POLICY GRADIENTS Y MÉTODOS DE TIPO ACTOR-CRÍTICO

EL7021: Seminario de robótica y sistemas autónomos

Francisco Leiva[2] Javier Ruiz-del-Solar[1,2]

[1]Departamento de Ingeniería Eléctrica, Universidad de Chile

[2]Advanced Mining Technology Center (AMTC), Universidad de Chile

Abril, 2023

Objetivo del aprendizaje reforzado

- Recordemos el objetivo del aprendizaje reforzado:

$$\underbrace{p_\pi(s_1, a_1, ..., s_T, a_T)}_{p_\pi(\tau)} = p(s_1) \prod_{t=1}^{T} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

$$J_{\text{RL}}(\pi) = \mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t) \right]$$

Objetivo del aprendizaje reforzado

- Recordemos el objetivo del aprendizaje reforzado:

$$\underbrace{p_\pi(s_1, a_1, ..., s_T, a_T)}_{p_\pi(\tau)} = p(s_1) \prod_{t=1}^{T} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

$$J_{\text{RL}}(\pi) = \mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t) \right]$$

- Considerando una aproximación $\pi_\theta(a|s)$ para la política, el problema se convierte en encontrar los parámetros θ^* tal que:

$$\theta^* = \arg\max_\theta J_{\text{RL}}(\pi_\theta)$$

Recordatorio: Taxonomía de los algoritmos de RL

Model-based vs Model-free

Hacen o no uso de un modelo del ambiente.

Recordatorio: Taxonomía de los algoritmos de RL

Model-based vs Model-free

Hacen o no uso de un modelo del ambiente.

Value-based

Aproximan $V^*(s)$ o $Q^*(s, a)$ para derivar una política.

Recordatorio: Taxonomía de los algoritmos de RL

Model-based vs Model-free

Hacen o no uso de un modelo del ambiente.

Value-based

Aproximan $V^*(s)$ o $Q^*(s, a)$ para derivar una política.

Policy gradient

Buscan $\pi(a|s)$ a través de la optimización directa de $J_{\mathrm{RL}}(\pi)$.

Recordatorio: Taxonomía de los algoritmos de RL

Model-based vs Model-free

Hacen o no uso de un modelo del ambiente.

Value-based

Aproximan $V^*(s)$ o $Q^*(s,a)$ para derivar una política.

Policy gradient

Buscan $\pi(a|s)$ a través de la optimización directa de $J_{\mathrm{RL}}(\pi)$.

Actor-Critic

Aproximan conjuntamente $V^*(s)$ o $Q^*(s,a)$ y una política $\pi(a|s)$.

Policy Gradients

- Una forma de optimizar los parámetros θ cuando la política es explícitamente representada por $\pi_\theta(a|s)$, consiste en hacerlo mediante gradiente ascendente, según $\nabla J_{\mathrm{RL}}(\pi_\theta)$.

Policy Gradients

- Una forma de optimizar los parámetros θ cuando la política es explícitamente representada por $\pi_\theta(a|s)$, consiste en hacerlo mediante gradiente ascendente, según $\nabla J_{\text{RL}}(\pi_\theta)$.

- Denotando $R(\tau) = \sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)$:

Policy Gradients

- Una forma de optimizar los parámetros θ cuando la política es explícitamente representada por $\pi_\theta(a|s)$, consiste en hacerlo mediante gradiente ascendente, según $\nabla J_{\mathrm{RL}}(\pi_\theta)$.

- Denotando $R(\tau) = \sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)$:

$$J_{\mathrm{RL}}(\pi_\theta) = \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} \left[\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t) \right]$$

Policy Gradients

- Una forma de optimizar los parámetros θ cuando la política es explícitamente representada por $\pi_\theta(a|s)$, consiste en hacerlo mediante gradiente ascendente, según $\nabla J_{\text{RL}}(\pi_\theta)$.

