$\check{H}^m_\delta(t) = (m\delta)^{1/2}\sum_{t_k \le t}\left[\nabla\hat{\mathcal{L}}^m(X(t_{k-1}); \mathbf{U}^*_m(t_k)) - \nabla g(X(t_{k-1})\right]$$
$$= (m\delta)^{1/2}\sum_{k=1}^{[t/\delta]}\left[\int \nabla\ell(X((k-1)\delta); u)\hat{Q}^*_{mk}(du)\right.$$
$$\left. - \int \nabla\ell(X((k-1)\delta); u)Q(du)\right].$$

Note that

$$E\left[\int \nabla\ell(\theta; u)\hat{Q}^*_{mk}(du)\right] = \int \nabla\ell(\theta; u)Q(du),$$

$$\boldsymbol{\sigma}^2(\theta) = mVar\left[\int \nabla\ell(\theta;u)\hat{Q}^*_{mk}(du)\right] = \int[\nabla\ell(\theta;u)]^2 Q(du) - \left[\int \nabla\ell(\theta;u)Q(du)\right]^2,$$

which are the mean and variance of $\nabla\ell(\theta;U)$, respectively. Since $\mathbf{U}^*_{mk}$, $k = 1, 2, \cdots, [T/\delta]$, are independent, then $\check{H}^m_\delta(t)$ is a normalized partial sum process for independent random variables and weakly converges to $\int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)$. Indeed, its finite-dimensional distribution convergence can be easily established through Lyapunov's Central Limit Theorem with Assumptions A3 and A4 and Lemma 14 below and its tightness can be shown by the fact that for $r \le s \le t$,

$$E\left\{\left|\check{H}^m_\delta(t) - \check{H}^m_\delta(s)\right|^2 \left|\check{H}^m_\delta(s) - \check{H}^m_\delta(r)\right|^2\right\} \le [\Upsilon(t) - \Upsilon(r)]^2, \tag{5.66}$$

where $\Upsilon(\cdot)$ is a continuous non-decreasing function on $[0, T]$ (Billingsley, 1999, Equation 13.14 & Theorem 13.5). To establish (5.66), because of independence, we have

$$\begin{aligned} &E\left\{\left|\check{H}^m_\delta(t) - \check{H}^m_\delta(s)\right|^2 \left|\check{H}^m_\delta(s) - \check{H}^m_\delta(r)\right|^2\right\} = E\left\{\left|\check{H}^m_\delta(t) - \check{H}^m_\delta(s)\right|^2\right\} E\left\{\left|\check{H}^m_\delta(s) - \check{H}^m_\delta(r)\right|^2\right\} \\ &= \delta^2 \sum_{s<k\delta\le t} tr[\boldsymbol{\sigma}^2(X((k-1)\delta))] \sum_{r<k\delta\le s} tr[\boldsymbol{\sigma}^2(X((k-1)\delta))] \\ &\sim \int_s^t tr[\boldsymbol{\sigma}^2(X(u))]du \int_r^s tr[\boldsymbol{\sigma}^2(X(u))]du. \end{aligned}$$

Since $X(t)$ is a deterministic bounded continuous curve, and $\boldsymbol{\sigma}^2(\theta)$ is a continuous positive definite matrix,

$$\int_s^t tr[\boldsymbol{\sigma}^2(X(u))]du \int_r^s tr[\boldsymbol{\sigma}^2(X(u))]du \le \left[\int_r^t tr[\boldsymbol{\sigma}^2(X(u))]du\right]^2 \equiv [\Upsilon(t) - \Upsilon(r)]^2.$$

We have shown that as $\delta \to 0$ and $m \to \infty$, $\check{H}^m_\delta(t)$ weakly converges to $H(t)$. By the limit theorem for stochastic processes (Jacod and Shiryaev, 2003, Theorem 3.11 in Chapter VIII), we obtain that the quadratic variation $[\check{H}^m_\delta, \check{H}^m_\delta]_t$ converges in probability to $[H, H]_t$ for $t \in [0, T]$.

The Lipschitz of $\nabla\ell(\theta;u,Q)$ in θ implies the Lipschitz of $\nabla\hat{\mathcal{L}}^m(\theta;\mathbf{U}^*_{mk},Q)$ (which is proved at the beginning of the proof of Lemma 11 below), and Lemma 12 below indicates that as $\delta \to 0$ and $m \to \infty$, $x^m_k - X(k\delta)$ converges to zero in probability (with order $\delta + m^{-1/2}\delta^{1/2}$) uniformly over $1 \le k \le T/\delta$. These two results along with the Lipschitz of $\nabla g(\theta)$ immediately show that

$$\max_{t\le T}\left|[\check{H}^m_\delta, \check{H}^m_\delta]_t - [H^m_\delta, H^m_\delta]_t\right| = O_P\left((m\delta)\ \delta^{-1}[\delta + m^{-1/2}\delta^{1/2}]\right) = o_P(1),$$

and, hence, quadratic variation $[H^m_\delta, H^m_\delta]_t$ also converges in probability to $[H, H]_t$ for $t \in [0, T]$. An application of the limit theorem for stochastic processes (Jacod and Shiryaev, 2003, Theorem 3.11 in Chapter VIII) leads to the conclusion that as $\delta \to 0$ and $m \to \infty$, $H^m_\delta(t)$ weakly converges to $H(t)$—that is, $\check{H}^m_\delta(t)$ and $H^m_\delta(t)$ share the same weak convergence limit $H(t)$. ■

Lemma 11 *We have*

$$\max_{k\le T/\delta}|x_k^m - x_k| = O_P(m^{-1/2}),$$

where x_k and x_k^m are defined by (2.2) and (4.18), respectively.

Proof. Let $\zeta(\mathbf{U}_{mk}^*) = \frac{1}{m}\sum_{i=1}^m h_1(U_{ik}^*)$, which converges in probability to $E[h_1(U)]$ as $m \to \infty$. Then, we obtain

$$\begin{aligned}
&|\nabla\hat{\mathcal{L}}^m(\theta;\mathbf{U}_{mk}^*) - \nabla\hat{\mathcal{L}}^m(\vartheta;\mathbf{U}_{mk}^*)| \le \zeta(\mathbf{U}_{mk}^*)|\theta-\vartheta|,\ \ |\mathbb{H}\hat{\mathcal{L}}^m(\theta;\mathbf{U}_{mk}^*)| \le \zeta(\mathbf{U}_{mk}^*),\\
&|\check{\theta} - x_k^m| \le |\check{\theta} - x_{k-1}^m| + \delta|\nabla\hat{\mathcal{L}}^m(x_{k-1}^m;\mathbf{U}_{mk}^*) - \nabla\hat{\mathcal{L}}^m(\check{\theta};\mathbf{U}_{mk}^*)| + \delta|\nabla\hat{\mathcal{L}}^m(\check{\theta};\mathbf{U}_{mk}^*)|\\
&\le (1+\delta\zeta(\mathbf{U}_{mk}^*))|\check{\theta} - x_{k-1}^m| + \delta|\nabla\hat{\mathcal{L}}^m(\check{\theta};\mathbf{U}_{mk}^*)|\\
&\le \Big(1+\delta E[h_1(U)] + O_P(\delta m^{-1/2})\Big)^k\\
&+ \Big(1+\delta E[h_1(U)] + O_P(\delta m^{-1/2})\Big)^k \delta\sum_{j=1}^k[|\nabla g(\check{\theta})| + m^{-1/2}|R^m(\check{\theta};\mathbf{U}_{mj}^*)|]\\
&\le e^{TE[h_1(U)]}[1 + |\nabla g(\check{\theta})| + O_P(m^{-1/2})] = e^{TE[h_1(U)]}[1+O_P(m^{-1/2})],
\end{aligned}$$

namely, x_k^m is bounded uniformly over $k \le T/\delta$. On the other hand, we have

$$\begin{aligned}
x_k^m - x_k &= x_{k-1}^m - x_{k-1} - \delta[\nabla\hat{\mathcal{L}}^m(x_{k-1}^m;\mathbf{U}_{mk}^*) - \nabla g(x_{k-1})]\\
&= x_{k-1}^m - x_{k-1} - \delta[\nabla\hat{\mathcal{L}}^m(x_{k-1}^m;\mathbf{U}_{mk}^*) - \nabla\hat{\mathcal{L}}^m(x_{k-1};\mathbf{U}_{mk}^*)] - \delta m^{-1/2}R^m(x_{k-1};\mathbf{U}_{mk}^*)\\
&= (x_{k-1}^m - x_{k-1})[1 - \delta\mathbb{H}\hat{\mathcal{L}}^m(x_{\xi,k-1}^m;\mathbf{U}_{mk}^*)] - \delta m^{-1/2}R^m(x_{k-1};\mathbf{U}_{mk}^*)\\
&= -\delta m^{-1/2}\sum_{j=1}^k[1 - \delta\mathbb{H}\hat{\mathcal{L}}^m(x_{\xi,j-1}^m;\mathbf{U}_{mj}^*)]^j R^m(x_{j-1};\mathbf{U}_{mj}^*),
\end{aligned}$$

where $x_{\xi,j-1}^m$ is between x_{j-1} and x_{j-1}^m. Using $\zeta(\mathbf{U}_{mj}^*) \to E[h_1(U)]$ and Assumption A4, we obtain for $j,k \le T/\delta$,

$$\begin{aligned}
&|[1 - \delta\mathbb{H}\hat{\mathcal{L}}^m(x_{\xi,j-1}^m;\mathbf{U}_{mj}^*)]^j| \le [1+\delta\zeta(\mathbf{U}_{mj}^*)]^{T/\delta} \le e^{TE[h_1(U)]}[1+O_P(m^{-1/2})],\\
&R^m(x_{j-1};\mathbf{U}_{mj}^*) \sim \boldsymbol{\sigma}(x_{j-1})\mathbf{Z} = O_P(1),\\
&|x_k^m - x_k| \le \delta m^{-1/2}\sum_{j=1}^k|1+\delta\zeta(\mathbf{U}_{mj}^*)]^{T/\delta}|R^m(x_{j-1};\mathbf{U}_{mj}^*)| = O_P(k\delta m^{-1/2}) = O_P(m^{-1/2}).\ \blacksquare
\end{aligned}$$

Lemma 12

$$\max_{k\le T/\delta}|X(k\delta) - x_k^m| = O_P(\delta + m^{-1/2}\delta^{1/2}),$$

where $X(t)$ and x_k^m are defined by (2.3) and (4.18), respectively.

Proof. For $k = 1,\cdots,T/\delta$,

$$\int\nabla\ell(x_{k-1}^m;u)\hat{Q}_{mk}^*(du) - \int\nabla\ell(x_{k-1}^m;u)Q(du)$$

are martingale differences with conditional mean zero and conditional variance $\boldsymbol{\sigma}^2(x_{k-1}^m)/m$. Since x_k in (2.2) is the Euler approximation of solution $X(t)$ of ODE (2.3), the standard ODE theory shows

$$\max_{k\le T/\delta} |x_k - X(k\delta)| = O(\delta). \tag{5.67}$$

By Lemma 11, we have that with probability tending to one, $x_{k-1}^m, k = 1, \cdots, T/\delta$, fall within the neighborhood of the solution curve of ODE (2.3); thus, the maximum of $\boldsymbol{\sigma}^2(x_{k-1}^m)$, $k = 1, \cdots, T/\delta$, is bounded. Applying Burkholder's inequality (Chow and Teicher, 1997; He et al., 1992; Jacod and Shiryaev, 2003), we obtain

$$\max_{1\le k\le T/\delta} \left| \sqrt{m} \sum_{\ell=1}^{k} \left[\int \nabla \ell(x_{\ell-1}^m; u) \hat{Q}_{m\ell}^*(du) - \int \nabla \ell(x_{\ell-1}^m; u) Q(du) \right] \right| = O_P(\delta^{-1/2}),$$

that is,

$$\max_{k\le T/\delta} \left| m^{-1/2} \sum_{\ell=1}^{k} R^m(x_{\ell-1}^m; \mathbf{U}_{m\ell}^*) \right| = O_P(m^{-1/2}\delta^{-1/2}).$$

Therefore, for $k = 1, \cdots, T/\delta$,

$$\begin{aligned} x_k^m &= x_0 - \delta \sum_{\ell=1}^{k} \nabla g(x_{\ell-1}^m) - m^{-1/2}\delta \sum_{\ell=1}^{k} R^m(x_{\ell-1}^m; \mathbf{U}_{m\ell}^*) \\ &= x_0 - \delta \sum_{\ell=1}^{k} \nabla g(x_{\ell-1}^m) - O_P(m^{-1/2}\delta^{1/2}). \end{aligned}$$

and with the same initial value x_0, comparing the expressions for x_k and x_k^m, we obtain

$$\begin{aligned} x_k^m - x_k &= x_{k-1}^m - x_{k-1} - \delta[\nabla g(x_{k-1}^m) - \nabla g(x_{k-1})] - \delta m^{-1/2} R^m(x_{k-1}^m; \mathbf{U}_{mk}^*) \\ &= \delta \sum_{\ell=1}^{k} [\nabla g(x_{\ell-1}) - \nabla g(x_{\ell-1}^m)] - \delta m^{-1/2} \sum_{\ell=1}^{k} R^m(x_{\ell-1}^m; \mathbf{U}_{m\ell}^*). \end{aligned}$$

Using the L-Lipschitz assumption on $\nabla g(\cdot)$, we conclude for $k = 1, \cdots, T/\delta$,

$$\begin{aligned} |x_k^m - x_k| &\le L\delta \sum_{\ell=1}^{k} |x_{\ell-1}^m - x_{\ell-1}| + \delta m^{-1/2} \left| \sum_{\ell=1}^{k} R^m(x_{\ell-1}^m; \mathbf{U}_{m\ell}^*) \right| \\ &\le LT \max_{1\le \ell\le k} |x_{\ell-1}^m - x_{\ell-1}| + \delta m^{-1/2} \left| \sum_{\ell=1}^{k} R^m(x_{\ell-1}^m; \mathbf{U}_{m\ell}^*) \right|. \end{aligned}$$

Finally, we can easily show by induction that

$$\max_{k\le T/\delta} |x_k^m - x_k| = O_P(m^{-1/2}\delta^{1/2}).$$

The lemma is a consequence of the above result and (5.67). ■

The following lemma refines the order regarding $m^{-1/2}\delta^{1/2}$ in Lemma 12.

Lemma 13 *We have*

$$\max_{k\le T/\delta}|x_k^m - X_\delta^m(k\delta)| = o_P(m^{-1/2}\delta^{1/2}) + O_P(\delta + \delta m^{-1/2}|\log\delta|^{1/2}),$$

$$\max_{t\le T}|x_\delta^m(t) - X_\delta^m(t)| = o_P(m^{-1/2}\delta^{1/2}) + O_P(\delta|\log\delta|^{1/2}),$$

where $X_\delta^m(t)$ is given by (4.21), and x_k^m and $x_\delta^m(t)$ are defined by (4.18) and (4.19), respectively.

Proof. With weak convergence of $H_\delta^m(t)$ to $H(t)$ in Lemma 10, by Skohorod's representation, we realize $H_\delta^m(t)$ and $H(t)$ on some common probability spaces, such that as $\delta\to 0$ and $m\to\infty$, under the metric in $D([0,T])$, $H_\delta^m(t)-H(t)$ is $o_P(1)$. We may use arguments based on Lemma 37 (in Section 5.7) and stochastic equi-continuity to establish the convergence of $H_\delta^m(t) - H(t)$ under the maximum norm. Here, we adopt a direct approach. Consider linear interpolation $\tilde H_\delta^m(t)$ between the values of $H_\delta^m(k\delta)$, $k = 1,\cdots,T/\delta$, which satisfies

$$\max_{t\le T}|\tilde H_\delta^m(t) - H_\delta^m(t)| \le \delta^{1/2}\max_{k\le T/\delta}|R^m(x_{k-1}^m;\mathbf{U}_{mk}^*)|.$$

By Assumptions A1 and A2, we have

$$\begin{aligned}&|[\nabla\ell(x_{k-1}^m;U_{ik}^*) - \nabla g(x_{k-1}^m)] - [\nabla\ell(X((k-1)\delta);U_{ik}^*) - \nabla g(X((k-1)\delta))]|\\ &\le [h_1(U_{ik}^*) + L]|x_{k-1}^m - X((k-1)\delta)|,\end{aligned}$$

and then

$$\begin{aligned}&|R^m(x_{k-1}^m;\mathbf{U}_{mk}^*) \le |R^m(X((k-1)\delta);\mathbf{U}_{mk}^*)|\\ &+ m^{-1/2}\sum_{i=1}^m[h_1(U_{ik}^*) + L]|x_{k-1}^m - X((k-1)\delta)|,\\ &\max_{t\le T}|\tilde H_\delta^m(t) - H_\delta^m(t)| \le \delta^{1/2}\max_{k\le T/\delta}|R^m(X((k-1)\delta);\mathbf{U}_{mk}^*)|\\ &+ \delta^{1/2}\max_{k\le T/\delta}\left\{\frac{1}{m}\sum_{i=1}^m h_1(U_{ik}^*) + L\right\}m^{1/2}\max_{k\le T/\delta}|x_{k-1}^m - X((k-1)\delta)|.\end{aligned}$$

Lemma 12 implies $m^{1/2}\max_{k\le T/\delta}|x_{k-1}^m - X((k-1)\delta)| = m^{1/2}O_P(\delta + m^{-1/2}\delta^{1/2}) = O_P(m^{1/2}\delta + \delta^{1/2}) = o_P(1)$; by Lemma 14 below, we derive that $\max_{t\le T}|\tilde H_\delta^m(t) - H_\delta^m(t)| = o_P(\delta^{1/4}|\log\delta|)$. Thus, $\tilde H_\delta^m(t)$ weakly converges to $H(t)$ in $D([0,T])$. As both $\tilde H_\delta^m(t)$ and $H(t)$ live in $C([0,T])$, the weak convergence of $\tilde H_\delta^m(t)$ to $H(t)$ holds in $C([0,T])$. Again, by Skorokhod's representation theorem, we realize $\tilde H_\delta^m(t)$ and $H(t)$ on some common probability spaces, such that as $\delta\to 0$ and $m\to\infty$, $\max_{t\le T}|\tilde H_\delta^m(t) - H(t)| = o_P(1)$ and, hence, $\max_{t\le T}|H_\delta^m(t) - H(t)| = o_P(1)$.

Note that for $1\le k\le T/\delta$,

$$\delta\nabla\hat{\mathcal{L}}^m(x_{k-1}^m;\mathbf{U}_{mk}^*) = \delta\nabla g(x_{k-1}^m) + m^{-1/2}\delta^{1/2}[H_\delta^m(k\delta) - H_\delta^m((k-1)\delta)],$$

$$x_k^m - x_{k-1}^m = -\delta \nabla g(x^m(t_{k-1})) - m^{-1/2}\delta^{1/2}[H_\delta^m(k\delta) - H_\delta^m((k-1)\delta)],$$
$$x_k^m = x_0 - \delta \sum_{\ell=1}^{k} \nabla g(x_{\ell-1}^m) - m^{-1/2}\delta^{1/2} H_\delta^m(k\delta)$$
$$= x_0 - \delta \sum_{\ell=1}^{k} \nabla g(x_{\ell-1}^m) - m^{-1/2}\delta^{1/2} H(k\delta) + o_P(m^{-1/2}\delta^{1/2}).$$

Define $\check{x}_0^m = x_0$, and

$$\check{x}_k^m - \check{x}_{k-1}^m = -\delta \nabla g(\check{x}_{k-1}^m) - m^{-1/2}\delta^{1/2}[H(k\delta) - H((k-1)\delta)]. \tag{5.68}$$

Then, the situation is the same as that in the last proof part of Lemma 12, and the same argument can be used to derive a recursive expression for $x_k^m - \check{x}_k^m$ and prove by induction that

$$\max_{k \le T/\delta} |x_k^m - \check{x}_k^m| = o_P(m^{-1/2}\delta^{1/2}).$$

The lemma is a consequence of the above result and Lemma 15 below. ■

Lemma 14

$$\sup E\{|R^m(X(t); \mathbf{U}_{mk}^*)|^4 : t \in [0,T], k = 1, \cdots, T/\delta\} < \infty,$$

$$\delta^{1/2} \max_{k \le T/\delta} \left\{ \frac{1}{m} \sum_{i=1}^{m} h_1(U_{ik}^*) - E[h_1(U)] \right\} = O_P(\delta^{1/4} |\log \delta|),$$

$$\delta^{1/2} \max_{k \le T/\delta} |R^m(X((k-1)\delta); \mathbf{U}_{mk}^*)| = O_P(\delta^{1/4} |\log \delta|).$$

Proof. Direct calculations lead to

$$P\left(\delta^{1/2} \max_{k \le T/\delta} |R^m(X((k-1)\delta); \mathbf{U}_{mk}^*)| > \delta^{1/4} |\log \delta|\right)$$
$$= 1 - \prod_{k \le T/\delta} P\left(\delta^{1/4} |R^m(X((k-1)\delta); \mathbf{U}_{mk}^*)| \le |\log \delta|\right)$$
$$\le 1 - \prod_{k \le T/\delta} \left[1 - \delta E\left\{|R^m(X((k-1)\delta); \mathbf{U}_{mk}^*)|^4\right\} / |\log \delta|^4\right]$$
$$\le 1 - \exp\left[-2T\tau / |\log \delta|^4\right] \sim 2T\tau / |\log \delta|^4 \to 0,$$

where we use Chebyshev's inequality, $\log(1-u) \ge -2u$ for $0 < u < 0.75$, and $\tau = \sup_{t,k} E\{|R^m(X(t); \mathbf{U}_{mk}^*)|^4\} \equiv \sup E\{|R^m(X(t); \mathbf{U}_{mk}^*)|^4 : t \in [0,T], k = 1, \cdots, T/\delta\}$ the finiteness of which will be shown below. Indeed, it is sufficient to show that each component of $R^m(X(t); \mathbf{U}_{mk}^*)$ has finite fourth-moment uniformly over $t \in [0,T], k = 1, \cdots, T/\delta$ and, thus, we need to prove it only in the one-dimensional case with a gradient equal to the

partial derivative. With this simple set-up, we have

$$
\begin{aligned}
&|R^m(X(t);\mathbf{U}^*_{mk})|^4 = m^{-2}\left[\sum_{i=1}^m\{\nabla\ell(X(t);U^*_{ik})-\nabla g(X(t))\}\right]^4\\
&= m^{-2}\sum_{i\neq j}\{\nabla\ell(X(t);U^*_{ik})-\nabla g(X(t))\}^2\{\nabla\ell(X(t);U^*_{jk})-\nabla g(X(t))\}^2\\
&+ m^{-2}\sum_{i=1}^m\{\nabla\ell(X(t);U^*_{ik})-\nabla g(X(t))\}^4 + \text{odd power terms},\\
&E\{|R^m(X(t);\mathbf{U}^*_{mk})|^4\} = m^{-2}\sum_{i=1}^m E[\{\nabla\ell(X(t);U^*_{ik})-\nabla g(X(t))\}^4]\\
&+ m^{-2}\sum_{i\neq j}E[\{\nabla\ell(X(t);U^*_{ik})-\nabla g(X(t))\}^2]E[\{\nabla\ell(X(t);U^*_{jk})-\nabla g(X(t))\}^2]\\
&\leq \{E[\{\nabla\ell(X(t);U^*_{1k})-\nabla g(X(t))\}^2]\}^2 + E[\{\nabla\ell(X(t);U^*_{1k})-\nabla g(X(t))\}^4]/m\\
&\leq \{E[\{\nabla\ell(X(t);U_{1k})-\nabla g(X(t))\}^2]\}^2 + E[\{\nabla\ell(X(t);U_{1k})-\nabla g(X(t))\}^4]/m,
\end{aligned}
$$

where we use the fact that all odd power terms have mean zero factor $\nabla\ell(X(t);U^*_{ik}) - \nabla g(X(t))$ and, thus, their expectations are equal to zero. By Assumption A1, we have

$$
\begin{aligned}
&\sup_{t,k}E[\{\nabla\ell(X(t);U_{1k})-\nabla g(X(t))\}^2] \leq 2\sup_{t\geq 0}\left\{|X(t)-x_0|^2\right\}E[h_1^2(U)]\\
&+ 2E[\{\nabla\ell(x_0,U)\}^2] + 2\sup_{t\geq 0}\{[\nabla g(X(t))]^2\},
\end{aligned}
$$

$$
\begin{aligned}
&\sup_{t,k}E[\{\nabla\ell(X(t);U_{1k})-\nabla g(X(t))\}^4] \leq 64\sup_{t\geq 0}\left\{|X(t)-x_0|^4\right\}E[h_1^4(U)]\\
&+ 64E[\{\nabla\ell(x_0,U)\}^4] + 8\sup_{t\geq 0}\{[\nabla g(X(t))]^4\},
\end{aligned}
$$

which are finite because $X(t)$ is deterministic and bounded. Thus, we obtain that $\tau = \sup_{t,k}E\{|R^m(X(t);\mathbf{U}^*_{mk})|^4\}$ is finite.

