On the Fast Convergence of Unstable Reinforcement Learning Problems Supplementary Material, ICLR 2023

ORGANIZATION OF APPENDIX

In the Appendix, we provided experiments, proofs and discussions to support the main paper. Here is the organization:

- Appendix A gives an example to distinguish between I/O and IS stability.
- Appendix B compares the gradient normalization vs logarithmic mapping.
- Appendix C shows extra LQR experiments under different instabilities.
- Appendix D shows other experiments under unstable RL problems with neural network based policy parameterization.
- Appendix E provides proofs supporting the theoretical analysis in the main paper.

A EXAMPLE ON STABILITY DEFINITION

We use the illustrative example from Sontag (2008). Consider a n dimension linear system $\dot{x} = Ax + Bu$ where $A \in \mathbb{R}^{n\times n}$ being full rank matrix, $x \in \mathbb{R}^n$ with initial condition $x(0) = x_0$, $B \in \mathbb{R}^{n\times m}$ and $u = u(t) \in \mathbb{R}^m$. By solving inhomogeneous ODE, the solution is

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau.$$

Lemma A.1. *The system is ISS if all the eigenvalues of A are strictly negative.*

Proof. let $\beta(x_0, t)$ be $\|e^{At}\|\|x_0\|$ and $\gamma(x_0)$ be $\|B\|\int_0^\infty \|e^{A\tau}\|\,\mathrm{d}\tau$. With all the eigenvalues of A being strictly negative, both $\|e^{At}\|$ and $\|B\|\int_0^\infty \|e^{A\tau}\|\,\mathrm{d}\tau$ are bounded. $\|x(t, x_0, u)\| \le \|e^{At}\|\|x_0\| + \|B\|\int_0^\infty \|e^{A\tau}\|\,\mathrm{d}\tau\|u\|_\infty$, therefore satisfies 5. □

Consider $y(x) := x$ itself, the system is I/O stable. While if we take $y(x) := \frac{1}{\|x\|}$, then the system is ISS but not I/O stable with $y \to \infty$ with $x_0, u \to 0$.

Now suppose A has non-negative eigenvalues and AB is not an empty matrix, then $\int_0^t e^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau$ is not bounded by γ function since $\|B\|\int_0^\infty \|e^{A\tau}\|\,\mathrm{d}\tau$ when $t \to \infty$, which also means the effect of previous action u will grow or at least not vanish along the time trajectory. Then the system is not ISS. But the system could be I/O stable if we take trivial output like $y(\cdot) := 0$.

In this paper, we consider I/O stability, regardless of the problem being ISS or not. While in many of the real-world RL applications such as target tracking, the output function $y(\cdot)$ is correlated to the norm of x such as using distance to target as cost function. In this case, I/O stability is dependent on ISS. Specifically, for LQR problems, the eigenvalues of the system matrix A determine the system ISS and also I/O stability (discrete LQR requires the eigenvalue within the unit circle and continuous LQR requires eigenvalues to the left half-plane). Therefore, in the discrete LQR experiments, we use matrix A to manipulate I/O stability. Since we are dealing with I/O stability, RL scenarios with ISS but not I/O stable system is beyond the scope of this paper, for instance, an unstable invert pendulum problem with cost clipped to $[0, 1]$.

B COMPARING LOGARITHMIC MAPPING WITH GRADIENT NORMALIZATION

Normalizing gradient is a classical approach to speed up the convergence, where we have the follow update step:

$$\theta \leftarrow \theta - \eta \frac{\nabla_\theta V_T(\theta)}{\|\nabla_\theta V_T(\theta)\|},$$

Compared with logarithmic mapping, the gradient normalization has similar theoretical performance in deterministic case by controlling the spectral radius of the optimization step. In the stochastic case, the updating step consists of a summation of gradient over the mini-batch followed by a normalization process. In logarithmic mapping, the log function is applied on individual examples ahead of summation. Therefore, the outliers with relatively large noise can be "normalized" to prevent them from dominating sampling summation. Besides, the portion of unstable examples with large loss are expected to drop during optimization, it is necessary to map the exponentially growing effect of these unstable cases into linear forms. In practice, our logarithmic mapping outperforms the gradient normalization in the convergence speed, as shown in the experiment section (Section 4).

Figure 2 are the comparisons between logarithmic mapping and normalizing gradient, where learning rate $1e$-1 and $1e0$ will crash their optimization respectively. The plots of $\eta = 1e$-2 for logarithmic mapping and $\eta = 1e$-1 for normalizing gradient effectively show similar convergence rate with minor fluctuation at the beginning. The logarithmic mapping eventually reaches a slightly better performance due to normalizing gradient's trapping in the local minimum. Noticeably, if both methods are coupled, the initial fluttering disappears and plots are smoother. We also tested using logarithmic mapping ahead of the average of each trajectory, where $loss = log(1/b \sum_j [v_T(s_j, \theta)])$, instead of following Equation (12). The results are similar.

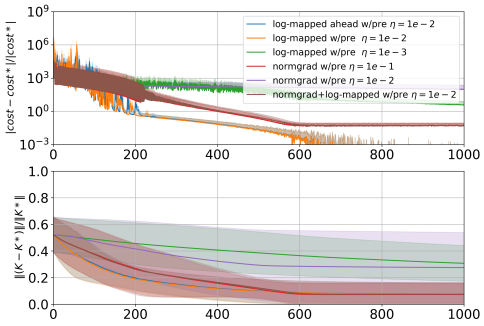

Figure 2: LQR cost difference to optimal: normalizing gradient vs log mapping, $\rho_{max}(A) = 5$

C MORE EXPERIMENTS OF UNSTABLE LQR WITH DIFFERENT SPECTRAL RADIUS

We include additional results for unstable LQR in Figure 3 both with pre-process enabled. $\rho_{max}(A) = 2$ is a relatively moderate case, the vanilla method could use a learning rate of $\eta = 1e - 4$ and slowly converge to optimal. In $\rho_{max}(A) = 10$ case, the vanilla method crashes for $\eta > 1\text{e} - 11$ and the optimization stagnates for $\eta = 1\text{e} - 12$. The logarithmic mapping has similar performance in $\rho_{max}(A) = 2$ case and converges faster than the latter in $\rho_{max}(A) = 10$ case.

D GENERAL UNSTABLE RL

Figure 4 shows 3 customized unstable environments: unstable cart-pole, unstable mountain car and Planar Vertical Take-off and Landing (PVTOL) aircraft. We use a single hidden layer neural network with 64 hidden neurons and ReLU activation functions. The input layer and output layer has the same dimension of environment state and action space respectively. Similar to LQR experiments, we

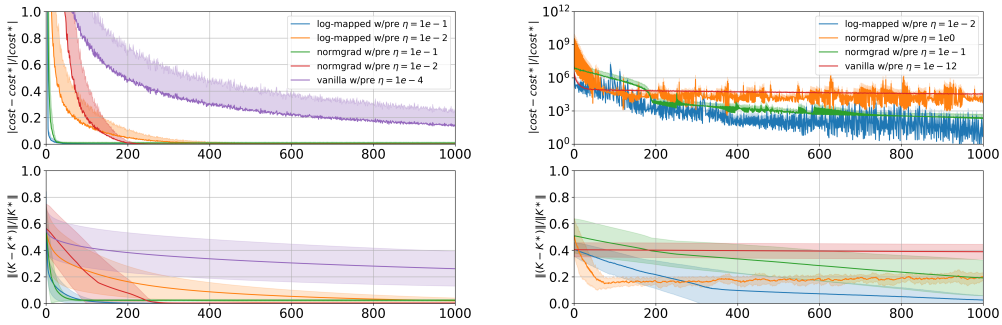

Figure 3: LQR cost difference to optimal, left: $\rho_{max}(A) = 2$, right: $\rho_{max}(A) = 10$

search a largest learning rate without crashing the optimization. Each experiment is performed under 3 random seeds. The lower half of variance is omitted for visualization in log-scale plots.

