A Proof of Theorem 4.6

Preliminaries. Note that the expected cumulative reward is equivalent to

$$\begin{aligned}
J(\theta) &= V_\theta^{(0)}(s_0) \\
V_\theta^{(t)}(s) &= R_\theta(s) + \mathbb{E}_{p(\zeta)}\left[V_\theta^{(t+1)}(f_\theta(s)+\zeta)\right] \qquad (\forall t \in \{0,1,...,T-1\}) \\
V_\theta^{(T)}(s) &= 0
\end{aligned}$$

and the expected model-based policy gradient is

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta V_\theta^{(0)}(s_0) \\
\nabla_\theta V_\theta^{(t)}(s) &= \nabla_\theta R_\theta(s) + \mathbb{E}_{p(\zeta)}\left[\nabla_\theta V_\theta^{(t+1)}(f_\theta(s)+\zeta) + \nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta)\nabla_\theta f_\theta(s)\right] \\
\nabla_s V_\theta^{(t)}(s) &= \nabla_s R_\theta(s) + \mathbb{E}_{p(\zeta)}\left[\nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta)\nabla_s f_\theta(s)\right] \\
\nabla_\theta V_\theta^{(T)}(s) &= \nabla_s V_\theta^{(T)}(s) = 0.
\end{aligned}$$

Similarly, given a sample $\vec{\zeta} \sim p(\vec{\zeta})$, the stochastic approximation of the expected cumulative reward is

$$\begin{aligned}
\hat{J}(\theta;\vec{\zeta}) &= \hat{V}_\theta^{(0)}(s_0;\vec{\zeta}) \\
\hat{V}_\theta^{(t)}(s;\vec{\zeta}) &= R_\theta(s) + \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) \qquad (\forall t \in \{0,1,...,T-1\}) \\
\hat{V}_\theta^{(T)}(s;\vec{\zeta}) &= 0
\end{aligned}$$

and the stochastic approximation of the model-based policy gradient is

$$\begin{aligned}
\nabla_\theta \hat{J}(\theta;\vec{\zeta}) &= \nabla_\theta \hat{V}_\theta^{(0)}(s_0;\vec{\zeta}) \\
\nabla_\theta \hat{V}_\theta^{(t)}(s;\vec{\zeta}) &= \nabla_\theta R_\theta(s) + \nabla_\theta \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) + \nabla_s \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta})\nabla_\theta f_\theta(s) \\
\nabla_s \hat{V}_\theta^{(t)}(s;\vec{\zeta}) &= \nabla_s R_\theta(s) + \nabla_s \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta})\nabla_s f_\theta(s) \\
\nabla_\theta \hat{V}_\theta^{(T)}(s;\vec{\zeta}) &= \nabla_s \hat{V}_s^{(T)}(s;\vec{\zeta}) = 0.
\end{aligned}$$

Bounding the deviation of $\nabla_\theta \hat{V}_\theta^{(t)}$ from $\nabla\theta V_\theta^{(t)}$. We claim that for $t \in \{0,1,...,T\}$, we have

$$\begin{aligned}
\|\nabla_\theta \hat{V}_\theta^{(t)}(s;\vec{\zeta}) - \nabla_\theta V_\theta^{(t)}(s)\| &\le B_0^{(t)}(\vec{\zeta}) \\
\|\nabla_s \hat{V}_\theta^{(t)}(s;\vec{\zeta}) - \nabla_s V_\theta^{(t)}(s)\| &\le B_1^{(t)}(\vec{\zeta})
\end{aligned}$$

for all $\theta \in \Theta$ and $s \in S$, where

$$\begin{aligned}
B_0^{(t)}(\vec{\zeta}) &= \sum_{i=t}^{T-1} L_{f_\theta} B_1^{(i+1)}(\vec{\zeta}) + L_{\nabla V}^{(i+1)}(L_{f_\theta}+1)(\|\zeta_i\| + \sigma_\zeta\sqrt{d_S}) \\
B_1^{(t)}(\vec{\zeta}) &= \sum_{i=t}^{T-1} L_{\nabla V}^{(i+1)} L_{f_\theta}^{i-t+1}(\|\zeta_i\| + \sigma_\zeta\sqrt{d_S}) \\
B_0^{(T)}(\vec{\zeta}) &= B_1^{(T)}(\vec{\zeta}) = 0,
\end{aligned}$$

where $L_{\nabla V}^{(t)}$ is a Lipschitz constant for $\nabla V_\theta^{(t)}$. The base case $t = T$ follows trivially. Note that $\sigma_\zeta\sqrt{d_S} \geq \sqrt{\mathbb{E}_{p(\zeta)}[\|\zeta\|^2]} \geq \mathbb{E}_{p(\zeta)}[\|\zeta\|]$. Then, for $t \in \{0, 1, ..., T-1\}$, we have

$$\begin{aligned}
\|\nabla_\theta \hat{V}_\theta^{(t)}(s;\vec{\zeta}) - \nabla_\theta V_\theta^{(t)}(s)\| \leq& \left\|\nabla_\theta \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) - \mathbb{E}_{p(\zeta)}\left[\nabla_\theta V_\theta^{(t+1)}(f_\theta(s)+\zeta)\right]\right\| \\
&+ L_{f_\theta}\left\|\nabla_s \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) - \mathbb{E}_{p(\zeta)}\left[\nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta)\right]\right\| \\
\leq& \left\|\nabla_\theta \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) - \nabla_\theta V_\theta^{(t+1)}(f_\theta(s)+\zeta_t)\right\| \\
&+ \mathbb{E}_{p(\zeta)}\left[\left\|\nabla_\theta V_\theta^{(t+1)}(f_\theta(s)+\zeta_t) - \nabla_\theta V_\theta^{(t+1)}(f_\theta(s)+\zeta)\right\|\right] \\
&+ L_{f_\theta}\left\|\nabla_s \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) - \nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta_t)\right\| \\
&+ L_{f_\theta}\mathbb{E}_{p(\zeta)}\left[\left\|\nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta_t) - \nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta)\right\|\right] \\
\leq& B_0^{(t+1)}(\vec{\zeta}) + L_{\nabla V}^{(t+1)}(\|\zeta_t\| + \sigma_\zeta\sqrt{d_S}) + L_{f_\theta}B_1^{(t+1)}(\vec{\zeta}) + L_{f_\theta}L_{\nabla V}^{(t+1)}(\|\zeta_t\| + \sigma_\zeta\sqrt{d_S}) \\
=& B_0^{(t+1)}(\vec{\zeta}) + L_{f_\theta}B_1^{(t+1)}(\vec{\zeta}) + L_{\nabla V}^{(t+1)}(L_{f_\theta}+1)(\|\zeta_t\| + \sigma_\zeta\sqrt{d_S}) \\
=& B_0^{(t)}(\vec{\zeta}).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\|\nabla_s \hat{V}_\theta^{(t)}(s;\vec{\zeta}) - \nabla_s V_\theta^{(t)}(s)\| \leq& L_{f_\theta}\left\|\nabla_s \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) - \mathbb{E}_{p(\zeta)}\left[\nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta)\right]\right\| \\
\leq& L_{f_\theta}\left\|\nabla_s \hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) - \nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta_t)\right\| \\
&+ L_{f_\theta}\mathbb{E}_{p(\zeta)}\left[\left\|\nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta_t) - \nabla_s V_\theta^{(t+1)}(f_\theta(s)+\zeta)\right\|\right] \\
\leq& L_{f_\theta}\left(B_1^{(t+1)}(\vec{\zeta}) + L_{\nabla V}^{(t+1)}(\|\zeta_t\| + \sigma_\zeta\sqrt{d_S})\right) \\
=& B_1^{(t)}(\vec{\zeta}).
\end{aligned}$$

The claim follows.

Bounding the deviation of $\nabla_\theta \hat{J}$ from $\nabla_\theta J$. We claim that

$$\|\nabla_\theta \hat{J}(\theta;\vec{\zeta}) - \nabla_\theta J(\theta)\| \leq 132T^7\bar{L}_{R_\theta}\bar{L}_{f_\theta}^{5T}(E + \sigma_\zeta\sqrt{d_S}),$$

where $E = T^{-1}\sum_{t=0}^{T-1}\|\zeta_t\|$. To this end, letting $L_{\nabla V} = \arg\max_{t\in\{0,1,...,T\}} L_{\nabla V}^{(t)}$, note that

$$B_1^{(t)} \leq TL_{\nabla V}\bar{L}_{f_\theta}^{T-1}(E + \sigma_\zeta\sqrt{d_S})$$

for $t \in \{1, 2, ..., T\}$, so

$$\begin{aligned}
\|\nabla_\theta \hat{J}(\theta;\vec{\zeta}) - \nabla_\theta J(\theta)\| \leq B_0^{(0)}(\vec{\zeta}) =& \sum_{i=0}^{T-1} L_{f_\theta}B_1^{(i+1)}(\vec{\zeta}) + L_{\nabla V}(L_{f_\theta}+1)(\|\zeta_i\| + \sigma_\zeta\sqrt{d_S}) \\
&\leq T^2L_{\nabla V}\bar{L}_{f_\theta}^T(E + \sigma_\zeta\sqrt{d_S}) + TL_{\nabla V}(L_{f_\theta}+1)(E + \sigma_\zeta\sqrt{d_S}) \\
&\leq 3T^2L_{\nabla V}\bar{L}_{f_\theta}^T(E + \sigma_\zeta\sqrt{d_S}) \\
&\leq 132T^7\bar{L}_{R_\theta}\bar{L}_{f_\theta}^{5T}(E + \sigma_\zeta\sqrt{d_S}),
\end{aligned}$$

where the last step follows from our bound on $L_{\nabla V}^{(t)}$ in Lemma D.2.

Upper bound on sample complexity of $\nabla_\theta \hat{J} - \nabla_\theta J$. Note that $E \leq \|\vec{\zeta}\|_1$, where we think of $\vec{\zeta}$ as the length Td_S concatenation of the vectors $\zeta_0, \zeta_1, ..., \zeta_{T-1}$, so $\vec{\zeta}$ is σ_ζ-sub-Gaussian. We apply Lemma G.7 with

$$\begin{aligned}
Y &= \nabla_\theta \hat{J}(\theta; \vec{\zeta}) - \nabla_\theta J(\theta) \\
X &= E \\
A &= 132 T^7 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{5T} \\
B &= A \sigma_\zeta \sqrt{d_S}.
\end{aligned}$$

Thus, Y is σ_{MB}-sub-Gaussian, where

$$\begin{aligned}
\sigma_{\text{MB}} &= \max\{10 A \sigma_\zeta T d_S \log(T d_S), 5 A \sigma_\zeta \sqrt{d_S}\} \\
&= 10 A \sigma_\zeta T d_S \log(T d_S) \\
&\leq 1320 T^8 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{5T} \sigma_\zeta d_S \log(T d_S).
\end{aligned}$$

Thus, by Lemma G.6, the sample complexity of $\nabla_\theta \hat{J}(\theta) - \nabla_\theta J(\theta)$ is

$$\begin{aligned}
\sqrt{n_{\text{MB}}(\epsilon, \delta)} &= \frac{\sigma_{\text{MB}} \sqrt{2 \log(2 d_S / \delta)}}{\epsilon} \\
&= O\left(\frac{T^8 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{5T} \sigma_\zeta d_S \log(T) \log(d_S)^{3/2} \log(1/\delta)^{1/2}}{\epsilon} \right).
\end{aligned}$$

The claim follows.