Similarly, as $h_1(U)$ has the fourth moment, we have

$$
E\left\{\left|m^{-1/2}\sum_{i=1}^m\{h_1(U^*_{ik})-E[h_1(U^*_{ik})]\}\right|^4\right\} \leq [Var(h_1(U))]^2 + E\left[\{h_1(U)-E[h_1(U)]\}^4\right] \equiv \tau_1,
$$

$$
\begin{aligned}
&P\left(\delta^{1/2}\max_{k\leq T/\delta}\left|m^{-1}\sum_{i=1}^m h_1(U^*_{ik})-E[h_1(U^*_{ik})]\right| > \delta^{1/4}|\log\delta|\right)\\
&\leq 1-\prod_{k\leq T/\delta}\left[1-\delta E\left\{\left|m^{-1}\sum_{i=1}^m h_1(U^*_{ik})-E[h_1(U^*_{ik})]\right|^4\right\}/|\log\delta|^4\right]\\
&\leq 1-\exp\left[-2T\tau_1/|\log\delta|^4\right] \sim 2T\tau_1/|\log\delta|^4 \to 0, \text{ as } \delta\to 0,
\end{aligned}
$$

which together with $E[h_1(U_{ik}^*)] = E[h_1(U)]$ imply

$$\delta^{1/2} \max_{k \le T/\delta} \{\sum_{i=1}^{m} h_1(U_{ik}^*)\}/m = \delta^{1/2} E[h_1(U)] + O_P(\delta^{1/4} |\log \delta|). \blacksquare$$

Lemma 15

$$\max_{t \in [0,T]} |\check{x}_k^m - X_\delta^m(k\delta)| = O_P(\delta + \delta m^{-1/2} |\log \delta|^{1/2}),$$
$$\max_{0 \le t-s \le \delta} |X_\delta^m(t) - X_\delta^m(s)| = O_P(\delta |\log \delta|^{1/2}),$$

where $\check{x}_k^m$ and $X_\delta^m(t)$ are defined by (5.68) and (4.21), respectively.

Proof. By (4.21) we have

$$|X_\delta^m(t) - X_\delta^m(s)| \le \int_s^t |\nabla g(X_\delta^m(u))| du + m^{-1/2}\delta^{1/2} \left| \int_s^t \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u) \right|$$
$$= O_P(\delta + m^{-1/2}\delta |\log \delta|^{1/2}),$$

where we use the fact that uniformly over $0 \le t - s \le \delta$,

$$\int_s^t |\nabla g(X_\delta^m(u))| du = O_P(\delta), \quad \int_s^t \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u) = O_P(\delta^{1/2} |\log \delta|^{1/2}),$$

and the order for the Brownian term is derived by the law of the iterated logarithm for Brownian motion.

Note that $\check{x}_k^m$ is the Euler approximation of SDE (4.21). The first result follows from the standard argument for the Euler approximation. Let $D(k) = |\check{x}_k^m - X_\delta^m(k\delta)|$. As $\check{x}_0^m = X_\delta^m(0) = x_0$, we have

$$\check{x}_1^m - X_\delta^m(\delta) = \int_0^\delta \nabla g(X_\delta^m(u)) du - \delta \nabla g(x_0),$$
$$D(1) = |\check{x}_1^m - X_\delta^m(\delta)| = | \int_0^\delta [\nabla g(X_\delta^m(u)) - \nabla g(x_0)] du|$$
$$\le C\delta \max_{0 \le u \le \delta} |X_\delta^m(u) - x_0| = O_P(\delta^2 + m^{-1/2}\delta^2 |\log \delta|^{1/2}),$$

where we use the fact that for $u \in [0, \delta]$,

$$|X_\delta^m(u) - x_0| \le \int_0^u |\nabla g(X_\delta^m(v))| dv + m^{-1/2}\delta^{1/2} \left| \int_0^u \boldsymbol{\sigma}(X(v)) d\mathbf{B}(v) \right|$$
$$= O_P(\delta + m^{-1/2}\delta |\log \delta|^{1/2}).$$

For the general k, we obtain

$$D(k) = \left| \int_0^{k\delta} \nabla g(X_\delta^m(u)) du - \delta \sum_{\ell=1}^{k} \nabla g(\check{x}_{\ell-1}^m) \right|$$
$$\le D(k-1) + \left| \int_{(k-1)\delta}^{k\delta} \nabla g(X_\delta^m(u)) du - \delta \nabla g(\check{x}_{k-1}^m) \right|,$$

$$\int_{(k-1)\delta}^{k\delta} \nabla g(X_\delta^m(u))du - \delta\nabla g(\check{x}_{k-1}^m) = \int_{(k-1)\delta}^{k\delta} [\nabla g(X_\delta^m(u)) - \nabla g(X_\delta^m((k-1)\delta))]du$$
$$+ \delta[\nabla g(X((k-1)\delta)) - \nabla g(\check{x}_{k-1}^m)],$$
$$|\nabla g(X((k-1)\delta)) - \nabla g(\check{x}_{k-1}^m)| \le C|X((k-1)\delta) - \check{x}_{k-1}^m| = CD(k-1),$$
$$|\nabla g(X_\delta^m(u)) - \nabla g(X_\delta^m((k-1)\delta))| = |\boldsymbol{H}g(X_\delta^m(u_*))[X_\delta^m(u) - X_\delta^m((k-1)\delta)]|$$
$$\le C\int_{(k-1)\delta}^{u} |\nabla g(X_\delta^m(v))|dv + Cm^{-1/2}\delta^{1/2}\left|\int_{(k-1)\delta}^{u} \boldsymbol{\sigma}(X(v))d\mathbf{B}(v)\right|$$
$$= O_P(\delta + m^{-1/2}\delta|\log\delta|^{1/2}),$$

and, thus, we conclude that

$$D(k) \le D(k-1) + C\delta D(k-1) + O_P(\delta^2 + m^{-1/2}\delta^2|\log\delta|^{1/2}),$$

which shows that for $k \le T/\delta$,

$$D(k) \le (1+C\delta)^{k-1}D(1) + O_P(k\delta^2 + km^{-1/2}\delta^2|\log\delta|^{1/2}) = O_P(\delta + m^{-1/2}\delta|\log\delta|^{1/2}). \blacksquare$$

Lemma 16

$$\max_{t\le T}|X_\delta^m(t) - X(t)| \le Cm^{-1/2}\delta^{1/2}\max_{t\le T}\left|\int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right| = O_P(m^{-1/2}\delta^{1/2}),$$

where $X(t)$ and $X_\delta^m(t)$ are defined by (2.3) and (4.21), respectively.

Proof. With the same initial value for $X(t)$ and X_δ^m, from (2.3) and (4.21) we have

$$|X_\delta^m(t) - X(t)| \le \int_0^t |\nabla g(X_\delta^m(u)) - \nabla g(X(u))|du + m^{-1/2}\delta^{1/2}\left|\int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|$$
$$\le C\int_0^t |X_\delta^m(u) - X(u)|du + m^{-1/2}\delta^{1/2}\left|\int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|.$$

Applying the Gronwall inequality, we obtain

$$|X_\delta^m(t) - X(t)| \le m^{-1/2}\delta^{1/2}\left[\left|\int_0^t \boldsymbol{\sigma}(X(t))d\mathbf{B}(u)\right| + C\int_0^t e^{C(t-s)}\left|\int_0^s \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|ds\right],$$

which implies

$$\max_{t\le T}|X_\delta^m(t) - X(t)| \le Cm^{-1/2}\delta^{1/2}\max_{t\le T}\left|\int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right| = O_P(m^{-1/2}\delta^{1/2}),$$

where the last equality is due to Burkholder's inequality. $\blacksquare$

Lemma 17

$$\max_{t\le T}|X_\delta^m(t) - \check{X}_\delta^m(t)| = O_P(m^{-1}\delta).$$

where $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ are the solutions of (4.21) and (4.22), respectively.

Proof. We have

$$
\begin{aligned}
&|X_\delta^m(t) - \check{X}_\delta^m(t)| \le \int_0^t |\nabla g(X_\delta^m(u)) - \nabla g(\check{X}_\delta^m(u))| du \\
&+ m^{-1/2}\delta^{1/2} \left| \int_0^t [\boldsymbol{\sigma}(X(u)) - \boldsymbol{\sigma}(\check{X}_\delta^m(u))] d\mathbf{B}(u) \right| \\
&\le C \int_0^t |X_\delta^m(u) - X(u)| du + m^{-1/2}\delta^{1/2} \left| \int_0^t [\boldsymbol{\sigma}(X(u)) - \boldsymbol{\sigma}(\check{X}_\delta^m(u))] d\mathbf{B}(u) \right|. \\
&E[|X_\delta^m(t) - \check{X}_\delta^m(t)|^2] \le C \int_0^t E[|X_\delta^m(u) - \check{X}_\delta^m(u)|^2] du \\
&+ 2m^{-1}\delta E \left[\left| \int_0^t [\boldsymbol{\sigma}(X(u)) - \boldsymbol{\sigma}(\check{X}_\delta^m(u))] d\mathbf{B}(u) \right|^2 \right] \\
&\le C \int_0^t E[|X_\delta^m(u) - \check{X}_\delta^m(u)|^2] du + 2m^{-1}\delta \int_0^t E[|\boldsymbol{\sigma}(X(u)) - \boldsymbol{\sigma}(\check{X}_\delta^m(u))|^2] du \\
&\le C \int_0^t E[|X_\delta^m(u) - \check{X}_\delta^m(u)|^2] du + C_1 m^{-1}\delta \int_0^t E[|X(u) - X_\delta^m(u)|^2] du \\
&+ C_1 m^{-1}\delta \int_0^t E[|X_\delta^m(u) - \check{X}_\delta^m(u)|^2] du,
\end{aligned}
$$

where the last inequality is due to

$$
|\boldsymbol{\sigma}(X(u)) - \boldsymbol{\sigma}(\check{X}_\delta^m(u))| \le C|X(u) - \check{X}_\delta^m(u)| \le C|X(u) - X_\delta^m(t)| + C|X_\delta^m(t) - \check{X}_\delta^m(t)|.
$$

The Gronwall inequality leads to

$$
E[|X_\delta^m(t) - \check{X}_\delta^m(t)|^2] \le Cm^{-1}\delta \max_{s \le t} E[|X(s) - X_\delta^m(s)|^2].
$$

Using Lemma 16, we have

$$
\begin{aligned}
&\max_{s \le t} E[|X(s) - X_\delta^m(s)|^2] \le Cm^{-1}\delta E \left[\max_{s \le t} \left| \int_0^s \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u) \right|^2 \right] \\
&\le Cm^{-1}\delta E \left[\int_0^t [\boldsymbol{\sigma}(X(u))]^2 du \right],
\end{aligned}
$$

where the last inequality is from Burkholder's inequality. Hence

$$
E[|X_\delta^m(t) - \check{X}_\delta^m(t)|^2] \le Cm^{-2}\delta^2 E \left[\int_0^t [\boldsymbol{\sigma}(X(u))]^2 du \right],
$$

and we can adopt the same argument to establish it for t as a bounded stopping time. Finally, we prove the lemma by applying Lenglart's inequality for semi-martingale with $\eta_i = D_i m^{-1}\delta$ for some positive constants D_i,

$$
\begin{aligned}
&P\left(\max_{s \le t} |X(s) - X_\delta^m(s)| > \eta_1 \right) \le \frac{Cm^{-2}\delta^2 \int_0^t [\boldsymbol{\sigma}(X(u))]^2 du}{\eta_1^2} \\
&+ P\left(Cm^{-2}\delta^2 \int_0^t [\boldsymbol{\sigma}(X(u))]^2 du > \eta_2^2 \right) \to 0, \text{ as } D_i \to \infty. \blacksquare
\end{aligned}
$$

Lemma 18 *As $\delta \to 0$ and $m, n \to \infty$, we have $V_\delta^m(t)$ and $\check{V}_\delta^m(t)$ both weakly converge to $V(t)$. Moreover, if $m(n\delta)^{1/2} \to 0$, and $m^{1/2}\delta|\log\delta|^{1/2} \to 0$, $(m/\delta)^{1/2}[x_\delta^m(t) - X(t)]$ weakly converges to $V(t)$.*

Proof. As the solutions of (4.21) and (4.22) have difference of order $m^{-1}\delta$, they have the same asymptotic distribution, and we can easily establish the result for $\check{V}_\delta^m(t)$ by that for $V_\delta^m(t)$ and Lemma 13.

Let us consider the easier one for $X_\delta^m(t)$. From (4.21) and (2.3), we have

$$d[X_\delta^m(t) - X(t)] = -[\nabla g(X_\delta^m(t) - \nabla g(X(t))]dt - m^{-1/2}\delta^{1/2}\boldsymbol{\sigma}(X(t))d\mathbf{B}(t),$$

and for $t \in [0, T]$,

$$X_\delta^m(t) - X(t) = -\int_0^t [\mathbb{H}g(X_\xi)][X_\delta^m(u) - X(u)]du - m^{-1/2}\delta^{1/2}\int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u),$$

where X_ξ is between $X(u)$ and $X_\delta^m(u)$ and, thus, Lemma 16 shows that uniformly over $[0, T]$,

$$|X_\xi - X(u)| \le |X_\delta^m(u) - X(u)| = O_P(m^{-1/2}\delta).$$

Then, we obtain

$$V_\delta^m(t) = -\int_0^t [\mathbb{H}g(X_\xi)]V_\delta^m(u)du - \int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u). \tag{5.69}$$

First as $\delta \to 0$, $m, n \to \infty$, equation (5.69) converges to (4.24).

We need to show stochastic equicontinuity for $V_\delta^m(t)$. From (5.69), we obtain

$$|V_\delta^m(t)| \le C\int_0^t |V_\delta^m(u)|du + \left|\int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|,$$

and by the Gronwall inequality, we have

$$\max_{t\le T}|V_\delta^m(t)| \le C\max_{t\le T}\left|\int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|,$$

that is $V_\delta^m(t)$ is bounded in probability uniformly over $[0, T]$. Again, (5.69) indicates that for any $s, t \in [0, T]$, and $t \in [s, s+\gamma]$,

$$\begin{aligned}
&V_\delta^m(t) - V_\delta^m(s) = -\int_s^t [\mathbb{H}g(X_\xi)]V_\delta^m(u)du - \int_s^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u),\\
&|V_\delta^m(t) - V_\delta^m(s)| \le C\int_s^t |V_\delta^m(u)|du + \left|\int_s^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|\\
&\le C\int_s^t |V_\delta^m(u) - V_\delta^m(s)|du + C(t-s)|V_\delta^n(s)| + \left|\int_s^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|.
\end{aligned}$$

Finally, applying the Gronwall inequality, we obtain uniformly for $t \in [s, s+\gamma]$,

$$|V_\delta^m(t) - V_\delta^m(s)| \le C\gamma|V_\delta^n(s)| + C\max_{s\le t\le s+\gamma}\left|\int_s^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right| = O_P(\gamma + \gamma^{1/2}|\log\gamma|^{1/2}),$$

which proves stochastic equicontinuity for $V_\delta^n(t)$. ■

5.5 Proof of Theorem 7

Theorem 7 can be proved by the same proof argument of Theorem 4, except for changing the step size from δ to $\delta^{1/2}$. ■

5.6 Proof of Theorem 8

We prove Theorem 8 in two subsections, with one for solutions of the second-order SDEs (4.35) and (4.36) and one for weak convergence of $V_\delta^m(t)$.

5.6.1 The Unique Solution of the Second-Order SDEs

In this subsection, we prove Lemma 25 below that the second-order SDEs (4.35) (with fixed δ and m) and (4.36) have unique (weak) solutions in the distributional sense.

Due to the similarity, we provide representative proof arguments only for the following second-order SDE,

$$\ddot{V}(t) + \frac{3}{t}\dot{V}(t) + [\nabla g(X(t))]V(t) + \boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0, \tag{5.70}$$

where initial conditions $V(0) = c$ and $\dot{V}(0) = 0$, $\mathbf{B}(t)$ is a standard Brownian motion, $\dot{V}(t)$ and $\ddot{V}(t)$ are the first and second derivatives of $V(t)$, respectively, $\dot{\mathbf{B}}(t) = \frac{dB(t)}{dt}$ is white noise in the sense that for any smooth function $h(t)$ with compact support,

$$\int h(t)\dot{\mathbf{B}}(t)dt = \int h(t)dB(t),$$

and the right-hand side is an Itô integral.

The second-order SDE (5.70) is equivalent to

$$Y(t) = V(t) + \frac{t}{2}\dot{V}(t), \;\; \dot{Y}(t) = -\frac{t}{2}[\nabla g(X(t))]V(t) - \frac{t}{2}\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t), \tag{5.71}$$

where $V(0) = c, \dot{V}(0) = 0$, and $Y(0) = V(0) = c$. Denote by $V_\eta(t)$ the solution of the smoothed second-order SDE

$$\ddot{V}_\eta(t) + \frac{3}{t \vee \eta}\dot{V}_\eta(t) + [\nabla g(X(t))]V_\eta(t) + \boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0, \tag{5.72}$$

with initial conditions $V_\eta(0) = c$ and $\dot{V}_\eta(0) = 0$.

Recall the notation $M_a(s,t;Y)$ defined in (5.53). In the proofs of Theorems 1 and 4, we have employed $M_a(s,t;Y)$ with $a = 1$, as curves and processes are solutions of ODE and, thus, differentiable. For this protion of proofs, we need to handle the Brownian motion and SDEs, and the related processes have less than 1/2-derivatives, so we fix $a \in (0, 1/2)$ and consider $M_a(s,t;Y)$ with $a < 1/2$.

Lemma 19

$$|\nabla g(X(t))| \le |\nabla g(X(s))| + L(t-s)|\dot{X}(s)| + LM_a(s,t;X)(t-s)^{1+a}/(1+a),$$
$$|\nabla g(X(t))V_\eta(t) - \nabla g(X(s))V_\eta(s)| \le L|V_\eta(s)|(t-s)|\dot{X}(s)| + |\nabla g(X(t))|(t-s)|\dot{V}_\eta(s)|$$
$$+ [L|V_\eta(s)|M_a(s,t;X) + |\nabla g(X(t))|M_a(s,t;V_\eta)](t-s)^{1+a}/(1+a).$$

Proof. We prove the lemma by the following direct calculation

$$
\begin{aligned}
&|\nabla g(X(t))V_\eta(t) - \nabla g(X(s))V_\eta(s)| \le |\nabla g(X(t))||V_\eta(t) - V_\eta(s)| \\
&+ |\nabla g(X(t)) - \nabla g(X(s))||V_\eta(s)| \\
&\le |\nabla g(X(t))||V_\eta(t) - V_\eta(s)| + L|V_\eta(s)||X(t) - X(s)| \\
&\le L|V_\eta(s)||\int_s^t [\dot{X}(v) - \dot{X}(s)]dv + (t-s)\dot{X}(s)| \\
&+ |\nabla g(X(t))||\int_s^t [\dot{V}_\eta(v) - \dot{V}_\eta(s)]dv + (t-s)\dot{V}_\eta(s)| \\
&\le L|V_\eta(s)|(t-s)|\dot{X}(s)| + |\nabla g(X(t))|(t-s)|\dot{V}_\eta(s)| \\
&+ L|V_\eta(s)||\int_s^t (v-s)^a \frac{\dot{X}(v) - \dot{X}(s)}{(v-s)^a}dv| + |\nabla g(X(t))||\int_s^t (v-s)^a \frac{\dot{V}_\eta(v) - \dot{V}_\eta(s)}{(v-s)^a}dv| \\
&\le L|V_\eta(s)|(t-s)|\dot{X}(s)| + |\nabla g(X(t))|(t-s)|\dot{V}_\eta(s)| \\
&+ [L|V_\eta(s)|M_a(s,t;X) + |\nabla g(X(t))|M_a(s,t;V_\eta)(t-s)^{1+a}/(1+a), \\
&|\nabla g(X(t))| \le |\nabla g(X(s))| + L|X(t) - X(s)| \\
&\le |\nabla g(X(s))| + L(t-s)|\dot{X}(s)| + LM_a(s,t;X)](t-s)^{1+a}/(1+a). \blacksquare
\end{aligned}
$$

Lemma 20 *There exists $\eta_0 > 0$, such that for $\eta \in (0, \eta_0]$, $1 - |\nabla g(X(0))|\eta^2/[(1+a)(2+a)] - LM_a(0,\eta;X)\eta^{3+a}/[(1+a)^2(3+2a)]$ is bounded below from zero. Then, for $\eta \in (0,\eta_0]$, we have*

$$
M_a(0,\eta;V_\eta) \le \frac{1}{1 - |\nabla g(X(0))|\eta^2/[(1+a)(2+a)] - LM_a(0,\eta;X)\eta^{3+a}/[(1+a)^2(3+2a)]}
$$

$$
\left[|\nabla g(X(0))V_\eta(0)|\eta^{1-a} + \frac{L|V_\eta(0)|M_a(0,\eta;X)\eta^2}{(1+a)(2+a)} + \max_{t\in(0,\eta]}\left|\frac{1}{t^a}e^{-3t/\eta}\int_0^t e^{3u/\eta}\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|\right].
$$

Proof. As $\nabla g(X(0))$ and $M_a(0,\eta;X)$ for each η are deterministic and finite, and $M_a(0,\eta;X)$ is continuous and increasing in η, we easily show that $|\nabla g(X(0))|\eta^2/[(1+a)(2+a)] + LM_a(0,\eta;X)\eta^{3+a}/[(1+a)^2(3+2a)]$ approaches zero as $\eta \to 0$, which leads to the existence of η_0. Then, Lemma 19 indicates

$$
\begin{aligned}
&|\nabla g(X(u))V_\eta(u) - \nabla g(X(0))V_\eta(0)| \\
&\le [L|V_\eta(0)|M_a(0,u;X) + |\nabla g(X(u))|M_a(0,u;V_\eta)]u^{1+a}/(1+a), \\
&|\nabla g(X(u))| \le |\nabla g(X(0))| + LM_a(0,u;X)u^{1+a}/(1+a).
\end{aligned}
$$

For $t \in (0,\eta]$, V_η satisfies

$$
\ddot{V}_\eta(t) + \frac{3}{\eta}\dot{V}_\eta(t) + [\nabla g(X(t))]V_\eta(t) + \boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0,
$$

which is equivalent to

$$
\left[\dot{V}_\eta(t)e^{3t/\eta}\right]' = -e^{3t/\eta}[\nabla g(X(t))]V_\eta(t) - e^{3t/\eta}\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t),
$$

$$\dot{V}_\eta(t)e^{3t/\eta} = -\int_0^t e^{3u/\eta}[\nabla g(X(u))]V_\eta(u)]du - \int_0^t e^{3u/\eta}\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du$$
$$= -\nabla g(X(0))V_\eta(0)\int_0^t e^{3u/\eta}du - \int_0^t e^{3u/\eta}[\nabla g(X(u))V_\eta(u) - \nabla g(X(0))V_\eta(0)]du$$
$$- \int_0^t e^{3u/\eta}\boldsymbol{\sigma}(X(u))d\mathbf{B}(u).$$