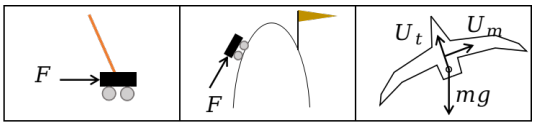

Figure 4: Unstable RL examples: modified cart-pole, modified mountain car, PVTOL aircraft

D.1 Modified cart-pole

Compared with standard cart-pole problem from OpenAI Gym package, we use a continuous force input and enlarged its force magnitude to introduce more instability to the input-output system (a small amount of control feedback could dramatically change the system behavior). Besides, we allow the agent to simulate fixed 20 time steps instead of terminating the episode if the agent runs into an undesired zone. The cost function is defined in the quadratic form of the distance between current state towards target position, instead of using the 0/1 reward depending on whether the episode is done or not.

Figure 5 shows the cost against epochs for cart-pole problem with and without pre-process. For vanilla loss without pre-process, $\eta = 1e\text{-}8$ is the maximum allowed learning rate and there is a significant difference in convergence speed compared with other two. The logarithmic mapped cost is higher but close to vanilla loss with normalizing gradient. With a pre-processed policy, the system is more stable at the beginning and therefore larger learning rates are allowed. All three methods could reach the optimal. To remark on the cart-pole problem, the instability mostly comes from the large force magnitude instead of the unbounded state space because there exists local equilibrium when the pole sticks downward. Therefore, compared to the following 2 environments, it is less challenging and could be addressed with vanilla loss with a simple pre-process.

D.2 PVTOL Aircraft

The Planar Vertical Take-off and Landing (PVTOL) aircraft (Lin et al., 1999) is a simplified 2D model of realistic aircraft maneuver. The aircraft state includes the lateral/vertical displacement of the gravity center and roll angle. The control feedback U_t and U_m are longitudinal thrust and lateral rolling force. Notice U_m provides both force and rolling moment to the airplane. The target is to control and airplane to a certain state and the cost function is also in the quadratic form.

Similar to the unstable mountain car example, both vanilla loss and normalizing gradient show slow convergence when pre-process is not engaged. The logarithmic mapping is capable to achieve optimal results regardless of the pre-treatment.

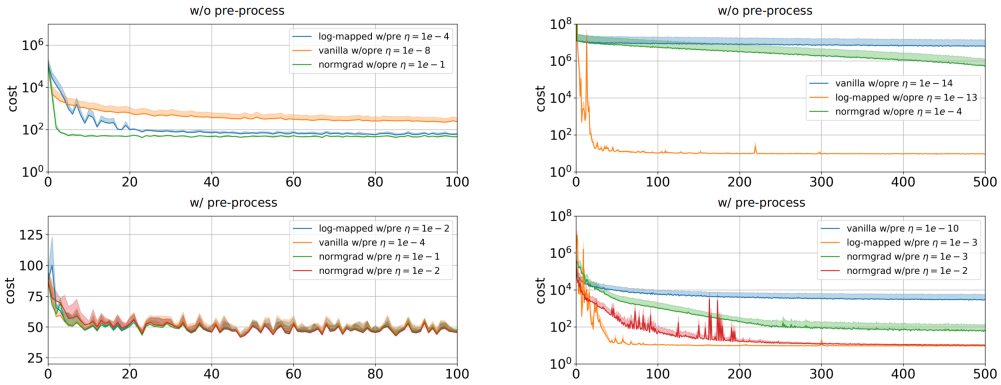

Figure 5: Cost loss: left: unstable cart-pole, right: PVTOL

D.3 MODIFIED MOUNTAIN CAR

Similar to the cart-pole treatment, we remove the terminal conditions and re-define the cost function in the quadratic form. The control target is to drive the car to a certain location and stabilize it. We manipulate a steep slope by adding an acceleration term proportional to the cube of horizontal displacement from the peak, and there does not exist any local equilibrium point.
The results are shown in Figure 6. Both vanilla loss and normalizing gradient require small learning rate, the logarithmic mapping outperforms the other two methods. When pre-processing is engaged, the vanilla loss still converges slowly, the other two methods share similar performance and achieve a smaller cost compared with the optimal results without the pre-process.

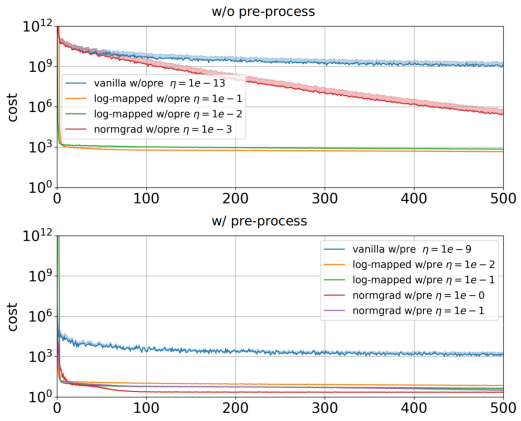

Figure 6: Cost loss: Unstable mountain car

E MORE THEORETICAL RESULTS ON UNSTABLE RL

Lemma 3.7 (Restated)**.** Update the value function $V_T(\theta)$ by policy gradient method with $\theta \leftarrow \theta - \eta \nabla_\theta V_T(\theta)$, choose step size $\eta < 2/\max_\xi \rho_{max}(J_T(\theta - \xi\eta\nabla J_T(\theta)))$ for $\xi \in [0, 1]$, then $V_T(\theta)$ is monotonically decreasing..

Proof. Let $\xi \in [0,1]$ be a scalar, denote $s = -\eta \nabla V_T(\theta), g(\xi) = V_T(\theta + \xi s)$, we have

$$\begin{aligned} V_T(\theta + \xi s) - V_T(\theta) = g(1) - g(0) &= \int_0^1 \frac{dg}{d\xi}\, d\xi \\ &= \int_0^1 s^\top \nabla V_T(\theta + \xi s)\, d\xi \\ &\leq \int_0^1 s^\top \nabla V_T(\theta)\, d\xi + |\int_0^1 s^\top (\nabla V_T(\theta) - \nabla V_T(\theta + \xi s))\, d\xi| \\ &\leq s^\top \nabla V_T(\theta) + \int_0^1 \|s\| \|(\nabla V_T(\theta) - \nabla V_T(\theta + \xi s))\|\, d\xi \\ &\leq s^\top \nabla V_T(\theta) + \|s\|^2 \max_\xi \rho_{max}(J_T(\theta + \xi s))/2. \end{aligned}$$

substitute $s = -\eta \nabla V_T(\theta)$ into the equation, we have

$$V_T(\theta + \xi s) - V_T(\theta) \leq -\eta(1 - \frac{\eta}{2} \max_\xi \rho_{max}(J_T(\theta - \xi \eta \nabla J_T(\theta))))\|\nabla V_T(\theta)\|^2 < 0.$$

when $(1 - \frac{\eta}{2} \max_\xi \rho_{max}(J_T(\theta - \eta \nabla J_T(\theta))))$ term is negative.

