7. Appendix

7.1. Why SVRPG does not work

Recall the importance weight from Section 5.2, which is defined in (Papini et al., 2018)

$$w(\theta^t, \tilde{\theta}; \tau) := \frac{p(\tau|\pi_{\tilde{\theta}})}{p(\tau|\pi_{\theta^t})} = \prod_{h=1}^{H} \frac{\pi_{\tilde{\theta}}(a_h|s_h)}{\pi_{\theta^t}(a_h|s_h)}, \tag{28}$$

and the SVRPG gradient estimator

$$\mathbf{g}_{vr}^t := \tilde{\mathbf{g}} + \mathbf{g}(\theta^t; \mathcal{M}) - \frac{1}{|\mathcal{M}|} \sum_{\tau \in \mathcal{M}} w(\theta^t, \tilde{\theta}; \tau) \mathbf{g}(\tilde{\theta}; \{\tau\}), \tag{29}$$

where $\tilde{\theta}$ and $\tilde{\mathbf{g}}$ are the reference point and its corresponding unbiased estimator respectively, and $\mathcal{M}$ is a mini-batch of trajectories sampled from $p(\cdot|\pi_{\theta^t})$.

While this importance sampling technique removes the bias, the variance of estimator (29) cannot be properly bounded since

$$\begin{aligned}
&\mathbb{E}_{\mathcal{M}} \|\mathbf{g}_{vr}^t - \nabla J(\theta^t)\|^2 \\
\leq &\frac{1}{|\mathcal{M}|} \mathbb{E}_{\tau} \|\mathbf{g}(\theta^t; \{\tau\}) - w(\theta^t, \tilde{\theta}; \tau) \mathbf{g}(\tilde{\theta}; \{\tau\})\|^2 \\
= &\frac{1}{|\mathcal{M}|} \int_{\tau} \frac{1}{p(\tau; \pi_{\theta^t})} \|p(\tau; \pi_{\theta^t}) \cdot \mathbf{g}(\theta^t; \{\tau\}) - p(\tau; \pi_{\tilde{\theta}}) \cdot \mathbf{g}(\tilde{\theta}; \{\tau\})\|^2 \mathbf{d}\tau,
\end{aligned}$$

and the term $\frac{1}{p(\tau;\pi_{\theta^t})}$ in the integral can be infinity large. The lack of proper variance control deprives SVRPG of its high sample-efficiency. Even under the strong assumption that the variance of the importance weight $w(\theta^t, \tilde{\theta}; \tau)$ is bounded (Assumption 4.3 in (Papini et al., 2018)), $\mathcal{O}(\frac{1}{\epsilon^4})$ random trajectories are still required by SVRPG to achieve an ϵ-FOSP (4) by scrutinizing the convergence result, which is the same as the original policy-gradient type method.

7.2. Derivation of Policy Gradient and Policy Hessian

Let $\tau = \{s_1, a_1, \ldots, s_H, a_H\}$ be a trajectory sampled according to $p(\tau; \pi_\theta)$ and define $\tau_h := \{s_1, a_1, \ldots, s_h, a_h\}$ for any $h \in [H]$. For simplicity of notation we will denote

$$\ell_\theta^{\tau_h} := \log p(\tau_h; \pi_\theta), \quad \bar{\mathcal{R}}_\gamma^{\tau_h} := \gamma^h \bar{\mathcal{R}}(a_h|s_h)$$

in the following discussion. From (3) and (2), we have

$$J(\theta) = \sum_{h=1}^{H} \mathbb{E}_{\tau \sim p(\tau;\pi_\theta)}[\bar{\mathcal{R}}_\gamma^{\tau_h}] = \sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim p(\tau_h;\pi_\theta)}[\bar{\mathcal{R}}_\gamma^{\tau_h}],$$

where we replace τ by τ_h since $\bar{\mathcal{R}}_\gamma^{\tau_h}$ is independent of the randomness after a_h. To compute the policy gradient

$$\nabla J(\theta) = \sum_{h=1}^{H} \int_{\tau_h} \bar{\mathcal{R}}_\gamma^{\tau_h} \nabla p(\tau_h; \pi_\theta) \mathbf{d}\tau_h = \sum_{h=1}^{H} \int_{\tau_h} \bar{\mathcal{R}}_\gamma^{\tau_h} p(\tau_h; \pi_\theta) \nabla \ell_\theta^{\tau_h} \mathbf{d}\tau_h,$$

where we use the log-trick in the second equation

$$\nabla p(\tau_h; \pi_\theta) = p(\tau_h; \pi_\theta) \nabla \log p(\tau_h; \pi_\theta) = p(\tau_h; \pi_\theta) \nabla \ell_\theta^{\tau_h}.$$

