Analysis of Online Learning Algorithms in Machine Learning

Ziheng Wang
The Queen's College
University of Oxford

A thesis presented for the degree of
Doctor of Philosophy

Trinity 2024

Acknowledgements

First and foremost, I am deeply grateful to those who have financially supported my PhD studies. My funding comes from the Engineering and Physical Sciences Research Council (EPSRC), Centre for Doctoral Training (CDT) in Mathematics of Random Systems, which includes support from The Queen's College, University of Oxford, and industry funding from HSBC, London, UK.

I would like to express my deepest gratitude to my supervisor, Prof. Justin Sirignano. Being your first student at University of Oxford has been a tremendous honor, and without your guidance, completing this thesis would not have been possible. Thank you for your unwavering support, understanding, and patience over the past four years. I am immensely thankful for the freedom you have given me to explore my future career path, which has illuminated my lifelong professional journey and will benefit me for years to come. The most important lesson you have imparted to me is: "It is a hundred times stronger to do one thing and be the best in the world at it than to do ten things simultaneously and be just okay at each." I will carry your teachings with me always.

Last but not least, I want to extend my heartfelt thanks to my parents, Junqiang Wang and Lili Cai, for raising me with a love for learning from a young age and for giving me unlimited freedom and constant support to pursue my true passions. My deepest appreciation also goes to my best friend in Oxford, Dr. Wolfgang Stockinger, for his unwavering care and support in both my studies and my life. I am profoundly grateful to my esteemed teachers and dear friends: Ruhong Jin from the University of Oxford for his insightful suggestions in my research, Dr. Keru Wu and Rihui Ou from Duke University for sharing their expertise in statistical machine learning, which has illuminated my future career path. Additionally, I wish to thank my collaborators and colleagues from the Mathematical and Computational Finance research group: Samuel Chun-Hei Lam, Dr. Yufei Zhang, Dr. Leandro Sanchez-Betancourt, Prof. Alvaro Cartea, Prof. Rama Cont and Prof. Ben Hambly for their invaluable experience and support throughout my studies. Your contributions have been indispensable to my academic journey.

Statement of Originality

I declare that this thesis contains no material which has been accepted or is currently being submitted for any other degree, diploma, certificate or other qualifications at the University of Oxford or elsewhere.

Chapter 2 and chapter 3 of this thesis are based on preprints [143] and [43]. Chapter 4 is based on a paper published in the Mathematical Finance, Special issue on Machine Learning in Finance [144]. Chapter 5 is based on a preprint [145]. All four papers are co-authored with my supervisor Prof. Justin Sirignano, and [43] is also co-authored with Samuel Chun-Hei Lam from University of Oxford.

Abstract

In this thesis, we consider the problem that optimizes the parameter in the stationary distribution of markov decision process, stochastic differential equations (SDEs) and stochastic partial differential equations (SPDEs). First, we study the online Actor-critic algorithms in Reinforcement Learning with tabular parametrization and prove that, under a time rescaling, the algorithm converges to ordinary differential equations (ODEs) as the number of updates becomes large. The convergence and convergence rate to the optimal strategies are given by using a two time-scale analysis which asymptotically decouples the critic ODE from the actor ODE. Next, under the same framework, we show that when both the actor and critic are parameterized by single-layer neural networks, the Actor-critic algorithm will converge in distribution to a system of ODEs with random initial conditions as the number of hidden units and the number of training steps goes to infinity. The convergence to a stationary point of the limit actor network is also established. Further, we develop a new continuous-time stochastic gradient descent method for optimizing over the stationary distribution of SDE models. The novel idea of our algorithm is that the gradient estimate is simultaneously updated using forward propagation of the SDE state derivatives, which asymptotically converges to the direction of steepest descent. We rigorously prove convergence of the online forward propagation algorithm for linear SDE models and present its numerical results to a range of mathematical finance applications. Finally, we establish the convergence of our algorithm for a class of nonlinear dissipative SDEs whose drift and volatility functions both depend upon the parameters which are being optimized. We also show the application of our algorithm in Neural SPDEs.

Contents

List of Figures

Chapter 1

Introduction

The classical supervised learning problems [66, 71] in machine learning [19, 110] aim to solve the following optimization problem

$$\begin{aligned} &\min_{\theta} J(\theta), \\ &J(\theta) = \mathbb{E}_{X\sim\pi} f(\theta, X). \end{aligned} \tag{1.0.1}$$

Here $X \sim \pi$ means that random variable X satisfies the distribution π, which does not consist of parameters that need to be optimized. At step t, the stochastic gradient descent (SGD) [11, 88, 122] use the i.i.d. samples X_t that is generated from distribution π to update the parameter:

$$\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta f(\theta_t, X_t). \tag{1.0.2}$$

The convergence of SGD (see [14, 27] for details) follows from the decomposition

$$\begin{aligned} &\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta J(\theta_t) + \alpha_t \epsilon_t, \\ &\text{where} \quad \epsilon_t = \nabla_\theta J(\theta_t) - \nabla_\theta f(\theta_t, X_t), \end{aligned} \tag{1.0.3}$$

and the stochastic estimate of the direction of steepest descent is unbiased:

$$\mathbb{E}\left[-\nabla_\theta J(\theta_t) + \epsilon_t \middle| \theta_t\right] = -\nabla_\theta J(\theta_t). \tag{1.0.4}$$

Broadly speaking, this thesis is focused on solving optimization problem:

$$\min_{\theta} \mathbb{E}_{Y\sim\pi_\theta} f(Y), \tag{1.0.5}$$

where π_θ is the stationary distribution of some Markov chains, Markov decision process or stochastic differential equations (SDEs), which is parameterized by θ. As an important generation of optimization problem (1.0.1), (1.0.5) can be derived from Ergodic Stochastic Control [6], Parameter Estimation in SDE [116, 128], Reinforcement Learning [136], Bayesian statistics [37, 68], Markov Chain Monte Carlo [97], etc.

In the first part of this thesis, we first study the Actor-critic (AC) algorithms, which are widely used in reinforcement learning. The difficulty of analysing AC is brought by the online arrival of non-i.i.d. data samples. What's more, the distribution of the data samples dynamically changes as the model is updated, which introduces a complex feedback loop between the data distribution and the reinforcement learning algorithm. We prove that, under a time rescaling,

- The online actor-critic algorithm with tabular parametrization converges to ODEs as the number of updates becomes large.
- The online actor-critic algorithm with single layer neural network parametrization converges to ODEs with random initialization as the numbers of iterations and hidden units becomes large.

The proof first establishes the geometric ergodicity of the data samples under a fixed policy. Then we prove that the fluctuations of the data samples around a dynamic probability measure, which is a function of the evolving actor model, vanish as the number of updates becomes large. Using the Poisson equation and weak convergence techniques, the ODE limit can be derived and then we study its convergence properties using a two time-scale analysis which asymptotically de-couples the critic ODE from the actor ODE. The convergence of the critic to the solution of the Bellman equation can be established for both tabular and neural network cases. And for the tabular parametrization, the actor will converge to the optimal policy while for the neural network case, the convergence to a stationary point is established.

Then in the second part, we develop a new continuous-time SGD method for optimizing over the stationary distribution of SDE models. The algorithm solves an SDE, derived using forward differentiation, which provides an asymptotically unbiased stochastic estimate for the gradient. The algorithm continuously updates the SDE model's parameters and the gradient estimate simultaneously. We rigorously prove convergence of the online forward propagation algorithm for:

- Multi-dimensional Ornstein-Uhlenbeck process by analysing the closed form formula of its probability density function.
- Nonlinear dissipative SDEs by characterizing the convergence rates of the transition semi-group and its derivatives.

The challenge of the proof comes from the fluctuations of the parameter evolution around the direction of steepest descent. We prove bounds for the solutions of a new class of Poisson partial differential equations (PDEs), which are then used to analyse the parameter fluctuations. We also present the application of our algorithm into a range of mathematical finance applications, including statistical calibration of SDE models, high-dimensional stochastic optimal control for long time horizons, training stochastic point process models of limit order book events and calibrate SPDE models.

1.1 Online Actor-Critic in Reinforcement Learning

1.1.1 Background of Reinforcement Learning

Let $\boldsymbol{\mathcal{M}} = (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}, p, \mu, r, \gamma)$ be an Markov decision process, where $\boldsymbol{\mathcal{X}}$ is a finite discrete state space, $\boldsymbol{\mathcal{A}}$ is a finite discrete action space, $p(x'|x, a)$ is the transition probability function, μ is the initial

probability distribution of the Markov chain, $r(x,a)$ is the reward function, and the discount factor is $\gamma \in (0,1)$. Let the policy $f(x,a)$ be the probability of selecting action a in state x. The state and action-value functions $V^f(\cdot) : \boldsymbol{\mathcal{X}} \to \mathbb{R}$ and $V^f(\cdot,\cdot) : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ are defined as the expected discounted sum of future rewards when actions are selected from the policy f:

$$V^f(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot r\left(x_k, a_k\right) \mid x_0 = x\right], \quad V^f(x,a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot r\left(x_k, a_k\right) \mid x_0 = x, a_0 = a\right], \tag{1.1.1}$$

where $a_k \sim f\left(x_k, \cdot\right)$, and $x_{k+1} \sim p\left(\cdot \mid x_k, a_k\right)$ for all $k \in \mathbb{Z}^+$. Note that the transition kernel p and policy f induce a Markov chain on the state-action space $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$. Then for any $(x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$, define the state and state-action visiting measures respectively as ν_μ^f and σ_μ^f, where

$$\nu_\mu^f(x) = \sum_{k=0}^{\infty} \gamma^k \cdot \mathbb{P}\left(x_k = x\right), \quad \sigma_\mu^f(x,a) = \sum_{k=0}^{\infty} \gamma^k \cdot \mathbb{P}\left(x_k = x, a_k = a\right) \tag{1.1.2}$$

and $x_0 \sim \mu(\cdot), a_k \sim f\left(x_k, \cdot\right)$, $x_{k+1} \sim p\left(\cdot \mid x_k, a_k\right)$ for all $k \geq 0$. One can define a new MDP $\widetilde{\boldsymbol{\mathcal{M}}} = (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}, \widetilde{p}, \mu, r, \gamma)$ with the transition probability function

$$\widetilde{p}\left(x' \mid x, a\right) = \gamma \cdot p\left(x' \mid x, a\right) + (1-\gamma) \cdot \mu\left(x'\right), \tag{1.1.3}$$

and [83] prove that the stationary distribution of $\widetilde{\boldsymbol{\mathcal{M}}}$ under policy f is the $\frac{1}{1-\gamma}\sigma_\mu^f$ in (1.1.2).

The goal of reinforcement learning is to learn the optimal policy f^* which maximizes the expected discounted sum of the future rewards:

$$J(f) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot r\left(x_k, a_k\right)\right] = \sum_{x \in \boldsymbol{\mathcal{X}}} \mu(x) V^f(x) = \sum_{(x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} \sigma_\mu^f(x,a) r(x,a). \tag{1.1.4}$$

Policy-based reinforcement learning method optimizes the objective function over a class of policies $\{f_\theta \mid \theta \in \boldsymbol{\mathcal{B}}\}$ using the policy gradient theorem [136], which states that

$$\nabla_\theta J\left(f_\theta\right) = \mathbb{E}_{\sigma_\mu^{f_\theta}}\left[V^{f_\theta}(x,a) \cdot \nabla_\theta \log f_\theta(x,a)\right], \tag{1.1.5}$$

where $\sigma_\mu^{f_\theta}$ is the state-action visitation measure defined in (1.1.2). In practice, the value function in the policy gradient theorem (1.1.5) is unknown and must therefore also be estimated by a statistical learning algorithm.

1.1.2 Online Actor-Critic Algorithm

Online actor-critic algorithms simultaneously estimate the value function using a critic model and the optimal policy using an actor model. In this thesis, we specifically study a class of online actor-critic algorithms with tabular and neural network parametrization.

Tabular parametrization: The "actor" is a tabular softmax policy

$$f_\theta(x,a) = \frac{e^{\theta(x,a)}}{\sum_{a' \in \boldsymbol{\mathcal{A}}} e^{\theta(x,a')}}, \tag{1.1.6}$$

with parameters $\theta = \big(\theta(x,a)\big)_{(x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}$. The "critic" $Q = (Q(x,a))_{(x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}$, acting as the approximation of the unknown state-action value function for the optimal policy (approximated by the actor model), is also a tabular with separate parameter for each state-action pair. The policy $f_\theta(x) = \big(f_\theta(x,a)\big)_{a\in\boldsymbol{\mathcal{A}}}$ is a probability distribution on the set of actions $\boldsymbol{\mathcal{A}}$.

In our algorithm, at the learning step k, we use θ_k to denote the estimate for the policy parameters and Q_k as the estimate for the value function under the policy f_{θ_k}. At step k, the sample $(\widetilde{x}_k.\widetilde{a}_k)$, used to update the actor parameters θ_k, is generated from MDP $\widetilde{\boldsymbol{\mathcal{M}}}$ by policy f_{θ_k}. Then we use vanilla policy gradient theorem (1.1.5) to update the actor and get new policy $f_{\theta_{k+1}}$. The sample (x_k, a_k) is generated from MDP $\boldsymbol{\mathcal{M}}$ by the exploration policy g_{θ_k} defined in (1.1.10). An exploration policy is used to guarantee that the policy will have a positive probability to visit all states and actions. We then update the critic by temporal difference learning [146] to obtain the new critic approximation Q_{k+1}. For notational convenience, we will sometimes use f_k and g_k to denote f_{θ_k} and g_{θ_k}. In summary,

- The samples $\{\widetilde{x}_k, \widetilde{a}_k\}_{k\geq 1}$ for the actor model are sampled from $\widetilde{\boldsymbol{\mathcal{M}}}$ under the policy f_k:
$$\widetilde{x}_0, \widetilde{a}_0 \xrightarrow{\widetilde{p}(\cdot|\widetilde{x}_0,\widetilde{a}_0)} \widetilde{x}_1 \xrightarrow{f_0(\widetilde{x}_1,\cdot)} \widetilde{a}_1 \xrightarrow{\widetilde{p}(\cdot|\widetilde{x}_1,\widetilde{x}_1)} \widetilde{x}_2 \xrightarrow{f_1(\widetilde{x}_2,\cdot)} \widetilde{a}_2 \xrightarrow{\widetilde{p}(\cdot|\widetilde{x}_2,\widetilde{a}_2)} \widetilde{x}_3 \cdots . \tag{1.1.7}$$

- The samples $\{x_k, a_k\}_{k\geq 1}$ used to train the critic model are sampled from $\boldsymbol{\mathcal{M}}$ under the exploration policy g_k:
$$x_0, a_0 \xrightarrow{p(\cdot|x_0,a_0)} x_1 \xrightarrow{g_0(x_1,\cdot)} a_1 \xrightarrow{p(\cdot|x_1,a_1)} x_2 \xrightarrow{g_1(x_2,\cdot)} a_2 \xrightarrow{p(\cdot|x_2,a_2)} x_3 \cdots . \tag{1.1.8}$$

And θ_k, Q_k are updated according to the actor-critic algorithm:

$$\begin{aligned} Q_{k+1}(x,a) &= Q_k(x,a) + \frac{\alpha}{N}\Big(r(x_k,a_k) + \gamma Q_k(x_{k+1},a_{k+1}) - Q_k(x_k,a_k)\Big)\partial_{x,a}Q_k(x_k,a_k) \\ \theta_{k+1}(x,a) &= \theta_k(x,a) + \frac{\zeta_k^N}{N} Q_k(\widetilde{x}_k,\widetilde{a}_k)\partial_{x,a}\log f_k(\widetilde{x}_k,\widetilde{a}_k), \end{aligned} \tag{1.1.9}$$

for $k = 0, 1, \ldots, TN$. The actions a_k in (1.1.9) are selected from the distribution

$$g_{\theta_k}(x,a) = \frac{\eta_k^N}{|\boldsymbol{\mathcal{A}}|} + (1-\eta_k^N)\cdot f_{\theta_k}(x,a), \quad \forall (x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}, \tag{1.1.10}$$

where $0 \leq \eta_k^N < 1$. That is, with probability η_k^N, we select an action uniformly at random and, with probability $1-\eta_k^N$, we select an action from the current estimate for the optimal policy. We let the exploration rate decay during training, i.e., $\eta_k^N \to 0$ as $k\to\infty$.

Neural Network Parametrization: Now the "actor" is defined as

$$f_\theta^N(x,a) = \mathrm{Softmax}(P_\theta^N(x,a)) = \frac{\exp(P_\theta^N(x,a))}{\sum\limits_{a'}\exp(P_\theta^N(x,a'))} \tag{1.1.11}$$

where $P_\theta^N(x,a)$ is the actor network:

$$P_\theta^N(x,a) = \frac{1}{\sqrt{N}}\sum_{i=1}^N B^i\sigma\left(U^i\cdot(x,a)\right) \tag{1.1.12}$$

parameterized by the parameters $\theta = (B^1, \dots, B^N, U^1, \dots U^N)$ with $B^i \in \mathbb{R}$ and $U^i \in \mathbb{R}^{|\mathcal{X}|+|\mathcal{A}|}$. The "critic" is another network

$$Q_\omega^N(x,a) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} C^i \sigma\left(W^i \cdot (x,a)\right), \tag{1.1.13}$$

parameterized by the parameters $\omega = (C^1, \dots, C^N, W^1, \dots W^N)$, where $C^i \in \mathbb{R}$ and $W^i \in \mathbb{R}^{|\mathcal{X}|+|\mathcal{A}|}$.

Similar to the framework for the online tabular Actor-Critic algorithm (1.1.7), the data samples used to train the actor and critic networks are generated as follows.

- The "actor" process:

$$(\widetilde{\boldsymbol{\mathcal{M}}}, \mathsf{Ac}) : (\widetilde{x}_0, \widetilde{a}_0) \sim \rho_0 \xrightarrow{\tilde{p}(\cdot|\widetilde{x}_0,\widetilde{a}_0)} \widetilde{x}_1 \xrightarrow{g_0^N(\widetilde{x}_1,\cdot)} \widetilde{a}_1 \xrightarrow{\tilde{p}(\cdot|\widetilde{x}_1,\widetilde{a}_1)} \widetilde{x}_2 \xrightarrow{g_1^N(\widetilde{x}_2,\cdot)} \widetilde{a}_2 \xrightarrow{\tilde{p}(\cdot|\widetilde{x}_2,\widetilde{a}_2)} \widetilde{x}_3 \xrightarrow{g_2^N(\widetilde{x}_3,\cdot)} \widetilde{a}_3 \cdots, \tag{1.1.14}$$

- The "critic" process:

$$(\boldsymbol{\mathcal{M}}, \mathsf{Cr}) : (x_0, a_0) \sim \rho_0 \xrightarrow{p(\cdot|x_0,a_0)} x_1 \xrightarrow{g_0^N(x_1,\cdot)} a_1 \xrightarrow{p(\cdot|x_1,a_1)} x_2 \xrightarrow{g_1^N(x_2,\cdot)} a_2 \xrightarrow{p(\cdot|x_2,a_2)} x_3 \xrightarrow{g_2^N(x_3,\cdot)} a_3 \cdots. \tag{1.1.15}$$

In order to make sure the convergence, unlike (1.1.7), here both the actor and critic processes are sampled by the exploration policy g_k^N, which is defined as

$$g_k^N(\xi) = \frac{\eta_k^N}{|\boldsymbol{\mathcal{A}}|} + (1-\eta_k^N) \cdot f_k^N(\xi), \quad \xi = (x,a), \tag{1.1.16}$$

and $(\eta_k^N)_{k\geq 0}$ is a sequence of exploration rates such that $0 < \eta_k^N \leq 1$ and $\eta_k^N \overset{k\to\infty}{\to} 0$. The parameters in the actor and critic networks are updated by:

$$\begin{aligned}
C_{k+1}^i &= C_k^i + \frac{\alpha^N}{\sqrt{N}} \left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right) \sigma\left(W_k^i \cdot \xi_k\right), \\
W_{k+1}^i &= W_k^i + \frac{\alpha^N}{\sqrt{N}} \left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right) C_k^i \sigma'\left(W_k^i \cdot \xi_k\right) \xi_k, \\
B_{k+1}^i &= B_k^i + \frac{\zeta_k^N}{N\sqrt{N}} \mathsf{clip}(Q_k^N(\tilde{\xi}_k)) \left(\sigma(U^i \cdot (\tilde{\xi}_k)) - \sum_{a''} f_k^N(\tilde{x}_k, a'') \sigma(U^i \cdot (\tilde{x}_k, a''))\right), \\
U_{k+1}^i &= U_k^i + \frac{\zeta_k^N}{N\sqrt{N}} \mathsf{clip}(Q_k^N(\tilde{\xi}_k)) \left(B_k^i \sigma'(U_k^i \cdot (\tilde{\xi}_k))(\tilde{\xi}_k) - \sum_{a''} f_k^N(\tilde{x}_k, a'') B_k^i \sigma'(U_k^i \cdot (\tilde{x}_k, \tilde{a}_k))(\tilde{x}_k, a'')\right),
\end{aligned} \tag{1.1.17}$$

where

$$\mathsf{clip}(x) = \max(\min(x, 2), 0) \tag{1.1.18}$$

is to ensure that the convergence when $N \to \infty$.

1.1.3 Our new contribution

In Chapter 2, we prove that, under a time rescaling, the online actor-critic algorithm (1.1.9) converges to an ordinary differential equations (ODEs) as the number of updates becomes large.

Theorem 1.1.1 (Limit Equations). *For any $T > 0$,*

$$\lim_{N\to\infty} \mathbb{E} \sup_{t\in[0,T]} \left[\left\|\theta_{\lfloor Nt\rfloor} - \bar{\theta}_t\right\| + \left\|Q_{\lfloor Nt\rfloor} - \bar{Q}_t\right\|\right] = 0, \tag{1.1.19}$$

where $\bar{Q}_t$ and $\bar{\theta}_t$ satisfy the nonlinear system of ODEs:

$$\begin{aligned}
\frac{d\bar{Q}_t}{dt}(x,a) &= \alpha\pi^{g_{\bar{\theta}_t}}(x,a)\left(r(x,a) + \gamma\sum_{z,a''}\bar{Q}_t(z,a'')g_{\bar{\theta}_t}(z,a'')p(z|x,a) - \bar{Q}_t(x,a)\right) \\
\frac{d\bar{\theta}_t}{dt}(x,a) &= \zeta_t\sigma_\mu^{f_{\bar{\theta}_t}}(x,a)\left[\bar{Q}_t(x,a) - \sum_{a'}\bar{Q}_t(x,a')f_{\bar{\theta}_t}(x,a')\right],
\end{aligned} \tag{1.1.20}$$

with initial condition $(\bar{Q}_0, \bar{\theta}_0) = (Q_0, \theta_0)$.

Using a two time-scale analysis, we provide the proof for the convergence of the critic ODE to the solution of the Bellman equation and the actor ODE to the optimal policy.

Theorem 1.1.2 (Global Convergence). *The limit critic model converges to the value function:*

$$\left\|\bar{Q}_t - V^{f_{\bar{\theta}_t}}\right\| = O\left(\frac{1}{\log^2 t}\right). \tag{1.1.21}$$

For the sufficient exploration initial distribution $\mu(x) > 0, \forall x \in \mathcal{X}$, the limit actor model converges to the optimal policy:

$$J(f^*) - J(f_{\bar{\theta}_t}) = O\left(\frac{1}{\log t}\right), \tag{1.1.22}$$

where f^* is any optimal policy.

Then in Chapter 3, we show algorithm (1.1.17) converges in distribution to the solution of a nonlinear ODE system as the number of hidden units for the neural networks and the number of training steps $\to \infty$. Define the empirical measures

$$\mu_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{B_k^i, U_k^i}, \quad \nu_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{C_k^i, W_k^i}. \tag{1.1.23}$$

In addition, we define the following time-rescaled processes for any $\xi = (x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$

$$\begin{aligned}
&P_t^N(\xi) = P_{\lfloor Nt\rfloor}^N(\xi), \quad f_t^N(\xi) = f_{\lfloor Nt\rfloor}^N(\xi), \quad g_t^N(\xi) = g_{\lfloor Nt\rfloor}^N(\xi), \\
&Q_t^N(\xi) = Q_{\lfloor Nt\rfloor}^N(\xi), \quad \mu_t^N = \mu_{\lfloor Nt\rfloor}^N, \quad \nu_t^N = \nu_{\lfloor Nt\rfloor}^N.
\end{aligned} \tag{1.1.24}$$

Using Assumptions 3.2.9 and 3.3.1, we know that $\mu_0^N, \nu_0^N \xrightarrow{d} \mu_0, \nu_0$ and $P_0^N, Q_0^N \xrightarrow{d} \mathcal{G}, \mathcal{H}$ as $N \to \infty$, where μ_0, ν_0 are mean-zero distribution and $\mathcal{G}, \mathcal{H}$ are mean-zero Gaussian random variables by the law of large numbers and central limit theorem for i.i.d. random variables, respectively. Define the state space for the time-rescaled process $\left(\mu_t^N, \nu_t^N, P_t^N, Q_t^N\right)$:

$$E = \mathcal{M}(\mathbb{R}^{1+d}) \times \mathcal{M}(\mathbb{R}^{1+d}) \times \mathbb{R}^M \times \mathbb{R}^M, \quad d = |\mathcal{X}| + |\mathcal{A}|, \quad M = |\mathcal{X} \times \mathcal{A}|, \tag{1.1.25}$$

where $\mathcal{M}(\mathbb{R}^{1+d})$ is the set of all probability measures on $\mathbb{R}^{1+d}$. Define the space

$$D_E([0,T]) = \{\text{càdlàg paths } f : [0,T] \to E\}. \tag{1.1.26}$$

Then the following convergence theorem holds.

Theorem 1.1.3. *Let Assumptions 3.2.9 and 3.3.1 hold, and the learning rate for the critic parameter updates be $\alpha^N = \alpha/N$ for an $\alpha > 0$. Then, the process $\left(\mu_t^N, \nu_t^N, P_t^N, Q_t^N\right)$ converges weakly in the space $D_E([0,T])$ as $N \to \infty$ to the process (μ_t, ν_t, P_t, Q_t), so that for any $t \in [0,T]$, any $\xi = (x,a) \in \mathcal{X} \times \mathcal{A}$, and for every $\varphi, \bar{\varphi} \in C_b^2\left(\mathbb{R}^{1+d}\right)$, the limit process (μ_t, ν_t, P_t, Q_t) satisfies the random ODE:*

$$\begin{aligned}
\frac{dQ_t}{dt}(\xi) &= \alpha \sum_{\xi'=(x',a')} A_{\xi,\xi'} \left(r(\xi') + \gamma \sum_{z,a''} Q_t(z,a'') g_t(z,a'') p(z|\xi') - Q_t(\xi') \right) \pi^{g_t}(\xi'), \\
\frac{dP_t}{dt}(\xi) &= \sum_{\xi'=(x',a')} \zeta_t \mathsf{clip}(Q_t(\xi')) \left[A_{\xi,\xi'} - \sum_{a''} f_t(x',a'') A_{\xi,x',a''} \right] \sigma_{\rho_0}^{g_t}(\xi'), \\
P_0(\xi) &= \mathcal{G}(\xi), \quad Q_0(\xi) = \mathcal{H}(\xi) \\
\langle \varphi, \mu_t \rangle &= \langle \bar{\varphi}, \nu_0 \rangle, \quad \langle \varphi, \nu_t \rangle = \langle \varphi, \nu_0 \rangle,
\end{aligned} \tag{1.1.27}$$

where

$$A_{\xi,\xi'} = \int \left(\sigma\left(w \cdot \xi'\right) \sigma(w \cdot \xi) + c^2 \sigma'\left(w \cdot \xi'\right) \sigma'(w \cdot \xi)(\xi \cdot \xi') \right) \nu_0(dc, dw), \tag{1.1.28}$$

$\mathcal{G}, \mathcal{H}$ are the weak limits of P_0^N and Q_0^N, which are mean-zero Gaussian random variables, and

$$f_t(\xi) = \mathrm{Softmax}(P_t(\xi)), \quad g_t(\xi) = \frac{\eta_t}{|\mathcal{A}|} + (1 - \eta_t) f_t(\xi).$$

Analysis of the limit ODE (1.1.27) shows that the limit critic network will converge to the true value function, which will provide the actor an asymptotically unbiased estimate of the policy gradient. We also prove that the limit actor network will converge to a stationary point.

Theorem 1.1.4. *If the actor network P_t and critic network Q_t evolved according to the limit ODE (1.1.27), then under assumptions 3.2.9 and 3.2.10, the critic network converges globally to the value function of the policy $f_t = \mathrm{Softmax}(P_t)$ as $t \to \infty$:*

$$\left\| Q_t - V^{f_t} \right\|_\infty = \max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_t(\xi) - V^{f_t}(\xi)| = O(\eta_t). \tag{1.1.29}$$

Moreover, the actor network converges to a stationary point:

$$\nabla_P J(f_t) \overset{t \to \infty}{\to} 0. \tag{1.1.30}$$

1.2 Stochastic Gradient Descent in Continuous time

1.2.1 Background of Stochastic Gradient Descent in Continuous time

[130] proposed a "stochastic gradient descent in continuous time" (SGDCT) algorithm for estimating parameters in an SDE from continuous observations. Consider a diffusion $X_t \in \mathcal{X} = \mathbb{R}^m$:

$$dX_t = f^*\left(X_t\right) dt + \sigma dW_t. \tag{1.2.1}$$

The goal is to statistically estimate a model $f(x,\theta)$ for $f^*(x)$ where $\theta \in \mathbb{R}^n$ where the function $f^*(x)$ is unknown and $W_t \in \mathbb{R}^m$ is a standard Brownian motion. The diffusion term W_t represents any random behavior of the system or environment. The functions $f(x,\theta)$ and $f^*(x)$ may be non-convex.

The authors assume that X_t is sufficiently ergodic and that it has some well-behaved $\pi(dx)$ as its unique invariant measure. Define the loss function

$$g(x,\theta) = \frac{1}{2}\left\|f(x,\theta) - f^*(x)\right\|^2_{\sigma\sigma^\top} = \frac{1}{2}\left\langle f(x,\theta) - f^*(x), \left(\sigma\sigma^\top\right)^{-1}\left(f(x,\theta) - f^*(x)\right)\right\rangle, \tag{1.2.2}$$

and the objective function to estimate the parameter θ to be

$$J(\theta) = \int_{\mathcal{X}} g(x,\theta)\pi(dx). \tag{1.2.3}$$

In [130], the SGDCT follows the SDE:

$$d\theta_t = \alpha_t\left[\nabla_\theta f\left(X_t,\theta_t\right)\left(\sigma\sigma^\top\right)^{-1} dX_t - \nabla_\theta f\left(X_t,\theta_t\right)\left(\sigma\sigma^\top\right)^{-1} f\left(X_t,\theta_t\right)dt\right], \tag{1.2.4}$$

where $\nabla_\theta f\left(X_t;\theta_t\right)$ is matrix valued and α_t is the learning rate. The parameter update (1.2.4) can be used for both statistical estimation given previously observed data as well as online learning (i.e., statistical estimation in real-time as data becomes available). SGDCT follows a noisy descent direction along a continuous stream of data produced by X_t and it is expected that θ_t will tend towards the minimum of the objective funtion (1.2.3). [130] proves convergence of the SGDCT to a stationary point.

1.2.2 Our new contribution

Unlike [17, 72, 126, 128, 130, 132, 135] that estimate the parameters for the SDE from continuous observations with fixed dynamic, in Chapter 4 and Chapter 5, we design a novel continuous time stochastic gradient algorithm to solve the different optimization problem that aim to optimize the parameters in the stationary distribution of ergodic SDEs.

Consider a parameterized process $X_t^{\theta,x} \in \mathbb{R}^d$ which satisfies the stochastic differential equation (SDE):

$$\begin{aligned} dX_t^{\theta,x} &= \mu(X_t^{\theta,x},\theta)dt + \sigma(X_t^{\theta,x},\theta)dW_t, \\ X_0^{\theta,x} &= x, \end{aligned} \tag{1.2.5}$$

where $\theta \in \mathbb{R}^\ell$ and W_t is a d-dimensional standard Brownian motion. Suppose $X_t^{\theta,x}$ is ergodic and our goal is to select θ such that the stationary distribution of X_t^θ matches certain target statistics, that is minimizing the objective function

$$J(\theta) = \sum_{i=1}^{N}\left(\mathbb{E}_{Y\sim\pi_\theta}\left[f_n(Y)\right] - \beta_n\right)^2, \tag{1.2.6}$$

where f_n are known functions and β_n are the target quantities [1]. The novel online forward propa-

[1] Without loss of generality, we will set $N = 1$ and $\beta_1 = \beta$ in (1.2.6).

gation algorithm for optimizing (1.2.6) is:

$$\begin{aligned}
\frac{d\theta_t}{dt} &= -2\alpha_t \left(f(\bar{X}_t) - \beta\right) \left(\nabla f(X_t)\tilde{X}_t\right)^\top, \\
d\tilde{X}_t &= \left(\mu_x(X_t,\theta_t)\tilde{X}_t + \mu_\theta(X_t,\theta_t)\right) dt + \left(\sigma_x(X_t,\theta_t)\tilde{X}_t + \sigma_\theta(X_t,\theta_t)\right) dW_t, \\
dX_t &= \mu(X_t,\theta_t)dt + \sigma(X_t,\theta_t)dW_t, \\
d\bar{X}_t &= \mu(\bar{X}_t,\theta_t)dt + \sigma(\bar{X}_t,\theta_t)d\bar{W}_t,
\end{aligned} \tag{1.2.7}$$

where W_t and $\bar{W}_t$ are independent Brownian motions and α_t is the learning rate.

In Chapter 4 and Chapter 5, we rigorously prove convergence of the online forward propagation algorithm (1.2.7) for the multi-dimensional Ornstein-Uhlenbeck (OU) process and a class of nonlinear dissipative SDEs whose drift and volatility functions both depend upon the parameters which are being optimized.

- The OU process we focus on in Chapter 4 is

$$\begin{aligned}
dX_t^\theta &= \left(g(\theta) - h(\theta)X_t^\theta\right) dt + \sigma dW_t, \\
X_0^\theta &= x,
\end{aligned} \tag{1.2.8}$$

  where $h(\theta)$ is uniformly positive definite.

- In Chapter 5, we study the SDE (1.2.5) with the drift and diffusion terms satisfy the following dissipativity assumption:

$$\langle \mu(x_1,\theta) - \mu(x_2,\theta),\ x_1 - x_2\rangle + \frac{7}{2}\left|\sigma(x_1,\theta) - \sigma(x_2,\theta)\right|^2 \le -\beta|x_1 - x_2|^2, \tag{1.2.9}$$

  where $\langle a,\ b\rangle := b^\top a$.

The main convergence result in Chapter 4 and Chapter 5 is the following:

Theorem 1.2.1. *For the Ornstein–Uhlenbeck process (4.2.1) and nonlinear dissipative SDE under Assumptions (A5.2.1) - (A5.2.5), the online algorithm (1.2.7) will converge to a stationary point almost surely:*

$$\lim_{t\to\infty} |\nabla_\theta J(\theta_t)| \overset{a.s.}{=} 0. \tag{1.2.10}$$

In addition to the proof of convergence, in Chapter 4, we show the applications of our algorithm to a range of mathematical finance problems involving statistical calibration of SDE models, stochastic high-dimensional optimal control for long time horizons. The applications of our algorithm in control SPDEs is also provided in Chapter 5.

1.3 Outline

The main contributions of this thesis can be summarised as follows:

- Derivations of ODE limit of fully online tabular Actor-Critic algorithm by using the Poisson equation;

- Proof of the convergence and convergence rate to the optimal policy for the ODE limit of Tabular Actor-Critic;
- Derivations of ODE limit of online Neural Actor-Critic by techniques of Poisson equation and weak convergence;
- Proof of the convergence to the stationary point for the ODE limit of Neural Actor-Critic;
- Develop the novel online forward propagation algorithm to optimize the parameter in the stationary distribution of SDEs;
- Prove the convergence of our algorithm for multi-dimensional Ornstein-Uhlenbeck process and nonlinear dissipative SDEs;
- Presentation of the numerical results of our algorithm to a range of mathematical finance applications.

In Chapter 2, we prove that, under a time rescaling, the online actor-critic algorithm with tabular parametrization converges to ordinary differential equations (ODEs) as the number of updates becomes large. The proof first establishes the geometric ergodicity of the data samples under a fixed actor policy. Then, using a Poisson equation, we prove that the fluctuations of the data samples around a dynamic probability measure, which is a function of the evolving actor model, vanish as the number of updates become large. Once the ODE limit has been derived, we study its convergence properties using a two time-scale analysis which asymptotically de-couples the critic ODE from the actor ODE. The convergence of the critic to the solution of the Bellman equation and the actor to the optimal policy are proven. In addition, a convergence rate to this global minimum is also established. Our convergence analysis holds under specific choices for the learning rates and exploration rates in the actor-critic algorithm, which could provide guidance for the implementation of actor-critic algorithms in practice.

In Chapter 3, we prove that a single-layer neural network trained with the online actor critic algorithm converges in distribution to a random ordinary differential equation (ODE) as the number of hidden units and the number of training steps $\to \infty$. In the online actor-critic algorithm, the distribution of the data samples dynamically changes as the model is updated, which is a key challenge for any convergence analysis. We establish the geometric ergodicity of the data samples under a fixed actor policy. Then, using a Poisson equation, we prove that the fluctuations of the model updates around the limit distribution due to the randomly-arriving data samples vanish as the number of parameter updates $\to \infty$. Using the Poisson equation and weak convergence techniques, we prove that the actor neural network and critic neural network converge to the solutions of a system of ODEs with random initial conditions. Analysis of the limit ODE shows that the limit critic network will converge to the true value function, which will provide the actor an asymptotically

unbiased estimate of the policy gradient. We then prove that the limit actor network will converge to a stationary point.

In Chapter 4, we develop a new continuous-time stochastic gradient descent method for optimizing over the stationary distribution of stochastic differential equation (SDE) models. The algorithm continuously updates the SDE model's parameters using an estimate for the gradient of the stationary distribution. The gradient estimate is simultaneously updated using forward propagation of the SDE state derivatives, asymptotically converging to the direction of steepest descent. We rigorously prove convergence of the online forward propagation algorithm for linear SDE models (i.e., the multi-dimensional Ornstein-Uhlenbeck process) and present its numerical results for nonlinear examples. The proof requires analysis of the fluctuations of the parameter evolution around the direction of steepest descent. Bounds on the fluctuations are challenging to obtain due to the online nature of the algorithm (e.g., the stationary distribution will continuously change as the parameters change). We prove bounds for the solutions of a new class of Poisson partial differential equations (PDEs), which are then used to analyze the parameter fluctuations in the algorithm. Our algorithm is applicable to a range of mathematical finance applications involving statistical calibration of SDE models and stochastic optimal control for long time horizons where ergodicity of the data and stochastic process is a suitable modeling framework. Numerical examples explore these potential applications, including learning a neural network control for high-dimensional optimal control of SDEs and training stochastic point process models of limit order book events.

In Chapter 5, we study the convergence of the forward propagation algorithm, which is invented in Chapter 4 for a class of nonlinear SDEs with non-constant volatility coefficients. Both the drift and volatility functions in the SDE may depend upon the parameters which are being optimized. The nonlinear SDE is assumed to satisfy standard dissipativity conditions, which allows us to leverage the ergodicity of nonlinear dissipative SDEs to characterize the convergence rate of the transition semigroup and its derivatives. Then, we prove bounds on the solution of a Poisson partial differential equation (PDE) for the expected time integral of the algorithm's stochastic fluctuations around the direction of steepest descent. We then re-write the algorithm using the PDE solution in order to characterize the parameter evolution around the direction of steepest descent. Our main result is a convergence theorem for the forward propagation algorithm for nonlinear dissipative SDEs.

Chapter 2

Global Convergence of Online Tabular Actor-Critic Algorithms

2.1 Introduction

Actor-critic (AC) algorithms [81, 85] have become some of the most successful and widely-used methods in reinforcement learning (RL) [136]. AC algorithms are typically implemented in two ways: batch and online. In the batch setting, the actor's one update in the outer loop is followed by the critic's numerous updates in the inner loop to get a good approximation of the value function. The convergence of batch AC has been carefully studied recently [86, 142, 150]. An online, two time-scale AC algorithm was first proposed in [81], where the actor and critic are updated simultaneously with i.i.d. data samples. In this chapter, we study a class of online actor-critic [78, 147, 149] algorithms where the data samples arrive from a Markov chain [83] (instead of i.i.d. data samples) and prove the actor/critic converge to the solution of an ODE as the number learning steps becomes large. It is then proven that the solution of the ODE converges to the optimal policy.

We consider an actor-critic algorithm where the actor and critic are updated simultaneously at each new time step by using the data samples from simultaneous simulations of two different Markov decision processes (MDPs). Specifically, the data samples used to update the critic are from the original MDP while the samples for the actor are from an artificial MDP with a slightly different transition probability (will be clearly defined in Section 2.2) such that the update direction of the actor asymptotically convergences to the unbiased policy gradient direction (see the algorithm in [150] for details). The data samples from the MDPs are non-i.i.d. and the transition probability function depends upon the action selected at each time step. Actions are selected using the actor's current policy. Therefore, the stationary distributions of the MDPs change as the actor evolves during learning. In order for the critic to converge to the value function, an exploration component is included in the selection of the actions, where the exploration decays to zero as the number of learning steps becomes large. We find that carefully choosing the decay rate for the exploration as well as the learning rate is crucial for proving global convergence of the limit ODEs to the optimal policy.

2.1.1 Related literature

Policy gradient Policy gradient (PG) method [138] is one of the most important concepts in RL and has achieved great effective empirical success [124, 125]. However, PG algorithms face non-convex optimization problems with respect to standard tabular policy parametrizations [1] and are thus hard to analyse mathematically. Recently, [1, 15, 79, 103, 104] have established the convergence and convergence rate to global optimum for the vanilla PG method by assuming the value function is known. [15] proved that projected PG on the simplex does not suffer from spurious local optima. [1] proves that with softmax tabular policy all the stationary points of PG are actually the global optimum and natural PG converges at rate $O\left(\frac{1}{t}\right)$. [104] provides the convergence rate $O\left(\frac{1}{t}\right)$ of PG method with softmax tabular policy. [79, 103] prove the Natural PG algorithm with softmax tabular policy indeed has asymptotically geometric global convergence.

Actor-critic The AC algorithm was first given in [138] and then extended to Natural AC in [117]. Batch AC algorithms [86, 142, 150, 151] involve a "double for loop" where the outer iteration updates the actor and, for each update of the actor, there is a large sub-iteration to solve the critic. [150] studied the global convergence of AC algorithms under the Linear Quadratic Regulator. [151] analyzed the finite-sample performance of batched AC. [86] considered the sample complexity for the "decoupled" AC methods under i.i.d. data samples. [142], under the over-parametrized two-layer neural-network proved that the neural AC algorithm converges to a global optimum at a sub-linear rate. In online AC [78, 81, 147, 149], actor and critic update simultaneously but with two time-scale. Actor updates at a slower rate while critic updates faster to provide the actor an accurate policy gradient. [81] studies an online AC algorithm with markovian data samples without using the ODE method and get the convergence to stationary point. [147] prove that two time-scale algorithms with non-i.i.d. data samples and linear function approximation finds an ϵ-stationary point with $O(\epsilon^{-\frac{5}{2}})$ samples, where ϵ measures the squared norm of the policy gradient. [149], under the compatibility condition [75, 138] between actor and critic, show that two time-scale AC requires sample complexity at order $O(\epsilon^{-2.5}\log^3(\epsilon^{-1}))$ to attain ϵ-stationary point. By carefully decreasing the exploration rate, [78] show that the two time-scale natural AC algorithm has sample complexity of $O\left(\delta^{-6}\right)$ for convergence to the global optimum. For natural AC, [41] develop a critic that employs n-step TD-learning algorithm in the off-policy natural actor-critic algorithm with linear function approximation and we establish a sample complexity of $O(\epsilon^{-3})$. [77] proposes an off-policy variant of the natural AC algorithm based on Importance Sampling, where they use Q-trace algorithm for the critic and provide a sample complexity of $O\left(\epsilon^{-3}\log\left(\frac{1}{\epsilon}\right)\right)$.

Stochastic approximation in RL Stochastic approximation [21, 23, 26] can be seen as a general framework in RL. Two time-scale stochastic approximation [51, 64] are one of the most popular

methods for AC style [25, 78, 85, 147, 149] algorithms. [21, 23, 26] establish the classical ODE method and use it for the stability and convergence of the (two time-scale) stochastic approximation where the stochastic error is a martingale difference sequence. [51, 64] proved convergence rate and finite time analysis for the two time-scale linear stochastic approximations in RL under i.i.d. assumption. [25, 85] use the ODE method for two time-scale stochastic approximations in AC algorithms where the actor is updated by policy iteration algorithm.

We study a different class of algorithms than previous literature. We consider the global convergence of the ODE limit for the online tabular AC algorithm. First, we use a time re-scaling [131] of the algorithm (2.2) to map it into a time interval $[0, T]$, and the mathematical analysis required for the convergence to ODE limit are different from the classical ODE method in stochastic approximation theory [21, 23, 26]. Second, unlike the batch AC with nested loop structure [86, 142], our online algorithm updates the actor and critic simultaneously under dynamic Markovian sampling, which has much fewer tuning parameters and thus is easier to implement. Third, [81, 147, 149] also studies an online AC algorithm with non-i.i.d. data samples. However, they only prove convergence to a stationary point while we set up the global convergence for tabular AC algorithm by analysing the limit ODE.

In this chapter, we include exploration in the policy so that the Markov chain visits all states and actions. The exploration decays to zero at a certain rate as the number of learning steps becomes large. A careful choice of the exploration rate and the learning rate is necessary in order to prove global convergence. In particular, the exploration rate does not satisfy the standard conditions (sum of the squares is finite) in stochastic approximation theory in [21, 23, 25, 26, 85]. However, by using the time-rescaling limit, we are still able to establish an ODE limit for a class of actor-critic algorithms.

2.2 Online Tabular Actor-Critic Algorithms

Let $\boldsymbol{\mathcal{M}} = (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}, p, \mu, r, \gamma)$ be an MDP, where $\boldsymbol{\mathcal{X}}$ is a finite discrete state space, $\boldsymbol{\mathcal{A}}$ is a finite discrete action space, $p(x'|x, a)$ is the transition probability function, μ is the initial probability distribution of the Markov chain, $r(x, a)$ is a bounded reward function, and the discount factor is $\gamma \in (0, 1)$. Let the policy $f(x, a)$ be the probability of selecting action a in state x. The state and action-value functions $V^f(\cdot) : \boldsymbol{\mathcal{X}} \to \mathbb{R}$ and $V^f(\cdot, \cdot) : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ are defined as the expected discounted sum of future rewards when actions are selected from the policy f:

$$V^f(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot r\left(x_k, a_k\right) \mid x_0 = x\right], \quad V^f(x, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot r\left(x_k, a_k\right) \mid x_0 = x, a_0 = a\right], \tag{2.2.1}$$

where $a_k \sim f\left(x_k, \cdot\right)$, and $x_{k+1} \sim p\left(\cdot \mid x_k, a_k\right)$ for all $k \in \mathbb{Z}^+$. [1] Note that the transition kernel p

[1] Note that the series in equation (2.2.1) converge since $\gamma \in (0, 1)$ and $r(x, a)$ is bounded.

and policy f induce a Markov chain on the state-action space $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$. Then for any $(x, a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$, define the state and state-action visiting measures respectively as ν_μ^f and σ_μ^f, where

$$\nu_\mu^f(x) = \sum_{k=0}^{\infty} \gamma^k \cdot \mathbb{P}\left(x_k = x\right), \quad \sigma_\mu^f(x, a) = \sum_{k=0}^{\infty} \gamma^k \cdot \mathbb{P}\left(x_k = x, a_k = a\right) \tag{2.2.2}$$

and $x_0 \sim \mu(\cdot), a_k \sim f\left(x_k, \cdot\right), x_{k+1} \sim p\left(\cdot \mid x_k, a_k\right)$ for all $k \geq 0$. The goal of reinforcement learning is to learn the optimal policy f^* which maximizes the expected discounted sum of the future rewards:

$$\max_f J(f),$$

where the objective function $J(f)$ is defined as

$$J(f) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot r\left(x_k, a_k\right)\right] = \sum_{x \in \boldsymbol{\mathcal{X}}} \mu(x) V^f(x) = \sum_{(x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} \sigma_\mu^f(x, a) r(x, a). \tag{2.2.3}$$

Policy-based reinforcement learning method optimize the objective function over a class of policies $\{f_\theta \mid \theta \in \boldsymbol{\mathcal{B}}\}$ using the policy gradient theorem [136]. In practice, the value function in the policy gradient theorem is unknown and must therefore also be estimated by a statistical learning algorithm. Online actor-critic algorithms simultaneously estimate the value function using a critic model and the optimal policy using an actor model. In this chapter, we specifically study a class of online actor-critic algorithms where the "actor" is a tabular softmax policy

$$f_\theta(x, a) = \frac{e^{\theta(x,a)}}{\sum_{a' \in \boldsymbol{\mathcal{A}}} e^{\theta(x,a')}}, \tag{2.2.4}$$

with parameters $\theta = \big(\theta(x, a)\big)_{(x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}}$. The "critic" $Q = (Q(x, a))_{(x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}}$ is also tabular with a separate parameter for each state-action pair. The policy $f_\theta(x) = \big(f_\theta(x, a)\big)_{a \in \boldsymbol{\mathcal{A}}}$ is a probability distribution on the set of actions $\boldsymbol{\mathcal{A}}$.

Define a new MDP $\widetilde{\boldsymbol{\mathcal{M}}} = (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}, \widetilde{p}, \mu, r, \gamma)$ with the transition probability function

$$\widetilde{p}\left(x' \mid x, a\right) = \gamma \cdot p\left(x' \mid x, a\right) + (1 - \gamma) \cdot \mu\left(x'\right), \tag{2.2.5}$$

Note that (2.2.5) is similar to the transition probability of MDP $\boldsymbol{\mathcal{M}}$ except that with probability $1-\gamma$ the state will be randomly re-initialized with distribution μ [83, 142, 149]. [83] proved that the stationary distribution of $\widetilde{\boldsymbol{\mathcal{M}}}$ under policy f is the $\frac{1}{1-\gamma}\sigma_\mu^f$ in (2.2.2). At the learning step k, we use θ_k to denote the estimate for the policy parameters while Q_k is the estimate for the value function under the policy f_{θ_k}. At step k, the sample $(\widetilde{x}_k.\widetilde{a}_k)$ used to update the actor parameters θ_k is generated from MDP $\widetilde{\boldsymbol{\mathcal{M}}}$ by policy f_{θ_k}. Then we use vanilla policy gradient theorem [138] to update the actor and get new policy $f_{\theta_{k+1}}$. The sample (x_k, a_k) is sampled from MDP $\boldsymbol{\mathcal{M}}$ by the exploration policy g_{θ_k} (see equation (2.2.9)). We then update the critic by temporal difference learning [146] to obtain the new critic approximation Q_{k+1}. An exploration policy is used to guarantee that the policy will have a positive probability to visit all states and actions. For notational convenience, we will sometimes use f_k and g_k to denote f_{θ_k} and g_{θ_k}.

In summary, the samples $\{x_k, a_k\}_{k\geq 1}$ used to train the critic model are sampled from $\boldsymbol{\mathcal{M}}$ under the exploration policy g_k:

$$x_0, a_0 \overset{p(\cdot|x_0,a_0)}{\longrightarrow} x_1 \overset{g_0(x_1,\cdot)}{\longrightarrow} a_1 \overset{p(\cdot|x_1,a_1)}{\longrightarrow} x_2 \overset{g_1(x_2,\cdot)}{\longrightarrow} a_2 \overset{p(\cdot|x_2,a_2)}{\longrightarrow} x_3 \cdots \tag{2.2.6}$$

The samples $\{\widetilde{x}_k, \widetilde{a}_k\}_{k\geq 1}$ for the actor model are sampled from $\widetilde{\boldsymbol{\mathcal{M}}}$ under the policy f_k:

$$\widetilde{x}_0, \widetilde{a}_0 \overset{\widetilde{p}(\cdot|\widetilde{x}_0,\widetilde{a}_0)}{\longrightarrow} \widetilde{x}_1 \overset{f_0(\widetilde{x}_1,\cdot)}{\longrightarrow} \widetilde{a}_1 \overset{\widetilde{p}(\cdot|\widetilde{x}_1,\widetilde{x}_1)}{\longrightarrow} \widetilde{x}_2 \overset{f_1(\widetilde{x}_2,\cdot)}{\longrightarrow} \widetilde{a}_2 \overset{\widetilde{p}(\cdot|\widetilde{x}_2,\widetilde{a}_2)}{\longrightarrow} \widetilde{x}_3 \cdots \tag{2.2.7}$$

and θ_k, Q_k are updated according to the actor-critic algorithm:

$$\begin{aligned} Q_{k+1}(x,a) &= Q_k(x,a) + \frac{\alpha}{N}\Big(r(x_k,a_k) + \gamma Q_k(x_{k+1},a_{k+1}) - Q_k(x_k,a_k)\Big)\partial_{x,a} Q_k(x_k,a_k) \\ \theta_{k+1}(x,a) &= \theta_k(x,a) + \frac{\zeta_k^N}{N} Q_k(\widetilde{x}_k,\widetilde{a}_k)\partial_{x,a}\log f_k(\widetilde{x}_k,\widetilde{a}_k), \end{aligned} \tag{2.2.8}$$

for $k = 0, 1, \ldots, TN$. The actions a_k in (2.2.8) are selected from the distribution

$$g_{\theta_k}(x,a) = \frac{\eta_k^N}{d_A} + (1-\eta_k^N)\cdot f_{\theta_k}(x,a), \quad \forall (x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}. \tag{2.2.9}$$

where $0 \leq \eta_k^N < 1$ and $d_A = |\boldsymbol{\mathcal{A}}|$. That is, with probability η_k^N, we select an action uniformly at random and, with probability $1-\eta_k^N$, we select an action from the current estimate for the optimal policy. We let the exploration rate decay during training, i.e., $\eta_k^N \to 0$ as $k \to \infty$. Note that the step-size for the online actor-critic algorithm (2.2.8) is $\frac{1}{N}$ and the number of learning steps is TN. We will later show that as $N \to \infty$ the critic and actor models converge to the solution of an ODE on the time interval $[0, T]$. Here we highlight that in order for the Q-learning algorithm converge, the policy needs to have positive probability to choose every action (see [105, 139] for details) and this is why we add exploration in the policy used to generate data samples.

Challenges for mathematical analysis: Convergence analysis of the actor-critic algorithm (2.2.8) must address several technical challenges. The data samples are non-i.i.d. and their distribution depends upon the actor model, which changes as the parameters are updated. Actions are selected using the actor model, which influences the states visited in the Markov chain and affects the actor model's evolution in the learning algorithm. Thus, actor-critic algorithms introduce a complex feedback loop between the distribution of the data samples and the model updates. Another challenge is that the learning algorithm is not guaranteed to update the model in a decent direction for the objective function, which is an obstacle for proving global convergence to the optimal policy. Finally, due to the softmax policy, the objective function is non-convex.

Overview of the proof: In our mathematical approach, we prove that the actor-critic algorithm (2.2.8) converges to an ODE under an appropriate time re-scaling. We address the challenge of non-i.i.d. data depending upon the actor model in two steps. The proof first establishes the geometric ergodicity of the data samples to a stationary distribution π^{f_θ} under a fixed actor policy f_θ. Then, using a Poisson equation, we prove that the fluctuations of the data samples around a dynamic

probability measure $\pi^{f_{\theta_k}}$, which is a function of the evolving actor model, vanish as the number of updates become large.

Once the ODE limit has been derived, we study its convergence properties using a two time-scale analysis which asymptotically de-couples the critic ODE from the actor ODE. The convergence of the critic to the solution of the Bellman equation and the actor to the optimal policy are proven. In addition, a convergence rate to this global minimum is also established. In order to prove the global convergence, the learning rate and exploration rate for the actor-critic algorithm must be carefully chosen.

2.3 Main Result

We prove that the actor and critic models converge to the solution of a nonlinear ODE system as the learning steps become large. Our results are proven under the following assumptions.

Assumption 2.3.1. The reward function r is bounded in $[0,1]$. $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{A}}$ are finite, discrete spaces.

In addition, an assumption regarding the ergodicity of the Markov chains (2.2.6) and (2.2.7) is required.

Assumption 2.3.2. For any finite θ, the Markov chain (X, A) for the MDP $\boldsymbol{\mathcal{M}}$ under exploration policy g_θ and the Markov chain $(\widetilde{X}, \widetilde{A})$ for the MDP $\widetilde{\boldsymbol{\mathcal{M}}}$ under policy f_θ are irreducible and non-periodic. Their stationary distributions π^f, σ_μ^f (which exist and are unique by Section 1.3.3 of [89]) are globally Lipschitz in policy f.

The global convergence proof also requires a careful choice for the learning rate and exploration rate.

Assumption 2.3.3. The learning rate and exploration rate are:

$$\begin{aligned} \zeta_k^N = \frac{1}{1+\frac{k}{N}}, \quad \eta_k^N = \frac{1}{1+\log^2(\frac{k}{N}+1)}, \\ \text{thus} \quad \zeta_{\lfloor Nt \rfloor}^N \to \zeta_t = \frac{1}{1+t}, \quad \eta_{\lfloor Nt \rfloor}^N \to \eta_t = \frac{1}{1+\log^2(t+1)}. \end{aligned} \tag{2.3.1}$$

Remark 2.3.4. The learning rate and exploration rate in (2.3.1) satisfy the following properties for any integer $n \in \mathbb{N}$:

$$\int_0^\infty \zeta_s ds = \infty, \quad \int_0^\infty \zeta_t^2 dt < \infty, \quad \int_0^\infty \zeta_s \eta_s ds < \infty, \quad \lim_{t\to\infty} \frac{\zeta_t}{\eta_t^n} = 0. \tag{2.3.2}$$

These properties are verified in the Appendix A.1.

The main results of this chapter are the following theorems.

Theorem 2.3.5 (Limit Equations)**.** *For any $T > 0$,*

$$\lim_{N\to\infty} \mathbb{E} \sup_{t\in[0,T]} \left[\left\| \theta_{\lfloor Nt \rfloor} - \bar{\theta}_t \right\| + \left\| Q_{\lfloor Nt \rfloor} - \bar{Q}_t \right\| \right] = 0, \tag{2.3.3}$$

where $\bar{Q}_t$ *and* $\bar{\theta}_t$ *satisfy the nonlinear system of ODEs:*

$$\begin{aligned}\frac{d\bar{Q}_t}{dt}(x,a) &= \alpha\pi^{g_{\bar{\theta}_t}}(x,a)\left(r(x,a)+\gamma\sum_{z,a''}\bar{Q}_t(z,a'')g_{\bar{\theta}_t}(z,a'')p(z|x,a)-\bar{Q}_t(x,a)\right)\\ \frac{d\bar{\theta}_t}{dt}(x,a) &= \zeta_t\sigma_\mu^{f_{\bar{\theta}_t}}(x,a)\left[\bar{Q}_t(x,a)-\sum_{a'}\bar{Q}_t(x,a')f_{\bar{\theta}_t}(x,a')\right],\end{aligned} \tag{2.3.4}$$

with initial condition $(\bar{Q}_0,\bar{\theta}_0)=(Q_0,\theta_0)$.

Thus, the critic converges to the limit variable $\bar{Q}_t$ while the actor converges to the limit variable $\bar{\theta}_t$, where $\bar{Q}_t$ and $\bar{\theta}_t$ are solutions to a nonlinear system of ODEs. We then prove the convergence of the limit ODEs (2.3.4) to the value function and optimal policy. The convergence analysis also allows us to obtain convergence rates.

Theorem 2.3.6 (Global Convergence)**.** *The limit critic model converges to the value function:*

$$\left\|\bar{Q}_t - V^{f_{\bar{\theta}_t}}\right\| = O\left(\frac{1}{\log^2 t}\right). \tag{2.3.5}$$

For the sufficient exploration initial distribution $\mu(x)>0,\forall x\in\mathcal{X}$*, the limit actor model converges to the optimal policy:*

$$J(f^*)-J(f_{\bar{\theta}_t}) = O\left(\frac{1}{\log t}\right), \tag{2.3.6}$$

where f^* *is any optimal policy.*

2.4 Derivation of the limit ODEs

We use the following steps to prove convergence to the limit ODEs:

- Prove *a priori* bounds for the actor and critic models.
- Derive random ODEs for the evolution of the actor and critic models. The ODEs will contain stochastic remainder terms from the non-i.i.d. data samples.
- Use a Poisson equation to estimate the fluctuations of the remainder terms around zero.
- Use Gronwall's inequality to obtain the convergence to the limit ODEs.

2.4.1 A Priori Bounds

In order to prove convergence to the limit equation, we first establish some *priori* bounds for the parameters. In our proof, we use C, C_0 and C_T to denote generic constants. For notational convenience, we will sometimes use ξ,ξ' and $\xi_k,\widetilde{\xi}_k$ to denote the elements $(x,a),(x',a')$ and data samples $(x_k,a_k),(\widetilde{x}_k,\widetilde{a}_k)$ respectively.

First, we establish a priori estimates for the actor and critic models.

Lemma 2.4.1. *For any fixed $T > 0, N \in \mathbb{N}$, there exists a constant C_T which only depends on T such that*

$$\begin{aligned} \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |Q_k(x,a)| \le C_T < \infty, \quad \forall k \le NT \\ \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |\theta_k(x,a)| \le C_T < \infty, \quad \forall k \le NT. \end{aligned} \tag{2.4.1}$$

Proof. For the update algorithm in (2.2.8)

$$Q_{k+1}(\xi) = Q_k(\xi) + \frac{\alpha}{N}\left(r(\xi_k) + \gamma Q_k(x_{k+1}, a_{k+1}) - Q_k(\xi_k)\right)\partial_\xi Q_k(\xi_k), \tag{2.4.2}$$

we have the bound

$$\sup_{\xi\in\mathcal{X}\times\mathcal{A}} |Q_{k+1}(\xi)| \le \sup_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_k(\xi)| + \frac{C}{N} \sup_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_k(\xi)| + \frac{C}{N}. \tag{2.4.3}$$

Then, using a telescoping series, we have

$$\begin{aligned} \sup_{\xi\in\mathcal{X}\times\mathcal{A}} |Q_k(\xi)| &= \sup_{\xi\in\mathcal{X}\times\mathcal{A}} |Q_0(\xi)| + \sum_{j=1}^{k} \left(\sup_{\xi\in\mathcal{X}\times\mathcal{A}} |Q_j(\xi)| - \sup_{\xi\in\mathcal{X}\times\mathcal{A}} |Q_{j-1}(\xi)| \right) \\ &\le \sup_{\xi\in\mathcal{X}\times\mathcal{A}} |Q_0(\xi)| + \sum_{j=1}^{k} \left(\frac{C}{N} \sup_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_{j-1}(\xi)| + \frac{C}{N} \right) \\ &\le \sup_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_0(\xi)| + \frac{C}{N} \sum_{j=1}^{k} \sup_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_{j-1}(\xi)| + C \\ &\le C + \frac{C}{N} \sum_{j=1}^{k} \sup_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_{j-1}(\xi)|, \end{aligned} \tag{2.4.4}$$

where the last inequality follows from the fact that Q_0 is a fixed finite vector. Then, by the discrete Gronwall's lemma and using $\frac{k}{N} \le T$, we have

$$\sup_{(x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_k(x,a)| \le C \exp\left(\frac{Ck}{N}\right) \le C\exp(CT) = C_T, \quad \forall k \le NT. \tag{2.4.5}$$

Recall that the update for the actor model is

$$\begin{aligned} \theta_{k+1}(\xi) &= \theta_k(\xi) + \frac{\zeta_k^N}{N} Q_k(\widetilde{\xi}_k) \partial_\xi \log f_k(\widetilde{\xi}_k) \\ &= \theta_k(\xi) + \frac{\zeta_k^N}{N} Q_k(\widetilde{\xi}_k) \mathbb{1}_{\{\widetilde{x}_k = x\}} \left[\mathbb{1}_{\{\widetilde{a}_k = a\}} - f_k(\widetilde{x}_k, a) \right], \end{aligned} \tag{2.4.6}$$

which together with the bound for the critic in (2.4.5) leads to

$$\sup_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |\theta_{k+1}(\xi)| \le \sup_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |\theta_k(\xi)| + \frac{C_T}{N}. \tag{2.4.7}$$

Then, using a telescoping series, we immediately obtain the bound in the statement of the lemma. □

2.4.2 Evolution of the Pre-limit Process

From their definitions in (2.2.1), $V^f(x)$ and $V^f(x,a)$ are related via the formula

$$V^f(x) = \sum_a V^f(x,a) f(x,a). \tag{2.4.8}$$

Define the state and state-action visiting measures, respectively, as ν_μ^f and σ_μ^f, where

$$\nu_\mu^f(x) = \sum_{k=0}^{\infty} \gamma^k \cdot \mathbb{P}(x_k = x), \quad \sigma_\mu^f(x,a) = \sum_{k=0}^{\infty} \gamma^k \cdot \mathbb{P}(x_k = x, a_k = a), \tag{2.4.9}$$

where $x_0 \sim \mu(\cdot), a_k \sim f(x_k, \cdot)$ and $x_{k+1} \sim p(\cdot \mid x_k, a_k)$ for all $k \geq 0$. By definition, we have $\sigma_\mu^f(x,a) = f(x,a) \cdot \nu_\mu^f(x)$ and, by [83], the stationary distribution of $\widetilde{\boldsymbol{\mathcal{M}}}$ is the corresponding visitation measure of $\mathcal{M}$.

Notation We first clarify some notations.

(a) For any $k \geq 0$, let $\mathbb{P}_{\theta_k}$ denote the transition probability for the Markov chain (X, A) induced by $\boldsymbol{\mathcal{M}}$ under softmax policy f_k and let Π_{θ_k} denote the transition probability for the Markov chain $(\widetilde{X}, \widetilde{A})$ induced by $\widetilde{\boldsymbol{\mathcal{M}}}$ under exploration policy g_k. That is,

$$\begin{aligned} \mathbb{P}_{\theta_k}(x,a;x',a') &= p(x'|x,a) g_k(x',a'), \\ \Pi_{\theta_k}(x,a;x',a') &= \widetilde{p}(x'|x,a) f_k(x',a'). \end{aligned} \tag{2.4.10}$$

(b) Let $\sigma_\mu^{f_k}$ and π^{g_k} denote the stationary distributions (whose existence and uniqueness are given by Assumption 2.3.2) for the transition probability Π_{θ_k} and $\mathbb{P}_{\theta_k}$, respectively.

(c) Define the σ-field of events generated by the samples $\xi_1, \cdots, \xi_n, \widetilde{\xi}_1, \ldots, \widetilde{\xi}_n$ in (2.2.6) and (2.2.7) to be $\mathscr{F}_n$. Then, for any Borel function $h(\theta, \xi)$,

$$\mathbb{E}\left[h\left(\theta_n, \widetilde{\xi}_{n+1}\right) \mid \mathscr{F}_n\right] = \sum_{y \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} h(\theta_n, y) \Pi_{\theta_n}(\xi_n; y). \tag{2.4.11}$$

For any function $h(\theta, \xi)$, we shall denote the partial mapping $\xi \to h(\theta, \xi)$ by h_θ and define the function

$$\Pi_\theta h_\theta(\xi) := \sum_{y \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} h(\theta, y) \Pi_\theta(\xi; y).$$

Using the visiting measures in (2.4.9), the policy gradient can be evaluated using the following formula.

Theorem 2.4.2 (Policy Gradient Theorem [138]). *For the MDP starting from μ, the policy gradient for f_θ is*

$$\nabla_\theta J(f_\theta) = \sum_{x,a} \sigma_\mu^{f_\theta}(x,a) V^{f_\theta}(x,a) \nabla_\theta \log f_\theta(x,a), \tag{2.4.12}$$

Let the advantage function of policy f denoted by

$$A^f(x,a) = V^f(x,a) - V^f(x), \quad \forall (x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}, \tag{2.4.13}$$

and the gradient $\nabla_\theta J(f_\theta)$ can be evaluated using the following formula when f_θ satisfies the softmax policy (2.2.4).

Lemma 2.4.3. *Define $\partial_{x,a} J(f_\theta) := \frac{\partial J(f_\theta)}{\partial \theta(x,a)}$ and then for the tabular policy (2.2.4), by policy gradient theorem (2.4.12), we have*

$$\partial_{x,a} J(f_\theta) = \sigma_\mu^{f_\theta}(x,a) A^{f_\theta}(x,a). \tag{2.4.14}$$

Proof. By the policy gradient theorem, we have

$$\begin{aligned}
\partial_{x,a} J(f_\theta) &= \sum_{x',a'} \nu_\mu^{f_\theta}(x') f_\theta(x',a') \mathbb{1}_{\{x'=x\}} \left[\mathbb{1}_{\{a'=a\}} - f_\theta(x',a)\right] V^{f_\theta}(x',a') \\
&= \sum_{a'} \nu_\mu^{f_\theta}(x) f_\theta(x,a') \left[\mathbb{1}_{\{a'=a\}} - f_\theta(x,a)\right] V^{f_\theta}(x,a') \\
&= \nu_\mu^{f_\theta}(x) f_\theta(x,a) V^{f_\theta}(x,a) - \nu_\mu^{f_\theta}(x) f_\theta(x,a) \left[\sum_{a'} f_\theta(x,a') V^{f_\theta}(x,a')\right] \\
&= \nu_\mu^{f_\theta}(x) f_\theta(x,a) A^{f_\theta}(x,a) \\
&= \sigma_\mu^{f_\theta}(x,a) A^{f_\theta}(x,a).
\end{aligned} \tag{2.4.15}$$

□

Using a telescoping series and the update equation for the actor (2.2.8),

$$\theta_{\lfloor Nt \rfloor}(x,a) = \theta_0(x,a) + \frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \zeta_k^N Q_k(\widetilde{x}_k, \widetilde{a}_k) \partial_{x,a} \log f_k(\widetilde{x}_k, \widetilde{a}_k). \tag{2.4.16}$$

Note that $\xi = (x,a), \widetilde{\xi}_k = (\widetilde{x}_k, \widetilde{a}_k)$ and define

$$M_t^N(\xi) = \frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \zeta_k^N Q_k(\widetilde{\xi}_k) \partial_\xi \log f_k(\widetilde{\xi}_k) - \frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \sum_{\xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} \zeta_k^N Q_k(\xi') \partial_\xi \log f_k(\xi') \sigma_\mu^{f_k}(\xi'), \tag{2.4.17}$$

where $\sigma_\mu^{f_k}$ is the visiting measure for $\mathcal{M}$ under policy f_k. Combining (2.4.16) and (2.4.17), we obtain the following pre-limit equation for the actor parameters:

$$\begin{aligned}
&\theta_{\lfloor Nt \rfloor}(x,a) - \theta_0(x,a) \\
=&\frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \sum_{\xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} \zeta_k^N Q_k(\xi') \partial_\xi \log f_k(\xi') \sigma_\mu^{f_k}(\xi') + M_t^N(x,a) \\
=&\frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \zeta_k^N \sum_{a'} \nu_\mu^{f_k}(x) f_k(x,a') \left[\mathbb{1}_{\{a'=a\}} - f_k(x,a)\right] Q_k(x,a') + M_t^N(x,a) \\
=&\frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \zeta_k^N \sigma_\mu^{f_k}(x,a) \left[Q_k(x,a) - \sum_{a'} Q_k(x,a') f_k(x,a')\right] + M_t^N(x,a) \\
\overset{(a)}{=}& \int_0^t \zeta_{\lfloor Ns \rfloor}^N \sigma_\mu^{f_{\lfloor Ns \rfloor}}(x,a) \left[Q_{\lfloor Ns \rfloor}(x,a) - \sum_{a'} Q_{\lfloor Ns \rfloor}(x,a') f_{\lfloor Ns \rfloor}(x,a')\right] ds + M_t^N(x,a) + O(N^{-1}),
\end{aligned} \tag{2.4.18}$$

where step (a) uses the a priori bound for the critic Q_k in Lemma 2.4.1.

Similarly, we can show that the critic model satisfies

$$Q_{\lfloor Nt \rfloor}(\xi) = Q_0(\xi) + \frac{\alpha}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \left[r(\xi_k) + \gamma Q_k(x_{k+1}, a_{k+1}) - Q_k(\xi_k)\right] \partial_{x,a} Q_k(x_k, a_k).$$

Define

$$
\begin{aligned}
M_t^{1,N}(\xi) &= -\frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} Q_k(\xi_k)\partial_\xi Q_k(\xi_k) + \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{\xi'} Q_k(\xi')\partial_\xi Q_k(\xi')\pi^{g_k}(\xi'),\\
M_t^{2,N}(\xi) &= \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} r(\xi_k)\partial_\xi Q_k(\xi_k) - \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{\xi'} r(\xi')\partial_\xi Q_k(\xi')\pi^{g_k}(\xi'),\\
M_t^{3,N}(\xi) &= \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} \gamma Q_k(x_{k+1},a_{k+1})\partial_\xi Q_k(\xi_k)\\
&\quad - \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{\xi'}\sum_{z,a''}\gamma Q_k(z,a'')g_k(z,a'')\partial_\xi Q_k(\xi')\pi^{g_k}(\xi')p(z|\xi').
\end{aligned}
\tag{2.4.19}
$$

where π^{g_k} is the stationary distribution of Markov chain (X,A) induced by $\mathcal{M}$ under policy g_k. Note that

$$\partial_{x,a}Q_k(x_k,a_k) = \mathbb{1}_{\{x_k=x,a_k=a\}}.$$

Then, we obtain the following pre-limit equation for the critic:

$$
\begin{aligned}
Q_{\lfloor Nt\rfloor}(\xi) &= Q_0(\xi) + \alpha\int_0^t \pi^{g_{\lfloor Ns\rfloor}}(\xi)\left[r(\xi) + \gamma\sum_{z,a''} Q_{\lfloor Ns\rfloor}(z,a'')g_{\lfloor Ns\rfloor}(z,a'')p(z|\xi) - Q_{\lfloor Ns\rfloor}(\xi)\right]ds\\
&\quad + \alpha\left(M_t^{1,N}(\xi) + M_t^{2,N}(\xi) + M_t^{3,N}(\xi)\right) + O(N^{-1}).
\end{aligned}
\tag{2.4.20}
$$

2.4.3 Poisson Equations

Now we rigorously derive the limit ODEs by using a Poisson equation to bound the fluctuations of the non-i.i.d data samples around the trajectory of the limit ODE. In fact, we first prove

$$\lim_{N\to\infty}\mathbb{E}\sup_{t\in[0,T]}\left|M_t^N(x,a)\right| = 0, \quad \forall (x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}. \tag{2.4.21}$$

Using a similar method, we can also prove the convergence of M_t^1, M_t^2, and M_t^3.

It is known that a finite state Markov chain which is irreducible and non-periodic has a geometric convergence rate to its stationary distribution [106]. We are able to prove a uniform geometric convergence rate for the Markov chains in this chapter under the *time-evolving actor policy updated using the actor-critic algorithm* (2.2.8).

Lemma 2.4.4. *Let $\Pi_{\theta_k}^n$ denote the n-step transition matrix under the policy f_{θ_k}. Then, for any fixed $T>0$, there exists an integer n_0 such that the following uniform estimates hold for all $\{\theta_k\}_{0\le k\le NT}$ and $N\in\mathbb{N}$ for the algorithm* (2.2.8).

- *Lower bound for the stationary distribution:*

$$\inf_{k\le NT}\sigma_\mu^{f_k}(x,a) \ge C\epsilon_T^{n_0}, \quad \forall (x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}, \tag{2.4.22}$$

  where $C,\epsilon_T>0$ are positive constants.

- *Uniform geometric ergodicity:*

$$\sup_{k\leq NT}\|\Pi_{\theta_k}^n(\xi;\cdot)-\sigma_\mu^{f_k}(\cdot)\|\leq(1-\beta_T)^{\lfloor\frac{n}{n_0}\rfloor}\quad\forall\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}},\tag{2.4.23}$$

where $\beta_T\in(0,1)$ is a positive constant.

Proof. By Assumption 2.3.2 and Lemma 1.8.2 of [112], for any fixed $\widetilde{\theta}\in\mathbb{R}^d$, there exists an $n_0=n_0(\widetilde{\theta})\in\mathbb{N}$ such that

$$\Pi_{\widetilde{\theta}}^{n_0}(\xi;\xi')>0\quad\forall\xi,\xi'.\tag{2.4.24}$$

For any (ξ,ξ'),

$$\begin{aligned}\Pi_{\widetilde{\theta}}^{n_0}(\xi;\xi')&=\sum_{\xi_1,\cdots,\xi_{n_0-1}}\Pi_{\widetilde{\theta}}(\xi;\xi_1)\cdots\Pi_{\widetilde{\theta}}(\xi_{n_0-1};\xi')\\&=\sum_{\xi_1,\cdots,\xi_{n_0-1}}\widetilde{p}(x_1|x,a)f_{\widetilde{\theta}}(x_1,a_1)\cdots\widetilde{p}(x'|x_{n_0-1},a_{n_0-1})f_{\widetilde{\theta}}(x',a'),\end{aligned}\tag{2.4.25}$$

where the constant C is defined as

$$C=C(n_0):=\inf_{x,a,x'}\sum_{\xi_1,\cdots,\xi_{n_0-1}}\widetilde{p}(x_1|x,a)\cdots\widetilde{p}(x'|x_{n_0-1},a_{n_0-1})>0,\tag{2.4.26}$$

where $C>0$ is because (2.4.24).

Due to f_θ being a softmax policy and the bound from Lemma 2.4.1, there exists a constant $\epsilon_T>0$ such that

$$\inf_{k\leq NT}f_k(x,a)>\epsilon_T,\quad\forall(x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}.\tag{2.4.27}$$

Then, using similar analysis as in (2.4.25) with constant $n_0=n_0(\widetilde{\theta})$ and $C=C(n_0)$, we have for all $k\leq NT$

$$\Pi_{\theta_k}^{n_0}(\xi;\xi')\geq C\epsilon_T^{n_0},\quad\forall\xi,\xi'.\tag{2.4.28}$$

Thus, we can derive a lower bound for the stationary distribution

$$\begin{aligned}\inf_{k\leq NT}\sigma_\mu^{f_k}(x',a')&=\inf_{k\leq NT}\sum_{x,a}\sigma_\mu^{f_k}(x,a)\Pi_{\theta_k}^{n_0}(x,a;x',a')\\&\geq\inf_{k\leq NT}\sum_{x,a}\sigma_\mu^{f_k}(x,a)C\epsilon_T^{n_0}\\&\overset{(a)}{=}C\epsilon_T^{n_0}\\&>0,\end{aligned}\tag{2.4.29}$$

where step (a) is because σ^{f_θ} is a probability and thus the summation equals to 1. We can now establish the uniform geometric ergodicity of the Markov chain. Let us choose $\beta_T=\inf_{k\leq NT}\min_{\xi,\xi'}\Pi_{\theta_k}^{n_0}(\xi,\xi')>0$ in (2.4.23), where $\beta_T>0$ is by (2.4.28). Thus, for $\forall k\leq NT$, the Markov chain with transition probability Π_{θ_k} satisfies Doeblin's condition. In particular, we can show that

$$\Pi_{\theta_k}^{n_0}(\xi,\xi')\geq\beta_T>0,\quad\forall\xi,\xi'.\tag{2.4.30}$$

Since n_0 and β_T are independent of θ_k, we can apply Theorem 16.2.4 of [106] to prove that for all $k\leq NT$

$$\|\Pi_{\theta_k}^n(\xi;\cdot)-\sigma_\mu^{f_k}(\cdot)\|\leq(1-\beta_T)^{\lfloor\frac{n}{n_0}\rfloor}\quad\forall\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}},\tag{2.4.31}$$

which proves the uniform geometric ergodicity (2.4.23). □

Then, using the same method as in Lemma 2.4.4, we can prove a similar result for the MDP $\boldsymbol{\mathcal{M}}$ with exploration policy g_k.

Corollary 2.4.5. *Let $\mathbb{P}^n_{\theta_k}$ denote the n-step transition matrix under policy g_k. Then, for any fixed $T < \infty$, there exists an integer n_0 and a constant*

$$C = C(n_0) := \inf_{x,a,x'} \sum_{\xi_1,\cdots,\xi_{n_0-1}} p(x_1|x,a)\cdots p(x'|x_{n_0-1}, a_{n_0-1}) > 0, \tag{2.4.32}$$

such that the following uniform estimate holds for all $\{\theta_k\}_{0\le k\le NT}$ and $N \in \mathbb{N}$ for the update algorithm (2.2.8):

- *Lower bound for the stationary distribution:*

$$\inf_{k\le NT} \pi^{g_k}(x,a) \ge C\left(\eta^N_{\lfloor NT\rfloor}\right)^{n_0}, \quad \forall (x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}. \tag{2.4.33}$$

- *Uniform geometric ergodicity:*

$$\sup_{k\le NT} \|\mathbb{P}^n_{\theta_k}(\xi;\cdot) - \pi^{g_k}(\cdot)\| \le (1-\beta_T)^{\lfloor \frac{n}{n_0}\rfloor} \quad \forall \xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}, \tag{2.4.34}$$

  where $\beta_T = C\left(\eta^N_{\lfloor NT\rfloor}\right)^{n_0} \in (0,1)$ is a positive constant.

Remark 2.4.6. Without loss of generality, we suppose the integer n_0 in Lemma 2.4.4 and Corollary 2.4.5 are the same. The proof of Corollary 2.4.5 is the same as the proof of Lemma 2.4.4 and the detailed proof can be found in Appendix A.2.

In order to prove the stochastic fluctuation term vanishes as $N \to \infty$, we first introduce a Poisson equation with a uniformly bounded solution.

Lemma 2.4.7. *For any $N \in \mathbb{N}$, state-action pair $\xi = (x,a)$, $T > 0$ and $k \le NT$, the Poisson equation*

$$\nu_{\theta_k}(\xi') - \Pi_{\theta_k}\nu_{\theta_k}(\xi') = \mathbb{1}_{\{\xi'=\xi\}} - \sigma^{f_k}(\xi), \quad \xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}} \tag{2.4.35}$$

has a solution[2]

$$\nu_{\theta_k}(\xi') := \sum_{n\ge 0} \left[\Pi^n_{\theta_k}(\xi';\xi) - \sigma^{f_k}(\xi)\right], \tag{2.4.36}$$

and there exists a constant C_T (which only depends on T) such that

$$\sup_{k\le NT} |\nu_{\theta_k}(\xi')| \le C_T, \quad \forall \xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}. \tag{2.4.37}$$

Proof. Due to the uniform geometric convergence rate (2.4.23) for all $k \le NT$ in Lemma 2.4.4, there exists a $\beta_T > 0$ (independent with k) such that for any $\xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$

$$\left|\Pi^n_{\theta_k}(\xi';\xi) - \sigma^{f_k}_{\mu}(\xi)\right| \le (1-\beta_T)^{\lfloor \frac{n}{n_0}\rfloor}, \quad \forall k \le NT \tag{2.4.38}$$

[2]We do not prove uniqueness of the solution to the Poisson equation (2.4.35). For the purposes of our later analysis, it is only necessary to find a uniformly bounded solution ν_θ which satisfies (2.4.36).

which can be used to show the convergence of the series in (2.4.36). Consequently, ν_{θ_k} is well-defined. The uniform bound (2.4.37) follows from

$$|\nu_{\theta_k}(\xi')| \leq \sum_{n\geq 0} \left|\Pi^n_{\theta_k}(\xi';\xi) - \sigma^{f_k}_\mu(\xi)\right| \leq \sum_{n\geq 0}(1-\beta_T)^{\lfloor \frac{n}{n_0}\rfloor} \leq C_T. \tag{2.4.39}$$

Finally, we can verify that ν_θ is a solution to the Poisson equation by observing that

$$\begin{aligned}
\Pi_{\theta_k}\nu_{\theta_k}(\xi') &= \sum_y \nu_{\theta_k}(y)\Pi_{\theta_k}(\xi';y) \\
&= \sum_y \left(\sum_{n\geq 0}\left[\Pi^n_{\theta_k}(y;\xi) - \sigma^{f_k}_\mu(\xi)\right]\right)\Pi_{\theta_k}(\xi';y) \\
&\overset{(a)}{=} \sum_{n\geq 0}\left(\sum_y \left[\Pi^n_{\theta_k}(y;\xi) - \sigma^{f_k}_\mu(\xi)\right]\Pi_{\theta_k}(\xi';y)\right) \\
&= \sum_{n\geq 1}\left[\Pi^n_{\theta_k}(\xi';\xi) - \sigma^{f_k}_\mu(\xi)\right] \\
&= \nu_{\theta_k}(\xi') - (\mathbb{1}_{\{\xi'=\xi\}} - \sigma^{f_k}_\mu(\xi)),
\end{aligned} \tag{2.4.40}$$

where the step (a) uses (2.4.38) and the Dominated Convergence Theorem. □

Using the Poisson equation (2.4.7), we can prove that the fluctuations of the data samples around a dynamic visiting measure $\sigma^{f_k}_\mu$ decay when the iteration steps become large.

Lemma 2.4.8. *For any fixed state action pair $\xi = (x,a)$ and $T > 0$,*

$$\lim_{N\to\infty} \mathbb{E}\left|\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\left[\mathbb{1}_{\{\widetilde{\xi}_k=\xi\}} - \sigma^{f_k}_\mu(\xi)\right]\right| = 0. \tag{2.4.41}$$

Proof. We define the error ϵ_k to be

$$\begin{aligned}
\epsilon_k :=& \mathbb{1}_{\{\widetilde{\xi}_{k+1}=\xi\}} - \sigma^{f_k}_\mu(\xi) \\
=& \nu_{\theta_k}\left(\widetilde{\xi}_{k+1}\right) - \Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_{k+1}\right) \\
=& \left[\nu_{\theta_k}(\widetilde{\xi}_{k+1}) - \Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_k\right)\right] + \left[\Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_k\right) - \Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_{k+1}\right)\right],
\end{aligned} \tag{2.4.42}$$

where we have used the definition of the Poisson equation (2.4.35). Let

$$\psi_\theta(y) = \Pi_\theta\nu_\theta(y). \tag{2.4.43}$$

Then, we have that

$$\begin{aligned}
\sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon_k &= \sum_{k=0}^{\lfloor NT\rfloor-1}\left[\nu_{\theta_k}(\widetilde{\xi}_{k+1}) - \Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_k\right)\right] + \sum_{k=0}^{\lfloor NT\rfloor-1}\left[\left(\psi_{\theta_k}\left(\widetilde{\xi}_k\right) - \psi_{\theta_k}\left(\widetilde{\xi}_{k+1}\right)\right)\right] \\
&= \sum_{k=0}^{\lfloor NT\rfloor-1}\left[\nu_{\theta_k}(\widetilde{\xi}_{k+1}) - \Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_k\right)\right] + \sum_{k=1}^{\lfloor NT\rfloor-1}\left[\psi_{\theta_k}\left(\widetilde{\xi}_k\right) - \psi_{\theta_{k-1}}\left(\widetilde{\xi}_k\right)\right] \\
&\quad + \psi_{\theta_0}\left(\widetilde{\xi}_0\right) - \psi_{\theta_{\lfloor NT\rfloor-1}}\left(\widetilde{\xi}_{\lfloor NT\rfloor}\right)
\end{aligned} \tag{2.4.44}$$

Define the error term

$$\sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon_k = \sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon^{(1)}_k + \sum_{k=1}^{\lfloor NT\rfloor-1}\epsilon^{(2)}_k + \rho_{\lfloor NT\rfloor;0}, \tag{2.4.45}$$

where

$$\begin{aligned}\epsilon_k^{(1)} &= \zeta_k^N \left[\nu_{\theta_k}\left(\widetilde{\xi}_{k+1}\right) - \Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_k\right)\right], \\ \epsilon_k^{(2)} &= \zeta_k^N \left[\psi_{\theta_k}\left(\widetilde{\xi}_k\right) - \psi_{\theta_{k-1}}\left(\widetilde{\xi}_k\right)\right], \\ \rho_{\lfloor NT\rfloor;0} &= \zeta_0^N \psi_{\theta_0}\left(\widetilde{\xi}_0\right) - \zeta_{\lfloor NT\rfloor-1}^N \psi_{\theta_{\lfloor NT\rfloor-1}}\left(\widetilde{\xi}_{\lfloor NT\rfloor}\right).\end{aligned} \tag{2.4.46}$$

To prove the convergence (2.4.41), it suffices to appropriately bound the fluctuation term $\left|\sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon_k\right|$. Actually, the first term can be bound due to the martingale property while the second term can be bounded using the uniform geometric ergodicity and Lipschitz continuity. The third and fourth terms are uniformly bounded by (2.4.37).

For the first term in (2.4.45), note that

$$\mathbb{E}\left\{\nu_{\theta_k}\left(\widetilde{\xi}_{k+1}\right) \mid \mathscr{F}_k\right\} = \Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_k\right), \tag{2.4.47}$$

thus

$$\left\{Z_n = \sum_{k=0}^{n-1}\gamma_k^{(1)},\ \mathscr{F}_n\right\}_{n\geq 0}$$

is a martingale and since the conditional expectation is a contraction in L^2, we have

$$\mathbb{E}\left|\Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_k\right)\right|^2 \leq \mathbb{E}\left|\nu_{\theta_k}\left(\widetilde{\xi}_{k+1}\right)\right|^2. \tag{2.4.48}$$

Then,

$$\begin{aligned}\mathbb{E}\left|\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon_k^{(1)}\right|^2 &= \frac{1}{N^2}\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbb{E}\left|\Pi_{\theta_k}\nu_{\theta_k}\left(\widetilde{\xi}_k\right) - \nu_{\theta_k}\left(\widetilde{\xi}_{k+1}\right)\right|^2 \\ &\leq \frac{4}{N^2}\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbb{E}\left|\nu_{\theta_k}\left(\widetilde{\xi}_{k+1}\right)\right|^2 \\ &\overset{(a)}{\leq} \frac{4C_T}{N},\end{aligned} \tag{2.4.49}$$

where the step (a) is by the uniform boundedness (2.4.37). Thus, for any $T > 0$,

$$\lim_{N\to\infty}\mathbb{E}\left|\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon_k^{(1)}\right| = 0. \tag{2.4.50}$$

For the second term of (2.4.45), by the uniform geometric ergodicity (2.4.23), for any fixed $\gamma_0 > 0$ we can choose N_0 large enough such that

$$\sup_{k\leq NT}\sum_{n=\lfloor N_0 T\rfloor}^{\infty}\left|\Pi_{\theta_k}^n(y,\xi) - \sigma_\mu^{f_k}(\xi)\right| < \gamma_0, \quad \forall y \in \boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}} \tag{2.4.51}$$

$$
\begin{aligned}
&\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\epsilon_k^{(2)}\right| \\
=&\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\left[\psi_{\theta_k}\left(\widetilde{\xi}_k\right)-\psi_{\theta_{k-1}}\left(\widetilde{\xi}_k\right)\right]\right| \\
\leq&\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\left[\sum_{n=1}^{\lfloor N_0T\rfloor-1}\left[\Pi_{\theta_k}^n\left(\widetilde{\xi}_k,\xi\right)-\sigma_\mu^{f_k}(\xi)\right]-\sum_{n=1}^{\lfloor N_0T\rfloor-1}\left[\Pi_{\theta_{k-1}}^n\left(\widetilde{\xi}_k,\xi\right)-\sigma_\mu^{f_{k-1}}(\xi)\right]\right]\right|+2C_T\gamma_0 \\
\leq&\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\sum_{n=1}^{\lfloor N_0T\rfloor-1}\left[\Pi_{\theta_k}^n\left(\widetilde{\xi}_k,\xi\right)-\Pi_{\theta_{k-1}}^n\left(\widetilde{\xi}_k,\xi\right)\right]\right|+\frac{\lfloor N_0T\rfloor}{N}\left|\sum_{k=1}^{\lfloor NT\rfloor-1}\left[\sigma_\mu^{f_k}(\xi)-\sigma_\mu^{f_{k-1}}(\xi)\right]\right|+2C_T\gamma_0 \\
:=&I_1^N+I_2^N+2C_T\gamma_0.
\end{aligned}
\tag{2.4.52}
$$

By Lemma 2.4.1, for any $k\leq NT$ we have

$$
\|\theta_k-\theta_{k-1}\|\leq\sum_{x,a\in\mathcal{X}\times\mathcal{A}}|\theta_k(x,a)-\theta_{k-1}(x,a)|\leq\frac{C_T}{N}
$$

For any finite n, Π_θ^n is Lipschitz continuous in θ. Consequently,

$$
\begin{aligned}
I_1^N&\leq\frac{\lfloor N_0T\rfloor}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\zeta_k^N C\left\|\theta_k-\theta_{k-1}\right\|\leq\frac{C_T}{N}, \\
I_2^N&\leq\frac{\lfloor N_0T\rfloor}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\zeta_k^N C\left\|\theta_k-\theta_{k-1}\right\|\leq\frac{C_T}{N},
\end{aligned}
\tag{2.4.53}
$$

where the constant C_T only depends on the fixed N_0,T. Thus, when N is large enough,

$$
\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\epsilon_k^{(2)}\right|\leq 4C_T\gamma_0 \tag{2.4.54}
$$

Since γ_0 is arbitrary,

$$
\lim_{N\to\infty}\mathbf{E}\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\epsilon_k^{(2)}\right|=0 \tag{2.4.55}
$$

Finally, for the last term of (2.4.45) by the boundedness in (2.4.37) we have

$$
\lim_{N\to\infty}\frac{1}{N}\rho_{\lfloor NT\rfloor;0}=0,
$$

which together with (2.4.50) and (2.4.55) derive the convergence of $\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon_k$ and therefore proving (2.4.41). □

2.4.4 Identification of the Limit ODEs

We next prove the convergence of M_t^N, which will allow us to prove the convergence to the limit ODEs (2.3.4).

Lemma 2.4.9. *For any $\xi=(x,a)$ and the stochastic error M_t^N defined in (2.4.17), we have*

$$
\lim_{N\to\infty}\sup_{t\in(0,T]}\mathbb{E}\left|M_t^N(\xi)\right|=0. \tag{2.4.56}
$$

Proof. The proof of (2.4.56) consists of two parts. We first set up a bound for the difference of the actor's update. Actually, define

$$H_{\xi,\xi',k} := \zeta_k Q_k(\xi')\partial_\xi \log f_k(\xi') \tag{2.4.57}$$

and we aim to prove

$$\left|H^N_{\xi,\xi',k+1} - H^N_{\xi,\xi',k}\right| \le \frac{C_T}{N} \tag{2.4.58}$$

Then we can use Lemma 2.4.8 to prove that as the training step becomes large, the fluctuations of the data samples around the stationary distribution will disappear, which concludes the result.

