# Métodos de gradiente de política (continuación)

Fernando Lozano

Universidad de los Andes

23 de mayo de 2023

# Advantage

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_\pi \left[ \psi_t \nabla_{\boldsymbol{\theta}} \ln \left( \pi(A_t \mid S_t, \boldsymbol{\theta}) \right) \right]$$

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_{\pi}\left[\psi_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)\right]$$

- $\psi_t = G_t \rightarrow$ varianza $\gg$

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_\pi \left[\psi_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)\right]$$

- $\psi_t = G_t \rightarrow$ varianza $\gg$
- $\psi_t = G_t - \hat{v}(S_t, \mathbf{w})$

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_\pi \left[\psi_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)\right]$$

- $\psi_t = G_t \rightarrow$ varianza $\gg$
- $\psi_t = G_t - \hat{v}(S_t, \mathbf{w}) \rightarrow$ disminuir varianza

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_{\pi}\left[\psi_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)\right]$$

- $\psi_t = G_t \rightarrow$ varianza $\gg$
- $\psi_t = G_t - \hat{v}(S_t, \mathbf{w}) \rightarrow$ disminuir varianza
- $\psi_t = \delta_t$

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_{\pi}\left[\psi_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)\right]$$

- $\psi_t = G_t \rightarrow$ varianza $\gg$
- $\psi_t = G_t - \hat{v}(S_t, \mathbf{w}) \rightarrow$ disminuir varianza
- $\psi_t = \delta_t \rightarrow$ disminuir varianza, aumenta sesgo

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_\pi \left[\psi_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)\right]$$

- $\psi_t = G_t \rightarrow$ varianza $\gg$
- $\psi_t = G_t - \hat{v}(S_t, \mathbf{w}) \rightarrow$ disminuir varianza
- $\psi_t = \delta_t \rightarrow$ disminuir varianza, aumenta sesgo
- Menor varianza posible:

$$a_\pi(S_t, A_t) = q_\pi(S_t, A_t) - v_\pi(S_t)$$

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_\pi \left[\psi_t \nabla_{\boldsymbol{\theta}} \ln\left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)\right]$$

- $\psi_t = G_t \rightarrow$ varianza $\gg$
- $\psi_t = G_t - \hat{v}(S_t, \mathbf{w}) \rightarrow$ disminuir varianza
- $\psi_t = \delta_t \rightarrow$ disminuir varianza, aumenta sesgo
- Menor varianza posible:

$$a_\pi(S_t, A_t) = q_\pi(S_t, A_t) - v_\pi(S_t)$$

  - Incrementa probabilidad de acciones mejor que el promedio.
  - Decrementa probabilidad de acciones peor que el promedio.

# Advantage

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_\pi \left[\psi_t \nabla_{\boldsymbol{\theta}} \ln \left(\pi(A_t \mid S_t, \boldsymbol{\theta})\right)\right]$$

- $\psi_t = G_t \to$ varianza $\gg$
- $\psi_t = G_t - \hat{v}(S_t, \mathbf{w}) \to$ disminuir varianza
- $\psi_t = \delta_t \to$ disminuir varianza, aumenta sesgo
- Menor varianza posible:

$$a_\pi(S_t, A_t) = q_\pi(S_t, A_t) - v_\pi(S_t)$$

  - Incrementa probabilidad de acciones mejor que el promedio.
  - Decrementa probabilidad de acciones peor que el promedio.
- Estimar $a_\pi(S_t, A_t)$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

**for** k=1,2,... **do**

Recolecte trayectorias $\mathcal{D}_k = \{\tau_k\}$ usando $\pi$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

**for** k=1,2,… **do**

Recolecte trayectorias $\mathcal{D}_k = \{\tau_k\}$ usando $\pi$

Calcule retornos $G_t$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

**for** k=1,2,... **do**

Recolecte trayectorias $\mathcal{D}_k = \{\tau_k\}$ usando $\pi$

Calcule retornos $G_t$

Estime ventajas $\hat{A}_t$ usando $\hat{v}(s, \mathbf{w})$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