- Denotando $R(\tau) = \sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)$:

$$\begin{aligned} J_{\text{RL}}(\pi_\theta) &= \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} \left[\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t) \right] \\ &= \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} \left[R(\tau) \right] \end{aligned}$$

Policy Gradients

- Una forma de optimizar los parámetros θ cuando la política es explícitamente representada por $\pi_\theta(a|s)$, consiste en hacerlo mediante gradiente ascendente, según $\nabla J_{\mathrm{RL}}(\pi_\theta)$.

- Denotando $R(\tau) = \sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)$:

$$
\begin{aligned}
J_{\mathrm{RL}}(\pi_\theta) &= \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} \left[\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t) \right] \\
&= \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} \left[R(\tau) \right] \\
&= \int p_{\pi_\theta}(\tau) R(\tau) \mathrm{d}\tau
\end{aligned}
$$

Policy Gradients

- Con esto, el gradiente de $J_{\text{RL}}(\pi_\theta)$ queda dado por:

$$\nabla_\theta J_{\text{RL}}(\pi_\theta) = \int \nabla_\theta p_{\pi_\theta}(\tau) R(\tau) \mathrm{d}\tau$$

Policy Gradients

- Con esto, el gradiente de $J_{\mathrm{RL}}(\pi_\theta)$ queda dado por:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) = \int \nabla_\theta p_{\pi_\theta}(\tau) R(\tau) \mathrm{d}\tau$$

- Notando la siguiente relación:

$$\begin{aligned} \nabla_\theta p_{\pi_\theta}(\tau) &= p_{\pi_\theta}(\tau) \frac{\nabla_\theta p_{\pi_\theta}(\tau)}{p_{\pi_\theta}(\tau)} \\ &= p_{\pi_\theta}(\tau) \nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) \end{aligned}$$

Policy Gradients

- Con esto, el gradiente de $J_{\mathrm{RL}}(\pi_\theta)$ queda dado por:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) = \int \nabla_\theta p_{\pi_\theta}(\tau) R(\tau) \mathrm{d}\tau$$

- Notando la siguiente relación:

$$\begin{aligned} \nabla_\theta p_{\pi_\theta}(\tau) &= p_{\pi_\theta}(\tau) \frac{\nabla_\theta p_{\pi_\theta}(\tau)}{p_{\pi_\theta}(\tau)} \\ &= p_{\pi_\theta}(\tau) \nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) \end{aligned}$$

- Entonces $\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)$ puede escribirse como sigue:

$$\begin{aligned} \nabla_\theta J_{\mathrm{RL}}(\pi_\theta) &= \int p_{\pi_\theta}(\tau) \nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) R(\tau) \mathrm{d}\tau \\ &= \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) R(\tau)\right] \end{aligned}$$

Policy Gradients

- Por la definición de $p_{\pi_\theta}(\tau)$:

$$\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) = \nabla_\theta \left[\log\left(p(s_1)\prod_{t=1}^{T}\pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t)\right)\right]$$

Policy Gradients

- Por la definición de $p_{\pi_\theta}(\tau)$:

$$\begin{aligned}\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) &= \nabla_\theta \left[\log\left(p(s_1)\prod_{t=1}^{T}\pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t)\right)\right]\\ &= \nabla_\theta \left[\log(p(s_1)) + \sum_{t=1}^{T}\left(\log\left(\pi_\theta(a_t|s_t)\right) + \log\left(p(s_{t+1}|s_t,a_t)\right)\right)\right]\end{aligned}$$

Policy Gradients

- Por la definición de $p_{\pi_\theta}(\tau)$:

$$\begin{aligned}
\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) &= \nabla_\theta \left[\log\left(p(s_1)\prod_{t=1}^{T}\pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t)\right)\right] \\
&= \nabla_\theta \left[\log(p(s_1)) + \sum_{t=1}^{T}\left(\log\left(\pi_\theta(a_t|s_t)\right) + \log\left(p(s_{t+1}|s_t,a_t)\right)\right)\right] \\
&= \sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta(a_t|s_t)\right)
\end{aligned}$$

