# Reinforcement Learning

## 1 Value-function based methods

### 1.1 Deep Q-Networks (DQN)

$$Q^{\pi}(s,a) = \mathbb{E}[R_{t+1}|s_t = s, a_t = a, \pi]$$

$$Y_t^{DQN} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \boldsymbol{\theta}_t^-)$$

$$\nabla_{\theta_t} L_t(\theta_t) = \mathbb{E}\left[\left(Y_t^{DQN} - Q(s,a;\theta_t)\right)\nabla_{\theta_t} Q(s,a;\theta_t)\right]$$

### 1.2 Double Deep Q-Networks (DDQN)

$$Q^{\pi}(s,a) = \mathbb{E}[R_{t+1}|s_t = s, a_t = a, \pi]$$

$$Y_t^{DDQN} \equiv R_{t+1} + \gamma Q\left(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t^-\right)$$

$$\nabla_{\theta_t} L_t(\theta_t) = \mathbb{E}\left[\left(Y_t^{DDQN} - Q(s,a;\theta_t)\right)\nabla_{\theta_t} Q(s,a;\theta_t)\right]$$

### 1.3 Dueling Deep Q-Networks

$$Q^{\pi}(s,a) = \mathbb{E}[R_{t+1}|s_t = s, a_t = a, \pi]$$

$$V^{\pi}(s,a) = \mathbb{E}[Q^{\pi}(s,a)]$$

$$Q^{\pi}(s,a;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}) = V^{\pi}(s;\boldsymbol{\theta},\boldsymbol{\beta}) + A^{\pi}(s,a;\boldsymbol{\theta},\boldsymbol{\alpha})$$

$$Q^{\pi}(s,a;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}) = V^{\pi}(s;\boldsymbol{\theta},\boldsymbol{\beta}) + \left(A^{\pi}(s,a;\boldsymbol{\theta},\boldsymbol{\alpha}) - \max_{a' \in |\mathcal{A}|} A^{\pi}(s,a;\boldsymbol{\theta},\boldsymbol{\alpha})\right)$$

$$Q^{\pi}(s,a;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}) = V^{\pi}(s;\boldsymbol{\theta},\boldsymbol{\beta}) + \left(A^{\pi}(s,a;\boldsymbol{\theta},\boldsymbol{\alpha}) - \frac{1}{|\mathcal{A}|}\Sigma_{a'} A^{\pi}(s,a;\boldsymbol{\theta},\boldsymbol{\alpha})\right) \text{ (Alternative)}$$

$$Y_t^{Q} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \boldsymbol{\theta}_t^-)$$

$$\nabla_{\theta_t} L_t(\theta_t) = \mathbb{E}\left[\left(Y_t^{Q} - Q(s,a;\theta_t)\right)\nabla_{\theta_t} Q(s,a;\theta_t)\right]$$

## 2 Model-based methods

### 2.1 Simulated Policy Learning (SimPle)

## 3 Policy-based methods

### 3.1 REINFORCE

### 3.2 Actor-Critic

### 3.3 Off-Policy Policy Gradient

## 4 Hierarchical methods