Convergence of actor-critic

Consider

$$\max_{\theta} V^{\theta}(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(X_k, U_k) \middle| X_0 = x\right]$$

$$\text{s.t.} \quad X_{k+1} \sim f(x_{k+1} | x_k, u_k)$$

$$U_k = \rho^{\theta}(X_k)$$

Recall the HJB:

$$V^*(x) = \max_u \left\{ r(x,u) + \gamma \mathbb{E}\left[V^*(X_+^u)\right] \right\}$$

Recall the Q-function:

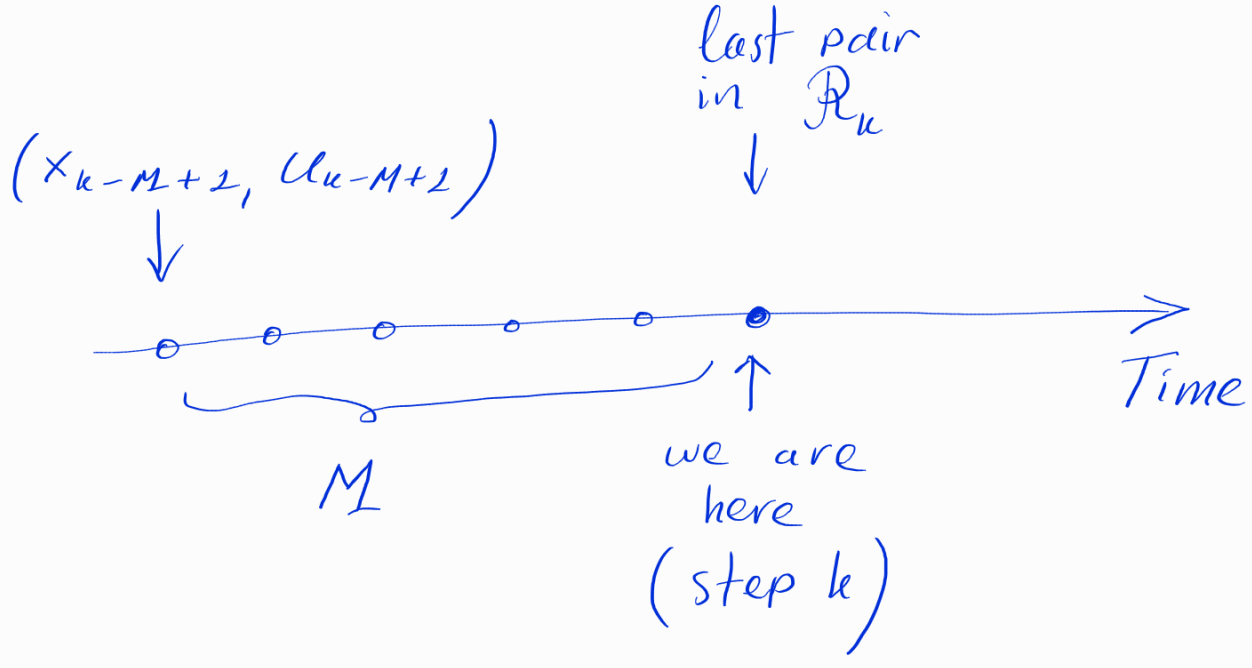
$$Q^*(x,u) = r(x,u) + \gamma \mathbb{E}\left[V^*(X_+^u)\right]$$

Assuming a NN model of the Q-function $\hat{Q}^w(x,u)$, state the TD:

$$e_{TD}^w(x,u,x_+,u_+) = \hat{Q}^w(x,u) - r(x,u) - \hat{Q}^w(x_+,u_+)$$

Let's say we have a data buffer (experience replay):

$$\mathcal{R}_k := \left\{ (x_{k-j+1}, u_{k-j+1}) \right\}_{j=1}^{M}$$



First, let's pick a model:

$$\hat{Q}^w(x,u) = w^T \varphi(x,u)$$

↑ feature vector

Getting back to the TD in terms of this model:

$$e_{TD}^w(x_t, u_t, x_{t+1}, u_{t+1}) = w^T \varphi(x_t, u_t) - r(x_t, u_t) - w^T \varphi(x_{t+1}, u_{t+1})$$

By the approximation theorem,

$$Q^*(x,u) = w^{*T}\varphi(x,u) + e_Q(x,u)$$

↑ ideal weights (say, in terms of sup norm)

