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

Yazhen Wang
University of Wisconsin-Madison

Abstract

This paper investigates asymptotic behaviors of gradient descent algorithms (particularly accelerated gradient descent and stochastic gradient descent) in the context of stochastic optimization arose in statistics and machine learning where objective functions are estimated from available data. We show that these algorithms can be modeled by continuous-time ordinary or stochastic differential equations, and their asymptotic dynamic evolutions and distributions are governed by some linear ordinary or stochastic differential equations, as the data size goes to infinity. We illustrate that our study can provide a novel unified framework for a joint computational and statistical asymptotic analysis on dynamic behaviors of these algorithms with the time (or the number of iterations in the algorithms) and large sample behaviors of the statistical decision rules (like estimators and classifiers) that the algorithms are applied to compute, where the statistical decision rules are the limits of the random sequences generated from these iterative algorithms as the number of iterations goes to infinity. The analysis results may shed light on the phenomenon of escaping from saddle points and converging to good local minimizers when stochastic gradient descent algorithms are applied to solve non-convex optimization problems in deep learning.

Key words: Accelerated gradient descent and stochastic gradient descent, asymptotic distribution and weak convergence, bootstrap, mini-batch, (stochastic) optimization, ordinary or stochastic differential equation, second order stochastic differential equation, and stationary distribution.

1 Introduction

Optimization plays an important role in scientific fields ranging from physical sciences to machine learning and statistics to engineering. Numerous algorithms and methods have been proposed to solve optimization problems. Examples include Newton's methods, gradient and subgradient descent, conjugate gradient methods, trust region methods, and interior point methods (see Polyak, 1987; Boyd and Vandenberghe, 2004; Nocedal and Wright, 2006; Ruszczynski, 2006; Boyd et al., 2011; Shor, 2012; Goodfellow et al. (2016) for expositions). Practical problems arose in fields like statistics and machine learning usually involve optimization settings where the objective functions are empirically estimated from available data (a training sample or a statistical sample) with the form of a sum of differentiable functions. We refer such optimization problems with random objective functions as stochastic optimization. As data sets in practical problems grow rapidly in scale and complexity, methods such as stochastic gradient descent can scale to the enormous size of big data and have been very popular. There have been great recent research interest and work on the theory and practice of gradient descent and its extensions and variants. For example, a number of recent papers were devoted to investigate stochastic gradient descent and its variants for solving complex optimization problems (Chen et al. (2016), Li et al. (2017), Mandt et al. (2016), Kawaguchi (2016), Keskar et al. (2017), Lee et al. (2016), Ge et al. (2015), Jin et al. (2017)). Su et al. (2016) showed that the continuous-time limit of Nesterov's accelerated gradient descent is a second-order ordinary differential equation that can be used to understand and analyze the acceleration phenomenon and generalize Nesterov's scheme; and Wibisono et al. (2016) studied acceleration from a continuous-time variational point of view and further developed a systematic approach to understand acceleration phenomenon and produce acceleration algorithms from continuous-time differential equations generated by a so-called Bregman Lagrangian. In spite of compelling theoretical and numerical evidence on the value of the stochastic approximation idea and acceleration phenomenon, yet there remains some conceptual mystery in the acceleration and stochastic approximation schemes.

This paper establishes asymptotic theory for gradient descent, stochastic gradient descent, and accelerated gradient descent in the stochastic optimization setup. We derive continuous-time ordinary or stochastic differential equations to model the dynamic behaviors of these gradient descent algorithms and investigate their asymptotic distributions as data size goes to infinity. Specifically 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 convex 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 statistical 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. We show that random sequences generated from these gradient descent algorithms and their continuous time ordinary or stochastic differential equations for the stochastic optimization setting converge to ordinary differential equations for the corresponding deterministic optimization set-up, with asymptotic distributions governed by some linear ordinary or stochastic differential equations. Moreover, our analysis may offer a novel unified framework to carry out a joint asymptotic analysis for computational algorithms and statistical decision rules that the algorithms are applied to compute. As iterated computational methods, these gradient descent algorithms generate sequences of numerical values that converge to the exact decision rule or the true parameter value for the corresponding optimization problems, as 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 asymptotic distributions of statistical decision rules as the sample size goes to infinity. In other words, the asymptotic analysis can be done with both time and data size, where the time direction corresponds to the computational asymptotics on dynamic behaviors of algorithms, and the data size direction associates with usual statistical asymptotics on the statistical behaviors of decision rules such as estimators and classifiers. To the best of our knowledge, this is the first paper to provide rigorous treatments between stochastic gradient descent and stochastic differential equations, discover the second order stochastic differential equations for the accelerated case, and offer the unified asymptotic analysis. The continuous-time modeling and the joint asymptotic analysis may shed some light via large deviation theory and limiting stationary distribution 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.

The rest of the paper is proceeded as follows. Section 2 introduces 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 carry out a joint analysis on computational asymptotics with time (or iteration) for the gradient descent algorithms and statistical asymptotics with sample size for statistical decision rules that the algorithms are applied to compute. 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 illustrate a joint analysis on computational asymptotics with time for the stochastic gradient descent algorithms and statistical asymptotics with sample size for statistical decision rules. Section 5 features an example. All technical proofs are relegated in the appendix section.
We adopt the following notations and conventions. For the stochastic optimization problem considered in Sections 3 and 4, we add a super index 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 any such subscripts or superscripts are for sequences and functions corresponding to the deterministic optimization problem given in Section 2.

2 Ordinary differential equations for gradient descent algorithms

Consider the following minimization problem

$$\min_{\theta\in\Theta} g(\theta), \tag{1}$$

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

$$x_k = x_{k-1} - \delta\nabla g(x_{k-1}), \tag{2}$$

where ∇ denotes gradient operator, and δ is a positive constant which 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 $x_k \approx X(k\sqrt{\delta})$ as follows. Define step function $x_\delta(t) = x_k$ for $(k-1)\delta < t \leq k\delta$, and as $\delta \to 0$, $x_\delta(t)$ approches to $X(t)$ satisfying

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

where $\dot{X}(t)$ denotes the derivative of $X(t)$, and initial value $X(0) = x_0$.
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

$$x_k = y_{k-1} - \delta\nabla g(y_{k-1}), \qquad y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}), \tag{4}$$

where δ is a positive constant. Using (4) we derive a recursive relationship between consecutive increments

$$\frac{x_{k+1} - x_k}{\sqrt{\delta}} = \frac{k-1}{k+2}\frac{x_k - x_{k-1}}{\sqrt{\delta}} - \sqrt{\delta}\nabla g(y_k). \tag{5}$$

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 step function $x_\delta(t) = x_k$ for $(k-1)\sqrt{\delta} < t \leq k\sqrt{\delta}$, 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. For $t = k\sqrt{\delta}$, as $\delta \to 0$, we have $x_k = x_{t/\sqrt{\delta}} = X(t)$, $x_{k+1} = x_{(t+\sqrt{\delta})/\sqrt{\delta}} = X(t + \sqrt{\delta})$, and

$$\begin{aligned} y_k &= X(t) + \frac{t/\sqrt{\delta} - 1}{t/\sqrt{\delta} + 2}[X(t) - X(t - \sqrt{\delta})] \\ &= X(t) + \left(1 - \frac{3\sqrt{\delta}}{t + 2\sqrt{\delta}}\right)[X(t) - X(t - \sqrt{\delta})] \\ &= X(t) + O(\sqrt{\delta}). \end{aligned}$$

Applying Taylor expansion and using L-Lipshitz continuous gradients we obtain

$$\frac{x_{k+1} - x_k}{\sqrt{\delta}} = \frac{X(t + \sqrt{\delta}) - X(t)}{\sqrt{\delta}} = \dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{\delta} + o(\sqrt{\delta}),$$

$$\frac{x_k - x_{k-1}}{\sqrt{\delta}} = \frac{X(t) - X(t - \sqrt{\delta})}{\sqrt{\delta}} = \dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{\delta} + o(\sqrt{\delta}),$$

$$\sqrt{\delta}\nabla g(y_k) = \sqrt{\delta}\nabla g(X(t)) + o(\sqrt{\delta}),$$

where $\ddot{X}(t)$ denotes the second derivative of $X(t)$. Substituting above results into equation (5) we obtain

$$\dot{X}(t)+\frac{1}{2}\ddot{X}(t)\sqrt{\delta}+o(\sqrt{\delta}) = \left(1 - \frac{3\sqrt{\delta}}{t+2\sqrt{\delta}}\right)(\dot{X}(t)-\frac{1}{2}\ddot{X}(t)\sqrt{\delta}+o(\sqrt{\delta}))-\sqrt{\delta}\nabla g(X(t))+o(\sqrt{\delta}).$$

Re-arranging the terms and dividing $\sqrt{\delta}$ on both sides lead to

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla g(X(t)) + o(1) = 0.$$

Note that $X(t)$ is free of δ. As $\delta \to 0$, the equation becomes

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla g(X(t)) = 0, \tag{6}$$

with the initial conditions $X(0) = x_0$ and $\dot{X}(0) = 0$. As the coefficient $3/t$ in (6) is singular at $t = 0$, classical ordinary differential equation theory is not applicable to establish the existence or uniqueness of the solution to (6). Su et al. (2016) derived (6) and proved that it 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) has shown that, because of the difference, the accelerated gradient descent algorithm moves much faster than the plain gradient descent algorithm along the curve $X(t)$. See also Wibisono et al. (2016) for more elaborate explanation on the acceleration phenomenon.

3 Gradient descent for stochastic optimization

Let $\theta = (\theta_1, ..., \theta_p)'$ be the parameter that we are interested in, and U be an relevant random element on a probability space with a given 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 M-estimation, we may treat U as a sample observation and $\ell(\theta, u)$ a ρ-function; in nonparametric function estimation and machine learning, we may choose U an observation and

$\ell(\theta; u)$ equal to a loss function plus some penalty. For these problems we need to consider the corresponding minimization problem (1) for characterizing the true parameter value, but practically, because $g(\theta)$ is usually unavailable, we have to employ its empirical version and consider a stochastic optimization problem, described as follows:

$$\min_{\theta \in \Theta} \mathcal{L}^n(\theta; \mathbf{U}_n), \tag{7}$$

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 training sample or statistical sample, and we assume $U_1, \cdots, U_n$ are i.i.d. and follow distribution Q.
Minimization problem (1) characterizes the true value of parameter θ in the statistical model such as an underlying M-functional in the M-estimation setup. As the true target function $g(\theta)$ is usually unknown in practices, we often solve stochastic minimization problem (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 obtain 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 (7) approaches that of (1).

3.1 Plain gradient descent algorithm

Applying the plain gradient descent scheme to the minimization problem (7) with initial value x_0^n, we obtain the following iterative algorithm to compute the solution of (7),

$$x_k^n = x_{k-1}^n - \delta \nabla \mathcal{L}^n(x_{k-1}^n; \mathbf{U}_n), \tag{8}$$

where $\delta > 0$ is a step size or learning rate, and $\mathcal{L}^n$ is the objective function in minimization problem (7).
Following the continuous curve approximation described in Section 2 we define step function $x_\delta^n(t) = x_k^n$ for $(k-1)\delta < t \leq k\delta$, and for each n, as $\delta \to 0$, $x_\delta^n(t)$ approaches a smooth curve $X^n(t)$, $t \geq 0$, given by

$$\dot{X}^n(t) + \nabla \mathcal{L}^n(X^n(t); \mathbf{U}_n) = 0, \tag{9}$$

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 $X^n(0) = x_0^n$.
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 (9) in a 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 $\mathbf{U}_n$ is identically distributed as $\mathbf{U}_n$, and $(\mathbf{U}_n^\dagger, X_\dagger^n(t))$ satisfies (9), and $X_\dagger^n(t)$ is called a (weak) solution of equation (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 variables $U_i^\dagger$ whose distributions are given by Q. The weak solution definition, which shares the same spirit as that for stochastic differential equations (see Ikeda and Watanabe (1981) and more in Section 4), will be very handy in facilitating our asymptotic analysis in this paper. For simplicity we drop 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 minimization problem (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 (7),

$$x_k^n = y_{k-1}^n - \delta \nabla \mathcal{L}^n(y_{k-1}^n; \mathbf{U}_n), \qquad y_k^n = x_k^n + \frac{k-1}{k+2}(x_k^n - x_{k-1}^n). \tag{10}$$

Using the continuous curve approach described in Section 2 we can define step function $x_\delta^n(t) = x_k$ for $(k-1)\sqrt{\delta} < t \le k\sqrt{\delta}$, 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

$$\ddot{X}^n(t) + \frac{3}{t}\dot{X}^n(t) + \nabla \mathcal{L}^n(X^n(t); \mathbf{U}_n) = 0, \tag{11}$$

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. Again we define a solution $X^n(t)$ of equation (11) in a weak sense that there exist a process $X^n(t)$ and 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 (11).

3.3 Asymptotic ordinary differential equations

To make equations (9) and (11) and their solutions to be well defined and study their asymptotics we need to impose the following conditions.

A0. Assume 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. $E[\nabla^\kappa \ell(\theta; U)] = \nabla^\kappa g(\theta)$, $\kappa = 0, 1, 2$. On the parameter space Θ, $g(\cdot)$ is continuously twice differentiable and strongly convex, and $\nabla g(\cdot)$ is L-Lipschitz for some $L > 0$, with linear growth, where $\Delta = \nabla^2$ is the Laplacian operator.

A3. $\mathrm{Cov}[\frac{\partial}{\partial\theta_i}\ell(\theta; U), \frac{\partial}{\partial\theta_j}\ell(\theta; U)] = \sigma_{ij}(\theta)$, $i, j = 1, 2, ...p$, and $\mathrm{Var}[\nabla\ell(\theta; U)] = \boldsymbol{\sigma}^2(\theta) = (\sigma_{ij}(\theta))_{1\leq i,j\leq p}$ is positive definite, continuously differentiable, and L-Lipschitz with linear growth in θ.

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 a normal distribution with mean zero and variance $\boldsymbol{\sigma}^2(\theta)$ uniformly over $\theta \in \Theta_X$, where $\boldsymbol{\sigma}^2(\theta)$ is given in A3, Θ_X is a bounded subset of Θ, and the interior of Θ_X contains solutions $X(t)$ of ordinary differential equations (3) and (6) connecting the initial value x_0 and the minimizer of $g(\theta)$.

Conditions A1-A2 are often used to make optimization problems and differential equations to be well defined, and the stochastic optimization (7) corresponds to optimization (1), and Conditions A3-A4 guarantee that the solution of (7) and its associated differential equations provide large sample approximations of those for (1). Condition A4 is quite reasonable, which can be easily justified by empirical processes with common assumptions such as that $\nabla\ell(\theta; U)$, $\theta \in \Theta_X$, form a Donsker class (van der Vaart and Wellner (2000)), since solution curves $X(t)$ of ordinary differential equations (3) and (6) are deterministic and bounded, and Θ_X is bounded.
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 equations (3) and (9) [or equations (6) and (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 3.1. *Under conditions A0-A4, as $n \to \infty$, $V^n(t)$ weakly converges to $V(t) = \Pi(t)\mathbf{Z}$ on $C([0, T])$, where $\mathbf{Z} \sim N_p(0, \mathbf{I}_p)$, and matrix $\Pi(t)$ is the unique solution of the following linear differential equation*

$$\dot{\Pi}(t) + [\Delta g(X(t))]\Pi(t) + \boldsymbol{\sigma}(X(t)) = 0,\ \ \Pi(0) = 0, \tag{12}$$

for the plain gradient descent case, and

$$\ddot{\Pi}(t) + \frac{3}{t}\dot{\Pi}(t) + [\Delta g(X(t))]\Pi(t) + \boldsymbol{\sigma}(X(t)) = 0, \;\; \Pi(0) = \dot{\Pi}(0) = 0, \tag{13}$$

for the accelerated gradient descent case, where $X(t)$ in (12) and (13) are the solutions of ordinary differential equations (3) and (6), respectively.

Remark 3.1. *As we discussed early 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 minimization problems (1) and (7) should be very close to each other. We may heuristically illustrate the derivation of Theorem 3.1 as follows. Central limit theorem may lead us to see that as $n \to \infty$, $\nabla\mathcal{L}^n(\theta; \mathbf{U}_n)$ is asymptotically distributed as $\nabla g(\theta) + n^{-1/2}\boldsymbol{\sigma}(\theta)\mathbf{Z}$, where random vector $\mathbf{Z} \sim N(0, \mathbf{I}_p)$. Then asymptotically differential equations (9) and (11) are, respectively, equivalent to*

$$\dot{X}^n(t) + \nabla g(X^n(t)) + n^{-1/2}\boldsymbol{\sigma}(X^n(t))\mathbf{Z} = 0, \tag{14}$$

$$\ddot{X}^n(t) + \frac{3}{t}\dot{X}^n(t) + \nabla g(X^n(t)) + n^{-1/2}\boldsymbol{\sigma}(X^n(t))\mathbf{Z} = 0. \tag{15}$$

Appplying the perturbation method for solving ordinary differential equations, we write approximation solutions of (14) and (15) as $X^n(t) = X(t) + n^{-1/2}V(t) + o(n^{-1/2})$ and substitute it into (14) and (15). With $X(t)$ satisfying (3) or (6), ignoring higher order terms, we obtain the following equations for the limit $V(t)$ of $V_n(t)$ in the two cases, respectively,

$$\dot{V}(t) + [\Delta g(X(t))]V(t) + \boldsymbol{\sigma}(X(t))\mathbf{Z} = 0, \tag{16}$$

$$\ddot{V}(t) + \frac{3}{t}\dot{V}(t) + [\Delta g(X(t))]V(t) + \boldsymbol{\sigma}(X(t))\mathbf{Z} = 0, \tag{17}$$

where $X(t)$ is a solution of the corresponding equation (3) or (6), random variable $\mathbf{Z}$ follows the p-variate standard normal distribution $N_p(0, \mathbf{I}_p)$, and initial conditions $V(0) = \dot{V}(0) = 0$. Using linear scaling we show that (16) and (17) have unique solutions $V(t) = \Pi(t)\mathbf{Z}$, where $\Pi(t)$ are unique solutions of (12) and (13).

As step function $x_\delta^n(t)$ is used to model x_k^n generated from gradient descent algorithms (8) and (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 3.2. *Under assumption A0-A4, as $\delta \to 0$ and $n \to \infty$, we have*

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

where $x_\delta^n(t)$ are continuous-time step processes for discrete x_k^n generated from algorithms (8) and (10), with continuous curves $X^n(t)$ defined by (9) and (11), for the cases of plain and accelerated gradient descent algorithms, respectively. In particular, if we take δ such that as $\delta \to 0$ and $n \to \infty$, $n\delta \to 0$, then on $D([0,T])$, $n^{1/2}[x_\delta^n(t) - X(t)]$ weakly converges to $\Pi(t)\mathbf{Z}$, where $X(t)$ is the solution of (3) or (6), and $\Pi(t)$ and $\mathbf{Z}$ are defined in Theorem 3.1. That is, $x_\delta^n(t)$ and $X^n(t)$ share the same weak convergence limit.

Remark 3.2. *There are two types of asymptotic analyses in the set up. One type is to employ continuous differential equations to model discrete sequences generated from gradient descent algorithms, which is associated with δ treated as step size between consecutive sequence points. Another type involves the use of random objective functions in stochastic optimization, which are estimated from sample data of size n. We refer the first and second types as computational (modeling) and statistical asymptotics, respectively. The computational asymptotic analysis is that for each n, differential equations (9) and (11)[or (14) and (15)] provide continuous solutions as the limits of discrete sequences generated from algorithms (8) and (10), respectively, when δ is allowed to go to zero. Theorem 3.1 provides the statistical asymptotic analysis to describe the behavior difference between the data based solutions $X^n(t)$ and the true solutions $X(t)$, as the sample size n goes to infinity. Theorem 3.2 involves both types of asymptotics and shows that as $\delta \to 0$ and $n \to \infty$, $x_\delta^n(t) - X^n(t)$ is of order $\delta^{1/2}$. As a computational modeling parameter, δ can be any arbitrary sequence approaching zero, and we may let it depend on n and choose $\delta = \delta_n$ that goes to zero fast enough 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 A framework to unify computational and statistical asymptotic analysis

The two types of asymptotics associated with δ and n seem to be quite different, with one for computational algorithms and one for statistical inferences. This section will elaborate further about these analyses and provide a framework to unify both point of views. Denote the solutions of optimization problems (1) and (7) by $\check{\theta}$ and $\hat{\theta}_n$, respectively. In the statistical set-up, $\check{\theta}$ and $\hat{\theta}_n$ represent the true parameter value

and estimator of the parameter θ, respectively. Then using the definitions of $\check{\theta}$ and $\hat{\theta}_n$ and Taylor expansion, we have

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

and the law of large number implies that $\Delta \mathcal{L}^n(\check{\theta}; \mathbf{U}_n)$ converges in probability to $\Delta g(\check{\theta})$ as $n \to \infty$. On the other hand, Assumption 4 indicates that

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

where $\mathbf{Z}$ stands for a standard normal random vector. Thus, $n^{1/2}(\hat{\theta}_n - \check{\theta})$ is asymptotically distributed as $[\Delta g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$. On the other hand, gradient descent algorithms generate sequences corresponding to $X(t)$ and $X^n(t)$ are expected to approach the solutions of the two optimization problems (1) and (7), respectively, hence $X(t)$ and $X^n(t)$ must move towards $\check{\theta}$ and $\hat{\theta}_n$, respectively, and $V_n(t)$ and $V(t)$ are reaching their corresponding targets $n^{1/2}(\hat{\theta}_n - \check{\theta})$ and $[\Delta g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$. Below we will provide a framework to connect $(X^n(t), X(t))$ with $(\hat{\theta}_n, \check{\theta})$ and $(V^n(t), V(t))$ with $n^{1/2}(\hat{\theta}_n - \check{\theta})$ and $[\Delta g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$.

