Implicit Model-Based RL via HJB Bias

Yulun Zhuang
yulunz@umich.edu

November 25, 2023

Enhanced PPO

Value Function Objective

$$V(x_k) = \max_{u_k}\ r(x_k, u_k) + \gamma V(x_{k+1})$$

$$\begin{aligned}
V(x_k) &= \sum_{t=0}^{\infty} \gamma^t r(x_{k+t}, u_{k+t}) \\
&= r(x_k, u_k) + \gamma \sum_{t=0}^{\infty} \gamma^t r(x_{k+t+1}, u_{k+t+1}) \\
&= r(x_k, u_k) + \gamma V(x_{k+1}) \\
&= r(x_k, u_k) + \gamma V(F(x_k, u_k)) \Leftarrow F(x_k, u_k) \coloneqq x_{k+1}
\end{aligned}$$

Partial Derivatives of Value Objective

$$\begin{aligned}
\frac{\partial V(x_k)}{\partial x_k} &= r_x(x_k, u_k) + \gamma F_x^T(x_k, u_k) \frac{\partial V(x_{k+1})}{\partial x_{k+1}} \\
\frac{\partial V(x_k)}{\partial x_k} &= r_u(x_k, u_k) + \gamma F_u^T(x_k, u_k) \frac{\partial V(x_{k+1})}{\partial x_{k+1}} = \mathbf{0} \\
\lambda_k &= r_x(x_k, u_k) + \gamma F_x^T(x_k, u_k) \lambda_{k+1} \Leftarrow \lambda_k \coloneqq \frac{\partial V(x_k)}{\partial x_k} \\
\mathbf{0} &= r_u(x_k, u_k) + \gamma F_u^T(x_k, u_k) \lambda_{k+1}
\end{aligned}$$

Enhanced Value Loss

$$\begin{aligned}
L^V = &\ \| r(x_k, u_k) + \gamma V(x_{k+1}) - V(x_k) \|^2 \\
&+ \| r_x(x_k, u_k) + \gamma F_x^T(x_k, u_k) \lambda_{k+1} - \lambda_k \|^2 \\
&+ \| r_u(x_k, u_k) + \gamma F_u^T(x_k, u_k) \lambda_{k+1} \|^2
\end{aligned}$$

Combined PPO Objective

$$L_t^{CLIP+V+S}(\theta) = \hat{\mathbb{E}}_t \left[L_t^{CLIP}(\theta) - c_1 L_t^V(\theta) + c_2 S\left[\pi_\theta\right](s_t) \right]$$

Generalized Advantage Estimation for State Derivatives

$$\begin{aligned}
&\hat{A}_t = -V\left(s_t\right) + r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V\left(s_T\right) \\
&\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + \cdots + (\gamma\lambda)^{T-t+1} \delta_{T-1} \\
&\quad \text{where} \quad \delta_t = r_t + \gamma V\left(s_{t+1}\right) - V\left(s_t\right)
\end{aligned}$$

Explicit Dynamics Models

Pendulum

$$\mathbf{x} = [\cos\theta,\ \sin\theta,\ \dot{\theta}]^T$$
$$\dot{\mathbf{x}} = \begin{bmatrix} -\sin\theta \\ \cos\theta \\ \frac{3g}{2l}\sin\theta + \frac{3}{ml^2}u \end{bmatrix}$$

Equation of Motion

$$\begin{aligned} x_{k+1} &= x_k + \dot{x}_k \mathrm{d}t \\ &= F(x_k, u_k) \\ &= \begin{bmatrix} \cos\theta\cos(\dot{\theta}\mathrm{d}t) - \sin\theta\sin(\dot{\theta}\mathrm{d}t) \\ \sin\theta\cos(\dot{\theta}\mathrm{d}t) + \cos\theta\sin(\dot{\theta}\mathrm{d}t) \\ \dot{\theta} + \left(\frac{3g}{2l}\sin\theta + \frac{3}{ml^2}u\right)\mathrm{d}t \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \frac{\partial F}{\partial x_k} &= F_x(x_k, u_k) \\ &= \begin{bmatrix} \cos(\dot{\theta}\mathrm{d}t) & -\sin(\dot{\theta}\mathrm{d}t) & -\mathrm{d}t\cos\theta\sin(\dot{\theta}\mathrm{d}t) - \mathrm{d}t\sin\theta\cos(\dot{\theta}\mathrm{d}t) \\ \sin(\dot{\theta}\mathrm{d}t) & \cos(\dot{\theta}\mathrm{d}t) & -\mathrm{d}t\sin\theta\sin(\dot{\theta}\mathrm{d}t) + \mathrm{d}t\cos\theta\cos(\dot{\theta}\mathrm{d}t) \\ 0 & \frac{3g}{2l}\mathrm{d}t & 1 \end{bmatrix} \\ \frac{\partial F}{\partial u_k} &= F_u(x_k, u_k) \\ &= \begin{bmatrix} 0 \\ 0 \\ \frac{3\mathrm{d}t}{ml^2} \end{bmatrix} \end{aligned}$$