□

Theorem 3.9 (Restated). *If $V_T(\theta)$ satisfies Assumption 3.2 and Assumption 3.4, using the vanilla gradient descent algorithm from Equation* (8)*, then $\rho_{max}(J_T(\theta)) < \sum_{k=1}^m d_k [L_1 (\sum_{t=0}^T t \phi_k(\theta)^{t-1}) + {L_2}^2 (\sum_{t=0}^T t(t-1)\phi_k(\theta)^{t-2}]$*

Proof.

$$\begin{aligned} \nabla_\theta V_T(\theta) &= \sum_{k=1}^m d_k (\sum_{t=0}^T t \phi_k(\theta)^{t-1}) \frac{\partial \phi_k(\theta)}{\partial \theta}, \\ \text{Jacobian } J_T(\theta) &= \nabla_\theta^2 V_T(\theta) \\ &= \sum_{k=1}^m J_T^k(\theta). \end{aligned}$$

where

$$\begin{aligned} J_T^k(\theta) = d_k [(\sum_{t=0}^T t \phi_k(\theta)^{t-1}) \frac{\partial^2 \phi_k(\theta)}{\partial \theta^2} \\ + (\sum_{t=0}^T t(t-1)\phi_k(\theta)^{t-2}) \frac{\partial \phi_k(\theta)}{\partial \theta} \frac{\partial \phi_k(\theta)}{\partial \theta}^\top]. \end{aligned} \tag{14}$$

Denote the eigenvalues of $\frac{\partial^2 \phi_k(\theta)}{\partial \theta^2}$, $\frac{\partial \phi_k(\theta)}{\partial \theta} \frac{\partial \phi_k(\theta)}{\partial \theta}^\top$, $J_T^k(\theta)$, as $\mu_1^k > ... > \mu_n^k$, $\upsilon_1^k > ... > \upsilon_n^k$, $\nu_1^k > ... > \nu_n^k$ respectively. Notice $\frac{\partial \phi_k(\theta)}{\partial \theta} \frac{\partial \phi_k(\theta)}{\partial \theta}^\top$ is positive semi-definite and has same non-zero eigenvalue with $\frac{\partial \phi_k(\theta)}{\partial \theta}^\top \frac{\partial \phi_k(\theta)}{\partial \theta}$, then $\upsilon_1^k \leq {L_2}^2$ by Lipschitz condition. To bound the eigenvalues of $\frac{\partial^2 \phi_k(\theta)}{\partial \theta^2}$,

$$\begin{aligned} |\mu_i^k| &\leq \|\frac{\partial^2 \phi_k(\theta)}{\partial \theta^2} v\| / \|v\| \\ &= \lim_{h \to 0} \frac{\|\nabla \phi_k(\theta + hv) - \nabla \phi_k(\theta)\|}{|h| \|v\|} \\ &\leq \frac{L_1 \|hv\|}{|h| \|v\|} \qquad \leq L_1. \end{aligned}$$

Notice $\frac{\partial^2 \phi_k(\theta)}{\partial \theta^2}$ and $\frac{\partial \phi_k(\theta)}{\partial \theta} \frac{\partial \phi_k(\theta)}{\partial \theta}^\top$ are Hermitian, by Weyl's inequality to bound:

$$\nu_1^k \leq d_k[L_1(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}) + {L_2}^2(\sum_{t=0}^{T} t(t-1)\phi_k(\theta)^{t-2})],$$
$$\nu_n^k \geq d_k[-L_1(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1})].$$

$$\begin{aligned}\rho_{max}(J_T^k(\theta)) &= \max(|\nu_1^k|, |\nu_n^k|) \\ &\leq d_k[L_1(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}) + {L_2}^2(\sum_{t=0}^{T} t(t-1)\phi_k(\theta)^{t-2})].\end{aligned}$$

$$\rho_{max}(J_T(\theta)) \leq \sum_{k=1}^{m} d_k[L_1(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}) + {L_2}^2(\sum_{t=0}^{T} t(t-1)\phi_k(\theta)^{t-2})].$$

□

Theorem 3.10 (Restated). *Suppose $V_T(\theta)$ satisfies Assumption 3.2 and Assumption 3.4, using the vanilla gradient descent algorithm from Equation* (8)*, if $\eta < 1/\sum_{k=1}^{m} d_k[L_1(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}) + {L_2}^2(\sum_{t=0}^{T} t(t-1)\phi_k(\theta)^{t-2}]$, then the update step requirement in Lemma 3.7 for monotonic decrease of value function is satisfied.*

The proof is completed by substituting Theorem 3.9 into Lemma 3.7 and taking $\xi = 0$.

Lemma E.1. *If function $f(x) : \mathbb{R}^m \to \mathbb{R}+$ is L_1 smooth and L_2 Lipschitz and non-negative for $x \in \mathbb{S} \subset \mathbb{R}^m$, then its polynomial $f(x)^n$ is $(n\overline{f_{\mathbb{S}}}^{n-1} L_1 + n(n-1)\overline{f_{\mathbb{S}}}^{n-2} {L_2}^2)$ smooth on $\mathbb{S}$, where $\overline{f_{\mathbb{S}}} = \max_{x \in \mathbb{S}}[f(x)]$*

Proof. With function $f(x)$ being L_1 smooth, equivalently

$$\begin{aligned}&\|\nabla f(x) - \nabla f(y)\| \leq L_1 \|x - y\| \\ \Longleftrightarrow &g(x) = \frac{L_1}{2} x^\top x - f(x) \text{ is convex} \\ \Longleftrightarrow &L_1 I \succeq \frac{\partial^2 f(x)}{\partial x^2}.\end{aligned}$$

With function $f(x)$ being L_2 Lipschitz, $\|\nabla f(x)\| \leq L_2$,

for polynomial $f(x)^n$,

$$\begin{aligned}\frac{\partial^2 [f(x)^n]}{\partial x^2} &= nf(x)^{n-1} \frac{\partial^2 f(x)}{\partial x^2} + n(n-1) f(x)^{n-2} \nabla f(x) \nabla f(x)^\top \\ &\preceq (n\overline{f_{\mathbb{S}}}^{n-1} L_1 + n(n-1)\overline{f_{\mathbb{S}}}^{n-2} {L_2}^2) I.\end{aligned}$$

where the ${L_2}^2$ term comes from the fact that $\nabla f(x) \nabla f(x)^\top$ has same non-zero eigenvalue with $\nabla f(x)^\top \nabla f(x)$.

