$$s \in S, \quad a \in A, \quad \pi(a|s), \quad p(s'|s,a), \quad r(s,a)$$

$$\tau = \{ s_0, a_0, r_0, s_1, a_1, r_1, \ldots \}$$

$$p(\tau) = p(s_0)\pi(a_0|s_0)p(s_1|s_0,a_0)\pi(a_1|s_1) \ldots$$

$$R_0 = r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots \qquad \gamma \in (0,1)$$

$$\mathbb{E}_{p(\tau|\pi)} R_0 \to \max_{\pi}$$

$$V^{\pi}(s) = \mathbb{E}[R_0 \mid s_0 = s]$$

$$Q^{\pi}(s,a) = \mathbb{E}[R_0 \mid s_0 = s, a_0 = a]$$

$$V^{*}(s) = \max_{\pi} V^{\pi}(s)$$

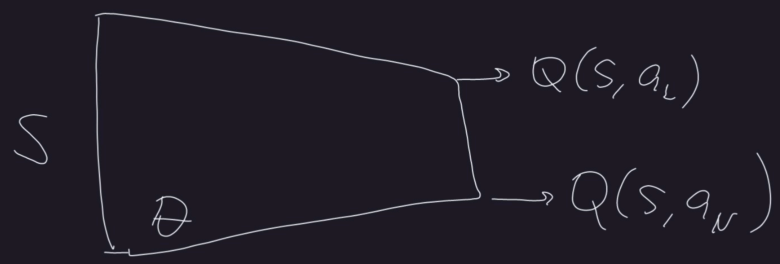
$$Q^{*}(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

$$\forall s,a \quad Q^{*}(s,a) = r(s,a) + \gamma \, \mathbb{E}_{p(s' \mid s,a)} \max_{a'} Q^{*}(s',a')$$

$$\pi(a \mid s) = \arg\max_{a} Q^{*}(s,a)$$

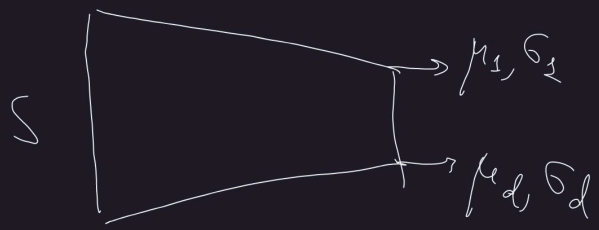
$$Q(s, a \mid \theta)$$

$$\text{Loss}(\theta) = \mathbb{E}_{s,a,s',r}\left(Q(s,a \mid \theta) - \left(r + \gamma \max_{a'} Q(s',a' \mid \bar{\theta})\right)\right)^2 \to \min_{\theta}$$



Policy Gradient RL

$$\pi(a \mid s, \theta) = \prod_{j=1}^{d} \mathcal{N}\left(a_j \mid \mu_j(s,\theta), \log\left(1+\exp\left(\sigma_j(s,\theta)\right)\right)\right)$$



$$\pi(a \mid s, \theta) = \mathrm{Softmax}(\mathrm{outputs}(s,\theta))$$

$$J(\theta) = \mathbb{E}_{p(\tau|\theta)} R(\tau) \to \max_{\theta}$$

Log-deriv. trick:

$$\nabla_\theta J(\theta) =$$

$$= \nabla_\theta \int p(\tau|\theta) R(\tau) d\tau =$$

$$= \int \nabla_\theta p(\tau|\theta) R(\tau) d\tau =$$

$\leftarrow \nabla_\theta \log p(\tau|\theta)$

$$= \int p(\tau|\theta) \frac{\nabla_\theta p(\tau|\theta)}{p(\tau|\theta)} R(\tau) d\tau =$$

$$= \mathbb{E}_{p(\tau|\theta)} \nabla_\theta \log p(\tau|\theta) \cdot R(\tau)$$

Reparameterization trick:

$$\nabla_\mu \mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)} f(x) =$$

$$= \nabla_\mu \mathbb{E}_{\mathcal{N}(\varepsilon|0,1)} f(\mu + \sigma\varepsilon) =$$

$$= \mathbb{E}_{\mathcal{N}(\varepsilon|0,1)} \nabla_\mu f(\mu + \sigma\varepsilon)$$

$$\tau = \{ s_0, a_0, r_0, \ldots \quad s_{T-1}, a_{T-1}, r_{T-1}, s_T \}$$

$$\mathbb{E}_{p(\tau|\theta)} \nabla_\theta \log p(\tau|\theta) \left[\sum_{t=0}^{T} \gamma^t \cdot r_t \right] =$$

$$= \left\{ p(\tau|\theta) = p(s_0)\, \pi(a_0|s_0, \theta)\, p(s_1|s_0, a_0)\, \pi(a_1|s_1, \theta) \ldots \right\} =$$

$$= \mathbb{E}_{p(\tau|\theta)} \left[\sum_{u=0}^{T} \nabla \log \pi(a_u|s_u, \theta) \right] \left[\sum_{t=0}^{T} \gamma^t r_t \right] =$$

$$= \sum_{u,t} \mathbb{E}_{p(\tau|\theta)} \nabla_\theta \log \pi(a_u|s_u, \theta) \cdot \gamma^t r_t \qquad u > t$$