Lower bound on sample complexity of $\nabla_\theta \hat{J} - \nabla_\theta J$. Consider a linear dynamical system with $S = A = \mathbb{R}$, time-invariant deterministic transitions $f(s, a) = \beta s + a$ (where $\beta \in \mathbb{R}$), time-varying noise

$$p_t(\zeta) = \begin{cases} \mathcal{N}(\zeta \mid 0, \sigma_\zeta^2) & \text{if } t = 0 \\ \delta(0) & \text{otherwise,} \end{cases}$$

where $\sigma_\zeta \in \mathbb{R}$, initial state $s_0 = 0$, time-varying rewards

$$R_t(s, a) = \begin{cases} s & \text{if } t = T - 1 \\ 0 & \text{otherwise,} \end{cases}$$

control policy class $\pi_\theta(s) = \theta s$, and current parameters $\theta = 0$. Note that

$$s_t = \begin{cases} 0 & \text{if } t = 0 \\ (\beta + \theta)^{t-1} \zeta & \text{otherwise,} \end{cases}$$

where $\zeta = \zeta_0$ is the noise on the first step. Thus, we have

$$\hat{J}(\theta; \zeta) = s_{T-1} = (\beta + \theta)^{T-2} \zeta,$$

so

$$\nabla_\theta \hat{J}(\theta; \zeta) = (T - 2)(\beta + \theta)^{T-3} \zeta.$$

Also, note that

$$\nabla_\theta J(\theta) = \mathbb{E}_{p(\zeta)}[\nabla_\theta \hat{J}(0; \zeta)] = \mathbb{E}_{p(\zeta)}[(T-2)(\beta + \theta)^{T-3} \zeta] = 0.$$

Next, note that for n i.i.d. samples $\zeta^{(1)}, ..., \zeta^{(n)} \sim \mathcal{N}(0, \sigma_\zeta^2)$, we have

$$\hat{D}_{\text{MB}}(0) - \nabla_\theta J(0) = \frac{1}{n} \sum_{i=1}^{n} (T - 2) \beta^{T-3} \zeta^{(i)} \sim \mathcal{N}\left(0, \frac{\sigma_{\text{MB}}^2}{n}\right),$$

where

$$\sigma_{\text{MB}} = \sigma_\zeta^2 (T-2)^2 \beta^{2(T-3)}.$$

Thus, by Lemma G.8, for

$$n < \frac{\sigma_{\text{MB}}^2 \left(\log\left(\sqrt{\frac{e}{2\pi}}\right) + \log(1/\delta)\right)}{\epsilon^2},$$

we have

$$\Pr\left[|\hat{D}_{\text{MB}}(0) - \nabla_\theta J(0)| \geq \epsilon\right] = \Pr_{x\sim\mathcal{N}(0,\sigma_{\text{MB}}^2/n)}\left[|x| \geq \epsilon\right] \geq \sqrt{\frac{e}{2\pi}} \cdot e^{-n\epsilon^2/\sigma_{\text{MB}}^2} > \delta.$$

Thus, the sample complexity of $\hat{D}_{\text{MB}}(0) - \nabla_\theta J(0)$ satisfies

$$n_{\text{MB}}(\epsilon, \delta) \geq \frac{\sigma_\zeta^2 (T-2)^2 \beta^{2(T-3)} \cdot \left(\log\left(\sqrt{\frac{e}{2\pi}}\right) + \log(1/\delta)\right)}{\epsilon^2}.$$

Note that the numerator is positive as long as $\delta \leq 1/2$. The claim follows, as does the theorem statement. □

B Proof of Theorem 4.7

Preliminaries. Recall the form of the policy gradient based on Theorem 3.1:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tilde{p}_\theta(\zeta)}\left[\sum_{t=0}^{T-1} \hat{A}_\theta^{(t)}(\zeta) \nabla_\theta \log \tilde{\pi}_\theta(a_t \mid s_t)\right],$$

where, for $t \in \{0, 1, ..., T-1\}$, we have

$$\hat{A}_\theta^{(t)}(\alpha) = \hat{Q}_\theta^{(t)}(\alpha) - \tilde{V}_\theta^{(t)}(s_t),$$

where

$$\begin{aligned} \hat{Q}_\theta^{(t)}(\alpha) &= R(s_t, a_t) + \hat{Q}_\theta^{(t+1)}(\alpha) \\ \tilde{V}_\theta^{(t)}(s) &= \mathbb{E}_{p_\xi(\xi), p(\zeta)}[\tilde{R}_\theta(s) + \tilde{V}_\theta^{(t+1)}(\tilde{f}_\theta(s, \xi) + \zeta)] \\ \hat{Q}_\theta^{(T)}(\alpha) &= \tilde{V}_\theta^{(T)}(s) = 0. \end{aligned}$$

The stochastic approximation of $\nabla_\theta J(\theta)$ for a single sampled rollout $\alpha \sim \tilde{p}(\alpha)$ is

$$\hat{D}_{\text{PG}}(\theta; \alpha) = \sum_{t=0}^{T-1} \hat{A}_\theta^{(t)}(\alpha) \nabla_\theta \log \tilde{\pi}_\theta(a_t \mid s_t).$$

Bounding $\hat{Q}_\theta^{(t)} - \tilde{V}_\theta^{(t)}$. We claim that

$$\|\hat{Q}_\theta^{(t)}(\zeta) - \tilde{V}_\theta^{(t)}(s_t)\| \leq B^{(t)}(\zeta),$$

where

$$B^{(t)}(\zeta) = \sum_{i=t}^{T-1} (L_R + L_{\tilde{V}}^{(i+1)} L_f)(\|\xi_t\| + \sigma_\zeta \sqrt{d}) + L_{\tilde{V}}^{(i+1)}(\|\zeta_t\| + \sigma_\zeta \sqrt{d}),$$

where $L_{\tilde{V}}^{(t)}$ is a Lipschitz constant for $\tilde{V}_\theta^{(t)}$. We prove by induction. The base case $t = T$ is trivial. Note that $\sigma_\zeta \sqrt{d} \geq \sqrt{\mathbb{E}_{p(\zeta)}[\|\zeta\|^2]} \geq \mathbb{E}_{p(\zeta)}[\|\zeta\|]$, and similarly $\sigma_\zeta \sqrt{d} \geq \sqrt{\mathbb{E}_{p_\xi(\xi)}[\|\xi\|^2]} \geq \mathbb{E}_{p_\xi(\xi)}[\|\xi\|]$. Then, for $t \in$

$\{0, 1, ..., T-1\}$, we have

$$\begin{aligned}\|\hat{Q}_\theta^{(t)}(\zeta) - \tilde{V}_\theta^{(t)}(s_t)\| \leq& \mathbb{E}_{p_\xi(\xi)}\left[\|R(s_t, \pi_\theta(s_t) + \xi_t) - R(s_t, \pi_\theta(s_t) + \xi)\|\right] \\ &+ \|\hat{Q}_\theta^{(t+1)}(\zeta) - \tilde{V}_\theta^{(t+1)}(s_{t+1})\| \\ &+ \mathbb{E}_{p_\xi(\xi), p(\zeta)}\left[\|\tilde{V}_\theta^{(t+1)}(f(s_t, \pi_\theta(s_t) + \xi_t) + \zeta_t) - \tilde{V}_\theta^{(t+1)}(f(s_t, \pi_\theta(s_t) + \xi) + \zeta)\|\right] \\ \leq& L_R(\|\xi_t\| + \sigma_\zeta\sqrt{d}) + B^{(t+1)}(\zeta) + L_{\tilde{V}}^{(t+1)}(\|\zeta_t\| + \sigma_\zeta\sqrt{d}) + L_{\tilde{V}}^{(t+1)} L_f(\|\xi_t\| + \sigma_\zeta\sqrt{d}) \\ =& B^{(t)}(\zeta).\end{aligned}$$

The claim follows.

Bounding $\log\tilde{\pi}_\theta(a \mid s)$**.** We claim that

$$\|\nabla_\theta \log\tilde{\pi}_\theta(a \mid s)\| \leq \frac{L_\pi}{\sigma_\zeta^2} \cdot \|\xi\|,$$

where $\xi = a - \pi_\theta(s)$. Recall that $p_\xi(\xi) = \mathcal{N}(\vec{0}, \sigma_\zeta^2 I_{d_A})$. Thus, we have

$$\log\tilde{\pi}_\theta(a \mid s) = \log p_\xi(a - \pi_\theta(s)) = \log\mathcal{N}(a - \pi_\theta(s) \mid 0, \sigma_\zeta^2 I_{d_A}) = -\frac{1}{2}\log(2\pi\sigma_\zeta^2) - \frac{1}{2\sigma_\zeta^2} \cdot \|a - \pi_\theta(s)\|^2.$$

Thus, we have

$$\|\nabla_\theta \log\tilde{\pi}_\theta(a \mid s)\| = \frac{1}{2\sigma_\zeta^2} \cdot \left\|\nabla_\theta\|a - \pi_\theta(s)\|^2\right\| = \frac{1}{\sigma_\zeta^2} \cdot \left\|\nabla_\theta \pi_\theta(s)^\top (a - \pi_\theta(s))\right\| \leq \frac{L_\pi}{\sigma_\zeta^2} \cdot \|\xi\|,$$

as claimed.

Bounding the deviation of $\hat{D}_{\mathbf{PG}}$ **from** $\nabla_\theta J$**.** We claim that

$$\|\hat{D}_{\rm PG}(\theta; \zeta) - \nabla_\theta J(\theta)\| \leq 3T^4(L_R + L_{\tilde{R}_\theta})\bar{L}_f L_\pi \bar{L}_{\tilde{f}_\theta}^T d \cdot \left(4d + \frac{\tilde{E} + E + 2\sigma_\zeta\sqrt{d}}{\sigma_\zeta^2}\right),$$

where $L_{\tilde{V}} = \arg\max_{t \in \{1, ..., T\}} L_{\tilde{V}}^{(t)}$, $E = T^{-1}\sum_{t=0}^{T-1}\|\zeta_t\|$, and $\tilde{E} = T^{-1}\sum_{t=0}^{T-1}\|\xi_t\|$. First, note that

$$\begin{aligned}\|\hat{Q}_\theta^{(t)}(\zeta) - \tilde{V}_\theta^{(t)}(s_t)\| &\leq T\left((L_R + L_{\tilde{V}} L_f)(\tilde{E} + \sigma_\zeta\sqrt{d}) + L_{\tilde{V}}(E + \sigma_\zeta\sqrt{d})\right) \\ &\leq 3T^3(L_R + L_{\tilde{R}_\theta})\bar{L}_f \bar{L}_{\tilde{f}_\theta}^{T-1}(\tilde{E} + E + 2\sigma_\zeta\sqrt{d}),\end{aligned}$$

where the last step follows from the bound on $L_{\tilde{V}}^{(t)}$ in Lemma D.3. Then, we have

$$\begin{aligned}\|\hat{D}_{\rm PG}(\theta; \zeta)\| &= \left\|\sum_{t=0}^{T-1}(\hat{Q}_\theta^{(t)}(\zeta) - \tilde{V}_\theta^{(t)}(s_t))\nabla_\theta \log\tilde{\pi}_\theta(a_t \mid s_t)\right\| \\ &\leq \sum_{t=0}^{T-1}\|\hat{Q}_\theta^{(t)}(\zeta) - \tilde{V}_\theta^{(t)}(s_t)\| \cdot \|\nabla_\theta \log\tilde{\pi}_\theta(a_t \mid s_t)\| \\ &\leq \sum_{t=0}^{T-1} 3T^3(L_R + L_{\tilde{R}_\theta})\bar{L}_f \bar{L}_{\tilde{f}_\theta}^T(\tilde{E} + E + 2\sigma_\zeta\sqrt{d}) \cdot \frac{L_\pi}{\sigma_\zeta^2} \cdot \|\xi_t\| \\ &= 3T^4(L_R + L_{\tilde{R}_\theta})\bar{L}_f L_\pi \bar{L}_{\tilde{f}_\theta}^T \cdot \frac{(E + \tilde{E} + 2\sigma_\zeta\sqrt{d})\tilde{E}}{\sigma_\zeta^2}.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\|\nabla_\theta J(\theta)\| &\leq \mathbb{E}_{\tilde{p}_\theta(\zeta)}[\|\hat{D}_{\rm PG}(\theta; \zeta)\|] \\ &\leq \mathbb{E}_{\tilde{p}_\theta(\zeta)}\left[3T^4(L_R + L_{\tilde{R}_\theta})\bar{L}_f L_\pi \bar{L}_{\tilde{f}_\theta}^T \cdot \frac{(E + \tilde{E} + 2\sigma_\zeta\sqrt{d})\tilde{E}}{\sigma_\zeta^2}\right] \\ &= 12T^4(L_R + L_{\tilde{R}_\theta})\bar{L}_f L_\pi \bar{L}_{\tilde{f}_\theta}^T d,\end{aligned}$$

where we have used the fact that $\mathbb{E}_{p(\vec{\zeta})}[E] = T^{-1}\sum_{t=0}^{T-1}\mathbb{E}_{p(\zeta_t)}[\|\zeta_t\|] \le \sigma_\zeta\sqrt{d}$, and similarly $\mathbb{E}_{p_\xi(\xi)}[\tilde{E}] = T^{-1}\sum_{t=0}^{T-1}\mathbb{E}_{p_\xi(\xi)}[\|\xi_t\|] \le \sigma_\zeta\sqrt{d}$. Therefore, we have

$$\|\hat{D}_{\text{PG}}(\theta;\zeta) - \nabla_\theta J(\theta)\| \le \|\hat{D}_{\text{PG}}(\theta;\zeta)\| + \|\nabla_\theta J(\theta)\| \le 3T^4(L_R + L_{\tilde{R}_\theta})\bar{L}_f L_\pi \bar{L}_{\tilde{f}_\theta}^T d\cdot\left(4d + \frac{(\tilde{E}+E+2\sigma_\zeta\sqrt{d})\tilde{E}}{\sigma_\zeta^2}\right),$$

as claimed.