**for** k=1,2,... **do**

Recolecte trayectorias $\mathcal{D}_k = \{\tau_k\}$ usando $\pi$

Calcule retornos $G_t$

Estime ventajas $\hat{A}_t$ usando $\hat{v}(s, \mathbf{w})$

Estime gradiente:

$$g_k = \frac{1}{|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \ln \pi(a \mid s, \boldsymbol{\theta}) \hat{A}_t$$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$
Inicialice $\boldsymbol{\theta}, \mathbf{w}$
**for** k=1,2,... **do**
Recolecte trayectorias $\mathcal{D}_k = \{\tau_k\}$ usando $\pi$
Calcule retornos $G_t$
Estime ventajas $\hat{A}_t$ usando $\hat{v}(s, \mathbf{w})$
Estime gradiente:

$$g_k = \frac{1}{|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \ln \pi(a \mid s, \boldsymbol{\theta}) \hat{A}_t$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_{\mathbf{w}} g_k$$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

**for** k=1,2,… **do**

Recolecte trayectorias $\mathcal{D}_k = \{\tau_k\}$ usando $\pi$

Calcule retornos $G_t$

Estime ventajas $\hat{A}_t$ usando $\hat{v}(s, \mathbf{w})$

Estime gradiente:

$$g_k = \frac{1}{|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \ln \pi(a \mid s, \boldsymbol{\theta}) \hat{A}_t$$

$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_{\mathbf{w}} g_k$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

# Gradiente de política "vainilla"

**Require:** $\pi(a \mid s, \boldsymbol{\theta})$, $\hat{v}(s, \mathbf{w})$ $\alpha_{\boldsymbol{\theta}}, \alpha_{\mathbf{w}} > 0$

Inicialice $\boldsymbol{\theta}, \mathbf{w}$

**for** k=1,2,... **do**

Recolecte trayectorias $\mathcal{D}_k = \{\tau_k\}$ usando $\pi$

Calcule retornos $G_t$

Estime ventajas $\hat{A}_t$ usando $\hat{v}(s, \mathbf{w})$

Estime gradiente:

$$g_k = \frac{1}{|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \ln \pi(a \mid s, \boldsymbol{\theta}) \hat{A}_t$$

$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_{\mathbf{w}} g_k$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathbf{w}} \delta \hat{v}(S, \mathbf{w})$

**end for**

# Trust Region policy Optimization (TRPO)

# Trust Region policy Optimization (TRPO)

- Criterio:

$$\eta(\pi) = \mathbb{E}_{\tilde{\pi}}\left[G_t \mid s_0 \sim \rho_0\right] \qquad G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

# Trust Region policy Optimization (TRPO)

- Criterio:

$$\eta(\pi) = \mathbb{E}_{\tilde{\pi}}\left[G_t \mid s_0 \sim \rho_0\right] \qquad G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- Comparando con política $\tilde{\pi}$:

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t a_\pi(S_t, A_t)\right]$$

# Trust Region policy Optimization (TRPO)

- Criterio:

$$\eta(\pi) = \mathbb{E}_{\tilde{\pi}}\left[G_t \mid s_0 \sim \rho_0\right] \qquad G_t = \sum_{k=0}^{\infty} {\color{red}\gamma}^k R_{t+k+1}$$

- Comparando con política $\tilde{\pi}$:

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t a_\pi(S_t, A_t)\right]$$

  - Ventaja de $A_t \sim \tilde{\pi}$ con respecto a $\mathbb{E}_\pi\left[v_\pi(S_t)\right]$

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t a_\pi(S_t, A_t)\right]$$

$$
\begin{aligned}
\eta(\tilde{\pi}) &= \eta(\pi) + \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t a_\pi(S_t, A_t)\right] \\
&= \eta(\pi) + \sum_{t=0}^{\infty} \sum_{s} \mathbf{P}\left\{S_t = s \mid \tilde{\pi}\right\} \sum_{a} \tilde{\pi}(a \mid S_t) \gamma^t a_\pi(S_t, A_t)
\end{aligned}
$$

$$
\begin{aligned}
\eta(\tilde{\pi}) &= \eta(\pi) + \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t a_\pi(S_t, A_t)\right] \\
&= \eta(\pi) + \sum_{t=0}^{\infty}\sum_{s}\mathbf{P}\left\{S_t = s \mid \tilde{\pi}\right\}\sum_{a}\tilde{\pi}(a \mid S_t)\gamma^t a_\pi(S_t, A_t) \\
&= \eta(\pi) + \sum_{s}\sum_{t=0}^{\infty}\gamma^t\mathbf{P}\left\{S_t = s \mid \tilde{\pi}\right\}\sum_{a}\tilde{\pi}(a \mid S_t) a_\pi(S_t, A_t)
\end{aligned}
$$