The policy gradient can be further simplified:

$$\begin{aligned}
\nabla J(\theta) &= \sum_{h=1}^{H} \int_{\tau_h} \bar{\mathcal{R}}_\gamma^{\tau_h} p(\tau_h;\pi_\theta) \nabla \ell_\theta^{\tau_h} \mathbf{d}\tau_h \\
&= \sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim p(\tau_h;\pi_\theta)} [\bar{\mathcal{R}}_\gamma^{\tau_h} \sum_{i=1}^{h} \nabla \log \pi_\theta(a_i|s_i)] \\
&= \sum_{h=1}^{H} \sum_{i=1}^{h} \mathbb{E}_{\tau_h \sim p(\tau_h;\pi_\theta)} [\bar{\mathcal{R}}_\gamma^{\tau_h} \nabla \log \pi_\theta(a_i|s_i)] \\
&= \sum_{h=1}^{H} \sum_{i=1}^{h} \mathbb{E}_{\tau \sim p(\tau;\pi_\theta)} [\bar{\mathcal{R}}_\gamma^{\tau_h} \nabla \log \pi_\theta(a_i|s_i)],
\end{aligned}$$

where in the last equality we use that $\bar{\mathcal{R}}_\gamma^{\tau_h} \nabla \log \pi_\theta(a_i|s_i)$ with $i \le h$ is independent of the randomness after a_h. Exchange the summation over i and h to obtain

$$\begin{aligned}
\nabla J(\theta) &= \sum_{i=1}^{H} \sum_{h=i}^{H} \mathbb{E}_{\tau \sim p(\tau;\pi_\theta)} [\bar{\mathcal{R}}_\gamma^{\tau_h} \nabla \log \pi_\theta(a_i|s_i)] \\
&= \sum_{i=1}^{H} \mathbb{E}_{\tau \sim p(\tau;\pi_\theta)} [\left(\sum_{h=i}^{H} \bar{\mathcal{R}}_\gamma^{\tau_h} \right) \nabla \log \pi_\theta(a_i|s_i)] \\
&= \sum_{i=1}^{H} \mathbb{E}_{\tau \sim p(\tau;\pi_\theta)} [\Psi_i(\tau) \nabla \log \pi_\theta(a_i|s_i)],
\end{aligned}$$

where $\Psi_i := \sum_{h=i}^{H} \gamma^h \bar{\mathcal{R}}(a_h|s_h)$ is the discounted reward after action a_i given state s_i. Let

$$\Phi(\theta;\tau) = \sum_{i=1}^{H} \Psi_i(\tau) \log p(a_i|s_i;\pi_\theta).$$

Using such notation, we have

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim p(\tau;\pi_\theta)} \nabla \Phi(\theta;\tau) = \int_\tau p(\tau;\pi_\theta) \nabla \Phi(\theta;\tau) \mathbf{d}\tau.$$

The second order derivative can be computed by

$$\begin{aligned}
\nabla^2 J(\theta) &= \int_\tau \nabla \Phi(\theta;\tau) \nabla p(\tau;\pi_\theta)^\top + p(\tau;\pi_\theta) \nabla^2 \Phi(\theta;\tau) \mathbf{d}\tau \\
&= \int_\tau p(\tau;\pi_\theta) \left[\nabla \Phi(\theta;\tau) \nabla \log p(\tau;\pi_\theta)^\top + \nabla^2 \Phi(\theta;\tau) \right] \mathbf{d}\tau \\
&= \mathbb{E}_{\tau \sim p(\tau;\pi_\theta)} \left[\nabla \Phi(\theta;\tau) \nabla \log p(\tau;\pi_\theta)^\top + \nabla^2 \Phi(\theta;\tau) \right].
\end{aligned}$$

7.3. Detail Hyper-parameter Settings

We present the Hyper-parameter settings in Table 1. The code for our experiments are available in https://github.com/m1zju/HAPG.

Table 1. Hyper-parameter Settings

	CartPole	Swimmer	Reacher	Walker2d	Humanoid	HumanoidStandup
Horizon	100	500	50	500	500	500
Baseline	No	Linear	Linear	Linear	Linear	Linear
Number of timesteps	$5 \cdot 10^5$	10^7	10^7	10^7	10^7	10^7
NN sizes	8	32x32	32x32	64x64	64x64	64x64
REINFORCE learning rate	0.01	0.01	0.01	0.01	0.01	0.01
REINFORCE batchsize	50	100	100	100	100	100
HAPG learning rate	0.01	0.01	0.01	0.01	0.01	0.01
HAPG $\lvert\mathcal{M}_0\rvert$	50	100	100	100	100	100
HAPG $\lvert\mathcal{M}\rvert$	10	10	10	10	10	10
HAPG p	5	10	10	10	10	10