Upper bound on the sample complexity of $\hat{D}_{\text{PG}} - \nabla_\theta J$. We have $E' = (\tilde{E}+E+2\sigma_\zeta\sqrt{d})\tilde{E} \le \|\phi\|_1$, where we think of ϕ as the $T^2(d_A+d_S+1)d_A$ values $\xi_{t,i}\xi_{t',i'}$, $\zeta_{t,j}\xi_{t',i'}$, and $2\sigma_\zeta\sqrt{d}\xi_{t',i'}$, for all $t,t'\in\{0,1,...,T-1\}$, $i,i'\in[d_A]$, and $j\in[d_S]$. Since ξ_t and ζ_t are σ_ζ-sub-Gaussian for each $t\in T$, by Lemma H.6, ϕ is (τ,b)-sub-exponential, where $\tau,b = O(d\sigma_\zeta^2)$. Thus, we can apply Lemma H.7 with

$$\begin{aligned}
Y &= \hat{D}_{\text{PG}}(\theta;\zeta) - \nabla_\theta J(\theta)\\
X &= E'\\
A &= \frac{3T^4(L_R + L_{\tilde{R}_\theta})\bar{L}_f L_\pi \bar{L}_{\tilde{f}_\theta}^T d}{\sigma_\zeta^2}\\
B &= 0.
\end{aligned}$$

Thus, Y is $(\tau_{\text{PG}}, b_{\text{PG}})$-sub-exponential, where

$$\tau_{\text{PG}}, b_{\text{PG}} = O(A(\tau+b)d\log d + B) = O\left(T^6(L_R + L_{\tilde{R}_\theta})\bar{L}_f L_\pi \bar{L}_{\tilde{f}_\theta}^T d^4\log(Td)\right).$$

Thus, by Lemma G.6, the sample complexity of $\hat{D}_{\text{PG}}(\theta) - \nabla_\theta J(\theta)$ is

$$\begin{aligned}
\sqrt{n_{\text{PG}}(\epsilon,\delta)} &= \frac{\tau_{\text{PG}}\sqrt{2\log(2Td_A/\delta)}}{\epsilon}\\
&= O\left(\frac{T^6(L_R + L_{\tilde{R}_\theta})\bar{L}_f L_\pi \bar{L}_{\tilde{f}_\theta}^T d^4\log(T)\log(d)^{3/2}\log(1/\delta)^{1/2}}{\epsilon}\right),
\end{aligned}$$

for all $\epsilon \le d\tau_{\text{PG}}^2/b_{\text{PG}}$. The claim follows.

Lower bound on the sample complexity of $\hat{D}_{\text{PG}} - \nabla_\theta J$. Consider a linear dynamical system with $S = A = \mathbb{R}$, time-varying deterministic transitions

$$f_t(s,a) = \begin{cases}\beta(s+a) & \text{if } s = 0\\ \beta s & \text{otherwise,}\end{cases}$$

zero noise $p_t(\zeta) = \delta(0)$ (i.e., $\sigma_\zeta = 0$), initial state $s_0 = 0$, time-varying rewards

$$R_t(s,a) = \begin{cases} s & \text{if } t = T-1\\ 0 & \text{otherwise,}\end{cases}$$

control policy class $\pi_\theta(s) = \theta$, current parameters $\theta = 0$, and action noise p_ξ. Note that

$$a_t = \theta + \tau_\xi\xi_t,$$

where $\xi_t \sim p_\xi(\xi)$ i.i.d., so

$$s_t = \begin{cases} 0 & \text{if } t = 0\\ \beta^{t-1}(\theta+\tau_\xi\xi) & \text{otherwise,}\end{cases}$$

where $\xi = \xi_0$ is the action noise on the first step. Note that

$$\hat{Q}_\theta^{(t)}(\xi) = \beta^{T-2}(\theta+\tau_\xi\xi),$$

and

$$\tilde{V}_\theta^{(t)}(s) = \begin{cases} \mathbb{E}_{p_\xi(\xi)}[\beta^{T-2}(s+\theta+\tau_\xi\xi)] = 0 & \text{if } t=0 \\ \beta^{T-t-2}s & \text{otherwise.} \end{cases}$$

In particular, note that

$$\hat{Q}_\theta^{(t)}(\xi) - \tilde{V}_\theta^{(t)}(s_t) = \begin{cases} \beta^{T-2}(\theta+\tau_\xi\xi) & \text{if } t=0 \\ 0 & \text{otherwise.} \end{cases}$$

Also, note that $\nabla_\theta J(\theta) = \beta^{T-2}$. Therefore, we have

$$\nabla_\theta \log \tilde{\pi}(a \mid s) = \nabla_\theta \log p_\xi\left(\frac{a-\theta}{\tau_\xi}\right) = -\frac{\nabla_\xi p_\xi\left(\frac{a-\theta}{\tau_\xi}\right)}{\tau_\xi \cdot p_\xi\left(\frac{a-\theta}{\tau_\xi}\right)} = -\frac{1}{\tau_\xi}\cdot\nabla_\xi \log p_\xi\left(\frac{a-\theta}{\tau_\xi}\right).$$

Thus, for i.i.d. samples $\xi^{(1)}, ..., \xi^{(n)} \sim p_\xi(\xi)$, we have

$$\begin{aligned}
\hat{D}_{\text{PG}}(0) - \nabla_\theta J(0) &= \frac{1}{n}\sum_{i=1}^n \left(\hat{Q}_\theta^{(t)}(\xi^{(i)}) - \tilde{V}_\theta^{(t)}(s_t^{(i)})\right)\cdot\left(-\nabla_\theta \log\tilde{\pi}(a_t^{(i)} \mid s_t^{(i)})\right) - \beta^{T-2} \\
&= \frac{1}{n}\sum_{i=1}^n \beta^{T-2}\tau_\xi\xi^{(i)}\cdot\left(-\frac{1}{\tau_\xi}\cdot\nabla_\xi\log p_\xi(\xi^{(i)})\right) - \beta^{T-2} \\
&= -\beta^{T-2}\left[1+\frac{1}{n}\sum_{i=1}^n \xi^{(i)}\cdot\nabla_\xi\log p_\xi(\xi^{(i)})\right].
\end{aligned}$$

Note that for $p_\xi(\xi)$ satisfying our conditions (differentiable on $\mathbb{R}$ and satisfying $\lim_{\xi\to\pm\infty}\xi\cdot p_\xi(\xi) = 0$), we have

$$\mathbb{E}_{p_\xi(\xi)}[\xi\cdot\nabla_\xi\log p_\xi(\xi)] = \int_{-\infty}^{\infty}\xi\cdot\nabla_\xi p_\xi(\xi)d\xi = -\int_{-\infty}^{\infty}p_\xi(\xi)d\xi = -1, \tag{2}$$

where the second-to-last step follows from integration by parts. Thus, by the definition of the sample complexity,

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n \xi^{(i)}\cdot\nabla_\xi\log p_\xi(\xi^{(i)}) + 1\right| \geq \epsilon\right] > \delta$$

for any $n < n_\xi(\epsilon,\delta)$, so we have

$$\Pr\left[|\hat{D}_{\text{PG}}(0) - \nabla_\theta J(0)| \geq \epsilon\right] = \Pr\left[\beta^{T-2}\left|\frac{1}{n}\sum_{i=1}^n \xi^{(i)}\cdot\nabla_\xi\log p_\xi(\xi^{(i)}) + 1\right| \geq \beta^{T-2}\epsilon\right] > \delta.$$

for any $n < n_\xi(\epsilon/\beta^{T-2},\delta)$. Thus, we have

$$n_{\text{PG}}(\epsilon,\delta) \geq n_\xi(\epsilon/\beta^{T-2},\delta).$$

Next, consider the case where $p_\xi(\xi) = \mathcal{N}(\xi \mid 0, \sigma^2)$, for any $\sigma \in \mathbb{R}_+$. Then, we have

$$\nabla_\xi \log p_\xi(\xi) = \nabla_\xi\left(-\log\sqrt{2\pi} - \frac{\|\xi\|^2}{2\sigma^2}\right) = -\frac{\xi}{\sigma^2},$$

so

$$\hat{D}_{\text{PG}}(0) - \nabla_\theta J(0) = \beta^{T-2}\left[-1+\frac{1}{n\sigma^2}\sum_{i=1}^n(\xi^{(i)})^2\right] = \beta^{T-2}\left[-1+\frac{1}{n}\sum_{i=1}^n(x^{(i)})^2\right],$$

where $x^{(i)} \sim \mathcal{N}(0,1)$ are i.i.d. standard Gaussian random variables for $i \in [n]$. By Lemma H.8, letting $x = n^{-1}\sum_{i=1}^{n}(x^{(i)})^2$ (so $\mu_x = \mathbb{E}_{p(x)} = 1$), for

$$n \leq \min\left\{\frac{2\beta^{T-2}\left(\frac{1}{2}\log(1/\delta) + \log(1/e^2\sqrt{2})\right)}{\epsilon}, \frac{1}{\delta}\right\},$$

we have

$$\Pr\left[\hat{D}_{\text{PG}}(0) - \nabla_\theta J(0) \geq \epsilon\right] = \Pr_{p(x)}\left[x \geq \mu_x + \frac{\epsilon}{\beta^{T-2}}\right] \geq \frac{1}{\sqrt{n}} \cdot \frac{1}{e^2\sqrt{2}} e^{-\frac{n\epsilon}{2\beta^{T-2}}} \geq \sqrt{\delta}\cdot\sqrt{\delta} = \delta.$$

Thus, the sample complexity of $\hat{D}_{\text{PG}} - \nabla_\theta J(\theta)$ satisfies

$$n_{\text{PG}}(\epsilon,\delta) \geq \min\left\{\frac{2\beta^{T-2}\left(\frac{1}{2}\log(1/\delta) + \log(1/e^2\sqrt{2})\right)}{\epsilon}, \frac{1}{\delta}\right\}.$$

Note that the numerator is positive as long as $\delta \leq 1/12$. The claim follows, as does the theorem statement. □

C Proof of Theorem 4.11

Preliminaries. Note that the expected cumulative reward is equivalent to

$$\begin{aligned} J(\theta) &= V_\theta^{(0)}(s_0) \\ V_\theta^{(t)}(s) &= R_\theta(s) + \mathbb{E}_{p(\zeta)}\left[V_\theta^{(t+1)}(f_\theta(s) + \zeta)\right] \qquad (\forall t \in \{0,1,...,T-1\}) \\ V_\theta^{(T)}(s) &= 0. \end{aligned}$$

Similarly, given a sample $\vec{\zeta} \sim p(\vec{\zeta})$, the stochastic approximation of the expected cumulative reward is

$$\begin{aligned} \hat{J}(\theta;\vec{\zeta}) &= \hat{V}_\theta^{(0)}(s_0;\vec{\zeta}) \\ \hat{V}_\theta^{(t)}(s;\vec{\zeta}) &= R_\theta(s) + \hat{V}_\theta^{(t+1)}(f_\theta(s) + \zeta_t;\vec{\zeta}) \qquad (\forall t \in \{0,1,...,T-1\}) \\ \hat{V}_\theta^{(T)}(s;\vec{\zeta}) &= 0. \end{aligned}$$

The finite difference approximation of $\nabla_\theta J(\theta)$ is

$$D_{\text{FD}}(\theta) = \sum_{k=1}^{d_\Theta} \frac{J(\theta + \lambda\nu^{(k)}) - J(\theta - \lambda\nu^{(k)})}{2\lambda} \cdot \nu^{(k)},$$

where $\nu^{(k)}$ is a basis vector for $k \in [d]$ and d_Θ is the dimension of the parameter space $\Theta = \mathbb{R}^d$. Finally, an estimate of the finite difference approximation for two samples $\zeta, \eta \sim \tilde{p}(\zeta)$ is

$$\hat{D}_{\text{FD}}(\theta;\vec{\zeta},\vec{\eta}) = \sum_{k=1}^{d_\Theta} \frac{\hat{J}(\theta + \lambda\nu^{(k)};\vec{\zeta}) - \hat{J}(\theta - \lambda\nu^{(k)};\vec{\eta})}{2\lambda} \cdot \nu^{(k)},$$

where $\hat{J}(\theta;\vec{\zeta})$ is as defined in the proof of Theorem 4.6.