$$
\begin{aligned}
\eta(\tilde{\pi}) &= \eta(\pi) + \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t a_\pi(S_t, A_t)\right] \\
&= \eta(\pi) + \sum_{t=0}^{\infty} \sum_{s} \mathbf{P}\left\{S_t = s \mid \tilde{\pi}\right\} \sum_{a} \tilde{\pi}(a \mid S_t)\gamma^t a_\pi(S_t, A_t) \\
&= \eta(\pi) + \sum_{s} \sum_{t=0}^{\infty} \gamma^t \mathbf{P}\left\{S_t = s \mid \tilde{\pi}\right\} \sum_{a} \tilde{\pi}(a \mid S_t) a_\pi(S_t, A_t) \\
&= \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a \mid s) a_\pi(s, a)
\end{aligned}
$$

$$
\begin{aligned}
\eta(\tilde{\pi}) &= \eta(\pi) + \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t a_\pi(S_t, A_t)\right] \\
&= \eta(\pi) + \sum_{t=0}^{\infty} \sum_{s} \mathbf{P}\left\{S_t = s \mid \tilde{\pi}\right\} \sum_{a} \tilde{\pi}(a \mid S_t)\gamma^t a_\pi(S_t, A_t) \\
&= \eta(\pi) + \sum_{s} \sum_{t=0}^{\infty} \gamma^t \mathbf{P}\left\{S_t = s \mid \tilde{\pi}\right\} \sum_{a} \tilde{\pi}(a \mid S_t) a_\pi(S_t, A_t) \\
&= \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a \mid s) a_\pi(s, a)
\end{aligned}
$$

donde

$$
\rho_\pi(s) = \mathbf{P}\left\{S_0 = s\right\} + \gamma \mathbf{P}\left\{S_1 = s\right\} + \gamma^2 \mathbf{P}\left\{S_2 = s\right\} + \dots
$$

con $a \sim \pi$

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a \mid s) a_{\pi}(s, a)$$

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a \mid s) a_{\pi}(s, a)$$

- Si $\forall s \sum_a \tilde{\pi}(a \mid s) a_{\pi}(s, a) \geq 0 \Rightarrow \eta(\tilde{\pi}) \geq \eta(\pi)$

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a \mid s) a_{\pi}(s, a)$$

- Si $\forall s \sum_a \tilde{\pi}(a \mid s) a_{\pi}(s, a) \geq 0 \Rightarrow \eta(\tilde{\pi}) \geq \eta(\pi)$
- Caso tabular $\rightarrow$ Teorema de mejoramiento de política.

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a \mid s) a_\pi(s, a)$$

- Si $\forall s \sum_a \tilde{\pi}(a \mid s) a_\pi(s, a) \geq 0 \Rightarrow \eta(\tilde{\pi}) \geq \eta(\pi)$
- Caso tabular $\rightarrow$ Teorema de mejoramiento de política.
- Con aproximación $\exists s$ para los que $\sum_a \tilde{\pi}(a \mid s) a_\pi(s, a) < 0$

- Aproximación local (de primer orden) :

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\pi(s)} \sum_a \tilde{\pi}(a \mid s) a_\pi(s, a)$$

- Aproximación local (de primer orden) :

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\pi(s)} \sum_a \tilde{\pi}(a \mid s) a_\pi(s, a)$$

- Visitas de acuerdo a $\pi$.