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

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

Solutions, $X(t)$, $X^n(t)$, $V(t)$, $V^n(t)$ of ordinary differential equations (3), (6), (9), (11)-(17) all live on $C(\mathbb{R}_+)$, and we can study their weak convergence on $C(\mathbb{R}_+)$. Similarly we may adopt the Skorokhod space $D(\mathbb{R}_+)$ equipped with the Skorokhod metric for the weak convergence study of $x^n_\delta(t)$ (see Billingsely (1999)). The following theorem establishes the weak convergence of these processes on $D(\mathbb{R}_+)$ and studies their asymptotic behaviors as $t \to \infty$.

Theorem 3.3. *Suppose that the assumption A0-A4 are met, $\Delta g(\check{\theta})$ is positive definite, all eigenvalues of $\int_0^t \Delta g(X(s))ds$ diverge as $t \to \infty$, and $n\delta \to 0$ 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 $\sqrt{n}[x^n_\delta(t) - X(t)]$ weakly converge to $V(t)$, $t \in [0, +\infty)$.*

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

(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 equations (2)-(4) and (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 (8)-(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 $[\Delta g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})$, where $V(t)$, defined in (16) and (17), is the weak convergence limit of $V^n(t)$ as $n \to \infty$.

Remark 3.3. *Denote the limits of the processes in Theorem 3.3 as $t, k \to \infty$ by the corresponding processes with t and k replacing by ∞. Then Theorem 3.3 shows that for the plain gradient descent case, $x_\infty = x_\delta(\infty) = X(\infty) = \check{\theta}$, $x_\infty^n = x_\delta^n(\infty) = X^n(\infty) = \hat{\theta}_n$, $V^n(\infty) = \sqrt{n}[X^n(\infty) - X(\infty)] = \sqrt{n}[x_\delta^n(\infty) - X(\infty)] = \sqrt{n}(\hat{\theta}_n - \check{\theta})$, $V(\infty) = [\Delta g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\mathbf{Z} = [\Delta g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$, $V(t)$ 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 process $V^n(t)$ is indexed by n and t, its limits are the same regardless the order of $n \to \infty$ and $t \to \infty$. Also as $\check{\theta} = X(\infty)$ is the minimizer of convex function $g(\cdot)$, the positive definite assumption $\Delta g(\check{\theta}) = \Delta g(X(\infty)) > 0$ is very reasonable; since the limit, $\Delta g(X(\infty))$, of $\Delta g(X(t))$ as $t \to \infty$ has all positive eigenvalues, it is natural to expect that $\int_0^\infty \Delta g(X(s))ds$ 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 $X(t)$, $x_\delta(t)$, $X^n(t)$, $x_\delta^n(t)$, $V(t)$, and $V^n(t)$ on $t \in [0, \infty]$ and derive the limits of $V^n(t)$ and $\sqrt{n}[x_\delta^n(t) - X(t)]$ on $[0, \infty]$ in Theorem 3.2. As $\delta \to 0$ and $n \to \infty$, the limiting distributions of $V^n(t) = \sqrt{n}[X^n(t) - X(t)]$ and $\sqrt{n}[x_\delta^n(t) - X(t)]$ are $V(t)$ for $t \in [0, \infty]$, where $(V^n(t), V(t))$ describe the dynamic evolution of gradient descent algorithms for $t \in [0, \infty)$ and the statistical distribution of $\sqrt{n}(\hat{\theta}_n - \check{\theta})$ for $t = \infty$. The joint asymptotic analysis provides a unified framework to describe distribution limits of $X^n(t)$ and $x_\delta^n(t)$ from both computation and statistical points of view as follows. For $t \in [0, \infty)$, $X(t)$ and $V(t)$ gives the limiting behaviors of $X^n(t)$ and $x_\delta^n(t)$ corresponding to computational algorithms, and $X(\infty)$ and $V(\infty)$ illustrate their limiting behaviors of the corresponding statistical decision rule $\hat{\theta}_n$ (or the exact solutions of the corresponding optimization problems (1) and (7) that the algorithms are designed to compute). We use the following simple example to explicitly illus-

trate 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 $N(\theta_1, \tau^2)$ 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 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 $\boldsymbol{\sigma}^2(\theta) = Var(U_1 - \theta) = \text{diag}(\tau^2, \check{\theta}_2)$, where $\bar{U}_n = (\bar{U}_{1n}, \bar{U}_{2n})'$ is the sample mean. It is easy to see that the minimization problems corresponding to (1) and (7) has 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), (8), (4) and (10) yield recursive formulas $x_k = x_{k-1} + \delta(\check{\theta} - x_{k-1})$, and $x_k^n = x_{k-1}^n + \delta(\bar{U}_n - x_{k-1}^n)$ for the plain gradient descent case; and $x_k = x_{k-1} + \delta(\check{\theta} - y_{k-1})$, $y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$, $x_k^n = x_{k-1}^n + \delta(\bar{U}_n - y_{k-1}^n)$, $y_k^n = x_k^n + \frac{k-1}{k+2}(x_k^n - x_{k-1}^n)$ 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 will 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. Ordinary differential equations (3) and (9) admit simple solutions

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

$$V^n(t) = (V_1^n(t), V_2^n(t))' = \sqrt{n}[(\bar{U}_n - \check{\theta}) - (x_0^n - x_0)][1 - e^{-t}].$$

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 3.1, 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 (12) in this case. Then for $t \in [0, \infty)$,

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

which confirms that $V^n(t)$ converges to $V(t)$, as shown in Theorem 3.1. Furthermore, 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 $V^n(t) \to V^n(\infty) \sim V(\infty) = \Pi(\infty)\mathbf{Z} = (\tau Z_1, \check{\theta}_2 Z_2)'$, which gives the asymptotic distribution of 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 gradient descent algorithms when applied to solve stochastic optimization problems. For example, as

functions of t, $X(t)$ and $X^n(t)$ can be used to describe how the generated sequences from the algorithms evolves 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; and their relationship with n and t can be used to investigate the joint dynamic effect of data size and algorithm iterations on the computation and statistical errors in the sequences generated by the algorithms. The key signature in this exmaple 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 (6) admits an expression via the Bessel function,

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

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 $u \to \infty$

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

Ordinary differential equation (11) has the solution

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

As in Theorem 3.1, 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 (13) in this case. Then for $t \in [0, \infty)$,

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

The result matches the weak convergence of $V^n(t)$ to $V(t)$ shown in Theorem 3.1, 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 $V^n(t) \to V^n(\infty) \sim V(\infty) = \Pi(\infty)\mathbf{Z} = (\tau Z_1, \check{\theta}_2 Z_2)'$, which indicates the asymptotic distribution of 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 generated sequences from the algorithm evolve along iterations (via $X(t)$ and $X^n(t)$ as functions of t), and converge to the target optimization solutions (or estimators) (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 $n \to \infty$). The major difference for 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 3.4. *Solving problems with large scale data often require some tradeoffs between statistical efficiency and computational efficiency, and thus need to handle 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 gradient descent algorithms for solving (7) based on sampled data, and $\check{\theta}$ is the exact solution of (1) corresponding to the true value of θ, with $\hat{\theta}_n$ the exact solution of (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 to the solution $\check{\theta}$ of optimization problem (1), and in fact numerically $\check{\theta}$ can be only evaluated by $X(t)$ and its corresponding algorithms, using $X(t)$ as a proxy of $\check{\theta}$ we may treat $x_\delta^n(t) - X(t)$ as a surrogate of the total error, and asymptotic differential equations for its asymptotic distribution $V(t)$ may be useful for the analysis of the total error.

4 Stochastic differential equations for stochastic gradient descent

Solving (7) by algorithms (8) and (10) requires evaluating the sum-gradient for all training data, that is, it requires expensive evaluations of the gradients $\nabla \ell(\theta; U_i)$ from summand functions $\ell(\theta; U_i)$ with all (training or sample) data U_i, $i = 1, \cdots, n$.

For big data problems, data are enormous, such evaluation of the sums of gradients for all data becomes prohibitively expensive. To overcome the computational burden, stochastic gradient descent uses a so-called mini-batch of data to evaluate a 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 corresponds to the m out of n (with or without replacement) bootstraps for gradients (Bickel et al. (1997)). 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 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 of with replacement, $U_1^*, \cdots, U_m^*$ are an i.i.d. sample taken from $\hat{Q}_n$, and $\hat{Q}_n$ is the empirical distribution of $U_1, \cdots, U_n$. The main computational idea in the algorithm is to replace $\mathcal{L}^n(\theta; \mathbf{U}_n)$ in (8) and (10) by a smaller sample version $\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*)$ at each iteration, where

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

4.1 Stochastic gradient descent

The stochastic gradient descent scheme replaces $\nabla \mathcal{L}^n(x_{k-1}^n; \mathbf{U}_n)$ in (8) by a smaller sample version at each iteration to obtain

$$x_k^m = x_{k-1}^m - \delta \nabla \hat{\mathcal{L}}^m(x_{k-1}^m; \mathbf{U}_{mk}^*), \tag{18}$$

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 (18) [and (26) for the accelerated case in Section 4.2 later] are designed for the computational purpose, they do not correspond to any optimization problem with a well-defined objective function like $g(\theta)$ in (1) or $\mathcal{L}^n(\theta; \mathbf{U}_n)$ in (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 quite different from these studied in Section 3. Here we consider the stochastic gradient descent case, and may define a 'pseudo objective function' as follows.

Define a mini-batch process $\mathbf{U}_m^*(t) = (U_1^*(t), \cdots, U_p^*(t))'$ and a step process $x_\delta^m(t)$, $t \geq 0$, for x_k^m in (18) as follows,

$$\mathbf{U}_m^*(t) = \mathbf{U}_{km}^* \text{ and } x_\delta^m(t) = x_k^m \text{ for } (k-1)\delta < t \leq k\delta. \tag{19}$$

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 $(k-1)\delta < t \leq k\delta$. $\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*(t))$ may be treated as a counterpart of $\mathcal{L}^n(\theta; \mathbf{U}_n)$. As $m, n \to \infty$, $\hat{\mathcal{L}}^m(\theta; \mathbf{U}_m^*(t))$ approaches to $g(\theta)$ for each δ, and the stochastic gradient descent algorithm (18) can still solve the optimization problem (7) numerically. But 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, n), since mini-batches change as the algorithm iterates, and the number of the mini-batches involved is determined by time t and step size δ. There is no single bona fide objective function here, and the 'pseudo objective function' $\mathcal{L}^m(\theta; \mathbf{U}_m^*(t))$ can't serve the role of genuine objective functions like $g(\theta)$ and $\mathcal{L}^n(\theta; \mathbf{U}_n)$. The approach in Section 2 and Section 3 can not be directly applied to obtain an ordinary differential equation like (9). In fact as we will see below, there exists no such analog ordinary differential equation. Instead we will derive asymptotic stochastic differential equations for (18). The new asymptotic stochastic differential equation may be considered as a counterpart to an asymptotic version (14) corresponding to (9), but the key difference is that the asymptotic stochastic differential equations must depend on step size δ as well as m to account for the mini-batch effect (see more detail later after equations (21) and (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 $m, n \to \infty$.

We need the following usual sample size condition to guarantee the validity of mini-batch subsampling.

A5. Mini-batch size m satisfies that as $n \to \infty$, we choose $m \to \infty$ and $m/n \to 0$. Also assume 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 4.1. *Define a partial sum process*

$$H_\delta^m(t) = (m\delta)^{1/2} \sum_{t_k \leq t} \left[\nabla\hat{\mathcal{L}}^m(x_\delta^m(t_{k-1}); \mathbf{U}_m^*(t_k)) - \nabla g(X(t_k)) \right], \ t \geq 0, \tag{20}$$

where $t_k = k\delta$, $k = 0, 1, 2, \cdots$. Under Conditions A1-A5, as $\delta \to 0$ and $m, n \to \infty$, 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 Condition A3, and* $X(t)$ *is the solution of (3).*

Remark 4.1. *As we have discussed early, due to mini-batches used in algorithm (18), there is no corresponding optimization problem with a well-defined objective function. As a result, we do not have any* δ*-free differential equation analog to (14). 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 4.1 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 (20) is needed 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 (14) we need to find some kind of continuous limit for* $\nabla \mathcal{L}^m(\theta; \mathbf{U}_m^*(t)) - \nabla g(\theta)$*. As* $\mathbf{U}_m^*(t)$ *is an empirical process for (conditionally) independent subsamples* $\mathbf{U}_{mk}^*$*, thus* $\nabla \mathcal{L}^m(\theta; \mathbf{U}_m^*(t)) - \nabla g(\theta)$ *may behave like a sort of discrete-time weighted white noise (in fact a martingale difference sequence). Therefore, a possible continuous 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 4.1, 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 (20) is needed to normalize a partial sum process with* $[T/\delta]$ *random variables for obtaining a weak convergence limit, from the white noise point of view, here we need a normalized factor* $\delta^{1/2}$ *to move from a discrete white noise to a continuous white noise. As a matter of fact, the weak convergence is very natural from the viewpoint of limit theorems 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 (18), the continuous modeling for stochastic gradient descent will be* δ*-dependent, which will be given below.*

We recast algorithm (18) as

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

We approximate $x_\delta^m(t)$ by a continuous process $X_\delta^m(t)$, and Theorem 4.1 suggests an approximation of $H_\delta^m(t)$ by continuous process $H(t)$. We take step size δ as dt,

and above difference equation becomes

$$\frac{dX_\delta^m(t)}{dt} = -\nabla g(X_\delta^m(t)) - (\delta/m)^{1/2}\,\frac{dH(t)}{dt},$$

which leads to the following stochastic differential equation

$$dX_\delta^m(t) = -\nabla g(X_\delta^m(t))dt - (\delta/m)^{1/2}\boldsymbol{\sigma}(X(t))d\mathbf{B}(t), \tag{21}$$

where $X(t)$ is the solution of (3), and $\mathbf{B}(t)$ is a p-dimensional standard Brownian motion. The solution $X_\delta^m(t)$ of (21) may be considered as a continuous approximation of x_k^m [or $x_\delta^m(t)$] generated from the stochastic gradient descent algorithm (18) [or (19)]. Since $X_\delta^m(t)$ is expected to be close to $X(t)$, and the Brownian term in (21) is of higher order, we may replace $X(t)$ in (21) by $X_\delta^m(t)$ to better mimic the recursive relationship in (18). That is, we consider the following stochastic differential equation

$$d\check{X}_\delta^m(t) = -\nabla g(\check{X}_\delta^m(t))dt - (\delta/m)^{1/2}\boldsymbol{\sigma}(\check{X}_\delta^m(t))d\mathbf{B}(t). \tag{22}$$

As our interest is on their distributional behaviors, we consider solutions of (21) and (22) in the weak sense that for each fixed δ and m, there exist continuous process $X_\delta^m(t)$ (or $\check{X}_\delta^m(t)$) and Brownian motion $\mathbf{B}(t)$ on some probability space to satisfy equation (21) (or (22)) (see Ikeda and Watanabe (1981)). Some versions of stochastic differential equations (such as vague or approximate matrices for the diffusion variance) are informally used in the deep learning and stochastic gradient descent literature based on some heuristic or loose reasoning without rigorous justification (see Chen et al. (2016), Li et al. (2015), Mandt et al. (2016), Sirignano and Spiliopoulos (2017)). As we will see, this is the first paper to provide explicit stochastic differential equations and establish rigorous weak convergence for stochastic gradient descent algorithms.

The stochastic Brownian terms in (21) and (22) are employed to account for the random fluctuations due to the use of min-batches for gradient estimation from iteration to iteration in the stochastic gradient descent algorithm (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 step size). At each iteration we resort to a mini-batch for gradient estimation, so 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 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 $\check{V}_\delta^m(t) = (m/\delta)^{1/2}[\check{X}_\delta^m(t) - X(t)]$. Treat

them as random elements in $C([0,T])$, we derive their weak convergence limit in the following theorem.

Theorem 4.2. *Under Conditions A1-A5, as $\delta \to 0$ and $m \to \infty$, we have*

$$\max_{0\le t\le T} |X_\delta^m(t) - \check{X}_\delta^m(t)| = O_P(m^{-1}\delta), \tag{23}$$

and both $V_\delta^m(t)$ and $\check{V}_\delta^m(t)$, $t \in [0,T]$, weakly converge to $V(t)$ satisfying

$$dV(t) = -[\Delta g(X(t))]V(t)dt - \boldsymbol{\sigma}(X(t))d\mathbf{B}(t), \;\; V(0) = 0, \tag{24}$$

where $X(t)$ is the solution of (3).

Remark 4.2. *Theorem 4.2 shows 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 seen from the fact that the difference between stochastic differential equations (21) and (22) is at the high order Brownian term with $X_\delta^m(t)$ replaced by its limit $X(t)$. Linear stochastic differential equation (24) indicates that the limiting process $V(t)$ is a time-dependent Ornstein-Uhlenbeck process with an explicit expression for $t \in [0,T]$,*

$$V(t) = -\int_0^t \exp\left[-\int_u^t \Delta g(X(v))dv\right] \boldsymbol{\sigma}(X(u))d\mathbf{B}(u). \tag{25}$$

Step process $x_\delta^m(t)$ in (19) is the empirical process for x_k^m generated from the stochastic gradient descent algorithm (18). Treating $x_\delta^m(t)$ as a random element in $D([0,T])$ we consider its asymptotic distribution in the follow theorem.

Theorem 4.3. *Under assumption A1-A5, as $\delta \to 0$, and $m, n \to \infty$, we have*

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

where $x_\delta^m(t)$ and $X_\delta^m(t)$ are defined by (19) and (21), respectively. In particular if we choose (δ, m, n) such that as $\delta \to 0$ and $m, n \to \infty$, $m(n\delta)^{-1/2} \to 0$, and $m^{1/2}\delta|\log\delta|^{1/2} \to 0$, then $(m/\delta)^{1/2}[x_\delta^m(t) - X(t)]$ weakly converges to $V(t)$, governed by stochastic differential equation (24).

Remark 4.3. *Theorem 4.3 indicates that sequences x_k^m generated from the stochastic gradient descent algorithm (18) can be very close to the continuous curves $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ governed by (21) and (22), respectively, and with proper choices of (δ, m, n) 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 $\check{X}_\delta^m(t)$.*

Remark 4.4. *We may consider stochastic gradient descent with momentum and/or diminishing learn rate and obtain the corresponding stochastic differential equations. For example, δ in (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 will lead us to stochastic differential equations like (21) and (22) with extra factor $(t+1)^{-\alpha}$,*

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

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

For the momentum case, we need to add an extra linear term in $X_\delta^m(t)$ (or $\check{X}_\delta^m(t)$) in the drifts.

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 (10) with a subsampled version at each iteration as follows,

$$x_k^m = y_{k-1}^m - \delta\nabla\hat{\mathcal{L}}^m(y_{k-1}^m; \mathbf{U}_{mk}^*), \qquad y_k^m = x_k^m + \frac{k-1}{k+2}(x_k^m - x_{k-1}^m), \tag{26}$$

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 (26) is conceptually in parallel with the case for the stochastic gradient descent algorithm (18) in Section 4.1, but the tricky part is on the technical side that we face many mathematical difficulties in multiple steps related to singularity in the second order stochastic differential equations involved. As we illustrate the continuous modeling of x_k^m generated from algorithm (26), it is easy to see 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 $m, n \to \infty$. Similar to the cases in Section 2 and Section 3.2, we define step function

$$x_\delta^m(t) = x_k^m, \;\; y_\delta^m(t) = y_k^m, \;\; \text{for } \; (k-1)\sqrt{\delta} < t \le k\sqrt{\delta}, \tag{27}$$

and approximate $x_\delta^m(t)$ by a smooth curve $X_\delta^m(t)$ given by (35) below. Note the step size difference that the step size is δ and $\delta^{1/2}$ for the plain and accelerated cases, respectively, as pointed out at the end of Section 2.

Theorem 4.4. *Define a partial sum process*

$$H_\delta^m(t) = (m^2\delta)^{1/4} \sum_{t_k \le t} \left[\nabla\hat{L}^m(y_\delta^m(t_{k-1}); \mathbf{U}_m^*(t_k)) - \nabla g(X(t_k))\right], \; t \ge 0, \tag{28}$$

where $t_k = k\delta^{1/2}$, $k = 0, 1, 2, \cdots$. Under Conditions A1-A5, as $\delta \to 0$ and $m, n \to \infty$, 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 Condition A3, and $X(t)$ is the solution of (3).