Bounding the deviation of $\hat{V}_\theta^{(t)}$ from $V_\theta^{(t)}$. We claim that for $t \in \{0,1,...,T\}$, we have

$$\|\hat{V}_\theta^{(t)}(s;\vec{\zeta}) - V_\theta^{(t)}(s)\| \leq B^{(t)}(\vec{\zeta})$$

for all $\theta \in \Theta$ and $s \in S$, where

$$B^{(t)}(\vec{\zeta}) = \sum_{i=t}^{T-1} L_V^{(i+1)}(\|\zeta_i\| + \sigma_\zeta\sqrt{d_A}),$$

where $L_V^{(t)}$ is a Lipschitz constant for $V_\theta^{(t)}$. The base case $t = T$ follows trivially. Note that $\sigma_\zeta\sqrt{d_A} \geq \sqrt{\mathbb{E}_{p(\zeta)}[\|\zeta\|^2]} \geq \mathbb{E}_{p(\zeta)}[\|\zeta\|]$. Then, for $t \in \{0, 1, ..., T-1\}$, we have

$$\begin{aligned}\|\hat{V}_\theta^{(t)}(s;\vec{\zeta}) - V_\theta^{(t)}(s)\| &= \left\|\hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) - \mathbb{E}_{p(\zeta)}\left[V_\theta^{(t+1)}(f_\theta(s)+\zeta)\right]\right\| \\ &\leq \|\hat{V}_\theta^{(t+1)}(f_\theta(s)+\zeta_t;\vec{\zeta}) - V_\theta^{(t+1)}(f_\theta(s)+\zeta_t)\| \\ &\quad + \mathbb{E}_{p(\zeta)}\left[\|V_\theta^{(t+1)}(f_\theta(s)+\zeta_t) - V_\theta^{(t+1)}(f_\theta(s)+\zeta)\|\right] \\ &\leq B^{(t+1)}(\vec{\zeta}) + L_V^{(t+1)}(\|\zeta_t\| + \sigma_\zeta\sqrt{d_A}) \\ &= B^{(t)}(\vec{\zeta}).\end{aligned}$$

The claim follows.

Bounding the deviation of $\hat{D}_{\mathbf{FD}}$ from $D_{\mathbf{FD}}$. Let

$$D_{\mathrm{FD}}(\theta)\mathbb{E}_{p(\vec{\zeta}),p(\vec{\eta})}[\hat{D}_{\mathrm{FD}}(\theta)].$$

Then, letting $L_{\nabla V} = \arg\max_{t\in\{0,1,...,T\}} L_{\nabla V}^{(t)}$, note that

$$\|\hat{J}(\theta;\vec{\zeta}) - J(\theta)\| \leq B^{(0)}(\vec{\zeta}) = \sum_{i=0}^{T-1} L_V^{(i+1)}(\|\zeta_i\| + \sigma_\zeta\sqrt{d_A}) \leq 3T^3 L_{R_\theta}\bar{L}_{f_\theta}^T(E + \sigma_\zeta\sqrt{d_A}),$$

where $E = T^{-1}\sum_{t=0}^{T-1}\|\zeta_t\|$. Thus, we have

$$\begin{aligned}\|\hat{D}_{\mathrm{FD}}(\theta;\zeta,\eta)_k - D_{\mathrm{FD}}(\theta)_k\| &= \left\|\frac{\hat{J}(\theta+\lambda\nu^{(k)};\vec{\zeta}) - \hat{J}(\theta-\lambda\nu^{(k)};\vec{\eta})}{2\lambda}\cdot\nu^{(k)} - \frac{J(\theta+\lambda\nu^{(k)}) - J(\theta-\lambda\nu^{(k)})}{2\lambda}\cdot\nu^{(k)}\right\| \\ &\leq \frac{\|\hat{J}(\theta+\lambda\nu^{(k)};\vec{\zeta}) - J(\theta+\lambda\nu^{(k)})\| + \|\hat{J}(\theta-\lambda\nu^{(k)};\vec{\eta}) - J(\theta-\lambda\nu^{(k)})\|}{2\lambda} \\ &\leq \frac{3T^3 L_{R_\theta}\bar{L}_{f_\theta}^T(E + \tilde{E} + 2\sigma_\zeta\sqrt{d_A})}{2\lambda}\end{aligned}$$

for $k \in [d_\Theta]$, where $\tilde{E} = T^{-1}\sum_{t=0}^{T-1}\|\eta_t\|$.

Upper bound on the sample complexity of $\hat{D}_{\mathbf{FD}} - D_{\mathbf{FD}}$. Note that $E + \tilde{E} \leq \|E'\|_1$, where $E' = \vec{\zeta}\circ\vec{\eta}$ is the length $2Td_S$ concatenation of the vectors $\zeta_0, \zeta_1, ..., \zeta_{T-1}, \eta_0, \eta_1, ..., \eta_{T-1}$, so E' is σ_ζ-sub-Gaussian. We apply Lemma G.7 with

$$\begin{aligned}Y &= \hat{D}_{\mathrm{FD}}(\theta;\vec{\zeta},\vec{\eta})_k - D_{\mathrm{FD}}(\theta)_k \\ X &= E' \\ A &= \frac{3T^3 L_{R_\theta}\bar{L}_{f_\theta}^T}{\lambda} \\ B &= A\sigma_\zeta\sqrt{d_A}.\end{aligned}$$

Thus, Y is σ_{FD}-sub-Gaussian, where

$$\begin{aligned}\sigma_{\mathrm{FD}} &= \max\{10A\sigma(2Td_A)\log(2Td_A), 5A\sigma_\zeta\sqrt{d_A})\} \\ &= 20A\sigma_\zeta Td_A\log(Td_A) \\ &\leq \frac{60T^4 L_{R_\theta}\bar{L}_{f_\theta}^T\sigma_\zeta d_A\log(Td_A)}{\lambda}.\end{aligned}$$

Thus, by Lemma G.6, for $k \in [d_\Theta]$, the sample complexity of $\hat{D}_{\mathrm{FD}}(\theta)_k - D_{\mathrm{FD}}(\theta)_k$ is

$$\begin{aligned}\sqrt{\tilde{n}_{\mathrm{FD}}(\tilde{\epsilon},\tilde{\delta})} &= \frac{\sigma_{\mathrm{FD}}\sqrt{2\log(2d_A/\tilde{\delta})}}{\tilde{\epsilon}} \\ &= O\left(\frac{T^4 L_{R_\theta}\bar{L}_{f_\theta}^T\sigma_\zeta d_A\log(T)\log(d_A)^{3/2}\log(1/\tilde{\delta})^{1/2}}{\lambda\tilde{\epsilon}}\right).\end{aligned}$$

Upper bound on the sample complexity of $\hat{D}_{\mathbf{FD}} - \nabla_\theta J(\theta)$. By Theorem 3.3, we have

$$\nabla_\theta J(\theta) = D_{\mathrm{FD}}(\theta) + \Delta,$$

where

$$\|\Delta\| \le L_{\nabla J} d_A \lambda \le 44T^5 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{4T} d_A \lambda,$$

where the second inequality follows from the fact that $L_{\nabla J} = L_{\nabla V}^{(0)}$ and the bound on $L_{\nabla V}^{(0)}$ in Lemma D.2. Now, taking

$$\begin{aligned}
\lambda &= \frac{\epsilon}{88T^5 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{4T} d_A} \\
\tilde{\epsilon} &= \frac{\epsilon}{2\sqrt{d_\Theta}} \\
\tilde{\delta} &= \frac{\delta}{d_\Theta},
\end{aligned}$$

then with probability $1-\delta$, we have

$$\|\hat{D}_{\mathrm{FD}}(\theta) - \nabla_\theta J(\theta)\| \le \|\hat{D}_{\mathrm{FD}}(\theta) - D_{\mathrm{FD}}(\theta)\| + \|\Delta\| \le \epsilon,$$

so the sample complexity of $\hat{D}_{\mathrm{FD}}(\theta) - \nabla_\theta J(\theta)$ is

$$\sqrt{n_{\mathrm{FD}}(\epsilon,\delta)} = = O\left(\frac{T^9 \bar{L}_{R_\theta}^2 \bar{L}_{f_\theta}^{5T} \sigma_\zeta d_A^2 \sqrt{d_\Theta} \log(T) \log(d_A)^{3/2} \log(d_\Theta)^{1/2} \log(1/\tilde{\delta})^{1/2}}{\epsilon^2}\right).$$

The claim follows.

Lower bound on the sample complexity of $\hat{D}_{\mathbf{FD}} - \nabla_\theta J(\theta)$. Consider a linear dynamical system with $S = \mathbb{R}^2$, $A = \mathbb{R}$, time-varying deterministic transitions

$$f_t((s,s'),a) = \begin{cases} \beta(s, s'+a) & \text{if } s = 0 \\ \beta(s,s') & \text{otherwise,} \end{cases}$$

time-varying noise

$$p_t((\zeta,0)) = \begin{cases} \mathcal{N}(\zeta \mid 0, \sigma_\zeta^2) & \text{if } t = 0 \\ \delta(0) & \text{otherwise,} \end{cases}$$

where $\sigma_\zeta \in \mathbb{R}$, initial state $s_0 = (0,0)$, time-varying rewards

$$R_t((s,s'),a) = \begin{cases} s + \phi(s') & \text{if } t = T-1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is defined by

$$\phi(x) = \begin{cases} 2x - 1 & \text{if } x \ge 1 \\ x^2 & \text{if } -1 \le x < 1 \\ 2x + 1 & \text{if } x < -1, \end{cases}$$

control policy class $\pi_\theta((s,s')) = \theta$, and current parameters $\theta = 0$. Note that technically, R is not twice continuously differentiable, so it does not satisfy Assumption 4.2. However, the only place in the proof of Theorem 4.11 where we need this assumption is to apply Lemma F.2 in Lemma D.2. By the discussion in the proof of Lemma F.2, the lemma still applies, so our theorems still apply to this dynamical system. Now, we have

$$s_t = \begin{cases} 0 & \text{if } t = 0 \\ \beta^{t-1}(\zeta,\theta) & \text{otherwise,} \end{cases}$$

where $\zeta = \zeta_0$ is the noise on the first step. Thus, we have

$$\hat{J}(\theta;\zeta) = s_{T-1} + s'_{T-1} = \beta^{T-2}\zeta + \phi(\beta^{T-2}\theta).$$

Also, note that

$$\nabla_\theta J(\theta) = \mathbb{E}_{p(\zeta)}[\nabla_\theta \hat{J}(0;\zeta)] = \phi'(\beta^{T-2}\theta)\cdot\beta^{T-2},$$

so $\nabla_\theta J(0) = 0$, since $\phi'(0) = 0$.

