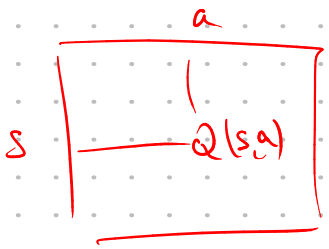
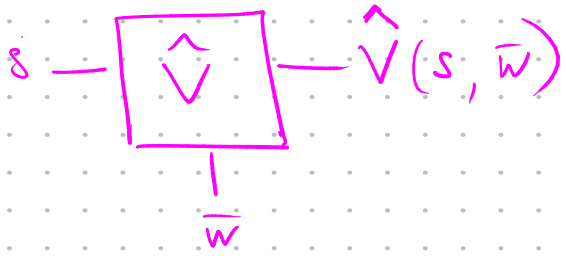
$$Q(s,a) = Q(s,a) + \alpha\left(r + \max_{a'} Q(s',a') - Q(s,a)\right)$$



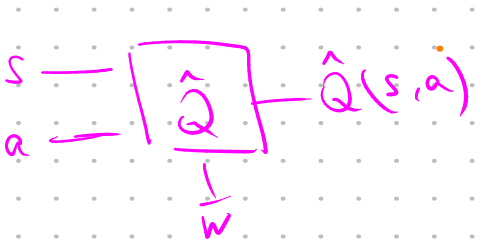
tabular RL

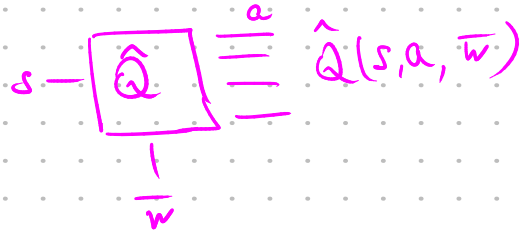
$V(s)$



Approximate RL

$Q(s,a)$





$$\overline{VE}(\bar{w}) = \sum_s \mu(s)\left(V_\pi(s) - \hat{V}_\pi(s,\bar{w})\right)^2 \longrightarrow \min_{\bar{w}}$$

$\mu(s)$: $Pr[s|b]$ ↑ Behaviour

SGD: $S_1, S_2, \ldots, S_t, \ldots$ $\quad V_\pi(S_t)$

Gradient Approx. RL

$$\bar{w}_{t+1} = \bar{w}_t + \alpha\left(V_\pi(S_t) - \hat{V}_\pi(S_t,\bar{w}_t)\right)\cdot \nabla_{\bar{w}}\hat{V}(S_t,\bar{w}_t)$$

$$\bar{w}_{t+1} = \bar{w}_t + \alpha\left(V'_t - \hat{V}_\pi(S_t,\bar{w}_t)\right)\cdot \nabla_{\bar{w}}\hat{V}(S_t,\bar{w}_t)$$

1) Gradient MC estimation $\quad V(S_t) \approx V'_t = G_t$

- $S_0, A_0, R_1, S_1, \ldots, R_T, S_T$
- $t = T-1, \ldots, 0$:

$$\bar{w} := \bar{w} + \alpha\left(G_t - \hat{V}(S_t,\bar{w})\right)\nabla_{\bar{w}}\hat{V}(S_t,\bar{w})$$

2) Semi-gradient TD(0) estimation $\quad V(S_t) \approx V'_t = R_{t+1} + \gamma V(S_{t+1})$

- (S,A,S',r):

$$\bar{w} := \bar{w} + \alpha\left(r + \gamma\hat{V}(s',\bar{w}) - \hat{V}(s,\bar{w})\right)\nabla_{\bar{w}}\hat{V}(s,\bar{w})$$

3) Semi-gradient TD control

$$\bar{w} := \bar{w} + \alpha\left(Q' - \hat{Q}(s,a,\bar{w})\right)\nabla_{\bar{w}}\hat{Q}(s,a,\bar{w})$$