It is not obvious that we can directly adopt the simple arguments in Sections 2 and 4.1 to derive the second order stochastic differential equation corresponding to algorithm (26). We illustrate an alternative approach instead. First, note that the second order ordinary differential equation (6) can be equivalently written as

$$\begin{cases} dX(t) = Z(t)dt, \\ dZ(t) = -\left[\frac{3}{t} Z(t) + \nabla g(X(t))\right] dt, \end{cases} \tag{29}$$

where $Z(t) = \dot{X}(t)$; and algorithm (5) is equivalent to

$$\begin{cases} x_{k+1} = x_k + \sqrt{\delta}\, z_k, \\ z_{k+1} = \left[1 - \frac{3}{k+3}\right] z_k - \sqrt{\delta}\, \nabla g \left(x_k + \frac{2k+3}{k+3}\sqrt{\delta}\, z_k\right), \end{cases} \tag{30}$$

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

$$\begin{cases} \frac{x_{k+1} - x_k}{\sqrt{\delta}} = z_k, \\ \frac{z_{k+1} - z_k}{\sqrt{\delta}} = -\frac{3}{t_k + 3\sqrt{\delta}} z_k - \nabla g \left(x_k + \frac{2k+3}{k+3}\sqrt{\delta}\, z_k\right), \end{cases} \tag{31}$$

where we take $t_k = k\sqrt{\delta}$. We approximate (x_k, z_k) by continuous curves $(X(t), Z(t))$. Noting that as $\delta \to 0$, $3\sqrt{\delta} \to 0$ and $\frac{2k+3}{k+3}\sqrt{\delta}\, z_k \to 0$ in (31), which are negligible relative to t_k and x_k. We take step size $\sqrt{\delta}$ as dt and turn discrete difference equations (31) into continuous differential equations (29).
Second, we replace (x_k, z_k) in (30) by (x_k^m, z_k^m), where $z_k^m = (x_{k+1}^m - x_k^m)/\sqrt{\delta}$, and write (26) in the following equivalent forms

$$\begin{cases} x_{k+1}^m = x_k^m + \sqrt{\delta}\, z_k^m, \\ z_{k+1}^m = \left[1 - \frac{3}{k+3}\right] z_k^m - \sqrt{\delta}\, \nabla g \left(x_k^m + \frac{2k+3}{k+3}\sqrt{\delta}\, z_k^m\right) - \frac{\delta^{1/4}}{\sqrt{m}}[H_\delta^m(t_k) - H_\delta^m(t_k)], \end{cases} \tag{32}$$

or equivalently,

$$\begin{cases} \frac{x_{k+1}^m - x_k^m}{\sqrt{\delta}} = z_k^m, \\ \frac{z_{k+1}^m - z_k^m}{\sqrt{\delta}} = -\frac{3}{t_k + 3\sqrt{\delta}} z_k^m - \nabla g \left(x_k^m + \frac{2k+3}{k+3}\sqrt{\delta}\, z_k^m\right) - \frac{\delta^{1/4}}{\sqrt{m}} \frac{H_\delta^m(t_k) - H_\delta^m(t_k)}{\sqrt{\delta}}, \end{cases} \tag{33}$$

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 4.4 suggests to 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 $\frac{2k+3}{k+3}\sqrt{\delta}\, z_k$, and taking step size $\sqrt{\delta}$ as dt we move from discrete difference equations (33) to the following stochastic differential equation system,

$$\begin{cases} dX_\delta^m(t) = Z_\delta^m(t)dt, \\ dZ_\delta^m(t) = -\left[\frac{3}{t}Z_\delta^m(t) + \nabla g(X_\delta^m(t))\right]dt - \frac{\delta^{1/4}}{\sqrt{m}}\boldsymbol{\sigma}(X(t))d\mathbf{B}(t), \end{cases} \tag{34}$$

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

$$\ddot{X}_\delta^m(t) + \frac{3}{t}\dot{X}_\delta^m(t) + \nabla g(X_\delta^m(t)) + (\delta/m^2)^{1/4}\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0, \tag{35}$$

where initial conditions $X_\delta^m(0) = x_0^m$ and $\dot{X}_\delta^m(0) = 0$, $X(t)$ is defined by (6), $\mathbf{B}(t)$ is a p-dimensional Brownian motion, and white noise $\dot{\mathbf{B}}(t)$ is the derivative of $\mathbf{B}(t)$ in the sense of generalized functions.

As we have discussed and demonstrated for the stochastic gradient descent case in Section 4.1, similar to the stochastic differential equations (21) and (22) for the stochastic gradient descent algorithm, the second order stochastic differential equations (34) and (35) depend on δ and m through the stochastic Brownian terms. They are used to account for the random fluctuation due to the use of min-batches for gradient estimation from iteration to iteration in algorithm (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 will show that the second order stochastic differential equation (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 (35). As in Section 4.1, process $X_\delta^m(t)$ provides a continuous approximation of x_k^m given by (26). As $\delta \to 0$ and $m \to \infty$, the Brownian term in (35) disappears, and $X_\delta^m(t)$ approaches to $X(t)$ defined by (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 $C([0, T])$, in the following theorem we derive a weak convergence limit of $V_\delta^m(t)$.

Theorem 4.5. *Under conditions A1-A5, the second order stochastic differential equation (35) has a unique solution in the weak sense, and as $\delta \to 0$, $m \to \infty$,*

$V_\delta^m(t)$ weakly converges to $V(t)$ on $C([0,T])$, where $V(t)$ is the unique solution of the following linear second order stochastic differential equation,

$$\ddot{V}(t) + \frac{3}{t}\dot{V}(t) + [\Delta g(X(t))]V(t) + \boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t) = 0, \tag{36}$$

where $\Delta = \nabla^2$ is Laplacian operator, $X(t)$ is a solution of equation (6), $\mathbf{B}(t)$ is a p-dimensional standard Brownian motion, and initial conditions $V(0) = \dot{V}(0) = 0$.

Remark 4.5. *As mentioned before, similar to the stochastic gradient descent case, the continuous modeling depends on both δ and m, and Theorems 4.4-4.5 are in parallel with Theorems 4.1-4.3. However, for the accelerated case, the challenges are largely on the technical proofs. For example, we need to handle second order stochastic differential equations like (35) with singularity (similar to the singularity case for ordinary differential equations (6) and (13)); the lack of adequate theory and technical tools for handling well-behaved second order stochastic differential equations, let alone the singularity difficulty; it is hard to analyze the complex recursive relationship in the accelerated stochastic gradient descent algorithm (26). Step process $x_\delta^m(t)$ in (27) is the empirical process for x_k^m generated from algorithm (26). Treating $x_\delta^m(t)$ as a random element in $D([0,T])$ we conjecture that $x_\delta^m(t)$ and $X_\delta^m(t)$ share the same weak convergence limit $V(t)$.*

Below we study the example considered in Section 3.4 under the stochastic gradient descent case.

Example 1(continue). With $\nabla g(\theta) = \theta - \check{\theta}$, $\Delta g(\theta) = 1$, $\boldsymbol{\sigma}(\theta) = \mathrm{diag}(\tau, \check{\theta}_2)$, we have for the plain case, $X(t) = \check{\theta} + (x_0 - \check{\theta})e^{-t}$,

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

where $V(t)$ is the Ornstein-Uhlenbeck process as the solution of (24). It is easy to see 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 (6) has the form

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

Below we will give solutions of stochastic differential equations (35) and (36) in this case. First we consider the solution $V(t)$ of stochastic differential equation (36). It is easy to check 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 of the first and second kind of order one, respectively. Since in this case, ∇g is linear, $\Delta g = 1$, and stochastic differential equations (35) and (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 condition in A5, it is clear 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 algorithms (18) and (26), and the weak convergence results established in Theorems 4.1-4.5 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 $t \to \infty$, 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 4.6. *Suppose that the assumptions A1-A5 are met, $\Delta g(\check{\theta})$ is positive definite, all eigenvalues of $\int_0^t \Delta g(X(s))ds$ diverge as $t \to \infty$, and assume $m(n\delta)^{-1/2} \to 0$ and $m^{1/2}\delta|\log\delta|^{1/2} \to 0$, as $n \to \infty$, $m, n \to \infty$. We obtain the following results.*

(i) As $\delta \to 0$, and $m, n \to \infty$, $V_\delta^m(t) = (m/\delta)^{1/2}[X_\delta^m(t) - X(t)]$ and $(m/\delta)^{1/2}[x_\delta^m(t) - X(t)]$ weakly converge to $V(t)$ on $D(\mathbb{R}_+)$.
(ii) Stochastic differential equation (24) admits a unique stationary distribution denoted by $V(\infty)$, where $V(\infty)$ follwos a normal distribution with mean zero and covariance matrix $\Gamma(\infty)$ satisfying

$$\Gamma(\infty)\Delta g(X(\infty)) + \Delta g(X(\infty))\Gamma(\infty) = \boldsymbol{\sigma}(X(\infty))\boldsymbol{\sigma}(X(\infty))'. \tag{37}$$

(iii) Further assume that there exists a unique stationary distribution, denoted by $X_\delta^m(\infty)$, for stochastic differential equation (21). Then as $\delta \to 0$, and $m, n \to \infty$, $V_\delta^m(\infty) = (m/\delta)^{1/2}[X_\delta^m(\infty) - X(\infty)]$ converges in distribution to $V(\infty)$.

Remark 4.6. *Similar to Theorem 3.3 and Remark 3.3, Theorem 4.6 indicates that for the stochastic gradient descent case, as $\delta \to 0$, $m, n \to \infty$, $X_\delta^m(\infty)$ approach $X(\infty) = \check{\theta}$, $V_\delta^m(t) = \sqrt{m/\delta}[X_\delta^m(t) - X(t)]$ 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 (21) may admit a stationary distribution $X_\delta^m(\infty)$ as the limiting distribution of $X_\delta^m(t)$ when $t \to \infty$ (see Da Prato and Zabczyk (1996) and Gardiner (2009) for the existence of stationary distributions). Naturally $X_\delta^m(\infty)$ corresponds to $V(\infty)$. Mandt et al. (2017) essentially takes these results as its major model assumptions for treating stochastic gradient descent as a statistical estimation procedure in the Bayesian framework.*

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 at each iteration. It is not clear whether we have known statistical estimation methods corresponding 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 Example 1 considered in Sections 3.4 and 4.2.
Example 1(continue). First we evaluate

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

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 $t \to \infty$. Solving

stochastic differential equation (21) we have

$$X_\delta^m(t) = x_0^m e^{-t} + \check{\theta}(1-e^{-t}) - \sqrt{\tfrac{\delta}{m}} \int_0^t e^{u-t} \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u)$$
$$= x_0^m e^{-t} + \check{\theta}(1-e^{-t}) + \sqrt{\tfrac{\delta}{m}} \Lambda(t), \tag{38}$$

where $\Lambda(t) = -(\int_0^t e^{u-t} dB_2(u), \int_0^t e^{u-t} dB_2(u))$ is the Ornstein-Uhlenbeck process whose stationary distribution is a bivariate normal distribution with mean zero and variance equal to the half of identity matrix. 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 (38) we conclude that as $t \to \infty$, $X_\delta^m(t)$ converges in distribution to $X_\delta^m(\infty) = \check{\theta} + (\delta/m)^{1/2} \mathrm{diag}(\tau, \check{\theta}_2)\mathbf{Z}/\sqrt{2}$. If the initial values satisfy $x_0^m - x_0 = o((\delta/m)^{1/2})$, then $V_\delta^m(t)$ 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, algorithm (18) gives

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

where we consider the case that mini-batches $\mathbf{U}_{mk}^* = (U_{1k}^*, \cdots, U_{mk}^*)$, $k \geq 1$, are sampled from large training data set $\mathbf{U}_n$, and $\bar{U}_{mk}^*$ is the bootstrap 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 stochastic optimization (7) based on all data, and $x_k = x_{k-1} + \delta(\check{\theta} - x_{k-1})$ for optimization (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 $x_k^m = (1-\delta)x_{k-1}^m + \delta \bar{U}_{mk}^*$, and obtain

$$x_\delta^m(t) = x_0^m(1-\delta)^{[t/\delta]} + \delta \sum_{k\delta \leq t} (1-\delta)^{k-1} \bar{U}_{mk}^*. \tag{39}$$

Similarly, we have

$$x_\delta^n(t) = x_0^n(1-\delta)^{[t/\delta]} + \bar{U}_n \delta \sum_{k\delta \leq t} (1-\delta)^{k-1}, \quad x_\delta(t) = x_0(1-\delta)^{[t/\delta]} + \check{\theta}\delta \sum_{k\delta \leq t} (1-\delta)^{k-1}.$$

Letting $t \to \infty$, we get

$$x_\delta^m(\infty) = \delta \sum_{k=1}^{\infty} (1-\delta)^{k-1} \bar{U}_{mk}^*, \quad x_\delta^n(\infty) = \bar{U}_n \delta \sum_{k=1}^{\infty} (1-\delta)^{k-1} = \bar{U}_n, \quad x_\delta(\infty) = \check{\theta},$$

where we can clearly see 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, with enormous n and relatively large m, $x^m_\delta(t)$ can be naturally approximated by its 'limit' $x^m_0 e^{-t} + \check{\theta}(1 - e^{-t}) - (\delta/m)^{1/2} \int_0^t e^{u-t} \boldsymbol{\sigma}(X(u)) d\mathbf{B}(u)$, which is equal to $X^m_\delta(t)$, where the last term of the right hand side of (39) after centered with $\check{\theta}$ 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

$$x^n_\delta(t) = x_\delta(t) + (\bar{U}_n - \check{\theta}) \left[1 - (1 - \delta)^{[t/\delta]+1}\right], \tag{40}$$

$$\begin{aligned} x^m_\delta(t) &= x^n_\delta(t) + \delta \sum_{k\delta \le t} (1 - \delta)^{k-1} (\bar{U}^*_{mk} - \bar{U}_n) \\ &= x_\delta(t) + (\bar{U}_n - \check{\theta}) \left[1 - (1 - \delta)^{[t/\delta]+1}\right] + \delta \sum_{k\delta \le t} (1 - \delta)^{k-1} (\bar{U}^*_{mk} - \bar{U}_n). \end{aligned} \tag{41}$$

The KMT strong approximation (Komlós et al. (1975, 1976)) and its related bootstrap strong approximation (Csörgö et al. (1999) and Csörgö and Mason (1989)) lead to

$$\bar{U}^*_{mk} - \bar{U}_n = m^{-1/2} A_{mk} + O_P(m^{-1} \log m), \;\; \bar{U}_n - \check{\theta} = n^{-1/2} D_n + O_P(n^{-1} \log n), \tag{42}$$

where A_{mk}, $k = 1, 2, \cdots$, are nearly 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. The second and third terms on the right hand side of (41) account for, respectively, the variability due to statistical estimation and the random fluctuation due to the use of min-batches (or bootstrap samples) for gradient estimation from iteration to iteration in the stochastic gradient descent algorithm. From (42) and $m/n \to 0$, we easily conclude that the second term on the right hand side of (41) is of order higher than the third term, where the the third term represents the cumulative min-batch-subsampling (or bootstrapping) effect up to the $k = [t/\delta]$-th iteration, with the second term for the statistical estimation error. (40) and (41) show that as $m, n \to \infty$, $x^n_\delta(t)$ and $x^m_\delta(t)$ approach $\check{\theta}$, and on average both gradient descent and stochastic gradient descent algorithms stay on target, the difference is their random variabilities. Theorems 3.1 and 3.2 establish an order of $n^{-1/2}\mathbf{Z}$ for the random variability of the gradient descent algorithm using all data, while Theorems 4.2 and 4.3 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 stochastic differential equation (24) (or its expression (25)) is a time-dependent Ornstein-Uhlenbeck process that may admit a stationary distribution with mean zero and variance $\boldsymbol{\sigma}(X(\infty))/[2\Delta g(X(\infty))]$, factor $m^{-1/2}$ accounts for the effect of each mini-batch (or bootstrap sample) of size m, and factor $\delta^{1/2}$ represents the effect of the total number of mini-batches (or bootstrap samples) that is proportional to $1/\delta$. The normalized factor $(\delta/m)^{1/2}$ means that while each mini-batch (or bootstrap sample) of size m is not as efficient as full data sample of size n, but repetitive use of min-batch subsampling (or bootstrapping) 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 min-batches (or bootstrap samples) 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 (or bootstrap sample) of small size at each iteration. Intuitively, it means that the stochastic gradient descent algorithm invokes different min-batches (or bootstrap samples) resulted some random fluctuation when moving from one iteration to another, and as the number of iterations increases, subsampling improves efficiency with factor $(\delta/m)^{1/2}$ instead of $m^{-1/2}$, to make up loss from $n^{-1/2}$ to $m^{-1/2}$, that is, updating with the use of many mini-batches (or bootstrap samples) 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 some implication 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 minimizers (Keskar et al. (2017)). We will provide some rigorous analysis and heuristic intuition to shed some light on the phenomenon. First not that Theorems 4.1-4.3 do not need convexity assumption on the objective function $g(\theta)$, and Theorem 4.6 can be easily adopted to non-convex optimization with $\check{\theta}$ being a critical point of $g(\theta)$. Suppose that stochastic gradient descent processes converge to the critical point $\check{\theta}$. Applying large deviation theory to stochastic differential equations (21) and (22) corresponding to the gradient descent algorithm, we obtain 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 (proportional to $(m/\delta)^{1/2}\log(m/\delta)$) (see Kifer (1981) and Li et al. (2017)); while, if the critical point is a local mini-

mizer of $g(\theta)$, the continuous processes will take an exponential time (proportional to $\exp\{c(m/\delta)^{1/2}\}$ for some generic constant c) to get out a neighborhood of the local minimizer (see Dembo and Zeitouni (2010) and Li et al. (2017)). We may also explain the phenomenon from the limiting distribution point of view. Theorem 4.2 indicates that continuous processes $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ generated from stochastic differential equations (21) and (22) are asymptotically the same as the deterministic solution $X(t)$ of ordinary differential equation (6) plus $(\delta/m)^{1/2}V(t)$, where $V(t)$ is the solution of stochastic differential equation (24). The limiting process $V(t)$ is a time-dependent Ornstein-Uhlenbeck process with explicit expression given by (25). We have 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 4.7. *Suppose that the assumptions A1-A5 (except for the convexity of $g(\cdot)$) are met, and gradient descent process $X(t)$ given by (3) converges to a critical point $\check{\theta}$ of $g(\cdot)$. Then we have the following results.*

$$g(X_\delta^m(t)) = g(X(t)) + (\delta/m)^{1/2}\nabla g(X(t))V(t) + \frac{\delta}{2m}\Delta g(X(t))[V(t)]^2 + o_P(\delta/m)$$
$$= g(\check{\theta}) + \frac{1}{2}\Delta g(\check{\theta})\left\{[X(t)-\check{\theta}]^2 + 2(\delta/m)^{1/2}[X(t)-\check{\theta}]V(t) + \delta[V(t)]^2/m\right\} + o_P(\delta/m),$$

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 solutions of differential equations (3), (21) and (22), respectively, and the equalities hold in the sense that we may consider $X_\delta^m(t)$ (or $\check{X}_\delta^m(t)$) and $V(t)$ on some common probability spaces through Skorokhod's representation.
For local minimizer $\check{\theta}$ with positive definite $\Delta g(\check{\theta})$, as $t \to \infty$, $V(t)$ has a limiting stationary distribution with mean zero and covariance matrix $\Gamma(\infty)$ satisfying (37) with $X(\infty) = \check{\theta}$. For saddle point $\check{\theta}$, $V(t)$ does not have any limiting distribution.

Theorem 4.7 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, and $V(t)$ plays a dominant role in the stochastic gradient descent algorithm. If the critical point $\check{\theta}$ is a saddle point of $g(\theta)$, $\Delta g(\cdot)$ is non-positive definite around the saddle point, and time-dependent Ornstein-Uhlenbeck process $V(t)$ does not have any stationary distribution, and 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 $\Delta g(\cdot)$ positive definite, around the local minimizer. Then $V(t)$ has a stationary distribution, and

all the processes maintain stable stochastic behaviors and tend to stay around the local minimizer. However, after further analyzing the behaviors of X_δ^m and $\check{X}_\delta^m$ around a local minimizer $\check{\theta}$, we find that stochastic gradient descent behaves quite differently depending on factor $(\delta/m)^{1/2}$ and the local geometry of $g(\cdot)$ around $\check{\theta}$. Keskar et al. (2017) considers two kinds of local minimizers: sharp and flat local minimizers. We characterize the sharpness of a local minimizer $\check{\theta}$ by the Hessian matrix $\Delta g(\check{\theta})$ and range index ρ such that the whole local minimizer well falls inside $\{\theta : |\theta - \check{\theta}| < \rho\}$, or equivalently, gradient descent process $X(t)$ will move away from the local minimizer $\check{\theta}$ if $X(t)$ starts outside $\{\theta : |\theta - \check{\theta}| < \rho\}$. We can easily see that the smaller ρ and larger $\Delta g(\check{\theta})$ are, the bigger $\nabla g(\cdot)$ and steeper $g(\cdot)$ are around $\check{\theta}$, while the larger ρ and smaller $\Delta g(\check{\theta})$ are, the smaller $\nabla g(\cdot)$ and flatter $g(\cdot)$ are around $\check{\theta}$. From Theorem 4.7, stochastic component $(\delta/m)^{1/2}V(t)$ play a key role in determining the behaviors of $g(X_\delta^m(t))$ and $g(\check{X}_\delta^m(t))$. First note that stochastic component decreases as the batch size increases. Second, for a local minimizer $\check{\theta}$ with a larger (or smaller) $\Delta g(\check{\theta})$, the corresponding stationary distribution of $V(t)$ has a smaller (or larger) variance $\Gamma(\infty)$ given by (37), and thus $V(t)$ tend to produce values of smaller (or larger) magnitude. Therefore, for a sharp local minimizer with given ρ, we need to choose small batch size m to yield a large enough stochastic component $(\delta/m)^{1/2}V(t)$ for stochastic gradient descent processes X_δ^m and $\check{X}_\delta^m$ to get out $\{\theta : |\theta - \check{\theta}| < \rho\}$, while large batch size m produces very small stochastic component $m^{-1/2}V(t)$, which tends to keep X_δ^m and $\check{X}_\delta^m$ inside $\{\theta : |\theta - \check{\theta}| < \rho\}$. This indicates that stochastic gradient descent with larger batch size has a tendency to stay around sharp minimizers, while stochastic gradient descent with smaller batch size can move away from sharp local minimizers and tends to settle around flat local minimizers.

Our findings are in consistent with empirical studies shown in Keskar et al. (2017) that stochastic gradient descent with small batch size often leads to flat local minimizers with good generalization errors, while stochastic gradient descent with large batch size tends to converge to sharp local minimizers. Our numerical results also confirm these findings and will be reported in the follow-up work.

5 An example

This section considers a simple example with some numerical study to illustrate the approximation of gradient descent algorithms by ordinary or stochastic differential equations.