$\underline{u > t}$

$$\mathbb{E}_{s_0, a_0, s_1, a_1, \dots} \nabla \log \pi(a_u | s_u, \theta) \gamma^t r_t =$$

$$= \mathbb{E}_{s_0, a_0, \dots, s_{u-1}, a_{u-1}, s_u} \gamma^t r_t \underbrace{\mathbb{E}_{\pi(a_u | s_u, \theta)} \nabla_\theta \log \pi(a_u | s_u, \theta)}_{\shortparallel}$$

$r(s, a)$

$$\sum_{a_u} \pi(a_u | s_u, \theta) \nabla \log \pi(a_u | s_u, \theta) =$$

$$= \sum_{a_u} \cancel{\pi(a_u | s_u, \theta)} \frac{\nabla_\theta \pi(a_u | s_u, \theta)}{\cancel{\pi(a_u | s_u, \theta)}} = \nabla_\theta \overbrace{\sum_{a_u} \pi(a_u | s_u, \theta)}^{1} =$$

$$= 0$$

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T} \mathbb{E}_{p(\tau|\theta)} \nabla_\theta \log \pi(a_t|s_t, \theta) \cdot \left(\sum_{u=t}^{T} \gamma^u \cdot r_u \right)$$

weighted ML maximization

REINFORCE algorithm

Init. θ

For episodes:

Init. s_0

For $t = 0, 1, \ldots, T$:

$a_t \sim \pi(a_t|s_t, \theta)$

get r_t, s_{t+1}

$R := 0,\ L_\theta := 0$

For $t = T, T-1, \ldots, 0$:

$R = r_t + \gamma R$

$L_\theta = L_\theta - \log \pi(a_t|s_t, \theta) \cdot \gamma^t \cdot R$

$\theta = \text{opt\_step}(\nabla_\theta L)$

Limitations:

- full episodes
- huge variance of SG

$$J(\theta) = \mathbb{E}_{p(\tau|\theta)} R(\tau) = \mathbb{E}_{p(s_0)} V^{\pi_\theta}(s_0)$$

$$\mu^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t p(s_0 \to s \mid t, \pi_\theta)$$

<u>Actor-Critic</u>

$\pi(a|s, \theta_\pi)$ - actor

$Q(s, a|\theta_Q)$ - critic

<u>Th</u> (PG theorem)

$$J(\theta) = \mathbb{E}_{p(s_0)} \mathbb{E}_{\pi(a_0|s_0,\theta)} Q^{\pi_\theta}(s_0, a_0)$$

$$\Rightarrow \nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{p(s_0)} \mathbb{E}_{s\sim\mu^\pi(s)} \mathbb{E}_{a\sim\pi(a|s,\theta)} \nabla_\theta \log \pi(a|s, \theta) Q^{\pi_\theta}(s, a)$$

$$\mathbb{E}_{\pi(a|s,\theta)} \nabla_\theta \log \pi(a|s,\theta) \left(Q(s,a) - \overset{\text{baseline}}{B(s)} \right)$$

$$\mathbb{E}_{\pi(a|s,\theta)} \nabla_\theta \log \pi(a|s,\theta) \, B(s) = B(s) \cdot 0$$

Baseline choice $B(s) = \mathbb{E}_{\pi(a|s,\theta)} Q^{\pi}(s,a) = V^{\pi}(s)$

Example $\mathbb{E}_a f_1(a) \cdot f_2(a)$

$$a: \begin{array}{cc} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{array} \qquad f_1 = \begin{cases} 1, & \text{if } a = 0 \\ -1, & \text{if } a = 1 \end{cases} \qquad f_2 = \begin{cases} 100, & \text{if } a = 0 \\ 101, & \text{if } a = 1 \end{cases}$$

$$\mathbb{E}_a f_1 = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (-1) = 0$$

$$\mathbb{E}_a f_1 \cdot f_2 = \frac{1}{2}\left(1 \cdot 100\right) + \frac{1}{2}\left((-1) \cdot 101\right) = -\frac{1}{2}$$

$$\mathbb{E}_a\left(f_1 \cdot f_2 - \mathbb{E} f_1 f_2\right)^2 = \mathbb{E}_a\left(f_1 \cdot f_2 + \frac{1}{2}\right)^2 =$$