Next, note that for $2n$ i.i.d. samples $\zeta^{(1)}, ..., \zeta^{(n)}, \eta^{(1)}, ..., \eta^{(n)} \sim \mathcal{N}(0, \sigma_\zeta^2)$, we have

$$\begin{aligned}\hat{D}_{\text{FD}}(0) - \nabla_\theta J(0) &= \frac{1}{2\lambda}\left[\frac{1}{n}\sum_{i=1}^n \hat{J}(\lambda;\zeta^{(i)}) - \frac{1}{n}\sum_{i=1}^n \hat{J}(-\lambda;\eta^{(i)})\right] \\ &= \frac{1}{2\lambda}\cdot\frac{1}{n}\sum_{i=1}^n\left[\beta^{T-2}\zeta^{(i)} - \beta^{T-2}\eta^{(i)}\right] + \frac{1}{2\lambda}\left[\phi(\beta^{T-2}\lambda) - \phi(-\beta^{T-2}\lambda)\right].\end{aligned}$$

Letting $\zeta^{(n+i)} = -\eta^{(i)}$ for $i \in [n]$, and using the fact that $\phi(-x) = -\phi(x)$, we have

$$\hat{D}_{\text{FD}}(0) - \nabla_\theta J(0) = \frac{1}{2\lambda n}\sum_{i=1}^{2n}\beta^{T-2}\zeta^{(i)} + \frac{1}{\lambda}\cdot\phi(\beta^{T-2}\lambda) \sim \mathcal{N}\left(\mu_{\text{FD}}, \frac{\sigma_{\text{FD}}}{n}\right).$$

where

$$\begin{aligned}\mu_{\text{FD}} &= \phi(\beta^{T-2}\lambda) \\ \sigma_{\text{FD}} &= \frac{\beta^{T-2}\sigma_\zeta}{\lambda}.\end{aligned}$$

Thus, by Lemma G.8, for

$$n \le \frac{\sigma_{\text{FD}}^2\left(\log\left(\sqrt{\frac{e}{2\pi}}\right) + \log(1/\tilde{\delta})\right)}{\epsilon^2},$$

and recalling that $D_{\text{FD}}(\theta) = \mathbb{E}_{p_\theta(\alpha)}[\hat{D}_{\text{FD}}(\theta;\alpha)] = \mu_{\text{FD}}$, we have

$$\Pr\left[\hat{D}_{\text{FD}}(0) - D_{\text{FD}}(0) \ge \tilde{\epsilon}\right] = \Pr_{x\sim\mathcal{N}(0,\sigma_{\text{FD}}^2/n)}[|x| \ge \tilde{\epsilon}] \ge \sqrt{\frac{e}{2\pi}}\cdot e^{-n\epsilon^2/\sigma_{\text{FD}}^2} \ge \tilde{\delta}.$$

Thus, the sample complexity of $\hat{D}_{\text{FD}}(0) - D_{\text{FD}}(0)$ satisfies

$$\tilde{n}_{\text{FD}}(\tilde{\epsilon},\tilde{\delta}) \ge \frac{\sigma_{\text{FD}}^2\left(\log\left(\sqrt{\frac{e}{2\pi}}\right) + \log(1/\tilde{\delta})\right)}{\tilde{\epsilon}^2}.$$

Now, recall that $\nabla_\theta J(0) = 0$, so

$$\Pr\left[\hat{D}_{\text{FD}}(0) - \nabla_\theta J(0) \ge \epsilon\right] = \Pr\left[\hat{D}_{\text{FD}}(0) \ge \epsilon\right] = \Pr\left[\hat{D}_{\text{FD}}(0) - D_{\text{FD}}(0) \ge \epsilon - \mu_{\text{FD}}\right].$$

Thus, using our assumption $\delta \le 1/2$, then we need to have $\mu_{\text{FD}} \le \epsilon$ for $\Pr\left[\hat{D}_{\text{FD}}(0) - \nabla_\theta J(0) \ge \epsilon\right] \le \delta$ to hold. As a consequence, using our assumption $\epsilon \le 1$, we must have

$$\epsilon \ge \mu_{\text{FD}} = \phi(\beta^{T-2}\lambda) = \beta^{2(T-2)}\lambda^2,$$

where the last step follows since $0 \le \phi(\beta^{T-2}\lambda) \le 1$ implies $\phi(x) = x^2$. Thus, we have $\lambda \le \sqrt{\frac{\epsilon}{\beta^{2(T-2)}}}$, so we have $\sigma_{\text{FD}} \ge \beta^{4(T-2)}\sigma_\zeta^2/\epsilon$. Finally, we have

$$\Pr\left[\hat{D}_{\text{FD}}(0) - \nabla_\theta J(0)) \ge \epsilon\right] \ge \Pr\left[\hat{D}_{\text{FD}}(0) - D_{\text{FD}}(0) \ge \epsilon\right],$$

so the sample complexity of $\hat{D}_{\text{FD}}(0) - \nabla_\theta J(\theta)$ satisfies

$$\begin{aligned} n_{\text{FD}}(\epsilon, \delta) \geq \tilde{n}_{\text{FD}}(\epsilon, \delta) &\geq \frac{\sigma_{\text{FD}}^2 (T-2)^2 \beta^{2(T-3)} \cdot \left(\log\left(\sqrt{\frac{e}{2\pi}}\right) + \log(1/\delta)\right)}{\epsilon^2} \\ &\geq \frac{(T-2)^2 \beta^{6(T-3)} \sigma_\zeta^2 \cdot \left(\log\left(\log(1/\delta) + \sqrt{\frac{e}{2\pi}}\right)\right)}{\epsilon^4}. \end{aligned}$$

Finally, for any $d_\Theta \in \mathbb{N}$, we can consider d_Θ independent copies of this dynamical system. Then, estimating the gradient $\nabla_\theta J(\theta)$ is equivalent to estimating $\frac{dJ}{d\theta_i}(\theta)$ for each $i \in [d_\Theta]$. Thus, we have

$$n_{\text{FD}}(\epsilon, \delta) \geq \tilde{n}_{\text{FD}}(\epsilon, \delta) \geq \frac{(T-2)^2 \beta^{6(T-3)} \sigma_\zeta^2 d_\Theta \cdot \left(\log\left(\log(1/\delta) + \sqrt{\frac{e}{2\pi}}\right)\right)}{\epsilon^4}.$$

The claim follows, as does the theorem statement. □

D Bounds on Lipschitz Constants

We prove bounds on the Lipschitz constants $L_V^{(t)}$ for $V_\theta^{(t)}$, $L_{\nabla V}^{(t)}$ for $\nabla V_\theta^{(t)}$, and $L_{\tilde{V}}^{(t)}$ for $\tilde{V}_\theta^{(t)}$. We use implicitly use the commonly known results in Appendix F throughout these proofs.

Lemma D.1. *We claim that for $t \in \{0, 1, ..., T\}$, $V_\theta^{(t)}$ is $L_V^{(t)}$-Lipschitz, where*

$$L_V^{(t)} \leq 3T^2 L_{R_\theta} \bar{L}_{f_\theta}^{T-t-1}.$$

Proof. First, we show that $V_\theta^{(t)}$ is $L_{V,\theta}^{(t)}$-Lipschitz in θ and $L_{V,s}^{(t)}$-Lipschitz in s, where

$$\begin{aligned} L_{V,\theta}^{(t)} &= \sum_{i=t}^{T-1} (L_{R_\theta} + L_{f_\theta} L_{V,s}^{(i+1)}) \\ L_{V,s}^{(t)} &= \sum_{i=t}^{T-1} L_{f_\theta}^{i-t} L_{R_\theta}, \end{aligned}$$

We prove by induction. The base case $t = T$ is trivial. Then, for $t \in \{0, 1, ..., T-1\}$, note that $V_\theta^{(t)}$ is $(L_{V,\theta}^{(t)})'$-Lipschitz in θ, where

$$(L_{V,\theta}^{(t)})' = L_{R_\theta} + L_{V,\theta}^{(t+1)} + L_{f_\theta} L_{V,s}^{(t+1)} = L_{V,\theta}^{(t)}.$$

Similarly, note that $V_\theta^{(t)}$ is $(L_{V,s}^{(t)})'$-Lipschitz in s, where

$$(L_{V,s}^{(t)})' = L_{R_\theta} + L_{f_\theta} L_{V,s}^{(t+1)} = L_{V,s}^{(t)},$$

as was to be shown. Finally, note that

$$L_{V,s}^{(t)} \leq T L_{R_\theta} \bar{L}_{f_\theta}^{T-t-1},$$

so

$$L_{V,\theta}^{(t)} \leq T(L_{R_\theta} + L_{f_\theta} \cdot T L_{R_\theta} \bar{L}_{f_\theta}^{T-t-2}) \leq 2T^2 L_{R_\theta} \bar{L}_{f_\theta}^{T-t-1}.$$

Thus, $V_\theta^{(T)}$ is $(L_V^{(t)})'$-Lipschitz, where

$$(L_V^{(t)}) \leq L_{V,\theta}^{(t)} + L_{V,s}^{(t)} \leq 3T^2 L_{R_\theta} \bar{L}_{f_\theta}^{T-t-1} = L_V^{(t)}.$$

The claim follows. □

Lemma D.2. *We claim that for $t \in \{0,1,...,T\}$, $\nabla V_\theta^{(t)}$ is $L_{\nabla V}^{(t)}$-Lipschitz, where*

$$L_{\nabla V}^{(t)} = 44T^5 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{4(T-t-1)}.$$

Proof. First, we show that $\nabla_\theta V_\theta^{(t)}$ is $L_{\nabla V,\theta,\theta}^{(t)}$-Lipschitz in θ and $L_{\nabla V,\theta,s}^{(t)}$-Lipschitz in s, and that $\nabla_s V_\theta^{(t)}$ is $L_{\nabla V,\theta,s}^{(t)}$-Lipschitz in θ and $L_{\nabla V,s,s}^{(t)}$-Lipschitz in s, where

$$\begin{aligned}
L_{\nabla V,\theta,\theta}^{(t)} &= \sum_{i=t}^{T-1} (L_{\nabla R_\theta} + 2L_{f_\theta} L_{\nabla V,\theta,s}^{(i+1)} + L_{f_\theta}^2 L_{\nabla V,s,s}^{(i+1)} + L_{\nabla f_\theta} L_V^{(i+1)}) \\
L_{\nabla V,\theta,s}^{(t)} &= \sum_{i=t}^{T-1} L_{f_\theta}^{i-t} (L_{\nabla R_\theta} + L_{f_\theta}^2 L_{\nabla V,s,s}^{(i+1)} + L_{\nabla f_\theta} L_V^{(i+1)}) \\
L_{\nabla V,s,s}^{(t)} &= \sum_{i=t}^{T-1} L_{f_\theta}^{2(i-t)} (L_{\nabla R_\theta} + L_{\nabla f_\theta} L_V^{(i+1)}) \\
L_{\nabla V,\theta,\theta}^{(T)} &= L_{\nabla V,\theta,s}^{(T)} = L_{\nabla V,s,s}^{(T)} = 0.
\end{aligned}$$

We prove by induction. The base case $t = T$ is trivial. First, for $t \in \{0,1,...,T-1\}$, note that $\nabla_\theta V_\theta^{(t)}$ is $(L_{\nabla V,\theta,\theta}^{(t)})'$-Lipschitz in θ, where

$$(L_{\nabla V,\theta,\theta}^{(t)})' = L_{\nabla R_\theta} + L_{\nabla V,\theta,\theta}^{(t+1)} + L_{f_\theta} L_{\nabla V,\theta,s}^{(t+1)} + L_{f_\theta}(L_{\nabla V,\theta,s}^{(t+1)} + L_{f_\theta} L_{\nabla V,s,s}^{(t+1)}) + L_{\nabla f_\theta} L_V^{(t+1)} = L_{\nabla V,\theta,\theta}^{(t)}.$$

Second, note that $\nabla_\theta V_\theta^{(t)}$ is $(L_{\nabla V,\theta,s}^{(t)})'$-Lipschitz in s, where

$$(L_{\nabla V,\theta,s}^{(t)})' = L_{\nabla R_\theta} + L_{f_\theta} L_{\nabla V,\theta,s}^{(t+1)} + L_{f_\theta}^2 L_{\nabla V,s,s}^{(t+1)} + L_{\nabla f_\theta} L_V^{(t+1)} = L_{\nabla V,\theta,s}^{(t)}.$$

Third, note that $\nabla_s V_\theta^{(t)}$ is $(L_{\nabla V,s,\theta}^{(t)})'$-Lipschitz in θ, where

$$(L_{\nabla V,s,\theta}^{(t)})' = L_{\nabla R_\theta} + L_{f_\theta}(L_{\nabla V,\theta,s}^{(t+1)} + L_{f_\theta} L_{\nabla V,s,s}^{(t+1)}) + L_{\nabla f_\theta} L_V^{(t+1)} = L_{\nabla V,\theta,s}^{(t)}.$$