Example 2. Consider the following simple linear regression model

$$U_{1i} = U_{2i}'\theta + \varepsilon_i, \quad i = 1, \cdots, n, \tag{43}$$

where U_{1i} and U_{2i} are response and covariate, respectively, parameter $\theta = (\theta_1, \theta_2)'$, random errors ε_i are i.i.d. normal random variables with mean zero and variance τ^2. We consider both fixed and random designs. For the random design case, we assume that U_{2i} and ε_i are independent, and U_{2i} are i.i.d. mean zero bivariate normal random vectors. For the fixed design case, we set U_{2i} to be deterministic instead of bivariate normal random variables, where observations U_{1i} are not i.i.d, and we need to make some obvious modification.

First consider the fixed design case. Denote by $\check{\theta}$ the true value of parameter θ in the regression model. Let $U_1 = (U_{11}, \cdots, U_{1n})'$, $U_2 = (U_{21}, \cdots, U_{2n})'$, and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)$. Assume that we have an orthogonal design so that $U_2'U_2/n$ is equal to the identity matrix. Then we may write regression model in a matrix form $U_1 = U_2\theta + \varepsilon$, and define $\mathcal{L}^n(\theta; \mathbf{U}_n) = (U_1 - U_2\theta)'(U_1 - U_2\theta)/(2n)$, and $g(\theta) = E[\mathcal{L}^n(\theta; \mathbf{U}_n)]$. Simple calculations show $g(\theta) = [(\theta - \check{\theta})'(\theta - \check{\theta}) + \tau^2]/2$, $\nabla g(\theta) = \theta - \check{\theta}$, $\nabla\mathcal{L}^n(\theta; \mathbf{U}_n) = -U_2'(U_1 - U_2\theta)/n = \nabla g(\theta) - n^{-1}\tau\mathbf{Z}$, where $n^{-1/2}U_2'\varepsilon/\tau$ follows a standard bivariate normal distribution, and we denote it by $\mathbf{Z}$.

It is easy to see that the minimization problems corresponding to (1) and (7) have explicit solutions in this case: $g(\theta)$ has the minimizer being the true parameter value $\check{\theta}$, and $\mathcal{L}^n(\theta; \mathbf{U}_n)$ has the minimizer equal to the least squares estimator $\hat{\theta}_n$.

The differential equations (9) and (14) [or (11) and (15)] are identical in this case, and as $n \to \infty$, their limits are given by (3) and (6). Specially, the differential equations admit solutions with the following expressions

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

for the plain gradient descent case. For the accelerated case,

$$X^n(t) = \begin{pmatrix} X_1^n(t) \\ X_2^n(t) \end{pmatrix} = \left(\check{\theta} + n^{-1/2}\tau\mathbf{Z} + \frac{2(x_0^n - \check{\theta} - n^{-1/2}\tau\mathbf{Z})}{t}J_1(t)\right)_{i=1,2},$$

$$X(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix} = \left(\check{\theta} + \frac{2(x_0 - \check{\theta})}{t}J_1(t)\right)_{i=1,2},$$

where x_0^n and x_0 are initial values of $X^n(t)$ and $X(t)$, respectively, and $J_1(u)$ is the Bessel function of the first kind of order one. The results are identical to those for

Example 1 given in Section 3.4. For the stochastic gradient descent case, the situation is also the same as the part of Example 1 considered in Section 4.2 with explicit forms for both gradient descent and stochastic gradient descent cases. Numerical results were computed from these explicit expressions to illustrate gradient descent algorithms and the corresponding differential equations as illustrated in Figure 1.
Now consider the random design case. Denote the covariance matrix of U_{2i} by $\boldsymbol{\alpha} = E[U_{2i}U_{2i}'] = (\alpha_{ij})_{i,j=1,2}$, and let $\check{\theta}$ be the true value of parameter θ in the regression model, and define $\ell(\theta; U_i) = (U_{1i} - U_{2i}'\theta)^2/2$. Then $\mathcal{L}^n(\theta; \mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n (U_{1i} - U_{2i}'\theta)^2/2$ is the half mean residual square error, $g(\theta) = E[\ell(\theta; U_i)] = \tau^2/2 + (\theta - \check{\theta})'\boldsymbol{\alpha}(\theta - \check{\theta})^2/2$, $\nabla g(\theta) = \boldsymbol{\alpha}(\theta - \check{\theta})$, $\nabla\ell(\theta; U_i) = U_{2i}(U_{2i}'\theta - U_{1i})$, and $\nabla\mathcal{L}^n(\theta; \mathbf{U}_n) = \frac{1}{n}\sum_{i=1}^n U_{2i}(U_{2i}'\theta - U_{1i})$. Also from the regression model (43) we have

$$\nabla\ell(\theta; U_i) = U_{2i}U_{2i}'(\theta - \check{\theta}) - U_{2i}\,\varepsilon_i, \quad E[\nabla\ell(\theta; U_i)] = \boldsymbol{\alpha}(\theta - \check{\theta}) = \nabla g(\theta),$$

$$\sigma(\theta) = Var[\nabla\ell(\theta; U_i)] = E[U_{21}U_{21}'(\theta - \check{\theta})(\theta - \check{\theta})'U_{21}U_{21}'] + \tau^2\boldsymbol{\alpha} - \boldsymbol{\alpha}(\theta - \check{\theta})(\theta - \check{\theta})'\boldsymbol{\alpha}',$$

where we set $U_{21} = (H_1, H_2)'$ and $\beta = (\beta_1, \beta_2)' = \theta - \check{\theta}$, and compute $E(H_1^4) = 3\alpha_{11}^2$, $E(H_1^3 h_2) = 3\alpha_{11}\alpha_{12}$, $E(H_1^2 H_2^2) = \alpha_{11}\alpha_{22} + 2\alpha_{12}^2$, $E(H_1 H_2^3) = 3\alpha_{22}\alpha_{21}$, $E(H_2^4) = 3\alpha_{22}^2$, and $U_{21}U_{21}'(\theta - \check{\theta})(\theta - \check{\theta})'U_{21}U_{21}' =$

$$\begin{pmatrix} H_1^4\beta_1^2 + 2H_1^3H_2\beta_1\beta_2 + H_1^2H_2^2\beta_2^2 & H_1^3H_2\beta_1^2 + 2H_1^2H_2^2\beta_1\beta_2 + H_1H_2^3\beta_2^2 \\ H_1^3H_2\beta_1^2 + 2H_1^2H_2^2\beta_1\beta_2 + H_1H_2^3\beta_2^2 & H_1^2H_2^2\beta_1^2 + 2H_1H_2^3\beta_1\beta_2 + H_2^4\beta_2^2 \end{pmatrix}.$$

Again it is easy to see that the minimization problems corresponding to (1) and (7) have explicit solutions in this case: $g(\theta)$ has the minimizer being the true parameter value $\check{\theta}$, and $\mathcal{L}^n(\theta; \mathbf{U}_n)$ has the minimizer equal to the least squares estimator $\hat{\theta}_n$.
Take $\boldsymbol{\alpha} = \begin{pmatrix} 0.02 & 0 \\ 0 & 0.005 \end{pmatrix}$, $\tau = 0.1$, and $\check{\theta} = (0, 0)'$. Then $g(\theta) = 0.02\theta_1^2 + 0.005\theta_2^2 + 1$, and

$$\boldsymbol{\sigma}^2(\theta) = \begin{pmatrix} 2\alpha_{11}^2\theta_1^2 + \alpha_{11}\alpha_{22}\theta_2^2 + \tau^2\alpha_{11} & \alpha_{11}\alpha_{22}\theta_1\theta_2 \\ \alpha_{11}\alpha_{22}\theta_1\theta_2 & \alpha_{11}\alpha_{22}\theta_1^2 + 2\alpha_{22}\theta_2^2 + \tau^2\alpha_{22} \end{pmatrix}.$$

Unlike the fixed design case, there lack of simple explicit expressions for accelerated (or stochastic) gradient descent algorithms and ordinary (or stochastic) differential equations. We applied gradient descent, accelerated gradient descent, and stochastic gradient descent algorithms and solved the corresponding ordinary or stochastic differential equations by the Euler scheme for various initial values and (m, n, δ). Figure 1 illustrates sample paths of sequences generated from accelerated gradient

descent (based on all data) and stochastic gradient descent algorithms and their corresponding ordinary or stochastic differential equations. As explicitly demonstrated in the fixed design case, the results show that both algorithms and ordinary or stochastic differential equations lead to solutions of the corresponding minimization problems, and whole sample paths for the solutions of stochastic optimization are random sequences or curves distributed around those for the corresponding deterministic optimization.

REFERENCES

Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On lower and upper bounds for smooth and strongly convex optimization problems. ArXiv preprint arXiv:1503.06833, 2015.

Hedy Attouch, Juan Peypouquet, and Patrick Redont. On the fast convergence of an inertial gradient-like system with vanishing viscosity. ArXiv preprint arXiv:1507.04782, 2015.

Patrick 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.

M. Bogdan, E. V. D. Berg, C. Sabatti, W. Su, and E. J. Candés. SLOPE adaptive variable selection via convex optimization. The Annals of Applied Statistics, 9(3):1103-1140, 2015.

Léon Bottou. Online learning and stochastic approximations. On-line learning in neural networks, 17(9):142, 1998.

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):1-122, 2011.

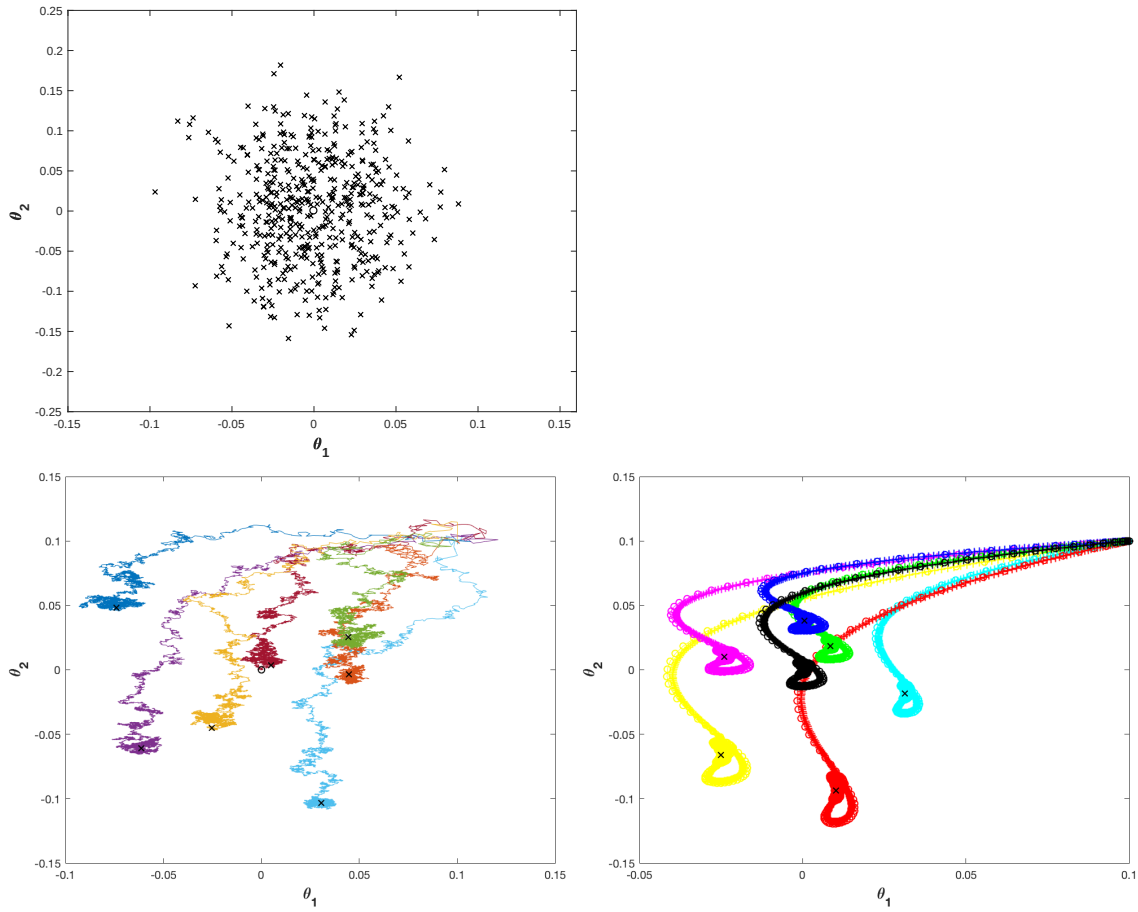

Figure 1: The scatter plot of estimators $\hat{\theta}_n$ and the plots of sample paths for accelerated and stochastic gradient descent algorithms and their corresponding ordinary or stochastic differential equations, where $\delta = 0.05$, $n = 1000$, $m = 200$, initial values $x_0 = x_0^n = x_0^m = (0.1, 0.1)'$. The top panel is the scatter plot of 500 simulated $\hat{\theta}_n$, and the bottom panels are sample paths of gradient descent algorithms to compute $\hat{\theta}_n$, with the bottom left and right panels for the stochastic gradient descent case, and the accelerated gradient descent case based on all data, respectively. In both bottom panels, $\times$ denotes the estimator $\hat{\theta}_n$ which is the solution of optimization (7), with $\circ$ for the true parameter value $\check{\theta} = (0, 0)$ which is the solution of optimization (1); in the bottom left panel, color curves are different sample paths of stochastic gradient descent; and in the bottom right panel, color curves with $+$ and $\circ$ correspond to sample paths of algorithm (10) and their corresponding differential equation (11), respectively, for accelerated gradient descent based on all data, with the black curve for the sample path of algorithm (4) and ordinary differential equation (6).

A. A. Brown and M. C. Bartholomew-Biggs. Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations. Journal of Optimization Theory and Applications, 62(2):211-224, 1989.

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

Sébastien Bubeck and Nicoló Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends in Machine Learning, 5(1):1-122, 2012.

J. C. Butcher. Numerical Methods for Ordinary Differential Equations. John Wiley. 2008.

Venkat Chandrasekarana and Michael I. Jordan (2012). Computational and statistical tradeoffs via convex relaxation. PNAS 110, no. 13, E1181-E1190.

Changyou Chen, David Carlson, Zhe Gan, Chunyuan Li, and Lawrence Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. arXiv:1512.07962v3, 2016.

Changyou Chen, Nan Ding, Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. arXiv:1610.06665v1, 2016.

Csörgö, M., Horváth, L., and Kokoszka, P. Approximation for bootstrapped empirical processes. Proceedings of the American Mathematical Society 128, 2457-2464. 1999.

Csörgö, S. and Mason, D. M. Bootstrapping empirical functions. Ann. Statist. 17, 1447-1471. 1989.

G. Da Prato and J. Zabczyk. Ergodicity for Infinite Dimensional Systems. Cambridge University Press, Cambridge; 1996.

Nicolas Flammarion and Francis R. Bach. From averaging to acceleration, there is only a step-size. In Proceedings of the 28th Conference on Learning Theory (COLT), 2015.

Crispin W. Gardiner. Stochastic Methods: A Handbook for the Natural and Social

Sciences. Springer, 4th edition. 2009.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In COLT, 2015.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming, 156(1):59-99, 2015.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT press. 2016.

He, S. W., Wang, J. G. and Yan, J. A. Semimartingale Theory and Stochastic Calculus. Science Press and CRC Press. 1992.

Chonghai Hu, James T. Kwok, and Weike Pan. Accelerated gradient methods for stochastic optimization and online learning. In Advances in Neural Information Processing Systems (NIPS) 22, 2009.

N. Ikeda and S. Watanabe. Stochastic Differential Equations and Diffusion Processes, Volume 24 (North-Holland Mathematical Library). 1981.

Jean Jacod and Albert Shiryaev. Limit Theorems for Stochastic Processes. Spriner. 2nd Edition. 2003.

Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In Proceedings of the 26th International Conference on Machine Learning (ICML), 2009.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, Michael I. Jordan. How to escape saddle points efficiently. arXiv:1703.00887v1

Vladimir Jojic, Stephen Gould, and Daphne Koller. Accelerated dual decomposition for MAP inference. In Proceedings of the 27th International Conference on Machine Learning (ICML), 2010.

Anatoli Juditsky. Convex Optimization II: Algorithms. Lecture Notes, 2013.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. Stoch. Syst., 1(1):17-58, 2011.

Kenji Kawaguchi. Deep learning without poor local minima. In Advances In Neural Information Processing Systems, pages 586-594, 2016.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang (2017). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. arXiv:1609.04836v2.

J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent R.V.'s and the sample DF. I. Z. Wahrschein. Verw. 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. Z. Wahrschein. Verw. Gebiete 34, 33-58. 1976.

Walid Krichene, Alexandre Bayen, and Peter Bartlett. Accelerated mirror descent in continuous and discrete time. In Advances in Neural Information Processing Systems (NIPS) 29, 2015.

Walid Krichene, Syrine Krichene, and Alexandre Bayen. Efficient Bregman projections onto the simplex. In 54th IEEE Conference on Decision and Control, 2015.

Guanghui Lan. An optimal method for stochastic composite optimization. Mathematical Programming, 133(1-2):365-397, 2012.

Guanghui Lan, Zhaosong Lu, and Renato Monteiro. Primal-dual first-order methods with $O(1/\epsilon)$ iteration-complexity for cone programming. Mathematical Programming, 126(1):1-29, 2011.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In Conference on Learning Theory, pages 1246-1257, 2016.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. SIAM Journal on Optimization, 26(1):57-95, 2016.

Chris Junchi Li, Lei Li, Junyang Qian, Jian-Guo Liu Batch Size Matters: A Diffusion Approximation Framework on Nonconvex Stochastic Gradient Descent. arXiv:1705.07562v1.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In Advances in Neural Information Processing Systems (NIPS) 28, 2015.

Qianxiao Li, Cheng Tai, Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. arXiv:1511.06251v3, 2015.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. A variational analysis of stochastic gradient algorithms. arXiv:1602.02666v1, 2016.

Stephan Mandt, Matthew D. Hoffman, David M. Blei. Stochastic gradient descent as approximate Bayesian inference. arXiv:1704.04289v1, 2017.

Pascal Massart (1989). Strong Approximation for Multivariate Empirical and Related Processes, Via KMT Constructions. Ann. Probab. 17, 266-291.

Qi Meng, Wei Chen, Jingcheng Yu,Taifeng Wang, Zhi-Ming Ma, Tie-Yan Liu. Asynchronous Accelerated Stochastic Gradient Descent. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence ((IJCAI-16). 2016.

R. Monteiro, C. Ortiz, and B. Svaiter. An adaptive accelerated first-order method for convex optimization. Technical report, ISyE, Gatech, 2012.

Arkadi Nemirovskii and David Yudin. Problem Complexity and Method Efficiency in Optimization. John Wiley & Sons, 1983.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 27(2):372-376, 1983.

Yurii Nesterov. Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization. Kluwer, 2004.

Yurii Nesterov. Smooth minimization of non-smooth functions. Mathematical Programming, 103(1):127-152, 2005.

Yurii Nesterov. Accelerating the cubic regularization of Newton's method on convex

problems. Mathematical Programming, 112(1):159-181, 2008.

Yurii Nesterov. Gradient methods for minimizing composite functions. Mathematical Programming, 140(1):125-161, 2013.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton's method and its global performance. Mathematical Programming, 108(1):177-205, 2006.

J. Nocedal and S. Wright. Numerical Optimization. Springer Science & Business Media, 2006.

Brendan O'Donoghue and Emmanuel Candés. Adaptive restart for accelerated gradient schemes. Foundations of Computational Mathematics, 15(3):715-732, 2015.

B. O'Donoghue and E. J. Candés. Adaptive restart for accelerated gradient schemes. Found. Comput. Math., 2013.

B. T. Polyak. Introduction to optimization. Optimization Software New York, 1987.

M. Raginsky and J. Bouvrie. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In CDC 2012, pages 6793-6800, 2012.

Alexander Rakhlin, Ohad Shamir, Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. arXiv:1109.5647v7, 2012.

Emmanuel Rio (1993a). Strong Approximation for Set-Indexed Partial Sum Processes Via KMT Constructions I. Ann. Probab. 21, 759-790.

Emmanuel Rio (1993b). Strong Approximation for Set-Indexed Partial-Sum Processes, Via KMT Constructions II. Ann. Probab. 21, 1706-1727.

R. T. Rockafellar. Convex Analysis. Princeton Landmarks in Mathematics. Princeton University Press, 1997. Reprint of the 1970 original.

Sebastian Ruder. An overview of gradient descent optimization algorithms. arXiv:1609.04747v1, 2016.

A. P. Ruszczyński. Nonlinear Optimization. Princeton University Press, 2006.

N. Z. Shor. Minimization Methods for Non-Differentiable Functions. Springer Science & Business Media, 2012.

Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time. arXiv:1611.05545v3, 2017.

Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time: A central limit theorem. arXiv:1710.04273v2, 2017.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In Proceedings of the 30th International Conference on Machine Learning, pages 1139-1147, 2013.

Weijie Su, Stephen Boyd, and Emmanuel J. Candés. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In Advances in Neural Information Processing Systems (NIPS) 27, 2014.

Weijie Su, Stephen Boyd, and Emmanuel J. Candés. A differential equation for modeling Nesterov's accelerated gradient method: Theory and Insights. Journal of Machine Learning Research 17, 1-43. 2016.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. SIAM Journal on Optimization, 2008.

P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. Mathematical Programming, 125(2):263-295, 2010.

A. W. van der Vaart and Jon Wellner. Weak Convergence and Empirical Processes With Applications to Statistics. Springer. 2000.

G. N. Watson. A Treatise on the Theory of Bessel Functions. Cambridge Mathematical Library. Cambridge University Press, 1995. Reprint of the second (1944) edition.

A. Wibisono, A. Wilson, and M. Jordan. A variational perspective on accelerated methods in optimization. Proc Natl Acad Sci 113, E7351-E7358. 2016.

