

Appendix—Supplementary Materials

In the appendix, we give proofs of the theorems. First, we give some preliminaries.

If $X \sim \chi^2(k)$, then the non-central moments are given by

$$\mathbb{E}[X^n] = 2^n \frac{\Gamma(n + k/2)}{\Gamma(k/2)} = k(k+2) \cdots (k+2n-2),$$

where $\Gamma(z)$ is the Gamma function defined as

$$\Gamma(z) := \int_0^{+\infty} t^{z-1} e^{-t} dt.$$

The Gamma function satisfies $\Gamma(z+1) = z\Gamma(z)$, $\Gamma(1/2) = \sqrt{\pi}$, and $\Gamma(1) = 1$.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, central absolute moments (the moments of $|X - \mu|$) are given by

$$\mathbb{E}[|x - \mu|^p] = \begin{cases} \sigma^p (p-1)!! \sqrt{2/\pi}, & p \text{ is odd,} \\ \sigma^p (p-1)!! & p \text{ is even,} \end{cases}$$

where $n!!$ denotes the double factorial defined by

$$n!! := \begin{cases} n \cdot (n-2) \cdots 5 \cdot 3 \cdot 1 & n \text{ is positive odd,} \\ n \cdot (n-2) \cdots 6 \cdot 4 \cdot 2 & n \text{ is positive even,} \\ 1 & n = 1 \text{ or } 0. \end{cases}$$

A Proof of Theorem 1

For notational brevity, we denote the i -th component of $\mathbf{f}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})$ and the i -th component of $\mathbf{g}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\tau}} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})$ as

$$\begin{aligned} f_i(\boldsymbol{\theta}) &= \nabla_{\eta_i} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}) = \frac{\theta_i - \eta_i}{\tau_i^2}, \\ g_i(\boldsymbol{\theta}) &= \nabla_{\tau_i} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}) = \frac{(\theta_i - \eta_i)^2 - \tau_i^2}{\tau_i^3}. \end{aligned}$$

Proof. According to Eq.(1), we have

$$\begin{aligned} \text{Var}[R(h)\mathbf{f}(\boldsymbol{\theta})] &\leq \sum_{i=1}^{\ell} \mathbb{E}[(Rf_i)^2] \\ &= \sum_{i=1}^{\ell} \int p(\theta_i) \left(\sum_{t=1}^T \gamma^{t-1} r(\mathbf{s}_t, a_t, \mathbf{s}_{t+1}) \right)^2 \left(\frac{\theta_i - \eta_i}{\tau_i^2} \right)^2 d\theta_i \\ &\leq \sum_{i=1}^{\ell} \int p(\theta_i) \left(\sum_{t=1}^T \gamma^{t-1} \beta \right)^2 \left(\frac{\theta_i - \eta_i}{\tau_i^2} \right)^2 d\theta_i \\ &= \sum_{i=1}^{\ell} \int p(\theta_i) \left(\frac{\beta(1 - \gamma^T)}{1 - \gamma} \right)^2 \left(\frac{\theta_i - \eta_i}{\tau_i^2} \right)^2 d\theta_i \\ &= \sum_{i=1}^{\ell} \frac{\beta^2(1 - \gamma^T)^2}{\tau_i^2(1 - \gamma)^2} \mathbb{E} \left[\left(\frac{\theta_i - \eta_i}{\tau_i} \right)^2 \right]. \end{aligned}$$

Let $\psi_i = ((\theta_i - \eta_i)/\tau_i)^2$ for $i = 1, \dots, \ell$. We could know that $\psi_i \sim \chi^2(1)$ and $\mathbb{E}[\psi_i] = 1$ since $\theta_i \sim \mathcal{N}(\eta_i, \tau_i^2)$, and thus

$$\text{Var}[R(h)\mathbf{f}(\boldsymbol{\theta})] \leq \frac{\beta^2(1 - \gamma^T)^2 B}{(1 - \gamma)^2}.$$

Hence the first part of Theorem 1 follows due to

$$\mathbf{Var} \left[\nabla_{\eta} \hat{J}(\rho) \right] = \frac{1}{N} \mathbf{Var} [R(h) \mathbf{f}(\theta)].$$