Fourth, note that $\nabla_s V_\theta^{(t)}$ is $(L_{\nabla V,s,s}^{(t)})'$-Lipschitz in s, where

$$(L_{\nabla V,s,s}^{(t)})' = L_{\nabla R_\theta} + L_{f_\theta}^2 L_{\nabla V,s,s}^{(t+1)} + L_{\nabla f_\theta} L_V^{(t+1)} = L_{\nabla V,s,s}^{(t)},$$

as was to be shown. Finally, note that

$$L_{\nabla V,s,s}^{(t)} \leq T\bar{L}_{f_\theta}^{2(T-t-1)} (L_{\nabla R_\theta} + L_{\nabla f_\theta} \cdot 3T^2 L_{R_\theta} \bar{L}_{f_\theta}^{T-t-2}) \leq 4T^3 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{3(T-t-1)},$$

so

$$L_{\nabla V,\theta,s}^{(t)} \leq T\bar{L}_{f_\theta}^{T-t-1} (L_{\nabla R_\theta} + L_{f_\theta}^2 \cdot 4T^3 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{3(T-t-2)} + L_{\nabla f_\theta} \cdot 3T^2 L_{R_\theta} \bar{L}_{f_\theta}^{T-t-2}) \leq 8T^4 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{4(T-t-1)}$$

so

$$\begin{aligned}
L_{\nabla V,\theta,\theta}^{(t)} &\leq T(L_{\nabla R_\theta} + 2L_{f_\theta} \cdot 8T^4 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{4(T-t-2)} + L_{f_\theta}^2 \cdot 4T^3 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{3(T-t-2)} + L_{\nabla f_\theta} \cdot 3T^2 L_{R_\theta} \bar{L}_{f_\theta}^{T-t-2}) \\
&\leq 24T^5 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{4(T-t-1)}.
\end{aligned}$$

Thus, $\nabla V_\theta^{(t)}$ is $(L_{\nabla V}^{(t)})'$-Lipschitz, where

$$(L_{\nabla V}^{(t)})' = L_{\nabla V,\theta,\theta} + 2L_{\nabla V,\theta,s} + L_{\nabla V,s,s} \leq 44T^5 \bar{L}_{R_\theta} \bar{L}_{f_\theta}^{4(T-t-1)} = L_{\nabla V}^{(t)}.$$

The claim follows. □

Lemma D.3. *We claim that for $t \in \{0,1,...,T\}$, $\tilde{V}_\theta^{(t)}$ is $L_{\tilde{V}}^{(t)}$-Lipschitz, where*

$$L_{\tilde{V}}^{(t)} = 3T^2 L_{\tilde{R}_\theta} \bar{L}_{\tilde{f}_\theta}^{T-t-1}.$$

Proof. Note that $\tilde{V}_\theta^{(t)}$ is exactly equal to $V_\theta^{(t)}$ with R_θ replaced with $\tilde{R}_\theta$ and f_θ replaced with $\tilde{f}_\theta$. Thus, the claim follows by the same argument as for Lemma D.1. □

E Proof of Theorem 3.3

Theorem E.1. *(Taylor's theorem) Let $f : \mathbb{R} \to \mathbb{R}$ be an everywhere differentiable function with $L_{f'}$-Lipschitz derivative. Then, for any $x, \epsilon \in \mathbb{R}$, we have*

$$f(x+\epsilon) = f(x) + f'(x) \cdot \epsilon + \Delta,$$

where

$$|\Delta| \leq \frac{L_{f'}\epsilon^2}{2}.$$

Proof. The claim follows from Theorem 5.15 in Rudin et al. (1976), together with Lemma F.2, which implies that $|f''(x)| \leq L_{f'}$ for all $x \in \mathbb{R}$. □

Now, we prove Theorem 3.3. By Taylor's theorem, we have

$$f(x+\mu) = f(x) + \langle \nabla f(x), \mu \rangle + \Delta(\mu),$$

where

$$\|\Delta(\mu)\| \leq \frac{1}{2} L_{\nabla f} \|\mu\|^2.$$

Thus, we have

$$\begin{aligned}
&\sum_{k=1}^{d} \frac{f(x+\lambda\nu^{(k)}) - f(x-\lambda\nu^{(k)})}{2\lambda} \cdot \nu^{(k)} \\
&= \sum_{k=1}^{d} \frac{(f(x) + \langle \nabla f(x), \lambda\nu^{(k)} \rangle + \Delta(\lambda\nu^{(k)})) - (f(x) - \langle \nabla f(x), \lambda\nu^{(k)} \rangle + \Delta(-\lambda\nu^{(k)}))}{2\lambda} \cdot \nu^{(k)} \\
&= \sum_{k=1}^{d} \langle \nabla f(x), \nu^{(k)} \rangle \cdot \nu^{(k)} + \frac{\Delta(\lambda\nu^{(k)}) - \Delta(-\lambda\nu^{(k)})}{2\lambda} \cdot \nu^{(k)} \\
&= \sum_{k=1}^{d} \nu^{(k)}((\nu^{(k)})^\top \nabla f(x)) + \sum_{k=1}^{d} \frac{\Delta(\lambda\nu^{(k)}) - \Delta(-\lambda\nu^{(k)})}{2\lambda} \cdot \nu^{(k)} \\
&= \nabla f(x) + \sum_{k=1}^{d} \frac{\Delta(\lambda\nu^{(k)}) - \Delta(-\lambda\nu^{(k)})}{2} \cdot \nu^{(k)}
\end{aligned}$$

Therefore, we have

$$\Delta = \sum_{k=1}^{d} \frac{\Delta(\lambda\nu^{(k)}) - \Delta(-\lambda\nu^{(k)})}{2\lambda} \cdot \nu^{(k)},$$

so

$$\|\Delta\| \leq \sum_{k=1}^{d} \left\| \frac{\Delta(\lambda\nu^{(k)}) - \Delta(-\lambda\nu^{(k)})}{2\lambda} \cdot \nu^{(k)} \right\| \leq \frac{1}{2} L_{\nabla f} \lambda \cdot \|\nu^{(k)}\|^3 \leq L_{\nabla f} d\lambda,$$

as claimed. □

F Technical Lemmas (Lipschitz Constants)

We define Lipschitz continuity (for the L_2 norm), and prove a number of standard results.

Definition F.1. A function $f : \mathcal{X} \to \mathcal{Y}$ (where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^{d'}$) is L_f-*Lipschitz continuous* if for all $x, x' \in \mathcal{X}$,

$$\|f(x) - f(x')\| \leq L_f \|x - x'\|. \tag{3}$$

If $\mathcal{X}$ is a space of matrices or tensors, we assume x and x' are unrolled into vectors. in (3).

Lemma F.2. *If $f : \mathcal{X} \to \mathcal{Y}$ is L_f-Lipschitz and continuously differentiable, then for all $x \in \mathcal{X}$,*

$$\|\nabla f(x)\| \leq L_f.$$

Proof. Note that

$$\nabla f(x) = \lim_{\|\epsilon\| \to 0} \frac{f(x+\epsilon) - f(x)}{\|\epsilon\|},$$

so

$$\|\nabla f(x)\| = \lim_{\|\epsilon\| \to 0} \frac{\|f(x+\epsilon) - f(x)\|}{\|\epsilon\|} \leq \lim_{\|\epsilon\| \to 0} \frac{L_f \|\epsilon\|}{\|\epsilon\|} = L_f,$$

as claimed. Note that the result holds even if each component f_i is continuously differentiable except on a finite set X. In particular, for each point $x \in X$, we can use the standard definition $(\nabla f(x))_i = (f'_{i,+}(x) + f'_{i,-}(x))/2$, where $f'_{i,+}(x)$ is the right derivative and $f'_{i,-}(x)$ is the left deriviative. Letting $(\nabla_+ f(x))_i = f'_{i,+}(x)$ and $(\nabla_- f(x))_i = f'_{i,-}(x)$, then $\nabla f(x) = (\nabla_+ f(x) + \nabla_- f(x))/2$. Then, we have

$$\|\nabla f(x)\| \leq \frac{\|\nabla_+ f(x)\| + \|\nabla_- f(x)\|}{2} \leq L_f,$$

as claimed. ∎

Lemma F.3. *If $f, g : \mathcal{X} \to \mathcal{Y}$ are L_f- and L_g-Lipschitz, respectively, then $h(x) = f(x) + g(x)$ is L_h-Lipschitz, where $L_h = L_f + L_g$.*

Proof. Note that

$$\|h(x) - h(x')\| \leq \|f(x) - f(x')\| + \|g(x) - g(x')\| \leq (L_f + L_g)\|x - x'\| = L_h \|x - x'\|,$$

as claimed. ∎

Lemma F.4. *If $f, g : \mathcal{X} \to \mathcal{Y}$ where f is L_f-Lipschitz and bounded by M_f (i.e., $|f(x)| \leq M_f$ for all $x \in \mathcal{X}$), and g is L_g-Lipschitz and bounded by M_g. Then $h(x) = f(x) \cdot g(x)$ is L_h-Lipschitz, where $L_h = M_g L_f + M_f L_g$.*

Proof. Note that

$$\begin{aligned}\|h(x) - h(x')\| &\leq \|(f(x) - f(x'))g(x)\| + \|(g(x) - g(x'))f(x')\| \\ &\leq M_g L_f \|x - x'\| + M_f L_g \|x - x'\| \\ &= L_h \|x - x'\|,\end{aligned}$$

as claimed. ∎

Lemma F.5. *If $f : \mathcal{X} \to \mathcal{Y}$ is L_f-Lipschitz and $g : \mathcal{Y} \to \mathcal{Z}$ is L_g-Lipschitz, then $h(x) = g(f(x))$ is L_h-Lipschitz, where $L_h = L_g L_f$.*

Proof. Note that

$$\|g(f(x)) - g(f(x'))\| \leq L_g \|f(x) - f(x')\| \leq L_g L_f \|x - x'\| \leq L_h \|x - x'\|,$$

as claimed. ∎

Lemma F.6. *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be $L_{f,x}$-Lipschitz in $\mathcal{X}$ (for all $y \in \mathcal{Y}$) and $L_{f,y}$-Lipschitz in $\mathcal{Y}$ (for all $x \in \mathcal{X}$). Then, f is L_f-Lipschitz in $\mathcal{X} \times \mathcal{Y}$, where $L_f = L_{f,x} + L_{f,y}$.*

Proof. Note that

$$\begin{aligned}
\|f(x,y) - f(x',y')\| &\le \|f(x,y) - f(x',y)\| + \|f(x',y) - f(x',y')\| \\
&\le L_{f,x}\|x - x'\| + L_{f,y}\|y - y'\| \\
&\le L_{f,x}\|(x,y) - (x',y')\| + L_{f,y}\|(x,y) - (x',y')\| \\
&\le (L_{f,x} + L_{f,y})\|(x,y) - (x',y')\| \\
&= L_f\|(x,y) - (x',y')\|,
\end{aligned}$$

as claimed. □

Lemma F.7. *Let $f : \mathcal{X} \to \mathcal{Y}$ be L_f-Lipschitz, and $g : \mathcal{X} \to \mathcal{Z}$ be L_g-Lipchitz. Then, $h(x) = (f(x), g(x))$ is L_h-Lipschitz, where $L_h = L_f + L_g$.*

Proof. Note that

$$\begin{aligned}
\|h(x) - h(x')\| &\le \|(f(x) - f(x'), g(x) - g(x'))\| \\
&= \sqrt{\sum_{i=1}^{d_{\mathcal{Y}}}(f_i(x) - f_i(x'))^2 + \sum_{j=1}^{d_{\mathcal{Z}}}(g_i(x) - g_i(x'))^2} \\
&\le \sqrt{\sum_{i=1}^{d_{\mathcal{Y}}}(f_i(x) - f_i(x'))^2} + \sqrt{\sum_{j=1}^{d_{\mathcal{Z}}}(g_i(x) - g_i(x'))^2} \\
&= \|f(x) - f(x')\| + \|g(x) - g(x')\| \\
&\le L_f\|x - x'\| + L_g\|x - x'\| \\
&\le (L_f + L_g)\|x - x'\| \\
&= L_h\|x - x'\|,
\end{aligned}$$

as claimed. □

Lemma F.8. *Let $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be L_f-Lipschitz. Then, $g(x) = \mathbb{E}_{p(z)}[f(x,z)]$ (where $p(z)$ is a distribution over $\mathcal{Z}$) is L_g-Lipschitz, where $L_g = L_f$.*

Proof. Note that

$$\|g(x) - g(x')\| \le \mathbb{E}_{p(z)}\left[\|f(x,z) - f(x',z)\|\right] \le L_f\|x - x'\| = L_g\|x - x'\|,$$

as claimed. □

G Technical Lemmas (Sub-Gaussian Random Variables)

We define sub-Gaussian random variables, and prove a number of standard results. We also prove Lemma G.7, a key lemma that enables us to infer a sub-Gaussian constant for a random variable bounded Y in norm by a sub-Gaussian random variable X, i.e., $\|Y\| \le A\|X\|_1 + B$ (where $\|\cdot\|$ is the L_2 norm). This lemma is a key step in the proofs of our upper bounds for the model-based and finite-difference policy gradient estimators. Finally, we also prove Lemma G.8, which is a key step in the proof of our lower bounds.

