Reinforcement Learning (Tabular Methods) Reference Sheet

Jian W Dong, based on Sutton & Barto: Reinforcement Learning - An Introduction 2E

Algorithm	On/Off Policy	Model-Free/ Model-Based	Control/ Prediction	Policy Update	Objective	Bootstrap	Target	Update Rule	When to Use
DYNAMIC PROGRAMMING (Chapter 4)									
Policy Eval	On	Model	Pred	None	V^π	Yes	$\sum_a \pi(a\|s) \sum_{s',r} p(s',r\|s,a)[r + \gamma V(s')]$	$V(s) \leftarrow$ Target	Known environment, need policy evaluation
Policy Iter	On	Model	Ctrl	Greedy	V^*	Yes	Same as Policy Eval	Policy Eval + $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s',r\|s,a)[r + \gamma V(s')]$	Known environment, guaranteed optimal policy
State-Value Iter	On	Model	Ctrl	Greedy	V^*	Yes	$\max_a \sum_{s',r} p(s',r\|s,a)[r + \gamma V(s')]$	$V(s) \leftarrow$ Target	Known environment, faster than policy iteration
Action-Value Iter	On	Model	Ctrl	Greedy	Q^*	Yes	$\sum_{s',r} p(s',r\|s,a)[r + \gamma \max_{a'} Q(s',a')]$	$Q(s,a) \leftarrow$ Target	Known environment, directly learns action values
MONTE CARLO METHODS (Chapter 5)									
First-Visit MC	On	Free	Pred	None	V^π	No	$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$ (from first visit)	$V(S_t) \leftarrow V(S_t) + \alpha[\text{Target} - V(S_t)]$	Episodic tasks, unbiased estimates
Every-Visit MC	On	Free	Pred	None	V^π	No	$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$ (from each visit)	$V(S_t) \leftarrow V(S_t) + \alpha[\text{Target} - V(S_t)]$	Episodic tasks, more data per episode
MC Exploring	On	Free	Ctrl	Greedy	Q^*	No	$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[\text{Target} - Q(S_t,A_t)]$	Episodic, can ensure exploring starts
On-policy MC	On	Free	Ctrl	ϵ-gr	Q^π	No	$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[\text{Target} - Q(S_t,A_t)]$	Episodic, practical exploration
Off-policy MC Pred	Off	Free	Pred	None	V^π	No	$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$	Weighted IS: $C(S_t) \leftarrow C(S_t) + \rho_{t:T-1}$, $V(S_t) \leftarrow V(S_t) + \frac{\rho_{t:T-1}}{C(S_t)}[G_t - V(S_t)]$	Learn about different policy than behavior
Off-policy MC Ctrl	Off	Free	Ctrl	Greedy	Q^*	No	$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$	Weighted IS: $C(S_t,A_t) \leftarrow C(S_t,A_t) + \rho_{t+1:T-1}$, $Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \frac{\rho_{t+1:T-1}}{C(S_t,A_t)}[G_t - Q(S_t,A_t)]$	Learn optimal policy from suboptimal data
TEMPORAL-DIFFERENCE LEARNING (Chapter 6)									
TD(0)	On	Free	Pred	None	V^π	Yes	$R_{t+1} + \gamma V(S_{t+1})$	$V(S_t) \leftarrow V(S_t) + \alpha[\text{Target} - V(S_t)]$	Online learning, fast updates
SARSA	On	Free	Ctrl	ϵ-gr	Q^π	Yes	$R_{t+1} + \gamma Q(S_{t+1},A_{t+1})$	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[\text{Target} - Q(S_t,A_t)]$	Safe online control, conservative
Q-learning	Off	Free	Ctrl	ϵ-gr	Q^*	Yes	$R_{t+1} + \gamma \max_{a'} Q(S_{t+1},a')$	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[\text{Target} - Q(S_t,A_t)]$	Learn optimal policy, exploration/exploitation
Expected SARSA	On/Off	Free	Ctrl	ϵ-gr	Q^π	Yes	$R_{t+1} + \gamma \sum_{a'} \pi(a'\|S_{t+1}) Q(S_{t+1},a')$	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[\text{Target} - Q(S_t,A_t)]$	Lower variance than SARSA
Double Q-learning	Off	Free	Ctrl	ϵ-gr	Q^*	Yes	Alternate: $R_{t+1} + \gamma Q_B(S_{t+1}, \arg\max_{a'} Q_A(S_{t+1},a'))$ or $R_{t+1} + \gamma Q_A(S_{t+1}, \arg\max_{a'} Q_B(S_{t+1},a'))$	Randomly update $Q_A(S_t,A_t)$ or $Q_B(S_t,A_t)$	Avoid overestimation bias
n-STEP BOOTSTRAPPING (Chapter 7)									
n-step TD	On	Free	Pred	None	V^π	Yes	$G_{t:t+n} = \sum_{i=0}^{n-1} \gamma^i R_{t+i+1} + \gamma^n V(S_{t+n})$	$V(S_t) \leftarrow V(S_t) + \alpha[\text{Target} - V(S_t)]$	Bridge MC and TD, tune bias/variance
n-step SARSA	On	Free	Ctrl	ϵ-gr	Q^π	Yes	$G_{t:t+n} = \sum_{i=0}^{n-1} \gamma^i R_{t+i+1} + \gamma^n Q(S_{t+n},A_{t+n})$	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[\text{Target} - Q(S_t,A_t)]$	Multi-step lookahead, on-policy
n-step Tree Backup	Off	Free	Ctrl	Any	Q^π	Yes	$G^{tree}_{t:t+n} = R_{t+1} + \gamma[\sum_{a \neq A_{t+1}} \pi(a\|S_{t+1}) Q(S_{t+1},a) + \pi(A_{t+1}\|S_{t+1}) G^{tree}_{t+1:t+n}]$	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[\text{Target} - Q(S_t,A_t)]$	Off-policy without importance sampling
n-step $Q(\sigma)$	On/Off	Free	Ctrl	Any	Q^π	Yes	σ-weighted combination of SARSA and Tree Backup	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[\text{Target} - Q(S_t,A_t)]$	Unify on/off-policy methods
Off-policy n-step	Off	Free	Ctrl	Any	Q^π	Yes	$G_{t:t+n} = \sum_{i=0}^{n-1} \gamma^i R_{t+i+1} + \gamma^n Q(S_{t+n},A_{t+n})$	$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha \rho_{t+1:t+n-1}[\text{Target} - Q(S_t,A_t)]$	Off-policy with multi-step returns
PLANNING AND LEARNING WITH TABULAR METHODS (Chapter 8)									