Therefore, $f(x)^n$ is locally $(n\overline{f_{\mathbb{S}}}^{n-1} L_1 + n(n-1)\overline{f_{\mathbb{S}}}^{n-2} {L_2}^2)$ smooth on the support $\mathbb{S}$. □

Lemma E.2.

$$\frac{V_T(\theta) - V_T(\theta_*)}{\|\nabla_\theta V_T(\theta)\|^2} \geq \frac{1}{2L'}.$$

where

$$L' = \sum_{k=1}^{m} d_k \sum_{t=0}^{T} [t\overline{\phi_k}(\theta)^{t-1} L_1 + t(t-1)\overline{\phi_k}(\theta)^{t-2} {L_2}^2].$$

where

$$\overline{\phi_k}(\theta) = \max_{\xi} [\phi_k(\theta_* + \xi(\theta - \theta_*))] \textit{ for } \xi \in [0, 1].$$

Proof. With Assumption 3.4 on $\phi_k(\theta)$, apply Lemma E.1 on the straight line from θ to θ_*, $V_T(\theta) \sim \sum_{k=1}^{m} d_k \sum_{t=0}^{T} \phi_k(\theta)^t$ is L' smooth on the straight line.

$$\begin{aligned} V_T(\theta_*) &\leq \min_{\xi} V_T(\theta - \xi \nabla V_T(\theta)) \\ &\leq \min_{\xi} [V_T(\theta) - \xi \|\nabla V_T(\theta)\|^2 + \frac{L'}{2} \xi^2 \|\nabla V_T(\theta)\|^2] \\ &\leq \min_{\xi} [V_T(\theta) + \|\nabla V_T(\theta)\|^2 (\frac{L'}{2} (\xi - \frac{1}{L'})^2 - \frac{1}{2L'})] \\ &\leq V_T(\theta) - \frac{1}{2L'} \|\nabla V_T(\theta)\|^2. \end{aligned}$$

where second inequality comes from the L' smoothness on the straight line from θ to θ_*.
therefore,

$$\frac{V_T(\theta) - V_T(\theta_*)}{\|\nabla_\theta V_T(\theta)\|^2} \geq \frac{1}{2L'}.$$

□

Remark E.3. If $\phi_k(\theta_* + \xi(\theta - \theta_*))$ is monotonically increasing on ξ, then

$$\overline{\phi_k}(\theta) \leq \phi_k(\theta),$$
$$L' \leq \sum_{k=1}^{m} d_k \sum_{t=0}^{T} [\phi_k(\theta)^{t-1} L_1 + t(t-1)\phi_k(\theta)^{t-2} {L_2}^2].$$

then the smoothness along the updated step is bounded by the spectral radius of the Hessian on θ, as $\rho_{max}(J_T(\theta))$ in Theorem 3.9.

Theorem 3.12 (Restated)**.** *Assume $\phi_k(\theta)$ is local strong convex as stated in Assumption 3.5 and the decreasing learning rate η satisfies the conditions in Theorem 3.10, then if we run gradient descent for $V_T(\theta)$, it yields a solution:*

$$\|\theta_l - \theta_*\|^2 \leq q^l \|\theta_0 - \theta_*\|^2, \tag{15}$$

where $\omega^* = \min_{\theta} \sum_{k=0}^{m} d_k [(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1})]$, $\sqrt{q}$ denotes the convergence rate and its square q is lower bounded s.t. $q \geq (1 - \frac{2\omega^* \alpha}{\rho_{max}(J_T(\theta_0))})$

Proof. by Proposition 3.11, $\frac{\partial^2 \phi_k(\theta)}{\partial \theta^2} \succcurlyeq \alpha I$. From equation 14,

$$\begin{aligned} J_T^k(\theta) &\succcurlyeq d_k [(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1})] \alpha I, \\ J_T(\theta) &\succcurlyeq \sum_{k=0}^{m} d_k [(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1})] \alpha I \\ &\succcurlyeq \min_{\theta} \sum_{k=0}^{m} d_k [(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1})] \alpha I = \omega^* \alpha I. \end{aligned}$$

by Proposition 3.11, $V_T(\theta)$ is $\omega^*\alpha$ strong convex:

$$V_T(\theta_1+\theta_2) \geq V_T(\theta_1) + \theta_2^\top \nabla_\theta V_T(\theta_1) + \frac{\omega^*\alpha}{2}\|\theta_2\|^2. \tag{16}$$

$$\begin{aligned}
&\|\theta_{l+1}-\theta_*\|^2 \\
&= \|\theta_l - \eta_l \nabla_\theta V_T(\theta_l) - \theta_*\|^2 \\
&= \|\theta_l-\theta_*\|^2 - 2\eta_l \nabla_\theta V_T(\theta_l)^\top(\theta_l-\theta_*) + \eta_l^2\|\nabla_\theta V_T(\theta_l)\|^2 \\
&\overset{(16)}{\leq} \|\theta_l-\theta_*\|^2(1-\eta_l\omega^*\alpha) - 2\eta_l(V_T(\theta_l)-V_T(\theta_*)) \\
&\quad + \eta_l^2\|\nabla_\theta V_T(\theta_l)\|^2 \\
&\leq \|\theta_l-\theta_*\|^2(1-\eta_l\omega^*\alpha) \text{ when } \eta_l < 2\frac{V_T(\theta_l)-V_T(\theta_*)}{\|\nabla_\theta V_T(\theta_l)\|^2}
\end{aligned} \tag{17}$$

where the inequality condition is satisfied with our analysis in Lemma E.2 and Remark E.3 when

$$\begin{aligned}
\eta_l &< 1/\sum_{k=1}^{m} d_k[L_1(\sum_{t=0}^{T} t\phi_k(\theta_l)^{t-1}) + {L_2}^2(\sum_{t=0}^{T} t(t-1)\phi_k(\theta_l)^{t-2}] \text{ (conditions in Theorem 3.10)} \\
&\leq \frac{1}{L'} \\
&\leq 2\frac{V_T(\theta_l)-V_T(\theta_*)}{\|\nabla_\theta V_T(\theta_l)\|^2}.
\end{aligned}$$

Since we have a non-increasing step size, η_l is upper bounded by $\frac{2}{\rho_{max}(J_T(\theta_0))}$, $(1-\eta_l\omega^*\alpha)$ is greater than $(1-\frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_0))})$.