Definition G.1. A random variable X over $\mathbb{R}$ is *σ_X-sub-Gaussian* if $\mathbb{E}[X] = 0$, and for all $t \in \mathbb{R}$, we have $\mathbb{E}[e^{tX}] \le e^{\sigma_X^2 t^2/2}$.

Lemma G.2. *If a random variable X over $\mathbb{R}$ is σ_X-sub-Gaussian, then $\mathbb{E}[|X|^2] \le \sigma_X^2$.*

Proof. See Stromberg (1994). □

Lemma G.3. *(Hoeffding's inequality) Let $x_1, ..., x_n \sim p_X(x)$ be i.i.d. σ_X-sub-Gaussian random variables over $\mathbb{R}$. Then,*

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_n\right| \geq \epsilon\right] \leq 2e^{-\frac{n\epsilon^2}{2\sigma_X^2}}.$$

Proof. See Proposition 2.1 of Wainwright (2019). □

Definition G.4. A random vector X over $\mathbb{R}^d$ is σ_X-sub-Gaussian if each X_i is σ_X-sub-Gaussian.

Lemma G.5. *If a random vector X over $\mathbb{R}^d$ is σ_X-sub-Gaussian, then $\mathbb{E}[\|X\|] \leq \sigma_X\sqrt{d}$.*

Proof. Note that

$$\mathbb{E}[\|X\|] = \mathbb{E}\left[\sqrt{\sum_{i=1}^{d}\|X_i\|^2}\right] \leq \sqrt{\sum_{i=1}^{d}\mathbb{E}[\|X_i\|^2]} \leq \sigma_X\sqrt{d},$$

where the first inequality follows from Jensen's inequality. □

Lemma G.6. *Let X be random vector over $\mathbb{R}^d$ with mean $\mu_X = \mathbb{E}[X]$, such that $X - \mu_X$ is σ_X-sub-Gaussian. Then, given $\epsilon, \delta \in \mathbb{R}_+$, the sample complexity of X satisfies*

$$n_X(\epsilon, \delta) \leq \frac{2\sigma_X^2 \log(2d/\delta)}{\epsilon^2},$$

i.e., given $x_1, ..., x_n \sim p_X(x)$ i.i.d. samples of X with empirical mean $x = n^{-1}\sum_{i=1}^{n} x_n$, then $Pr[\|x - \mu_X\| \geq \epsilon] \leq \delta$.

Proof. Note that

$$\Pr[\|x - \mu_X\| \geq \epsilon] \leq \Pr[\|x - \mu_X\|_1 \geq \epsilon] \leq \sum_{i=1}^{d} \Pr\left[|x_i - \mu_{X,i}| \geq \frac{\epsilon}{d}\right] \leq 2de^{-\frac{nt^2}{2\sigma_X^2}} \leq \delta,$$

as claimed. □

Lemma G.7. *Let X be a σ_X-sub-Gaussian random vector over $\mathbb{R}^d$, and let Y be a random vector over $\mathbb{R}^{d'}$ satisfying*

$$\|Y\| \leq A\|X\|_1 + B,$$

where $A, B \in \mathbb{R}_+$. Then Y is σ_Y-sub-Gaussian, where

$$\sigma_Y = \max\{10A\sigma_X d \log d, 5B\}.$$

Proof. We first prove that $|Y_i|$ is bounded for each $i \in [d]$, and then use this fact to prove that Y_i is sub-Gaussian. In particular, we claim that for any $i \in [d]$ and any $t \in \mathbb{R}_+$, we have

$$\Pr[|Y_i| \geq t] \leq 2e^{-\frac{t^2}{2\tilde{\sigma}_Y^2}},$$

where

$$\tilde{\sigma}_Y = \max\left\{4A\sigma_X d\sqrt{\log d}, 2B\right\}.$$

To this end, note that by Theorem 5.1 in Lattimore and Szepesvári (2018), for any $i \in [d]$ and any $t \in \mathbb{R}_+$, we have

$$\Pr[|X_i| \geq t] \leq 2e^{-\frac{t^2}{2\sigma_X^2}}.$$

Now, note that

$$\Pr[|Y_i| \geq t] \leq \Pr[\|Y\| \geq t] \leq \Pr\left[\|X\|_1 \geq \frac{t-B}{A}\right] \leq \sum_{i=1}^{d} \Pr\left[|X_i| \geq \frac{t-B}{Ad}\right] \leq 2de^{-\frac{(t-B)^2}{(Ad\sigma_X\sqrt{2})^2}}.$$

We consider three cases. First, suppose that $t \geq \max\{4A\sigma_X d\sqrt{\log d}, 2B\}$. Then, $(t-B)^2 \geq (t/2)^2$, so

$$\Pr[|Y_i| \geq t] \leq 2de^{-\frac{t^2}{(Ad\sigma_X\sqrt{8})^2}} = 2e^{-\frac{t^2-(Ad\sigma_X\sqrt{8})^2\log d}{(Ad\sigma_X\sqrt{8})^2}}.$$

Furthermore, $t^2 - (Ad\sigma_X\sqrt{8})^2 \log d \geq (t^2/2)$, so

$$\Pr[|Y_i| \geq t] \leq 2e^{-\frac{t^2-(Ad\sigma_X\sqrt{8})^2\log d}{(Ad\sigma_X\sqrt{8})^2}} \leq 2e^{-\frac{t^2}{2(Ad\sigma_X\sqrt{8})^2}} \leq 2e^{-\frac{t^2}{2\tilde{\sigma}_Y^2}}.$$

Second, if $t \leq 2B$, then

$$2e^{-\frac{t^2}{2\tilde{\sigma}_Y^2}} \geq 2e^{-\frac{(2B)^2}{2\tilde{\sigma}_Y^2}} = 2e^{-1/2} > 1,$$

so

$$\Pr[|Y_i| \geq t] \leq 1 \leq 2e^{-\frac{t^2}{2\tilde{\sigma}_Y^2}}.$$

Third, if $t \leq 4A\sigma_X d\sqrt{\log d}$, then

$$2e^{-\frac{t^2}{2\tilde{\sigma}_Y^2}} \geq 2e^{-\frac{(4A\sigma_X d\sqrt{\log d})^2}{2\tilde{\sigma}_Y^2}} \geq 2e^{-1/2} > 1,$$

so

$$\Pr[|Y_i| \geq t] \leq 1 \leq 2e^{-\frac{t^2}{2\tilde{\sigma}_Y^2}}.$$

As a consequence, by Note 5.4.2 in Lattimore and Szepesvári (2018), Y_i is $\tilde{\sigma}_Y\sqrt{5}$-sub-Gaussian. Note that $\sigma_Y \geq \tilde{\sigma}_Y\sqrt{5}$, so the theorem follows. ☐

Lemma G.8. *Given $\sigma \in \mathbb{R}_+$,*

$$Pr_{x\sim\mathcal{N}(0,\sigma^2)}[|x| \geq t] \geq \sqrt{\frac{e}{2\pi}} \cdot e^{-t^2/\sigma^2}.$$

Proof. By Theorem 2 in Chang et al. (2011), we have

$$1 - \Phi(t) \geq \frac{1}{2}\sqrt{\frac{e}{2\pi}} \cdot e^{-t^2},$$

where $\Phi(t)$ is the cumulative distribution function of $\mathcal{N}(0,1)$. Thus, for $\epsilon \in \mathbb{R}_+$, we have

$$\Pr_{x\sim\mathcal{N}(0,\sigma^2)}[|x| \geq t] = \Pr_{z\sim\mathcal{N}(0,1)}\left[|z| \geq \frac{t}{\sigma}\right] = 2\left(1 - \Phi\left(\frac{t}{\sigma}\right)\right) \geq \sqrt{\frac{e}{2\pi}} \cdot e^{-t^2/\sigma^2} \geq \delta.$$

The claim follows. ☐

H Technical Lemmas (Sub-Exponential Random Variables)

We define sub-exponential random variables, and prove a number of standard results. Additionally, we prove Lemma H.7 (an analog of Lemma G.7), a key lemma that enables us to infer a sub-exponential constant for a random variable bounded Y in norm by a sub-exponential random variable X, i.e., $\|Y\| \leq A\|X\|_1 + B$ (where $\|\cdot\|$ is the L_2 norm). This lemma is a key step in the proof of our upper bound in Theorem 4.7. Finally, we also prove Lemma H.8, which is a key step in the proof of our lower bound in Theorem 4.7.

Definition H.1. A random variable X over $\mathbb{R}$ is (τ_X, b_X)-*sub-exponential* if $\mathbb{E}[X] = 0$, and for all $t \in \mathbb{R}$ satisfying $|t| \leq b_X^{-1}$, we have $\mathbb{E}[e^{tX}] \leq e^{\tau_X^2 t^2/2}$.

Lemma H.2. *Let $x_1, ..., x_n \sim p_X(x)$ be i.i.d. (τ_X, b_X)-sub-exponential random variables over $\mathbb{R}$. Then, we have*

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_n\right| \geq \epsilon\right] \leq \begin{cases} 2e^{-\frac{n\epsilon^2}{2\tau_X^2}} & \text{if } |\epsilon| \leq \tau_X^2/b_X \\ 2e^{-\frac{n\epsilon}{2b_X}} & \text{otherwise.} \end{cases}$$

Proof. See (2.20) in Wainwright (2019). ◻

Definition H.3. A random vector X over $\mathbb{R}^d$ is (τ_X, b_X)-sub-exponential if each X_i is (τ_X, b_X)-sub-exponential.

Lemma H.4. *Let X be a random vector over $\mathbb{R}^d$ with mean $\mu_X = \mathbb{E}[X]$, such that $X - \mu_X$ is (τ_X, b_X)-sub-exponential. Then, given $\epsilon, \delta \in \mathbb{R}_+$ such that $\epsilon \leq d\tau_X^2/b_X$, the sample complexity of X satisfies*

$$n_X(\epsilon, \delta) = \frac{2\tau_X^2 \log(2d/\delta)}{\epsilon^2},$$

i.e., given $x_1, ..., x_n \sim p_X(x)$ i.i.d. samples of X with empirical mean $x = n^{-1}\sum_{i=1}^{n} x_n$, then $Pr[\|x - \mu_X\| \geq \epsilon] \leq \delta$.