Similarly,

$$\begin{aligned} \mathbf{Var} [R(h) \mathbf{g}(\theta)] &\leq \sum_{i=1}^{\ell} \mathbb{E} [(Rg_i)^2] \\ &\leq \sum_{i=1}^{\ell} \frac{\beta^2(1-\gamma^T)^2}{\tau_i^2(1-\gamma)^2} \mathbb{E} \left[\left(\left(\frac{\theta_i - \eta_i}{\tau_i} \right)^2 - 1 \right)^2 \right]. \end{aligned}$$

Let $\psi_i = ((\theta_i - \eta_i)/\tau_i)^2$ for $i = 1, \dots, \ell$. Since $\theta_i \sim \mathcal{N}(\eta_i, \tau_i^2)$, we could know that

$$\mathbb{E} [(\psi_i - 1)^2] = \mathbb{E} [\psi_i^2] - 2\mathbb{E}[\psi_i] + 1 = 2.$$

Hence

$$\mathbf{Var} [R(h) \mathbf{g}(\theta)] \leq \frac{2\beta^2(1-\gamma^T)^2 B}{(1-\gamma)^2}.$$

Notice that

$$\mathbf{Var} \left[\nabla_{\tau} \hat{J}(\rho) \right] = \frac{1}{N} \mathbf{Var} [R(h) \mathbf{g}(\theta)],$$

which completes the proof. \square

B Proof of Theorem 2

To begin with, we note that $\boldsymbol{\mu}$ is a vector and σ is a scalar in REINFORCE. We denote the i -th component of $\mathbf{f}(h) = \sum_{t=1}^T \nabla_{\boldsymbol{\mu}} \log p(a_t | \mathbf{s}_t, \boldsymbol{\theta})$ and the scalar function $g(h)$ as

$$\begin{aligned} f_i(h) &= \sum_{t=1}^T \nabla_{\mu_i} \log p(a_t | \mathbf{s}_t, \boldsymbol{\theta}) = \sum_{t=1}^T \frac{a_t - \boldsymbol{\mu}^\top \mathbf{s}_t}{\sigma^2} s_{t,i}, \\ g(h) &= \sum_{t=1}^T \nabla_{\sigma} \log p(a_t | \mathbf{s}_t, \boldsymbol{\theta}) = \sum_{t=1}^T \frac{(a_t - \boldsymbol{\mu}^\top \mathbf{s}_t)^2 - \sigma^2}{\sigma^3}, \end{aligned}$$

where all functions above are parameterized by $\boldsymbol{\theta}$.

Proof. Since

$$\begin{aligned} \mathbf{Var} [\nabla_{\boldsymbol{\mu}} \hat{J}(\boldsymbol{\theta})] &= \frac{1}{N} \mathbf{Var} [R(h) \mathbf{f}(h)], \\ \mathbf{Var} [\nabla_{\sigma} \hat{J}(\boldsymbol{\theta})] &= \frac{1}{N} \mathbf{Var} [R(h) g(h)], \end{aligned}$$

we can just focus on the bounds of $\mathbf{Var} [R(h) \mathbf{f}(h)]$ and $\mathbf{Var} [R(h) g(h)]$.

The upper bound of $\mathbf{Var} [R(h) \mathbf{f}(h)]$:

$$\begin{aligned} \mathbf{Var} [R(h) \mathbf{f}(h)] &\leq \sum_{i=1}^{\ell} \mathbb{E} [(Rf_i)^2] \\ &= \mathbb{E} [R^2 \mathbf{f}^\top \mathbf{f}] \\ &= \int_h p(h) \left(\sum_{t=1}^T \gamma^{t-1} r(\mathbf{s}_t, a_t, \mathbf{s}_{t+1}) \right)^2 \left(\sum_{t=1}^T \frac{a_t - \boldsymbol{\mu}^\top \mathbf{s}_t}{\sigma^2} \mathbf{s}_t \right)^\top \left(\sum_{t=1}^T \frac{a_t - \boldsymbol{\mu}^\top \mathbf{s}_t}{\sigma^2} \mathbf{s}_t \right) dh \\ &\leq \frac{\beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2} \mathbb{E} \left[\left(\sum_{t,t'=1}^T \frac{(a_t - \boldsymbol{\mu}^\top \mathbf{s}_t)(a_{t'} - \boldsymbol{\mu}^\top \mathbf{s}_{t'})}{\sigma^2} \mathbf{s}_t^\top \mathbf{s}_{t'} \right) \right]. \end{aligned}$$