Thus, for $t \in (0, \eta]$, we have

$$\left|\frac{\dot{V}_\eta(t)}{t^a}\right| \le \frac{1}{t^a}e^{-3t/\eta}|[\nabla g(X(0))]V_\eta(0)|\int_0^t e^{3u/\eta}du + \frac{1}{t^a}e^{-3t/\eta}\left|\int_0^t e^{3u/\eta}\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|$$
$$+ \frac{1}{(1+a)t^a}e^{-3t/\eta}\int_0^t [L|V_\eta(0)|M_a(0,u;X) + |\nabla g(X(u))|M_a(0,u;V_\eta)]u^{1+a}e^{3u/\eta}du,$$

$$\le t^{1-a}|\nabla g(X(0))V_\eta(0)| + \frac{[L|V_\eta(0)|M_a(0,t;X) + |\nabla g(X(0))|M_a(0,t,V_\eta)]\eta^2}{(1+a)(2+a)}$$
$$+ \frac{LM_a(0,t;X)M_a(0,t,V_\eta)\eta^{3+a}}{(1+a)^2(3+2a)} + \frac{1}{t^a}e^{-3t/\eta}\left|\int_0^t e^{3u/\eta}\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|.$$

Taking the maximum over $t \in (0, \eta]$ on both sides of the above inequality, and using the definition of $M_a(0,t;\cdot)$ (which is increasing in t), we can easily prove the lemma through simple algebra manipulation (which is also employed in the proof of Lemma 4). ■

Lemma 21 *There exists $\eta_0 > 0$, such that for $\eta \in (0, \eta_0]$ and $\eta < t < \eta + \eta_0$, $1 - \frac{(t-\eta)^2}{(1+a)(2+a)}|\nabla g(X(\eta))| - \frac{LM_a(\eta,t;X)(t-\eta)^{3+a}}{(1+a)^2(3+2a)}$ is bounded below from zero. Then, for $\eta \in (0, \eta_0]$ and $\eta < t < \eta + \eta_0$, we have*

$$M_a(\eta,t;V_\eta)\left[1 - \frac{(t-\eta)^2}{(1+a)(2+a)}|\nabla g(X(\eta))| - \frac{LM_a(\eta,t;X)(t-\eta)^{3+a}}{(1+a)^2(3+2a)}\right]$$
$$\le C_1 M_a(0,\eta;V_\eta) + C_2|\nabla g(X(\eta))V_\eta(\eta)|$$
$$+ \frac{(t-\eta)^{2-a}}{2}[L(|V_\eta(\eta)|+1)|\dot{X}(\eta)| + |\nabla g(X(\eta)|] + \frac{(t-\eta)^3}{(1+a)(3+a)}L|\dot{V}_\eta(\eta)|M_a(\eta,t;X)$$
$$+ \frac{(t-\eta)^2}{(1+a)(2+a)}L|V_\eta(\eta)|M_a(\eta,t;X) + \max_{t_0\in(\eta,t]}\left|\frac{1}{t_0^3(t_0-\eta)^a}\int_\eta^{t_0} u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|.$$

Proof. Since $\nabla g(X(\eta))$ and $M_a(\eta,t;X)$ are deterministic and continuous in η, their maximum over η in a neighborhood of 0 is finite. As $t - \eta \to 0$, $\frac{(t-\eta)^2}{(1+a)(2+a)}|\nabla g(X(\eta))| + \frac{LM_a(\eta,t;X)(t-\eta)^{3+a}}{(1+a)^2(3+2a)}$ approaches zero, and the existence of η_0 is obvious. For $t > \eta$, V_η satisfies

$$\ddot{V}_\eta(t) + \frac{3}{t}\dot{V}_\eta(t) + [\nabla g(X(t))]V_\eta(t) + \boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0,$$

which is equivalent to

$$\left[t^3\dot{V}_\eta(t)\right]' = -t^3[\nabla g(X(t))]V_\eta(t) - t^3\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t),$$

and

$$t^3\dot{V}_\eta(t) = \eta^3\dot{V}_\eta(\eta) - \int_\eta^t u^3[\nabla g(X(u))]V_\eta(u)du - \int_\eta^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du$$
$$= \eta^3\dot{V}_\eta(\eta) - \int_\eta^t u^3[\nabla g(X(u))V_\eta(u) - \nabla g(X(\eta))V_\eta(\eta)]du - \int_\eta^t u^3[\nabla g(X(\eta))]V_\eta(\eta)du$$
$$- \int_\eta^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du.$$

Thus, we obtain

$$\frac{|\dot{V}_\eta(t) - \dot{V}_\eta(\eta)|}{(t-\eta)^a} \leq \frac{(t^3-\eta^3)\eta^a}{t^3(t-\eta)^a}\frac{|\dot{V}_\eta(\eta)|}{\eta^a} + \frac{t^4-\eta^4}{4t^3(t-\eta)^a}|\nabla g(X(\eta))V_\eta(\eta)|$$
$$+ \frac{1}{t^3(t-\eta)^a}\int_\eta^t [L|V_\eta(\eta)|(u-\eta)|\dot{X}(\eta)| + |\nabla g(X(u)|(u-\eta)|\dot{V}_\eta(\eta)|]u^3du$$
$$+ \frac{1}{(1+a)t^3(t-\eta)^a}\int_\eta^t [L|V_\eta(\eta)|M_a(\eta,u;X) + |\nabla g(X(u))|M_a(\eta,u;V_\eta)]u^3(u-\eta)^{1+a}du$$
$$+ \frac{1}{t^3(t-\eta)^a}\left|\int_\eta^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|$$
$$\leq \frac{(t^3-\eta^3)\eta^a}{t^3(t-\eta)^a}M_a(0,\eta;V_\eta) + \frac{t^4-\eta^4}{4t^3(t-\eta)^a}|\nabla g(X(\eta))V_\eta(\eta)|$$
$$+ \frac{(t-\eta)^{2-a}}{2}[L(|V_\eta(\eta)|+1)|\dot{X}(\eta)| + |\nabla g(X(\eta)|] + \frac{(t-\eta)^3}{(1+a)(3+a)}L|\dot{V}_\eta(\eta)|M_a(\eta,t;X)$$
$$+ \frac{(t-\eta)^2}{(1+a)(2+a)}[L|V_\eta(\eta)|M_a(\eta,t;X) + |\nabla g(X(\eta))|M_a(\eta,t;V_\eta)]$$
$$+ \frac{LM_a(\eta,t;X)M_a(\eta,t,V_\eta)(t-\eta)^{3+a}}{(1+a)^2(3+2a)} + \frac{1}{t^3(t-\eta)^a}\left|\int_\eta^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|.$$

As in the proof of Lemma 20, replacing t by u in the above inequality, taking the maximum over $u \in (\eta, t]$ on both sides, and using the definition of $M_a(\eta,t;\cdot)$ (which is increasing in t), we conclude that

$$M_a(\eta,t;V_\eta) \leq C_1M_a(0,\eta;V_\eta) + C_2|\nabla g(X(\eta))V_\eta(\eta)|$$
$$+ \frac{(t-\eta)^{2-a}}{2}[L(|V_\eta(\eta)|+1)|\dot{X}(\eta)| + |\nabla g(X(\eta)|] + \frac{(t-\eta)^3}{(1+a)(3+a)}L|\dot{V}_\eta(\eta)|M_a(\eta,t;X)$$
$$+ \frac{(t-\eta)^2}{(1+a)(2+a)}[L|V_\eta(\eta)|M_a(\eta,t;X) + |\nabla g(X(\eta))|M_a(\eta,t;V_\eta)]$$
$$+ \frac{LM_a(\eta,t;X)M_a(\eta,t;V_\eta)(t-\eta)^{3+a}}{(1+a)^2(3+2a)} + \max_{t_0\in(\eta,t]}\left|\frac{1}{t_0^3(t_0-\eta)^a}\int_\eta^{t_0} u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|,$$

which leads to the lemma. ■

Lemma 22 *There exists $\eta_0 > 0$, such that for $\eta \in (0, \eta_0]$ and $\eta < s < t < \eta + s \le T$, $1 - \frac{(t-s)^2}{(1+a)(2+a)}|\nabla g(X(s))| - \frac{LM_a(s,t;X)(t-s)^{3+a}}{(1+a)^2(3+2a)}$ is bounded below from zero. Then, for $\eta \in (0, \eta_0]$ and $\eta < s < t < \eta + s$, we have*

$$
\begin{aligned}
&M_a(s,t;V_\eta)\left[1 - \frac{(t-s)^2}{(1+a)(2+a)}|\nabla g(X(s))| - \frac{LM_a(s,t;X)(t-s)^{3+a}}{(1+a)^2(3+2a)}\right] \\
&\le C_1 M_a(0,s;V_\eta) + C_2|\nabla g(X(s))V_\eta(s)| \\
&+ \frac{(t-s)^{2-a}}{2}[L(|V_\eta(s)|+1)|\dot{X}(s)| + |\nabla g(X(s)|] + \frac{(t-s)^3}{(1+a)(3+a)}L|\dot{V}_\eta(s)|M_a(s,t;X) \\
&+ \frac{(t-s)^2}{(1+a)(2+a)}L|V_\eta(s)|M_a(s,t;X) + \max_{t_0\in(s,t]}\left|\frac{1}{t_0^3(t_0-s)^a}\int_s^{t_0} u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|.
\end{aligned}
$$

Proof. Note that for $s, t > \eta$, V_η satisfies

$$
\left[t^3\dot{V}_\eta(t)\right]' = -t^3[\nabla g(X(t))]V_\eta(t) - t^3\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t),
$$

and

$$
\begin{aligned}
&t^3\dot{V}_\eta(t) = s^3\dot{V}_\eta(s) - \int_s^t u^3[\nabla g(X(u))]V_\eta(u)du - \int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du \\
&= s^3\dot{V}_\eta(s) - \int_s^t u^3[\nabla g(X(u))V_\eta(u) - \nabla g(X(s))V_\eta(s)]du - \int_s^t u^3[\nabla g(X(s))]V_\eta(s)du \\
&- \int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du.
\end{aligned}
$$

Then, we work on $\frac{|\dot{V}_\eta(t)-\dot{V}_\eta(s)|}{(t-s)^a}$. The reminder of the proof argument is the same as in the proof of Lemma 21 with η replaced by s. ■

Lemma 23 *We have*

$$
P\left(\max_{v\in(s,t]}\left|\frac{1}{v^3(v-s)^a}\int_s^v u^3\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right| < \infty \text{ for all } 0 < s < t\right) = 1.
$$

Proof. We need to show that the Gaussian process $\int_s^v u^3\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)$ has the a-th derivative. Indeed, we have

$$
\begin{aligned}
&\frac{1}{v^3(v-s)^a}\int_s^v u^3\boldsymbol{\sigma}(X(u))d[\mathbf{B}(u)-\mathbf{B}(s)] = \boldsymbol{\sigma}(X(v))\frac{[\mathbf{B}(v)-\mathbf{B}(s)]}{(v-s)^a} \\
&- \frac{1}{v^3(v-s)^a}\int_s^v \frac{d[u^3\boldsymbol{\sigma}(X(u))]}{du}[\mathbf{B}(u)-\mathbf{B}(s)]du \\
&= \frac{[\mathbf{B}(v)-\mathbf{B}(s)]}{(v-s)^a}\boldsymbol{\sigma}(X(v)) - \frac{1}{v^3}\int_s^v \frac{d[u^3\boldsymbol{\sigma}(X(u))]}{du}\frac{(u-s)^a}{(v-s)^a}\frac{\mathbf{B}(u)-\mathbf{B}(s)}{(u-s)^a}du,
\end{aligned}
$$

which is a.s. finite, due to the fact that $0 < (u-s)^a/(v-s)^a \le 1$, $X(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ are continuously differentiable and Lipschitz, and the Brownian motion has a well-known property that for all $u > s > 0$, $\sup_{s<u}\frac{|\mathbf{B}(u)-\mathbf{B}(s)|}{(u-s)^a}$ is a.s. finite. ■

Lemma 24 *For any given $T > 0$, $V_\eta(t)$ is stochastically equicontinuous and stochastically bounded on $[0, T]$ uniformly over η.*

Proof. Take η_* to be the smallest η_0 defined in Lemmas 20-22. Divide the interval $[0, T]$ into $N = [T/\eta_* + 1]$ number of subintervals with length almost equal to η_* (except for the last one), and denote by $\mathcal{I}_i = [s_{i-1}, s_i]$, $i = 1, \cdots, N$ (with $s_0 = 0$, $s_N = T$, $\mathcal{I}_1 = [0, T/N]$, $1/N < \eta_*/T$, $\mathcal{I}_N = [s_{N-1}, T]$). First, for $t \in \mathcal{I}_1$, we have

$$|\dot{V}_\eta(t)| \leq |\mathcal{I}_1|^a M_a(\mathcal{I}_1; V_\eta), \;\; |V_\eta(t)| \leq |V_\eta(0)| + \int_{\mathcal{I}_1} |\dot{V}_\eta(u)| du,$$

and the upper bounds on $\dot{V}_\eta(t)$ and $V_\eta(t)$ over $\mathcal{I}_1$ are a.s. finite uniformly over η, which implies that $V_\eta(t)$ is stochastically equicontinuous and stochastically bounded over $\mathcal{I}_1$.

For $t \in \mathcal{I}_i$, $i = 2, \cdots, N$, we have

$$|\dot{V}_\eta(t) - \dot{V}_\eta(s_{i-1})| \leq |\mathcal{I}_i|^a M_a(\mathcal{I}_i; V_\eta),$$

and

$$|V_\eta(t)| \leq |V_\eta(s_{i-1})| + |\mathcal{I}_i||\dot{V}_\eta(s_{i-1})| + \int_{\mathcal{I}_i} |\dot{V}_\eta(u) - \dot{V}_\eta(s_{i-1})| du.$$

Note that N is free of η. We use the above two inequalities to prove by induction that the upper bounds of $V_\eta(t)$ and $\dot{V}_\eta(t)$ on $[0, T]$ are a.s. finite uniformly over η. Assume that the upper bounds of $V_\eta(t)$ and $\dot{V}_\eta(t)$ on $\cup_{j=1}^{i-1} \mathcal{I}_j$ are a.s. finite uniformly over η. The above-mentioned two inequalities immediately show that their upper bounds on $\mathcal{I}_i$ are also a.s. finite uniformly over η. This implies that the uniform finite bounds of $V_\eta(t)$ and $\dot{V}_\eta(t)$ on $\cup_{j=1}^{N} \mathcal{I}_j = [0, T]$ and, thus, $V_\eta(t)$ is stochastically equicontinuous and stochastically bounded on $[0, T]$. ■

Lemma 25 *For fixed (δ, m), the second-order SDEs (4.35) and (4.36) have unique solutions in the distributional sense.*

Proof. Due to the similarity, we provide proof arguments for (5.70) only. We take a decreasing sequence of η in the following manner: η_k, $k = 1, 2, \cdots$, are decreasing, and as $k \to \infty$, $\eta_k \to 0$. Lemma 24 implies that $\{V_{\eta_k}(t), k = 1, 2, \ldots, \}$ is tight and, thus, there exists a subsequence that has a weak limit process $V_\dagger(t)$. We show that $V_\dagger(t)$ satisfies (4.36). Without loss of generality, we may assume that $V_{\eta_k}(t)$ weakly converges to $V_\dagger(t)$. Further, using Skorokhod's representation theorem, we may assume that $V_{\eta_k}(t)$ converges to $V_\dagger(t)$ a.s.. $V_{\eta_k}(t)$ obeys the initial condition $V_{\eta_k}(0) = \dot{V}_{\eta_k}(0) = 0$; thus, $V_\dagger(0) = 0$, and

$$\frac{|V_\dagger(t) - V_\dagger(0)|}{t} = \lim_{k\to\infty} \frac{|V_{\eta_k}(t) - V_{\eta_k}(0)|}{t} = \lim_{k\to\infty} |\dot{V}_{\eta_k}(\xi_k)| \leq \limsup_{k\to\infty} [t^a M_a(0, t, V_{\eta_k})].$$

Since $M_a(0, t, V_{\eta_k})$ is a.s. finite uniformly over η_k, taking $t \to 0$, we obtain $\dot{V}_\dagger(0) = 0$. For $t > \eta_k$, the second-order SDE (5.72) is equivalent to the following smoothed stochastic

differential equation system,

$$\begin{aligned}\dot{V}_{\eta_k}(t) &= \frac{2}{t}\,Y_{\eta_k}(t) - \frac{2}{t}\,V_{\eta_k}(t)\\ \dot{Y}_{\eta_k}(t) &= -\frac{t}{2}\,[\nabla g(X(t))]V_{\eta_k}(t) - \frac{t}{2}\,\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t).\end{aligned}$$

Its inherited initial conditions are $V_{\eta_k}(0) = Y_{\eta_k}(0) = c$ and $\dot{V}_{\eta_k}(0) = 0$. The right-hand side of the second equation in the above system implies that as $k \to \infty$, $Y_{\eta_k}(t)$ converges to $Y(t)$ defined by

$$\dot{Y}(t) = -\frac{t}{2}\,[\nabla g(X(t))]V_\dagger(t) - \frac{t}{2}\,\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t), \quad Y(0) = c,$$

which in turn shows that $\dot{V}_{\eta_k}(t)$ converges to $\dot{V}_*(t)$, given by

$$\dot{V}_*(t) = \frac{2}{t}\,Y(t) - \frac{2}{t}\,V_\dagger(t).$$

Since $V_{\eta_k}(t)$ converges to $V_\dagger(t)$, $\dot{V}_*(t) = \dot{V}_\dagger(t)$. Thus, $V_\dagger(t)$ satisfies

$$\dot{V}_\dagger(t) = \frac{2}{t}\,Y(t) - \frac{2}{t}\,V_\dagger(t),$$

which impies that $V_\dagger(t)$ obeys

$$\ddot{V}_\dagger(t) + \frac{3}{t}\dot{V}_\dagger(t) + [\nabla g(X(t))]V_\dagger(t) + \boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0.$$

Suppose that the equation has two solutions $(V(t), \mathbf{B}(t))$ and $(V_*(t), \mathbf{B}_*(t))$. Then, we may realize both solutions on some common probability space such that $\mathbf{B}(t) = \mathbf{B}_*(t)$. Hence, $U(t) = V(t) - V_*(t)$ obeys

$$\ddot{U}(t) + \frac{3}{t}\dot{U}(t) + [\nabla g(X(t))]U(t) = 0, \qquad U(0) = \dot{U}(0) = 0,$$

which has a unique solution zero, as it is a second-order ODE similar to ODEs (2.6) and (3.15). Thus $V(t) = V_*(t)$—that is, the two solutions have an identical distribution, which proves the unique solution. ■

Remark 15 *As in Section 5.1, the inhomogeneous linear SDE (4.36) has a corresponding homogeneous linear ODE, and its solution $V(t)$ enjoys an explicit expression in terms of the solution for the homogeneous linear ODE. We may prove the unique solution result by analyzing the ODE and using the explicit expression. In fact, denote by $\Pi_1(t)$ an invertible solution of the following 2nd order linear matrix ODE,*

$$\ddot{\Pi}_1(t) - \dot{\Pi}_1(t)\left[\frac{3}{t} + \nabla\log \boldsymbol{I\!H}g(X(t))\right] + \Pi_1(t)[\boldsymbol{I\!H}g(X(t))] = 0,$$

and by $\Pi_2(t)$ the solution of the following matrix ODE,

$$\dot{\Pi}_2(t) = \Pi_2(t)\left[\frac{\dot{\Pi}_1(t)}{\boldsymbol{I\!H}g(X(t))}\right]^{-1}\Pi_1(t), \qquad \Pi_2(0) = \mathbf{I}.$$

Let

$$\Pi_*(t) = \frac{\dot{\Pi}_1(t)}{\mathbb{H}g(X(t))}.$$

Then, $(\Pi_1(t), \Pi_*(t), \Pi_2(t))$ *satisfies the 1st order linear ODE system*

$$\begin{cases} d\Pi_1(t) = \Pi_*(t)\mathbb{H}g(X(t))dt, \\ d\Pi_*(t) = \left[\frac{3}{t}\Pi_*(t) - \Pi_1(t)\right] dt, \\ d\Pi_2(t) = \Pi_2(t)[\Pi_*(t)]^{-1}\Pi_1(t), \ \ \Pi_2(0) = \mathbf{I}. \end{cases}$$

Direct calculations with Itô lemma lead to

$$\Pi_2(t)V(t) = -\int_0^t \left\{\int_s^t \Pi_2(v)\left[\Pi_*(v)\right]^{-1} dv\right\}\left[\Pi_*(s)\right]^{-1}\boldsymbol{\sigma}((X(s))d\mathbf{B}(s).$$

Thus, the solution of SDE (4.36) has the following expression,

$$V(t) = -[\Pi_2(t)]^{-1}\int_0^t \Pi_2(v)\left[\frac{\dot{\Pi}_1(v)}{\mathbb{H}g(X(v))}\right]^{-1}\left\{\int_0^v \left[\frac{\dot{\Pi}_1(u)}{\mathbb{H}g(X(u))}\right]^{-1}\boldsymbol{\sigma}((X(u))d\mathbf{B}(u)\right\} dv.$$

5.6.2 Weak Convergence of $V_\delta^m(t)$

Lemma 26 *For* $X(t)$, $X_\delta^m(t)$, *and* $V_\delta^m(t)$, *we have*

$$M_1(s,t;X) \le \frac{1}{1-L(t-s)^2/6}\left[\left(\frac{3}{s}+\frac{L(t-s)}{2}\right)|\dot{X}(s)| + |\nabla g(X(s))|\right], \text{ if } t-s < \sqrt{\frac{3}{L}},$$

$$M_a(s,t;X_\delta^m) \le \frac{1}{1-L(t-s)^2/[(a+1)(a+2)]}\left[(t-s)^{1-a}\left(\frac{3}{s}+\frac{L(t-s)}{2}\right)|\dot{X}_\delta^m(s)|\right.$$
$$\left.+(t-s)^{1-a}|\nabla g(X_\delta^m(s))| + \max_{v\in(s,t]}\frac{\delta^{1/4}m^{-1/2}}{4v^3(v-s)^a}\left|\int_s^v u^3\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|\right],$$

$$M_a(s,t;V_\delta^m) \le \frac{1}{1-L(t-s)^2/[(a+1)(a+2)]}$$
$$\left[(t-s)^{1-a}\left\{2L|V_\delta^m(s)| + [3/s + L(t-s)]|\dot{V}_\delta^m(s)|\right\}\right.$$
$$\left.+\max_{v\in(s,t]}\frac{1}{v^3(v-s)^a}\left|\int_s^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\right],$$

when $s > 0$ *and* $t - s < \sqrt{(a+1)(a+2)/(2L)}$. *In particular, for* $s = 0$ *we have*

$$M_1(0,t;X) \le \frac{|\nabla g(x_0)|}{1-Lt^2/6},$$

$$M_a(0,t;X_\delta^m) \le \frac{t^{1-a}|\nabla g(x_0)| + \max_{v\in(s,t]}\frac{\delta^{1/4}(mT)^{-1/2}}{4v^{3+a}}\left|\int_0^v u^3\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|}{1-Lt^2/[(a+1)(a+2)]},$$

$$M_a(0,t;V_\delta^m) \le \frac{1}{1-Lt^2/[(a+1)(a+2)]}\max_{v\in(0,t]}\left[\frac{1}{v^{3+a}}\left|\int_0^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\right].$$