□

Theorem 3.15 (Restated)**.** *Assume $\phi_k(\theta)$ is local strong convex as stated in Assumption 3.5 and the decreasing learning rate $\eta_l < \frac{C}{L_1 T + {L_2}^2 T(T-1)}$, then if we run gradient descent for logarithmic mapped $V_T(\theta)$ with Equation* (13)*, it yields a solution:*

$$\|\theta_{l+1}-\theta_*\|^2 \leq q_l\|\theta_l-\theta_*\|^2.$$

where the the square of the step convergence rate q_l has a varying lower bound s.t. $q_l \geq (1-\frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_l))})$

Proof.

$$\begin{aligned}
&\|\theta_{l+1}-\theta_*\|^2 \\
&= \|\theta_l - \frac{\eta_l}{V_T(\theta_l)}\nabla_\theta V_T(\theta_l) - \theta_*\|^2 \\
&= \|\theta_l-\theta_*\|^2 - 2\frac{\eta_l}{V_T(\theta_l)}\nabla_\theta V_T(\theta_l)^\top(\theta_l-\theta_*) + \frac{\eta_l}{V_T(\theta_l)}^2\|\nabla_\theta V_T(\theta_l)\|^2 \\
&\leq \|\theta_l-\theta_*\|^2(1-\frac{\eta_l}{V_T(\theta_l)}\omega^*\alpha) - 2\frac{\eta_l}{V_T(\theta_l)}(V_T(\theta)-V_T(\theta_*)) + \frac{\eta_l}{V_T(\theta_l)}^2\|\nabla_\theta V_T(\theta_l)\|^2 \\
&\leq \|\theta_l-\theta_*\|^2(1-\frac{\eta_l}{V_T(\theta_l)}\omega^*\alpha) \text{ when } \frac{\eta_l}{V_T(\theta_l)} \leq 2\frac{V_T(\theta_l)-V_T(\theta_*)}{\|\nabla_\theta V_T(\theta_l)\|^2}
\end{aligned}$$

where the inequality condition is satisfied with our analysis in Lemma E.2 and Remark E.3 when $\frac{\eta_l}{V_T(\theta_l)} < 1/\sum_{k=1}^{m} d_k[L_1(\sum_{t=0}^{T} t\phi_k(\theta_l)^{t-1}) + {L_2}^2(\sum_{t=0}^{T} t(t-1)\phi_k(\theta_l)^{t-2}]$. The latter is valid

because $V_T(\theta_l)/\sum_{k=1}^m d_k[L_1(\sum_{t=0}^T t\phi_k(\theta_l)^{t-1}) + {L_2}^2(\sum_{t=0}^T t(t-1)\phi_k(\theta_l)^{t-2}]$ is lower bounded by a constant

$$\begin{aligned}
&\min_l \frac{V_T(\theta_l)}{\sum_{k=1}^m d_k[L_1(\sum_{t=0}^T t\phi_k(\theta_l)^{t-1}) + {L_2}^2(\sum_{t=0}^T t(t-1)\phi_k(\theta_l)^{t-2}]} \\
&= \min_l \frac{\sum_{k=1}^m d_k[(\sum_{t=0}^T \phi_k(\theta_l)^t)}{\sum_{k=1}^m d_k[L_1(\sum_{t=0}^T t\phi_k(\theta_l)^{t-1}) + {L_2}^2(\sum_{t=0}^T t(t-1)\phi_k(\theta_l)^{t-2})} \\
&\geq \min_l \min_k \frac{\phi_k(\theta_l)}{L_1 T + {L_2}^2 T(T-1)} \\
&> \frac{\underline{C}}{L_1 T + {L_2}^2 T(T-1)} > \eta_l.
\end{aligned}$$

Because

$$\begin{aligned}
\frac{\eta_l}{V_T(\theta_l)} &< 1/\sum_{k=1}^m d_k[L_1(\sum_{t=0}^T t\phi_k(\theta_l)^{t-1}) + {L_2}^2(\sum_{t=0}^T t(t-1)\phi_k(\theta_l)^{t-2}] \\
&< \frac{2}{\rho_{max}(J_T(\theta_l))}. \qquad (18)
\end{aligned}$$

we have $(1 - \frac{\eta_l}{V_T(\theta_l)}\omega^*\alpha) > (1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_l))})$. □

Assumption E.4. assume $\widehat{\phi}_{j,k}(\theta)$ and its gradient has small i.i.d. random noise $\epsilon_{1,j,k}$ and $\epsilon_{2,j,k}$ s.t.:

$$\begin{aligned}
\widehat{\phi}_{j,k}(\theta) &= \phi_k(\theta) + \epsilon_{1,j,k}, \\
\frac{\partial \widehat{\phi}_{j,k}(\theta)}{\partial \theta} &= \frac{\partial \phi_k(\theta)}{\partial \theta} + \epsilon_{2,j,k}.
\end{aligned}$$

Let

$$\widehat{V}_{b,T}(\theta) = \frac{1}{b}\sum_{j=1}^b v_T(s_j, \theta).$$

be the value function approximation by stochastic sampling over mini-batch with size b. where $v_T(s_j, \theta)$ is parameterized by $\sum_{k=1}^m d_k \sum_{t=0}^T \widehat{\phi}_{j,k}(\theta)^t$.

Lemma E.5. *if the realizations of $\widehat{\phi}_{j,k}$ and its gradient has random noise following Assumption E.4, the variance of stochastic gradient descent for vanilla method is bounded by*

$$\frac{1}{b^2}[\|\sum_{k=1}^m d_k \sum_{t=0}^T t(t-1)\phi_k(\theta)^{t-2}\frac{\partial \phi_k(\theta)}{\partial \theta}\|^2 \mathbb{E}[\epsilon_{1,j,k}^2] + (\sum_{k=1}^m d_k \sum_{t=0}^T t\phi_k(\theta)^{t-1})^2 \mathbb{E}[\|\epsilon_{2,j,k}\|^2].$$

Proof. the approximated gradient is:

$$\begin{aligned}
&\nabla_\theta \widehat{V}_{b,T}(\theta) \\
&= \frac{1}{b}\sum_{j=1}^b \sum_{k=1}^m d_k (\sum_{t=0}^T t\widehat{\phi}_{j,k}(\theta)^{t-1}) \frac{\partial \widehat{\phi}_{j,k}(\theta)}{\theta} \\
&= \frac{1}{b}\sum_{j=1}^b \sum_{k=1}^m d_k (\sum_{t=0}^T t(\phi_k(\theta) + \epsilon_{1,j,k})^{t-1})(\frac{\partial \phi_k(\theta)}{\partial \theta} + \epsilon_{2,j,k}).
\end{aligned}$$

The variance of vanilla method is:

$$
\begin{aligned}
&\mathbb{E}[\|\nabla_\theta V_T(\theta) - \nabla_\theta \widehat{V}_{b,T}(\theta)\|^2] \\
&= \mathbb{E}[\|\frac{1}{b}\sum_{j=1}^{b}\sum_{k=1}^{m} d_k[(\sum_{t=0}^{T} t(\phi_k(\theta)+\epsilon_{1,j,k})^{t-1})(\frac{\partial \phi_k(\theta)}{\partial \theta}+\epsilon_{2,j,k}) \\
&- (\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}\frac{\partial \phi_k(\theta)}{\partial \theta})]\|^2] \\
&\text{neglect high order terms} \\
&\approx \mathbb{E}[\|\frac{1}{b}\sum_{j=1}^{b}\sum_{k=1}^{m} d_k[(\sum_{t=0}^{T} t((t-1)\phi_k(\theta)^{t-2}\epsilon_{1,j,k}\frac{\partial \phi_k(\theta)}{\partial \theta} \\
&+ \phi_k(\theta)^{t-1}\epsilon_{2,j,k})\|^2] \\
&= \frac{1}{b^2}[\|\sum_{k=1}^{m} d_k \sum_{t=0}^{T} t(t-1)\phi_k(\theta)^{t-2}\frac{\partial \phi_k(\theta)}{\partial \theta}\|^2\, \mathbb{E}[\epsilon_{1,j,k}^2] \\
&+ (\sum_{k=1}^{m} d_k \sum_{t=0}^{T} t\phi_k(\theta)^{t-1})^2\, \mathbb{E}[\|\epsilon_{2,j,k}\|^2].
\end{aligned}
$$