Let $\xi_t = (a_t - \boldsymbol{\mu}^\top \mathbf{s}_t)/\sigma$ for $t = 1, \dots, T$. Then, ξ_1, \dots, ξ_T are independent standard normal variables because of $a_t \sim \mathcal{N}(\boldsymbol{\mu}^\top \mathbf{s}_t, \sigma^2)$. Since all $\nabla_{\boldsymbol{\mu}} \log p(a_t | \mathbf{s}_t, \boldsymbol{\theta})$ in $\mathbf{f}(h)$ are parameterized by the states \mathbf{s}_t , and the stochasticity of ξ_t comes only from a_t , it is sufficient to consider fixed states. Given $\{\mathbf{s}_t\}_{t=1}^T$, $\xi_1 \mathbf{s}_1, \dots, \xi_T \mathbf{s}_T$ are ℓ -dimensional independent normal variables with zero means, that is, $\mathbb{E}[\xi_t \mathbf{s}_t] = \mathbf{0}$. Hence,

$$\begin{aligned} \mathbb{E} \left[\left(\sum_{t,t'=1}^T \frac{(a_t - \boldsymbol{\mu}^\top \mathbf{s}_t)(a_{t'} - \boldsymbol{\mu}^\top \mathbf{s}_{t'})}{\sigma^2} \mathbf{s}_t^\top \mathbf{s}_{t'} \right) \right] &= \mathbb{E} \left[\left(\sum_{t,t'=1}^T \xi_t \xi_{t'} \mathbf{s}_t^\top \mathbf{s}_{t'} \right) \right] \\ &= \sum_{t=1}^T \mathbb{E} [\xi_t^2 \mathbf{s}_t^\top \mathbf{s}_t] + \sum_{t,t'=1, t \neq t'}^T \mathbb{E} [\xi_t \mathbf{s}_t]^\top \mathbb{E} [\xi_{t'} \mathbf{s}_{t'}] \\ &= \sum_{t=1}^T \|\mathbf{s}_t\|^2 \mathbb{E} [\xi_t^2]. \end{aligned}$$

Since $\xi_t \sim \mathcal{N}(0, 1)$, we have $\xi_t^2 \sim \chi^2(1)$ and $\mathbb{E}[\xi_t^2] = 1$. Consequently,

$$\begin{aligned} \text{Var}[R(h)\mathbf{f}(h)] &\leq \frac{\beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2} \sum_{t=1}^T \|\mathbf{s}_t\|^2 \mathbb{E} [\xi_t^2] \\ &= \frac{\beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2} \sum_{t=1}^T \|\mathbf{s}_t\|^2 \\ &\leq \frac{D_T \beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2}, \end{aligned}$$

with probability at least $(1-\delta)^{1/2N}$.

The upper bound of $\text{Var}[R(h)g(h)]$:

$$\begin{aligned} \text{Var}[R(h)g(h)] &\leq \mathbb{E} [(Rg)^2] \\ &= \int_h p(h) \left(\sum_{t=1}^T \gamma^{t-1} r(\mathbf{s}_t, a_t, \mathbf{s}_{t+1}) \right)^2 \left(\sum_{t=1}^T \frac{(a_t - \boldsymbol{\mu}^\top \mathbf{s}_t)^2 - \sigma^2}{\sigma^3} \right)^2 dh \\ &\leq \frac{\beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2} \mathbb{E} \left[\left(\sum_{t=1}^T \left(\frac{a_t - \boldsymbol{\mu}^\top \mathbf{s}_t}{\sigma} \right)^2 - T \right)^2 \right]. \end{aligned}$$

Let $\xi_t = (a_t - \boldsymbol{\mu}^\top \mathbf{s}_t)/\sigma$ for $t = 1, \dots, T$. Then ξ_1, \dots, ξ_T are independent standard normal variables. Let $\kappa = \sum_{t=1}^T \xi_t^2$. Then we have $\kappa \sim \chi^2(T)$ and

$$\mathbb{E} [(\kappa - T)^2] = \mathbb{E} [\kappa^2] - 2T\mathbb{E}[\kappa] + T^2 = 2T.$$

Hence

$$\text{Var}[R(h)g(h)] \leq \frac{2T\beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2}.$$

