Sutton & Barto RL Cheatsheet

Pierluca D'Oro

Monte Carlo Exploring Starts

Initialize, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
$\quad Q(s,a) \leftarrow$ arbitrary
$\quad \pi(s) \leftarrow$ arbitrary
$\quad \mathit{Returns}(s,a) \leftarrow$ empty list

Repeat forever:
$\quad$ Choose $S_0 \in \mathcal{S}$ and $A_0 \in \mathcal{A}(S_0)$ randomly
$\quad$ Generate episode starting from S_0, A_0, using π
$\quad$ For each pair s, a appearing in episode:
$\quad\quad G \leftarrow$ return following first s, a occurrence
$\quad\quad$ Append G to $\mathit{Returns}(s,a)$
$\quad\quad Q(s,a) \leftarrow \text{average}(\mathit{Returns}(s,a))$
$\quad$ For each s in the episode:
$\quad\quad \pi(s) \leftarrow \text{argmax}_a Q(s,a)$

On policy first-visit MC control

Initialize, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$:
$\quad Q(s,a) \leftarrow$ arbitrary
$\quad \mathit{Returns}(s,a) \leftarrow$ empty list
$\quad \pi(a|s) \leftarrow$ arbitrary $\epsilon-$soft policy

Repeat forever:
$\quad$ (a) Generate an episode using π
$\quad$ (b) For each pair s, a appearing in the episode:
$\quad\quad\quad G \leftarrow$ return following first s, a occurrence
$\quad\quad\quad$ Append G to $\mathit{Returns}(s,a)$
$\quad\quad\quad Q(s,a) \leftarrow \text{average}(\mathit{Returns}(s,a))$
$\quad$ (c) For each s in the episode:
$\quad\quad\quad A^* \leftarrow \text{argmax}_a Q(s,a)$
$\quad\quad\quad$ For all $a \in \mathbf{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1-\epsilon+\epsilon/|\mathcal{A}(s)|, & \text{if } a = A^* \\ \epsilon/|\mathcal{A}(s)|, & \text{if } a \neq A^* \end{cases}$$

Monte Carlo Policy Gradient (REINFORCE)

Given $\pi_{\boldsymbol{\theta}}(a|s)$, initialize $\boldsymbol{\theta} \in \mathbb{R}^{d'}$
Repeat forever:
$\quad$ Generate an episode following π
$\quad$ For each step t to T in the episode:
$\quad\quad G \leftarrow$ return from step t
$\quad\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\gamma^t G \nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(A_t|S_t)$

Q-learning (off-policy TD control)

Initialize $Q(s,a)$ arbitrarily, $Q(\mathit{terminal\text{-}state}, -) = 0$

Repeat (for each episode):
$\quad$ Initialize S
$\quad$ Repeat (for each step of episode):
$\quad\quad$ Choose A in S with ϵ-greedy policy from Q
$\quad\quad$ Take action A, observe R, S'
$\quad\quad Q(S,A) \leftarrow (1-\alpha)Q(S,A) +$
$\quad\quad\quad\quad\quad \alpha(R + \gamma \max_a Q(S',a))$
$\quad\quad S \leftarrow S'$
$\quad$ Until S is terminal

Sarsa (on-policy TD control)

Initialize $Q(s,a)$ arbitrarily, $Q(\mathit{terminal\text{-}state}, -) = 0$

Repeat (for each episode):
$\quad$ Initialize S
$\quad$ Choose A in S with ϵ-greedy policy from Q
$\quad$ Repeat (for each step of episode):
$\quad\quad$ Take action A, observe R, S'
$\quad\quad$ Choose A' in S' with ϵ-greedy policy from Q
$\quad\quad Q(S,A) \leftarrow (1-\alpha)Q(S,A) +$
$\quad\quad\quad\quad\quad \alpha(R + \gamma Q(S',A'))$
$\quad\quad S \leftarrow S'$
$\quad\quad A \leftarrow A'$
$\quad$ Until S is terminal

One-step Actor-Critic

Given $\pi_{\boldsymbol{\theta}}(a|s)$, $\hat{v}_{\boldsymbol{\omega}}(s)$, initialize $\boldsymbol{\theta} \in \mathbb{R}^{d'}$, $\boldsymbol{\omega} \in \mathbb{R}^{d}$

Repeat forever:
$\quad$ Initialize S
$\quad I \leftarrow 1$
$\quad$ While S is not terminal:
$\quad\quad A \sim \pi_{\boldsymbol{\theta}}(A|S)$
$\quad\quad$ Take action A, observe S', R
$\quad\quad \delta \leftarrow R + \gamma\hat{v}_{\boldsymbol{\omega}}(S') - \hat{v}_{\boldsymbol{\omega}}(S)$
$\quad\quad \boldsymbol{\omega} \leftarrow \boldsymbol{\omega} + \alpha I \delta \nabla_{\boldsymbol{\omega}} \hat{v}_{\boldsymbol{\omega}}(S)$
$\quad\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta I \delta \nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(A|S)$
$\quad\quad I \leftarrow \gamma I$
$\quad\quad S \leftarrow S'$

Value Iteration

Initialize array V arbitrarily

Repeat
$\quad \Delta \leftarrow 0$
$\quad$ For each $s \in \mathcal{S}$:
$\quad\quad v \leftarrow V(s)$
$\quad\quad V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$
$\quad\quad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \epsilon$

Output a deterministic policy $\pi \approx \pi_*$:
$\quad \pi(s) = \text{argmax}_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$

Policy Iteration

(1) Initialization
$V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all s $\in \mathcal{S}$

(2) Policy Evaluation
Repeat
$\quad \Delta \leftarrow 0$
$\quad$ For each $s \in \mathcal{S}$:
$\quad\quad v \leftarrow V(s)$
$\quad\quad V(s) \leftarrow \sum_{s',r} p(s',r|s,\pi(s))[r + \gamma V(s')]$
$\quad\quad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \epsilon$

(3) Policy Improvement
policy-stable $\leftarrow$ *true*
For each $s \in \mathcal{S}$:
$\quad$ *old-action* $\leftarrow \pi(s)$
$\quad \pi(s) \leftarrow \text{argmax}_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$
$\quad$ If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
If *policy-stable*, then stop, else go to (2)