(i) To bound the difference (2.4.58), note that

$$\begin{aligned}
&|H_{\xi,\xi',k+1} - H_{\xi,\xi',k}| \\
\le& \left|\zeta^N_{k+1} - \zeta^N_k\right| \cdot |Q_{k+1}(\xi')\partial_\xi \log f_{k+1}(\xi')| \\
&+\zeta^N_k \cdot |Q_{k+1}(\xi')\partial_\xi \log f_{k+1}(\xi') - Q_k(\xi')\partial_\xi \log f_k(\xi')| \\
:=& I^N_3 + I^N_4.
\end{aligned} \tag{2.4.59}$$

For the first term,

$$I^N_3 \le C\left|\zeta^N_{k+1} - \zeta^N_k\right| \le C\left(\frac{1}{1+\frac{k}{N}} - \frac{1}{1+\frac{k+1}{N}}\right) = \frac{C}{N\left(1+\frac{k}{N}\right)\left(1+\frac{k+1}{N}\right)} \le \frac{C}{N}. \tag{2.4.60}$$

Then for the second term, using the bound in Lemma 2.4.1, we have for any $k \le TN$

$$\begin{aligned}
&\sup_{\xi' \in \mathcal{X}\times\mathcal{A}} |Q_k(\xi')| \le C_T, \\
&\sup_{\xi'} |Q_k(\xi') - Q_{k-1}(\xi')| \le \frac{C_T}{N}.
\end{aligned} \tag{2.4.61}$$

Noting that

$$\partial_\xi \log f_k(\xi') = \mathbb{1}_{\{x'=x\}}\left[\mathbb{1}_{\{a'=a\}} - f_k(x',a)\right],$$

then by the Lipschitz continuity of the softmax transformation and (2.4.61) we have

$$\begin{aligned}
&|Q_k(\xi')\partial_\xi \log f_k(\xi') - Q_{k-1}(\xi')\partial_\xi \log f_{k-1}(\xi')| \\
=& \left|\left[Q_k(\xi') - Q_{k-1}(\xi')\right]\partial_\xi \log f_k(\xi')\right| + \left|Q_{k-1}(\xi')\left[\partial_\xi \log f_k(\xi') - \partial_\xi \log f_{k-1}(\xi')\right]\right| \\
\le& \frac{C_T}{N} + C_T\left|\partial_\xi \log f_k(\xi') - \partial_\xi \log f_{k-1}(\xi')\right| \\
\le& \frac{C_T}{N} + C_T\|\theta_k - \theta_{k-1}\| \le \frac{C_T}{N}.
\end{aligned} \tag{2.4.62}$$

(ii) Now we can prove the convergence (2.4.56). Actually, for any $K \in \mathbb{N}$ and $\Delta = \frac{t}{K}$, we have

$$\begin{aligned}
M_t^N(\xi) &= \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=j\lfloor \Delta N \rfloor}^{(j+1)\lfloor \Delta N \rfloor - 1} \left(H_{\xi,\tilde{\xi}_k,k} - \sum_{\xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} H_{\xi,\xi',k} \sigma_\mu^{f_k}(\xi') \right) + o(1) \\
&= \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=j\lfloor \Delta N \rfloor}^{(j+1)\lfloor \Delta N \rfloor - 1} \left(H_{\xi,\tilde{\xi}_k,j\lfloor \Delta N \rfloor} - \sum_{\xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} H_{\xi,\xi',j\lfloor \Delta N \rfloor} \sigma_\mu^{f_k}(\xi') \right) \\
&+ \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=j\lfloor \Delta N \rfloor}^{(j+1)\lfloor \Delta N \rfloor - 1} \Bigg[\left(H_{\xi,\tilde{\xi}_k,k} - \sum_{\xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} H_{\xi,\xi',k} \sigma_\mu^{f_k}(\xi') \right) \\
&- \left(H_{\xi,\tilde{\xi}_k,j\lfloor \Delta N \rfloor} - \sum_{\xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} H_{\xi,\xi',j\lfloor \Delta N \rfloor} \sigma_\mu^{f_k}(\xi') \right) \Bigg] + o(1) \\
&:= \sum_{j=0}^{K-1} \Delta I_{1,j}^N + \sum_{j=0}^{K-1} \Delta I_{2,j}^N + o(1),
\end{aligned} \tag{2.4.63}$$

where the term $o(1)$ goes to zero, at least, in L^1 as $N \to \infty$.

To prove the convergence of the first term, note that

$$H_{\xi,\tilde{\xi}_k,j\lfloor \Delta N \rfloor} - \sum_{\xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} H_{\xi,\xi',j\lfloor \Delta N \rfloor} \sigma_\mu^{f_k}(\xi') = \sum_{\xi'} H_{\xi,\xi',j\lfloor \Delta N \rfloor} \left[\mathbb{1}_{\{\xi_k = \xi'\}} - \sigma_\mu^{f_k}(\xi') \right]. \tag{2.4.64}$$

Thus, for any $j \in \{0, 1, \ldots, K\}$,

$$\begin{aligned}
\left| I_{1,j}^N \right| &= \left| \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=j\lfloor \Delta N \rfloor}^{(j+1)\lfloor \Delta N \rfloor - 1} \sum_{\xi'} H_{\xi,\xi',j\lfloor \Delta N \rfloor} \left[\mathbb{1}_{\{\xi_k = \xi'\}} - \sigma_\mu^{f_k}(\xi') \right] \right| \\
&= \left| \sum_{\xi'} H_{\xi,\xi',j\lfloor \Delta N \rfloor} \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=j\lfloor \Delta N \rfloor}^{(j+1)\lfloor \Delta N \rfloor - 1} \left[\mathbb{1}_{\{\xi_k = \xi'\}} - \sigma_\mu^{f_k}(\xi') \right] \right| \\
&\leq C \sum_{\xi'} \left| \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=j\lfloor \Delta N \rfloor}^{(j+1)\lfloor \Delta N \rfloor - 1} \left[\mathbb{1}_{\{\xi_k = \xi'\}} - \sigma_\mu^{f_k}(\xi') \right] \right|,
\end{aligned} \tag{2.4.65}$$

which together with Lemma 2.4.8 derive

$$\lim_{N \to \infty} \mathbb{E} \left| I_{1,j}^N \right| = 0. \tag{2.4.66}$$

Thus

$$\sum_{j=0}^{K-1} \Delta I_{1,j}^N = \Delta \sum_{j=0}^{K-1} O(1) = t \frac{\sum_{j=0}^{K-1} O(1)}{K}, \tag{2.4.67}$$

which proves the convergence of the first term.

For the second term, use 2.4.58 and we have for any $j \in 0, 1, \cdots, K-1$ and any $k \in [j\lfloor \Delta N \rfloor, (j+1)\lfloor \Delta N \rfloor - 1]$,

$$\left| H_{\xi,\xi',k} - H_{\xi,\xi',j\lfloor \Delta N \rfloor} \right| \leq \frac{C(k - j\lfloor \Delta N \rfloor)}{N}. \tag{2.4.68}$$

Therefore,

$$\begin{aligned}
\sum_{j=0}^{K-1} \Delta I_{2,j}^N &\leq C \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=j\lfloor \Delta N \rfloor}^{(j+1)\lfloor \Delta N \rfloor - 1} \frac{k - j\lfloor \Delta N \rfloor}{N} \\
&= C \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=0}^{\lfloor \Delta N \rfloor - 1} \frac{k}{N} \\
&\leq C \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \frac{\lfloor \Delta N \rfloor^2}{N} \\
&\leq C \sum_{j=0}^{K-1} \Delta \frac{\lfloor \Delta N \rfloor}{N} \\
&\leq C \sum_{j=0}^{K-1} \Delta^2 \\
&\leq C\Delta.
\end{aligned} \tag{2.4.69}$$

Collecting our results, we have shown that

$$\lim_{N\to\infty} \sup_{t\in(0,T]} \mathbb{E} \left| M_t^N \right| \leq C \frac{T}{K} \tag{2.4.70}$$

Note that K was arbitrary. Consequently, we obtain

$$\lim_{N\to\infty} \sup_{t\in(0,T]} \mathbb{E} \left| M_t^N \right| = 0, \tag{2.4.71}$$

□

Following the same method, we can finish proving the convergence of the stochastic fluctuation terms and the detailed proof can be found in Appendix A.3.

Lemma 2.4.10. *For $t \in [0,T]$, $M_t^{1,N}$, $M_t^{2,N}$, $M_t^{3,N} \xrightarrow{L^1} 0$ as $N \to \infty$.*

Using Lemma 2.4.9 and 2.4.10, we can now finish the derivation of the limit ODEs.

Proof of Theorem 2.3.5: Due to Assumption 2.3.2 and the Lipschitz continuity of softmax transformation, we know $\sigma_\mu^{f_{\bar{\theta}_t}}, \pi^{g_{\bar{\theta}_t}}$ is Lipschitz continuous in $\bar{\theta}_t$. By Theorem 2.2 and Theorem 2.17 of [141], for any initial value, there exists a unique solution on $(0, +\infty)$ for the ODE system (2.3.4). Let $\bar{Q}_t(x,a)$, $\bar{\theta}_t(x,a)$ be the solution of (2.3.4) with initial value Q_0, θ_0. Using the bound in Lemma 2.4.1 and the Lipschitz continuity from Assumption 2.3.2, we have for $t \in [0,T]$

$$\begin{aligned}
&\left| \sigma_\mu^{f_{\lfloor Nt \rfloor}}(x,a) Q_{\lfloor Nt \rfloor}(x,a) - \sigma_\mu^{f_t}(x,a) Q_t(x,a) \right| \\
\leq &\left| \sigma_\mu^{f_{\lfloor Nt \rfloor}}(x,a) - \sigma_\mu^{f_t}(x,a) \right| \cdot \left| Q_{\lfloor Nt \rfloor}(x,a) \right| + \sigma_\mu^{f_t}(x,a) \left| Q_{\lfloor Nt \rfloor}(x,a) - Q_t(x,a) \right| \\
\leq &C_T \left[\left\| \theta_{\lfloor Nt \rfloor} - \bar{\theta}_t \right\| + \left\| Q_{\lfloor Nt \rfloor} - \bar{Q}_t \right\| \right],
\end{aligned} \tag{2.4.72}$$

and we can also show for the exploration policy from (2.2.9) that

$$\begin{aligned}
&\left|g_{\lfloor Nt\rfloor}(x,a)-g_t(x,a)\right| \\
\leq&\frac{\left|\eta^N_{\lfloor Nt\rfloor}-\eta_t\right|}{d_A}+\left|(1-\eta^N_{\lfloor Nt\rfloor})\cdot f_{\theta_{\lfloor Nt\rfloor}}(x,a)-(1-\eta_t)\cdot f_{\theta_t}(x,a)\right| \\
=&\frac{\left|\eta^N_{\lfloor Nt\rfloor}-\eta_t\right|}{d_A}+\left|f_{\theta_{\lfloor Nt\rfloor}}(x,a)-f_{\theta_t}(x,a)\right|+\left|\eta^N_{\lfloor Nt\rfloor}f_{\theta_{\lfloor Nt\rfloor}}(x,a)-\eta_t f_{\theta_t}(x,a)\right| \\
\leq&C\left|\eta^N_{\lfloor Nt\rfloor}-\eta_t\right|+C\left\|\theta_{\lfloor Nt\rfloor}-\bar\theta_t\right\|.
\end{aligned} \tag{2.4.73}$$

Combining (2.4.18), (2.4.20), and (2.3.4) and using the same decomposition method as in (2.4.72), we have for $t\in[0,T]$

$$\begin{aligned}
&\left\|\theta_{\lfloor Nt\rfloor}-\bar\theta_t\right\|+\left\|Q_{\lfloor Nt\rfloor}-\bar Q_t\right\| \\
\leq&\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left[\left|\theta_{\lfloor Nt\rfloor}(x,a)-\bar\theta_t(x,a)\right|+\left|Q_{\lfloor Nt\rfloor}(x,a)-\bar Q_t(x,a)\right|\right] \\
\leq&C_T\int_0^t\left[\left\|\theta_{\lfloor Ns\rfloor}-\bar\theta_t\right\|+\left\|Q_{\lfloor Ns\rfloor}-\bar Q_t\right\|\right]ds+\left|M^N_t\right|+\sum_{i=1}^3\left|M^{i,N}_t\right|+O(N^{-1}) \\
&+C_T\int_0^t\left[\left|\zeta^N_{\lfloor Ns\rfloor}-\zeta_s\right|+\left|\eta^N_{\lfloor Ns\rfloor}-\eta_s\right|\right]ds.
\end{aligned} \tag{2.4.74}$$

Define

$$\begin{aligned}
\varphi^N_t&:=\left\|\theta_{\lfloor Nt\rfloor}-\bar\theta_t\right\|+\left\|Q_{\lfloor Nt\rfloor}-\bar Q_t\right\| \\
B^N_t&:=\left|M^N_t\right|+\sum_{i=1}^3\left|M^{i,N}_t\right|+O(N^{-1})+C_T\int_0^t\left[\left|\zeta^N_{\lfloor Ns\rfloor}-\zeta_s\right|+\left|\eta^N_{\lfloor Ns\rfloor}-\eta_s\right|\right]ds.
\end{aligned} \tag{2.4.75}$$

Due to Lemma 2.4.9 and 2.4.10,

$$\lim_{N\to\infty}\mathbb{E}\sup_{t\in[0,T]}B^N_t=0. \tag{2.4.76}$$

Taking the supremum and expectation of (2.4.74),

$$\mathbb{E}\sup_{s\in[0,t]}\varphi^N_s\leq C_T\int_0^t\mathbb{E}\sup_{r\in[0,s]}\varphi^N_r ds+\mathbb{E}\sup_{s\in[0,t]}B^N_s,\quad\forall t\in[0,T] \tag{2.4.77}$$

By Gronwall's lemma, we have

$$\mathbb{E}\sup_{t\in[0,T]}\varphi^N_t\leq\mathbb{E}\sup_{t\in[0,T]}B^N_t+C_T\int_0^T\mathbb{E}\sup_{s\in[0,t]}B^N_s dt\leq C_T\mathbb{E}\sup_{t\in[0,T]}B^N_t, \tag{2.4.78}$$

which together with (2.4.76) proves the convergence (2.3.3). □

2.5 Convergence of Limit ODEs

We now study the convergence of the limit actor-critic algorithm, which satisfies the ODE system (2.3.4).

2.5.1 Critic convergence

Now we prove convergence of the critic (2.3.5), which states that the critic model will converge to the state-action value function during training. We first derive an ODE for the difference between the critic and the value function. Then, we use a comparison lemma, a two time-scale analysis, and

the properties of the learning and exploration rates (2.3.2) to prove the convergence of the critic to the value function.

Recall that the value function V^{g_t} satisfies the Bellman equation

$$r(x,a)+\gamma\sum_{z,a''}V^{g_{\bar{\theta}_t}}(z,a'')g_{\bar{\theta}_t}(z,a'')p(z|x,a)-V^{g_{\bar{\theta}_t}}(x,a)=0. \tag{2.5.1}$$

Define the difference

$$\phi_t=\bar{Q}_t-V^{g_{\bar{\theta}_t}}. \tag{2.5.2}$$

As a first step, we prove an a priori uniform bound for the critic in the update (2.3.4). Without loss of generality, we initialize the ODE as $\bar{Q}_0=0$ (we can always define $\bar{Q}'_t=\bar{Q}_t-\bar{Q}_0$ and prove the uniform bound for Q'_t).

Lemma 2.5.1. *For any state x and action a, we have*

$$\max_{x,a}\left|\bar{Q}_t(x,a)\right|\leq\frac{2}{1-\gamma},\quad t\geq 0. \tag{2.5.3}$$

Proof. We first prove $\max_{x,a}\bar{Q}_t(x,a)$ cannot become larger than $\frac{2}{1-\gamma}$. Actually, if $\max_{x,a}\bar{Q}_t(x,a)$ ever attains $\frac{2}{1-\gamma}$, that is for some $t_0\geq 0$

$$\max_{x,a}\bar{Q}_{t_0}(x,a)=\frac{2}{1-\gamma}, \tag{2.5.4}$$

then for any state-action pair (x_0,a_0) such that $Q_{t_0}(x_0,a_0)=\frac{2}{1-\gamma}$ we have

$$\frac{d\bar{Q}_t}{dt}(x_0,a_0)\bigg|_{t=t_0}\leq\alpha\pi^{g_{\bar{\theta}_{t_0}}}(x_0,a_0)\left[1+2\frac{\gamma}{1-\gamma}-\frac{2}{1-\gamma}\right]=-\alpha\pi^{g_{\bar{\theta}_{t_0}}}(x_0,a_0)\leq 0, \tag{2.5.5}$$

and therefore $\max_{x,a}\bar{Q}_t(x,a)$ can never exceed $\frac{2}{1-\gamma}$. Similarly, we can prove

$$\min_{x,a}\bar{Q}_t(x,a)\geq-\frac{2}{1-\gamma},\quad t\geq 0, \tag{2.5.6}$$

which concludes the proof of the lemma. □

We now develop an ODE comparison principle which will help us to prove the convergence (2.3.5).

Lemma 2.5.2. *Suppose a non-negative function Y_t satisfies*

$$\frac{dY_t}{dt}\leq-\frac{C}{\log^{2n_0}t}Y_t+\frac{1}{t},\quad t\geq t_0, \tag{2.5.7}$$

where C,n_0 are constant and $t_0\geq 0$. Then,

$$Y_t=O\left(\frac{1}{\log^4 t}\right). \tag{2.5.8}$$

Proof. First, we establish a comparison principle with the following ODE:

$$\begin{aligned}\frac{dZ_t}{dt}&=-\frac{C}{\log^{2n_0}t}Z_t+\frac{1}{t}\quad t\geq t_0,\\ Z_{t_0}&=Y_{t_0}.\end{aligned} \tag{2.5.9}$$

Define

$$V_t=Y_t-Z_t.$$

Then, we have $V_{t_0} = 0$ and for any $t \geq t_0$

$$\begin{aligned}\frac{dV_t}{dt} &= \frac{dY_t}{dt} - \frac{dV_t}{dt} \\ &\leq -\frac{C}{\log^{2n_0} t} Y_t + \frac{1}{t} - \left(-\frac{C}{\log^{2n_0} t} Z_t + \frac{1}{t}\right) \\ &= -\frac{C}{\log^{2n_0} t}(Y_t - Z_t) \\ &= -\frac{C}{\log^{2n_0} t} V_t.\end{aligned} \tag{2.5.10}$$

Then, using an integrating factor,

$$\frac{d}{dt}\left[\exp\left\{\int_{t_0}^{t} \frac{C}{\log^{2n_0}\tau} d\tau\right\} V_t\right] = \exp\left\{\int_{t_0}^{t} \frac{C}{\log^{2n_0}\tau} d\tau\right\}\left[\frac{dV_t}{dt} + \frac{C}{\log^{2n_0} t} V_t\right] \leq 0. \tag{2.5.11}$$

Thus we have $V_t \leq \exp\left\{-\int_{t_0}^{t} \frac{C}{\log^{2n_0}\tau} d\tau\right\} V_{t_0} = 0, \quad t \geq t_0$. Therefore,

$$Y_t \leq Z_t \quad t \geq t_0. \tag{2.5.12}$$

Then, if we can establish a convergence rate for Z_t, we have a convergence rate for Y_t.

To solve the ODE (2.5.9), note that

$$\frac{d}{dt}\left[\exp\left\{\int_{t_0}^{t} \frac{C}{\log^{2n_0}\tau} d\tau\right\} Z_t\right] = \exp\left\{\int_{t_0}^{t} \frac{C}{\log^{2n_0}\tau} d\tau\right\}\left[\frac{dZ_t}{dt} + \frac{C}{\log^{2n_0} t} Z_t\right] = \frac{1}{t}\exp\left\{\int_{t_0}^{t} \frac{C}{\log^{2n_0}\tau} d\tau\right\}. \tag{2.5.13}$$

Then,

$$\begin{aligned} Z_t &= \frac{Z_{t_0}}{\exp\left\{\int_{t_0}^{t} \frac{C}{\log^{2n_0}\tau} d\tau\right\}} + \frac{\int_{t_0}^{t} \frac{1}{s}\exp\left\{\int_{t_0}^{s} \frac{C}{\log^{2n_0}\tau} d\tau\right\} ds}{\exp\left\{\int_{t_0}^{t} \frac{C}{\log^{2n_0}\tau} d\tau\right\}} \\ &:= I_t^3 + I_t^4. \end{aligned} \tag{2.5.14}$$

Note that for any integer n and constant $\gamma > 0$,

$$\lim_{t\to\infty} \frac{\log^n t}{t^\gamma} = 0. \tag{2.5.15}$$

Thus, without loss of generality, we can suppose t_0 is large enough such that

$$\log^{2n_0} t \leq t, \quad t \geq t_0. \tag{2.5.16}$$

Then, we can show that

$$I_t^3 \leq \frac{Z_{t_0}}{\exp\left\{\int_{t_0}^{t} \frac{C}{\tau} d\tau\right\}} = \frac{Z_{t_0} t_0^C}{t^C}. \tag{2.5.17}$$

By L'Hospital's Rule, we have

$$\begin{aligned}\lim_{t\to\infty}\log^4 t\cdot I_t^4 &= \lim_{t\to\infty}\frac{\frac{4\log^{2n_0+3}t}{Ct}\int_{t_0}^t\frac{1}{s}\exp\left\{\int_{t_0}^s\frac{C}{\log^{2n_0}\tau}d\tau\right\}ds}{\exp\left\{\int_{t_0}^t\frac{C}{\log^{2n_0}\tau}d\tau\right\}}+\lim_{t\to\infty}\frac{\log^{2n_0+4}t}{Ct}\\ &\overset{(a)}{=}\lim_{t\to\infty}\frac{\int_{t_0}^t\frac{1}{s}\exp\left\{\int_{t_0}^s\frac{C}{\log^{2n_0}\tau}d\tau\right\}ds}{\exp\left\{\int_{t_0}^t\frac{C}{\log^{2n_0}\tau}d\tau\right\}}\\ &=\lim_{t\to\infty}\frac{\frac{1}{t}\exp\left\{\int_{t_0}^t\frac{C}{\log^{2n_0}\tau}d\tau\right\}}{\exp\left\{\int_{t_0}^t\frac{C}{\log^{2n_0}\tau}d\tau\right\}\frac{C}{\log^{2n_0}t}}\\ &=\lim_{t\to\infty}\frac{\log^{2n_0}t}{Ct}\\ &=0,\end{aligned}\tag{2.5.18}$$

where step (a) is by (2.5.15). Therefore, we can let t_0 be large enough such that

$$I_t^4\le\frac{1}{\log^4 t},\quad\forall t\ge t_0.\tag{2.5.19}$$

Combining our results, we have

$$Y_t\le Z_t\le\frac{Y_{t_0}t_0^C}{t^C}+\frac{1}{\log^4 t},\quad t\ge t_0,\tag{2.5.20}$$

which together with (2.5.15) proves (2.5.8). □

Using Lemma 2.5.2, now we prove the critic convergence (2.3.5).

Proof of (2.3.5)*:* Combining (2.3.4) and (2.5.1),

$$\frac{d\phi_t}{dt}(x,a)=-\alpha\pi^{g_{\bar\theta_t}}(x,a)\phi_t(x,a)+\alpha\gamma\pi^{g_{\bar\theta_t}}(x,a)\sum_{z,a''}\phi_t(z,a'')g_{\bar\theta_t}(z,a'')p(z|x,a)+\frac{dV^{g_{\bar\theta_t}}}{dt}(x,a).\tag{2.5.21}$$

Let $\odot$ denote element-wise multiplication. Then,

$$\frac{d\phi_t}{dt}=-\alpha\pi^{g_{\bar\theta_t}}\odot\phi_t+\alpha\gamma\pi^{g_{\bar\theta_t}}\odot\Gamma_t+\frac{dV^{g_{\bar\theta_t}}}{dt},\tag{2.5.22}$$

where $\Gamma_t(x',a')=\sum_{z,a''}\phi_t(z,a'')g_{\bar\theta_t}(z,a'')p(z|x',a')$.

Define the process

$$Y_t=\frac{1}{2}\phi_t^\top\phi_t.\tag{2.5.23}$$

Differentiating yields

$$\frac{dY_t}{dt}=\phi_t^\top\frac{d\phi_t}{dt}=-\alpha\phi_t^\top\pi^{g_{\bar\theta_t}}\odot\phi_t+\alpha\gamma\phi_t^\top\pi^{g_{\bar\theta_t}}\odot\Gamma_t+\phi_t^\top\frac{dV^{g_{\bar\theta_t}}}{dt}.\tag{2.5.24}$$

The second term on the last line of (2.5.24) becomes:

$$
\begin{aligned}
&\left|\phi_t^\top \pi^{g_{\bar\theta_t}} \odot \Gamma_t\right| \\
=&\left|\sum_{x',a'} \phi_t(x',a')\pi^{g_{\bar\theta_t}}(x',a') \sum_{z,a''} \phi_t(z,a'')g_{\bar\theta_t}(z,a'')p(z|x',a')\right| \\
=&\left|\sum_{x',a'}\sum_{z,a''} \phi_t(z,a'')\phi_t(x',a')g_{\bar\theta_t}(z,a'')p(z|x',a')\pi^{g_{\bar\theta_t}}(x',a')\right| \\
\le&\sum_{x',a'}\sum_{z,a''} \left|\phi_t(z,a'')\phi_t(x',a')\right| g_{\bar\theta_t}(z,a'')p(z|x',a')\pi^{g_{\bar\theta_t}}(x',a') \\
\le&\frac{1}{2}\sum_{x',a'}\sum_{z,a''} \Big(\phi_t(z,a'')^2+\phi_t(x',a')^2\Big) g_{\bar\theta_t}(z,a'')p(z|x',a')\pi^{g_{\bar\theta_t}}(x',a') \\
=&\frac{1}{2}\sum_{z,a''} \phi_t(z,a'')^2 \sum_{x',a'} g_{\bar\theta_t}(z,a'')p(z|x',a')\pi^{g_{\bar\theta_t}}(x',a') \\
+&\frac{1}{2}\sum_{x',a'} \phi_t(x',a')^2\pi^{g_{\bar\theta_t}}(x',a') \sum_{z,a''} g_{\bar\theta_t}(z,a'')p(z|x',a') \\
=&\frac{1}{2}\sum_{z,a''} \phi_t(z,a'')^2\pi^{g_{\bar\theta_t}}(z,a'') + \frac{1}{2}\sum_{x',a'} \phi_t(x',a')^2\pi^{g_{\bar\theta_t}}(x',a') \\
=&\sum_{x',a'} \phi_t(x',a')^2\pi^{g_{\bar\theta_t}}(x',a').
\end{aligned}
$$

where we have used Young's inequality, the fact that $\sum_{z,a''} g_{\bar\theta_t}(z,a'')p(z|x',a') = 1$ for each (x',a'), and $\sum_{x',a'} g_{\bar\theta_t}(z,a'')p(z|x',a')\pi^{g_{\bar\theta_t}}(x',a') = \pi^{g_{\bar\theta_t}}(z,a'')$. Therefore,

$$
\frac{dY_t}{dt} \le -\alpha(1-\gamma)\pi^{g_{\bar\theta_t}}\cdot\phi_t^2 + \phi_t^\top \frac{dV^{g_{\bar\theta_t}}}{dt}, \tag{2.5.25}
$$

where ϕ_t^2 is an element-wise square.

By the limit ODEs in (2.3.4) and the uniform boundedness in Lemma 2.5.1, we have for any (x,a)

$$
\left|\frac{d\bar\theta_t}{dt}(x,a)\right| = \left|\zeta_t \sigma_\mu^{f_{\bar\theta_t}}(x,a)\left[\bar Q_t(x,a) - \sum_{a'} \bar Q_t(x,a')f_{\bar\theta_t}(x,a')\right]\right| \le C\zeta_t \tag{2.5.26}
$$

For any state x_0, define $\partial_{x,a}V^{f_\theta}(x_0) = \frac{\partial V^{f_\theta}(x_0)}{\partial\theta(x,a)}$. Then, for the exploration policy (2.2.9), by the policy gradient theorem (2.4.12) we have

$$
\begin{aligned}
|\partial_{x,a}V^{g_{\bar\theta_t}}(x_0)| &= \left|\sum_{x',a'} \sigma_{x_0}^{g_{\bar\theta_t}}(x',a')V^{g_{\bar\theta_t}}(x',a')\partial_{x,a}\log g_{\bar\theta_t}(x',a')\right| \\
&\le C\sum_{x',a'} \left|\partial_{x,a}\log g_{\bar\theta_t}(x',a')\right| \\
&= C(1-\eta_t)\sum_{x',a'} \frac{f_{\bar\theta_t}(x',a')}{g_{\bar\theta_t}(x',a')}\left|\partial_{x,a}\log f_{\bar\theta_t}(x',a')\right| \\
&\overset{(a)}{\le} C,
\end{aligned} \tag{2.5.27}
$$

where step (a) is by

$$\frac{f_{\bar{\theta}_t}(x',a')}{g_{\bar{\theta}_t}(x',a')} = \frac{f_{\bar{\theta}_t}(x',a')}{\frac{\eta_t}{d_A} + (1-\eta_t)\cdot f_{\bar{\theta}_t}(x',a')} \leq C \tag{2.5.28}$$

and

$$\left|\partial_{x,a} \log f_{\bar{\theta}_t}(x',a')\right| = \left|\mathbb{1}_{\{x'=x\}}\left[\mathbb{1}_{\{a'=a\}} - f_{\bar{\theta}_t}(x',a)\right]\right| \leq 2. \tag{2.5.29}$$

The relationship between the value functions

$$V^{f_{\bar{\theta}_t}}(x_0,a_0) = r(x_0,a_0) + \gamma \sum_{x'} V^{f_{\bar{\theta}_t}}(x')p(x'|x_0,a_0), \quad \forall (x_0,a_0), \tag{2.5.30}$$

can be combined with (2.5.27) to derive

$$\|\nabla_\theta V^{g_{\bar{\theta}_t}}(x,a)\| \leq C, \quad \forall (x,a). \tag{2.5.31}$$

Combining (2.5.26) and (2.5.31),

$$\left|\frac{dV^{g_{\bar{\theta}_t}}}{dt}(x,a)\right| = \left|\nabla_\theta V^{g_{\bar{\theta}_t}}(x,a)\cdot\frac{d\bar{\theta}_t}{dt}\right| \leq \|\nabla_\theta V^{g_{\bar{\theta}_t}}(x,a)\|\cdot\left\|\frac{d\bar{\theta}_t}{dt}\right\| \leq C\zeta_t, \tag{2.5.32}$$

where $C>0$ is a constant independent of T.

Combining (2.5.25), (2.5.32) and (2.4.33), we have

$$\begin{aligned}
\frac{dY_t}{dt} &\leq -\alpha(1-\gamma)\min_{x,a}\{\pi^{g_{\bar{\theta}_t}}(x,a)\}Y_t + C\phi_t^\top\zeta_t \\
&\leq -\alpha C\eta_t^{n_0}(1-\gamma)Y_t + C\phi_t^\top\zeta_t \\
&\leq -C\eta_t^{n_0}Y_t + \frac{\eta_t^{n_0}}{\eta_t^{n_0}}\|\phi_t\|C\zeta_t \\
&\leq -C\eta_t^{n_0}Y_t + \|\phi_t\|^2\eta_t^{2n_0} + \frac{C\zeta_t^2}{\eta_t^{2n_0}} \\
&= -\eta_t^{n_0}(C-2\eta_t^{n_0})Y_t + \frac{C\zeta_t}{\eta_t^{2n_0}}\zeta_t.
\end{aligned} \tag{2.5.33}$$

Since $\frac{\zeta_t}{\eta_t^{2n_0}} \to 0$ as $t\to\infty$, there exists $t_0 \geq 2$ such that $\forall t \geq t_0$

$$\frac{dY_t}{dt} \leq -C\eta_t^{n_0}Y_t + \zeta_t \leq -\frac{C}{\log^{2n_0} t}Y_t + \frac{1}{t}, \tag{2.5.34}$$

where the C is a constant independent with t. Then, by Lemma 2.5.2, there exists $t_1 \geq t_0$ such that

$$Y_t = O\left(\frac{1}{\log^4 t}\right) = O\left(\eta_t^2\right). \tag{2.5.35}$$

By the policy gradient theorem (2.4.12), we have

$$\frac{\partial V^f(x_0)}{\partial_{f(x,a)}} = V^f(x,a)\sigma_{x_0}^f(x). \tag{2.5.36}$$

Thus, by the relationship (2.5.30),

$$\frac{\partial V^{f_{\bar{\theta}_t}}(x_0,a_0)}{\partial_{f(x,a)}} = \gamma\sum_{x'} V^{f_{\bar{\theta}_t}}(x,a)\sigma_{x'}^{f_{\bar{\theta}_t}}(x)p(x'|x_0,a_0) \leq C. \tag{2.5.37}$$

Then, for any $(x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}$, there exists $\widetilde{t}\in[0,1]$ such that

$$\left|V^{g_{\bar{\theta}_t}}(x,a) - V^{f_{\bar{\theta}_t}}(x,a)\right| = \left|\nabla_f V^{\widetilde{t}f_{\bar{\theta}_t}+(1-\widetilde{t})g_{\bar{\theta}_t}}(x,a)\cdot\left[g_{\bar{\theta}_t} - f_{\bar{\theta}_t}\right]\right| \leq C\eta_t, \tag{2.5.38}$$

Finally, combining (2.5.35) and (2.5.38), we obtain (2.3.5).

□

2.5.2 Actor convergence

2.5.2.1 Convergence to stationary point

In order to prove global convergence, we first show that the actor converges to a stationary point. We introduce the following notation:

$$\begin{aligned}\widehat{\nabla}_\theta J(f_{\bar{\theta}_t}) &:= \sum_{x,a} \sigma_\mu^{f_{\bar{\theta}_t}}(x,a)\bar{Q}_t(x,a)\nabla_\theta \log f_{\bar{\theta}_t}(x,a),\\ \widehat{\partial}_{x,a} J(f_{\bar{\theta}_t}) &:= \sum_{x,a} \sigma_\mu^{f_{\bar{\theta}_t}}(x,a)\bar{Q}_t(x,a)\partial_{x,a} \log f_{\bar{\theta}_t}(x,a).\end{aligned} \tag{2.5.39}$$

Then, the limit ode for θ in (2.3.4) can be written as

$$\frac{d\bar{\theta}_t}{dt} = \zeta_t \widehat{\nabla}_\theta J(f_{\bar{\theta}_t}). \tag{2.5.40}$$

By direct calculations,

$$\begin{aligned}\nabla_\theta \log f_\theta(x,a) =& \nabla_\theta \left[\theta(x,a) - \log \sum_{a'} e^{\theta(x,a')}\right]\\ =& \nabla_\theta\theta(x,a) - \frac{\sum_{a'} e^{\theta(x,a')}\nabla_\theta\theta(x,a')}{\sum_{a'} e^{\theta(x,a')}}\\ =& \nabla_\theta\theta(x,a) - \sum_{a'} f_\theta(x,a')\nabla_\theta\theta(x,a')\\ =& \nabla_\theta\theta(x,a) - \mathbb{E}_{a'\sim f_\theta(x,\cdot)}[\nabla_\theta\theta(x,a')]\\ =& e_{x,a} - \sum_{a'} e_{x,a'} f_\theta(x,a'),\end{aligned} \tag{2.5.41}$$

where $e_{x,a}$ is the unit vector where only the x,a element is 1 and all other elements are 0. Then, the difference is

$$\begin{aligned}\nabla_\theta J(f_{\bar{\theta}_t}) - \widehat{\nabla}_\theta J(f_{\bar{\theta}_t}) &= \sum_{x,a} \sigma^{f_{\bar{\theta}_t}}(x,a)\left(\bar{Q}_t(x,a) - V^{f_{\bar{\theta}_t}}(x,a)\right)\nabla_\theta \log f_{\bar{\theta}_t}(x,a),\\ &= \sum_{x,a} \sigma^{f_{\bar{\theta}_t}}(x,a)\left(\bar{Q}_t(x,a) - V^{f_{\bar{\theta}_t}}(x,a)\right)\left(e_{x,a} - \sum_{a'} e_{x,a'} f_{\bar{\theta}_t}(x,a')\right),\end{aligned} \tag{2.5.42}$$

which together with (2.3.5) derives

$$\|\nabla_\theta J(\bar{\theta}_t) - \widehat{\nabla}_\theta J(\bar{\theta}_t)\|_2 \le C\|\bar{Q}_t - V^{f_{\bar{\theta}_t}}\|_2 \le C\eta_t. \tag{2.5.43}$$

Thus we re-write the gradient flow (2.5.40) as

$$\frac{d\bar{\theta}_t}{dt} = \zeta_t \nabla_\theta J(f_{\bar{\theta}_t}) + \zeta_t \sum_{x,a} \sigma^{f_{\bar{\theta}_t}}(x,a)\left[\left(\bar{Q}_t(x,a) - V^{f_{\bar{\theta}_t}}(x,a)\right)\cdot \nabla_\theta \log f_{\bar{\theta}_t}(x,a)\right]. \tag{2.5.44}$$

Now we can adapt the proof in [14] to show the gradient flow converges to a stationary point. We first provide a useful lemma.

Lemma 2.5.3. *Let Y_t, W_t and Z_t be three functions such that W_t is nonnegative. Asuume that*

$$\frac{dY_t}{dt} \ge W_t + Z_t, \quad t \ge 0 \tag{2.5.45}$$

and that $\int_0^\infty Z_t dt$ *converges. Then, either* $Y_t \to \infty$ *or else* Y_t *converges to a finite value and* $\int_0^\infty W_t dt < \infty$.

Proof. For any $\bar{t} > 0$. By integrating the relationship $\frac{dY_t}{dt} \geq Z_t$ from $\bar{t}$ to $t \geq \bar{t}$ and taking the limit inferior as $t \to \infty$, we obtain

$$\liminf_{t\to\infty} Y_t \geq Y_{\bar{t}} + \int_{\bar{t}}^\infty Z_t dt > -\infty. \tag{2.5.46}$$

By taking the limit superior of the right-hand side as $\bar{t} \to \infty$ and using the fact $\lim_{\bar{t}\to\infty} \int_{\bar{t}}^\infty Z_t dt = 0$, we obtain

$$\liminf_{t\to\infty} Y_t \geq \limsup_{\bar{t}\to\infty} Y_t > -\infty. \tag{2.5.47}$$

This proves that either $Y_t \to \infty$ or Y_t converges to a finite value. If Y_t converges to a finite value, we can integrate the relationship (2.5.45) to show that

$$\int_0^t W_s ds \leq Y_t - Y_0 - \int_0^t Z_s ds, \tag{2.5.48}$$

which implies that $\int_0^\infty W_s ds \leq \lim_{t\to\infty} Y_t - Y_0 - \int_0^\infty Z_s ds < \infty$. □

Next we can prove convergence to the stationary point under the learning rate (2.3.1).

Theorem 2.5.4. *Suppose the learning rate* ζ_t *satisfies* (2.3.1). *Then, for the gradient flow* (2.5.40), *we have that* $J(\bar{\theta}_t)$ *converges to a finite value and*

$$\lim_{t\to+\infty} \nabla_\theta J(f_{\bar{\theta}_t}) = 0. \tag{2.5.49}$$

Proof. First we note that by the proof of Lemma 7 in [104], we know that the eigenvalues of the Hessian matrix of $J(f_\theta)$ are smaller than $L := \frac{8}{(1-\gamma)^3}$ and thus $\nabla_\theta J(f_\theta)$ is L-Lipschitz continuous with respect to θ.

Then, by the gradient flow (2.5.40), (2.5.43), and chain rule, we can show that

$$\begin{aligned}
\frac{dJ(f_{\bar{\theta}_t})}{dt} &= \zeta_t \nabla_\theta J(f_{\bar{\theta}_t}) \widehat{\nabla}_\theta J(f_{\bar{\theta}_t}) \\
&= \zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 + \zeta_t \nabla_\theta J(f_{\bar{\theta}_t}) \left(\widehat{\nabla}_\theta J(f_{\bar{\theta}_t}) - \nabla_\theta J(f_{\bar{\theta}_t})\right) \\
&\geq \zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 - C\zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\| \cdot \|Q_t(\cdot,\cdot) - V^{f_{\bar{\theta}_t}}(\cdot,\cdot)\|_2 \\
&\overset{(a)}{\geq} \zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 - C\zeta_t \eta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\| \\
&\overset{(b)}{\geq} (\zeta_t - C\zeta_t\eta_t)\|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 - C\zeta_t\eta_t \\
&\overset{(c)}{\geq} C\zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 - C\zeta_t\eta_t.
\end{aligned} \tag{2.5.50}$$

where the step (a) follows (2.5.43). Step (b) is by using the relationship $\|\nabla_\theta J(f_{\bar{\theta}_t})\| \leq 1 + \|\nabla_\theta J(f_{\bar{\theta}_t})\|_2^2$ and step (c) is because $\eta_t \to 0$ and C_1, C_2 are some sufficiently large enough constants. Then, by Lemma 2.5.3 and the assumption in (2.3.1), we can show that either $J(f_{\bar{\theta}_t}) \to \infty$ or $J(f_{\bar{\theta}_t})$ converges to a finite value and

$$\int_0^{+\infty} \zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 dt < \infty. \tag{2.5.51}$$

Note that $J(f_\theta) = \mathbb{E}_{f_\theta}\left[\sum_{k=0}^{+\infty}\gamma^k r(x_k, a_k)\right]$. Therefore, the objective function J is bounded by Assumption 2.3.1 and thus we know $J(\bar{\theta}_t)$ converges to a finite value and (2.5.51) is valid.

If there existed an $\epsilon_0 > 0$ and $\bar{t} > 0$ such that $\|\nabla_\theta J(f_{\bar{\theta}_t})\| \geq \epsilon_0$ for all $t \geq \bar{t}$, we would have

$$\int_{\bar{t}}^{+\infty} \zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 dt \geq \epsilon_0^2 \int_{\bar{t}}^{+\infty} \zeta_t dt = \infty, \tag{2.5.52}$$

which contradicts (2.5.51). Therefore, $\liminf_{t\to\infty} \|\nabla_\theta J(f_{\bar{\theta}_t})\| = 0$. To show that $\lim_{t\to\infty} \|\nabla_\theta J(f_{\bar{\theta}_t})\| = 0$, assume the contrary; that is $\limsup_{t\to\infty} \|\nabla_\theta J(f_{\bar{\theta}_t})\| > 0$. Then we can find a constant $\epsilon_1 > 0$ and two increasing sequences $\{a_n\}_{n\geq 1}, \{b_n\}_{n\geq 1}$ such that

$$\begin{aligned} &a_1 < b_1 < a_2 < b_2 < a_3 < b_3 < \cdots, \\ &\|\nabla_\theta J(f_{\bar{\theta}_{a_n}})\| < \frac{\epsilon_1}{2}, \quad \|\nabla_\theta J(f_{\bar{\theta}_{b_n}})\| > \epsilon_1. \end{aligned} \tag{2.5.53}$$

Define the following cycle of stopping times:

$$\begin{aligned} t_n &:= \sup\{s | s \in (a_n, b_n), \|\nabla_\theta J(f_{\bar{\theta}_s})\| < \frac{\epsilon_1}{2}\}, \\ i(t_n) &:= \inf\{s | s \in (t_n, b_n), \|\nabla_\theta J(f_{\bar{\theta}_s})\| > \epsilon_1\}. \end{aligned} \tag{2.5.54}$$

Note that $\|\nabla_\theta J(f_{\bar{\theta}_t})\|$ is continuous against t, thus we have

$$\begin{aligned} &a_n \leq t_n < i(t_n) \leq b_n \\ &\|\nabla_\theta J(f_{\bar{\theta}_{t_n}})\| = \frac{\epsilon_1}{2}, \quad \|\nabla_\theta J(f_{\bar{\theta}_{i(t_n)}})\| = \epsilon_1 \\ &\frac{\epsilon_1}{2} \leq \|\nabla_\theta J(f_{\bar{\theta}_s})\| \leq \epsilon_1, \quad s \in (t_n, i(t_n)). \end{aligned} \tag{2.5.55}$$

Then, by the L-Lipschitz property of the gradient, we have for any t_n

$$\begin{aligned} \frac{\epsilon_1}{2} &= \|\nabla_\theta J(f_{\bar{\theta}_{i(t_n)}})\| - \|\nabla_\theta J(f_{\bar{\theta}_{t_n}})\| \\ &\leq \|\nabla_\theta J(f_{\bar{\theta}_{i(t_n)}}) - \nabla_\theta J(f_{\bar{\theta}_{t_n}})\| \\ &\leq L\|\bar{\theta}_{i(t_n)} - \bar{\theta}_{t_n}\| \\ &\leq L\int_{t_n}^{i(t_n)} \zeta_s \|\nabla_\theta J(f_{\bar{\theta}_s})\| ds + L\int_{t_n}^{i(t_n)} \zeta_s \|\widehat{\nabla}_\theta J(f_{\bar{\theta}_s}) - \nabla_\theta J(f_{\bar{\theta}_s})\| ds \\ &\leq L\epsilon_1 \int_{t_n}^{i(t_n)} \zeta_s ds + CL\int_{t_n}^{i(t_n)} \zeta_s \eta_s ds. \end{aligned} \tag{2.5.56}$$

From this and by (2.3.2) it follows that

$$\frac{1}{2L} \leq \liminf_{n\to\infty} \int_{t_n}^{i(t_n)} \zeta_s ds. \tag{2.5.57}$$

Using (2.5.50) and (2.5.55), we see that

$$J(f_{\bar{\theta}_{i(t_n)}}) - J(f_{\bar{\theta}_{t_n}}) \geq C_1(\frac{\epsilon_1}{2})^2 \int_{t_n}^{i(t_n)} \zeta_s ds - C_2 \int_{t_n}^{i(t_n)} \zeta_s \eta_s ds. \tag{2.5.58}$$

Due to the convergence of $J(f_{\theta_{t_n}})$ and the assumption of the learning rate, this implies that

$$\lim_{n\to\infty} \int_{t_n}^{i(t_n)} \zeta_s ds = 0, \tag{2.5.59}$$

which contradicts (2.5.57) and thus the convergence to the stationary point is proven. □

2.5.2.2 Global convergence

We now prove the global convergence rate (2.3.6) for the actor dynamic using the following steps:

- Derive non-uniform Lojasiewicz inequalities.
- Adapt the method in [1] to obtain the global convergence.
- Set up the uniform Lojasiewicz inequalities and the ODE for actor convergence.
- Analyse the ODE by a comparison lemma to get the convergence rate.

Since the objective function $J(f_\theta)$ is non-concave, the convergence to a stationary point in Theorem 2.5.4 does not guarantee global convergence to the optimal policy. As a first step, we establish the following non-uniform Lojasiewicz inequalities that show that the gradient of the objective function for any parameter value dominates the sub-optimality of the parameter. Actually, (2.5.60) is used for the case that the best action at any state x is unique, while (2.5.63) is for the non-unique optimal action case.

Lemma 2.5.5 (Non-uniform Lojasiewicz Bound)**.** *Choose any deterministic optimal policy f^*.*

- *Suppose for any state $\forall x \in \boldsymbol{\mathcal{X}}$, there exists unique optimal action, then we have*

$$\|\nabla_\theta J(f_\theta)\| \geq \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}} \cdot \left\| \frac{\nu_\mu^{f^*}}{\nu_\mu^{f_\theta}} \right\|_\infty^{-1} \cdot \min_x f_\theta\left(x, a^*(x)\right) \cdot \left[J(f^*) - J(f_\theta)\right] \tag{2.5.60}$$

  where $a^*(x) = \arg\max_a V^{f^*}(x,a), \forall x \in \boldsymbol{\mathcal{X}}$.

- *When under some state $x \in \boldsymbol{\mathcal{X}}$, there is an "optimal action set":*

$$\boldsymbol{\mathcal{A}}^*(x) := \left\{ a^*(x) \in \boldsymbol{\mathcal{A}} : V^{f^*}(x, a^*(x)) = \max_a V^{f^*}(x,a) \right\}, \tag{2.5.61}$$

  i.e. all actions $a^*(x) \in \boldsymbol{\mathcal{A}}^*(x)$ are the greedy action w.r.t. the optimal state-action value functin V^{f^*}. Given any policy f_θ, construct the following optimal policy

$$f_\theta^*(x,a) = \begin{cases} \frac{f_\theta(x,a)}{\sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a')}, & \text{if } a \in \boldsymbol{\mathcal{A}}^*(x), \\ 0, & \text{otherwise} \end{cases} \tag{2.5.62}$$

  It is obvious that f_θ^* is an optimal policy, since for all $x \in \mathcal{X}$,

$$\sum_{a \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta^*(x,a) = \frac{\sum_{a \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a)}{\sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta\left(x,a'\right)} = 1.$$

  Now we have

$$\|\nabla_\theta J(f_\theta)\| \geq \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \cdot \left\| \frac{\nu_\mu^{f_\theta^*}}{\nu_\mu^{f_\theta}} \right\|_\infty^{-1} \cdot \left[\min_x \sum_{a^*(x) \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x, a^*(x)) \right] \cdot \left[J(f^*) - J(f_\theta)\right]. \tag{2.5.63}$$

Remark 2.5.6. As the proof of Lemma 2.5.5 is similar as in [104], we move the detailed proof into Appendix A.4.

Lemma 2.5.5 is not sufficient to prove a global convergence rate (or even global convergence). For example, the term $\min_{x\in\boldsymbol{\mathcal{X}}} f_{\bar{\theta}_t}(x, a^*(x))$ in (2.5.60) could converge to zero as $t \to \infty$. Thus to obtain (2.3.6), we follow the steps.

(i) Prove the global convergence

$$J(f^*) - J(f_{\bar{\theta}_t}) \to 0, \quad t \to \infty, \tag{2.5.64}$$

This global convergence can be proven by adapting the method in [1] to the setting in this chapter.

(ii) Due to the convergence (2.5.64), if for each state x the best action $a^*(x)$ is unique, we will have

$$\lim_{t\to\infty} f_{\bar{\theta}_t}(x, a^*(x)) = 1, \quad \forall x \in \boldsymbol{\mathcal{X}} \tag{2.5.65}$$

and thus

$$\inf_{x\in\mathcal{X}, t\geq 0} f_{\bar{\theta}_t}(x, a^*(x)) > 0. \tag{2.5.66}$$

If for some state x, the best action is not unique, then the convergence (2.5.64) implies that

$$\lim_{t\to\infty} \sum_{a^*(x)\in\boldsymbol{\mathcal{A}}^*(x)} f_{\bar{\theta}_t}(x, a^*(x)) = 1, \quad \forall x \in \boldsymbol{\mathcal{X}} \tag{2.5.67}$$

and thus

$$\inf_{x\in\mathcal{X}, t\geq 0} \sum_{a^*(x)\in\boldsymbol{\mathcal{A}}^*(x)} f_{\bar{\theta}_t}(x, a^*(x)) > 0. \tag{2.5.68}$$

(iii) The lower bound for $\min_x f_{\bar{\theta}_t}(x, a^*(x))$, $\min_x \sum_{a^*(x)\in\boldsymbol{\mathcal{A}}^*(x)} f_{\bar{\theta}_t}(x, a^*(x))$ and (2.5.60), (2.5.63) can be used to derive the uniform Lojasiewicz inequality for MDP with unique or non-unique optimal action. By analysing the gradient flow, we can prove the convergence rate (2.3.6).

Now we adapt the method in [1] to obtain the global convergence (2.5.64). For the gradient flow

$$\frac{d\bar{\theta}_t}{dt} = \zeta_t \widehat{\nabla}_\theta J(f_{\bar{\theta}_t}), \tag{2.5.69}$$

where $\widehat{\nabla}_\theta J(\bar{\theta}_t) := \sum_{x,a} \sigma_\mu^{f_{\bar{\theta}_t}}(x, a)\bar{Q}_t(x, a)\nabla_\theta \log f_{\bar{\theta}_t}(x, a)$, with the similar calculations in (2.4.15), it can be shown that

$$\begin{aligned}
\frac{d}{dt}\bar{\theta}_t(x, a) &= \zeta_t \widehat{\partial}_{x,a} J(f_{\bar{\theta}_t}) \\
&= \zeta_t \sum_{x',a'} \nu_\mu^{f_\theta}(x') f_{\bar{\theta}_t}(x', a') \mathbb{1}_{\{x'=x\}} \left[\mathbb{1}_{\{a'=a\}} - f_{\bar{\theta}_t}(x', a)\right] \bar{Q}_t(x', a') \\
&= \zeta_t \sum_{a'} \nu_\mu^{f_{\bar{\theta}_t}}(x) f_{\bar{\theta}_t}(x, a') \left[\mathbb{1}_{\{a'=a\}} - f_{\bar{\theta}_t}(x, a)\right] \bar{Q}_t(x, a') \\
&= \zeta_t \nu_\mu^{f_{\bar{\theta}_t}}(x) f_{\bar{\theta}_t}(x, a)\bar{Q}_t(x, a) - \zeta_t \nu_\mu^{f_{\bar{\theta}_t}}(x) f_{\bar{\theta}_t}(x, a) \left[\sum_{a'} f_{\bar{\theta}_t}(x, a')\bar{Q}_t(x, a')\right] \\
&= \zeta_t \sigma_\mu^{f_{\bar{\theta}_t}}(x, a) \left[\bar{Q}_t(x, a) - \sum_{a'} \bar{Q}_t(x, a') f_{\bar{\theta}_t}(x, a')\right].
\end{aligned} \tag{2.5.70}$$

The following lemma is important in our proof.

Lemma 2.5.7 (The performance difference lemma [74]). *For all policies f, f' and state x_0,*

$$V^{f}(x_0) - V^{f'}(x_0) = \sum_{x,a} \sigma^{f}_{x_0}(x,a) A^{f'}(x,a), \tag{2.5.71}$$

where $\sigma^{f}_{x_0}$ is the visiting measure for the MDP $\boldsymbol{\mathcal{M}}$ with initial distribution δ_{x_0} and policy f.

We first prove the following convergence lemma for value functions $V^{f_{\bar{\theta}_t}}(x)$ and $V^{f_{\bar{\theta}_t}}(x,a)$.

Lemma 2.5.8. *There exists value $V^{\infty}(x)$ and $V^{\infty}(x,a)$ for every state x and action a such that*

$$\lim_{t\to\infty} V^{f_{\bar{\theta}_t}}(x) = V^{\infty}(x), \quad \lim_{t\to\infty} V^{f_{\bar{\theta}_t}}(x,a) = V^{\infty}(x,a).$$

Then, by the critic convergence (2.3.5), we immediately have when $t \to \infty$

$$\begin{aligned} &\bar{Q}_t(x,a) \to V^{\infty}(x,a) \\ &\bar{Q}_t(x) := \sum_{a} \bar{Q}_t(x,a) f_{\bar{\theta}_t}(x,a) \to V^{\infty}(x). \end{aligned} \tag{2.5.72}$$

Define

$$\Delta = \min_{\{x,a | A^{\infty}(x,a)\neq 0\}} |A^{\infty}(x,a)|, \tag{2.5.73}$$

where $A^{\infty}(x,a) = V^{\infty}(x,a) - V^{\infty}(x)$. Then there exists a T_0 such that $\forall t > T_0, (x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$, we have

$$V^{\infty}(x,a) - \frac{\Delta}{4} \leq Q_t(x,a) \leq V^{\infty}(x,a) + \frac{\Delta}{4}. \tag{2.5.74}$$

Remark 2.5.9. Here we can suppose that $\Delta > 0$ because if $\Delta = 0$, then we have for any states and actions $A^{\infty}(x,a) = 0$. By Lemma 2.5.7,

$$\begin{aligned} &\lim_{t\to\infty} [J(f^*) - J(f_{\bar{\theta}_t})] \\ =& \lim_{t\to\infty} \sum_{x_0} \mu(x_0) \left[V^{f^*}(x_0) - V^{f_{\bar{\theta}_t}}(x_0) \right] \\ =& \lim_{t\to\infty} \sum_{x_0} \mu(x_0) \left[\sum_{x,a} \sigma^{f^*}_{x_0}(x,a) \left[V^{f_{\bar{\theta}_t}}(x,a) - V^{f_{\bar{\theta}_t}}(x) \right] \right] \\ =& \lim_{t\to\infty} \sum_{x,a} \sigma^{f^*}_{\mu}(x,a) A^{f_{\bar{\theta}_t}}(x,a) \\ =& 0, \end{aligned} \tag{2.5.75}$$

which immediately concludes the global convergence.

Proof. For any fixed state x_0, treat the state value $V^{f_\theta}(x_0)$ as the objective function for an MDP whose initial distribution is δ_{x_0} and, by the policy gradient theorem (2.4.12), we have

$$\nabla_\theta V^{f_{\bar{\theta}_t}}(x_0) = \sum_{x,a} \sigma^{f_{\bar{\theta}_t}}_{x_0}(x,a) V^{f_{\bar{\theta}_t}}(x,a) \nabla_\theta \log f_{\bar{\theta}_t}(x,a), \tag{2.5.76}$$

where $\sigma^{f_{\bar{\theta}_t}}_{x_0}(x,a)$ denotes the visiting measure of the MDP starting from x_0 under the policy $f_{\bar{\theta}_t}$. Thus, using the same calculations as in (2.4.14), we have

$$\frac{\partial}{\partial_{\theta(x,a)}} V^{f_{\bar{\theta}_t}}(x_0) = \sigma^{f_{\bar{\theta}_t}}_{x_0}(x,a) A^{f_{\bar{\theta}_t}}(x,a) \tag{2.5.77}$$

Let $\beta_t(x,a) = \bar{Q}_t(x,a) - V^{f_{\bar{\theta}_t}}(x,a)$ denote the critic error. Due to (2.3.5), we know that for any state-action pair (x,a), $|\beta_t(x,a)| \leq C\eta_t$. Combining (2.5.70) with (2.5.77) and using the chain rule, we have

$$\begin{aligned}
\frac{d}{dt}V^{f_{\bar{\theta}_t}}(x_0) &= \nabla_\theta V^{f_{\bar{\theta}_t}}(x_0) \cdot \frac{d}{dt}\bar{\theta}_t \\
&= \sum_{x,a} \frac{\partial}{\partial_{\theta(x,a)}} V^{f_{\bar{\theta}_t}}(x_0) \frac{d}{dt}\bar{\theta}_t(x,a) \\
&= \zeta_t \sum_{x,a} \sigma_{x_0}^{f_{\bar{\theta}_t}}(x,a) A^{f_{\bar{\theta}_t}}(x,a) \sigma_\mu^{f_{\bar{\theta}_t}}(x,a) \left[\bar{Q}_t(x,a) - \sum_{a'} \bar{Q}_t(x,a') f_{\bar{\theta}_t}(x,a') \right] \\
&= \zeta_t \sum_{x,a} \sigma_{x_0}^{f_{\bar{\theta}_t}}(x,a) A^{f_{\bar{\theta}_t}}(x,a) \sigma_\mu^{f_{\bar{\theta}_t}}(x,a) \left[\beta_t(x,a) - \sum_{a'} \beta_t(x,a') f_{\bar{\theta}_t}(x,a') + A^{f_{\bar{\theta}_t}}(x,a) \right] \\
&\geq \zeta_t \sum_{x,a} \sigma_{x_0}^{f_{\bar{\theta}_t}}(x,a) \sigma_\mu^{f_{\bar{\theta}_t}}(x,a) (A^{f_{\bar{\theta}_t}}(x,a))^2 - C\zeta_t\eta_t,
\end{aligned} \tag{2.5.78}$$

where the last inequality follows from (2.3.5). Thus, by Lemma 2.5.3 and the boundedness of the value functions, we obtain the convergence for the state value function. Then, due to

$$V^{f_{\bar{\theta}_t}}(x,a) = r(x,a) + \gamma \sum_{x'} V^{f_{\bar{\theta}_t}}(x') p(x'|x,a), \tag{2.5.79}$$

the convergence for the state action value function is concluded. The convergence for Q_t is immediately follows from the critic convergence (2.5.33). Combining the convergence for value functions, $\Delta > 0$, and the finiteness of the action space, we obtain (2.5.74). □

Next, partition the action space $\mathcal{A}$ into three sets according to the value $V^\infty(x)$ and $V^\infty(x,a)$,

$$\begin{aligned}
I_0^x &:= \{a | V^\infty(x,a) = V^\infty(x)\} \\
I_+^x &:= \{a | V^\infty(x,a) > V^\infty(x)\} \\
I_-^x &:= \{a | V^\infty(x,a) < V^\infty(x)\}.
\end{aligned} \tag{2.5.80}$$

The following steps can be used to prove the global convergence (2.5.64).

- Show that the probabilities

$$\lim_{t\to\infty} f_{\bar{\theta}_t}(x,a) = 0, \quad \forall a \in I_+^x \cup I_-^x.$$

- Show that for actions $a \in I_-^x$, $\lim_{t\to\infty} \bar{\theta}_t(x,a) = -\infty$ and, for all actions $a \in I_+^x$, $\bar{\theta}_t(x,a)$ is bounded below as $t \to \infty$.

- Prove that the set I_+^x is empty by contradiction for all states x and conclude the global convergence (2.5.64).

Lemma 2.5.10. *Define the advantage function for the critic as*

$$A_t(x,a) := \bar{Q}_t(x,a) - \bar{Q}_t(x). \tag{2.5.81}$$

Then, there exists a T_1 such that $\forall t \geq T_1, x \in \boldsymbol{\mathcal{X}}$, we have

$$A_t(x,a) < -\frac{\Delta}{4} \quad \forall a \in I_-^x; \quad A_t(x,a) > \frac{\Delta}{4} \quad \forall a \in I_+^x. \tag{2.5.82}$$

Proof. Since $\bar{Q}_t(x) \to V^\infty(x)$, we have that there exists $T_1 > T_0$ such that for all $t \geq T_1$,

$$V^\infty(x) - \frac{\Delta}{4} < Q_t(x) < V^\infty(x) + \frac{\Delta}{4}. \tag{2.5.83}$$

Then, for any actions $a \in I_-^x$, we have for any $t \geq T_1 > T_0$

$$\begin{aligned} A_t(x,a) &= \bar{Q}_t(x,a) - \bar{Q}_t(x) \\ &\overset{(a)}{\leq} V^\infty(x,a) + \frac{\Delta}{4} - \bar{Q}_t(x) \\ &\overset{(b)}{\leq} V^\infty(x,a) + \frac{\Delta}{4} - V^\infty(x) + \frac{\Delta}{4} \\ &\overset{(c)}{\leq} -\Delta + \frac{\Delta}{2} \\ &< -\frac{\Delta}{4}, \end{aligned} \tag{2.5.84}$$

where step (a) is by (2.5.74), step (b) is by (2.5.83) and step (c) is by the definition of I_-^x in (2.5.80) and Δ in (2.5.73). Similarly, for $a \in I_+^x$,

$$\begin{aligned} A_t(x,a) &= \bar{Q}_t(x,a) - \bar{Q}_t(x) \\ &\geq V^\infty(x,a) - \frac{\Delta}{4} - \bar{Q}_t(x) \\ &\geq V^\infty(x,a) - \frac{\Delta}{4} - V^\infty(x) - \frac{\Delta}{4} \\ &\geq \Delta - \frac{\Delta}{2} \\ &> \frac{\Delta}{4}. \end{aligned} \tag{2.5.85}$$

□

Lemma 2.5.11. *For any state action pair $(x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$, we have $\lim_{t\to\infty} \widehat{\partial}_{x,a} J(f_{\bar{\theta}_t}) = 0$. This implies that*

$$\lim_{t\to\infty} f_{\bar{\theta}_t}(x,a) = 0, \quad \forall a \in I_+^x \cup I_-^x,$$

and thus

$$\lim_{t\to\infty} \sum_{a \in I_0^x} f_{\bar{\theta}_t}(x,a) = 1. \tag{2.5.86}$$

Lemma 2.5.12 (Monotonicity in $\bar{\theta}_t(x,a)$)**.** *For all $a \in I_+^x$, $\bar{\theta}_t(x,a)$ is strictly increasing for $t \geq T_1$. For all $a \in I_-^x$, $\bar{\theta}_t(x,a)$ is strictly decreasing for $t \geq T_1$.*

Lemma 2.5.13. *For any state x with the set $I_+^x \neq \emptyset$, we have:*

$$\max_{a \in I_0^x} \bar{\theta}_t(x,a) \to \infty, \quad \min_{a \in \mathcal{A}} \bar{\theta}_t(x,a) \to -\infty. \tag{2.5.87}$$

The proofs of Lemmas 2.5.11, 2.5.12, and 2.5.13 are the same as in [1] and therefore are omitted.

Lemma 2.5.14. *For all states x with the set $I_+^x \neq \emptyset$, choose any $a_+ \in I_+^x$. Then, for any $a \in I_0^x$, if there exists $t \geq T_0$ such that $f_{\bar{\theta}_t}(x,a) \leq f_{\bar{\theta}_t}(x,a_+)$, we have*

$$f_{\bar{\theta}_\tau}(x,a) \leq f_{\bar{\theta}_\tau}(x,a_+), \quad \forall \tau \geq t. \tag{2.5.88}$$

Proof. If $f_{\bar{\theta}_t}(x,a) \leq f_{\bar{\theta}_t}(x,a_+)$, we know $\bar{\theta}_t(x,a) \leq \bar{\theta}_t(x,a_+)$ and there exists a small $\epsilon_0 > 0$ such that $f_{\bar{\theta}_t}(x,a_+) \geq \epsilon_0$. Therefore,

$$\begin{aligned}
\widehat{\partial}_{x,a} J(f_{\bar{\theta}_t}) &= \nu_\mu^{f_{\bar{\theta}_t}}(x) f_{\bar{\theta}_t}(x,a) \left[\bar{Q}_t(x,a) - \bar{Q}_t(x)\right] \\
&\overset{(a)}{\leq} \nu_\mu^{f_{\bar{\theta}_t}}(x) f_{\bar{\theta}_t}(x,a_+) \left[\bar{Q}_t(x,a_+) - \bar{Q}_t(x) - \frac{\Delta}{4}\right] \\
&\leq \nu_\mu^{f_{\bar{\theta}_t}}(x) f_{\bar{\theta}_t}(x,a_+) \left[\bar{Q}_t(x,a_+) - \bar{Q}_t(x)\right] - \epsilon_0 \nu_\mu^{f_{\bar{\theta}_t}}(x) \left[\bar{Q}_t(x,a_+) - \bar{Q}_t(x)\right] \\
&\leq \widehat{\partial}_{x,a_+} J(f_{\bar{\theta}_t}) - \nu_\mu^{f_{\bar{\theta}_t}}(x) \frac{\Delta \epsilon_0}{4},
\end{aligned} \tag{2.5.89}$$

where the step (a) follows from $t > T_0$, $a \in I_0^x$ and $a_+ \in I_+^x$,

$$\bar{Q}_t(x,a_+) \geq V^\infty(x,a_+) - \frac{\Delta}{4} \geq V^\infty(x) + \Delta - \frac{\Delta}{4} = V^\infty(x,a) + \frac{3}{4}\Delta > \bar{Q}_t(x,a) + \frac{\Delta}{4}, \tag{2.5.90}$$

and the fact that $\beta_t(x,a)$ decay exponentially. Let $C = \nu_\mu^{f_{\bar{\theta}_t}}(x) \frac{\Delta \epsilon_0}{4}$ and note that

$$\widehat{\partial}_{x,a_+} J(f_{\bar{\theta}_t}) - C \geq 0. \tag{2.5.91}$$

Then, we have

$$\bar{\theta}_t'(x,a) \leq \bar{\theta}_t'(x,a_+) - C\zeta_t. \tag{2.5.92}$$

By the gradient flow (2.5.69), Theorem 2.5.4, and (2.5.43), we have for any action a

$$\frac{\frac{d}{dt}\bar{\theta}_t(x,a)}{\zeta_t} = \widehat{\partial}_{x,a} J(f_{\bar{\theta}_t}) \to 0, \quad t \to \infty. \tag{2.5.93}$$

Thus, without loss of generality, we can suppose that constant T_0 is large enough such that for any $t \geq T_0$ and any action $a \in \mathcal{A}$,

$$-\frac{C}{3}\zeta_t \leq \bar{\theta}_t'(x,a) \leq \frac{C}{3}\zeta_t. \tag{2.5.94}$$

Thus, for any $s > t > T_0$,

$$\begin{aligned}
\bar{\theta}_s'(x,a) &= \bar{\theta}_s'(x,a) - \bar{\theta}_t'(x,a) + \bar{\theta}_t'(x,a) \\
&\overset{(a)}{\leq} \frac{C}{3}\zeta_t + \frac{C}{3}\zeta_t + \bar{\theta}_t'(x,a_+) - C\zeta_t \\
&\leq \bar{\theta}_s'(x,a_+) - \frac{C}{3}\zeta_t,
\end{aligned} \tag{2.5.95}$$

where step (a) use ζ_t is decreasing. Finally, we have for any $T_0 < t \leq \tau$,

$$\begin{aligned}
\bar{\theta}_\tau(x,a) &= \bar{\theta}_t(x,a) + \int_t^\tau \bar{\theta}_s'(x,a) ds \\
&\leq \bar{\theta}_t(x,a_+) + \int_t^\tau \bar{\theta}_s'(x,a_+) ds \\
&= \bar{\theta}_\tau(x,a_+).
\end{aligned} \tag{2.5.96}$$

and therefore (2.5.88) is true. □

For any $a_+ \in I_+^x$, we divide the set I_0^x into two sets $B_0^x(a_+)$ and $\bar{B}_0^x(a_+)$ as follows: $B_0^x(a_+)$

is the set of all $a \in I_0^x$ such that for all $t \geq T_0, f_{\bar{\theta}_t}(x, a_+) < f_{\bar{\theta}_t}(x, a)$ and $\bar{B}_0^x(a_+)$ contains the remainder of the actions from I_0^x. By the definition of $B_0^x(a_+)$, we immediately have two Lemmas.

Lemma 2.5.15. *Suppose for a state $x \in \boldsymbol{\mathcal{X}}$, $I_+^x \neq \emptyset$. Then, $\forall a_+ \in I_+^x$ we have that $B_0^x(a_+) \neq \emptyset$ and that*

$$\lim_{t\to\infty} \sum_{a\in B_0^x(a_+)} f_{\bar{\theta}_t}(x, a) = 1, \tag{2.5.97}$$

which also derives

$$\max_{a\in B_0^x(a_+)} \bar{\theta}_t(x, a) \to \infty. \tag{2.5.98}$$

Lemma 2.5.16. *Consider any x with $I_+^x \neq \emptyset$. Then, for any $a_+ \in I_+^x$, there exists an T_{a_+} such that for all $a \in \bar{B}_0^x(a_+)$*

$$f_{\bar{\theta}_t}(x, a_+) \geq f_{\bar{\theta}_t}(x, a), \quad \forall t > T_{a_+}.$$

The proofs of Lemmas 2.5.15 and 2.5.16 are the same as in [1] and therefore are omitted.

Lemma 2.5.17. *For all actions $a \in I_+^x$, we have that $\bar{\theta}_t(x, a)$ is bounded from below as $t \to \infty$. For all actions $a \in I_-^x$, we have that $\bar{\theta}_t(x, a) \to -\infty$ as $t \to \infty$.*

Proof. From Lemma 2.5.12, we know that when $t \geq T_1$ and for any $a \in I_+^x$, $\bar{\theta}_t(x, a)$ is strictly increasing. Thus $\bar{\theta}_t(x, a)$ is bounded from below for any $a \in I_+^x$. For the second claim, from Lemma 2.5.12 we know that when $t \geq T_1$, $\bar{\theta}_t(x, a)$ is strictly decreasing for $a \in I_-^x$. Therefore, by monotone convergence theorem, $\lim_{t\to\infty} \bar{\theta}_t(x, a)$ exists and is either $-\infty$ or some constant ϵ_0. Next, we prove the convergence to $-\infty$ by contradiction.

Suppose for some $a \in I_-^x$ that there exists a ϵ_0 such that $\bar{\theta}_t(x, a) > \epsilon_0, \forall t \geq T_1$. By Lemma 2.5.13, we know that there exists an action $a' \in \boldsymbol{\mathcal{A}}$ such that

$$\liminf_{t\to\infty} \bar{\theta}_t(x, a') = -\infty. \tag{2.5.99}$$

Choose a constant $\delta > 0$ such that $\bar{\theta}_{T_1}(x, a') \geq \epsilon_0 - \delta$. Then, we can find an increasing sequence $\{t_n\}_{n\geq 0}$ larger than T_1 and converging to ∞ such that

$$\theta_{t_n}(x, a') < \epsilon_0 - \delta, \quad \lim_{n\to\infty} \bar{\theta}_{t_n}(x, a') = -\infty. \tag{2.5.100}$$

Define τ_n as

$$\tau_n := \sup\{s | s \in [T_1, t_n],\ \bar{\theta}_s(x, a') \geq \epsilon_0 - \delta\} \tag{2.5.101}$$

where

$$\mathcal{T}^{(n)} := \{s | s \in (\tau_n, t_n),\ \widehat{\partial}_{x,a'} J(f_{\bar{\theta}_s}) < 0\} \tag{2.5.102}$$

By the continuity of $\widehat{\nabla}_\theta J(f_\theta)$, we know $\mathcal{T}^{(n)}$ is a Lebesgue measurable set. Note that the Lebesgue measure of $\mathcal{T}^{(n)}$ should be positive for all n. Suppose there is a constant n such that $\mathcal{L}(\mathcal{T}^{(n)}) = 0$,

then by $\bar{\theta}_{\tau_n}(x,a') \geq \epsilon_0 - \delta$, we will have

$$\begin{aligned}\bar{\theta}_{t_n}(x,a') &= \bar{\theta}_{\tau_n}(x,a') + \int_{\tau_n}^{t_n} \zeta_s \widehat{\partial}_{x,a'} J(f_{\bar{\theta}_s}) ds \\ &= \bar{\theta}_{\tau_n}(x,a') + \int_{(\tau_n,t_n)\setminus\mathcal{T}^{(n)}} \zeta_s \widehat{\partial}_{x,a'} J(f_{\bar{\theta}_s}) ds \\ &\geq \bar{\theta}_{\tau_n}(x,a') \\ &\geq \epsilon_0 - \delta,\end{aligned} \tag{2.5.103}$$

which contradicts (2.5.100).

Define the sequence $\{Z_n\}_{n\geq 0}$ as

$$Z_n := \int_{\mathcal{T}^{(n)}} \zeta_s \widehat{\partial}_{x,a'} J(f_{\bar{\theta}_s}) ds.$$

Then,

$$Z_n \leq \int_{\tau_n}^{t_n} \zeta_s \widehat{\partial}_{x,a'} J(f_{\bar{\theta}_s}) ds \leq \bar{\theta}_{t_n}(x,a') - (\epsilon_0 - \delta). \tag{2.5.104}$$

By (2.5.100), this implies that

$$\lim_{n\to\infty} Z_n = -\infty \tag{2.5.105}$$

For the positive measure set $\mathcal{T}^{(n)}$, we have for any $t' \in \mathcal{T}^{(n)}$,

$$\left|\frac{\widehat{\partial}_{x,a} J(f_{\bar{\theta}_{t'}})}{\widehat{\partial}_{x,a'} J(f_{\bar{\theta}_{t'}})}\right| = \left|\frac{f_{\bar{\theta}_{t'}}(x,a) A_{t'}(x,a)}{f_{\bar{\theta}_{t'}}(x,a') A_{t'}(x,a')}\right| \geq \exp\left(\epsilon_0 - \bar{\theta}_{t'}(x,a')\right) \frac{(1-\gamma)\Delta}{2} \geq \exp(\delta)\frac{(1-\gamma)\Delta}{2} \tag{2.5.106}$$

where we have used that $|A^{f_{\bar{\theta}_{t'}}}(x,a')| \leq \frac{1}{1-\gamma}$, $|A^{f_{\bar{\theta}_{t'}}}(x,a') - A_{t'}(x,a')| \to 0$ and $|A^{f_{\bar{\theta}_{t'}}}(x,a)| \geq \frac{\Delta}{4}$ for all $t' > T_1$ (from Lemma 2.5.10). Note that since $\widehat{\partial}_{x,a} J(f_{\bar{\theta}_{t'}}) < 0$ and $\widehat{\partial}_{x,a'} J(f_{\bar{\theta}_{t'}}) < 0$ for all $t' \in \mathcal{T}^{(n)}$, we have

$$\widehat{\partial}_{x,a} J(f_{\bar{\theta}_{t'}}) \leq \exp(\delta)\frac{(1-\gamma)\Delta}{2} \widehat{\partial}_{x,a'} J(f_{\bar{\theta}_{t'}}). \tag{2.5.107}$$

Thus

$$\begin{aligned}\bar{\theta}_{t_n}(x,a) &= \bar{\theta}_{T_1}(x,a) + \int_{T_1}^{t_n} \zeta_s \widehat{\partial}_{x,a} J(f_{\bar{\theta}_s}) ds \\ &\overset{(a)}{\leq} \bar{\theta}_{T_1}(x,a) + \int_{\mathcal{T}^{(n)}} \zeta_s \widehat{\partial}_{x,a} J(f_{\bar{\theta}_s}) ds \\ &\overset{(b)}{\leq} \bar{\theta}_{T_1}(x,a) + \exp(\delta)\frac{(1-\gamma)\Delta}{2} \int_{\mathcal{T}^{(n)}} \zeta_s \widehat{\partial}_{x,a'} J(f_{\bar{\theta}_s}) ds \\ &= \bar{\theta}_{T_1}(x,a) + \exp(\delta)\frac{(1-\gamma)\Delta}{2} Z_n.\end{aligned} \tag{2.5.108}$$

where the step (a) follows from $\widehat{\partial}_{x,a} J(f_{\bar{\theta}_s}) < 0$ for any $s \geq T_1$ (Lemma 2.5.12) and step (b) is from (2.5.107). Since (2.5.105) and (2.5.108) contradict that $\bar{\theta}_t(x,a)$ is bounded from below, the proof is complete. □

Lemma 2.5.18. *Consider any state x with $I_+^x \neq \emptyset$. We have for any $a_+ \in I_+^x$,*

$$\lim_{t\to\infty} \sum_{a\in B_0^x(a_+)} \bar{\theta}_t(x,a) = \infty \tag{2.5.109}$$

Proof. By definition of $B_0^x(a_+)$, we know when $t \geq T_0$,

$$f_{\bar{\theta}_t}(x, a_+) < f_{\bar{\theta}_t}(x, a), \quad \forall a \in B_0^x(a_+),$$

which implies $\bar{\theta}_t(x, a_+) < \bar{\theta}_t(x, a)$. By Lemma 2.5.17, we know $\bar{\theta}_t(x, a_+)$ is lower bounded as $t \to \infty$, and thus for all $a \in B_0^x(a_+)$, $\bar{\theta}_t(x, a)$ is lower bounded as $t \to \infty$, which together with $\max_{a \in B_0^x(a_+)} \bar{\theta}_t(x, a) \to \infty$ in Lemma 2.5.15 derive (2.5.109). □

We are now ready to prove the global convergence of the tabular actor-critic algorithm by following the same method in [1].

Lemma 2.5.19 (Global convergence)**.** *For any optimal policy f^*,*

$$J(f^*) - J(f_{\bar{\theta}_t}) \to 0, \quad t \to \infty. \tag{2.5.110}$$

Proof. We only need to prove I_+^x is empty for any x. If so, by (2.5.75)

$$0 \leq \lim_{t\to\infty} [J(f^*) - J(f_{\bar{\theta}_t})] = \lim_{t\to\infty} \sum_{x,a} \sigma_\mu^{f^*}(x, a) A^{f_{\bar{\theta}_t}}(x, a) = \sum_{x,a} \sigma_\mu^{f^*}(x, a) [V^\infty(x, a) - V^\infty(x)] \leq 0, \tag{2.5.111}$$

which implies the global convergence (2.5.110).

Now we prove $I_+^x = \emptyset, \forall x \in \boldsymbol{\mathcal{X}}$ by contradiction. Suppose I_+^x is non-empty for some state $x \in \boldsymbol{\mathcal{X}}$ and let $a_+ \in I_+^x$. Then, from Lemma 2.5.18, we must have

$$\sum_{a \in B_0^x(a_+)} \bar{\theta}_t(x, a) \to \infty. \tag{2.5.112}$$

By Lemma 2.5.17, we know for any $a \in I_-^x$, $\bar{\theta}_t(x, a) \to -\infty$ and $\bar{\theta}_t(x, a_+)$ is bounded from below. Thus we have

$$\frac{f_{\bar{\theta}_t}(x, a)}{f_{\bar{\theta}_t}(x, a_+)} = \exp\{\bar{\theta}_t(x, a) - \bar{\theta}_t(x, a_+)\} \to 0, \tag{2.5.113}$$

and there exists $T_2 > T_0$ such that $\forall t \geq T_2$

$$\frac{f_{\bar{\theta}_t}(x, a)}{f_{\bar{\theta}_t}(x, a_+)} < \frac{(1-\gamma)\Delta}{16|\boldsymbol{\mathcal{A}}|}, \tag{2.5.114}$$

or equivalently

$$-\sum_{a \in I_-^x} \frac{f_{\bar{\theta}_t}(x, a)}{1-\gamma} > -f_{\bar{\theta}_t}(x, a_+) \frac{\Delta}{16}. \tag{2.5.115}$$

Noting that $\bar{B}_0^x \subset I_0^x$, we have

$$\lim_{t\to\infty} A_t(x, a) = 0, \quad \forall a \in \bar{B}_0^x(a_+). \tag{2.5.116}$$

By Lemma 2.5.16,

$$\frac{f_{\bar{\theta}_t}(x, a_+)}{f_{\bar{\theta}_t}(x, a)} \geq 1, \quad \forall t > T_{a_+},$$

which together (2.5.116) derives that there exists $T_3 > T_2, T_{a_+}$ such that

$$|A_t(x, a)| < \frac{f_{\bar{\theta}_t}(x, a_+)}{f_{\bar{\theta}_t}(x, a)} \frac{\Delta}{16|\boldsymbol{\mathcal{A}}|}, \quad \forall t \geq T_3. \tag{2.5.117}$$

Thus we have

$$\sum_{a\in\bar{B}_0^x(a_+)} f_{\bar{\theta}_t}(x,a)\,|A_t(x,a)| < f_{\bar{\theta}_t}(x,a_+)\frac{\Delta}{16}, \tag{2.5.118}$$

or equivalently

$$-f_{\bar{\theta}_t}(x,a_+)\frac{\Delta}{16} < \sum_{a\in\bar{B}_0^x(a_+)} f_{\bar{\theta}_t}(x,a)A_t(x,a) < f_{\bar{\theta}_t}(x,a_+)\frac{\Delta}{16}. \tag{2.5.119}$$

Then, we have for $t > T_3$,

$$\begin{aligned}
0 &\overset{(a)}{=} \sum_{a\in I_0^x} f_{\bar{\theta}_t}(x,a)A_t(x,a) + \sum_{a\in I_+^x} f_{\bar{\theta}_t}(x,a)A_t(x,a) + \sum_{a\in I_-^x} f_{\bar{\theta}_t}(x,a)A_t(x,a)\\
&\overset{(b)}{\geq} \sum_{a\in B_0^x(a_+)} f_{\bar{\theta}_t}(x,a)A_t(x,a) + \sum_{a\in\bar{B}_0^x(a_+)} f_{\bar{\theta}_t}(x,a)A_t(x,a) + f_{\bar{\theta}_t}(x,a_+)A_t(x,a_+) + \sum_{a\in I_-^x} f_{\bar{\theta}_t}(x,a)A_t(x,a)\\
&\overset{(c)}{\geq} \sum_{a\in B_0^x(a_+)} f_{\bar{\theta}_t}(x,a)A_t(x,a) + \sum_{a\in\bar{B}_0^x(a_+)} f_{\bar{\theta}_t}(x,a)A_t(x,a) + f_{\bar{\theta}_t}(x,a_+)\frac{\Delta}{4} - \sum_{a\in I_-^x}\frac{2f_{\bar{\theta}_t}(x,a)}{1-\gamma}\\
&\overset{(d)}{>} \sum_{a\in B_0^x(a_+)} f_{\bar{\theta}_t}(x,a)A_t(x,a) - f_{\bar{\theta}_t}(x,a_+)\frac{\Delta}{16} + f_{\bar{\theta}_t}(x,a_+)\frac{\Delta}{4} - f_{\bar{\theta}_t}(x,a_+)\frac{\Delta}{8}\\
&> \sum_{a\in B_0^x(a_+)} f_{\bar{\theta}_t}(x,a)A_t(x,a),
\end{aligned} \tag{2.5.120}$$

where step (a) is from (2.5.70) and in the step (b) we used $A_t(x,a) > 0$ for all actions $a \in I_+^x$ for $t > T_3 > T_1$ from Lemma 2.5.10. Step (c) follows from $A_t(x,a_+) \geq \frac{\Delta}{4}$ for $t > T_3 > T_1$ from Lemma 2.5.10, the fact $A^{f_{\bar{\theta}_t}}(x,a) \geq -\frac{1}{1-\gamma}$ and the critic convergence $|A^{f_{\bar{\theta}_t}}(x,a) - A_t(x,a)| \to 0$, while step (d) is by (2.5.115) and the left inequality in (2.5.119). This implies that for all $t > T_3$

$$\sum_{a\in B_0^x(a_+)} \widehat{\partial}_{x,a} J(f_{\bar{\theta}_t}) < 0.$$

Then,

$$\lim_{t\to\infty} \sum_{a\in B_0^x(a_+)} \left(\bar{\theta}_t(x,a) - \bar{\theta}_{T_3}(x,a)\right) \leq \int_{T_3}^{\infty} \zeta_t \sum_{a\in B_0^x(a_+)} \widehat{\partial}_{x,a} J(f_{\bar{\theta}_t}) dt < \infty, \tag{2.5.121}$$

which contradicts (2.5.112). Therefore, the set I_+^x must be empty for all $x \in \boldsymbol{\mathcal{X}}$ and then the proof is completed. □

The global convergence in Lemma 2.5.19 can also allow one to prove the global convergence of the policy.

Lemma 2.5.20. *For any deterministic optimal policy f^*, let $a^*(x) = \arg\max_a f^*(x,a)$, $\forall x \in \mathcal{X}$. Recall that the optimal actions set*

$$\boldsymbol{\mathcal{A}}^*(x) := \left\{a^*(x) \in \boldsymbol{\mathcal{A}} : V^{f^*}(x,a^*(x)) = \max_a V^{f^*}(x,a)\right\}, \quad \forall x \in \boldsymbol{\mathcal{X}}.$$

Then, by the convergence (2.5.110), if for each state x the best action $a*(x)$ is unique, we will have

$$\lim_{t\to\infty} f_{\bar{\theta}_t}(x,a^*(x)) = 1, \quad \forall x \in \boldsymbol{\mathcal{X}} \tag{2.5.122}$$

and thus

$$\inf_{x\in\boldsymbol{\mathcal{X}},t\geq 0} f_{\bar{\theta}_t}\left(x,a^*(x)\right)>0. \tag{2.5.123}$$

If for some state x, the best action is not unique, then the convergence (2.5.110) will imply

$$\lim_{t\to\infty}\sum_{a^*(x)\in\boldsymbol{\mathcal{A}}^*(x)} f_{\bar{\theta}_t}(x,a^*(x))=1,\quad \forall x\in\boldsymbol{\mathcal{X}} \tag{2.5.124}$$

and thus

$$\inf_{x\in\boldsymbol{\mathcal{X}},t\geq 0}\sum_{a^*(x)\in\boldsymbol{\mathcal{A}}^*(x)} f_{\bar{\theta}_t}(x,a^*(x))>0. \tag{2.5.125}$$

Proof. As in (2.5.75), we have

$$\begin{aligned} J(f^*)-J(f_{\bar{\theta}_t}) &= \sum_x \nu_\mu^{f^*}(x)\sum_a f^*(x,a)A^{f_{\bar{\theta}_t}}(x,a)\\ &= \sum_x \nu_\mu^{f^*}(x)A^{f_{\bar{\theta}_t}}(x,a^*(x))\\ &= \sum_x \nu_\mu^{f^*}(x)\left[V^{f_{\bar{\theta}_t}}(x,a^*(x))-\sum_{a'}V^{f_{\bar{\theta}_t}}(x,a')f_{\bar{\theta}_t}(x,a')\right] \end{aligned} \tag{2.5.126}$$

By (2.5.110), we have the convergence

$$0=\lim_{t\to\infty}\left[J(f^*)-J(f_{\bar{\theta}_t})\right]=\sum_x \mu(x)\left[V^{f^*}(x)-V^{f_{\bar{\theta}_t}}(x)\right], \tag{2.5.127}$$

which together with $\mu(x)>0, V^{f^*}(x)-V^{f_{\bar{\theta}_t}}(x)\geq 0, \forall x\in\boldsymbol{\mathcal{X}}$ and the relationship (2.5.79) leads to

$$\begin{aligned} &\lim_{t\to\infty} V^{f^*}(x)-V^{f_{\bar{\theta}_t}}(x)=0,\quad \forall x\in\boldsymbol{\mathcal{X}}\\ &\lim_{t\to\infty} V^{f^*}(x,a)-V^{f_{\bar{\theta}_t}}(x,a)=0,\quad \forall (x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}. \end{aligned} \tag{2.5.128}$$

Combining (2.5.126) and (2.5.128), we have

$$\begin{aligned} 0&=\lim_{t\to\infty}\left[J(f^*)-J(f_{\bar{\theta}_t})\right]\\ &=\lim_{t\to\infty}\sum_x \nu_\mu^{f^*}(x)\left[V^{f_{\bar{\theta}_t}}(x,a^*(x))-\sum_{a'}V^{f_{\bar{\theta}_t}}(x,a')f_{\bar{\theta}_t}(x,a')\right]\\ &=\lim_{t\to\infty}\sum_x \nu_\mu^{f^*}(x)\left[\max_a V^{f^*}(x,a)-\sum_{a'}V^{f^*}(x,a')f_{\bar{\theta}_t}(x,a')\right]\\ &\overset{(a)}{\geq}\lim_{t\to\infty}\sum_x \mu(x)\left[\max_a V^{f^*}(x,a)-\sum_{a'}V^{f^*}(x,a')f_{\bar{\theta}_t}(x,a')\right] \end{aligned} \tag{2.5.129}$$

where step (a) is due to

$$\max_a V^{f^*}(x,a)-\sum_{a'}V^{f^*}(x,a')f_{\bar{\theta}_t}(x,a')\geq 0,\quad \forall x\in\boldsymbol{\mathcal{X}}.$$

Then we have

$$\lim_{t\to\infty}\left[V^{f^*}(x,a^*(x))-\sum_{a'}V^{f^*}(x,a')f_{\bar{\theta}_t}(x,a')\right]=0,\quad \forall x\in\boldsymbol{\mathcal{X}}. \tag{2.5.130}$$

Thus if the best action $a^*(x)$ for any state $x\in\boldsymbol{\mathcal{X}}$ is unique, (2.5.130) derives

$$\lim_{t\to\infty} f_{\bar{\theta}_t}\left(x,a^*(x)\right)=1,\quad \forall x\in\boldsymbol{\mathcal{X}}.$$

When there exist multiple optimal actions in $\boldsymbol{\mathcal{A}}^*(x)$, (2.5.130) derives

$$\lim_{t\to\infty} \sum_{a^*(x)\in\boldsymbol{\mathcal{A}}^*(x)} f_{\bar{\theta}_t}(x, a^*(x)) = 1, \quad \forall x \in \boldsymbol{\mathcal{X}}.$$

Finally, noting that f_θ being a softmax policy and the bound in Lemma 2.4.1, for any finite $t > 0$, the policy is positive. Thus (2.5.123) and (2.5.125) are direct corollary of (2.5.122) and (2.5.124). □

Finally, combining Lemma 2.5.5 and Lemma 2.5.20, we can obtain the uniform Lojasiewicz inequality, which will prove the convergence rate (2.3.6).

Proof of (2.3.6)*:* Define the actor error

$$Y_t := J(f^*) - J(f_{\bar{\theta}_t}).$$

Then, by chain rule,

$$\begin{aligned}
\frac{dY_t}{dt} &= -\zeta_t \nabla_\theta J(f_{\bar{\theta}_t}) \widehat{\nabla}_\theta J(f_{\bar{\theta}_t}) \\
&= -\zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 + \zeta_t \nabla_\theta J(f_{\bar{\theta}_t}) \left(\nabla_\theta J(f_{\bar{\theta}_t}) - \widehat{\nabla}_\theta J(f_{\bar{\theta}_t}) \right) \\
&\leq -\zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 + C\zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\| \cdot \|\bar{Q}_t - V^{f_{\bar{\theta}_t}}\|_2 \\
&\leq -\zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 + C\zeta_t \eta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\| \\
&\leq -(\zeta_t - C\zeta_t\eta_t) \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 + C\zeta_t\eta_t \\
&\leq -C\zeta_t \|\nabla_\theta J(f_{\bar{\theta}_t})\|^2 + C\zeta_t\eta_t.
\end{aligned} \tag{2.5.131}$$

By Lemma 2.5.5 and Lemma 2.5.20, there exists a constant $C > 0$ such that

$$\|\nabla_\theta J(f_{\bar{\theta}_t})\| \geq C \left[J(f^*) - J(f_{\bar{\theta}_t}) \right] = CY_t, \tag{2.5.132}$$

which together with (2.5.131) derives

$$\begin{aligned}
\frac{dY_t}{dt} &\leq -C\zeta_t Y_t^2 + C\zeta_t\eta_t \\
&< -\frac{C}{t} Y_t^2 + \frac{C}{t\log^2 t}.
\end{aligned} \tag{2.5.133}$$

Consider the comparison ODE:

$$\begin{aligned}
&\frac{dZ_t}{dt} = -\frac{C}{t} Z_t^2 + \frac{C}{t\log^2 t}, \quad t \geq 2. \\
&Z_2 > Y_2,
\end{aligned} \tag{2.5.134}$$

By the Basic Comparison Theorem in [101], we have

$$0 \leq Y_t < Z_t \quad t \geq 2. \tag{2.5.135}$$

Then, if we can establish a convergence rate for Z_t, we will have a convergence rate for Y_t.

Without loss of generality, we suppose the constant $C = 1$ and define function

$$0 \leq X_t = Z_t \log t, \quad t \geq 2.$$

Thus,

$$\begin{aligned}\frac{dX_t}{dt} &= \frac{1}{t}Z_t + \log t\left(-\frac{1}{t}Z_t^2 + \frac{1}{t\log^2 t}\right)\\ &= \frac{1}{t\log t}\left(Z_t\log t - Z_t^2\log^2 t + 1\right)\\ &= \frac{1}{t\log t}\left(X_t - X_t^2 + 1\right), \quad t \geq 2.\end{aligned} \tag{2.5.136}$$

Noting that $\frac{1-\sqrt{5}}{2}$ and $\frac{1+\sqrt{5}}{2}$ are two stationary solution of (2.5.136), the solution X_t will decrease if it is larger than $\frac{1+\sqrt{5}}{2}$ and it will increase for $X_t \in [0, \frac{1+\sqrt{5}}{2}]$. Thus, for a solution X_t starting from $X_2 \geq 0$, there are two cases:

(1) If the starting point $X_2 \geq \frac{1+\sqrt{5}}{2}$, the solution X_t will decrease and always be larger than $\frac{1+\sqrt{5}}{2}$ by the uniqueness theorem for ODEs (Theorem 2.2 of [141]).

(2) If the starting point $X_2 \in [0, \frac{1+\sqrt{5}}{2}]$, the solution X_t will increase and always be smaller than $\frac{1+\sqrt{5}}{2}$ by the uniqueness theorem for ODEs (Theorem 2.2 of [141]).

Thus, no matter where X_t starts from, we always have

$$0 \leq X_t \leq \max\{X_2, \frac{1+\sqrt{5}}{2}\}, \quad t \geq 2, \tag{2.5.137}$$

which shows that

$$0 \leq Y_t < Z_t \leq \frac{C}{\log t}, \quad t \geq 2, \tag{2.5.138}$$

and therefore the convergence rate (2.3.6) is proven. □

Chapter 3

Convergence of Online Neural Network Actor-Critic Algorithms

3.1 Introduction

Neural network actor-critic algorithms are one of the most popular methods in deep reinforcement learning. A neural network model is trained to select the policy (the "actor") while another neural network (the "critic") is simultaneously trained to learn the value function given the actor's policy. Specifically, the actor selects an action and, given the action, a new state transition occurs according to a Markov stochastic process and a reward (a measurement of the success/failure) is observed. The critic must learn to approximate the value function – the solution to the Bellman equation – given the actor's policy. Given the critic's estimate for the value function of the current policy, the actor must be updated to improve the value function (i.e., increase the expected reward). Actor-critic algorithms are well-established methods in reinforcement learning [85, 83]; the key recent advance is using (deep) neural networks as the actor/critic and training their parameters using gradient descent methods [108, 53, 107, 16, 117].

Analysis of neural network actor-critic algorithms is challenging due to: (1) the non-convexity of the neural networks, (2) the complex feedback loop between the actor and critic (the actor determines the sequence of data samples which are used to train the critic and the critic is used to train the actor), and (3) the simultaneous online updates of both the actor and critic which lead to (3A) the distribution of the data, which depends upon the actor, constantly evolving in time and (3B) the actor being updated with a noisy, biased estimate of the value function.