The lower bound of $\text{Var}[R(h)f(h)]$: By the same technique used in the corresponding upper bound, we can prove that with probability at least $(1-\delta)^{1/2N}$,

$$\sum_{i=1}^{\ell} \mathbb{E} [(Rf_i)^2] \geq \frac{C_T \alpha^2 (1-\gamma^T)^2}{\sigma^2(1-\gamma)^2}.$$

On the other hand, based on the existence of $\{d_t\}_{t=1}^T$, there must be $\{d_{t,i}\}_{t=1}^T$ for $i = 1, \dots, \ell$, such that $d_t^2 = \sum_{i=1}^{\ell} d_{t,i}^2$ and the inequality $|s_{t,i}| \leq d_{t,i}$ holds with probability at least $(1-\delta)^{1/2N\ell}$. Let $\xi_{t,i} = \text{sgn}(s_{t,i})(a_t - \boldsymbol{\mu}^\top \mathbf{s}_t)/\sigma$ for $t = 1, \dots, T$ and $i = 1, \dots, \ell$. Then all $\xi_{t,i}$ are independent standard normal variables. Let $\kappa_i = \sum_{t=1}^T \xi_{t,i}^2$ and $\zeta_i = \sum_{t=1}^T \xi_{t,i} d_{t,i}$. Then $\kappa_i \sim \mathcal{N}(0, \sum_{t=1}^T s_{t,i}^2)$

for fixed $s_{1,i}, \dots, s_{T,i}, \zeta_i \sim \mathcal{N}(0, \sum_{t=1}^T d_{t,i}^2)$, and $\mathbb{E}[|\kappa_i| \mid s_{1,i}, \dots, s_{T,i}] \leq \mathbb{E}[|\zeta_i|]$ holds with probability at least $(1 - \delta)^{1/2N\ell}$ over the choice of $s_{1,i}, \dots, s_{T,i}$ according to the underlying $p(h)$. When $\int_h p(h) R f_i dh > 0$, with probability at least $(1 - \delta)^{1/2N\ell}$,

$$\begin{aligned}
\int_h p(h) R f_i dh &\leq \int_{\{h \mid f_i(h) > 0\}} p(h) R f_i dh \\
&\leq \frac{\beta(1 - \gamma^T)}{1 - \gamma} \int_{\{h \mid f_i(h) > 0\}} p(h) f_i dh \\
&= \frac{\beta(1 - \gamma^T)}{1 - \gamma} \int_{\{h \mid \sum_{t=1}^T \xi_{t,i} |s_{t,i}| > 0\}} p(h) \sum_{t=1}^T \xi_{t,i} |s_{t,i}| dh \\
&= \frac{\beta(1 - \gamma^T)}{1 - \gamma} \int_0^{+\infty} p(\kappa_i) \kappa_i d\kappa_i \\
&= \frac{\beta(1 - \gamma^T)}{1 - \gamma} \left(\frac{1}{2} \mathbb{E}[|\kappa_i|] \right) \\
&= \frac{\beta(1 - \gamma^T)}{1 - \gamma} \left(\frac{1}{2} \mathbb{E}_{s_{1,i}, \dots, s_{T,i}} \left[\mathbb{E}_{\kappa_i}[|\kappa_i| \mid s_{1,i}, \dots, s_{T,i}] \right] \right) \\
&\leq \frac{\beta(1 - \gamma^T)}{1 - \gamma} \left(\frac{1}{2} \mathbb{E}[|\zeta_i|] \right) \\
&= \frac{\beta(1 - \gamma^T)}{1 - \gamma} \frac{\sqrt{\sum_{t=1}^T d_{t,i}^2}}{\sqrt{2\pi}}.
\end{aligned}$$

When $\int_h p(h) R f_i dh < 0$, with probability at least $(1 - \delta)^{1/2N\ell}$,

$$\int_h p(h) R f_i dh \geq -\frac{\beta(1 - \gamma^T)}{1 - \gamma} \frac{\sqrt{\sum_{t=1}^T d_{t,i}^2}}{\sqrt{2\pi}}.$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{\ell} (\mathbb{E}[R f_i])^2 &= \sum_{i=1}^{\ell} \left(\int_h p(h) R f_i dh \right)^2 \\
&\leq \sum_{i=1}^{\ell} \frac{\beta^2(1 - \gamma^T)^2}{\sigma^2(1 - \gamma)^2} \frac{\sum_{t=1}^T d_{t,i}^2}{2\pi} \\
&= \frac{\beta^2(1 - \gamma^T)^2}{2\pi\sigma^2(1 - \gamma)^2} \sum_{t=1}^T \sum_{i=1}^{\ell} d_{t,i}^2 \\
&= \frac{\beta^2(1 - \gamma^T)^2}{2\pi\sigma^2(1 - \gamma)^2} \sum_{t=1}^T d_t^2 \\
&= \frac{D_T \beta^2(1 - \gamma^T)^2}{2\pi\sigma^2(1 - \gamma)^2},
\end{aligned}$$

with probability at least $(1 - \delta)^{1/2N}$.