□

Lemma E.6. *If the realizations of $\widehat{\phi}_{j,k}$ and its gradient has random noise following Assumption E.4, the variance of log mapping policy gradient is bounded by $\frac{1}{b^2}(2{L_2}^2\frac{T^4}{\underline{C}^4}\,\mathbb{E}[\epsilon_{1,j,k}^2] + \frac{T^2}{\underline{C}^2}\,\mathbb{E}[\|\epsilon_{2,j,k}\|^2])$, where $\underline{C}$ is the lower bound of $\phi_k(\theta)$ in Assumption 3.2*

Proof. Variance of stochastic gradient descent for logarithmic mapped $\widetilde{V}_T(\theta)$:

let

$$
\widehat{\widetilde{V}}_{b,T}(\theta) = \frac{1}{b}\sum_{j=1}^{b} log(v_T(s_j,\theta)).
$$

be the value function approximation by stochastic sampling over mini-batch with size b.

$$
\begin{aligned}
&\nabla_\theta \widehat{\widetilde{V}}_{b,T}(\theta) \\
&= \frac{1}{b}\sum_{j=1}^{b} \frac{\sum_{k=1}^{m} d_k(\sum_{t=0}^{T} t\widehat{\phi}_{j,k}(\theta)^{t-1})\frac{\partial \widehat{\phi}_{j,k}(\theta)}{\theta}}{\sum_{k=1}^{m} d_k(\sum_{t=0}^{T}\widehat{\phi}_{j,k}(\theta)^t)} \\
&= \frac{1}{b}\sum_{j=1}^{b} \frac{\sum_{k=1}^{m} d_k(\sum_{t=0}^{T} t(\phi_k(\theta)+\epsilon_{1,j,k})^{t-1})(\frac{\partial \phi_k(\theta)}{\partial \theta}+\epsilon_{2,j,k})}{\sum_{k=1}^{m} d_k(\sum_{t=0}^{T}(\phi_k(\theta)+\epsilon_{1,j,k})^t}.
\end{aligned}
$$

The gradient variance of logarithmic mapping is:

$$
\begin{aligned}
&\mathbb{E}[\|\nabla_\theta \widetilde{V}_T(\theta) - \nabla_\theta \widehat{\widetilde{V}}_{b,T}(\theta)\|^2] \\
&= \mathbb{E}[\|\frac{1}{b}\sum_{j=1}^{b}(\frac{\sum_{k=1}^{m} d_k(\sum_{t=0}^{T} t(\phi_k(\theta)+\epsilon_{1,j,k})^{t-1})(\frac{\partial \phi_k(\theta)}{\partial \theta}+\epsilon_{2,j,k})}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T}(\phi_k(\theta)+\epsilon_{1,j,k})^t} - \frac{\sum_{k=1}^{m} d_k(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1})\frac{\partial \phi_k(\theta)}{\partial \theta}}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T}\phi_k(\theta)^t})\|^2] \\
&\text{neglect high order terms} \\
&\approx \mathbb{E}[\|\frac{1}{b}\sum_{j=1}^{b}(\frac{\sum_{k=1}^{m} d_k(\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}\frac{\partial \phi_k(\theta)}{\partial \theta} + t(t-1)\phi_k(\theta)^{t-2}\epsilon_{1,j,k}\frac{\partial \phi_k(\theta)}{\partial \theta} + t\phi_k(\theta)^{t-1}\epsilon_{2,j,k})}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T}(\phi_k(\theta)^t + t\phi_k(\theta)^{t-1}\epsilon_{1,j,k})}
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{\sum_{k=1}^{m} d_k (\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}) \frac{\partial \phi_k(\theta)}{\partial \theta}}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T} \phi_k(\theta)^t}) \|^2] \\
&= \mathbb{E}[\| \frac{1}{b} \sum_{j=1}^{b} \frac{\sum_{k=1}^{m} d_k (\sum_{t=0}^{T} t(t-1)\phi_k(\theta)^{t-2} \epsilon_{1,j,k} \frac{\partial \phi_k(\theta)}{\partial \theta} + t\phi_k(\theta)^{t-1} \epsilon_{2,j,k})}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T} \phi_k(\theta)^t} \\
&- \frac{\sum_{k=1}^{m} d_k (\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}) \frac{\partial \phi_k(\theta)}{\partial \theta}}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T} \phi_k(\theta)^t} \frac{\sum_{k=1}^{m} d_k \sum_{t=0}^{T} (t-1)\phi_k(\theta)^{t-1} \epsilon_{1,j,k}}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T} \phi_k(\theta)^t} \|^2] \\
&\leq \frac{1}{b^2} \frac{1}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T} \phi_k(\theta)^t}^2 (\sum_{k=1}^{m} \| d_k (\sum_{t=0}^{T} t(t-1)\phi_k(\theta)^{t-2}) \frac{\partial \phi_k(\theta)}{\partial \theta} \|^2 \mathbb{E}[\epsilon_{1,j,k}^2] + (\sum_{k=1}^{m} d_k \sum_{t=0}^{T} t\phi_k(\theta)^{t-1})^2 \mathbb{E}[\|\epsilon_{2,j,k}\|^2] \\
&+ (\frac{\| \sum_{k=1}^{m} d_k (\sum_{t=0}^{T} t\phi_k(\theta)^{t-1}) \frac{\partial \phi_k(\theta)}{\partial \theta} \|}{\sum_{k=1}^{m} d_k \sum_{t=0}^{T} \phi_k(\theta)^t})^2 \sum_{k=1}^{m} (d_k (\sum_{t=0}^{T} (t-1)\phi_k(\theta)^{t-1}))^2 \mathbb{E}[\epsilon_{1,j,k}^2]) \\
&\leq \frac{1}{b^2} (L_2{}^2 \frac{T^4}{\underline{C}^4} \mathbb{E}[\epsilon_{1,j,k}^2] + \frac{T^2}{\underline{C}^2} \mathbb{E}[\|\epsilon_{2,j,k}\|^2] + L_2{}^2 \frac{T^4}{\underline{C}^4} \mathbb{E}[\epsilon_{1,j,k}^2]) \\
&= \frac{1}{b^2} (2L_2{}^2 \frac{T^4}{\underline{C}^4} \mathbb{E}[\epsilon_{1,j,k}^2] + \frac{T^2}{\underline{C}^2} \mathbb{E}[\|\epsilon_{2,j,k}\|^2]).
\end{aligned}
$$