6 Appendix: Proofs

Denote by C generic constant free of (δ, m, n) whose value may change from appearance to appearance. For simplicity we take initial values $x_0^n = x_0^m = x_0$. In

appendix sections of theorem proofs, 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 put 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 notation. Because of this convention and Skorokhod's representation theorem, we often switch between "convergence in probability" and "convergence in distribution." Also because of the convention, when no confusion occurs, we try 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, X's for solutions of ordinary differential equations (ODEs) and stochastic differential equations (SDEs).

Convention 2. As described at the end of Section 1, for gradient descent algorithms to solve optimization (7), we add super indices n and m to notations for the associated processes and sequences based on all data in Section 3 and based on mini-batches (or bootstrap samples) in Section 4, respectively, while notations without any superscript are for sequences and functions corresponding to optimization (1).

Convention 3. We reserve V's for normalized solutions difference between differential equations associated with optimization (1) and optimization (7) under the cases for all data and mini-batches (bootstrap samples), while we reserve V without any superscript as their corresponding weak convergence limits.

Convention 4. We add an extra label Q to the objective functions $\ell(\theta; U_i)$ and $\mathcal{L}(\theta; \mathbf{U}_n)$, and write them as $\ell(\theta; U_i, Q)$ and $\mathcal{L}(\theta; \mathbf{U}_n, Q)$ so that Q is clearly specified as the distribution of U_i.

Convention 5. As described at the end of Section 1, we add a superscript $*$

to notations U's associated with mini-batches (or bootstrap samples), and as in Convention 2, their corresponding process notations have a superscript m.

Convention 6. We denote by $|\Psi|$ the absolute value of scalar Ψ, the Euclidean norm of vector Ψ, or the spectral norm of matrix Ψ.

6.1 Proofs of Theorem 3.1

First we provide detailed arguments for the accelerated case, as results for the plain case are relatively easier to show and will be establish later.

6.1.1 Differential equation derivation

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

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

Then

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

For the accelerated case, we can re-express ODE (11) as

$$\ddot{X}^n(t) + \frac{3}{t}\dot{X}^n(t) + \nabla g(X^n(t)) + \frac{1}{\sqrt{n}}R^n(X^n(t); \mathbf{U}_n, Q) = 0. \tag{44}$$

By Lemma 6.4 below we obtain that $X^n(t)$ converges 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 $n \to \infty$, $R^n(X^n(t); \mathbf{U}_n, Q) = O_P(1)$, and $n^{-1/2}R^n(X^n(t); \mathbf{U}_n, Q)| \to 0$. Hence, ODEs (11) and (44) both converge to ODE (6).

By Skorohod's representation theorem, there exist $\mathbf{U}_\dagger$ and $\mathbf{Z}_\dagger$ defined on some common probability space with $\mathbf{Z}_\dagger \sim N_p(0, \mathbf{I}_p)$ and $\mathbf{U}_\dagger$ identically distributed as $\mathbf{U}_n$ such

that as $n \to \infty$, $R^n(\theta; \mathbf{U}_\dagger, Q) - \boldsymbol{\sigma}(\theta)\mathbf{Z}_\dagger = o(1)$ uniformly over $\theta \in \Theta_X$. Thus we have that the solution $X^n(t)$ of equations (11) is identically distributed as the solution $X^n_\dagger(t)$ of

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

which in turn may be written as

$$\ddot{X}^n_\dagger(t) + \frac{3}{t}\dot{X}^n_\dagger(t) + \nabla g(X^n_\dagger(t)) + \frac{1}{\sqrt{n}} \boldsymbol{\sigma}(X^n_\dagger(t))\mathbf{Z}_\dagger + o\left(n^{-1/2}\right) = 0. \tag{45}$$

In particular (45) is equivalent to (15) up to the order of $n^{-1/2}$, which implies that as $n \to \infty$, ODEs (11), (15), and (45) all converge to ODE (6), and $X^n_\dagger(t)$ almost surely converges to $X(t)$. Since the solutions of equations (11), (15) and (45) are defined in the distribution sense, when there is no confusion, with a little abuse of notations we may drop index $\dagger$ and write equation (45) as

$$\ddot{X}^n(t) + \frac{3}{t}\dot{X}^n(t) + \nabla g(X^n(t)) + \frac{1}{\sqrt{n}} \boldsymbol{\sigma}(X^n(t))\mathbf{Z} + o\left(n^{-1/2}\right) = 0, \tag{46}$$

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 (44) for each n. We can employ the same arguments with $\nabla g(\cdot)$ replaced by $\Delta g(X(t))\,\Pi(t) + \boldsymbol{\sigma}(X(t))$ or $\Delta g(X(t))\,V(t) + \boldsymbol{\sigma}(X(t))\mathbf{Z}$ to show that linear differential equations (13) and (17) have unique solutions.

For the plain gradient descent case, Lemma 6.1 below shows that $X^n(t)$ converges to $X(t)$ uniformly over any finite interval. Similarly we can establish that ODE (9) is asymptotically equivalent to ODE (14), and the standard ODE theory shows that they have unique solutions.

6.1.2 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 plus uniform tightness (or stochastic equicontinuity) (see Kim and Pollard (1990, Theorem 2.3), Pollard (1988) and Van der Vaart and Wellner (2000)). We establish finite-dimensional convergence below.
For the accelerated case, taking a difference between ODEs (6) and (45), 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

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

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

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

where $X(t)$ is the solution of equation (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 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 (3) and (14) can establish finite-dimensional convergence.

Now we show 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

$$\limsup_{n\to\infty} P\left[\sup_{(t_1,t_2)\in\mathcal{T}(T,\delta)} |V_n(t_1) - V_n(t_2)| > \eta\right] < \varepsilon, \tag{47}$$

where $\mathcal{T}(T,\eta) = \{(t_1,t_2), t_1, t_2 \in I\!R_+, \max(t_1,t_2) \le T, |t_1 - t_2| < \eta\}$. The tightness of $V_n(t)$ on $I\!R_+$ requires above result for any $T < \infty$.

Note that as (47) requires only some probability evaluation, with the abuse of notations we have dropped index † and work on equation (46).

6.1.3 Weak convergence proof for the plain gradient descent case

Lemma 6.1. *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 (3) and (9) we have

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

and using assumption A1-A2 we obtain

$$|\nabla g(X^n(t)) - \nabla g(X(t))| \le L|X^n(t) - X(t)|,$$

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

Combining them together we arrive at

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

and an application of Gronwall's inequality leads to

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

which implies

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

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

Lemma 6.2. *For any given $T > 0$, $V^n(t)$ is stochastically equicontinuous on [0,T].*

Proof. Lemma 6.2 has shown $\max_{t \in [0,T]} |V^n(t)| = O_P(1)$. From ODEs (3) and (9) we have

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

Lemma 6.1 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 $|\sup_t R^n(X^n(t); \mathbf{U}_n, Q)| \sim \sup_t |\boldsymbol{\sigma}(X^n(t))\mathbf{Z}| = O_P(1)$. 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 6.1, we immediately establish the lemma.

Proof of Theorem 3.1 for the plain gradient descent case The same perturbation argument in Section 6.1.2 can be used to show finite-dimensional distribution convergence of $V^n(t)$ to $V(t)$ for simple ODE (9) in the plain gradient descent case. With the tightness of $V^n(t)$ shown in Lemma 6.2 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.

6.1.4 Weak Convergence proof for the accelerated case

We can use the same proof in Su et al. (2016, Theorem 1) to show that ODE (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 such as 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. We will first extend and refine some 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 $a \in (0, 1]$,

$$M_a(s, t; Y) = M_a(\mathcal{I}; Y) = \sup_{u\in[s,t]} \left| \frac{\dot{Y}(u) - \dot{Y}(s)}{(u-s)^a} \right|. \tag{48}$$

In the proof of Theorem 3.1 we take $a = 1$ and use $M_1(s, t; Y)$. We will need $M_a(s, t; Y)$ with $a < 1$ later in the proof of Theorems 4.5.

Lemma 6.3. *For $X(t)$ and $X^n(t)$ we have the following inequalities,*

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

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

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

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

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

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

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 $s=0$,

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

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

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

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

$$\ddot{V}^n(t)+\frac{3}{t}\dot{V}^n(t)+\sqrt{n}\nabla[g(X^n(t))-g(X(t))]+R^n(X^n(t);\mathbf{U}_n,Q)=0. \tag{49}$$

Let

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

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

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

As in the proof of Lemma 6.1, using assumptions A1-A2 we obtain

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

Putting together these results we get

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

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 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

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

ODE (49) is equivalent to

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

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

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

As above inequality holds for any $t > s$, using the definition of $M_1(s,t;V^n)$ we have

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

and solving for $M_1(s,t;V^n)$ leads to

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

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 above inequality, and $V^n(0) = \dot{V}^n(0) = 0$, $X^n(0) = X(0) = x_0$. Then

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

which in particular implies that

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

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

Lemma 6.4. *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 6.3 we have

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

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

Assumption A4 implies that $R^n(x_0; \mathbf{U}_n, Q) = O_P(1)$, and $R^n(X(s_1); \mathbf{U}_n, Q) = O_P(1)$, and 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 6.3 we have

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

and

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

We will use 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, by induction we have shown that the upper bounds of $X^n(t) - X(t)$ and $\dot{X}^n(t) - \dot{X}(t)$ over $t \leq s_{i-1}$ 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 $R^n(X^n(s_{i-1}); \mathbf{U}_n, Q) = O_P(1)$, and $\sup_{u\geq 0} |R^n(X(u); \mathbf{U}_n, Q)| = O_P(1)$. Above two inequalities immediately show 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 6.5. *$V^n(t)$ is stochastically equicontinuous on $[0, T]$.*

Proof. Lemma 6.4 shows 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 3.1 Lemma 6.5 together with the finite distribution convergence immediately lead to that as $n \to \infty$, $V^n(t)$ weakly converges to $V(t)$.

6.2 Proof of Theorem 3.2

6.2.1 Proof for the plain gradient descent case

Lemma 6.6. *For the case of plain gradient descent algorithm, we have*

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

where $\{x_k^n\}$ is generated from algorithm (8), with $x_\delta^n(t)$ its continuous-time step process, and $X^n(t)$ is the solution of ODE (9).

Proof. Algorithm (8) is the Euler scheme for solving ODE (9), and we will apply the standard ODE theory to obtain the global approximation error for the Euler scheme. First by assumption A1 we have that $\nabla\mathcal{L}^n(\theta; \mathbf{U}_n, Q)$ 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 derivative on both sides of ODE (9), we obtain

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

Using Lemma 6.1, we conclude that for large n, $X^n(t)$ falls into Θ_X, and thus assumption A4 indicates that $\sup_t |\nabla^\kappa \mathcal{L}^n(X^n(t); \mathbf{U}_n, Q)| \sim \sup_t |\nabla^\kappa g(X^n(t)) + n^{-1/2}\boldsymbol{\sigma}_k(X^n(t))\mathbf{Z}_\kappa| = O_P(1)$, where $\mathbf{Z}_\kappa$ are standard normal random variables. Combining these results together we get $\sup_{t\in[0,T]} |\ddot{X}^n(t)| = O_P(1)$. An application of the standard ODE theory for the Euler scheme leads to

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

Proof of Theorem 3.2. Lemma 6.6 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 3.1.

6.2.2 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 (4) and (10), respectively, and $X(t)$ and $X^n(t)$ are respective solutions of ODEs (6) and (11), and $X_\eta(t)$ and $X_\eta^n(t)$ the corresponding smoothed ODEs with t replaced by $t \vee \eta$, and $X_{k,\eta}$ and $X_{k,\eta}^n$ the approximation sequences generated by the Euler scheme for the corresponding smoothed ODEs, where η is a small positive constant.

Lemma 6.7.

$$\max_{k\leq T\delta^{-1/2}}|x_k - X_{k,\eta}| = O(\eta+\delta^{1/2}),\quad \max_{k\leq T\delta^{-1/2}}|x_k^n - X_{k,\eta}^n| = O_P([\eta+\delta^{1/2}][1+n^{-1/2}]). \tag{50}$$

Proof. In the proof of Lemma 21 in Su et al (2016) we can analyze the approximation errors and find order of $x_k - X_{k,\eta}$ in terms of (δ,η) to be $\eta+\delta^{1/2}$. Indeed, for example in the derivation of the approximation error of Su et al. (2016, equation (36)), in terms of our parameter notations δ (step size in continuous modeling) and η (parameter to smooth $1/t$ at zero in the smoothed ODE), we may find the approximation error order from the derivation of equation (36) in Su et al. (2016) is bounded by

$$\begin{aligned}&\frac{1}{2L}\Big[\{C_1L\delta+C_2+(C_1+C_2)\sqrt{L\delta}\}(1+\sqrt{L\delta})^{[\eta/\sqrt{\delta}]}\\&+\{C_1L\delta+C_2-(C_1+C_2)\sqrt{L\delta}\}(1-\sqrt{L\delta})^{[\eta/\sqrt{\delta}]}-2C_2\Big],\end{aligned}$$

where C_1 and C_2 are constants. Note that for small δ,

$$(1\pm\sqrt{L\delta})^{[\eta/\sqrt{\delta}]} = e^{\pm\eta\sqrt{L}+O(L\eta\delta^{1/2})}.$$

We obtain the order for the approximation error

$$\begin{aligned}&\frac{C_1\delta}{2}\left[e^{\eta\sqrt{L}+O(L\eta\delta^{1/2})}+e^{-\eta\sqrt{L}+O(L\eta\delta^{1/2})}\right]+\frac{C_2}{2L}\left[e^{\eta\sqrt{L}+O(L\eta\delta^{1/2})}+e^{-\eta\sqrt{L}+O(L\eta\delta^{1/2})}-2\right]\\&+\frac{(C_1+C_2)\sqrt{L\delta}}{2}\left[e^{\eta\sqrt{L}+O(L\eta\delta^{1/2})}-e^{-\eta\sqrt{L}+O(L\eta\delta^{1/2})}\right]\\&=O(\delta)+O(\eta^2+\eta\delta^{1/2})+O(\eta\delta^{1/2})=O(\eta^2+\eta\delta^{1/2}+\delta)=O(\eta+\delta^{1/2}),\end{aligned} \tag{51}$$

where C_1 and C_2 are constants.

For the case of $(x_k^n, X_{k,\eta}^n)$, the only modification needed is to replace $\nabla g(\theta)$ for the case of $(x_k, X_{k,\eta})$ by $\nabla g(\theta)+n^{-1/2}R^n(\theta;\mathbf{U}_n,Q)$. Note that $\nabla g(\theta)$ is L-Lipschitz,

the proof of Lemma 6.1 indicates that $n^{-1/2}R^n(\theta; \mathbf{U}_n, Q)$ is $(n^{-1}\sum_{i=1}^n h_1(U_i) + L)$-Lipschitz, and thus

$$\begin{aligned}
&|\nabla g(\theta) + n^{-1/2}R^n(\theta; \mathbf{U}_n, Q) - \nabla g(\vartheta) - n^{-1/2}R^n(\vartheta; \mathbf{U}_n, Q)| \\
&\leq |\nabla g(\theta) - \nabla g(\vartheta)| + n^{-1/2}|R^n(\theta; \mathbf{U}_n, Q) - R^n(\vartheta; \mathbf{U}_n, Q)| \\
&\leq \left(n^{-1}\sum_{i=1}^n h_1(U_i) + 2L \right) |\theta - \vartheta|.
\end{aligned}$$

The same argument for the case of $(x_k, X_{k,\eta})$ can be applied to the case of $(x_k^n, X_{k,\eta}^n)$, with Lipschitz constant L replaced by $L_n = n^{-1}\sum_{i=1}^n h_1(U_i) + 2L$. As $n \to \infty$, $n^{-1}\sum_{i=1}^n h_1(U_i) = E[h_1(U)] + O_P(n^{-1/2})$, and $L_n = 2L + E[h_1(U)] + O_P(n^{-1/2})$. Therefore, substituting L by L_n on the right hand side of (51) we obtain the approximation error bound with order $O_P(\eta + \delta^{1/2} + (\eta + \delta^{1/2})n^{-1/2})$ for the case of $(x_k, X_{k,\eta})$.

Lemma 6.8.

$$\max_{k \leq T\delta^{-1/2}} |X_{k,\eta} - X_\eta(k\delta^{1/2})| = O(\eta + \delta^{1/2}), \quad \max_{k \leq T\delta^{-1/2}} |X_{k,\eta}^n - X_\eta^n(k\delta^{1/2})| = O_P([\eta + \delta^{1/2}][1 + n^{-1/2}]). \tag{52}$$

Proof. Again we adopt the proof argument of Lemma 21 in Su et al (2016) to derive the approximation error order. First consider the case of $X_{k,\eta} - X_\eta(k\delta^{1/2})$. We rewrite ODE (6) for $X(t)$ and its corresponding smoothed ODE for $X_\eta(t)$ as

$$\dot{X}(t) = Z(t), \;\; \dot{Z}(t) = -\frac{3}{t}Z(t) - \nabla g(X(t)), \;\; X(0) = x_0, \dot{X}(0) = 0, \tag{53}$$

$$\dot{X}_\eta(t) = Z_\eta(t), \;\; \dot{Z}_\eta(t) = -\frac{3}{\eta \vee t}Z_\eta(t) - \nabla g(X_\eta(t)), \;\; X_\eta(0) = x_0, \dot{X}_\eta(0) = 0, \tag{54}$$

and the Euler scheme for the smoothed ODE (54) is given by

$$X_{k+1,\eta} = X_{k,\eta} + \delta^{1/2} Z_{k,\eta}, \tag{55}$$

$$Z_{k+1,\eta} = \left(1 - \frac{3\delta^{1/2}}{\eta \vee (k\delta^{1/2})}\right) Z_{k,\eta} - \delta^{1/2} \nabla g(X_{k,\eta}), \quad X_{0,\eta} = x_0, Z_{0,\eta} = 0. \tag{56}$$

From (53) we get

$$\begin{aligned} X(t_{k+1}) &= X(0) + \int_0^{t_{k+1}} Z(u)du = X(t_k) + \int_{t_k}^{t_{k+1}} Z(u)du \\ &= X(t_k) + \delta^{1/2} Z(t_k) + \int_0^{\delta^{1/2}} [Z(v + t_k) - Z(t_k)]dv, \end{aligned} \tag{57}$$

$$\begin{aligned} Z(t_{k+1}) &= Z(0) - \int_0^{t_{k+1}} \frac{3}{u} Z(u)du - \int_0^{t_{k+1}} \nabla g(X(u))du \\ &= \left(1 - \frac{3\delta^{1/2}}{t_k}\right) Z(t_k) - \delta^{1/2} \nabla g(X(t_k)) \\ &- \int_{t_k}^{t_{k+1}} \left[\frac{3}{u} Z(u) - \frac{3}{t_k} Z(t_k)\right] du - \int_{t_k}^{t_{k+1}} [\nabla g(X(u)) - \nabla g(X(t_k))]du. \end{aligned} \tag{58}$$

Since $X(t)$ and $Z(t) = \dot{X}(t)$ have bounded derivatives on $[0, T]$, and $Z(0) = \dot{X}(0) = 0$, we have $|Z(t_k)| \le C t_k$,

$$\left| \int_0^{\delta^{1/2}} [Z(v + t_k) - Z(t_k)]dv \right| \le \int_0^{\delta^{1/2}} |Z(v + t_k) - Z(t_k)|dv = O(\delta), \tag{59}$$

$$\left| \int_{t_k}^{t_{k+1}} [\nabla g(X(u)) - \nabla g(X(t_k))]du \right| \le L \int_{t_k}^{t_{k+1}} |X(u) - X(t_k)|du = O(\delta), \tag{60}$$

$$\left|\int_{t_k}^{t_{k+1}}\left[\frac{3}{u}Z(u)-\frac{3}{t_k}Z(t_k)\right]du\right| \le \left|\int_{t_k}^{t_{k+1}}\left[\frac{3}{u}Z(u)-\frac{3}{u}Z(t_k)\right]du\right|$$
$$+\left|\int_{t_k}^{t_{k+1}}\left[\frac{3}{u}Z(t_k)-\frac{3}{t_k}Z(t_k)\right]du\right|$$
$$\le \frac{3}{t_k}\int_{t_k}^{t_{k+1}}|Z(u)-Z(t_k)|du+|Z(t_k)|(t_{k+1}-t_k)\left[\frac{3}{t_k}-\frac{3}{t_{k+1}}\right]du$$
$$\le \frac{3}{k\delta^{1/2}}O(\delta)+Ct_k\frac{3(t_{k+1}-t_k)^2}{t_kt_{k+1}}=O(\delta^{1/2}k^{-1}). \tag{61}$$

Denote by $a_k = |X_{k,\eta} - X(t_k)|$ and $b_k = |Z_{k,\eta} - Z(t_k)|$, whose initial values are $a_0 = 0$ and $b_0 = |\nabla g(x_0)|\delta^{1/2}$. Then as in Su et al (2016), checking the similar recursive expressions (55) and (57) and using (59) we obtain $a_{k+1} \le a_k + \delta^{1/2}b_k + C\delta$, and $a_k \le \delta^{1/2}S_{k-1} + Ck\delta$, where $S_k = \sum_{i=0}^{k} b_i$, and examining the similar recursive expressions (56) and (58) and utilizing (60) and (61) we conclude for $k+1 \le T\delta^{-1/2}$,

$$b_{k+1} \le \left|1-\frac{3\delta^{1/2}}{\eta\vee t_k}\right|b_k+\delta^{1/2}La_k+3\delta^{1/2}\left|\frac{1}{t_k}-\frac{1}{\eta\vee t_k}\right||Z(t_k)|+C\left(\delta+\delta^{1/2}k^{-1}\right)$$
$$\le b_k+L\delta^{1/2}a_k+3\delta^{1/2}\left|\frac{1}{t_k}-\frac{1}{\eta\vee t_k}\right|Ct_k+C\left(\delta+\delta^{1/2}k^{-1}\right)$$
$$\le b_k+L\delta^{1/2}(\delta^{1/2}S_{k-1}+Ck\delta)+3C\delta^{1/2}+C\left(\delta+\delta^{1/2}k^{-1}\right)$$
$$\le b_k+L\delta S_{k-1}+C_2\delta^{1/2}.$$

Since the recursive relationship for b_k is the same as in the proof of Lemma 21 in Su et al (2016), the same argument leads to the approximation error bound given by (51) for the maximum of $|X_{k,\eta} - X(k\delta^{1/2})|$ over $k \le T\delta^{-1/2}$, and thus we obtain the same order $\eta + \delta^{1/2}$ for the approximation error.