Finally, with probability at least $(1 - \delta)^{1/N}$, we have

$$\begin{aligned}
\mathbf{Var}[R(h) \mathbf{f}(h)] &= \sum_{i=1}^{\ell} \mathbb{E}[(R f_i)^2] - (\mathbb{E}[R f_i])^2 \\
&\geq \frac{(1 - \gamma^T)^2}{\sigma^2(1 - \gamma)^2} \mathcal{L}(T).
\end{aligned}$$

□

C Proof of Theorem 3

Proof. According to Theorem 1 and Theorem 2, we could know that if there exists T_0 such that

$$\frac{(1 - \gamma^T)^2}{N\sigma^2(1 - \gamma)^2} \mathcal{L}(T_0) \geq \frac{\beta^2(1 - \gamma^T)^2 B}{N(1 - \gamma)^2},$$

we could get

$$\mathcal{L}(T_0) \geq \beta^2 B \sigma^2.$$

Under our assumption that $\mathcal{L}(T) > 0$ and $\mathcal{L}(T)$ is monotonically increasing with respect to T , we will have that whenever

$$\exists T_0, \mathcal{L}(T_0) \geq \beta^2 B \sigma^2,$$

there must be

$$\forall T > T_0, \mathbf{Var}[\nabla_{\mu} \hat{J}(\theta)] > \mathbf{Var}[\nabla_{\eta} \hat{J}(\rho)]. \quad \square$$

D Proof of Theorem 4

We denote $\mathbf{f}(\theta)$ and its i -th component $\mathbf{f}_i(\theta)$ as

$$\begin{aligned} \mathbf{f}(\theta) &= (\nabla_{\eta} \log p(\theta \mid \rho)^\top, \nabla_{\tau} \log p(\theta \mid \rho)^\top)^\top = \nabla_{\rho} \log p(\theta \mid \rho), \\ \mathbf{f}_i(\theta) &= (\nabla_{\eta_i} \log p(\theta \mid \rho), \nabla_{\tau_i} \log p(\theta \mid \rho))^\top = \nabla_{\rho_i} \log p(\theta \mid \rho). \end{aligned}$$

Note that we still have

$$\begin{aligned} \mathbf{Var}[\nabla_{\rho} \hat{J}^b(\rho)] &= \mathbf{Var}[\nabla_{\eta} \hat{J}^b(\rho)] + \mathbf{Var}[\nabla_{\tau} \hat{J}^b(\rho)] \\ &= \frac{1}{N} \mathbf{Var}[(R(h) - b) \nabla_{\eta} \log p(\theta \mid \rho)] + \frac{1}{N} \mathbf{Var}[(R(h) - b) \nabla_{\tau} \log p(\theta \mid \rho)] \\ &= \frac{1}{N} \mathbf{Var}[(R(h) - b) \mathbf{f}(\theta)]. \end{aligned}$$

Proof. According to Eq.(1), we have

$$\begin{aligned} \mathbf{Var}[(R(h) - b) \mathbf{f}(\theta)] &= \sum_{i=1}^{\ell} \mathbb{E}[(R - b)^2 \mathbf{f}_i^\top \mathbf{f}_i] - (\mathbb{E}[(R - b) \mathbf{f}_i])^\top (\mathbb{E}[(R - b) \mathbf{f}_i]) \\ &= \sum_{i=1}^{\ell} \mathbb{E}[R^2 \mathbf{f}_i^\top \mathbf{f}_i] - 2\mathbb{E}[Rb \mathbf{f}_i^\top \mathbf{f}_i] + \mathbb{E}[b^2 \mathbf{f}_i^\top \mathbf{f}_i] \\ &\quad - (\mathbb{E}[R \mathbf{f}_i] - \mathbb{E}[b \mathbf{f}_i])^\top (\mathbb{E}[R \mathbf{f}_i] - \mathbb{E}[b \mathbf{f}_i]). \end{aligned}$$