Let's introduce the weight error:

$$\tilde{w} := w - w^*$$

Get back to the TD once again:

$$e_{TD}^{w}(x,u,x_+,u_+) = w^T\varphi(x,u) - r(x,u) - w^T\varphi(x_+,u_+)$$

$$= w^T(\varphi(x,u) - \varphi(x_+,u_+)) - r(x,u)$$

$$= \tilde{w}^T(\varphi(x,u) - \varphi(x_+,u_+)) - r(x,u)$$
$$+ w^{*T}\varphi(x,u) - w^{*T}\varphi(x_+,u_+)$$

$$= \tilde{w}^T(\varphi(x,u) - \varphi(x_+,u_+)) - r(x,u)$$
$$+ Q^*(x,u) - e_Q(x,u)$$
$$- Q^*(x_+,u_+) + e_Q(x_+,u_+)$$

Let's denote

$$e^*(x,u,x_+,u_+) := Q^*(x,u) - r(x,u) - Q^*(x_+,u_+)$$

If we followed the optimal policy precisely, e^* would equal zero.

So, all in all

$$e^{\omega}_{TD}(x,u,x_+,u_+) = \tilde{\omega}^T \underbrace{d(x,u,x_+,u_+)}_{\text{data vector}}$$

$$- e_Q(x,u) + e_Q(x_+,u_+)$$

$$+ e^*(x,u,x_+,u_+)$$

Now, finally, the critic loss:

$$J^c(\omega \mid \mathcal{R}_k) = \frac{1}{2} \sum_{j=k-M+1}^{k-1} \frac{\left(e^{\omega}_{TD}(x_j,u_j,x_{j+1},u_{j+1})\right)^2}{\left(d_j^T d_j + 1\right)^2}$$

shorthand $d_j := d(x_j,u_j,x_{j+1},u_{j+1})$

Let's learn weights via gradient descent on the critic loss:

$$w_{k+1} := w_k - \alpha \nabla_w J^c(w_k | \mathcal{R}_k)$$

α: learning rate

Let's work out the gradient:

$$\nabla_w J^c(w_k | \mathcal{R}_k) = \sum_{j=k-M+1}^{k-1} \frac{e_{TD}^{w_k}(x_j, u_j, x_{j+1}, u_{j+1}) \nabla_w e^{w_k}(x_j, u_j, x_{j+1}, u_{j+1})}{(d_j^T d_j + 1)^2}$$

In turn,

$$\nabla_w e^{w_k}(x_j, u_j, x_{j+2}, u_{j+1}) = d_j$$

So,

$$\nabla_w J^c(w_k | \mathcal{R}_k) = \sum_{j=k-M+1}^{k-1} \frac{e_{TD}^{w_k}(x_j, u_j, x_{j+1}, u_{j+1}) d_j}{(d_j^T d_j + 1)^2}$$

$$\nabla_w J^c(w_k / R_k) =$$

$$\sum_{j=k-M+1}^{k-1} \frac{d_j d_j^T}{(d_j^T d_j + 1)^2} \tilde{w}_k +$$

$$\sum_{j=k-M+1}^{k-1} \frac{d_j \left(e_Q(x_{j+1}, u_{j+1}) - e_Q(x_j, u_j) + e^*(x_j, u_j, x_{j+1}, u_{j+1}) \right)}{(d_j^T d_j + 1)^2}$$

Denote a data matrix:

$$\mathcal{E}_k := \sum_{j=k-M+1}^{k-1} \frac{d_j d_j^T}{(d_j^T d_j + 1)^2} \geq 0$$

Let's pretend that e_Q was zero and e^* was zero

We have the evolution of the weights described by:

$$w_{k+1} = w_k - \alpha \mathcal{E}_k \tilde{w}_k$$

Evidently

$$\tilde{w}_{k+1} = \tilde{w}_k - \alpha \mathcal{E}_k \tilde{w}_k$$

We will compare now how the weight error changes from a time step to a time step in norm square:

$$\tilde{w}_{k+1}^T \tilde{w}_{k+1} - \tilde{w}_k^T \tilde{w}_k =$$

$$\left(\tilde{w}_k^T - \alpha \mathcal{E}_k \tilde{w}_k^T\right)\left(\tilde{w}_k - \alpha \mathcal{E}_k \tilde{w}_k\right) - \tilde{w}_k^T \tilde{w}_k =$$

$$-2\alpha \mathcal{E}_k \|\tilde{w}_k\|^2 + \alpha^2 \mathcal{E}_k^2 \|\tilde{w}_k\|^2$$