□

Lemma E.7 (Bertsekas (2011)). *Let Y_k, Z_k, and W_k, $k = 0, 1, ...$, be three sequences of random variables and let $F_k, k \geq 0$ be a filtration, that is, σ-algebras such that $\{F_k\} \subset F_{k+1}$ for all k. Suppose that:*

- *The random variables Y_k, Z_k, and W_k are non-negative, and F_k-measurable.*
- *For each k, we have $\mathbb{E}[Y_{k+1}|F_k] \leq Y_k - Z_k + W_k$*
- *There holds, w.p.1,*

$$\sum_{k=0}^{\infty} W_k < \infty.$$

then we have w.p.1,

$$\sum_{k=0}^{\infty} Z_k < \infty \text{ and } Y_k \to Y \geq 0.$$

Theorem E.8. *If the realizations of $\widehat{\phi}_{j,k}$ and its gradient has random noise following Assumption E.4 with decreasing learning rate η_l s.t.*

$$0 < \eta_l \leq \frac{2}{\rho_{max}(J_T(\theta_0))}, \sum_{l=0}^{\infty} \eta_l = \infty \text{ and } \sum_{l=0}^{\infty} \eta_l^2 < \infty,$$

then we have $\|\theta_l - \theta_*\| \to 0$ w.p.1

Proof. In the stochastic case, the gradient is replaced by $\nabla_\theta \widehat{V}_{b,T}(\theta)$ approximated by sampling. From Lemma E.5,

$$\mathbb{E}[\|\nabla_\theta \widehat{V}_{b,T}(\theta) - \nabla_\theta V_T(\theta_l)\|^2] = \frac{N(\theta)}{b^2}.$$

where

$$
\begin{aligned}
N(\theta) = [\| &\sum_{k=1}^{m} d_k \sum_{t=0}^{T} t(t-1)\phi_k(\theta)^{t-2} \frac{\partial \phi_k(\theta)}{\partial \theta} \|^2 \mathbb{E}[\epsilon_{1,j,k}^2] \\
&+ (\sum_{k=1}^{m} d_k \sum_{t=0}^{T} t\phi_k(\theta)^{t-1})^2 \mathbb{E}[\|\epsilon_{2,j,k}\|^2].
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[\|\theta_{l+1}-\theta_*\|^2] \\
&= \mathbb{E}[\|\theta_l - \eta_l \nabla_\theta \widehat{V}_{b,T}(\theta_l) - \theta_*\|^2] \\
&= \mathbb{E}[\|\theta_l - (\nabla_\theta V_T(\theta_l) + (\nabla \widehat{V}_{b,T}(\theta_l) - \nabla_\theta V_T(\theta_l))) - \theta_*\|^2] \\
&= \|\theta_l - \theta_*\|^2 - 2\eta_l \nabla_\theta V_T(\theta_l)^\top (\theta_l - \theta_*) + \eta_l^2 \|\nabla_\theta V_T(\theta_l)\|^2 + \eta_l^2 \,\mathbb{E}[\|\nabla_\theta \widehat{V}_{b,T}(\theta) - \nabla_\theta V_T(\theta_l)\|^2] \\
&= \|\theta_l - \theta_*\|^2 - 2\eta_l \nabla_\theta V_T(\theta_l)^\top (\theta_l - \theta_*) + \eta_l^2 \|\nabla_\theta V_T(\theta_l)\|^2 + \eta_l^2 \frac{N(\theta_l)}{b^2} \\
&\overset{(16)}{\leq} \|\theta_l - \theta_*\|^2 (1 - \eta_l \omega^* \alpha) - 2\eta_l (V_T(\theta_l) - V_T(\theta_*)) + \eta_l^2 \|\nabla_\theta V_T(\theta_l)\|^2 + \eta_l^2 \frac{N(\theta_l)}{b^2} \\
&\leq \|\theta_l - \theta_*\|^2 (1 - \eta_l \omega^* \alpha) + \eta_l^2 \frac{N(\theta_l)}{b^2} \text{ when } \eta_l < 2\frac{V_T(\theta_l) - V_T(\theta_*)}{\|\nabla_\theta V_T(\theta_l)\|^2}.
\end{aligned} \tag{19}
$$

Similar to the treatment of Theorem 1 in Nguyen et al. (2018), apply Lemma E.7 with $\sum_{l=0}^{\infty} \eta_l^2 \frac{N(\theta_l)}{b^2} < \infty$ and $\eta_l \leq \frac{2}{\rho_{max}(J_T(\theta_0))}$, then we have w.p.1,

$$
\|\theta_l - \theta_*\|^2 \to W \geq 0,
$$
$$
\text{and } \sum_{l=0}^{\infty} \|\theta_l - \theta_*\|^2 \frac{2\omega^* \alpha}{\rho_{max}(J_T(\theta_0))} < \infty.
$$

Suppose there exists $\epsilon > 0$ and l_0, s.t. $\|\theta_l - \theta_*\|^2 > \epsilon$ for $l > l_0$, then

$$
\sum_{l=0}^{\infty} \|\theta_l - \theta_*\|^2 \frac{2\omega^* \alpha}{\rho_{max}(J_T(\theta_0))} > \sum_{l=l_0}^{\infty} \|\theta_l - \theta_*\|^2 \frac{2\omega^* \alpha}{\rho_{max}(J_T(\theta_0))} > \sum_{l=l_0}^{\infty} \epsilon \frac{2\omega^* \alpha}{\rho_{max}(J_T(\theta_0))} = \infty.
$$

by contradiction, $\|\theta_l - \theta_*\|^2 \to 0$ w.p.1. □

Theorem E.9. *If the realizations of $\widehat{\phi}_{j,k}$ and its gradient has random noise following Assumption E.4 with decreasing learning rate η_l s.t.*

$$
0 < \eta_l \leq \frac{\underline{C}}{{L_1}T + {L_2}^2 T(T-1)}, \sum_{l=0}^{\infty} \eta_l = \infty \text{ and } \sum_{l=0}^{\infty} \eta_l^2 < \infty.
$$

where $\underline{C}$ is the lower bound of in Assumption 3.2.

then we have $\|\theta_l - \theta_*\| \to 0$ w.p.1.

Proof. For the logarithmic mapping stochastic case, the stochastic gradient is $\widehat{\widetilde{V}}_{b,T}(\theta) = \frac{1}{b}\sum_{j=1}^{b} log(v_T(s_j, \theta))$, let $\widetilde{N} = \frac{1}{b^2}(2{L_2}^2 \frac{T^4}{\underline{C}^4} \mathbb{E}[\epsilon_{1,j,k}^2] + \frac{T^2}{\underline{C}^2} \mathbb{E}[\|\epsilon_{2,j,k}\|^2])$ be its variance upper bound in Lemma E.6.