- Aproximación local (de primer orden) :

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\pi(s)} \sum_a \tilde{\pi}(a \mid s) a_\pi(s, a)$$

- Visitas de acuerdo a $\pi$.
- Si $\pi_{\boldsymbol{\theta}} = \pi(a \mid s, \boldsymbol{\theta})$
  - $L_{\pi_{\boldsymbol{\theta}_0}}(\pi_{\boldsymbol{\theta}_0}) = \eta(\pi_{\boldsymbol{\theta}_0})$
  - $\nabla_{\boldsymbol{\theta}} L_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \nabla_{\boldsymbol{\theta}} \eta(\pi_{\boldsymbol{\theta}})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$

- Aproximación local (de primer orden) :

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\pi(s)} \sum_a \tilde{\pi}(a \mid s) a_\pi(s, a)$$

- Visitas de acuerdo a $\pi$.
- Si $\pi_{\boldsymbol{\theta}} = \pi(a \mid s, \boldsymbol{\theta})$
  - $L_{\pi_{\boldsymbol{\theta}_0}}(\pi_{\boldsymbol{\theta}_0}) = \eta(\pi_{\boldsymbol{\theta}_0})$
  - $\nabla_{\boldsymbol{\theta}} L_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \nabla_{\boldsymbol{\theta}} \eta(\pi_{\boldsymbol{\theta}})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$
- Aproximación válida si $\tilde{\pi}$ es cercana a $\pi$

# Distancia entre distribuciones

- Variación total:

$$D_{TV} = \frac{1}{2}\sum_i |p_i - q_i|$$

# Distancia entre distribuciones

- Variación total:

$$D_{TV} = \frac{1}{2} \sum_i |p_i - q_i|$$

- Entropía relativa:

$$D_{KL} = \sum_i p_i \log \frac{pi}{q_i}$$

# Distancia entre distribuciones

- Variación total:

$$D_{TV} = \frac{1}{2}\sum_i |p_i - q_i|$$

- Entropía relativa:

$$D_{KL} = \sum_i p_i \log \frac{pi}{q_i}$$

- Teoría:

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - 4\frac{\epsilon\gamma}{(1-\gamma^2)}\alpha^2$$

  - $\alpha = \text{máx}\, D_{TV}(\tilde{\pi}, \pi)$
  - $\epsilon = \text{máx}_{s,a} |a_\pi(s,a)|$

- Sugiere:

$$\begin{aligned} \max \quad & L_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{\theta}) \\ \text{sujeto a} \quad & \mathbb{E}_{\pi}\left[KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t))\right] \leq \delta \end{aligned}$$

- Sugiere:

$$\begin{aligned} \text{máx} \quad & L_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{\theta}) \\ \text{sujeto a} \quad & \mathbb{E}_{\pi}\left[KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t))\right] \leq \delta \end{aligned}$$

- Donde $KL(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log \frac{pi}{q_i}$.

- Sugiere:

$$\begin{array}{rl} \text{máx} & L_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{\theta}) \\ \text{sujeto a} & \mathbb{E}_{\pi}\left[KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t))\right] \leq \delta \end{array}$$

- Donde $KL(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log \frac{pi}{q_i}$.
- $\boldsymbol{\theta}$ no es muy lejana a $\boldsymbol{\theta}_{\text{old}}$

# En la práctica

$$
\begin{aligned}
\text{máx} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\
\text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta
\end{aligned}
$$

# En la práctica

$$\begin{aligned} \text{máx} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\ \text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta \end{aligned}$$

- $\hat{\mathbb{E}}_t [\ldots]$ promedio empírico sobre un batch de muestras.

# En la práctica

$$
\begin{aligned}
\text{máx} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\
\text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta
\end{aligned}
$$

- $\hat{\mathbb{E}}_t \left[ \dots \right]$ promedio empírico sobre un batch de muestras.
- $\hat{A}_t$ es estimativo empírico de la ventaja.

# En la práctica

$$\begin{aligned} \text{máx} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\ \text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta \end{aligned}$$

- $\hat{\mathbb{E}}_t [\dots]$ promedio empírico sobre un batch de muestras.
- $\hat{A}_t$ es estimativo empírico de la ventaja.
- Muestreo por importancia.

# En la práctica

$$
\begin{aligned}
\text{máx} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\
\text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta
\end{aligned}
$$

- $\hat{\mathbb{E}}_t [\ldots]$ promedio empírico sobre un batch de muestras.
- $\hat{A}_t$ es estimativo empírico de la ventaja.
- Muestreo por importancia.
- Alternar muestreo y optimización.

# Deep RL Asincrónico

- $DQN$ y variantes $\rightarrow$ Experience replay.

# Deep RL Asincrónico

- $DQN$ y variantes $\rightarrow$ Experience replay.
  - Datos no correlacionados.