Recall:

$$\mathcal{E}_k = \sum_{j=k-M+1}^{k-1} \frac{d_j d_j^T}{(d_j^T d_j + 1)^2}$$

First, $\|\mathcal{E}_k\|^2 \leq \frac{M}{4}$

Second, we need a

persistence of excitation condition to hold:

$$\forall k \quad \mathcal{E}_k \gtrsim \varepsilon I_{N_c}$$

$\downarrow$

# of critic features

Then,

$$\tilde{w}_{k+1}^T \tilde{w}_{k+1} - \tilde{w}_k^T \tilde{w}_k \leq$$

$$-2\alpha\varepsilon\|\tilde{w}_k\|^2 + \alpha^2 \frac{M}{4}\|\tilde{w}_k\|^2$$

For $\|\tilde{w}_k\|^2$ to „squeeze", we need to ensure

$$\text{(need)} \quad -2\alpha\varepsilon\|\tilde{w}_k\|^2 + \alpha^2\frac{M}{4}\|\tilde{w}_k\|^2 < 0$$

$$\Rightarrow \text{(need)} \quad -2\alpha\varepsilon + \alpha^2\frac{M}{4} < 0$$

$$\Rightarrow \text{(need)} \quad -2\varepsilon + \alpha\frac{M}{4} < 0$$

$\Rightarrow$ (need) $\quad \alpha \frac{M}{\gamma} < 2\varepsilon$

$\Rightarrow \quad \alpha < \frac{2\varepsilon}{M}$

This is an ideal scenario, but reality there are errors e_Q, e^*.

How do they influence the learning?

Looking at

$$\sum_{j=k-M+1}^{k-1} \frac{d_j \left(e_Q(x_{j+1}, u_{j+1}) - e_Q(x_j, u_j) + e^*(x_j, u_j, x_{j+1}, u_{j+1}) \right)}{(d_j^T d_j + 1)^2}$$

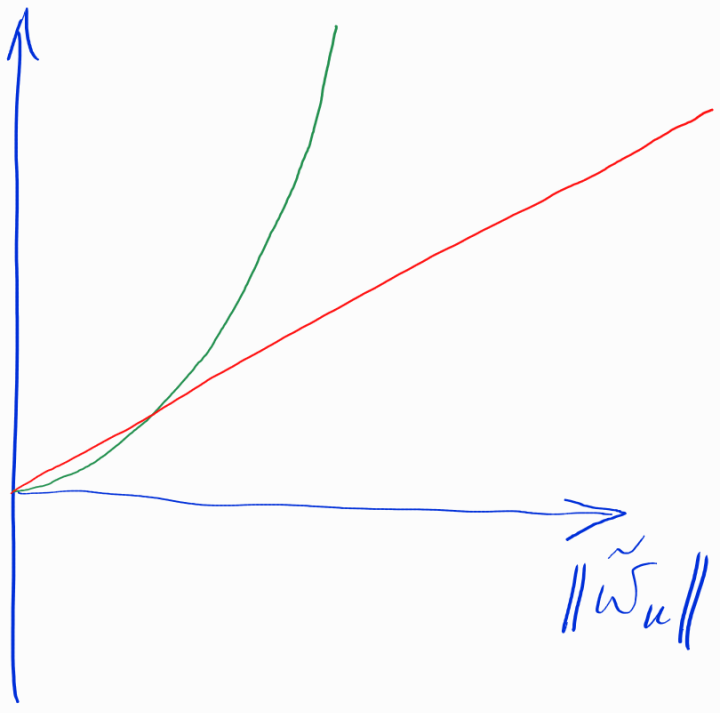
we may write

$$\tilde{w}_{k+1}^T \tilde{w}_{k+1} - \tilde{w}_k^T \tilde{w}_k =$$

NB! NB!

$$\left(\tilde{w}_k^T - \alpha \varepsilon_k \tilde{w}_k^T + C \right) \left(\tilde{w}_k - \alpha \varepsilon_k \tilde{w}_k + C \right) - \tilde{w}_k^T \tilde{w}_k \leq$$

$$-2\alpha\varepsilon \|\tilde{w}_k\|^2 + \alpha^2 \frac{M}{\gamma} \|\tilde{w}_k\|^2 + O(\alpha \|\tilde{w}_k\|)$$



Actor: $\rho^\theta(x) = \theta^T \psi(x)$

$$\max_\theta \hat{Q}^\omega(x, \theta^T \psi(x))$$

$$\theta_{k+1} := \theta_k + \beta \nabla_\theta \hat{Q}^{\omega_k}(x_k, \theta_k^T \psi(x_k))$$

$$= \theta_k + \beta \nabla_\theta \omega_k^T \varphi(x_k, \theta_k^T \psi(x_k))$$