Proof. Because of similarity, we provide proof arguments only for $M_1(s,t;V_\delta^m$. Let $H(t;V_\delta^m)$ $= \delta^{-1/4}m^{1/2}[\nabla g(X_\delta^m(t)) - \nabla g(X(t))]$, and $J(s,t;H,V_\delta^m) = \int_s^t u^3[H(u;V_\delta^m) - H(s;V_\delta^m)]du$. Then, we obtain

$$
\begin{aligned}
&|H(t;V_\delta^m)| \le L\delta^{-1/4}m^{1/2}|X_\delta^m(t) - X(t)| = L|V_\delta^m(t)|,\\
&|H(t;V_\delta^m) - H(s;V_\delta^m)| = \delta^{-1/4}m^{1/2}|\nabla[g(X_\delta^m(t)) - g(X_\delta^m(s)) - g(X(t)) + g(X(s))]|\\
&\le L\delta^{-1/4}m^{1/2}|X_\delta^m(t) - X(t)| + L\delta^{-1/4}m^{1/2}|X_\delta^m(s) - X(s)| = L|V^n(t)| + L|V_\delta^m(s)|,\\
&V_\delta^m(t) = \int_s^t \dot V_\delta^m(u)du + V_\delta^m(s) = \int_s^t [\dot V_\delta^m(u) - \dot V^n(s)]du + V_\delta^m(s) + (t-s)\dot V^n(s),\\
&|H(t;V_\delta^m) - H(s;V_\delta^m)| \le L\int_s^t |\dot V_\delta^m(u) - \dot V_\delta^m(s)|du + L[2|V_\delta^m(s)| + |(t-s)\dot V_\delta^m(s)|],\\
&\int_s^t |\dot V_\delta^m(u) - \dot V_\delta^m(s)|du \le \int_s^t (u-s)^a \frac{|\dot V_\delta^m(u) - \dot V_\delta^m(s)|}{(u-s)^a}du \le \int_s^t (u-s)^a M_a(s,t;V_\delta^m)du\\
&= \frac{M_a(s,t;V_\delta^m)(t-s)^{a+1}}{a+1},\\
&\frac{L}{a+1}\int_s^t M_a(s,u;V_\delta^m)u^3(u-s)^{a+1}du \le \frac{LM_a(s,t;V_\delta^m)t^3(t-s)^{a+2}}{(a+1)(a+2)},\\
&|J(s,t;H,V_\delta^m)| \le \frac{Lt^3(t-s)^{a+2}}{(a+1)(a+2)}M_a(s,t;V_\delta^m) + L[2|V_\delta^m(s)| + (t-s)|\dot V_\delta^m(s)|]t^3(t-s).
\end{aligned}
$$

SDE (4.35) is equivalent to

$$
\begin{aligned}
&\frac{t^3\dot V_\delta^m(t)}{dt} = -t^3H(t;V_\delta^m) - t^3\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t), \text{ which implies that}\\
&t^3\dot V_\delta^m(t) - s^3\dot V_\delta^m(s) = -\int_s^t u^3H(u;V_\delta^m)du - \int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\\
&= -\frac{t^4-s^4}{4}H(s;V_\delta^m) - J(s,t;H,V_\delta^m) - \int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du,\\
&\frac{\dot V_\delta^m(t) - \dot V_\delta^m(s)}{t-s} = -\frac{t^3-s^3}{t^3(t-s)}\dot V_\delta^m(s) - \frac{t^4-s^4}{4t^3(t-s)}H(s;V_\delta^m) - \frac{J(s,t;H,V_\delta^m)}{t^3(t-s)}\\
&- \frac{1}{t^3(t-s)}\int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du,
\end{aligned}
$$

and using the upper bounds of $H(s;V_\delta^m)$ and $J(s,t;H,V_\delta^m)$ and algebraic manipulations, we obtain

$$
\begin{aligned}
&\frac{|\dot V_\delta^m(t) - \dot V_\delta^m(s)|}{t-s} \le \frac{t^3-s^3}{t^3(t-s)}|\dot V_\delta^m(s)| + \frac{t^4-s^4}{4t^3(t-s)}|H(s;V_\delta^m)| + \frac{|J(s,t;H,V_\delta^m)|}{t^3(t-s)}\\
&+ \frac{1}{t^3(t-s)}\left|\int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\\
&\le \frac{t^2+st+s^2}{t^3}|\dot V_\delta^m(s)| + \frac{(t^2+s^2)(t+s)}{2t^3}L|V_\delta^m(s)| + M_a(s,t;V_\delta^m)\frac{L(t-s)^{a+1}}{(a+1)(a+2)}\\
&+ L[2|V_\delta^m(s)| + (t-s)|\dot V_\delta^m(s)|] + \frac{1}{t^3(t-s)}\left|\int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|.
\end{aligned}
$$

As the above inequality holds for any $s < t$, an application of the definition of $M_a(s,t;V_\delta^m)$ leads to

$$M_a(s,t;V_\delta^m) \le (t-s)^{1-a}\left\{\frac{3}{s}|\dot{V}_\delta^m(s)| + L[4|V_\delta^m(s)| + (t-s)|\dot{V}_\delta^m(s)|]\right\}$$
$$+ M_a(t,s;V_\delta^m)\frac{L(t-s)^2}{(a+1)(a+2)} + \max_{v\in(s,t]}\frac{1}{v^3(v-s)^a}\left|\int_s^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|,$$

and solving for $M_a(s,t:V_\delta^m)$ to obtain

$$M_a(s,t;V_\delta^m) \le \frac{1}{1-L(t-s)^2/[(a+1)(a+2)]}$$
$$\left[(t-s)^{1-a}\left\{4L|V_\delta^m(s)| + [3/s + L(t-s)]|\dot{V}_\delta^m(s)|\right\}\right.$$
$$\left. + \max_{v\in(s,t]}\frac{1}{v^3(v-s)^a}\left|\int_s^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\right],$$

when $s > 0$ and $t - s < \sqrt{(a+1)(a+2)/(2L)}$. If $s = 0$, we replace the coefficient $3/s$ by $1/t$ in above inequality, and $V_\delta^m(0) = \dot{V}_\delta^m(0) = 0$, $X_\delta^m(0) = X(0) = x_0$. Then, we conclude that

$$M_a(0,t;V_\delta^m) \le \frac{1}{1-Lt^2/[(a+1)(a+2)]}\max_{v\in(0,t]}\left[\frac{1}{v^{3+a}}\left|\int_0^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\right],$$

which proves the lemma. ■

Lemma 27 *For any given $T > 0$, we have*

$$\max_{t\in[0,T]}|V_\delta^m(t)| = O_P(1), \quad \max_{t\in[0,T]}|X_\delta^m(t) - X(t)| = O_P(\delta^{1/4}m^{-1/2}),$$

$$\max_{t\in[0,T]}|\dot{V}_\delta^m(t)| = O_P(1), \quad \max_{t\in[0,T]}|\dot{X}_\delta^m(t) - \dot{X}(t)| = O_P(\delta^{1/4}m^{-1/2}).$$

Proof. As $V_\delta^m(t) = \delta^{-1/4}m^{1/2}[X_\delta^m(t) - X(t)]$, we need to establish the results for $V_\delta^m(t)$ only. Divide interval $[0,T]$ into $N = \left\lceil T\sqrt{2L/\{(a+1)(a+2)\}}\right\rceil + 1$ number of subintervals with length $\sqrt{(a+1)(a+2)/(2L)}$ (except for the last one), and denote the subintervals by $\mathcal{I}_i = [s_{i-1}, s_i]$, $i = 1, \cdots, N$ (with $s_0 = 0$, $s_N = T$, $\mathcal{I}_1 = [0, \sqrt{3/L}]$, $\mathcal{I}_N = [s_{N-1}, T]$). First, for $t \in \mathcal{I}_1$, from Lemma 26 we have

$$|\dot{V}_\delta^m(t)| \le |\mathcal{I}_1|^a M_a(\mathcal{I}_1;V_\delta^m) \le C\max_{v\in(0,s_1]}\left[\frac{1}{v^{3+a}}\left|\int_0^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\right],$$

$$|V_\delta^m(t)| \le |V_\delta^m(0)| + \int_{\mathcal{I}_1}|\dot{V}_\delta^m(u)|du \le C\max_{v\in(0,s_1]}\left[\frac{1}{v^{3+a}}\left|\int_0^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\right].$$

The upper bounds of $V_\delta^m(t)$ and $\dot{V}_\delta^m(t)$ on $\mathcal{I}_1$ are a.s. finite uniformly over (δ, m).

For $t \in \mathcal{I}_i$, $i = 2, \cdots, N$, from Lemma 26 we have

$$|\dot{V}_\delta^m(t) - \dot{V}_\delta^m(s_{i-1})| \le |\mathcal{I}_i|^a M_a(\mathcal{I}_i, V_\delta^m) \le C\left[4L|V_\delta^m(s_{i-1})| + (3/s_1 + Ls_1)|\dot{V}_\delta^m(s_{i-1})|\right]$$
$$+ C \max_{v\in(s_{i-1},s_i]} \frac{1}{v^3(v-s_{i-1})^a}\left|\int_{s_{i-1}}^v u^3 \boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|,$$
$$|V_\delta^m(t)| \le |V_\delta^m(s_{i-1})| + |\mathcal{I}_i||\dot{V}_\delta^m(s_{i-1})| + \int_{\mathcal{I}_i} |\dot{V}_\delta^m(u) - \dot{V}_\delta^m(s_{i-1})|du$$
$$\le |V_\delta^m(s_{i-1})| + \sqrt{3/L}|\dot{V}_\delta^m(s_{i-1})| + C\left[4L|V_\delta^m(s_{i-1})| + (3/s_1 + Ls_1)|\dot{V}_\delta^m(s_{i-1})|\right]$$
$$+ C \max_{v\in(s_{i-1},s_i]} \frac{1}{v^3(v-s_{i-1})^a}\left|\int_{s_{i-1}}^v u^3 \boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|.$$

We use the above two inequalities to prove by induction that the upper bounds of $V_\delta^m(t)$ and $\dot{V}_\delta^m(t)$ on $[0,T]$ are a.s. finite uniformly over (m,δ). Assume that the upper bounds of $V_\delta^m(t)$ and $\dot{V}_\delta^m(t)$ on $\cup_{j=1}^{i-1}\mathcal{I}_j$ are a.s. finite uniformly over (m,δ). Note that $\max_{v\in(s_{i-1},s_i]} \frac{1}{v^3(v-s_{i-1})^a}\left|\int_{s_{i-1}}^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|$ is a.s. finite, and N is free of (m,δ). The above-mentioned two inequalities immediately show that the upper bounds of $V_\delta^m(t)$ and $\dot{V}_\delta^m(t)$ on $\mathcal{I}_i$ are also a.s. finite uniformly over (m,δ). This implies that the uniform finite bounds of $V_\delta^m(t)$ and $\dot{V}_\delta^m(t)$ on $\cup_{j=1}^N \mathcal{I}_j = [0,T]$. ■

Lemma 28 *For any given $T > 0$, as $\delta \to 0$ and $m \to \infty$, $V_\delta^m(t)$ is stochastically equicontinuous on $[0,T]$.*

Proof. Lemma 27 proves that $\max_{t\in[0,T]}|V_\delta^m(t)| = O_P(1)$ and $\max_{t\in[0,T]}|\dot{V}_\delta^m(t)| = O_P(1)$, which implies that $V_\delta^m(t)$ is stochastically equicontinuous on $[0,T]$. ■

Proof of Theorem 8. Lemma 25 indicates the unique solutions of SDEs. Moreover, (4.36) is a linear SDE, and its constant term linearly depends on $\mathbf{B}(t)$; thus, its solution $V(t)$ is Gaussian. As in Section 5.1.2, we can easily establish finite distribution convergence for $V_\delta^m(t)$. Lemma 28 along with the finite distribution convergence immediately lead to the conclusion that as $\delta \to 0$ and $m \to \infty$, $V_\delta^m(t)$ weakly converges to $V(t)$. ■

5.7 Proof of Theorem 9

Recall that sequences $\{x_k, y_k\}$ and $\{x_k^m, y_k^m\}$ are defined by algorithms (2.4) and (4.26), respectively, with initial values $x_0^m = y_0^m = x_0$; $X_\delta^m(t)$ and $X(t)$ are the solutions of ODE (2.6) and SDE (4.35), respectively.

We discretize SDE (4.34), which is equivalent to (4.35), to define a new sequence in the following manner. Let $\{\check{x}_k^m, \check{y}_k^m\}$ be the sequence, with initial values $\check{x}_0^m = \check{y}_0^m = x_0$, generated by

$$\check{x}_k^m = \check{y}_{k-1}^m - \delta\nabla g(\check{y}_{k-1}^m) - m^{-1/2}\delta^{3/4}[H(t_k) - H(t_{k-1})], \check{y}_k^m = \check{x}_k^m + \frac{k-1}{k+2}(\check{x}_k^m - \check{x}_{k-1}^m), \quad (5.73)$$

where $H(t) = \int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u)$ and $t_k = k\delta^{1/2}$.

We rewrite algorithm (4.26) to generate $\{x_k^m, y_k^m\}$ as follows:

$$x_k^m = y_{k-1}^m - \delta \nabla g(y_{k-1}^m) - m^{-1/2}\delta^{3/4}[H_\delta^m(t_k) - H_\delta^m(t_{k-1})], \;\; y_k^m = x_k^m + \frac{k-1}{k+2}(x_k^m - x_{k-1}^m). \tag{5.74}$$

Note that (5.73) and (5.74) share the same recursive structure with the only difference being between $H_\delta^m(t)$ and $H(t)$. The approach to our proof is that (i) Lemma 34 below reveals that $\{x_k^m, y_k^m\}$ and $\{\check{x}_k^m, \check{y}_k^m\}$ can be realized on certain probability spaces within a small order distance; (ii) Lemma 38 below derives an order bound for the discretization error $\check{x}_k^m - X_\delta^m(t_k)$; (iii) the theorem is proved by combining two lemmas in (i) and (ii).

Lemma 29

$$\max_{k \le T\delta^{-1/2}} |x_k - X(t_k)| = O(\delta^{1/2}|\log\delta|), \quad \max_{k \le T\delta^{-1/2}} |z_k - \dot{X}(t_k)| = O(\delta^{1/2}|\log\delta|),$$

$$\max_{k \le T\delta^{-1/2}} |y_k - x_k| = O(\delta^{1/2}).$$

Proof. As $X(t)$ is the solution of ODE (2.6), it is shown that $X(t)$, $\dot{X}(t)$, and $\nabla g(X(t))$ are uniformly bounded on $[0,T]$, and Lemma 4 further indicates that $\dot{X}(t)$ is Lipschitz. Let $Z(t) = \dot{X}(t)$. With deterministic sequence $\{x_k, y_k\}$ given by algorithm (2.4), we define $z_0 = 0$, $z_k = (x_k - x_{k-1})/\delta^{1/2}$. Lemma 8 indicates that

$$\max_{k \le T\delta^{-1/2}} |x_k - X(t_k)| = O(\delta^{1/2}|\log\delta|), \quad \max_{k \le T\delta^{-1/2}} |z_k - Z(t_k)| = O(\delta^{1/2}|\log\delta|),$$

and $y_k - x_k = \frac{3k+4}{k+3}\delta^{1/2} z_k$, which implies that

$$\max_{k \le T\delta^{-1/2}} |y_k - x_k| \le 3\delta^{1/2}\left(\max_{k \le T\delta^{-1/2}} |z_k - Z(t_k)| + \max_{k \le T\delta^{-1/2}} |Z(t_k)|\right) = O(\delta^{1/2}). \blacksquare$$

As in the proofs of Theorems 4-6, we use notations $R^m(\theta; \mathbf{U}_m^*(t)) = (R_1^m(\theta; \mathbf{U}_m^*(t)), \cdots, R_p^m(\theta; \mathbf{U}_m^*(t)))'$, where

$$R_j^m(\theta; \mathbf{U}_m^*(t)) = \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^m \frac{\partial}{\partial\theta_j}\ell(\theta; U_i^*(t)) - \frac{\partial}{\partial\theta_j}g(\theta)\right], \;\; j = 1, \cdots, p.$$

Lemma 30

$$\max_{k \le T\delta^{-1/2}} E[|R^m(X(t_{k-1}); \mathbf{U}_{mk}^*)|^4] \le C.$$

Proof. For simplicity, we write $R_i = R^m(X(t_i); \mathbf{U}_{m(i+1)}^*)$, and $\mathrm{r}_q = \nabla\ell(X(t_i); U_{q(i+1)}^*) - \nabla g(X(t_i))$. Then, $R_i = m^{-1/2}\sum_{q=1}^m \mathrm{r}_q$. Since $X(t)$ is deterministic, and U_{qk}^*, $q = 1, 2, \cdots, m$, are independent, we have that $\mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{r}_m$ are independent with mean 0, and

$$|\mathrm{r}_q| \le h_1(U_{q(i+1)}^*)|X(t_i) - \theta_0| + |\nabla\ell(\theta_0; U_{q(i+1)}^*)| + |\nabla g(X(t_i))|,$$

$$E|\mathrm{r}_q|^2 \le 3 \cdot (E[h_1^2(U)]|X(t_i)-\theta_0|^2 + E|\nabla \ell(\theta_0;U)|^2 + |\nabla g(X(t_i))|^2),$$
$$E|\mathrm{r}_q|^4 \le 27 \cdot (E[h_1^4(U)]|X(t_i)-\theta_0|^4 + E|\nabla \ell(\theta_0;U)|^4 + |\nabla g(X(t_i))|^4).$$

Note that $\sup_t |X(t)-\theta_0|$ and $\sup_t |\nabla g(X(t))|$ are bounded, and Assumption A1 implies that $E|\mathrm{r}_q|^2$ and $E|\mathrm{r}_q|^4$ are uniformly bounded. Therefore, we obtain

$$E|R_i|^2 = m^{-1}E\left|\sum_{q=1}^m \mathrm{r}_q\right|^2 = m^{-1}\sum_{q=1}^m E|\mathrm{r}_q|^2 \le C,$$

$$\begin{aligned}
E|R_i|^4 &= m^{-2}E\left|\sum_{q=1}^m \mathrm{r}_q\right|^4 \\
&= m^{-2}E\left[\left(\sum_{q=1}^m \mathrm{r}_q'\right)\left(\sum_{q=1}^m \mathrm{r}_q\right)\left(\sum_{q=1}^m \mathrm{r}_q'\right)\left(\sum_{q=1}^m \mathrm{r}_q\right)\right] \\
&= m^{-2}E\left[\sum_{q=1}^m |\mathrm{r}_q|^4 + \sum_{p<q}(4(\mathrm{r}_p'\mathrm{r}_q)^2 + 2|\mathrm{r}_p|^2|\mathrm{r}_q|^2)\right] \\
&\le m^{-2}\left[\sum_{q=1}^m E|\mathrm{r}_q|^4 + \sum_{p<q} 6E|\mathrm{r}_p|^2 E|\mathrm{r}_q|^2\right] \\
&\le m^{-2}(mC + 3m^2C^2) \le C + 3C^2,
\end{aligned}$$

where we use the inequality $(\mathrm{r}_p'\mathrm{r}_q)^2 \le |\mathrm{r}_p|^2|\mathrm{r}_q|^2$. In other words, we show that $E(|R_i|^2)$ and $E(|R_i|^4)$ are uniformly bounded. ■

Lemma 31

$$\max_{t\in[0,T]} E[|G_\delta^m(t) - \check{G}_\delta^m(t)|^2] \le C \max_{k\le k_T} E[|y_k^m - X(t_k)|^2],$$

where $G_\delta^m(t)$ and $\check{G}_\delta^m(t)$ are defined in (5.78) and (5.80) below, respectively.

Proof. Define filtration $\mathcal{F}_t = \sigma(y_\delta^m(s), \mathbf{U}_m^*(s); s \le t)$, where $y_\delta^m(t)$ and $\mathbf{U}_m^*(t)$ are given by (4.27). Then, for $i > j$, we obtain

$$\begin{aligned}
&E\{(R^m(y_i^m;\mathbf{U}_{m(i+1)}^*) - R^m(X(t_i);\mathbf{U}_{m(i+1)}^*))'(R^m(y_j^m;\mathbf{U}_{m(j+1)}^*) - R^m(X(t_j);\mathbf{U}_{m(j+1)}^*)\} \\
&= E\{E[(R^m(y_i^m;\mathbf{U}_{m(i+1)}^*) - R^m(X(t_i);\mathbf{U}_{m(i+1)}^*))' \\
&\qquad (R^m(y_j^m;\mathbf{U}_{m(j+1)}^*) - R^m(X(t_j);\mathbf{U}_{m(j+1)}^*))|\mathcal{F}_{t_j}]\} \\
&= E\{(R^m(y_i^m;\mathbf{U}_{m(i+1)}^*) - R^m(X(t_i);\mathbf{U}_{m(i+1)}^*))' \\
&\qquad E[R^m(y_j^m;\mathbf{U}_{m(j+1)}^*) - R^m(X(t_j);\mathbf{U}_{m(j+1)}^*)|\mathcal{F}_{t_j}]\} \\
&= 0.
\end{aligned}$$

Set $r_{qi} = \nabla \ell(y_i^m; U^*_{q(i+1)}) - \nabla g(y_i^m) - (\nabla \ell(X(t_i); U^*_{q(i+1)}) - \nabla g(X(t_i))$. Then, we have

$$D_i \triangleq R^m(y_i^m; \mathbf{U}^*_{m(i+1)}) - R^m(X(t_i); \mathbf{U}^*_{m(i+1)}) = m^{-1/2} \sum_{q=1}^{m} r_{qi},$$

and for $q \neq s$,

$$E(r'_{qi} r_{si}) = E(E(r'_{qi} r_{si} | \mathcal{F}_{t_i})) = E(E(r'_{qi} | \mathcal{F}_{t_i}) E(r_{si} | \mathcal{F}_{t_i})) = 0.$$

On the other hand, we obtain $|r_{qi}| \leq (h_1(U_{q(i+1)}) + L)|y_i^m - X(t_i)|$,

$$E(|r_{qi}|^2) \leq E(E[(h_1(U_{q(i+1)}) + L)^2 |y_i^m - X(t_i)|^2 | \mathcal{F}_{t_i}]) \leq C \cdot E|y_i^m - X(t_i)|^2,$$

and, thus, we arrive at

$$E|D_i|^2 \leq C \cdot E|y_i^m - X(t_i)|^2.$$

Direct calculations show that for $t_{k+1} \leq t < t_{k+2}$,

$$E|G_\delta^m(t) - \check{G}_\delta^m(t)|^2 = \frac{\delta^{1/2}}{c_k^2} E \left| \sum_{i=1}^{k} c_i D_i \right|^2 = \frac{\delta^{1/2}}{c_k^2} \sum_{i=1}^{k} c_i^2 E|D_i|^2$$
$$\leq k\delta^{1/2} C \cdot \max_{1 \leq i \leq k} E|y_i^m - X(t_i)|^2 \leq C \cdot \max_{1 \leq i \leq k} E|y_i^m - X(t_i)|^2,$$

and, therefore, we conclude that

$$\max_{t \in [0,T]} E|G_\delta^m(t) - \check{G}_\delta^m(t)|^2 \leq C \cdot \max_{k \leq k_T} E|y_k^m - X(t_k)|^2. \blacksquare$$

Lemma 32

$$\max_{k \leq T\delta^{-1/2}} E[|y_k^m - y_k|^4] = O(m^{-2}), \quad \max_{k \leq T\delta^{-1/2}} E[|y_k^m - X(t_k)|^4] = O(m^{-2} + \delta^{1/2} |\log \delta|).$$

Proof. The second result can be easily established from the first one and Lemma 29. We prove the first result below.