Policy Gradients

- Por la definición de $p_{\pi_\theta}(\tau)$:

$$\begin{aligned}\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) &= \nabla_\theta \left[\log\left(p(s_1)\prod_{t=1}^{T}\pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t)\right)\right] \\ &= \nabla_\theta \left[\log(p(s_1)) + \sum_{t=1}^{T}\left(\log\left(\pi_\theta(a_t|s_t)\right) + \log\left(p(s_{t+1}|s_t,a_t)\right)\right)\right] \\ &= \sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta(a_t|s_t)\right)\end{aligned}$$

- Reemplazando en $\nabla_\theta J_{\text{RL}}(\pi_\theta)$:

$$\nabla_\theta J_{\text{RL}}(\pi_\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) R(\tau)\right]$$

Policy Gradients

- Por la definición de $p_{\pi_\theta}(\tau)$:

$$\begin{aligned}
\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) &= \nabla_\theta \left[\log\left(p(s_1)\prod_{t=1}^{T}\pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t)\right)\right] \\
&= \nabla_\theta \left[\log(p(s_1)) + \sum_{t=1}^{T}\left(\log\left(\pi_\theta(a_t|s_t)\right) + \log\left(p(s_{t+1}|s_t,a_t)\right)\right)\right] \\
&= \sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta(a_t|s_t)\right)
\end{aligned}$$

- Reemplazando en $\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)$:

$$\begin{aligned}
\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) &= \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) R(\tau)\right] \\
&= \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta(a_t|s_t)\right)\right)\left(\sum_{t=1}^{T}\gamma^{t-1} r(s_t,a_t)\right)\right]
\end{aligned}$$

Policy Gradients

- ¿Cómo aproximar este gradiente?

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) = \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)}\left[\left(\sum_{t=1}^{T} \nabla_\theta \log\left(\pi_\theta(a_t|s_t)\right)\right)\left(\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)\right)\right]$$

Policy Gradients

- ¿Cómo aproximar este gradiente?

$$
\begin{aligned}
\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) &= \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)}\left[\left(\sum_{t=1}^{T} \nabla_\theta \log\left(\pi_\theta(a_t|s_t)\right)\right)\left(\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)\right)\right] \\
&\approx \frac{1}{N}\sum_{k=1}^{N}\left[\left(\sum_{t=1}^{T} \nabla_\theta \log\left(\pi_\theta\left(a_t^{(k)}|s_t^{(k)}\right)\right)\right)\left(\sum_{t=1}^{T} \gamma^{t-1} r\left(s_t^{(k)}, a_t^{(k)}\right)\right)\right]
\end{aligned}
$$

REINFORCE

Algoritmo 1: REINFORCE

Inicializar $\pi_\theta(a|s)$ con parámetros θ
for *i=1, M* **do**
    for *k=1, N* **do**
        Obtener s_1
        for *t=1, T-1* **do**
            Ejecutar acción $a_t \sim \pi_\theta(a_t|s_t)$, observar r_t y s_{t+1}
            Guardar transición (s_t, a_t, r_t) en $\tau^{(k)}$
        end
    end
    Calcular $\nabla_\theta J_{\text{RL}} \approx \frac{1}{N} \sum_{k=1}^{N} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log \left(\pi_\theta \left(a_t^{(k)} | s_t^{(k)} \right) \right) \right) \left(\sum_{t=1}^{T} \gamma^{t-1} r \left(s_t^{(k)}, a_t^{(k)} \right) \right) \right]$
    Actualizar $\pi_\theta(a|s)$ con gradiente ascendente según $\nabla_\theta J_{\text{RL}}(\pi_\theta)$
end

Policy Gradients

- Uno de los problemas de policy gradient:
  - Alta varianza (en estimación del gradiente).