Noticing that

$$\begin{aligned} \mathbb{E}[b \mathbf{f}_i] &= \int p(\theta_i \mid \rho_i) b \nabla_{\rho_i} \log p(\theta_i \mid \rho_i) d\theta_i \\ &= \int b \nabla_{\rho_i} p(\theta_i \mid \rho_i) d\theta_i \\ &= b \nabla_{\rho_i} \int p(\theta_i \mid \rho_i) d\theta_i \\ &= b \nabla_{\rho_i} 1 \\ &= b (\nabla_{\eta_i} 1, \nabla_{\tau_i} 1)^\top \\ &= (0, 0)^\top, \end{aligned}$$

we have

$$\mathbf{Var}[(R(h) - b) \mathbf{f}(\theta)] = \mathbb{E}[R^2 \mathbf{f}^\top \mathbf{f}] - 2\mathbb{E}[Rb \mathbf{f}^\top \mathbf{f}] + \mathbb{E}[b^2 \mathbf{f}^\top \mathbf{f}] - \mathbb{E}[R \mathbf{f}]^\top \mathbb{E}[R \mathbf{f}].$$

The optimal baseline is obtained by minimizing the variance, so that differentiating it with respect to b and setting the result to zero will give us the optimal baseline for PGPE:

$$b_{\text{PGPE}}^* = \frac{\mathbb{E}[R\mathbf{f}^\top \mathbf{f}]}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]}.$$

Subsequently,

$$\begin{aligned} & \mathbf{Var}[(R - b)\mathbf{f}] - \mathbf{Var}[(R - b_{\text{PGPE}}^*)\mathbf{f}] \\ &= -2\mathbb{E}[Rb\mathbf{f}^\top \mathbf{f}] + \mathbb{E}[b^2\mathbf{f}^\top \mathbf{f}] + 2\mathbb{E}[Rb_{\text{PGPE}}^*\mathbf{f}^\top \mathbf{f}] - \mathbb{E}[b_{\text{PGPE}}^{*2}\mathbf{f}^\top \mathbf{f}] \\ &= -2\mathbb{E}[Rb\mathbf{f}^\top \mathbf{f}] + \mathbb{E}[b^2\mathbf{f}^\top \mathbf{f}] + 2\frac{\mathbb{E}[R\mathbf{f}^\top \mathbf{f}]}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]} \mathbb{E}[R\mathbf{f}^\top \mathbf{f}] - \left(\frac{\mathbb{E}[R\mathbf{f}^\top \mathbf{f}]}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]}\right)^2 \mathbb{E}[\mathbf{f}^\top \mathbf{f}] \\ &= b^2\mathbb{E}[\mathbf{f}^\top \mathbf{f}] - 2b\mathbb{E}[R\mathbf{f}^\top \mathbf{f}] + \frac{(\mathbb{E}[R\mathbf{f}^\top \mathbf{f}])^2}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]} \\ &= \left(b - \frac{\mathbb{E}[R\mathbf{f}^\top \mathbf{f}]}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]}\right)^2 \mathbb{E}[\mathbf{f}^\top \mathbf{f}] \\ &= (b - b_{\text{PGPE}}^*)^2 \mathbb{E}[\mathbf{f}^\top \mathbf{f}], \end{aligned}$$

which leads to

$$\begin{aligned} \mathbf{Var}[\nabla_{\rho} \hat{J}^b(\rho)] - \mathbf{Var}[\nabla_{\rho} \hat{J}^{b_{\text{PGPE}}^*}(\rho)] &= \frac{1}{N} \mathbf{Var}[(R - b)\mathbf{f}] - \frac{1}{N} \mathbf{Var}[(R - b_{\text{PGPE}}^*)\mathbf{f}] \\ &= \frac{(b - b_{\text{PGPE}}^*)^2}{N} \mathbb{E}[\mathbf{f}^\top \mathbf{f}]. \end{aligned} \quad \square$$

E Proof of Theorem 5

We denote the i -th component of $\mathbf{f}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})$ as

$$f_i(\boldsymbol{\theta}) = \nabla_{\eta_i} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}) = \frac{\theta_i - \eta_i}{\tau_i^2}.$$

Proof. By the same technique used in the proof of Theorem 4, we know, when the baseline $b = 0$,