Recall that $d_0 = 0$, $d_k = x_k - x_{k-1}$, $d_0^m = 0$, $d_k^m = x_k^m - x_{k-1}^m$, $a_k = |x_k - x_k^m|$, $b_k = |d_k - d_k^m|$. We have $a_0 = 0$, $b_0 = 0$,

$$a_k \leq |x_{k-1} - x_{k-1}^m| + |d_k - d_k^m| = a_{k-1} + b_k \leq S_k,$$

where $S_k = b_0 + b_1 + \cdots + b_k$. Moreover, we obtain

$$d_{k+1} = \frac{k-1}{k+2} d_k - \delta \nabla g(y_k),$$

$$d_{k+1}^m = \frac{k-1}{k+2} d_k^m - \delta \nabla g(y_k^m) - \delta m^{-1/2} R^m(y_k^m; \mathbf{U}^*_{m(k+1)}),$$

$$|y_k^m - y_k| \leq a_k + b_k.$$

Since $\max_{k\le k_T}|y_k - X(t_k)| = O(\delta^{1/2}|\log\delta|)$, we have $\max_{k\le k_T}|y_k| = O(1)$, and as in Lemma 30, $\max_{k\le T\delta^{-1/2}} E[|R^m(y_k;\mathbf{U}^*_{m(k+1)})|^4] = O(1)$. For simplicity, we let $R_k = |R^m(y_k;\mathbf{U}^*_{m(k+1)})|$. Recall that

$$\begin{aligned}|R^m(y_k^m;\mathbf{U}^*_{m(k+1)})| \le& R_k + m^{-1/2}\sum_{i=1}^m [h_1(U^*_{i(k+1)}) + L]|y_k^m - y_k| \\ \le& R_k + m^{-1/2}\left|\sum_{i=1}^m [h_1(U^*_{i(k+1)}) - E(h_1(U))]\right|\cdot|y_k^m - y_k| \\ &+ m^{1/2}(E(h_1(U)) + L)\cdot|y_k^m - y_k|.\end{aligned}$$

Let $h_k = m^{-1/2}\left|\sum_{i=1}^m [h_1(U^*_{i(k+1)}) - E(h_1(U))]\right|$. Then, we have

$$\max_{k\le T\delta^{-1/2}} E[h_k^4] = O(1),$$

$$b_{k+1} \le b_k + L\delta(a_k + b_k) + \delta m^{-1/2}R_k + (\delta m^{-1/2}h_k + C\delta)(a_k + b_k),$$

and using $a_k + b_k \le 2S_k$, we obtain

$$b_{k+1} \le b_k + C\delta(1 + m^{-1/2}h_k)S_k + \delta m^{-1/2}R_k.$$

Define a sequence b'_k that satisfies $b'_0 = 0$,

$$b'_{k+1} = b'_k + C\delta(1 + m^{-1/2}h_k)S'_k + \delta m^{-1/2}R_k.$$

Then, $b_k \le b'_k$, b'_k is non-decreasing, and since $k\delta^{1/2} \le T$,

$$b'_{k+1} \le b'_k + C\delta(1+m^{-1/2}h_k)kb'_k + \delta m^{-1/2}R_k \le (1+C\delta^{1/2})(1+C\delta^{1/2}m^{-1/2}h_k)b'_k + \delta m^{-1/2}R_k.$$

Define another sequence b^*_k that satisfies $b^*_0 = 0$,

$$b^*_{k+1} = (1 + C\delta^{1/2})(1 + C\delta^{1/2}m^{-1/2}h_k)b^*_k + \delta m^{-1/2}R_k.$$

Then, $b'_k \le b^*_k$, and

$$\begin{aligned}b^*_k &= \sum_{i=0}^{k-1}\left\{(1 + C\delta^{1/2})^{k-i-1}\left[\prod_{j=i+1}^{k-1}(1 + C\delta^{1/2}m^{-1/2}h_j)\right]\delta m^{-1/2}R_i\right\} \\ &\le C\delta m^{-1/2}\sum_{i=0}^{k-1}\left\{\left[\prod_{j=i+1}^{k-1}(1 + C\delta^{1/2}m^{-1/2}h_j)\right]R_i\right\}.\end{aligned}$$

Since for $r = C\delta^{1/2}m^{-1/2} < 1$,

$$E(1 + rh_j)^4 \le 1 + 4rEh_j + 6rEh_j^2 + 4rEh_j^3 + rEh_j^4 \le 1 + Cr,$$

and $R_i, h_{i+1}, \ldots, h_{k-1}$ are independent, we obtain

$$\begin{aligned}
\max_{k\le k_T} Ea_k^4 &\le E(k_T b_{k_T}^*)^4 \\
&\le C\delta^2 m^{-2} k_T^3 \sum_{i=0}^{k_T-1} E\left\{\left[\prod_{j=i+1}^{k_T-1}(1+C\delta^{1/2}m^{-1/2}h_j)\right] R_i\right\}^4 \\
&\le C\delta^{1/2} m^{-2} \sum_{i=0}^{k_T-1}\left\{\left[\prod_{j=i+1}^{k_T-1} E(1+C\delta^{1/2}m^{-1/2}h_j)^4\right] ER_i^4\right\} \\
&\le C\delta^{1/2} m^{-2} k_T (C(1+C\delta^{1/2}m^{-1/2})^{k_T}) \\
&= O(m^{-2}).
\end{aligned}$$

Finally, we conclude that

$$\max_{k\le k_T} E[|y_k^m - y_k|^4] \le \max_{k\le k_T} E[(a_k+b_k)^4] \le C \max_{k\le k_T} E[a_k^4 + b_k^4] = O(m^{-2}). \blacksquare$$

Lemma 33

$$\max_{k\le T\delta^{-1/2}} E[|R^m(y_{k-1}^m; \mathbf{U}_{mk}^*)|^4] \le C.$$

Proof. Define filtration $\mathcal{F}_t = \sigma(y_\delta^m(s), \mathbf{U}_m^*(s); s \le t)$, where $y_\delta^m(t)$ and $\mathbf{U}_m^*(t)$ are given by (4.27). For simplicity, we write $R_k^m = R^m(y_k^m; \mathbf{U}_{m(k+1)}^*)$, and $r_{qk} = \nabla\ell(y_k^m, U_{q(k+1)}^*) - \nabla g(y_k^m)$. Then, given $\mathcal{F}_{t_k}$, $r_{1k}, \ldots, r_{qk}$ are conditionally independent with conditional mean 0,

$$R_k^m = m^{-1/2}\sum_{q=1}^m r_{qk},$$

$$E|r_{qk}|^4 \le C\left(E[|h_1(U_{q(k+1)}^*)|^4|y_k^m-\theta_0|^4] + E|y_k^m-\theta_0|^4 + E|\nabla\ell(\theta_0, U_{q(k+1)}^*)|^4 + |\nabla g(\theta_0)|^4\right),$$

which is bounded uniformly over $1 \le k \le k_T$, since $E[|y_k^m-\theta_0|^4] \le E[|y_k^m-y_k|^4]+|y_k-\theta_0|^4] \le C$ (implied by Lemmas 29 and 32), and

$$E[|h_1(U_{q(k+1)}^*)|^4|y_k^m-\theta_0|^4] = E[E[|h_1(U_{q(k+1)}^*)|^4|y_k^m-\theta_0|^4|\mathcal{F}_{t_k}]] = E(|h_1(U)|^4)E[|y_k^m-\theta_0|^4] \le C.$$

Finally, we conclude that

$$\begin{aligned}
E|R_k^m|^4 &= m^{-2}E\left[E\left[\left(\sum_{q=1}^m r'_{qk}\right)\left(\sum_{q=1}^m r_{qk}\right)\left(\sum_{q=1}^m r'_{qk}\right)\left(\sum_{q=1}^m r_{qk}\right)\middle|\mathcal{F}_{t_k}\right]\right] \\
&\le m^{-2}\left(\sum_{q=1}^m E|r_{qk}|^4 + 6\sum_{p<q}(E|r_{pk}|^4\cdot E|r_{qk}|^4)^{\frac{1}{2}}\right) \\
&\le C. \blacksquare
\end{aligned}$$

Lemma 34 *There exist simultaneous realizations* $\{\tilde{x}_k^m, \tilde{y}_k^m\}$, $\{\tilde{\check{x}}_k^m, \tilde{\check{y}}_k^m\}$, *standard Brownian motion* $\tilde{B}$, *and* $\tilde{H}(t) = \int_0^t \boldsymbol{\sigma}(X(u))d\tilde{\mathbf{B}}(u)$ *on some common probability spaces, such that sequences* $\{\tilde{x}_k^m, \tilde{y}_k^m, k \le T/\delta^{1/2}\}$ *have the same distribution as* $\{x_k^m, y_k^m, k \le T/\delta^{1/2}\}$, *and sequences* $\{\tilde{\check{x}}_k^m, \tilde{\check{y}}_k^m, k \le T/\delta^{1/2}\}$ *are generated from* $\tilde{H}(t)$ *in the same manner as* $\{\check{x}_k^m, \check{y}_k^m, k \le T/\delta^{1/2}\}$ *generated from* $H(t)$ *according to (5.73), and as* $\delta \to 0, m \to \infty$,

$$\max_{k \le T/\delta^{1/2}} |\tilde{x}_k^m - \tilde{\check{x}}_k^m| = o_p(m^{-1/2}\delta^{1/4}). \tag{5.75}$$

Proof. For $k \ge 1$, let $\check{d}_k^m = \check{x}_{k+1}^m - \check{x}_k^m$,

$$\check{d}_0^m = -\delta\nabla g(x_0) - m^{-1/2}\delta^{3/4}(H(t_1) - H(t_0)),$$

and rewrite (5.73) as

$$\check{x}_{k+1}^m = \check{x}_k^m + \frac{k-1}{k+2}(\check{x}_k^m - \check{x}_{k-1}^m) - \delta\nabla g(\check{y}_k^m) - m^{-1/2}\delta^{3/4}(H(t_{k+1}) - H(t_k)).$$

Then, we obtain

$$\begin{aligned}
\check{d}_k^m &= \frac{k-1}{k+2}\check{d}_{k-1}^m - \delta\nabla g(\check{y}_k^m) - m^{-1/2}\delta^{3/4}(H(t_{k+1}) - H(t_k)) \\
&= \frac{k-1}{k+2}\left(\frac{k-2}{k+1}\check{d}_{k-2}^m - \delta\nabla g(\check{y}_{k-1}^m) - m^{-1/2}\delta^{3/4}(H(t_k) - H(t_{k-1}))\right) \\
&\quad -\delta\nabla g(\check{y}_k^m) - m^{-1/2}\delta^{3/4}(H(t_{k+1}) - H(t_k)) \\
&= -\sum_{i=1}^{k}\left(\frac{k-1}{k+2}\cdot\frac{k-2}{k+1}\cdot\cdots\cdot\frac{i}{i+3}\right)(\delta\nabla g(\check{y}_i^m) + m^{-1/2}\delta^{3/4}(H(t_{i+1}) - H(t_i))) \\
&= -\sum_{i=1}^{k}\frac{(i+2)(i+1)i}{(k+2)(k+1)k}(\delta\nabla g(\check{y}_i^m) + m^{-1/2}\delta^{3/4}(H(t_{i+1}) - H(t_i))).
\end{aligned} \tag{5.76}$$

Similarly, let $d_k^m = x_{k+1}^m - x_k^m$,

$$d_0^m = -\delta\nabla\hat{\mathcal{L}}^m(x_0; \mathbf{U}_{m1}^*) = -\delta\nabla g(x_0) - \delta(\nabla\hat{\mathcal{L}}^m(x_0; \mathbf{U}_{m1}^*) - \nabla g(x_0)),$$

and we have

$$\begin{aligned}
d_k^m &= -\sum_{i=1}^{k}\frac{(i+2)(i+1)i}{(k+2)(k+1)k}\delta\nabla\hat{\mathcal{L}}^m(y_i^m; \mathbf{U}_{m(i+1)}^*) \qquad (5.77) \\
&= -\sum_{i=1}^{k}\frac{(i+2)(i+1)i}{(k+2)(k+1)k}\left(\delta\nabla g(y_i^m) + \delta\left(\nabla\hat{\mathcal{L}}^m(y_i^m; \mathbf{U}_{m(i+1)}^*) - \nabla g(y_i^m)\right)\right).
\end{aligned}$$

Set $c_i = (i+2)(i+1)i$,

$$R^m(\theta; \mathbf{U}_{mk}^*) = m^{1/2}\left(\nabla\hat{\mathcal{L}}^m(\theta; \mathbf{U}_{mk}^*) - \nabla g(\theta)\right),$$

and define càdlàg processes $G_\delta^m(t)$ and $G_\delta(t)$ as follows:

$$G_\delta^m(t) = \begin{cases} 0, & 0 \le t < t_1, \\ \delta^{1/4} R^m(x_0; \mathbf{U}_{m1}^*), & t_1 \le t < t_2, \\ \delta^{1/4} \dfrac{1}{c_k} \displaystyle\sum_{i=1}^{k} c_i R^m(y_i^m; \mathbf{U}_{m(i+1)}^*), & t_{k+1} \le t < t_{k+2}, \end{cases} \tag{5.78}$$

$$G_\delta(t) = \begin{cases} 0, & 0 \le t < t_1, \\ H(t_1) - H(t_0), & t_1 \le t < t_2, \\ \dfrac{1}{c_k} \displaystyle\sum_{i=1}^{k} c_i (H(t_{i+1}) - H(t_i)), & t_{k+1} \le t < t_{k+2}. \end{cases} \tag{5.79}$$

By Assumption A4, $R^m(\theta; \mathbf{U}_{mk}^*)$ weakly converges to $N(0, \boldsymbol{\sigma}^2(\theta))$ uniformly over θ as $m \to \infty$. Note that $H(t_{i+1}) - H(t_i)$ follows $N(0, \int_{t_i}^{t_{i+1}} \boldsymbol{\sigma}^2(X(u))du)$, and $\mathrm{Var}(\delta^{1/4} R^m(y_i^m; \mathbf{U}_{m(i+1)}^*))$ is approximately equal to $\int_{t_i}^{t_{i+1}} \boldsymbol{\sigma}^2(X(u))du$. According to Lemma 35 below, there exist $\tilde{G}_\delta^m(t)$ and $\tilde{H}(t) = \int_0^t \boldsymbol{\sigma}(X(u))d\tilde{\mathbf{B}}(u)$ on some common probability spaces, such that $\tilde{G}_\delta^m(t)$ and $G_\delta^m(t)$ are identically distributed, $\tilde{G}_\delta(t)$ is generated by $\tilde{H}(t)$ in the same manner as $G_\delta(t)$ by $H(t)$ via scheme (5.79), and as $\delta \to 0, m \to \infty$,

$$\max_{t \le T} |\tilde{G}_\delta^m(t) - \tilde{G}_\delta(t)| = o_p(1).$$

Using $\tilde{G}_\delta^m(t)$, we define associated sequences $\{\tilde{R}_k\}$ and $\{\tilde{x}_k^m, \tilde{y}_k^m\}$ as follows:

$$\tilde{R}_0 = \delta^{-1/4} \tilde{G}_\delta^m(t_1), \tilde{x}_0^m = \tilde{y}_0^m = x_0,$$

$$\tilde{x}_k^m = \tilde{y}_{k-1}^m - \delta \nabla g(\tilde{y}_{k-1}^m) - \delta m^{-1/2} \tilde{R}_{k-1}, \ \tilde{y}_k^m = \tilde{x}_k^m + \frac{k-1}{k+2}(\tilde{x}_k^m - \tilde{x}_{k-1}^m),$$

$$\tilde{R}_k = \delta^{-1/4} \tilde{G}_\delta^m(t_{k+1}) - \frac{1}{c_k} \sum_{i=1}^{k-1} c_i \tilde{R}_i.$$

Since $\tilde{G}_\delta^m(t_1), ..., \tilde{G}_\delta^m(t_k)$ have the same distribution as $G_\delta^m(t_1), ...G_\delta^m(t_k)$, we easily conclude that $\{\tilde{x}_k^m, \tilde{y}_k^m\}$ are identically distributed as $\{x_k^m, y_k^m\}$, and $\tilde{d}_k^m = \tilde{x}_{k+1}^m - \tilde{x}_k^m$ satisfies

$$\tilde{d}_0^m = -\delta \nabla g(x_0) - m^{-1/2} \delta^{3/4} \tilde{G}_\delta^m(t_1), \ \tilde{d}_k^m = -\sum_{i=1}^{k} \frac{c_i}{c_k} \delta \nabla g(\tilde{y}_i^m) - m^{-1/2} \delta^{3/4} \tilde{G}_\delta^m(t_{k+1}).$$

Similarly, we define $\{\tilde{\tilde{x}}_k^m, \tilde{\tilde{y}}_k^m\}$ by $\tilde{G}_\delta(t)$, and set $\tilde{\tilde{d}}_k^m = \tilde{\tilde{x}}_{k+1}^m - \tilde{\tilde{x}}_k^m$, so that

$$\tilde{\tilde{d}}_0^m = -\delta \nabla g(x_0) - m^{-1/2} \delta^{3/4} \tilde{G}_\delta(t_1), \tilde{\tilde{d}}_k^m = -\sum_{i=1}^{k} \frac{c_i}{c_k} \delta \nabla g(\tilde{\tilde{y}}_i^m) - m^{-1/2} \delta^{3/4} \tilde{G}_\delta(t_{k+1}).$$

Let $a_k = |\tilde{x}_k^m - \tilde{\tilde{x}}_k^m|, b_k = |\tilde{d}_k^m - \tilde{\tilde{d}}_k^m|$, $S_k = b_0 + \cdots + b_k$, and $\mathcal{Y} = m^{-1/2}\delta^{3/4} \max_{t\le T} |\tilde{G}_\delta^m(t) - \tilde{G}_\delta(t)|$. Then, we have $b_0 \le \mathcal{Y}$,

$$a_k = |\tilde{x}_{k-1}^m + \tilde{d}_{k-1}^m - \tilde{\tilde{x}}_{k-1}^m - \tilde{\tilde{d}}_{k-1}^m| \le a_{k-1} + b_{k-1} \le b_0 + b_1 + \cdots + b_{k-1} = S_{k-1},$$

$$|\tilde{y}_k^m - \tilde{\tilde{y}}_k^m| = \left| \tilde{x}_k^m + \frac{k-1}{k+2}\tilde{d}_{k-1} - \tilde{\tilde{x}}_k^m - \frac{k-1}{k+2}\tilde{\tilde{d}}_{k-1} \right| \le a_k + b_{k-1} \le S_{k-1} + b_{k-1},$$

$$b_k \le L\delta \sum_{i=1}^{k} |\tilde{y}_i^m - \tilde{\tilde{y}}_i^m| + \mathcal{Y} \le L\delta \sum_{i=1}^{k} (S_{i-1} + b_{i-1}) + \mathcal{Y} \le 2L\delta k S_{k-1} + \mathcal{Y} \le C\delta^{1/2} S_{k-1} + \mathcal{Y}.$$

Let $b_0^* = \mathcal{Y}, b_k^* = C\delta^{1/2} S_{k-1}^* + \mathcal{Y}$, where $S_k^* = b_0^* + \ldots + b_k^*$. Then, by induction we easily conclude that

$$b_0 \le b_0^*, b_k \le C\delta^{1/2} S_{k-1} + \mathcal{Y} \le C\delta^{1/2} S_{k-1}^* + \mathcal{Y} = b_k^*, S_k \le S_k^*.$$

Since $b_{k+1}^* = C\delta^{1/2} S_k^* + \mathcal{Y}$ leads to $b_{k+1}^* - b_k^* = C\delta^{1/2} b_k^*$ for all $k \ge 0$, we immediately obtain the geometric sequence $b_k^* = (1 + C\delta^{1/2})^k \mathcal{Y}$ and find its sum S_k^*. Finally, we conclude that

$$\max_{k \le T/\delta^{1/2}} |\tilde{x}_k^m - \tilde{\tilde{x}}_k^m| \le S_{\lfloor T/\delta^{1/2} \rfloor - 1} \le S^*_{\lfloor T/\delta^{1/2} \rfloor - 1} \le T/\delta^{1/2} (1 + C\delta^{1/2})^{T/\delta^{1/2}} \mathcal{Y}$$

$$\le C\mathcal{Y}/\delta^{1/2} = o_p(m^{-1/2}\delta^{1/4}). \blacksquare$$

Lemma 35 *Given that $G_\delta^m(t)$ and $G_\delta(t)$ are defined by (5.78) and (5.79), respectively, we can show that there exist $\tilde{G}_\delta^m(t)$ and $\tilde{H}(t) = \int_0^t \boldsymbol{\sigma}(X(u)) d\tilde{\mathbf{B}}(u)$ on some common probability spaces, such that $\tilde{G}_\delta^m(t)$ and $G_\delta^m(t)$ are identically distributed, $\tilde{G}_\delta(t)$ are generated by $\tilde{H}(t)$ in the same manner as $G_\delta(t)$ by $H(t)$ via scheme (5.79), and as $\delta \to 0, m \to \infty$,*

$$\max_{t \le T} |\tilde{G}_\delta^m(t) - \tilde{G}_\delta(t)| = o_p(1).$$

Proof. Define a càdlàg process

$$\check{G}_\delta^m(t) = \begin{cases} 0, & 0 \le t < t_1, \\ \delta^{1/4} R^m(x_0; \mathbf{U}_{m1}^*), & t_1 \le t < t_2, \\ \delta^{1/4} \dfrac{1}{c_k} \displaystyle\sum_{i=1}^{k} c_i R^m(X(t_i); \mathbf{U}_{m(i+1)}^*), & t_{k+1} \le t < t_{k+2}. \end{cases} \tag{5.80}$$

Note that the only change in (5.80) is to replace y_i^m in (5.78) by $X(t_i)$. Define $G(t) = \frac{1}{t^3} \int_0^t u^3 \boldsymbol{\sigma}(X(u)) d\mathrm{B}(u)$. We prove that $\check{G}_\delta^m(t)$ weakly converges to $G(t)$. Set $\mathcal{C}_i^\delta = \delta^{3/2} c_i = t_i t_{i+1} t_{i+2}$, and for any fixed $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_l \le T$, let $k_j^\delta = \max(0, \lfloor \tau_j/\delta^{1/2} \rfloor - 1)$. Then, $\mathcal{C}_{k_j^\delta}^\delta = \delta^{3/2} k_j^\delta (k_j^\delta + 1)(k_j^\delta + 2) \to \tau_j^3$, as $\delta \to 0$. Using the definition of $\check{G}_\delta^m(t)$, we have

$$\check{G}_\delta^m(\tau_j) = \delta^{1/4} \frac{1}{\mathcal{C}_{k_j^\delta}^\delta} \sum_{i=1}^{k_j^\delta} \mathcal{C}_i^\delta R^m(X(t_i); \mathbf{U}_{m(i+1)}^*),$$

$$\mathcal{C}^{\delta}_{k^{\delta}_{j+1}}\check{G}^{m}_{\delta}(\tau_{j+1}) - \mathcal{C}^{\delta}_{k^{\delta}_{j}}\check{G}^{m}_{\delta}(\tau_j) = \delta^{1/4}\sum_{i=k^{\delta}_{j}+1}^{k^{\delta}_{j+1}} \mathcal{C}^{\delta}_{i} R^{m}(X(t_i); \mathbf{U}^{*}_{m(i+1)}).$$

The right-hand side of the above equation is the sum of independent random variables, by Assumption A4, as $m \to \infty, \delta \to 0$, $\mathcal{C}^{\delta}_{k^{\delta}_{j+1}}\check{G}^{m}_{\delta}(\tau_{j+1}) - \mathcal{C}^{\delta}_{k^{\delta}_{j}}\check{G}^{m}_{\delta}(\tau_j)$ converges in distribution to a normal distribution with mean 0 and variance $\int_{\tau_j}^{\tau_{j+1}} u^6\sigma^2(X(u))du$. Because of independence between consecutive differences, we can easily arrive at that

$$\left(\mathcal{C}^{\delta}_{k^{\delta}_{1}}\check{G}^{m}_{\delta}(\tau_1), \mathcal{C}^{\delta}_{k^{\delta}_{2}}\check{G}^{m}_{\delta}(\tau_2) - \mathcal{C}^{\delta}_{k^{\delta}_{1}}\check{G}^{m}_{\delta}(\tau_1), \ldots, \mathcal{C}^{\delta}_{k^{\delta}_{l}}\check{G}^{m}_{\delta}(\tau_l) - \mathcal{C}^{\delta}_{k^{\delta}_{l-1}}\check{G}^{m}_{\delta}(\tau_{l-1})\right)$$

converges in distribution to

$$\left(\tau_1^3 G(\tau_1), \tau_2^3 G(\tau_2) - \tau_1^3 G(\tau_1), \ldots, \tau_l^3 G(\tau_l) - \tau_{l-1}^3(\tau_{l-1})\right),$$

which immediately shows that $\left(\mathcal{C}^{\delta}_{k^{\delta}_{1}}\check{G}^{m}_{\delta}(\tau_1), \mathcal{C}^{\delta}_{k^{\delta}_{2}}\check{G}^{m}_{\delta}(\tau_2), \ldots, \mathcal{C}^{\delta}_{k^{\delta}_{l}}\check{G}^{m}_{\delta}(\tau_l)\right)$ converges in distribution to $\left(\tau_1^3 G(\tau_1), \tau_2^3 G(\tau_2), \ldots, \tau_l^3 G(\tau_l)\right)$. Since $\mathcal{C}^{\delta}_{k^{\delta}_{j}} \to \tau_j^3$ as $\delta \to 0$, we conclude that $\left(\check{G}^{m}_{\delta}(\tau_1), \check{G}^{m}_{\delta}(\tau_2), \ldots, \check{G}^{m}_{\delta}(\tau_l)\right)$ converges in distribution to $(G(\tau_1), G(\tau_2), \ldots, G(\tau_l))$. Thus, we prove the finite-dimensional distribution convergence of $\check{G}^{m}_{\delta}(t)$.