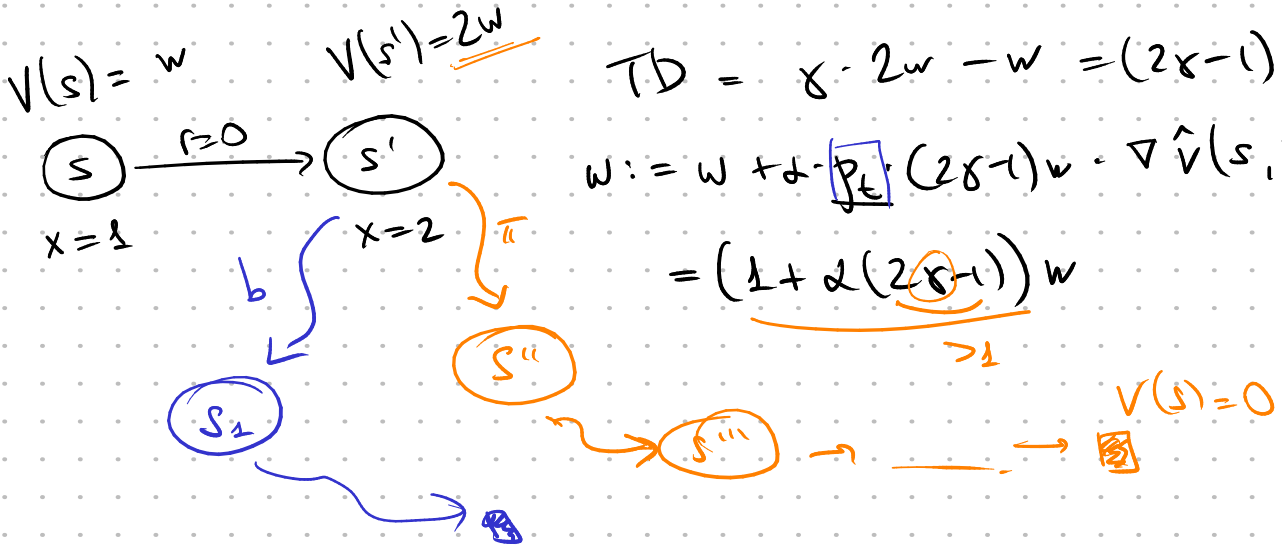
Sarsa: $\bar{w} := \bar{w} + \alpha\left(r + \gamma\hat{Q}(s',a',\bar{w}) - \hat{Q}(s,a,\bar{w})\right)\nabla_{\bar{w}}\hat{Q}(s,a,\bar{w})$

n-step: $R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{k-1}R_{t+k} + \gamma^{k}\hat{Q}(S_{t+k+1}, A_{t+k+1}, \bar{w})$

off-policy $\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$ importance weights

TD(0): $\bar{w} := \bar{w} + \alpha\rho_t\left(r + \gamma\hat{V}(s',\bar{w}) - \hat{V}(s,\bar{w})\right)\nabla_{\bar{w}}\hat{V}(s,\bar{w})$

Expected Sarsa $\bar{w} := \bar{w} + \alpha\left(r + \gamma\cdot\sum_{a'}\pi(a'|s')\hat{Q}(s',a',\bar{w}) - \hat{Q}(s,a,\bar{w})\right)\nabla_{\bar{w}}\hat{Q}$

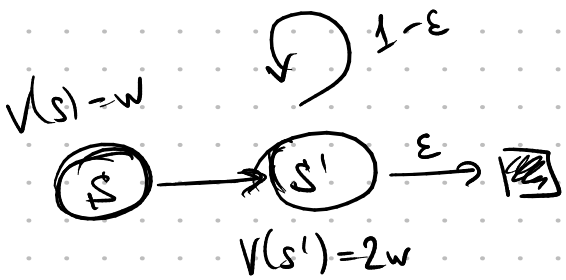


$$TD = \gamma\cdot 2w - w = (2\gamma - 1)w$$

$$w := w + \alpha\cdot\rho_t\cdot(2\gamma-1)w\cdot\nabla\hat{v}(s,\bar{w}) =$$

$$= \left(1 + \alpha(2\gamma - 1)\right)w$$

>1



$$w_{k+1} = \operatorname{argmin}_{\bar{w}}\sum_{s}\left(\hat{v}(s,\bar{w}) - \mathbb{E}[r + \gamma\hat{V}(s',\bar{w})]\right)^2$$

$$= \operatorname{argmin}_{\bar{w}}\left[\left(w - \gamma\cdot 2w_k\right)^2 + \left(2w - (1-\varepsilon)\cdot\gamma\cdot 2w_k\right)^2\right]$$