Policy Gradients

- Uno de los problemas de policy gradient:
  - Alta varianza (en estimación del gradiente).

- ¿Qué representa realmente esta expresión?

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) = \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log\left(\pi_\theta(a_t|s_t)\right) \right) R(\tau) \right]$$

Policy Gradients

Reward to go

- ¿Cómo abordar el problema de la alta varianza?
- La acción ejecutada en t no afecta la recompensa obtenida en t' si $t' < t$.

Policy Gradients

Reward to go

- ¿Cómo abordar el problema de la alta varianza?
- La acción ejecutada en t no afecta la recompensa obtenida en t' si $t' < t$.
- *Reward to go*:

$$R(\tau_t) = \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})$$

Policy Gradients
Reward to go

- ¿Cómo abordar el problema de la alta varianza?
- La acción ejecutada en t no afecta la recompensa obtenida en t' si $t' < t$.
- *Reward to go*:

$$R(\tau_t) = \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})$$

- Luego:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) \approx \frac{1}{N} \sum_{k=1}^{N} \left[\sum_{t=1}^{T} \nabla_\theta \log \left(\pi_\theta \left(a_t^{(k)} | s_t^{(k)} \right) \right) \left(\sum_{t'=t}^{T} \gamma^{t'-t} r \left(s_{t'}^{(k)}, a_{t'}^{(k)} \right) \right) \right]$$

Policy Gradients

Baselines

- ¿Otra posible mejora?

Policy Gradients

Baselines

- ¿Otra posible mejora?
- *Baselines*:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) \approx \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta\left(a_t^{(k)}|s_t^{(k)}\right)\right)\left(\sum_{t'=t}^{T}\gamma^{t'-t} r\left(s_{t'}^{(k)}, a_{t'}^{(k)}\right) - b\right)\right]$$

donde b podría ser, por ejemplo, igual a $\frac{1}{N}\sum_{k=1}^{N} R^{(k)}(\tau_t)$

Policy Gradients
Baselines

- ¿Otra posible mejora?
- *Baselines*:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) \approx \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta\left(a_t^{(k)}|s_t^{(k)}\right)\right)\left(\sum_{t'=t}^{T}\gamma^{t'-t} r\left(s_{t'}^{(k)}, a_{t'}^{(k)}\right) - b\right)\right]$$

donde b podría ser, por ejemplo, igual a $\frac{1}{N}\sum_{k=1}^{N} R^{(k)}(\tau_t)$

- ¿Por qué es posible emplear *baselines*?

Policy Gradient

Baselines

- Tenemos la siguiente expresión:

$$\nabla_\theta J_{\mathsf{RL}}(\pi_\theta) = \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t) - b\right)\right]$$

Policy Gradient

Baselines

- Tenemos la siguiente expresión:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t) - b\right)\right]$$

- Notemos que:

$$\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) b] = \int p_{\pi_\theta}\nabla_\theta \log(p_{\pi_\theta}(\tau))b\mathrm{d}\tau$$

Policy Gradient

Baselines

- Tenemos la siguiente expresión:

$$\nabla_\theta J_{\text{RL}}(\pi_\theta) = \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t) - b\right)\right]$$

- Notemos que:

$$\begin{aligned}
\mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)}[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right) b] &= \int p_{\pi_\theta} \nabla_\theta \log(p_{\pi_\theta}(\tau)) b \mathrm{d}\tau \\
&= b \nabla_\theta \int p_{\pi_\theta}(\tau) \mathrm{d}\tau
\end{aligned}$$

Policy Gradient

Baselines

- Tenemos la siguiente expresión:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) = \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} \left[\nabla_\theta \log \left(p_{\pi_\theta}(\tau) \right) \left(R(\tau_t) - b \right) \right]$$

