Remarks on deep RL

Problem setup

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

s.t. $X_{k+1} \sim f(x_{k+2} \,|\, x_k, u_k)$, $x \in \mathbb{X} \subseteq \mathbb{R}^n$

$U_k = \rho^{\theta}(X_k)$, $u \in \mathbb{U} \subseteq \mathbb{R}^m$

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

$X_+^u \sim f(x_+ \,|\, x, u)$

Q-function:

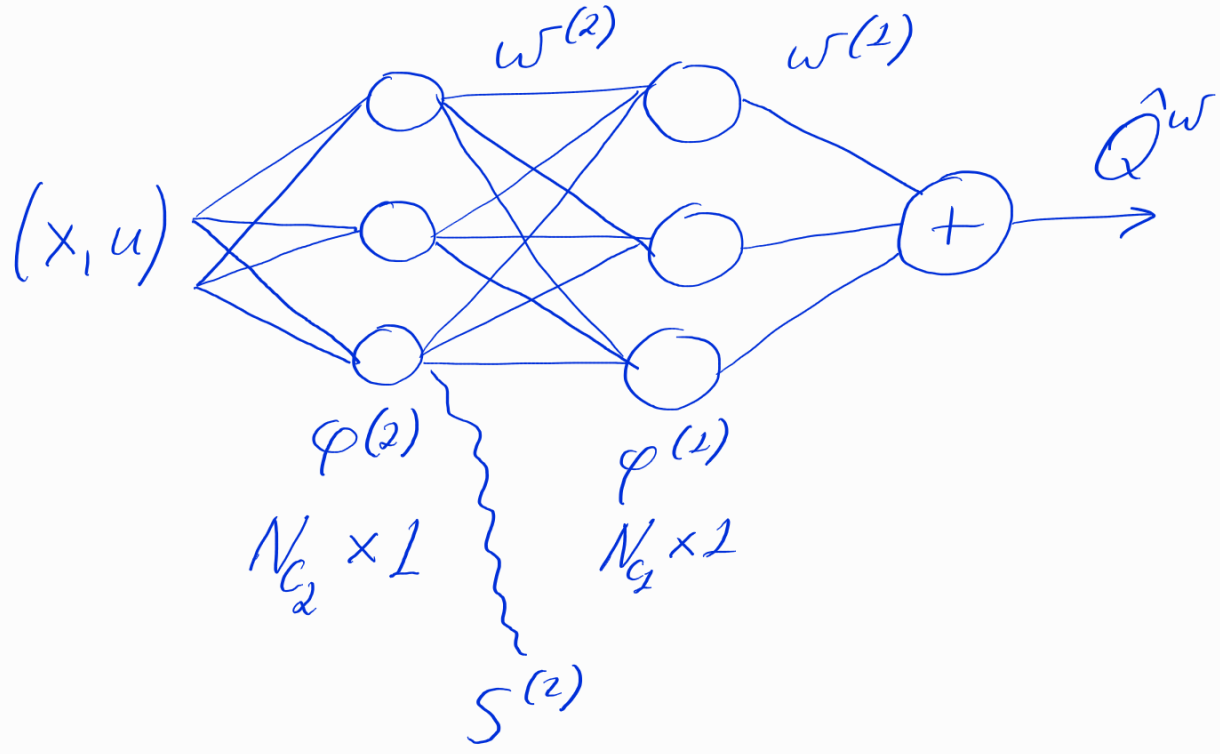
$$Q^*(x,u) = r(x,u) + \gamma \mathbb{E}\left[V^*(X_+^u)\right]$$

TD:

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

where

$$\hat{Q}^{w}(x,u) = w^{(1)T} \varphi^{(1)}\left(w^{(2)T} \varphi^{(2)}(x,u)\right)$$



$w^{(1)}$ is $N_{c_1} \times 1$

$w^{(2)}$ is $N_{c_1} \times N_{c_2}$

$z := (x, u)$

So,

$$\mathcal{E}_{TD}^{w}(z, z_+) = w^{(1)T}\varphi^{(1)}\left(w^{(2)T}\varphi^{(2)}(z)\right) - r(z) - w^{(1)T}\varphi^{(1)}\left(w^{(2)T}\varphi^{(2)}(z_+)\right)$$

Let's linearize the TD wrt weights.

$$\underbrace{\nabla_{w^{(1)}} e_{TD}^{w}(z, z_+)}_{N_{c_1} \times 1} = \varphi^{(1)}\left(w^{(2)^T} \varphi^{(1)}(z)\right) - \varphi^{(1)}\left(w^{(2)^T} \varphi^{(1)}(z_+)\right)$$

Now, gradient wrt hidden layer weights.