3.1.1 Convergence Analysis of Actor-critic Algorithms

ODE Methods Various versions of actor-critic algorithms have been studied under the framework of stochastic approximation algorithms, see [84, 20, 83, 82] and the associated references for an extensive discussion and literature review. A common way of analysing the stability and convergence of this class of algorithms would be to show that the algorithm converges to the limit set of an associated ODE [11, 21, 22]. As a result, the algorithm can be studied by characterizing the limit

set of the ODE [20, 38]. The references [11, 24] provide general overviews of this method. We note that the stability of the actor-critic algorithm can be established via a pure martingale argument [82].

Although our approach also connects the actor-critic algorithm with an ODE, the analysis and convergence theorem are different. Here we establish the *pathwise uniform* convergence of the time-rescaled trajectory of the actor-critic algorithm using weak convergence techniques [56] as the number of hidden units and training steps $\to \infty$. The convergence to the limit ODE with a neural network actor and a neural network critic as the number of hidden units $\to \infty$ was not previously considered in the ODE literature discussed above.

Finite time analysis Non-asymptotic convergence rates can also be established for the actor-critic algorithm using finite-time analysis approaches. These results establish a convergence rate to a time when the optimality gap is arbitrarily small. Finite-time convergence rates for actor-critic algorithms with linear approximators for the action value function have been proven in [149, 148, 86].

Recent advances using neural tangent kernel (NTK) analysis [121, 120, 70, 90] has enabled finite-time analysis on various versions of the neural network actor-critic algorithm. Building upon the NTK results [121, 120, 70, 90], [142, 39] study a "batch" version of the actor-critic algorithm where a large number of critic parameter updates are required for each actor update to ensure accurate approximation of the action-value function. A convergence result is proven when the ratio of critic updates for each actor updates $\to \infty$. In particular, [142, 39] establish that the batch actor-critic algorithm can become arbitrarily close to a stationary point within a large but finite number of iterations. These results do not guarantee the convergence of the actor-critic algorithm as the training time $\to \infty$, as the parameters could escape from the global/local minimum of the loss function.

While [142, 39] study the batch version of the actor-critic algorithm – where the number of critic updates for each actor update $\to \infty$ at each iteration – we develop a convergence analysis for *online* neural network actor-critic algorithm where there is a single actor and a single critic update at each iteration. The advantage of the online algorithm is that a much larger number of optimization iterations can be completed in the same computational time. The online updates introduce key mathematical differences to the analysis. The learning rates for both the actor and critic must be carefully selected in order to guarantee convergence in the online setting. In addition, the exploration policy for the actor must also be carefully designed. A two-timescale analysis to separate the timescales of the actor and critic must be applied in combination with the NTK methods. Due to the online updates, a Poisson equation must be used to analyze the fluctuations of the algorithm around its limit trajectory. The main mathematical result is also different; we characterize the limit of the neural network actor-critic algorithm as the number of training steps and hidden units $\to \infty$,

proving that it converges to the solution of a system of ODEs using weak convergence techniques. Finally, we prove that the limit ODE converges to a stationary point of the expected reward as the training time $\to \infty$. Similar to [142, 39], this also implies that there is a finite training time such that the pre-limit algorithm converges arbitrarily close to a stationary point of the objective function.

3.1.2 Our Mathematical Approach

We prove that the *trajectory* of the time-rescaled neural network outputs converges *pathwise* weakly to an ODE with random initialisation as the number of hidden units $\to \infty$. We then prove that the limit critic converges to the value function and the actor converges to a stationary point of the objective function as the training time $\to \infty$. In particular, we show that both

- the *Bellman error* for the critic model and
- the norm of the gradient of the objective function with respect to the actor

converge to zero as the training time tends to infinity. These results are stated formally in Section 3.3. Our results are strictly stronger than the classical ODE approaches in [20, 38] as it provides information about the training trajectory. We prove that the trained limit neural network *converges* to a stationary point as the training time $t \to \infty$. In this chapter, a constant learning rate is used for the critic and a logarithmic learning rate is used for the actor, which asymptotically yields accurate value function estimates for the online actor update. These learning rates are non-standard in the classical approach (see [24, 21, 84, 82]).

The convergence to a limit ODE is a result of the neural network parameters remaining within a small neighborhood of their initial values as they train. This result is referred to as the Neural Tangent Kernel (NTK) result and was discovered in [90] for feedforward networks in supervised learning. The NTK analysis has been widely-used to study neural networks, including for reinforcement learning algorithms [133, 143]. Therefore, the evolving neural network outputs (during training) can be linearized around the initial empirical distribution of the parameters. In the reinforcement learning setting, convergence to the limit ODE with the NTK kernel requires the analysis of non-i.i.d. data samples whose distribution depends upon the neural network parameters (since the distribution of the Markov chain depends upon the actor). The actor parameter updates themselves depend upon the data samples, introducing a complex feedback loop. Our analysis requires constructing an appropriate Poisson equation to address this challenge.

We first establish the geometric ergodicity of the data samples under a fixed actor policy. Then, using the Poisson equation, we prove that the fluctuations of the model updates around the limit

distribution due to the randomly-arriving data samples vanish as the number of parameter updates $\to \infty$. Using the Poisson equation and weak convergence techniques, we prove that the actor neural network and critic neural network converge to the solutions of a system of ODEs with random initial conditions. Unlike in the classic NTK analysis of feedforward neural networks which produces a *linear* limit ODE, the limit ODE for the actor-critic algorithm is nonlinear. Leveraging the two timescales for the actor and critic ODEs (due to their respective learning rates), we are able to prove that the critic ODE converges to the true value function (the solution of the Bellman equation) as the training time $t \to \infty$, which provides the actor with an asymptotically unbiased estimate of the policy gradient. We then prove that the limit actor network will converge to a stationary point of the objective function as $t \to \infty$. Therefore, although in the pre-limit actor-critic algorithm the critic provides a noisy, biased (i.e., there is error) estimate of the value function, we are able to prove that asymptotically the critic will converge sufficiently rapidly such that the actor also converges.

3.1.3 Organisation of the analysis

Section 3.2 describes the class of actor-critic algorithms that we study. Section 3.3 states the main convergence results that are proven. The proof of the main result is presented in Section 3.5. Finally, we analyse the limit ODE as $t \to \infty$ in Section 3.5 to establish the convergence of critic network to the true action-value function and the convergence of actor network to a stationary point of the expected discounted future reward.

3.2 Actor-Critic Algorithms

3.2.1 Markov Decision Processes

We will study a neural network actor-critic algorithm for the following Markov decision process (MDP).

Definition 3.2.1 (Markov decision process (MDP))**.** A Markov decision process $\boldsymbol{\mathcal{M}} = (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}, p, \rho_0, r, \gamma)$ consists of the following:

- $\boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^{d_x}$, the space of all possible states of the MDP (the *state space*);
- $\boldsymbol{\mathcal{A}} \subseteq \mathbb{R}^{d_a}$, the space of all actions of the MDP (the *action space*);
- $p(x'|x, a)$, the transition kernel that gives the probability of next state being x' if the current state is x and the action a is taken;
- ρ_0, the distribution that characterises how the initial state and action are chosen,
- $r(x, a)$, the reward gained by taking action a at state x, and
- $\gamma \in (0, 1)$ being the *discount factor*.

Here $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}} \subset \mathbb{R}^d$, where $d = d_x + d_a$. Any elements $\xi := (x, a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$ are called *state-action* pairs.

We make the same assumptions on the MDP as the ones made in [143]:

Assumption 3.2.2 (Basic assumptions on the MDP)**.**

- Finite state space: we assume that the state space $\boldsymbol{\mathcal{X}}$ is discrete and finite with size $\#\boldsymbol{\mathcal{X}}$,
- Finite action space: we assume that the action space $\boldsymbol{\mathcal{A}}$ is discrete and finite with size $\#\boldsymbol{\mathcal{A}}$, and
- The reward function r is bounded in $[-1, 1]$.

We denote the size of the state-action space $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$ as $M = \#\boldsymbol{\mathcal{X}} \times \#\boldsymbol{\mathcal{A}}$.

3.2.2 Policy in the MDP

A policy $f = f(x, a)$ specifies the probability of selecting action a at state x. The policy f acts on the MDP $\boldsymbol{\mathcal{M}}$ to induce the following Markov chain on the state-action pair $\xi_k := (x_k, a_k)$:

$$(\boldsymbol{\mathcal{M}}, f) : \xi_0 := (x_0, a_0) \sim \rho_0 \xrightarrow[=p(\cdot|\xi_0)]{p(\cdot|x_0,a_0)} x_1 \xrightarrow{f(x_1,\cdot)} a_1 \xrightarrow[=p(\cdot|\xi_1)]{p(\cdot|x_1,a_1)} x_2 \xrightarrow{f(x_2,\cdot)} a_2 \xrightarrow[=p(\cdot|\xi_2)]{p(\cdot|x_2,a_2)} x_3 \xrightarrow{f(x_3,\cdot)} a_3 \cdots, \tag{3.2.1}$$

which is time-homogeneous with initial distribution ρ_0 and transition kernel

$$f(x_{k+1}, a_{k+1})\, p(x_{k+1} \,|\, x_k, a_k)$$

from $\xi_k = (x_k, a_k)$ to $\xi_{k+1} = (x_{k+1}, a_{k+1})$.

The overall reward for a policy f in the MDP $\boldsymbol{\mathcal{M}}$ is evaluated by the following state and action-value functions:

Definition 3.2.3 (State and action-value functions)**.** The state and action-value functions of a policy f acting on MDP $\boldsymbol{\mathcal{M}}$ is defined as follows:

- the *state*-value function $V^f : \boldsymbol{\mathcal{X}} \to \mathbb{R}$ is the expected discounted sum of future awards when the MDP is started from a certain state x and there is a fixed policy f for all timesteps:

$$V^f(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(\xi_k) \mid x_0 = x\right], \tag{3.2.2}$$

  and

- the *action*-value function $V^f : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ is the expected discounted sum of future awards when the MDP is started from a certain state-action pair (x, a) and there is a fixed policy f for all timesteps:

$$V^f(x, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(\xi_k) \mid x_0 = x, a_0 = a\right]. \tag{3.2.3}$$

Both expectations are taken with respect to the Markov chain $(\boldsymbol{\mathcal{M}}, f) := (\xi_k)_{k\geq 0} = (x_k, a_k)_{k\geq 0}$.

Remark 3.2.4. These expectations are well-defined as $\gamma \in (0,1)$ and $r(\cdot)$ are bounded; see the remarks at the beginning of Section 2 of [143].

We define further the state and state-action visiting measures of a policy f:

Definition 3.2.5 (State and state-action visiting measures, see e.g. [137, 83] and Section 2 of [143])**.** Let $(\boldsymbol{\mathcal{M}}, f) := (x_k, a_k)_{k\geq 0}$ be the Markov chain induced when the policy f acts on the MDP $\boldsymbol{\mathcal{M}}$. Let $\xi = (x, a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$ be a state-action pair of the MDP $\boldsymbol{\mathcal{M}}$. Let

- $\mathbb{P}(x_k = x)$ be the probability that $x_k = x$ for $(\boldsymbol{\mathcal{M}}, f)$, and
- $\mathbb{P}(x_k = x, a_k = a) := \mathbb{P}(x_k = x) f(x, a)$ be the probability that $x_k = x$ and $a_k = a$ for $(\boldsymbol{\mathcal{M}}, f)$.

Then, we define the state and state-action visiting measures respectively as $\nu^f_{\rho_0}$ and $\sigma^f_{\rho_0}$, such that

$$\nu^f_{\rho_0}(\{x\}) = \sum_{k=0}^{\infty} \gamma^k \mathbb{P}\left(x_k = x\right), \quad \sigma^f_{\rho_0}(\{(x,a)\}) = \sum_{k=0}^{\infty} \gamma^k \mathbb{P}\left(x_k = x, a_k = a\right), \tag{3.2.4}$$

Remark 3.2.6.

- Both $(1-\gamma)\nu^f_{\rho_0}(\cdot)$ and $(1-\gamma)\sigma^f_{\rho_0}(\cdot)$ are probability measures.
- Define the auxiliary Markov chain induced when the policy f acts on the MDP $\boldsymbol{\mathcal{M}}$ in terms of the state-action pair $\tilde{\xi}_k := (\tilde{x}_k, \tilde{a}_k)$:

$$(\boldsymbol{\mathcal{M}}, f)_{\mathsf{aux}} : (\tilde{x}_0, \tilde{a}_0) \sim \rho_0 \xrightarrow[=\tilde{p}(\cdot|\tilde{\xi}_0)]{\tilde{p}(\cdot|\tilde{x}_0,\tilde{a}_0)} \tilde{x}_1 \xrightarrow{f(\tilde{x}_1,\cdot)} \tilde{a}_1 \xrightarrow[=\tilde{p}(\cdot|\tilde{\xi}_1)]{\tilde{p}(\cdot|\tilde{x}_1,\tilde{a}_1)} \tilde{x}_2 \xrightarrow{f(\tilde{x}_2,\cdot)} \tilde{a}_2 \xrightarrow[=\tilde{p}(\cdot|\tilde{\xi}_2)]{\tilde{p}(\cdot|\tilde{x}_2,\tilde{a}_2)} \tilde{x}_3 \xrightarrow{f(\tilde{x}_3,\cdot)} \tilde{a}_3 \cdots, \tag{3.2.5}$$

  where

$$\tilde{p}\left(\tilde{x}' \mid \tilde{x}, \tilde{a}\right) = \gamma p\left(\tilde{x}' \mid \tilde{x}, \tilde{a}\right) + (1-\gamma)\rho_0\left(\tilde{x}'\right), \quad \forall\left(\tilde{x}, \tilde{a}, \tilde{x}'\right) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{X}} \tag{3.2.6}$$

  sample from the initial distribution ρ_0 with probability $1-\gamma$ at each transition to a new state. Then $(1-\gamma)\sigma^f_{\rho_0}$ is the stationary measure of the auxiliary Markov chain $(\boldsymbol{\mathcal{M}}, f)_{\mathsf{aux}}$. This is proven on page 36 of [83].

We make the assumption on the transition p of an MDP $\boldsymbol{\mathcal{M}}$ to ensure ergodicity for the Markov chains $(\boldsymbol{\mathcal{M}}, f)$ and $(\boldsymbol{\mathcal{M}}, f)_{\mathsf{aux}}$. The assumption is stated in terms of the total variation (TV) distance. The TV distance between two probability distributions on $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$ with masses p_1 and p_2 are defined as

$$d_{\mathsf{TV}}(p_1, p_2) = \frac{1}{2} \sum_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |p_1(\xi) - p_2(\xi)|. \tag{3.2.7}$$

Assumption 3.2.7 (Ergodicity of the MDP)**.** We assume that the Markov chains $(\boldsymbol{\mathcal{M}}, f)$ and $(\boldsymbol{\mathcal{M}}, f)_{\mathsf{aux}}$ are both ergodic (irreducible and aperiodic) whenever f selects every action with positive probability. As a result, both $(\boldsymbol{\mathcal{M}}, f)$ and $(\boldsymbol{\mathcal{M}}, f)_{\mathsf{aux}}$ have a unique stationary distribution (see section 1.3.3 of [89] and page 36 of [83]), denoted as π^f and $\sigma^f_{\rho_0}$ respectively. Furthermore, we assume that both π^f and $\sigma^f_{\rho_0}$ are globally Lipschitz of f with respect to the TV distance, so that there exists $C > 0$ such that for any policies f, f',

$$\max(d_{\mathsf{TV}}(\pi^f, \pi^{f'}), d_{\mathsf{TV}}(\sigma^f_{\rho_0}, \sigma^{f'}_{\rho_0})) \leq d_{\mathsf{TV}}(f, f'). \tag{3.2.8}$$

3.2.3 Online Neural Network Actor-critic Algorithm

The main goal of reinforcement learning is to learn the optimal policy f^* which maximizes the expected discounted sum of the future rewards:

$$\max_f J(f), \tag{3.2.9}$$

where the objective function $J(f)$ is the state-value function, weighted by the initial state-action pair:

$$J(f) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot r\left(x_k, a_k\right)\right] = \sum_{x \in \boldsymbol{\mathcal{X}}} \rho_0(x) V^f(x) = \sum_{\xi=(x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} \sigma^f_{\rho_0}(\xi) r(\xi), \tag{3.2.10}$$

and see also equation (2.3) of [143]. Policy-based reinforcement learning methods optimize the objective function over a class of policies $\{f_\theta \mid \theta \in \boldsymbol{\Theta}\}$ based on the policy gradient theorem [136]. In practice, the value functions are unknown and must therefore also be estimated. In this chapter, we study the *online* actor-critic algorithms, which simultaneously estimate the action-value function using a *critic* model and the optimal policy using an *actor* model at every time step of the MDP:

- The *actor model*, acting as the approximation of an optimal policy, is defined as

$$f^N_\theta(\xi) = \mathrm{Softmax}(P^N_\theta(\xi)) = \frac{\exp(P^N_\theta(x,a))}{\sum_{a'} \exp(P^N_\theta(x,a'))}, \quad \xi = (x,a) \tag{3.2.11}$$

  where $P^N_\theta(\xi)$ is the *actor network*:

$$P^N_\theta(\xi) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} B^i \sigma\left(U^i \cdot \xi\right), \tag{3.2.12}$$

  parameterized by the parameters $\theta = (B^1, \ldots, B^N, U^1, \ldots U^N)$, where $B^i \in \mathbb{R}$ and $U^i \in \mathbb{R}^d$.

- The *critic model*, acting as the approximation of the unknown state-action value function for the optimal policy (approximated by the actor model), is the *critic network*

$$Q^N_\omega(\xi) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} C^i \sigma\left(W^i \cdot \xi\right), \tag{3.2.13}$$

  parameterized by the parameters $\omega = (C^1, \ldots, C^N, W^1, \ldots W^N)$, where $C^i \in \mathbb{R}$ and $W^i \in \mathbb{R}^d$.

Remark 3.2.8. We emphasise that

- The outputs of actor and critic networks P_k^N, Q_k^N could be viewed as either functions on $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$ or as vectors in $\mathbb{R}^M$ indexed by elements in $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$, and
- f_k^N refers to the actor model (i.e., the *probability distribution* output of the actor network), which could be viewed as either a function of $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$ or as a vector in $\mathbb{R}^M$ indexed by elements in $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$.

These interpretations are interchangeable.

Assumption 3.2.9 (Activation function)**.** The scalar function $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$, known as the *activation function*, is assumed to be

- twice continuously differentiable (i.e. in $C_b^2(\mathbb{R})$) with outputs and derivatives bounded by 1, and
- slowly increasing, such that for any $a > 0$,

$$\lim_{x \to \pm\infty} \frac{\sigma(x)}{x^a} \to 0.$$

An example would be the standard sigmoid function $\sigma(x) = (1 + e^{-x})^{-1}$.

$\theta_k = (B_k^1, \ldots, B_k^N, U_k^1, \ldots U_k^N)$ and $\omega_k = (C_k^1, \ldots, C_k^N, W_k^1, \ldots W_k^N)$ are the trained parameters of the actor and critic networks after k training updates. We also define $P_k^N := P_{\theta_k}^N$, $f_k^N := \text{Softmax}(P_k^N)$ and $Q_k^N := Q_{\omega_k}^N$.

Our Actor-critic algorithm is online, which means that the policies used to sample state-action pairs in the MDP will change at each iteration. Similar to the coupled system in [144, 145], our version of the Actor-critic algorithm will sample two parallel sequences of state-action pairs:

- the "actor" process:

$$(\boldsymbol{\mathcal{M}}, \mathsf{Ac}) : (\widetilde{x}_0, \widetilde{a}_0) \sim \rho_0 \xrightarrow[=\tilde{p}(\cdot|\tilde{\xi}_0)]{\tilde{p}(\cdot|\tilde{x}_0,\tilde{a}_0)} \widetilde{x}_1 \xrightarrow{g_0^N(\widetilde{x}_1,\cdot)} \widetilde{a}_1 \xrightarrow[=\tilde{p}(\cdot|\tilde{\xi}_1)]{\tilde{p}(\cdot|\tilde{x}_1,\tilde{a}_1)} \widetilde{x}_2 \xrightarrow{g_1^N(\widetilde{x}_2,\cdot)} \widetilde{a}_2 \xrightarrow[=\tilde{p}(\cdot|\tilde{\xi}_2)]{\tilde{p}(\cdot|\tilde{x}_2,\tilde{a}_2)} \widetilde{x}_3 \xrightarrow{g_2^N(\widetilde{x}_3,\cdot)} \widetilde{a}_3 \cdots, \tag{3.2.14}$$

  and

- the "critic" process:

$$(\boldsymbol{\mathcal{M}}, \mathsf{Cr}) : (x_0, a_0) \sim \rho_0 \xrightarrow[=p(\cdot|\xi_0)]{p(\cdot|x_0,a_0)} x_1 \xrightarrow{g_0^N(x_1,\cdot)} a_1 \xrightarrow[=p(\cdot|\xi_1)]{p(\cdot|x_1,a_1)} x_2 \xrightarrow{g_1^N(x_2,\cdot)} a_2 \xrightarrow[=p(\cdot|\xi_2)]{p(\cdot|x_2,a_2)} x_3 \xrightarrow{g_2^N(x_3,\cdot)} a_3 \cdots, \tag{3.2.15}$$

where the *exploration policy* g_k^N is defined as

$$g_k^N(\xi) = \frac{\eta_k^N}{\#\boldsymbol{\mathcal{A}}} + (1 - \eta_k^N) \cdot f_k^N(\xi), \quad \xi = (x, a), \tag{3.2.16}$$

and $(\eta_k^N)_{k\geq 0}$ is a sequence of exploration rates such that $0 < \eta_k^N \leq 1$ and $\eta_k^N \overset{k\to\infty}{\to} 0$. This ensures that each action in $\boldsymbol{\mathcal{A}}$ is selected with probability at least $\eta_k^N/\#\boldsymbol{\mathcal{A}} > 0$, and so by Assumption 3.2.7 the induced Markov chains $(\boldsymbol{\mathcal{M}}, g_k^N)$ and $(\boldsymbol{\mathcal{M}}, g_k^N)_{\mathsf{aux}}$ are both ergodic, and the stationary measures $\pi^{g_\theta^N}$ and $\sigma_{\rho_0}^{g_\theta^N}$ are well-defined (exist and are unique). This will be made precise in Lemma 3.4.12.

We will now describe the two main steps of the online actor-critic algorithm at each optimization iteration.

Step 1: Update of the critic network: We first update the critic network's parameters by temporal difference learning [146]. Temporal difference learning aims to take a stochastic gradient descent step at the sample *critic loss* with respect to the critic network parameters ω_k:

$$L^{\theta_k}(\omega_k) := \sum_{\xi} \left[Y_k(\xi) - Q_k^N(\xi)\right]^2 \pi^{f_k^N}, \tag{3.2.17}$$

where the "target" Y_k is defined as

$$Y_k(\xi) := r(\xi) + \gamma \sum_{x'} \left[\sum_{a'} Q_k^N(x', a') f_k^N(x', a')\right] p(x'|\xi) \tag{3.2.18}$$

and $\pi^{f_\theta^N}$ is the unique stationary distribution of the Markov chain $(\boldsymbol{\mathcal{M}}, f_\theta^N)$ as specified in Assumption 3.2.2. In fact, if $\pi^{f_k^N}(\xi) > 0$ for all $\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$ and $L^{\theta_k}(\omega_k) = 0$, then $Q_k^N(\xi)$ satisfies the Bellman equation and hence is a value function of f_k^N.

Unfortunately the stationary distribution $\pi^{f_k^N}(\xi)$ is inaccessible, so we use $\xi_k = (x_k, a_k)$ from the critic process $(\boldsymbol{\mathcal{M}}, \mathsf{Cr})$ as a *sample* of $\pi^{f_k^N}$ to estimate and evaluate the gradient over the sample critic loss

$$\ell^{\theta_k}(\omega_k) := \left[Y_k(\xi_k) - Q_{\omega_k}^N(\xi_k)\right]^2. \tag{3.2.19}$$

We emphasise that the critic process $(\boldsymbol{\mathcal{M}}, \mathsf{Cr})$ evolves as the following for any $k \geq 1$:

$$x_{k+1} \sim p(\cdot \,|\, \xi_k) = p(\cdot \,|\, x_k, a_k), \quad a_{k+1} \sim g_k^N(x_k, \cdot). \tag{3.2.20}$$

Further note that the term $Y_k(\xi_k)$ involves an expectation of $Q_{\omega_k}^N(\cdot,\cdot)$ with respect to the distribution $f_{\theta_k}^N(\cdot,\cdot)p(\cdot|\xi_k)$, which could be replaced by the estimate $Q(\xi_{k+1})$. Treating the target $Y^{\theta_k}(\xi_k)$ as constant, we have the following gradient-descent-like update for the critic parameters

$$\begin{aligned} C_{k+1}^i &= C_k^i + \frac{\alpha^N}{\sqrt{N}} \left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right) \sigma\left(W_k^i \cdot \xi_k\right), \\ W_{k+1}^i &= W_k^i + \frac{\alpha^N}{\sqrt{N}} \left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right) C_k^i \sigma'\left(W_k^i \cdot \xi_k\right) \xi_k, \end{aligned} \tag{3.2.21}$$

where $\alpha^N = 1/N$ is the scaling of the step size of parameter updates, chosen so that the parameter updates converge to a limiting ODE as $N \to \infty$.

Step 2: Update of the actor network: We then use the policy gradient theorem [137] to update the actor network's parameters. The policy gradient theorem states that if a policy f_θ is parameterized by θ, then

$$\nabla_\theta V^{f_\theta}(x) = \sum_x \left(\sum_{k\geq 0} \mathbb{P}(x_k = x \,|\, x_0) \right) \sum_a \nabla_\theta f_\theta(x,a) V^{f_\theta}(x,a). \tag{3.2.22}$$

Therefore

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{x_0} \left[\sum_x \left(\sum_{k\geq 0} \mathbb{P}(x_k = x \,|\, x_0) \right) \sum_a \nabla_\theta f_\theta(x,a) V^{f_\theta}(x,a) \right] \rho_0(x_0) \\
&= \sum_{x,a,x_0} \left(\sum_{k\geq 0} f_\theta(x,a) \mathbb{P}(x_k = x \,|\, x_0) \rho_0(x_0) \right) \frac{1}{f_\theta(x,a)} \nabla_\theta f_\theta(x,a) V^{f_\theta}(x,a) \\
&= \sum_{x,a} \sigma_{\rho_0}^{f_\theta}(\{(x,a)\}) \nabla_\theta (\ln f_\theta(x,a)) V^{f_\theta}(x,a)
\end{aligned} \tag{3.2.23}$$

This is an expectation of the quantity $\nabla_\theta(\ln f_\theta(x,a)) V^{f_\theta}(x,a)$ with respect to the visiting measure $\sigma_{\rho_0}^{f_\theta}(\cdot)$. Since we do not have access to the visiting measure $\sigma_{\rho_0}^{f_\theta}(\cdot)$, we estimate this gradient as in [142, 149] by evaluating the quantity $\nabla_\theta(\ln f_\theta(\tilde{\xi}_k)) V^{f_\theta}(\tilde{\xi}_k)$, where $\tilde{\xi}_k := (\tilde{x}_k, \tilde{a}_k)$ is taken from the actor process $(\boldsymbol{\mathcal{M}}, \mathsf{Ac})$ as a sample from the visiting measure $\sigma_{\rho_0}^{f_k^N}(\cdot)$. The actor process $(\boldsymbol{\mathcal{M}}, \mathsf{Ac})$ evolves as follows for any $k \geq 1$:

$$\tilde{x}_{k+1} \sim \tilde{p}(\cdot \,|\, \tilde{\xi}_k) = \tilde{p}(\cdot \,|\, \tilde{x}_k, \tilde{a}_k), \quad \tilde{a}_{k+1} \sim g_k^N(\tilde{x}_{k+1}, \cdot). \tag{3.2.24}$$

The partial derivatives of the actor model $f_\theta^N = \mathrm{Softmax}(P_\theta^N)$ with respect to the parameters θ are:

$$\begin{aligned}
\frac{d}{dB^i} \ln f_\theta^N(x,a) &= \frac{d}{dB^i} \left(P_\theta^N(x,a) - \ln \left(\sum_{a'} \exp\left(P_\theta^N(x,a')\right) \right) \right) \\
&= \frac{d}{dB^i} \left(f_\theta^N(x,a) \right) - \frac{\sum_{a'} \frac{d}{dB^i} \exp\left(P_\theta^N(x,a')\right)}{\sum_{a''} \exp\left(P_\theta^N(x,a'')\right)} \\
&= \frac{d}{dB^i} \left(f_\theta^N(x,a) \right) - \frac{\sum_{a'} \exp\left(P_\theta^N(x,a')\right) \frac{d}{dB^i} P_\theta^N(x,a')}{\sum_{a''} \exp\left(P_\theta^N(x,a'')\right)} \\
&= \frac{1}{\sqrt{N}} \sigma(U^i \cdot (x,a)) - \sum_{a'} \left(\frac{\exp\left(P_\theta^N(x,a')\right)}{\sum_{a''} \exp\left(P_\theta^N(x,a'')\right)} \frac{1}{\sqrt{N}} \sigma(U^i \cdot (x,a')) \right) \\
&= \frac{1}{\sqrt{N}} \left(\sigma(U^i \cdot (x,a)) - \sum_{a'} f_\theta^N(x,a') \sigma(U^i \cdot (x,a')) \right),
\end{aligned} \tag{3.2.25}$$

and

$$\nabla_{U^i}(\ln f_\theta^N(x,a)) = \frac{1}{\sqrt{N}} \left(B^i \sigma'(U^i \cdot (x,a))(x,a) - \sum_{a'} f_\theta^N(x,a') B^i \sigma'(U^i \cdot (x,a'))(x,a') \right).$$

In our online actor-critic algorithm, we will replace the action-value function $V^{f_{\theta_k}}(x,a)$ by its estimate, i.e. the clipped critic $\mathsf{clip}(Q_k^N(\cdot,\cdot))$, where

$$\mathsf{clip}(x) = \max(\min(x, 2), 0). \tag{3.2.26}$$

The clipping is here to ensure that the magnitude of updates for parameters B_k^i and U_k^i are bounded. Clipping is a common technique used in practice in deep learning and is also necessary for our convergence analysis as $N \to \infty$.

Therefore, the actor network's parameters are updated according to:

$$B_{k+1}^i = B_k^i + \frac{\zeta_k^N}{N\sqrt{N}} \mathsf{clip}(Q_k^N(\tilde{\xi}_k)) \left(\sigma(U^i \cdot (\tilde{\xi}_k)) - \sum_{a''} f_k^N(\tilde{x}_k, a'') \sigma(U^i \cdot (\tilde{x}_k, a'')) \right),$$

$$U_{k+1}^i = U_k^i + \frac{\zeta_k^N}{N\sqrt{N}} \mathsf{clip}(Q_k^N(\tilde{\xi}_k)) \left(B_k^i \sigma'(U_k^i \cdot (\tilde{\xi}_k))(\tilde{\xi}_k) - \sum_{a''} f_k^N(\tilde{x}_k, a'') B_k^i \sigma'(U_k^i \cdot (\tilde{x}_k, \tilde{a}_k))(\tilde{x}_k, a'') \right), \tag{3.2.27}$$

where ζ_k^N/N is the learning rate.

The complete online Actor-Critic algorithm – for simultaneously training both the actor and critic networks – is summarised in Algorithm 1 below.

Algorithm 1 Online Actor-Critic Algorithm with Neural Network Approximation

procedure ONLINENAC($\boldsymbol{\mathcal{M}}, N, T, \nu_0, \mu_0$) ▷ Hyperparameters: MDP, network size, running time, initial distributions of critic and actor parameters.
  initialise neural network parameters: $\forall i, (C_0^i, W_0^i) \overset{\text{iid}}{\sim} \nu_0$ and $(B_0^i, W_0^i) \overset{\text{iid}}{\sim} \mu_0$.
  set $k = 0$
  initialise states/actions $\xi_0 = (x_0, a_0) \sim \rho_0$ and $\tilde{\xi}_0 = (\tilde{x}_0, \tilde{a}_0) \sim \rho_0$,
  while $k \leq NT$ **do**
    simulate $x_{k+1} \sim p(\cdot \,|\, \xi_k)$ and $\tilde{x}_{k+1} \sim p(\cdot \,|\, \xi_k)$
    simulate $a_{k+1} \sim g_k^N(x_{k+1}, \cdot)$ and $\tilde{a}_{k+1} \sim g_k^N(\tilde{x}_{k+1}, \cdot)$
    for all $i \in \{1, 2, ..., N\}$ **do**
      update (C_{k+1}^i, W_{k+1}^i) according to (3.2.21) using $\xi_k = (x_k, a_k)$, $\xi_{k+1} = (x_{k+1}, a_{k+1})$ and (C_k^i, W_k^i)
      update (B_{k+1}^i, U_{k+1}^i) according to (3.2.27) using $\tilde{\xi}_k = (\tilde{x}_k, \tilde{a}_k)$ and (B_k^i, U_k^i)
    end for
  end while
end procedure

The main contribution of this chapter is to prove that the evolution of the "actor" and "critic" networks trained with this online Actor-Critic algorithm weakly converge to the solution of a limiting ODE as $N \to \infty$. We then study the evolution of the limiting ODE to characterise the convergence of the online Actor-Critic algorithm. Specifically, we are able to prove that as training time $t \to \infty$ (A) the limit critic network converges to the true value function for the actor's policy and (B) the limit actor network converges to a stationary point of the objective function.

Assumption 3.2.10. In practical implementation, both the "actor" and "critic" networks should contain bias parameters, and should be written in the form

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} C^i \sigma(\mathsf{weight}^i \cdot (x, a) + \mathsf{bias}^i), \tag{3.2.28}$$

where $\mathsf{bias}^i \in \mathbb{R}$. The bias parameter could be incorporated into the weight vectors by introducing

an additional column of 1 in the state vector x, so that the networks could be expressed as

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} C^i \sigma(\widetilde{\mathsf{weight}}^i \cdot (x', a)), \quad x' = (x, 1). \tag{3.2.29}$$

We make this as an assumption of the MDP state space $\boldsymbol{\mathcal{X}}$. We further assume that the elements in $\mathcal{X} \times \mathcal{A}$ are in distinct directions (as defined on page 192 of [69]).

3.3 Main Result

Our results are proven under some assumptions for the neural networks, MDP and learning rates.

Assumption 3.3.1. For the actor network in (3.2.12) and critic network in (3.2.13), we assume:

- The randomly initialised parameters $\left(C_0^i, W_0^i, B_0^i, U_0^i\right)$ are independent and identically distributed (i.i.d.) mean-zero random variables for all i with distribution $\nu_0(dc, dw) \otimes \mu_0(db, dw)$, where $\otimes$ refers to the product of measures.
- ν_0 and μ_0 are absolutely continuous with respect to the Lebesgue measure,
- for each i, C_0^i, W_0^i, B_0^i, U_0^i are mutually independent, and
- $\max_i(|C_0^i|, |B_0^i|, \mathbb{E}_{\nu_0}\|W_0^i\|, \mathbb{E}_{\nu_0}\|U_0^i\|) \leq 1$ and $\mathbb{E}[C_0^i] = \mathbb{E}[B_0^i] = 0$.

We assume further that $\nu_0 = \mu_0$ for simplicity, although this additional assumption could be easily removed.

Our convergence proof also requires a careful choice for the learning rate and exploration rate.

Assumption 3.3.2. The learning rate and exploration rate are:

$$\begin{aligned} \zeta_k^N = \frac{1}{1+\frac{k}{N}}, \quad \eta_k^N = \frac{1}{1+\log^2(\frac{k}{N}+1)}, \\ \text{thus, as } N \to \infty, \quad \zeta_{\lfloor Nt \rfloor}^N \to \zeta_t = \frac{1}{1+t}, \quad \eta_{\lfloor Nt \rfloor}^N \to \eta_t = \frac{1}{1+\log^2(t+1)}. \end{aligned} \tag{3.3.1}$$

The learning rate and exploration rate in (3.3.1) satisfy the following properties for any integer $n \in \mathbb{N}$:

$$\int_0^\infty \zeta_s ds = \infty, \quad \int_0^\infty \zeta_t^2 dt < \infty, \quad \int_0^\infty \zeta_s \eta_s ds < \infty, \quad \lim_{t\to\infty} \frac{\zeta_t}{\eta_t^n} = 0. \tag{3.3.2}$$

We prove that the outputs of the actor and critic models converge to the solution of a nonlinear ODE system as the number of hidden units for the neural networks $N \to \infty$. We define the empirical measures

$$\mu_k^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{B_k^i, U_k^i}, \quad \nu_k^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{C_k^i, W_k^i}. \tag{3.3.3}$$

In addition, we define the following time-rescaled processes for any $\xi = (x, a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$

$$\begin{aligned} &P_t^N(\xi) = P_{\lfloor Nt \rfloor}^N(\xi), \quad f_t^N(\xi) = f_{\lfloor Nt \rfloor}^N(\xi), \quad g_t^N(\xi) = g_{\lfloor Nt \rfloor}^N(\xi), \\ &Q_t^N(\xi) = Q_{\lfloor Nt \rfloor}^N(\xi), \quad \mu_t^N = \mu_{\lfloor Nt \rfloor}^N, \quad \nu_t^N = \nu_{\lfloor Nt \rfloor}^N. \end{aligned} \tag{3.3.4}$$

Using Assumptions 3.2.9 and 3.3.1, we know that $\mu_0^N, \nu_0^N \xrightarrow{d} \nu_0$ and $P_0^N, Q_0^N \xrightarrow{d} \mathcal{G}, \mathcal{H}$ as $N \to \infty$, where $\mathcal{G}, \mathcal{H}$ are mean-zero Gaussian random variables by the law of large numbers and central limit theorem for i.i.d. random variables, respectively.

Define the state space for the time-rescaled process $\left(\mu_t^N, \nu_t^N, P_t^N, Q_t^N\right)$:

$$E = \mathcal{M}(\mathbb{R}^{1+d}) \times \mathcal{M}(\mathbb{R}^{1+d}) \times \mathbb{R}^M \times \mathbb{R}^M, \quad d = d_x + d_a, \quad M = |\mathcal{X} \times \mathcal{A}|, \tag{3.3.5}$$

where $\mathcal{M}(\mathbb{R}^{1+d})$ is the set of all probability measures on $\mathbb{R}^{1+d}$. Define the space

$$D_E([0,T]) = \{\text{càdlàg paths } f : [0,T] \to E\}. \tag{3.3.6}$$

We will study the convergence of the time-rescaled process (3.3.4) in the space $D_E([0,T])$ as $N \to \infty$.

The following definitions will also be used in our analysis:

- The inner-product of a measure ν and a function f is:

$$\langle f, \nu \rangle = \int f \, d\nu, \tag{3.3.7}$$

- The kernel matrix that will appears in our limit ODE in theorem 3.3.3 is:

$$A_{\xi,\xi'} = \left\langle \sigma\left(w \cdot \xi'\right) \sigma(w \cdot \xi) + c^2 \sigma'\left(w \cdot \xi'\right) \sigma'(w \cdot \xi)(\xi \cdot \xi'), \nu_0(dc, dw) \right\rangle, \tag{3.3.8}$$

  where $\xi' = (x', a')$.

The convergence of the online actor-critic algorithm is characterised by the following theorems:

Theorem 3.3.3. *Let Assumptions 3.2.9 and 3.3.1 hold, and let the learning rate for the critic parameter updates be $\alpha^N = \alpha/N$ for an $\alpha > 0$. Then, the process $\left(\mu_t^N, \nu_t^N, P_t^N, Q_t^N\right)$ converges weakly in the space $D_E([0,T])$ as $N \to \infty$ to the process (μ_t, ν_t, P_t, Q_t), so that for any $t \in [0,T]$, any $(x,a) \in \mathcal{X} \times \mathcal{A}$, and for every $\varphi, \bar{\varphi} \in C_b^2\left(\mathbb{R}^{1+d}\right)$, the limit process (μ_t, ν_t, P_t, Q_t) satisfies the random ODE:*

$$\begin{aligned}
\frac{dQ_t}{dt}(\xi) &= \alpha \sum_{\xi'=(x',a')} A_{\xi,\xi'} \left(r(\xi') + \gamma \sum_{z,a''} Q_t(z,a'') g_t(z,a'') p(z|\xi') - Q_t(\xi') \right) \pi^{g_t}(\xi'), \\
\frac{dP_t}{dt}(\xi) &= \sum_{\xi'=(x',a')} \zeta_t \mathsf{clip}(Q_t(\xi')) \left[A_{\xi,\xi'} - \sum_{a''} f_t(x',a'') A_{\xi,x',a''} \right] \sigma_{\rho_0}^{g_t}(\xi'), \\
P_0(\xi) &= \mathcal{G}(\xi), \quad Q_0(\xi) = \mathcal{H}(\xi) \\
\langle \varphi, \mu_t \rangle &= \langle \bar{\varphi}, \nu_0 \rangle, \quad \langle \varphi, \nu_t \rangle = \langle \varphi, \nu_0 \rangle,
\end{aligned} \tag{3.3.9}$$

where $\mathcal{G}, \mathcal{H}$ are the weak limits of P_0^N and Q_0^N, which are mean-zero Gaussian random variables, and

$$f_t(\xi) = \mathrm{Softmax}(P_t(\xi)), \quad g_t(\xi) = \frac{\eta_t}{\#\boldsymbol{\mathcal{A}}} + (1 - \eta_t) f_t(\xi).$$

We note the following property of the matrix A shown in the section 7 of [133]:

Lemma 3.3.4. *Under Assumptions 3.2.9 and 3.2.10, the matrix A is positive definite.*

Due to the matrix A being positive definite, we can prove that the limit critic network converges to the state-action value function and the limit actor network converges to a stationary point of the objective function:

Theorem 3.3.5. *If the actor network P_t and critic network Q_t evolved according to the limit ODE (3.3.9), then under assumptions 3.2.9 and 3.2.10, the critic network converges globally to the value function of the policy $f_t = \mathrm{Softmax}(P_t)$ as $t \to \infty$:*

$$\left\|Q_t - V^{f_t}\right\|_\infty = \max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_t(\xi) - V^{f_t}(\xi)| = O(\eta_t). \tag{3.3.10}$$

Moreover, the actor network converges to a stationary point:

$$\nabla_P J(f_t) \overset{t \to \infty}{\to} 0. \tag{3.3.11}$$

Remark 3.3.6. We note that the choice of norm/distance to study the pre-limit processes (P_t^N, Q_t^N) in Theorem 3.3.3 does not matter as $(P_t^N, Q_t^N) \in \mathbb{R}^{2M}$ is finite dimensional. The choice of norm for Theorem 3.3.5 does not matter for the same reason. We will use $\|\cdot\|_\infty$ as the supremum norm as defined in (3.3.10)

$$\|P - \tilde{P}\|_\infty = \max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |P(\xi) - \tilde{P}(\xi)|$$

and the usual Euclidean norm

$$\|P - \tilde{P}\| = \left(\sum_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |P(\xi) - \tilde{P}(\xi)| \right)^{1/2}$$

Note that the Softmax function is Lipschitz in the following sense: there exist constants $C, C' > 0$ such that for $P, \tilde{P} \in \mathbb{R}^M$,

$$d_{\mathsf{TV}}(\mathrm{Softmax}(P), \mathrm{Softmax}(\tilde{P})) \leq C' \|P - \tilde{P}\|_\infty. \tag{3.3.12}$$

3.4 Derivation of the limit ODEs

We use the following steps to prove convergence to the limit ODE system:

1. We first derive a pre-limit evolution process for the outputs of the actor and critic networks, and a-priori bounds on the magnitude of changes to the parameters at each update step. The pre-limit process will contain stochastic remainder terms with dependence on non-i.i.d. data samples.

2. We prove the relative compactness of the pre-limit process, which requires proof of the compact containment and regularity of the sample paths.

3. We then use the Poisson equation to prove the stochastic remainder terms in the pre-limit process vanish as $N \to +\infty$.

4. We prove the existence and uniqueness of the limits ODEs.

5. Finally, we combine the above results to prove the convergence in Theorem 3.3.3.

3.4.1 Evolution of the Pre-limit Processes

Before presenting the technical details of the proof, we first highlight some important details for the derivation of the limit ODE system of the neural actor-critic algorithm (algorithm 1).

Definition 3.4.1. For a random variable Z_N,

- $Z_N = O_p(\beta_N)$ if Z_N/β_N is *stochastically* bounded, i.e. for any $\epsilon > 0$, there exists $M < \infty$ and some $N_0 < \infty$ such that
$$\mathbb{P}\left(\left|\frac{Z_N}{\beta_N}\right| > M\right) < \epsilon, \quad \forall N > N_0.$$
- The notation $Z_N = O(\beta_N)$ means there exists a constant $C < \infty$ independent of N such that
$$|Z_N| \leq C|\beta_N|, \quad \forall N.$$

In the following proofs, constants C, C_T denote generic constants and we will sometimes we use $\xi, \xi_k, \xi'_k, \tilde{\xi}_k$ to denote the state-action pairs $(x,a), (x_k,a_k), (x'_k,a'_k), (\tilde{x}_k,\tilde{a}_k)$, respectively. For the learning rate $\alpha^N = 1/N$, the online actor-critic algorithm (algorithm 1) could therefore be written as:

$$\begin{aligned}
B^i_{k+1} &= B^i_k + \frac{\zeta^N_k}{N^{3/2}}\mathsf{clip}(Q^N_k(\tilde{\xi}_k))\left(\sigma(U^i\cdot(\tilde{\xi}_k)) - \sum_{a''} f^N_k(\tilde{x}_k,a'')\sigma(U^i\cdot(\tilde{x}_k,a''))\right),\\
U^i_{k+1} &= U^i_k + \frac{\zeta^N_k}{N^{3/2}}\mathsf{clip}(Q^N_k(\tilde{\xi}_k))\left(B^i_k\sigma'(U^i_k\cdot\tilde{\xi}_k)\tilde{\xi}_k - \sum_{a''} f^N_k(\tilde{x}_k,a'')B^i_k\sigma'(U^i_k\cdot\tilde{\xi}_k)(\tilde{x}_k,a'')\right)\\
C^i_{k+1} &= C^i_k + \frac{\alpha}{N^{3/2}}\left(r(\xi_k) + \gamma Q^N_k(\xi_{k+1}) - Q^N_k(\xi_k)\right)\sigma\left(W^i_k\cdot\xi_k\right),\\
W^i_{k+1} &= W^i_k + \frac{\alpha}{N^{3/2}}\left(r(\xi_k) + \gamma Q^N_k(\xi_{k+1}) - Q^N_k(\xi_k)\right)C^i_k\sigma'\left(W^i_k\cdot\xi_k\right)\xi_k.
\end{aligned} \tag{3.4.1}$$

The evolution of the actor and critic network Q^N_k can be studied by using Taylor expansions. For the critic network, one has:

$$\begin{aligned}
Q^N_{k+1}(\xi) &= Q^N_k(\xi) + \frac{1}{\sqrt{N}}\sum_{i=1}^N\left[(C^i_{k+1}-C^i_k)\sigma(W^i_{k+1}\cdot\xi) + (\sigma(W^i_{k+1}\cdot\xi)-\sigma(W^i_k\cdot\xi))C^i_k\right]\\
&= Q^N_k(\xi) + \frac{1}{\sqrt{N}}\sum_{i=1}^N\Bigg[(C^i_{k+1}-C^i_k)\left(\sigma(W^i_k\cdot\xi) + \sigma'(W^{i,*}_k\cdot\xi)(W^i_{k+1}-W^i_k)\cdot\xi\right)\\
&\quad + C^i_k\left(\sigma'(W^i_k\cdot\xi)(W^i_{k+1}-W^i_k)\cdot\xi + \frac{1}{2}\sigma''(W^{i,**}_k\cdot\xi)((W^i_{k+1}-W^i_k)\cdot\xi)^2\right)\Bigg],\\
&= Q^N_k(\xi) + \frac{1}{\sqrt{N}}\sum_{i=1}^N\Bigg[(C^i_{k+1}-C^i_k)\left(\sigma(W^i_k\cdot\xi) + C^i_k\sigma'(W^i_k\cdot\xi)(W^i_{k+1}-W^i_k)\cdot\xi\right) + \mathsf{error\ term},
\end{aligned} \tag{3.4.2}$$

where $W_k^{i,*}$ and $W_k^{i,**}$ are points in the line segment connecting the points W_k^i and W_{k+1}^i. Substituting the parameter updates (3.4.1), we have the following pre-limit evolution equation:

$$\begin{aligned} Q_{k+1}^N(\xi) &= Q_k^N(\xi) + \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[(C_{k+1}^i - C_k^i)\sigma(W_k^i \cdot \xi) + \sigma'(W_k^i \cdot \xi) C_k^i (W_{k+1}^i - W_k^i) \cdot \xi \right] + \text{error term} \\ &= Q_k^N(\xi) + \frac{\alpha}{N^2} \left[r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k) \right] \\ &\quad \times \sum_{i=1}^N \left(\sigma(W_k^i \cdot \xi_k)\sigma(W_k^i \cdot \xi) + (C_k^i)^2 \sigma'(W_k^i \cdot \xi)\sigma(W_k^i \cdot \xi_k)(\xi \cdot \xi_k) \right) + \text{error term}. \end{aligned} \tag{3.4.3}$$

If we let

$$\begin{aligned} \mathsf{B}_{\xi,\xi',k}^N &= \frac{1}{N} \sum_{i=1}^N \left[\sigma(W_k^i \cdot \xi')\sigma(W_k^i \cdot \xi) + (C_k^i)^2 \sigma'(W_k^i \cdot \xi')\sigma'(W_k^i \cdot \xi)(\xi' \cdot \xi) \right] \\ &= \langle \sigma(w \cdot \xi')\sigma(w \cdot \xi) + c^2 \sigma'(w \cdot \xi')\sigma(w \cdot \xi)(\xi' \cdot \xi), \nu_k^N \rangle, \end{aligned} \tag{3.4.4}$$

then

$$Q_{k+1}^N(\xi) = Q_k^N(\xi) + \frac{\alpha}{N} \left[r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k) \right] \mathsf{B}_{\xi,\xi_k,k}^N + \text{error term}. \tag{3.4.5}$$

For the actor network, one has

$$\begin{aligned} P_{k+1}^N(\xi) &= P_k^N(\xi) + \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[(B_{k+1}^i - B_k^i)\sigma(U_{k+1}^i \cdot \xi) + (\sigma(U_{k+1}^i \cdot \xi) - \sigma(U_k^i \cdot \xi)) B_k^i \right] \\ &= P_k^N(\xi) + \frac{1}{\sqrt{N}} \sum_{i=1}^N \Bigg[(B_{k+1}^i - B_k^i) \left(\sigma(U_k^i \cdot \xi) + \sigma'(U_k^{i,*} \cdot \xi)(U_{k+1}^i - U_k^i) \cdot \xi \right) \\ &\quad + B_k^i \left(\sigma'(U_k^i \cdot \xi)(U_{k+1}^i - U_k^i) \cdot \xi + \frac{1}{2} \sigma''(U_k^{i,**} \cdot \xi)((U_{k+1}^i - U_k^i) \cdot \xi)^2 \right) \Bigg] \\ &= P_k^N(\xi) + \sum_{i=1}^N \frac{\zeta_k^N}{N^2} \mathsf{clip}(Q_k^N(\tilde{\xi}_k)) \Bigg[\sigma(U_k^i \cdot \xi) \left(\sigma(U_k^i \cdot \tilde{\xi}_k) - \sum_{a''} f_k^N(\tilde{x}_k, a'')\sigma(U_k^i \cdot (\tilde{x}_k, a'')) \right) \\ &\quad + (B_k^i)^2 \sigma'(U_k^i \cdot \xi) \left(\sigma'(U_k^i \cdot \tilde{\xi}_k) \xi^\top \tilde{\xi}_k - \sum_{a''} f_k^N(\tilde{x}_k, a'')\sigma'(U_k^i \cdot (\tilde{x}_k, a''))((\tilde{x}_k, a'') \cdot \xi) \right) \Bigg] \\ &\quad + \text{error term}, \end{aligned} \tag{3.4.6}$$

where $U_k^{i,*}$ and $U_k^{i,**}$ are points in the line segment connecting the points U_k^i and U_{k+1}^i. We define the tensor

$$\begin{aligned} \bar{\mathsf{B}}_{\xi,\xi',k}^N &= \frac{1}{N} \sum_{i=1}^N \left[\sigma(U_k^i \cdot \xi')\sigma(U_k^i \cdot \xi) + (B_k^i)^2 \sigma'(U_k^i \cdot \xi')\sigma'(U_k^i \cdot \xi)(\xi' \cdot \xi) \right] \\ &= \langle \sigma(u \cdot \xi')\sigma(u \cdot \xi) + b^2 \sigma'(u \cdot \xi')\sigma'(u \cdot \xi)(\xi' \cdot \xi), \mu_k^N \rangle. \end{aligned} \tag{3.4.7}$$

Then,

$$P_{k+1}^N(\xi) = P_k^N(\xi) + \frac{\zeta_k^N}{N} \mathsf{clip}(Q_k^N(\tilde{\xi}_k)) \left[\bar{\mathsf{B}}_{\xi,\tilde{\xi}_k,k}^N - \sum_{a''} f_k^N(\tilde{x}_k, a'') \bar{\mathsf{B}}_{\xi,(\tilde{x}_k,a''),k}^N \right] + \text{error term}. \tag{3.4.8}$$

There are several error terms in the above evolution equations, which we will precisely define and analyse in the next section of this chapter. Specifically, we will:

- prove that the increments of the parameters at each update step are bounded,

- prove a-priori L^2 bounds for the outputs of the actor and critic networks,
- analyse the size of the error terms in the pre-limit evolution equation,
- rewrite the pre-limit evolution in terms of fluctuation terms, and
- study the evolution of the empirical measure of the parameters.

3.4.1.1 Bounds for the increments of the parameters

Lemma 3.4.2 (A-priori bounds of size of increments of parameters)**.** *For any fixed $T > 0$, any k such that $k \leq TN$ and $i \in [N] = \{1, ..., N\}$, there exists a constant $C_T < \infty$ that only depends on T such that*

$$\max\left(\left|C_k^i\right|, \mathbb{E}\|W_k^i\|, |B_k^i|, \mathbb{E}\|U_k^i\|\right) < C_T, \tag{3.4.9}$$

and that

$$\max\left(\left|C_{k+1}^i - C_k^i\right|, \left\|W_{k+1}^i - W_k^i\right\|\right) \leq \frac{C_T}{N}. \tag{3.4.10}$$

Moreover,

$$\max\left(\left|B_{k+1}^i - B_k^i\right|, \left\|U_{k+1}^i - U_k^i\right\|\right) < \frac{C_T}{N^{3/2}} \tag{3.4.11}$$

Proof. As σ is bounded by 1 by assumption 3.2.9, we have

$$\max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_k^N(\xi)| = \max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N C_k^i \sigma(W_k^i \cdot \xi) \right| \leq \frac{1}{\sqrt{N}} \sum_{i=1}^N |C_k^i| \tag{3.4.12}$$

We may then obtain a recursive bound for $|C_k^i|$:

$$\begin{aligned}
\left|C_{k+1}^i - C_k^i\right| &\leq \frac{\alpha^N}{\sqrt{N}} \left|r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right| \cdot \left|\sigma(W_k^i \cdot \xi_k)\right| \\
&\leq \frac{\alpha}{N^{3/2}} \left(|r(\xi_k)| + (1+\gamma) \max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} \left|Q_k^N(\xi)\right| \right) \cdot \left|\sigma(W_k^i \cdot \xi_k)\right| \\
&\leq \frac{\alpha}{N^{3/2}} \left(1 + (\gamma+1) \frac{1}{\sqrt{N}} \sum_{i=1}^N |C_k^i| \right) \\
&= \frac{\alpha}{N^{3/2}} + \frac{\alpha}{N^2} \sum_{i=1}^N |C_k^i|.
\end{aligned} \tag{3.4.13}$$

By recursively using the triangle inequality, and recalling that C_0^i is a bounded random variable, we have

$$\begin{aligned}
|C_k^i| \leq |C_0^i| + \sum_{j=1}^k (|C_j^i - C_{j-1}^i|) &\leq 1 + \sum_{j=1}^k \left(\frac{\alpha}{N^{3/2}} + \frac{\alpha}{N^2} \sum_{i=1}^N |C_{j-1}^i| \right) \\
&= 1 + \frac{\alpha}{N^{1/2}} + \frac{\alpha}{N^2} \sum_{j=1}^k \sum_{i=1}^N |C_{j-1}^i|.
\end{aligned} \tag{3.4.14}$$

Define

$$m_k^N = \frac{1}{N} \sum_{i=1}^N |C_k^i|. \tag{3.4.15}$$

Then

$$m_k^N \leq \frac{1}{N}\sum_{i=1}^{N}\left(1+\frac{\alpha}{N^{1/2}}+\frac{\alpha}{N^2}\sum_{j=1}^{k}\sum_{l=1}^{N}|C_{j-1}^l|\right) \leq C+\frac{\alpha}{N}\sum_{j=1}^{k}m_{j-1}^N. \tag{3.4.16}$$

By the discrete Gronwall's lemma and using $k \leq TN$,

$$m_k^N \leq C\exp\left(\frac{\alpha k}{N}\right) \leq C_T$$

Plugging this into (3.4.14) yields

$$\left|C_k^i\right| \leq \left|C_0^i\right| + \frac{C}{N^{1/2}} + \frac{C}{N}\sum_{j=1}^{k} m_{j-1}^N \leq \left|C_0^i\right| + \frac{C}{N^{1/2}} + C_T \leq C_T, \tag{3.4.17}$$

We could bootstrap with this a-priori bound to show that

$$|C_{k+1}^i - C_k^i| \leq \frac{C}{N^{3/2}} + N \times \frac{C}{N^2} \times C_T \leq \frac{C_T}{N}. \tag{3.4.18}$$

We can similarly get the bound for $\left\|W_k^i\right\|$. In fact,

$$\begin{aligned}\left\|W_{k+1}^i - W_k^i\right\| &\leq \frac{\alpha^N}{\sqrt{N}}\left|r(\xi_k)+\gamma Q_k^N(\xi_{k+1})-Q_k^N(\xi_k)\right|\cdot\left|C_k^i\sigma'\left(W_k^i\cdot\xi_k\right)\right|\left\|\xi_k\right\| \\ &\leq \frac{C_T}{N^{\frac{3}{2}}}\left(C+(\gamma+1)N^{-\frac{1}{2}}\sum_{i=1}^{N}|C_k^i|\right) \overset{(3.4.17)}{\leq} \frac{C_T}{N},\end{aligned} \tag{3.4.19}$$

Taking expectation and using assumptions 3.2.9 and 3.3.1 yields

$$\mathbb{E}\left\|W_k^i\right\| \leq \mathbb{E}\left\|W_0^i\right\| + \sum_{j=0}^{k-1}\mathbb{E}\left\|W_{k+1}^i - W_k^i\right\| \leq C_T. \tag{3.4.20}$$

For the boundedness of parameters in the actor network, observe that

$$|B_{k+1}^i - B_k^i| \leq \zeta_k^N N^{-\frac{3}{2}}|\mathsf{clip}(Q_k^N(\tilde{\xi}_k))|\cdot\sup_{a''}|\sigma(\tilde{x}_k,a'')|\cdot\left(1+\sum_{a''}f_k^N(\tilde{x}_k,a'')\right) < \frac{C}{N^{3/2}} \tag{3.4.21}$$

then by telescoping series, we have for all $k \leq NT$

$$|B_k^i| \leq |B_0^i| + C\frac{k}{N^{\frac{3}{2}}} \leq C + C\frac{T}{N^{\frac{1}{2}}} \leq C_T. \tag{3.4.22}$$

As the state-action space is finite, we also have

$$\left\|U_{k+1}^i - U_k^i\right\| \leq \zeta_k^N N^{-\frac{3}{2}}\left|\mathsf{clip}(Q_k^N(\tilde{\xi}_k))\right||B_k^i|\left(1+\sum_{a''}f_k^N(\tilde{x}_k,a'')\right)\cdot\sup_{\xi\in\mathcal{X}\times\mathcal{A}}\|\xi\| \leq \frac{C_T}{N^{3/2}}, \tag{3.4.23}$$

which yields

$$\mathbb{E}\left\|U_k^i\right\| \leq \mathbb{E}\left\|U_0^i\right\| + C_T\frac{k}{N^{\frac{3}{2}}} \leq C + \frac{C_T}{N^{\frac{1}{2}}} \leq C_T, \quad \forall k \leq TN. \tag{3.4.24}$$

□

Lemma 3.4.3 (Increments of entries in the pre-limit kernels)**.** *For all $k \leq NT$,*

$$\max_{\xi,\xi'\in\mathcal{X}\times\mathcal{A}}\max\left[\left|\mathsf{B}_{\xi,\xi',k+1}^N - \mathsf{B}_{\xi,\xi',k}^N\right|, \left|\bar{\mathsf{B}}_{\xi,\xi',k+1}^N - \bar{\mathsf{B}}_{\xi,\xi',k}^N\right|\right] \leq \frac{C_T}{N}, \tag{3.4.25}$$

where the kernels $\mathsf{B}_{\xi,\xi',k}^N$, $\bar{\mathsf{B}}_{\xi,\xi',k}^N$ are defined in (3.4.4) and (3.4.7) respectively. Consequently, one could show by method of telescoping series that for all $k \leq NT$

$$\max_{\xi,\xi'\in\mathcal{X}\times\mathcal{A}}\max\left[\left|\mathsf{B}_{\xi,\xi',k}^N\right|, \left|\bar{\mathsf{B}}_{\xi,\xi',k}^N\right|\right] \leq C_T, \tag{3.4.26}$$

Proof. The proof for the case of kernel $\bar{\mathsf{B}}^N$ is exactly the same with the proof for the case of B^N, for which we could utilize our a-priori bound of increments $\max(|C^i_{k+1} - C^i_k|, \left\|W^i_{k+1} - W^i_k\right\|) \leq C_T/N$ to prove our result. To the end, for all $\xi, \xi' \in \mathcal{X} \times \mathcal{A}$, we have

$$\begin{aligned}
&\left|\left\langle \sigma(w \cdot \xi')\sigma(w \cdot \xi),\, \nu^N_{k+1} - \nu^N_k \right\rangle\right| \\
&\leq \frac{1}{N}\sum_{i=1}^N \left|\sigma(W^i_{k+1} \cdot \xi')\sigma(W^i_{k+1} \cdot \xi) - \sigma(W^i_k \cdot \xi')\sigma(W^i_k \cdot \xi)\right| \\
&\leq \frac{1}{N}\sum_{i=1}^N \left[\left|\sigma(W^i_{k+1} \cdot \xi') - \sigma(W^i_k \cdot \xi')\right| \left|\sigma(W^i_{k+1} \cdot \xi)\right| + \left|\sigma(W^i_{k+1} \cdot \xi) - \sigma(W^i_k \cdot \xi)\right| \left|\sigma(W^i_k \cdot \xi')\right|\right] \\
&\leq \frac{1}{N}\sum_{i=1}^N (|\xi'| + |\xi|) \left\|W^i_{k+1} - W^i_k\right\| \leq \frac{C_T}{N}.
\end{aligned} \tag{3.4.27}$$

Similarly,

$$\begin{aligned}
&\left|\left\langle c^2\sigma'(w \cdot \xi')\sigma'(w \cdot \xi),\, \nu^N_{k+1} - \nu^N_k \right\rangle\right| \\
&\leq \frac{1}{N}\sum_{i=1}^N \left|(C^i_{k+1})^2\sigma(W^i_{k+1} \cdot \xi')\sigma(W^i_{k+1} \cdot \xi) - (C^i_k)^2\sigma(W^i_k \cdot \xi')\sigma(W^i_k \cdot \xi)\right| \\
&\leq \frac{1}{N}\sum_{i=1}^N \Big[\left|(C^i_{k+1})^2 - (C^i_k)^2\right| \left|\sigma(W^i_{k+1} \cdot \xi')\right| \left|\sigma(W^i_{k+1} \cdot \xi)\right| \\
&\quad + (C^i_k)^2 \left|\sigma(W^i_{k+1} \cdot \xi') - \sigma(W^i_k \cdot \xi')\right| \left|\sigma(W^i_{k+1} \cdot \xi)\right| + (C^i_k)^2 \left|\sigma(W^i_{k+1} \cdot \xi) - \sigma(W^i_k \cdot \xi)\right| \left|\sigma(W^i_k \cdot \xi')\right| \Big]
\end{aligned} \tag{3.4.28}$$

We have the control

$$\left|(C^i_{k+1})^2 - (C^i_k)^2\right| \leq \left|C^i_{k+1} - C^i_k\right|^2 + 2|C^i_k||C^i_{k+1} - C^i_k| \leq \frac{C_T^2}{N^2} + \frac{2C_T^2}{N} \leq \frac{C_T}{N}. \tag{3.4.29}$$

By combining this with our previous analyses, we have

$$\left|\left\langle c^2\sigma'(w \cdot \xi')\sigma'(w \cdot \xi),\, \nu^N_{k+1} - \nu^N_k \right\rangle\right| \leq \frac{1}{N}\sum_{i=1}^N \left[\frac{C_T}{N} + C_T \times \frac{C_T}{N} + C_T \times \frac{C_T}{N}\right] = \frac{C_T}{N}. \tag{3.4.30}$$

Summing up (3.4.27) and (3.4.30) yields $|\mathsf{B}^N_{\xi,\xi',k+1} - \mathsf{B}^N_{\xi,\xi',k}| \leq C_T/N$, uniformly in ξ, ξ'. It remains for us to show that there is a $C > 0$, independent of T, such that $|B^N_{\xi,\xi',0}| \leq C$. This is clearly true by the sure boundedness of $\sigma(\cdot), \sigma'(\cdot)$ and C^i_0 as guaranteed in assumption 3.2.9 and 3.3.1. Therefore, we could consider the telescoping sum

$$\left|\mathsf{B}^N_{\xi,\xi',k}\right| \leq \left|\mathsf{B}^N_{\xi,\xi',0}\right| + \sum_{j=0}^{k-1} \left|\mathsf{B}^N_{\xi,\xi',j+1} - \mathsf{B}^N_{\xi,\xi',j}\right| \leq C + N \times \frac{C_T}{N} \leq C_T, \tag{3.4.31}$$

which completes our proof. □

3.4.1.2 L^2 bounds of network outputs

Using lemma 3.4.2 and 3.4.3, we can now establish the bound for the neural networks.

Lemma 3.4.4 (A-priori L^2 bound for the outputs of the critic network)**.** *For all k such that $k \leq TN$,*

there is a $C_T < \infty$ such that

$$\mathbb{E}\left[\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} |Q_k^N(x,a)|^2\right] < C_T. \tag{3.4.32}$$

Proof. We first prove the statement for $k = 0$. Since C_0^i and $\sigma(W_0^i \cdot \xi)$ are both bounded by 1, we have

$$\begin{aligned}\mathbb{E}\left[\max_{\xi\in\mathcal{X}\times\mathcal{A}} |Q_0^N(\xi)|^2\right] \le \mathbb{E}\left[\sum_{\xi\in\mathcal{X}\times\mathcal{A}} |Q_0^N(\xi)|^2\right] &\le \sum_{\xi\in\mathcal{X}\in\mathcal{A}} \mathbb{E}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^N C_0^i\sigma(W_0^i\cdot\xi)\right]^2 \\ &\le \frac{C}{N}\sum_{i=1}^N \mathbb{E}\left[C_0^i\sigma(W_0^i\cdot\xi)\right]^2 \le C < \infty, \end{aligned} \tag{3.4.33}$$

We now provide an L^2 control over the maximum increments of the outputs $Q_k^N(\xi)$. Recall that

$$Q_{k+1}^N(\xi) - Q_k^N(\xi) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \left[(C_{k+1}^i - C_k^i)\sigma(W_{k+1}^i\cdot\xi) + C_k^i(\sigma(W_{k+1}^i\cdot\xi) - \sigma(W_{k+1}^i\cdot\xi))\right], \tag{3.4.34}$$

so

$$\begin{aligned}&\left|Q_{k+1}^N(\xi) - Q_k^N(\xi)\right|^2 \\ \overset{(\mathsf{CS})}{\le}& \frac{2}{N}\left[\left(\sum_{i=1}^N (C_{k+1}^i - C_k^i)\sigma(W_{k+1}^i\cdot\xi)\right)^2 + \left(\sum_{i=1}^N C_k^i(\sigma(W_{k+1}^i\cdot\xi) - \sigma(W_k^i\cdot\xi))\right)^2\right] \\ \overset{(\mathsf{CS})}{\le}& \frac{2}{N}\left[\sum_{i=1}^N (C_{k+1}^i - C_k^i)^2 \sum_{i=1}^N \left(\sigma(W_{k+1}^i\cdot\xi)\right)^2 + \sum_{i=1}^N (C_k^i)^2 \sum_{i=1}^N \left(\sigma(W_{k+1}^i\cdot\xi) - \sigma(W_k^i\cdot\xi)\right)^2\right] \\ \le& 2\left[\sum_{i=1}^N (C_{k+1}^i - C_k^i)^2 + C_T\sum_{i=1}^N \left(\sigma(W_{k+1}^i\cdot\xi) - \sigma(W_k^i\cdot\xi)\right)^2\right]. \end{aligned} \tag{3.4.35}$$

Hence

$$\begin{aligned}\left|C_{k+1}^i - C_k^i\right|^2 &\le \frac{(\alpha^N)^2}{N}\left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right)^2\left(\sigma(W_k^i\cdot\xi_k)\right)^2 \\ &\overset{(\mathsf{CS})}{\le} \frac{3\alpha}{N^3}\left[(r(\xi_k))^2 + \gamma^2(Q_k^N(\xi_{k+1})))^2 + \left(Q_k^N(\xi_k)\right)^2\right] \\ &\le \frac{3\alpha}{N^3}\left(1 + (1+\gamma^2)\max_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_k^N(\xi)|^2\right) \end{aligned} \tag{3.4.36}$$

Making use of the mean-value inequality (and the fact that $|\sigma'| \le 1$ by assumption 3.2.9), one could show similarly

$$\begin{aligned}&|\sigma(W_{k+1}^i\cdot\xi) - \sigma(W_k^i\cdot\xi)|^2 \\ \le& |(W_{k+1}^i - W_k^i)\cdot\xi|^2 \\ \le& \frac{(\alpha^N)^2}{N}\left(r(\xi_k) + \gamma\sum_{a''} Q_k^N(x_{k+1},a'')g_k^N(x_{k+1},a'') - Q_k^N(\xi_k)\right)^2\left(\sigma(W_k^i\cdot\xi_k)\right)^2(C_k^i)^2(\xi_k\cdot\xi)^2 \\ \le& \frac{3\alpha C_T^2}{N^3}\left(1 + (1+\gamma^2)\max_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_k^N(\xi)|^2\right), \end{aligned} \tag{3.4.37}$$

noting that $(\xi_k\cdot\xi)^2$ is bounded by some constant C as ξ, ξ_k are elements from the finite set $\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}$. Substituting into (3.4.35) yields

$$\left|Q_{k+1}^N(\xi) - Q_k^N(\xi)\right|^2 \le \frac{C_T}{N^2}\left(1 + (1+\gamma^2)\max_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} |Q_k^N(\xi)|^2\right). \tag{3.4.38}$$

Therefore for any ξ and $k \leq NT$,

$$
\begin{aligned}
|Q_k^N(\xi)|^2 &= \left(Q_0^N(\xi) + \sum_{j=0}^{N-1} (Q_{j+1}^N(\xi) - Q_j^N(\xi)) \right)^2 \\
&\overset{(\mathsf{CS})}{=} 2\left(Q_0^N(\xi)\right)^2 + 2\left(\sum_{j=0}^{k-1} (Q_{j+1}^N(\xi) - Q_j^N(\xi)) \right)^2 \\
&\overset{(\mathsf{CS})}{\leq} 2\left(Q_0^N(\xi)\right)^2 + NT \sum_{j=0}^{k-1} (Q_{j+1}^N(\xi) - Q_j^N(\xi))^2 \\
&\leq 2 \max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_0^N(\xi)|^2 + \frac{C_T}{N} \sum_{j=0}^{k-1} \left(1 + (1+\gamma^2) \max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_j^N(\xi)|^2 \right). \qquad (3.4.39)
\end{aligned}
$$

Taking maximum then expectation yields

$$
\begin{aligned}
\mathbb{E}\left[\max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_k^N(\xi)|^2 \right] &\leq 2\mathbb{E}\left[\max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_0^N(\xi)|^2 \right] + \frac{C_T}{N} + \frac{C_T}{N} \sum_{j=0}^{k-1} \mathbb{E}\left[\max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_j^N(\xi)|^2 \right] \\
&\leq C_T + \frac{C_T}{N} \sum_{j=0}^{k-1} \mathbb{E}\left[\max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_j^N(\xi)|^2 \right]. \qquad (3.4.40)
\end{aligned}
$$

We conclude by discrete Gronwall's lemma that for all $k \leq TN$:

$$
\mathbb{E}\left[\max_{\xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}} |Q_k^N(\xi)|^2 \right] \leq C_T \exp\left(C_T \frac{k}{N} \right) \leq C_T < +\infty. \qquad (3.4.41)
$$

□

Lemma 3.4.5 (A-priori L^2 bound for the outputs of the actor network)**.** *For all k such that $k \leq NT$, there is a $C_T < \infty$ such that*

$$
\mathbb{E}\left[\max_{(x,a) \in \mathcal{X} \times \mathcal{A}} \left| P_k^N(x,a) \right|^2 \right] < C_T. \qquad (3.4.42)
$$

Proof. Again we first prove the statement for $k = 0$. Since B_0^i and $\sigma(W_0^i \cdot \xi)$ are bounded by 1,

$$
\begin{aligned}
\mathbb{E}\left[\max_{\xi \in \mathcal{X} \times \mathcal{A}} |P_0^N(\xi)|^2 \right] \leq \mathbb{E}\left[\sum_{\xi \in \mathcal{X} \times \mathcal{A}} |P_0^N(\xi)|^2 \right] &\leq \sum_{\xi \in \mathcal{X} \in \mathcal{A}} \mathbb{E}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^N B_0^i \sigma(U_0^i \cdot \xi) \right]^2 \\
&\leq \frac{C}{N} \sum_{i=1}^N \mathbb{E}\left[B_0^i \right]^2 \leq C < \infty. \qquad (3.4.43)
\end{aligned}
$$

The increments could again be controlled by noting

$$
\begin{aligned}
|P_{k+1}^N(\xi) - P_k^N(\xi)| &\leq \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[(B_{k+1}^i - B_k^i) \sigma(U_{k+1}^i \cdot \xi) + (\sigma(U_{k+1}^i \cdot \xi) - \sigma(U_k^i \cdot \xi)) B_k^i \right] \\
&\leq \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[\left| B_{k+1}^i - B_k^i \right| \left| \sigma(U_{k+1}^i \cdot \xi) \right| + \left| \sigma(U_{k+1}^i \cdot \xi) - \sigma(U_k^i \cdot \xi) \right| \left| B_k^i \right| \right]
\end{aligned}
$$

By the mean-value inequality and the fact that both σ and σ' are bounded by 1 by assumption 3.2.9,

$$
|P_{k+1}^N(\xi) - P_k^N(\xi)| \leq \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{C_T}{N^{3/2}} = \frac{C_T}{N^2}. \qquad (3.4.44)
$$

Therefore for all ξ,

$$|P_k^N(\xi)|^2 = \left(P_0^N(\xi) + \sum_{j=0}^{k-1}(P_{j+1}^N(\xi) - P_j^N(\xi))\right)^2 \leq 2\max_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}|P_0^N(\xi)|^2 + 2N\sum_{j=0}^{k-1}(P_{j+1}^N(\xi) - P_j^N(\xi))^2$$
$$\leq 2\max_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}|P_0^N(\xi)|^2 + \frac{C_T}{N^2}. \tag{3.4.45}$$

Taking supremum then expectation yields the result. □

3.4.1.3 Pre-limit evolution of the network outputs

We can now control the unspecified error terms in the pre-limit evolutions of the actor and critic networks.

Proposition 3.4.6 (Evolution of the actor and critic networks)**.** *For $k \leq NT$, the evolution of the critic network yields,*

$$\mathbb{E}\left[\max_{\xi}\left|Q_{k+1}^N(\xi) - Q_k^N(\xi) - \frac{\alpha}{N}\left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right)\mathsf{B}_{\xi,\xi_k,k}^N\right|\right] \leq \frac{C_T}{N^{5/2}},$$

while the evolution of the actor network yields

$$\max_{\xi}\left|P_{k+1}^N(\xi) - P_k^N(\xi) - \frac{\zeta_k^N}{N}\mathsf{clip}(Q_k^N(\tilde{\xi}_k))\left(\bar{\mathsf{B}}_{\xi,\tilde{\xi}_k,k}^N - \sum_{a''} f_k^N(\tilde{x}_k, a'')\bar{\mathsf{B}}_{\xi,(\tilde{x}_k,a''),k}^N\right)\right| \leq \frac{C_T}{N^{5/2}}.$$

Proof. We begin by noting for all ξ,

$$\left|Q_{k+1}^N(\xi) - Q_k^N(\xi) - \frac{\alpha}{N}\left(r(\xi_k) + \gamma\sum_{a''}Q_k^N(x_{k+1}, a'')g_k^N(x_{k+1}, a'') - Q_k^N(\xi_k)\right)\mathsf{B}_{\xi,\xi_k,k}^N\right|$$
$$= \frac{1}{\sqrt{N}}\left|\sum_{i=1}^N \sigma'(W_k^{i,*}\cdot\xi)(C_{k+1}^i - C_k^i)(W_{k+1}^i - W_k^i)\cdot\xi + \frac{\sigma''(W_k^{i,**}\cdot\xi)C_k^i}{2}((W_{k+1}^i - W_k^i)\cdot\xi)^2\right|$$
$$\overset{(\mathsf{CS})}{\leq} \frac{1}{\sqrt{N}}\sum_{i=1}^N\left[\left|C_{k+1}^i - C_k^i\right|\left\|W_{k+1}^i - W_k^i\right\|\|\xi\| + C_T\left\|W_{k+1}^i - W_k^i\right\|^2\|\xi\|^2\right]$$
$$\leq \frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{C_T}{N^3}\left(1 + (1+\gamma)\max_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}|Q_k^N(\xi)|\right)^2 \leq \frac{C_T}{N^{5/2}}\left(1 + (1+\gamma)^2\max_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}|Q_k^N(\xi)|^2\right) \tag{3.4.46}$$

Taking maximum and expectation yields

$$\mathbb{E}\left[\max_{\xi}\left|Q_{k+1}^N(\xi) - Q_k^N(\xi) - \frac{\alpha}{N}\left(r(\xi_k) + \gamma\sum_{a''}Q_k^N(x_{k+1}, a'')g_k^N(x_{k+1}, a'') - Q_k^N(\xi_k)\right)\mathsf{B}_{\xi,\xi_k,k}^N\right|\right]$$
$$\leq \frac{C_T}{N^{5/2}}\mathbb{E}\left[1 + (1+\gamma)^2\max_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}|Q_k^N(\xi)|^2\right] \leq \frac{C_T}{N^{5/2}}. \tag{3.4.47}$$

Similarly, for all ξ,

$$\left|P_{k+1}^N(\xi) - P_k^N(\xi) - \frac{\zeta_k^N}{N}\mathsf{clip}(Q_k^N(\tilde{\xi}_k))\left(\bar{\mathsf{B}}_{\xi,\tilde{\xi}_k,k}^N - \sum_{a''} f_k^N(\tilde{x}_k, a'')\bar{\mathsf{B}}_{\xi,(\tilde{x}_k,a''),k}^N\right)\right|$$
$$= \frac{1}{\sqrt{N}}\left|\sum_{i=1}^N \sigma'(U_k^{i,*}\cdot\xi)(B_{k+1}^i - B_k^i)(U_{k+1}^i - U_k^i)\cdot\xi + \frac{\sigma''(U_k^{i,**}\cdot\xi)C_k^i}{2}((U_{k+1}^i - U_k^i)\cdot\xi)^2\right|$$
$$\overset{(\mathsf{CS})}{\leq} \frac{1}{\sqrt{N}}\sum_{i=1}^N\left[\left|B_{k+1}^i - B_k^i\right|\left\|U_{k+1}^i - U_k^i\right\|\|\xi\| + C_T\left\|U_{k+1}^i - U_k^i\right\|^2\|\xi\|^2\right] \leq \frac{C_T}{N^{5/2}}.$$

This completes the proof. □

Using the notation introduced in definition 3.4.1, one could write

$$Q^N_{k+1}(\xi) = Q^N_k(\xi) + \frac{\alpha}{N}\left[r(\xi_k) + \gamma \sum_{a''} Q^N_k(x_{k+1}, a'') g^N_k(x_{k+1}, a'') - Q^N_k(\xi_k)\right] \mathsf{B}^N_{\xi,\xi_k,k} + O_p(N^{-5/2}). \tag{3.4.48}$$

$$P^N_{k+1}(\xi) = P^N_k(\xi) + \frac{\zeta^N_k}{N}\mathsf{clip}(Q^N_k(\tilde{\xi}_k))\left[\bar{\mathsf{B}}^N_{\xi,\tilde{\xi}_k,k} - \sum_{a''} f^N_k(\tilde{x}_k, a'')\bar{\mathsf{B}}^N_{\xi,(\tilde{x}_k,a''),k}\right] + O(N^{-5/2}). \tag{3.4.49}$$

Network evolution We recall that $P^N_t(\xi) = P^N_{\lfloor Nt\rfloor}$, $f^N_t(\xi) = f^N_{\lfloor Nt\rfloor}(\xi)$, $g^N_t(\xi) = g^N_{\lfloor Nt\rfloor}(\xi)$, $Q^N_t(\xi) = Q^N_{\lfloor Nt\rfloor}$, and define $\mathsf{B}^N_{\xi,\xi',s} = \mathsf{B}^N_{\xi,\xi',\lfloor Ns\rfloor}$ and $\bar{\mathsf{B}}^N_{\xi,\xi',s} = \bar{\mathsf{B}}^N_{\xi,\xi',\lfloor Ns\rfloor}$. We further define the fluctuation terms:

$$\begin{aligned}
M^{1,N}_t(\xi) &= -\frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} Q^N_k(\xi_k)\mathsf{B}^N_{\xi,\xi_k,k} + \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{\xi'} Q^N_k(\xi')\mathsf{B}^N_{\xi,\xi',k}\pi^{g^N_k}(\xi'),\\
M^{2,N}_t(\xi) &= \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} r(\xi_k)\mathsf{B}^N_{\xi,\xi_k,k} - \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{\xi'} r(\xi')\mathsf{B}^N_{\xi,\xi',k}\pi^{g^N_k}(\xi'),\\
M^{3,N}_t(\xi) &= \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1} \gamma Q^N_k(\xi_{k+1})\mathsf{B}^N_{\xi,\xi_k,k} - \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{\xi'}\sum_{z,a''} \gamma Q^N_k(z,a'')g^N_k(z,a'')\mathsf{B}^N_{\xi,\xi',k}p(z|\xi')\pi^{g^N_k}(\xi'),
\end{aligned} \tag{3.4.50}$$

then

$$\begin{aligned}
&Q^N_t(\xi) = Q^N_0(\xi) + \sum_{k=0}^{\lfloor Nt\rfloor-1}[Q^N_{k+1}(\xi) - Q^N_k(\xi)]\\
=&Q^N_0(\xi) + \frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\left[r(\xi_k) + \gamma Q^N_k(\xi_{k+1}) - Q^N_k(\xi_k)\right]\mathsf{B}^N_{\xi,\xi_k,k} + O_p(N^{-3/2})\\
=&Q^N_0(\xi) + \frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\sum_{\xi'}\mathsf{B}^N_{\xi,\xi',k}\pi^{g^N_k}(\xi')\Big(-Q^N_k(\xi') + r(\xi') + \gamma\sum_{z,a''} Q^N_k(z,a'')g^N_k(z,a'')p(z|\xi')\Big)\\
&+\alpha\left(M^{1,N}_t(\xi) + M^{2,N}_t(\xi) + M^{3,N}_t(\xi)\right) + O_p(N^{-3/2})\\
=&Q^N_0(\xi) + \frac{\alpha}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{k/N}^{k+1/N}\sum_{\xi'}\mathsf{B}^N_{\xi,\xi',\lfloor Ns\rfloor}\pi^{g^N_{\lfloor Ns\rfloor}}(\xi')\Big(r(\xi') + \gamma\sum_{z,a''} Q^N_{\lfloor Ns\rfloor}(z,a'')g^N_{\lfloor Ns\rfloor}(z,a'')p(z|\xi')\\
&-Q^N_{\lfloor Ns\rfloor}(\xi')\Big)ds + \alpha\left(M^{1,N}_t(\xi) + M^{2,N}_t(\xi) + M^{3,N}_t(\xi)\right) + O_p(N^{-3/2})\\
=&Q^N_0(\xi) + \alpha\int_0^t\sum_{\xi'}\mathsf{B}^N_{\xi,\xi',s}\pi^{g^N_s}(\xi')\left[r(\xi') + \gamma\sum_{z,a''} Q^N_s(z,a'')g^N_s(z,a'')p(z|\xi') - Q^N_s(\xi')\right]ds\\
&+\alpha\left(M^{1,N}_t(\xi) + M^{2,N}_t(\xi) + M^{3,N}_t(\xi)\right) + O_p(N^{-3/2}).
\end{aligned} \tag{3.4.51}$$

Similarly, define the fluctuation terms

$$
\begin{aligned}
M_t^N(\xi) = &\frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\zeta_k^N \mathsf{clip}(Q_k^N(\tilde{\xi}_k))\left[\bar{\mathsf{B}}^N_{\xi,\tilde{\xi}_k,k} - \sum_{a''} f_k^N(\tilde{x}_k,a'')\bar{\mathsf{B}}^N_{\xi,(\tilde{x}_k,a''),k}\right] \\
&- \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\zeta_k^N \sum_{\xi'}\mathsf{clip}(Q_k^N(\xi'))\left[\bar{\mathsf{B}}^N_{\xi,\xi',k} - \sum_{a''} f_k^N(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),k}\right]\sigma_{\rho_0}^{g_k^N}(\xi'),
\end{aligned}
\tag{3.4.52}
$$

where $\sigma_{\rho_0}^{g_k^N}(\xi')$ is the visiting measure of the Markov chain as defined in (3.2.14). Then:

$$
\begin{aligned}
P_t^N(\xi) &= P_0^N(\xi) + \sum_{k=0}^{\lfloor Nt\rfloor-1}(P_{k+1}^N(\xi) - P_k^N(\xi)) \\
&= P_0^N(\xi) + \sum_{k=0}^{\lfloor Nt\rfloor-1}\frac{\zeta_k^N}{N}\mathsf{clip}(Q_k^N(\tilde{\xi}_k))\left[\bar{\mathsf{B}}_{\xi,\tilde{\xi}_k,k} - \sum_{a''} f_k^N(\tilde{x}_k,a'')\bar{\mathsf{B}}_{\xi,(\tilde{x}_k,a''),k}\right] + O(N^{-3/2}) \\
&= P_0^N(\xi) + \sum_{k=0}^{\lfloor Nt\rfloor-1}\zeta_k^N\sum_{\xi'}\mathsf{clip}(Q_k^N(\xi'))\left[\bar{\mathsf{B}}^N_{\xi,\xi',k} - \sum_{a''} f_k^N(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),k}\right]\sigma_{\rho_0}^{g_k^N}(\xi') \\
&\quad + \alpha M_t^N(x,a) + O(N^{-3/2}) \\
&= P_0^N(\xi) + \sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{k/N}^{(k+1)/N}\zeta^N_{\lfloor Ns\rfloor}\sum_{\xi'}\mathsf{clip}(Q^N_{\lfloor Ns\rfloor}(\xi'))\Bigg[\bar{\mathsf{B}}^N_{\xi,\xi',\lfloor Ns\rfloor} \\
&\quad - \sum_{a''} f^N_{\lfloor Ns\rfloor}(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),\lfloor Ns\rfloor}\Bigg]\sigma_{\rho_0}^{g^N_{\lfloor Ns\rfloor}}(\xi') + \alpha M_t^N(x,a) + O(N^{-3/2}) \\
&= P_0^N(\xi) + \int_0^t \zeta^N_{\lfloor Ns\rfloor}\sum_{\xi'}\sigma_{\rho_0}^{g_s^N}(\xi')\,\mathsf{clip}(Q_s^N(\xi'))\left[\bar{\mathsf{B}}^N_{\xi,\xi',s} - \sum_{a''} f_s^N(x',a'')\bar{\mathsf{B}}_{\xi,(x',a''),s}\right]ds \\
&\quad + M_t^N(\xi) + O(N^{-3/2}).
\end{aligned}
\tag{3.4.53}
$$

3.4.1.4 Evolution of empirical measure

The evolution of the empirical measure ν_k^N can be characterised in terms of their projection onto test functions $\varphi \in C_b^2(\mathbb{R}^{1+M})$, by Taylor's expansion

$$
\begin{aligned}
&\langle\varphi,\nu_{k+1}^N\rangle - \langle\varphi,\nu_k^N\rangle \\
=&\frac{1}{N}\sum_{i=1}^N(\varphi(C_{k+1}^i,W_{k+1}^i) - \varphi(C_k^i,W_k^i)) \\
=&\frac{1}{N}\sum_{i=1}^N\Big[\partial_c\varphi(C_k^i,W_k^i)(C_{k+1}^i - C_k^i) + \partial_w\varphi(C_k^i,W_k^i)\cdot(W_{k+1}^i - W_k^i) \\
&+ \frac{1}{2}\Big(\partial_c^2\varphi(C_k^{i,*},W_k^{i,*})(C_{k+1}^i - C_k^i)^2 + (C_{k+1}^i - C_k^i)\partial_{cw}^2\varphi(C_k^{i,**},W_k^{i,**})(W_{k+1}^i - W_k^i) \\
&+ (W_{k+1}^i - W_k^i)\cdot\partial_w^2\varphi(C_k^{i,***},W_k^{i,***})(W_{k+1}^i - W_k^i)\Big)\Big],
\end{aligned}
\tag{3.4.54}
$$

where $\left(C_k^{i,*}, W_k^{i,*}\right), \left(C_k^{i,**}, W_k^{i,**}\right), \left(C_k^{i,***}, W_k^{i,***}\right)$ are points lying on the line segments connecting between $\left(C_k^i, W_k^i\right)$ and $\left(C_{k+1}^i, W_{k+1}^i\right)$. Substituting (3.2.21) into (3.4.54), we have

$$\begin{aligned}
&\left\langle \varphi, \nu_{k+1}^N \right\rangle - \left\langle \varphi, \nu_k^N \right\rangle \\
=&\frac{1}{N}\sum_{i=1}^N \left[\partial_c \varphi(C_k^i, W_k^i)(C_{k+1}^i - C_k^i) + \partial_w \varphi(C_k^i, W_k^i)\cdot(W_{k+1}^i - W_k^i)\right] + O_p(N^{-2}) \\
=&\alpha N^{-\frac{5}{2}}\left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right) \\
&\times \sum_{i=1}^N \left(\partial_c \varphi(C_k^i, W_k^i)\sigma(W_k^i\cdot\xi_k)) + C_k^i \sigma'(W_k^i\cdot\xi_k)\partial_w\varphi(C_k^i, W_k^i)\xi_k\right) + O_p(N^{-2}) \\
=&\alpha N^{-\frac{3}{2}}\left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right) \\
&\times \left\langle \partial_c\varphi(c,w)\sigma(w\cdot\xi_k) + c\sigma'(w\cdot\xi_k)\partial_w\varphi(c,w)\xi_k, \nu_k^N\right\rangle + O_p(N^{-3}).
\end{aligned} \tag{3.4.55}$$

Therefore, the time-rescaled empirical measure $\nu_t^N := \nu_{\lfloor Nt\rfloor}^N$ satisfies

$$\begin{aligned}
\left\langle \varphi, \nu_t^N \right\rangle - \left\langle \varphi, \nu_0^N \right\rangle = \alpha N^{-\frac{3}{2}} \sum_{k=0}^{\lfloor Nt\rfloor - 1} &\left(r(\xi_k) + \gamma Q_k^N(\xi_{k+1}) - Q_k^N(\xi_k)\right) \\
&\times \left\langle \partial_c\varphi(c,w)\sigma(w\cdot\xi_k) + c\sigma'(w\cdot\xi_k)\partial_w\varphi(c,w)\xi_k, \nu_k^N\right\rangle + O_p(N^{-2}).
\end{aligned} \tag{3.4.56}$$

We can similarly characterise the evolution of the empirical measure μ_k^N in terms of their projection onto any test functions $\varphi \in C_b^2(\mathbb{R}^{1+M})$:

$$\begin{aligned}
&\left\langle \varphi, \mu_{k+1}^N \right\rangle - \left\langle \varphi, \mu_k^N \right\rangle \\
=&\frac{1}{N}\sum_{i=1}^N \left[\partial_b \varphi(B_k^i, U_k^i)(B_{k+1}^i - B_k^i) + \partial_u \varphi(B_k^i, U_k^i)\cdot(U_{k+1}^i - U_k^i)\right] + O_p(N^{-2}) \\
=&\frac{1}{N^{\frac{5}{2}}}\sum_{i=1}^N \zeta_k^N \mathsf{clip}(Q_k^N(\tilde{\xi}_k))\Bigg[\sigma(U_k^i\cdot\tilde{\xi}_k)(\partial_b\varphi(B_k^i, U_k^i) - B_k^i\partial_w\varphi(B_k^i, U_k^i)\cdot\xi_k) \\
&- \sum_{a''} f_k^N(\tilde{x}_k, a'')\sigma'(U_k^i\cdot(\tilde{x}_k, a''))\left(\partial_b\varphi(B_k^i, U_k^i) - B_k^i\partial_w\varphi(B_k^i, U_k^i)\cdot(\tilde{x}_k, a'')\right)\Bigg] + O_p(N^{-2}) \\
=&\frac{1}{N^{\frac{3}{2}}}\xi_k^N \mathsf{clip}(Q_k^N(\tilde{\xi}_k))\Big[\left\langle \sigma(u\cdot\tilde{\xi}_k)(\partial_b\varphi(b,u) - b\partial_w\varphi(b,u)\cdot\xi_k), \mu_k^N\right\rangle \\
&- \sum_{a''} f_k^N(\tilde{x}_k, a'')\left\langle \sigma'(u\cdot(\tilde{x}_k, a''))\left(\partial_b\varphi(b,u) - b\partial_w\varphi(b,u)\cdot(\tilde{x}_k, a'')\right), \mu_k^N\right\rangle\Big] + O_p(N^{-2}),
\end{aligned} \tag{3.4.57}$$

and hence

$$\left\langle \varphi, \mu_t^N \right\rangle - \left\langle \varphi, \mu_0^N \right\rangle \tag{3.4.58}$$

$$= \frac{1}{N^{\frac{3}{2}}}\sum_{k=0}^{\lfloor Nt\rfloor - 1} \xi_k^N \mathsf{clip}(Q_k^N(\tilde{\xi}_k))\Big[\left\langle \sigma(u\cdot\tilde{\xi}_k)(\partial_b\varphi(b,u) - b\partial_w\varphi(b,u)\cdot\xi_k), \mu_k^N\right\rangle \tag{3.4.59}$$

$$- \sum_{a''} f_k^N(\tilde{x}_k, a'')\left\langle \sigma'(u\cdot(\tilde{x}_k, a''))\left(\partial_b\varphi(b,u) - b\partial_w\varphi(b,u)\cdot(\tilde{x}_k, a'')\right), \mu_k^N\right\rangle\Big] + O_p(N^{-1}). \tag{3.4.60}$$

3.4.2 Relative Compactness

In this section, we prove the family of processes $\left(\mu_t^N, \nu_t^N, P_t^N, Q_t^N\right)$ are relatively compact under the choice of scaling of critic parameter updates $\alpha^N = 1/N$. Section 3.4.2.1 proves compact containment and Section 3.4.2.2 proves needed regularity. Section 3.4.2.3 combines these results to prove the relative compactness.