- Notemos que:

$$\begin{aligned} \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} [\nabla_\theta \log \left(p_{\pi_\theta}(\tau) \right) b] &= \int p_{\pi_\theta} \nabla_\theta \log (p_{\pi_\theta}(\tau)) b \mathrm{d}\tau \\ &= b \nabla_\theta \int p_{\pi_\theta}(\tau) \mathrm{d}\tau \\ &= b \nabla_\theta 1 \end{aligned}$$

Policy Gradient

Baselines

- Tenemos la siguiente expresión:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) = \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)} \left[\nabla_\theta \log \left(p_{\pi_\theta}(\tau) \right) \left(R(\tau_t) - b \right) \right]$$

- Notemos que:

$$\begin{aligned}
\mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)}[\nabla_\theta \log \left(p_{\pi_\theta}(\tau) \right) b] &= \int p_{\pi_\theta} \nabla_\theta \log(p_{\pi_\theta}(\tau)) b \mathrm{d}\tau \\
&= b \nabla_\theta \int p_{\pi_\theta}(\tau) \mathrm{d}\tau \\
&= b \nabla_\theta 1 \\
&= 0
\end{aligned}$$

Policy Gradient

Baselines

- ¿Cuál es el mejor *baseline*?

$$\mathrm{Var}\left[\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)\right] = \mathrm{Var}\left[\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t) - b\right)\right]\right]$$

Policy Gradient
Baselines

- ¿Cuál es el mejor *baseline*?

$$
\begin{aligned}
\mathrm{Var}\left[\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)\right] &= \mathrm{Var}\left[\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right]\right] \\
&= \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\left(\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right)^2\right] \\
&- \left(\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right]\right)^2
\end{aligned}
$$

Policy Gradient

Baselines

- ¿Cuál es el mejor *baseline*?

$$
\begin{aligned}
\text{Var}\left[\nabla_\theta J_{\text{RL}}(\pi_\theta)\right] &= \text{Var}\left[\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right]\right] \\
&= \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\left(\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right)^2\right] \\
&- \left(\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right]\right)^2
\end{aligned}
$$

$$
\frac{\text{dVar}\left[\nabla_\theta J_{\text{RL}}(\pi_\theta)\right]}{\text{d}b} = \frac{\text{d}}{\text{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\left(\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right)^2\right]
$$

Policy Gradient

Baselines

- ¿Cuál es el mejor *baseline*?

$$\begin{aligned}
\mathrm{Var}\left[\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)\right] &= \mathrm{Var}\left[\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right]\right] \\
&= \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\left(\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right)^2\right] \\
&- \left(\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right]\right)^2
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}\mathrm{Var}\left[\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)\right]}{\mathrm{d}b} &= \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\left(\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)\left(R(\tau_t)-b\right)\right)^2\right] \\
&= \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t)^2\right] \\
&- 2\frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t) b\right] \\
&+ \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 b^2\right]
\end{aligned}$$

Policy Gradient

Baselines

- Luego:

$$\begin{aligned}\frac{\mathrm{dVar}\left[\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)\right]}{\mathrm{d}b} &= \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t)^2\right] \\ &- 2\frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t) b\right] \\ &+ \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 b^2\right]\end{aligned}$$

Policy Gradient

Baselines

- Luego:

$$
\begin{aligned}
\frac{\mathrm{d}\mathrm{Var}\left[\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)\right]}{\mathrm{d}b} &= \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t)^2\right] \\
&- 2\frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t)b\right] \\
&+ \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 b^2\right] \\
&= -2\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t)\right] \\
&+ 2b\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2\right]
\end{aligned}
$$

Policy Gradient

Baselines

- Luego:

$$\begin{aligned}\frac{\mathrm{dVar}\left[\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)\right]}{\mathrm{d}b} &= \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t)^2\right] \\ &- 2\frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t) b\right] \\ &+ \frac{\mathrm{d}}{\mathrm{d}b}\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 b^2\right] \\ &= -2\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t)\right] \\ &+ 2b\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2\right]\end{aligned}$$

- Igualando a cero:

$$b = \frac{\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2 R(\tau_t)\right]}{\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\nabla_\theta \log\left(p_{\pi_\theta}(\tau)\right)^2\right]}$$

Policy Gradients

- ¿Como hacer una versión *off-policy* de Policy Gradients?