Proof. Note that

$$\Pr[\|x - \mu_X\| \geq \epsilon] \leq \Pr[\|x - \mu_X\|_1 \geq \epsilon] \leq \sum_{i=1}^{d} \Pr\left[|x_i - \mu_{X,i}| \geq \frac{\epsilon}{d}\right] \leq 2de^{-\frac{nt^2}{2\tau_X^2}} \leq \delta,$$

as claimed. ◻

Lemma H.5. *Let X be σ_X-sub-Gaussian. Then, X^2 is (τ_X, b_X)-sub-exponential, where $\tau_X, b_X = O(\sigma_X^2)$.*

Proof. The result follows from Lemma 5.5, Lemma 5.14, and the discussion preceding Definition 5.13 in Vershynin (2010). In particular, using the notation in Vershynin (2010), by Lemma 5.5, we have that X satisfies $\|X\|_{\psi_2} = O(\sigma_X)$. Then, by Lemma 5.14, we have that $\|X^2\|_{\psi_1} = 2\|X\|_{\psi_2}^2 = O(\sigma_X^2)$. Finally, by the discussion preceding Definition 5.13, we have that X^2 is (τ_X, b_X)-sub-exponential with parameters $\tau_X, b_X = O(\|X^2\|_{\psi_1}) = O(\sigma_X^2)$. The claim follows. ◻

Lemma H.6. *Let X and Y be σ_X-sub-Gaussian, respectively. Then, $Z = XY$ is (τ_Z, b_Z)-sub-exponential, where $\tau_Z, b_Z = O(\sigma_X^2)$.*

Proof. Note that

$$Z = XY = \frac{(X+Y)^2 - (X-Y)^2}{4}.$$

By Lemma H.5, we have $X + Y$ and $X - Y$ are (τ, b)-sub-exponential for $\tau, b = O(\sigma_X^2)$, so Z is τ_Z, b_Z-sub-exponential, for $\tau_Z, b_Z = O(\tau + b) = O(\sigma_X^2)$, as claimed. ◻

Lemma H.7. *Let X be a (τ_X, b_X)-sub-exponential random vector over $\mathbb{R}^d$, and let Y be a random vector over $\mathbb{R}^{d'}$ satisfying*

$$\|Y\| \leq A\|X\|_1 + B,$$

where $A, B \in \mathbb{R}_+$. Then Y is (τ_Y, b_Y)-sub-exponential, where $\tau_Y, b_Y = O(A(\tau_X + b_X)d \log d + B)$.

Proof. We use Lemma 5.14 and the discussion preceding Definition 5.13 in Vershynin (2010). In particular, let $\tilde{\tau}_X = \max\{\tau_X, b_X\}$; then, from the definition of sub-exponential random variables with $t = \tilde{\tau}_X^{-1}$, we have

$$\mathbb{E}\left[e^{\frac{X_i}{\tilde{\tau}}}\right] \leq \mathbb{E}\left[e^{\frac{t^2}{2\tilde{\tau}_X^2}}\right] \leq e$$

for each $i \in [d]$. Thus, using the notation in Vershynin (2010), so by the discussion preceding the Definition 5.13 in Vershynin (2010), we have X_i satisfies $\|X_i\|_{\psi_1} = O(\tilde{\tau}_X)$, and furthermore satisfies

$$\Pr[|X_i| \geq t] \leq 3e^{-t/K}$$

for all $t \in \mathbb{R}_+$, where $K = O(\|X_i\|_{\psi_1}) = O(\tilde{\tau}_X)$. Thus, for each $i \in [d]$, we have

$$\Pr[|Y_i| \geq t] \leq \Pr\left[\|X\|_1 \geq \frac{t-B}{A}\right] \leq \sum_{i=1}^{d} \Pr\left[|X_i| \geq \frac{t-B}{Ad}\right] \leq de^{1-\frac{t-B}{AKd}}.$$

Now, let

$$\tilde{\tau}_Y = \max\{4AKd\log d, 2B\}.$$

We consider three cases. First, suppose that $t \geq \max\{4AKd\log d, 2B\}$. Then, $t - B \geq t/2$, so

$$\Pr[|Y_i| \geq t] \leq de^{1-\frac{t}{2AKd}} = e^{1-\frac{t-2AKd\log d}{2AKd}}.$$

Furthermore, $t - 2AKd\log d \geq t/2$, so

$$\Pr[|Y_i| \geq t] \leq e^{1-\frac{t-2AKd\log d}{2AKd}} \leq e^{1-\frac{t}{4AKd}} \leq e^{1-\frac{t}{\tilde{\tau}_Y}}.$$

Second, if $t \leq 2B$, then

$$e^{1-\frac{t}{\tilde{\tau}_Y}} \geq e^{1-\frac{2B}{\tilde{\tau}_Y}} \geq 1,$$

so

$$\Pr[|Y_i| \geq t] \leq 1 \leq e^{1-\frac{t}{\tilde{\tau}_Y}}.$$

Third, if $t \leq 4AKd\log d$, then

$$e^{1-\frac{t}{\tilde{\tau}_Y}} \geq e^{1-\frac{4AKd\log d}{\tilde{\tau}_Y}} \geq 1,$$

so

$$\Pr[|Y_i| \geq t] \leq 1 \leq e^{1-\frac{t}{\tilde{\tau}_Y}}.$$

As a consequence, by the discussion preceding Definition 5.13 in Vershynin (2010), we have Y_i satisfies $\|Y_i\|_{\psi_1} = O(\tilde{\tau}_Y)$. Thus, by Lemma 5.15 in Vershynin (2010), we have that Y_i is (τ_Y, b_Y)-sub-exponential, where

$$\tau_Y, b_Y = O(\|Y_i\|_{\psi_1}) = O(\tilde{\tau}_Y) = O(AKd\log d + B) = O(A\tilde{\tau}_X d\log d + B) = O(A(\tau_X + b_X)d\log d + B).$$

The claim follows. □

Lemma H.8. *Given $\sigma \in \mathbb{R}_+$, let*

$$x = \frac{(x^{(1)})^2 + ... + (x^{(n)})^2}{n},$$

where $x^{(1)}, ..., x^{(n)} \sim \mathcal{N}(0, \sigma^2)$ i.i.d., and let $\mu_x = \mathbb{E}_{p(x)}[x] = \sigma^2$. Then, we have

$$Pr_{p(x)}[x \geq \mu_x + \epsilon] \geq \frac{1}{e^2\sqrt{2n}} e^{-\frac{n\epsilon}{2\sigma^2}}.$$

Proof. Let $z = (z^{(1)})^2 + ... + (z^{(n)})^2$ be the sum of the squares of n i.i.d. standard Gaussian random variables $z^{(1)}, ..., z^{(n)} \sim \mathcal{N}(0, 1)$. We assume that $n = 2k$ is even. Then, z is distributed according to the χ^2_{2k} distribution, which has density function

$$p_{2k}(z) = \frac{1}{2^k(k-1)!} z^{k-1} e^{-z},$$

and mean $\mu_{2k} = 2k$. For $z \geq \mu_{2k} = 2k$, we have

$$p_{2k}(z) \geq \frac{1}{2^k(k-1)!}(2k)^{k-1}e^{-z/2} = \frac{1}{2} \cdot \frac{k^{k-1}}{(k-1)!}e^{-z/2} \geq \frac{1}{2} \cdot \frac{k^{k-1}}{(k-1)^{k-1/2}e^{-k+2}}e^{-z/2} \geq \frac{1}{2e^2\sqrt{k}}e^{k-z/2},$$

where the second inequality follows from a result

$$n! \leq n^{n+1/2}e^{1-n}$$

based on Stirling's approximation Robbins (1955). Thus, for any $\epsilon \in \mathbb{R}_+$, we have

$$\mathrm{Pr}_{z\sim\chi^2_{2k}}[z \geq \mu_{2k} + \epsilon] \geq \int_{\mu_{2k}+\epsilon}^{\infty} \frac{1}{2e^2\sqrt{k}}e^{k-z/2} = \frac{1}{2e^2\sqrt{k}}e^{k-(\mu_{2k}+\epsilon)/2} = \frac{1}{2e^2\sqrt{k}}e^{-\epsilon/2}.$$

Finally, for $x = ((x^{(1)})^2 + ... + (x^{(n)})^2)/n$, where $x^{(1)}, ..., x^{(n)} \sim \mathcal{N}(0, \sigma^2)$ i.i.d., note that $x = \frac{\sigma^2 z}{n}$ and

$$\mu_x = \mathbb{E}_{p(x)}[x] = \frac{\sigma^2\mu_n}{n} = \sigma^2,$$

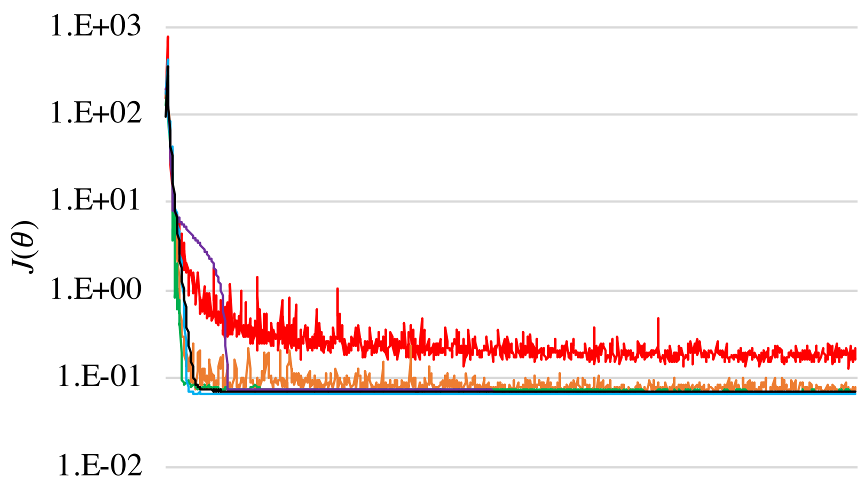
so we have

$$\mathrm{Pr}_{p(x)}[x \geq \mu_x + \epsilon] = \mathrm{Pr}_{z\sim\chi^2_n}\left[z \geq \mu_n + \frac{n\epsilon}{\sigma^2}\right] \geq \frac{1}{e^2\sqrt{2n}}e^{-\frac{n\epsilon}{2\sigma^2}}.$$

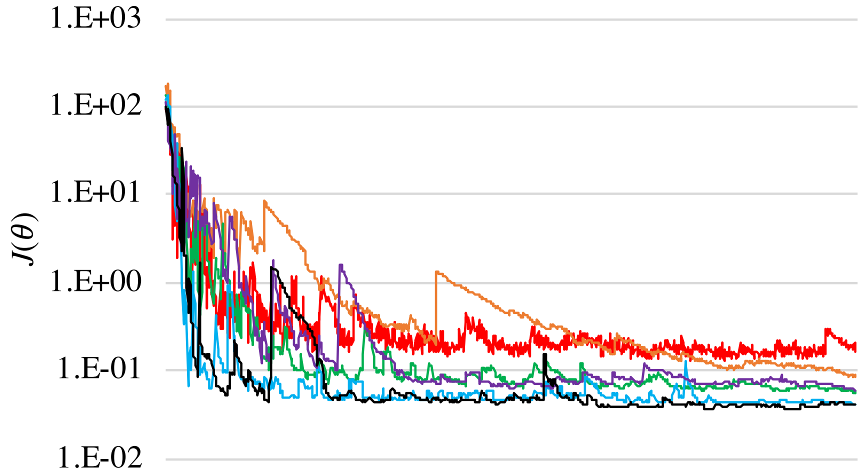
The claim follows. □

I Experimental Results

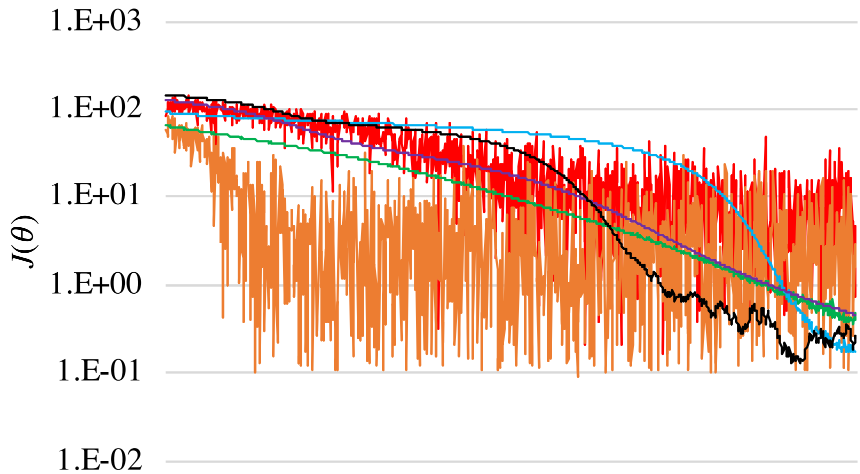
We show enlarged versions of the plots from Figure 1:



Model-Based Algorithm



Finite-Differences Algorithm



Policy Gradient Theorem Algorithm