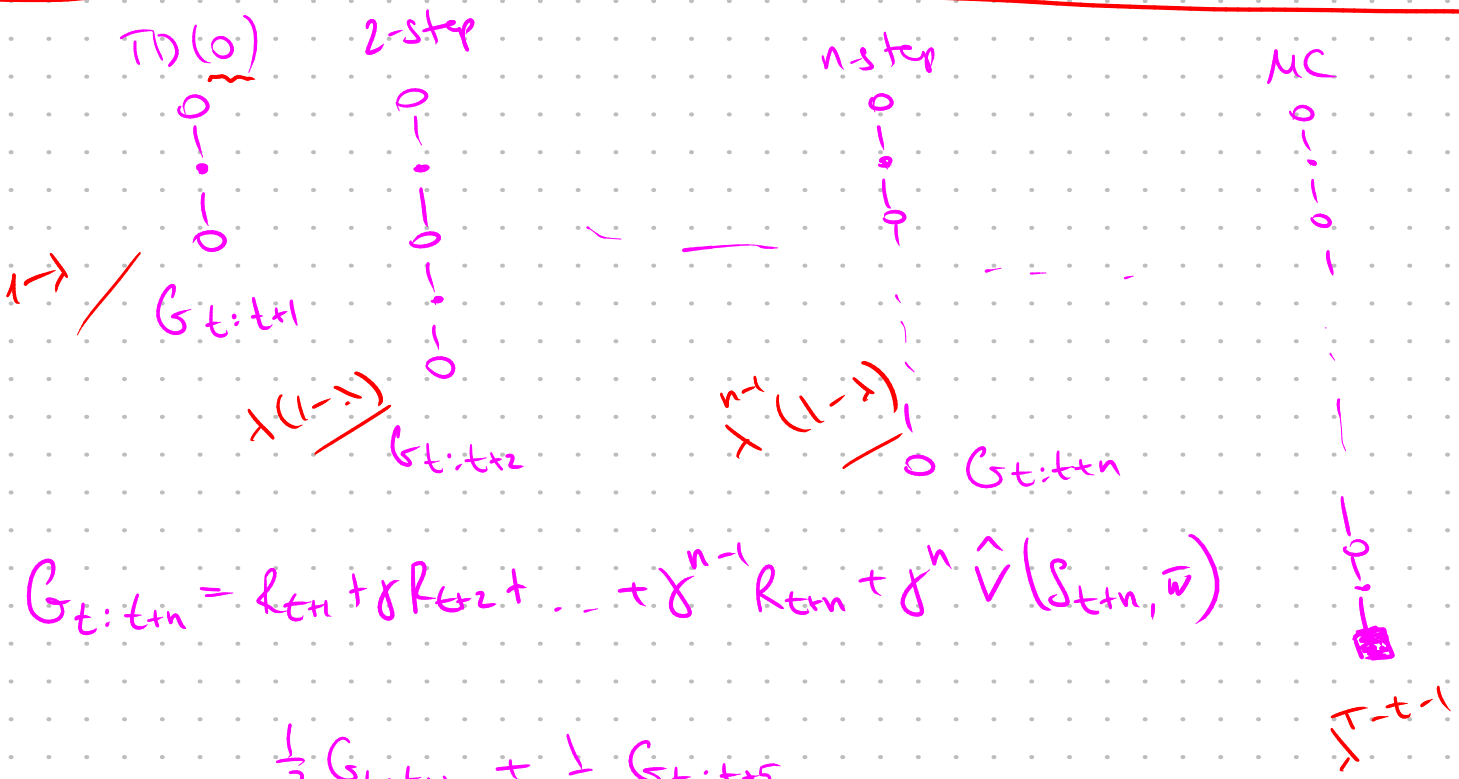
$$= \operatorname{argmin}_{\bar{w}}\left(w^2 + 4w^2 - 2w\left(2w_k\gamma + 4w_k\gamma(1-\varepsilon)\right) + \dots\right]$$

$$\gamma > \frac{5}{6-4\varepsilon}$$

$$w_{k+1} = \frac{6-4\varepsilon}{5}w_k\gamma$$

The Deadly Triad

- function approximation
- bootstrapping
- off-policy training



$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{v}(S_{t+n}, \bar{w})$$

$$\frac{1}{2} G_{t:t+1} + \frac{1}{2} G_{t:t+5}$$

Eligibility traces $TD(\lambda)$

$$G_t^\lambda = (1-\lambda) G_{t:t+1} + (1-\lambda)\cdot\lambda\cdot G_{t:t+2} +$$

$$+ (1-\lambda)\lambda^2 G_{t:t+3} + \dots =$$

$$= (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n}$$