Policy Gradients

- ¿Como hacer una versión *off-policy* de Policy Gradients?
  - *Importance Sampling*

Policy Gradients

- ¿Como hacer una versión *off-policy* de Policy Gradients?
  - *Importance Sampling*

$$\mathbb{E}_{x \sim p(x)}[f(s)] = \int p(x) f(x) \mathrm{d}x$$

Policy Gradients

- ¿Como hacer una versión *off-policy* de Policy Gradients?
  - *Importance Sampling*

$$
\begin{aligned}
\mathbb{E}_{x\sim p(x)}[f(s)] &= \int p(x)f(x)\mathrm{d}x \\
&= \int \frac{q(x)}{q(x)}p(x)f(x)\mathrm{d}x = \int q(x)\frac{p(x)}{q(x)}f(x)\mathrm{d}x
\end{aligned}
$$

Policy Gradients

- ¿Como hacer una versión *off-policy* de Policy Gradients?
  - *Importance Sampling*

$$
\begin{aligned}
\mathbb{E}_{x\sim p(x)}[f(s)] &= \int p(x)f(x)\mathrm{d}x \\
&= \int \frac{q(x)}{q(x)}p(x)f(x)\mathrm{d}x = \int q(x)\frac{p(x)}{q(x)}f(x)\mathrm{d}x \\
&= \mathbb{E}_{x\sim q(x)}\left[\frac{p(x)}{q(x)}f(x)\right]
\end{aligned}
$$

Policy Gradients

- ¿Como hacer una versión *off-policy* de Policy Gradients?
  - *Importance Sampling*

$$\begin{aligned}\mathbb{E}_{x\sim p(x)}[f(s)] &= \int p(x)f(x)\mathrm{d}x \\ &= \int \frac{q(x)}{q(x)}p(x)f(x)\mathrm{d}x = \int q(x)\frac{p(x)}{q(x)}f(x)\mathrm{d}x \\ &= \mathbb{E}_{x\sim q(x)}\left[\frac{p(x)}{q(x)}f(x)\right]\end{aligned}$$

- Luego:

Policy Gradients

- ¿Como hacer una versión *off-policy* de Policy Gradients?
  - *Importance Sampling*

$$
\begin{aligned}
\mathbb{E}_{x\sim p(x)}[f(s)] &= \int p(x)f(x)\mathrm{d}x \\
&= \int \frac{q(x)}{q(x)}p(x)f(x)\mathrm{d}x = \int q(x)\frac{p(x)}{q(x)}f(x)\mathrm{d}x \\
&= \mathbb{E}_{x\sim q(x)}\left[\frac{p(x)}{q(x)}f(x)\right]
\end{aligned}
$$

- Luego:

$$
J_{\mathrm{RL}}(\pi_\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[R(\tau)\right]
$$

Policy Gradients

- ¿Como hacer una versión *off-policy* de Policy Gradients?
  - *Importance Sampling*

$$
\begin{aligned}
\mathbb{E}_{x\sim p(x)}[f(s)] &= \int p(x)f(x)\mathrm{d}x \\
&= \int \frac{q(x)}{q(x)}p(x)f(x)\mathrm{d}x = \int q(x)\frac{p(x)}{q(x)}f(x)\mathrm{d}x \\
&= \mathbb{E}_{x\sim q(x)}\left[\frac{p(x)}{q(x)}f(x)\right]
\end{aligned}
$$

- Luego:

$$
\begin{aligned}
J_{\mathrm{RL}}(\pi_\theta) &= \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[R(\tau)\right] \\
&= \mathbb{E}_{\tau\sim \overline{p}(\tau)}\left[\frac{p_{\pi_\theta}(\tau)}{\overline{p}(\tau)}R(\tau)\right]
\end{aligned}
$$