3.4.2.1 Compact Containment

The L^2 bounds for the actor and critic networks in Lemma 3.4.4 and 3.4.5 enable us to prove that the process $(\mu_t^N, \nu_t^N, P_t^N, Q_t^N)$ is compactly bounded. As a reminder, we now treat P_t^N, Q_t^N are vectors of size $d = |\mathcal{X} \times \mathcal{A}|$, thanks to the assumption of the state-action space being finite. Letting $E = \mathcal{M}(\mathbb{R}^{1+d}) \times \mathcal{M}(\mathbb{R}^{1+d}) \times \mathbb{R}^d \times \mathbb{R}^d$, we have

Lemma 3.4.7 (Compact Containment)**.** *For any $\eta > 0$, there is a compact subset $\mathcal{K}$ of E such that*

$$\sup_{N \in \mathbb{N}, 0 \le t \le T} \mathbb{P}\left[\left(\mu_t^N, \nu_t^N, P_t^N, Q_t^N\right) \notin \mathcal{K}\right] < \eta. \tag{3.4.61}$$

Proof. Let $K_L = [-L, L]^{1+d}$ denote a compact subset in $\mathbb{R}^{1+d}$. We then see that for any $t \ge 0$ and $N \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[\nu_t^N\left(\mathbb{R}^{1+d} \setminus K_L\right)\right] &= \frac{1}{N}\sum_{i=1}^N \mathbb{P}\left(\left(C^i_{\lfloor Nt \rfloor}, W^i_{\lfloor Nt \rfloor}\right) \in \mathbb{R}^{1+d} \setminus K_L\right) \\
&\le \frac{1}{N}\sum_{i=1}^N \mathbb{P}\left(\left|C^i_{\lfloor Nt \rfloor}\right| + \left\|W^i_{\lfloor Nt \rfloor}\right\| \ge L\right) \le \frac{C_T}{L},
\end{aligned} \tag{3.4.62}$$

where the final step is by $\left|C^i_{\lfloor Nt \rfloor}\right| + \left\|W^i_{\lfloor Nt \rfloor}\right\|$ is integrable (from Lemma 3.4.2) and Chebyshev's inequality. We define the following subset of $\mathcal{M}\left(\mathbb{R}^{1+d}\right)$

$$\hat{K}_L = \overline{\left\{\nu \in \mathcal{M}\left(\mathbb{R}^{1+d}\right) \,\middle|\, \nu\left(\mathbb{R}^{1+d} \setminus K_{(L+j)^2}\right) < \frac{1}{\sqrt{L+j}} \text{ for all } j\right\}}, \tag{3.4.63}$$

which is a closure of a tight family of measures and thus being a compact subset of $\mathcal{M}(\mathbb{R}^{1+d})$. Observe that

$$\begin{aligned}
\mathbb{P}\left(\nu_t^N \notin \hat{K}_L\right) &\le \mathbb{P}\left(\exists j \text{ s.t. } \nu_t^N(\mathbb{R}^{1+d} \setminus K_{(L+j)^2}) > \frac{1}{\sqrt{L+j}}\right) \\
&\le \sum_{j=1}^\infty \mathbb{P}\left(\nu_t^N(\mathbb{R}^{1+d} \setminus K_{(L+j)^2}) > \frac{1}{\sqrt{L+j}}\right) \\
&\overset{(a)}{\le} \sum_{j=1}^\infty \frac{\mathbb{E}\left[\nu_t^N(\mathbb{R}^{1+d} \setminus K_{(L+j)^2})\right]}{(L+j)^{-1/2}} \\
&\overset{(b)}{\le} \sum_{j=1}^\infty \frac{C_T}{(L+j)^{3/2}} < \infty.
\end{aligned}$$

where step (a) is from Chebyshev's inequality and step (b) from (3.4.62). By dominated convergence theorem for infinite sum, we see that $\sum_{j \ge 1}(L+j)^{-3/2} \to 0$ as $L \to +\infty$, thus for any $\eta > 0$ there

is an L such that

$$\sup_{N\in\mathbb{N},t\in[0,T]}\mathbb{P}\left(\nu_t^N\notin\hat{K}_L\right)<\frac{\eta}{4}.$$

With the exact same argument, we can also make L large enough such that

$$\sup_{N\in\mathbb{N},t\in[0,T]}\mathbb{P}\left(\mu_t^N\notin\hat{K}_L\right)<\frac{\eta}{4}.$$

As we have shown in Lemma 3.4.4 and 3.4.5 that the L^2 norm of P and Q are locally bounded, so by Chebyshev's inequality we know for each $\eta>0$, there exists $B>0$ such that

$$\sup_{N\in\mathbb{N},t\in[0,T]}\mathbb{P}\left(Q_t^N\notin[-B,B]^M\right)<\frac{\eta}{4},$$

and

$$\sup_{N\in\mathbb{N},t\in[0,T]}\mathbb{P}\left(P_t^N\notin[-B,B]^M\right)<\frac{\eta}{4}.$$

Therefore, for each $\eta>0$, there is a compact set $\mathcal{K}:=\hat{K}_L\times\hat{K}_L\times[-B,B]^M\times[-B,B]^M\subseteq E$ such that

$$\sup_{N\in\mathbb{N},0\le t\le T}\mathbb{P}\left[\left(\mu_t^N,\nu_t^N,P_t^N,Q_t^N\right)\notin\mathcal{K}\right]<\eta,$$

which completes the proof. □

3.4.2.2 Regularity

Now we establish some regularity results for the sample paths of the process $\left(\mu_t^N,\nu_t^N,P_t^N,Q_t^N\right)$. As in [133], we clarify the following notations:

- $q(z_1,z_2)=|z_1-z_2|\wedge 1$ for any $z_1,z_2\in\mathbb{R}$.
- $\mathcal{F}_t^N$ be the σ-algebra generated by $\left\{(C_0^1,W_0^i)\right\}_{i=1}^N$ and $\left\{\left(\xi_j,\tilde{\xi}_j\right)\right\}_{j=0}^{\lfloor Nt\rfloor-1}$.

Lemma 3.4.8. *Let $f\in C_b^2(\mathbb{R}^{1+d})$. For any $\delta\in(0,1)$, there is a constant $C_T<\infty$ such that for $u\in[0,\delta]$, $t\in[0,T]$,*

$$\mathbb{E}\left[q\left(\left\langle f,\nu_{t+u}^N\right\rangle,\left\langle f,\nu_t^N\right\rangle\right)\mid\mathcal{F}_t^N\right]\le C_T\delta+\frac{C_T}{N^{3/2}} \tag{3.4.64}$$

$$\mathbb{E}\left[q\left(\left\langle f,\mu_{t+u}^N\right\rangle,\left\langle f,\mu_t^N\right\rangle\right)\mid\mathcal{F}_t^N\right]\le C_T\delta+\frac{C_T}{N^{3/2}} \tag{3.4.65}$$

Proof. We start by the following Taylor's expansion for $0\le s<t\le T$:

$$\begin{aligned}
&\left|\left\langle f,\nu_t^N\right\rangle-\left\langle f,\nu_s^N\right\rangle\right|\\
&=\left|\left\langle f,v_{\lfloor Nt\rfloor}^N\right\rangle-\left\langle f,v_{\lfloor Ns\rfloor}^N\right\rangle\right|\\
&\le\frac{1}{N}\sum_{i=1}^N\left|f\left(C_{\lfloor Nt\rfloor}^i,W_{\lfloor Nt\rfloor}^i\right)-f\left(C_{\lfloor Ns\rfloor}^i,W_{\lfloor Ns\rfloor}^i\right)\right|\\
&\le\frac{1}{N}\sum_{i=1}^N\left|\partial_c f\left(\bar{C}_{\lfloor Nt\rfloor}^i,\bar{W}_{\lfloor Nt\rfloor}^i\right)\right|\left|C_{\lfloor Nt\rfloor}^i-C_{\lfloor Ns\rfloor}^i\right|+\frac{1}{N}\sum_{i=1}^N\left\|\partial_w f\left(\bar{C}_{N,s,t}^i,\bar{W}_{N,s,t}^i\right)\right\|\left\|W_{\lfloor Nt\rfloor}^i-W_{\lfloor Ns\rfloor}^i\right\|,
\end{aligned} \tag{3.4.66}$$

where $\bar{C}^i_{N,s,t}, \bar{W}^i_{N,s,t}$ are in the segments connecting $C^i_{\lfloor Ns\rfloor}$ to $C^i_{\lfloor Nt\rfloor}$ and $W^i_{\lfloor Ns\rfloor}$ to $W^i_{\lfloor Nt\rfloor}$ respectively.

Let's now establish a bound on $\left|C^i_{\lfloor Nt\rfloor} - C^i_{\lfloor Ns\rfloor}\right|$ for $s < t \le T$ with $0 < t - s \le \delta < 1$.

$$\begin{aligned}
&\mathbb{E}\left[\left|C^i_{\lfloor Nt\rfloor} - C^i_{\lfloor Ns\rfloor}\right| \mid \mathcal{F}^N_s\right] = \mathbb{E}\left[\left|\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\left(C^i_{k+1} - C^i_k\right)\right| \mid \mathcal{F}^N_s\right] \\
\le&\mathbb{E}\left[\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\frac{\alpha}{N^{\frac{3}{2}}}\left|r(\xi_k) + \gamma Q^N_k(\xi_{k+1}) - Q^N_k(\xi_k)\right| \cdot \left|\sigma(W^i_k \cdot \xi_k)\right| \mid \mathcal{F}^N_s\right] \\
\le&\frac{\alpha C}{N^{\frac{3}{2}}}\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\left(C + (\gamma+1)\mathbb{E}\left[\sup_{\xi\in\mathcal{X}\times\mathcal{A}}|Q^N_k(\xi)|\right]\right) \\
\overset{(a)}{\le}&\frac{C(\lfloor Nt\rfloor - \lfloor Ns\rfloor)}{N^{\frac{3}{2}}}\left(C + (\gamma+1)C_T^{1/2}\right) \\
\le&\frac{C_T(N(t-s)+1)}{N^{\frac{3}{2}}} \le \frac{C_T}{\sqrt{N}}\delta + \frac{C_T}{N^{\frac{3}{2}}}.
\end{aligned} \tag{3.4.67}$$

where step (a) is by Lemma 3.4.4. Similarly for $\left\|W^i_{\lfloor Nt\rfloor} - W^i_{\lfloor Ns\rfloor}\right\|$ for any $s < t \le T$ with $0 < t - s \le \delta < 1$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|W^i_{\lfloor Nt\rfloor} - W^i_{\lfloor Ns\rfloor}\right\| \mid \mathcal{F}^N_s\right] = \mathbb{E}\left[\left\|\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\left(W^i_{k+1} - W^i_k\right)\right\| \mid \mathcal{F}^N_s\right] \\
\le&\mathbb{E}\left[\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\frac{\alpha}{N^{\frac{3}{2}}}\left|r(\xi_k) + \gamma Q^N_k(\xi_{k+1}) - Q^N_k(x_k,a_k)\right| \cdot \left|C^i_k\right| \cdot \left|\sigma'(W^i_k \cdot \xi_k)\right| \mid \mathcal{F}^N_s\right] \\
\le&\frac{\alpha C_T}{N^{\frac{3}{2}}}\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\left(C + (\gamma+1)\mathbb{E}\left[\sup_{\xi\in\mathcal{X}\times\mathcal{A}}|Q^N_k(\xi)|\right]\right) \le \frac{C_T}{\sqrt{N}}\delta + \frac{C_T}{N^{3/2}},
\end{aligned} \tag{3.4.68}$$

where we have used the bound in Lemma 3.4.2 and 3.4.4 again. Combine (3.4.67), (3.4.68) and (3.4.66), we have for any $0 \le s < t \le T$ with $0 < t - s \le \delta < 1$

$$\mathbb{E}\left[\left|\left\langle f, \nu^N_t\right\rangle - \left\langle f, \nu^N_s\right\rangle\right|\right] \le \frac{C_T}{\sqrt{N}}\delta + \frac{C_T}{N^{3/2}} \le C_T\delta + \frac{C_T}{N^{3/2}}. \tag{3.4.69}$$

Similarly for μ^N_t, we have by Taylor's expansion that for $0 \le s < t \le T$ with $0 \le s < t \le T$ that

$$\begin{aligned}
&\left|\left\langle f, \mu^N_t\right\rangle - \left\langle f, \mu^N_s\right\rangle\right| \\
=&\left|\left\langle f, \mu^N_{\lfloor Nt\rfloor}\right\rangle - \left\langle f, \mu^N_{\lfloor Ns\rfloor}\right\rangle\right| \\
\le&\frac{1}{N}\sum_{i=1}^N\left|f\left(B^i_{\lfloor Nt\rfloor}, U^i_{\lfloor Nt\rfloor}\right) - f\left(B^i_{\lfloor Ns\rfloor}, U^i_{\lfloor Ns\rfloor}\right)\right| \\
\le&\frac{1}{N}\sum_{i=1}^N\left|\partial_b f\left(\bar{B}^i_{\lfloor Nt\rfloor}, \bar{U}^i_{\lfloor Nt\rfloor}\right)\right|\left|B^i_{\lfloor Nt\rfloor} - B^i_{\lfloor Ns\rfloor}\right| + \frac{1}{N}\sum_{i=1}^N\left\|\partial_u f\left(\bar{B}^i_{\lfloor Nt\rfloor}, \bar{U}^i_{\lfloor Nt\rfloor}\right)\right\|\left\|U^i_{\lfloor Nt\rfloor} - U^i_{\lfloor Ns\rfloor}\right\|,
\end{aligned} \tag{3.4.70}$$

and

$$\mathbb{E}\left[\left|B^i_{\lfloor Nt\rfloor}-B^i_{\lfloor Ns\rfloor}\right|\mid\mathcal{F}^N_s\right]\le\frac{C}{N^{\frac{3}{2}}}\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\mathbb{E}|\mathsf{clip}(Q^N_k(\tilde{x}_k,\tilde{a}_k))|\le\frac{C(N(t-s)+1)}{N^{\frac{3}{2}}}\le\frac{C_T}{\sqrt{N}}\delta+\frac{C_T}{N^{\frac{3}{2}}}$$
$$\mathbb{E}\left[\left\|U^i_{\lfloor Nt\rfloor}-U^i_{\lfloor Ns\rfloor}\right\|\mid\mathcal{F}^N_s\right]\le\mathbb{E}\left[\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}CN^{-\frac{3}{2}}|\mathsf{clip}(Q^N_k(\tilde{\xi}_k))||B^i_k|\right]\le\frac{C_T}{\sqrt{N}}\delta+\frac{C_T}{N^{\frac{3}{2}}},\tag{3.4.71}$$

where $\bar{B}^i_{N,s,t},\bar{U}^i_{N,s,t}$ are in the segments connecting $B^i_{\lfloor Ns\rfloor}$ to $B^i_{\lfloor Nt\rfloor}$ and $U^i_{\lfloor Ns\rfloor}$ to $U^i_{\lfloor Nt\rfloor}$ respectively. With the fact that the terms $\left|\partial_b f(\bar{B}^i_{\lfloor Nt\rfloor},\bar{U}^i_{\lfloor Nt\rfloor})\right|$ and $\left\|\partial_w f(\bar{B}^i_{\lfloor Nt\rfloor},\bar{U}^i_{\lfloor Nt\rfloor})\right\|$ are bounded in expectation, we have that that for $0\le s<t\le T$ with $0<t-s\le\delta<1$

$$\mathbb{E}\left[\left|\left\langle f,\mu^N_t\right\rangle-\left\langle f,\mu^N_s\right\rangle\right|\right]\le\frac{C_T}{\sqrt{N}}\delta+\frac{C_T}{N^{3/2}}\le C_T\delta+\frac{C_T}{N^{3/2}}.\tag{3.4.72}$$

□

Finally, we prove the regularity of the process $\left(P^N_t,Q^N_t\right)$ by the same method. For our convenience, we abuse notation and define $q(z_1,z_2)=\|z_1-z_2\|_\infty\wedge 1$, where for $z:=(z^1,...,z^M)\in\mathbb{R}^M$ with $M=|\mathcal{X}\times\mathcal{A}|$, we have $\|z\|_\infty=\max_{i=1}^M|z^i|$ is the infinity norm of the vector. [1]

Lemma 3.4.9. *We have*

$$\sup_{k\le NT}\max\left(\mathbb{E}\left[\max_\xi|Q^N_{k+1}(\xi)-Q^N_k(\xi)|\right],\max_\xi|P^N_{k+1}(\xi)-P^N_k(\xi)|\right)\le\frac{C_T}{N}.\tag{3.4.73}$$

With a more delicate analysis, we could show that for any $\delta\in(0,1)$, there is a $C_T<\infty$ such that for $0\le u\le\delta<1$, $t\in[0,T]$,

$$\mathbb{E}\left(q\left(Q^N_{t+u},Q^N_t\right)\mid\mathcal{F}^N_t\right)\le C_T\delta+\frac{C_T}{N},\tag{3.4.74}$$
$$\mathbb{E}\left(q\left(P^N_{t+u},P^N_t\right)\mid\mathcal{F}^N_t\right)\le C_T\delta+\frac{C_T}{N}.\tag{3.4.75}$$

Proof. Recalling the assumption that the state-action space is finite, it suffices to prove a uniform bound for the increments of the outputs $P^N(\xi),Q^N(\xi)$. In particular, by (3.4.48) we have

$$\mathbb{E}\left[\max_\xi|Q^N_{k+1}(\xi)-Q^N_k(\xi)|\right]\le\frac{\alpha}{N}\mathbb{E}\left|r(\xi_k)+\gamma Q^N_k(\xi_{k+1})-Q^N_k(\xi_k)\right||\mathsf{B}^N_{\xi,\xi_k,k}|+\frac{C_T}{N^{3/2}}\le\frac{C_T}{N}+\frac{C_T}{N^{3/2}},\tag{3.4.76}$$

and that by (3.4.8) we have

$$\max_\xi|P^N_{k+1}(\xi)-P^N_k(\xi)|\le\frac{\zeta^N_k}{N}|\mathsf{clip}(Q^N_k(\tilde{\xi}_k))|\left|\bar{\mathsf{B}}_{\xi,\tilde{\xi}_k,k}-\sum_{a''}f^N_k(\tilde{x}'_k,a'')\bar{\mathsf{B}}_{\xi,(\tilde{x}_k,a''),k}\right|+\frac{C_T}{N^{3/2}}\le\frac{C_T}{N}+\frac{C_T}{N^{3/2}}.\tag{3.4.77}$$

[1] The choice of the norm does not matter here as the process (P^N_t,Q^N_t) lives in a finite-dimensional space.

In fact, one could prove a stronger inequality.

$$\begin{aligned}
&\mathbb{E}\left[\max_{\xi}\left|Q_t^N(\xi)-Q_s^N(\xi)\right|\right] \le \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\mathbb{E}\left[\max_{\xi}|Q_{k+1}(\xi)-Q_k(\xi)|\right]\\
&\le \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\sup_{\xi}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left|(C_{k+1}^i-C_k^i)\sigma(W_k^i\cdot\xi)+\sigma'(W_k^i\cdot\xi)\xi^\top(W_{k+1}^i-W_k^i)C_k^i\right|+O_p(N^{-5/2})\right]\\
&\le \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\left[\frac{C}{\sqrt{N}}\sum_{i=1}^{N}\left(|C_{k+1}^i-C_k^i|+\left\|W_{k+1}^i-W_k^i\right\|\right)+O_p(N^{-5/2})\right].
\end{aligned} \tag{3.4.78}$$

Taking expectations and using the bounds (3.4.67) and (3.4.68), we have

$$\begin{aligned}
&\mathbb{E}\left[\max_{\xi}\left|Q_t^N(\xi)-Q_s^N(\xi)\right|\,|\,\mathcal{F}_s^N\right]\\
\le&\sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\left[\frac{C}{\sqrt{N}}\sum_{i=1}^{N}\left(\mathbb{E}\left[|C_{k+1}^i-C_k^i|+\left\|W_{k+1}^i-W_k^i\right\|\,|\,\mathcal{F}_s^N\right]\right)+\mathbb{E}[O_p(N^{-5/2})]\right]\\
\le&\frac{C}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{C_T}{\sqrt{N}}\delta+\frac{C_T}{N^{3/2}}\right)\\
\le&C_T\delta+\frac{C_T}{N}.
\end{aligned} \tag{3.4.79}$$

With exactly the same arguments, we can derive

$$\begin{aligned}
\left|P_t^N(\xi)-P_s^N(\xi)\right| &= \left|P_{\lfloor Nt\rfloor}^N(\xi)-P_{\lfloor Ns\rfloor}^N(\xi)\right|\\
&\le \sum_{k=\lfloor Ns\rfloor}^{\lfloor Nt\rfloor-1}\left[\frac{C}{\sqrt{N}}\sum_{i=1}^{N}\left(|B_{k+1}^i-B_k^i|+\left\|U_{k+1}^i-U_k^i\right\|\right)+O(N^{-5/2})\right],
\end{aligned}$$

which together with (3.4.71) derive

$$\mathbb{E}\left[\max_{\xi}\left|P_t^N(\xi)-P_s^N(\xi)\right|\,|\,\mathcal{F}_s^N\right]\le C_T\delta+\frac{C_T}{N}.$$

□

3.4.2.3 Proof of Relative Compactness

Theorem 8.6 and Remark 8.7 in [56] provides a criterion for us to prove the relative compactness of a general stochastic process with càdlàg sample paths, for which we will state without proof.

Theorem 3.4.10. *Let E be a metric space equipped with the metric r. Denote $q=r\wedge 1$, and let (X_t^N). be a sequence of E-valued stochastic processes with càdlàg sample paths. Write $\mathcal{F}_t^N$ as the natural filtration generated by the random variables (X_t^N). Then $(X_t^N)_{t\ge 0}$ is relatively compact if the following conditions hold:*

1. *(Compact containment) For any $\eta>0$ and (rational) $t>0$, there is a compact subset $\mathcal{K}:=\mathcal{K}_{\eta,t}$ of E such that*

$$\sup_{N\in\mathbb{N}}\mathbb{P}(X_t^N\notin\mathcal{K})<\eta. \tag{3.4.80}$$

2. *(Regularity of paths) For each $T > 0$, there is a family of non-negative random variables $\{\gamma_N(\delta) : \delta \in (0,1)\}$ satisfying*

$$\mathbb{E}\left[q(X^N_{t+u}, X^N_t) \,|\, \mathcal{F}^N_t\right] \le \mathbb{E}\left[\gamma_N(\delta) \,|\, \mathcal{F}^N_t\right], \quad t \in [0,T],\ u \in [0,\delta], \tag{3.4.81}$$

such that

$$\lim_{\delta\to 0} \limsup_{N\to\infty} \mathbb{E}[\gamma_N(\delta)] = 0. \tag{3.4.82}$$

We will therefore prove condition 1 and 2 in the Section 3.4.2.1 and Section 3.4.2.2 respectively.

Lemma 3.4.11. *The family of processes $\left(\mu^N_t, \nu^N_t, P^N_t, Q^N_t\right)_{N\in\mathbb{N}}$ is relative compact in $D_E([0,T])$.*

Proof. Combining the two lemmas above, we see that the process (μ^N, ν^N, P^N, Q^N) satisfies condition 2 with $\gamma_N(\delta)$ being a $O(\delta)$ term plus a $o(1)$ term with respect to N. All conditions in theorem 3.4.10 are satisfied, and hence the sequence of processes (μ^N, ν^N, P^N, Q^N) is relatively compact. □

3.4.3 Identification of the Limit

With the relative compactness result in Section 3.4.2, we can conclude that (μ^N, ν^N, P^N, Q^N) contains a subsequence that converges weakly. To prove the convergence in Theorem 3.3.3, we need to identify the potential limit points, which involves showing the error terms $M^N_t, M^{i,N}_t \overset{N\to\infty}{\to} 0$ in probability for $i = 1,2,3$. Then the desired convergence comes from the uniqueness of the limit ODEs.

We begin by some notations.

- For any $k \ge 0$, we let $\mathbb{P}^N_k$ and Π_k be the transition kernel of $(\boldsymbol{\mathcal{M}}, g^N_k)$ and $(\boldsymbol{\mathcal{M}}, g^N_k)_{\mathsf{aux}}$ respectively, so that

$$\begin{aligned} \mathbb{P}^N_k((x,a)\to(x',a')) &= p(x'|x,a)g^N_k(x',a'), \\ \Pi^N_k((x,a)\to(x',a')) &= \tilde{p}(x'|x,a)g^N_k(x',a'). \end{aligned} \tag{3.4.83}$$

  We highlight the superscript N in transition probability $\mathbb{P}^N_k, \Pi^N_k$ comes from the pre-limit neural network P^N_k.

- Let $\pi^{g^N_k}$ and $\sigma^{g^N_k}_{\rho_0}$ denote the stationary distributions of $(\boldsymbol{\mathcal{M}}, g^N_k)$ and $(\boldsymbol{\mathcal{M}}, g^N_k)_{\mathsf{aux}}$ respectively, whose existence and uniqueness are given by Assumption 3.2.7. The initial distribution ρ_0 in $\sigma^{g^N_k}_{\rho_0}$ may be omitted when the context is clear.

- Define the σ-field of events generated by the joint Actor and Critic processes up to n-th step be

$$\mathscr{F}_n = \sigma(\xi_k, \tilde{\xi}_k)_{k\le n}, \quad (\xi_k)_{k\ge 0} \sim (\boldsymbol{\mathcal{M}}, \mathsf{Cr}),\ (\tilde{\xi}_k)_{k\ge 0} \sim (\boldsymbol{\mathcal{M}}, \mathsf{Ac}). \tag{3.4.84}$$

  Then $\mathbb{P}^N_k$ and Γ^N_k each induces an operator acting on any Borel function $h(\cdot) : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$

$$\begin{aligned} \mathbb{P}^N_k h(\xi) &:= \sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} h(\xi')\mathbb{P}^N_k(\xi\to\xi') \\ \Pi^N_k h(\xi) &:= \sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} h(\xi')\Pi^N_k(\xi\to\xi'), \end{aligned} \tag{3.4.85}$$

3.4.3.1 Poisson Equations

Now we rigorously derive the limit ODEs by using a Poisson equation [113, 143, 144, 145], which can be comprehended as the limit of the Kolmogorov forward equation (Fokker-Planck equation [95, 96, 115]) for stochastic process, to bound the fluctuations terms around the trajectory of the limit ODE. Such analysis is needed as the fluctuation terms evolve as the actor and critic networks evolve, which further depend on the non-i.i.d data samples from the Markov chains (3.2.14) and (3.2.15). We first prove

$$\lim_{N\to\infty} \mathbb{E} \sup_{t\in[0,T]} \left|M_t^N(x,a)\right| = 0, \quad \forall (x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}. \tag{3.4.86}$$

Using a similar method, we can also prove the convergence of $M_t^{1,N}, M_t^{2,N}$, and $M_t^{3,N}$.

It is known that a finite state Markov chain which is irreducible and non-periodic has a geometric convergence rate to its stationary distribution [106]. We are able to prove a uniform geometric convergence rate for the Markov chains in this chapter under the *time-evolving* actor policy updated using the actor-critic algorithm (1).

Lemma 3.4.12. *Let $\Pi_k^{N,n}$ denote the n-step transition matrix under derived from transition probability Π_k^N with $\Pi_k^{N,0}(\xi,\xi') = \mathbb{1}_{\xi'=\xi}$. Then, for any fixed $T>0$, there exists an integer n_0 such that the following uniform estimates hold for all policies $\{g_k^N\}_{0\le k\le NT}$ and $N\in\mathbb{N}$ for the algorithm (1).*

- *Lower bound for the stationary distribution:*

$$\inf_{k\le NT} \sigma^{g_k^N}(x,a) \ge C\epsilon_T^{n_0}, \quad \forall (x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}, \tag{3.4.87}$$

  where $C, \epsilon_T > 0$ are positive constants.

- *Uniform geometric ergodicity:*

$$\sup_{k\le NT} \|\Pi_k^{N,n}(\xi\to\cdot) - \sigma^{g_k^N}(\cdot)\| \le (1-\beta_T)^{\lfloor \frac{n}{n_0} \rfloor} \quad \forall \xi \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}, \tag{3.4.88}$$

  where $\beta_T \in (0,1)$ is a positive constant, and the norm $\|\cdot\|$ is the usual total variation norm.

The proof of the above lemma is exactly the same as the lemma A.4 of [133].Then, using the same method as in Lemma 3.4.12, we can prove a similar result for the MDP $\boldsymbol{\mathcal{M}}$ with exploration policy g_k^N.

Corollary 3.4.13. *Let $\mathbb{P}_k^{N,n}$ denote the n-step transition matrix under policy g_k^N with $\mathbb{P}_k^{N,0}(\xi,\xi') = \mathbb{1}_{\{\xi'=\xi\}}$. Then, for any fixed $T<\infty$, there exists an integer n_0 and a constant*

$$C = C(n_0) := \inf_{x,a,x'} \sum_{\xi_1,\cdots,\xi_{n_0-1}} p(x_1|x,a)\cdots p(x'|x_{n_0-1},a_{n_0-1}) > 0, \tag{3.4.89}$$

such that the following uniform estimate holds for all $\{g_k^N\}_{0\le k\le NT}$ and $N\in\mathbb{N}$ for the update algorithm (1):

- *Lower bound for the stationary distribution:*

$$\inf_{k\leq NT} \pi^{g_k^N}(x,a) \geq C\left(\eta^N_{\lfloor NT\rfloor}\right)^{n_0}, \quad \forall (x,a)\in \boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}. \tag{3.4.90}$$

- *Uniform geometric ergodicity:*

$$\sup_{k\leq NT} \|\mathbb{P}_k^{N,n}(\xi\to\cdot) - \pi^{g_k^N}(\cdot)\| \leq (1-\beta_T)^{\lfloor \frac{n}{n_0}\rfloor} \quad \forall \xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}, \tag{3.4.91}$$

where $\beta_T = C\left(\eta^N_{\lfloor NT\rfloor}\right)^{n_0} \in (0,1)$ *is a positive constant.*

Without loss of generality, we assume that the value of n_0 in the lemma 3.4.12 and 3.4.13 are the same. In order to prove the stochastic fluctuation term vanishes as $N\to\infty$, we solve the system of Poisson equations associated with the Markov chains $(\boldsymbol{\mathcal{M}}, g_k^N)$ and $(\boldsymbol{\mathcal{M}}, g_k^N)_{\mathsf{aux}}$, which relates their transition kernels with their unique stationary distributions. We will only analyse the Markov chain $(\boldsymbol{\mathcal{M}}, g_k^N)_{\mathsf{aux}}$ here as the analysis for $(\boldsymbol{\mathcal{M}}, g_k^N)$ is identical. The system of Poisson equations associated with $(\boldsymbol{\mathcal{M}}, g_k^N)_{\mathsf{aux}}$ is defined as followed:

Definition 3.4.14 (Poisson equations)**.** Let $N\in\mathbb{N}$, $T>0$ and $k\leq NT$. The Poisson equations corresponding to the chain induced by transition kernel Π_k^N state-action seeks a function $\nu^N_{k,\xi}(\cdot) : \boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}\to\mathbb{R}$ for each state-action pairs $\xi=(x,a)$, such that

$$\nu^N_{k,\xi}(\xi') - \Pi_k^N \nu^N_{k,\xi}(\xi') = \mathbb{1}_{\{\xi'=\xi\}} - \sigma^{g_k^N}(\xi), \quad \forall \xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}. \tag{3.4.92}$$

Lemma 3.4.15 (Existence of solution to the Poisson equations)**.** *The Poisson equations* (3.4.92) *admits a uniformly bounded solution*

$$\nu^N_{k,\xi}(\xi') := \sum_{n\geq 0}\left[\Pi_k^{N,n}(\xi'\to\xi) - \sigma^{g_k^N}(\xi)\right], \tag{3.4.93}$$

and there exists a constant C_T *(which only depends on* T*) such that*

$$\sup_{k\leq NT}\max_{\xi,\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} \left|\nu^N_{k,\xi}(\xi')\right| \leq C_T. \tag{3.4.94}$$

Remark 3.4.16. For the purposes of our later analysis, it is enough to find a uniformly bounded solution ν_θ which satisfies (3.4.93). Therefore, we do not establish the uniqueness of the solution to the Poisson equation (3.4.92) here.

Proof. (of lemma 3.4.15). Due to the uniform geometric convergence rate (3.4.88) for all $k\leq NT$ in Lemma 3.4.12, there exists a $\beta_T>0$ (independent with k) such that for any $\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}$

$$\left|\Pi_k^{N,n}(\xi'\to\xi) - \sigma^{g_k^N}(\xi)\right| \leq (1-\beta_T)^{\lfloor\frac{n}{n_0}\rfloor}, \quad \forall k\leq NT \tag{3.4.95}$$

which can be used to show the convergence of the series in (3.4.93). Consequently, $\nu^N_{k,\xi}$ is well-defined and uniformly bounded as in (3.4.94). In fact,

$$\left|\nu^N_{k,\xi}(\xi')\right| \leq \sum_{n\geq 0}\left|\Pi_k^{N,n}(\xi'\to\xi) - \sigma^{g_k^N}(\xi)\right| \leq \sum_{n\geq 0}(1-\beta_T)^{\lfloor\frac{n}{n_0}\rfloor} \leq C_T. \tag{3.4.96}$$

Finally, we can verify that $\nu^N_{k,\xi}$ is a solution to the Poisson equation (3.4.92) by observing that

$$\begin{aligned}
\Pi^N_k \nu^N_{k,\xi}(\xi') &= \sum_y \nu^N_{k,\xi}(y)\Pi^N_k(\xi' \to y) \\
&= \sum_y \left(\sum_{n\geq 0} \left[\Pi^{N,n}_k(y \to \xi) - \sigma^{g^N_k}(\xi) \right] \right) \Pi^N_k(\xi' \to y) \\
&\overset{(a)}{=} \sum_{n\geq 0} \left(\sum_y \left[\Pi^{N,n}_k(y \to \xi) - \sigma^{g^N_k}(\xi) \right] \Pi^N_k(\xi' \to y) \right) \\
&= \sum_{n\geq 0} \left[\Pi^{N,n+1}_k(\xi' \to \xi) - \sigma^{g^N_k}(\xi) \right] \\
&= \nu^N_{k,\xi}(\xi') - (\mathbb{1}_{\{\xi'=\xi\}} - \sigma^{g^N_k}(\xi)),
\end{aligned}$$

where the step (a) uses (3.4.95) and the Dominated Convergence Theorem. □

Using the Poisson equation (3.4.15), we can prove that the fluctuations of the data samples around a dynamic visiting measure $\sigma^{g^N_k}$ decay when the iteration steps become large.

Lemma 3.4.17. *Let $(\tilde{\xi}_k)_{k\geq 0}$ be the Actor process $(\boldsymbol{\mathcal{M}}, \mathsf{Ac})$. Then for any fixed state action pair $\xi = (x, a)$ and $T > 0$,*

$$\lim_{N\to\infty} \mathbb{E} \left| \frac{1}{N} \sum_{k=0}^{c(T,N)} \left[\mathbb{1}_{\{\tilde{\xi}_k = \xi\}} - \sigma^{g^N_k}(\xi) \right] \right|^2 = 0, \tag{3.4.97}$$

where $c(T, N)$ is a positive integer that depends on T and N such that $c(T, N) \leq \lfloor NT \rfloor - 1$.

The proof of Lemma 3.4.17 is similar to Lemma 2.4.8 and thus omitted. The detailed proof can be found in Lemma 4.17 of [43]. Then we can show the convergence of the stochastic fluctuation terms from the actor update.

Lemma 3.4.18. *For any $\xi = (x, a)$ and the stochastic error M^N_t defined in (3.4.52), we have*

$$\lim_{N\to\infty} \sup_{t\in(0,T]} \mathbb{E} \left| M^N_t(\xi) \right| = 0. \tag{3.4.98}$$

Proof. The proof of (3.4.98) consists of two parts. We first set up a bound for the difference of the actor's update. Define

$$\bar{H}^N_{\xi,\xi',k} := \zeta^N_k \, \mathsf{clip}(Q^N_k(\xi')) \left[\bar{\mathsf{B}}^N_{\xi,\xi',k} - \sum_{a''} f^N_k(x', a'') \bar{\mathsf{B}}^N_{\xi,(x',a''),k} \right]. \tag{3.4.99}$$

If we can prove

$$\left| \bar{H}^N_{\xi,\xi',k+1} - \bar{H}^N_{\xi,\xi',k} \right| \leq \frac{C_T}{N} \tag{3.4.100}$$

Then we can use Lemma 3.4.17 to prove that as the training step becomes large, the fluctuations of the data samples around the stationary distribution will disappear, completing our proof.

(i) To bound the difference (3.4.100), note that

$$
\begin{aligned}
&\left|\bar{H}^N_{\xi,\xi',k+1} - \bar{H}^N_{\xi,\xi',k}\right| \\
\leq &|\zeta^N_{k+1} - \zeta^N_k| \left|\mathsf{clip}(Q^N_{k+1}(\xi')) \left[\bar{\mathsf{B}}^N_{\xi,\xi',k+1} - \sum_{a''} f^N_{k+1}(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),k+1}\right]\right| \\
+&\zeta^N_k \left|\mathsf{clip}(Q^N_{k+1}(\xi')) - \mathsf{clip}(Q^N_k(\xi'))\right| \left|\bar{\mathsf{B}}^N_{\xi,\xi',k+1} - \sum_{a''} f^N_{k+1}(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),k+1}\right| \\
+&\zeta^N_k \left|\mathsf{clip}(Q^N_k(\xi'))\right| \left|\left[\bar{\mathsf{B}}^N_{\xi,\xi',k+1} - \sum_{a''} f^N_{k+1}(x',a'')\bar{B}^N_{\xi,(x',a''),k+1}\right] - \left[\bar{\mathsf{B}}^N_{\xi,\xi',k} - \sum_{a''} f^N_k(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),k}\right]\right| \\
:=&I^N_1 + I^N_2 + I^N_3.
\end{aligned} \tag{3.4.101}
$$

For the first term,

$$
I^N_1 \leq C_T|\zeta^N_{k+1} - \zeta^N_k| \leq C_T\left(\frac{1}{1+\frac{k}{N}} - \frac{1}{1+\frac{k+1}{N}}\right) = \frac{C_T}{N\left(1+\frac{k}{N}\right)\left(1+\frac{k+1}{N}\right)} \leq \frac{C_T}{N}. \tag{3.4.102}
$$

Then noting that the function $\mathsf{clip}(.)$ is 1-Lipschitz (i.e. $|\mathsf{clip}(x) - \mathsf{clip}(y)| \leq |x-y|$), we have

$$
I^N_2 \leq \frac{C_T}{N}\left|\bar{\mathsf{B}}^N_{\xi,\xi',k+1} - \sum_{a''} f^N_k(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),k+1}\right| \leq \frac{C_T}{N}. \tag{3.4.103}
$$

Finally, by lemma 3.4.3 we know that for any $k \leq NT$,

$$
\sup_{\xi,\xi'\in\mathcal{X}\times\mathcal{A}} \left|\bar{\mathsf{B}}^N_{\xi,\xi',k+1} - \bar{\mathsf{B}}^N_{\xi,\xi',k}\right| \leq \frac{C_T}{N}. \tag{3.4.104}
$$

Hence,

$$
\begin{aligned}
I^N_3 &\leq C\left|\left[\bar{\mathsf{B}}^N_{\xi,\xi',k+1} - \sum_{a''} f^N_{k+1}(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),k+1}\right] - \left[\bar{\mathsf{B}}^N_{\xi,\xi',k} - \sum_{a''} f^N_k(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),k}\right]\right| \\
&\leq C\Bigg[\left|\bar{\mathsf{B}}^N_{\xi,\xi',k+1} - \bar{\mathsf{B}}^N_{\xi,\xi',k}\right| + \sum_{a''}\left|f^N_{k+1}(x',a'') - f^N_k(x',a'')\right| \cdot \left|\bar{\mathsf{B}}^N_{\xi,(x',a''),k+1}\right| \\
&\quad + \sum_{a''} f^N_k(x',a'')\left|\bar{\mathsf{B}}^N_{\xi,(x',a''),k+1} - \bar{\mathsf{B}}^N_{\xi,(x',a''),k}\right|\Bigg] \\
&\leq C\left(1 + \sum_{a''} f^N_k(x',a'')\right)\sup_{\xi'\in\mathcal{X}\times\mathcal{A}}\left|\bar{\mathsf{B}}^N_{\xi,\xi',k+1} - \bar{\mathsf{B}}^N_{\xi,\xi',k}\right| + C\left\|P^N_{k+1} - P^N_k\right\| \leq \frac{C_T}{N}.
\end{aligned} \tag{3.4.105}
$$

Combining (3.4.102), (3.4.103) and (3.4.105), we can conclude (3.4.100).

(ii) Now we can prove the convergence (3.4.98). We let $K := K(N) \in \mathbb{N}$, such that $1 \ll K(N) \ll N$ (i.e. $K(N) \to +\infty$ and $K(N)/N \to 0$ as $N \to \infty$). We further define $\Delta = t/K$. Then

$$
\begin{aligned}
M^N_t(\xi) &= \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor - 1}\left(\bar{H}^N_{\xi,\tilde{\xi}_k,k} - \sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} \bar{H}^N_{\xi,\xi',k}\sigma^{g^N_k}(\xi')\right) \\
&= \frac{1}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor - 1}\left(\bar{H}^N_{\xi,\tilde{\xi}_k,k} - \sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} \bar{H}^N_{\xi,\xi',k}\sigma^{g^N_k}(\xi')\right) + r^N_t(\xi),
\end{aligned}
$$

where

$$r_t^N(\xi) = \frac{1}{N} \sum_{k=K\lfloor N\Delta\rfloor}^{\min((K+1)\lfloor N\Delta\rfloor-1,\lfloor Nt\rfloor-1)} \left(\bar{H}^N_{\xi,\tilde{\xi}_k,k} - \sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} \bar{H}^N_{\xi,\xi',k} \sigma^{g_k^N}(\xi') \right).$$

The terms $H^N_{\xi,\xi',k}$ are bounded by some constant $C_T > 0$ as the kernel entries $|\bar{\mathsf{B}}^N_{\xi,\xi',k}|$ are bounded, so are the summands. Thus

$$|r_t^N(\xi)| \le \frac{\lfloor N\Delta\rfloor}{N} C_T \le \frac{TC_T}{K}. \tag{3.4.106}$$

We could further break down $M_t^N(\xi)$ as followed:

$$\begin{aligned} M_t^N(\xi) - r_t^N(\xi) = \frac{1}{N} \sum_{j=0}^{K-1} \sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1} \Bigg[& \left(\bar{H}^N_{\xi,\tilde{\xi}_k,k} - \bar{H}^N_{\xi,\tilde{\xi}_k,j\lfloor N\Delta\rfloor} \right) \\ & + \left(\bar{H}^N_{\xi,\tilde{\xi}_k,j\lfloor N\Delta\rfloor} - \sum_{\xi'} \bar{H}^N_{\xi,\xi',j\lfloor N\Delta\rfloor} \sigma^{g_k^N}(\xi') \right) + \sum_{\xi'} \left(\bar{H}^N_{\xi,\xi',j\lfloor N\Delta\rfloor} - \bar{H}^N_{\xi,\xi',k} \right) \sigma^{g_k^N}(\xi') \Bigg] \\ = J^N_{1,t}(\xi) + J^N_{2,t}(\xi) + J^N_{3,t}(\xi), & \end{aligned} \tag{3.4.107}$$

where

$$\begin{aligned} J^N_{1,t}(\xi) &= \frac{1}{N} \sum_{j=0}^{K-1} \sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1} \left(\bar{H}^N_{\xi,\tilde{\xi}_k,k} - \bar{H}^N_{\xi,\tilde{\xi}_k,j\lfloor N\Delta\rfloor} \right) \\ J^N_{2,t}(\xi) &= \frac{1}{N} \sum_{j=0}^{K-1} \sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1} \left(\bar{H}^N_{\xi,\tilde{\xi}_k,j\lfloor N\Delta\rfloor} - \sum_{\xi'} \bar{H}^N_{\xi,\xi',j\lfloor N\Delta\rfloor} \sigma^{g_k^N}(\xi') \right) \\ J^N_{3,t}(\xi) &= \frac{1}{N} \sum_{j=0}^{K-1} \sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1} \sum_{\xi'} \left(\bar{H}^N_{\xi,\xi',j\lfloor N\Delta\rfloor} - \bar{H}^N_{\xi,\xi',k} \right) \sigma^{g_k^N}(\xi'). \end{aligned}$$

Using (3.4.100), we have

$$\begin{aligned} \max\left(\left| \bar{H}^N_{\xi,\tilde{\xi}_k,k} - \bar{H}^N_{\xi,\tilde{\xi}_k,j\lfloor\Delta N\rfloor} \right|, \sum_{\xi'} \left| \bar{H}^N_{\xi,\xi',k} - \bar{H}^N_{\xi,\xi',j\lfloor\Delta N\rfloor} \right| \sigma^{g_k^N}(\xi') \right) &\le \sup_{\xi,\xi'} |\bar{H}^N_{\xi,\xi',k} - \bar{H}^N_{\xi,\xi',j\lfloor\Delta N\rfloor}| \\ &\le \frac{C_T(k - j\lfloor\Delta N\rfloor)}{N}. \end{aligned} \tag{3.4.108}$$

Therefore,

$$\begin{aligned} \max(J^N_{1,t}(\xi), J^N_{3,t}(\xi)) &\le \frac{1}{N} \sum_{j=0}^{K-1} \sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor\Delta N\rfloor-1} C_T \frac{k - j\lfloor N\Delta\rfloor}{N} \\ &= \frac{1}{N} \sum_{j=0}^{K-1} \sum_{k=0}^{\lfloor\Delta N\rfloor-1} \frac{C_T k}{N} \\ &\le \frac{C_T}{N} \sum_{j=0}^{K-1} \frac{\lfloor N\Delta\rfloor^2}{N} \\ &= \frac{KC_T\lfloor N\Delta\rfloor^2}{N^2} \le KC_T\Delta^2 = C_T K \left(\frac{t}{K}\right)^2 \le \frac{C_T}{K}. \end{aligned} \tag{3.4.109}$$

To control $J_{2,t}^N(\xi)$, we note that

$$\bar{H}^N_{\xi,\tilde{\xi}_k,j\lfloor N\Delta\rfloor} - \sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} \bar{H}^N_{\xi,\xi',j\lfloor N\Delta\rfloor}\sigma^{g_k^N}(\xi') = \sum_{\xi'} \bar{H}^N_{\xi,\xi',j\lfloor N\Delta\rfloor}\left[\mathbb{1}_{\{\tilde{\xi}_k=\xi'\}} - \sigma^{g_k^N}(\xi')\right], \tag{3.4.110}$$

so one could control $J_{2,t}^N(\xi)$ by the uniform boundedness of $\bar{H}^N_{\xi,\xi',j\lfloor\Delta N\rfloor}$ and lemma 3.4.17. Indeed,

$$\begin{aligned}
\left|J_{2,t}^N(\xi)\right| &= \left|\frac{1}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}\bar{H}^N_{\xi,\xi',j\lfloor N\Delta\rfloor}\left[\mathbb{1}_{\{\xi_k=\xi'\}} - \sigma^{g_k^N}(\xi')\right]\right| \\
&= \left|\frac{1}{N}\sum_{j=0}^{K-1}\sum_{\xi'}\bar{H}^N_{\xi,\xi',j\lfloor N\Delta\rfloor}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\left[\mathbb{1}_{\{\xi_k=\xi'\}} - \sigma^{g_k^N}(\xi')\right]\right| \\
&\le C_T\sum_{\xi'}\left|\frac{1}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\left[\mathbb{1}_{\{\xi_k=\xi'\}} - \sigma^{g_k^N}(\xi')\right]\right|, \\
&= C_T\sum_{\xi'}\left|\frac{1}{N}\sum_{k=0}^{K\lfloor N\Delta\rfloor-1}\left[\mathbb{1}_{\{\xi_k=\xi'\}} - \sigma^{g_k^N}(\xi')\right]\right|,
\end{aligned} \tag{3.4.111}$$

which together with Lemma 3.4.17 derive

$$\lim_{N\to\infty}\mathbb{E}\left|J_{2,t}^N(\xi)\right|^2 = 0.$$

Collecting our results, we have shown that

$$\sup_{t\in(0,T]}\mathbb{E}\left|M_t^N(\xi)\right| \le \frac{C_T}{K(N)} \overset{N\to\infty}{\to} 0 \tag{3.4.112}$$

by the assumption that $1 \ll K(N)$. □

Following the same method, we can finish proving the convergence of the stochastic fluctuation terms from the dynamics of the critic network.

Lemma 3.4.19. *For any $\xi = (x,a)$ and the stochastic error $M_t^{i,N}, i = 1,2,3$ defined in (3.4.50), we have*

$$\lim_{N\to\infty}\sup_{t\in(0,T]}\mathbb{E}\left|M_t^{i,N}(\xi)\right| = 0, \quad i = 1,2,3. \tag{3.4.113}$$

Proof. As in the proof for the decay of M_t^N, we use two steps to prove the result.

(i) Prove that the fluctuations of the data samples around a dynamic stationary distribution π^{g_k} decay when the number of iteration steps becomes large. Actually, with exactly the same approach as in Lemma 3.4.17, we can prove for any fixed state action pair $\xi = (x,a), \forall T > 0$

$$\lim_{N\to 0}\mathbb{E}\left|\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\left[\mathbb{1}_{\{\xi_k=\xi\}} - \pi^{g_k}(\xi)\right]\right|^2 = 0. \tag{3.4.114}$$

(ii) Use the same method as in Lemma 3.4.18 to prove the stochastic fluctuation terms vanish as $N\to\infty$.

We first look at $M_t^{3,N}$ and the proof for $M_t^{1,N}, M_t^{2,N}$ is the same. Recalling the notation in (3.4.85), we have

$$\begin{aligned} M_t^{3,N}(\xi) &= \frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\gamma\left[Q_k^N(\xi_{k+1})-\mathbb{P}_k^N Q_k^N(\xi_k)\right]B^N_{\xi,\xi_k,k} \\ &\quad+\frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor-1}\gamma\left[\mathbb{P}_k^N Q_k^N(\xi_k)B^N_{\xi,\xi_k,k}-\sum_{\xi'}\mathbb{P}_k^N Q_k^N(\xi')B^N_{\xi,\xi',k}\pi^{g_k^N}(\xi')\right] \\ &:= I_t^{1,N}(\xi)+I_t^{2,N}(\xi). \end{aligned} \tag{3.4.115}$$

To control $I_t^{1,N}(\xi)$, we first define

$$\epsilon_k := \left[Q_k^N(\xi_{k+1})-\mathbb{P}_k^N Q_k^N(\xi_k)\right]B^N_{\xi,\xi_k,k}. \tag{3.4.116}$$

Since

$$\mathbb{E}\left[Q_k^N(\xi_{k+1})\mid\mathscr{F}_k\right]=\mathbb{P}_k^N Q_k^N(\xi_k), \tag{3.4.117}$$

hence

$$\sum_{k=0}^{n-1}\epsilon_k$$

is a martingale with respect to the filtration $\mathscr{F}_n$. Since the conditional expectation is a contraction in L^2, we have

$$\mathbb{E}\left|\mathbb{P}_k^N Q_k^N(\xi_k)\right|^2\le\mathbb{E}\left|Q_k^N(\xi_{k+1})\right|^2. \tag{3.4.118}$$

Then,

$$\begin{aligned} \mathbb{E}\left|\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon_k\right|^2 &= \frac{1}{N^2}\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbb{E}\left|\mathbb{P}_k^N Q_k^N(\xi_k)-Q_k^N(\xi_{k+1})\right|^2 \\ &\le \frac{4}{N^2}\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbb{E}\left|Q_k^N(\xi_{k+1})\right|^2\overset{(a)}{\le}\frac{C_T}{N}, \end{aligned} \tag{3.4.119}$$

where step (a) follows from (3.4.26) and Lemma 3.4.4. Thus, for any $T>0$,

$$\lim_{N\to\infty}\mathbb{E}\left|I_t^{1,N}\right|=\lim_{N\to\infty}\gamma\mathbb{E}\left|\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\epsilon_k\right|=0. \tag{3.4.120}$$

For $I_t^{2,N}$, we define as in the proof of Lemma 3.4.18

$$H^N_{\xi,\xi',k} := \mathbb{P}_k^N Q_k^N(\xi')B^N_{\xi,\xi',k}=\sum_{z,a''}Q_k^N(z,a'')g_k^N(z,a'')p(z|\xi')B^N_{\xi,\xi',k}. \tag{3.4.121}$$

By Lemma 3.4.3 and 3.4.4, we have the bound

$$\sup_{0\le k\le\lfloor TN\rfloor}\sup_{\xi'\in\mathcal{X}\times\mathcal{A}}\mathbb{E}\left|H^N_{\xi,\xi',k}\right|^2\le C_T. \tag{3.4.122}$$

Furthermore, by Lemma 3.4.3 and 3.4.9,

$$\begin{aligned}\mathbb{E}\left|H^N_{\xi,\xi',k+1}-H^N_{\xi,\xi',k}\right|^2 &\le \sum_{z,a''}\mathbb{E}\left|Q^N_{k+1}(z,a'')g^N_{k+1}(z,a'')\mathsf{B}^N_{\xi,\xi',k+1}-Q^N_k(z,a'')g^N_k(z,a'')\mathsf{B}^N_{\xi,\xi',k}\right|^2\\ &\le 3\sum_{z,a''}\left|\left(Q^N_{k+1}(z,a'')-Q^N_k(z,a'')\right)g^N_{k+1}(z,a'')\mathsf{B}^N_{\xi,\xi',k+1}\right|^2\\ &+3\sum_{z,a''}\left|Q^N_k(z,a'')\mathsf{B}^N_{\xi,\xi',k+1}\left(g^N_{k+1}(z,a'')-g^N_k(z,a'')\right)\right|^2\\ &+3\sum_{z,a''}\left|Q^N_k(z,a'')g^N_k(z,a'')\left(\mathsf{B}^N_{\xi,\xi',k+1}-\mathsf{B}^N_{\xi,\xi',k}\right)\right|^2\\ &\le \frac{C_T}{N^2},\end{aligned} \tag{3.4.123}$$

so

$$\sup_{0\le k\le\lfloor TN\rfloor-1}\sup_{\xi'\in\mathcal{X}\times\mathcal{A}}\mathbb{E}\left|H^N_{\xi,\xi',k+1}-H^N_{\xi,\xi',k}\right|\le\left(\sup_{0\le k\le\lfloor TN\rfloor-1}\sup_{\xi'\in\mathcal{X}\times\mathcal{A}}\mathbb{E}\left|H^N_{\xi,\xi',k+1}-H^N_{\xi,\xi',k}\right|^2\right)^{\frac{1}{2}}\le\frac{C_T}{N}. \tag{3.4.124}$$

Then following the step (ii) in the proof of Lemma 3.4.18, now we can prove the convergence $I^{2,N}_t(\xi)$. We let $K:=K(N)\in\mathbb{N}$ such that $1\ll K\ll N$ and define $\Delta=t/K$. Then, we can decompose $I^{2,N}_t(\xi)$ into the following terms:

$$I^{2,N}_t(\xi)=J^N_{1,t}(\xi)+J^N_{2,t}(\xi)+J^N_{3,t}(\xi)+r^N_t(\xi), \tag{3.4.125}$$

where

$$\begin{aligned}J^N_{1,t}(\xi)&=\frac{1}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\left(H^N_{\xi,\tilde{\xi}_k,k}-H^N_{\xi,\tilde{\xi}_k,j\lfloor N\Delta\rfloor}\right)\\ J^N_{2,t}(\xi)&=\frac{1}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\left(H^N_{\xi,\tilde{\xi}_k,j\lfloor N\Delta\rfloor}-\sum_{\xi'}H^N_{\xi,\xi',j\lfloor N\Delta\rfloor}\pi^{g^N_k}(\xi')\right)\\ J^N_{3,t}(\xi)&=\frac{1}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}\left(H^N_{\xi,\xi',j\lfloor N\Delta\rfloor}-H^N_{\xi,\xi',k}\right)\pi^{g^N_k}(\xi')\\ r^N_t(\xi)&=\frac{1}{N}\sum_{k=K\lfloor N\Delta\rfloor}^{\min((K+1)\lfloor N\Delta\rfloor-1,\lfloor Nt\rfloor-1)}\left(H^N_{\xi,\tilde{\xi}_k,k}-\sum_{\xi'}H^N_{\xi,\xi',k}\pi^{g^N_k}(\xi')\right).\end{aligned}$$

Again, we have

$$\begin{aligned}|r^N_t(\xi)|^2&\le\frac{\lfloor N\Delta\rfloor}{N^2}\sum_{k=K\lfloor N\Delta\rfloor}^{\min((K+1)\lfloor N\Delta\rfloor-1,\lfloor Nt\rfloor-1)}\left(H^N_{\xi,\tilde{\xi}_k,k}-\sum_{\xi'}H^N_{\xi,\xi',k}\pi^{g^N_k}(\xi')\right)^2\\ &\le\frac{2\Delta}{N}\sum_{k=K\lfloor N\Delta\rfloor}^{\min((K+1)\lfloor N\Delta\rfloor-1,\lfloor Nt\rfloor-1)}\left[\left(H^N_{\xi,\tilde{\xi}_k,k}\right)^2+\sum_{\xi'}\left(H^N_{\xi,\xi',k}\pi^{g^N_k}(\xi')\right)^2\right]\\ &\le\frac{2\Delta}{N}\sum_{k=K\lfloor N\Delta\rfloor}^{\min((K+1)\lfloor N\Delta\rfloor-1,\lfloor Nt\rfloor-1)}\left[\left(H^N_{\xi,\tilde{\xi}_k,k}\right)^2+\sum_{\xi'}\left(H^N_{\xi,\xi',k}\right)^2\pi^{g^N_k}(\xi')\right],\end{aligned}$$

so by (3.4.122),

$$\mathbb{E}|r_t^N(\xi)|^2 \leq \frac{C_T\Delta\lfloor N\Delta\rfloor}{N} \leq C_T\Delta^2 \leq \frac{C_T}{K^2}. \tag{3.4.126}$$

Moreover,

$$\begin{aligned}
\mathbb{E}[J_{1,t}^N(\xi)]^2 &\leq \frac{K\lfloor N\Delta\rfloor}{N^2}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\mathbb{E}\left[H^N_{\xi,\tilde{\xi}_k,k} - H^N_{\xi,\tilde{\xi}_k,j\lfloor N\Delta\rfloor}\right]^2 \\
&\leq \frac{T}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\left(\frac{C_T(k-j\lfloor N\Delta\rfloor)}{N}\right)^2 \\
&\leq \frac{T}{N}\sum_{j=0}^{K-1}\sum_{k=0}^{\lfloor N\Delta\rfloor-1}\left(\frac{kC_T}{N}\right)^2 \\
&\leq \frac{TC_T^2}{3N^3}\sum_{j=0}^{K-1}\lfloor N\Delta\rfloor^3 \leq KC_T\Delta^3 \leq \frac{C_T}{K^2}.
\end{aligned} \tag{3.4.127}$$

We can similarly control $J_{3,t}^N(\xi)$ as followed:

$$\begin{aligned}
\mathbb{E}[J_{3,t}^N(\xi)]^2 &\leq \frac{K\lfloor N\Delta\rfloor}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\mathbb{E}\left[\sum_{\xi'}\left(H^N_{\xi,\xi',j\lfloor N\Delta\rfloor} - H^N_{\xi,\xi',k}\right)\pi^{g_k^N}(\xi')\right]^2 \\
&\leq \frac{K\lfloor N\Delta\rfloor}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\mathbb{E}\left[\sum_{\xi'}\left(H^N_{\xi,\xi',j\lfloor N\Delta\rfloor} - H^N_{\xi,\xi',k}\right)^2\pi^{g_k^N}(\xi')\right] \\
&\leq \frac{T}{N}\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\left(\frac{C_T(k-j\lfloor N\Delta\rfloor)}{N}\right)^2 \leq \frac{C_T}{K^2}.
\end{aligned} \tag{3.4.128}$$

Finally, note that

$$H^N_{\xi,\tilde{\xi}_k,j\lfloor\Delta N\rfloor} - \sum_{\xi'\in\mathcal{X}\times\mathcal{A}} H^N_{\xi,\xi',j\lfloor\Delta N\rfloor}\pi^{g_k^N}(\xi') = \sum_{\xi'} H^N_{\xi,\xi',j\lfloor\Delta N\rfloor}\left[\mathbb{1}_{\{\xi_k=\xi'\}} - \pi^{g_k^N}(\xi')\right]. \tag{3.4.129}$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\left|J_{2,t}^N(\xi)\right| \\
=&\frac{1}{N}\mathbb{E}\left|\sum_{j=0}^{K-1}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}H_{\xi,\xi',j\lfloor N\Delta\rfloor}^N\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k^N}(\xi')\right]\right| \\
\leq&\frac{1}{N}\mathbb{E}\left|\sum_{j=0}^{K-1}\left(\max_{\xi'}H_{\xi,\xi',j\lfloor N\Delta\rfloor}^N\right)\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k^N}(\xi')\right]\right| \\
\overset{\text{(CS)}}{\leq}&\frac{1}{N}\mathbb{E}\left[\left(\sum_{j=0}^{K-1}\left(\max_{\xi'}H_{\xi,\xi',j\lfloor N\Delta\rfloor}^N\right)^2\right)^{1/2}\left(\sum_{j=0}^{K-1}\left(\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k^N}(\xi')\right]\right)^2\right)^{1/2}\right] \\
\overset{\text{(CS)}}{\leq}&\frac{1}{N}\left[\mathbb{E}\left(\sum_{j=0}^{K-1}\left(\max_{\xi'}H_{\xi,\xi',j\lfloor N\Delta\rfloor}^N\right)^2\right)\mathbb{E}\left(\sum_{j=0}^{K-1}\left(\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k^N}(\xi')\right]\right)^2\right)\right]^{1/2} \\
\overset{(3.4.122)}{\leq}&\frac{KC_T}{N}\left[\mathbb{E}\left[\frac{1}{K}\sum_{j=0}^{K-1}\left(\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k^N}(\xi')\right]\right)^2\right]\right]^{1/2} \\
=&\frac{KC_T\lfloor N\Delta\rfloor}{N}\left[\mathbb{E}\left[\frac{1}{K}\sum_{j=0}^{K-1}\left(\frac{1}{\lfloor N\Delta\rfloor}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k^N}(\xi')\right]\right)^2\right]\right]^{1/2} \\
\leq&TC_T\underbrace{\left[\mathbb{E}\left[\frac{1}{K}\sum_{j=0}^{K-1}\left(\frac{1}{\lfloor N\Delta\rfloor}\sum_{k=j\lfloor N\Delta\rfloor}^{(j+1)\lfloor N\Delta\rfloor-1}\sum_{\xi'}\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k^N}(\xi')\right]\right)^2\right]\right]^{1/2}}_{\to 0} \\
\overset{n\to\infty}{\to}&0,
\end{aligned}
\tag{3.4.130}
$$

where step (CS) is by Cauchy-Schwarz inequality. Combining (3.4.114), (3.4.122) and (3.4.130), we have

$$
\lim_{N\to\infty}\mathbb{E}\left|I_{3,j}^N\right|=0. \tag{3.4.131}
$$

Consequently $\mathbb{E}|I_t^{2,N}(\xi)|\to 0$, and so is $M_t^{3,N}(\xi)$. The proof of the convergence for $M_t^{1,N}, M_t^{2,N}$ are exactly the same for $M_t^{3,N}$. The proof is completed. □

Let ρ^N denotes the probability measure of $\left(\mu_t^N,\nu_t^N,P_t^N,Q_t^N\right)_{0\leq t\leq T}$, which takes value in the set of probability measures $\mathcal{M}\left(D_E([0,T])\right)$. From the relative compactness result in Section 3.4.2, we know that the sequence of measures $\{\rho^N\}_{N\in\mathbb{N}}$ contains a subsequence ρ^{N_k} that converges weakly. Now we can prove the limit points of any convergence subsequence ρ^{N_k} will satisfy the limiting ODEs (3.3.9).

Lemma 3.4.20. *Let ρ^N be the probability measure of (μ^N,ν^N,P^N,Q^N). We restrict ourselves to a convergent subsequence ρ^{N_k} which converges to some limit point $\rho=(\mu,\nu,P,Q)$. Then ρ is a Dirac measure on $D_E([0,T])$ such that (μ,ν,P,Q) satisfies the limiting ODEs (3.3.9).*

Proof. For any sequence of time-points $0 \leq s_1 < s_2 < ... < s_p \leq t$, functions $\varphi, \bar{\varphi} \in C_b^2(\mathbb{R}^{1+d})$, $\phi_1, ..., \phi_p, \bar{\phi}_1, ..., \bar{\phi}_p \in C_b(\mathbb{R}^{1+d})$ and $\psi_1, ..., \psi_p, \bar{\psi}_1, ..., \bar{\psi}_p \in C_b(\mathcal{X} \times \mathcal{A})$, and consider a map $F : D_E([0,T]) \to \mathbb{R}^+$, defined as

$$F(\mu, \nu, P, Q) = F_1(\mu) + F_2(\nu) + F_3(\mu, \nu, P, Q) + F_4(\mu, \nu, P, Q), \tag{3.4.132}$$

where

$$F_1(\mu) = \left| (\langle \bar{\varphi}, \mu_t \rangle - \langle \bar{\varphi}, \mu_0 \rangle) \times \prod_{j=1}^{p} \langle \bar{\phi}_j, \mu_{s_j} \rangle \right|, \tag{3.4.133}$$

$$F_2(\nu) = \left| (\langle \varphi, \nu_t \rangle - \langle \varphi, \nu_0 \rangle) \times \prod_{j=1}^{p} \langle \phi_j, \nu_{s_j} \rangle \right|, \tag{3.4.134}$$

$$\begin{aligned} &F_3(\mu, \nu, P, Q) \\ &= \sum_{\xi \in \mathcal{X} \times \mathcal{A}} \left| Q_t(\xi) - Q_0(\xi) - \alpha \int_0^t \sum_{\xi' = (x', a')} \left(r(\xi') + \gamma \sum_{z, a''} Q_s(z, a'') g_s(z, a'') p(z|\xi') - Q_s(x', a') \right) \right. \\ &\quad \left. \times \left(\left\langle \sigma(w \cdot \xi') \sigma(w \cdot \xi) + c^2 \sigma'(w \cdot \zeta') \sigma'(w \cdot \zeta) \zeta \cdot \zeta', \nu_s \right\rangle \right) \pi^{g_s}(\xi') \, ds \right| \times \prod_{j=1}^{p} |\psi_j(Q_{s_j})|. \end{aligned} \tag{3.4.135}$$

$$\begin{aligned} &F_4(\mu, \nu, P, Q) \\ &= \sum_{\xi \in \mathcal{X} \times \mathcal{A}} \left| P_t(\xi) - P_0(\xi) - \int_0^t \sum_{\xi' = (x', a')} \zeta_s Q_s(\xi') \sigma^{g_s}(x', a') \right. \\ &\quad \cdot \Big(\left\langle \sigma(w \cdot \zeta') \sigma(w \cdot \zeta) + c^2 \sigma'(w \cdot \xi) \sigma(w \cdot \xi) (\xi' \cdot \xi), \mu_s \right\rangle \\ &\quad \left. - \sum_{a''} f_s(x', a'') \left\langle \sigma(w \cdot \zeta') \sigma(w \cdot \zeta) + c^2 \sigma'(w \cdot \xi) \sigma(w \cdot \xi) (\xi' \cdot \xi), \mu_s \right\rangle \Big) \, ds \right| \times \prod_{j=1}^{p} |\bar{\psi}_j(P_{s_j})|, \end{aligned} \tag{3.4.136}$$

where

$$f_t = \text{Softmax}(P_t), \quad g_t = \frac{\eta_t}{d_a} + (1 - \eta_t) f_t \tag{3.4.137}$$

Then we have

$$\mathbb{E}_{\rho^N}[F(\mu, \nu, P, Q)] = \mathbb{E}\left[F(\mu^N, \nu^N, P^N, Q^N)\right] \tag{3.4.138}$$

Let us analyse each term of $\mathbb{E}\left[F(\mu^N, \nu^N, P^N, Q^N)\right]$ one by one. Firstly, (3.4.72) and the boundedness of $\bar{\phi}_j$ yields

$$\mathbb{E}[F_1(\mu^N)] \leq C\mathbb{E}\left|\langle \bar{\varphi}, \mu_t^N \rangle - \langle \bar{\varphi}, \mu_0^N \rangle\right| \leq \frac{C_T}{\sqrt{N}} + \frac{C_T}{N^{3/2}} \overset{N \to \infty}{\to} 0.$$

Similarly, (3.4.72) and the boundedness of ϕ_j yields

$$\mathbb{E}[F_2(\nu^N)] \leq C\mathbb{E}\left|\langle \varphi, \nu_t^N \rangle - \langle \varphi, \nu_0^N \rangle\right| \leq \frac{C_T}{\sqrt{N}} + \frac{C_T}{N^{3/2}} \overset{N \to \infty}{\to} 0.$$

To study the next two term, we define

$$f_t^N = \text{Softmax}(P_t^N), \quad \tilde{g}_t^N = \frac{\eta_t}{d_a} + (1 - \eta_t) f_t^N, \tag{3.4.139}$$

$$E_t^{1,N}(\xi) = \int_0^t \sum_{\xi'} \mathsf{B}^N_{\xi,\xi',s}(\pi^{\tilde{g}_s^N}(\xi') - \pi^{g_s^N}(\xi')) \left[r(\xi') + \gamma \sum_{z,a''} Q_s^N(z,a'')\tilde{g}_s^N(z,a'')p(z|\xi') - Q_s^N(\xi') \right] ds, \tag{3.4.140}$$

$$E_t^{2,N}(\xi) = \int_0^t \sum_{\xi'} \mathsf{B}^N_{\xi,\xi',s}\pi^{g_s^N}(\xi')\gamma \sum_{z,a''} Q_s^N(z,a'')(\tilde{g}_s^N(z,a'') - g_s^N(z,a''))p(z|\xi')\, ds. \tag{3.4.141}$$

Then by (3.4.51):

$$\begin{aligned}
&F_3(\mu^N,\nu^N,P^N,Q^N) \\
&= \sum_{\xi\in\mathcal{X}\times\mathcal{A}} \Bigg| Q_t^N(\xi) - Q_0^N(\xi) - \alpha \int_0^t \sum_{\xi'} \mathsf{B}^N_{\xi,\xi',s}\pi^{\tilde{g}_s^N}(\xi') \\
&\cdot \left[r(\xi') + \gamma \sum_{z,a''} Q_s^N(z,a'')\tilde{g}_s^N(z,a'')p(z|\xi') - Q_s^N(\xi') \right] ds \Bigg| \times \prod_{j=1}^p |\psi_j(Q_{s_j})| \\
&= \sum_{\xi\in\mathcal{X}\times\mathcal{A}} \Bigg| Q_t^N(\xi) - Q_0^N(\xi) - \alpha \int_0^t \sum_{\xi'} \mathsf{B}^N_{\xi,\xi',s}\pi^{g_s^N}(\xi') \\
&\cdot \left[r(\xi') + \gamma \sum_{z,a''} Q_s^N(z,a'')g_s^N(z,a'')p(z|\xi') - Q_s^N(\xi') \right] ds + E_t^{1,N}(\xi) + E_t^{2,N}(\xi) \Bigg| \times \prod_{j=1}^p |\psi_j(Q_{s_j})| \\
&\overset{(3.4.51)}{=} \sum_{\xi\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}} \alpha \left| M_t^{1,N}(\xi) + M_t^{2,N}(\xi) + M_t^{3,N}(\xi) + E_t^{1,N}(\xi) + E_t^{2,N}(\xi) + O_p(N^{-1/2}) \right| \times \prod_{j=1}^p |\psi_j(Q_{s_j})|.
\end{aligned} \tag{3.4.142}$$

Recall by Assumption 3.2.7 that the stationary measures π^g are globally Lipschitz in g, so for any ξ' and $s \le NT$

$$\begin{aligned}
|\pi^{\tilde{g}_s^N}(\xi') - \pi^{g_s^N}(\xi')| &\le C \sup_{\xi'} |\tilde{g}_s^N(\xi') - g_s^N(\xi')| \\
&\le C|\eta^N_{\lfloor Ns \rfloor} - \eta^N_s| \\
&= \frac{C}{1+\log^2(\frac{\lfloor Ns \rfloor}{N}+1)} - \frac{C}{1+\log^2(s+1)} \\
&\le C\left(\log^2(s+1) - \log^2\left(\frac{\lfloor Ns \rfloor}{N} + 1 \right) \right) \\
&\le C\left(\log^2\left(\frac{\lfloor Ns \rfloor + 1}{N} + 1 \right) - \log^2\left(\frac{\lfloor Ns \rfloor}{N} + 1 \right) \right) \le \frac{C}{N},
\end{aligned} \tag{3.4.143}$$

owing to the fact that $\log^2(\cdot)$ is 1-Lipschitz. We therefore have

$$\begin{aligned}
\mathbb{E}[E_t^{1,N}(\xi)] &\le \frac{C}{N}\mathbb{E}\left[\int_0^t \sum_{\xi'} |\mathsf{B}^N_{\xi,\xi',s}| \left[|r(\xi')| + \gamma \sum_{z,a''} |Q_s^N(z,a'')||g_s^N(z,a'')|p(z|\xi') - Q_s^N(\xi') \right] ds \right] \\
&\le \frac{1}{N}\mathbb{E}\left[TC_T \sup_{\xi} |Q_s^N(\xi)| \right] \le \frac{C_T}{N},
\end{aligned}$$

and

$$E_t^{2,N}(\xi) = \frac{C}{N}\mathbb{E}\left[\int_0^t \sum_{\xi'} \mathsf{B}^N_{\xi,\xi',s}\pi^{g_s^N}(\xi')\gamma \sum_{z,a''} |Q_s^N(z,a'')|p(z|\xi')\, ds \right] \le \frac{1}{N}\mathbb{E}\left[TC_T \sup_{\xi} |Q_s^N(\xi)| \right] \le \frac{C_T}{N}.$$

Finally, we have

$$\begin{aligned}&\mathbb{E}[F_3(\mu^N,\nu^N,P^N,Q^N)]\\ \leq& C\sum_{\xi}\left[\mathbb{E}|M_t^{1,N}(\xi)|+\mathbb{E}|M_t^{2,N}(\xi)|+\mathbb{E}|M_t^{3,N}(\xi)|+\mathbb{E}|E_t^{1,N}(\xi)|+\mathbb{E}|E_t^{2,N}(\xi)|\right]\\ \overset{N\to\infty}{\to}& 0.\end{aligned}$$

To study the final term, we define

$$E_t^{3,N}(\xi)=\int_0^t\zeta_s\sum_{\xi'}(\sigma_{\rho_0}^{\tilde{g}_s^N}(\xi')-\sigma_{\rho_0}^{g_s^N}(\xi'))\,\mathsf{clip}(Q_s^N(\xi'))\left[\bar{\mathsf{B}}^N_{\xi,\xi',s}-\sum_{a''}f_s^N(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),s}\right]ds, \tag{3.4.144}$$

$$E_t^{4,N}(\xi)=\int_0^t(\zeta^N_{\lfloor Ns\rfloor}-\zeta_s)\sum_{\xi'}\sigma_{\rho_0}^{g_s^N}(\xi')\,\mathsf{clip}(Q_s^N(\xi'))\left[\bar{\mathsf{B}}^N_{\xi,\xi',s}-\sum_{a''}f_s^N(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),s}\right]ds, \tag{3.4.145}$$

Then

$$\begin{aligned}&F_4(\mu^N,\nu^N,P^N,Q^N)\\ =&\sum_{\xi\in\mathcal{X}\times\mathcal{A}}\left|P_t^N(\xi)-P_0^N(\xi)-\int_0^t\sum_{\xi'}\zeta_sQ_s^N(\xi')\sigma^{\tilde{g}_s^N}(\xi')\left[\bar{\mathsf{B}}^N_{\xi,\xi',s}-\sum_{a''}f_s^N(x',a'')\bar{\mathsf{B}}_{\xi,(x',a''),s}\right]ds\right|\\ &\times\prod_{j=1}^p|\bar{\psi}_j(P_{s_j})|,\\ =&\sum_{\xi\in\mathcal{X}\times\mathcal{A}}\left|P_t^N(\xi)-P_0^N(\xi)-\int_0^t\sum_{\xi'}\zeta_{\lfloor Ns\rfloor}Q_s^N(\xi')\sigma^{g_s^N}(\xi')\left[\bar{\mathsf{B}}^N_{\xi,\xi',s}-\sum_{a''}f_s^N(x',a'')\bar{\mathsf{B}}_{\xi,(x',a''),s}\right]ds\right.\\ &\left.+E_t^{3,N}(\xi)+E_t^{4,N}(\xi)\right|\times\prod_{j=1}^p|\bar{\psi}_j(P_{s_j})|,\\ =&\sum_{\xi\in\mathcal{X}\times\mathcal{A}}|E_t^{3,N}(\xi)+E_t^{4,N}(\xi)+M_t^N(\xi)+O(N^{-1/2})|\times\prod_{j=1}^p|\bar{\psi}_j(P_{s_j})|.\end{aligned}$$

Notice that the stationary measures σ^g are globally Lipschitz in g by Assumption 3.2.7, so using a similar argument, we prove that

$$|\sigma^{\tilde{g}_s^N}(\xi')-\sigma^{g_s^N}(\xi')|\leq\frac{C}{N}. \tag{3.4.146}$$

In addition, we have

$$\sup_{\xi,\xi'}\left|\bar{\mathsf{B}}^N_{\xi,\xi',s}-\sum_{a''}f_s^N(x',a'')\bar{\mathsf{B}}^N_{\xi,(x',a''),s}\right|\leq\sup_{\xi,\xi'}\left[\left|\bar{\mathsf{B}}^N_{\xi,\xi',s}\right|+\sum_{a''}f_s^N(x',a'')\left|\bar{\mathsf{B}}_{\xi,(x',a''),s}\right|\right]\leq C_T$$

as a result of $\bar{B}^N_{\xi,\xi',s}$ being uniformly bounded by Lemma 3.4.3 whenever $s\leq T$. Therefore for any $t\leq T$,

$$E_t^{3,N}(\xi)\leq T\times\frac{C_T}{N}\times 2\times C_T=\frac{C_T}{N}.$$

Similarly,

$$\begin{aligned}\left|E_t^{4,N}(\xi)\right| &\le C_T \int_0^T \left|\zeta^N_{\lfloor Ns\rfloor} - \zeta_s\right| ds \\ &\le \sum_{k=0}^{\lfloor NT\rfloor -1} \int_{k/N}^{(k+1)/N} \left|\frac{1}{1+k/N} - \frac{1}{1+s}\right| ds \\ &\le C_T \sum_{k=0}^{\lfloor NT\rfloor -1} \frac{1}{N^2} = \frac{C_T}{N}.\end{aligned}$$

Combining with the boundedness of $\tilde{\phi}_p$, we have

$$F_4(\mu^N,\nu^N,P^N,Q^N) \le C\sum_{\xi}\left[\mathbb{E}|E_t^{3,N}(\xi)| + \mathbb{E}|E_t^{4,N}(\xi)| + \mathbb{E}|M_t^N(\xi)| + O(N^{-1/2})\right] \overset{N\to\infty}{\to} 0. \tag{3.4.147}$$

Combining the above analysis yields:

$$\mathbb{E}_{\rho^N}[F(\mu,\nu,P,Q)] \overset{N\to\infty}{\to} 0.$$

But since F is uniformly bounded, by bounded convergence theorem, we have

$$\mathbb{E}_{\rho}[F(\mu,\nu,P,Q)] = 0.$$

This holds for any choice of the test functions $\varphi, \bar{\varphi}, \phi_j, \bar{\phi}_j, \psi_j, \bar{\psi}_j$, so we know that ρ is a Dirac measure concentrated on a solution that satisfies the evolution equation. □

3.4.4 Existence and uniqueness of solutions to limit ODEs

To complete the proof, it suffices to show that there exists a unique solution for the ODEs (3.3.9). Here we treat (Q,P) as a vector of size $2M$ with $M = \#\boldsymbol{\mathcal{X}} \times \#\boldsymbol{\mathcal{A}}$ as defined in assumption 3.2.2.

$$\frac{d}{dt}\begin{pmatrix}Q_t\\P_t\end{pmatrix} = F(t,Q_t,P_t) = \begin{pmatrix}F_1(t,Q_t,P_t)\\F_2(t,Q_t,P_t)\end{pmatrix} \tag{3.4.148}$$

where the first M entries $F(Q,P)$ are specified as

$$\begin{aligned}&F_1(t,Q,P)(x,a)\\ =&\alpha\sum_{x',a'}\bar{A}_{x,a,x',a'}\pi^{g_t(P)}(x',a')\left(r(x',a') + \gamma\sum_{z,a''}Q(z,a'')[g_t(P)](z,a'')p(z|x',a') - Q(x',a')\right)\end{aligned}$$

and the remaining M entries are specified as

$$F_2(t,Q,P)(x,a) = \sum_{x',a'}\zeta_t\mathsf{clip}(Q(x',a'))\left[A_{x,a,x',a'} - \sum_{a''}[f(P)](x',a'')A_{x,a,x',a''}\right]\sigma^{g_t(P)}(x',a').$$

Here the notation $f(P)$ and $g_t(P)$ denote the (probability) vectors in $\mathbb{R}^M$:

$$\begin{aligned}[f(P)](x,a) = \mathrm{Softmax}(P)(x,a) &= \frac{\exp\left(P(x,a)\right)}{\sum_{a''}\exp\left(P(x,a)\right)}\\ [g_t(P)](x,a) &= \frac{\eta_t}{|\mathcal{A}|} + (1-\eta_t)[f(P)](x,a).\end{aligned}$$

We will show the global existence of solutions for $t \in [0,\infty)$ by taking the usual route of showing that $F(Q,P)$ is locally Lipschitz and linearly bounded.

Lemma 3.4.21. *Let $\|\cdot\|_\infty$ be the infinity norm as defined in remark 3.3.6. Then for all $R > 0$, there is a constant $C_R > 0$ that only depends on R such that for all $(Q, P), (\tilde{Q}, \tilde{P})$ lying in the open R-ball, we have*

$$\left\|F(t,Q,P) - F(t,\tilde{Q},\tilde{P})\right\|_\infty \leq C_R \left\|(Q,P) - (\tilde{Q},\tilde{P})\right\|_\infty, \quad \forall t \geq 0. \tag{3.4.149}$$

Moreover, there is a constant $C > 0$ such that for all Q, P, we have

$$\|F(t,Q,P)\|_\infty \leq C \|(Q,P)\|_\infty + C, \quad \forall t \geq 0. \tag{3.4.150}$$

Therefore, F is locally Lipschitz and linearly bounded and for any fixed starting point (Q_0, P_0), there exists the unique solution for ODE (3.4.148).

We emphasise that the above lemma will also be true for any other norms on $\mathbb{R}^{2M}$, as pointed out in remark 3.3.6, as any norms in $\mathbb{R}^{2M}$ are equivalent with $\|\cdot\|_\infty$.

Proof. Let us first prove equation (3.4.150). Note that the tensor $\bar{A}_{\xi,\xi'}$ is uniformly bounded by assumptions 3.2.9 and 3.3.1. Thus

$$\begin{aligned}
|F_1(t,Q,P)(x,a)| &\leq C \sum_{x',a'} \pi^{g_t(P)}(x',a') \left(|r(x',a')| + \gamma \sum_{z,a''} |Q(z,a'')| g(z,a'') p(z|x',a') + |Q(x',a')| \right) \\
&\leq C \sup_{x',a'} |r(x',a')| + C\gamma \sup_{z,a''} |Q(z,a'')| + C\gamma \sup_{x',a'} |Q(x',a')| \\
&\leq C + C \|(Q,P)\|_\infty .
\end{aligned}$$

It is also clear that

$$|F_2(t,Q,P)(x,a)| \leq C \sup_{x,a} |\mathsf{clip}(Q(x,a))| \leq C$$

This shows that F is linearly bounded.

To prove the local Lipschitz condition (3.4.149), note that for all x, a,

$$\begin{aligned}
&\left|F_1(t,Q,P)(x,a) - F_1(t,\tilde{Q},\tilde{P})(x,a)\right| \\
\leq &\alpha \sum_{x',a'} |A_{x,a,x',a'}| \left|\pi^{g_t(P)}(x',a') - \pi^{g_t(\tilde{P})}(x',a')\right| \\
&\cdot \underbrace{\left| r(x',a') + \gamma \sum_{z,a''} Q(z,a'')[g_t(P)](z,a'') p(z|x',a') - Q(x',a') \right|}_{\leq C + (\gamma+1)R} \\
&+ \alpha \sum_{x',a'} |A_{x,a,x',a'}| \pi^{g_t(\tilde{P})}(x',a') \left| \sum_{z,a''} \gamma (Q(z,a'')[g_t(P)](z,a'') \right. \\
&\left. - \tilde{Q}(z,a'')[g_t(\tilde{P})](z,a'')) p(z|x',a') - (Q(x',a') - \tilde{Q}(x',a')) \right|.
\end{aligned} \tag{3.4.151}$$

Using the Lipschitz continuity of the softmax function and Assumption 3.2.7, we know

$$\begin{aligned}\sup_{x,a}|[\pi^{g_t(P)}](x,a)-[\pi^{g_t(\tilde{P})}](x,a)| &\le C\sup_{x,a}|[g_t(P)](x,a)-[g_t(\tilde{P})](x,a)| \\ &= C\sup_{x,a}|[f(P)](x,a)-[f(\tilde{P})](x,a)| \\ &\le C\left\|P-\tilde{P}\right\|_\infty .\end{aligned} \tag{3.4.152}$$

Note that for all z, a''

$$\begin{aligned}&\left|Q(z,a'')[g_t(P)](z,a'')-\tilde{Q}(z,a'')[g_t(\tilde{P})](z,a'')\right| \\ \le&|Q(z,a'')|\cdot\left|[g_t(P)](z,a'')-[g_t(\tilde{P})](z,a'')\right|+[g_t(\tilde{P})](z,a'')\cdot\left|Q(z,a'')-\tilde{Q}(z,a'')\right| \\ \le&CR\left(\sup_{z,a''}\left|P(z,a'')-\tilde{P}(z,a'')\right|\right)+\sup_{z,a''}\left|Q(z,a'')-\tilde{Q}(z,a'')\right| \\ \le&CR\left\|(Q,P)-(\tilde{Q},\tilde{P})\right\|_\infty .\end{aligned} \tag{3.4.153}$$

Combining (3.4.151), (3.4.152) and (3.4.153), we have

$$\left|[F_1(t,Q,P)](x,a)-[F_1(t,\tilde{Q},\tilde{P})]\right|\le C_R\left\|(Q,P)-(\tilde{Q},\tilde{P})\right\|_\infty . \tag{3.4.154}$$

Similarly for F_2,

$$\begin{aligned}&\left|[F_2(t,Q,P)](x,a)-[F_2(t,\tilde{Q},\tilde{P})](x,a)\right| \\ \le&\sum_{x',a'}\zeta_t\left|\mathsf{clip}(Q(x',a'))-\mathsf{clip}(\tilde{Q}(x',a'))\right|\sigma^{g_t(P)}(x',a')\left|A_{x,a,x',a'}-\sum_{a''}[f(P)](x',a'')A_{x,a,x',a''}\right| \\ +&\sum_{x',a'}\zeta_t\left|\mathsf{clip}(\tilde{Q}(x',a'))\right|\left|\sigma^{g_t(P)}(x',a')-\sigma^{g_t(\tilde{P})}(x',a')\right|\left|A_{x,a,x',a'}-\sum_{a''}[f(P)](x',a'')A_{x,a,x',a''}\right| \\ +&\sum_{x',a'}\zeta_t\left|\mathsf{clip}(\tilde{Q}(x',a'))\right|\sigma^{g_t(\tilde{P})}(x',a')\left|\sum_{a''}([f(P)](x',a'')-[f(\tilde{P})](x',a''))A_{x,a,x',a''}\right| \\ \le&C\left\|(Q,P)-(\tilde{Q},\tilde{P})\right\|_\infty .\end{aligned} \tag{3.4.155}$$

We therefore show that F is locally Lipschitz if we restrict (Q,P) to be inside a R-ball for any $R<\infty$.

The linear boundedness of F can guarantee that the solution grows almost exponentially. In fact, we have

$$\|(Q_t,P_t)\|\le\|(Q_0,P_0)\|+\int_0^t(C+\|(Q_s,P_s)\|C)\,ds\le(\|(Q_0,P_0)\|+Ct)+C\int_0^t\|(Q_s,P_s)\|\,ds. \tag{3.4.156}$$

which, together with Grönwall's inequality, implies

$$\|(Q_t,P_t)\|\le(\|(Q_0,P_0)\|+Ct)e^{Ct}. \tag{3.4.157}$$

Suppose the above evolution equation possesses two solutions $(Q,P)_t, (\tilde{Q},\tilde{P})_t$ that satisfies $Q_0=\tilde{Q}_0$ and $P_0=\tilde{P}_0$. Then we have

$$\frac{d}{dt}\left\|(Q_t,P_t)-(\tilde{Q}_t,\tilde{P}_t)\right\|^2\le 2\left\|(Q_t,P_t)-(\tilde{Q}_t,\tilde{P}_t)\right\|\cdot\left\|F(t,Q_t,P_t)-F(t,\tilde{Q}_t,\tilde{P}_t)\right\|.$$

Using (3.4.152), (3.4.155), (3.4.157) and replacing R in (3.4.152) by the norm $\|(Q_t, P_t)\|$ in (3.4.157), we can show that

$$\frac{d}{dt}\left\|(Q_t,P_t)-(\tilde{Q}_t,\tilde{P}_t)\right\|^2 \leq \underbrace{(C+(C+Ct)e^{Ct})}_{H(t)}\left\|(Q_t,P_t)-(\tilde{Q}_t,\tilde{P}_t)\right\|^2. \tag{3.4.158}$$

Therefore, by Gronwall's inequality, we have

$$\left\|(Q_t,P_t)-(\tilde{Q}_t,\tilde{P}_t)\right\|^2 \leq \left\|(Q_0,P_0)-(\tilde{Q}_0,\tilde{P}_0)\right\|^2 \exp\left(\int_0^t H(s)\,ds\right)=0,$$

which guarantees uniqueness. □

3.4.5 Proof of convergence

With the above preparations, now we can finish the proof of Theorem 3.3.3. Rrecall the sequence of probability measure ρ^N being the law of $\left(\mu_t^N,\nu_t^N,P_t^N,Q_t^N\right)_{0\leq t\leq T}$. We have shown by relative compactness that every subsequence of ρ^N posesses a further subsequence that weakly converges to the $\rho=(\mu,\nu,P,Q)$, which is the unique solution of the limit ODEs (3.3.9). Therefore by Prokhorov's Theorem (see [18, 56] for details), ρ^N weakly converges to ρ, and thus we can conclude that the process $\left(\mu_t^N,\nu_t^N,P_t^N,Q_t^N\right)_{0\leq t\leq T}$ weakly converges to ρ.

3.5 Analysis of the limiting ODE

We have already set up the limit ODEs for the algorithm (1) and now we study the convergence of the limit ODEs (3.3.9). To improve the readability, we first clarify some notations.

- From their definitions in (3.2.3), $V^f(x)$ and $V^f(x,a)$ are related via the formula

$$V^f(x)=\sum_a V^f(x,a)f(x,a). \tag{3.5.1}$$

- Recalling the state and state-action visiting measures ν^f and σ^f defined in (3.2.4), we have $\sigma_\mu^f(x,a)=f(x,a)\cdot\nu_\mu^f(x)$. By [83], the stationary distribution of $\widetilde{\boldsymbol{\mathcal{M}}}$ is the corresponding visitation measure of $\boldsymbol{\mathcal{M}}$. And for the MDP start from a fixed state x_0, the visiting measures are denoted by $\nu_{x_0}^f(\cdot),\sigma_{x_0}^f(\cdot,\cdot)$

- Let the advantage function of policy f denoted by

$$A^f(x,a)=V^f(x,a)-V^f(x),\quad \forall(x,a)\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}, \tag{3.5.2}$$

We recall that the gradient of a policy f parameterized by some parameter θ can be evaluated in terms of the visiting measure (3.2.4) according to the policy gradient theorem (3.2.23):

$$\nabla_\theta J(f_\theta)=\sum_{x,a}\sigma^{f_\theta}(x,a)V^{f_\theta}(x,a)\nabla_\theta\log f_\theta(x,a), \tag{3.5.3}$$

Assume that $f = \text{softmax}(P)$ be the softmax policy parameterized directly by the values $P(x,a)$, so that

$$f(x,a) = \frac{\exp\left(P(x,a)\right)}{\sum_{a''} \exp\left(P(x,a)\right)}. \tag{3.5.4}$$

Then the gradient $\nabla_P J(f)$ can be evaluated using the following formula.

Lemma 3.5.1. *Define* $\partial_{x,a} J(f) := \frac{\partial J(f)}{\partial P(x,a)}$ *and then for the policy* (3.5.4)*, by policy gradient theorem* (3.5.3)*, we have*

$$\partial_{x,a} J(f) = \sigma^f_{\rho_0}(x,a) A^f(x,a). \tag{3.5.5}$$

Proof. By the policy gradient theorem (3.5.3), we have

$$\begin{aligned}
\partial_{x,a} J(f) &= \sum_{x',a'} \nu^{f_\theta}_{\rho_0}(x') f(x',a') \mathbb{1}_{\{x'=x\}} \left[\mathbb{1}_{\{a'=a\}} - f(x',a)\right] V^f(x',a') \\
&= \sum_{a'} \nu^f_{\rho_0}(x) f(x,a') \left[\mathbb{1}_{\{a'=a\}} - f(P)(x,a)\right] V^f(x,a') \\
&= \nu^f_{\rho_0}(x) f_\theta(x,a) V^f(x,a) - \nu^f_{\rho_0}(x) f(x,a) \left[\sum_{a'} f(x,a') V^f(x,a')\right] \\
&= \nu^f_{\rho_0}(x) f(x,a) A^f(x,a) \\
&= \sigma^f_{\rho_0}(x,a) A^f(x,a).
\end{aligned} \tag{3.5.6}$$

□

3.5.1 Critic Convergence

Now we prove convergence of the critic (3.3.10), which states that the critic model will converge to the state-action value function during training. We first derive an ODE for the difference between the critic and the value function. Then, we use a comparison lemma, a two time-scale analysis, and the properties of the learning and exploration rates (3.3.2) to prove the convergence of the critic to the value function.

Recall that the value function V^{g_t} satisfies the Bellman equation

$$r(x,a) + \gamma \sum_{z,a''} V^{g_t}(z,a'') g_t(z,a'') p(z|x,a) - V^{g_t}(x,a) = 0. \tag{3.5.7}$$

Define the difference

$$\phi_t = Q_t - V^{g_t}. \tag{3.5.8}$$

Without loss of generality, we initialise the ODE as $\bar{Q}_0 = 0$. We can then finish the proof for the convergence for the critic.