We establish the tightness of $\check{G}^{m}_{\delta}(t)$ by proving that for any $0 \le r \le s \le t \le T$,

$$E\{|\check{G}^{m}_{\delta}(t) - \check{G}^{m}_{\delta}(s)|^2 |\check{G}^{m}_{\delta}(s) - \check{G}^{m}_{\delta}(r)|^2\} \le C(t-r)^2. \tag{5.81}$$

To simplify some notation, we let $R_i = R^m(X(t_i); \mathbf{U}^{*}_{m(i+1)})$. First, we show that for any fixed $1 \le j < k < l$,

$$E\left\{\left|\frac{1}{c_k}\sum_{i=1}^{k} c_i R_i - \frac{1}{c_j}\sum_{i=1}^{j} c_i R_i\right|^2 \left|\frac{1}{c_l}\sum_{i=1}^{l} c_i R_i - \frac{1}{c_k}\sum_{i=1}^{k} c_i R_i\right|^2\right\} \le C(l-j)^2, \tag{5.82}$$

where C is a generic constant free of the choice of (j,k,l).

Lemma 30 implies that $E(|R_i|^2)$ and $E(|R_i|^4)$ are uniformly bounded over $1 \le i \le T\delta^{-1/2}$. Since $R_0, R_1, R_2, \ldots$ are independent with mean 0, we have

$$E\left|\sum_{i=1}^{k} c_i R_i\right|^2 = \sum_{i=1}^{k} c_i^2 E|R_i|^2 \le Ckc_k^2,$$

$$E\left|\sum_{i=1}^{k} c_i R_i\right|^4 \le \sum_{i=1}^{k} c_i^4 E|R_i|^4 + \sum_{i<j} 6c_i^2 c_j^2 E|R_i|^2 E|R_j|^2 \le Ck^2 c_k^4,$$

where we recall the convention that C denotes any generic constant free of (δ, m, n) and (i,j,k,l), and its value may change from appearance to appearance.

Let

$$D_1 = \sum_{i=1}^{j} c_i R_i, D_2 = \sum_{i=j+1}^{k} c_i R_i, D_3 = \sum_{i=k+1}^{l} c_i R_i.$$

Then, D_1, D_2, and D_3 are independent, and similarly we can show that

$$E|D_2|^2 \le C(k-j)c_k^2, E|D_2|^4 \le C(k-j)^2 c_k^4,$$

$$E|D_3|^2 \le C(l-k)c_l^2, E|D_3|^4 \le C(l-k)^2 c_l^4,$$

$$\frac{(c_k-c_j)^2}{c_k^2} \le \frac{c_k-c_j}{c_k} = 1 - \frac{j(j+1)(j+2)}{k(k+1)(k+2)} \le 1 - \frac{j^3}{k^3} \le 3\left(\frac{k-j}{k}\right).$$

Therefore, we establish (5.82) as follows:

$$\begin{aligned}
& E\left\{\left|\frac{1}{c_k}\sum_{i=1}^{k} c_i R_i - \frac{1}{c_j}\sum_{i=1}^{j} c_i R_i\right|^2 \left|\frac{1}{c_l}\sum_{i=1}^{l} c_i R_i - \frac{1}{c_k}\sum_{i=1}^{k} c_i R_i\right|^2\right\} \\
= \;& E\left\{\left|\frac{D_1+D_2}{c_k} - \frac{D_1}{c_j}\right|^2 \left|\frac{D_1+D_2+D_3}{c_l} - \frac{D_1+D_2}{c_k}\right|^2\right\} \\
= \;& E\left\{\left|\frac{D_2}{c_k} - \frac{c_k-c_j}{c_k c_j}D_1\right|^2 \left|\frac{D_3}{c_l} - \frac{c_l-c_k}{c_l c_k}(D_1+D_2)\right|^2\right\} \\
\le \;& C\cdot E\left\{\left(\frac{|D_2|^2}{c_k^2} + \frac{(c_k-c_j)^2}{c_k^2 c_j^2}|D_1|^2\right)\left(\frac{|D_3|^2}{c_l^2} + \frac{(c_l-c_k)^2}{c_l^2 c_k^2}(|D_1|^2+|D_2|^2)\right)\right\} \\
\le \;& 9C\cdot E\left\{\left(\frac{|D_2|^2}{c_k^2} + \frac{k-j}{kc_j^2}|D_1|^2\right)\left(\frac{|D_3|^2}{c_l^2} + \frac{l-k}{lc_k^2}(|D_1|^2+|D_2|^2)\right)\right\} \\
\le \;& C\left(\frac{(k-j)c_k^2(l-k)c_l^2}{c_k^2 c_l^2} + \frac{(k-j)c_k^2(l-k)jc_j^2}{c_k^2 l c_k^2} + \frac{(l-k)(k-j)^2 c_k^4}{lc_k^4}\right) \\
& + C\left(\frac{(k-j)jc_j^2(l-k)c_l^2}{kc_j^2 c_l^2} + \frac{(k-j)(l-k)j^2 c_j^4}{kc_j^2 l c_k^2} + \frac{(k-j)jc_j^2(l-k)(k-j)c_k^2}{kc_j^2 l c_k^2}\right) \\
\le \;& C(l-j)^2.
\end{aligned}$$

Second, we prove

$$E\left\{\left|\frac{1}{c_k}\sum_{i=1}^{k} c_i R_i - \frac{1}{c_j}\sum_{i=1}^{j} c_i R_i\right|^2 \left|\frac{1}{c_j}\sum_{i=1}^{j} c_i R_i\right|^2\right\} \le Ck^2. \tag{5.83}$$

Indeed, similar direct calculations lead to

$$
\begin{aligned}
& E\left\{\left|\frac{1}{c_k}\sum_{i=1}^{k} c_i R_i - \frac{1}{c_j}\sum_{i=1}^{j} c_i R_i\right|^2 \left|\frac{1}{c_j}\sum_{i=1}^{j} c_i R_i\right|^2\right\} \\
= \;& E\left\{\left|\frac{D_1+D_2}{c_k} - \frac{D_1}{c_j}\right|^2 \left|\frac{D_1}{c_j}\right|^2\right\} \\
\le \;& C\cdot E\left\{\left(\frac{|D_2|^2}{c_k^2} + \frac{(c_k-c_j)^2}{c_k^2 c_j^2}|D_1|^2\right)\frac{|D_1|^2}{c_j^2}\right\} \\
\le \;& C\left(\frac{(k-j)c_k^2 j c_j^2}{c_k^2 c_j^2} + \frac{j^2 c_j^4}{c_j^4}\right) \\
\le \;& Ck^2.
\end{aligned}
$$

Third, for any $0 \le r \le s \le t \le T$, we may choose (j,k,l) such that $t_{j+1} \le r < t_{j+2}, t_{k+1} \le s < t_{k+2}, t_{l+1} \le t < t_{l+2}$. If $j = k$ or $k = l$, then $r = s$ or $s = t$, and (5.81) is obvious. Assume that $j < k < l$, and we prove (5.81) for each scenario. If $j = -1$ and $k = 0$, then

$$
\begin{aligned}
& E\left\{|\check{G}_\delta^m(t) - \check{G}_\delta^m(s)|^2 |\check{G}_\delta^m(s) - \check{G}_\delta^m(r)|^2\right\} \\
= \;& \delta E\left\{\left|\frac{1}{c_l}\sum_{i=1}^{l} c_i R_i - R_0\right|^2 |R_0|^2\right\} \\
\le \;& C\delta l^2 = C(t_{l+1} - t_1)^2 \le C(t-r)^2.
\end{aligned}
$$

If $j = -1$ and $k \ge 1$, then

$$
\begin{aligned}
& E\left\{|\check{G}_\delta^m(t) - \check{G}_\delta^m(s)|^2 |\check{G}_\delta^m(s) - \check{G}_\delta^m(r)|^2\right\} \\
= \;& \delta E\left\{\left|\frac{1}{c_l}\sum_{i=1}^{l} c_i R_i - \frac{1}{c_k}\sum_{i=1}^{k} c_i R_i\right|^2 \left|\frac{1}{c_k}\sum_{i=1}^{k} c_i R_i\right|^2\right\} \\
\le \;& C\delta l^2 = C(t_{l+1} - t_1)^2 \le C(t-r)^2.
\end{aligned}
$$

If $j = 0$, then

$$
\begin{aligned}
& E\left\{|\check{G}_\delta^m(t) - \check{G}_\delta^m(s)|^2 |\check{G}_\delta^m(s) - \check{G}_\delta^m(r)|^2\right\} \\
= \;& \delta E\left\{\left|\frac{1}{c_l}\sum_{i=1}^{l} c_i R_i - \frac{1}{c_k}\sum_{i=1}^{k} c_i R_i\right|^2 \left|\frac{1}{c_k}\sum_{i=1}^{k} c_i R_i - R_0\right|^2\right\} \\
\le \;& C\delta l^2 \le 4C(t_{l+1} - t_2)^2 \le 4C(t-r)^2.
\end{aligned}
$$

If $j \ge 1$, then

$$
\begin{aligned}
& E\left\{|\check{G}_\delta^m(t) - \check{G}_\delta^m(s)|^2 |\check{G}_\delta^m(s) - \check{G}_\delta^m(r)|^2\right\} \\
= \;& \delta E\left\{\left|\frac{1}{c_l}\sum_{i=1}^{l} c_i R_i - \frac{1}{c_k}\sum_{i=1}^{k} c_i R_i\right|^2 \left|\frac{1}{c_k}\sum_{i=1}^{k} c_i R_i - \frac{1}{c_j}\sum_{i=1}^{j} c_i R_i\right|^2\right\} \\
\le \;& C\delta (l-j)^2 \le 4C(t_{l+1} - t_{j+2})^2 \le 4C(t-r)^2.
\end{aligned}
$$

Now, with the established finite-dimensional distribution convergence and tightness for $\check{G}^m_\delta(t)$, we conclude that $\check{G}^m_\delta(t)$ weakly converges to $G(t)$.

Note that the only difference between $\check{G}^m_\delta(t)$ and $G^m_\delta(t)$ is y^m_i and $X(t_i)$ used in $R^m(\cdot; \mathbf{U}^*_{mi})$. By Lemmas 31 and 32, we immediately show the finite-dimensional distribution convergence of $G^m_\delta(t)$ to $G(t)$.

The same argument for deriving the tightness of $\check{G}^m_\delta(t)$ can be used to establish the tightness of $G^m_\delta(t)$ by proving that for any $0 \le r \le s \le t \le T$,

$$E\{|G^m_\delta(t) - G^m_\delta(s)|^2 |G^m_\delta(s) - G^m_\delta(r)|^2\} \le C(t-r)^2. \tag{5.84}$$

Again, for simplicity, we let $R_k = R^m(y^m_k; \mathbf{U}^*_{m(k+1)})$, and we show that for any fixed $1 \le j < k < l$,

$$E\left\{ \left| \frac{1}{c_k} \sum_{i=1}^k c_i R_i - \frac{1}{c_j} \sum_{i=1}^j c_i R_i \right|^2 \left| \frac{1}{c_l} \sum_{i=1}^l c_i R_i - \frac{1}{c_k} \sum_{i=1}^k c_i R_i \right|^2 \right\} \le C(l-j)^2. \tag{5.85}$$

Indeed, recall that $c_i = i(i+1)(i+2)$, and define $S_k = \sum_{i=1}^k c_i R^m_i$. Then, there exists a constant $C_1 = \gamma^2 C$, $\gamma > 1$, such that $E|S_k|^4 \le C_1 k^2 c_k^4$, which we prove by induction. Lemma 33 implies that it holds for $k = 1$. Assume that it holds for $k-1$; then, using Lemma 33, we obtain

$$\begin{aligned}
E|S_k|^4 =& E\left[\left(\sum_{i=1}^k c_i (R^m_i)' \right) \left(\sum_{i=1}^k c_i R^m_i \right) \left(\sum_{i=1}^k c_i (R^m_i)' \right) \left(\sum_{i=1}^k c_i R^m_i \right) \right] \\
\le& E|S_{k-1}|^4 + c_k^4 E|R^m_k|^4 + 4 \sum_{i=1}^{k-1} c_i c_k^3 E(|R^m_i| \cdot |R^m_k|^3) + 6 c_k^2 E(|S_{k-1}|^2 |R^m_k|^2) \\
&+ 4 \sum_{i,j,l<k} c_i c_j c_l c_k E((R^m_i)' R^m_j (R^m_l)' R^m_k) \\
\le& C_1 (k-1)^2 c_k^4 + c_k^4 C + 4(k-1) c_k^4 C + 6 c_k^2 \sqrt{C_1}(k-1) c_k^2 \sqrt{C} \\
\le& c_k^4 (C_1 (k-1)^2 + (4 + 6\gamma) C k) \\
\le& c_k^4 (C_1 (k-1)^2 + \gamma^2 C k) \\
\le& C_1 k^2 c_k^4,
\end{aligned}$$

where we take $\gamma = 7$ so that $4 + 6\gamma < \gamma^2$, and we employ the Cauchy-Schwarz inequality multiple times. Moreover, in the above derivation, we use the fact that

$$E(|R^m_i| \cdot |R^m_k|^3) \le (E|R^m_i|^4)^{\frac{1}{4}} \cdot (E|R^m_k|^4)^{\frac{3}{4}} \le C,$$

$$E(|S_{k-1}|^2 |R^m_k|^2) \le \sqrt{E|S_{k-1}|^4 \cdot E|R^m_k|^4} \le \sqrt{C_1}(k-1) c_k^2 \sqrt{C}$$

and the zero conditional mean for $i, j, l < k$,

$$E((R^m_i)' R^m_j (R^m_l)' R^m_k) = E[E[(R^m_i)' R^m_j (R^m_l)' R^m_k | \mathcal{F}_{t_k}]] = E[(R^m_i)' R^m_j (R^m_l)' E[R^m_k | \mathcal{F}_{t_k}]] = 0.$$

Note that all we have used in proving $E|S_k|^4 \le C_1 k^2 c_k^4$ are the above zero conditional mean and $E|R_k^m|^4 \le C$ implied by Lemma 33. Applying the argument to $S_k - S_j$, we obtain $E|S_k - S_j|^4 \le C_1(k-j)^2 c_k^4$. Since

$$\frac{(c_k - c_j)^2}{c_k^2} \le \frac{c_k - c_j}{c_k} = 1 - \frac{j(j+1)(j+2)}{k(k+1)(k+2)} \le 1 - \frac{j^3}{k^3} \le 3\left(\frac{k-j}{k}\right),$$

direct calculations show

$$\begin{aligned} E\left|\frac{S_k}{c_k} - \frac{S_j}{c_j}\right|^4 &= E\left|\frac{S_k - S_j}{c_k} - \frac{c_k - c_j}{c_j c_k} S_j\right|^4 \\ &\le 8\left(\frac{E|S_k - S_j|^4}{c_k^4} + \frac{E|S_j|^4 (c_k - c_j)^4}{c_j^4 c_k^4}\right) \\ &\le 8\left(C_1(k-j)^2 + \frac{9C_1 j^2 (k-j)^2}{k^2}\right) \\ &\le C_2(k-j)^2. \end{aligned}$$

Hence, for $j < k < l$, we conclude that

$$E\left[\left|\frac{S_l}{c_l} - \frac{S_k}{c_k}\right|^2 \left|\frac{S_k}{c_k} - \frac{S_j}{c_j}\right|^2\right] \le \left(E\left|\frac{S_l}{c_l} - \frac{S_k}{c_k}\right|^4\right)^{\frac{1}{2}} \left(E\left|\frac{S_k}{c_k} - \frac{S_j}{c_j}\right|^4\right)^{\frac{1}{2}} \le C_2(l-j)^2,$$

which proves (5.85). The remaining arguemnts for establishing (5.84) are easy and pretty much the same as those for establishing (5.81).

The finite-dimensional distribution convergence and (5.84) show that $G_\delta^m(t)$ has the same weak convergence limit, $G(t)$, as $\check{G}_\delta^m(t)$. Skorokhod's representation theorem indicates that there exist $\tilde{G}_\delta^m(t)$ and $\tilde{G}(t)$ on some common probability spaces, such that $\tilde{G}_\delta^m(t)$ and $G_\delta^m(t)$ are identically distributed, $\tilde{G}(t)$ and $G(t)$ are identically distributed, and as $\delta \to 0$ and $m \to \infty$, under the metric d in $D[0,T]$,

$$d(\tilde{G}_\delta^m(t), \tilde{G}(t)) = o_p(1).$$

By Lemma 37 below, we obtain that if we further prove the tightness of $\tilde{G}(t)$, then the above-mentioned $o_p(1)$ result under the metric in $D[0,T]$ can be strengthened to the maximum norm—that is,

$$\max_{t \le T} |\tilde{G}_\delta^m(t) - \tilde{G}(t)| = o_p(1). \tag{5.86}$$

We must establish the tightness of $\tilde{G}(t)$. Because $G(t)$ and $\tilde{G}(t)$ are identically distributed, if we show that for any $0 \le r \le s \le t \le T$,

$$E\left\{|G(t) - G(s)|^2 |G(s) - G(r)|^2\right\} \le C(t-r)^2. \tag{5.87}$$

Then, $\tilde{G}$ also satisfies the above inequality and both $G(t)$ and $\tilde{G}(t)$ are tight.

We prove (5.87). If $r > 0$, let

$$D_1 = \int_0^r u^3 \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u), D_2 = \int_r^s u^3 \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u), D_3 = \int_s^t u^3 \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u).$$

By Assumption A3, we have

$$\|\boldsymbol{\sigma}^2(X(t))\| \leq \|\boldsymbol{\sigma}^2(\theta_0)\| + L|X(t) - \theta_0| \leq C.$$

D_2 follows a normal distribution with mean 0 and variance $\Sigma = \int_r^s u^6 \boldsymbol{\sigma}^2(X(u))du$, and

$$\|\Sigma\| \leq \int_r^s u^6 \|\boldsymbol{\sigma}^2(X(u))\|du \leq C(s-r)s^6.$$

Taking eigen-matrix decomposition $\Sigma = \Gamma'\Lambda\Gamma$, we obtain that ΓD_2 follows a normal distribution with mean 0 and variance matrix Λ, and

$$E|D_2|^2 = E|\Gamma D_2|^2 = tr(\Lambda) \leq C(s-r)s^6,$$
$$E|D_2|^4 \leq CE|\Gamma D_2|^4 \leq Ctr(\Lambda^2) \leq C(s-r)^2 s^{12}.$$

Similarly, we have

$$E|D_1|^2 \leq Cr^7, E|D_1|^4 \leq Cr^{14},$$
$$E|D_3|^2 \leq C(t-s)t^6, E|D_3|^4 \leq C(t-s)^2 t^{12}.$$

Putting them together, we arrive at

$$\begin{aligned}
& E\left\{|G(s) - G(r)|^2 |G(t) - G(s)|^2\right\} \\
= \; & E\left\{\left|\frac{D_1 + D_2}{s^3} - \frac{D_1}{r^3}\right|^2 \left|\frac{D_1 + D_2 + D_3}{t^3} - \frac{D_1 + D_2}{s^3}\right|^2\right\} \\
\leq \; & C \cdot E\left\{\left(\frac{|D_2|^2}{s^6} + \frac{(s^3 - r^3)^2}{s^6 r^6}|D_1|^2\right)\left(\frac{|D_3|^2}{t^6} + \frac{(t^3 - s^3)^2}{t^6 s^6}(|D_1|^2 + |D_2|^2)\right)\right\} \\
\leq \; & 9C \cdot E\left\{\left(\frac{|D_2|^2}{s^6} + \frac{s-r}{sr^6}|D_1|^2\right)\left(\frac{|D_3|^2}{t^6} + \frac{t-s}{ts^6}(|D_1|^2 + |D_2|^2)\right)\right\} \\
\leq \; & C\left(\frac{(s-r)s^6(t-s)t^6}{s^6 t^6} + \frac{(s-r)s^6(t-s)r^7}{ts^{12}} + \frac{(t-s)(s-r)^2 s^{12}}{ts^{12}}\right) \\
& + C\left(\frac{(s-r)r^7(t-s)t^6}{sr^6 t^6} + \frac{(s-r)(t-s)r^{14}}{sr^6 ts^6} + \frac{(s-r)r^7(t-s)(s-r)s^6}{sr^6 ts^6}\right) \\
\leq \; & C(t-r)^2.
\end{aligned}$$

In addition, similar arguments show that for $0 < r < s$,

$$\begin{aligned}
& E\left\{|G(s) - G(r)|^2 |G(r)|^2\right\} \\
= \; & E\left\{\left|\frac{D_1 + D_2}{s^3} - \frac{D_1}{r^3}\right|^2 \left|\frac{D_1}{r^3}\right|\right\} \\
\leq \; & C \cdot E\left\{\left(\frac{|D_2|^2}{s^6} + \frac{s-r}{sr^6}|D_1|^2\right)\frac{|D_1|^2}{r^6}\right\} \\
\leq \; & C\left(\frac{(s-r)s^6 r^7}{s^6 r^6} + \frac{(s-r)r^{14}}{sr^{12}}\right) \\
\leq \; & Cs^2.
\end{aligned}$$

With $\tilde{H}(t) = \int_0^t \boldsymbol{\sigma}(X(u))d\tilde{\mathbf{B}}(u)$, $\tilde{G}_\delta(t)$ generated by $\tilde{H}(t)$ via scheme (5.79), $\tilde{G}(t) = \frac{1}{t^3}\int_0^t u^3 d\tilde{H}(u)$. Lemma 36 below indicates that as $\delta \to 0$,

$$\max_{t\le T} |\tilde{G}_\delta(t) - \tilde{G}(t)| = o_p(1).$$

Finally, combining the above result with (5.86), we conclude

$$\max_{t\le T} |\tilde{G}^m_\delta(t) - \tilde{G}_\delta(t)| \le \max_{t\le T} |\tilde{G}^m_\delta(t) - \tilde{G}(t)| + \max_{t\le T} |\tilde{G}_\delta(t) - \tilde{G}(t)| = o_p(1). \blacksquare$$

Lemma 36 *Given Brownian motion $B(t)$, we define $G(t) = t^{-3}\int_0^t u^3\boldsymbol{\sigma}(X(u))dB(u)$ and $G_\delta(t)$ by (5.79), as in the proof of Lemma 35. Then, we have*

$$\max_{t|leqT} |G_\delta(t) - G(t)| = o_p(1).$$

Proof. Denote by $\boldsymbol{\Sigma}_k$ the variance of $G_\delta(t_k) - G(t_k)$. Then

$$G_\delta(t_1) - G(t_1) = \frac{1}{t_1^3}\int_0^{t_1} (t_1^3 - u^3)\boldsymbol{\sigma}(X(u))d\mathbf{B}(u),$$

$$\boldsymbol{\Sigma}_1 = \frac{1}{t_1^6}\int_0^{t_1} (t_1^3 - u^3)^2\boldsymbol{\sigma}^2(X(u))du,$$

$$\|\boldsymbol{\Sigma}_1\| \le \frac{Ct_1^7}{t_1^6} \le C\delta^{1/2}.$$

Let $C_i^\delta = t_i t_{i+1} t_{i+2}$. We have for $k \ge 1$,

$$G_\delta(t_{k+1}) - G(t_{k+1}) = \frac{1}{C_k^\delta t_{k+1}^3}\sum_{i=0}^k \int_{t_i}^{t_{i+1}} (C_i^\delta t_{k+1}^3 - C_k^\delta u^3)\boldsymbol{\sigma}(X(u))d\mathbf{B}(u),$$

$$\boldsymbol{\Sigma}_{k+1} = \frac{1}{t_k^2 t_{k+1}^2 t_{k+2}^2 t_{k+1}^6}\sum_{i=0}^k \int_{t_i}^{t_{i+1}} (t_i t_{i+1} t_{i+2} t_{k+1}^3 - t_k t_{k+1} t_{k+2} u^3)^2\boldsymbol{\sigma}^2(X(u))du.$$

Since $|t_i t_{i+1} t_{i+2} t_{k+1}^3 - t_k t_{k+1} t_{k+2} u^3| \le C t_{k+1}^5 \delta^{1/2}$, $t_k t_{k+2} \ge t_{k+1}^2/2$, we obtain

$$\|\boldsymbol{\Sigma}_{k+1}\| \le \frac{C}{t_{k+1}^{12}}\sum_{i=0}^k \int_{t_i}^{t_{i+1}} t_{k+1}^{10}\delta du \le \frac{C\delta}{t_{k+1}} \le C\delta^{1/2}.$$

In other words, $\|\boldsymbol{\Sigma}_k\| \le C\delta^{1/2}$ uniformly over $k \le T\delta^{-1/2}$. As both $G_\delta(t)$ and $G(t)$ are normally distributed, we get

$$E|G_\delta(t_k) - G(t_k)|^4 \le C\delta,$$

and, hence, for any $\eta > 0$, we have

$$P(\max_{k\le k_T} |G_\delta(t_k) - G(t_k)| > \eta) \le \sum_{k=0}^{k_T} P(|G_\delta(t_k) - G(t_k)| > \eta)$$
$$\le \sum_{k=0}^{k_T} \frac{E|G_\delta(t_k) - G(t_k)|^4}{\eta^4} \le \sum_{k=0}^{k_T} \frac{C\delta}{\eta^4} \le \frac{CT\delta^{1/2}}{\eta^4} \to 0.$$

Finally, the tightness of $G(t)$ implies that

$$\max_{s,t\le T, |t-s|\le\delta^{1/2}} |G(t) - G(s)| = o_p(1),$$

and, thus, we conclude that

$$\max_{t\le T} |G_\delta(t) - G(t)| \le \max_{k\le k_T} |G_\delta(t_k) - G(t_k)| + \max_{k\le k_T, t_k\le t<t_{k+1}} |G(t) - G(t_k)| = o_p(1). \blacksquare$$

The following lemma is a known result, but we state it explicitly in our context.