First,

$$\underbrace{g(w^{(2)}, z, z_+)}_{N_{c_1} \times 1} := \varphi^{(1)}\left(w^{(2)^T} \varphi^{(1)}(z)\right) - \varphi^{(1)}\left(w^{(2)^T} \varphi^{(1)}(z_+)\right)$$

With this at hand, we can write:

$$e_{TD}^{w}(z, z_+) = \sum_{\ell=1}^{N_{c_1}} w_\ell^{(1)} g_\ell(w^{(2)}, z, z_+) - r(z)$$

ℓth entry

Let's introduce an operator:

$\mathrm{vec}(\bullet)$ — conversion of a matrix into a column vector

by stacking up the matrix columns

Then,

$$\nabla_{vec(w^{(2)})} \ell_{TD}^{w}(z, z_+) = \sum_{\ell=1}^{N_{c_1}} \nabla_{vec(w^{(2)})} g_\ell(w^{(2)}, z, z_+) \, w_\ell^{(1)}$$

$N_{c_2} \cdot N_{c_1} \times 1$

$$\left(\nabla_{vec(w^{(2)})} g_1(\dots) \right) w_1^{(1)} + \dots + \left(\dots \right) w_{N_{c_1}}^{(1)}$$

$$G = \left(\nabla \quad \dots \quad \nabla \right)$$

$$w^{(1)} = \begin{pmatrix} w_1^{(1)} \\ \vdots \\ w_{N_{c_1}}^{(1)} \end{pmatrix}$$

$G w^{(1)}$

$$\nabla_{vec(w^{(2)})} \ell_{TD}^{w}(z, z_+) = G(w^{(2)}, z, z_+) w^{(1)}$$

Notice that G depends (in general) non-linearly on $w^{(2)}$

With these gradients at hand, we can now linearize the TD:

first, the gradient reads:

$$\nabla_{vec(w)} e_{TD}^{w}(z, z_+) = \begin{pmatrix} \overbrace{\varphi^{(1)}\left(w^{(1)^T}\varphi^{(2)}(z)\right) - \varphi^{(1)}\left(w^{(1)^T}\varphi^{(2)}(z_+)\right)}^{g(w^{(1)}, z, z_+)} \\ G\left(w^{(2)}, z, z_+\right) w^{(1)} \end{pmatrix}$$

Back to linearization:

$$e^{w}_{TD}(z, z_+) = \overbrace{w^{*(1)^T} g(w^{*(2)}, z, z_+)}^{\text{"pivot"}} - r(z) -$$
$$\left(\nabla_{vec(w)} e^{w}_{TD}(z, z_+)\Big|_{w=w}\right)^T vec(\tilde{w}) +$$
$$\mathcal{O}\left(\|vec(\tilde{w})\|^2\right)$$

$$w = \{w^{(1)}, w^{(2)}\}$$
$$\rightsquigarrow \quad vec(w) = (w^{(1)}, vec(w^{(2)}))$$
$$vec(\tilde{w}) := vec(w) - vec(w^*)$$

Following the recipe of the lecture on shallow actor-critic convergence, we can extract NN errors and error related to the „quality" of the policy.

$$e^{w}_{TD}(z, z_+) = \left(\nabla_{vec(w)} e^{w}_{TD}(z, z_+)\Big|_{w}\right)^T vec(\tilde{w}) -$$
$$e_Q(z) + e_Q(z_+) + e^*(z, z_+) +$$
$$\mathcal{O}\left(\|vec(\tilde{w})\|^2\right) !$$

Now, data vector

$$d(z, z_+, w) := \nabla_{vec(w)} e^w_{TD}(z, z_+)\big|_w$$

$$\Rightarrow$$

$$e^w_{TD}(z, z_+) = d^T(z, z_+, w)\, vec(\tilde{w}) - \dots$$

↑ Is this bad?

< repeating the setup from the shallow NN case >

experience replay ↓

$$\nabla_w J_c(w_u \mid \mathcal{R}_u) =$$

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

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

$$\sum_{j=k-M+2}^{k-1} \frac{d_j \, O\left(\|vec(\tilde{w}_u)\|^2\right)}{(d_j^T d_j + 1)^2}$$

$$\text{vec}(\tilde{\omega}_{k+1})^T \text{vec}(\tilde{\omega}_{k+1}) - \text{vec}(\tilde{\omega}_k)^T \text{vec}(\tilde{\omega}_k) \leq$$

$$\underbrace{\leq -c \|\text{vec}(\tilde{\omega}_k)\|^2}$$

$$- 2\alpha\varepsilon \|\text{vec}(\tilde{\omega}_k)\|^2 + \alpha^2 \frac{M}{q} \|\text{vec}(\tilde{\omega}_k)\|^2 +$$

$$O(\alpha \|\text{vec}(\omega_k)\|) + O(\alpha \|\text{vec}(\tilde{\omega}_k)\|^3)$$

* c comes from the learning rate choice for convergence