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}} \hat{J}(\rho)] - \mathbf{Var}[\nabla_{\boldsymbol{\eta}} \hat{J}^{b_{\text{PGPE}}^*}(\rho)] = \frac{(\mathbb{E}[R\mathbf{f}^\top \mathbf{f}])^2}{N\mathbb{E}[\mathbf{f}^\top \mathbf{f}]}.$$

On one hand,

$$\begin{aligned} \mathbb{E}[\mathbf{f}^\top \mathbf{f}] &= \sum_{i=1}^{\ell} \mathbb{E}[f_i^2] \\ &= \sum_{i=1}^{\ell} \mathbb{E}\left[\left(\frac{\theta_i - \eta_i}{\tau_i^2}\right)^2\right] \\ &= \sum_{i=1}^{\ell} \frac{1}{\tau_i^2} \mathbb{E}\left[\left(\frac{\theta_i - \eta_i}{\tau_i}\right)^2\right]. \end{aligned}$$

Let $\psi_i = ((\theta_i - \eta_i)/\tau_i)^2$ for $i = 1, \dots, \ell$. We could know that $\psi_i \sim \chi^2(1)$ and $\mathbb{E}[\psi_i] = 1$ since $\theta_i \sim \mathcal{N}(\eta_i, \tau_i^2)$, and thus

$$\mathbb{E}[\mathbf{f}^\top \mathbf{f}] = \sum_{i=1}^{\ell} \frac{1}{\tau_i^2} = B.$$

On the other hand, when $\mathbb{E}[R\mathbf{f}^\top \mathbf{f}] > 0$, we have

$$\begin{aligned}\mathbb{E}[R\mathbf{f}^\top \mathbf{f}] &= \sum_{i=1}^{\ell} \int p(\theta_i) R \left(\frac{\theta_i - \eta_i}{\tau_i^2} \right)^2 d\theta_i \\ &\leq \sum_{i=1}^{\ell} \frac{\beta(1 - \gamma^T)}{\tau_i^2(1 - \gamma)} \int p(\theta_i) \left(\frac{\theta_i - \eta_i}{\tau_i} \right)^2 d\theta_i \\ &= \sum_{i=1}^{\ell} \frac{\beta(1 - \gamma^T)}{\tau_i^2(1 - \gamma)} \mathbb{E}[\psi_i] \\ &= \frac{\beta(1 - \gamma^T)B}{(1 - \gamma)},\end{aligned}$$

while $\mathbb{E}[R\mathbf{f}^\top \mathbf{f}] < 0$, we have

$$\mathbb{E}[R\mathbf{f}^\top \mathbf{f}] \geq -\frac{\beta(1 - \gamma^T)B}{(1 - \gamma)}.$$

Hence,

$$\frac{(\mathbb{E}[R\mathbf{f}^\top \mathbf{f}])^2}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]} \leq \frac{\beta^2(1 - \gamma^T)^2 B}{(1 - \gamma)^2}.$$

Similarly,

$$\frac{(\mathbb{E}[R\mathbf{f}^\top \mathbf{f}])^2}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]} \geq \frac{\alpha^2(1 - \gamma^T)^2 B}{(1 - \gamma)^2},$$

which completes the proof. \square

F Proof of Theorem 6

We denote $\mathbf{f}(h) = \sum_{t=1}^T \nabla_{\boldsymbol{\mu}} \log p(a_t \mid \mathbf{s}_t, \boldsymbol{\theta})$.

Proof. It is easy to prove that, when $b = 0$,

$$\mathbf{Var}[\nabla_{\boldsymbol{\mu}} \hat{J}(\boldsymbol{\theta})] - \mathbf{Var}[\nabla_{\boldsymbol{\mu}} \hat{J}_{\text{REINFORCE}}^{b*}(\boldsymbol{\theta})] = \frac{(\mathbb{E}[R\mathbf{f}^\top \mathbf{f}])^2}{N\mathbb{E}[\mathbf{f}^\top \mathbf{f}]}.$$

From the proof of Theorem 2, we could have

$$\mathbb{E}[\mathbf{f}^\top \mathbf{f}] = \frac{1}{\sigma^2} \sum_{t=1}^T \|\mathbf{s}_t\|^2.$$