For the case of $X^n_{k,\eta} - X^n_\eta(k\delta^{1/2})$, as in the proof of Lemma 6.7 we need to replace $\nabla g(\theta)$ for the first case by $\nabla g(\theta) + n^{-1/2}R^n(\theta; \mathbf{U}_n, Q)$ and change Lipschitz constant L to $L_n = 2L + E[h_1(U)] + O_P(n^{-1/2})$ and use the same argument to establish order $O_P(\eta + \delta^{1/2} + (\eta + \delta^{1/2})n^{-1/2})$ for $X^n_{k,\eta} - X^n_\eta(k\delta^{1/2})$.

Lemma 6.9.

$$\sup_{u\in[0,T]} |R^n(X^n(t);\mathbf{U}_n,Q) - R^n(X^n_\eta(t);\mathbf{U}_n,Q)| = o_P(1).$$

Proof. Lemma 6.4 shows that $X^n(t)$ converges in probability to $X(t)$ and thus for large n, $X^n(t)$ will fall into Θ_X. Similarly, as $n\to\infty$, $X^n_\eta(t)$ converges in probability to $X_\eta(t)$, and as $\eta\to 0$, $X_\eta(t)$ converges to $X(t)$, and hence $\max_{t\in[0,T]}|X^n_\eta(t) - X^n(t)| = o_p(1)$, and for large n and small η, $X^n(t)$ and $X^n_\eta(t)$ fall into Θ_X. By assumptions A3-A4, we obtain that uniformly over $t\in[0,T]$,

$$|R^n(X^n(t);\mathbf{U}_n,Q)-\boldsymbol{\sigma}(X^n(t))\mathbf{Z}| = o_P(1),\ \ |R^n(X^n_\eta(t);\mathbf{U}_n,Q)-\boldsymbol{\sigma}(X^n_\eta(t))\mathbf{Z}| = o_P(1).$$

$$|\boldsymbol{\sigma}(X^n(t)) - \boldsymbol{\sigma}(X^n_\eta(t))| = o_P(1).$$

Finally we have

$$\begin{aligned}&|R^n(X^n(t);\mathbf{U}_n,Q) - R^n(X^n_\eta(t);\mathbf{U}_n,Q)| \le |R^n(X^n(t);\mathbf{U}_n,Q) - \boldsymbol{\sigma}(X^n(t))\mathbf{Z}|\\ &+ |R^n(X^n_\eta(t);\mathbf{U}_n,Q) - \boldsymbol{\sigma}(X^n_\eta(t))\mathbf{Z}| + |\boldsymbol{\sigma}(X^n(t)) - \boldsymbol{\sigma}(X^n_\eta(t))||\mathbf{Z}| = o_P(1).\end{aligned}$$

Lemma 6.10.

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

Proof. Since Lemma 6.11 below implies that both $M_1(0,\eta;X^n_\eta)$ and $M_1(0,\eta;X^n)$ are a.s. finite, we have for $t\in[0,\eta]$,

$$|\dot X^n_\eta(t) - \dot X^n(t)| \le \eta[M_1(0,\eta,X^n_\eta) + M_1(0,\eta;X^n)] = O(\eta).$$

On the other hand,

$$X^n_\eta(t)-X^n(t) = X^n_\eta(0)-X^n(0)+\int_0^t [\dot X^n_\eta(u)-\dot X^n_\eta(0)+\dot X^n(u)-\dot X^n(0)]du + t[\dot X^n_\eta(0)-\dot X^n(0)],$$

as X_η^n and X^n have the same initial value, $X_\eta^n(0) - X^n(0) = \dot{X}_\eta^n(0) - \dot{X}^n(0) = 0$, and we get

$$|X_\eta^n(t) - X^n(t)| \le \int_0^\eta u[M_1(0,\eta;X_\eta^n) + M_1(0,\eta;X^n)]du \le \frac{\eta^2}{2}[M_1(0,\eta;X_\eta^n) + M_1(0,\eta;X^n)].$$

For $t \in [\eta, \sqrt{3/L}]$, using Lemma 6.9 we obtain

$$\begin{aligned}
&|[\dot{X}^n(t) - \dot{X}_\eta^n(t)] - [\dot{X}^n(\eta) - \dot{X}_\eta^n(\eta)]| \le (t-\eta)M_1(\eta, t, X^n - X_\eta^n) \\
&\le C_1|\dot{X}^n(\eta) - \dot{X}_\eta^n(\eta)| + C_2|X^n(\eta) - X_\eta^n(\eta)| \\
&+ C_3 n^{-1/2} \sup_{u\in[0,T]} |R^n(X^n(t); \mathbf{U}_n, Q) - R^n(X_\eta^n(t); \mathbf{U}_n, Q)| \\
&= O_P(\eta + \eta^2) + o_P(n^{-1/2}),
\end{aligned}$$

and

$$|\dot{X}^n(t) - \dot{X}_\eta^n(t)| = O_P(\eta) + o_P(n^{-1/2}).$$

$$\begin{aligned}
&X_\eta^n(t) - X^n(t) = X_\eta^n(\eta) - X^n(\eta) + \int_\eta^t [\dot{X}_\eta^n(u) - \dot{X}_\eta^n(\eta) - \dot{X}^n(u) + \dot{X}^n(\eta)]du \\
&+ (t-\eta)[\dot{X}_\eta^n(\eta) - \dot{X}^n(\eta)],
\end{aligned}$$

$$\begin{aligned}
|X_\eta^n(t) - X^n(t)| &\le |X_\eta^n(\eta) - X^n(\eta)| + (t-\eta)|\dot{X}_\eta^n(\eta) - \dot{X}^n(\eta)| + \int_\eta^t (u-\eta)M_1(\eta, u; X^n - X_\eta^n)du \\
&= O_P(\eta^2 + \eta) + o_P(n^{-1/2}) = O(\eta) + o_P(n^{-1/2}).
\end{aligned}$$

Divide interval $[0,T]$ into $N = T\sqrt{L/3} + 1$ subintervals $[s_{i-1}, s_i]$, $i = 1, \cdots, N$, and we will show by induction that $X_\eta^n(t) - X^n(t)$ has order $O(\eta) + o_P(n^{-1/2})$ on $[0,T]$. Note that N is a fixed generic constant free of (n, δ, η).
We have already shown that for $t \in [s_0, s_1]$, both $X_\eta^n(t) - X^n(t)$ and $\dot{X}^n(t) - \dot{X}_\eta^n(t)$ are of order $O(\eta) + o_P(n^{-1/2})$. Assume the order result is true up to all $i-1$ subintervals.

For $t \in [s_{i-1}, s_i]$,

$$|[\dot{X}^n(t) - \dot{X}^n_\eta(t)] - [\dot{X}^n(s_{i-1}) - \dot{X}^n_\eta(s_{i-1})]| \le (s_i - s_{i-1}) M_1(s_{i-1}, s_i, X^n - X^n_\eta),$$

$$|\dot{X}^n(t) - \dot{X}^n_\eta(t)| \le C_1 |\dot{X}^n(s_{i-1}) - \dot{X}^n_\eta(s_{i-1})| + C_2 |X^n(s_{i-1}) - X^n_\eta(s_{i-1})|$$

$$+ C_3 n^{-1/2} \sup_{u \in [0,T]} |R^n(X^n(t); \mathbf{U}_n, Q) - R^n(X^n_\eta(t); \mathbf{U}_n, Q)|$$

$$= O_P(\eta) + o_P(n^{-1/2}),$$

$$X^n_\eta(t) - X^n(t) = X^n_\eta(s_{i-1}) - X^n(s_{i-1}) + \int_{s_{i-1}}^t [\dot{X}^n_\eta(u) - \dot{X}^n_\eta(s_{i-1}) - \dot{X}^n(u) + \dot{X}^n(s_{i-1})] du$$

$$+ (t - s_{i-1})[\dot{X}^n_\eta(s_{i-1}) - \dot{X}^n(s_{i-1})],$$

$$|X^n_\eta(t) - X^n(t)| \le |X^n_\eta(s_{i-1}) - X^n(s_{i-1})| + (t - s_{i-1}) |\dot{X}^n_\eta(s_{i-1}) - \dot{X}^n(s_{i-1})|$$

$$+ \int_{s_{i-1}}^t M_1(s_{i-1}, s_i; X^n - X^n_\eta) du = O(\eta) + o_P(n^{-1/2}),$$

that is, the order result is true for the i-th subinterval. By induction we conclude the order result is true for all N subintervals, which implies the lemma.

Lemma 6.11. *For $X_\eta(t)$ and $X^n_\eta(t)$ we have the following inequalities, if $\eta < \sqrt{6/L}$,*

$$M_1(0, \eta; X_\eta) \le \frac{|\nabla g(x_0)|}{1 - L\eta^2/6}, \quad M_1(0, \eta; X^n_\eta) \le \frac{|\nabla g(x_0)| + 3n^{-1/2} \sup_{u \in [0,\eta]} |R^n(X^n_\eta(u); \mathbf{U}_n, Q)|}{1 - L\eta^2/6},$$

and if $\eta < \sqrt{6/L}$ and $\eta < t < \sqrt{12/L}$,

$$M_1(0, t; X_\eta) \le \frac{(5 - L\delta^2/6)|\nabla g(x_0)|}{4(1 - L\eta^2/6)(1 - Lt^2/12)},$$

$$M_1(0, t; X^n_\eta) \le \frac{(5 - L\delta^2/6)[|\nabla g(x_0)| + 3n^{-1/2} \sup_{u \in [0,t]} |R^n(X^n_\eta(u); \mathbf{U}_n, Q)|]}{4(1 - L\eta^2/6)(1 - Lt^2/12)}.$$

For $s, t \geq \eta$ and $t - s < \sqrt{6/L}$,

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

$$M_1(s,t;X^n - X^n_\eta) \leq \frac{1}{1 - L(t-s)^2/6} \left[\left(\frac{3}{s} + \frac{L(t-s)}{2} \right) |\dot{X}^n(s) - \dot{X}^n_\eta(s)| \right.$$
$$\left. + 3L|X^n(s) - X^n_\eta(s)| + 3n^{-1/2} \sup_{u \in [s,t]} |R^n(X^n(u); \mathbf{U}_n, Q) - R^n(X^n_\eta(u); \mathbf{U}_n, Q)| \right].$$

Proof. $M_1(s,t;X)$, $M_1(s;t,X^n)$, and $M_1(0,t;X_\eta)$ have been established, and the same arguments can be applied to easily obtain $M_1(0,t;X^n_\eta)$. The difference between $M_1(s,t;X - X_\eta)$ and $M_1(s,t;X^n - X^n_\eta)$ is an extra term with $R^n(\theta;\mathbf{U}_n,Q)$. We will derive $M_1(s,t;X^n - X^n_\eta)$ only. Note that $X^n_\eta(t)$ satisfies the following ODE with random coefficients

$$\ddot{X}^n_\eta(t) + \frac{3}{t \vee \eta} \dot{X}^n_\eta(t) + \nabla g(X^n_\eta(t)) + \frac{1}{\sqrt{n}} R^n(X^n_\eta(t); \mathbf{U}_n, Q) = 0,$$

which has the same form as ODE for $X^n(t)$ for $t \geq \eta$. Define

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

and $J(s,t;H,X^n-X^n_\eta)=\int_s^t u^3[H(u;X^n-X^n_\eta)-H(s;X^n-X^n_\eta)]du$. Then

$$
\begin{aligned}
&|H(t;X^n-X^n_\eta)-H(s;X^n-X^n_\eta)|\le|[\nabla g(X^n(t))-\nabla g(X^n_\eta(t))]\\
&-[\nabla g(X^n(s))-\nabla g(X^n_\eta(s))]|+n^{-1/2}|[R^n(X^n(t);\mathbf{U}_n,Q)-R^n(X^n_\eta(t);\mathbf{U}_n,Q)]\\
&-[R^n(X^n(s);\mathbf{U}_n,Q)-R^n(X^n_\eta(s);\mathbf{U}_n,Q)]|,\\
&|[\nabla g(X^n(t))-\nabla g(X^n_\eta(t))]-[\nabla g(X^n(s))-\nabla g(X^n_\eta(s))]|\\
&\le L|X^n(t)-X^n_\eta(t)|+L|X^n(s)-X^n_\eta(s)|,
\end{aligned}
$$

$$
\begin{aligned}
&X^n(t)-X^n_\eta(t)=\int_s^t[\dot X^n(u)-\dot X^n_\eta(u)]du\\
&=\int_s^t\{[\dot X^n(u)-\dot X^n_\eta(u)]-[\dot X^n(s)-\dot X^n_\eta(s)]\}du+(t-s)[\dot X^n(s)-\dot X^n_\eta(s)].
\end{aligned}
$$

Putting these results together we get

$$
\begin{aligned}
&|H(u;X^n-X^n_\eta)-H(s;X^n-X^n_\eta)|\le L\int_s^u|[\dot X^n(v)-\dot X^n_\eta(v)]-[\dot X^n(s)-\dot X^n_\eta(s)]|dv\\
&+L(u-s)|\dot X^n(s)-\dot X^n_\eta(s)|+2L|X^n(s)-X^n_\eta(s)|\\
&+2n^{-1/2}\sup_{v\in[s,u]}|R^n(X^n(v);\mathbf{U}_n,Q)-R^n(X^n_\eta(v);\mathbf{U}_n,Q)|.
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
&\int_s^u|[\dot X^n(v)-\dot X^n_\eta(v)]-[\dot X^n(s)-\dot X^n_\eta(s)]|dv\\
&\le\int_s^u(v-s)\frac{|[\dot X^n(v)-\dot X^n_\eta(v)]-[\dot X^n(s)-\dot X^n_\eta(s)]|}{v-s}dv\\
&\le\int_s^u(v-s)M_1(s,u;X^n-X^n_\eta)dv=\frac{M_1(s,u;X^n-X^n_\eta)(u-s)^2}{2},\\
&L\int_s^t M_1(s,u;X^n-X^n_\eta)u^3(u-s)^2du/2\le LM_1(s,t;X^n-X^n_\eta)t^3(t-s)^3/6.
\end{aligned}
$$

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

$$
\begin{aligned}
&|J(s,t;H,X^n - X^n_\eta)| \le LM_1(s,t;X^n - X^n_\eta)t^3(t-s)^3/6 + L|\dot{X}^n(s) - \dot{X}^n_\eta(s)|t^3(t-s)^2/2 \\
&+ 2L|X^n(s) - X^n_\eta(s)|t^3(t-s) \\
&+ 2t^3(t-s)n^{-1/2} \sup_{u\in[s,t]} |R^n(X^n(u); \mathbf{U}_n, Q) - R^n(X^n_\eta(u); \mathbf{U}_n, Q)|.
\end{aligned}
$$

The ODE difference between $X^n(t)$ and $X^n_\eta(t)$ for $t \ge \eta$ is equivalent to

$$
\begin{aligned}
&\frac{d\{t^3[\dot{X}^n(t) - \dot{X}^n_\eta(t)]\}}{dt} = -t^3 H(t; X^n - X^n_\eta), \text{ which implies} \\
&t^3[\dot{X}^n(t) - \dot{X}^n_\eta(t)] - s^3[\dot{X}^n(s) - \dot{X}^n_\eta(s)] = -\int_s^t u^3 H(u; X^n - X^n_\eta)du \\
&= -\frac{t^4 - s^4}{4} H(s; X^n - X^n_\eta) - J(s,t;H,X^n - X^n_\eta), \\
&\frac{[\dot{X}^n(t) - \dot{X}^n_\eta(t)] - [\dot{X}^n(s) - \dot{X}^n_\eta(s)]}{t-s} = -\frac{t^3 - s^3}{t^3(t-s)}[\dot{X}^n(s) - \dot{X}^n_\eta(s)] \\
&- \frac{t^4 - s^4}{4t^3(t-s)} H(s; X^n - X^n_\eta) - \frac{J(s,t;H,X^n - X^n_\eta)}{t^3(t-s)}
\end{aligned}
$$

and using the upper bound of $|J(s,t; H, X^n - X^n_\eta)|$ and algebraic manipulation we

get

$$\frac{|[\dot{X}^n(t) - \dot{X}^n_\eta(t)] - [\dot{X}^n(s) - \dot{X}^n_\eta(s)]|}{t-s} \leq \frac{t^3 - s^3}{t^3(t-s)}|\dot{X}^n(s) - \dot{X}^n_\eta(s)|$$
$$+ \frac{t^4 - s^4}{4t^3(t-s)}|H(s; X^n - X^n_\eta)| + \frac{|J(s,t; H, X^n - X^n_\eta)|}{t^3(t-s)}$$
$$\leq \frac{t^2 + st + s^2}{t^3}|\dot{X}^n(s) - \dot{X}^n_\eta(s)| + \frac{(t^2+s^2)(t+s)}{4t^3}|H(s; X^n - X^n_\eta)|$$
$$+ \frac{L}{6}M_1(s,t; X^n - X^n_\eta)(t-s)^2 + \frac{L}{2}|\dot{X}^n(s) - \dot{X}^n_\eta(s)|(t-s) + 2L|X^n(s) - X^n_\eta(s)|$$
$$+ 2n^{-1/2} \sup_{u\in[s,t]} |R^n(X^n(u); \mathbf{U}_n, Q) - R^n(X^n_\eta(u); \mathbf{U}_n, Q)|.$$

As above inequality holds for any $t > s$, using the definition of $M_1(s,t; X^n - X^n_\eta)$ we have

$$M_1(s,t; X^n - X^n_\eta) \leq \frac{3}{s}|\dot{X}^n(s) - \dot{X}^n_\eta(s)| + |H(s; X^n - X^n_\eta)|$$
$$+ \frac{L}{6}M_1(s,t; X^n - X^n_\eta)(t-s)^2 + \frac{L}{2}|\dot{X}^n(s) - \dot{X}^n_\eta(s)|(t-s) + 2L|X^n(s) - X^n_\eta(s)|$$
$$+ 2n^{-1/2} \sup_{u\in[s,t]} |R^n(X^n(u); \mathbf{U}_n, Q) - R^n(X^n_\eta(u); \mathbf{U}_n, Q)|,$$

and solving for $M_1(s,t; X^n - X^n_\eta)$ we obtain

$$M_1(s,t; X^n - X^n_\eta) \leq \frac{1}{1 - L(t-s)^2/6}\left[\left(\frac{3}{s} + \frac{L(t-s)}{2}\right)|\dot{X}^n(s) - \dot{X}^n_\eta(s)|\right.$$
$$+ |H(s; X^n - X^n_\eta)| + 2L|X^n(s) - X^n_\eta(s)|$$
$$\left. +2n^{-1/2} \sup_{u\in[s,t]} |R^n(X^n(u); \mathbf{U}_n, Q) - R^n(X^n_\eta(u); \mathbf{U}_n, Q)|\right],$$

where

$$|H(s; X^n - X^n_\eta)| \leq L|X^n(s) - X^n_\eta(s)| + n^{-1/2}|R^n(X^n(s); \mathbf{U}_n, Q) - R^n(X^n_\eta(s); \mathbf{U}_n, Q)|.$$

Lemma 6.12.

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

Proof. Note that

$$X^n(t) - X^n(s) = \int_s^t [\dot{X}^n(u) - \dot{X}^n(s)]du + (t-s)\dot{X}^n(s).$$

Since $M_1(s,t;X^n)$ is finite, and $\dot{X}^n(t)$ has a finite bound on $[0,T]$, we have order $\delta^{1/2}$ for the variation of $X^n(t)$ over any interval of length $\delta^{1/2}$. Divide $[0,T]$ into subintervals $[(k-1)\delta^{1/2}, k\delta^{1/2}]$. For any $t \in [0,T]$ there is an interval $[(k-1)\delta^{1/2}, k\delta^{1/2}]$ containing t for which $x_\delta^n(t) = x_k^n$ and $|X^n(k\delta^{1/2}) - X^n(t)| = O_P(\delta^{1/2})$, and thus

$$\max_{t\in[0,T]} |x_\delta^n(t) - X^n(t)| \le \max_{k\le T\delta^{-1/2}} |x_k^n - X^n(k\delta^{1/2})| + \max_{k\le T\delta^{-1/2}} \max_{k-1\le t\delta^{-1/2}\le k} |X^n(t) - X^n(k\delta^{1/2})|$$

$$= O_P(\eta + \delta^{1/2}) + o_P(n^{-1/2}),$$

where we use the fact that

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

which is obtained by combining Lemmas 6.7-6.8 and 6.10.
Both $x_\delta^n(t)$ and $X^n(t)$ are free of smoothing parameter η, which can be any sequence approaching zero, so we may take $\eta \le \delta^{1/2}$.
Proof of Theorem 3.2. Lemma 6.12 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 3.1.