Quadratic Rewards

$$\begin{aligned} r_k &= -(\theta^2 + 0.1\dot{\theta}^2 + 0.001u^2) \\ \frac{\partial r_k}{\partial x_k} &= r_x = \begin{bmatrix} 2\theta\sin\theta \\ -2\theta\cos\theta \\ -0.2\dot{\theta} \end{bmatrix} \\ \frac{\partial r_k}{\partial u_k} &= r_u = \begin{bmatrix} -0.002u \end{bmatrix} \end{aligned}$$

Cart-Pole

$$\begin{aligned} \mathbf{x} &= [x, \dot{x}, \theta, \dot{\theta}]^T \\ \dot{\mathbf{x}} &= \begin{bmatrix} \dot{x} \\ \frac{u + m_p l\left(\dot{\theta}^2\sin\theta - \ddot{\theta}\cos\theta\right)}{m_c + m_p} \\ \dot{\theta} \\ \frac{g\sin\theta + \cos\theta\left(\frac{-u - m_p l\dot{\theta}^2\sin\theta}{m_c + m_p}\right)}{l\left(\frac{4}{3} - \frac{m_p\cos^2\theta}{m_c + m_p}\right)} \end{bmatrix} \\ \mathbf{s} &\coloneqq [x, \dot{x}, \theta, \dot{\theta}, u]^T \end{aligned}$$

Equation of Motion

$$
\begin{aligned}
x_{k+1} &= x_k + \dot{x}_k \mathrm{d}t \\
&= F(x_k, u_k) \\
&= \begin{bmatrix} x + \dot{x}\mathrm{d}t \\ \dot{x} + \frac{u + m_p l\left(\dot{\theta}^2 \sin\theta - \ddot{\theta}\cos\theta\right)}{m_c + m_p}\mathrm{d}t \\ \theta + \dot{\theta}\mathrm{d}t \\ \dot{\theta} + \frac{g\sin\theta + \cos\theta\left(\frac{-u - m_p l \dot{\theta}^2 \sin\theta}{m_c + m_p}\right)}{l\left(\frac{4}{3} - \frac{m_p \cos^2\theta}{m_c + m_p}\right)}\mathrm{d}t \end{bmatrix}
\end{aligned}
$$

$$
\frac{\partial F}{\partial s_k} = \begin{bmatrix} 1 & \mathrm{d}t & 0 & 0 & 0 \\ 0 & 1 & \frac{\partial F_2}{\partial \theta} & \mathrm{d}t\left(\frac{2\,l\,m_p\,\dot{\theta}\,\sin(\theta)}{m_c+m_p} - \frac{2\,l\,{m_p}^2\,\dot{\theta}\cos(\theta)^2\,\sin(\theta)}{\sigma_3}\right) & \mathrm{d}t\left(\frac{1}{m_c+m_p} - \frac{\sigma_6}{\sigma_3}\right) \\ 0 & 0 & 1 & \mathrm{d}t & 0 \\ 0 & 0 & -\frac{\mathrm{d}t\,\sigma_2}{l\,\sigma_5} - \frac{2\,\mathrm{d}t\,m_p\,\cos(\theta)\,\sin(\theta)\,\sigma_1}{l\,(m_c+m_p)\,{\sigma_5}^2} & \frac{2\,\mathrm{d}t\,m_p\,\dot{\theta}\,\cos(\theta)\,\sin(\theta)}{(m_c+m_p)\,\sigma_5} + 1 & \frac{\mathrm{d}t\,\cos(\theta)}{l\,(m_c+m_p)\,\sigma_5} \end{bmatrix}
$$

where

$$
\begin{aligned}
\frac{\partial F_2}{\partial \theta} &= \mathrm{d}t\left(\frac{m_p\,\cos(\theta)\,\sigma_2}{(m_c+m_p)\,\sigma_5} - \frac{m_p\,\sin(\theta)\,\sigma_1}{(m_c+m_p)\,\sigma_5} + \frac{l\,m_p\,\dot{\theta}^2\,\cos(\theta)}{m_c+m_p} + \frac{2\,{m_p}^2\cos(\theta)^2\,\sin(\theta)\,\sigma_1}{(m_c+m_p)^2\,{\sigma_5}^2}\right) \\
\sigma_1 &= g\,\sin(\theta) - \frac{\cos(\theta)\,\sigma_4}{m_c+m_p} \\
\sigma_2 &= g\,\cos(\theta) + \frac{\sin(\theta)\,\sigma_4}{m_c+m_p} - \frac{l\,m_p\,\dot{\theta}^2\cos(\theta)^2}{m_c+m_p} \\
\sigma_3 &= (m_c+m_p)^2\,\sigma_5 \\
\sigma_4 &= l\,m_p\,\sin(\theta)\,\dot{\theta}^2 + u \\
\sigma_5 &= \frac{\sigma_6}{m_c+m_p} - \frac{4}{3} \\
\sigma_6 &= m_p\cos(\theta)^2
\end{aligned}
$$

Quadratic Rewards

$$
\begin{aligned}
r_k &= -diag(1, 0.1, 1, 0.1)\ \mathbf{x}^2 + 0.001\mathbf{u}^2 \\
&= -(x^2 + 0.1\dot{x}^2 + \theta^2 + 0.1\dot{\theta}^2 + 0.001 * u^2) \\
\frac{\partial r_k}{\partial s_k} &= \begin{bmatrix} -2x \\ -0.2\dot{x} \\ -2\theta \\ -0.2\dot{\theta} \\ -0.002u \end{bmatrix}
\end{aligned}
$$