# Deep RL Asincrónico

- $DQN$ y variantes $\rightarrow$ Experience replay.
  - Datos no correlacionados.
  - Costo computacional.

# Deep RL Asincrónico

- $DQN$ y variantes $\rightarrow$ Experience replay.
  - Datos no correlacionados.
  - Costo computacional.
  - Off-Policy.

# Deep RL Asincrónico

- $DQN$ y variantes $\rightarrow$ Experience replay.
  - Datos no correlacionados.
  - Costo computacional.
  - Off-Policy.
- RL Asíncrono:
  - Múltiples agentes asíncronos en paralelo.

# Deep RL Asincrónico

- $DQN$ y variantes $\rightarrow$ Experience replay.
  - Datos no correlacionados.
  - Costo computacional.
  - Off-Policy.
- RL Asíncrono:
  - Múltiples agentes asíncronos en paralelo.
  - Diferentes agentes experimentan diferentes estados $\rightarrow$ decorrelación de datos.

# Deep RL Asincrónico

- $DQN$ y variantes $\rightarrow$ Experience replay.
  - Datos no correlacionados.
  - Costo computacional.
  - Off-Policy.
- RL Asíncrono:
  - Múltiples agentes asíncronos en paralelo.
  - Diferentes agentes experimentan diferentes estados $\rightarrow$ decorrelación de datos.
  - Ejecución en CPU, mucho más rápida.

# Deep RL Asincrónico

- $DQN$ y variantes $\rightarrow$ Experience replay.
  - Datos no correlacionados.
  - Costo computacional.
  - Off-Policy.
- RL Asíncrono:
  - Múltiples agentes asíncronos en paralelo.
  - Diferentes agentes experimentan diferentes estados $\rightarrow$ decorrelación de datos.
  - Ejecución en CPU, mucho más rápida.
  - Algoritmos Off-Policy, On-policy, Policy Gradient.

# A3C: Asynchronous advantage actor critic

**Algorithm S3** Asynchronous advantage actor-critic - pseudocode for each actor-learner thread.

*// Assume global shared parameter vectors $\theta$ and $\theta_v$ and global shared counter $T = 0$*
*// Assume thread-specific parameter vectors $\theta'$ and $\theta'_v$*
Initialize thread step counter $t \leftarrow 1$
**repeat**
  Reset gradients: $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$.
  Synchronize thread-specific parameters $\theta' = \theta$ and $\theta'_v = \theta_v$
  $t_{start} = t$
  Get state $s_t$
  **repeat**
    Perform $a_t$ according to policy $\pi(a_t|s_t;\theta')$
    Receive reward $r_t$ and new state $s_{t+1}$
    $t \leftarrow t+1$
    $T \leftarrow T+1$
  **until** terminal $s_t$ **or** $t - t_{start} == t_{max}$
  $R = \begin{cases} 0 & \text{for terminal } s_t \\ V(s_t, \theta'_v) & \text{for non-terminal } s_t \text{// Bootstrap from last state} \end{cases}$
  **for** $i \in \{t-1, \ldots, t_{start}\}$ **do**
    $R \leftarrow r_i + \gamma R$
    Accumulate gradients wrt $\theta'$: $d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i|s_i;\theta')(R - V(s_i;\theta'_v))$
    Accumulate gradients wrt $\theta'_v$: $d\theta_v \leftarrow d\theta_v + \partial\left(R - V(s_i;\theta'_v)\right)^2 / \partial\theta'_v$
  **end for**
  Perform asynchronous update of $\theta$ using $d\theta$ and of $\theta_v$ using $d\theta_v$.
**until** $T > T_{max}$

- NN compartida para $\pi$ y $\hat{v}$.