Proof of (3.3.10). We first prove the convergence of $\|Q_t - V^{g_t}\|$ and then by the decay of the exploration rate ϵ_t we can get the convergence of $\|Q_t - V^{f_t}\|$. Combining (3.3.9) and (3.5.7), we get

the ODE for ϕ_t

$$\begin{aligned}\frac{d\phi_t}{dt}(x,a) &= -\alpha \sum_{x',a'} A_{x,a,x',a'} \pi^{g_t}(x',a')\phi_t(x',a') \\ &\quad + \alpha\gamma \sum_{x',a'} A_{x,a,x',a'} \pi^{g_t}(x',a') \sum_{z,a''} \phi_t(z,a'')g_t(z,a'')p(z|x',a') \\ &\quad - \frac{d}{dt}V^{g_t}(x,a).\end{aligned} \tag{3.5.9}$$

Let $\odot$ denote element-wise multiplication. Then,

$$\frac{d\phi_t}{dt} = -\alpha A(\pi^{g_t} \odot \phi_t) + \alpha\gamma A(\pi^{g_t} \odot \Gamma_t) + \frac{\partial V^{g_t}}{\partial g}\frac{dg_t}{dt}, \tag{3.5.10}$$

where $\Gamma_t(x',a') = \sum_{z,a''} \phi_t(z,a'')g_t(z,a'')p(z|x',a')$. Define the process

$$Y_t = \frac{1}{2}\phi_t^\top A^{-1}\phi_t. \tag{3.5.11}$$

Differentiating yields

$$\begin{aligned}\frac{dY_t}{dt} &= \phi_t^\top A^{-1}\frac{d\phi_t}{dt} \\ &= -\alpha\phi_t^\top \pi^{g_t} \odot \phi_t + \alpha\gamma\phi_t^\top \pi^{g_t} \odot \Gamma_t + \phi_t^\top A^{-1}\frac{\partial V^{g_t}}{\partial g}\frac{dg_t}{dt}.\end{aligned} \tag{3.5.12}$$

The second term on the last line of (3.5.12) becomes:

$$\begin{aligned}&\left|\phi_t^\top \pi^{g_t} \odot \Gamma_t\right| \\ =&\left|\sum_{x',a'} \phi_t(x',a')\pi^{g_t}(x',a') \sum_{z,a''} \phi_t(z,a'')g_t(z,a'')p(z|x',a')\right| \\ =&\left|\sum_{x',a'}\sum_{z,a''} \phi_t(z,a'')\phi_t(x',a')g_t(z,a'')p(z|x',a')\pi^{g_t}(x',a')\right| \\ \le&\sum_{x',a'}\sum_{z,a''} \left|\phi_t(z,a'')\phi_t(x',a')\right| g_t(z,a'')p(z|x',a')\pi^{g_t}(x',a') \\ \le&\frac{1}{2}\sum_{x',a'}\sum_{z,a''} \left(\phi_t(z,a'')^2 + \phi_t(x',a')^2\right) g_t(z,a'')p(z|x',a')\pi^{g_t}(x',a') \\ =&\frac{1}{2}\sum_{z,a''} \phi_t(z,a'')^2 \sum_{x',a'} g_t(z,a'')p(z|x',a')\pi^{g_t}(x',a') + \frac{1}{2}\sum_{x',a'} \phi_t(x',a')^2\pi^{g_t}(x',a') \sum_{z,a''} g_t(z,a'')p(z|x',a') \\ =&\frac{1}{2}\sum_{z,a''} \phi_t(z,a'')^2\pi^{g_t}(z,a'') + \frac{1}{2}\sum_{x',a'} \phi_t(x',a')^2\pi^{g_t}(x',a') \\ =&\sum_{x',a'} \phi_t(x',a')^2\pi^{g_t}(x',a').\end{aligned}$$

where we have used Young's inequality, the fact that $\sum_{z,a''} g_t(z,a'')p(z|x',a') = 1$ for each (x',a'), and $\sum_{x',a'} g_t(z,a'')p(z|x',a')\pi^{g_t}(x',a') = \pi^{g_t}(z,a'')$. Therefore,

$$\frac{dY_t}{dt} \le -\alpha(1-\gamma)\pi^{g_t} \cdot \phi_t^2 + \phi_t^\top A^{-1}\frac{\partial V^{g_t}}{\partial g}\frac{dg_t}{dt}, \tag{3.5.13}$$

where ϕ_t^2 is an element-wise square. By the limit ODEs in (3.3.9), we have for any (x, a)

$$\left|\frac{dP_t}{dt}(x,a)\right| = \left|\sum_{x',a'} \zeta_t \mathsf{clip}(Q_t(x',a')) \left[A_{x,a,x',a'} - \sum_{a''} f_t(x',a'') A_{x,a,x',a''}\right] \sigma^{f_t}(x',a')\right| \leq C\zeta_t \quad (3.5.14)$$

For any state x_0, define

$$\partial_{P(x,a)} V^f(x_0) := \frac{\partial V^f(x_0)}{\partial P(x,a)}.$$

Then, for the exploration policy (3.2.16), by the policy gradient theorem we have

$$\begin{aligned}
\left|\partial_{P(x,a)} V^{g_t}(x_0)\right| &= \left|\sum_{x',a'} \sigma_{x_0}^{g_t}(x',a') V^{g_t}(x',a') \partial_{P(x,a)} \log g_t(x',a')\right| \\
&\leq C \sum_{x',a'} \left|\partial_{P(x,a)} \log g_t(x',a')\right| \\
&= C(1-\eta_t) \sum_{x',a'} \frac{f_t(x',a')}{g_t(x',a')} \left|\partial_{P(x,a)} \log f_t(x',a')\right| \\
&\overset{(a)}{\leq} C,
\end{aligned} \quad (3.5.15)$$

where step (a) is by

$$\frac{f_t(x',a')}{g_{\bar{\theta}_t}(x',a')} = \frac{f_t(x',a')}{\frac{\eta_t}{d_A} + (1-\eta_t) \cdot f_t(x',a')} \leq C \quad (3.5.16)$$

and

$$\left|\partial_{P(x,a)} \log f_t(x',a')\right| = \left|\mathbb{1}_{\{x'=x\}} \left[\mathbb{1}_{\{a'=a\}} - f_t(x',a)\right]\right| \leq 2. \quad (3.5.17)$$

The relationship between the value functions

$$V^{f_t}(x_0,a_0) = r(x_0,a_0) + \gamma \sum_{x'} V^{f_t}(x') p(x'|x_0,a_0), \quad \forall (x_0,a_0), \quad (3.5.18)$$

can be combined with (3.5.15) to derive

$$\|\nabla_P V^{g_t}(x,a)\| \leq C, \quad \forall (x,a). \quad (3.5.19)$$

Combining (3.5.14) and (3.5.19),

$$\left|\frac{dV^{g_t}}{dt}(x,a)\right| = \left|\nabla_P V^{g_t}(x,a) \cdot \frac{dP_t}{dt}\right| \leq \|\nabla_P V^{g_t}(x,a)\| \cdot \left\|\frac{dP_t}{dt}\right\| \leq C\zeta_t, \quad (3.5.20)$$

where $C > 0$ is a constant independent of T.

Combining (3.5.13), (3.5.20), we have

$$\begin{aligned}
\frac{dY_t}{dt} &\leq -\alpha(1-\gamma) \min_{x,a}\{\pi^{g_t}(x,a)\} Y_t + C\phi_t^\top \zeta_t \\
&\leq -\alpha C \eta_t^{n_0} (1-\gamma) Y_t + C\phi_t^\top \zeta_t \\
&\leq -C\eta_t^{n_0} Y_t + \frac{\eta_t^{n_0}}{\eta_t^{n_0}} \|\phi_t\| C\zeta_t \\
&\leq -C\eta_t^{n_0} Y_t + \|\phi_t\|^2 \eta_t^{2n_0} + \frac{C\zeta_t^2}{\eta_t^{2n_0}} \\
&= -\eta_t^{n_0}(C - 2\eta_t^{n_0}) Y_t + \frac{C\zeta_t}{\eta_t^{2n_0}} \zeta_t.
\end{aligned} \quad (3.5.21)$$

Since $\eta_t^{n_0} \to 0$ and $\frac{\zeta_t}{\eta_t^{2n_0}} \to 0$ as $t \to \infty$, there exists $t_0 \geq 2$ such that

$$\frac{dY_t}{dt} \leq -C\eta_t^{n_0} Y_t + \zeta_t, \quad t \geq t_0, \tag{3.5.22}$$

where the C is a constant independent with t. Noting that $\frac{\zeta_t}{\eta^{n_0}} \to 0$ as $t \to \infty$, we know for any $\epsilon_0 > 0$, there exists $t_0 \geq t_0$ such that

$$\frac{d(Y_t - \epsilon_0)}{dt} \leq -C\eta_t^{n_0} \left(Y_t - \frac{\zeta_t}{\eta_t^{n_0}} \right) \leq -C\eta_t^{n_0} \left(Y_t - \epsilon_0 \right), \quad t \geq t_0, \tag{3.5.23}$$

By multiplying the integral factor $\exp\left\{\int_{t_0}^t C\eta_s^{n_0} ds\right\}$, we get

$$\frac{d}{dt}\left(\exp\left\{\int_{t_0}^t C\eta_s^{n_0} ds\right\} \cdot (Y_t - \epsilon_0)\right) \leq \exp\left\{\int_{t_0}^t C\eta_s^{n_0} ds\right\} \cdot \left(\frac{d(Y_t - \epsilon_0)}{dt} + C\eta_t^{n_0} (Y_t - \epsilon_0)\right) \leq 0, \quad t \geq t_0,$$

which derives

$$Y_t - \epsilon_0 \leq \exp\left\{-\int_{t_0}^t C\eta_s^{n_0} ds\right\} \cdot (Y_{t_0} - \epsilon_0) \to 0, \quad \text{as} \quad t \to \infty. \tag{3.5.24}$$

Thus we get for any $\epsilon_0 > 0$, there exists $t_0 > 0$, such that $Y_t \leq 2\epsilon_0$ for any $t \geq t_0$, which brings us the desired convergence for ϕ_t.

By the policy gradient theorem, we have

$$\frac{\partial V^f(x_0)}{\partial_{f(x,a)}} = V^f(x,a)\sigma_{x_0}^f(x). \tag{3.5.25}$$

Thus, by the relationship (3.5.18),

$$\frac{\partial V^f(x_0,a_0)}{\partial_{f(x,a)}} = \gamma \sum_{x'} V^{f(x,a)} \sigma_{x'}^f(x) p(x'|x_0,a_0) \leq C. \tag{3.5.26}$$

Then, for any $(x,a) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}$, there exists $\widetilde{t} \in [0,1]$ such that

$$\left|V^{g_t}(x,a) - V^{f_t}(x,a)\right| = \left|\nabla_f V^{\widetilde{t}f_t + (1-\widetilde{t})g_t}(x,a) \cdot [g_t - f_t]\right| \leq C\eta_t, \tag{3.5.27}$$

Finally, combining (3.3.10) and (3.5.27), we obtain (3.3.10). □

3.5.2 Actor Convergence

Now we show that the actor converges to a stationary point. We introduce the following notation:

$$\begin{aligned} \widehat{\nabla}_P J(f_t) &:= \sum_{x,a} \sigma_{\rho_0}^{g_t}(x,a) \bar{Q}_t(x,a) \nabla_P \log f_t(x,a), \\ \widehat{\partial}_{P(x,a)} J(f_t) &:= \sum_{x,a} \sigma_{\rho_0}^{g_t}(x,a) \bar{Q}_t(x,a) \partial_{P(x,a)} \log f_t(x,a). \end{aligned} \tag{3.5.28}$$

By the policy gradient theorem, using the similar approach as in Lemma 3.5.1 for the softmax policy $f = \text{softmax}(P)$ we have

$$\begin{aligned} \widehat{\partial}_{P(x,a)} J(f_t) =& \sum_{x',a'} Q_t(x',a') \sigma_\rho^{g_t}(x',a') \widehat{\partial}_{P(x,a)} J(f_t) \log f_t(x',a') \\ =& \sigma_{\rho_0}^{g_t}(x,a) \left[Q_t(x,a) - \sum_{a'} Q_t(x,a') f(x,a') \right] \end{aligned} \tag{3.5.29}$$

By the same method in [143] and the following lemmas, we can prove $\|\nabla_P J(f_t)\| \to 0, \quad t \to \infty$.

Lemma 3.5.2. *Let Y_t, W_t and Z_t be three functions such that W_t is nonnegative. Assume there exists $t_0 \geq 0$ such that*

$$\frac{dY_t}{dt} \geq W_t + Z_t, \quad t \geq t_0 \tag{3.5.30}$$

and that $\int_{t_0}^{\infty} Z_t dt$ converges. Then either $Y_t \to \infty$ or else Y_t converges to a finite value and $\int_0^{\infty} W_t dt < \infty$.

We may modify the above lemma so that the dichotomy holds whenever (3.5.30) holds for $t \geq T$. Now we can prove the convergence for the actor.

Proof of theorem 3.3.5. Let f be the softmax policy in (3.5.4), by the proof of Lemma 7 in [104], we know that the eigenvalues of the Hessian matrix of $J(f_\theta)$ w.r.t. P are smaller than $L := \frac{8}{(1-\gamma)^3}$ and thus $\nabla_P J(f)$ is L-Lipschitz continuous with respect to P.

For the limit ode of P_t in (3.3.9), define

$$Y_t := A^{-1} P_t$$

Then

$$\begin{aligned}
&\frac{dY_t}{dt}(x,a) \\
=&\sum_{z,b} \left(A^{-1}\right)_{x,a,z,b} \frac{dP_t}{dt}(z,b) \\
=&\sum_{z,b} \left(A^{-1}\right)_{x,a,z,b} \sum_{x',a'} \zeta_t \mathsf{clip}(Q_t(x',a')) \left[A_{z,b,x',a'} - \sum_{a''} f_t(x',a'') A_{z,b,x',a''}\right] \sigma_{\rho_0}^{g_t}(x',a') \\
=&\zeta_t \sum_{x',a'} \mathsf{clip}(Q_t(x',a')) \sigma_{\rho_0}^{g_t}(x',a') \left[\sum_{z,b} \left(A^{-1}\right)_{x,a,z,b} A_{z,b,x',a'} - \sum_{z,b,a''} f_t(x',a'') \left(A^{-1}\right)_{x,a,z,b} A_{z,b,x',a''}\right] \\
=&\zeta_t \sum_{x',a'} \mathsf{clip}(Q_t(x',a')) \sigma_{\rho_0}^{g_t}(x',a') \left[\mathbb{1}_{\{x'=x,\ a'=a\}} - \sum_{a''} f_t(x',a'') \mathbb{1}_{\{x'=x,\ a''=a\}}\right] \\
=&\zeta_t \mathsf{clip}(Q_t(x,a)) \sigma_{\rho_0}^{g_t}(x,a) - \sum_{a'} \mathsf{clip}(Q_t(x,a')) \sigma_{\rho_0}^{g_t}(x,a') f_t(x,a) \\
=&\zeta_t \sigma_{\rho_0}^{g_t}(x,a) \left[\mathsf{clip}(Q_t(x,a)) - \sum_{a'} \mathsf{clip}(Q_t(x,a')) f_t(x,a')\right].
\end{aligned} \tag{3.5.31}$$

Thus we get the ode for Y_t:

$$\frac{dY_t}{dt}(x,a) = \zeta_t \sigma_{\rho_0}^{g_t}(x,a) \left[\mathsf{clip}(Q_t(x,a)) - \sum_{a'} \mathsf{clip}(Q_t(x,a')) f_t(x,a')\right] \tag{3.5.32}$$

Since we know that $\left\|Q_t - V^{f_t}\right\| \to 0$, we know that there is a T for which $\mathsf{clip}(Q_t) = Q_t$ whenever $t \geq T$. Thus we have

$$\frac{dP_t}{dt} = \zeta_t A \hat{\nabla}_P J(f_t) \quad t \geq T \tag{3.5.33}$$

By chain rule and note that A is a positive definiteness matrix, we get for all $t \geq T$:

$$\frac{d}{dt} J(f_t) = \nabla_P J(f_t) \cdot \frac{dP_t}{dt} \geq C\zeta_t \lambda_1 \|\nabla_P J(f_t)\|^2 - C\zeta_t \eta_t \tag{3.5.34}$$

Then, by Lemma 3.5.2 and the assumption in (3.3.1), we can show that either $J(f_t) \to \infty$ or $J(f_t)$ converges to a finite value and

$$\int_0^{+\infty} \zeta_t \|\nabla_P J(f_t)\|^2 dt < \infty. \tag{3.5.35}$$

Note that $J(f) = \mathbb{E}_f \left[\sum_{k=0}^{+\infty} \gamma^k r(x_k, a_k)\right]$. Therefore, the objective function J is bounded by Assumption 3.2.2 and thus we know $J(f_t)$ converges to a finite value and (3.5.35) is valid.

If there existed an $\epsilon_0 > 0$ and $\bar{t} > 0$ such that $\|\nabla_P J(f_t)\| \geq \epsilon_0$ for all $t \geq \bar{t}$, we would have

$$\int_{\bar{t}}^{+\infty} \zeta_t \|\nabla_P J(f_t)\|^2 dt \geq \epsilon_0^2 \int_{\bar{t}}^{+\infty} \zeta_t dt = \infty, \tag{3.5.36}$$

which contradicts (3.5.35). Therefore, $\liminf_{t\to\infty} \|\nabla_P J(f_t)\| = 0$. To show that $\lim_{t\to\infty} \|\nabla_P J(f_t)\| = 0$, assume the contrary; that is $\limsup_{t\to\infty} \|\nabla_P J(f_t)\| > 0$. Then we can find a constant $\epsilon_1 > 0$ and two increasing sequences $\{a_n\}_{n\geq 1}, \{b_n\}_{n\geq 1}$ such that

$$\begin{aligned} &a_1 < b_1 < a_2 < b_2 < a_3 < b_3 < \cdots, \\ &\|\nabla_P J(f_{a_n})\| < \frac{\epsilon_1}{2}, \quad \|\nabla_P J(f_{b_n})\| > \epsilon_1. \end{aligned} \tag{3.5.37}$$

Define the following cycle of stopping times:

$$\begin{aligned} t_n &:= \sup\{s | s \in (a_n, b_n),\ \|\nabla_P J(f_s)\| < \frac{\epsilon_1}{2}\}, \\ i(t_n) &:= \inf\{s | s \in (t_n, b_n),\ \|\nabla_\theta J(f_s)\| > \epsilon_1\}. \end{aligned} \tag{3.5.38}$$

Note that $\|\nabla_P J(f_t)\|$ is continuous against t, thus we have

$$\begin{aligned} &a_n \leq t_n < i(t_n) \leq b_n \\ &\|\nabla_P J(f_{t_n})\| = \frac{\epsilon_1}{2}, \quad \|\nabla_P J(f_{i(t_n)})\| = \epsilon_1 \\ &\frac{\epsilon_1}{2} \leq \|\nabla_P J(f_s)\| \leq \epsilon_1, \quad s \in (t_n, i(t_n)). \end{aligned} \tag{3.5.39}$$

Then, by the L-Lipschitz property of the gradient, we have for any t_n

$$\begin{aligned} \frac{\epsilon_1}{2} &= \|\nabla_P J(f_{i(t_n)})\| - \|\nabla_P J(f_{t_n})\| \\ &\leq \|\nabla_P J(f_{i(t_n)}) - \nabla_P J(f_{t_n})\| \\ &\leq L\|P_{i(t_n)} - P_{t_n}\| \\ &\leq C \int_{t_n}^{i(t_n)} \zeta_s \|\nabla_P J(f_s)\| ds + C \int_{t_n}^{i(t_n)} \zeta_s \|\widehat{\nabla}_P J(f_s) - \nabla_P J(f_s)\| ds \\ &\leq C\epsilon_1 \int_{t_n}^{i(t_n)} \zeta_s ds + C \int_{t_n}^{i(t_n)} \zeta_s \eta_s ds. \end{aligned} \tag{3.5.40}$$

From this and by (3.3.2) it follows that

$$\frac{1}{2L} \leq \liminf_{n\to\infty} \int_{t_n}^{i(t_n)} \zeta_s ds. \tag{3.5.41}$$

Using (3.5.39), we see that

$$J(f_{\bar{\theta}_{i(t_n)}}) - J(f_{\bar{\theta}_{t_n}}) \geq C_1 (\frac{\epsilon_1}{2})^2 \int_{t_n}^{i(t_n)} \zeta_s ds - C_2 \int_{t_n}^{i(t_n)} \zeta_s \eta_s ds. \tag{3.5.42}$$

Due to the convergence of $J(f_{\theta_{t_n}})$ and the assumption of the learning rate, this implies that

$$\lim_{n\to\infty} \int_{t_n}^{i(t_n)} \zeta_s ds = 0, \tag{3.5.43}$$

which contradicts (3.5.41) and thus the convergence to the stationary point is proven. □

Chapter 4

Online SDE Optimization: Linear Case

4.1 Introduction

Consider a parametric process $X_t^\theta \in \mathbb{R}^d$ which satisfies the stochastic differential equation (SDE):

$$\begin{aligned} dX_t^\theta &= \mu(X_t^\theta, \theta)dt + \sigma(X_t^\theta, \theta)dW_t, \\ X_0^\theta &= x, \end{aligned} \tag{4.1.1}$$

where $\theta \in \mathbb{R}^\ell, \mu \in \mathbb{R}^d, \sigma \in \mathbb{R}^{d\times d}$, and W_t is a standard Brownian motion. Suppose X_t^θ is ergodic with the stationary distribution π_θ.[1]

Our goal is to select the parameters θ which minimize the objective function

$$J(\theta) = \sum_{n=1}^{N} \left(\mathbb{E}_{\pi_\theta}\left[f_n(Y)\right] - \beta_n\right)^2, \tag{4.1.2}$$

where Y is a random variable with distribution π_θ, f_n are known functions, and β_n are the target quantities. Thus, we are interested in optimizing the parameterized SDEs (4.1.1) such that their stationary distribution matches, as closely as possible, the target statistics β_n. In practice, the target statistics may be data from real-world observations which are then used to calibrate the SDE model (4.1.1).

4.1.1 Existing methods to optimize over the stationary distribution of SDEs

The stationary distribution π_θ is typically unknown and therefore it is challenging to optimize over $J(\theta)$. The quantity $\mathbb{E}_{Y\sim\pi_\theta}\left[f_n(Y)\right]$ as well as its gradient $\nabla_\theta \mathbb{E}_{Y\sim\pi_\theta}\left[f_n(Y)\right]$ must be estimated in order to minimize $J(\theta)$. $\mathbb{E}_{Y\sim\pi_\theta}\left[f_n(Y)\right]$ can be evaluated using the forward Kolmogorov equation

$$\mathcal{L}_x^{\theta,*} p_\infty(x,\theta) = 0, \tag{4.1.3}$$

[1] Sufficient conditions ([114]) for the existence and uniqueness of π_θ are: (1) both coefficients μ and σ are assumed to be bounded and σ is uniformly continuous with respect to x variable, (2) $\lim_{|x|\to\infty} \sup_\theta \mu(x,\theta)x = -\infty$, and (3) there exist two constants $0 < \lambda < \Lambda < \infty$ such that $\lambda I_d \le \sigma\sigma^\top(x,\theta) \le \Lambda I_d$ where I_d is the $d \times d$ identity matrix.

where $\mathcal{L}_x^\theta$ is the infinitesimal generator of the process X_t^θ and $\mathcal{L}_x^{\theta,*}$ is the adjoint operator of $\mathcal{L}_x^\theta$. $\nabla_\theta \mathbb{E}_{Y\sim\pi_\theta}[f_n(Y)]$ can be calculated using an appropriate adjoint PDE for (4.1.3) [5, 28, 59, 76]. However, if the dimension of d for X_t^θ is large, solving the forward Kolmogorov equation and its adjoint PDE becomes extremely computationally expensive. In the special case where the drift function μ is the gradient of a scalar function and the volatility function σ is constant, there exists a closed-form formula for the stationary distribution [115].

Alternatively, $\mathbb{E}_{Y\sim\pi_\theta}\left[f_n(Y)\right]$ can be approximated by simulating (4.1.1) over a long time $[0,T]$. Similar to [34], the gradient descent algorithm would be:

- Simulate $X_t^{\theta_k}$ for $t\in[0,T]$.
- Evaluate the gradient of $J_T(\theta_k) := \sum_{n=1}^{N}\left(\frac{1}{T}\int_0^T f_n(X_t^{\theta_k})dt - \beta_n\right)^2$.
- Update the parameter as $\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta J_T(\theta_k)$,

where α_k is the learning rate. This gradient descent algorithm will be slow; a long simulation time T will be required for each optimization iteration. A second disadvantage is that $J_T(\theta)$ is an approximation to $J(\theta)$ and therefore error is introduced into the algorithm, i.e. $\nabla_\theta J_T(\theta) \neq \nabla_\theta J(\theta)$.

4.1.2 An Online Optimization Algorithm

We propose a new continuous-time stochastic gradient descent algorithm which allows for computationally efficient optimization of (4.1.2). The algorithm uses **online forward propagation** to asymptotically estimate the gradient of the objective function with respect to the parameters. For notational convenience (and without loss of generality), we will set $N=1$ and $\beta_1=\beta$. The online forward propagation algorithm for optimizing (4.1.2) is:

$$\begin{aligned}
\frac{d\theta_t}{dt} &= -2\alpha_t\left(f(\bar{X}_t)-\beta\right)\left(\nabla f(X_t)\tilde{X}_t\right)^\top,\\
d\tilde{X}_t &= \left(\nabla_x\mu(X_t,\theta_t)\tilde{X}_t + \nabla_\theta\mu(X_t,\theta_t)\right)dt + \left(\nabla_x\sigma(X_t,\theta_t)\tilde{X}_t + \nabla_\theta\sigma(X_t,\theta_t)\right)dW_t,\\
dX_t &= \mu(X_t,\theta_t)dt + \sigma(X_t,\theta_t)dW_t,\\
d\bar{X}_t &= \mu(\bar{X}_t,\theta_t)dt + \sigma(\bar{X}_t,\theta_t)d\bar{W}_t,
\end{aligned} \tag{4.1.4}$$

where W_t and $\bar{W}_t$ are independent Brownian motions and α_t is the learning rate. Before proceeding with our analysis, we first clarify the notation in (4.1.4). In this chapter, the Jacobian matrix of a vector value function $f: x\in\mathbb{R}^n \to \mathbb{R}^m$ is an $m\times n$ matrix, i.e. $\nabla_x f(x)\in\mathbb{R}^{n\times m}$. When the function has only one variable, we may omit the subscript in the gradient. For example, we may use $\nabla f(x)$ to denote $\nabla_x f(x)$. For functions of several variables, we use the subscript in the gradient to denote the partial derivative with respect to a subset of variables. For example, we will use $\nabla_x\mu(X_t^\theta,\theta)$ to denote $\nabla_x\mu(x,\theta)\Big|_{x=X_t^\theta}$. Therefore, the variables have the following dimensions:

$$\tilde{X}_t\in\mathbb{R}^{d\times\ell},\quad \nabla_x\mu\in\mathbb{R}^{d\times d},\quad \nabla_\theta\mu\in\mathbb{R}^{d\times\ell},\quad \nabla_x\sigma\in\mathbb{R}^{d\times d\times d},\quad \nabla_\theta\sigma\in\mathbb{R}^{d\times d\times\ell}.$$

Let $\tilde{X}_t^i$ denote the i-th row of $\tilde{X}_t$ and then the dynamics of $\tilde{X}_t$ in (4.1.4) are:

$$d\tilde{X}_t^i = \left(\nabla_x \mu_i(X_t, \theta_t)\tilde{X}_t + \nabla_\theta \mu_i(X_t, \theta_t)\right) dt + \sum_{j=1}^{d} \left(\nabla_x \sigma_{i,j}(X_t, \theta_t)\tilde{X}_t + \nabla_\theta \sigma_{i,j}(X_t, \theta_t)\right) dW_t^j.$$

In (4.1.4), $\bar{X}_t$ and X_t have the same dynamics, although they are driven by independent Brownian motions. The role of $\bar{X}_t$ will be explained in detail later in this section. The learning rate α_t in (4.1.4) must be chosen such that $\int_0^\infty \alpha_s ds = \infty$ and $\int_0^\infty \alpha_s^2 ds < \infty$. (An example is $\alpha_t = \frac{C}{1+t}$.) $\tilde{X}_t$ estimates the derivative of X_t with respect to θ_t. The parameter θ_t is continuously updated using $\left(f(\bar{X}_t) - \beta\right)\left(\nabla f(X_t)\tilde{X}_t\right)^\top$ as a stochastic estimate for $\nabla_\theta J(\theta_t)$. Deterministic gradient descent in continuous-time is often referred to as a "gradient flow"; therefore, the proposed algorithm can be viewed as a "stochastic gradient flow".

To better understand the algorithm (4.1.4), let us first rewrite the gradient of the objective function using the ergodicity of X_t^θ:

$$\begin{aligned} \nabla_\theta J(\theta) &= 2\left(\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \beta\right) \nabla_\theta \mathbb{E}_{Y\sim\pi_\theta} f(Y) \\ &\overset{a.s.}{=} 2\left(\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_t^\theta)dt - \beta\right) \cdot \nabla_\theta \left(\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_t^\theta)dt\right). \end{aligned} \tag{4.1.5}$$

If the derivative and the limit can be interchanged, the gradient can be expressed as

$$\nabla_\theta J(\theta) = 2\left(\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_t^\theta)dt - \beta\right) \cdot \lim_{T\to\infty} \frac{1}{T}\int_0^T \nabla f(X_t^\theta)\nabla_\theta X_t^\theta dt. \tag{4.1.6}$$

Define $\tilde{X}_t^\theta = \nabla_\theta X_t^\theta$ and, under mild regularity conditions for the coefficients (see for example [123, 145]), $\tilde{X}_t^\theta$ will satisfy

$$d\tilde{X}_t^\theta = \left(\nabla_x \mu(X_t^\theta, \theta)\tilde{X}_t^\theta + \nabla_\theta \mu(X_t^\theta, \theta)\right) dt + \left(\nabla_x \sigma(X_t^\theta, \theta)\tilde{X}_t^\theta + \nabla_\theta \sigma(X_t^\theta, \theta)\right) dW_t. \tag{4.1.7}$$

Note that $\tilde{X}_t$ and $\tilde{X}_t^\theta$ satisfy the same equations, except θ is a fixed constant for $\tilde{X}_t^\theta$ while θ_t is updated continuously in time for $\tilde{X}_t$. Then, we have that

$$\nabla_\theta J(\theta) = 2\left(\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_t^\theta)dt - \beta\right) \cdot \lim_{T\to\infty} \frac{1}{T}\int_0^T \nabla f(X_t^\theta)\tilde{X}_t^\theta dt. \tag{4.1.8}$$

The formula (4.1.8) can be used to evaluate $\nabla_\theta J(\theta)$ and thus allows for optimization via a gradient descent algorithm. However, as highlighted in Section 4.1.1, X_t^θ must be simulated for a large time period $[0, T]$ for each optimization iteration, which is computationally costly. A natural alternative is to develop a *continuous-time* stochastic gradient descent algorithm which updates θ using a stochastic estimate $G(\theta_t)$ for $\nabla_\theta J(\theta_t)$, where $G(\theta_t)$ asymptotically converges to an unbiased estimate for the direction of steepest descent $\nabla_\theta J(\theta_t)$. (The random variable $G(\theta_t)$ is called an unbiased estimate for $\nabla_\theta J(\theta_t)$ if $\mathbb{E}[G(\theta_t)|\theta_t] = \nabla_\theta J(\theta_t)$.) The online algorithm (4.1.4) does exactly this using $G(\theta_t) = 2\left(f(\bar{X}_t) - \beta\right)\nabla f(X_t)\tilde{X}_t$ as a stochastic estimate for $\nabla_\theta J(\theta_t)$.

For large t, we expect that

$$\mathbb{E}\left[f(\bar{X}_t) - \beta\right] \approx \mathbb{E}_{Y\sim\pi_{\theta_t}}\left[f(Y) - \beta\right], \quad \mathbb{E}\left[\nabla f(X_t)\tilde{X}_t\right] \approx \nabla_\theta\left(\mathbb{E}_{Y\sim\pi_{\theta_t}}\left[f(X) - \beta\right]\right).$$

since θ_t is changing very slowly as t becomes large due to $\lim_{t\to\infty}\alpha_t = 0$. Here we highlight that for random variables X and Y, it is not typically true that $\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y$ unless X and Y are independent. This is the reason why the process $\bar{X}_t$ is introduced. Since $\bar{X}_t$ and X_t are driven by independent Brownian motions, we expect that $\mathbb{E}\left[2\left(f(\bar{X}_t) - \beta\right)\nabla f(X_t)\tilde{X}_t\right] \approx \nabla_\theta J(\theta_t)$ for large t due to $\bar{X}_t$ and $(X_t, \tilde{X}_t)$ becoming asymptotically independent since θ_t will be changing very slowly for large t. Thus, we expect that for large t, the stochastic sample $G(\theta_t) = 2\left(f(\bar{X}_t) - \beta\right)\nabla f(X_t)\tilde{X}_t$ will provide an asymptotically unbiased estimate for the direction of steepest descent $\nabla_\theta J(\theta_t)$ and $\|\nabla_\theta J(\theta_t)\|$ will converge to zero as $t \to \infty$.

4.1.3 Contributions of this chapter

We rigorously prove the convergence of the algorithm (4.1.4) when $\mu(\cdot)$ is linear and for constant σ. Even in the linear case, the distribution of $(X_t, \bar{X}_t, \tilde{X}_t, \theta_t)$ will be non-Gaussian and convergence analysis is non-trivial. Unlike in the traditional stochastic gradient descent algorithm, the data is not i.i.d. (i.e., X_t is correlated with X_s for $s \neq t$) and, for a finite time t, the stochastic update direction $G(\theta_t)$ is not an unbiased estimate of $\nabla_\theta J(\theta_t)$. One must show that asymptotically $G(\theta_t)$ becomes an unbiased estimate of the direction of steepest descent $\nabla_\theta J(\theta_t)$. Furthermore, it must be proven that the stochastic fluctuations of $G(\theta_t)$ around the direction of steepest descent vanish in an appropriate way as $t \to \infty$.

The proof therefore requires analysis of the fluctuations of the stochastic update direction $G(\theta_t)$ around $\nabla_\theta J(\theta_t)$. Bounds on the fluctuations are challenging to obtain due to the online nature of the algorithm. The stationary distribution π_{θ_t} will continuously change as the parameters θ_t evolve. We prove bounds on a new class of Poisson partial differential equations, which are then used to analyze the parameter fluctuations in the algorithm. The fluctuations are re-written in terms of the solution to the Poisson PDE using Ito's Lemma, the PDE solution bounds are subsequently applied, and then we can show asymptotically that the fluctuations vanish. Our main theorem proves for the multi-dimensional Ornstein-Uhlenbeck process that:

$$\lim_{t\to\infty}|\nabla_\theta J(\theta_t)| \overset{a.s.}{=} 0. \tag{4.1.9}$$

In the numerical section of this chapter, we evaluate the performance of our online algorithm (4.1.4) for a variety of linear and nonlinear examples. In these examples, we show that the algorithm can also perform well in practice for nonlinear SDEs. We also demonstrate that the online algorithm can optimize over path-dependent SDEs and pathwise statistics of SDEs such as the auto-covariance. In addition, we also demonstrate the applications of the online optimization algorithm to mathemat-

ical finance problems, such as SDE model calibration, parameter estimation for partially-observed SDEs, high dimensional stochastic control problems, and limit order book models.

4.1.4 Literature Review

In this chapter we show that, if α_t is appropriately chosen, then $\nabla_\theta J(\theta_t) \to 0$ as $t \to \infty$ with probability 1. Similar results have been previously proven for stochastic gradient descent (SGD) in discrete time. [14] proves the convergence of SGD with i.i.d. data samples. [12] proves the convergence of SGD in discrete time with the correlated data samples under stronger conditions than [14]. We refer readers to [12, 14, 27, 63, 88] for a thorough review of the very large literature on SGD and similar stochastic optimization algorithms (e.g., SGD with momentum, Adagrad, ADAM, and RMSprop). However, these articles do not study stochastic gradient descent methods for optimizing over the stationary distribution of stochastic models, which is the focus of this chapter.

Recent articles such as [17, 126, 130, 132, 135] have studied continuous-time stochastic gradient descent. [130] proposed a "stochastic gradient descent in continuous time" (SGDCT) algorithm for estimating parameters θ in an SDE X_t^θ from continuous observations of $X_t^{\theta^*}$ where θ^* is the true parameter. [130] proves convergence of the algorithm to a stationary point. [17] extended SGDCT to estimate the drift parameter of a continuous-time jump-diffusion process. [132] analyzed proved a central limit theorem for the SGDCT algorithm and a convergence rate for strongly convex objective functions. [126] established the almost sure convergence of two-timescale stochastic gradient descent algorithms in continuous time. [135] designed an online learning algorithm for estimating the parameters of a partially observed diffusion process and studied its convergence. [128] proposes an online estimator for the parameters of the McKean-Vlasov SDE and proves that this estimator converges in L_1 to the stationary points of the asymptotic log-likelihood.

This chapter has several important differences as compared to [17, 126, 128, 130, 132, 135]. These previous papers estimate the parameter θ for the SDE X_t^θ from observations of $X_t^{\theta^*}$ where θ^* is the true parameter. In this chapter, our goal is to select θ such that the stationary distribution of X_t^θ matches certain target statistics. Therefore, unlike the previous papers, we are directly optimizing over the stationary distribution of X_t^θ. The presence of the X process in SGDCT makes the mathematical analysis challenging as the X term introduces correlation across times, and this correlation does not disappear as time tends to infinity. In order to prove convergence, [130, 132] use an appropriate Poisson PDE [62, 113, 114] associated with X to describe the evolution of the parameters for large times and analyze the fluctuations of the parameter around the direction of steepest descent. However, the theoretical results from [113, 114] do not apply to the PDE considered in this chapter since the diffusion term in our PDE is not uniformly elliptic. This is a direct result of the process $\tilde{X}_t$ in (4.1.4), which shares the same Brownian motion with the process X_t. In the case of constant σ, the PDE operator will not be uniformly elliptic and, furthermore, the coefficient for

derivatives such as $\frac{\partial^2}{\partial \tilde{x}^2}$ is zero. Consequently, we must analyze a new class of Poisson PDEs which is different than the class of Poisson PDEs studied in [113, 114]. We prove there exists a solution to this new class of Poisson PDEs which satisfies polynomial bounds. The polynomial bounds are crucial for analyzing the fluctuations of the parameter evolution in the algorithm (4.1.4).

4.1.5 Organization of the chapter

The chapter is organized into three main sections. In Section 4.2, we present the assumptions and the main theorem. Section 4.3 rigorously proves the convergence of our algorithm for multi-dimensional linear SDEs. Section 4.4 studies the numerical performance of our algorithm for a variety of linear and nonlinear SDEs, including McKean-Vlasov and path-dependent SDEs. Applications of the online optimization algorithm in mathematical finance are discussed, including SDE model calibration, parameter estimation for partially-observed SDE models, stochastic optimal control, and mean-field games. Numerical examples demonstrate how the method can be used to numerically solve high-dimensional stochastic optimal control problems and high-dimensional stochastic models of limit order book events.

4.2 Main Result

In this section, we rigorously prove convergence of the algorithm (4.1.4) for the following multi-dimensional Ornstein–Uhlenbeck process:

$$\begin{aligned} dX_t^\theta &= \left(g(\theta) - h(\theta)X_t^\theta\right) dt + \sigma dW_t, \\ X_0^\theta &= x, \end{aligned} \tag{4.2.1}$$

where $\theta \in \mathbb{R}^\ell$, $g(\theta) \in \mathbb{R}^d$, $h(\theta) \in \mathbb{R}_+^{d\times d}$, $W_t \in \mathbb{R}^d$, $X_t^\theta \in \mathbb{R}^d$, and σ is a scalar constant. Since $h(\theta)$ is positive definite, the solution to the SDE (4.2.1) is

$$X_t^\theta = e^{-h(\theta)t}x + (h(\theta))^{-1}\left(I_d - e^{-h(\theta)t}\right) g(\theta) + e^{-h(\theta)t}\int_0^t e^{h(\theta)s}\sigma dW_s, \tag{4.2.2}$$

where I_d is the $d \times d$ identity matrix. Let π_θ be the stationary distribution of X_t^θ. (π_θ exists and is unique; for example, see [115].) Our goal is to solve the optimization problem

$$\min_\theta J(\theta) = \min_\theta \left(\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \beta\right)^2, \tag{4.2.3}$$

where β is a constant. To solve (4.2.3), our online algorithm (4.1.4) becomes:

$$\begin{aligned} \frac{d\theta_t}{dt} &= -2\alpha_t\left(f(\bar{X}_t) - \beta\right)\nabla f(X_t)\tilde{X}_t, \\ dX_t &= (g(\theta_t) - h(\theta_t)X_t)dt + \sigma dW_t, \\ \frac{d\tilde{X}_t}{dt} &= \nabla_\theta g(\theta_t) - \nabla_\theta h(\theta_t)X_t - h(\theta_t)\tilde{X}_t, \\ d\bar{X}_t &= (g(\theta_t) - h(\theta_t)\bar{X}_t)dt + \sigma d\bar{W}_t, \end{aligned} \tag{4.2.4}$$

where W_t and $\bar{W}_t$ are independent Brownian motions, $\nabla_\theta g(\theta_t) \in \mathbb{R}^{d\times\ell}$, $\nabla_\theta h(\theta_t) \in \mathbb{R}^{d\times d\times\ell}$ and $\tilde{X}_t \in \mathbb{R}^{d\times\ell}$ is the gradient process for X_t. The element (i,j) of the process $\tilde{X}_t$ satisfies:

$$\frac{d}{dt}\tilde{X}_t^{i,j} = \frac{\partial g_i(\theta_t)}{\partial\theta_j} - \sum_{k=1}^{d}\frac{\partial h_{ik}(\theta_t)}{\partial\theta_j}X_t^k - \sum_{k=1}^{d} h_{ik}(\theta_t)\tilde{X}_t^{k,j}, \quad i\in\{1,2,\cdots,d\}, \quad j\in\{1,2,\cdots,\ell\}. \tag{4.2.5}$$

For the rest of this article, we will use C, C_k, C_p to denote generic constants. Our convergence theorem will require the following assumptions.

Assumption 4.2.1. (1) $g(\theta)$, $\nabla_\theta^i g(\theta)$, $h(\theta)$ and $\nabla_\theta^i h(\theta)$ are uniformly bounded functions for $i = 1, 2$.

(2) h is symmetric and uniformly positive definite, i.e. there exists a constant $c > 0$ such that

$$\min\left\{x^\top h(\theta)x\right\} \geq c|x|^2, \quad \forall\theta\in\mathbb{R}^\ell, x\in\mathbb{R}^d.$$

(3) $f, \nabla^i f, i = 1, 2, 3$ are polynomially bounded[2]:

$$|f(x)| + \sum_{i=1}^{3}|\nabla^i f(x)| \leq C(1+|x|^{\hat{m}}), \quad \forall x\in\mathbb{R}^d \tag{4.2.6}$$

for some constant $C, \hat{m} > 0$.

(4) The learning rate α_t satisfies $\int_0^\infty \alpha_t dt = \infty$, $\int_0^\infty \alpha_t^2 dt < \infty$, $\int_0^\infty |\alpha_s'|\, ds < \infty$, and there is a $\hat{p} > 0$ such that $\lim_{t\to\infty} \alpha_t^2 t^{\frac{1}{2}+2\hat{p}} = 0$.

Under these assumptions, we are able to prove the following convergence result.

Theorem 4.2.2. *Under Assumption 4.2.1 and for the Ornstein–Uhlenbeck process (4.2.1), the algorithm (4.2.4) will converge to a stationary point almost surely:*

$$\lim_{t\to\infty}|\nabla_\theta J(\theta_t)| \overset{a.s.}{=} 0. \tag{4.2.7}$$

4.3 Proof of Theorem 4.2.2

In this section, we present the proof of Theorem 4.2.2. We begin by decomposing the evolution of θ_t in (4.2.4) into several terms:

$$\begin{aligned}
\frac{d\theta_t}{dt} &= -2\alpha_t(f(\bar{X}_t)-\beta)\left(\nabla f(X_t)\tilde{X}_t\right)^\top \\
&= -2\alpha_t(\mathbb{E}_{Y\sim\pi_{\theta_t}}f(Y)-\beta)\left(\nabla f(X_t)\tilde{X}_t\right)^\top - 2\alpha_t\left(f(\bar{X}_t)-\mathbb{E}_{Y\sim\pi_{\theta_t}}f(Y)\right)\left(\nabla f(X_t)\tilde{X}_t\right)^\top \\
&= \underbrace{-\alpha_t\nabla_\theta J(\theta_t)}_{\text{Direction of Steepest Descent}} \underbrace{-2\alpha_t(\mathbb{E}_{Y\sim\pi_{\theta_t}}f(Y)-\beta)\left(\nabla f(X_t)\tilde{X}_t - \nabla_\theta\mathbb{E}_{Y\sim\pi_{\theta_t}}f(Y)\right)^\top}_{\text{Fluctuation term 1}} \\
&\quad \underbrace{-2\alpha_t\left(f(\bar{X}_t)-\mathbb{E}_{Y\sim\pi_{\theta_t}}f(Y)\right)\left(\nabla f(X_t)\tilde{X}_t\right)^\top}_{\text{Fluctuation term 2}}.
\end{aligned} \tag{4.3.1}$$

[2] $|\cdot|$ denotes the Euclidean norm. Sometimes for a square matrix x, $|x|$ will be used to denote its spectral norm which is equivalent to the Euclidean norm.

Define the error terms

$$\begin{aligned} Z_t^1 &= \left(\mathbb{E}_{Y\sim\pi_{\theta_t}} f(Y) - \beta\right)\left(\nabla f(X_t)\tilde{X}_t - \nabla_\theta \mathbb{E}_{Y\sim\pi_{\theta_t}} f(Y)\right)^\top, \\ Z_t^2 &= \left(f(\bar{X}_t) - \mathbb{E}_{Y\sim\pi_{\theta_t}} f(Y)\right)\left(\nabla f(X_t)\tilde{X}_t\right)^\top. \end{aligned} \tag{4.3.2}$$

We have therefore decomposed the evolution of θ_t into the direction of steepest descent $-\alpha_t \nabla_\theta J(\theta_t)$ and the two fluctuation terms $2\alpha_t Z_t^1$ and $2\alpha_t Z_t^2$.

As in [130], we study a cycle of stopping times to control the time periods where $|\nabla_\theta J(\theta_t)|$ is close to zero and away from zero. Let us select an arbitrary constant $\kappa > 0$ and also define $\mu = \mu(\kappa) > 0$ (to be chosen later). Then set $\sigma_0 = 0$ and define the cycles of random times

$$0 = \sigma_0 \le \tau_1 \le \sigma_1 \le \tau_2 \le \sigma_2 \le \dots,$$

where for $k = 1, 2, \dots$

$$\begin{aligned} &\tau_k = \inf\left\{t > \sigma_{k-1} : |\nabla_\theta J(\theta_t)| \ge \kappa\right\} \\ &\sigma_k = \sup\left\{t > \tau_k : \frac{|\nabla_\theta J(\theta_{\tau_k})|}{2} \le |\nabla_\theta J(\theta_s)| \le 2|\nabla_\theta J(\theta_{\tau_k})| \text{ for all } s \in [\tau_k, t] \text{ and } \int_{\tau_k}^t \alpha_s ds \le \mu\right\}. \end{aligned} \tag{4.3.3}$$

We define the random time intervals $J_k = [\sigma_{k-1}, \tau_k)$ and $I_k = [\tau_k, \sigma_k)$. We introduce $\eta > 0$ which will be chosen to be sufficiently small later. We first seek to control

$$\Delta^i_{\tau_k, \sigma_k+\eta} := \int_{\tau_k}^{\sigma_k+\eta} \alpha_s Z_s^i ds, \quad i = 1, 2 \tag{4.3.4}$$

and, as in [130], we will use a Poisson equation to bound the online fluctuation terms $\Delta^i_{\tau_k,\sigma_k+\eta}$ where the ergodic properties of X_t^θ will be leveraged in the analysis.

In this chapter, we focus on the Ornstein–Uhlenbeck process (4.2.1). As in (4.1.7), its gradient process $\tilde{X}_t^\theta := \nabla_\theta X_t^\theta = \left(\frac{\partial X_t^{\theta,i}}{\partial \theta_j}\right)_{i,j} \in \mathbb{R}^{d\times\ell}$ now satisfies the SDE:

$$\frac{d\tilde{X}_t^\theta}{dt} = \nabla_\theta g(\theta) - \nabla_\theta h(\theta) X_t^\theta - h(\theta)\tilde{X}_t^\theta, \tag{4.3.5}$$

which can be equivalently written as

$$\frac{d}{dt}\frac{\partial X_t^{\theta,i}}{\partial \theta_j} = \frac{\partial g_i(\theta)}{\partial \theta_j} - \sum_{k=1}^d \frac{\partial h_{ik}(\theta)}{\partial \theta_j} X_t^{\theta,k} - \sum_{k=1}^d h_{ik}(\theta)\frac{\partial X_t^{\theta,k}}{\partial \theta_j}, \tag{4.3.6}$$

for $i \in \{1, 2, \cdots, d\}$ and $j \in \{1, 2, \cdots, \ell\}$. Thus, we know the solution of (4.3.5) with initial point $\tilde{x}$ is

$$\tilde{X}_t^\theta = e^{-h(\theta)t}\tilde{x} + e^{-h(\theta)t}\int_0^t e^{h(\theta)s}\left(\nabla_\theta g(\theta) - \nabla_\theta h(\theta) X_s^\theta\right) ds. \tag{4.3.7}$$

The independent Ornstein–Uhlenbeck process used to obtain the asymptotic unbiased gradient is

$$\begin{aligned} d\bar{X}_t^\theta &= (g(\theta) - h(\theta)\bar{X}_t^\theta)dt + \sigma d\bar{W}_t, \\ \bar{X}_0^\theta &= \bar{x}, \end{aligned} \tag{4.3.8}$$

where $\bar{W}_t$ is another Brownian motion independent of W_t. For the processes $X_t^\theta, \tilde{X}_t^\theta, \bar{X}_t^\theta$ in (4.2.1), (4.3.5), and (4.3.8), we can prove the following convergence results.

Proposition 4.3.1. *Let $p_t(x, x', \theta)$ and $p_\infty(x', \theta)$ denote the transition probability and invariant density of the multi-dimensional Ornstein–Uhlenbeck process (4.2.1). Under Assumption 4.2.1, we have the following ergodic result:*

(i) For any $m > 0$, there exists a constant $C = C(m)$ such that

$$\left|\nabla_\theta^i p_\infty\left(x', \theta\right)\right| \leq \frac{C}{1 + |x'|^m}, \quad i = 0, 1, 2. \tag{4.3.9}$$

(ii) For any m', k there exist constants C, m such that for any $t > 1$

$$\left|\nabla_\theta^i p_t\left(x, x', \theta\right) - \nabla_\theta^i p_\infty\left(x', \theta\right)\right| \leq \frac{C\left(1 + |x|^m\right)}{\left(1 + |x'|^{m'}\right)(1 + t)^k}, \quad i = 0, 1, 2. \tag{4.3.10}$$

(iii) For any m', k there exist constants C, m such that for any $t > 1$

$$\left|\nabla_x^j \nabla_\theta^i p_t\left(x, x', \theta\right)\right| \leq \frac{C\left(1 + |x|^m\right)}{\left(1 + |x'|^{m'}\right)(1 + t)^k}, \quad i = 0, 1, \quad j = 1, 2. \tag{4.3.11}$$

(iv) For any $m > 0$, there exists a constant $C = C(m)$ such that for any $t \geq 0$

$$\mathbb{E}_x \left|X_t^\theta\right|^m \leq C(1 + |x|^m), \quad \mathbb{E}_{x,\tilde{x}} \left|\tilde{X}_t^\theta\right|^m \leq C(1 + |x|^m + |\tilde{x}|^m). \tag{4.3.12}$$

Here $\mathbb{E}_x$ denotes that the initial condition for the process X_t^θ is x, i.e. $X_0^\theta = 0$. $\mathbb{E}_{x,\tilde{x}}$ denotes that the initial conditions of the processes $(X_t^\theta, \tilde{X}_t^\theta)$ in (4.1.7) are $(x, \tilde{x})$, i.e. $X_0^\theta = x$ and $\tilde{X}_0^\theta = \tilde{x}$.

(v) For any function f satisfying (4.2.6), there exists constants C, m such that for any $t \in [0, 1]$

$$\left|\nabla_x^j \nabla_\theta^i \mathbb{E}_x f(X_t^\theta)\right| \leq C(1 + |x|^m), \quad i = 0, 1, \quad j = 0, 1, 2. \tag{4.3.13}$$

Remark 4.3.2. Proposition 4.3.1 is similar to Theorem 1 in [114]. However, the assumption of uniform boundedness in [113] does not hold for the multi-dimensional Ornstein–Uhlenbeck process (4.2.1). Thus we give a brief proof by direct calculations in Section B.1.

We must analyze the fluctuation terms Z_t^1 and Z_t^2. In order to do this, we prove a polynomially-bounded solution exists to a new class of Poisson PDEs. The polynomial bound is in the spatial coordinates and, importantly, the bound is uniform in the parameter θ. A Poisson PDE was also used in [130]. However, several key innovations are required for the online optimization algorithm (4.2.4) that we consider in this chapter. Unlike in [130], $\tilde{X}_t^\theta$ in (4.3.5) does not have a diffusion term, which means $(X_t^\theta, \tilde{X}_t^\theta)$ is a degenerate diffusion process and its generator $\mathcal{L}_{x,\tilde{x}}^\theta$ is not a uniformly elliptic operator. Thus we cannot use the results from [113, 114]. Instead, we must prove existence and bounds for this new class of Poisson PDEs.

Lemma 4.3.3. *Define the error function*

$$G^1(x, \tilde{x}, \theta) = \left(\mathbb{E}_{Y \sim \pi_\theta} f(Y) - \beta\right)\left(\nabla f(x)\tilde{x} - \nabla_\theta \mathbb{E}_{Y \sim \pi_\theta} f(Y)\right)^\top \tag{4.3.14}$$

and

$$v^1(x,\tilde{x},\theta) = -\int_0^\infty \mathbb{E}_{x,\tilde{x}} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta)dt, \tag{4.3.15}$$

where $\mathbb{E}_{x,\tilde{x}}$ *is a conditional expectation given* $X_0^\theta = x$ *and* $\tilde{X}_0^\theta = \tilde{x}$*. Then, under Assumption 4.2.1,* $v^1(x,\tilde{x},\theta)$ *is the classical solution of the Poisson equation*

$$\mathcal{L}_{x,\tilde{x}}^\theta u(x,\tilde{x},\theta) = G^1(x,\tilde{x},\theta), \tag{4.3.16}$$

where $u = (u_1,\dots,u_\ell)^\top \in \mathbb{R}^\ell$ *is a vector,* $\mathcal{L}_{x,\tilde{x}}^\theta u(x,\tilde{x},\theta) = (\mathcal{L}_{x,\tilde{x}}^\theta u_1(x,\tilde{x},\theta),\dots,\mathcal{L}_{x,\tilde{x}}^\theta u_\ell(x,\tilde{x},\theta))^\top$*, and* $\mathcal{L}_{x,\tilde{x}}^\theta$ *is the infinitesimal generator of the process* $(X_\cdot^\theta, \tilde{X}_\cdot^\theta)$*, i.e. for any test function* φ

$$\mathcal{L}_{x,\tilde{x}}^\theta \varphi(x,\tilde{x}) = \mathcal{L}_x^\theta \varphi(x,\tilde{x}) + tr\left(\nabla_{\tilde{x}}\varphi(x,\tilde{x})^\top \left(\nabla_\theta g(\theta) - \nabla_\theta h(\theta)x - h(\theta)\tilde{x}\right)\right). \tag{4.3.17}$$

Furthermore, there exist an integer m' *and a constant* $C = C(m')$ *which do not depend upon* $(x,\tilde{x},\theta)$ *such that the solution* v^1 *satisfies the bound*

$$\left|v^1(x,\tilde{x},\theta)\right| + \left|\nabla_\theta v^1(x,\tilde{x},\theta)\right| + \left|\nabla_x v^1(x,\tilde{x},\theta)\right| + \left|\nabla_{\tilde{x}} v^1(x,\tilde{x},\theta)\right| \le C\left(1 + |x|^{m'} + |\tilde{x}|^{m'}\right). \tag{4.3.18}$$

The proof of Lemma 4.3.3 is in Appendix B.2. We will next study the fluctuation terms Z_t^i. It will be necessary to prove bounds on the moments of X_t and $\tilde{X}_t$ in order to analyze the error term $\Delta^i_{\tau_k,\sigma_k+\eta}$.

Lemma 4.3.4. *For any* $p > 0$*, there exists a constant* C_p *that only depends on* p *such that the processes* $X_t, \tilde{X}_t$ *from (4.2.4) satisfy*

$$\mathbb{E}_x|X_t|^p \le C_p\left(1 + |x|^p\right), \quad \mathbb{E}_{x,\tilde{x}}|\tilde{X}_t|^p \le C_p\left(1 + |x|^p + |\tilde{x}|^p\right). \tag{4.3.19}$$

Furthermore, we have the bounds

$$\begin{aligned} \mathbb{E}_x\left(\sup_{0\le t'\le t}|X_{t'}|^p\right) &= O(\sqrt{t}) \quad \text{as } t\to\infty, \\ \mathbb{E}_{x,\tilde{x}}\left(\sup_{0\le t'\le t}|\tilde{X}_{t'}|^p\right) &= O(\sqrt{t}) \quad \text{as } t\to\infty. \end{aligned} \tag{4.3.20}$$

Proof. By adapting the method in [57], we first prove (4.3.19) for $p \ge 2$ and then the result for $0 < p < 2$ follows from Hölder's inequality. Let $p = 2m$ and applying Itô's formula to $e^{m\alpha t}|X_t|^{2m}$, we have for any $t \ge 0$,

$$\begin{aligned} e^{p\alpha t/2}|X_t|^p - |X_0|^p \le &\int_0^t p\left(\frac{\alpha}{2}|X_t|^2 + \langle X_s, g(\theta_s) - h(\theta_s)X_s\rangle\right)e^{p\alpha s/2}|X_s|^{p-2}\,ds \\ &+ \int_0^t \frac{p(p-1)d}{2}e^{p\alpha s/2}|X_s|^{p-2}\,ds + \int_0^t pe^{p\alpha s/2}|X_s|^{p-2}\langle X_s,\ dW_s\rangle, \end{aligned} \tag{4.3.21}$$

where $\langle a,\ b\rangle := a^T b$. By Assumption 4.2.1, we know there exists constants $\alpha > 0, \beta > d$ such that for any θ

$$\langle x, g(\theta) - h(\theta)x\rangle \le -\alpha|x|^2 + \beta. \tag{4.3.22}$$

Thus by taking expectations on both sides of (4.3.21) and using (4.3.22), we obtain

$$\mathbb{E}_x\left[e^{p\alpha t/2}|X_t|^p\right] - |x|^p \le \int_0^t -\frac{p\alpha}{2}\mathbb{E}_x\left[e^{p\alpha s/2}|X_s|^p\right]ds + \int_0^t \mathbb{E}_x\left[\frac{p(p+1)\beta}{2}e^{p\alpha s/2}|X_s|^{p-2}\right]ds.$$

Young's inequality implies that

$$\frac{p(p+1)\beta}{2}\mathrm{e}^{p\alpha s/2}\left|X_s\right|^{p-2} \leq \frac{p\alpha}{2}e^{p\alpha s/2}\left|X_s\right|^p + c_p e^{p\alpha s/2}$$

where $c_p = \left(\frac{p-2}{p\alpha}\right)^{p/2-1}(\beta(p+1))^{p/2}$. Therefore, we obtain

$$\mathbb{E}_x\left[e^{p\alpha t/2}\left|X_t\right|^p\right] - |x|^p \leq \int_0^t c_p \mathrm{e}^{p\alpha s/2} ds$$

and

$$\mathbb{E}_x\left|X_t\right|^p \leq \frac{2c_p}{p\alpha} + e^{-p\alpha t/2}\left|x\right|^p \leq C_p\left(1+|x|^p\right).$$

Using the moment bound for X_t, we can derive the moment bound for $\tilde{X}_t$. From (4.3.7) and (4.2.4) we know

$$\tilde{X}_t = e^{-\int_0^t h(\theta_u)du}\tilde{X}_0 + \int_0^t e^{-\int_s^t h(\theta_u)du}\left(\nabla_\theta g(\theta_s) - \nabla_\theta h(\theta_s)X_s\right)ds \tag{4.3.23}$$

and thus

$$\begin{aligned}
\mathbb{E}_{x,\tilde{x}}|\tilde{X}_t|^p &\leq 2|\tilde{x}|^p + 2\mathbb{E}_{x,\tilde{x}}\left|\int_0^t \left|e^{-\int_s^t h(\theta_u)du}\right| \cdot |\nabla_\theta g(\theta_s) - \nabla_\theta h(\theta_s)X_s|\, ds\right|^p \\
&\overset{(a)}{\leq} 2|\tilde{x}|^p + C_p\mathbb{E}_{x,\tilde{x}}\left|\int_0^t e^{-c(t-s)}\left(1+|X_s|\right)ds\right|^p \\
&\leq 2|\tilde{x}|^p + C_p\mathbb{E}_x\left|\int_0^t \frac{e^{cs}}{e^{ct}-1}\left(1+|X_s|\right)ds\right|^p e^{-cpt}\left(e^{ct}-1\right)^p \\
&\overset{(b)}{\leq} 2|\tilde{x}|^p + C_p\mathbb{E}_x\left|\int_0^t \frac{e^{cs}}{e^{ct}-1}\left(1+|X_s|\right)^p ds\right| \\
&\leq C_p\left(1+|x|^p+|\tilde{x}|^p\right),
\end{aligned} \tag{4.3.24}$$

where step (a) is by Assumption 4.2.1 and the fact

$$\lambda_{\max}\left(e^{-\int_s^{t'} h(\theta_u)du}\right) = e^{-\lambda_{\min}\left(\int_s^{t'} h(\theta_u)du\right)} \leq e^{-c(t'-s)} \tag{4.3.25}$$

and step (b) is by Jensen's inequality.

To prove (4.3.20), we use a similar method as in [113]. By Itô's formula, we have for $p \geq 1$

$$\begin{aligned}
|X_t|^{2p} - |X_0|^{2p} &\leq \int_0^t 2p\left|X_s\right|^{2p-2}\langle X_s,\ g(\theta_s) - h(\theta_s)X_s\rangle\, ds + \int_0^t p(d+2p-1)\left|X_s\right|^{2p-2}ds \\
&\quad + 2p\int_0^t \left|X_s\right|^{2p-2}\langle X_s,\ dW_s\rangle \\
&\leq C_p\int_0^t \left|X_s\right|^{2p-2}ds + 2p\int_0^t \left|X_s\right|^{2p-2}\langle X_s,\ dW_s\rangle.
\end{aligned} \tag{4.3.26}$$

Using the Burkholder-Davis-Gundy inequality, there exists a constant C such that

$$\mathbb{E}_x\left(\sup_{t'\leq t}|X_{t'}|^{2p}\right) \leq |x|^{2p} + C_p\left(\mathbb{E}_x\int_0^t |X_s|^{4p-2}\,ds\right)^{1/2} + C_p\mathbb{E}_x\int_0^t |X_s|^{2p-2}\,ds, \tag{4.3.27}$$

which together with estimate (4.3.19) can be used to derive the bound

$$\mathbb{E}_x\left(\sup_{t'\leq t}|X_{t'}|^{2p}\right) \leq |x|^{2p} + C_p\left(t+t^{1/2}\right)\left(1+|x|^{2p-1}\right). \tag{4.3.28}$$

Furthermore, for $t \geq 1$,

$$\begin{aligned}\mathbb{E}_x\left(\sup_{t'\leq t}|X_{t'}|^p\right) &\overset{(a)}{\leq} \left(\mathbb{E}_x \sup_{t'\leq t}|X_{t'}|^{2p}\right)^{\frac{1}{2}} \\ &\leq \left(|x|^{2p} + C_p\left(t + t^{1/2}\right)\left(1+|x|^{2p-1}\right)\right)^{\frac{1}{2}} \\ &\leq |x|^p + C_p\left(1+|x|^{p-\frac{1}{2}}\right)\sqrt{t},\end{aligned} \tag{4.3.29}$$

where step (a) is by Hölder inequality. Similarly, we have for any $p' < p$ and $t \geq 1$ that

$$\mathbb{E}_x\left(\sup_{t'\leq t}|X_{t'}|^{p'}\right) \leq C|x|^{p'} + C\left(1+|x|^{p-\frac{1}{2}}\right)t^{\frac{p'}{2p}}, \tag{4.3.30}$$

and thus the result for X_t in (4.3.20) follows. Finally, similarly as in (4.3.24),

$$\begin{aligned}\mathbb{E}_{x,\tilde{x}}\sup_{t'\leq t}|\tilde{X}_{t'}|^p &\leq 2|\tilde{x}|^p + 2\mathbb{E}_{x,\tilde{x}}\sup_{t'\leq t}\left|\int_0^{t'}\left|e^{-\int_s^{t'}h(\theta_u)du}\right|\cdot|\nabla_\theta g(\theta_s) - \nabla_\theta h(\theta_s)X_s|\,ds\right|^p \\ &\leq 2|\tilde{x}|^p + C_p\mathbb{E}_x\sup_{t'\leq t}\left|\int_0^{t'}e^{-c(t'-s)}\left(1+|X_s|\right)ds\right|^p \\ &\leq 2|\tilde{x}|^p + C_p\mathbb{E}_x\sup_{t'\leq t}\left(1+|X_{t'}|^p\right).\end{aligned} \tag{4.3.31}$$

Combining (4.3.29), (4.3.30), and (4.3.31), we can prove the bound for $\tilde{X}_t$ in (4.3.20). □

Using the estimates in Lemma 4.3.3 and Lemma 4.3.4, we can now bound the first fluctuation term $\Delta^1_{\tau_k,\sigma_k+\eta}$ in (4.3.4).

Lemma 4.3.5. *Under Assumption 4.2.1, for any fixed $\eta > 0$*

$$\left|\Delta^1_{\tau_k,\sigma_k+\eta}\right| \to 0 \text{ as } k\to\infty, \quad a.s. \tag{4.3.32}$$

Proof. The idea is to use the Poisson equation in Lemma 4.3.3 to derive an equivalent expression for the term $\Delta^i_{\tau_k,\sigma_k+\eta}$ which we can appropriately control as k becomes large. Consider the function

$$G^1(x,\tilde{x},\theta) = \left(\mathbb{E}_{Y\sim\pi_\theta}f(Y) - \beta\right)\left(\nabla f(x)\tilde{x} - \nabla_\theta\mathbb{E}_{Y\sim\pi_\theta}f(Y)\right)^\top.$$

By Lemma 4.3.3, the Poisson equation $\mathcal{L}^\theta_{x\tilde{x}}u(x,\tilde{x},\theta) = G^1(x,\tilde{x},\theta)$ will have a unique smooth solution $v^1(x,\tilde{x},\theta)$ that grows at most polynomially in $(x,\tilde{x})$. Let us apply Itô's formula to the function

$$u^1(t,x,\tilde{x},\theta) := \alpha_t v^1(x,\tilde{x},\theta) \in \mathbb{R}^\ell,$$

evaluated on the stochastic process $(X_t,\tilde{X}_t,\theta_t)$. Recall that u_i denotes the i-th element of u for $i \in \{1,2,\cdots,\ell\}$. Then,

$$\begin{aligned}u^1_i\left(\sigma,X_\sigma,\tilde{X}_\sigma,\theta_\sigma\right) =& u^1_i\left(\tau,X_\tau,\tilde{X}_\tau,\theta_\tau\right) + \int_\tau^\sigma \partial_s u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right)ds + \int_\tau^\sigma \mathcal{L}^{\theta_s}_{x\tilde{x}}u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right)ds \\ &+ \int_\tau^\sigma \nabla_\theta u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right)d\theta_s + \int_\tau^\sigma \nabla_x u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right)\sigma dW_s.\end{aligned} \tag{4.3.33}$$

Rearranging the previous equation, we obtain the representation

$$\begin{aligned}\Delta^1_{\tau_k,\sigma_k+\eta} =& \int_{\tau_k}^{\sigma_k+\eta} \alpha_s G^1(X_s,\tilde{X}_s,\theta_s)ds = \int_{\tau_k}^{\sigma_k+\eta} \mathcal{L}^{\theta_s}_{x\tilde{x}} u^1\left(s,X_s,\tilde{X}_s,\theta_s\right) ds \\ =&\alpha_{\sigma_k+\eta} v^1\left(X_{\sigma_k+\eta},\tilde{X}_{\sigma_k+\eta},\theta_{\sigma_k+\eta}\right) - \alpha_{\tau_k} v^1\left(X_{\tau_k},\tilde{X}_{\tau_k},\theta_{\tau_k}\right) - \int_{\tau_k}^{\sigma_k+\eta} \alpha'_s v^1\left(X_s,\tilde{X}_s,\theta_s\right) ds \\ &+ \int_{\tau_k}^{\sigma_k+\eta} 2\alpha_s^2 \nabla_\theta v^1\left(X_s,\tilde{X}_s,\theta_s\right)\left(f(\bar{X}_s)-\beta\right)\left(\nabla f(X_s)\tilde{X}_s\right)^\top ds \\ &- \int_{\tau_k}^{\sigma_k+\eta} \alpha_s \nabla_x v^1\left(X_s,\tilde{X}_s,\theta_s\right) dW_s.\end{aligned} \tag{4.3.34}$$

The next step is to treat each term on the right hand side of (4.3.34) separately. For this purpose, let us first set

$$J^{1,1}_t = \alpha_t \sup_{s\in[0,t]} \left|v^1\left(X_s,\tilde{X}_s,\theta_s\right)\right|. \tag{4.3.35}$$

By (4.3.18) and (4.3.20), there exists a constant C that only depends on m' such that

$$\begin{aligned}\mathbb{E}\left|J^{1,1}_t\right|^2 &\le C\alpha_t^2 \mathbb{E}\left[1+\sup_{s\in[0,t]}|X_s|^{m'} + \sup_{s\in[0,t]}\left|\tilde{X}_s\right|^{m'}\right] \\ &= C\alpha_t^2\left[1+\sqrt{t}\frac{\mathbb{E}\sup_{s\in[0,t]}|X_s|^{m'} + \mathbb{E}\sup_{s\in[0,t]}\left|\tilde{X}_s\right|^{m'}}{\sqrt{t}}\right] \\ &\le C\alpha_t^2\sqrt{t}.\end{aligned} \tag{4.3.36}$$

Let $p>0$ be the constant in Assumption 4.2.1 such that $\lim_{t\to\infty}\alpha_t^2 t^{1/2+2p}=0$ and for any $\delta\in(0,p)$ define the event $A_{t,\delta}=\left\{J^{1,1}_t \ge t^{\delta-p}\right\}$. Then we have for t large enough such that $\alpha_t^2 t^{1/2+2p}\le 1$

$$\mathbb{P}\left(A_{t,\delta}\right) \le \frac{\mathbb{E}\left|J^{1,1}_t\right|^2}{t^{2(\delta-p)}} \le C\frac{\alpha_t^2 t^{1/2+2p}}{t^{2\delta}} \le C\frac{1}{t^{2\delta}}.$$

The latter implies that

$$\sum_{n\in\mathbb{N}} \mathbb{P}\left(A_{2^n,\delta}\right) < \infty.$$

Therefore, by the Borel-Cantelli lemma we have that for every $\delta\in(0,p)$ there is a finite positive random variable $d(\omega)$ and some $n_0<\infty$ such that for every $n\ge n_0$ one has

$$J^{1,1}_{2^n} \le \frac{d(\omega)}{2^{n(p-\delta)}}.$$

Thus, for $t\in\left[2^n,2^{n+1}\right)$ and $n\ge n_0$ one has for some finite constant $C<\infty$

$$J^{1,1}_t \le C\alpha_{2^{n+1}} \sup_{s\in(0,2^{n+1}]}\left|v^1\left(X_s,\tilde{X}_s,\theta_s\right)\right| \le C\frac{d(\omega)}{2^{(n+1)(p-\delta)}} \le C\frac{d(\omega)}{t^{p-\delta}},$$

which proves that for $t\ge 2^{n_0}$ with probability one

$$J^{1,1}_t \le C\frac{d(\omega)}{t^{p-\delta}} \to 0, \text{ as } t\to\infty. \tag{4.3.37}$$

Next we consider the term

$$J^{1,2}_{t,0} = \int_0^t \left|\alpha'_s v^1\left(X_s,\tilde{X}_s,\theta_s\right) - 2\alpha_s^2\nabla_\theta v^1\left(X_s,\tilde{X}_s,\theta_s\right)\left(f(\bar{X}_s)-\beta\right)\left(\nabla f(X_s)\tilde{X}_s\right)^\top\right| ds.$$

There exists a constant $0 < C < \infty$ (that may change from line to line) and $0 < m' < \infty$ such that

$$\begin{aligned}\sup_{t>0} \mathbb{E}\left|J^{1,2}_{t,0}\right| &\overset{(a)}{\leq} C\int_0^\infty \left(|\alpha'_s| + \alpha_s^2\right)\left(1 + \mathbb{E}\left|X_s\right|^{m'} + \mathbb{E}\left|\bar{X}_s\right|^{m'} + \mathbb{E}|\tilde{X}_t|^{m'}\right) ds \\ &\overset{(b)}{\leq} C\int_0^\infty \left(|\alpha'_s| + \alpha_s^2\right) ds \\ &\leq C,\end{aligned}$$

where step (a) is by Assumption 4.2.1 and (4.3.18) and in step (b) we use (4.3.19). Thus there is a finite random variable $J^{1,2}_{\infty,0}$ such that

$$J^{1,2}_{t,0} \to J^{1,2}_{\infty,0}, \text{as } t \to \infty \text{ with probability one.} \tag{4.3.38}$$

The last term we need to consider is the martingale term

$$J^{1,3}_{t,0} = \int_0^t \alpha_s \nabla_x v^1\left(X_s, \tilde{X}_s, \theta_s\right) dW_s.$$

By Doob's inequality, Assumption 4.2.1, (4.3.18), (4.3.19), and using calculations similar to the ones for the term $J^{1,2}_{t,0}$, we can show that for some finite constant $C < \infty$,

$$\sup_{t>0} \mathbb{E}\left|J^{1,3}_{t,0}\right|^2 \leq C\int_0^\infty \alpha_s^2 ds < \infty.$$

Thus, by Doob's martingale convergence theorem there is a square integrable random variable $J^{1,3}_{\infty,0}$ such that

$$J^{1,3}_{t,0} \to J^{1,3}_{\infty,0}, \quad \text{as } t \to \infty \text{ both almost surely and in } L^2. \tag{4.3.39}$$

Let us now return to (4.3.34). Using the terms $J^{1,1}_t, J^{1,2}_{t,0}$, and $J^{1,3}_{t,0}$ we can write

$$\left|\Delta^1_{\tau_k,\sigma_k+\eta}\right| \leq J^{1,1}_{\sigma_k+\eta} + J^{1,1}_{\tau_k} + \left|J^{1,2}_{\sigma_k+\eta,\tau_k}\right| + \left|J^{1,3}_{\sigma_k+\eta,\tau_k}\right|,$$

which together with (4.3.37), (4.3.38), and (4.3.39) prove the statement of the Lemma. □

Now we prove a similar convergence result for $\Delta^2_{\tau_k,\sigma_k+\eta}$. We first give an extension of Lemma 4.3.3 for the Poisson equation.

Lemma 4.3.6. *Define the error function*

$$G^2(x,\tilde{x},\bar{x},\theta) = \left[f(\bar{x}) - \mathbb{E}_{Y\sim\pi_\theta} f(Y)\right]\left(\nabla f(x)\tilde{x}\right)^\top. \tag{4.3.40}$$

and

$$v^2(x,\tilde{x},\bar{x},\theta) = -\int_0^\infty \mathbb{E}_{x,\tilde{x},\bar{x}} G^2(X^\theta_t, \tilde{X}^\theta_t, \bar{X}^\theta_t, \theta) dt, \tag{4.3.41}$$

where $\mathbb{E}_{x,\tilde{x},\bar{x}}$ is a conditional expectation given $X^\theta_0 = x$, $\tilde{X}^\theta_0 = \tilde{x}$, and $\bar{X}^\theta_0 = \bar{x}$. Under Assumption 4.2.1, $v^2(x,\tilde{x},\bar{x},\theta)$ is the classical solution of the Poisson equation

$$\mathcal{L}^\theta_{x,\tilde{x},\bar{x}} u(x,\tilde{x},\bar{x},\theta) = G^2(x,\tilde{x},\bar{x},\theta), \tag{4.3.42}$$

where $\mathcal{L}^\theta_{x,\tilde{x},\bar{x}}$ is generator of the process $(X^\theta_\cdot, \tilde{X}^\theta_\cdot, \bar{X}^\theta_\cdot)$, i.e. for any test function φ

$$\mathcal{L}^\theta_{x,\tilde{x},\bar{x}}\varphi(x,\tilde{x},\bar{x}) = \mathcal{L}^\theta_{x,\tilde{x}}\varphi(x,\tilde{x},\bar{x}) + \mathcal{L}^\theta_{\bar{x}}\varphi(x,\tilde{x},\bar{x}). \tag{4.3.43}$$

Furthermore, there exist an integer m' and a constant $C = C(m')$ which do not depend upon $(x, \tilde{x}, \bar{x}, \theta)$ such that the solution v^2 satisfies the bound

$$\begin{aligned}&\left|v^2(x,\tilde{x},\bar{x},\theta)\right| + \left|\nabla_{\bar{x}} v^2(x,\tilde{x},\bar{x},\theta)\right| + \left|\nabla_{\theta} v^2(x,\tilde{x},\bar{x},\theta)\right| + \left|\nabla_{x} v^2(x,\tilde{x},\bar{x},\theta)\right| \\ \leq & C\left(1 + |x|^{m'} + |\tilde{x}|^{m'} + |\bar{x}|^{m'}\right).\end{aligned} \tag{4.3.44}$$

The proof of Lemma 4.3.6 can be found in Appendix B.2.

Lemma 4.3.7. *Under Assumption 4.2.1, for any fixed $\eta > 0$, we have*

$$\left|\Delta^2_{\tau_k, \sigma_k + \eta}\right| \to 0, \text{ as } k \to \infty, \quad a.s.. \tag{4.3.45}$$

Proof. Consider the function

$$G^2(x,\tilde{x},\bar{x},\theta) = \left(f(\bar{x}) - \mathbb{E}_{Y\sim\pi_\theta} f(Y)\right)\left(\nabla f(x)\tilde{x}\right)^\top. \tag{4.3.46}$$

Let v^2 be the solution of (4.3.42) in Lemma 4.3.6. We apply Itô formula to the function $u^2(t,x,\tilde{x},\bar{x},\theta) = \alpha_t v^2(x,\tilde{x},\bar{x},\theta)$ evaluated on the stochastic process $(X_t, \tilde{X}_t, \bar{X}_t, \theta_t)$ and get for any $i \in \{1, 2, \cdots, \ell\}$

$$\begin{aligned}&u_i^2\left(\sigma, X_\sigma, \tilde{X}_\sigma, \bar{X}_\sigma, \theta_\sigma\right) - u_i^2\left(\tau, X_\tau, \tilde{X}_\tau, \bar{X}_\tau, \theta_\tau\right) \\ =&\int_\tau^\sigma \partial_s u_i^2\left(s, X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) ds + \int_\tau^\sigma \mathcal{L}^{\theta_s}_{x,\tilde{x}} u_i^2\left(s, X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) ds \\ +&\int_\tau^\sigma \mathcal{L}^{\theta_s}_{\bar{x}} u_i^2\left(s, X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) ds + \int_\tau^\sigma \nabla_\theta u_i^2\left(s, X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) d\theta_s \\ +&\int_\tau^\sigma \nabla_x u_i^2\left(s, X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) dW_s + \int_\tau^\sigma \nabla_{\bar{x}} u_i^2\left(s, X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) d\bar{W}_s.\end{aligned} \tag{4.3.47}$$

Rearranging the previous equation, we obtain the representation

$$\begin{aligned}\Delta^2_{\tau_k,\sigma_k+\eta} =& \int_{\tau_k}^{\sigma_k+\eta} \alpha_s G^2(X_s, \tilde{X}_s, \bar{X}_s, \theta_s) ds \\ =& \int_{\tau_k}^{\sigma_k+\eta} \mathcal{L}^{\theta_s}_{x,\tilde{x},\bar{x}} u^2\left(s, X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) ds \\ =& \alpha_{\sigma_k+\eta} v^2\left(X_{\sigma_k+\eta}, \tilde{X}_{\sigma_k+\eta}, \bar{X}_{\sigma_k+\eta}, \theta_{\sigma_k+\eta}\right) - \alpha_{\tau_k} v^2\left(X_{\tau_k}, \tilde{X}_{\tau_k}, \bar{X}_{\tau_k}, \theta_{\tau_k}\right) \\ &- \int_{\tau_k}^{\sigma_k+\eta} \alpha'_s v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) ds - \int_{\tau_k}^{\sigma_k+\eta} \alpha_s \nabla_{\bar{x}} v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) d\bar{W}_s \\ &+ \int_{\tau_k}^{\sigma_k+\eta} 2\alpha_s^2 \nabla_\theta v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right)\left(f(\bar{X}_s) - \beta\right)\left(\nabla f(X_s)\tilde{X}_s\right)^\top ds \\ &- \int_{\tau_k}^{\sigma_k+\eta} \alpha_s \nabla_x v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) dW_s.\end{aligned} \tag{4.3.48}$$

The next step is to treat each term on the right hand side of (4.3.48) separately. For this purpose, let us first set

$$J_t^{2,1} = \alpha_t \sup_{s\in[0,t]} \left|v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right)\right|. \tag{4.3.49}$$

Using the same approach as for X_t in Lemma 4.3.4, we can show that for any $p > 0$ there exists a constant C_p that only depends on p such that

$$\mathbb{E}_{\bar{x}}|\bar{X}_t|^p \leq C_p\left(1 + |\bar{x}|^p\right), \quad \mathbb{E}_{\bar{x}}\left(\sup_{0\leq t'\leq t} |\bar{X}_{t'}|^p\right) = O(\sqrt{t}) \quad \text{as } t \to \infty. \tag{4.3.50}$$

Combining Lemma 4.3.4, (4.3.44), and (4.3.50), we know that there exists a constant C such that

$$\begin{aligned}\mathbb{E}\left|J_t^{2,1}\right|^2 &\leq C\alpha_t^2\mathbb{E}\left[1+\sup_{s\in[0,t]}|X_s|^{m'}+\sup_{s\in[0,t]}\left|\tilde{X}_s\right|^{m'}+\sup_{s\in[0,t]}\left|\bar{X}_s\right|^{m'}\right]\\ &= C\alpha_t^2\left[1+\sqrt{t}\frac{\mathbb{E}\sup_{s\in[0,t]}|X_s|^{m'}+\mathbb{E}\sup_{s\in[0,t]}\left|\tilde{X}_s\right|^{m'}+\mathbb{E}\sup_{s\in[0,t]}\left|\bar{X}_s\right|^{m'}}{\sqrt{t}}\right] \qquad (4.3.51)\\ &\leq C\alpha_t^2\sqrt{t}.\end{aligned}$$

Let $p>0$ be the constant in Assumption 4.2.1 such that $\lim_{t\to\infty}\alpha_t^2 t^{1/2+2p}=0$ and for any $\delta\in(0,p)$ define the event $A_{t,\delta}=\left\{J_t^{2,1}\geq t^{\delta-p}\right\}$. Then we have for t large enough such that $\alpha_t^2 t^{1/2+2p}\leq 1$ and

$$\mathbb{P}\left(A_{t,\delta}\right)\leq\frac{\mathbb{E}\left|J_t^{2,1}\right|^2}{t^{2(\delta-p)}}\leq C\frac{\alpha_t^2 t^{1/2+2p}}{t^{2\delta}}\leq C\frac{1}{t^{2\delta}}.$$

The latter implies that

$$\sum_{n\in\mathbb{N}}\mathbb{P}\left(A_{2^n,\delta}\right)<\infty.$$

Therefore, by the Borel-Cantelli lemma we have that for every $\delta\in(0,p)$ there is a finite positive random variable $d(\omega)$ and some $n_0<\infty$ such that for every $n\geq n_0$ one has

$$J_{2^n}^{2,1}\leq\frac{d(\omega)}{2^{n(p-\delta)}}.$$

Thus for $t\in\left[2^n,2^{n+1}\right)$ and $n\geq n_0$ one has for some finite constant $C<\infty$

$$J_t^{2,1}\leq C\alpha_{2^{n+1}}\sup_{s\in(0,2^{n+1}]}\left|v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)\right|\leq C\frac{d(\omega)}{2^{(n+1)(p-\delta)}}\leq C\frac{d(\omega)}{t^{p-\delta}},$$

which derives that for $t\geq 2^{n_0}$ we have with probability one

$$J_t^{2,1}\leq C\frac{d(\omega)}{t^{p-\delta}}\to 0,\ \text{as } t\to\infty. \qquad (4.3.52)$$

Next we consider the term

$$J_{t,0}^{2,2}=\int_0^t\left|\alpha_s' v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)-2\alpha_s^2\nabla_\theta v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)\left(f(\bar{X}_s)-\beta\right)\left(\nabla f(X_s)\tilde{X}_s\right)^\top\right|ds$$

and thus we see that there exists a constant $0<C<\infty$ such that

$$\begin{aligned}\sup_{t>0}\mathbb{E}\left|J_{t,0}^{2,2}\right| &\overset{(a)}{\leq} C\int_0^\infty\left(|\alpha_s'|+\alpha_s^2\right)\left(1+\mathbb{E}\,|X_s|^{m'}+\mathbb{E}\,|X_s|^{m'}+\mathbb{E}\left|\bar{X}_s\right|^{m'}\right)ds\\ &\overset{(b)}{\leq} C\int_0^\infty\left(|\alpha_s'|+\alpha_s^2\right)ds\\ &\leq C,\end{aligned}$$

where in step (a) we use (4.3.44) and in step (b) we use Lemma 4.3.4 and (4.3.50). Thus we know there is a finite random variable $J_{\infty,0}^{2,2}$ such that

$$J_{t,0}^{2,2}\to J_{\infty,0}^{2,2},\ \text{as } t\to\infty \text{ with probability one.} \qquad (4.3.53)$$

The last term we need to consider is the martingale term

$$J_{t,0}^{2,3}=\int_0^t\alpha_s\nabla_x v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)dW_s+\int_0^t\alpha_s\nabla_{\bar{x}}v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)d\bar{W}_s.$$

Notice that Doob's inequality and the bounds of (4.3.44) (using calculations similar to the ones for the term $J^{2,2}_{t,0}$) give us that for some finite constant $K < \infty$, we have

$$\sup_{t>0} \mathbb{E}\left|J^{2,3}_{t,0}\right|^2 \leq K \int_0^\infty \alpha_s^2 ds < \infty.$$

Thus, by Doob's martingale convergence theorem there is a square integrable random variable $J^{(3)}_{\infty,0}$ such that

$$J^{2,3}_{t,0} \to J^{2,3}_{\infty,0}, \quad \text{as } t \to \infty \text{ both almost surely and in } L^2. \tag{4.3.54}$$

Let us now go back to (4.3.48). Using the terms $J^{2,1}_t$, $J^{2,2}_{t,0}$ and $J^{2,3}_{t,0}$ we can write

$$\left|\Delta^2_{\tau_k,\sigma_k+\eta}\right| \leq J^{2,1}_{\sigma_k+\eta} + J^{2,1}_{\tau_k} + J^{2,2}_{\sigma_k+\eta,\tau_k} + \left|J^{2,3}_{\sigma_k+\eta,\tau_k}\right|,$$

which together with (4.3.52), (4.3.53) and (4.3.54) prove the statement of the Lemma. □

Using (B.2.3) and the dominated convergence theorem, we can establish a bound for the objective function $J(\theta)$ from (4.2.3):

$$\left|\nabla^2_\theta J(\theta)\right| \leq C\left(\left|\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \beta\right|^2 + \left|\nabla_\theta \mathbb{E}_{Y\sim\pi_\theta} f(Y)\right|^2\right) \leq C, \tag{4.3.55}$$

and therefore the gradient $\nabla_\theta J(\theta)$ is Lipschitz continuous with respect to θ.

Lemma 4.3.8. *Under Assumption 4.2.1, choose $\mu > 0$ in (4.3.3) such that for the given $\kappa > 0$, one has $3\mu + \frac{\mu}{8\kappa} = \frac{1}{2L_{\nabla J}}$, where $L_{\nabla J}$ is the Lipschitz constant of $\nabla_\theta J(\theta)$ in (4.2.3). Then, for k large enough (where k can be random) and $\eta > 0$ small enough (potentially random depending on k), $\int_{\tau_k}^{\sigma_k+\eta} \alpha_s ds > \mu$ with probability one. In addition, we also have $\frac{\mu}{2} \leq \int_{\tau_k}^{\sigma_k} \alpha_s ds \leq \mu$ with probability one.*

Proof. We use a "proof by contradiction". Assume that $\int_{\tau_k}^{\sigma_k+\eta} \alpha_s ds \leq \mu$ and let $\delta > 0$ be such that $\delta < \mu/8$. Without loss of generality, we assume that for any k, η is small enough such that for any $s \in [\tau_k, \sigma_k + \eta]$ one has $|\nabla_\theta J(\theta_s)| \leq 3|\nabla_\theta J(\theta_{\tau_k})|$.

Combining (4.3.1) and (4.3.2) yields

$$\frac{d\theta_t}{dt} = -\alpha_t \nabla_\theta J(\theta_t) - 2\alpha_t Z^1_t - 2\alpha_t Z^2_t \tag{4.3.56}$$

and thus

$$\begin{aligned} |\theta_{\sigma_k+\eta} - \theta_{\tau_k}| &\leq \int_{\tau_k}^{\sigma_k+\eta} \alpha_t \left|\nabla_\theta J(\theta_t)\right| dt + 2\left|\int_{\tau_k}^{\sigma_k+\eta} \alpha_t Z^1_t dt\right| + 2\left|\int_{\tau_k}^{\sigma_k+\eta} \alpha_t Z^2_t dt\right| \\ &\leq 3\left|\nabla_\theta J(\theta_{\tau_k})\right| \mu + I_1 + I_2. \end{aligned} \tag{4.3.57}$$

By Lemmas 4.3.5 and 4.3.7, we have that for k large enough,

$$\begin{aligned} I_1 &\leq 2|\Delta^1_{\tau_k,\sigma_k+\eta}| \leq \delta < \mu/16 \\ I_2 &\leq 2|\Delta^2_{\tau_k,\sigma_k+\eta}| \leq \delta < \mu/16. \end{aligned} \tag{4.3.58}$$

In addition, we also have by definition that $\frac{\kappa}{|\nabla_\theta J(\theta(\tau_k))|} \leq 1$. Combining (4.3.57) and (4.3.58) yields

$$|\theta_{\sigma_k+\eta} - \theta_{\tau_k}| \leq \left|\nabla_\theta J\left(\theta_{\tau_k}\right)\right| \left(3\mu + \frac{\mu}{8\kappa}\right) = \frac{1}{2L_{\nabla J}} \left|\nabla_\theta J\left(\theta_{\tau_k}\right)\right|.$$

This means that

$$|\nabla_\theta J(\theta_{\sigma_k+\eta}) - \nabla_\theta J(\theta_{\tau_k})| \le L_{\nabla J}|\theta_{\sigma_k+\eta} - \theta_{\tau_k}| \le \frac{1}{2}|\nabla_\theta J(\theta_{\tau_k})|,$$

and thus

$$\frac{1}{2}|\nabla_\theta J(\theta_{\tau_k})| \le |\nabla_\theta J(\theta_{\sigma_k+\eta})| \le 2|\nabla_\theta J(\theta_{\tau_k})|.$$

However, this produces a contradiction since it implies $\int_{\tau_k}^{\sigma_k+\eta} \alpha_s ds > \mu$; otherwise, from the definition of σ_k in (4.3.3), we will have $\sigma_k+\eta \in [\tau_k, \sigma_k]$. This concludes the proof of the first part of the lemma.

The proof of the second part of the lemma is straightforward. By its definition in (4.3.3), we have that $\int_{\tau_k}^{\sigma_k} \alpha_s ds \le \mu$. It remains to show that $\int_{\tau_k}^{\sigma_k} \alpha_s ds \ge \frac{\mu}{2}$. We have shown that $\int_{\tau_k}^{\sigma_k+\eta} \alpha_s ds > \mu$. For k large enough and η small enough we can choose that $\int_{\sigma_k}^{\sigma_k+\eta} \alpha_s ds \le \frac{\mu}{2}$. The conclusion then follows. □

Lemma 4.3.9. *Under Assumption 4.2.1, suppose that there exists an infinite number of intervals $I_k = [\tau_k, \sigma_k)$. Then there is a fixed constant $\gamma_1 = \gamma_1(\kappa) > 0$ such that for k large enough (where k can be random),*

$$J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) \le -\gamma_1. \tag{4.3.59}$$

Proof. By chain rule, we have that

$$\begin{aligned} J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) &= -\int_{\tau_k}^{\sigma_k} \alpha_\rho |\nabla J(\theta_\rho)|^2 d\rho - 2\int_{\tau_k}^{\sigma_k} \alpha_\rho \langle \nabla J(\theta_\rho),\ Z_\rho^1\rangle d\rho - 2\int_{\tau_k}^{\sigma_k} \alpha_\rho \langle \nabla J(\theta_\rho),\ Z_\rho^2\rangle d\rho \\ &=: M_{1,k} + M_{2,k} + M_{3,k}. \end{aligned} \tag{4.3.60}$$

For $M_{1,k}$, note that for $\rho \in [\tau_k, \sigma_k]$ we have $\frac{|\nabla_\theta J(\theta_{\tau_k})|}{2} \le |\nabla_\theta J(\theta_\rho)| \le 2|\nabla_\theta J(\theta_{\tau_k})|$. Thus for sufficiently large k, we have by Lemma 4.3.8

$$M_{1,k} \le -\frac{|\nabla_\theta J(\theta_{\tau_k})|^2}{4}\int_{\tau_k}^{\sigma_k} \alpha_\rho d\rho \le -\frac{|\nabla_\theta J(\theta_{\tau_k})|^2}{8}\mu.$$

For $M_{2,k}$ and $M_{3,k}$, we can use the same method of Poisson equations as in Lemmas 4.3.5 and 4.3.7. Define

$$\begin{aligned} G^1(x,\tilde{x},\theta) &= \langle \nabla_\theta J(\theta),\ (\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \beta)(\nabla f(x)\tilde{x} - \nabla_\theta \mathbb{E}_{Y\sim\pi_\theta} f(Y))^\top \rangle \\ G^2(x,\tilde{x},\bar{x},\theta) &= \langle \nabla_\theta J(\theta),\ (f(\bar{x}) - \mathbb{E}_{Y\sim\pi_\theta} f(Y))(\nabla f(x)\tilde{x})^\top \rangle, \end{aligned} \tag{4.3.61}$$

and use the solution of the corresponding Poisson equations

$$\begin{aligned} \mathcal{L}^\theta_{x,\tilde{x}} v^1(x,\tilde{x},\theta) &= G^1(x,\tilde{x},\theta), \\ \mathcal{L}^\theta_{x,\tilde{x},\bar{x}} v^2(x,\tilde{x},\bar{x},\theta) &= G^2(x,\tilde{x},\bar{x},\theta), \end{aligned} \tag{4.3.62}$$

as in Lemmas 4.3.5 and 4.3.7 to prove $M_{2,k}, M_{3,k} \to 0$ as $k \to \infty$ almost surely.