Algorithm	**On/Off Policy**	**Model-Free/ Model-Based**	**Control/ Prediction**	**Policy Update**	**Objective**	**Bootstrap**	**Target**	**Update Rule**	**When to Use**
Dyna-Q	Off	Model	Ctrl	ϵ-gr	Q^*	Yes	$r + \gamma \max_{a'} Q(s', a')$ (same as Q-learning)	Q-learning update + model learning + planning	Combine learning and planning
Dyna-Q+	Off	Model	Ctrl	ϵ-gr	Q^*	Yes	$r + \kappa\sqrt{\tau} + \gamma \max_{a'} Q(s', a')$	Same as Dyna-Q with exploration bonus	Handle changing environments
Prioritized Sweeping	Off	Model	Ctrl	ϵ-gr	Q^*	Yes	$r + \gamma \max_{a'} Q(s', a')$	Updates prioritized by $\lvert \text{Target} - Q(s, a) \rvert > \theta$	Efficient planning, focus important updates
Trajectory Sampling	Off	Model	Ctrl	ϵ-gr	Q^*	Yes	$r + \gamma \max_{a'} Q(s', a')$	Sample long trajectories vs. uniform sweeping	Better state distribution for planning
Real-time DP	On	Model	Ctrl	Greedy	V^*	Yes	$\max_a \sum_{s',r} p(s', r \vert S_t, a)[r + \gamma V(s')]$	$V(S_t) \leftarrow$ Target only for visited states	Online DP, focus on relevant states
ELIGIBILITY TRACES (Chapter 12)									
Offline λ-return	On	Free	Pred	None	V^π	Yes	$G_t^\lambda = (1 - \lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t$	$V(S_t) \leftarrow V(S_t) + \alpha[G_t^\lambda - V(S_t)]$ (offline at episode end)	Theoretical foundation for TD(λ), offline learning
TD(λ)	On	Free	Pred	None	V^π	Yes	$R_{t+1} + \gamma V(S_{t+1})$ (TD error: δ_t)	$V(s) \leftarrow V(s) + \alpha \delta_t e_t(s)$ where $e_t(s) = \gamma \lambda e_{t-1}(s) + \mathbf{1}_{S_t=s}$	Credit assignment, faster learning
SARSA(λ)	On	Free	Ctrl	ϵ-gr	Q^π	Yes	$R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$ (TD error: δ_t)	$Q(s, a) \leftarrow Q(s, a) + \alpha \delta_t e_t(s, a)$	On-policy with eligibility traces
Q(λ)	Off	Free	Ctrl	ϵ-gr	Q^*	Yes	$R_{t+1} + \gamma \max_a Q(S_{t+1}, a)$ (TD error: δ_t)	Watkins's Q(λ): traces reset if non-greedy action	Off-policy with traces (limited)
True Online TD(λ)	On	Free	Pred	None	V^π	Yes	$R_{t+1} + \gamma V(S_{t+1})$ (TD error: δ_t)	Modified update with trace correction term	More accurate trace implementation

Notation: $V(s)$: State value, $Q(s, a)$: Action-value, $\pi(a|s)$: Policy, α: Learning rate, γ: Discount, ϵ: Exploration, G_t: Return, ρ: Importance ratio, $e_t(s)$: Eligibility trace, λ: Trace decay, δ_t: TD error, $C(\cdot)$: Cumulative sum of weights for IS.
Abbreviations: Model = Model-based, Free = Model-free, Ctrl = Control, Pred = Prediction, ϵ-gr = ϵ-greedy, IS = Importance Sampling.