# A3C: Asynchronous advantage actor critic

**Algorithm S3** Asynchronous advantage actor-critic - pseudocode for each actor-learner thread.

```
// Assume global shared parameter vectors θ and θ_v and global shared counter T = 0
// Assume thread-specific parameter vectors θ' and θ'_v
Initialize thread step counter t ← 1
repeat
    Reset gradients: dθ ← 0 and dθ_v ← 0.
    Synchronize thread-specific parameters θ' = θ and θ'_v = θ_v
    t_start = t
    Get state s_t
    repeat
        Perform a_t according to policy π(a_t|s_t; θ')
        Receive reward r_t and new state s_{t+1}
        t ← t + 1
        T ← T + 1
    until terminal s_t or t − t_start == t_max
    R = { 0            for terminal s_t
        { V(s_t, θ'_v)  for non-terminal s_t // Bootstrap from last state
    for i ∈ {t − 1, ..., t_start} do
        R ← r_i + γR
        Accumulate gradients wrt θ': dθ ← dθ + ∇_θ' log π(a_i|s_i; θ')(R − V(s_i; θ'_v))
        Accumulate gradients wrt θ'_v: dθ_v ← dθ_v + ∂(R − V(s_i; θ'_v))^2/∂θ'_v
    end for
    Perform asynchronous update of θ using dθ and of θ_v using dθ_v.
until T > T_max
```

- NN compartida para $\pi$ y $\hat{v}$.
- Añade entropía de $\pi$ a la función objetivo $\rightarrow$ hiperparámetro $\beta$.

# Proximal Policy Optimization (PPO)

- TRPO:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\ \text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta \end{aligned}$$

# Proximal Policy Optimization (PPO)

- TRPO:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\ \text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta \end{aligned}$$

- Versión con penalización:

$$\max_{\boldsymbol{\theta}} \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t - \beta KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right]$$

# Proximal Policy Optimization (PPO)

- TRPO:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\ \text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta \end{aligned}$$

- Versión con penalización:

$$\max_{\boldsymbol{\theta}} \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t - \beta KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right]$$

  - Cota inferior en mejora de la política.

# Proximal Policy Optimization (PPO)

- TRPO:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right] \\ \text{sujeto a} \quad & \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta \end{aligned}$$

- Versión con penalización:

$$\max_{\boldsymbol{\theta}} \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t - \beta KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right]$$

  - Cota inferior en mejora de la política.
  - En la práctica es difícil ajustar $\beta$

# Proximal Policy Optimization (PPO)

- TRPO:

$$\max_{\boldsymbol{\theta}} \quad \hat{\mathbb{E}}_t \left[ \frac{\pi_{\boldsymbol{\theta}}(a_t \mid s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t \right]$$

$$\text{sujeto a} \quad \hat{\mathbb{E}}_t \left[ KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right] \leq \delta$$

- Versión con penalización:

$$\max_{\boldsymbol{\theta}} \hat{\mathbb{E}}_t \left[ r_t(\boldsymbol{\theta}) \hat{A}_t - \beta KL(\pi_{\boldsymbol{\theta}_{\text{old}}}(. \mid s_t)), \pi_{\boldsymbol{\theta}}(. \mid s_t)) \right]$$

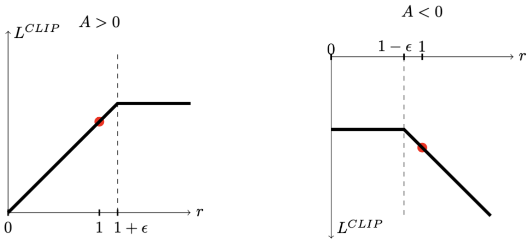
  - Cota inferior en mejora de la política.
  - En la práctica es difícil ajustar $\beta \longrightarrow \beta(t)$

- Objetivo recortado:

$$L^{\mathrm{CLIP}}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_t \left[ \min(r_t(\boldsymbol{\theta})\hat{A}_t, \mathrm{clip}(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$$

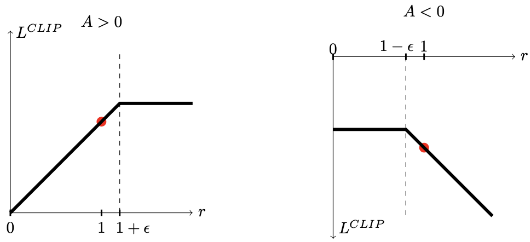
- Objetivo recortado:

$$L^{\text{CLIP}}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_t \left[ \text{mín}(r_t(\boldsymbol{\theta})\hat{A}_t, \text{clip}(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$$



- Objetivo recortado:

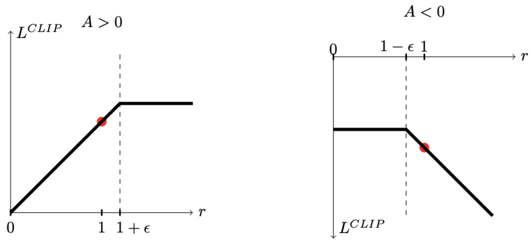
$$L^{\text{CLIP}}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_t \left[ \text{mín}(r_t(\boldsymbol{\theta})\hat{A}_t, \text{clip}(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$$




- Mínimo entre objetivo y objetivo recortado.