$$
\begin{aligned}
&\mathbb{E}[\|\theta_{l+1}-\theta_*\|^2] \\
&= \mathbb{E}[\|\theta_l - \eta_l \nabla_\theta \widehat{\widetilde{V}}_{b,T}(\theta_l) - \theta_*\|^2] \\
&= \mathbb{E}[\|\theta_l - \eta_l (\nabla_\theta \widetilde{V}_{b,T}(\theta_l) + (\nabla_\theta \widehat{\widetilde{V}}_{b,T}(\theta_l) - \nabla_\theta \widetilde{V}_{b,T}(\theta_l))) - \theta_*\|^2] \\
&= \|\theta_l - \theta_*\|^2 - 2\eta_l \nabla_\theta \widetilde{V}_{b,T}(\theta_l)^\top (\theta_l - \theta_*) + \eta_l^2 \|\nabla_\theta \widetilde{V}_{b,T}(\theta_l)\|^2 + \eta_l^2 \,\mathbb{E}[\|\nabla_\theta \widehat{\widetilde{V}}_{b,T}(\theta_l) - \nabla_\theta \widetilde{V}_{b,T}(\theta_l)\|^2] \\
&= \|\theta_l - \theta_*\|^2 - 2\eta_l \nabla_\theta \widetilde{V}_{b,T}(\theta_l)^\top (\theta_l - \theta_*) + \eta_l^2 \|\nabla_\theta \widetilde{V}_{b,T}(\theta_l)\|^2 + \eta_l^2 \frac{\widetilde{N}}{b^2} \\
&\leq \|\theta_l - \theta_*\|^2 (1 - \frac{\eta_l}{V_T(\theta_l)} \omega^* \alpha) - 2\frac{\eta_l}{V_T(\theta_l)} (V_T(\theta) - V_T(\theta_*)) + \frac{\eta_l}{V_T(\theta_l)}^2 \|\nabla_\theta V_T(\theta_l)\|^2 + \eta_l^2 \frac{\widetilde{N}}{b^2} \\
&\leq \|\theta_l - \theta_*\|^2 (1 - \frac{\eta_l}{V_T(\theta_l)} \omega^* \alpha) + \eta_l^2 \frac{\widetilde{N}}{b^2} \text{ when } \frac{\eta_l}{V_T(\theta_l)} \leq 2\frac{V_T(\theta_l) - V_T(\theta_*)}{\|\nabla_\theta V_T(\theta_l)\|^2}
\end{aligned} \tag{20}
$$

Using the same treatment in Theorem E.8 we have $\|\theta_l - \theta_*\|^2 \to 0$ w.p.1. □

Now Theorem E.8 and Theorem E.9 prove the convergence of both vanilla and logarithmic mapping method under stochastic case, we would like to further explore the convergence rate.

For the vanilla case, taking the expectation of Equation (19), we have

$$\begin{aligned}\mathbb{E}[\|\theta_{l+1} - \theta_*\|^2] &\leq \mathbb{E}[\|\theta_l - \theta_*\|^2](1 - \eta_l\omega^*\alpha) + \eta_l^2\frac{N(\theta_l)}{b^2} \\ &\leq \mathbb{E}[\|\theta_l - \theta_*\|^2](1 - \eta_l\omega^*\alpha) + \eta_l^2\frac{N(\theta_0)}{b^2}\end{aligned}$$

where η_l is bounded by $\frac{2}{\rho_{max}(J_T(\theta_0))}$. Substitute the maximum allowed $\eta_l = \frac{2}{\rho_{max}(J_T(\theta_0))}$ into above, then we have

$$\mathbb{E}[\|\theta_{l+1} - \theta_*\|^2] \leq \mathbb{E}[\|\theta_l - \theta_*\|^2](1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_0))}) + \frac{2}{\rho_{max}(J_T(\theta_0))}^2\frac{N(\theta_l)}{b^2}$$

By induction,

$$\begin{aligned}\mathbb{E}[\|\theta_l - \theta_*\|^2] &\leq \mathbb{E}[\|\theta_0 - \theta_*\|^2]\prod_{i=0}^{l-1}(1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_0))}) + \frac{N(\theta_0)}{b^2}\sum_{i=0}^{l-1}\frac{2}{\rho_{max}(J_T(\theta_0))}^2\prod_{j=i+1}^{l-1}(1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_0))}) \\ &= \mathbb{E}[\|\theta_0 - \theta_*\|^2](1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_0))})^l + \frac{N(\theta_0)}{b^2}\frac{2}{\rho_{max}(J_T(\theta_0))}\frac{1 - (1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_0))})^l}{\omega^*\alpha}\end{aligned} \tag{21}$$

where the 2nd term $\to \frac{N(\theta_0)}{b^2}\frac{2}{\rho_{max}(J_T(\theta_0))}\frac{1}{\omega^*\alpha}$ when $l \to \infty$

For the logarithmic mapping case, take the expectation of Equation (20),

$$\mathbb{E}[\|\theta_{l+1} - \theta_*\|^2] \leq \mathbb{E}[\|\theta_l - \theta_*\|^2](1 - \frac{\eta_l}{V_T(\theta_l)}\omega^*\alpha) + \eta_l^2\frac{\widetilde{N}}{b^2}$$

using the maximum allowed $\frac{\eta_l}{V_T(\theta_l)} = \frac{2}{\rho_{max}(J_T(\theta_l))}$ in Equation (18), then we have

$$\mathbb{E}[\|\theta_{l+1} - \theta_*\|^2] \leq \mathbb{E}[\|\theta_l - \theta_*\|^2](1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_l))}) + (\frac{2V_T(\theta_l)}{\rho_{max}(J_T(\theta_l))})^2\frac{\widetilde{N}}{b^2}$$

By induction,

$$\mathbb{E}[\|\theta_l - \theta_*\|^2] \leq \mathbb{E}[\|\theta_0 - \theta_*\|^2]\prod_{i=0}^{l-1}(1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_)})) + \frac{\widetilde{N}}{b^2}\sum_{i=0}^{l-1}\frac{2V_T(\theta_i)}{\rho_{max}(J_T(\theta_i))}^2\prod_{j=i+1}^{l-1}(1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_j))}) \tag{22}$$

Here we do not have an upper bound for $\frac{2V_T(\theta_i)}{\rho_{max}(J_T(\theta_i))}$ term because the denominator $\rho_{max}(J_T(\theta_i))$ is upper bounded by $\sum_{k=1}^m d_k[L_1(\sum_{t=0}^T t\phi_k(\theta_i)^{t-1}) + {L_2}^2(\sum_{t=0}^T t(t-1)\phi_k(\theta_i)^{t-2}]$. With $V_T(\theta_i) \sim \sum_{k=1}^m d_k\sum_{t=0}^T \phi_k(\theta_i)^t$, approximate $\frac{2V_T(\theta_i)}{\rho_{max}(J_T(\theta_i))} \sim \max\{\frac{\overline{C}}{L_1 T}, \frac{\overline{C}}{{L_2}^2 T(T-1)}\}$ is a constant, where $\overline{C} = \max_k \max_\theta \phi_k(\theta)$. The $\prod_{j=i+1}^{l-1}(1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_j))})$ term is a diminishing series with coefficient $(1 - \frac{2\omega^*\alpha}{\rho_{max}(J_T(\theta_i+1))})$ strictly less than 1. Therefore, the 2nd term in Equation (22) converges to a constant as well.

Now, comparing Equation (21) and Equation (22) both under maximum allowed fix learning rate, both have a constant second term and the first term of Equation (22) diminishes much faster than Equation (21) as similar to the deterministic case.