Combining the above results, we obtain that for k large enough such that $|M_{2,k}| + |M_{3,k}| \le \delta <$

$\frac{\mu}{16}\kappa^2$

$$\begin{aligned} J\left(\theta_{\sigma_k}\right) - J\left(\theta_{\tau_k}\right) &\leq -\frac{\left|\nabla J\left(\theta_{\tau_k}\right)\right|^2}{8}\mu + \delta \\ &\leq -\frac{\mu}{8}\kappa^2 + \frac{\mu}{16}\kappa^2 \\ &= -\frac{\mu}{16}\kappa^2. \end{aligned} \tag{4.3.63}$$

Let $\gamma_1 = \frac{\mu}{16}\kappa^2$, which concludes the proof of the lemma. □

Lemma 4.3.10. *Under Assumption 4.2.1, suppose that there exists an infinite number of intervals $I_k = [\tau_k, \sigma_k)$. Then, there is a fixed constant $\gamma_2 < \gamma_1$ such that for k large enough (where k can be random),*

$$J(\theta_{\tau_k}) - J(\theta_{\sigma_{k-1}}) \leq \gamma_2. \tag{4.3.64}$$

Proof. By chain rule, we have

$$\begin{aligned} &J(\theta_{\tau_k}) - J(\theta_{\sigma_{k-1}}) \\ &= -\int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \left|\nabla_\theta J(\theta_\rho)\right|^2 d\rho + \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \langle \nabla_\theta J(\theta_\rho),\ Z_\rho^1\rangle d\rho + \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \langle \nabla_\theta J(\theta_\rho),\ Z_\rho^2\rangle d\rho \\ &\leq \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \langle \nabla_\theta J(\theta_\rho),\ Z_\rho^1\rangle d\rho + \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \langle \nabla_\theta J(\theta_\rho),\ Z_\rho^2\rangle d\rho. \end{aligned} \tag{4.3.65}$$

As in the proof of Lemma 4.3.9 we get that for k large enough, the right hand side of the last display can be arbitrarily small, which concludes the proof of the lemma. □

Proof of Theorem 4.2.2: Recalling (4.3.3), we know τ_k is the first time $|\nabla_\theta J(\theta_t)| > \kappa$ when $t > \sigma_{k-1}$. Thus, if for any fixed $\kappa > 0$, there only exists a finite number of times τ_k, then there is a finite T^* such that $|\nabla_\theta J(\theta_t)| \leq \kappa$ for $t \geq T^*$ and the proof of (4.2.2) is complete. We now use a "proof by contradiction". Suppose there are an infinite number of times τ_k, then by Lemma 4.3.9 and 4.3.10, we have for sufficiently large k (integer k can be random) that

$$J\left(\theta_{\sigma_k}\right) - J\left(\theta_{\tau_k}\right) \leq -\gamma_1$$

$$J\left(\theta_{\tau_k}\right) - J\left(\theta_{\sigma_{k-1}}\right) \leq \gamma_2$$

with $0 < \gamma_2 < \gamma_1$. Choose N large enough so that the above relations hold simultaneously for $k \geq N$. Then for all $n \geq N$

$$\begin{aligned} J\left(\theta_{\tau_{n+1}}\right) - J\left(\theta_{\tau_N}\right) &= \sum_{k=N}^{n} \left[J\left(\theta_{\sigma_k}\right) - J\left(\theta_{\tau_k}\right) + J\left(\theta_{\tau_{k+1}}\right) - J\left(\theta_{\sigma_k}\right)\right] \\ &\leq \sum_{k=N}^{n} \left(-\gamma_1 + \gamma_2\right) \\ &< \left(n - N\right) \times \left(-\gamma_1 + \gamma_2\right). \end{aligned} \tag{4.3.66}$$

Letting $n \to \infty$, we observe that $J\left(\theta_{\tau_n}\right) \to -\infty$, which is a contradiction, since by definition $J(\theta_t) \geq 0$. Thus, there can be at most finitely many τ_k. Thus, there exists a finite random time T such that almost surely $|\nabla_\theta J(\theta_t)| < \kappa$ for $t \geq T$. Since κ is arbitrarily chosen, we have proven that $|\nabla_\theta J(\theta_t)| \to 0$ as $t \to \infty$ almost surely. □

4.4 Numerical Performance of the Online Algorithm

In this section, we will implement the continuous-time stochastic gradient descent algorithm (4.1.4) and evaluate its numerical performance. The algorithm is implemented for a variety of linear and nonlinear models. The algorithm is also implemented for the simultaneous optimization of both the drift and volatility functions, optimizing over a path-dependent SDE, and optimizing over the auto-covariance of an SDE. In our numerical experiments, we found that the performance of the algorithm can depend upon carefully selecting hyperparameters such as the learning rate and mini-batch size. The algorithm with mini-batch size N is

$$\begin{aligned}
\frac{d\theta_t}{dt} &= -2\alpha_t \left(\frac{1}{N} \sum_{i=1}^{N} \left(f(\bar{X}_t^{(i)}) - \beta \right) \right) \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \left(\nabla f \left(X_t^{(i)} \right) \tilde{X}_t^{(i)} \right)^\top \right), \\
d\tilde{X}_t^{(i)} &= \left(\nabla_x \mu \left(X_t^{(i)}, \theta_t \right) \tilde{X}_t^{(i)} + \nabla_\theta \mu \left(X_t^{(i)}, \theta_t \right) \right) dt + \left(\nabla_x \sigma \left(X_t^{(i)}, \theta_t \right) \tilde{X}_t^{(i)} + \nabla_\theta \sigma \left(X_t^{(i)}, \theta_t \right) \right) dW_t^{(i)}, \\
dX_t^{(i)} &= \mu \left(X_t^{(i)}, \theta_t \right) dt + \sigma \left(X_t^{(i)}, \theta_t \right) dW_t^{(i)}, \\
d\bar{X}_t^{(i)} &= \mu \left(\bar{X}_t^{(i)}, \theta_t \right) dt + \sigma \left(\bar{X}_t^{(i)}, \theta_t \right) d\bar{W}_t^{(i)},
\end{aligned} \tag{4.4.1}$$

for $i = 1, 2, \cdots, N$. The notation (i) indicates the i-th sample in the mini-batch.

$$\frac{1}{N} \sum_{i=1}^{N} \left(f \left(\bar{X}_t^{(i)} \right) - \beta \right), \quad \frac{1}{N} \sum_{i=1}^{N} \left(\nabla f \left(X_t^{(i)} \right) \tilde{X}_t^{(i)} \right)$$

are stochastic estimates of

$$\mathbb{E}_{\pi_{\theta_t}} \left[f(Y) - \beta \right], \quad \nabla_\theta \left(\mathbb{E}_{\pi_{\theta_t}} \left[f(X) - \beta \right] \right).$$

A larger mini-batch size reduces the noise in the estimation of the gradient descent direction. The learning rate must decay as $t \to \infty$, but it should not be decreased too rapidly and the initial magnitude should be large enough so that the algorithm converges quickly. In our examples where there is a unique global minimizer, our algorithm will always converge to the optimum if we choose the correct learning rate. For the examples with multiple global minimizers, the algorithm will converge to one of the global minimizers.

Remark 4.4.1. We discuss below some important aspects of the numerical implementation:

(a) Discretization of SDEs: To implement the algorithm (4.4.1), we use an Euler scheme with step size $\Delta = 10^{-3} - 10^{-2}$. For example, $X_t^{(i)}$ is simulated as:

$$\begin{aligned}
X_{(n+1)\Delta}^{(i)} = X_{n\Delta}^{(i)} &+ \left(\nabla_x \mu(X_{n\Delta}^{(i)}, \theta_{n\Delta}) \tilde{X}_{n\Delta}^{(i)} + \nabla_\theta \mu(X_{n\Delta}^{(i)}, \theta_{n\Delta}) \right) * \Delta \\
&+ \left(\nabla_x \sigma(X_{n\Delta}^{(i)}, \theta_{n\Delta}) \tilde{X}_{n\Delta}^{(i)} + \nabla_\theta \sigma(X_{n\Delta}^{(i)}, \theta_{n\Delta}) \right) * N(0,1) * \sqrt{\Delta},
\end{aligned} \tag{4.4.2}$$

(b) Learning Rate and mini-batch size: The learning rate can be chosen to be piecewise constant or gradually decreasing with learning rate schedule

$$\alpha_t = \frac{C}{1+t},$$

where C is also a hyper-parameter needs to be selected. The mini-batch size N that we use is of the order $10^2 - 10^4$.

(c) Initial Values for SDE simulations: In (4.4.10), the initial value of the gradient process $\tilde{X}_t$ must be zero. The choice of initial points for $X_t, \bar{X}_t$ is flexible. In our experiments, we usually choose $X_0 = \bar{X}_0 = 1$. θ_t can be randomly initialized or initialized at a deterministic point such as zero.

(d) Objective Function: For some simple examples, we can directly calculate the objective function in closed form. For those examples, we directly use that formula to compute the objective function during training. For the more complex examples (with no closed-form formula), we always approximate the objective function $J(\theta)$ using a time-average since, due to the ergodic theorem,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X_s^\theta) ds = \mathbb{E}_{Y\sim\pi_\theta} f(Y) \quad \text{a.s.} \tag{4.4.3}$$

4.4.1 One-Dimensional Ornstein–Uhlenbeck Process

We start with a simple case of a one-dimensional Ornstein–Uhlenbeck process $X_t^\theta \in \mathbb{R}$:

$$dX_t^\theta = (\theta - X_t^\theta)dt + dW_t. \tag{4.4.4}$$

We will use the algorithm (4.1.4) to learn the minimizer for

$$J(\theta) = (\mathbb{E}_{Y\sim\pi_\theta} Y - 2)^2. \tag{4.4.5}$$

Note that in this case we have the closed-form solution $\pi_\theta \sim N\left(\theta, \frac{1}{2}\right)$ and thus the global minimizer is $\theta^* = 2$. In Figure 4.1, several different sample paths generated by the online algorithm are plotted where all trained parameters converges to the global minimizer ($\theta^* = 2$).

Similarly, we use the algorithm (4.1.4) to learn the minimizer for

$$J(\theta) = (\mathbb{E}_{Y\sim\pi_\theta} Y^2 - 2)^2. \tag{4.4.6}$$

In this case, the two global minimizers are $\theta^* = \pm\sqrt{1.5}$. In Figure 4.2, the parameter trained by the online algorithm converges to a global minimizer. The global minimizer which the algorithm converges to depends on the initial value of θ_0.

We now consider a more general Ornstein–Uhlenbeck process with parameters $\theta = \left(\theta^1, \theta^2\right)$:

$$dX_t^\theta = \left(\theta^1 - \theta^2 X_t^\theta\right) dt + dW_t, \tag{4.4.7}$$

The online algorithm (4.1.4) is used to learn the minimizer for the objective function $J(\theta) =$

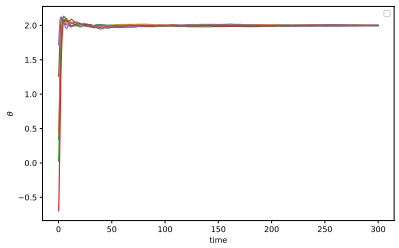

Figure 4.1: Online Algorithm for the objective function (4.4.5).

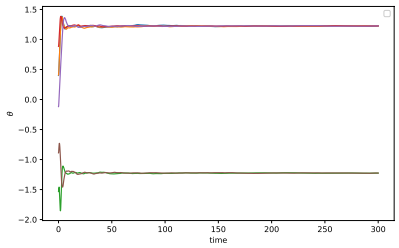

Figure 4.2: Online Algorithm for the objective function (4.4.6).

$(\mathbb{E}_{Y\sim\pi_\theta} Y^2 - 2)^2$. Algorithm (4.4.1) will be used:

$$\begin{aligned}
d\theta_t^1 &= -4\alpha_t \left(\frac{1}{N}\sum_{i=1}^{N} \left(\bar{X}_t^{(i)}\right)^2 - 2 \right) \cdot \left(\frac{1}{N}\sum_{i=1}^{N} X_t^{(i)} \tilde{X}_t^{1,(i)} \right) dt \\
d\theta_t^2 &= -4\alpha_t \left(\frac{1}{N}\sum_{i=1}^{N} \left(\bar{X}_t^{(i)}\right)^2 - 2 \right) \cdot \left(\frac{1}{N}\sum_{i=1}^{N} X_t^{(i)} \tilde{X}_t^{2,(i)} \right) dt \\
dX_t^{(i)} &= \left(\theta_t^1 - \theta_t^2 X_t^{(i)}\right) dt + dW_t^i \\
d\tilde{X}_t^{1,(i)} &= \left(1 - \theta_t^2 \tilde{X}_t^{1,(i)}\right) dt \\
d\tilde{X}_t^{2,(i)} &= \left(-X_t^{(i)} - \theta_t^2 \tilde{X}_t^{2,(i)}\right) dt \\
d\bar{X}_t^{(i)} &= \left(\theta_t^1 - \theta_t^2 \bar{X}_t^{(i)}\right) dt + d\bar{W}_t^i
\end{aligned} \tag{4.4.8}$$

for $i = 1, 2, \cdots, N$. To make the training more stable and accelerate the convergence rate, we choose the batch size $N = 10000$. Figure 4.3 and 4.4 show the dynamic of the parameters and objective function during training.

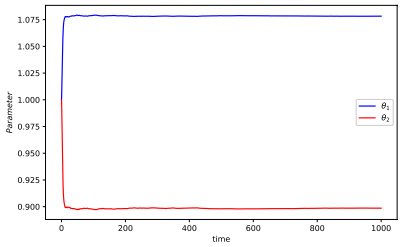

Figure 4.3: Parameters for algorithm (4.4.8).

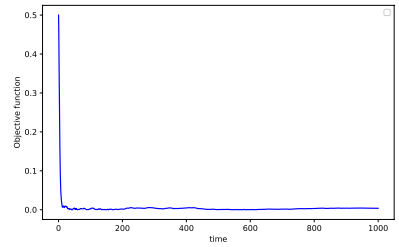

Figure 4.4: Objective function for algorithm (4.4.8).

4.4.2 One-Dimensional Nonlinear Process

We now use the online algorithm to optimize over the stationary distribution of a one-dimensional nonlinear process

$$dX_t^\theta = \left(\theta - X_t^\theta - \left(X_t^\theta\right)^3\right) dt + dW_t. \tag{4.4.9}$$

We use the algorithm (4.1.4) to learn the minimizer of $J(\theta) = \left(\mathbb{E}_{Y\sim\pi_\theta} Y^2 - 2\right)^2$. The mini-batch algorithm (4.4.10) is used:

$$\begin{aligned} d\theta_t &= -4\alpha_t \left(\frac{1}{N}\sum_{i=1}^{N}\left(\bar{X}_t^{(i)}\right)^2 - 2\right)\cdot\left(\frac{1}{N}\sum_{i=1}^{N} X_t^{(i)}\tilde{X}_t^{(i)}\right)dt \\ dX_t^{(i)} &= \left(\theta_t - X_t^{(i)} - \left(X_t^{(i)}\right)^3\right)dt + dW_t^{(i)} \\ d\tilde{X}_t^{(i)} &= \left(1 - \tilde{X}_t^{(i)} - 3\left(X_t^{(i)}\right)^2\tilde{X}_t^{(i)}\right)dt \\ d\bar{X}_t^{(i)} &= \left(\theta_t - \bar{X}_t^{(i)} - \left(\bar{X}_t^{(i)}\right)^3\right)dt + d\bar{W}_t^{(i)} \end{aligned} \tag{4.4.10}$$

for $i = 1, 2, \cdots, N$. Figure 4.5 shows the convergence of the parameter θ_t. In Figure 4.6, the objective function decays to zero (the global minimum) very quickly.

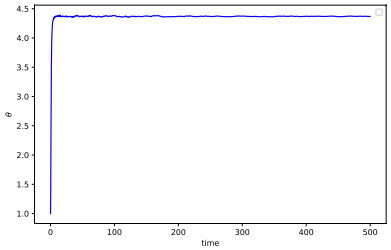

Figure 4.5: Parameter for algorithm (4.4.10).

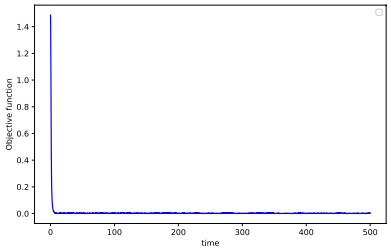

Figure 4.6: Objective function for algorithm (4.4.10).

4.4.3 Optimizing over the Drift and Volatility Coefficients

We now optimize over the drift and volatility functions of the process

$$dX_t^\theta = (\mu - X_t^\theta)dt + \sigma dW_t \tag{4.4.11}$$

with parameters $\theta = (\mu, \sigma)$. The online algorithm (4.1.4) is used to learn the minimizer of $J(\theta) = (\mathbb{E}_{Y\sim\pi_\theta} Y^2 - 2)^2$. The mini-batch algorithm (4.4.1) is used:

$$\begin{aligned} d\mu_t &= -4\alpha_t \left(\frac{1}{N}\sum_{i=1}^{N}\left(\bar{X}_t^{(i)}\right)^2 - 2\right)\cdot\left(\frac{1}{N}\sum_{i=1}^{N} X_t^{(i)}\tilde{X}_t^{1,(i)}\right)dt \\ d\sigma_t &= -4\alpha_t \left(\frac{1}{N}\sum_{i=1}^{N}\left(\bar{X}_t^{(i)}\right)^2 - 2\right)\cdot\left(\frac{1}{N}\sum_{i=1}^{N} X_t^{(i)}\tilde{X}_t^{2,(i)}\right)dt \\ dX_t^i &= \left(\mu_t - X_t^{(i)}\right)dt + \sigma_t dW_t^{(i)} \\ d\tilde{X}_t^{1,(i)} &= \left(1 - \tilde{X}_t^{1,(i)}\right)dt \\ d\tilde{X}_t^{2,(i)} &= -\tilde{X}_t^{2,(i)}dt + dW_t^{(i)} \\ d\bar{X}_t^{(i)} &= \left(\mu_t - \bar{X}_t^{(i)}\right)dt + \sigma_t d\bar{W}_t^{(i)} \end{aligned} \tag{4.4.12}$$

for $i = 1, 2, \cdots, N$. In Figure 4.7, the trained parameters μ_t, σ_t converge and in Figure 4.8 the objective function $J(\theta_t) \to 0$ very quickly.

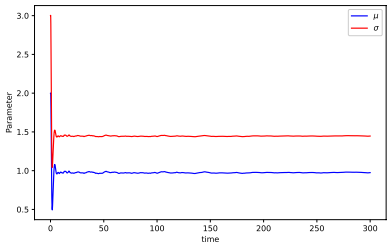

Figure 4.7: Parameters for algorithm (4.4.12).

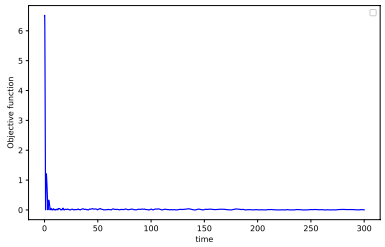

Figure 4.8: Objective function for algorithm (4.4.12).

We also implement the online algorithm for the nonlinear process

$$dX_t^\theta = \left(\mu - \left(X_t^\theta\right)^3\right) dt + \sigma X_t^\theta dW_t, \tag{4.4.13}$$

where $\theta = (\mu, \sigma)$ are the parameters and the objective function is $J(\theta) = \left(\mathbb{E}_{Y\sim\pi_\theta} Y^2 - 10\right)^2$. The mini-batch algorithm (4.4.1) now becomes:

$$\begin{aligned}
d\mu_t &= -4\alpha_t \left(\frac{1}{N}\sum_{i=1}^N \left(\bar{X}_t^{(i)}\right)^2 - 2\right) \cdot \left(\frac{1}{N}\sum_{i=1}^N X_t^{(i)} \tilde{X}_t^{1,(i)}\right) dt \\
d\sigma_t &= -4\alpha_t \left(\frac{1}{N}\sum_{i=1}^N (\bar{X}_t^{(i)})^2 - 2\right) \cdot \left(\frac{1}{N}\sum_{i=1}^N X_t^{(i)} \tilde{X}_t^{2,(i)}\right) dt \\
dX_t^i &= (\mu_t - \left(X_t^{(i)}\right)^3)dt + \sigma_t X_t^{(i)} dW_t^{(i)} \\
d\tilde{X}_t^{1,(i)} &= \left(1 - 3\left(X_t^{(i)}\right)^2 \tilde{X}_t^{1,(i)}\right) dt + \sigma_t \tilde{X}_t^{1,(i)} dW_t^{(i)} \\
d\tilde{X}_t^{2,(i)} &= -3\left(X_t^{(i)}\right)^2 \tilde{X}_t^{2,(i)} dt + \left(X_t^{(i)} + \sigma_t \tilde{X}_t^{2,(i)}\right) dW_t^{(i)} \\
d\bar{X}_t^{(i)} &= \left(\mu_t - \left(\bar{X}_t^{(i)}\right)^3\right) dt + \sigma_t \bar{X}_t^{(i)} d\bar{W}_t^{(i)}
\end{aligned} \tag{4.4.14}$$

for $i = 1, 2, \cdots, N$. In Figure 4.9, the trained parameters μ_t, σ_t converge and in Figure 4.10 the objective function $J(\theta_t) \to 0$ very quickly.

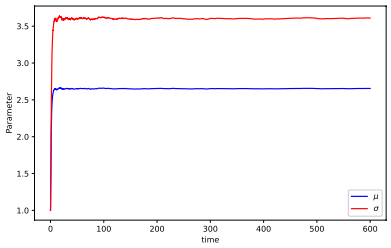

Figure 4.9: Parameters evolution for algorithm (4.4.14).

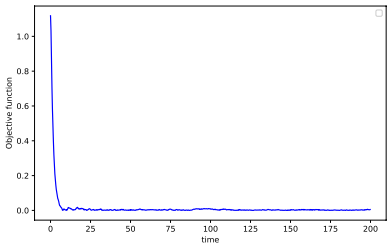

Figure 4.10: Objective function for algorithm (4.4.14).

4.4.4 Multi-Dimensional Independent Ornstein–Uhlenbeck Process

We next consider a simple multi-dimensional Ornstein–Uhlenbeck process which consists of m independent copies of (4.4.7). For the parameter $\theta = (\theta^1, \theta^2) \in R^{2m}$, let the m-dimensional Ornstein–Uhlenbeck process be

$$dX_t^\theta = \left(\theta^1 - \theta^2 \odot X_t^\theta\right) dt + dW_t, \tag{4.4.15}$$

where $X_t^\theta \in R^m$, $W_t \in R^m$, and $\odot$ is an element-wise product. The objective function is

$$J(\theta) := \left(\sum_{k=1}^{m} \mathbb{E}_{Y\sim\pi_\theta}|Y_k|^2 - 2m\right)^2. \tag{4.4.16}$$

The online algorithm (4.1.4) is

$$\begin{aligned}
d\theta_t^1 &= -4\alpha_t \left(|\bar{X}_t|^2 - 2\right) X_t \odot \tilde{X}_t^1 dt,\\
d\theta_t^2 &= -4\alpha_t \left(|\bar{X}_t|^2 - 2\right) X_t \odot \tilde{X}_t^2 dt,\\
dX_t &= \left(\theta_t^1 - \theta_t^2 \odot X_t\right) dt + dW_t^i,\\
d\tilde{X}_t^1 &= \left(1 - \theta_t^2 \odot \tilde{X}_t^1\right) dt,\\
d\tilde{X}_t^2 &= \left(-X_t - \theta_t^2 \odot \tilde{X}_t^2\right) dt,\\
d\bar{X}_t &= \left(\theta_t^1 - \theta_t^2 \odot \bar{X}_t\right) dt + d\bar{W}_t^i.
\end{aligned} \tag{4.4.17}$$

We implement the algorithm for $m = 3$ and $m = 10$. In Figures 4.11 and 4.12, the objective functions $J(\theta_t) \to 0$ as t becomes large.

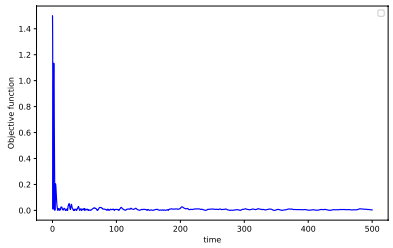

Figure 4.11: Objective function for (4.4.17) with $m = 3$.

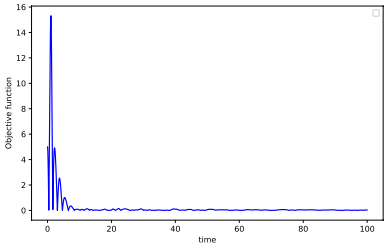

Figure 4.12: Objective function for (4.4.17) with $m = 10$.

4.4.5 Multi-Dimensional Correlated Ornstein–Uhlenbeck Process

For the parameters $\theta = (\mu, \sigma)$ with $\mu \in \mathbb{R}^m$, $\sigma \in \mathbb{R}^{m\times m}$, let the m-dimensional process X_t^θ satisfy

$$dX_t^\theta = \left(\mu - X_t^\theta\right) dt + \sigma dW_t, \tag{4.4.18}$$

where $W_t \in \mathbb{R}^m$. Let $X_t^{\theta,i}$ denote the i-th element of X_t^θ and define $\tilde{X}_t^\mu$ and $\tilde{X}_t^\sigma$ as the Jacobian matrices of X_t^θ with respect to μ and σ:

$$\begin{aligned} \tilde{X}_t^\mu &= \nabla_\mu X_t^\theta \in \mathbb{R}^{m\times m}, \quad \tilde{X}_t^{\mu,i} = \nabla_\mu X_t^{\theta,i} \in \mathbb{R}^m, \\ \tilde{X}_t^\sigma &= \nabla_\sigma X_t^\theta \in \mathbb{R}^{m\times m\times m}, \quad \tilde{X}_t^{\sigma,i} = \nabla_\sigma X_t^{\theta,i} \in \mathbb{R}^{m\times m}. \end{aligned} \tag{4.4.19}$$

Noting that for $i \in \{1, 2, \cdots, m\}$

$$dX_t^{\theta,i} = \left(\mu_i - X_t^{\theta,i}\right) dt + \sum_j \sigma_{i,j} dW_t^j,$$

now the algorithm (4.1.4) becomes

$$\begin{aligned} d\mu_t &= -4\alpha_t \left(|\bar{X}_t|^2 - 2m\right) \left(\sum_{k=1}^m X_t^k \tilde{X}_t^{\mu,k}\right) dt \\ d\lambda_t &= -4\alpha_t \left(|\bar{X}_t|^2 - 2m\right) \left(\sum_{k=1}^m X_t^k \tilde{X}_t^{\lambda,k}\right) dt \\ dX_t &= \left(\mu_t - X_t\right) dt + \sigma_t dW_t \\ d\bar{X}_t &= \left(\mu_t - \bar{X}_t\right) dt + \sigma_t d\bar{W}_t \\ d\tilde{X}_t^\mu &= \left(I_m - \tilde{X}_t^\mu\right) dt \\ d\tilde{X}_t^{\sigma,i} &= -\tilde{X}_t^{\sigma,i} dt + D_i\left(dW_t\right), \quad i \in \{1, \cdots, m\} \end{aligned} \tag{4.4.20}$$

where I_m is the $m \times m$ identity matrix and where $D_i(dW_t)$ is a $m \times m$ matrix with all elements equal to 0 except i-th column being dW_t. We examine the algorithm's performance for dimensions $m = 3, 10$. In Figures 4.13 and 4.14, the objective function $J(\theta_t) \to 0$.

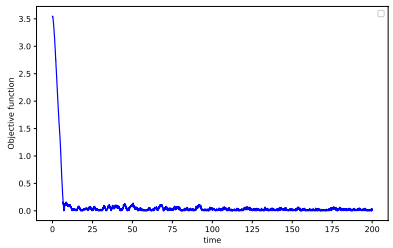

Figure 4.13: Object function for (4.4.20) with $m = 3$.

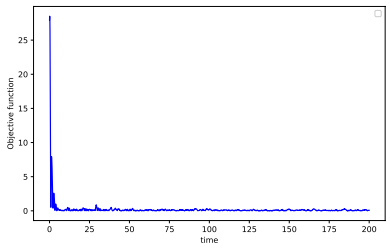

Figure 4.14: Objective function for (4.4.20) with $m = 10$.

4.4.6 Multi-dimensional Nonlinear SDE

In our next example, we optimize over the stationary distribution of a multi-dimensional nonlinear SDE:

$$dX_t^{\theta,i} = \left(\theta - \frac{1}{N}\sum_{j=1}^{N} X_t^{\theta,j} - (X_t^{\theta,i})^3\right) dt + dW_t^i, \quad i = 1, 2, \cdots, N, \tag{4.4.21}$$

and now N is the number of agents in the system (4.4.21) instead of mini-batch size as before. The objective function is

$$J(\theta) = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{Y\sim\pi_\theta} Y_i^2 - 2\right)^2. \tag{4.4.22}$$

The nonlinear SDE (4.4.21) has a mean-field limit as $N \to \infty$. Thus, for large N, our algorithm could also be used to optimize over the mean-field limit equation ([140]) for (4.4.21). The online algorithm for (4.4.21) is

$$\begin{aligned}
d\theta_t &= -4\alpha_t \left(\frac{1}{N}\sum_{i=1}^{N} \left(\bar{X}_t^i\right)^2 - 2\right) \times \left(\frac{1}{N}\sum_{i=1}^{N} X_t^i \tilde{X}_t^i\right) dt \\
dX_t^i &= \left(\theta_t - \frac{1}{N}\sum_{j=1}^{N} X_t^j - \left(X_t^i\right)^3\right) dt + dW_t^i \\
d\tilde{X}_t^i &= \left(1 - \frac{1}{N}\sum_{j=1}^{N} \tilde{X}_t^{1,j} - 3\left(X_t^i\right)^2 \tilde{X}_t^i\right) dt \\
d\bar{X}_t^i &= \left(\theta_t - \frac{1}{N}\sum_{j=1}^{N} \bar{X}_t^j - \left(\bar{X}_t^i\right)^3\right) dt + d\bar{W}_t^i
\end{aligned} \tag{4.4.23}$$

for $i = 1, 2, \cdots, N$. We will select $N = 1,000$ for our numerical experiment. Therefore, this is an example of high-dimensional SDE model calibration where the dimension of the SDE is $N = 1,000$. Figure 4.15 and 4.16 shows the convergence of parameter and objective function.

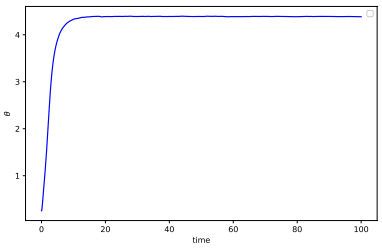

Figure 4.15: Parameter for algorithm (4.4.23).

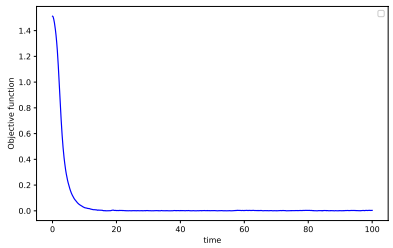

Figure 4.16: Objective function for algorithm (4.4.23).

4.4.7 Path-dependent SDE

We consider the path-dependent SDE

$$dX_t^\theta = \left(\theta - X_t^\theta - \frac{1}{t}\int_0^t X_s^\theta ds\right) dt + dW_t, \tag{4.4.24}$$

where $X_t^\theta, W_t \in \mathbb{R}$. Although path-dependent SDEs are not directly addressed by this article's convergence theory, this numerical example suggests that the online forward propagation algorithm can also be applied to path-dependent stochastic processes.

For this numerical example, the objective function is

$$J(\theta) = (\mathbb{E}_{Y\sim\pi_\theta} Y - 2)^2. \tag{4.4.25}$$

The SDE (4.4.24) does not fit the problem described in (4.1.1) and (4.1.2). However, our algorithm still can find the global optimum.

Now the online algorithm (4.1.4) is:

$$\begin{aligned}
d\theta_t &= -4\alpha_t(\bar{X}_t - 2)\tilde{X}_t dt \\
dX_t &= \left(\theta_t - X_t - \frac{1}{t}\int_0^t X_s ds\right) dt + dW_t \\
d\tilde{X}_t &= \left(1 - \tilde{X}_t - \frac{1}{t}\int_0^t \tilde{X}_s ds\right) dt \\
d\bar{X}_t &= \left(\theta_t - \bar{X}_t - \frac{1}{t}\int_0^t \bar{X}_s ds\right) dt + d\bar{W}_t.
\end{aligned} \tag{4.4.26}$$

In Figure 4.17, the trained parameter converges. The objective function $J(\theta_t)$ is approximated using a time-average. In Figure 4.18, the objective function $J(\theta_t)$ converges to 0 very quickly.

4.4.8 Optimizing over the Auto-Covariance of the Ornstein-Uhlenbeck Process

As our final numerical example, consider the Ornstein-Uhlenbeck process

$$dX_t^\theta = (\mu - \lambda X_t^\theta)dt + \sigma dW_t, \tag{4.4.27}$$

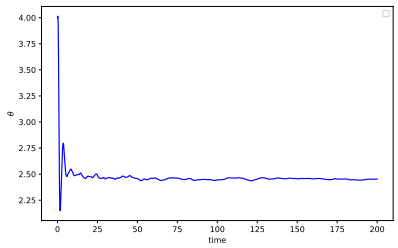

Figure 4.17: Parameter for algorithm (4.4.26).

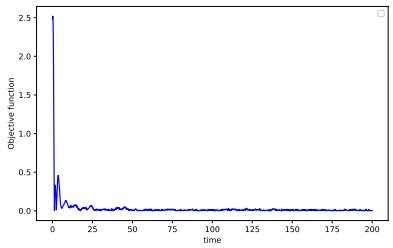

Figure 4.18: Objective function for algorithm (4.4.26).

where $\theta = (\mu, \lambda, \sigma)$. Define π_θ as the stationary distribution of X_t^θ and $\pi_{\theta,\tau}(dx, dx')$ as the stationary distribution of $(X_{t-\tau}^\theta, X_t^\theta)$. The objective function is

$$J(\theta) = \left(\mathbb{E}_{Y\sim\pi_\theta} Y - 1\right)^2 + \left(\mathbb{E}_{Y\sim\pi_\theta} Y^2 - 2\right)^2 + \left(\mathbb{E}_{Y,Y'\sim\pi_{\theta,\tau}} YY' - 1.6\right)^2, \tag{4.4.28}$$

where we will select $\tau = 0.1$ for our numerical experiment.

The online algorithm is

$$\begin{aligned}
d\mu_t &= -2\alpha_t \left[\left(\bar{X}_t - 1\right)\tilde{X}_t^1 + 2(\bar{X}_t^2 - 2)X_t\tilde{X}_t^1 + (\bar{X}_{t-\tau}\bar{X}_t - 1.6)\left(\tilde{X}_{t-\tau}^1 X_t + X_{t-\tau}\tilde{X}_t^1\right)\right] dt \\
d\lambda_t &= -2\alpha_t \left[(\bar{X}_t - 1)\tilde{X}_t^2 + 2(\bar{X}_t^2 - 2)X_t\tilde{X}_t^2 + (\bar{X}_{t-\tau}\bar{X}_t - 1.6)\left(\tilde{X}_{t-\tau}^2 X_t + X_{t-\tau}\tilde{X}_t^2\right)\right] dt \\
d\sigma_t &= -2\alpha_t \left[(\bar{X}_t - 1)\tilde{X}_t^3 + 2(\bar{X}_t^2 - 2)X_t\tilde{X}_t^3 + (\bar{X}_{t-\tau}\bar{X}_t - 1.6)\left(\tilde{X}_{t-\tau}^3 X_t + X_{t-\tau}\tilde{X}_t^3\right)\right] dt \\
dX_t &= (\mu_t - \lambda_t X_t)dt + \sigma_t dW_t \\
d\tilde{X}_t^1 &= (1 - \lambda_t\tilde{X}_t^1)dt \\
d\tilde{X}_t^2 &= (-X_t - \lambda_t\tilde{X}_t^2)dt \\
d\tilde{X}_t^3 &= -\lambda_t\tilde{X}_t^3 dt + dW_t \\
d\bar{X}_t &= (\mu_t - \lambda_t\bar{X}_t)dt + d\bar{W}_t.
\end{aligned} \tag{4.4.29}$$

Figures 4.19 - 4.22 display the trained parameters and the objective function. The trained parameters have $\sim 0.1 - 0.3\%$ relative error compared to the global minimizers. The objective function $J(\theta_t)$ is computed from the exact formula

$$J(\theta) = \left(\frac{\mu}{\lambda} - 1\right)^2 + \left(\left(\frac{\mu}{\lambda}\right)^2 + \frac{\sigma^2}{2\lambda} - 2\right)^2 + \left(\left(\frac{\mu}{\lambda}\right)^2 + \frac{\sigma^2 e^{-\lambda\tau}}{2\lambda} - 1.6\right)^2. \tag{4.4.30}$$

4.4.9 Applications to Mathematical Finance

In this section, we discuss several potential applications of the forward propagation algorithm (4.1.4) in mathematical finance. Our algorithm provides a new approach to estimate the parameters in SDE models in mathematical finance and financial econometrics [2, 35, 80, 91, 92, 93, 94, 153], including when the SDE is partially observed. Our algorithm is applicable for the calibration/estimation of SDE model parameters for long time series where ergodicity in the data is

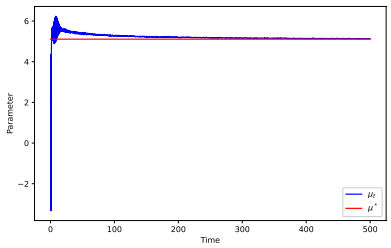

Figure 4.19: μ_t evolution in (4.4.29).

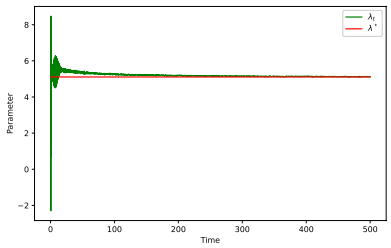

Figure 4.20: λ_t evolution in (4.4.29).

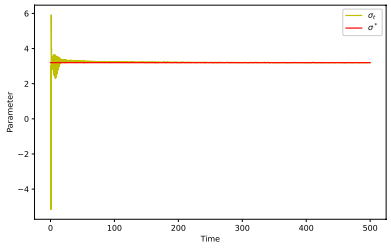

Figure 4.21: σ_t evolution in (4.4.29).

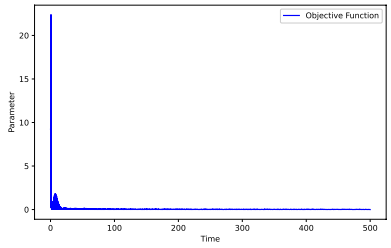

Figure 4.22: Objective function for (4.4.29).

expected. In section 4.4.10, we discuss parameter estimation in partially-observed SDE models [2, 127, 135], which are commonly used in financial econometrics [9, 42, 47, 48, 49, 54, 67].

In section 4.4.11, we discuss the application of our algorithm to solving stochastic optimal control problems for long time horizons where the ergodic framework is suitable; stochastic optimal control is important in many areas of mathematical finance such as optimal order execution and portfolio optimization [36, 65, 118, 152, 6, 8]. High-dimensional stochastic optimal control problems are computationally intractable for traditional numerical methods. Although the optimal control satisfies a Hamilton-Jacobi-Bellman (HJB) equation, finite difference methods cannot solve high-dimensional PDEs. We demonstrate that our online optimization algorithm can efficiently solve high-dimensional stochastic optimal control problems (in the ergodic setting). In order to evaluate the accuracy of our algorithm for solving stochastic optimal control problems, we implement it for several high-dimensional stochastic linear quadratic regulator (LQR) problems [58, 65, 13, 55, 152]. The LQR problem is selected since a closed-form solution is available (even in high dimensions) to evaluate the accuracy of our algorithm. (However, it should be highlighted that our online optimization algorithm can be used for the stochastic optimal control of any ergodic SDE, including nonlinear SDEs.) The online optimization algorithm learns a parametric control, either a linear function or a neural network (NN), to minimize the objective function. In both the linear and neural network cases, the algorithm can learn the optimal control. The optimal control function appears in the drift of the SDE. In the case of the neural network optimal control, the SDE is therefore a "neural network-SDE". Neural network-SDEs – sometimes referred to as neural-SDEs – are SDEs where the

drift and/or volatility of the SDE is a neural network. Neural-SDEs have recently become of great interest in mathematical finance [7, 44, 45, 46, 61, 111].

The online optimization algorithm can also be used to solve multi-agent stochastic control problems – e.g., mean-field games – which is a widely-researched topic in mathematical finance [8, 29, 30, 32, 33, 34] in the ergodic setting. The finite multi-agent stochastic optimal control problem is typically computationally intractable since the corresponding HJB equation is very high-dimensional. It will be an $N \times d$ dimensions PDE, where N is the number of agents and d is the dimension of each agent's state (i.e., SDE) process. The limit mean-field game, which approximates the finite case, may be computationally tractable to solve. However, if the state space of each agent is high-dimensional (e.g., dimension $d > 4$), the limit mean-field game will also be computationally intractable since it will be a PDE in d dimensions. In addition, the mean-field game limit may not be accurate for the finite-N case if N is not sufficiently large. Therefore, it is of interest to develop new methods for the computational solution of high-dimensional multi-agent stochastic optimal control problems in mathematical finance. As an example, we numerically implement the online optimization model for a simplified version of the multi-agent systemic risk model ([32]) in Section 4.4.12. There are N agents where each agent is modeled by an SDE. As $N \to \infty$, the system converges to a mean-field game limit. In the numerical example, we use the online optimization algorithm to solve the the high-dimensional stochastic optimal control problem corresponding to a large number of N SDEs ($N = 5,000$).

Finally, the online optimization algorithm can be used to train SDE models (including point process models) of limit order books [109] [10] [129] [99] [87]. Order books involve large numbers of high-frequency events ($\sim 10^5 - 10^6$ events per day per stock) and high-dimensional dynamics (many price levels, each with limit order submissions and cancellations, as well as market orders, hidden orders, and transactions). The large amounts of high-frequency high-dimensional data for limit order books makes this a very promising application area for the online forward propagation algorithm, which is able to asymptotically optimize general classes of models over the *entire history* of the order flow dataset (in contrast to standard methods can typically only optimize over much smaller sub-sequences).

4.4.10 Optimizing parameters in partially-observed SDE models

4.4.10.1 Two-dimensional Ornstein–Uhlenbeck Model

In this section, we focus on the following partially observed two-dimensional Ornstein–Uhlenbeck process [2] with parameters $\theta = (\alpha, \sigma_1, \sigma_2)$:

$$\begin{aligned} dX_t &= \kappa^1 (Y_t - X_t)\, dt + \sigma^1 dW_t^1 \\ dY_t &= \kappa^2 (\alpha - Y_t)\, dt + \sigma^2 dW_t^2, \end{aligned} \tag{4.4.31}$$

where the state process X_t is observable and Y_t is the latent (unobserved) process. As in Section 4.4, we can estimate the parameters by calibrating the model to the moments of the stationary distribution. In our numerical example, the objective function is

$$J(\theta) = \left(\mathbb{E}_{Y\sim\pi_\theta} Y - 1\right)^2 + \left(\mathbb{E}_{Y\sim\pi_\theta} Y^2 - 2\right)^2 + \left(\mathbb{E}_{Y\sim\pi_\theta} Y^3 - 4\right)^2. \tag{4.4.32}$$

The algorithm (4.1.4) becomes

$$\begin{aligned}
d\alpha_t &= -\alpha_t \left[\left(\bar{X}_t - 1\right)\tilde{X}_t^1 + 2\left(\bar{X}_t^2 - 2\right) X_t \tilde{X}_t^1 + 3\left(\bar{X}_t^3 - 4\right) X_t^2 \tilde{X}_t^1\right] dt \\
d\sigma_t^1 &= -\alpha_t \left[\left(\bar{X}_t - 1\right)\tilde{X}_t^2 + 2\left(\bar{X}_t^2 - 2\right) X_t \tilde{X}_t^2 + 3\left(\bar{X}_t^3 - 4\right) X_t^2 \tilde{X}_t^2\right] dt \\
d\sigma_t^2 &= -\alpha_t \left[\left(\bar{X}_t - 1\right)\tilde{X}_t^3 + 2\left(\bar{X}_t^2 - 2\right) X_t \tilde{X}_t^3 + 3\left(\bar{X}_t^3 - 4\right) X_t^2 \tilde{X}_t^3\right] dt \\
dX_t &= \kappa^1 \left(Y_t - X_t\right) dt + \sigma_t^1 dW_t^1 \\
dY_t &= \kappa^2 \left(\alpha_t - Y_t\right) dt + \sigma_t^2 dW_t^2 \\
d\tilde{X}_t^1 &= \kappa^1 \left(\tilde{Y}_t^1 - \tilde{X}_t^1\right) dt \\
d\tilde{Y}_t^1 &= \kappa^2 \left(1 - \tilde{Y}_t^1\right) dt \\
d\tilde{X}_t^2 &= -\kappa^1 \tilde{X}_t^2 dt + dW_t^1 \\
d\tilde{X}_t^3 &= \kappa^1 \left(\tilde{Y}_t^3 - \tilde{X}_t^3\right) dt \\
d\tilde{Y}_t^3 &= -\kappa^2 \tilde{Y}_t^3 dt + dW_t^2 \\
d\bar{X}_t &= \kappa^1 \left(\bar{Y}_t - \bar{X}_t\right) dt + \sigma_t^1 d\bar{W}_t^1 \\
d\bar{Y}_t &= \kappa^2 \left(\alpha_t - \bar{Y}_t\right) dt + \sigma_t^2 d\bar{W}_t^2.
\end{aligned} \tag{4.4.33}$$

Figures 4.23 and 4.24 display the parameter convergence and the objective function.

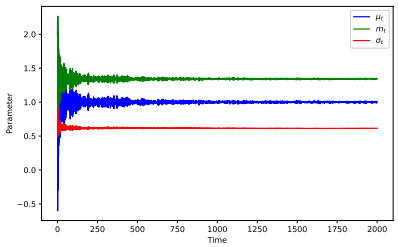

Figure 4.23: Parameters for algorithm (4.4.33).

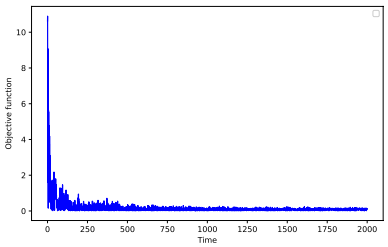

Figure 4.24: Objective function for algorithm (4.4.33).

4.4.11 Stochastic Optimal Control

The online optimization algorithm can be used to solve stochastic optimal control problems, including high-dimensional problems for which traditional numerical methods (e.g., solving the HJB equation with finite difference methods) are computationally expensive or intractable. As a numerical example we consider the classic LQR problem [4, 13, 152], which itself has many financial applications

such as optimal execution [3, 36, 35, 65]. Let $\{X_t\}_{t\geq 0}$ be the state process that satisfies the SDE

$$dX_t = (AX_t + BU_t)\,dt + \sigma dW_t, \tag{4.4.34}$$

where $X_0 = x_0, X_t \in \mathbb{R}^n$, matrix $A, \sigma \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $\{W_t\}_{t\geq 0}$ is an $\mathbb{R}^n$-valued standard Wiener process, and $\{U_t\}_{t\geq 0} \in \mathbb{R}^m$ denotes the control. The objective is to learn a control process $u.$ to minimize the following ergodic cost functional for system (4.4.34):

$$J(U.) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(X_t^{\mathrm{T}} Q X_t + U_t R U_t\right) dt, \tag{4.4.35}$$

where Q and R are positive definite matrices. It is well-known that the optimal control is given by [55]:

$$U = -R^{-1}B^\top K X, \tag{4.4.36}$$

where K is the unique solution of the following algebraic Riccati equation (ARE)

$$A^\top K + KA - KBR^{-1}B^\top K + Q = 0. \tag{4.4.37}$$

In order to evaluate the accuracy of our algorithm for solving stochastic optimal control problems, we numerically implement it for several high-dimensional stochastic (LQR) problems. The LQR problem is selected since a closed-form solution is available (even in high dimensions) to evaluate the accuracy of our algorithm. We present a series of numerical examples where the online optimization algorithm learns parametric controls for various LQR problems. The parametric control is either a linear function or a neural network.

4.4.11.1 One-dimensional Linear Control

As a first step, we implement the online optimization algorithm for the one-dimensional case with a linear control function. For simplicity, we assume that $A = -1$, $B = \sigma = Q = R = 1$ for (4.4.34):

$$\begin{aligned} dX_t^\theta &= \left(-X_t^\theta + \theta X_t^\theta\right) dt + dW_t, \\ J(\theta) &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(1+\theta^2\right) \left(X_t^\theta\right)^2 dt. \end{aligned} \tag{4.4.38}$$

The coupled system (4.4.1) becomes

$$\begin{aligned} d\theta_t &= -\alpha_t \left[\frac{1}{N}\sum_{i=1}^N \left(2\theta_t \left(X_t^{(i)}\right)^2 + 2\left(1+\theta_t^2\right) X_t^{(i)} \tilde{X}_t^{(i)}\right)\right] dt, \\ dX_t^{(i)} &= (\theta_t - 1) X_t^{(i)} dt + dW_t^{(i)}, \\ d\tilde{X}_t^{(i)} &= (X_t^{(i)} + (\theta_t - 1)\tilde{X}_t^{(i)}) dt, \end{aligned} \tag{4.4.39}$$

with $i = 1, 2, \cdots, N$. Solving the ARE (4.4.37) yields the optimal control $\theta^* = -0.41421$. Figure 4.25 shows that the parameter θ_t trained with the online optimization algorithm converges to θ^*.

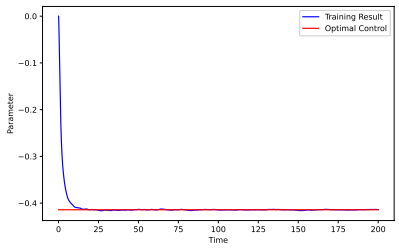

Figure 4.25: Parameter θ_t for algorithm (4.4.39)

4.4.11.2 Multi-dimensional Linear Control

We next solve a multi-dimensional LQR problem with a linear control function. For simplicity, we assume that $m = n,\ A = -I_n,\ B = \sigma = I_n$ in (4.4.34) where I_n is n dimensional identity matrix. That is,

$$\begin{aligned} dX_t^\theta &= \left(-X_t^\theta + \theta X_t^\theta\right) dt + dW_t, \\ J(\theta) &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(X_t^\theta\right)^\top \left(Q + \theta^\top R\theta\right) X_t^\theta dt, \end{aligned} \tag{4.4.40}$$

where $\theta \in \mathbb{R}^{n\times n}$. Let $X_t^{\theta,i}$ denote the i-th element of X_t^θ and define

$$\tilde{X}_t^\theta = \nabla_\theta X_t^\theta, \quad \tilde{X}_t^{\theta,i} = \nabla_\theta X_t^{\theta,i},\ \forall i \in \{1, 2, \cdots, n\}. \tag{4.4.41}$$

$\tilde{X}_t^\theta$ has dimensions $n \times n \times n$ and $\tilde{X}_t^{\theta,i}$ has dimensions $n \times n$. Note that when we are training over a mini-batch of size N, $\tilde{X}_t^\theta$ has dimensions $N \times n \times n \times n$.

We first discuss the methods necessary for the computationally efficient simulation of the gradient $\nabla_\theta X_t^\theta$. The state process from (4.4.40) satisfies

$$dX_t^{\theta,i} = \left(-X_t^{\theta,i} + \sum_{j=1}^n \theta_{i,j} X_t^{\theta,j}\right) dt + dW_t^i, \tag{4.4.42}$$

and therefore

$$d\tilde{X}_t^{\theta,i} = \left(-\tilde{X}_t^{\theta,i} + \sum_{j=1}^n \theta_{i,j} \tilde{X}_t^{\theta,j} + D_i(X_t^\theta)\right) dt, \tag{4.4.43}$$

where $D_i(X_t^\theta)$ is an $n \times n$ matrix whose elements are all zeros except for the i-th row, which has values X_t^θ. The gradient of the objective function in (4.4.40) is:

$$\begin{aligned} &\nabla_\theta \left[\left(X_t^\theta\right)^\top \left(Q + \theta^\top R\theta\right) X_t^\theta\right] \\ &= \sum_{i,j} \nabla_\theta \left(\delta_{i,j} + \sum_{k=1}^n \theta_{k,i}\theta_{k,j}\right) X_t^{\theta,i} X_t^{\theta,j} + 2\sum_{i,j} \nabla_\theta X_t^{\theta,i} \left(q_{i,j} + \theta_{:,i}^\top R\theta_{:,j}\right) X_t^{\theta,j} \\ &= \sum_{i,j} \left((R\theta)_{k,j} \mathbb{1}_{\{\ell=i\}} + (R\theta)_{k,i} \mathbb{1}_{\{\ell=j\}}\right)_{n\times n} X_t^{\theta,i} X_t^{\theta,j} + 2\sum_{i,j} \tilde{X}_t^{\theta,i} \left(q_{i,j} + \theta_{:,i}^\top R\theta_{:,j}\right) X_t^{\theta,j}, \end{aligned} \tag{4.4.44}$$

where $\left((R\theta)_{k,j} \mathbb{1}_{\{\ell=i\}} + (R\theta)_{k,i} \mathbb{1}_{\{\ell=j\}}\right)_{n\times n}$ denotes for a $n\times n$ matrix whose k-th row and ℓ-th column is $(R\theta)_{k,j} \mathbb{1}_{\{\ell=i\}} + (R\theta)_{k,i} \mathbb{1}_{\{\ell=j\}}$.

We now present the method for computationally efficient evaluation of the gradient process $\tilde{X}_t^\theta$.

For notational simplicity, we only discuss the case below without using a mini-batch. The method can be easily extended to the mini-batch case though. Let $\odot$ indicate element-wise multiplication *with broadcasting* [100]. The RHS of (4.4.43) can be evaluated using the following operations:

- To vectorize the term $\sum_{j=1}^{n} \theta_{i,j} \tilde{X}_t^{\theta,j}$ for $i \in \{1, 2, \cdots, n\}$, we need to perform an inner-product of the *second dimension* of the $n \times n \times 1 \times 1$ matrix θ with the $1 \times n \times n \times n$ matrix $\tilde{X}_t^{\theta}$.
- Note that the final output w is a tensor with dimensions $n \times n \times n$.
- To vectorize the term $D_i(X_t^{\theta})$, consider the $n \times n \times n$ tensor E where $E_{i,j,:} = \delta_{ij}$. Then $p = E \odot X_t^{\theta}$.
- Add w and p.

The objective function can be evaluated using a similar method:

- First vectorize the $(R\theta)_{k,j} \mathbb{1}_{\{\ell=i\}} + (R\theta)_{k,i} \mathbb{1}_{\{\ell=j\}}$ to be an $n \times n \times n \times n$ matrix, which can be achieved by broadcasting, and denote the output as D. Similarly, the matrix multiplication of $X_t^{\theta,i} X_t^{\theta,j}$ produces a $n \times n \times 1 \times 1$ which we denote X.
- Perform an inner-product of the *first and second dimension* of the $n \times n \times n \times n$ matrix D with the $n \times n \times 1 \times 1$ matrix X. Call this output z, which will be a tensor with dimensions $n \times n$.
- Perform the inner-product of the first dimension of the $n \times n \times n$ matrix $\tilde{X}_t^{\theta}$ and $n \times 1 \times 1$ matrix F, where $F_{i,:,:} = \sum_j \left(q_{i,j} + \theta_{:,i}^{\top} R \theta_{:,j} \right) X_t^{\theta,j}$. The output q is a tensor with dimensions $n \times n$.
- Add z and q.

Table 4.1 presents the numerical results for the online optimization algorithm for learning the optimal control of the LQR problem. The online optimization algorithm performs well even in high dimensions. Figure 4.26 and Figure 4.27 display the maximum and average errors for dimension 5 and 20 during training.

Table 4.1: Training Result for Linear Control

Dimension	**Ave Error**	**Max Error**	**Cost Error**
1	0.1%	0.1%	0.01%
5	0.2%	0.5%	0.05%
20	0.5%	1%	0.05%

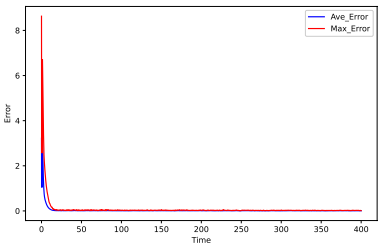

Figure 4.26: Training result for dim = 5

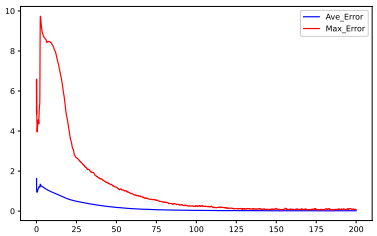

Figure 4.27: Training result for dim = 20

The error metrics in Table 4.1 are defined as:

$$\begin{aligned}
\text{Ave Error} &= \frac{\sum_{i,j=1}^{n} \left|\theta_{t,i,j} - \theta^*_{i,j}\right|}{\sum_{i,j=1}^{n} \left|\theta^*_{i,j}\right|} \\
\text{Max Error} &= \frac{\max_{i,j\in\{1,2,\cdots,n\}} \left|\theta_{t,i,j} - \theta^*_{i,j}\right|}{\frac{1}{n^2}\sum_{i,j=1}^{n} \left|\theta^*_{i,j}\right|} \\
\text{Cost Error} &= \frac{|J(\theta_T) - J(\theta^*)|}{|J(\theta^*)|},
\end{aligned} \tag{4.4.45}$$

where θ^* is the optimal control and θ_t is the parameter during training. $J(\theta_T)$ and $J(\theta^*)$ denote the objective function $J(\theta)$ in (4.4.40) with the parameters θ_T and θ^*, respectively.

4.4.11.3 One-dimensional Neural Network Control

We will now train a single-layer neural network control using the online optimization algorithm. The state process is:

$$dX_t^\theta = \left(-X_t^\theta + f_\theta\left(X_t^\theta\right)\right) dt + dW_t, \tag{4.4.46}$$

where the control $f_\theta(\cdot)$ is a single-layer neural network

$$f_\theta(x) = \sum_{i=1}^{m} c^i \sigma\left(w^i x + b^i\right), \tag{4.4.47}$$

with parameters $\theta = (c^i, w^i, b_i)_{i=1}^m$. The objective function is

$$J(\theta) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(X_t^\theta\right)^2 + \left(f_\theta(X_t^\theta)\right)^2 dt. \tag{4.4.48}$$

Define the gradient of X_t with respect to the parameters as:

$$\tilde{X}_t^w = \nabla_w X_t^\theta \in \mathbb{R}^m, \quad \tilde{X}_t^b = \nabla_b X_t^\theta \in \mathbb{R}^m, \quad \tilde{X}_t^c = \nabla_c X_t^\theta \in \mathbb{R}^m. \tag{4.4.49}$$

The coupled system (4.1.4) becomes

$$\begin{aligned}
dw_t &= -\alpha_t \left(2X_t\tilde{X}_t^w + 2f_{\theta_t}(X_t)\left(c_t \odot \sigma'(w_tX_t + b_t)X_t + f'_{\theta_t}(X_t)\tilde{X}_t^w\right)\right) dt,\\
db_t &= -\alpha_t \left(2X_t\tilde{X}_t^b + 2f_{\theta_t}(X_t)\left(c_t \odot \sigma'(W_tX_t + B_t) + f'_{\theta_t}(X_t)\tilde{X}_t^b\right)\right) dt,\\
dc_t &= -\alpha_t \left(2X_t\tilde{X}_t^c + 2f_{\theta_t}(X_t)\left(\sigma(w_tX_t + b_t) + f'_{\theta_t}(X_t)\tilde{X}_t^c\right)\right) dt,\\
dX_t &= (-X_t + f_{\theta_t}(X_t))dt + dW_t,\\
d\tilde{X}_t^w &= (-\tilde{X}_t^w + c_t \odot \sigma'(w_tX_t^i + b_t)X_t + f'_{\theta_t}(X_t)\tilde{X}_t^w)dt,\\
d\tilde{X}_t^b &= (-\tilde{X}_t^b + c_t \odot \sigma'(w_tX_t^i + b_t) + f'_{\theta_t}(X_t)\tilde{X}_t^b)dt,\\
d\tilde{X}_t^c &= (-\tilde{X}_t^c + \sigma(w_tX_t + b_t) + f'_{\theta_t}(X_t)\tilde{X}_t^c)dt,
\end{aligned} \tag{4.4.50}$$

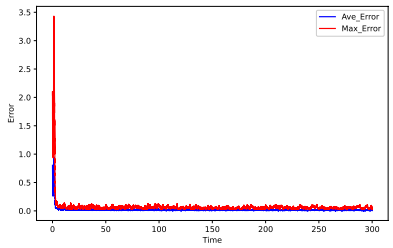

Figure 4.28: Training result for dim = 1

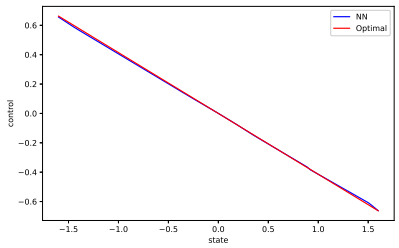

Figure 4.29: Neural Network output after training

The training result for 1 dimensional LQR with network network control is presented in Figure 4.28, Figure 4.29, and Table 4.2. The error metrics are defined as:[3]

$$\begin{aligned}
\text{Ave Error} &= \frac{\sum_{i=1}^{n} \left\| f_{\theta_t}(X^i) - \theta^* X^i \right\|}{\sum_{i=1}^{n} \|\theta^* X^i\|}\\
\text{Max Error} &= \frac{\max_{i\in\{1,2,\cdots,n\}} \left\| f_{\theta_t}(X^i) - \theta^* X^i \right\|}{\sum_{i=1}^{n} \|\theta^* X^i\|}\\
\text{Cost Error} &= \frac{|J(\theta_T) - J(\theta^*)|}{|C^*|},
\end{aligned} \tag{4.4.51}$$

where θ^* is the optimal control and θ_t is the trained parameter. $J(\theta_T)$ and $J(\theta^*)$ denote the objective function $J(\theta)$ in (4.4.48) with the parameters θ_T and θ^*, respectively. The points $\{X^i\}_{i=1}^n$ are uniformly sampled from $[-L, L]$ with L chosen such that $[-L, L]$ contains the optimally controlled process 99% of the time.

[3]Here the norm $\|\cdot\|$ denotes the L^1 norm, i.e. for a vector $Y = (y_1, y_2, \cdots, y_d) \in \mathbb{R}^d$, $\|Y\| = \sum_{i=1}^d |y_i|$.

4.4.11.4 Multi-dimensional Neural Network Control

We now optimize a single-layer neural network control for a high-dimensional state process:

$$\begin{aligned} dX_t^\theta &= \left(-X_t^\theta + f_\theta\left(X_t^\theta\right)\right) dt + dW_t, \\ J(\theta) &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(X_t^\theta\right)^\top Q X_t^\theta + \left(f_\theta\left(X_t^\theta\right)\right)^\top R f_\theta\left(X_t^\theta\right) dt, \end{aligned} \tag{4.4.52}$$

where $X_t^\theta \in \mathbb{R}^n$ and the single-layer neural network with m hidden units is:

$$f_\theta(x) = c\sigma\left(wx + b\right), \tag{4.4.53}$$

where $w \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}^{n\times m}$. As in (4.4.49), define

$$\begin{aligned} \tilde{X}_t^w &= \nabla_w X_t^\theta \in \mathbb{R}^{n\times m\times n}, \quad \tilde{X}_t^{w,i} = \nabla_w X_t^{\theta,i} \in \mathbb{R}^{m\times n}, \\ \tilde{X}_t^b &= \nabla_b X_t^\theta \in \mathbb{R}^{n\times m}, \quad \tilde{X}_t^{b,i} = \nabla_b X_t^{\theta,i} \in \mathbb{R}^{m}, \\ \tilde{X}_t^c &= \nabla_c X_t^\theta \in \mathbb{R}^{n\times n\times m}, \quad \tilde{X}_t^{c,i} = \nabla_c X_t^{\theta,i} \in \mathbb{R}^{n\times m}, \end{aligned} \tag{4.4.54}$$

for $i = 1, 2, \cdots, n$.

The online algorithm (4.1.4) becomes:

$$\begin{aligned} dw_t &= -\alpha_t \left[\nabla_w \left(f_{\theta_t}(X_t)^\top R f_{\theta_t}(X_t)\right) + \sum_{i=1}^n \frac{\partial}{\partial x_i}\left((X_t)^\top Q X_t + f_{\theta_t}(X_t)^\top R f_{\theta_t}(X_t)\right) \tilde{X}_t^{w,i}\right] dt \\ db_t &= -\alpha_t \left[\nabla_b \left(f_{\theta_t}(X_t)^\top R f_{\theta_t}(X_t)\right) + \sum_{i=1}^n \frac{\partial}{\partial x_i}\left((X_t)^\top Q X_t + f_{\theta_t}(X_t)^\top R f_{\theta_t}(X_t)\right) \tilde{X}_t^{b,i}\right] dt \\ dc_t &= -\alpha_t \left[\nabla_c \left(f_{\theta_t}(X_t)^\top R f_{\theta_t}(X_t)\right) + \sum_{i=1}^n \frac{\partial}{\partial x_i}\left((X_t)^\top Q X_t + f_{\theta_t}(X_t)^\top R f_{\theta_t}(X_t)\right) \tilde{X}_t^{c,i}\right] dt \\ dX_t &= (-X_t + f_{\theta_t}(X_t))dt + dW_t, \\ d\tilde{X}_t^{w,i} &= \left(-\tilde{X}_t^{w,i} + \sum_k c_{t,i,k}\sigma'\left(w_t X_t + b_t\right)_k \left(\sum_\ell w_{t,k,\ell}\tilde{X}_t^{w,\ell}\right) + \left(c_{t,i,:}\right)^\top \odot \sigma'\left(w_t X_t + b_t\right)\left(X_t\right)^\top\right) dt \\ d\tilde{X}_t^{b,i} &= \left(-\tilde{X}_t^{b,i} + \sum_k c_{t,i,k}\sigma'\left(w_t X_t + b_t\right)_k \left(\sum_\ell w_{t,k,\ell}\tilde{X}_t^{b,\ell}\right) + \left(c_{t,i,:}\right)^\top \odot \sigma'\left(w_t X_t + b_t\right)\right) dt \\ d\tilde{X}_t^{c,i} &= \left(-\tilde{X}_t^{c,i} + \sum_k c_{t,i,k}\sigma'\left(w_t X_t + b_t\right)_k \left(\sum_\ell w_{t,k,\ell}\tilde{X}_t^{c,\ell}\right) + D_i\left(\sigma\left(w_t X_t + b_t\right)\right)\right) dt \end{aligned} \tag{4.4.55}$$

for $i = 1, 2, \cdots, N$. In (4.4.55), $C_{t,i,:} \in \mathbb{R}^n$ denotes the i-th row of the matrix C_t and $D_i(X_t)$ is an $n \times n$ matrix whose elements are all zeros except for the i-th row, which has the vector value $\sigma\left(w_t X_t + b_t\right)$.

The numerical results for training the neural network SDE control with the online optimization algorithm are presented in Figure 4.30, Figure 4.31, and Table 4.2. In general, the trained neural network control performs well, even in high dimensions.

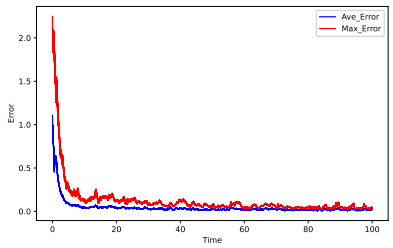

Figure 4.30: Training result for dim = 5

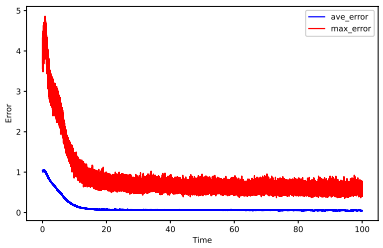

Figure 4.31: Training result for dim = 20

Table 4.2: Training Result for NN Control

Dimension	Ave Error	Max Error	Cost Error
1	0.1%	0.1%	0.01%
5	0.6%	1%	0.02%
20	1%	10%	0.1%

4.4.12 Applications to Multi-Agent and Mean-Field System Control

Finally, the online optimization algorithm can be used to solve multi-agent stochastic control problems – e.g., mean-field control and mean-field games, which are important topics in mathematical finance [8, 29, 30, 32, 33, 34] – in the ergodic setting. As an example, we numerically implement the online optimization model for a simplified version of the multi-agent systemic risk model ([32]) in Section 4.4.12. There are N agents where each agent is modeled by an SDE. As $N \to \infty$, the system converges to a mean-field game limit. In the numerical example, we use the online optimization algorithm to solve the high-dimensional stochastic optimal control problem corresponding to a large number of N SDEs ($N = 5,000$).

We consider the following multi-agent control problem, which is a simplified version of the systemic risk model in [32]:

$$dX_t^{\theta,i} = \left[a\left(\frac{1}{N}\sum_{j=1}^{N} X_t^{\theta,j} - X_t^{\theta,i} \right) + f_\theta(X_t^{\theta,i}) \right] dt + \sigma dW_t^i \tag{4.4.56}$$

for $i = 1, 2, \cdots N$ with the objective function

$$J^N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(X_t^{\theta,i}\right)^2 + f^2\left(X_t^{\theta,i}\right) dt. \tag{4.4.57}$$

This mean-field system has the following mean-field limit:

$$\begin{aligned} dX_t^\theta &= a\left[\left(EX_t^\theta - X_t^\theta\right) + f_\theta\left(X_t^\theta\right)\right] dt + \sigma dW_t \\ J(\theta) &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(X_t^\theta\right)^2 + f_\theta^2(X_t^\theta) dt. \end{aligned} \tag{4.4.58}$$

We describe how the online optimization algorithm can train both linear and neural network

controls for this mean-field system. The algorithm (4.1.4) to train the linear model becomes:

$$\begin{aligned}
d\theta_t &= -\alpha_t \left[\frac{1}{N} \sum_{i=1}^{N} \left(2\theta_t \left(X_t^i \right)^2 + 2 \left(1 + \theta_t^2 \right) X_t^i \tilde{X}_t^i \right) \right] dt \\
dX_t^i &= \left[a \left(\frac{1}{N} \sum_j X_t^j - X_t^i \right) + \theta_t X_t^i \right] dt + dW_t^i \\
d\tilde{X}_t^i &= \left[a \left(\frac{1}{N} \sum_j \tilde{X}_t^j - \tilde{X}_t^i \right) + X_t^i + \theta_t \tilde{X}_t^i \right] dt.
\end{aligned} \tag{4.4.59}$$

The training result for the linear control is displayed in Figure 4.32.

We next train a neural network for the control function $f_\theta(x) = c\sigma(wx + b)$ where $\theta = (c, w, b)$. The online optimization algorithm becomes:

$$\begin{aligned}
dw_t &= -\alpha_t \left[\frac{1}{N} \sum_{i=1}^{N} \left(2X_t^i \tilde{X}_t^{w,i} + 2f_{\theta_t}(X_t^i) \left(C_t \odot \sigma'(W_t X_t^i + B_t) X_t^i + f'_{\theta_t}(X_t^i) \tilde{X}_t^{w,i} \right) \right) \right] dt \\
db_t &= -\alpha_t \left[\frac{1}{N} \sum_{i=1}^{N} \left(2X_t^i \tilde{X}_t^{b,i} + 2f_{\theta_t}(X_t^i) \left(C_t \odot \sigma'(W_t X_t^i + B_t) + f'_{\theta_t}(X_t^i) \tilde{X}_t^{b,i} \right) \right) \right] dt \\
dc_t &= -\alpha_t \left[\frac{1}{N} \sum_{i=1}^{N} \left(2X_t^i \tilde{X}_t^{c,i} + 2f_{\theta_t}(X_t^i) \left(\sigma(W_t X_t^i + B_t) + f'_{\theta_t}(X_t^i) \tilde{X}_t^{c,i} \right) \right) \right] dt \\
dX_t^i &= \left[a \left(\frac{1}{N} \sum_j X_t^j - X_t^i \right) + f_{\theta_t}(X_t^i) \right] dt + dW_t^i \\
d\tilde{X}_t^{w,i} &= \left[a \left(\frac{1}{N} \sum_j \tilde{X}_t^{w,j} - \tilde{X}_t^{w,i} \right) + C_t \odot \sigma'(W_t X_t^i + B_t) X_t^i + f'_{\theta_t}(X_t^i) \tilde{X}_t^{w,i} \right] dt \\
d\tilde{X}_t^{b,i} &= \left[a \left(\frac{1}{N} \sum_j \tilde{X}_t^{b,j} - \tilde{X}_t^{b,i} \right) + C_t \odot \sigma'(W_t X_t^i + B_t) + f'_{\theta_t}(X_t^i) \tilde{X}_t^{b,i} \right] dt \\
d\tilde{X}_t^{c,i} &= \left[a \left(\frac{1}{N} \sum_j \tilde{X}_t^{c,j} - \tilde{X}_t^{c,i} \right) + \sigma(W_t X_t^i + B_t) + f'_{\theta_t}(X_t^i) \tilde{X}_t^{c,i} \right] dt.
\end{aligned} \tag{4.4.60}$$

The trained neural network control is also displayed in Figure 4.32; the controls learned by the linear model and neural network are similar.

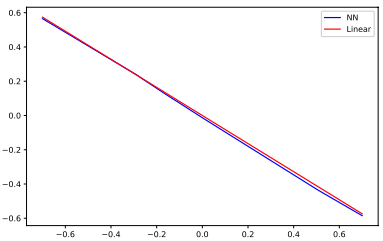

Figure 4.32: Training result for (4.4.59) and (4.4.60).

4.4.13 Models of Order Book Dynamics

Order books involve large numbers of high-frequency events ($\sim 10^5 - 10^6$ events per day per stock) and high-dimensional dynamics (many price levels, each with limit order submissions and cancellations, as well as market orders, hidden orders, and transactions). Due to the size of the datasets and the high-dimensionality, calibrating simulation models of order book dynamics to data is computationally challenging. Recent examples of such model frameworks for the simulation of the order books include [10, 87, 99, 109, 129], where they develop stochastic point process models to model the event-by-event dynamics in order books.

For more complex stochastic models, it is computationally intractable for many traditional calibration methods to optimize over the entire order flow history (even for a few days of events) to estimate the model parameters from the data. The online forward propagation optimization algorithm proposed in this chapter provides a tractable computational method to optimize over the entire order flow history. In particular, the online forward propagation optimization algorithm asymptotically minimizes the objective function over the stationary distribution of the entire order flow process (instead of optimizing over only small subsets of the data, which can lead to a sub-optimal model parameter calibration). In principle, our online optimization algorithm could be used to calibrate a general class of point process models to event-by-event order book data. Such a large-scale data project is outside of the scope of this chapter, which is focused on developing a convergence theory. However, in order to demonstrate the applicability of our method to point process models, we present two simple numerical examples below. Synthetic data is simulated from a standard Hawkes process with stochastic intensity

$$d\lambda_t^* = -\alpha^*(\mu^* - \lambda_t^*)dt + \kappa^* dN_t^*, \tag{4.4.61}$$

where N_t^* is the number of events that have occurred by time t. Events arrive with stochastic intensity λ_t, i.e. $\lim_{\Delta \to 0} \frac{\mathbb{P}[N_{t+\Delta}^* - N_t^* = 1 | \mathcal{F}_t]}{\Delta} = \lambda_t^*$. For example, N_t^* could be the number of limit orders submitted to the order book by time t. Multi-dimensional point process models can model the dynamics of the entire order book (e.g., limit order submissions, cancellations, market orders, hidden orders, and transactions) [10, 109].

Model parameters for point process models can be calibrated from event data. The data consists of only the observed process N_t^*; the stochastic intensity λ_t^* is unobserved. Note that (4.4.61) is an ergodic process with a stationary distribution. Hawkes process models have been widely used in the financial literature for modeling order book events (for example, see [109]). Using the event data N_t^* simulated from (4.4.61), we will calibrate point process models using the online forward propagation optimization algorithm.

First, we consider calibrating a standard Hawkes model using the online optimization algorithm. The model is

$$d\lambda_t^\theta = -\alpha(\mu - \lambda_t^\theta)dt + \kappa dN_t^\theta, \tag{4.4.62}$$

where $\theta = (\alpha, \mu, \kappa)$ are the parameters that must be trained and the time-averaged log-likelihood objective function is

$$L_T(\theta) = -\frac{1}{T}\int_0^T \hat{\lambda}_t^\theta dt + \frac{1}{T}\int_0^T \log(\hat{\lambda}_t^\theta) dN_t^*, \tag{4.4.63}$$

where $\hat{\lambda}_t^\theta$ is the intensity process (4.4.62) conditioned on the event observations $(N_{t'}^*)_{t' \leq t}$, i.e. $d\hat{\lambda}_t^\theta = -\alpha(\mu - \hat{\lambda}_t^\theta)dt + \kappa dN_t^*$. Using our online optimization algorithm, we train the parameters θ_t to maximize the objective function $L_T(\theta)$. Figure 4.33 displays the results from the training and demonstrate the numerical convergence of the method. The training converges to a global minimizer; the objective function evaluated at the trained parameters matches the objective function evaluated at the true parameters $\theta^* = (\alpha^*, \mu^*, \kappa^*) = (\frac{1}{10}, 1, \frac{1}{10})$.

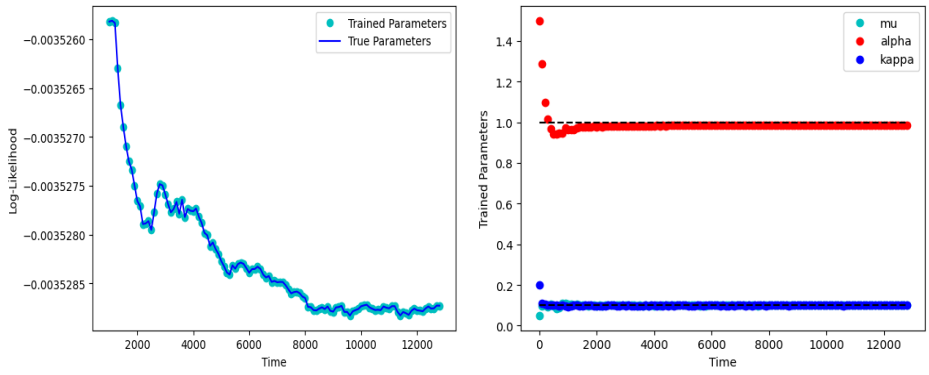

Figure 4.33: Objective function (left) and trained parameters (right).

We now consider a slightly more complex model where the intensity dynamics are given by a neural network. Neural network (or "neural SDEs") have been widely studied in the financial mathematics literature [7, 44, 45, 46, 61, 111]. Neural network Hawkes processes (or "neural Hawkes processes' ') have also been recently studied and implemented in a number of papers for modeling order book data [87, 99, 129]. We consider the following neural SDE:

$$d\bar{\lambda}_t^\theta = f(\bar{\lambda}_t^\theta; \theta)dt + \kappa dN_t^\theta, \tag{4.4.64}$$

where, for this simplified numerical experiment, we set $\kappa = \kappa^*$ and $\lambda_t^\theta = |\bar{\lambda}_t^\theta| + \epsilon$ where $\epsilon > 0$. $f(\lambda; \theta)$ is a single-layer neural network with 25 hidden units. The neural network parameters θ are trained

with the online forward propagation optimization algorithm:

$$\begin{aligned} d\tilde{\lambda}_t &= \left(\frac{\partial f}{\partial \lambda}(\bar{\lambda}_t;\theta_t)\tilde{\lambda}_t + \frac{\partial f}{\partial \theta}(\bar{\lambda}_t;\theta_t)\right)dt, \\ d\bar{\lambda}_t &= f(\bar{\lambda}_t;\theta_t)dt + \kappa dN_t^*, \\ d\theta_t &= \alpha_t\left(-\frac{\partial \lambda_t}{\partial \bar{\lambda}_t}\tilde{\lambda}_t dt + (\lambda_t)^{-1}\frac{\partial \lambda_t}{\partial \bar{\lambda}_t}\tilde{\lambda}_t dN_t^*\right), \end{aligned} \tag{4.4.65}$$

where $\lambda_t = |\bar{\lambda}_t| + \epsilon$ and α_t is the learning rate. The data N_t^* which the model (4.4.64) is trained on is generated using (4.4.61) with the "true parameters" $\theta^* = (\frac{1}{10}, 1, \frac{1}{10})$. The training and out-of-sample test results are displayed in Figure 4.34. The plots display the value of the objective function (4.4.63) evaluated using the "true" process (4.4.61) with the true parameters θ^* (which is the global minimum) as compared to the value of the objective function (4.4.63) for the trained model (4.4.64). The neural network point process model (4.4.64), trained with the online forward propagation algorithm, is able to achieve a nearly identical value for the objective function as the exact global minimizer (with $\sim 10^{-4}$ relative error), indicating that the trained model converges to a global minimizer.

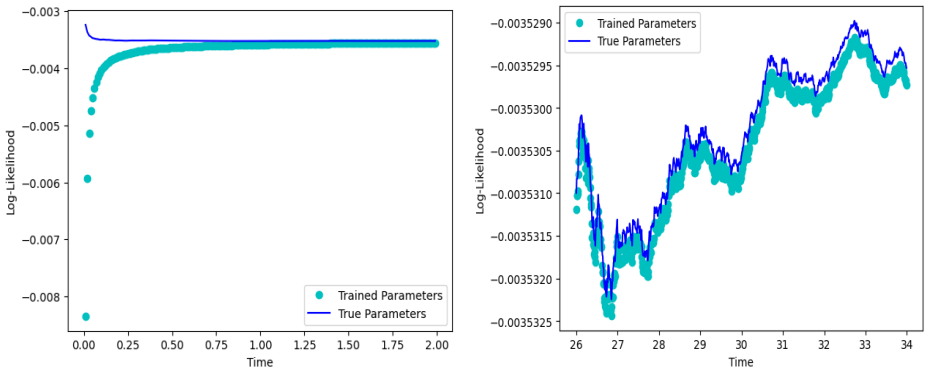

Figure 4.34: Objective function during the initial time period of training (left) and out-of-sample objective function (right).

The above numerical examples use the well-known time-scaling method for simulation of the point processes with an Euler discretization with time step size 10^{-2}, which introduces discretization error. Therefore, in practice, the above numerical examples can be viewed as optimizations over discrete-time stochastic models.

We conclude by highlighting that – although outside of the scope of this chapter – a more general multi-dimensional model for the entire order book (see [109]) could also be calibrated to real order book data using the online forward propagation algorithm. General classes of multi-dimensional neural SDE models can be optimized using our method. For example, "recurrent neural SDEs", where the dynamics (4.4.62) depend upon the evolution of a "hidden" neural SDE, can also be

calibrated using the online forward propagation method, such as:

$$d\tilde{\lambda}_t^\theta = f(\tilde{\lambda}_t^\theta, S_t^\theta; \theta)dt + \kappa(\tilde{\lambda}_t^\theta, S_t^\theta; \theta)dN_t^\theta,$$
$$dS_t^\theta = g(\tilde{\lambda}_t^\theta, S_t^\theta; \theta)dt + h(\tilde{\lambda}_t^\theta, S_t^\theta; \theta)dN_t^\theta, \tag{4.4.66}$$

where f, g, h, κ are neural networks with collective parameters θ and where $\tilde{\lambda}_t^\theta, N_t^\theta$, and S_t^θ can be multidimensional. Recurrent neural networks Hawkes models for order books have been investigated in [87, 129]. Recurrent neural network Hawkes processes have recently received significant interest in the broader machine learning community [102]. General classes of continuous-time recurrent network SDEs have also been proposed in [50]. A more general class of continuous-time recurrent network point processes has also been developed in [40]; (4.4.66) is an example from the general framework in [40]. The unique capability provided by the algorithm is to asymptotically optimize such models over the *entire history* of the order flow dataset, while standard methods can typically only optimize over much smaller sub-sequences.

Chapter 5

Online SDE Optimization: Nonlinear Case

5.1 Introduction

Optimizing over the stationary distribution of a stochastic process is a challenging mathematical and computational problem. Consider a parameterized ergodic process $X_t^{\theta,x} \in \mathbb{R}^d$ which satisfies the stochastic differential equation (SDE):

$$\begin{aligned} dX_t^{\theta,x} &= \mu(X_t^{\theta,x}, \theta)dt + \sigma(X_t^{\theta,x}, \theta)dW_t, \\ X_0^{\theta,x} &= x, \end{aligned} \tag{5.1.1}$$

where $\theta \in \mathbb{R}^\ell$ and W_t is a d-dimensional standard Brownian motion. Our goal is to select the parameters θ which minimize the objective function

$$J(\theta) = \sum_{n=1}^{N} \left(\mathbb{E}_{Y \sim \pi_\theta} \left[f_n(Y) \right] - \beta_n \right)^2, \tag{5.1.2}$$

where f are known functions and β are the target quantities [1].

In chapter 4, a new online algorithm was developed to optimize over the stationary distribution of SDEs such as (5.1.1). The online algorithm simultaneously simulates (5.1.1) while continuously updating the parameter θ_t using a stochastic estimate for the gradient $\nabla_\theta J(\theta_t)$. The stochastic estimate for the gradient $\nabla_\theta J(\theta_t)$ is based upon a *forward propagation* SDE for the gradient of $X_t^{\theta,x}$ with respect to θ. The online forward propagation algorithm for optimizing (5.1.2) is:[2]

$$\begin{aligned} \frac{d\theta_t}{dt} &= -2\alpha_t \left(f(\bar{X}_t) - \beta \right) \left(\nabla f(X_t) \tilde{X}_t \right)^\top, \\ d\tilde{X}_t &= \left(\mu_x(X_t, \theta_t)\tilde{X}_t + \mu_\theta(X_t, \theta_t) \right) dt + \left(\sigma_x(X_t, \theta_t)\tilde{X}_t + \sigma_\theta(X_t, \theta_t) \right) dW_t, \\ dX_t &= \mu(X_t, \theta_t)dt + \sigma(X_t, \theta_t)dW_t, \\ d\bar{X}_t &= \mu(\bar{X}_t, \theta_t)dt + \sigma(\bar{X}_t, \theta_t)d\bar{W}_t, \end{aligned} \tag{5.1.3}$$

where W_t and $\bar{W}_t$ are independent Brownian motions and α_t is the learning rate.

In chapter 4, we prove the convergence of the online forward propagation algorithm (5.1.3) for a class of linear SDEs with constant volatility coefficients. Numerical experiments demonstrate that

[1] For notational convenience (and without loss of generality), we will set $N = 1$ and $\beta_1 = \beta$ in (5.1.2) in the later discussion.

[2] In this chapter's notation, the Jacobian matrix of a vector valued function $f : x \in \mathbb{R}^n \to \mathbb{R}^m$ is an $m \times n$ matrix.

the forward propagation algorithm also converges for nonlinear SDEs. *In this chapter, we rigorously prove the convergence of the forward propagation algorithm for a class of nonlinear SDEs with non-constant volatility coefficients.* The mathematical approach uses a Poisson partial differential equation (PDE), such as in [114, 113], to rewrite the fluctuation terms in terms of the solution of the PDEs. The fluctuation terms can be appropriately bounded by proving bounds on the solution to the PDEs. We leverage recent methods from [123] to characterize the convergence rate of the transition semigroup for (5.1.1) and its derivatives with respect to the initial condition x and the parameter θ, which combined with the moment stability for (5.1.1), allow us to prove there exists appropriately bounded solutions to the PDE. Once the fluctuation terms have been bounded, using the moment stability for the coupled system (5.1.3), we can prove convergence of the forward propagation algorithm using the cycle of stopping times argument [14, 130, 134].

Organization of Chapter This chapter is organized into four main sections. In Section 5.2, we present the assumptions and the main theorem. Section 5.3 rigorously proves the convergence of the online forward propagation algorithm for nonlinear dissipative SDEs. Application of our algorithm in SPDEs parameter estimation is give in Section 5.4.

5.2 Main Results

We will study the convergence of the algorithm (5.1.3) for a class of nonlinear SDEs satisfying the following conditions, which are similar to the assumptions in [123].

Assumption 5.2.1. There exist constants $C, \beta > 0$ such that the following conditions hold for all $x_1, x_2 \in \mathbb{R}^d,\ \theta_1, \theta_2 \in \mathbb{R}^\ell$:

- Lipschitz continuity:

$$|\mu(x_1, \theta_1) - \mu(x_2, \theta_2)| + |\sigma(x_1, \theta_1) - \sigma(x_2, \theta_2)| \leq C\left(|x_1 - x_2| + |\theta_1 - \theta_2|\right). \tag{5.2.1}$$

- Dissipativity:

$$\langle \mu(x_1, \theta) - \mu(x_2, \theta),\ x_1 - x_2 \rangle + \frac{7}{2}\left|\sigma(x_1, \theta) - \sigma(x_2, \theta)\right|^2 \leq -\beta |x_1 - x_2|^2, \tag{5.2.2}$$

  where $\langle a,\ b \rangle := b^\top a$.

- A uniform bound for the SDE coefficients in the variable θ:

$$\sup_{\theta \in \mathbb{R}^\ell} \max\{|\mu(0, \theta)|, |\sigma(0, \theta)|\} \leq C. \tag{5.2.3}$$

Assumption 5.2.2. $\mu(x,\theta)$, $\mu(x,\theta)$, $\sigma(x,\theta)$, and $\sigma(x,\theta)$ are twice differentiable. For any $x_1, x_2 \in \mathbb{R}^d$,

$$\sup_{\theta\in\mathbb{R}^\ell} |\nabla_x \mu(x_1,\theta) - \nabla_x \mu(x_2,\theta)| \le C|x_1 - x_2|, \tag{5.2.4}$$

$$\sup_{\theta\in\mathbb{R}^\ell} |\nabla_\theta \mu(x_1,\theta) - \nabla_\theta \mu(x_2,\theta)| \le C|x_1 - x_2|, \tag{5.2.5}$$

$$\sup_{\theta\in\mathbb{R}^\ell} |\nabla_x \sigma(x_1,\theta) - \nabla_x \sigma(x_2,\theta)| \le C|x_1 - x_2|, \tag{5.2.6}$$

$$\sup_{\theta\in\mathbb{R}^\ell} |\nabla_\theta \sigma(x_1,\theta) - \nabla_\theta \sigma(x_2,\theta)| \le C|x_1 - x_2|. \tag{5.2.7}$$

Assumption 5.2.3. For any $x_1, x_2 \in \mathbb{R}^d$,

$$\sup_{\theta\in\mathbb{R}^\ell} \left|\nabla_x^2 \mu(x_1,\theta) - \nabla_x^2 \mu(x_2,\theta)\right| \le C|x_1 - x_2|, \tag{5.2.8}$$

$$\sup_{\theta\in\mathbb{R}^\ell} \left|\partial_\theta^2 \mu(x_1,\theta) - \nabla_\theta^2 \mu(x_2,\theta)\right| \le C|x_1 - x_2|, \tag{5.2.9}$$

$$\sup_{\theta\in\mathbb{R}^\ell} |\nabla_x \nabla_\theta \mu(x_1,\theta) - \nabla_x \nabla_\theta \mu(x_2,\theta)| \le C|x_1 - x_2|, \tag{5.2.10}$$

$$\sup_{\theta\in\mathbb{R}^\ell} \left|\nabla_x^2 \nabla_\theta \mu(x_1,\theta) - \nabla_x^2 \nabla_\theta \mu(x_2,\theta)\right| \le C|x_1 - x_2|. \tag{5.2.11}$$

We assume that the Lipschitz properties (5.2.8) - (5.2.11) also hold for σ. In addition,

$$\begin{aligned} \sup_{(x,\theta)\in\mathbb{R}^d\times\mathbb{R}^\ell} \max\left\{\left|\nabla_x^2 \mu(x,\theta)\right|, \left|\nabla_\theta^2 \mu(x,\theta)\right|, |\nabla_x \nabla_\theta \mu(x,\theta)|, \left|\nabla_x^2 \nabla_\theta \mu(x,\theta)\right|\right\} \le C, \\ \sup_{(x,\theta)\in\mathbb{R}^d\times\mathbb{R}^\ell} \max\left\{\left|\nabla_x^2 \sigma(x,\theta)\right|, \left|\nabla_\theta^2 \sigma(x,\theta)\right|, |\nabla_x \nabla_\theta \sigma(x,\theta)|, \left|\nabla_x^2 \nabla_\theta \sigma(x,\theta)\right|\right\} \le C, \end{aligned} \tag{5.2.12}$$

Assumption 5.2.4. The function f in the objective function is continuously differentiable and has uniformly bounded derivatives, i.e. there exists a constant C such that

$$\left|\nabla^i f(x)\right| \le C, \quad \forall x \in \mathbb{R}^d, \ i = 1,2,3. \tag{5.2.13}$$

Assumption 5.2.5. The learning rate α_t satisfies $\int_0^\infty \alpha_t dt = \infty$, $\int_0^\infty \alpha_t^2 dt < \infty$, $\int_0^\infty |\alpha_s'|\, ds < \infty$, and there is a $p > 0$ such that $\lim_{t\to\infty} \alpha_t^2 t^{\frac{1}{2}+2p} = 0$.

Our assumptions are standard in the mathematical literature which studies ergodic nonlinear SDEs [123]. Under these assumptions, we are able to prove the convergence of the online forward algorithm (5.1.3). We briefly comment on these assumptions before presenting our main theoretical result.

A sufficient condition for the dissipative SDE (5.1.1) to satisfy Assumptions A5.2.1-A5.2.3 is that the first-, second-, and third-order derivatives of μ, σ are uniformly bounded and for any $x, y \in \mathbb{R}^d, \theta \in \mathbb{R}^\ell$

$$y^\top \nabla_x \mu(x,\theta) y \le -C|y|^2, \quad |\sigma(x,\theta) - \sigma(y,\theta)| \le L|x-y|, \tag{5.2.14}$$

where $C, L > 0$ are constants and $\frac{3}{2}L^2 < C$. A classic example is the Langevin Equation, where the drift term is the gradient of some convex potential. That is $\mu(x,\theta) = -\nabla V(x,\theta)$ with $V(x,\theta)$ being convex with respect to x. See [115, 116] for a detailed discussion. Under Assumption (A5.2.1), there

will exist a unique invariant measure π_θ for (5.1.1) such that[3]

$$\int_{\mathbb{R}^d} |x| \pi_\theta(dx) \leq C < \infty. \tag{5.2.15}$$

Theorem 5.2.6. *Under Assumptions (A5.2.1) - (A5.2.5), the online algorithm (5.1.3) converges almost surely:*

$$\lim_{t\to\infty} |\nabla_\theta J(\theta_t)| \overset{a.s.}{=} 0. \tag{5.2.16}$$

5.3 Proof

The SDE system (5.1.3) has a unique strong solution.[4] In equation (**??**), we decomposed the evolution of θ_t into the direction of steepest descent $-\alpha_t \nabla_\theta J(\theta_t)$ and two fluctuation terms. Define the fluctuation terms as

$$\begin{aligned} Z_t^1 &= \left(\mathbb{E}_{\pi_{\theta_t}} f(Y) - \beta\right) \left(\nabla f(X_t)\tilde{X}_t - \nabla_\theta \mathbb{E}_{\pi_{\theta_t}} f(Y)\right)^\top, \\ Z_t^2 &= \left(f(\bar{X}_t) - \mathbb{E}_{\pi_{\theta_t}} f(Y))\right) \left(\nabla f(X_t)\tilde{X}_t\right)^\top. \end{aligned} \tag{5.3.1}$$

As in [130], we will study a cycle of stopping times to control the time periods where $|\nabla_\theta J(\theta_t)|$ is close to zero and away from zero. Let us select an arbitrary constant $\kappa > 0$ and also define $\mu = \mu(\kappa) > 0$ (to be chosen later). Then, set $\sigma_0 = 0$ and define the cycles of random times

$$0 = \sigma_0 \leq \tau_1 \leq \sigma_1 \leq \tau_2 \leq \sigma_2 \leq \dots,$$

where the stopping times are defined as

$$\begin{aligned} &\tau_n = \inf\left\{t > \sigma_{n-1} : |\nabla_\theta J\left(\theta_t\right)| \geq \kappa\right\} \\ &\sigma_n = \sup\left\{t > \tau_n : \frac{|\nabla_\theta J\left(\theta_{\tau_n}\right)|}{2} \leq |\nabla_\theta J\left(\theta_s\right)| \leq 2\,|\nabla_\theta J\left(\theta_{\tau_n}\right)| \text{ for all } s \in [\tau_n, t] \text{ and } \int_{\tau_n}^t \alpha_s ds \leq \mu\right\}. \end{aligned} \tag{5.3.2}$$

We define the random time intervals $J_n = [\sigma_{n-1}, \tau_n)$ and $I_n = [\tau_n, \sigma_n)$. We introduce the constant $\eta > 0$ which will be chosen to be sufficiently small later. In order to prove convergence, we will have to show that the fluctuation terms become small as $t \to \infty$. In particular, the following integral of the fluctuation term will be crucial to the convergence analysis:

$$\Delta^i_{\tau_n, \sigma_n+\eta} := \int_{\tau_n}^{\sigma_n+\eta} \alpha_s Z_s^i ds, \quad i = 1, 2. \tag{5.3.3}$$

We will begin our analysis by first presenting several lemmas regarding Lipschitz continuity, moment bounds, and ergodicity. The proofs are the same as in [123] and thus we omit them.