- Objetivo recortado:

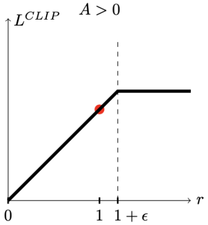
$$L^{\text{CLIP}}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_t \left[ \text{mín}(r_t(\boldsymbol{\theta})\hat{A}_t, \text{clip}(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$$

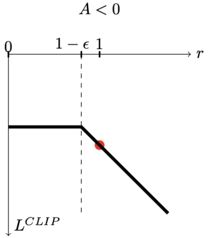



  - Mínimo entre objetivo y objetivo recortado.
  - Hiper parámetro $\epsilon$.

- Objetivo recortado:

$$L^{\text{CLIP}}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_t \left[ \text{mín}(r_t(\boldsymbol{\theta})\hat{A}_t, \text{clip}(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$$

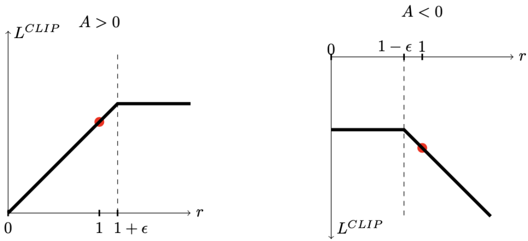



  - Mínimo entre objetivo y objetivo recortado.
  - Hiper parámetro $\epsilon$.
  - Cota inferior (pesimista) en el objetivo de TRPO.

- Objetivo recortado:

$$L^{\text{CLIP}}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_t \left[ \text{mín}(r_t(\boldsymbol{\theta})\hat{A}_t, \text{clip}(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$$




  - Mínimo entre objetivo y objetivo recortado.
  - Hiper parámetro $\epsilon$.
  - Cota inferior (pesimista) en el objetivo de TRPO.

- NN compartida:

$$L(\boldsymbol{\theta}) = \hat{\mathbb{E}}_t \left[ L^{\text{CLIP}}(\boldsymbol{\theta}) - c_1 (v_{\boldsymbol{\theta}}(S_t) - V_t^{\text{targ}})^2 + c_2 H(\pi_{\boldsymbol{\theta}})(S_t) \right]$$

# PPO

**for** iteración=1,2,. . . , **do**

# PPO

**for** iteración=1,2,. . . , **do**

  **for** actor=1,2,. . . ,N **do**

# PPO

**for** iteración=1,2,. . . , **do**

  **for** actor=1,2,. . . ,N **do**

    Corra política $\pi_{\boldsymbol{\theta}_{\text{old}}}$ por $T$ pasos

# PPO

**for** iteración=1,2,. . . , **do**

**for** actor=1,2,. . . ,N **do**

Corra política $\pi_{\boldsymbol{\theta}_{\text{old}}}$ por $T$ pasos

Calcule estimativos $\hat{A}_t$, $t = 1, \ldots, T$

# PPO

**for** iteración=1,2,. . . , **do**

  **for** actor=1,2,. . . ,N **do**

    Corra política $\pi_{\boldsymbol{\theta}_{\text{old}}}$ por $T$ pasos

    Calcule estimativos $\hat{A}_t$, $t = 1, \ldots, T$

  **end for**

  Optimice $L(\boldsymbol{\theta})$ con $K$ épocas y minibatch de tamaño $M \leq NT$

# PPO

**for** iteración=1,2,. . . , **do**
    **for** actor=1,2,. . . ,N **do**

        Corra política $\pi_{\boldsymbol{\theta}_{\text{old}}}$ por $T$ pasos
        Calcule estimativos $\hat{A}_t$, $t = 1, \ldots, T$
    **end for**
    Optimice $L(\boldsymbol{\theta})$ con $K$ épocas y minibatch de tamaño $M \leq NT$
    $\boldsymbol{\theta}_{\text{old}} = \boldsymbol{\theta}$
**end for**