Lemma 37 *Let $D[0,T]$ be the space of all càdlàg functions on $[0,T]$, equipped with metric $d(X(t), Y(t))$ given by*

$$d(X(t), Y(t)) = \inf\left\{\delta : \exists \textit{one to one map } \Gamma \textit{ on } [0,T] \textit{such that } \sup_{t\le T} |\Gamma(t) - t| \le \delta, \right.$$
$$\left. \sup_{t\le T} |X(\Gamma(t)) - Y(t)| \le \delta\right\}.$$

For processes $X_n(t)$ and $X(t)$ in $D[0,T]$, assume that $X(t)$ is tight, and as $n \to \infty$, $d(X_n(t), X(t)) = o_p(1)$ under the metric in $D[0,T]$. Then, we have

$$\sup_{t\le T} |X_n(t) - X(t)| = o_p(1).$$

Proof. For any $\varepsilon > 0, \eta > 0$, by the tightness of $X(t)$, there exists $\delta < \eta/2$ such that

$$P\left(\sup_{s,t\le T, |t-s|\le\delta} |X(t) - X(s)| > \eta/2\right) < \varepsilon/2.$$

Let

$$A_n = \left\{\sup_{s,t\le T, |t-s|\le\delta} |X(t) - X(s)| \le \eta/2\right\} \cap \left\{d(X_n(t), X(t)) < \delta\right\},$$

$$B_n = \left\{\sup_{t\le T} |X_n(t) - X(t)| \le \eta\right\}.$$

Then, $A_n \subseteq B_n$. Indeed, if $d(X_n(t), X(t)) < \delta$, then there exists a one-to-one map Γ on $[0,T]$, such that $\sup_{t\le T} |\Gamma(t) - t| \le \delta$ and $\sup_{t\le T} |X_n(t) - X(\Gamma(t))| \le \delta$. If we also have $\sup_{|t-s|\le\delta} |X(t) - X(s)| \le \eta/2$, then

$$\sup_{t\le T} |X(\Gamma(t)) - X(t)| \le \eta/2,$$

$$\sup_{t\le T}|X_n(t)-X(t)|\le \sup_{t\le T}|X_n(t)-X(\Gamma(t))|+\sup_{t\le T}|X(\Gamma(t))-X(t)|\le \delta+\eta/2\le\eta.$$

Hence, we have $P(B_n^C)\le P(A_n^C)$, and

$$P\left(\sup_{t\le T}|X_n(t)-X(t)|>\eta\right)\le P\left(\sup_{s,t\le T,|t-s|\le\delta}|X(t)-X(s)|>\eta/2\right)+P(d(X_n(t),X(t))\ge\delta).$$

Since $d(X_n(t),X(t))=o_p(1)$, $\exists N$, for $n>N$, $P(d(X_n(t),X(t))\ge\delta)<\varepsilon/2$, then

$$P\left(\sup_{t\le T}|X_n(t)-X(t)|>\eta\right)<\varepsilon/2+\varepsilon/2=\varepsilon.$$

This completes the proof. ■

Lemma 38

$$\max_{k\le T\delta^{-1/2}}|\check{x}_k^m-X_\delta^m(t_k)|=O_p(\delta^{1/2}|\log\delta|),$$

where $\check{x}_k^m$ and X_δ^m are defined by (5.73) and (4.35), respectively.

Proof. The same proof argument of Lemma 23 can be easily used to show

$$\mathbf{\Psi}_a=\sup_{0\le s<v\le T}\left|\frac{1}{(v-s)^a}\int_s^v\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)\right|\text{ is a.s. finite.}\tag{5.88}$$

By Lemma 27, we have

$$\max_{t\in[0,T]}|\dot{X}_\delta^m(t)|\le\max_{t\in[0,T]}|\dot{X}_\delta^m(t)-\dot{X}(t)|+\max_{t\in[0,T]}|\dot{X}(t)|=O_p(1),$$

$$\max_{t\in[0,T]}|X_\delta^m(t)|\le\max_{t\in[0,T]}|X_\delta^m(t)-X(t)|+\max_{t\in[0,T]}|X(t)|=O_p(1),$$

$$\max_{t\in[0,T]}|\nabla g(X_\delta^m(t))|\le|\nabla g(\theta_0)|+L\cdot\max_{t\in[0,T]}|X_\delta^m(t)-\theta_0|=O_p(1).$$

Let

$$\Upsilon_\delta^m=\max\left\{\mathbf{\Psi}_a,\max_{t\in[0,T]}|\dot{X}_\delta^m(t)|,\max_{t\in[0,T]}|X_\delta^m(t)|,\max_{t\in[0,T]}|\nabla g(X_\delta^m(t))|\right\}.$$

Then, $\Upsilon_\delta^m=O_p(1)$. For simplicity, we continue to use notation Υ_δ^m to denote it after multiplying and adding some generic constant C or adding random variable $\mathbf{\Psi}_a$ in (5.88), as long as it is $O_p(1)$.

For a fixed $a<1/2$, set $\xi=\sqrt{(a+1)(a+2)/2L}$. By Lemma 26, we have for $t<\xi$,

$$M_a(0,t;X_\delta^m)\le 2(t^{1-a}|\nabla g(x_0)|+\delta^{1/4}m^{-1/2}\Upsilon_\delta^m),$$

$$|\dot{X}_\delta^m(t)|\le C(t+t^a\delta^{1/4}m^{-1/2}\Upsilon_\delta^m),$$

and for $s>0$ and $t-s<\xi$,

$$M_a(s,t;X_\delta^m)\le$$
$$2\left\{(t-s)^{1-a}\left(\frac{3}{s}+\frac{L(t-s)}{2}\right)|\dot{X}_\delta^m(s)|+(t-s)^{1-a}|\nabla g(X_\delta^m(s))|+\delta^{1/4}m^{-1/2}\Upsilon_\delta^m\right\}.$$

If further $t-s \le s$, then for $s < \xi$,

$$\begin{aligned}
& (t-s)^{1-a}\left(\frac{3}{s}+\frac{L(t-s)}{2}\right)|\dot{X}_\delta^m(s)| \\
\le\ & C(t-s)^{1-a}\left(\frac{3}{s}+\frac{L(t-s)}{2}\right)(s+s^a\delta^{1/4}m^{-1/2}\Upsilon_\delta^m) \\
\le\ & C(t-s)^{1-a}\left(3+\frac{L(t-s)s}{2}\right)+C(t-s)^{1-a}\left(3s^{a-1}+\frac{L(t-s)s^a}{2}\right)\delta^{1/4}m^{-1/2}\Upsilon_\delta^m \\
\le\ & C(t-s)^{1-a}+C\left[3\left(\frac{t-s}{s}\right)^{1-a}+\frac{L(t-s)^{2-a}s^a}{2}\right]\delta^{1/4}m^{-1/2}\Upsilon_\delta^m \\
\le\ & C(t-s)^{1-a}+C\delta^{1/4}m^{-1/2}\Upsilon_\delta^m,
\end{aligned}$$

and for $s \ge \xi$,

$$(t-s)^{1-a}\left(\frac{3}{s}+\frac{L(t-s)}{2}\right)|\dot{X}_\delta^m(s)| \le C(t-s)^{1-a}\Upsilon_\delta^m.$$

Putting them together, we conclude that

$$\begin{aligned}
& |\dot{X}_\delta^m(t)-\dot{X}_\delta^m(s)| \le C(t-s)(\Upsilon_\delta^m+1)+C(t-s)^a\delta^{1/4}m^{-1/2}\Upsilon_\delta^m \\
& \le C(\Upsilon_\delta^m+1)\left[(t-s)+(t-s)^a\delta^{1/4}m^{-1/2}\right]=\Upsilon_\delta^m\left[(t-s)+(t-s)^a\delta^{1/4}m^{-1/2}\right],
\end{aligned}$$

where we use the notation convention noted early to write Υ_δ^m for $C(\Upsilon_\delta^m+1)$.

The theorem assumption implies that $\delta^{a/2-1/4}m^{-1/2} < C_0$ for some generic constant C_0. For $\delta^{1/2} < \xi$, if $t-s \le \delta^{1/2}$ and $t-s \le s$, we obtain

$$|\dot{X}_\delta^m(t)-\dot{X}_\delta^m(s)| \le \left[\delta^{1/2}+\delta^{a/2}\delta^{1/4}m^{-1/2}\right]\Upsilon_\delta^m \le \delta^{1/2}(1+C_0)\Upsilon_\delta^m \le \delta^{1/2}\Upsilon_\delta^m,$$

and if $t \le \delta^{1/2}$,

$$|\dot{X}_\delta^m(t)| \le \left[\delta^{1/2}+\delta^{a/2}\delta^{1/4}m^{-1/2}\right]\Upsilon_\delta^m \le \delta^{1/2}(1+C_0)\Upsilon_\delta^m \le \delta^{1/2}\Upsilon_\delta^m.$$

Recall that $t_k = k\delta^{1/2}$ for any $k \ge 1$, and $t_{k+1}-t_k = \delta^{1/2} \le t_k$. Then, for any $t \in [t_k, t_{k+1}]$, we have

$$|\dot{X}_\delta^m(t)-\dot{X}_\delta^m(t_k)| \le \delta^{1/2}\Upsilon_\delta^m,\ \ |\dot{X}_\delta^m(t_{k+1})-\dot{X}_\delta^m(t)| \le \delta^{1/2}\Upsilon_\delta^m,$$

and, thus, we obtain

$$|\dot{X}_\delta^m(t_k)| \le |\dot{X}_\delta^m(t_1)|+|\dot{X}_\delta^m(t_2)-\dot{X}_\delta^m(t_1)|+\cdots+|\dot{X}_\delta^m(t_k)-\dot{X}_\delta^m(t_{k-1})| \le k\delta^{1/2}\Upsilon_\delta^m = t_k\Upsilon_\delta^m,$$

$$|X_\delta^m(t_{k+1})-X_\delta^m(t_k)| \le \int_{t_k}^{t_{k+1}}|\dot{X}_\delta^m(t)|dt \le \delta^{1/2}\Upsilon_\delta^m.$$

Define $\check{z}_0^m = 0$, $\check{z}_k^m = (\check{x}_k^m - \check{x}_{k-1}^m)/\delta^{1/2}$. Using the definition of $\check{x}_k^m$ and $\check{y}_k^m$ in (5.73), we obtain

$$\check{z}_1^m = (\check{x}_1^m - x_0)/\delta^{1/2} = -\delta^{1/2}\nabla g(x_0) - m^{-1/2}\delta^{1/4}(H(t_1)-H(t_0)).$$

Then, using (5.88), we have

$$|\check{z}_1^m| \leq \delta^{1/2}|\nabla g(x_0)| + m^{-1/2}\delta^{1/4}\delta^{a/2}\mathbf{\Psi}_a \leq \delta^{1/2}(|\nabla g(x_0)| + C_0\mathbf{\Psi}_a).$$

Again, we use the notation convention to write Υ_δ^m for $\Upsilon_\delta^m + |\nabla g(x_0)| + C_0\mathbf{\Psi}_a$ (which is still $O_p(1)$). Using the above result, the notation convention and the definition of $\check{x}_k^m$ and $\check{y}_k^m$ in (5.73), we obtain

$$|\check{x}_1^m - x_0| \leq \delta^{1/2}\delta^{1/2}\Upsilon_\delta^m = \delta\Upsilon_\delta^m, \;\; |X_\delta^m(t_1) - x_0| \leq \int_0^{t_1} |\dot{X}_\delta^m(t)|dt \leq \delta\Upsilon_\delta^m.$$

Let $a_k = |\check{x}_k^m - X_\delta^m(t_k)|$. Then, $a_0 = 0$, $a_1 \leq 2\delta\Upsilon_\delta^m$. For $k \geq 2$,

$$\begin{aligned} X_\delta^m(t_k) &= X_\delta^m(t_{k-1}) + \int_{t_{k-1}}^{t_k} \dot{X}_\delta^m(t)dt \\ &= X_\delta^m(t_{k-1}) + \delta^{1/2}\dot{X}_\delta^m(t_k) + \int_{t_{k-1}}^{t_k} (\dot{X}_\delta^m(t) - \dot{X}_\delta^m(t_k))dt. \end{aligned}$$

Set $Z_\delta^m(t) = \dot{X}_\delta^m(t)$, $b_k = |\check{z}_k^m - Z_\delta^m(t_k)|$. Then, $b_0 = 0$, $b_1 \leq 2\delta^{1/2}\Upsilon_\delta^m$. Combining above equality with the definition of $\check{z}_k^m$ (i.e. $\check{x}_k^m = \check{x}_{k-1}^m + \delta^{1/2}\check{z}_k^m$), we conclude that

$$\begin{aligned} a_k &= |\check{x}_k^m - X_\delta^m(t_k)| \\ &\leq |\check{x}_{k-1}^m - X_\delta^m(t_{k-1})| + \delta^{1/2}|\check{z}_k^m - Z_\delta^m(t_k)| + \int_{t_{k-1}}^{t_k} |\dot{X}_\delta^m(t_k) - \dot{X}_\delta^m(t))|dt \\ &\leq a_{k-1} + \delta^{1/2}b_k + \delta\Upsilon_\delta^m \\ &\leq a_1 + \delta^{1/2}(b_2 + \cdots + b_k) + (k-1)\delta\Upsilon_\delta^m \\ &\leq \delta^{1/2}S_k + k\delta\Upsilon_\delta^m, \end{aligned} \tag{5.89}$$

where $S_k = b_1 + \cdots + b_k$. Note that $Z_\delta^m(t) = \dot{X}_\delta^m(t)$ obeys

$$dZ_\delta^m(t) = -\frac{3}{t}Z_\delta^m(t)dt - \nabla g(X_\delta^m(t))dt - m^{-1/2}\delta^{1/4}dH(t),$$

and, thus, we arrive at

$$\begin{aligned} &Z_\delta^m(t_{k+1}) = Z_\delta^m(t_k) - \int_{t_k}^{t_{k+1}} \frac{3}{t}Z_\delta^m(t)dt - \int_{t_k}^{t_{k+1}} \nabla g(X_\delta^m(t))dt \\ &-m^{-1/2}\delta^{1/4}(H(t_{k+1}) - H(t_k)) \\ &= Z_\delta^m(t_k) - \frac{3\delta^{1/2}}{t_k}Z_\delta^m(t_k) - \int_{t_k}^{t_{k+1}} \left[\frac{3}{t}Z_\delta^m(t)dt - \frac{3}{t_k}Z_\delta^m(t_k)\right]dt - \delta^{1/2}\nabla g(X_\delta^m(t_k)) \\ &- \int_{t_k}^{t_{k+1}} [\nabla g(X_\delta^m(t)) - \nabla g(X_\delta^m(t_k))]dt - m^{-1/2}\delta^{1/4}(H(t_{k+1}) - H(t_k)). \end{aligned} \tag{5.90}$$

For $k \geq 1$, we have

$$\begin{aligned}
& \left| \int_{t_k}^{t_{k+1}} \left[\frac{3}{t} Z_\delta^m(t) - \frac{3}{t_k} Z_\delta^m(t_k) \right] dt \right| \\
\leq & \int_{t_k}^{t_{k+1}} \left| \frac{3}{t} [Z_\delta^m(t) - Z_\delta^m(t_k)] \right| du + \int_{t_k}^{t_{k+1}} \left| \left(\frac{3}{t} - \frac{3}{t_k} \right) Z_\delta^m(t_k) \right| dt \\
\leq & \frac{3\delta \Upsilon_\delta^m}{t_k} + \frac{3(t_{k+1} - t_k)^2}{t_k t_{k+1}} t_k \Upsilon_\delta^m \leq 6\delta^{1/2} k^{-1} \Upsilon_\delta^m,
\end{aligned}$$

and

$$\left| \int_{t_k}^{t_{k+1}} [\nabla g(X_\delta^m(t)) - \nabla g(X_\delta^m(t_k))] du \right| \leq L \int_{t_k}^{t_{k+1}} |X_\delta^m(t) - X_\delta^m(t_k)| du \leq L\delta \Upsilon_\delta^m.$$

Recall (5.76) and note that $\check{z}_k^m = \check{d}_{k-1}^m / \delta^{1/2}$; then, we have

$$\check{z}_{k+1}^m = \frac{k-1}{k+2} \check{z}_k^m - \delta^{1/2} \nabla g(\check{y}_k^m) - m^{-1/2} \delta^{1/4} (H(t_{k+1}) - H(t_k)).$$

Using the above equality and (5.90), we arrive at

$$\begin{aligned}
b_{k+1} = & |\check{z}_{k+1}^m - Z_\delta^m(t_{k+1})| \leq \left(1 - \frac{3}{k+2} \right) |\check{z}_k^m - Z_\delta^m(t_k)| + \frac{6}{k(k+2)} |Z_\delta^m(t_k)| \\
& + \left| \int_{t_k}^{t_{k+1}} \left[\frac{3}{t} Z_\delta^m(t) dt - \frac{3}{t_k} Z_\delta^m(t_k) \right] dt \right| + \delta^{1/2} |\nabla g(X_\delta^m(t_k)) - \nabla g(\check{y}_k^m)| \\
& \left| \int_{t_k}^{t_{k+1}} [\nabla g(X_\delta^m(t)) - \nabla g(X_\delta^m(t_k))] dt \right| \\
\leq & b_k + 12\delta^{1/2} k^{-1} \Upsilon_\delta^m + L\delta^{1/2} \left| X_\delta^m(t_k) - \check{x}_k^m - \frac{k-1}{k+2} \delta^{1/2} \check{z}_k^m \right| + L\delta \Upsilon_\delta^m \\
\leq & b_k + 12\delta^{1/2} k^{-1} \Upsilon_\delta^m + L\delta^{1/2} \left(a_k + \delta^{1/2} (|Z_\delta^m(t_k)| + |\check{z}_k^m - Z_\delta^m(t_k)|) \right) + L\delta \Upsilon_\delta^m \\
\leq & b_k + 12\delta^{1/2} k^{-1} \Upsilon_\delta^m + L\delta^{1/2} (\delta^{1/2} S_k + k\delta \Upsilon_\delta^m + \delta^{1/2} (\Upsilon_\delta^m + b_k)) + L\delta \Upsilon_\delta^m \\
\leq & b_k + C\delta S_k + C\delta^{1/2} k^{-1} \Upsilon_\delta^m,
\end{aligned}$$

where we use the fact $\delta \leq T\delta^{1/2} k^{-1}$.