Posibles mejoras para policy gradients

- Al incorporar el uso de *baseline*:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) \approx \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta\left(a_t^{(k)}|s_t^{(k)}\right)\right)\left(Q^\pi\left(s_t^{(k)}, a_t^{(k)}\right) - b\right)\right]$$

Posibles mejoras para policy gradients

- Al incorporar el uso de *baseline*:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) \approx \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta\left(a_t^{(k)}|s_t^{(k)}\right)\right)\left(Q^\pi\left(s_t^{(k)}, a_t^{(k)}\right) - b\right)\right]$$

- Pero, ¿qué *baseline* utilizar?

Posibles mejoras para policy gradients

- Al incorporar el uso de *baseline*:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) \approx \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta\left(a_t^{(k)}|s_t^{(k)}\right)\right)\left(Q^\pi\left(s_t^{(k)}, a_t^{(k)}\right) - b\right)\right]$$

- Pero, ¿qué *baseline* utilizar?
- Una opción: $b = \frac{1}{N}\sum_{k=1}^{N} Q^\pi\left(s_t^{(k)}, a_t^{(k)}\right)$

Posibles mejoras para policy gradients

- Al incorporar el uso de *baseline*:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) \approx \frac{1}{N} \sum_{k=1}^{N} \left[\sum_{t=1}^{T} \nabla_\theta \log \left(\pi_\theta \left(a_t^{(k)} | s_t^{(k)} \right) \right) \left(Q^\pi \left(s_t^{(k)}, a_t^{(k)} \right) - b \right) \right]$$

- Pero, ¿qué *baseline* utilizar?
- Una opción: $b = \frac{1}{N} \sum_{k=1}^{N} Q^\pi \left(s_t^{(k)}, a_t^{(k)} \right)$
- Recordemos la definición de la función de valor $V^\pi(s)$:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_{\tau_t \sim p_\pi(\tau_t)} \left[\sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'}) \middle| s_t = s \right] \\ &= \mathbb{E}_{a \sim \pi(a|s)} Q^\pi(s, a) \end{aligned}$$

Posibles mejoras para policy gradients

- Con esto:

$$\nabla_\theta J_{\mathrm{RL}}(\pi_\theta) \approx \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta\left(a_t^{(k)}|s_t^{(k)}\right)\right)\underbrace{\left(Q^\pi\left(s_t^{(k)},a_t^{(k)}\right)-V^\pi\left(s_t^{(k)}\right)\right)}_{A^\pi\left(s_t^{(k)},a_t^{(k)}\right)}\right]$$

- $A^\pi(s,a)$ es usualmente denominada “*advantage function*”.

Aprendizaje de la función de ventaja

- Recordemos la siguiente relación:

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} V^{\pi}(s')$$

Aprendizaje de la función de ventaja

- Recordemos la siguiente relación:

$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} V^\pi(s')$$

- Por lo tanto:

$$\begin{aligned} A^\pi(s,a) &= Q^\pi(s,a) - V^\pi(s) \\ &= r(s,a) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} V^\pi(s') - V^\pi(s) \\ &\approx r(s,a) + \gamma V^\pi(s') - V^\pi(s) \end{aligned}$$

Aprendizaje de la función de ventaja

- ¿Cómo aproximar $V^{\pi}(s)$?

$$V^{\pi}(s_t) \approx \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})$$

Aprendizaje de la función de ventaja

- ¿Cómo aproximar $V^{\pi}(s)$?

$$V^{\pi}(s_t) \approx \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})$$

- Considerando una aproximación de $V_{\phi}(s)$ para $V^{\pi}(s)$, es posible ajustar los parámetros ϕ como en un problema de regresión, al minimizar el costo:

$$L(\phi) = \frac{1}{TN} \sum_{k=1}^{N} \sum_{t=1}^{T} \left(V_{\phi}\left(s_t^{(k)}\right) - \sum_{t'=t}^{T} \gamma^{t'-t} r\left(s_{t'}^{(k)}, a_{t'}^{(k)}\right) \right)^2$$