6.3 Proof of Theorem 3.3

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

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

where $\mathbf{Z}$ stands for the standard normal random vector, the second equality is due to Assumptions 2 and 4, and the Skorokhod representation, and the law of large number, and the third equality is from $\nabla g(\check{\theta}) = 0$. As $\hat{\theta}_n$ is the minimizer of $\mathcal{L}^n(\theta, \mathbf{U}_n, Q)$, $\hat{\vartheta}_n = n^{1/2}(\hat{\theta}_n - \check{\theta})$ asymptotically minimizes $\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}\vartheta + \Delta g(\check{\theta})\vartheta^2/2$ over ϑ, and thus has an asymptotic distribution $[\Delta g(\check{\theta})]^{-1}\boldsymbol{\sigma}(\check{\theta})\mathbf{Z}$. 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 (see Billingsely (1999) and Jacod and Shiryaev (2002)). Treating $X(t)$, $X^n(t)$, $V(t)$, $V^n(t)$, and $x^n_\delta(t)$ as random elements in $D(\mathbb{R}_+)$, since the weak convergence results established in Theorems 3.1 and 3.2 hold for $X^n(t)$ and $x^n_\delta(t)$ on $D([0,T])$ for any $T > 0$, thus we may 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) and (4) converge to the solution $\check{\theta}$ of (1) with speed of $(\delta k)^{-1}$ and $(\sqrt{\delta}k)^{-2}$, respectively, while as $t \to \infty$, their corresponding continuous curves $X(t)$ as the solutions of ODEs (3) and (6) approach $\check{\theta}$ with speed of t^{-1} and t^{-2}, respectively (see Nesterov (1983) and Su et al. (2016)). Similarly for fixed n, as $k, t \to \infty$,

x_k^n and $x_\delta^n(t)$ from algorithms (8) and (10) and $X^n(t)$ from ODEs (9) and (11) approach the solution $\hat{\theta}_n$ of (7). For the weak limit $V(t)$ governed by (12) or (16), as $t \to \infty$, both ODEs lead to $[\Delta g(X(\infty))]V(\infty) + \boldsymbol{\sigma}(X(\infty))\mathbf{Z} = 0$, or equivalently, $V(\infty) = [\Delta g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\mathbf{Z}$. In fact, the solutions of (12) and (16) admit simple explicit expressions, for example,

$$V(t) = \int_0^t \exp\left[-\int_s^t \Delta g(X(u))du\right]\boldsymbol{\sigma}(X(s))ds\mathbf{Z}, \tag{62}$$

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

$$\int_{t_0}^t \exp\left[-\int_s^t \Delta g(X(u))du\right]\boldsymbol{\sigma}(X(s))ds = \int_{t_0}^t \exp\left[-\int_s^t \Delta g(X(u))du\right]\Delta g(X(s))$$

$$\left\{[\Delta g(X(s))]^{-1}\boldsymbol{\sigma}(X(s)) - [\Delta g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\right\}ds$$

$$+ \int_{t_0}^t \exp\left[-\int_s^t \Delta g(X(u))du\right]\Delta g(X(s))ds[\Delta g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty)). \tag{63}$$

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

$$
\begin{aligned}
&\left|\int_0^{t_0} \exp\left[-\int_s^t \Delta g(X(u))du\right]\boldsymbol{\sigma}(X(s))ds\right| \\
&\leq \left|\exp\left[-\int_{t_0}^t \Delta g(X(u))du\right]\right| \int_0^{t_0} |\boldsymbol{\sigma}(X(s))|ds \to 0, \\
&\int_{t_0}^t \exp\left[-\int_s^t \Delta g(X(u))du\right]\Delta g(X(s))ds = 1 - \exp\left[-\int_{t_0}^t \Delta g(X(s))ds\right] \to 1, \\
&\int_{t_0}^t \left|\exp\left[-\int_s^t \Delta g(X(u))du\right]\Delta g(X(s))\right| \left|[\Delta g(X(s))]^{-1}\boldsymbol{\sigma}(X(s))\right. \\
&\left.-[\Delta g(X(\infty))]^{-1}\boldsymbol{\sigma}(X(\infty))\right| ds \\
&\leq \epsilon - \epsilon\left|\exp\left[-\int_{t_0}^t \Delta g(X(s))ds\right]\right| \leq \epsilon,
\end{aligned}
$$

which goes to zero, as we let $\epsilon \to 0$. Combining these results with (62) and (63) we conclude that as $t \to \infty$,

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

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

6.4 Proofs of Theorems 4.1-4.3

Theorem 4.1 is proved by Lemma 6.13, with Theorems 4.2 and 4.3 shown in Lemma 6.21, where both lemmas are established in this section below.

For simplicity we will prove the case that 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 can be handled through strong approximation (Csörgö and Mason (1989), Csörgö et al. (1999), Massart (1989), Rio (1993a, b)).

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

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

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

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

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

We have

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

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

Lemma 6.13. *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

$$\check{H}^m_\delta(t) = (m\delta)^{1/2}\sum_{k=1}^{[t/\delta]}\left[\int \nabla\ell(X((k-1)\delta); u, Q)\hat{Q}^*_{mk}(du)\right.$$
$$\left. - \int \nabla\ell(X((k-1)\delta); u, Q)Q(du)\right].$$

Note that

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

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

which are the mean and variance of $\nabla \ell(\theta; U, Q)$, 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 $\int_0^t \boldsymbol{\sigma}(X(t))d\mathbf{B}(u)$. Indeed, its finite-dimensional distribution convergence can be easily established through assumptions A3-A4, and its tightness can be shown by the fact that for $r \leq s \leq t$,

$$E\left\{\left|\check{H}^m_\delta(t) - \check{H}^m_\delta(s)\right|^2 \left|\check{H}^m_\delta(s) - \check{H}^m_\delta(r)\right|^2\right\} \leq [\Upsilon(t) - \Upsilon(r)]^2, \tag{64}$$

where $\Upsilon(\cdot)$ is a continuous non-decreasing function on $[0, T]$ (Billingsley (1999, equation (13.14) & theorem 13.5)). To establish (64), we have that, because of independence,

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

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

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

We have shown that as $\delta \to 0$ and $m \to \infty$, $\check{H}^m_\delta(t)$ weakly converges to $H(t)$. Since $\nabla \ell(\theta; u, Q)$ is Lipschitz in θ, and Lemma 6.15 below indicates that as $\delta \to 0$ and $m \to \infty$, $x^m_k - X(k\delta)$ converges to zero in probability uniformly over $1 \le k \le T/\delta$, $\check{H}^m_\delta(t)$ and $H^m_\delta(t)$ share the same weak convergence limit $H(t)$.

Lemma 6.14. *We have*

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

where x_k and x^m_k are defined by (2) and (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

$$\begin{aligned}
&|\nabla \mathcal{L}^m(\theta; \mathbf{U}^*_{mk}, Q) - \nabla \mathcal{L}^m(\vartheta; \mathbf{U}^*_{mk}, Q)| \le \zeta(\mathbf{U}^*_{mk})|\theta - \vartheta|, \;\; |\Delta \mathcal{L}(\theta; \mathbf{U}^*_{mk}, Q)| \le \zeta(\mathbf{U}^*_{mk}),\\
&|\check{\theta} - x^m_k| \le |\check{\theta} - x^m_{k-1}| + \delta|\nabla \mathcal{L}^m(x^m_{k-1}; \mathbf{U}^*_{mk}, Q) - \nabla \mathcal{L}^m(\check{\theta}; \mathbf{U}^*_{mk}, Q)| + \delta|\nabla \mathcal{L}^m(\check{\theta}; \mathbf{U}^*_{mk}, Q)|\\
&\le (1 + \delta\zeta(\mathbf{U}^*_{mk}))|\check{\theta} - x^m_{k-1}| + \delta|\nabla \mathcal{L}^m(\check{\theta}; \mathbf{U}^*_{mk}, Q)|\\
&\le \big(1 + \delta E[h_1(U)] + O_P(\delta m^{-1/2})\big)^k\\
&+ \big(1 + \delta E[h_1(U)] + O_P(\delta m^{-1/2})\big)^k \delta \sum_{j=1}^k [|\nabla g(\check{\theta})| + m^{-1/2}|R^m(\check{\theta}; \mathbf{U}^*_{mj}, Q)|]\\
&= e^{E[h_1(U)]}[1 + |\nabla g(\check{\theta})| + O_P(m^{-1/2})],
\end{aligned}$$

namely, x^m_k are bounded uniformly over $k \le T/\delta$. On the other hand, we have

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

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

$$|[1 - \delta\Delta\mathcal{L}^m(x^m_{\xi,j-1}; \mathbf{U}^*_{mj}, Q)]^j| \le [1 + \delta\zeta(\mathbf{U}^*_{mj})]^{T/\delta} \le e^{TE[h_1(U)]}[1 + O_P(m^{-1/2})],$$
$$R^m(x_{j-1}; \mathbf{U}^*_{mj}, Q) \sim \boldsymbol{\sigma}(x_{j-1})\mathbf{Z} = O_P(1),$$
$$|x_k - x^m_k| \le \delta m^{-1/2} \sum_{j=1}^{k} |1 + \delta\zeta(\mathbf{U}^*_{mj})]^{T/\delta} |R^m(x_{j-1}; \mathbf{U}^*_{mj}, Q)| = O_P(k\delta m^{-1/2}) = O_P(m^{-1/2}).$$

Lemma 6.15.

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

where $X(t)$ and x^m_k are defined by (3) and (18), respectively.

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

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

are martingale differences with conditional mean zero and conditional variance $\boldsymbol{\sigma}^2(x^m_{k-1})$. Since x_k in (2) is the Euler approximation of solution $X(t)$ of ODE (3), the standard ODE theory shows

$$\max_{k \le T/\delta} |x_k - X(k\delta)| = O(\delta). \tag{65}$$

By Lemma 6.14 we have that with probability tending to one, $x^m_{k-1}, k = 1, \cdots, T/\delta$, fall inside a neighborhood of solution curve of ODE (3), and thus the maximum of $\boldsymbol{\sigma}^2(x^m_{k-1})$, $k = 1, \cdots, T/\delta$, is bounded. Applying Burkholder's inequality we obtain

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

that is,

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

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

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

With the same initial value x_0, comparing the expressions for x_k and x^m_k we can show by induction that

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

The lemma is a consequence of above result and (65).
The following lemma refines the order regarding $m^{-1/2}\delta^{1/2}$ in Lemma 6.15.

Lemma 6.16. *We have*

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

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

where $X^m_\delta(t)$ is given by (21), and x^m_k and $x^m_\delta(t)$ are defined by (18) and (19), respectively.

Proof. With weak convergence of $H^m_\delta(t)$ to $H(t)$ in Lemma 6.13, by Skohorod's representation we may realize $H^m_\delta(t)$ and $H(t)$ on some common probability spaces such that as $\delta\to 0$ and $m\to\infty$, under the metric in $D([0,T])$, $H^m_\delta(t)-H(t)$ is $o_P(1)$. Also we may consider linear interpolation $\tilde{H}^m_\delta(t)$ between the values of

$H_\delta^m(k\delta)$, $k = 1, \cdots, T/\delta$, which satisfies

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

By assumptions A1-A2, we have

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

and then

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

Lemma 6.15 implies $m^{1/2} \max_{k\le T/\delta} |x_{k-1}^m - X((k-1)\delta)| = m^{1/2} O_P(\delta + m^{-1/2}\delta^{1/2}) = O_P(m^{1/2}\delta + \delta^{1/2}) = o_P(1)$, and by Lemma 6.17 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 Skohorod's representation we may 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 $1 \le k \le T/\delta$,

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

$$x_k^m - x_{k-1}^m = -\delta \nabla g(x^m(t_{k-1})) - m^{-1/2}\delta^{1/2}[H_\delta^m(k\delta) - H_\delta^m((k-1)\delta)],$$

$$x_k^m = x_0 - \delta \sum_{\ell=1}^{k} \nabla g(x_{\ell-1}^m) - m^{-1/2}\delta^{1/2} H_\delta^m(k\delta)$$

$$= x_0 - \delta \sum_{\ell=1}^{k} \nabla g(x_{\ell-1}^m) - m^{-1/2}\delta^{1/2} H(k\delta) + o_P(m^{-1/2}\delta^{1/2}).$$

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

$$\check{x}_k^m - \check{x}_{k-1}^m = -\delta \nabla g(\check{x}_{k-1}^m) - m^{-1/2}\delta^{1/2}[H(k\delta) - H((k-1)\delta)]. \tag{66}$$

Then comparing the expressions for x_k^m and $\check{x}_k^m$ we can 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 above result and Lemma 6.18 below.

Lemma 6.17.

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

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

Proof. Direct calculations lead to

$$P\left(\delta^{1/2} \max_{k \le T/\delta} |R^m(X((k-1)\delta); \mathbf{U}_{mk}^*, Q)| > \delta^{1/4} |\log \delta| \right)$$

$$= 1 - \prod_{k \le T/\delta} P\left(\delta^{1/2} |R^m(X((k-1)\delta); \mathbf{U}_{mk}^*, Q)| \le \delta^{1/4} |\log \delta| \right)$$

$$\le 1 - \prod_{k \le T/\delta} \left[1 - \delta E\left\{ |R^m(X((k-1)\delta); \mathbf{U}_{mk}^*, Q)|^4 \right\} / |\log \delta|^4 \right]$$

$$\le 1 - \exp\left[-2T\tau / |\log \delta|^4 \right] \sim 2T\tau / |\log \delta|^4 \to 0,$$

where we use Chebyshev's inequality, $\log(1-u) \geq -2u$ for $0 < u < 1$, and $\tau = \sup_{t,k} E\{|R^m(X(t); \mathbf{U}^*_{mk}, Q)|^4\}$ whose finiteness will be shown below. Indeed, we have

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

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

$$
\begin{aligned}
&\sup_{t,k} E[\{\nabla \ell(X(t); U_{1k}, Q) - \nabla g(X(t))\}^2] \leq 2 \sup_{t \geq 0} \left\{ |X(t) - x_0|^2 \right\} E[h_1^2(U)] \\
&+ 2E[\{\nabla \ell(x_0, U)\}^2] + 2 \sup_{t \geq 0} \{[\nabla g(X(t))]^2\}, \\
&\sup_{t,k} E[\{\nabla \ell(X(t); U_{1k}, Q) - \nabla g(X(t))\}^4] \leq 64 \sup_{t \geq 0} \left\{ |X(t) - x_0|^4 \right\} E[h_1^4(U)] \\
&+ 64E[\{\nabla \ell(x_0, U)\}^4] + 8 \sup_{t \geq 0} \{[\nabla g(X(t))]^4\},
\end{aligned}
$$

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

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

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

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

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

Lemma 6.18.

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

where $\check{x}_k^m$ and $X_\delta^m(t)$ are defined by (66) and (21), respectively.

Proof. By (21) we have

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

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

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

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

Note that $\check{x}_k^m$ are the Euler approximation of SDE (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

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

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

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

For the general k, we obtain

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

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

and thus

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

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

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

Lemma 6.19.

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

where $X(t)$ and $X_\delta^m(t)$ are defined by (3) and (21), respectively.

Proof. With the same initial value for $X(t)$ and X_δ^m, from (3) and (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 get

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

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.

Lemma 6.20.

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

where $X_\delta^m(t)$ and $\check{X}_\delta^m(t)$ are the solutions of (21) and (22), respectively.

Proof.

$$\begin{aligned}
&|X_\delta^m(t)-\check{X}_\delta^m(t)| \le \int_0^t |\nabla g(X_\delta^m(u))-\nabla g(\check{X}_\delta^m(u))|du\\
&+m^{-1/2}\delta^{1/2}\left|\int_0^t[\boldsymbol{\sigma}(X(u))-\boldsymbol{\sigma}(\check{X}_\delta^m(u))]d\mathbf{B}(u)\right|\\
&\le C\int_0^t |X_\delta^m(u)-X(u)|du+m^{-1/2}\delta^{1/2}\left|\int_0^t[\boldsymbol{\sigma}(X(u))-\boldsymbol{\sigma}(\check{X}_\delta^m(u))]d\mathbf{B}(u)\right|.\\
&E[|X_\delta^m(t)-\check{X}_\delta^m(t)|^2] \le C\int_0^t E[|X_\delta^m(u)-\check{X}_\delta^m(u)|^2]du\\
&+2m^{-1}\delta E\left[\left|\int_0^t[\boldsymbol{\sigma}(X(u))-\boldsymbol{\sigma}(\check{X}_\delta^m(u))]d\mathbf{B}(u)\right|^2\right]\\
&\le C\int_0^t E[|X_\delta^m(u)-\check{X}_\delta^m(u)|^2]du+2m^{-1}\delta\int_0^t E[|\boldsymbol{\sigma}(X(u))-\boldsymbol{\sigma}(\check{X}_\delta^m(u))|^2]du\\
&\le C\int_0^t E[|X_\delta^m(u)-\check{X}_\delta^m(u)|^2]du+C_1m^{-1}\delta\int_0^t E[|X(u)-X_\delta^m(u)|^2]du\\
&+C_1m^{-1}\delta\int_0^t E[|X_\delta^m(u)-\check{X}_\delta^m(u)|^2]du,
\end{aligned}$$

where the last inequality is due to

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

The Gronwall inequality leads to

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

Using Lemma 6.19 we have

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

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

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

and the lemma is a consequence of above result and Lenglart's inequality for semi-martingale.

Lemma 6.21. *As $\delta \to 0$, and $m,n \to \infty$, we have $V^m_\delta(t)$ and $\check{V}^m_\delta(t)$ both weakly converge to $V(t)$. Moreover, if $m(n\delta)^{1/2} \to 0$, and $m^{1/2}\delta|\log\delta|^{1/2} \to 0$, $(mT/\delta)^{1/2}[x^m_\delta(t)-X(t)]$ weakly converges to $V(t)$.*

Proof. As the solutions of (21) and (22) have difference of order $m^{-1}\delta$, they have the same asymptotic distribution, and we can easily establish the result for $\check{V}^m_\delta(t)$ by that for $V^m_\delta(t)$ and Lemma 6.16.
Let consider the easier one for $X^m_\delta(t)$. From (21) and (3), we have

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

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

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

where X_ξ is between $X(u)$ and $X_\delta^m(u)$ and thus Lemma 6.19 shows that uniformly over $[0,T]$,

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

Then

$$V_\delta^m(t) = -\int_0^t [\Delta g(X_\xi)]V_\delta^m(u)du - \int_0^t \boldsymbol{\sigma}(X(u))d\mathbf{B}(u). \tag{67}$$

First as $\delta \to 0$, $m,n \to \infty$, equation (67) converges to (24).
We need to show stochastic equicontinuity for $V_\delta^m(t)$. From (25) 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 (67) indicates that for any $s,t \in [0,T]$, and $t \in [s, s+\gamma]$,

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

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

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

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

6.5 Proof of Theorem 4.4

Theorem 4.4 can be proved by the same proof argument of Theorem 4.1 except for changing step size from δ to $\delta^{1/2}$.

6.6 Proof of Theorems 4.5

6.6.1 The unique solution of the second order SDEs

In this section we will prove Lemma 6.28 blow that the second order SDEs (35) (with fixed (δ, m)) and (36) have unique (weak) solutions in the distributional sense. Due to the similarity we provide proof arguments for (36) only. Consider the 2nd order SDE (36) with initial conditions $V(0) = c$ and $\dot{V}(0) = 0$, where $\mathbf{B}(t)$ 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 a white noise in a sense that for any smooth function $h(t)$ with compact support,

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

where the right hand side is Itô integral.
The second order SDE (36) is equivalent to

$$Y(t) = V(t) + \frac{t}{2}\dot{V}(t), \;\; \dot{Y}(t) = -\frac{t}{2}[\nabla g(X(t))]V(t) - \frac{t}{2}\boldsymbol{\sigma}(X(t))\dot{\mathbf{B}}(t), \tag{68}$$

where $V(0) = c, \dot{V}(0) = 0$, $Y(0) = V(0) = c$, $\dot{Y}(0) = \frac{3}{2}\dot{V}(0) + \frac{0}{2}\ddot{V}(0) = 0$. Denote

by $(V_\eta(t), Y_\eta(t))$ the solution of the smoothed stochastic equation system

$$\dot{V}(t) = \frac{2}{t \vee \eta} Y(t) - \frac{2}{t \vee \eta} V(t), \;\; \dot{Y}(t) = -\frac{t}{2} [\nabla g(X(t))] V(t) - \frac{t}{2} \boldsymbol{\sigma}(X(t)) \dot{\mathbf{B}}(t), \;\; (69)$$

where $V(0) = Y(0) = c, \dot{V}(0) = \dot{Y}(0) = 0$.
We need notation $M_a(s,t;Y)$ defined in (48). In the proofs of Theorems 3.1 and 4.1, we have employed $M_a(s,t;Y)$ with $a = 1$, as curves and processes are solutions of ODE and thus differentiable. For this part of proofs we need to handle 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 6.22.

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

Proof. We prove the lemma by the following direction calculation,

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

Lemma 6.23. *For $\eta < [(1+a)(2+a)/L]^{1/2}$, we have*

$$M_a(0,\eta;V_\eta) \le \frac{1}{1-L\eta^2/(1+a)(2+a)}\left[|\nabla g(X(0))V_\eta(0)| + \frac{LM_a(0,t;X)\eta^{1+a}}{(1+a)(2+a)}\right.$$

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

Proof.

$$|\nabla g(X(u))V_\eta(u) - \nabla g(X(0))V_\eta(0)| \le L|X(u)-X(0)| + L|V_\eta(u)-V_\eta(0)|$$
$$\le L|\int_0^u \dot{X}(v)dv| + L|\int_0^u \dot{V}_\eta(v)dv|$$
$$\le L|\int_0^u v^a\frac{\dot{X}(v)}{v^a}dv| + L|\int_0^u v^a\frac{\dot{V}_\eta(v)}{v^a}dv| \le L[M_a(0,u;X)+M_a(0,u;V_\eta)]u^{1+a}/(1+a).$$

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,

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

The lemma is proved by taking maximum over $t \in (0,\eta]$ with simple algebra manipulation of above inequality.