Lemma 5.3.1 (Lipschitz continuity)**.** *For any $t > 0$, $x_i \in \mathbb{R}^d$, and $\theta_i \in R^\ell$, we have*

$$\mathbb{E}\left|X_t^{\theta_1, x_1} - X_t^{\theta_2, x_2}\right|^2 \leq e^{-\beta t}|x_1 - x_2|^2 + C|\theta_1 - \theta_2|^2. \tag{5.3.4}$$

A proof can be found in Lemma 3.6 of [123].

[3] See Theorem 4.3.9 of [119].

[4] Existence and uniqueness can be proven using the standard method of a contraction map; see Theorem 1.2 of [31] for details.

Lemma 5.3.2 (Ergodicity). *For any $t \geq 0$, $x \in \mathbb{R}^d$, and $\theta \in \mathbb{R}^\ell$,*

$$\left|\mathbb{E} f(X_t^{\theta,x}) - \mathbb{E}_{\pi_\theta} f(Y)\right| \leq Ce^{-\frac{\beta t}{2}}(1+|x|). \tag{5.3.5}$$

A proof can be found in Proposition 3.7 of [123].

Lemma 5.3.3 (Moment Bound). *There exists a constant C such that*

$$\mathbb{E}\left|X_t^{\theta,x}\right|^2 \leq C(1+e^{-\beta t}|x|^2), \quad \forall x \in \mathbb{R}^d, t \geq 0. \tag{5.3.6}$$

Proof. Applying Itô's formula to $\left|X_t^{\theta,x}\right|^2$ yields

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left|X_t^{\theta,x}\right|^2 \\
=&\mathbb{E}\left[2\left\langle \mu(X_t^{\theta,x},\theta),\ X_t^{\theta,x}\right\rangle + \left|\sigma(X_t^{\theta,x},\theta)\right|^2\right] \\
\leq&\mathbb{E}\left[2\left\langle \mu(X_t^{\theta,x},\theta)-\mu(0,\theta),\ X_t^{\theta,x}\right\rangle + 2\left|\sigma(X_t^{\theta,x},\theta)-\sigma(0,\theta)\right|^2 + 2\left\langle \mu(0,\theta),\ X_t^{\theta,x}\right\rangle + 2\left|\sigma(0,\theta)\right|^2\right] \\
\overset{(b)}{\leq}& -2\beta\mathbb{E}\left|X_t^{\theta,x}\right|^2 + \left(\beta\mathbb{E}\left|X_t^{\theta,x}\right|^2 + \frac{1}{\beta}\left|\mu(0,\theta)\right|^2\right) + 2\left|\sigma(0,\theta)\right|^2 \\
\overset{(a)}{\leq}& -\beta\mathbb{E}\left|X_t^{\theta,x}\right|^2 + C,
\end{aligned} \tag{5.3.7}$$

where step (a) uses the dissipativity assumption (5.2.2) and Young's inequality and step (b) uses the bound (5.2.3). Therefore, using a comparison principle for ODEs,

$$\mathbb{E}\left|X_t^{\theta,x}\right|^2 \leq e^{-\beta t}|x|^2 + C. \tag{5.3.8}$$

□

Using similar calculations as in Proposition 4.1 of [123], several ergodicity results for X_t^θ can be proven.

Proposition 5.3.4. *Under Assumptions (A5.2.1) - (A5.2.4), we have the following ergodic bounds:*

(i) There exists a constant C such that for any $\theta \in \mathbb{R}^\ell, x \in \mathbb{R}^d$, and $t > 0$,

$$\left|\nabla_\theta^i \mathbb{E} f(X_t^{\theta,x}) - \nabla_\theta^i \mathbb{E}_{\pi_\theta} f(Y)\right| \leq Ce^{-\beta t}(1+|x|), \quad i = 0,1,2. \tag{5.3.9}$$

(ii) There exists a constant $C > 0$ such that for any $\theta \in \mathbb{R}^\ell$ and $i = 0,1,2$,

$$\left|\nabla_\theta^i \mathbb{E}_{\pi_\theta} f(Y)\right| \leq C. \tag{5.3.10}$$

(iii) There exists constants $C, \gamma > 0$ such that for any for any $\theta \in \mathbb{R}^\ell, x \in \mathbb{R}^d$, and $t > 0$,

$$\left|\nabla_x^j \nabla_\theta^i \mathbb{E} f(X_t^{\theta,x})\right| \leq Ce^{-\gamma t}, \quad i = 0,1, \quad j = 1,2. \tag{5.3.11}$$

The proof method for Proposition 5.3.4 is the same as in Proposition 4.1 of [123], although we need the convergence result for higher-order derivatives in (5.3.11). The detailed proof can be found in the Appendix A of [145].

We next prove that a solution exists to a Poisson equation for the fluctuation terms and, furthermore, that the solution satisfies certain linear bounds. We first introduce the process $\tilde{X}_t^{\theta,x,\tilde{x}}$, which satisfies the SDE:

$$\begin{cases} d\tilde{X}_t^{\theta,x,\tilde{x}} = \left[\nabla_x \mu(X_t^{\theta,x},\theta)\tilde{X}_t^{\theta,x,\tilde{x}} + \nabla_\theta \mu(X_t^{\theta,x},\theta)\right] dt + \left[\nabla_x \sigma(X_t^{\theta,x},\theta)\tilde{X}_t^{\theta,x,\tilde{x}} + \nabla_\theta \sigma(X_t^{\theta,x},\theta)\right] dW_t, \\ \tilde{X}_0^{\theta,x,\tilde{x}} = \tilde{x}, \end{cases} \tag{5.3.12}$$

where the Brownian is the same as in (5.1.1). It should be noted that $\tilde{X}_t^{\theta,x,0} = \nabla_\theta X_t^{\theta,x}$ almost surely.

Lemma 5.3.5. *Define the error function*

$$G^1(x,\tilde{x},\theta) = \left(\mathbb{E}_{\pi_\theta} f(Y) - \beta\right)\left(\nabla f(x)\tilde{x} - \nabla_\theta \mathbb{E}_{\pi_\theta} f(Y)\right)^\top \tag{5.3.13}$$

and the function

$$v^1(x,\tilde{x},\theta) = -\int_0^\infty \mathbb{E} G^1(X_t^{\theta,x},\tilde{X}_t^{\theta,x,\tilde{x}},\theta)dt. \tag{5.3.14}$$

Let $\mathcal{L}_{x,\tilde{x}}^\theta$ denote the infinitesimal generator of the process $(X_\cdot^{\theta,x},\tilde{X}_\cdot^{\theta,x,\tilde{x}})$, i.e. for any test function φ

$$\begin{aligned} \mathcal{L}_{x,\tilde{x}}^\theta \varphi(x,\tilde{x}) &= \mathcal{L}_x^\theta \varphi(x,\tilde{x}) + \sum_{k=1}^{\ell} \mathcal{L}_{\tilde{x}^{:,k}}^\theta \varphi(x,\tilde{x}) \\ &+ \sum_{j=1}^{\ell} tr\left(\nabla_{\tilde{x}^{:,j}} \nabla_x \varphi(x,\tilde{x})\sigma(x,\theta)\left(\nabla_x \sigma(x,\theta)\tilde{x}^{:,j} + \frac{\partial \sigma(x,\theta)}{\partial \theta_j}\right)^\top\right) \\ &+ \sum_{j<k} tr\left(\nabla_{\tilde{x}^{:,k}} \nabla_{\tilde{x}^{:,j}} \varphi(x,\tilde{x})\left(\nabla_x \sigma(x,\theta)\tilde{x}^{:,j} + \frac{\partial \sigma(x,\theta)}{\partial \theta_j}\right)\left(\nabla_x \sigma(x,\theta)\tilde{x}^{:,k} + \frac{\partial \sigma(x,\theta)}{\partial \theta_k}\right)^\top\right) \end{aligned}$$

where $\tilde{x}^{:,k}$ for $k \in \{1,\cdots,\ell\}$ is the k-th column of $\tilde{x}$.

Then, under Assumptions (A5.2.1) - (A5.2.4), $v^1(x,\tilde{x},\theta)$ is the classical solution of the Poisson equation

$$\mathcal{L}_{x,\tilde{x}}^\theta u(x,\tilde{x},\theta) = G^1(x,\tilde{x},\theta), \tag{5.3.15}$$

where $u = (u_1,\ldots,u_\ell)^\top \in \mathbb{R}^\ell$ is a vector, $\mathcal{L}_{x,\tilde{x}}^\theta u(x,\tilde{x},\theta) = (\mathcal{L}_{x,\tilde{x}}^\theta u_1(x,\tilde{x},\theta),\ldots,\mathcal{L}_{x,\tilde{x}}^\theta u_\ell(x,\tilde{x},\theta))^\top$. Furthermore, the solution v^1 satisfies the bound

$$\left|v^1(x,\tilde{x},\theta)\right| + \left|\nabla_\theta v^1(x,\tilde{x},\theta)\right| + \left|\nabla_x v^1(x,\tilde{x},\theta)\right| + \left|\nabla_{\tilde{x}} v^1(x,\tilde{x},\theta)\right| \le C\left(1 + |x| + |\tilde{x}|\right), \tag{5.3.16}$$

where $C > 0$ is a constant which does not depend upon $(x,\tilde{x},\theta)$.

Proof. We begin by proving that the integral (5.3.14) is finite. We divide (5.3.14) into two terms:

$$\begin{aligned} &v^1(x,\tilde{x},\theta) \\ =&\left(\mathbb{E}_{\pi_\theta} f(Y) - \beta\right)\int_0^\infty \left(\nabla_\theta \mathbb{E}_{\pi_\theta} f(Y) - \mathbb{E}\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right)^\top dt \\ =&\left(\mathbb{E}_{\pi_\theta} f(Y) - \beta\right)\left[\int_0^\infty \left(\nabla_\theta \mathbb{E}_{\pi_\theta} f(Y) - \nabla_\theta \mathbb{E} f(X_t^{\theta,x})\right)^\top dt + \int_0^\infty \left(\nabla_\theta \mathbb{E} f(X_t^{\theta,x}) - \mathbb{E}\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right)^\top dt\right] \\ =:&v^{1,1}(x,\theta) + v^{1,2}(x,\tilde{x},\theta). \end{aligned} \tag{5.3.17}$$

We first bound $v^{1,1}(x,\theta)$. Following the method in Lemma 3.3 of [144], we have by Proposition 5.3.4 and the dominated convergence theorem (DCT) that:

$$\begin{aligned}
\left|v^{1,1}(x,\theta)\right| &\le C\int_0^\infty \left|\nabla_\theta \mathbb{E}_{\pi_\theta} f(Y) - \nabla_\theta \mathbb{E} f(X_t^{\theta,x})\right| dt \le C(1+|x|),\\
\left|\nabla_\theta v^{1,1}(x,\theta)\right| &\le C\int_0^\infty \left|\nabla_\theta \mathbb{E}_{\pi_\theta} f(Y) - \nabla_\theta \mathbb{E} f(X_t^{\theta,x})\right| + C\int_0^\infty \left|\nabla_\theta^2 \mathbb{E}_{\pi_\theta} f(Y) - \nabla_\theta^2 \mathbb{E} f(X_t^{\theta,x})\right| \le C(1+|x|),\\
\left|\nabla_x^i v^{1,1}(x,\theta)\right| &\le C\int_0^\infty \left|\nabla_x^i \nabla_\theta \mathbb{E} f(X_t^{\theta,x})\right| dt \le C, \quad i=1,2.
\end{aligned} \tag{5.3.18}$$

For $v^{1,2}(x,\tilde{x},\theta)$, define

$$Z_t = \tilde{X}_t^{\theta,x.\tilde{x}_1} - \tilde{X}_t^{\theta,x,\tilde{x}_2}.$$

We can derive a differential inequality for $Z_t^{:,k}$, the k-th column of Z_t, using the inequality (**??**):

$$\frac{d}{dt}\mathbb{E}\left|Z_t^{:,k}\right|^2 \overset{(a)}{=} \mathbb{E}\left[2\left\langle \nabla_x \mu(X_t^{\theta,x_1},\theta) Z_t^{:,k},\ Z_t^{:,k}\right\rangle + \left|\nabla_x \sigma(X_t^{\theta,x_1},\theta) Z_t^{:,k}\right|^2\right] \le -\beta \mathbb{E}|Z_t^{:,k}|^2, \tag{5.3.19}$$

where step (a) is by using Itô's formula to $\left|Z_t^{:,k}\right|^2$. Therefore, we can prove the exponential decay:

$$\begin{aligned}
\mathbb{E}\left|\tilde{X}_t^{\theta,x,\tilde{x}_1} - \tilde{X}_t^{\theta,x,\tilde{x}_2}\right|^2 &\le Ce^{-\beta t}|\tilde{x}_1 - \tilde{x}_2|^2,\\
\mathbb{E}\left|\nabla_{\tilde{x}} \tilde{X}_t^{\theta,x,\tilde{x}}\right|^2 &\le Ce^{-\beta t}.
\end{aligned} \tag{5.3.20}$$

Let $\tilde{X}^{\theta,x,\tilde{x},:,k}$ denote the k-th column of the matrix $\tilde{X}_t^{\theta,x,\tilde{x}}$ and for any $m \in \{1,\cdots,d\}, n \in \{1,\cdots,\ell\}$, we know

$$d\frac{\partial \tilde{X}_t^{\theta,x,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}} = \nabla_x \mu(X_t^{\theta,x},\theta)\frac{\partial \tilde{X}_t^{\theta,x,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}} dt + \nabla_x \sigma(X_t^{\theta,x},\theta)\frac{\partial \tilde{X}_t^{\theta,x,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}} dW_t, \tag{5.3.21}$$

where $\tilde{x}^{m,n}$ denotes the (m,n) element of the matrix $\tilde{x}$. Let

$$\tilde{Z}_t^1 = \frac{\partial \tilde{X}_t^{\theta,x,\tilde{x}_1,:,k}}{\partial \tilde{x}^{m,n}} - \frac{\partial \tilde{X}_t^{\theta,x,\tilde{x}_2,:,k}}{\partial \tilde{x}^{m,n}}, \quad \tilde{Z}_t^2 = \frac{\partial \tilde{X}_t^{\theta,x_1,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}} - \frac{\partial \tilde{X}_t^{\theta,x_2,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}}.$$

Note that $\tilde{Z}_t^1$ satisfies the SDE

$$d\tilde{Z}_t^1 = \nabla_x \mu(X_t^{\theta,x},\theta)\tilde{Z}_t^1 dt + \nabla_x \sigma(X_t^{\theta,x},\theta)\tilde{Z}_t^1 dW_t \tag{5.3.22}$$

Similar to (5.3.19), we can get

$$\frac{d}{dt}\mathbb{E}\left|\tilde{Z}_t^1\right|^2 \le -\beta \mathbb{E}\left|\tilde{Z}_t^1\right|^2 \tag{5.3.23}$$

which derives

$$\begin{aligned}
\mathbb{E}\left|\nabla_{\tilde{x}} \tilde{X}_t^{\theta,x,\tilde{x}_1} - \nabla_{\tilde{x}} \tilde{X}_t^{\theta,x,\tilde{x}_2}\right|^2 &\le Ce^{-\beta t}|\tilde{x}_1 - \tilde{x}_2|^2,\\
\mathbb{E}\left|\nabla_{\tilde{x}}^2 \tilde{X}_t^{\theta,x,\tilde{x}}\right|^2 &\le Ce^{-\beta t}.
\end{aligned} \tag{5.3.24}$$

Then for $\tilde{Z}_t^2$

$$\begin{aligned}
d\tilde{Z}_t^2 = &\left(\nabla_x \mu(X_t^{\theta,x_1},\theta)\frac{\partial \tilde{X}_t^{\theta,x_1,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}} - \nabla_x \mu(X_t^{\theta,x_2},\theta)\frac{\partial \tilde{X}_t^{\theta,x_2,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}}\right) dt\\
&+ \left(\nabla_x \sigma(X_t^{\theta,x_1},\theta)\frac{\partial \tilde{X}_t^{\theta,x_1,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}} - \nabla_x \sigma(X_t^{\theta,x_2},\theta)\frac{\partial \tilde{X}_t^{\theta,x_2,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}}\right) dW_t,
\end{aligned} \tag{5.3.25}$$

and as in (**??**)

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left|\tilde{Z}_t^2\right|^2 =&\mathbb{E}\left[2\left\langle \nabla_x\mu(X_t^{\theta,x_1},\theta)\frac{\partial \tilde{X}_t^{\theta,x_1,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}} - \nabla_x\mu(X_t^{\theta,x_2},\theta)\frac{\partial \tilde{X}_t^{\theta,x_2,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}},\ \tilde{Z}_t^2\right\rangle\right]\\ &+\mathbb{E}\left[\left|\nabla_x\sigma(X_t^{\theta,x_1},\theta)\frac{\partial \tilde{X}_t^{\theta,x_1,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}} - \nabla_x\sigma(X_t^{\theta,x_2},\theta)\frac{\partial \tilde{X}_t^{\theta,x_2,\tilde{x},:,k}}{\partial \tilde{x}^{m,n}}\right|^2\right]\\ \leq&\mathbb{E}\left[2\left\langle \nabla_x\mu(X_t^{\theta,x_1},\theta)\tilde{Z}_t^2,\ \tilde{Z}_t^2\right\rangle + 2\left|\nabla_x\sigma(X_t^{\theta,x_1},\theta)\tilde{Z}_t^2\right|^2\right] + \beta\mathbb{E}|\tilde{Z}_t^2|^2 + C\mathbb{E}\left|X_t^{\theta,x_1}-X_t^{\theta,x_2}\right|^2\\ \leq&-\beta\mathbb{E}|\tilde{Z}_t^2|^2 + Ce^{-\beta t}|x_1-x_2|^2,\end{aligned} \tag{5.3.26}$$

which derives

$$\begin{aligned}\mathbb{E}\left|\nabla_{\tilde{x}}\tilde{X}_t^{\theta,x_1,\tilde{x}} - \nabla_{\tilde{x}}\tilde{X}_t^{\theta,x_2,\tilde{x}}\right|^2 &\leq Ce^{-\beta t}|x_1-x_2|^2,\\ \mathbb{E}\left|\nabla_x\nabla_{\tilde{x}}\tilde{X}_t^{\theta,x,\tilde{x}}\right|^2 &\leq Ce^{-\beta t}.\end{aligned} \tag{5.3.27}$$

Combining (5.3.20), (5.3.24) and (5.3.27) we can establish bounds on $v^{1,2}(x,\tilde{x},\theta)$.

$$\begin{aligned}\left|v^{1,2}(x,\tilde{x},\theta)\right| &\overset{(a)}{\leq} C\int_0^\infty \mathbb{E}\left|\nabla f(X_t^{\theta,x})\left(\tilde{X}_t^{\theta,x,\tilde{x}} - \tilde{X}_t^{\theta,x,0}\right)\right| dt \leq \int_0^\infty Ce^{-\frac{\beta}{2}t}|\tilde{x}|dt \leq C|\tilde{x}|\\ \left|\nabla_{\tilde{x}}^i v^{1,2}(x,\tilde{x},\theta)\right| &\leq C\int_0^\infty \mathbb{E}\left|\nabla_{\tilde{x}}^i\tilde{X}_t^{\theta,x,\tilde{x}}\right| dt \leq \int_0^\infty Ce^{-\frac{\beta}{2}t}dt \leq C,\quad i=1,2\\ \left|\nabla_x\nabla_{\tilde{x}}v^{1,2}(x,\tilde{x},\theta)\right| &\leq C\int_0^\infty \mathbb{E}\left[\left|\nabla_x^2 f(X_t^{\theta,x})\nabla_x X_t^{\theta,x}\right|\cdot\left|\nabla_{\tilde{x}}X_t^{\theta,x,\tilde{x}}\right|\right] dt + C\int_0^\infty \mathbb{E}\left|\nabla_x\nabla_{\tilde{x}}\tilde{X}_t^{\theta,x,\tilde{x}}\right| dt \leq C\end{aligned} \tag{5.3.28}$$

where in step (a) we use the fact $\nabla_\theta X_t^{\theta,x} \overset{a.s.}{=} \tilde{X}_t^{\theta,x,0}$.

The analysis of $\nabla_x^i v^{1,2}$ for $i=1,2$ and $\nabla_\theta v^{1,2}$ is similar to the calculations for $v^{1,1}$. Define

$$\bar{Z}_t = \nabla_\theta\tilde{X}_t^{\theta,x.\tilde{x}_1} - \nabla_\theta\tilde{X}_t^{\theta,x,\tilde{x}_2}.$$

$\bar{Z}_t$ satisfies the SDE:

$$\begin{aligned}d\bar{Z}_t =& \left(\left\langle\nabla_x^2\mu(X_t^{\theta,x},\theta)\nabla_\theta X_t^{\theta,x},\ Z_t\right\rangle + \nabla_x\nabla_\theta\mu(X_t^{\theta,x},\theta)Z_t + \nabla_x\mu(X_t^{\theta,x},\theta)\bar{Z}_t\right)dt,\\ &+\left(\left\langle\nabla_x^2\sigma(X_t^{\theta,x},\theta)\nabla_\theta X_t^{\theta,x},\ Z_t\right\rangle + \nabla_x\nabla_\theta\sigma(X_t^{\theta,x},\theta)Z_t + \nabla_x\sigma(X_t^{\theta,x},\theta)\bar{Z}_t\right)dW_t,\end{aligned} \tag{5.3.29}$$

where $\langle\ ,\ \rangle$ in the equation above is defined as:

$$\begin{aligned}\left\langle\nabla_x^2\mu(X_t^{\theta,x},\theta)\nabla_\theta X_t^{\theta,x},\ Z_t\right\rangle_{m,p,q} &= \left\langle\nabla_x^2\mu_m(X_t^{\theta,x},\theta)\frac{\partial X_t^{\theta,x}}{\partial\theta_q},\ Z_t^{:,p}\right\rangle,\\ \left\langle\nabla_x^2\sigma(X_t^{\theta,x},\theta)\nabla_\theta X_t^{\theta,x},\ Z_t\right\rangle_{m,n,p,q} &= \left\langle\nabla_x^2\sigma_{mn}(X_t^{\theta,x},\theta)\frac{\partial X_t^{\theta,x}}{\partial\theta_q},\ Z_t^{:,p}\right\rangle.\end{aligned} \tag{5.3.30}$$

Thus by the same calculations as in (**??**) and the uniform bounds for the derivatives of μ,σ, we can derive the differential inequality:

$$\frac{d}{dt}\mathbb{E}\left|\bar{Z}_t\right|^2 \leq -\beta\mathbb{E}|\bar{Z}_t|^2 + C\mathbb{E}|Z_t|^2. \tag{5.3.31}$$

Using an integrating factor, we have

$$\frac{d}{dt}\left(e^{\beta t}\mathbb{E}\left|\bar{Z}_t\right|^2\right) \leq Ce^{\beta t}\mathbb{E}\left|Z_t\right|^2, \tag{5.3.32}$$

which combined with (5.3.20) yields

$$\begin{aligned}\mathbb{E}|\nabla_\theta \tilde{X}_t^{\theta,x,\tilde{x}_1} - \nabla_\theta \tilde{X}_t^{\theta,x,\tilde{x}_2}|^2 &= \mathbb{E}\left|\bar{Z}_t\right|^2 \leq e^{-\beta t}\left|\tilde{x}_1 - \tilde{x}_2\right|^2 + e^{-\beta t}\int_0^t e^{\beta s}\mathbb{E}\left|Z_s\right|^2 ds \\ &\leq Ce^{-\frac{\beta}{2}t}|\tilde{x}_1 - \tilde{x}_2|^2. \end{aligned} \tag{5.3.33}$$

Consequently,

$$\begin{aligned}\left|\nabla_\theta v^{1,2}(x,\tilde{x},\theta)\right| \leq & C|\tilde{x}|\cdot|\nabla_\theta \mathbb{E}_{\pi_\theta} f(Y)| + C\int_0^\infty \mathbb{E}\left|\nabla f(X_t^{\theta,x})\left(\nabla_\theta \tilde{X}_t^{\theta,x,\tilde{x}} - \nabla_\theta \tilde{X}_t^{\theta,x,0}\right)\right| dt \\ & + \int_0^\infty \mathbb{E}\left|\left\langle \nabla^2 f(X_t^{\theta,x})\nabla_\theta X_t^{\theta,x},\ \tilde{X}_t^{\theta,x,\tilde{x}} - \tilde{X}_t^{\theta,x,0}\right\rangle\right| dt \\ \overset{(a)}{\leq} & C|\tilde{x}| + \int_0^\infty Ce^{-\frac{\beta}{4}t}|\tilde{x}|dt + \int_0^\infty Ce^{-\frac{\beta}{2}t}|\tilde{x}|dt \\ \leq & C|\tilde{x}|, \end{aligned} \tag{5.3.34}$$

where in step (a) we use the Cauchy-Schwarz inequality, (**??**), (5.3.20) and (5.3.33).

Finally, for the derivatives with respect to x, define

$$\hat{Z}_t = \nabla_x \tilde{X}_t^{\theta,x.\tilde{x}_1} - \nabla_x \tilde{X}_t^{\theta,x,\tilde{x}_2}$$

and as in (5.3.29) and (5.3.30) it satisfies the SDE

$$\begin{aligned} d\hat{Z}_t = & \left(\left\langle \nabla_x^2 \mu(X_t^{\theta,x},\theta)\nabla_x X_t^{\theta,x},\ Z_t\right\rangle + \nabla_x \mu(X_t^{\theta,x},\theta)\hat{Z}_t\right) dt \\ & + \left(\left\langle \nabla_x^2 \sigma(X_t^{\theta,x},\theta)\nabla_\theta X_t^{\theta,x},\ Z_t\right\rangle + \nabla_x \sigma(X_t^{\theta,x},\theta)\hat{Z}_t\right) dW_t, \end{aligned} \tag{5.3.35}$$

Similarly, we can derive the differential inequality

$$\frac{d}{dt}\mathbb{E}|\hat{Z}_t|^2 \leq -\beta\mathbb{E}|\hat{Z}_t|^2 + C\mathbb{E}|Z_t|^2.$$

Consequently,

$$\mathbb{E}\left|\nabla_x \tilde{X}_t^{\theta,x,\tilde{x}_1} - \nabla_x \tilde{X}_t^{\theta,x,\tilde{x}_2}\right|^2 \leq Ce^{-\frac{\beta}{2}t}\left|\tilde{x}_1 - \tilde{x}_2\right|^2. \tag{5.3.36}$$

Due to Lemma 5.3.1,

$$\mathbb{E}\left|X_t^{\theta,x_1} - X_t^{\theta,x_2}\right|^2 \leq e^{-\beta t}|x_1 - x_2|^2, \tag{5.3.37}$$

which, combined with the dominated convergence theorem, yields

$$\mathbb{E}\left|\nabla_x X_t^{\theta,x}\right|^2 \leq e^{-\beta t}. \tag{5.3.38}$$

Therefore,

$$\begin{aligned} & \left|\nabla_x v^{1,2}(x,\tilde{x},\theta)\right| \\ \leq & C\int_0^\infty \mathbb{E}\left|\nabla f(X_t^{\theta,x})\left(\nabla_x \tilde{X}_t^{\theta,x,\tilde{x}} - \nabla_x \tilde{X}_t^{\theta,x,0}\right)\right| dt \\ & + C\int_0^\infty \mathbb{E}\left|\left\langle \nabla^2 f(X_t^{\theta,x})\nabla_x X_t^{\theta,x},\ \tilde{X}_t^{\theta,x,\tilde{x}} - \tilde{X}_t^{\theta,x,0}\right\rangle\right| dt \\ \overset{(a)}{\leq} & \int_0^\infty Ce^{-\beta t}|\tilde{x}|dt + \int_0^\infty Ce^{-\frac{\beta}{2}t}|\tilde{x}|dt \\ \leq & C|\tilde{x}|, \end{aligned} \tag{5.3.39}$$

where step (a) is by Cauchy-Schwarz inequality, (5.3.38), (5.3.20) and (5.3.36). The bound for $\nabla_x^2 v^{1,2}$ follows from exactly the same method. Combining the bounds for $v^{1,1}$ and $v^{1,2}$ proves the desired bound (5.3.16). Using the same calculations as in Lemma 3.3 of [144], we can show that v^1

is the classical solution of PDE (5.3.15) and thus the proof is completed. □

We will also need bounds on the moments of X_t and $\tilde{X}_t$ in order to analyze the fluctuation term $\Delta^i_{\tau_n,\sigma_n+\eta}$.

Lemma 5.3.6. *There exists a constant $C>0$ such that the processes $X_t, \tilde{X}_t$ in (5.1.3) satisfy*

$$\mathbb{E}_x|X_t|^8 \le C\left(1+|x|^8\right), \quad \mathbb{E}_{x,\tilde{x}}|\tilde{X}_t|^8 \le C\left(1+|\tilde{x}|^8\right), \tag{5.3.40}$$

where $\mathbb{E}_{x,\tilde{x}}$ is the conditional expectation given that $X_0=x$ and $\tilde{X}_0=\tilde{x}$. Furthermore, we have the bounds

$$\mathbb{E}_x\left(\sup_{0\le t'\le t}|X_{t'}|^4\right) = O(\sqrt{t}) \quad \text{as } t\to\infty, \tag{5.3.41}$$

$$\mathbb{E}_{x,\tilde{x}}\left(\sup_{0\le t'\le t}|\tilde{X}_{t'}|^4\right) = O(\sqrt{t}) \quad \text{as } t\to\infty. \tag{5.3.42}$$

Proof. By Itô's formula, for any $m\ge 1$ we have

$$\begin{aligned} d|X_t|^{2m} &= 2m|X_t|^{2m-2}\langle \mu(X_t,\theta_t),\ X_t\rangle dt + 2m|X_t|^{2m-2}\langle \sigma(X_t,\theta_t),\ X_t\rangle dW_t \\ &\quad + m|X_t|^{2m-2}\cdot|\sigma(X_t,\theta_t)|^2 dt + 2m(m-1)|X_t|^{2m-4}\cdot|\langle X_t,\ \sigma(X_t,\theta_t)\rangle|^2. \end{aligned} \tag{5.3.43}$$

We use induction to prove the bound for the 8-th moment. First let $m=1$ in (5.3.43), by the same proof as in Lemma 5.3.3, we have

$$\frac{d}{dt}\mathbb{E}_x|X_t|^2 \le -\beta\mathbb{E}_x|X_t|^2 + \mathbb{E}_x\left(\frac{1}{\beta}|\mu(0,\theta_t)|^2 + 2|\sigma(0,\theta_t)|^2\right) \le -\beta\mathbb{E}_x|X_t|^2 + C, \tag{5.3.44}$$

which yields the bound for the second moment

$$\mathbb{E}_x|X_t|^2 \le C\left(1+|x|^2\right). \tag{5.3.45}$$

For $k\in\{1,2,\cdots,\ell\}$, let $\tilde{X}_t^{:,k}$ denote the k-th column of $\tilde{X}_t$. $\left|\tilde{X}_t^{:,k}\right|^2$ satisfies the following SDE:

$$\begin{aligned} d\left|\tilde{X}_t^{:,k}\right|^2 =& 2\left\langle \nabla_x\mu(X_t,\theta_t)\tilde{X}_t^{:,k} + \frac{\partial\mu(X_t,\theta_t)}{\partial\theta_k},\ \tilde{X}_t^{:,k}\right\rangle dt + \left|\nabla_x\sigma(X_t,\theta_t)\tilde{X}_t^{:,k} + \frac{\partial\sigma(X_t,\theta_t)}{\partial\theta_k}\right|^2 dt \\ &+2\left\langle \nabla_x\sigma(X_t,\theta_t)\tilde{X}_t^{:,k} + \frac{\partial\sigma(X_t,\theta_t)}{\partial\theta_k},\ \tilde{X}_t^{:,k}\right\rangle dW_t. \end{aligned} \tag{5.3.46}$$

Similar to (**??**), we can derive the differential inequality

$$\frac{d}{dt}\mathbb{E}_{x,\tilde{x}}\left|\tilde{X}_t^{:,k}\right|^2 \le -\beta\mathbb{E}_{x,\tilde{x}}\left|\tilde{X}_t^{:,k}\right|^2 + \mathbb{E}_{x,\tilde{x}}\left(\frac{1}{\beta}\left|\frac{\partial\mu(X_t,\theta_t)}{\partial\theta_k}\right|^2 + 2\left|\frac{\partial\sigma(X_t,\theta_t)}{\partial\theta_k}\right|^2\right) \le -\beta\mathbb{E}_{x,\tilde{x}}\left|\tilde{X}_t^{:,k}\right|^2 + C. \tag{5.3.47}$$

Therefore,

$$\mathbb{E}_{x,\tilde{x}}|\tilde{X}_t|^2 \le C\left(1+|\tilde{x}|^2\right). \tag{5.3.48}$$

Recalling that conditions (5.2.2) and (5.2.3), we can get there exists a constant $C>0$ such that, for any $x\in\mathbb{R}^d$ and $\theta\in\mathbb{R}^\ell$,

$$2\langle\mu(x,\theta),\ x\rangle + 7|\sigma(x,\theta)|^2 \le -\beta|x|^2 + C. \tag{5.3.49}$$

Actually,

$$\begin{aligned}
&2\langle \mu(x,\theta),\ x\rangle + 7|\sigma(x,\theta)|^2 \\
=&2\langle \mu(x,\theta)-\mu(0,\theta),\ x\rangle + 2\langle \mu(0,\theta),\ x\rangle + 7|\sigma(x,\theta)-\sigma(0,\theta)+\sigma(0,\theta)|^2 \\
\overset{(a)}{\leq}& -2\beta|x|^2 + 2\langle \mu(0,\theta),\ x\rangle + 7|\sigma(0,\theta)|^2 + 14|\sigma(x,\theta)-\sigma(0,\theta)|\cdot|\sigma(0,\theta)| \\
\overset{(b)}{\leq}& -\beta|x|^2 + C\left(|\mu(0,\theta)|^2 + |\sigma(0,\theta)|^2\right),
\end{aligned} \tag{5.3.50}$$

where step (a) uses the inequality (5.2.2) and step (b) uses Young's inequality and the inequality (5.2.1). Thus now let $m=2$ in (5.3.43) and use the bound (5.3.49),

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}_x|X_t|^4 &= 4\mathbb{E}_x\left(|X_t|^2\langle \mu(X_t,\theta_t),\ X_t\rangle\right)dt + \mathbb{E}_x\left(2|X_t|^2|\sigma(X_t,\theta_t)|^2 + 4\left|\langle \sigma(X_t,\theta_t),\ X_t\rangle\right|^2\right) \\
&\leq \mathbb{E}_x\left[|X_t|^2\left(4\langle \mu(X_t,\theta_t),\ X_t\rangle + 6|\sigma(X_t,\theta_t)|^2\right)\right] \\
&\leq -\beta\mathbb{E}_x|X_t|^4 + C\mathbb{E}_x|X_t|^2,
\end{aligned} \tag{5.3.51}$$

which together with (5.3.45) and Gronwall's inequality prove the bound for the fourth moment of X_t. Similarly, as in (5.3.50)

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}_{x,\tilde{x}}\left|\tilde{X}_t^{:,k}\right|^4 \\
\leq&\mathbb{E}_{x,\tilde{x}}\left[\left|\tilde{X}_t^{:,k}\right|^2\left(4\left\langle \nabla_x\mu(X_t,\theta_t)\tilde{X}_t^{:,k} + \frac{\partial\mu(X_t,\theta_t)}{\partial\theta_k},\ \tilde{X}_t^{:,k}\right\rangle + 6\left|\nabla_x\sigma(X_t,\theta_t)\tilde{X}_t^{:,k} + \frac{\partial\sigma(X_t,\theta_t)}{\partial\theta_k}\right|^2\right)\right] \\
\leq& -\beta\mathbb{E}_{x,\tilde{x}}\left|\tilde{X}_t^{:,k}\right|^4 + C\mathbb{E}_{x,\tilde{x}}\left|\tilde{X}_t^{:,k}\right|^2,
\end{aligned} \tag{5.3.52}$$

which together with (5.3.48) derives the estimate for $\tilde{X}_t$ in (5.3.40). By induction, we can prove the bound for the sixth and eighth moments of $(X_t,\tilde{X}_t)$ in (5.3.40).

Finally, as in (5.3.43) and use (5.3.49), we have

$$\begin{aligned}
|X_t|^8 &= |x|^8 + 8\int_0^t |X_s|^6\langle \mu(X_s,\theta_s),\ X_s\rangle\,ds + 8\int_0^t |X_s|^6\langle \sigma(X_s,\theta_s),\ X_s\rangle\,dW_s \\
&\quad + 24\int_0^t |X_t|^4\cdot|\langle \sigma(X_s,\theta_s),\ X_s\rangle|^2\,ds + 4\int_0^t |X_s|^6\cdot|\sigma(X_s,\theta_s)|^2\,ds \\
&\leq -4\beta\int_0^t |X_s|^8\,ds + C\int_0^t |X_s|^6 ds + 8\int_0^t |X_s|^6\langle \sigma(X_s,\theta_s),\ X_s\rangle dW_s,
\end{aligned} \tag{5.3.53}$$

which together with the Burkholder-Davis-Gundy inequality and (5.3.45) derive that there exists a constant C such that

$$\begin{aligned}
\mathbb{E}_x\sup_{0\leq t'\leq t}|X_{t'}|^8 &\leq |x|^8 + ct + C\mathbb{E}_x\left(\int_0^t |X_s|^{14}\cdot|\sigma(X_s,\theta_s)|^2 ds\right)^{\frac{1}{2}} \\
&\leq |x|^8 + ct + C\mathbb{E}_x\left(\sup_{0\leq t'\leq t}|X_{t'}|^8\cdot\int_0^t |X_s|^6\cdot|\sigma(X_s,\theta_s)|^2 ds\right)^{\frac{1}{2}} \\
&\overset{(a)}{\leq} |x|^8 + ct + \frac{1}{2}\mathbb{E}_x\sup_{0\leq t'\leq t}|X_{t'}|^8 + C\int_0^t \mathbb{E}_x\left[|X_s|^6 + |X_s|^8\right]ds,
\end{aligned} \tag{5.3.54}$$

where step (a) is by Young's inequality. Thus, combining (5.3.40) and (5.3.54) we obtain

$$\mathbb{E}_x\left(\sup_{0\leq t'\leq t}|X_{t'}|^8\right) = O(t) \quad \text{as } t\to\infty,$$

which derives (5.3.41). Similarly for (5.3.42), using Itô's formula for $\left|\tilde{X}_t\right|^8$ and the Burkholder-Davis-Gundy inequality,

$$\begin{aligned}\mathbb{E}_{x,\tilde{x}} \sup_{0\le t'\le t}\left|\tilde{X}_{t'}\right|^8 &\le |\tilde{x}|^8 + ct + C\mathbb{E}_{x,\tilde{x}}\left(\int_0^t |\tilde{X}_s|^{14}\cdot\left|\nabla_x\sigma(X_s,\theta_s)\tilde{X}_s + \nabla_\theta\sigma(X_s,\theta_s)\right|^2 ds\right)^{\frac{1}{2}} \\ &\le |\tilde{x}|^8 + ct + \frac{1}{2}\mathbb{E}_{x,\tilde{x}}\sup_{0\le t'\le t}\left|\tilde{X}_{t'}\right|^8 + C\int_0^t \mathbb{E}_{x,\tilde{x}}\left[\left|\tilde{X}_s\right|^6 + \left|\tilde{X}_s\right|^8\right] ds,\end{aligned} \tag{5.3.55}$$

which together with (5.3.40) derive (5.3.42).

□

We can now bound the first fluctuation term $\Delta^1_{\tau_k,\sigma_k+\eta}$ in (5.3.3) using the estimates from Lemma 5.3.5 and Lemma 5.3.6.

Lemma 5.3.7. *Under Assumptions A5.2.1 - A5.2.5, for any fixed $\eta > 0$,*

$$\left|\Delta^1_{\tau_n,\sigma_n+\eta}\right| \to 0 \text{ as } n\to\infty, \quad a.s. \tag{5.3.56}$$

Proof. We will express $\Delta^i_{\tau_n,\sigma_n+\eta}$ in terms of the Poisson equation in Lemma 5.3.5 and then prove it vanishes as n becomes large. Consider the function

$$G^1(x,\tilde{x},\theta) = \left(\mathbb{E}_{\pi_\theta} f(Y) - \beta\right)\left(\nabla f(x)\tilde{x} - \nabla_\theta \mathbb{E}_{\pi_\theta} f(Y)\right)^\top.$$

By Lemma 5.3.5, the Poisson equation $\mathcal{L}^\theta_{x\tilde{x}} u(x,\tilde{x},\theta) = G^1(x,\tilde{x},\theta)$ will have a unique smooth solution $v^1(x,\tilde{x},\theta)$ that grows at most linearly in $(x,\tilde{x})$. Let us apply Itô's formula to the function

$$u^1(t,x,\tilde{x},\theta) := \alpha_t v^1(x,\tilde{x},\theta) \in \mathbb{R}^\ell,$$

evaluated on the stochastic process $(X_t,\tilde{X}_t,\theta_t)$. Recall that u_i denotes the i-th element of u and $\tilde{X}^{:,k}_t$ be the k-th column of the matrix $\tilde{X}_t$ for $i,k\in\{1,2,\cdots,\ell\}$. Then,

$$\begin{aligned}u^1_i\left(\sigma,X_\sigma,\tilde{X}_\sigma,\theta_\sigma\right) =& u^1_i\left(\tau,X_\tau,\tilde{X}_\tau,\theta_\tau\right) + \int_\tau^\sigma \partial_s u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right) ds + \int_\tau^\sigma \mathcal{L}^{\theta_s}_{x\tilde{x}} u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right) ds \\ &+ \int_\tau^\sigma \nabla_\theta u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right) d\theta_s + \int_\tau^\sigma \nabla_x u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right)\sigma(X_s,\theta_s) dW_s \\ &+ \sum_{k=1}^\ell \int_\tau^\sigma \nabla_{\tilde{x}^{:,k}} u^1_i\left(s,X_s,\tilde{X}_s,\theta_s\right)\left(\nabla_x\sigma(X_s,\theta_s)\tilde{X}^{:,k}_s + \frac{\partial\sigma(X_s,\theta_s)}{\partial\theta_k}\right) dW_s.\end{aligned} \tag{5.3.57}$$

Rearranging the previous equation, we obtain the representation

$$\begin{aligned}
\Delta^1_{\tau_n,\sigma_n+\eta} &= \int_{\tau_n}^{\sigma_n+\eta} \alpha_s G^1(X_s,\tilde{X}_s,\theta_s)ds \\
&= \int_{\tau_k}^{\sigma_k+\eta} \mathcal{L}^{\theta_s}_{x\tilde{x}} u^1\left(s,X_s,\tilde{X}_s,\theta_s\right) ds \\
&= \alpha_{\sigma_n+\eta} v^1\left(X_{\sigma_n+\eta},\tilde{X}_{\sigma_n+\eta},\theta_{\sigma_n+\eta}\right) - \alpha_{\tau_n} v^1\left(X_{\tau_n},\tilde{X}_{\tau_n},\theta_{\tau_n}\right) - \int_{\tau_n}^{\sigma_n+\eta} \alpha'_s v^1\left(X_s,\tilde{X}_s,\theta_s\right) ds \\
&+ \int_{\tau_n}^{\sigma_n+\eta} 2\alpha_s^2 \nabla_\theta v^1\left(X_s,\tilde{X}_s,\theta_s\right)\left(f(\bar{X}_s)-\beta\right)\left(\nabla f(X_s)\tilde{X}_s\right)^\top ds \\
&- \int_{\tau_n}^{\sigma_n+\eta} \alpha_s \nabla_x v^1\left(X_s,\tilde{X}_s,\theta_s\right)\sigma(X_s,\theta_s)dW_s \\
&- \sum_{k=1}^{\ell}\int_{\tau_n}^{\sigma_n+\eta} \alpha_s \nabla_{\tilde{x}^{:,k}} v^1\left(X_s,\tilde{X}_s,\theta_s\right)\left(\nabla_x\sigma(X_s,\theta_s)\tilde{X}_s^{:,k} + \frac{\partial\sigma(X_s,\theta_s)}{\partial\theta_k}\right) dW_s.
\end{aligned} \tag{5.3.58}$$

The next step is to treat each term on the right hand side of (5.3.58) separately. For this purpose, let us first set

$$J_t^{1,1} = \alpha_t \sup_{s\in[0,t]}\left|v^1\left(X_s,\tilde{X}_s,\theta_s\right)\right|. \tag{5.3.59}$$

By (5.3.16) and Lemma 5.3.6, there exists a constant C such that

$$\begin{aligned}
\mathbb{E}\left|J_t^{1,1}\right|^2 &\le C\alpha_t^2\mathbb{E}\left[1+\sup_{s\in[0,t]}|X_s|^2+\sup_{s\in[0,t]}\left|\tilde{X}_s\right|^2\right] \\
&= C\alpha_t^2\left[1+\sqrt{t}\frac{\mathbb{E}\sup_{s\in[0,t]}|X_s|^2+\mathbb{E}\sup_{s\in[0,t]}\left|\tilde{X}_s\right|^2}{\sqrt{t}}\right] \\
&\le C\alpha_t^2\sqrt{t}.
\end{aligned} \tag{5.3.60}$$

Let $p>0$ be the constant in Assumption A5.2.5 such that $\lim_{t\to\infty}\alpha_t^2 t^{1/2+2p}=0$ and for any $\delta\in(0,p)$ define the event $A_{t,\delta}=\left\{J_t^{1,1}\ge t^{\delta-p}\right\}$. Then we have for t large enough such that $\alpha_t^2 t^{1/2+2p}\le 1$

$$\mathbb{P}\left(A_{t,\delta}\right)\le\frac{\mathbb{E}\left|J_t^{1,1}\right|^2}{t^{2(\delta-p)}}\le C\frac{\alpha_t^2 t^{1/2+2p}}{t^{2\delta}}\le C\frac{1}{t^{2\delta}}.$$

The latter implies that

$$\sum_{m\in\mathbb{N}}\mathbb{P}\left(A_{2^m,\delta}\right)<\infty.$$

Therefore, by the Borel-Cantelli lemma we have that for every $\delta\in(0,p)$ there is a finite positive random variable $d(\omega)$ and some $m_0<\infty$ such that for every $m\ge m_0$ one has

$$J_{2^n}^{1,1}\le\frac{d(\omega)}{2^{m(p-\delta)}}.$$

Thus, for $t\in\left[2^m,2^{m+1}\right)$ and $m\ge m_0$ one has for some finite constant $C<\infty$

$$J_t^{1,1}\le C\alpha_{2^{m+1}}\sup_{s\in(0,2^{m+1}]}\left|v^1\left(X_s,\tilde{X}_s,\theta_s\right)\right|\le C\frac{d(\omega)}{2^{(m+1)(p-\delta)}}\le C\frac{d(\omega)}{t^{p-\delta}},$$

which proves that for $t \geq 2^{m_0}$ with probability one

$$J_t^{1,1} \leq C\frac{d(\omega)}{t^{p-\delta}} \to 0, \text{ as } t \to \infty. \tag{5.3.61}$$

Next we consider the term

$$J_{t,0}^{1,2} = \int_0^t \left| \alpha_s' v^1\left(X_s, \tilde{X}_s, \theta_s\right) - 2\alpha_s^2 \nabla_\theta v^1\left(X_s, \tilde{X}_s, \theta_s\right)\left(f(\bar{X}_s) - \beta\right)\left(\nabla f(X_s)\tilde{X}_s\right)^\top \right| ds.$$

Noting that by the same approach for X_t in Lemma 5.3.6, we can prove that there exists a constant $C > 0$ such that

$$\mathbb{E}_{\bar{x}}|\bar{X}_t|^4 \leq C\left(1 + |\bar{x}|^4\right), \quad \mathbb{E}_{\bar{x}}\left(\sup_{0\leq t'\leq t}|\bar{X}_{t'}|^2\right) = O(\sqrt{t}) \quad \text{as } t \to \infty. \tag{5.3.62}$$

Thus

$$\begin{aligned}
\sup_{t>0}\mathbb{E}\left|J_{t,0}^{1,2}\right| &\overset{(a)}{\leq} C\int_0^\infty \left(|\alpha_s'| + \alpha_s^2\right)\left(1 + \mathbb{E}\left|X_s\right|^4 + \mathbb{E}|\tilde{X}_s|^4 + \mathbb{E}\left|\bar{X}_s\right|^2\right) ds \\
&\overset{(b)}{\leq} C\int_0^\infty \left(|\alpha_s'| + \alpha_s^2\right) ds \\
&\leq C,
\end{aligned}$$

where step (a) is by Assumption A5.2.4 and (5.3.16) and in step (b) we use (5.3.40). Thus there is a finite random variable $J_{\infty,0}^{1,2}$ such that

$$J_{t,0}^{1,2} \to J_{\infty,0}^{1,2}, \text{as } t \to \infty \text{ with probability one.} \tag{5.3.63}$$

The last term we need to consider is the martingale term

$$\begin{aligned}
J_{t,0}^{1,3} = &\int_0^t \alpha_s \nabla_x v^1\left(X_s, \tilde{X}_s, \theta_s\right)\sigma(X_s, \theta_s) dW_s \\
&+ \sum_{k=1}^{\ell}\int_0^t \alpha_s \nabla_{\tilde{x}^{:,k}} v^1\left(X_s, \tilde{X}_s, \theta_s\right)\left(\nabla_x \sigma(X_s, \theta_s)\tilde{X}_s^{:,k} + \frac{\partial\sigma(X_s, \theta_s)}{\partial\theta_k}\right) dW_s.
\end{aligned}$$

By Doob's inequality, Assumption A5.2.5, (5.3.16), (5.3.40), and using calculations similar to the ones for the term $J_{t,0}^{1,2}$, we can show that for some finite constant $C < \infty$,

$$\sup_{t>0}\mathbb{E}\left|J_{t,0}^{1,3}\right|^2 \leq C\int_0^\infty \alpha_s^2\left(1 + \mathbb{E}|X_t|^4 + \mathbb{E}|\tilde{X}_t|^4\right) ds < \infty$$

Thus, by Doob's martingale convergence theorem there is a square integrable random variable $J_{\infty,0}^{1,3}$ such that

$$J_{t,0}^{1,3} \to J_{\infty,0}^{1,3}, \quad \text{as } t \to \infty \text{ both almost surely and in } L^2. \tag{5.3.64}$$

Let us now return to (5.3.58). Using the terms $J_t^{1,1}, J_{t,0}^{1,2}$, and $J_{t,0}^{1,3}$ we can write

$$\left|\Delta_{\tau_n,\sigma_n+\eta}^1\right| \leq J_{\sigma_n+\eta}^{1,1} + J_{\tau_n}^{1,1} + \left|J_{\sigma_n+\eta,\tau_n}^{1,2}\right| + \left|J_{\sigma_n+\eta,\tau_n}^{1,3}\right|,$$

which together with (5.3.61), (5.3.63), and (5.3.64) prove the statement of the Lemma. □

We will next prove a similar convergence result for $\Delta_{\tau_n,\sigma_n+\eta}^2$. We must first prove an extension of Lemma 5.3.5 for the Poisson equation.

Lemma 5.3.8. *Define the error function*

$$G^2(x, \tilde{x}, \bar{x}, \theta) = \left[f(\bar{x}) - \mathbb{E}_{\pi_\theta} f(Y)\right]\left(\nabla f(x)\tilde{x}\right)^\top. \tag{5.3.65}$$

Under Assumptions (A5.2.1) - (A5.2.4), the function

$$v^2(x,\tilde{x},\bar{x},\theta) = -\int_0^\infty \mathbb{E}G^2(X_t^{\theta,x},\tilde{X}_t^{\theta,x,\tilde{x}},\bar{X}_t^{\theta,\bar{x}},\theta)dt \tag{5.3.66}$$

is the classical solution of the Poisson equation

$$\mathcal{L}^\theta_{x,\tilde{x},\bar{x}}u(x,\tilde{x},\bar{x},\theta) = G^2(x,\tilde{x},\bar{x},\theta), \tag{5.3.67}$$

where $\mathcal{L}^\theta_{x,\tilde{x},\bar{x}}$ is generator of the process $(X_\cdot^{\theta,x},\tilde{X}_\cdot^{\theta,x,\tilde{x}},\bar{X}_\cdot^{\theta,\bar{x}})$, i.e. for any test function φ

$$\mathcal{L}^\theta_{x,\tilde{x},\bar{x}}\varphi(x,\tilde{x},\bar{x}) = \mathcal{L}^\theta_{x,\tilde{x}}\varphi(x,\tilde{x},\bar{x}) + \mathcal{L}^\theta_{\bar{x}}\varphi(x,\tilde{x},\bar{x}). \tag{5.3.68}$$

Furthermore, this solution satisfies the bound

$$\begin{aligned}&\left|v^2(x,\tilde{x},\bar{x},\theta)\right| + \left|\nabla_{\bar{x}}v^2(x,\tilde{x},\bar{x},\theta)\right| + \left|\nabla_\theta v^2(x,\tilde{x},\bar{x},\theta)\right| + \left|\nabla_x v^2(x,\tilde{x},\bar{x},\theta)\right| + \left|\nabla_{\bar{x}}v^2(x,\tilde{x},\bar{x},\theta)\right| \\ \leq& C\left(1+|\bar{x}|\right)\left(1+|\tilde{x}|\right),\end{aligned} \tag{5.3.69}$$

where C is a constant independent of $(x,\tilde{x},\bar{x},\theta)$.

Proof. The proof is exactly the same as in Lemma 5.3.5 except for the presence of the dimension $\bar{x}$ and $\mathcal{L}_{\bar{x}}$. Since X_t^θ and $\bar{X}_t^\theta$ are i.i.d., the bounds from Proposition 5.3.4 are also true for $\bar{X}_t$. We first show that the integral (5.3.66) is finite. Note that

$$\begin{aligned}v^2(x,\tilde{x},\bar{x},\theta) &= \int_0^\infty \mathbb{E}\left[\left(\mathbb{E}_{\pi_\theta}f(Y) - f(\bar{X}_t^{\theta,\bar{x}})\right)\cdot\left(\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right)^\top\right]dt \\ &\overset{(a)}{=} \int_0^\infty \left(\mathbb{E}_{\pi_\theta}f(Y) - \mathbb{E}f(\bar{X}_t^{\theta,\bar{x}})\right)\cdot\mathbb{E}\left[\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right]^\top dt,\end{aligned} \tag{5.3.70}$$

where step (a) is due to the independence of $\bar{X}_\cdot^{\theta,\bar{x}}$ and $(X_\cdot^{\theta,x},\tilde{X}_\cdot^{\theta,x,\tilde{x}})$. As in (5.3.46) and (5.3.47), we can prove

$$\mathbb{E}\left|\tilde{X}_t^{\theta,x,\tilde{x}}\right|^2 \leq C\left(1+|\tilde{x}|^2\right)$$

and thus by Assumption A5.2.4

$$\left|\mathbb{E}\left[\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right| \leq C\mathbb{E}\left|\tilde{X}_t^{\theta,x,\tilde{x}}\right| \leq C\left(1+|\tilde{x}|\right), \tag{5.3.71}$$

which together with Proposition 5.3.4 yields

$$\left|v^2(x,\tilde{x},\bar{x},\theta)\right| \leq C\left(1+|\tilde{x}|\right)\cdot\int_0^\infty\left|\mathbb{E}f(\bar{X}_t^{\theta,\bar{x}}) - \mathbb{E}_{\pi_\theta}f(Y)\right|dt \leq C\left(1+|\tilde{x}|\right)\left(1+|\bar{x}|\right). \tag{5.3.72}$$

We next show that $v^2(x,\tilde{x},\bar{x},\theta)$ is differentiable with respect to $(x,\tilde{x},\bar{x},\theta)$. Similar to Lemma 5.3.5, we first change the order of differentiation and integration and show the corresponding integral exists. Then, we apply DCT to prove that the differentiation and integration can be interchanged. For the ergodic process $\bar{X}_\cdot^\theta$, by Proposition 5.3.4 and (5.3.71), we have the following bound for $i=1,2$:

$$\left|\nabla^i_{\bar{x}}v^2(x,\tilde{x},\bar{x},\theta)\right| \leq \int_0^\infty\left|\nabla^i_{\bar{x}}\mathbb{E}f(\bar{X}_t^{\theta,\bar{x}})\right|\cdot\left|\mathbb{E}\left[\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right|dt \leq C\left(1+|\tilde{x}|\right). \tag{5.3.73}$$

Note that

$$\nabla_\theta X_t^\theta = \tilde{X}_t^{\theta,x,0} \tag{5.3.74}$$

and thus, as in the proof of Proposition of 5.3.4 and Lemma 5.3.5, it is easy to prove the bounds

$$\begin{aligned}
&\sup_{\theta\in\mathbb{R}^\ell,x\in\mathbb{R}^d}\left|\nabla_x^i\tilde{X}_t^{\theta,x,\tilde{x}}\right|^2\le Ce^{-\beta t},\quad \sup_{\theta\in\mathbb{R}^\ell,x\in\mathbb{R}^d}\left|\nabla_{\tilde{x}}^i\tilde{X}_t^{\theta,x,\tilde{x}}\right|^2\le Ce^{-\beta t},\quad i=1,2,\\
&\sup_{\theta\in\mathbb{R}^\ell,x\in\mathbb{R}^d}\left|\nabla_\theta\tilde{X}_t^{\theta,x,\tilde{x}}\right|^2\le C,\quad \sup_{\theta\in\mathbb{R}^\ell,x\in\mathbb{R}^d}\left|\nabla_x\nabla_{\tilde{x}}\tilde{X}_t^{\theta,x,\tilde{x}}\right|^2\le Ce^{-\beta t},
\end{aligned}\tag{5.3.75}$$

which derives

$$\begin{aligned}
&\sum_{i=1}^2\left|\nabla_x^i\mathbb{E}\left[\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right|+\left|\nabla_\theta\mathbb{E}\left[\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right|\le C\left(1+|\tilde{x}|\right),\\
&\sum_{i=1}^2\left|\nabla_{\tilde{x}}^i\mathbb{E}\left[\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right|+\left|\nabla_x\nabla_{\tilde{x}}\mathbb{E}\left[\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right|\le C.
\end{aligned}\tag{5.3.76}$$

Therefore, for $i=1,2$,

$$\begin{aligned}
\left|\nabla_x^i v^2(x,\tilde{x},\bar{x},\theta)\right|&\le\int_0^\infty\left|\mathbb{E}_{\pi_\theta}f(Y)-\mathbb{E}f(\bar{X}_t^{\theta,\bar{x}})\right|\cdot\left|\nabla_x^i\mathbb{E}\left[\nabla f(X_t^\theta)\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right|dt\le C\left(1+|\bar{x}|\right)\left(1+|\tilde{x}|\right),\\
\left|\nabla_{\tilde{x}}^i v^2(x,\tilde{x},\bar{x},\theta)\right|&\le\int_0^\infty\left|\mathbb{E}_{\pi_\theta}f(Y)-\mathbb{E}f(\bar{X}_t^{\theta,\bar{x}})\right|\cdot\left|\nabla_{\tilde{x}}^i\mathbb{E}\left[\nabla f(X_t^\theta)\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right|dt\le C\left(1+|\bar{x}|\right)
\end{aligned}\tag{5.3.77}$$

and

$$\begin{aligned}
\left|\nabla_\theta v^2(x,\tilde{x},\bar{x},\theta)\right|&=\left|\int_0^\infty\nabla_\theta\left(\left[\mathbb{E}_{\pi_\theta}f(Y)-\mathbb{E}f(\bar{X}_t^{\theta,\bar{x}})\right]\cdot\mathbb{E}\left[\nabla f(X_t^\theta)\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right)dt\right|\\
&\le\left|\int_0^\infty\left[\nabla_\theta\mathbb{E}_{\pi_\theta}f(Y)-\nabla_\theta\mathbb{E}f(\bar{X}_t^{\theta,\bar{x}})\right]\cdot\mathbb{E}\left[\nabla f(X_t^\theta)\tilde{X}_t^{\theta,x,\tilde{x}}\right]dt\right|\\
&+\left|\int_0^\infty\left[\mathbb{E}_{\pi_\theta}f(Y)-\mathbb{E}f(\bar{X}_t^{\theta,\bar{x}})\right]\cdot\nabla_\theta\mathbb{E}\left[\nabla f(X_t^\theta)\tilde{X}_t^{\theta,x,\tilde{x}}\right]dt\right|\\
&\le C\left(1+|\bar{x}|\right)\left(1+|\tilde{x}|\right).
\end{aligned}\tag{5.3.78}$$

Finally,

$$\left|\nabla_x\nabla_{\tilde{x}}v^2(x,\tilde{x},\bar{x},\theta)\right|\le\int_0^\infty\left|\mathbb{E}_{\pi_\theta}f(Y)-\mathbb{E}f(\bar{X}_t^{\theta,\bar{x}})\right|\cdot\left|\nabla_x\nabla_{\tilde{x}}\mathbb{E}\left[\nabla f(X_t^{\theta,x})\tilde{X}_t^{\theta,x,\tilde{x}}\right]\right|dt\le C\left(1+|\bar{x}|\right).\tag{5.3.79}$$

By the same calculations as in Lemma 3.3 of [144], it can be shown that v^2 is the classical solution of PDE (5.3.67) and the bound (5.3.69) holds. □

Now we can bound the second fluctuation term Z_t^2. The proof is exactly the same as in Lemma 5.3.7.

Lemma 5.3.9. *Under Assumptions (A5.2.1) - (A5.2.5), for any fixed $\eta>0$,*

$$\left|\Delta^2_{\tau_n,\sigma_n+\eta}\right|\to 0,\ \text{as } n\to\infty,\quad a.s.\tag{5.3.80}$$

Proof. Consider the function

$$G^2(x,\tilde{x},\bar{x},\theta)=\left[f(\bar{x})-\mathbb{E}_{\pi_\theta}f(Y)\right]\left(\nabla f(x)\tilde{x}\right)^\top.\tag{5.3.81}$$

Let v^2 be the solution of (5.3.67) in Lemma 5.3.8. We apply Itô formula to the function $u^2(t,x,\tilde{x},\bar{x},\theta)=$

$\alpha_t v^2(x,\tilde{x},\bar{x},\theta)$ evaluated on the stochastic process $(X_t,\tilde{X}_t,\bar{X}_t,\theta_t)$ and get for any $i\in\{1,2,\cdots,\ell\}$

$$\begin{aligned}
&u_i^2\left(\sigma,X_\sigma,\tilde{X}_\sigma,\bar{X}_\sigma,\theta_\sigma\right)-u_i^2\left(\tau,X_\tau,\tilde{X}_\tau,\bar{X}_\tau,\theta_\tau\right)\\
&=\int_\tau^\sigma \partial_s u_i^2\left(s,X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)ds+\int_\tau^\sigma \mathcal{L}_{x,\tilde{x}}^{\theta_s}u_i^2\left(s,X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)ds\\
&+\int_\tau^\sigma \mathcal{L}_{\bar{x}}^{\theta_s}u_i^2\left(s,X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)ds+\int_\tau^\sigma \nabla_\theta u_i^2\left(s,X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)d\theta_s\\
&+\int_\tau^\sigma \nabla_x u_i^2\left(s,X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)dW_s+\int_\tau^\sigma \nabla_{\bar{x}} u_i^2\left(s,X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)d\bar{W}_s\\
&+\sum_{k=1}^{\ell}\int_\tau^\sigma \nabla_{\tilde{x}^{:,k}}u_i^2\left(s,X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)\left(\nabla_x\sigma(X_s,\theta_s)\tilde{X}_s^{:,k}+\frac{\partial\sigma(X_s,\theta_s)}{\partial\theta_k}\right)dW_s.
\end{aligned}\tag{5.3.82}$$

Rearranging the previous equation, we obtain the representation

$$\begin{aligned}
\Delta^2_{\tau_n,\sigma_n+\eta}=&\int_{\tau_n}^{\sigma_n+\eta}\alpha_s G^2(X_s,\tilde{X}_s,\bar{X}_s,\theta_s)ds\\
=&\int_{\tau_n}^{\sigma_n+\eta}\mathcal{L}_{x,\tilde{x},\bar{x}}^{\theta_s}u^2\left(s,X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)ds\\
=&\alpha_{\sigma_n+\eta}v^2\left(X_{\sigma_n+\eta},\tilde{X}_{\sigma_n+\eta},\bar{X}_{\sigma_n+\eta},\theta_{\sigma_n+\eta}\right)-\alpha_{\tau_n}v^2\left(X_{\tau_n},\tilde{X}_{\tau_n},\bar{X}_{\tau_n},\theta_{\tau_n}\right)\\
&-\int_{\tau_n}^{\sigma_n+\eta}\alpha_s' v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)ds-\int_{\tau_n}^{\sigma_n+\eta}\alpha_s\nabla_{\bar{x}}v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)d\bar{W}_s\\
&+\int_{\tau_n}^{\sigma_n+\eta}2\alpha_s^2\nabla_\theta v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)\left(f(\bar{X}_s)-\beta\right)\left(\nabla f(X_s)\tilde{X}_s\right)^\top ds\\
&-\int_{\tau_n}^{\sigma_n+\eta}\alpha_s\nabla_x v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)dW_s\\
&-\sum_{k=1}^{\ell}\int_{\tau_n}^{\sigma_n+\eta}\alpha_s\nabla_{\tilde{x}^{:,k}}v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)\left(\nabla_x\sigma(X_s,\theta_s)\tilde{X}_s^{:,k}+\frac{\partial\sigma(X_s,\theta_s)}{\partial\theta_k}\right)dW_s.
\end{aligned}\tag{5.3.83}$$

The next step is to treat each term on the right hand side of (5.3.83) separately. For this purpose, let us first set

$$J_t^{2,1}=\alpha_t\sup_{s\in[0,t]}\left|v^2\left(X_s,\tilde{X}_s,\bar{X}_s,\theta_s\right)\right|.\tag{5.3.84}$$

Combining Lemma 5.3.6, (5.3.69) and (5.3.62), we know that there exists a constant C such that

$$\begin{aligned}
\mathbb{E}\left|J_t^{2,1}\right|^2&\le C\alpha_t^2\mathbb{E}\left[1+\sup_{s\in[0,t]}\left|\tilde{X}_s\right|^4+\sup_{s\in[0,t]}\left|\bar{X}_s\right|^4\right]\\
&=C\alpha_t^2\left[1+\sqrt{t}\frac{\mathbb{E}\sup_{s\in[0,t]}\left|\tilde{X}_s\right|^4+\mathbb{E}\sup_{s\in[0,t]}\left|\bar{X}_s\right|^4}{\sqrt{t}}\right]\\
&\le C\alpha_t^2\sqrt{t}.
\end{aligned}\tag{5.3.85}$$

Let $p>0$ be the constant in Assumption A5.2.5 such that $\lim_{t\to\infty}\alpha_t^2 t^{1/2+2p}=0$ and for any $\delta\in(0,p)$ define the event $A_{t,\delta}=\left\{J_t^{2,1}\ge t^{\delta-p}\right\}$. Then we have for t large enough such that $\alpha_t^2 t^{1/2+2p}\le 1$ and

$$\mathbb{P}\left(A_{t,\delta}\right)\le\frac{\mathbb{E}\left|J_t^{2,1}\right|^2}{t^{2(\delta-p)}}\le C\frac{\alpha_t^2 t^{1/2+2p}}{t^{2\delta}}\le C\frac{1}{t^{2\delta}}.$$

The latter implies that

$$\sum_{m\in\mathbb{N}} \mathbb{P}\left(A_{2^m,\delta}\right) < \infty.$$

Therefore, by the Borel-Cantelli lemma we have that for every $\delta \in (0,p)$ there is a finite positive random variable $d(\omega)$ and some $m_0 < \infty$ such that for every $n \geq m_0$ one has

$$J^{2,1}_{2^n} \leq \frac{d(\omega)}{2^{m(p-\delta)}}.$$

Thus for $t \in \left[2^m, 2^{m+1}\right)$ and $m \geq m_0$ one has for some finite constant $C < \infty$

$$J^{2,1}_t \leq C\alpha_{2^{m+1}} \sup_{s\in(0,2^{m+1}]} \left|v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right)\right| \leq C\frac{d(\omega)}{2^{(m+1)(p-\delta)}} \leq C\frac{d(\omega)}{t^{p-\delta}},$$

which derives that for $t \geq 2^{m_0}$ we have with probability one

$$J^{2,1}_t \leq C\frac{d(\omega)}{t^{p-\delta}} \to 0, \text{ as } t \to \infty. \tag{5.3.86}$$

Next we consider the term

$$J^{2,2}_{t,0} = \int_0^t \left|\alpha'_s v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) - 2\alpha_s^2 \nabla_\theta v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right)\left(f(\bar{X}_s) - \beta\right)\left(\nabla f(X_s)\tilde{X}_s\right)^\top\right| ds$$

and thus we see that there exists a constant $0 < C < \infty$ such that

$$\begin{aligned}
\sup_{t>0} \mathbb{E}\left|J^{2,2}_{t,0}\right| &\overset{(a)}{\leq} C\int_0^\infty \left(|\alpha'_s| + \alpha_s^2\right)\left(1 + \mathbb{E}\left|\bar{X}_s\right|^4 + \mathbb{E}\left|\bar{X}_s\right|^4\right) ds \\
&\overset{(b)}{\leq} C\int_0^\infty \left(|\alpha'_s| + \alpha_s^2\right) ds \\
&\leq C,
\end{aligned}$$

where in step (a) we use (5.3.69) and in step (b) we use Lemma 5.3.6 and (5.3.62). Thus we know there is a finite random variable $J^{2,2}_{\infty,0}$ such that

$$J^{2,2}_{t,0} \to J^{2,2}_{\infty,0}, \text{ as } t \to \infty \text{ with probability one.} \tag{5.3.87}$$

The last term we need to consider is the martingale term

$$\begin{aligned}
J^{2,3}_{t,0} = &\int_0^t \alpha_s \nabla_x v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) dW_s + \int_0^t \alpha_s \nabla_{\bar{x}} v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right) d\bar{W}_s \\
&+ \sum_{k=1}^{\ell} \int_0^t \alpha_s \nabla_{\tilde{x}^{:,k}} v^2\left(X_s, \tilde{X}_s, \bar{X}_s, \theta_s\right)\left(\nabla_x \sigma(X_s, \theta_s)\tilde{X}^{:,k}_s + \frac{\partial \sigma(X_s, \theta_s)}{\partial \theta_k}\right) dW_s.
\end{aligned}$$

Notice that Doob's inequality and the bounds of (5.3.69) (using calculations similar to the ones for the term $J^{2,2}_{t,0}$) give us that for some finite constant $K < \infty$, we have

$$\sup_{t>0} \mathbb{E}\left|J^{2,3}_{t,0}\right|^2 \leq K\int_0^\infty \alpha_s^2 ds < \infty.$$

Thus, by Doob's martingale convergence theorem there is a square integrable random variable $J^{(3)}_{\infty,0}$ such that

$$J^{2,3}_{t,0} \to J^{2,3}_{\infty,0}, \quad \text{as } t \to \infty \text{ both almost surely and in } L^2. \tag{5.3.88}$$

Let us now go back to (5.3.83). Using the terms $J^{2,1}_t$, $J^{2,2}_{t,0}$ and $J^{2,3}_{t,0}$ we can write

$$\left|\Delta^2_{\tau_n,\sigma_n+\eta}\right| \leq J^{2,1}_{\sigma_n+\eta} + J^{2,1}_{\tau_n} + J^{2,2}_{\sigma_n+\eta,\tau_n} + \left|J^{2,3}_{\sigma_k+\eta,\tau_n}\right|,$$

which together with (5.3.86), (5.3.87) and (5.3.88) prove the statement of the Lemma. □

By (5.3.10), we know that

$$|\nabla_\theta J(\theta)| = 2\,|\mathbb{E}_{\pi_\theta} f(Y) - \beta| \cdot |\nabla_\theta \mathbb{E}_{\pi_\theta} f(Y)| \leq C. \tag{5.3.89}$$

Therefore, the objective function $J(\theta)$ is Lipschitz continuous with respect to θ. The following lemmas are the same as in [144] and thus we omit the proofs.

Lemma 5.3.10. *Under Assumptions (A5.2.1)-(A5.2.5), choose $\mu > 0$ in (5.3.2) such that for the given $\kappa > 0$, one has $3\mu + \frac{\mu}{8\kappa} = \frac{1}{2L_{\nabla J}}$, where $L_{\nabla J}$ is the Lipschitz constant of objective function J in (5.1.2). Then for n large enough and $\eta > 0$ small enough (potentially random depending on n), one has $\int_{\tau_n}^{\sigma_n+\eta} \alpha_s ds > \mu$. In addition we also have $\frac{\mu}{2} \leq \int_{\tau_n}^{\sigma_n} \alpha_s ds \leq \mu$ with probability one.*

Lemma 5.3.11. *Under Assumptions (A5.2.1)-(A5.2.5), suppose that there exists an infinite number of intervals $I_n = [\tau_n, \sigma_n)$. Then there is a fixed constant $\gamma_1 = \gamma_1(\kappa) > 0$ such that for n large enough,*

$$J(\theta_{\sigma_n}) - J(\theta_{\tau_n}) \leq -\gamma_1. \tag{5.3.90}$$

Lemma 5.3.12. *Under Assumptions (A5.2.1)-(A5.2.5), suppose that there exists an infinite number of intervals $I_n = [\tau_n, \sigma_n)$. Then, there is a fixed constant $\gamma_2 < \gamma_1$ such that for n large enough,*

$$J(\theta_{\tau_n}) - J(\theta_{\sigma_{n-1}}) \leq \gamma_2. \tag{5.3.91}$$

Proof of Theorem 5.2.6: Recalling (5.3.2), we know τ_n is the first time $|\nabla_\theta J(\theta_t)| > \kappa$ when $t > \sigma_{n-1}$. Thus, for any fixed $\kappa > 0$, if there are only a finite number of τ_n, then there is a finite T^* such that $|\nabla_\theta J(\theta_t)| \leq \kappa$ for $t \geq T^*$. We now use a "proof by contradiction". Suppose that there are infinitely many instances of τ_n. By Lemmas 5.3.11 and 5.3.12, we have for sufficiently large n that

$$J\left(\theta_{\sigma_n}\right) - J\left(\theta_{\tau_n}\right) \leq -\gamma_1$$

$$J\left(\theta_{\tau_n}\right) - J\left(\theta_{\sigma_{n-1}}\right) \leq \gamma_2,$$

where $0 < \gamma_2 < \gamma_1$. Choose N large enough so that the above relations hold simultaneously for $n \geq N$. Then,

$$\begin{aligned} J\left(\theta_{\tau_{m+1}}\right) - J\left(\theta_{\tau_N}\right) &= \sum_{n=N}^{m} \left[J\left(\theta_{\sigma_n}\right) - J\left(\theta_{\tau_n}\right) + J\left(\theta_{\tau_{n+1}}\right) - J\left(\theta_{\sigma_n}\right)\right] \\ &\leq \sum_{k=N}^{n} \left(-\gamma_1 + \gamma_2\right) \\ &< (m - N) \times \left(-\gamma_1 + \gamma_2\right). \end{aligned} \tag{5.3.92}$$

Letting $m \to \infty$, we observe that $J\left(\theta_{\tau_m}\right) \to -\infty$, which is a contradiction, since by definition $J(\theta_t) \geq 0$. Thus, there can be at most finitely many τ_n. Thus, there exists a finite time T such that almost surely $|\nabla_\theta J(\theta_t)| < \kappa$ for $t \geq T$. Since κ is arbitrarily chosen, we have proven that $|\nabla_\theta J(\theta_t)| \to 0$ as $t \to \infty$ almost surely.

□

5.4 Numerical Example

In this section, we use our algorithm in the parameter estimation for SPDEs. Consider the stochastic Burger's equation:

$$\frac{\partial u}{\partial t}(t,x;\theta) = \theta \frac{\partial^2 u}{\partial x^2}(t,x;\theta) - F - u(t,x;\theta)\frac{\partial u}{\partial x}(t,x;\theta) + \sigma \dot{W}(t), \tag{5.4.1}$$

where $x \in [0,1]$ and $\dot{W}(t)$ is the white noise [98]. The finite difference discretization of (5.4.1) satisfies a system of nonlinear stochastic differential equations ([52, 73]). We define its time average to be

$$\hat{u}(x,\theta) := \lim_{T\to+\infty} \frac{1}{T}\int_0^T u(t,x;\theta)dt. \tag{5.4.2}$$

And for a fixed target function $h(\cdot)$, the objective function to learn the parameter is

$$J(\theta) := \int_0^1 \left[\hat{u}(x,\theta) - h(x)\right]^2 dx \tag{5.4.3}$$

We discretize the spatial to be $(x_0, x_1, \cdots, x_N)$ where $x_0 = 0$, $x_N = 1$ and then the objective function is

$$J(\theta) = \sum_{i=0} \left[\hat{u}(x_i;\theta) - h(x_i)\right]^2. \tag{5.4.4}$$

Take the gradient,

$$\begin{aligned} \nabla_\theta J(\theta) &= \sum_{i=0} \nabla_\theta \left(\lim_{T\to+\infty} \frac{1}{T}\int_0^T u(t,x_i;\theta)dt - h(x_i) \right)^2 \\ &= 2\sum_{i=0} \left(\lim_{T\to+\infty} \frac{1}{T}\int_0^T u(t,x_i;\theta)dt - h(x_i) \right) \cdot \lim_{T\to+\infty} \frac{1}{T}\int_0^T \nabla_\theta u(t,x_i;\theta)dt. \end{aligned} \tag{5.4.5}$$

Then the coupled system (5.1.3) is

$$\begin{aligned} d\theta_t &= -\alpha_t \sum_{i=0}^N (\bar{u}(t,x_i;\theta_t) - h(x_i)) \cdot \tilde{u}(t,x_i;\theta_t)dt \\ du(t,x;\theta_t) &= \left(\theta_t \frac{\partial^2 u}{\partial x^2}(t,x;\theta_t) - F - u(t,x;\theta_t)\frac{\partial u}{\partial x}(t,x;\theta_t) \right) dt + \sigma dW(t), \\ d\bar{u}(t,x;\theta_t) &= \left(\theta_t \frac{\partial^2 \bar{u}}{\partial x^2}(t,x;\theta_t) - F - \bar{u}(t,x;\theta)\frac{\partial \bar{u}}{\partial x}(t,x;\theta_t) \right) dt + \sigma d\bar{W}(t), \\ d\tilde{u}(t,x;\theta_t) &= \left(\frac{\partial^2 u}{\partial x^2}(t,x;\theta_t) + \theta_t \frac{\partial^2 \tilde{u}}{\partial x^2}(t,x;\theta_t) - \tilde{u}(t,x;\theta_t)\frac{\partial u}{\partial x}(t,x;\theta_t) - u(t,x;\theta_t)\frac{\partial \tilde{u}}{\partial x}(t,x;\theta_t) \right) dt. \end{aligned}$$

For our numerical experiment, we use a spatial discretization of $\Delta x = 0.01$ and the following finite difference scheme for Burger's equation

$$\begin{aligned} du(t,x_i;\theta) = \theta &\frac{u(t,x_{i+1};\theta) - 2u(t,x_i;\theta) + u(t,x_{i-1};\theta)}{\Delta x^2}dt - F \\ &- u(t,x_i;\theta)\frac{u(t,x_{i+1};\theta) - u(t,x_{i-1};\theta)}{2\Delta x}dt + \sigma dW_t. \end{aligned} \tag{5.4.6}$$

(5.4.6) is simulated with the Euler scheme with a time step of 10^{-5} (i.e., we solve Burger's equation with explicit finite difference) and the small time step is needed to avoid instability in the scheme. We fix $\theta^* = 1$ and then simulate for long enough time T to make sure the convergence and get the

target function

$$u^*(x) = \frac{1}{T}\int_0^T u(t, x; \theta^*)dt, \tag{5.4.7}$$

which is plugged into the objective function (5.4.4). The numerical result of the system (5.4.5) are shown in figure

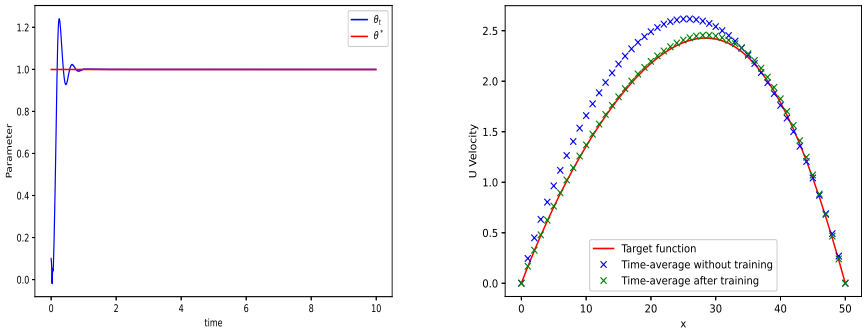

Figure 5.1: target function before and after training (left) and trained parameters (right).

Chapter 6

Conclusion and Future Research Direction

This thesis investigates the optimization of parameters in the stationary distribution of stochastic dynamics. In Chapter 2 and Chapter 3, we study online Actor-Critic algorithms in Reinforcement Learning with both tabular and neural network parameterization. We establish the convergence of these online algorithms to the limit ODEs for both parameterization methods. For the tabular case, we provide results on convergence and convergence rates to the optimal strategies. In the neural network case, we establish convergence to a stationary point of the limit actor network.

In Chapter 4 and Chapter 5, we propose a new online learning algorithm for computationally efficient optimization over the stationary distribution of ergodic SDEs. Specifically, the online forward propagation algorithm optimizes parameterized SDEs to minimize the distance between their stationary distribution and target statistics. By proving bounds for a new class of Poisson PDEs, we analyze parameter fluctuations during training and rigorously prove convergence to a stationary point for multi-dimensional Ornstein-Uhlenbeck processes and nonlinear dissipative SDE models. Additionally, we present the numerical performance of our algorithm across various mathematical finance applications, including statistical calibration of SDE and SPDE models, high-dimensional stochastic optimal control over long time horizons, and training stochastic point process models for the limit order book.

Several future research directions are worth exploring. Theoretically, we have only proven convergence to the stationary point of the limit actor network in Chapter 3. Investigating the global convergence and convergence rates of the online neural Actor-Critic algorithm would be valuable. Additionally, studying the convergence of our online algorithm in Chapter 4 for discrete-time stochastic processes is an interesting area for future research. Numerically, we could apply our algorithm to estimate parameters in partially observed systems by optimizing the asymptotic likelihood or solving nonlinear ergodic stochastic control problems. It would also be intriguing to compare the numerical performance (convergence speed) of our online algorithm with the standard backpropagation algorithm.

Appendix A

Addtional Proofs for Chapter 2

A.1 Verification of (2.3.2)

$$\begin{aligned}\int_0^\infty \zeta_s \eta_s ds &= \int_0^2 \zeta_s \eta_s ds + \int_2^\infty \zeta_s \eta_s ds \\ &\leq C + \int_2^\infty \frac{1}{t \log^2 t} dt \\ &= C - \frac{1}{\log t}\Big|_2^\infty < \infty, \\ \lim_{t\to\infty} \frac{\zeta_t}{\eta_t^n} &= \lim_{t\to\infty} \frac{\log^{2n} t}{t} \overset{(a)}{=} 0\end{aligned} \tag{A.1.1}$$

where step (a) is by L'Hospital's Rule.

A.2 Proof of Corollary 2.4.5

Proof. Recall the exploration policy in (2.2.9) with the decreasing exploration rate η_k^N. Then, we have for $\forall k \leq NT$,

$$g_k(x,a) \geq \frac{\eta^N_{\lfloor NT \rfloor}}{d_A}, \quad \forall x, a \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}. \tag{A.2.1}$$

Then, for any ξ, ξ' and $k \leq NT$, with the constant C from (2.4.32),

$$\begin{aligned}\mathbb{P}^{n_0}_{\theta_k}(\xi;\xi') &= \sum_{\xi_1,\cdots,\xi_{n_0-1}} \mathbb{P}_{\theta_k}(\xi;\xi_1)\cdots\mathbb{P}_{\theta_k}(\xi_{n_0-1};\xi') \\ &= \sum_{\xi_1,\cdots,\xi_{n_0-1}} p(x_1|x,a)g_k(x_1,a_1)\cdots p(x'|x_{n_0-1},a_{n_0-1})g_k(x',a') \\ &\geq C\left(\eta^N_{\lfloor NT \rfloor}\right)^{n_0}.\end{aligned} \tag{A.2.2}$$

Thus, we can derive a lower bound for the stationary distribution

$$\begin{aligned}\inf_{k\leq NT} \pi^{g_k}(x',a') &= \inf_{k\leq NT} \sum_{x,a} \pi^{g_k}(x,a)\mathbb{P}^{n_0}_{\theta_k}(x,a;x',a') \\ &\geq \inf_{k\leq NT} \sum_{x,a} \pi^{g_k}(x,a) C\left(\eta^N_{\lfloor NT \rfloor}\right)^{n_0} \\ &\overset{(a)}{=} C\left(\eta^N_{\lfloor NT \rfloor}\right)^{n_0} \\ &> 0,\end{aligned} \tag{A.2.3}$$

where the step (a) is because π^{g_k} is a probability and thus the summation equals 1. For the uniform geometric ergodicity, we can choose $\beta_T = \inf_{k\le NT}\min_{\xi,\xi'} \mathbb{P}_{\theta_k}^{n_0}(\xi,\xi') > 0$ in (2.4.34), where $\beta_T > 0$ is by (A.2.2). Thus for $\forall k \le NT$, the Markov chain with transition probability $\mathbb{P}_{\theta_k}$ satisfies the Doeblin's condition, then by Theorem 16.2.4 of [106], we can derive the uniform geometric ergodicity (2.4.34). □

A.3 Proof of Lemma 2.4.10

Proof. As in the proof for the decay of M_t^N, we use two steps to prove the result.

(i) Prove that the fluctuations of the data samples around a dynamic stationary distribution π^{g_k} decay when the number of iteration steps becomes large.

(ii) Use the same method as in Lemma 2.4.9 to prove the stochastic fluctuation terms vanish as $N \to \infty$.

(i) To prove that for any fixed state action pair $\xi = (x,a), \forall T > 0$

$$\lim_{N\to 0} \mathbb{E}\left| \frac{1}{N} \sum_{k=0}^{\lfloor NT\rfloor -1} \left[\mathbb{1}_{\{\xi_k=\xi\}} - \pi^{g_k}(\xi)\right]\right| = 0, \tag{A.3.1}$$

we first introduce a similar Poisson equation for any fixed state-action pair $\xi = (x,a)$, $N \in \mathbb{N}$, $T < \infty$ and $k \le NT$,

$$\bar{\nu}_{\theta_k}(\xi') - \mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi') = \mathbb{1}_{\{\xi'=\xi\}} - \pi^{g_k}(\xi), \quad \xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}. \tag{A.3.2}$$

A solution of (A.3.2) can be expressed as

$$\bar{\nu}_{\theta_k}(\xi') := \sum_{n\ge 0} \left[\mathbb{P}_{\theta_k}^n(\xi';\xi) - \pi^{g_k}(\xi)\right]. \tag{A.3.3}$$

By Corollary 2.4.5, there exists a constant C_T (which only depends on T) such that

$$\sup_{k\le NT} |\bar{\nu}_{\theta_k}(\xi')| \le C_T, \quad \forall \xi' \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{A}}. \tag{A.3.4}$$

Then, as in the proof of Lemma 2.4.8, we define the error $\bar{\epsilon}_k$ as

$$\begin{aligned} \bar{\epsilon}_k :=& \mathbb{1}_{\{\xi_{k+1}=\xi\}} - \pi^{g_k}(\xi) \\ =& \bar{\nu}_{\theta_k}(\xi_{k+1}) - \mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_{k+1}) \\ =& \left[\bar{\nu}_{\theta_k}(\xi_{k+1}) - \mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_k)\right] + \left[\mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_k) - \mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_{k+1})\right]. \end{aligned} \tag{A.3.5}$$

Let

$$\bar{\psi}_\theta(y) = \mathbb{P}_\theta \bar{\nu}_\theta(y). \tag{A.3.6}$$

Then, we have

$$\begin{aligned}\sum_{k=0}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k &= \sum_{k=0}^{\lfloor NT\rfloor-1}\left[\bar{\nu}_{\theta_k}(\xi_{k+1})-\mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_k)\right]+\sum_{k=0}^{\lfloor NT\rfloor-1}\left[\bar{\psi}_{\theta_k}(\xi_k)-\bar{\psi}_{\theta_k}(\xi_{k+1})\right]\\ &= \sum_{k=0}^{\lfloor NT\rfloor-1}\left[\bar{\nu}_{\theta_k}(\xi_{k+1})-\mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_k)\right]+\sum_{k=1}^{\lfloor NT\rfloor-1}\left[\bar{\psi}_{\theta_k}(\xi_k)-\bar{\psi}_{\theta_{k-1}}(\xi_k)\right]\\ &\quad+\bar{\psi}_{\theta_0}(\xi_0)-\bar{\psi}_{\theta_{\lfloor NT\rfloor-1}}\left(\xi_{\lfloor NT\rfloor}\right)\end{aligned} \tag{A.3.7}$$

Define the error term as

$$\sum_{k=0}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k=\sum_{k=0}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k^{(1)}+\sum_{k=1}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k^{(2)}+\bar{\rho}_{\lfloor NT\rfloor;0} \tag{A.3.8}$$

where

$$\begin{aligned}\bar{\epsilon}_k^{(1)}&=\bar{\nu}_{\theta_k}(\xi_{k+1})-\mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_k)\\ \bar{\epsilon}_k^{(2)}&=\bar{\psi}_{\theta_k}(\xi_k)-\bar{\psi}_{\theta_{k-1}}(\xi_k)\\ \bar{\rho}_{\lfloor NT\rfloor;0}&=\bar{\psi}_{\theta_0}(\xi_0)-\bar{\psi}_{\theta_{\lfloor NT\rfloor-1}}\left(\xi_{\lfloor NT\rfloor}\right).\end{aligned} \tag{A.3.9}$$

To prove the convergence (A.3.1), it suffices to appropriately bound the fluctuation term $\left|\sum_{k=0}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k\right|$. The first term can be bounded using the martingale property while the second term can be bounded using the uniform geometric ergodicity and Lipschitz continuity. The third term is bounded using (A.3.4).

For the first term in (A.3.8), note that

$$\mathbb{E}\left\{\bar{\nu}_{\theta_k}(\xi_{k+1})\mid\mathscr{F}_k\right\}=\mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_k). \tag{A.3.10}$$

Therefore,

$$\left\{\bar{Z}_n=\sum_{k=0}^{n-1}\bar{\epsilon}_k^{(1)},\ \mathscr{F}_n\right\}_{n\geq 0}$$

is a martingale and since the conditional expectation is a contraction in L^2, we have

$$\mathbb{E}\left|\mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_k)\right|^2\leq\mathbb{E}\left|\bar{\nu}_{\theta_k}(\xi_{k+1})\right|^2. \tag{A.3.11}$$

Then,

$$\begin{aligned}\mathbb{E}\left|\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k^{(1)}\right|^2&=\frac{1}{N^2}\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbb{E}\left|\bar{\nu}_{\theta_k}(\xi_{k+1})-\mathbb{P}_{\theta_k}\bar{\nu}_{\theta_k}(\xi_k)\right|^2\\ &\leq\frac{4}{N^2}\sum_{k=0}^{\lfloor NT\rfloor-1}\mathbb{E}\left|\bar{\nu}_{\theta_k}(\xi_{k+1})\right|^2\\ &\overset{(a)}{\leq}\frac{4C_T}{N},\end{aligned} \tag{A.3.12}$$

where the step (a) is by the uniform boundedness (A.3.4). Thus we have for any $T>0$

$$\lim_{N\to\infty}\mathbb{E}\left|\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k^{(1)}\right|=0. \tag{A.3.13}$$

For the second term of (A.3.8), by the uniform geometric ergodicity (2.4.34), for any fixed $\gamma_0>0$

we can choose N_0 large enough such that

$$\sup_{k\le NT}\sum_{n=\lfloor N_0T\rfloor}^{\infty}\left|\mathbb{P}^n_{\theta_k}(y,\xi)-\pi^{g_k}(\xi)\right|<\gamma_0,\quad \forall y\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}} \tag{A.3.14}$$

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k^{(2)}\right|\\
=&\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\left[\bar{\psi}_{\theta_k}(\xi_k)-\bar{\psi}_{\theta_{k-1}}(\xi_k)\right]\right|\\
\le&\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\left[\sum_{n=1}^{\lfloor N_0T\rfloor-1}\left[\mathbb{P}^n_{\theta_k}(\xi_k,\xi)-\pi^{g_k}(\xi)\right]-\sum_{n=1}^{\lfloor N_0T\rfloor-1}\left[\mathbb{P}^n_{\theta_{k-1}}(\xi_k,\xi)-\pi^{g_{k-1}}(\xi)\right]\right]\right|+2C_T\gamma_0\\
=&\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\sum_{n=1}^{\lfloor N_0T\rfloor-1}\left[\mathbb{P}^n_{\theta_k}(\xi_k,\xi)-\mathbb{P}^n_{\theta_{k-1}}(\xi_k,\xi)\right]\right|+\frac{\lfloor N_0T\rfloor}{N}\left|\sum_{k=1}^{\lfloor NT\rfloor-1}\left[\pi^{g_k}(\xi)-\pi^{g_{k-1}}(\xi)\right]\right|+2C_T\gamma_0\\
:=&\bar{I}_1^N+\bar{I}_2^N+2C_T\gamma_0.
\end{aligned} \tag{A.3.15}$$

With the exploration policy g_k in (2.2.9) and Lipschitz continuity in Assumption 2.3.2, we have

$$\|g_k-g_{k-1}\|\le\sum_{x,a\in\mathcal{X}\times\mathcal{A}}|g_k(x,a)-g_{k-1}(x,a)|\le C\left|\eta_k^N-\eta_{k-1}^N\right|+C\|\theta_k-\theta_{k-1}\|. \tag{A.3.16}$$

For any finite n,

$$\mathbb{P}^n_{\theta_k}(\xi;\xi')=\sum_{\xi_1,\cdots,\xi_{n-1}}p(x_1|x,a)g_k(x_1,a_1)\cdots p(x'|x_{n-1},a_{n-1})g_k(x',a'),\quad\forall\xi,\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}. \tag{A.3.17}$$

is Lipschitz continuous in the policy g_k. Then, there exists a constant C_T which only depends on the fixed N_0, T such that

$$\begin{aligned}
\bar{I}_1^N&\le\frac{\lfloor N_0T\rfloor}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}C\|g_k-g_{k-1}\|\le\frac{C_T}{N}\left[\eta_0^N+\sum_{k=1}^{\lfloor NT\rfloor-1}\|\theta_k-\theta_{k-1}\|\right]\overset{(a)}{\le}\frac{C_T}{N},\\
\bar{I}_2^N&\le\frac{\lfloor N_0T\rfloor}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}C\|g_k-g_{k-1}\|\le\frac{C_T}{N}\left[\eta_0^N+\sum_{k=1}^{\lfloor NT\rfloor-1}\|\theta_k-\theta_{k-1}\|\right]\overset{(a)}{\le}\frac{C_T}{N},
\end{aligned} \tag{A.3.18}$$

where step (a) is due to Lemma 2.4.1:

$$\|\theta_k-\theta_{k-1}\|\le\frac{C_T}{N},\quad\forall k\le NT.$$

Thus, when N is large enough,

$$\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k^{(2)}\right|\le 4C_T\gamma_0 \tag{A.3.19}$$

Since γ_0 is arbitrary,

$$\lim_{N\to\infty}\mathbf{E}\left|\frac{1}{N}\sum_{k=1}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k^{(2)}\right|=0 \tag{A.3.20}$$

Obviously, for the last term of (A.3.8) by the bound in (A.3.4) we have

$$\lim_{N\to\infty}\frac{1}{N}\bar{\rho}_{\lfloor NT\rfloor;0}=0,$$

which together with (A.3.13) and (A.3.20) derive the convergence of $\frac{1}{N}\sum_{k=0}^{\lfloor NT\rfloor-1}\bar{\epsilon}_k$ and (A.3.1).