On the other hand,

$$\begin{aligned}\mathbb{E}[R\mathbf{f}^\top \mathbf{f}] &= \int_h p(h) \left(\sum_{t=1}^T \gamma^{t-1} r(\mathbf{s}_t, a_t, \mathbf{s}_{t+1}) \right) \left(\sum_{t=1}^T \frac{a_t - \boldsymbol{\mu}^\top \mathbf{s}_t}{\sigma^2} \mathbf{s}_t \right)^\top \left(\sum_{t=1}^T \frac{a_t - \boldsymbol{\mu}^\top \mathbf{s}_t}{\sigma^2} \mathbf{s}_t \right) dh \\ &\leq \frac{\beta(1 - \gamma^T)}{\sigma^2(1 - \gamma)} \mathbb{E} \left[\left(\sum_{t, t'=1}^T \frac{(a_t - \boldsymbol{\mu}^\top \mathbf{s}_t)(a_{t'} - \boldsymbol{\mu}^\top \mathbf{s}_{t'})}{\sigma^2} \mathbf{s}_t^\top \mathbf{s}_{t'} \right) \right] \\ &= \frac{\beta(1 - \gamma^T)}{\sigma^2(1 - \gamma)} \sum_{t=1}^T \|\mathbf{s}_t\|^2.\end{aligned}$$

Similarly,

$$\mathbb{E}[R\mathbf{f}^\top \mathbf{f}] \geq \frac{\alpha(1 - \gamma^T)}{\sigma^2(1 - \gamma)} \sum_{t=1}^T \|\mathbf{s}_t\|^2.$$

Therefore,

$$\frac{\alpha^2(1-\gamma^T)^2 \sum_{t=1}^T \|\mathbf{s}_t\|^2}{\sigma^2(1-\gamma)^2} \leq \frac{(\mathbb{E}[R\mathbf{f}^\top \mathbf{f}])^2}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]} \leq \frac{\beta^2(1-\gamma^T)^2 \sum_{t=1}^T \|\mathbf{s}_t\|^2}{\sigma^2(1-\gamma)^2},$$

and subsequently, with probability at least $(1-\delta)^{1/N}$, we have

$$\frac{C_T \alpha^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2} \leq \frac{(\mathbb{E}[R\mathbf{f}^\top \mathbf{f}])^2}{\mathbb{E}[\mathbf{f}^\top \mathbf{f}]} \leq \frac{\beta^2(1-\gamma^T)^2 D_T}{\sigma^2(1-\gamma)^2}.$$

From this, the theorem follows. \square

G Proof of Theorem 7

Proof. According to Theorem 5, we know

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}} \hat{J}^{b_{\text{PGPE}}}(\boldsymbol{\rho})] \leq \mathbf{Var}[\nabla_{\boldsymbol{\eta}} \hat{J}(\boldsymbol{\rho})] - \frac{\alpha^2(1-\gamma^T)^2 B}{N(1-\gamma)^2}.$$

According to Theorem 1, we have

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}} \hat{J}(\boldsymbol{\rho})] \leq \frac{\beta^2(1-\gamma^T)^2 B}{N(1-\gamma)^2}.$$

Hence,

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}} \hat{J}^{b_{\text{PGPE}}}(\boldsymbol{\rho})] \leq \frac{(1-\gamma^T)^2}{N(1-\gamma)^2} (\beta^2 - \alpha^2) B.$$

According to Theorem 6, we know that

$$\mathbf{Var}[\nabla_{\boldsymbol{\mu}} \hat{J}^{b_{\text{REINFORCE}}}(\boldsymbol{\theta})] \leq \mathbf{Var}[\nabla_{\boldsymbol{\mu}} \hat{J}(\boldsymbol{\theta})] - \frac{C_T \alpha^2(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}$$

will hold with probability at least $(1-\delta)^{1/2}$. Furthermore, according to Theorem 2, we have the following upper bound with probability at least $(1-\delta)^{1/2}$:

$$\mathbf{Var}[\nabla_{\boldsymbol{\mu}} \hat{J}(\boldsymbol{\theta})] \leq \frac{D_T \beta^2(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}.$$

Eventually, we arrive at the upper bound for REINFORCE with the optimal baseline:

$$\mathbf{Var}[\nabla_{\boldsymbol{\mu}} \hat{J}^{b_{\text{REINFORCE}}}(\boldsymbol{\theta})] \leq \frac{(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2} (D_T \beta^2 - C_T \alpha^2),$$

with probability at least $1-\delta$. \square