Lemma 6.24. *For $t-\eta < [(1+a)(2+a)/L]^{1/2}$ and $\eta < t < 2[(1+a)(2+a)/L]^{1/2}$, we have*

$$M_a(0,t;V_\eta)\left[1-\frac{L(t-\eta)^2}{(1+a)(2+a)}\right] \leq \left[M_a(0,\eta;V_\eta) + \frac{t^{1-a}}{4}|\nabla g(X(0))V_\eta(0)|\right.$$
$$\left. + \frac{L(t-\eta)^2}{(1+a)(2+a)}M_a(0,t;X) + \max_{t_0\in(\eta,t]}\left|\frac{1}{(t_0-\eta)^a}\int_\eta^{t_0} u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\right].$$

Proof. 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

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

Thus,

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

Taking maximum over $t > \eta$, and using the fact that the right hand side of above inequality is increasing in t, we conclude that

$$M_a(0,t;V_\eta) \leq M_a(0,\eta;V_\eta) + \frac{t^{1-a}}{4}|\nabla g(X(0))V_\eta(0)|$$
$$+ \frac{(t-\eta)^2}{(1+a)(2+a)}[M_a(0,t;X) + M_a(0,t;V_\eta)] + \max_{t_0\in(\eta,t]}\left|\frac{1}{(t_0-\eta)^a}\int_\eta^{t_0} u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|,$$

which leads to the lemma.

Lemma 6.25. *For $t-s<[(1+a)(2+a)/L]^{1/2}$ and $s<t<2[(1+a)(2+a)/L]^{1/2}$, we have*

$$M_a(s,t;V_\eta)\le\left[1-\frac{L(t-s)^2}{(1+a)(2+a)}\right]^{-1}\left[\frac{3}{t^{1-a}}|\dot V_\eta(s)|+\frac{t^{1-a}}{4}|\nabla g(X(0))V_\eta(0)|\right.$$
$$+\frac{(t-s)^2}{(1+a)(2+a)}M_a(s,t;X)+\frac{(t-s)^{2-a}}{2}[L|V_\eta(s)||\dot X(s)|+|\nabla g(X(*))||\dot V_\eta(s)|]$$
$$\left.+\max_{v\in(s,t]}\left|\frac{1}{v^3(v-s)^a}\int_s^v u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|\right].$$

Proof. We consider the case of $s,t>\eta$. As 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

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

Thus,

$$\frac{|\dot{V}_\eta(t)-\dot{V}_\eta(s)|}{(t-s)^a} \le \frac{t^3-s^3}{t^3}\frac{|\dot{V}_\eta(s)|}{(t-s)^a} + \frac{t^4-s^4}{4t^3(t-s)^a}|\nabla g(X(s))V_\eta(s)|$$
$$+ \frac{L}{(1+a)t^3(t-s)^a}\int_s^t [M_a(s,u;X)+M_a(s,u;V_\eta)]u^3(u-s)^{1+a}du$$
$$+ \frac{1}{t^3(t-s)^a}\int_s^t [L|V_\eta(s)||\dot{X}(s)| + |\nabla g(X(u))||\dot{V}_\eta(s)|]u^3(u-s)du$$
$$+ \frac{1}{t^3(t-s)^a}|\int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du|.$$
$$\le \frac{3}{t^{1-a}}|\dot{V}_\eta(s)| + t^{1-a}|\nabla g(X(s))V_\eta(s)| + \frac{Lt^3[(t-s)^{2+a}]}{(1+a)(2+a)t^3(t-s)^a}[M_a(s,t;X)+M_a(s,t;V_\eta)]$$
$$+ \frac{(t-s)^{2-a}}{2}[L|V_\eta(s)||\dot{X}(s)| + |\nabla g(X(*))||\dot{V}_\eta(s)|] + \frac{1}{t^3(t-s)^a}|\int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du|$$
$$\le \frac{3}{t^{1-a}}|\dot{V}_\eta(s)| + \frac{t^{1-a}}{4}|\nabla g(X(0))V_\eta(0)| + \frac{L(t-s)^2}{(1+a)(2+a)}[M_a(s,t;X)+M_a(s,t;V_\eta)]$$
$$+ \frac{(t-s)^{2-a}}{2}[L|V_\eta(s)||\dot{X}(s)| + |\nabla g(X(*))||\dot{V}_\eta(s)|] + \frac{1}{t^3(t-s)^a}|\int_s^t u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du|.$$

Taking maximum over $t > \eta$, and using the fact that the right hand side of above inequality is increasing in t, we conclude that

$$M_a(s,t;V_\eta) \le \frac{3}{t^{1-a}}|\dot{V}_\eta(s)| + \frac{t^{1-a}}{4}|\nabla g(X(0))V_\eta(0)| + \frac{(t-s)^2}{(1+a)(2+a)}[M_a(s,t;X)+M_a(s,t;V_\eta)]$$

$$+ \frac{(t-s)^{2-a}}{2}[L|V_\eta(s)||\dot{X}(s)| + |\nabla g(X(*))||\dot{V}_\eta(s)|] + \max_{t_0\in(s,t]}\left|\frac{1}{t_0^3(t_0-s)^a}\int_s^{t_0} u^3\boldsymbol{\sigma}(X(u))\dot{\mathbf{B}}(u)du\right|,$$

which leads to the lemma.

Lemma 6.26. *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 Gaussian process $\int_s^v u^3\boldsymbol{\sigma}(X(u))d\mathbf{B}(u)$ has the a-th derivative. Indeed, note that all $u > s > 0$, we have that $\sup_{s<u} \frac{|\mathbf{B}(u)-\mathbf{B}(s)|}{(u-s)^a}$ is a.s. finite. Then

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

which is a.s. finite, since all random terms are a.s. finite, and $0 < (u-s)^a/(v-s)^a \leq 1$.

Lemma 6.27. *For any given $T > 0$, $V_\eta(t)$ is stochastically equicontinuous and stochastically bounded on $[0, T]$.*

Proof. Divide the interval $[0, T]$ into $N = [T\sqrt{L/(1+a)(2+a)}] + 1$ number of subintervals with length almost equal to $\sqrt{(1+a)(2+a)/L}$ (except for the last one), and denote by $\mathcal{I}_i = [s_{i-1}, s_i]$, $i = 1, \cdots, N$ (with $s_0 = 0$, $s_N = T$, $\mathcal{I}_1 = [0, T/N]$, $1/N < \sqrt{(1+a)(2+a)/L}$, $\mathcal{I}_N = [s_{N-1}, T]$). First for $t \in \mathcal{I}_1$, we have

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

and the upper bounds on $\dot{V}_\eta(t)$ and $V_\eta(t)$ over $\mathcal{I}_1$ are a.s. finite uniformly over δ, 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

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

Note that N is free of η. We will use 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. Above 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 6.28. *For fixed (δ, m), the second order SDEs (35) and (36) have unique solutions in the distributional sense.*

Proof. Due to the similarity we provide proof arguments for (36) only. Take a decreasing sequence of η as follows: η_k, $k = 1, 2, \cdots$, are decreasing, and as $k \to \infty$, $\eta_k \to 0$. Lemma 6.27 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 will show that $V_\dagger(t)$ satisfies (36). Without loss of generality, we may assume $V_{\eta_k}(t)$ weakly converges to $V_\dagger(t)$, and using SKorohod's representation theorem we may further assume that $V_{\eta_k}(t)$ converges to $V_\dagger(t)$ a.s. $V_{\eta_k}(t)$ obey the initial condition $V_{\eta_k}(0) = \dot{V}_{\eta_k}(0) = 0$, thus $V_\dagger(0) = 0$, and

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

Since $M_a(0,t,V_{\eta_k})$ is a.s. finite uniformly over η_k, taking $t \to 0$ we obtain $\dot{V}_\dagger(0) = 0$.

Note that $(V_{\eta_k}(t), Y_{\eta_k}(t))$ satisfy

$$\begin{aligned}
\dot{V}_{\eta_k}(t) &= \frac{2}{t \vee \eta_k} Y_{\eta_k}(t) - \frac{2}{t \vee \eta_k} V_{\eta_k}(t) \\
\dot{Y}_{\eta_k}(t) &= -\frac{t \vee \eta_k}{2} [\nabla g(X(t))] V_{\eta_k}(t) - \frac{t \vee \eta_k}{2} \boldsymbol{\sigma}(X(t)) \dot{\mathbf{B}}(t) \\
V_{\eta_k}(0) &= Y_{\eta_k}(0) = 0, \dot{V}_{\eta_k}(0) = \dot{Y}_{\eta_k}(0) = 0
\end{aligned}$$

The right hand side of the second equation implies that as $k \to \infty$, $Y_{\eta_k}(t)$ converges to $Y(t)$ defined by

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

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 $V_\dagger(t)$ 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 $\mathbf{B}(t) = \mathbf{B}_*(t)$. Then $U(t) = V(t) - V_*(t)$ obey

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

which has a unique solution zero, as it is a second order ODE similar to ODEs (6) and (13). Thus $V(t) = V_*(t)$, that is, the two solutions have an identical distribution, which proves the unique solution.

6.6.2 Weak convergence of $V_\delta^m(t)$

Lemma 6.29. *For $X(t)$, $X_\delta^m(t)$ and $V_\delta^m(t)$ we have the following inequalities,*

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

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

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

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

$$|H(t;V_\delta^m)| \le L\delta^{-1/4}m^{1/2}|X_\delta^m(t) - X(t)| = L|V_\delta^m(t)|,$$
$$|H(t;V_\delta^m) - H(s;V_\delta^m)| = \delta^{-1/4}m^{1/2}|\nabla[g(X_\delta^m(t)) - g(X_\delta^m(s)) - g(X(t)) + g(X(s))]|$$
$$\le L\delta^{-1/4}m^{1/2}|X_\delta^m(t) - X(t)| + L\delta^{-1/4}m^{1/2}|X_\delta^m(s) - X(s)| = L|V^n(t)| + L|V_\delta^m(s)|,$$
$$V_\delta^m(t) = \int_s^t \dot{V}_\delta^m(u)du + V_\delta^m(s) = \int_s^t [\dot{V}_\delta^m(u) - \dot{V}^n(s)]du + V_\delta^m(s) + (t-s)\dot{V}^n(s),$$

$$|H(t;V_\delta^m) - H(s;V_\delta^m)| \le L\int_s^t |\dot{V}_\delta^m(u) - \dot{V}_\delta^m(s)|du + L[2|V_\delta^m(s)| + |(t-s)\dot{V}_\delta^m(s)|],$$
$$\int_s^t |\dot{V}_\delta^m(u) - \dot{V}_\delta^m(s)|du \le \int_s^t (u-s)^a \frac{|\dot{V}_\delta^m(u) - \dot{V}_\delta^m(s)|}{(u-s)^a}du \le \int_s^t (u-s)^a M_a(s,t;V_\delta^m)du$$
$$= \frac{M_a(s,t;V_\delta^m)(t-s)^{a+1}}{a+1},$$
$$\frac{L}{a+1}\int_s^t M_a(s,u;V_\delta^m)u^3(u-s)^{a+1}du \le \frac{LM_a(s,t;V_\delta^m)t^3(t-s)^{a+2}}{(a+1)(a+2)},$$
$$|J(s,t;H,V_\delta^m)| \le \frac{Lt^3(t-s)^{a+2}}{(a+1)(a+2)}M_a(s,t;V_\delta^m) + L[2|V_\delta^m(s)| + (t-s)|\dot{V}_\delta^m(s)|]t^3(t-s).$$

The SDE (35) is equivalent to

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

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

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

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

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

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

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

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

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

which proves the lemma.

Lemma 6.30. *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 $N = \left[T\sqrt{2L/\{(a+1)(a+2)\}}\right]+1$ number of subintervals with length $\sqrt{(a+1)(a+2)/(2L)}$ (except for the last one) and denote by $\mathcal{I}_i = [s_{i-1}, s_i]$, $i = 1, \cdots, N$ (with $s_0 = 0$, $s_N = T$, $\mathcal{I}_1 = [0, \sqrt{3/L}]$, $\mathcal{I}_N = [s_{N-1}, T]$). First for $t \in \mathcal{I}_1$, from Lemma 6.29 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 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 6.29 we have

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

We will use 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). Above 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 6.31. *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 6.30 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 4.5. Lemma 6.28 shows the unique solutions of SDEs. As in Section 6.1.2, we can easily establish finite distribution convergence for $V_\delta^m(t)$. Lemma 6.31 together with the finite distribution convergence immediately lead to that as $\delta\to 0$ and $m\to\infty$, $V_\delta^m(t)$ weakly converges to $V(t)$.

6.7 Proof of Theorem 4.6

Part (i) can be proved by using the same argument for showing Theorem 3.3. First we will show parts (ii) and (iii) in one dimension. From solution (25) of SDE (24) we find that $V(t)$ follows a normal distribution with mean zero and variance

$$\Gamma(t)=\int_0^t\exp\left[-2\int_u^t\Delta g(X(v))dv\right]\boldsymbol{\sigma}^2(X(u))du.$$

It is easy to check that $\Gamma(t)$ satisfies ODE

$$\dot{\Gamma}(t)+2[\Delta g(X(t))]\Gamma(t)+\boldsymbol{\sigma}^2(X(t))=0,$$

and show that the limit $\Gamma(\infty)$ of $\Gamma(t)$ as $t \to \infty$ is equal to

$$\Gamma(\infty) = \boldsymbol{\sigma}^2(X(\infty))[2\Delta g(X(\infty))]^{-1}.$$

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

$$P(\theta) \propto \exp\left\{ -\frac{m}{\delta \boldsymbol{\sigma}^2(\check{\theta})} g(\theta) \right\}.$$

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

$$P_0(v) \propto \exp\left\{ -\frac{m}{\delta \boldsymbol{\sigma}^2(\check{\theta})} g\left(\check{\theta} + \sqrt{\delta/m}\, v \right) \right\} \sim \exp\left\{ -\frac{2m}{\delta \boldsymbol{\sigma}^2(\check{\theta})} \left[g(\check{\theta}) + \frac{\delta \Delta g(\check{\theta})}{2m} v^2 \right] \right\}$$
$$\propto \exp\left\{ -\frac{\Delta g(\check{\theta})}{\boldsymbol{\sigma}^2(\check{\theta})} v^2 \right\},$$

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 $N\left(0, \frac{\boldsymbol{\sigma}^2(\check{\theta})}{2\Delta(\check{\theta})}\right)$, and we conclude that $V_\delta^m(\infty)$ has a limiting normal distribution with mean zero and variance $\boldsymbol{\sigma}^2(\check{\theta})[2\Delta(\check{\theta})]^{-1} = \Gamma(\infty)$.

Similarly we can show parts (ii) and (iii) in the multivariate case by following matrix

arguments in Gardiner (2009, chapters 4 & 6) and Da Prato and Zabczyk (1996, chapter 9) as follows. Using the explicit solution (25) of SDE (24) we find that $V(t)$ follows a normal distribution with mean zero and variance matrix

$$\begin{aligned}\Gamma(t) &= \int_0^t \exp\left[-\int_u^t \Delta g(X(v))dv\right] \boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' \exp\left[-\int_u^t \Delta g(X(v))dv\right] du \\ &= \int_0^t \exp\left[-\int_u^t \Delta g(X(v))dv\right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_u^t \Delta g(X(v))dv\right] du + \zeta_t,\end{aligned} \tag{70}$$

where

$$\begin{aligned}\zeta_t = &\int_0^t \exp\left[-\int_u^t \Delta g(X(v))dv\right] \{\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\} \\ &\exp\left[-\int_u^t \Delta g(X(v))dv\right] du.\end{aligned}$$

Similar to the proof for Part 3 of Theorem 3.3, we can show that as $t \to \infty$, $|\zeta_t| \to 0$. Indeed, for any $\epsilon > 0$, there exists $t_0 > 0$ such that for any $u > t_0$,

$$\begin{aligned}&|\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'| < \epsilon, \quad \left|\Delta g(X(u))[\Delta g(X(\infty))]^{-1}\right| < 1 + \epsilon, \\ &\left|\int_0^{t_0} \exp\left[-\int_u^t \Delta g(X(v))dv\right] \{\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\}\right. \\ &\left.\exp\left[-\int_u^t \Delta g(X(v))dv\right] du\right| \\ &\le \left|\exp\left[-2\int_{t_0}^t \Delta g(X(v))dv\right]\right| \int_0^{t_0} |\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'|\, du \\ &\le C\left|\exp\left[-2\int_{t_0}^t \Delta g(X(v))dv\right]\right| \to 0,\end{aligned}$$

$$
\begin{aligned}
&\left| \int_{t_0}^{t} \exp\left[-\int_u^t \Delta g(X(v))dv \right] \{\boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' - \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'\} \right. \\
&\left. \exp\left[-\int_u^t \Delta g(X(v))dv \right] du \right| \\
&\leq \frac{\epsilon}{1+\epsilon} \int_{t_0}^{t} \left| \exp\left[-2\int_u^t \Delta g(X(v))dv \right] \Delta g(X(u)) \right| du |\Delta g(X(\infty))|^{-1} \\
&\leq \frac{\epsilon}{2(1+\epsilon)} \left| 1 - \exp\left[-2\int_{t_0}^t \Delta g(X(v))dv \right] \right| |\Delta g(X(\infty))|^{-1} \\
&\leq \frac{\epsilon}{2(1+\epsilon)} |\Delta g(X(\infty))|^{-1} \to 0, \text{ as we let } \epsilon \to 0,
\end{aligned}
$$

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 will verify the detailed balance condition using (70) and $\epsilon_t \to 0$. Direct algebraic manipulations show

$$
\begin{aligned}
&\Delta g(X(t))\Gamma(t) + \Gamma(t)\Delta g(X(t)) \\
&= \int_0^t \Delta g(X(t)) \exp\left[-\int_u^t \Delta g(X(v))dv \right] \boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' \exp\left[-\int_u^t \Delta g(X(v))dv \right] du \\
&+ \int_0^t \exp\left[-\int_u^t \Delta g(X(v))dv \right] \boldsymbol{\sigma}(X(u))[\boldsymbol{\sigma}(X(u))]' \exp\left[-\int_u^t \Delta g(X(v))dv \right] \Delta g(X(t))du \\
&= \int_0^t \Delta g(X(t)) \exp\left[-\int_u^t \Delta g(X(v))dv \right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_u^t \Delta g(X(v))dv \right] du \\
&+ \int_0^t \exp\left[-\int_u^t \Delta g(X(v))dv \right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_u^t \Delta g(X(v))dv \right] \Delta g(X(t))du \\
&+ \Delta g(X(t))\zeta_t + \zeta_t \Delta g(X(t)) \\
&= \int_0^t \frac{d}{du} \left\{ \exp\left[-\int_u^t \Delta g(X(v))dv \right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_u^t \Delta g(X(v))dv \right] \right\} du \\
&+ \Delta g(X(t))\zeta_t + \zeta_t \Delta g(X(t))
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \\
&- \exp\left[-\int_0^t \Delta g(X(v))dv\right] \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]' \exp\left[-\int_0^t \Delta g(X(v))dv\right] \\
&+ \Delta g(X(t))\zeta_t + \zeta_t \Delta g(X(t)),
\end{aligned}
$$

where by the assumption we have that as $t \to \infty$, $\int_0^t \Delta g(X(v))dv \to \infty$, which together with $\epsilon_t \to 0$ indicate that the last three terms on the right hand size of above expression go to zero. Hence we have shown that as $t \to \infty$, $\Delta g(X(t))\Gamma(t) + \Gamma(t)\Delta g(X(t)) \to \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'$, that is, their limits obey the following detailed balance condition,

$$
\Delta g(X(\infty))\Gamma(\infty) + \Gamma(\infty)\Delta g(X(\infty)) = \boldsymbol{\sigma}(X(\infty))[\boldsymbol{\sigma}(X(\infty))]'. \tag{71}
$$

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 (71) 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 $\Gamma(\infty)$.

6.8 Proof of Theorem 4.7

By Taylor expansion we have

$$
\begin{aligned}
&g(X_\delta^m(t)) = g(X(t)) + \delta m^{-1/2}\nabla g(X(t))V(t) + \frac{\delta^2}{2m}\Delta g(X(t))[V(t)]^2 + o_P(\delta^2/m),\\
&g(X(t)) \sim g(\check{\theta}) + \nabla g(\check{\theta})[X(t)-\check{\theta}] + \frac{1}{2}\Delta g(\check{\theta})[X(t)-\check{\theta}]^2 = g(\check{\theta}) + \frac{1}{2}\Delta g(\check{\theta})[X(t)-\check{\theta}]^2,\\
&\nabla g(X(t)) \sim \Delta g(\check{\theta})[X(t)-\check{\theta}],\ \ \Delta g(X(t)) \sim \Delta g(\check{\theta}),\\
&g(X_\delta^m(t)) \sim g(\check{\theta}) + \frac{1}{2}\Delta g(\check{\theta})[X_\delta^m(t)-\check{\theta}]^2\\
&= g(\check{\theta}) + \frac{1}{2}\Delta g(\check{\theta})\left\{[X(t)-\check{\theta}]^2 + 2[X(t)-\check{\theta}][X_\delta^m(t)-X(t)] + [X_\delta^m(t)-X(t)]^2\right\}\\
&= g(\check{\theta}) + \frac{1}{2}\Delta g(\check{\theta})\left\{[X(t)-\check{\theta}]^2 + 2\eta[X(t)-\check{\theta}]V(t) + \eta^2[V(t)]^2\right\}.
\end{aligned}
$$

Similar to the stationary distribution part of the proof for Theorem 4.6, we can derive the stationary distribution when $\Delta g(\check{\theta})$ is positive definite. For the saddle point case, for simplicity we assume that $\Delta g(\check{\theta})$ is diagonal with eigenvalues λ_i. Then $V(t)$ has covariance function

$$
[Cov(V(t),V(s))]_{ii} = \frac{\boldsymbol{\sigma}_{ii}(X(t))}{2\lambda_i}\left[e^{-\lambda_i|t+s|} - e^{-\lambda_i|t-s|}\right],
$$

which, for negative λ_i, diverge as $t, s \to \infty$. Thus, $V(t)$ does not have any limiting stationary distribution.