Aprendizaje de la función de ventaja

- ¿Otra forma de aproximar $V^{\pi}(s)$?

$$V^{\pi}(s) \approx r(s, a) + \gamma V^{\pi}(s')$$

Aprendizaje de la función de ventaja

- ¿Otra forma de aproximar $V^{\pi}(s)$?

$$V^{\pi}(s) \approx r(s,a) + \gamma V^{\pi}(s')$$

- Con esto:

$$L(\phi) = \frac{1}{TN} \sum_{k=1}^{N} \sum_{t=1}^{T} \left(V_{\phi}\left(s_t^{(k)}\right) - r\left(s_t^{(k)}, a_r^{(k)}\right) - \gamma V_{\phi}\left(s_{t+1}^{(k)}\right) \right)^2$$

Ejemplo de algoritmo actor-crítico

Algoritmo 2: Algoritmo actor-crítico simple

Inicializar $\pi_\theta(a|s)$ con parámetros θ
Inicializar $V_\phi(s)$ con parámetros ϕ
for *i=1, M* **do**
  for *k=1, N* **do**
    Obtener s_1
    for *t=1, T-1* **do**
      Ejecutar acción $a_t \sim \pi_\theta(a_t|s_t)$, observar r_t y s_{t+1}
      Guardar transición (s_t, a_t, r_t) en $\tau^{(k)}$
    end
  end
  Ajustar ϕ minimizando $L(\phi) = \frac{1}{TN}\sum_{k=1}^{N}\sum_{t=1}^{T}\left(V_\phi\left(s_t^{(k)}\right) - \sum_{t'=t}^{T}\gamma^{t'-t} r\left(s_{t'}^{(k)}, a_{t'}^{(k)}\right)\right)^2$
  Calcular $\nabla_\theta J_{\mathrm{RL}} \approx \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log\left(\pi_\theta\left(a_t^{(k)}|s_t^{(k)}\right)\right) A^\pi\left(s_t^{(k)}, a_t^{(k)}\right)\right]$
  Actualizar $\pi_\theta(a|s)$ con gradiente ascendente según $\nabla_\theta J_{\mathrm{RL}}(\pi_\theta)$
end

Lecturas adicionales sugeridas

- Deterministic Policy Gradient (DPG)[1]
- Deep Deterministic Policy Gradient (DDPG)[2]
- Twin Delayed Deterministic Policy Gradient (TD3)[3]
- Soft Actor Critic (SAC)[4]

---

[1] David Silver et al. «Deterministic policy gradient algorithms». En: *International conference on machine learning*. PMLR. 2014, págs. 387-395.

[2] Timothy P Lillicrap et al. «Continuous control with deep reinforcement learning». En: *arXiv preprint arXiv:1509.02971* (2015).

[3] Scott Fujimoto, Herke Hoof y David Meger. «Addressing function approximation error in actor-critic methods». En: *International conference on machine learning*. PMLR. 2018, págs. 1587-1596.

[4] Tuomas Haarnoja et al. «Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor». En: *International conference on machine learning*. PMLR. 2018, págs. 1861-1870.

Referencias

- Fujimoto, Scott, Herke Hoof y David Meger. «Addressing function approximation error in actor-critic methods». En: *International conference on machine learning*. PMLR. 2018, págs. 1587-1596.
- Haarnoja, Tuomas et al. «Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor». En: *International conference on machine learning*. PMLR. 2018, págs. 1861-1870.
- Lillicrap, Timothy P et al. «Continuous control with deep reinforcement learning». En: *arXiv preprint arXiv:1509.02971* (2015).
- Silver, David et al. «Deterministic policy gradient algorithms». En: *International conference on machine learning*. PMLR. 2014, págs. 387-395.