$$= \frac{1}{2}\left(1 \cdot 100 + \frac{1}{2}\right)^2 + \frac{1}{2}\left((-1) \cdot 101 + \frac{1}{2}\right)^2 = (100.5)^2$$

$$\mathbb{E}_a f_1\left(f_2 - \mathbb{E} f_2\right) = \mathbb{E} f_1 \cdot f_2$$

$$\mathbb{E} f_2 = \frac{1}{2} \cdot 100 + \frac{1}{2} \cdot 101 = 100.5$$

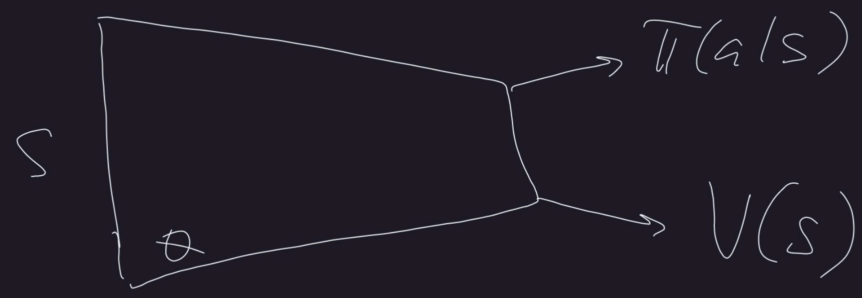
$$\mathrm{Var} = \mathbb{E}_a\left(f_1(f_2 - 100.5) + \frac{1}{2}\right)^2 =$$

$$= \frac{1}{2}\left(1 \cdot (100 - 100.5) + \frac{1}{2}\right)^2 +$$

$$+ \frac{1}{2}\left((-1)(101 - 100.5) + \frac{1}{2}\right)^2 = 0$$

$$A^{\pi}(s,a) := Q^{\pi}(s,a) - V^{\pi}(s)$$

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



A2C (Advantage Actor Critic)

Init. θ

For episodes:

$k = 0$

while $k < T$:

for $t = k, k+1, \ldots, k+n$:

$a_t \sim \pi(a_t | s_t, \theta)$

get r_t, s_{t+1}

$R = V(s_{k+n} | \bar{\theta})$, $L_{\theta_\pi} = 0$, $L_{\theta_v} = 0$

for $t = k+n-1, \ldots, k$:

$R = r_t + \gamma \cdot R$

$L_{\theta_v} = L_{\theta_v} + \frac{1}{2}(V(s_t | \theta) - R)^2$

$L_{\theta_\pi} = L_{\theta_\pi} - \log \pi(a_t | s_t, \theta)(R - V(s_t | \bar{\theta}))$

$\theta_v = \text{opt\_step}(\nabla L_{\theta_v})$

$\theta_\pi = \text{opt\_step}(\nabla L_{\theta_\pi})$

use parallel agent

$\pi(a | s, \theta)$ - learned policy

$b(s)$ - behaviour policy

If $b(s) = \pi(a | s, \theta)$

=> on-policy RL

Otherwise off-policy RL

<u>RLHF</u> (RL from Human Feedback)

$\tau_1 = \{ s_0^1, a_0^1, s_1^1, a_1^1, \ldots, s_n^1, a_n^1 \}$

$\tau_2 = \{ s_0^2, a_0^2, s_1^2, a_1^2, \ldots, s_n^2, a_n^2 \}$

$$\tau_1 \cup \tau_2 \rightarrow \begin{array}{ll} \mu = (1, 0) & \text{if } \tau_1 > \tau_2 \\ \mu = (0, 1) & \text{if } \tau_1 < \tau_2 \\ \mu = (\frac{1}{2}, \frac{1}{2}) & \text{if } \tau_1 \approx \tau_2 \end{array}$$

$r_\theta(s,a)$

$$p(\hat{\tau}_1 > \hat{\tau}_2 \mid \theta) = 0.9\,\frac{\exp\left(\sum_{s,a\in\hat{\tau}_1} r_\theta(s,a)\right)}{\exp\left(\sum_{s,a\in\hat{\tau}_1} r_\theta(s,a)\right) + \exp\left(\sum_{s,a\in\hat{\tau}_2} r_\theta(s,a)\right)} + 0.1\cdot\frac{1}{2} =$$

$$= \sigma\left(\sum_{s,a\in\hat{\tau}_1} r_\theta(s,a) - \sum_{s,a\in\hat{\tau}_2} r_\theta(s,a)\right)$$

$$L(\theta) = -\mathbb{E}_{(\hat{\tau}_1,\hat{\tau}_2,\mu)}\left(\mu(1)\log p(\hat{\tau}_1>\hat{\tau}_2\mid\theta) + \mu(2)\log p(\hat{\tau}_2>\hat{\tau}_1\mid\theta)\right) \to \min_\theta$$