(ii) Following the same method in Lemma 2.4.9, we can prove the convergence of the stochastic error $M_t^{i,N}$ for $i=1,2,3$.

For any $K\in\mathbb{N}$ and $\Delta=\frac{t}{K}$, we have

$$
\begin{aligned}
&-M_t^{1,N}(\xi)\\
&=\sum_{j=0}^{K-1}\Delta\frac{1}{\lfloor\Delta N\rfloor}\sum_{k=j\lfloor\Delta N\rfloor}^{(j+1)\lfloor\Delta N\rfloor-1}\left(Q_k(\xi_k)\partial_\xi Q_k(\xi_k)-\sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}Q_k(\xi')\partial_\xi Q_k(\xi')\pi^{g_k}(\xi')\right)+o(1)\\
&=\sum_{j=0}^{K-1}\Delta\frac{1}{\lfloor\Delta N\rfloor}\sum_{k=j\lfloor\Delta N\rfloor}^{(j+1)\lfloor\Delta N\rfloor-1}\left(Q_{j\lfloor\Delta N\rfloor}(\xi_k)\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi_k)-\sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}Q_{j\lfloor\Delta N\rfloor}(\xi')\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi')\pi^{g_k}(\xi')\right)\\
&+\sum_{j=0}^{K-1}\Delta\frac{1}{\lfloor\Delta N\rfloor}\sum_{k=j\lfloor\Delta N\rfloor}^{(j+1)\lfloor\Delta N\rfloor-1}\left[\left(Q_k(\xi_k)\partial_\xi Q_k(\xi_k)-\sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}Q_k(\xi')\partial_\xi Q_k(\xi')\pi^{g_k}(\xi')\right)\right.\\
&\left.-\left(Q_{j\lfloor\Delta N\rfloor}(\xi_k)\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi_k)-\sum_{\xi'\in\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{A}}}Q_{j\lfloor\Delta N\rfloor}(\xi')\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi')\pi^{g_k}(\xi')\right)\right]+o(1)\\
&:=\sum_{j=0}^{K-1}\Delta I_{5,j}^N+\sum_{j=0}^{K-1}\Delta I_{6,j}^N+o(1),
\end{aligned}
\tag{A.3.21}
$$

where the term $o(1)$ goes to zero, at least, in L^1 as $N\to\infty$.

To prove the convergence of the first term, note that

$$
\begin{aligned}
&Q_{j\lfloor\Delta N\rfloor}(\xi_k)\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi_k)-\sum_{\xi'}Q_{j\lfloor\Delta N\rfloor}(\xi')\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi')\pi^{g_k}(\xi')\\
&=\sum_{\xi'}Q_{j\lfloor\Delta N\rfloor}(\xi')\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi')\mathbb{1}_{\{\xi_k=\xi'\}}-\sum_{\xi'}Q_{j\lfloor\Delta N\rfloor}(\xi')\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi')\pi^{g_k}(\xi')\\
&=\sum_{\xi'}Q_{j\lfloor\Delta N\rfloor}(\xi')\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi')\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k}(\xi')\right].
\end{aligned}
\tag{A.3.22}
$$

Thus, for any $j\in 0,1,\dots,K$,

$$
\begin{aligned}
\left|I_{5,j}^N\right|&=\left|\frac{1}{\lfloor\Delta N\rfloor}\sum_{k=j\lfloor\Delta N\rfloor}^{(j+1)\lfloor\Delta N\rfloor-1}\sum_{\xi'}Q_{j\lfloor\Delta N\rfloor}(\xi')\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi')\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k}(\xi')\right]\right|\\
&=\left|\sum_{\xi'}Q_{j\lfloor\Delta N\rfloor}(\xi')\partial_\xi Q_{j\lfloor\Delta N\rfloor}(\xi')\frac{1}{\lfloor\Delta N\rfloor}\sum_{k=j\lfloor\Delta N\rfloor}^{(j+1)\lfloor\Delta N\rfloor-1}\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k}(\xi')\right]\right|\\
&\le C\sum_{\xi'}\left|\frac{1}{\lfloor\Delta N\rfloor}\sum_{k=j\lfloor\Delta N\rfloor}^{(j+1)\lfloor\Delta N\rfloor-1}\left[\mathbb{1}_{\{\xi_k=\xi'\}}-\pi^{g_k}(\xi')\right]\right|,
\end{aligned}
\tag{A.3.23}
$$

which together with Lemma 2.4.8 proves

$$
\lim_{N\to\infty}\mathbb{E}\left|I_{5,j}^N\right|=0.
\tag{A.3.24}
$$

Thus,

$$
\sum_{j=0}^{K-1}\Delta I_{5,j}^N=\Delta\sum_{j=0}^{K-1}O(1)=t\frac{\sum_{j=0}^{K-1}O(1)}{K},
\tag{A.3.25}
$$

which proves the convergence of the first term.

For the second term, by the bound in Lemma 2.4.1, for any $k \leq TN$ we have

$$\begin{aligned} \sup_{\xi' \in \mathcal{X} \times \mathcal{A}} |Q_k(\xi')| &\leq C, \\ \sup_{\xi' \in \mathcal{X} \times \mathcal{A}} |Q_k(\xi') - Q_{k-1}(\xi')| &\leq \frac{C}{N}. \end{aligned} \tag{A.3.26}$$

Note that

$$\partial_\xi Q_k(\xi') = \mathbb{1}_{\{\xi'=\xi\}}.$$

Then, by the Lipschitz continuity of the softmax transformation and the bound in Lemma 2.4.1,

$$|Q_k(\xi')\partial_\xi Q_k(\xi') - Q_{k-1}(\xi')\partial_\xi Q_{k-1}(\xi')| = \mathbb{1}_{\{\xi_k=\xi\}} |Q_k(\xi') - Q_{k-1}(\xi')| \leq \frac{C}{N}. \tag{A.3.27}$$

Then, for any $j \in 0, 1, \cdots, K-1$ and any $k \in [j\lfloor \Delta N \rfloor, (j+1)\lfloor \Delta N \rfloor - 1]$,

$$\left|Q_k(\xi')\partial_\xi Q_k(\xi') - Q_{j\lfloor \Delta N \rfloor}(\xi')\partial_\xi Q_{j\lfloor \Delta N \rfloor}(\xi')\right| \leq \frac{C(k - j\lfloor \Delta N \rfloor)}{N}. \tag{A.3.28}$$

Therefore,

$$\begin{aligned} \sum_{j=0}^{K-1} \Delta I_{6,j}^N &\leq C \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=j\lfloor \Delta N \rfloor}^{(j+1)\lfloor \Delta N \rfloor - 1} \frac{k - j\lfloor \Delta N \rfloor}{N} \\ &= C \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \sum_{k=0}^{\lfloor \Delta N \rfloor - 1} \frac{k}{N} \\ &\leq C \sum_{j=0}^{K-1} \Delta \frac{1}{\lfloor \Delta N \rfloor} \frac{\lfloor \Delta N \rfloor^2}{N} \\ &\leq C \sum_{j=0}^{K-1} \Delta \frac{\lfloor \Delta N \rfloor}{N} \\ &\leq C \sum_{j=0}^{K-1} \Delta^2 \\ &\leq C\Delta. \end{aligned} \tag{A.3.29}$$

Collecting our results, we have shown that

$$\lim_{N \to \infty} \sup_{t \in (0,T]} \mathbb{E} \left| M_t^{1,N} \right| \leq C \frac{T}{K} \tag{A.3.30}$$

Note that K was arbitrary. Consequently, we obtain

$$\lim_{N \to \infty} \sup_{t \in (0,T]} \mathbb{E} \left| M_t^{1,N} \right| = 0, \tag{A.3.31}$$

Using the same approach, one can prove the claim for $M_t^{2,N}$ and $M_t^{3,N}$. The details of the proof are omitted due to the similarity of the argument. □

A.4 Proof of Lemma 2.5.5

Proof. To prove (2.5.60), note that

$$\begin{aligned}
\|\nabla_\theta J(f_\theta)\| &= \left[\sum_{x,a}\left(\partial_{x,a} J(f_\theta)\right)^2\right]^{\frac{1}{2}} \\
&\geq \left[\sum_{x}\left(\frac{\partial J(f_\theta)}{\partial\theta\left(x,a^*(x)\right)}\right)^2\right]^{\frac{1}{2}} \\
&\overset{(a)}{\geq} \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}}\sum_{x}\left|\frac{\partial J(f_\theta)}{\partial\theta\left(x,a^*(x)\right)}\right| \\
&\overset{(b)}{=} \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}}\sum_{x}\left|\nu_\mu^{f_\theta}(x)\cdot f_\theta\left(x,a^*(x)\right)\cdot A^{f_\theta}\left(x,a^*(x)\right)\right| \\
&= \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}}\sum_{x}\nu_\mu^{f_\theta}(x)\cdot f_\theta\left(x,a^*(x)\right)\cdot\left|A^{f_\theta}\left(x,a^*(x)\right)\right|,
\end{aligned} \tag{A.4.1}$$

where step (a) is by Cauthy-Schwarz inequality and step (b) is by Lemma 2.4.3.

Define the coefficient as

$$\left\|\frac{\nu_\mu^{f^*}}{\nu_\mu^{f}}\right\|_\infty = \max_x \frac{\nu_\mu^{f^*}(x)}{\nu_\mu^*(x)}.$$

We then have the inequality:

$$\begin{aligned}
\|\nabla_\theta J(f_\theta)\| &\geq \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}}\sum_{x}\frac{\nu_\mu^{f_\theta}(x)}{\nu_\mu^{f^*}(x)}\cdot\nu_\mu^{f^*}(x)\cdot f_\theta\left(x,a^*(x)\right)\cdot\left|A^{f_\theta}\left(x,a^*(x)\right)\right| \\
&\geq \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}}\cdot\left\|\frac{\nu_\mu^{f^*}}{\nu_\mu^{f_\theta}}\right\|_\infty^{-1}\cdot\min_x f_\theta\left(x,a^*(x)\right)\cdot\sum_x\nu_\mu^{f^*}(x)\cdot\left|A^{f_\theta}\left(x,a^*(x)\right)\right| \\
&\geq \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}}\cdot\left\|\frac{\nu_\mu^{f^*}}{\nu_\mu^{f_\theta}}\right\|_\infty^{-1}\cdot\min_x f_\theta\left(x,a^*(x)\right)\cdot\sum_x\nu_\mu^{f^*}(x)\cdot A^{f_\theta}\left(x,a^*(x)\right) \\
&\overset{(a)}{=} \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}}\cdot\left\|\frac{\nu_\mu^{f^*}}{\nu_\mu^{f_\theta}}\right\|_\infty^{-1}\cdot\min_x f_\theta\left(x,a^*(x)\right)\cdot\sum_x\nu_\mu^{f^*}(x)\sum_a f^*(x,a)\cdot A^{f_\theta}(x,a) \\
&= \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}|}}\cdot\left\|\frac{\nu_\mu^{f^*}}{\nu_\mu^{f_\theta}}\right\|_\infty^{-1}\cdot\min_x f_\theta\left(x,a^*(x)\right)\cdot\left[J(f^*)-J(f_\theta)\right]
\end{aligned} \tag{A.4.2}$$

where step (a) uses the fact that f^* is deterministic and in state x selects $a^*(x)$ with probability one. The last equality uses Lemma 2.5.7.

To prove the second claim, given a policy f, recall the greedy action set for each state x:

$$\boldsymbol{\mathcal{A}}^*(x) = \left\{a^*(x)\in\boldsymbol{\mathcal{A}} : V^{f^*}(x,a^*(x)) = \max_a V^{f^*}(x,a)\right\},$$

By similar arguments as before, we can show that

$$
\begin{aligned}
\|\nabla_\theta J(f_\theta)\| &\geq \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \sum_{x,a} \left| \frac{\partial J(f_\theta)}{\partial \theta(x,a)} \right| \\
&= \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \sum_x \nu_\mu^{f_\theta}(x) \sum_a f_\theta(x,a) \cdot \left| A^{f_\theta}(x,a) \right| \qquad \text{(by Lemma 2.4.3)} \\
&\geq \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \sum_x \nu_\mu^{f_\theta}(x) \sum_{a^*(x) \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x, a^*(x)) \cdot \left| A^{f_\theta}(x, a^*(x)) \right| \\
&= \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \sum_x \nu_\mu^{f_\theta}(x) \sum_{a^*(x) \in \boldsymbol{\mathcal{A}}^*(x)} \frac{f_\theta(x, a^*(x))}{\sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a')} \cdot \left[\sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a') \right] \cdot \left| A^{f_\theta}(x, a^*(x)) \right| \\
&\overset{(a)}{=} \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \sum_x \nu_\mu^{f_\theta}(x) \sum_{a^*(x) \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta^*(x, a^*(x)) \cdot \left[\sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a') \right] \cdot \left| A^{f_\theta}(x, a^*(x)) \right| \\
&\geq \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \cdot \left\| \frac{\nu_\mu^{f_\theta^*}}{\nu_\mu^{f_\theta}} \right\|_\infty^{-1} \cdot \left[\min_x \sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a') \right] \cdot \sum_x \nu_\mu^{f_\theta^*}(x) \sum_{a^*(x) \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta^*(x, a^*(x)) A^{f_\theta}(x, a^*(x)) \\
&\overset{(b)}{=} \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \cdot \left\| \frac{\nu_\mu^{f_\theta^*}}{\nu_\mu^{f_\theta}} \right\|_\infty^{-1} \cdot \left[\min_x \sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a') \right] \cdot \sum_x \nu_\mu^{f_\theta^*}(x) \sum_{a \in \boldsymbol{\mathcal{A}}} f_\theta^*(x,a) A^{f_\theta}(x,a) \\
&\overset{(c)}{=} \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \cdot \left\| \frac{\nu_\mu^{f_\theta^*}}{\nu_\mu^{f_\theta}} \right\|_\infty^{-1} \cdot \left[\min_x \sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a') \right] \cdot \left[J(f_\theta^*) - J(f_\theta) \right] \\
&\overset{(d)}{=} \frac{1}{\sqrt{|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|}} \cdot \left\| \frac{\nu_\mu^{f_\theta^*}}{\nu_\mu^{f_\theta}} \right\|_\infty^{-1} \cdot \left[\min_x \sum_{a' \in \boldsymbol{\mathcal{A}}^*(x)} f_\theta(x,a') \right] \cdot \left[J(f^*) - J(f_\theta) \right],
\end{aligned}
\tag{A.4.3}
$$

where step (a) and (b) are by the definition of the optimal policy (2.5.62), step (c) by due to difference lemma 2.5.7 and step d is because of f_θ^* is optimal policy and thus $J(f_\theta^*) = J(f^*)$. □

Appendix B

Addtional Proofs for Chapter 4

B.1 Proof of Proposition 4.3.1

We first present a useful lemma before proving Proposition 4.3.1. The bound (B.1.1) will be frequently used in the proof of Proposition 4.3.1.

Lemma B.1.1. *For any $m', k \in \mathbb{R}_+$, there exist constants $C, m > 0$ such that for any $x, x' \in \mathbb{R}^d$,*

$$e^{-|x'-x|^2} \cdot |x'-x|^k \leq C \frac{1+|x|^m}{1+|x'|^{m'}}. \tag{B.1.1}$$

Proof. For any fixed $x \in \mathbb{R}^d$, when $\frac{|x'|}{2} \geq |x|$ we have

$$\begin{aligned} |x'-x| &\geq |x'| - |x| \geq \frac{|x'|}{2}, \\ |x'-x| &\leq |x'| + |x| \leq \frac{3|x'|}{2}. \end{aligned} \tag{B.1.2}$$

Therefore, we have that for any $m', k > 0$ there exists a constant $C_1 > 0$ such that

$$e^{-|x'-x|^2} \cdot |x'-x|^k \leq e^{-\frac{|x'|^2}{4}} \left(\frac{3}{2}|x'|\right)^k \overset{(a)}{\leq} \frac{C_1}{1+|x'|^{m'}}, \tag{B.1.3}$$

where the first inequality is due (B.1.2) and step (a) uses the fact that

$$\lim_{s \to +\infty} \frac{s^m}{e^s} = 0, \quad \forall m > 0. \tag{B.1.4}$$

When $\frac{|x'|}{2} < |x|$, for any $m', k > 0$ there exist constants $C_2, m > 0$ such that

$$e^{-|x'-x|^2} \cdot |x'-x|^k \leq (3|x|)^k \leq \frac{(3|x|)^k \cdot \left(1+|2x|^{m'}\right)}{1+|x'|^{m'}} \leq C_2 \frac{1+|x|^m}{1+|x'|^{m'}}. \tag{B.1.5}$$

Let us now choose $C = C_1 + C_2$ and then (B.1.1) holds. □

Proof of Proposition 4.3.1: The proof for the convergence results leverages the closed-form formula for the distribution. Let

$$f(t,x,\theta) = e^{-h(\theta)t}x + h(\theta)^{-1}\left(I_d - e^{-h(\theta)t}\right)g(\theta), \quad \Sigma_t(\theta) = \sigma^2(2h(\theta))^{-1}\left(I_d - e^{-2h(\theta)t}\right), \tag{B.1.6}$$

and from (4.2.2) we know that

$$X_t^\theta \sim N\left(f(t,x,\theta),\ \Sigma_t(\theta)\right). \tag{B.1.7}$$

Thus, the stationary distribution for X_t^θ is $N\left(h^{-1}(\theta)g(\theta),\ \sigma^2(2h(\theta))^{-1}\right)$. Since $h(\theta)$ is positive definite, there exists orthogonal matrix $Q(\theta)$ such that

$$h(\theta) = Q(\theta)^\top \Lambda(\theta) Q(\theta)$$

where $\Lambda(\theta) = \mathrm{diag}(\lambda_1(\theta), \cdots, \lambda_d(\theta))$ is a diagonal and all its eigenvalues are positive. Thus for $t > 0$

$$\Sigma_t(\theta) = \frac{\sigma^2}{2} Q(\theta)^T \Lambda^{-1}(\theta) \left(I_d - e^{-2\Lambda(\theta)t}\right) Q(\theta), \tag{B.1.8}$$

and the eigenvalues of $\Sigma_t(\theta)$ are $\left(\frac{\sigma^2\left(1-e^{-2\lambda_1(\theta)t}\right)}{2\lambda_1(\theta)}, \cdots, \frac{\sigma^2\left(1-e^{-2\lambda_d(\theta)t}\right)}{2\lambda_d(\theta)}\right)$. Then we know the covariance matrix $\Sigma_t(\theta)$ is also positive definite for any $t > 0$ and the density is

$$\begin{aligned} p_t(x, x', \theta) &= \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_t(\theta)\right|}} \exp\left\{-\frac{1}{2}(x' - f(t,x,\theta))^\top \Sigma_t^{-1}(\theta)(x' - f(t,x,\theta))\right\},\ t > 0 \\ p_\infty(x', \theta) &= \frac{1}{\sqrt{(2\pi)^d \left|\sigma^2(2h(\theta))^{-1}\right|}} \exp\left\{-(x' - h(\theta)^{-1}g(\theta))^\top \frac{h(\theta)}{\sigma^2}(x' - h(\theta)^{-1}g(\theta))\right\}. \end{aligned} \tag{B.1.9}$$

Proof of (i). Recall that (by assumption) $h(\theta)$ is uniformly positive definite and thus

$$p_\infty(x', \theta) \le C\sqrt{|h(\theta)|} \exp\left\{-c|x' - h(\theta)^{-1}g(\theta)|^2\right\} \overset{(a)}{\le} \frac{C}{1 + |x'|^m} \tag{B.1.10}$$

where step (a) uses the bounds for g, h in Assumption 4.2.1 and (B.1.4). Due to (B.1.9), we have for any $k \in \{1, 2, \cdots, \ell\}$ that

$$\begin{aligned} &\left|\frac{\partial}{\partial\theta_k} p_\infty(x', \theta)\right| \\ \le& C\left(\frac{\partial}{\partial\theta_k}\sqrt{|h(\theta)|}\right) \cdot \exp\left\{-c\left|x' - h(\theta)^{-1}g(\theta)\right|^2\right\} + C\sqrt{|h(\theta)|}\exp\left\{-c\left|x' - h(\theta)^{-1}g(\theta)\right|^2\right\} \\ &\cdot \left[\left|\left(x' - h(\theta)^{-1}g(\theta)\right)^\top \frac{\partial h(\theta)}{\partial\theta_k}(x' - h(\theta)^{-1}g(\theta))\right| + 2\left|\left(\frac{\partial\left(h(\theta)^{-1}g(\theta)\right)}{\partial\theta_k}\right)^\top h(\theta)\left(x' - h(\theta)^{-1}g(\theta)\right)\right|\right] \\ \overset{(a)}{\le}& C\exp\left\{-c\left|x' - h(\theta)^{-1}g(\theta)\right|^2\right\} + C\exp\left\{-c\left|x' - h(\theta)^{-1}g(\theta)\right|^2\right\} \\ &\cdot \left(\left|x' - h(\theta)^{-1}g(\theta)\right|^2 + \left|x' - h(\theta)^{-1}g(\theta)\right|\right) \\ \overset{(b)}{\le}& \frac{C}{1 + |x'|^m}, \end{aligned} \tag{B.1.11}$$

where step (a) is by the boundedness of $g(\theta)$, $\frac{\partial g(\theta)}{\partial\theta_k}$, $h(\theta)$, $\frac{\partial h(\theta)}{\partial\theta_k}$ and since $h(\theta)$ is positive definite due to Assumption 4.2.1. Step (b) is due to equation (B.1.1) with $x = h(\theta)^{-1}g(\theta)$ and equation (B.1.4). Using the same method as in (B.1.11), we can obtain the bound for $\nabla_\theta^2 p_\infty(x, \theta)$.

Proof of (ii) and (iii). We now prove (4.3.10). First let

$$X := x' - f(t, x, \theta), \quad Y := x' - h(\theta)^{-1}g(\theta),$$

and then since h is uniformly positive definite:

$$|X - Y| = \left|e^{-h(\theta)t}x - e^{-2h(\theta)t}h(\theta)^{-1}g(\theta)\right| \le Ce^{-ct}(1 + |x|). \tag{B.1.12}$$

We will use the following decomposition:

$$
\begin{aligned}
&|p_t(x,x',\theta)-p_\infty(x',\theta)|\\
\le& C\left(\frac{1}{\sqrt{\left|\left(I_d-e^{-2h(\theta)t}\right)\right|}}-1\right)\\
&+C\left|\exp\left\{-X^\top\frac{h(\theta)}{\sigma^2}\left(I_d-e^{-2h(\theta)t}\right)^{-1}X\right\}-\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}\left(I_d-e^{-2h(\theta)t}\right)^{-1}Y\right\}\right| \qquad \text{(B.1.13)}\\
&+\left|\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}\left(I_d-e^{-2h(\theta)t}\right)^{-1}Y\right\}-\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}Y\right\}\right|\\
=:& I_1+I_2+I_3
\end{aligned}
$$

For I_1, note that when $t>1$

$$
\frac{1}{\sqrt{\left|\left(I_d-e^{-2h(\theta)t}\right)\right|}}-1=\frac{1}{\sqrt{\prod_{k=1}^d\left(1-e^{-2\lambda_k(\theta)t}\right)}}-1\le C\left[1-\prod_{k=1}^d\left(1-e^{-2\lambda_k(\theta)t}\right)\right]\le Ce^{-2\lambda_1(\theta)t}\le Ce^{-2ct}, \qquad \text{(B.1.14)}
$$

where $\lambda_1(\theta)\le\lambda_2(\theta)\le\cdots\le\lambda_d(\theta)$ are the eigenvalues of the matrix $h(\theta)$. For I_3, similar to (B.1.8), we know the eigenvalues of $h(\theta)\left(\left(I_d-e^{-2h(\theta)t}\right)^{-1}-I_d\right)$ are $\frac{\lambda_i(\theta)e^{-2\lambda_i(\theta)t}}{1-e^{-2\lambda_i(\theta)t}}$, $i=1,\cdots,d$, which implies that $h(\theta)\left(\left(I_d-e^{-2h(\theta)t}\right)^{-1}-I_d\right)$ is also a positive definite matrix. When $t>1$, since $h(\theta)$ is uniformly positive definite, the eigenvalues will have a uniform upper bound:

$$
\frac{\lambda_i(\theta)e^{-2\lambda_i(\theta)t}}{1-e^{-2\lambda_i(\theta)t}}\le\frac{C}{1-e^{-c}}\le C,\quad t>1,\ \forall i\in\{1,\cdots,d\}. \qquad \text{(B.1.15)}
$$

Thus for any $m',k>0$, there exists a constant $C>0$ such that when $t>1$

$$
\begin{aligned}
&\left|\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}\left(I_d-e^{-2h(\theta)t}\right)^{-1}Y\right\}-\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}Y\right\}\right|\\
=&\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}Y\right\}\left|\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}\left(\left(I_d-e^{-2h(\theta)t}\right)^{-1}-I_d\right)Y\right\}-1\right|\\
\overset{(a)}{\le}&C\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}Y\right\}\left|Y^\top\frac{h(\theta)}{\sigma^2}\left(\left(I_d-e^{-2h(\theta)t}\right)^{-1}-I_d\right)Y\right|\\
\overset{(b)}{\le}&C\exp\left\{-Y^\top\frac{h(\theta)}{\sigma^2}Y\right\}|Y|^2\cdot\lambda_{\max}\left(h(\theta)\left(\left(I_d-e^{-2h(\theta)t}\right)^{-1}-I_d\right)\right) \qquad \text{(B.1.16)}\\
\overset{(c)}{\le}&C\exp\left\{-c\left|x'-h(\theta)^{-1}g(\theta)\right|^2\right\}\cdot\left|x'-h(\theta)^{-1}g(\theta)\right|^2\\
\overset{(d)}{\le}&Ce^{-ct}\frac{1}{1+|x'|^{m'}}\\
\le&C\frac{1}{(1+|x'|^{m'})(1+t)^k},
\end{aligned}
$$

where step (a) is by the positive definiteness of $\frac{h(\theta)}{\sigma^2}\left(\left(I_d-e^{-2h(\theta)t}\right)^{-1}-I_d\right)$, which means

$$
Y^\top\frac{h(\theta)}{\sigma^2}\left(\left(I_d-e^{-2h(\theta)t}\right)^{-1}-I_d\right)Y\ge 0,
$$

and the fact $0\le 1-e^{-s}\le s,\ \forall s\ge 0$. In step (b), $\lambda_{\max}$ denotes the largest eigenvalue and step (c) uses (B.1.15). Step (d) follows from (B.1.1) with $x=h(\theta)^{-1}g(\theta)$ and the boundedness of g,h.

For I_2, define the function on $F_t : \mathbb{R}^d \to \mathbb{R}$ for $t > 0$

$$F_t(x) := \exp\left\{-x^\top \frac{h(\theta)}{\sigma^2}\left(I_d - e^{-2h(\theta)t}\right)^{-1} x\right\}.$$

By mean value theorem,

$$F_t(x) - F_t(y) = \nabla F_t(x_0)^\top (x-y) = -\frac{2h(\theta)}{\sigma^2}\left(I_d - e^{-2h(\theta)t}\right)^{-1} F_t(x_0) x_0^\top (x-y), \tag{B.1.17}$$

where $x_0 = t_0 x + (1-t_0)y$ for some $t_0 \in [0,1]$. Thus for any $m', k > 0$ there exist constants $C, m > 0$ such that when $t > 1$

$$\begin{aligned} |F_t(x) - F_t(Y)| &\overset{(a)}{=} \frac{2}{\sigma^2}\left|\exp\left\{-(X_0)^\top \frac{h(\theta)}{\sigma^2}\left(I_d - e^{-2h(\theta)t}\right)^{-1} X_0\right\} X_0^\top (X-Y)\right| \\ &\overset{(b)}{\leq} e^{-c|X_0|^2}|X_0| C e^{-ct}(1+|x|) \\ &\overset{(c)}{\leq} C\frac{1+|x|^m}{(1+|x'|^{m'})(1+t)^k}, \end{aligned} \tag{B.1.18}$$

where in step (a)

$$X_0 = t_0 X + (1-t_0)Y = x' - t_0 f(t,x,\theta) - (1-t_0)h(\theta)^{-1} g(\theta), \tag{B.1.19}$$

for some $t_0 \in [0,1]$. Step (b) uses (B.1.12) and (B.1.15) and step (c) is by substituting in x in (B.1.1) to be the X_0 in (B.1.19). Combining (B.1.13), (B.1.14), (B.1.16), and (B.1.18), we have for $t > 1$

$$|p_t(x,x',\theta) - p_\infty(x',\theta)| \leq C\frac{1+|x|^m}{(1+|x'|^{m'})(1+t)^k}. \tag{B.1.20}$$

The proof of (4.3.10) for the case $i = 1, 2$ and (4.3.11) is the same as the proof for $|p_t(x,x',\theta) - p_\infty(x',\theta)|$ above (i.e., one uses the decomposition in (B.1.13) and (B.1.1) with different choices of x). The only challenge is establishing a bound for $\nabla_\theta e^{-h(\theta)t}$. $e^{-h(\theta)t}$ satisfies the ODE

$$\frac{d}{dt}e^{-h(\theta)t} = -h(\theta)e^{-h(\theta)t} \tag{B.1.21}$$

with initial value I_d.[1] Differentiating (B.1.21) with respect to $\theta_i, i \in \{1, \cdots, d\}$ yields an ODE for $\frac{\partial}{\partial \theta_i}e^{-h(\theta)t}$:

$$\frac{d}{dt}\frac{\partial}{\partial\theta_i}e^{-h(\theta)t} = -\frac{\partial h(\theta)}{\partial\theta_i}e^{-h(\theta)t} - h(\theta)\frac{\partial}{\partial\theta_i}e^{-h(\theta)t}, \tag{B.1.22}$$

with initial value 0. Using an integrating factor yields

$$\frac{d}{dt}\left(e^{h(\theta)t}\frac{\partial}{\partial\theta_i}e^{-h(\theta)t}\right) = -e^{h(\theta)t}\frac{\partial h(\theta)}{\partial\theta_i}e^{-h(\theta)t},$$

and thus

$$\frac{\partial}{\partial\theta_i}e^{-h(\theta)t} = e^{-h(\theta)t}\int_0^t e^{h(\theta)s}\frac{\partial h(\theta)}{\partial\theta_i}e^{-h(\theta)s}ds. \tag{B.1.23}$$

Since $e^{h(\theta)t}$ is invertible for any t, we know the matrices $e^{h(\theta)s}\frac{\partial h(\theta)}{\partial\theta_i}e^{-h(\theta)s}$ and $\frac{\partial h(\theta)}{\partial\theta_i}$ are similar and thus their eigenvalues are the same, which implies that their spectral norm are also the same. We therefore can show that

$$\left|\frac{\partial}{\partial\theta_i}e^{-h(\theta)t}\right| \leq C\left|e^{-h(\theta)t}\right|\int_0^t\left|\frac{\partial h(\theta)}{\partial\theta_i}\right|ds \overset{(a)}{\leq} Ce^{-ct}t, \tag{B.1.24}$$

[1] Here we use the fact that $\frac{\partial}{\partial y}e^{Ay} = Ae^{Ay} = e^{Ay}A$.

where step (a) is by the bound for $\nabla_\theta h(\theta)$ in Assumption 4.2.1. Using the same method, we also can show that

$$\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}e^{-h(\theta)t}\right| \le C\left|e^{-h(\theta)t}\right|\int_0^t\left|\frac{\partial h(\theta)}{\partial\theta_i\partial\theta_j}\right|ds \overset{(a)}{\le} Ce^{-ct}t, \quad \forall i,j\in\{1,\cdots,d\}. \tag{B.1.25}$$

Proof of (iv). The first part of (4.3.12) follows from the fact that X_t^θ has a multivariate normal distribution whose mean and variance are uniformly bounded. (4.3.12) is obvious when $t=0$. For $t>0$, as we know $\Sigma_t(\theta)$ is positive definite for $t>0$, thus the random variable

$$Y := \Sigma_t^{-\frac{1}{2}}(\theta)\left(X_t^\theta - f(t,x,\theta)\right) \tag{B.1.26}$$

has a d-dimensional standard normal distribution, where $\Sigma_t^{\frac{1}{2}}(\theta)$ denotes the square root matrix of $\Sigma_t(\theta)$. Since for any $m>0$ there exists a $C_m>0$ such that $\mathbb{E}\left|Y\right|^m = C_m < \infty$.

$$\mathbb{E}_x\left|X_t^\theta\right|^m = \mathbb{E}_x\left|\Sigma_t^{\frac{1}{2}}(\theta)Y + f(t,x,\theta)\right|^m \le C\left(\left|\Sigma_t^{\frac{1}{2}}(\theta)\right|^m\mathbb{E}_x\left|Y\right|^m + \left|f(t,x,\theta)\right|^m\right) \overset{(a)}{\le} C(1+|x|^m), \tag{B.1.27}$$

where step (a) is by the uniform bound for $g(\theta)$ and $h(\theta)$ in Assumption 4.2.1. For the second part of (4.3.12), we use (B.1.27) to develop the following bound:

$$\begin{aligned}
\mathbb{E}_{x,\tilde{x}}|\tilde{X}_t^\theta|^m &\le 2|\tilde{x}|^m + 2\mathbb{E}_{x,\tilde{x}}\left|\int_0^t\left|e^{-h(\theta)(t-s)}\right|\cdot\left|\nabla_\theta g(\theta) - \nabla_\theta h(\theta)X_s^\theta\right|ds\right|^m \\
&\overset{(a)}{\le} 2|\tilde{x}|^m + C_m\mathbb{E}_{x,\tilde{x}}\left|\int_0^t e^{-c(t-s)}\left(1+|X_s^\theta|\right)ds\right|^m \\
&\le 2|\tilde{x}|^m + C_m\mathbb{E}_x\left|\int_0^t\frac{e^{cs}}{e^{ct}-1}\left(1+|X_s^\theta|\right)ds\right|^m e^{-cmt}\left(e^{ct}-1\right)^m \\
&\overset{(b)}{\le} 2|\tilde{x}|^m + C_m\mathbb{E}_x\left|\int_0^t\frac{e^{cs}}{e^{ct}-1}\left(1+|X_s^\theta|\right)^m ds\right| \\
&\le C_m\left(1+|x|^m+|\tilde{x}|^m\right),
\end{aligned} \tag{B.1.28}$$

where step (a) is by Assumption 4.2.1 and the fact

$$\lambda_{\max}\left(e^{-h(\theta)(t-s)}\right) = e^{-\lambda_{\min}(h(\theta)(t-s))} \le e^{-c(t-s)} \tag{B.1.29}$$

and step (b) is by Jensen's inequality. In particular, let $p(s) = \frac{1}{c}\frac{e^{cs}}{e^{ct}-1}$ and we have $\int_0^t p(s)ds = 1$, and therefore $p(s)$ is a probability density function on $[0,t]$. By Jensen's inequality,

$$\left|\int_0^t\left(1+|X_s^\theta|\right)p(s)ds\right|^m \le \int_0^t\left(1+|X_s^\theta|\right)^m p(s)ds, \tag{B.1.30}$$

which we have used in step (b) of equation (B.1.28).

Proof of (v). For (4.3.13), the conclusion for $t=0$ is trivial. When $t>0$, by (4.2.6) and (4.3.12), we have for any polynomial bounded function f that

$$|\mathbb{E}_x f\left(X_t^\theta\right)| \le \mathbb{E}_x|f\left(X_t^\theta\right)| \le C\mathbb{E}_x(1+\left|X_t^\theta\right|^m) \le C(1+|x|^m). \tag{B.1.31}$$

For the derivatives, we will use the dominated convergence theorem. By (B.1.9), we have

$$\mathbb{E}_x f\left(X_t^\theta\right) = \int_{\mathbb{R}^d} f\left(f(t,x,\theta)+x'\right)\frac{1}{\sqrt{(2\pi)^d\left|\Sigma_t(\theta)\right|}}\exp\left\{-\frac{1}{2}x'^\top\Sigma_t^{-1}(\theta)x'\right\}dx'. \tag{B.1.32}$$

Let Z^θ denote a normal distribution

$$Z_t^\theta \sim N\left(0, \Sigma_t(\theta)\right)$$

and then

$$\mathbb{E}_x f(X_t^\theta) = \mathbb{E} f\left(f(t,x,\theta) + Z_t^\theta\right) = \mathbb{E} f\left(e^{-h(\theta)t}x + h(\theta)^{-1}\left(I_d - e^{-h(\theta)t}\right)g(\theta) + Z_t^\theta\right). \tag{B.1.33}$$

For $\nabla_x \mathbb{E}_x f(X_t^\theta)$, we change the order of ∇_x and $\mathbb{E}_x$ and obtain for $t \in (0,1]$

$$\begin{aligned}
&\mathbb{E}\left|\nabla_x f\left(e^{-h(\theta)t}x + h(\theta)^{-1}\left(I_d - e^{-h(\theta)t}\right)g(\theta) + Z_t^\theta\right)\right| \\
=&\mathbb{E}\left|e^{-h(\theta)t}\nabla f\left(e^{-h(\theta)t}x + h(\theta)^{-1}\left(I_d - e^{-h(\theta)t}\right)g(\theta) + Z_t^\theta\right)\right| \\
\leq& e^{-ct}\mathbb{E}_x\left|\nabla f(X_t^\theta)\right| \\
\leq& C(1+|x|^m).
\end{aligned} \tag{B.1.34}$$

Therefore, by DCT we have that

$$\begin{aligned}
\left|\nabla_x \mathbb{E}_x f(X_t^\theta)\right| &= \left|\mathbb{E}\nabla_x f\left(e^{-h(\theta)t}x + h(\theta)^{-1}\left(I_d - e^{-h(\theta)t}\right)g(\theta) + Z_t^\theta\right)\right| \\
&= \left|e^{-h(\theta)t}\mathbb{E}_x \nabla f(X_t^\theta)\right| \\
&\leq C(1+|x|^m).
\end{aligned} \tag{B.1.35}$$

Similarly for $\nabla_x^2 \mathbb{E}_x f(X_t^\theta)$, we have for $t \in (0,1]$

$$\begin{aligned}
\left|\nabla_x^2 \mathbb{E}_x f(X_t^\theta)\right| &= \left|\mathbb{E}e^{-h(\theta)t}\nabla^2 f\left(e^{-h(\theta)t}x + h(\theta)^{-1}\left(I_d - e^{-h(\theta)t}\right)g(\theta) + Z_t^\theta\right)e^{-h(\theta)t}\right| \\
&= \left|e^{-h(\theta)t}\mathbb{E}_x \nabla^2 f(X_t^\theta)e^{-h(\theta)t}\right| \\
&\leq C(1+|x|^m).
\end{aligned} \tag{B.1.36}$$

Finally, for $\nabla_\theta \nabla_x^2 \mathbb{E}_x f(X_t^\theta)$, by (B.1.24) we have for $t \in (0,1]$ that

$$\begin{aligned}
\left|\frac{\partial}{\partial\theta_i}\nabla_x^2 \mathbb{E}_x f(X_t^\theta)\right| &\leq 2\left|\frac{\partial}{\partial\theta_i}e^{-h(\theta)t}\right|\left|\mathbb{E}_x \nabla^2 f(X_t^\theta)\right|\left|e^{-h(\theta)t}\right| + \left|e^{-h(\theta)t}\frac{\partial}{\partial\theta_i}\mathbb{E}_x \nabla^2 f(X_t^\theta)e^{-h(\theta)t}\right| \\
&\leq C(1+|x|^m) + e^{-ct}\left|\frac{\partial}{\partial\theta_i}\mathbb{E}_x \nabla^2 f(X_t^\theta)\right|.
\end{aligned} \tag{B.1.37}$$

Thus, it remains to prove a bound for $\frac{\partial}{\partial\theta_i}\mathbb{E}_x f_0(X_t^\theta)$, where f_0 is any polynomial bounded function such that

$$|f_0(x)| + |\nabla f_0(x)| \leq C(1+|x|^m), \quad \forall x \in \mathbb{R}^d.$$

In order to establish this result, we need a bound for $\nabla_\theta \Sigma_t^{-1}(\theta)$ when $t \in [0,1]$. For $t \in (0,1]$,

$$\begin{aligned}
&\frac{\partial}{\partial\theta_i}\Sigma_t^{-1}(\theta) \\
=&2\sigma^2\frac{\partial}{\partial\theta_i}\left[h(\theta)\left(I_d - e^{-2h(\theta)t}\right)^{-1}\right] \\
\overset{(a)}{=}&2\sigma^2\left(\frac{\partial}{\partial\theta_i}h(\theta)\right)\left(I_d - e^{-2h(\theta)t}\right)^{-1} + 2\sigma^2\left(I_d - e^{-2h(\theta)t}\right)^{-1}h(\theta)\left(\frac{\partial}{\partial\theta_i}e^{-2h(\theta)t}\right)\left(I_d - e^{-2h(\theta)t}\right)^{-1} \\
=&2\sigma^2\left(I_d - e^{-2h(\theta)t}\right)^{-1}\left[I_d - e^{-2h(\theta)t} - 2e^{-2h(\theta)t}h(\theta)t\right]\left(\frac{\partial}{\partial\theta_i}h(\theta)t\right)\left(I_d - e^{-2h(\theta)t}\right)^{-1},
\end{aligned} \tag{B.1.38}$$

where in step (a) we change the order of $\left(I_d - e^{-2h(\theta)t}\right)^{-1}$ and $h(\theta)$ since $h(\theta)e^{h(\theta)t} = e^{h(\theta)t}h(\theta)$.

For $t \in [0,1]$, 0 is the only singular point for $\nabla_\theta \Sigma_t^{-1}(\theta)$. Therefore to prove the uniform bound, it suffices to prove the limit exists when $t \to 0^+$. As $t \to 0+$,

$$I_d - e^{-2h(\theta)t} - 2e^{-2h(\theta)t}h(\theta)t = I_d - 2h(\theta)t - \left(I_d - 2h(\theta)t + 2h^2(\theta)t^2 + o(t^2)\right) = -2h^2(\theta)t^2 + o(t^2). \tag{B.1.39}$$

Therefore,

$$\begin{aligned}
\lim_{t\to 0+} \frac{\partial}{\partial \theta_i}\Sigma_t^{-1}(\theta) &= 2\sigma^2 \lim_{t\to 0+} \left(2h(\theta)t + o(t)\right)^{-1} \left(-2h^2(\theta)t^2 + o(t^2)\right) \left(\frac{\partial}{\partial \theta_i} h(\theta)\right) \left(2h(\theta)t + o(t)\right)^{-1} \\
&= 2\sigma^2 \lim_{t\to 0+} \left(2h(\theta) + o(1)\right)^{-1} \left(-2h^2(\theta) + o(1)\right) \left(\frac{\partial}{\partial \theta_i} h(\theta)\right) \left(2h(\theta) + o(1)\right)^{-1} \\
&= -16h(\theta) \left(\frac{\partial}{\partial \theta_i} h(\theta)\right) h^{-1}(\theta),
\end{aligned} \tag{B.1.40}$$

which together with the bound for $h(\theta)$ from Assumption 4.2.1 yields

$$\left|\nabla_\theta \Sigma_t^{-1}(\theta)\right| \le C, \quad t \in [0,1]. \tag{B.1.41}$$

We will now analyze $\frac{\partial}{\partial \theta_i}\mathbb{E}_x f_0(X_t^\theta)$ for $t \in (0,1]$ using formula (B.1.32) and changing the order of $\frac{\partial}{\partial \theta_i}$ and $\mathbb{E}_x$.

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left| \frac{\partial}{\partial \theta_i} \left(f_0\left(f(t,x,\theta) + x'\right) \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_t(\theta)\right|}} \exp\left\{-x'^\top \Sigma_t^{-1}(\theta) x'\right\} \right) \right| dx' \\
\le &\int_{\mathbb{R}^d} \left| \left(\frac{\partial}{\partial \theta_i} f(t,x,\theta)\right)^\top \nabla f_0\left(f(t,x,\theta) + x'\right) \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_t(\theta)\right|}} \exp\left\{-x'^\top \Sigma_t^{-1}(\theta) x'\right\} \right| dx' \\
+ &\int_{\mathbb{R}^d} \left| f_0\left(f(t,x,\theta) + x'\right) \frac{\partial}{\partial \theta_i} \left(\frac{1}{\sqrt{(2\pi)^d \left|\Sigma_t(\theta)\right|}} \right) \exp\left\{-x'^\top \Sigma_t^{-1}(\theta) x'\right\} \right| dx' \\
+2 &\int_{\mathbb{R}^d} \left| f_0\left(f(t,x,\theta) + x'\right) \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_t(\theta)\right|}} \exp\left\{-x'^\top \Sigma_t^{-1}(\theta) x'\right\} x'^\top \frac{\partial}{\partial \theta_i} \left(\Sigma_t^{-1}(\theta)\right) x' \right| dx' \\
\overset{(a)}{\le} &C\mathbb{E}\left|\nabla f_0\left(f(t,x,\theta) + Z_t^\theta\right)\right| + C\mathbb{E}\left|f_0\left(f(t,x,\theta) + Z_t^\theta\right)\right| + C\mathbb{E}\left|f_0\left(f(t,x,\theta) + Z_t^\theta\right)\left(Z_t^\theta\right)^\top Z_t^\theta\right| \\
\le &C\mathbb{E}_x\left|\nabla f_0\left(X_t^\theta\right)\right| + \mathbb{E}_x\left|f_0\left(X_t^\theta\right)\right| + \mathbb{E}_x\left|f_0^2\left(X_t^\theta\right)\right| + \mathbb{E}\left|Z_t^\theta\right|^4 \\
\overset{(b)}{\le} &C(1 + |x|^m),
\end{aligned} \tag{B.1.42}$$

where step (a) is by (B.1.41) and the uniform bounds for $g(\theta), h(\theta)$ and step (b) is by (B.1.31) and the polynomial boudnds for $f_0^2, \nabla f_0$. Then, by the dominated convergence theorem,

$$\left|\nabla_\theta \mathbb{E}_x f_0(X_t^\theta)\right| \le C(1 + |x|^m), \quad t \in (0,1]. \tag{B.1.43}$$

Combining (B.1.37) and (B.1.43), we obtain the bound for $\nabla_\theta \nabla_x^2 \mathbb{E}_x f(X_t^\theta)$. The bound $\nabla_\theta^2 \nabla_x^2 \mathbb{E}_x f(X_t^\theta)$ can be obtained using similar calculations, which concludes the proof of the proposition. □

B.2 Poisson PDEs

In this section we give the detailed proof of the regularities for the solutions of Poisson PDEs. We first show the proof of Lemma 4.3.3.

Proof of Lemma 4.3.3: We begin by proving that the integral (4.3.15) is finite. We divide (4.3.15) into two terms:

$$\begin{aligned}
&v^1(x,\tilde{x},\theta)\\
=&(\mathbb{E}_{\pi_\theta} f(Y)-\beta)\int_0^\infty \left(\nabla_\theta \mathbb{E}_{\pi_\theta} f(Y)-\mathbb{E}_{x,\tilde{x}}\nabla f(X_t^\theta)\tilde{X}_t^\theta\right)^\top dt\\
=&(\mathbb{E}_{\pi_\theta} f(Y)-\beta)\left[\int_0^\infty \left(\nabla_\theta \mathbb{E}_{\pi_\theta} f(Y)-\nabla_\theta \mathbb{E}_x f(X_t^\theta)\right)^\top dt+\int_0^\infty \left(\nabla_\theta \mathbb{E}_x f(X_t^\theta)-\mathbb{E}_{x,\tilde{x}}\nabla f(X_t^\theta)\tilde{X}_t^\theta\right)^\top dt\right]\\
=:&v^{1,1}(x,\theta)+v^{1,2}(x,\tilde{x},\theta).
\end{aligned} \tag{B.2.1}$$

By Assumption 4.2.1 and (4.3.9),

$$\left|\int_{\mathbb{R}^d} f(x')\nabla_\theta^i p_\infty(x',\theta)dx'\right|\le C\int \frac{1+|x'|^m}{1+|x'|^{m'}}dx' \overset{(a)}{\le} C,\quad i=0,1,2, \tag{B.2.2}$$

where step (a) is by choosing $m'>m+d$. Thus by dominated convergence theorem (DCT):

$$\left|\nabla_\theta^i \mathbb{E}_{Y\sim\pi_\theta} f(Y)\right|=\left|\int_{\mathbb{R}^d} f(x')\nabla_\theta^i p_\infty(x',\theta)dx'\right|\le C,\quad i=0,1,2. \tag{B.2.3}$$

Similarly, we can bound $v^{1,1}$ as follows:

$$\begin{aligned}
&\left|v^{1,1}(x,\theta)\right|\\
\overset{(a)}{\le}&C\int_0^1 \left(1+\left|\nabla_\theta \mathbb{E}_x f(X_t^\theta)\right|\right)dt+C\int_1^\infty\int_R (1+|x'|^m)\left|\nabla_\theta p_\infty(x',\theta)-\nabla_\theta p_t(x,x',\theta)\right|dx'dt\\
\overset{(b)}{\le}&C+C\int_0^\infty\int_R (1+|x'|^m)\frac{1+|x|^{m'}}{(1+|x'|^{m''})(1+t)^2}dx'dt\\
\overset{(c)}{\le}&C\left(1+|x|^{m'}\right),
\end{aligned} \tag{B.2.4}$$

where steps (a) is by Assumption 4.2.1 and (B.2.3), step (b) by (4.3.10) and (4.3.13), and step (c) follows from selecting $m''>m+d$. For $v^{1,2}$, by Assumption 4.2.1 and (4.3.12) we have

$$\left|\mathbb{E}_{x,0}\nabla f(X_t^\theta)\tilde{X}_t^\theta\right|\le \mathbb{E}_{x,0}\left|\tilde{X}_t^\theta\right|^2+\mathbb{E}_x\left|\nabla f(X_t^\theta)\right|^2\le C<\infty. \tag{B.2.5}$$

Thus by DCT we have

$$\nabla_\theta \mathbb{E}_x f(X_t^\theta)=\mathbb{E}_x\nabla_\theta f(X_t^\theta)=\mathbb{E}_{x,0}\nabla f(X_t^\theta)\tilde{X}_t^\theta, \tag{B.2.6}$$

which together with (4.3.7) derives

$$\begin{aligned}
\nabla_\theta \mathbb{E}_x f(X_t^\theta)-\mathbb{E}_{x,\tilde{x}}\nabla f(X_t^\theta)\tilde{X}_t^\theta &= \mathbb{E}_{x,0}\nabla f(X_t^\theta)\tilde{X}_t^\theta-\mathbb{E}_{x,\tilde{x}}\nabla f(X_t^\theta)\tilde{X}_t^\theta\\
&= -\mathbb{E}_x\nabla f\left(X_t^\theta\right)e^{-h(\theta)t}\tilde{x}.
\end{aligned} \tag{B.2.7}$$

Thus, $v^{1,2}$ satisfies the bound

$$\begin{aligned}\left|v^{1,2}(x,\tilde{x},\theta)\right| &= \left|\left(\mathbb{E}_{Y\sim\pi_\theta} f(Y)-\beta\right)\int_0^\infty \mathbb{E}_x \nabla f\left(X_t^\theta\right) e^{-h(\theta)t}\tilde{x}dt\right| \\ &\overset{(a)}{\leq} C\int_0^\infty \left(1+\mathbb{E}_x\left|X_t^\theta\right|^m\right) e^{-ct}dt\cdot|\tilde{x}| \\ &\overset{(b)}{\leq} C\int_0^\infty \left(1+|x|^{m'}\right) e^{-ct}dt\cdot|\tilde{x}| \\ &\leq C\left(1+|x|^{m'}+|x'|^{m'}\right),\end{aligned} \tag{B.2.8}$$

where step (a) is by Assumption 4.2.1, (B.2.3) and $\lambda_{\max}\left(e^{-h(\theta)t}\right)\leq e^{-ct}$. Step (b) is by (4.3.12).

Next we show that $v^1(x,\tilde{x},\theta)$ is differentiable with respect to $x,\tilde{x}$, and θ. We can prove this using a version of the dominated convergence theorem (see Theorem 2.27 in [60]), where it suffices to show that the derivative of the integrand is bounded by an integrable function. Using the same analysis as in (B.2.8), we can show that

$$\left|\int_0^\infty e^{-h(\theta)t}\mathbb{E}_x\nabla f(X_t^\theta)dt\right| \leq C\int_0^\infty \left(1+\mathbb{E}_x\left|X_t^\theta\right|^m\right)e^{-ct}dt \leq C\left(1+|x|^{m'}\right). \tag{B.2.9}$$

Therefore, by the dominated convergence theorem, we know v^1 is differentiable with respect to $\tilde{x}$. Furthermore, we can change the order of $\nabla_{\tilde{x}}$ and the integral in v^1 and obtain

$$\left|\nabla_{\tilde{x}}v^1(x,\tilde{x},\theta)\right| = \left|\left(\mathbb{E}_{Y\sim\pi_\theta}f(Y)-\beta\right)\int_0^\infty \mathbb{E}_x\nabla f\left(X_t^\theta\right)e^{-h(\theta)t}dt\right| \overset{(a)}{\leq} C\left(1+|x|^{m'}\right), \tag{B.2.10}$$

where step (a) is by (B.2.3) and (B.2.9).

By (4.3.10), (B.2.3), and the same approach as in (B.2.4), we have

$$\begin{aligned}&\Bigg|\nabla_\theta\mathbb{E}_{\pi_\theta}f(Y)\int_1^\infty\left(\nabla_\theta\mathbb{E}_{\pi_\theta}f(Y)-\nabla_\theta\mathbb{E}_x f(X_t^\theta)\right)dt \\ &+\left(\mathbb{E}_{\pi_\theta}f(Y)-\beta\right)\int_1^\infty\left(\nabla_\theta^2\mathbb{E}_{\pi_\theta}f(Y)-\nabla_\theta^2\mathbb{E}_x f(X_t^\theta)\right)dt\Bigg| \\ \leq& C\int_1^\infty\int_{\mathbb{R}^d}\left|f(x')\nabla_\theta\left[p_\infty(x',\theta)-p_t(x,x',\theta)\right]\right|dx'dt \\ &+C\int_1^\infty\int_{\mathbb{R}^d}\left|f(x')\nabla_\theta^2\left[p_\infty(x',\theta)-p_t(x,x',\theta)\right]\right|dx'dt \\ \leq& C\left(1+|x|^{m'}\right).\end{aligned} \tag{B.2.11}$$

By (4.3.13) and (B.2.3),

$$\begin{aligned}&\Bigg|\nabla_\theta\mathbb{E}_{Y\sim\pi_\theta}f(Y)\int_0^1\nabla_\theta\mathbb{E}_{Y\sim\pi_\theta}f(Y)-\nabla_\theta\mathbb{E}_x f(X_t^\theta)dt \\ &+\left(\mathbb{E}_{Y\sim\pi_\theta}f(Y)-\beta\right)\int_0^1\nabla_\theta^2\mathbb{E}_{Y\sim\pi_\theta}f(Y)-\nabla_\theta^2\mathbb{E}_x f(X_t^\theta)dt\Bigg| \\ \leq& C\left(1+|x|^{m'}\right).\end{aligned} \tag{B.2.12}$$

By (B.2.11), (B.2.12), and DCT we know $v^{1,1}$ is differentiable with respect to θ and

$$\left|\nabla_\theta v^{1,1}(x,\theta)\right| \leq C\left(1+|x|^{m'}\right). \tag{B.2.13}$$

For $\nabla_\theta v^{1,2}$, by (B.2.7) we have for any $i \in \{1, 2, \cdots, \ell\}$

$$\begin{aligned}
&\left| \int_0^\infty \frac{\partial}{\partial \theta_i} \left(\nabla_\theta \mathbb{E}_x f(X_t^\theta) - \mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta) \tilde{X}_t^\theta \right) dt \right| \\
=& \left| \int_0^\infty \mathbb{E}_x \nabla f\left(X_t^\theta\right) \left(\frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \right) \tilde{x} dt + \int_0^\infty \left(\frac{\partial}{\partial \theta_i} \mathbb{E}_x \nabla f\left(X_t^\theta\right) \right) e^{-h(\theta)t} \tilde{x} dt \right| \\
\overset{(a)}{\leq}& |\tilde{x}| \cdot \int_0^\infty \left| \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \right| \cdot \left| \mathbb{E}_x \nabla f(X_t^\theta) \right| dt + |\tilde{x}| \cdot \int_0^\infty \left| e^{-h(\theta)t} \right| \cdot \left| \mathbb{E}_x \frac{\partial}{\partial \theta_i} \nabla f(X_t^\theta) \right| dt \\
=:& I_4 + I_5.
\end{aligned} \tag{B.2.14}$$

where in step (a) we use

$$\frac{\partial}{\partial \theta_i} \mathbb{E}_x \nabla f\left(X_t^\theta\right) = \mathbb{E}_x \frac{\partial}{\partial \theta_i} \nabla f(X_t^\theta), \tag{B.2.15}$$

which is due to (B.2.5) and (B.2.6).

By (B.1.24),

$$\begin{aligned}
I_4 \leq& C |\tilde{x}| \left(\left| \int_0^1 \mathbb{E}_x \nabla f(X_t^\theta) e^{-ct} t dt \right| + \left| \int_1^\infty \mathbb{E}_x \nabla f(X_t^\theta) e^{-ct} t dt \right| \right) \\
\overset{(a)}{\leq}& C|\tilde{x}| \Bigg(1 + \int_1^\infty e^{-ct} t \int_{\mathbb{R}^d} (1 + |x'|^m) \left| p_t(x, x', \theta) - p_\infty(x', \theta) \right| dx' dt \\
&+ \int_1^\infty e^{-ct} t \int_{\mathbb{R}^d} (1 + |x'|^m) \left| p_\infty(x', \theta) \right| dx' dt \Bigg) \\
\overset{(b)}{\leq}& C \left(1 + |x|^{m'} + |\tilde{x}|^{m'} \right),
\end{aligned} \tag{B.2.16}$$

where in step (a) we used (4.3.13) and step (b) is by (4.3.9), (4.3.10), and the same analysis as in (B.2.3) and (B.2.4). Similarly,

$$\begin{aligned}
I_5 \leq& C |\tilde{x}| \left(\int_0^1 \left| \mathbb{E}_x \frac{\partial}{\partial \theta_i} \nabla f(X_t^\theta) \right| e^{-ct} dt + \int_1^\infty \left| \mathbb{E}_x \frac{\partial}{\partial \theta_i} \nabla f(X_t^\theta) \right| e^{-ct} dt \right) \\
\leq& C|\tilde{x}| + C|\tilde{x}| \cdot \int_1^\infty e^{-ct} \int_{\mathbb{R}^d} (1 + |x'|^m) \left| \frac{\partial}{\partial \theta_i} p_t(x, x', \theta) - \frac{\partial}{\partial \theta_i} p_\infty(x', \theta) \right| dx' dt \\
&+ C|\tilde{x}| \cdot \int_1^\infty e^{-ct} \int_{\mathbb{R}^d} (1 + |x'|^m) \left| \frac{\partial}{\partial \theta_i} p_\infty(x', \theta) \right| dx' dt \\
\leq& C \left(1 + |x|^{m'} + |\tilde{x}|^{m'} \right).
\end{aligned} \tag{B.2.17}$$

Combining (B.2.8), (B.2.14), (B.2.16), (B.2.17), and DCT, we know $v^{1,2}$ is differentiable with respect to θ and for any $i \in \{1, 2, \cdots, \ell\}$

$$\begin{aligned}
\left| \frac{\partial v^{1,2}}{\partial \theta_i}(x, \tilde{x}, \theta) \right| \leq& \left| \frac{\partial}{\partial \theta_i} \mathbb{E}_{\pi_\theta} f(Y) \int_0^\infty (e^{-h(\theta)t} \tilde{x})^\top \mathbb{E}_x \nabla f\left(X_t^\theta\right) dt \right| + |\mathbb{E}_{\pi_\theta} f(Y) - \beta| \cdot (I_4 + I_5) \\
\leq& C \left(1 + |x|^{m'} + |\tilde{x}|^{m'} \right),
\end{aligned} \tag{B.2.18}$$

which together with (B.2.13) yields

$$\left| \nabla_\theta v^1(x, \tilde{x}, \theta) \right| \leq C \left(1 + |x|^{m'} + |\tilde{x}|^{m'} \right). \tag{B.2.19}$$

Similarly, by (4.3.11), (4.3.12), and (4.3.13),

$$\begin{aligned}
&\left|\int_0^\infty \nabla_x \left(\nabla_\theta \mathbb{E}_{Y\sim\pi_\theta} f(Y) - \nabla_\theta \mathbb{E}_x f(X_t^\theta)\right) dt\right| \\
=&\left|\int_0^1 \nabla_x \nabla_\theta \mathbb{E}_x f(X_t^\theta) dx' dt\right| + \left|\int_1^{+\infty} \int_{\mathbb{R}} f(x') \nabla_x \nabla_\theta p_t(x, x', \theta) dx' dt\right| \\
\leq& C\left(1 + |x|^{m'}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\left|\int_0^\infty \nabla_x \left(\nabla_\theta \mathbb{E}_x f(X_t^\theta) - \mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta) \tilde{X}_t^\theta\right) dt\right| \\
=&\left|\int_0^\infty \nabla_x \left(\mathbb{E}_x \nabla f\left(X_t^\theta\right)\right) e^{-h(\theta)t} \tilde{x} dt\right| \\
\leq& C\left(1 + |x|^{m'} + |\tilde{x}|^{m'}\right).
\end{aligned}$$

By DCT and (B.2.3),

$$\begin{aligned}
\left|\nabla_x v^{1,1}(x,\theta)\right| &= \left|(\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \beta) \int_0^\infty \int_{\mathbb{R}^d} f(x') \nabla_x \nabla_\theta p_t(x, x', \theta) dx' dt\right| \leq C\left(1 + |x|^{m'}\right), \\
\left|\nabla_x v^{1,2}(x,\tilde{x},\theta)\right| &= \left|(\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \beta) \int_0^\infty \nabla_x \left(\mathbb{E}_x \nabla f\left(X_t^\theta\right)\right) e^{-h(\theta)t} \tilde{x} dt\right| \leq C\left(1 + |x|^{m'} + |\tilde{x}|^{m'}\right).
\end{aligned} \tag{B.2.20}$$

Then, for $\nabla_x^2 v^1(x,\tilde{x},\theta)$, we have

$$\begin{aligned}
&\left|\int_0^\infty \nabla_x^2 \left(\nabla_\theta \mathbb{E}_{Y\sim\pi_\theta} f(Y) - \nabla_\theta \mathbb{E}_x f(X_t^\theta)\right) dt\right| \\
=&\left|\int_0^1 \nabla_x^2 \nabla_\theta \mathbb{E}_x f(X_t^\theta) dx' dt\right| + \left|\int_0^\infty \int_{\mathbb{R}^d} f(x') \nabla_x^2 \nabla_\theta p_t(x, x', \theta) dx' dt\right| \\
\leq& C\left(1 + |x|^{m'}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\left|\int_0^\infty \nabla_x^2 \left(\nabla_\theta \mathbb{E}_x f(X_t^\theta) - \mathbb{E}_{x,\tilde{x}} \nabla_x f(X_t^\theta) \tilde{X}_t^\theta\right) dt\right| \\
=&\left|\int_0^\infty \nabla_x^2 \left(\mathbb{E}_x \nabla f\left(X_t^\theta\right)\right) e^{-h(\theta)t} \tilde{x} dt\right| \\
\leq& C\left(1 + |x|^{m'} + |\tilde{x}|^{m'}\right).
\end{aligned}$$

By DCT and (B.2.3),

$$\begin{aligned}
\left|\nabla_x^2 v^{1,1}(x,\theta)\right| &= \left|(\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \beta) \int_0^\infty \int_{\mathbb{R}^d} f(x') \nabla_x^2 \nabla_\theta p_t(x, x', \theta) dx' dt\right| \leq C\left(1 + |x|^{m'}\right), \\
\left|\nabla_x^2 v^{1,2}(x,\tilde{x},\theta)\right| &= \left|(\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \beta) \int_0^\infty \nabla_x^2 \left(\mathbb{E}_x \nabla f\left(X_t^\theta\right)\right) e^{-h(\theta)t} \tilde{x} dt\right| \leq C\left(1 + |x|^{m'} + |\tilde{x}|^{m'}\right).
\end{aligned} \tag{B.2.21}$$

Finally, we verify that v^1 is a solution to the PDE (4.3.16). Note that

$$\begin{aligned}
\int_0^\infty \mathbb{E}_{x,\tilde{x}} \mathbb{E}_{X_s^\theta, \tilde{X}_s^\theta} \left|G^1(X_t^\theta, \tilde{X}_t^\theta, \theta)\right| dt &\overset{(a)}{=} \int_0^\infty \mathbb{E}_{x,\tilde{x}} \left|G^1(X_{t+s}^\theta, \tilde{X}_{t+s}^\theta, \theta)\right| dt \\
&\overset{(b)}{=} \int_s^\infty \mathbb{E}_{x,\tilde{x}} \left|G^1(X_t^\theta, \tilde{X}_t^\theta, \theta)\right| dt \overset{(c)}{<} \infty,
\end{aligned} \tag{B.2.22}$$

where step (a) is by the Markov property of the process $(X_\cdot^\theta, \tilde{X}_\cdot^\theta)$, step (b) by change of variables

and step (c) is by the convergence of v^1. By Fubini's theorem,

$$\mathbb{E}_{x,\tilde{x}} v^1(X_s^\theta, \tilde{X}_s^\theta, \theta) = \mathbb{E}_{x,\tilde{x}} \int_0^\infty \mathbb{E}_{X_s^\theta, \tilde{X}_s^\theta} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta) dt = \int_0^\infty \mathbb{E}_{x,\tilde{x}} \mathbb{E}_{X_s^\theta, \tilde{X}_s^\theta} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta) dt. \tag{B.2.23}$$

Combining (B.2.22) and (B.2.23), we have that

$$\begin{aligned}
&\frac{1}{s}\left[\mathbb{E}_{x,\tilde{x}} v^1(X_s^\theta, \tilde{X}_s^\theta, \theta) - v^1(x, \tilde{x}, \theta)\right] \\
=&\frac{1}{s}\left[-\int_0^\infty \mathbb{E}_{x,\tilde{x}} \mathbb{E}_{X_s^\theta, \tilde{X}_s^\theta} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta) dt + \int_0^\infty \mathbb{E}_{x,\tilde{x}} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta) dt\right] \\
=&\frac{1}{s}\left[-\int_0^\infty \mathbb{E}_{x,\tilde{x}} G^1(X_{t+s}^\theta, \tilde{X}_{t+s}^\theta, \theta) dt + \int_0^\infty \mathbb{E}_{x,\tilde{x}} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta) dt\right] \\
=&\frac{1}{s}\left[-\int_s^\infty \mathbb{E}_{x,\tilde{x}} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta) dt + \int_0^\infty \mathbb{E}_{x,\tilde{x}} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta) dt\right] \\
=&\frac{1}{s}\int_0^s \mathbb{E}_{x,\tilde{x}} G^1(X_t^\theta, \tilde{X}_t^\theta, \theta) dt,
\end{aligned} \tag{B.2.24}$$

Let $s \to 0^+$. By the definition of the infinitesimal generator and since $v^1(x, \tilde{x}, \theta)$ is twice differentiable with respect to x and once differentiable with respect to $\tilde{x}$, $v^1(x, \tilde{x}, \theta)$ is the classical solution of the Poisson PDE (4.3.16). □

Now we show the proof of Lemma 4.3.6.

Proof of Lemma 4.3.6: The proof is exactly the same as in Lemma 4.3.3 except for the presence of the dimension $\bar{x}$ and $\mathcal{L}_{\bar{x}}$. We first show that the integral in (4.3.41) converges. Note that

$$\begin{aligned}
v^2(x, \tilde{x}, \bar{x}, \theta) &= \int_0^\infty \mathbb{E}_{x,\tilde{x},\bar{x}} \left[\left(\mathbb{E}_{Y\sim\pi_\theta} f(Y) - f(\bar{X}_t^\theta)\right) \cdot \left(\nabla f(X_t^\theta) \tilde{X}_t^\theta\right)^\top\right] dt \\
&\overset{(a)}{=} \int_0^\infty \left(\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \mathbb{E}_{\bar{x}} f(\bar{X}_t^\theta)\right) \cdot \mathbb{E}_{x,\tilde{x}} \left(\nabla f(X_t^\theta) \tilde{X}_t^\theta\right)^\top dt,
\end{aligned} \tag{B.2.25}$$

where step (a) is by the independence of $\bar{X}_\cdot^\theta$ and $(X_\cdot^\theta, \tilde{X}_\cdot^\theta)$.

We now prove a uniform bound for $\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta) \tilde{X}_t^\theta$ and then by the ergodicity of $\bar{X}_\cdot^\theta$ in Lemma 4.3.1 we can show that the integrals converge.

$$\begin{aligned}
\left|\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta) \tilde{X}_t^\theta\right| &= \left|\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta) \tilde{X}_t^\theta - \nabla_\theta \mathbb{E}_x f(X_t^\theta) + \nabla_\theta \mathbb{E}_x f(X_t^\theta)\right| \\
&\overset{(a)}{\leq} \left|\mathbb{E}_x \nabla f(X_t^\theta) e^{-h(\theta)t} \tilde{x}\right| + \left|\nabla_\theta \mathbb{E}_x f(X_t^\theta)\right|,
\end{aligned} \tag{B.2.26}$$

where step (a) is by (B.2.7). Therefore, for any $t \in [0, 1]$, we can conclude

$$\left|\mathbb{E}_x \nabla f(X_t^\theta) e^{-h(\theta)t} \tilde{x}\right| + \left|\nabla_\theta \mathbb{E}_x f(X_t^\theta)\right| \leq C\left(1 + |x|^{m'} + |\tilde{x}|^{m'}\right), \tag{B.2.27}$$

where we have used Assumption 4.2.1 and equation (4.3.13). For $t > 1$, we have

$$\begin{aligned}
&\left|\mathbb{E}_x \nabla f(X_t^\theta) e^{-h(\theta)t}\tilde{x}\right| + \left|\nabla_\theta \mathbb{E}_x f(X_t^\theta)\right| \\
\overset{(a)}{\leq} & C\left(1 + \mathbb{E}_x\left|X_t^\theta\right|^m\right)\cdot|\tilde{x}| + C\int_{\mathbb{R}^d}\left(1+|x'|^m\right)\left|\nabla_\theta p_t(x,x',\theta) - \nabla_\theta p_\infty(x',\theta)\right|dx' \\
& + C\int_{\mathbb{R}^d}\left(1+|x'|^m\right)\left|\nabla_\theta p_\infty(x',\theta)\right|dx' \\
\overset{(b)}{\leq} & C\left(1+|x|^{m'}+|\tilde{x}|^{m'}\right),
\end{aligned} \tag{B.2.28}$$

where step (a) uses Assumption 4.2.1 and step (b) uses Proposition 4.3.1 and the same calculations as in (B.2.3) and (B.2.4). Combining (B.2.27) and (B.2.28), we have for any $t \geq 0$

$$\left|\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right| \leq C\left(1+|x|^{m'}+|\tilde{x}|^{m'}\right). \tag{B.2.29}$$

Thus, by (B.2.29) and the same derivation as in (B.2.4), we have

$$\begin{aligned}
\left|v^2(x,\tilde{x},\bar{x},\theta)\right| &\leq C\left(1+|x|^{m'}+|\tilde{x}|^{m'}\right)\cdot\int_0^\infty\left|\mathbb{E}_{\bar{x}} f(\bar{X}_t^\theta) - \mathbb{E}_{Y\sim\pi_\theta} f(Y)\right|dt \\
&\leq C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right).
\end{aligned} \tag{B.2.30}$$

We next show that $v^2(x,\tilde{x},\bar{x},\theta)$ is differentiable with respect to $x,\tilde{x},\bar{x},\theta$. Similar to Lemma 4.3.3, we first change the order of differentiation and integration and show the corresponding integral exists. Then, we apply DCT to prove that the differentiation and integration can be interchanged. For the ergodic process $\bar{X}_\cdot^\theta$, by (B.2.29), (4.3.11), and (4.3.13), we have the bounds

$$\begin{aligned}
&\int_0^\infty\int_{\mathbb{R}^d}\left|f(\bar{x}')\nabla_{\bar{x}} p_t(\bar{x},\bar{x}',\theta)\right|d\bar{x}'\cdot\left|\mathbb{E}_{x,\tilde{x}}\nabla f(X_t^\theta)\tilde{X}_t^\theta\right|dt \leq C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right), \\
&\int_0^\infty\int_{\mathbb{R}^d}\left|f(\bar{x}')\nabla^2_{\bar{x}} p_t(\bar{x},\bar{x}',\theta)\right|d\bar{x}'\cdot\left|\mathbb{E}_{x,\tilde{x}}\nabla f(X_t^\theta)\tilde{X}_t^\theta\right|dt \leq C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right),
\end{aligned} \tag{B.2.31}$$

and thus by the DCT

$$\sum_{i=1}^2\left|\nabla^i_{\bar{x}} v^2(x,\tilde{x},\bar{x},\theta)\right| \leq C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right). \tag{B.2.32}$$

To address $\nabla_x v^2, \nabla_x^2 v^2$, we first note that for any $i,j \in \{1,2,\cdots,d\}$

$$\begin{aligned}
\left|\nabla_x \mathbb{E}_{x,\tilde{x}}\nabla f(X_t^\theta)\tilde{X}_t^\theta\right| &\leq \left|\nabla_x \mathbb{E}_x \nabla f(X_t^\theta) e^{-h(\theta)t}\tilde{x}\right| + \left|\nabla_x\nabla_\theta \mathbb{E}_x f(X_t^\theta)\right| \\
&\overset{(a)}{\leq} C\left(1+|x|^{m'}+|\tilde{x}|^{m'}\right), \\
\left|\frac{\partial^2}{\partial x_i\partial x_j}\mathbb{E}_{x,\tilde{x}}\nabla f(X_t^\theta)\tilde{X}_t^\theta\right| &\leq \left|\frac{\partial^2}{\partial x_i\partial x_j}\mathbb{E}_x\nabla f(X_t^\theta)e^{-h(\theta)t}\tilde{x}\right| + \left|\frac{\partial^2}{\partial x_i\partial x_j}\nabla_\theta\mathbb{E}_x f(X_t^\theta)\right| \\
&\overset{(a)}{\leq} C\left(1+|x|^{m'}+|\tilde{x}|^{m'}\right),
\end{aligned} \tag{B.2.33}$$

where in step (a) we use (4.3.11) when $t > 1$ and (4.3.13) for $t \in [0,1]$. Thus we have $\forall i,j \in$

$\{1,2,\cdots,d\}$

$$\int_0^\infty \left|\left[\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \mathbb{E}_{\bar{x}} f(\bar{X}_t^\theta)\right]\right| \cdot \left|\nabla_x \mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right| dt \le C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right),$$
$$\int_0^\infty \left|\left[\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \mathbb{E}_{\bar{x}} f(\bar{X}_t^\theta)\right]\right| \cdot \left|\frac{\partial^2}{\partial x_i \partial x_j}\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right| dt \le C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right). \tag{B.2.34}$$

Then by DCT,

$$\sum_{i=1}^{2} \left|\nabla_x^i v^2(x,\tilde{x},\bar{x},\theta)\right| \le C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right). \tag{B.2.35}$$

Then for $\nabla_\theta v^2$, first we have for any $i \in \{1,2,\cdots,\ell\}$

$$\begin{aligned}
&\left|\frac{\partial}{\partial\theta_i}\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right| \\
\le& \left|\left(\frac{\partial}{\partial\theta_i}\mathbb{E}_x \nabla f(X_t^\theta)\right) e^{-h(\theta)t}\tilde{x}\right| + \left|\mathbb{E}_x \nabla f(X_t^\theta)\left(\frac{\partial}{\partial\theta_i} e^{-h(\theta)t}\right)\tilde{x}\right| + \left|\frac{\partial}{\partial\theta_i}\nabla_\theta \mathbb{E}_x f(X_t^\theta)\right| \\
\overset{(a)}{\le}& C\left(1+|x|^{m'}+|\tilde{x}|^{m'}\right),
\end{aligned} \tag{B.2.36}$$

where in step (a) we use (B.1.24) and the same analysis as in (B.2.17). Thus

$$\begin{aligned}
&\left|\int_0^\infty \frac{\partial}{\partial\theta_i}\left(\left[\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \mathbb{E}_{\bar{x}} f(\bar{X}_t^\theta)\right]\cdot \mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right) dt\right| \\
\le& \int_0^\infty \int_{\mathbb{R}^d} \left|f(x')\frac{\partial}{\partial\theta_i}\left(p_\infty(\bar{x}',\theta) - p_t(\bar{x},\bar{x}',\theta)\right)\right| d\bar{x}' \cdot \left|\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right| dt \\
&+ \int_0^\infty \int_{\mathbb{R}^d} \left|f(x')\left(p_\infty(\bar{x}',\theta) - p_t(\bar{x},\bar{x}',\theta)\right)\right| d\bar{x}' \cdot \left|\frac{\partial}{\partial\theta_i}\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right| dt \\
\le& C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right),
\end{aligned} \tag{B.2.37}$$

which together with the DCT derives

$$\left|\nabla_\theta v^2(x,\tilde{x},\bar{x},\theta)\right| \le C\left(1+|x|^{m'}+|\tilde{x}|^{m'}+|\bar{x}|^{m'}\right). \tag{B.2.38}$$

Finally, note that

$$\begin{aligned}
&\left|\int_0^\infty \nabla_{\tilde{x}}\left(\left[\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \mathbb{E}_{\bar{x}} f(\bar{X}_t^\theta)\right]\cdot \mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right) dt\right| \\
\le& \int_0^\infty \left|\left[\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \mathbb{E}_{\bar{x}} f(\bar{X}_t^\theta)\right]\right| \cdot \left|\nabla_{\tilde{x}}\mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta)\tilde{X}_t^\theta\right| dt \\
\le& C\int_0^\infty \left|\left[\mathbb{E}_{Y\sim\pi_\theta} f(Y) - \mathbb{E}_{\bar{x}} f(\bar{X}_t^\theta)\right]\right| \cdot \left|\mathbb{E}_x \nabla f(X_t^\theta) e^{-h(\theta)t}\right| dt \\
\le& C\left(1+|x|^{m'}+|\bar{x}|^{m'}\right)
\end{aligned} \tag{B.2.39}$$

and then by DCT

$$\left|\nabla_{\tilde{x}} v^2(x,\tilde{x},\bar{x},\theta)\right| \le C\left(1+|x|^{m'}+|\bar{x}|^{m'}\right). \tag{B.2.40}$$

By the same calculations as in (B.2.24), we know v^2 is the classical solution of PDE (4.3.42) and the bound (4.3.44) holds. □

Bibliography

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

[2] Yacine Aït-Sahalia, Chenxu Li, and Chen Xu Li. Maximum likelihood estimation of latent markov models using closed-form approximations. *Journal of Econometrics*, 2020.

[3] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

[4] Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[5] Mario Annunziato and Alfio Borzì. A fokker–planck control framework for multidimensional stochastic processes. *Journal of Computational and Applied Mathematics*, 237(1):487–507, 2013.

[6] Ari Arapostathis, Vivek S Borkar, and Mrinal K Ghosh. *Ergodic control of diffusion processes*, volume 143. Cambridge University Press, 2012.

[7] C. Arribas I.P., Salvi and L. Szpruch. Sig-sdes model for quantitative finance. *In Proceedings of the First ACM International Conference on AI in Finance*, 2020.

[8] Martino Bardi and Fabio S Priuli. Linear-quadratic n-person and mean-field games with ergodic cost. *SIAM Journal on Control and Optimization*, 52(5):3022–3052, 2014.

[9] David S Bates. Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *The Review of Financial Studies*, 9(1):69–107, 1996.

[10] Claudio Bellani, Damiano Brigo, Mikko Pakkanen, and Leandro Sanchez-Betancourt. Non-average price impact in order-driven markets. *arXiv preprint arXiv:2110.00771*, 2021.

[11] Albert. Benveniste, Michel. Metivier, and Pierre. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Stochastic Modelling and Applied Probability, 22. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 1990. edition, 1990.

[12] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

[13] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

[14] Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[15] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[16] Shalabh Bhatnagar, Mohammad Ghavamzadeh, Mark Lee, and Richard S Sutton. Incremental natural actor-critic algorithms. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[17] Theerawat Bhudisaksang and Álvaro Cartea. Online drift estimation for jump-diffusion processes. *Bernoulli*, 27(4):2494–2518, 2021.

[18] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[19] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:5–43, 2006.

[20] V. S. Borkar and S. P. Meyn. The o.d.e. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

[21] Vivek S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[22] Vivek S. Borkar. Asynchronous stochastic approximations. *SIAM Journal on Control and Optimization*, 36(3):840–851, 1998.

[23] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[24] Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint Second Edition*. Texts and Readings in Mathematics, 48. Hindustan Book Agency, Gurgaon, 1st ed. 2022. edition, 2022.

[25] Vivek S Borkar and Vijaymohan R Konda. The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22(4):525–543, 1997.

[26] Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

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

[28] MM Butt. Numerical solution to 3d bilinear fokker–planck control problem. *International Journal of Computer Mathematics*, 99(12):2466–2481, 2022.

[29] Haoyang Cao, Jodi Dianetti, and Giorgio Ferrari. Stationary discounted and ergodic mean field games with singular controls. *Mathematics of Operations Research*, 2022.

[30] Pierre Cardaliaguet and Cristian Mendico. Ergodic behavior of control and mean field games problems depending on acceleration. *Nonlinear Analysis*, 203:112185, 2021.

[31] René Carmona. *Lectures on BSDEs, stochastic control, and stochastic differential games with financial applications*. SIAM, 2016.

[32] Rene Carmona, Jean-Pierre Fouque, and Li-Hsien Sun. Mean field games and systemic risk. *arXiv preprint arXiv:1308.2172*, 2013.

[33] René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games i: the ergodic case. *SIAM Journal on Numerical Analysis*, 59(3):1455–1485, 2021.

[34] René Carmona and Mathieu Laurière. Deep learning for mean field games and mean field control with applications to finance. *arXiv preprint arXiv:2107.04568*, 2021.

[35] Alvaro Cartea and Sebastian Jaimungal. Incorporating order-flow into optimal execution. *Mathematics and Financial Economics*, 10:339–364, 2016.

[36] Álvaro Cartea, Sebastian Jaimungal, and José Penalva. *Algorithmic and high-frequency trading*. Cambridge University Press, 2015.

[37] George Casella. Empirical bayes gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

[38] Dotan Di Castro and Ron Meir. A convergent online single time scale actor critic algorithm. *Journal of Machine Learning Research*, 11(11):367–410, 2010.

[39] Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm, 2022.

[40] Ricky TQ Chen, Brandon Amos, and Maximilian Nickel. Neural spatio-temporal point processes. *arXiv preprint arXiv:2011.04583*, 2020.

[41] Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022.

[42] Peter Christoffersen, Steven Heston, and Kris Jacobs. The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well. *Management Science*, 55(12):1914–1932, 2009.

[43] Samuel Chun-Hei Lam, Justin Sirignano, and Ziheng Wang. Weak convergence analysis of online neural actor-critic algorithms. *arXiv e-prints*, pages arXiv–2403, 2024.

[44] Samuel N Cohen, Christoph Reisinger, and Sheng Wang. Arbitrage-free neural-sde market models. *arXiv preprint arXiv:2105.11053*, 2021.

[45] Samuel N Cohen, Christoph Reisinger, and Sheng Wang. Estimating risks of option books using neural-sde market models. *arXiv preprint arXiv:2202.07148*, 2022.

[46] Samuel N Cohen, Christoph Reisinger, and Sheng Wang. Hedging option books using neural-sde market models. *arXiv preprint arXiv:2205.15991*, 2022.

[47] Pierre Collin-Dufresne and Robert S Goldstein. Do bonds span the fixed income markets? theory and evidence for unspanned stochastic volatility. *The Journal of Finance*, 57(4):1685–1730, 2002.

[48] Pierre Collin-Dufresne, Christopher Jones, and Robert Goldstein. Can interest rate volatility be extracted from the cross section of bond yields? an investigation of unspanned stochastic volatility, 2004.

[49] Drew D Creal and Jing Cynthia Wu. Estimation of affine term structure models with spanned or unspanned stochastic volatility. *Journal of Econometrics*, 185(1):60–81, 2015.

[50] Shweta Dahale, Sai Munikoti, and Balasubramaniam Natarajan. A general framework for uncertainty quantification via neural sde-rnn. *arXiv preprint arXiv:2306.01189*, 2023.

[51] Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233. PMLR, 2018.

[52] A Davie and J Gaines. Convergence of numerical schemes for the solution of parabolic stochastic partial differential equations. *Mathematics of Computation*, 70(233):121–134, 2001.

[53] Raghuram Bharadwaj Diddigi, Prateek Jain, Prabuchandran K. J, and Shalabh Bhatnagar. Neural network compatible off-policy natural actor-critic algorithm. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2022.

[54] Darrell Duffie, Jun Pan, and Kenneth Singleton. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6):1343–1376, 2000.

[55] Tyrone E Duncan, Lei Guo, and Bozenna Pasik-Duncan. Adaptive continuous-time linear quadratic gaussian control. *IEEE Transactions on automatic control*, 44(9):1653–1662, 1999.

[56] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[57] Wei Fang and Michael B Giles. Adaptive euler-maruyama method for sdes with non-globally lipschitz drift: Part ii, infinite time interval. *arXiv preprint arXiv:1703.06743*, 2017.

[58] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

[59] Arthur Fleig and Roberto Guglielmi. Optimal control of the fokker–planck equation with space-dependent controls. *Journal of Optimization Theory and Applications*, 174:408–427, 2017.

[60] Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

[61] P. Gierjatowicz, M. Sabate-Vidales, D. Siska, L. Szpruch, and Z. Zuric. Robust pricing and hedging via neural sdes. *arXiv preprint arXiv:2007.04154*, 2020.

[62] David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*, volume 224. springer, 2015.

[63] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[64] Harsh Gupta, Rayadurgam Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

[65] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

[66] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[67] Steven L Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, 6(2):327–343, 1993.

[68] Peter D Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009.

[69] Yoshifusa Ito. Nonlinearity creates linear independence. *Advances in Computational Mathematics*, 5(1):189–203, 1996.

[70] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[71] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[72] Kexin Jin, Jonas Latz, Chenguang Liu, and Carola-Bibiane Schönlieb. A continuous-time stochastic gradient descent method for continuous data. *Journal of Machine Learning Research*, 24(274):1–48, 2023.

[73] Yuri Kabanov, Robert Liptser, Jordan Stoyanov, Aureli Alabert, and István Gyongy. On numerical approximation of stochastic burgers' equation. *From Stochastic Calculus to Mathematical Finance: The Shiryaev Festschrift*, pages 1–15, 2006.

[74] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

[75] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[76] Barbara Kaltenbacher and Barbara Pedretscher. Parameter estimation in sdes via the fokker–planck equation: Likelihood function and adjoint based gradient computation. *Journal of Mathematical Analysis and Applications*, 465(2):872–884, 2018.

[77] Sajad Khodadadian, Zaiwei Chen, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic algorithm. In *International Conference on Machine Learning*, pages 5420–5431. PMLR, 2021.

[78] Sajad Khodadadian, Thinh T Doan, Siva Theja Maguluri, and Justin Romberg. Finite sample analysis of two-time-scale natural actor-critic algorithm. *arXiv preprint arXiv:2101.10506*, 2021.

[79] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*, 2021.

[80] Yerkin Kitapbayev and Tim Leung. Mean reversion trading with sequential deadlines and transaction costs. *International Journal of Theoretical and Applied Finance*, 21(01):1850004, 2018.

[81] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer, 2000.

[82] Vijay R. Konda and John N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

[83] Vijaymohan R. Konda. Actor-critic algorithms (ph. d. thesis). *Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology*, 2002.

[84] Vijaymohan R. Konda and Vivek S. Borkar. Actor-critic type learning algorithms for markov decision processes. *SIAM Journal on Control and Optimization*, 38(1):94–123, 1999.

[85] Vijaymohan R. Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[86] Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine learning*, 2023.

[87] P. Kumar. Deep hawkes process for high-frequency market making. *arXiv preprint arXiv:2109.15110*, 2021.

[88] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[89] Gregory F Lawler. *Introduction to stochastic processes*. Chapman and Hall/CRC, 2018.

[90] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[91] Charles-Albert Lehalle and Eyal Neuman. Incorporating signals into optimal trading. *Finance and Stochastics*, 23:275–311, 2019.

[92] Tim Leung, Jiao Li, Xin Li, and Zheng Wang. Speculative futures trading under mean reversion. *Asia-Pacific Financial Markets*, 23:281–304, 2016.

[93] Tim Leung and Xin Li. Optimal mean reversion trading with transaction costs and stop-loss exit. *International Journal of Theoretical and Applied Finance*, 18(03):1550020, 2015.

[94] Tim Siu-tang Leung and Xin Li. *Optimal mean reversion trading: Mathematical analysis and practical applications*, volume 1. World Scientific, 2015.

[95] Jian-Guo Liu, Ziheng Wang, Yantong Xie, Yuan Zhang, and Zhennan Zhou. Investigating the integrate and fire model as the limit of a random discharge model: a stochastic analysis perspective. *Mathematical Neuroscience and Applications*, 1, 2021.

[96] Jian-Guo Liu, Ziheng Wang, Yuan Zhang, and Zhennan Zhou. Rigorous justification of the fokker–planck equations of neural networks based on an iteration perspective. *SIAM Journal on Mathematical Analysis*, 54(1):1270–1312, 2022.

[97] Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

[98] Wei Liu and Michael Röckner. *Stochastic partial differential equations: an introduction*. Springer, 2015.

[99] X. Lu and F. Abergel. High-dimensional hawkes processes for limit order books: modelling, empirical analysis and numerical calibration. *Quantitative Finance*, 18(2):249–264, 2018.

[100] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[101] Alex McNabb. Comparison theorems for differential equations. *Journal of mathematical analysis and applications*, 119(1-2):417–428, 1986.

[102] H. Mei and J. Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.

[103] Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[104] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[105] Francisco S Melo. Convergence of q-learning: A simple proof. *Institute Of Systems and Robotics, Tech. Rep*, pages 1–4, 2001.

[106] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[107] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016.

[108] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature (London)*, 518(7540):529–533, 2015.

[109] M. Morariu-Patrichi and M. Pakkanen. State-dependent hawkes processes and their application to limit order book modelling. *Quantitative Finance*, 22(3):563–583, 2022.

[110] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[111] H. Ni, L. Szpruch, M. Sabate-Vidales, B. Xiao, M. Wiese, and S. Liao. Sig-wasserstein gans for time series generation. *In Proceedings of the Second ACM International Conference on AI in Finance*, 2021.

[112] James R Norris, John Robert Norris, and James Robert Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

[113] E Pardoux and A Yu Veretennikov. On the poisson equation and diffusion approximation. i. *Annals of probability*, pages 1061–1085, 2001.

[114] E Pardoux and A Yu Veretennikov. On poisson equation and diffusion approximation 2. *The Annals of Probability*, 31(3):1166–1192, 2003.

[115] Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

[116] Grigorios A Pavliotis and AM Stuart. Parameter estimation for multiscale diffusions. *Journal of Statistical Physics*, 127(4):741–781, 2007.

[117] Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7):1180–1190, 2008. Progress in Modeling, Theory, and Application of Computational Intelligenc.

[118] Huyên Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61. Springer Science & Business Media, 2009.

[119] Claudia Prévôt and Michael Röckner. *A concise course on stochastic partial differential equations*, volume 1905. Springer, 2007.

[120] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[121] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[122] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[123] Michael Röckner, Xiaobin Sun, and Yingchao Xie. Strong convergence order for slow–fast mckean–vlasov stochastic differential equations. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 57, pages 547–576. Institut Henri Poincaré, 2021.

[124] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[125] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[126] Louis Sharrock and Nikolas Kantas. Two-timescale stochastic gradient descent in continuous time with applications to joint online parameter estimation and optimal sensor placement. *arXiv preprint arXiv:2007.15998*, 2020.

[127] Louis Sharrock and Nikolas Kantas. Joint online parameter estimation and optimal sensor placement for the partially observed stochastic advection-diffusion equation. *SIAM/ASA Journal on Uncertainty Quantification*, 10(1):55–95, 2022.

[128] Louis Sharrock, Nikolas Kantas, Panos Parpas, and Grigorios A Pavliotis. Parameter estimation for the mckean-vlasov stochastic differential equation. *arXiv preprint arXiv:2106.13751*, 2021.

[129] Z. Shi and J. Cartlidge. State dependent parallel neural hawkes process for limit order book event stream prediction and simulation. *In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1607–1615, 2022.

[130] Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time. *SIAM Journal on Financial Mathematics*, 8(1):933–961, 2017.

[131] Justin Sirignano and Konstantinos Spiliopoulos. Asymptotics of reinforcement learning with neural networks. *arXiv preprint arXiv:1911.07304*, 2019.

[132] Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time: A central limit theorem. *Stochastic Systems*, 10(2):124–151, 2020.

[133] Justin Sirignano and Konstantinos Spiliopoulos. Asymptotics of reinforcement learning with neural networks. *Stochastic Systems*, 2021.

[134] Justin Sirignano and Konstantinos Spiliopoulos. Online adjoint methods for optimization of pdes. *arXiv preprint arXiv:2101.09621*, 2021.

[135] Simone Carlo Surace and Jean-Pascal Pfister. Online maximum-likelihood estimation of the parameters of partially observed diffusion processes. *IEEE transactions on automatic control*, 64(7):2814–2829, 2018.

[136] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[137] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[138] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer, 1999.

[139] Csaba Szepesvári. The asymptotic convergence-rate of q-learning. *Advances in neural information processing systems*, 10, 1997.

[140] Alain-Sol Sznitman. Topics in propagation of chaos. *Ecole d'été de probabilités de Saint-Flour XIX—1989*, 1464:165–251, 1991.

[141] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

[142] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

[143] Ziheng Wang and Justin Sirignano. Global convergence of the ode limit for online actor-critic algorithms in reinforcement learning. *arXiv preprint arXiv:2108.08655*, 2021.

[144] Ziheng Wang and Justin Sirignano. Continuous-time stochastic gradient descent for optimizing over the stationary distribution of stochastic differential equations. *arXiv preprint arXiv:2202.06637*, 2022.

[145] Ziheng Wang and Justin Sirignano. A forward propagation algorithm for online optimization of nonlinear stochastic differential equations. *arXiv preprint arXiv:2207.04496*, 2022.

[146] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

[147] Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

[148] Yue Frank Wu, Weitong ZHANG, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17617–17628. Curran Associates, Inc., 2020.

[149] Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020.

[150] Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*, 2019.

[151] Zhuoran Yang, Kaiqing Zhang, Mingyi Hong, and Tamer Başar. A finite sample analysis of the actor-critic algorithm. In *2018 IEEE conference on decision and control (CDC)*, pages 2759–2764. IEEE, 2018.

[152] Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999.

[153] Jize Zhang, Tim Leung, and Aleksandr Aravkin. Mean reverting portfolios via penalized ou-likelihood estimation. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5795–5800. IEEE, 2018.