Let $b_1' = b_1$, $b_{k+1}' = b_k' + C\delta S_k' + C\delta^{1/2} k^{-1} \Upsilon_\delta^m$, where $S_k' = b_1' + b_2' + ... + b_k'$. Then, we prove by induction that $b_k \leq b_k'$. Indeed, if $b_j \leq b_j'$ for $j = 1, ..., k$, then $S_k \leq S_k'$,

$$b_{k+1} \leq b_k + C\delta S_k + C\delta^{1/2} k^{-1} \Upsilon_\delta^m \leq b_k' + C\delta S_k' + C\delta^{1/2} k^{-1} \Upsilon_\delta^m = b_{k+1}'.$$

Next, since $C\delta S_k' + C\delta^{1/2} k^{-1} \Upsilon_\delta^m \geq 0$, and $\{b_k'\}$ is non-decreasing, we obtain $S_k' \leq k b_k'$,

$$b_{k+1}' \leq b_k' + C\delta k b_k' + C\delta^{1/2} k^{-1} \Upsilon_\delta^m.$$

Similarly, let $b_1^* = b_1'$, $b_{k+1}^* = b_k^* + C\delta k b_k^* + C\delta^{1/2}k^{-1}\Upsilon_\delta^m$. The same argument leads to $b_k' \le b_k^*$. It is easy to derive from the definition that

$$\begin{aligned} b_{k+1}^* &= (1+C\delta k)b_k^* + C\delta^{1/2}k^{-1}\Upsilon_\delta^m \\ &= (1+C\delta k)((1+C\delta(k-1))b_{k-1}^* + C\delta^{1/2}(k-1)^{-1}\Upsilon_\delta^m) + C\delta^{1/2}k^{-1}\Upsilon_\delta^m \\ &= \cdots \\ &\le (1+C\delta k)^k \left(b_1^* + C\delta^{1/2}\Upsilon_\delta^m \sum_{j=1}^{k} j^{-1} \right) \\ &\le (1+C\delta k)^k \left(\delta^{1/2}\Upsilon_\delta^m + C\delta^{1/2}\Upsilon_\delta^m (1+\log(k)) \right). \end{aligned}$$

Let $k_T = \lfloor T/\delta^{1/2} \rfloor$. Then, using (5.89) and the above bound result, we conclude that

$$\begin{aligned} &\max_{k\le k_T} |\check{x}_k^m - X_\delta^m(t_k)| = \max_{k\le k_T} a_k \le \delta^{1/2} S_{k_T}^* + k_T\delta\Upsilon_\delta^m \\ &\le \delta^{1/2} k_T b_{k_T}^* + T\delta^{1/2}\Upsilon_\delta^m \\ &\le (1+C\delta T/\delta^{1/2})^{T/\delta^{1/2}} C\delta^{1/2}\Upsilon_\delta^m (1+\log(T/\delta^{1/2})) + T\delta^{1/2}\Upsilon_\delta^m \\ &\le CTe^{CT^2}\delta^{1/2}\Upsilon_\delta^m (1+T+|\log\delta|/2) + T\delta^{1/2}\Upsilon_\delta^m \\ &= O_p(\delta^{1/2}|\log\delta|). \blacksquare \end{aligned}$$

Proof of Theorem 9 As in Lemma 34, we realize x_k^m, $B(t)$, $H(t)$, $\check{x}_k^m$, and $X_\delta^m(t)$ [defined by $B(t)$ via (4.35)] on some common probability spaces and consider their versions $\tilde{x}_k^m$, $\tilde{B}(t)$, $\tilde{H}(t)$, $\tilde{\check{x}}_k^m$, and $\tilde{X}_\delta^m(t)$ on the probability spaces. An application of Lemma 38 leads to

$$\max_{k\le T\delta^{-1/2}} |\tilde{\check{x}}_k^m - \tilde{X}_\delta^m(t_k)| = O_P(\delta^{1/2}|\log\delta|).$$

Combining the above result with Lemma 34, we obtain

$$\max_{k\le T/\delta^{1/2}} |\tilde{x}_k^m - \tilde{X}_\delta^m(t_k)| = o_p(m^{-1/2}\delta^{1/4}) + O_P(\delta^{1/2}|\log\delta|).$$

For process $X_\delta^m(t)$, we have shown in the proof of Lemma 38 that

$$\max_{t-s\le\delta^{1/2}} |X_\delta^m(t) - X_\delta^m(s)| \le \delta^{1/2}\Upsilon_\delta^m = O_p(\delta^{1/2}),$$

and thus the same result also holds for $\tilde{X}_\delta^m(t)$. Therefore, we conclude that

$$\max_{t\le T} |\tilde{x}_\delta^m(t) - \tilde{X}_\delta^m(t)| = o_p(m^{-1/2}\delta^{1/4}) + O_p(\delta^{1/2}|\log\delta|),$$

where we use the fact that $x_\delta^m(t) = x_k^m$ for $t_k \le t < t_{k+1}$. With the theorem condition $m^{1/2}\delta^{1/4}|\log\delta| \to 0$, we immediately arrive at

$$m^{1/2}\delta^{-1/4} \max_{t\le T} |\tilde{x}_\delta^m(t) - \tilde{X}_\delta^m(t)| = o_p(1),$$

and Theorem 8 indicates that $m^{1/2}\delta^{-1/4}[x_\delta^m(t) - X(t)]$ weakly converges to $V(t)$. $\blacksquare$

Remark 16 *The proof arguments in fact also establish*

$$\max_{t\le T} |y_\delta^m(t) - X(t)| = O_p(m^{-1/2}\delta^{1/4} + \delta^{1/2}|\log\delta|).$$

5.8 Proof of Theorem 10

Part (i) can be proved by using the same argument for showing Theorem 3. First, we show parts (ii) and (iii) in one dimension. From solution (4.25) of SDE (4.24) we find that $V(t)$ follows a normal distribution with mean zero and variance

$$\Gamma(t) = \int_0^t \exp\left[-2\int_u^t \mathbb{H}g(X(v))dv\right]\boldsymbol{\sigma}^2(X(u))du.$$

It is easy to check that $\Gamma(t)$ satisfies ODE

$$\dot{\Gamma}(t) + 2[\mathbb{H}g(X(t))]\Gamma(t) - \boldsymbol{\sigma}^2(X(t)) = 0,$$

and show that the limit $\Gamma(\infty)$ of $\Gamma(t)$ as $t \to \infty$ is equal to

$$\Gamma(\infty) = \boldsymbol{\sigma}^2(X(\infty))[2\mathbb{H}g(X(\infty))]^{-1}.$$

Thus, as $t \to \infty$, $V(t)$ converges in distribution to $V(\infty) = [\Gamma(\infty)]^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ is a standard normal random variable.

Denote by $P(\theta;t)$ the probability distribution of $X_\delta^m(t)$ at time t. Then, from the Fokker-Planck equation, we have

$$\frac{\partial P(\theta;t)}{\partial t} = \nabla\left[-\nabla g(\theta)P(\theta;t) - \frac{\delta}{2m}\boldsymbol{\sigma}^2(X(t))\,\nabla P(\theta;t)\right],$$

and its stationary distribution $P(\theta)$ satisfies

$$0 = \nabla\left[-\nabla g(\theta)P(\theta) - \frac{\delta}{2m}\boldsymbol{\sigma}^2(X(\infty))\,\nabla P(\theta)\right],$$

which has solution

$$P(\theta) \propto \exp\left\{-\frac{2m}{\delta\boldsymbol{\sigma}^2(\check{\theta})}g(\theta)\right\}.$$

The corresponding stationary distribution $P_0(v)$ for $V_\delta^m(\infty) = (m/\delta)^{1/2}(X_\delta^m(\infty) - \check{\theta})$ takes the form

$$P_0(v) \propto \exp\left\{-\frac{m}{\delta\boldsymbol{\sigma}^2(\check{\theta})}g\left(\check{\theta} + \sqrt{\delta/m}\,v\right)\right\} \sim \exp\left\{-\frac{2m}{\delta\boldsymbol{\sigma}^2(\check{\theta})}\left[g(\check{\theta}) + \frac{\delta\mathbb{H}g(\check{\theta})}{2m}v^2\right]\right\}$$
$$\propto \exp\left\{-\frac{\mathbb{H}g(\check{\theta})}{\boldsymbol{\sigma}^2(\check{\theta})}v^2\right\},$$

where we use the fact that $\nabla g(\check{\theta}) = 0$, and the asymptotics are based on taking $\delta \to 0$, $m \to \infty$. Therefore, P_0 converges to $N\left(0, \frac{\boldsymbol{\sigma}^2(\check{\theta})}{2\mathbb{H}(\check{\theta})}\right)$, and we conclude that $V_\delta^m(\infty)$ has a limiting normal distribution with mean zero and variance $\boldsymbol{\sigma}^2(\check{\theta})[2\mathbb{H}g(\check{\theta})]^{-1} = \Gamma(\infty)$.

Similarly, we can show parts (ii) and (iii) in the multivariate case by following the matrix arguments given in Gardiner (2009, Chapters 4 & 6) and Da Prato and Zabczyk

(1996, Chapter 9) in the following manner. Using the explicit solution (4.25) of SDE (4.24), we find that $V(t)$ follows a normal distribution with mean zero and variance matrix

$$\begin{aligned}\Gamma(t) &= \int_0^t \exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] \boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' \exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] du \\ &= \int_0^t \exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] du + \zeta_t,\end{aligned} \tag{5.91}$$

where

$$\begin{aligned}\zeta_t = &\int_0^t \exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] \left\{\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\right\} \\ &\exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] du.\end{aligned}$$

Similar to the proof for Part 3 of Theorem 3, we show that as $t \to \infty$, $|\zeta_t| \to 0$. Indeed, for any $\epsilon > 0$, there exists $t_0 > 0$, such that for any $u > t_0$,

$$\begin{aligned}&\left|\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\right| < \epsilon, \ \left|\mathbb{H}g(X(u))[\mathbb{H}g(X(\infty))]^{-1}\right| > 1-\epsilon, \\ &\left|\int_0^{t_0} \exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] \left\{\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\right\}\right. \\ &\left.\exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] du\right| \\ &\le \left|\exp\left[-2\int_{t_0}^t \mathbb{H}g(X(v))dv\right]\right| \int_0^{t_0} \left|\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\right| du \\ &\le C\left|\exp\left[-2\int_{t_0}^t \mathbb{H}g(X(v))dv\right]\right| \to 0, \\ &\left|\int_{t_0}^t \exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] \left\{\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\right\}\right. \\ &\left.\exp\left[-\int_u^t \mathbb{H}g(X(v))dv\right] du\right| \\ &\le \frac{\epsilon}{1-\epsilon}\int_{t_0}^t \left|\exp\left[-2\int_u^t \mathbb{H}g(X(v))dv\right] \mathbb{H}g(X(u))\right| du |\mathbb{H}g(X(\infty))|^{-1} \\ &\le \frac{\epsilon}{2(1-\epsilon)}\left|1 - \exp\left[-2\int_{t_0}^t \mathbb{H}g(X(v))dv\right]\right| |\mathbb{H}g(X(\infty))|^{-1} \\ &\le \frac{\epsilon}{2(1-\epsilon)}|\mathbb{H}g(X(\infty))|^{-1} \to 0, \text{ as we let } \epsilon \to 0,\end{aligned}$$

and these results implies that the integral in ζ_t can be divided into two parts over $[0, t_0]$ and $[t_0, t]$, both of which go to zero as $t \to \infty$.

Now, we verify the detailed balance condition using (5.91) and $\zeta_t \to 0$. Direct algebraic manipulations show that

$$
\begin{aligned}
&\boldsymbol{H}g(X(t))\Gamma(t) + \Gamma(t)\boldsymbol{H}g(X(t)) \\
&= \int_0^t \boldsymbol{H}g(X(t)) \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] \boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] du \\
&+ \int_0^t \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] \boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] \boldsymbol{H}g(X(t))du \\
&= \int_0^t \boldsymbol{H}g(X(t)) \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] du \\
&+ \int_0^t \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] \boldsymbol{H}g(X(t))du \\
&+ \boldsymbol{H}g(X(t))\zeta_t + \zeta_t\boldsymbol{H}g(X(t)) \\
&= \int_0^t \frac{d}{du}\left\{\exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_u^t \boldsymbol{H}g(X(v))dv\right]\right\} du \\
&+ \boldsymbol{H}g(X(t))\zeta_t + \zeta_t\boldsymbol{H}g(X(t)) \\
&= \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \\
&- \exp\left[-\int_0^t \boldsymbol{H}g(X(v))dv\right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_0^t \boldsymbol{H}g(X(v))dv\right] \\
&+ \boldsymbol{H}g(X(t))\zeta_t + \zeta_t\boldsymbol{H}g(X(t)),
\end{aligned}
$$

where by assumption we have that as $t \to \infty$, $\int_0^t \boldsymbol{H}g(X(v))dv \to \infty$, which together with $\zeta_t \to 0$ indicates that the last three terms on the right-hand side of the above expression go to zero. Hence we have shown that as $t \to \infty$, $\boldsymbol{H}g(X(t))\Gamma(t) + \Gamma(t)\boldsymbol{H}g(X(t)) \to \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'$—that is, their limits obey the following detailed balance condition,

$$\boldsymbol{H}g(X(\infty))\Gamma(\infty) + \Gamma(\infty)\boldsymbol{H}g(X(\infty)) = \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'. \tag{5.92}$$

With the limit $\Gamma(\infty)$ of $\Gamma(t)$ as $t \to \infty$, we conclude that $V(t)$ converges in distribution to $V(\infty) = [\Gamma(\infty)]^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ is a standard normal random vector.

Denote by $P(\theta; t)$ the probability distribution of $X_\delta^m(t)$ at time t. Then, from the Fokker-Planck equation, we have

$$\frac{\partial P(\theta;t)}{\partial t} = \nabla\left[-\nabla g(\theta)P(\theta;t) - \frac{\delta}{2m}\boldsymbol{\sigma}(X(t))[\boldsymbol{\sigma}(X(t))]'\nabla P(\theta;t)\right]$$

and under the detailed balance condition (5.92), its stationary distribution $P(\theta)$ satisfies

$$0 = \nabla\left[-\nabla g(\theta)P(\theta) - \frac{\delta}{2m}\boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\nabla P(\theta)\right],$$

which corresponds to a normal stationary distribution $N(0, \Gamma(\infty))$ for $V_\delta^m(\infty) = (m/\delta)^{1/2}$ $(X_\delta^m(\infty) - \check{\theta})$. Thus, we conclude that $V_\delta^m(\infty)$ has a limiting normal distribution with mean zero and variance $\Gamma(\infty)$. ■

5.9 Proof of Theorem 11

As $\nabla g(\check{\theta}) = 0$, by Taylor expansion we have

$$g(X_\delta^m(t)) = g(X(t)) + (\delta/m)^{-1/2}\nabla g(X(t))V_\delta^m(t) + \frac{\delta}{2m}[V_\delta^m(t)]'\mathbb{H}g(X(t))V_\delta^m(t) + o_P(\delta/m),$$
$$\nabla g(X_\delta^m(t)) = \nabla g(X(t)) + (\delta/m)^{-1/2}\mathbb{H}g(X(t))V_\delta^m(t) + o_P((\delta/m)^{1/2}),$$
$$g(X(t)) \sim g(\check{\theta}) + \nabla g(\check{\theta})[X(t)-\check{\theta}] + \frac{1}{2}[X(t)-\check{\theta}]'\mathbb{H}g(\check{\theta})[X(t)-\check{\theta}]$$
$$= g(\check{\theta}) + \frac{1}{2}[X(t)-\check{\theta}]'\mathbb{H}g(\check{\theta})[X(t)-\check{\theta}],$$
$$\nabla g(X(t)) \sim \mathbb{H}g(\check{\theta})[X(t)-\check{\theta}], \quad \mathbb{H}g(X(t)) \sim \mathbb{H}g(\check{\theta}),$$
$$g(X_\delta^m(t)) \sim g(\check{\theta}) + \frac{1}{2}[X_\delta^m(t)-\check{\theta}]'\mathbb{H}g(\check{\theta})[X_\delta^m(t)-\check{\theta}],$$
$$X_\delta^m(t) - \check{\theta} = X(t) - \check{\theta} + (\delta/m)^{-1/2}V_\delta^m(t).$$

Thus, by Theorem 5, we have that $g(X_\delta^m(t))$ and $\nabla g(X_\delta^m(t))$ behave as, respectively,

$$g(X(t)) + (\delta/m)^{1/2}\nabla g(X(t))V(t) + \frac{\delta}{2m}[V(t)]'\mathbb{H}g(X(t))V(t),$$

$$\text{and} \quad \nabla g(X(t)) + (\delta/m)^{1/2}\mathbb{H}g(X(t))V(t).$$

Similar to the stationary distribution part of the proof for Theorem 10, when $\mathbb{H}g(\check{\theta})$ is positive definite, we can derive the stationary distribution of $V(t)$ to be a normal distribution with mean zero and variance $\Gamma(\infty)$ defined by (4.38). Thus, we have

$$E[V(t)] = 0, \qquad E\{[V(t)]'\mathbb{H}g(X(t))V(t)\} = tr[\Gamma(\infty)\mathbb{H}g(X(t))],$$
$$Var\{\mathbb{H}g(X(t))V(t)\} = tr[\Gamma(\infty)\{\mathbb{H}g(X(t))\}^2],$$

where the expectation is taken under the stationary distribution. Taking the trace on both sides of (4.38), we obtain

$$tr[\Gamma(\infty)\mathbb{H}g(X(\infty))] = tr[\mathbb{H}g(X(\infty))\Gamma(\infty)] = tr[\boldsymbol{\sigma}^2(X(\infty))]/2,$$

and multiplying $\mathbb{H}g(X(\infty))$ on both sides of (4.38) and then performing the trace operation, we arrive at

$$tr[\Gamma(\infty)\{\mathbb{H}g(X(\infty))\}^2] = tr[\mathbb{H}g(X(\infty))\Gamma(\infty)\mathbb{H}g(X(\infty))] = tr[\boldsymbol{\sigma}^2(X(\infty))\mathbb{H}g(X(\infty))]/2.$$

Putting these results together and using $\mathbb{H}g(X(t)) \to \mathbb{H}g(X(\infty))$ as $t \to \infty$, we prove (4.46) and (4.47).

For the saddle point case, for simplicity, we assume that $\mathbb{H}g(\check{\theta})$ is a diagonal matrix with eigenvalues λ_i, $i = 1, \cdots, p$. Then, $V(t)$ has covariance function (Gardiner, 2009)

$$[Cov(V(t), V(s))]_{ii} = \frac{\boldsymbol{\sigma}_{ii}(X(t))}{2\lambda_i}\left[e^{-\lambda_i|t+s|} - e^{-\lambda_i|t-s|}\right],$$

which, for negative λ_i, diverge as $t, s \to \infty$. Thus, $V(t)$ does not have any limiting stationary distribution. ■

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grants DMS-1528375, DMS-1707605, and DMS-1913149). The authors thank action editors Rina Foygel Barber and Ryan Tibshirani, an associate editor, and three anonymous referees for their helpful comments and suggestions, which led to significant improvements in both the substance and the style of the paper.

References

A. Ali, Z. Kolter, and R. Tibshirani. A continuous-time view of early stopping for least squares regression. *International Conference on Artificial Intelligence and Statistics*, 2019.

P. Billingsley. *Convergence of Probability Measures*. Wiley, 2nd Edition. 1999.

P. J. Bickel, F. Götze, and W. R. van Zwet. Resampling fewer than n observations: Gains, loses, and remedies for loses. *Statistica Sinica* 7, 1-31, 1997.

S. Blanes, F. Casas, J. A. Oteo, and J. Ros. The Magnus expansion and some of its applications. *Physics Reports* 470, 151-238, 2009.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press. 2004.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3, 1-122, 2011.

J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley. 2008.

C. Chen, D. Carlson, Z. Gan, C. Li, and L. Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. arXiv preprint arXiv:1512.07962, 2016.

C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. arXiv preprint arXiv:1610.06665, 2016.

X. Chen, J. D. Lee, X. T. Tong, Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *Annals of Statistics* 48, 251-273, 2020.

Y. S. Chow and H. Teicher. *Probability Theory. Independence, Interchangeability, Martingales*. Springer, 3rd Edition. 1997.

M. Csörgö, L. Horváth, and P. Kokoszka. Approximation for bootstrapped empirical processes. *Proceedings of the American Mathematical Society* 128, 2457-2464, 1999.

S. Csörgö, and D. M. Mason. Bootstrapping empirical functions. *Annals of Statistics* 17, 1447-1471, 1989.

G. Da Prato and J. Zabczyk. *Ergodicity for Infinite Dimensional Systems*. Cambridge University Press. 1996.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society B* 79, 651-676, 2017a.

A. S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. *Proceedings of Machine Learning Research* 65, 1-12, 2017b.

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications* 129, 5278-5311, 2019.

J. Fan, W. Gong, C. J. Li, and Q. Sun. Statistical sparse online regression: A diffusion approximation perspective. *International Conference on Artificial Intelligence and Statistics* 84, 1017-1026, 2018.

D. J. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, B. Woodworth. The complexity of making the gradient small in stochastic convex optimization. *Proceedings of Machine Learning Research* 99, 1-27, 2019.

A. Frieze, M. Jerrum, and R. Kannan. Learning linear transformations. FOCS 1996.

C. W. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer, 4th Edition. 2009.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. *International Conference on Learning Theory (COLT)*, 2015.

S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming* 156, 59-99, 2015.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT press. 2016.

T. Hida and S. Si. *Lectures on White Noise Functionals*. World Scientific. 2008.

S. W. He, J. G. Wang, and J. A. Yan. *Semimartingale Theory and Stochastic Calculus*. Science Press and CRC Press. 1992.

N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes*. North-Holland Mathematical Library. 1981.

J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer, 2nd Edition. 2003.

S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in SGD. arXiv preprint arXiv:1711.04623, 2018.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, M. I. Jordan. How to escape saddle points efficiently. arXiv preprint arXiv:1703.00887, 2017.

K. Kawaguchi. Deep learning without poor local minima. *Advances In Neural Information Processing Systems*, 586-594, 2016.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: generalization gap and sharp minima. *International Conference on Learning Representations*, 2017.

Y. Kifer. The exit problem for small random perturbations of dynamical systems with a hyperbolic fixed point. *Israel Journal of Mathematics* 40, 74-96, 1981.

J. Kim and D. Pollard. Cube root asymptotics. *Annals of Statistics* 18, 191-219, 1990.

P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer. 1992.

J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent R.V.'s and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 32, 111-131, 1975.

J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent R.V.'s and the sample DF. II. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 34, 33-58, 1976.

H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer. 2003.

J. D. Lee, M. Simchowitz, M. I Jordan, and B. Recht. Gradient descent only converges to minimizers. *Conference on Learning Theory*, 1246-1257, 2016.

T. Li, L. Liu, A. Kyrillidis, and C. Caramanis. Statistical Inference Using SGD. *AAAI Conference on Artificial Intelligence*, 3571-3578, 2018.

C. J. Li, Z. Wang, H. Liu. Online ICA: Understanding global dynamics of nonconvex optimization via diffusion processes. *Conference on Neural Information Processing Systems (NIPS)*, 2016.

C. J. Li, M. Wang, H. Liu, and T. Zhang. Diffusion approximations for online principal component estimation and global convergence. *Conference on Neural Information Processing Systems (NIPS)*, 2017a.

C. J. Li, L. Li, J. Qian, J. Liu. Batch size matters: A diffusion approximation framework on nonconvex stochastic gradient descent. arXiv preprint arXiv:1705.07562, 2017b.

H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Q. Li, C. Tai, W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. arXiv preprint arXiv:1511.06251, 2015.

V. Luo and Y. Wang. How many factors influencing minima in SGD? arXiv preprint arXiv:2009.11858, 2020.

Y. A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, c, and M. I. Jordan Is there an analog of Nesterov acceleration for MCMC? arXiv preprint arXiv:1902.00996, 2019.

S. Mandt, M. D. Hoffman, and D. M. Blei. A variational analysis of stochastic gradient algorithms. arXiv preprint arXiv:1602.02666, 2016.

S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research* 18, 1-35, 2017.

P. Massart. Strong approximation for multivariate empirical and related processes via KMT constructions. *Annals of Probability* 17, 266-291, 1989.

A. Nemirovskii and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons. 1983.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19, 1574-1609, 2009.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady* 27, 372-376, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer. 2004.

J. Nocedal and S. Wright. *Numerical Optimization*. Springer. 2006.

D. Pollard. *Empirical Processes: Theory and Applications*. CMS. 1988.

B. T. Polyak. *Introduction to Optimization*. Optimization Software New York. 1987.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* 30, 838, 1992.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. arXiv preprint arXiv:1109.5647, 2012.

E. Rio. Strong approximation for set-indexed partial sum processes via KMT constructions I. *Annals of Probability* 21, 759-790, 1993a.

E. Rio. Strong approximation for set-indexed partial-sum processes via KMT constructions II. *Annals of Probability* 21, 1706-1727, 1993b.

D. Ruppert. Efficient estimators from a slowly converging Robbins-Monro process. Technical report. 1988.

A. P. Ruszczyński. *Nonlinear Optimization*. Princeton University Press. 2006.

C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl. Measuring the effects of Data Parallelism on Neural Network Training. *Journal of Machine Learning Research* 20, 1-49, 2019.

N. Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer Science & Business Media. 2012.

J. Sirignano and K. Spiliopoulos. Stochastic gradient descent in continuous time: A central limit theorem. arXiv preprint arXiv:1710.04273, 2017.

W. Su, S. Boyd, and E. J. Candés. A differential equation for modeling Nesterov's accelerated gradient method: Theory and Insights. *Journal of Machine Learning Research* 17, 1-43, 2016.

P. Toulis, E. M. Airoldi, and J. Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. *International Conference on Machine Learning (ICML)*, 667-675, 2014.

P.Toulis and E. M. Airoldi. Scalable estimation strategies based on stochastic approximations: classical results and new insights. *Statistics and Computing* 25, 781-795, 2015.

P. Toulis and E. M. Airoldi. Stochastic gradient methods for principled estimation with large datasets. *In Handbook of Big Data*. CRC Press. 2016.

P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Annals of Statistics* 45, 1694-1727, 2017.

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer. 2000.

Y. Wang. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. ArXiv preprint arXiv:1711.09514, 2019.

G. N. Watson. *A Treatise on the Theory of Bessel Functions*. Cambridge Mathematical Library. Cambridge University Press. 1995.

A. Wibisono, A. Wilson, and M. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences* 113, E7351-E7358, 2016.

Z. A. Zhu. *Katyusha: The first direct acceleration of stochastic gradient methods. Journal of Machine Learning Research* 18, 1-51, 2018.