Reinforcement Learning Cheat Sheet

Summary of Notation

Math styles:

$\mathbb{P}$	\mathbb{P}
$\mathbf{P}$	\mathbf{P}
$\mathtt{P}$	\mathtt{P}
P	\mathit{P}
$\mathcal{P}$	\mathcal{P}
$\mathfrak{P}$	\mathfrak{P}
$\mathnormal{P}$	\mathnormal{P}
$\bar{P}$	\bar{P}
$\hat{P}$	\hat{P}
$\tilde{P}$	\tilde{P}
$\vec{P}$	\vec{P}
$\overline{P}$	\overline{P}

$\doteq$	equality relationship that is true by definition
$\approx$	approximately equal
$\propto$	proportional to
$\Pr\{X = x\}$	probability that a random variable X takes on value x
X $\sim$ p	random variable X selected from distribution $p(x) \doteq Pr\{X = x\}$
$\mathbb{E}[X]$	expected value of random variable X, i.e., $\mathbb{E}[X] \doteq \sum_x p(x) \cdot x$
$\arg\max_x f(x)$	x that maximizes $f(x)$
lnx	natural logarithm of x
e^x, $\exp(x)$	the base of the natural logarithm, $e \approx 2.71828$, carried to power x, $e^{lnx} = x$
$\mathbb{R}$	set of real numbers
$f : \mathcal{X} \to \mathcal{Y}$	function f from elements of set $\mathcal{X}$ to elements fo set $\mathcal{Y}$
$\leftarrow$	assignment operator
$(a, b]$	interval from a to b including b
ϵ	probability of taking a random action in an ϵ-greedy policy
α, β	step-size parameters
γ	discount factor
λ	decay-rare parameter for eligibility traces
$\mathbb{1}_{predicate}$	indicator function ($\mathbb{1}_{predicate} \doteq 1$ if the *predicate* is true, else 0)

In a multi-arm bandit problem:

k	number of actions(arms)
t	discrete time step or play number
$q_*(a)$	true value (expected reward) of action a
$Q_t(a)$	estimate at time t of $q_*(a)$
$N_t(a)$	number if times action a has been selected up prior to time t
$H_t(a)$	learned preference for selecting action a at time t
$\pi_t(a)$	probability of selecting action a at time t
$\bar{R}_t$	estimate at time t of the expected reward given π_t

In Markov Decision Processes:

s,s'	states
a	an actions
r	a reward
$\mathcal{S}$	set of all non-terminal states
$\mathcal{S}^+$	set of all states, including terminal states
$\mathcal{A}(s)$	set of all actions in state s
$\mathcal{R}$	set of all possible rewards, a finite subset of $\mathbb{R}$
$\subset$	subset of; $(\mathcal{R} \subset \mathbb{R})$
$\in$	is an element of; e.g. $(s \in \mathcal{S},\ r \in \mathcal{R})$
t	discrete time step
$T, T(t)$	final time step of an episode, or of the episode oncluding time step t
A_t	action at tim et
S_t	state at time t, typicaly due, stochastically, to S_{t-1} and A_{t-1}
R_t	state at time t, typicaly due, stochastically, to S_{t-1} and A_{t-1}
π	policy (decision-making-rule)
$\pi(s)$	action taken in state s under *deterministic* policy π
$\pi(a\|s)$	probability of taking action a in state s under *stochastic* policy π
G_t	return following time t
h	horizon, th etime step one look up to in a forward view
$G_{t:t+n}$	n-ste return from $t+1$ to $t+n$, or to h (discounted and corrected)
$\bar{G}_{t:h}$	flat return (undiscounted and uncorrected) from $t+1$ to h
G_t^λ	λ-return
$G_{t:h}^\lambda$	truncated, corrected λ-return
$G_t^{\lambda s}$, $G_t^{\lambda a}$	λ-return, corrected by estimate state, or action, values
$p(s', r\|s, a)$	probability of transitioning to state s' and receiving reward r when in state s and taking action a
$p(s'\|s, a)$	probability of transitioning to state s' when in state s and taking action a
$r(s, a)$	expected reward when in state s and taking action a
$r(s, a, s')$	expected reward when in state s and taking action a, and transitioning to state s'
$v_\pi(s)$	value of state s under policy π (expected-reward)
$v_*(s)$	value of state s under the optimal policy
$q_\pi(s, a)$	value of state s and action a under policy π (expected-reward)
$q_*(s, a)$	value of state s and action a under the optimal policy
V,V_t	array estimates of state values function v_π or v_*
Q,Q_t	array estimates of state-action values function q_π or q_*
$V_t(s)$	expected approximate action value; for example, $\bar{V}_t(s) \doteq \sum_a \pi(a\|a)Q_t(s, a)$
U_t	target for estimate at time t
δ_t	temporal-difference (TD) error at t (a random varibale)
δ_t^s, δ_t^a	state-and action-specific forms of the TD error
n	in n-step methods, n is th enumber of steps of bootstrapping
d	dimensionality – the number of components of
$\mathbf{w}$, $\mathbf{w}_t$	d-vector of weights underlying an approximate value function
w_i, $w_{t,i}$	the ith component of learnable wright vector $\mathbf{w}$
$v_{\mathbf{w}}(s)$	alternate notation for $\hat{v}(s, \mathbf{w})$
$\hat{q}(s, a, \mathbf{w})$	approximate value of state-action pair s,a given weight vector $\mathbf{w}$
$\nabla\hat{v}(s,' vbw)$	column vector of partial derivatives of $\hat{v}(s, \mathbf{w})$ with respect to $\mathbf{w}$
$\nabla\hat{q}(s, a, \mathbf{w})$	column vector of partial derivatives of $\hat{q}(s, a, \mathbf{w})$ with respect to $\mathbf{w}$
$\mathbf{x}(s)$	vector of features visible in state s
$\mathbf{x}(s, a)$	vector of featres visible when in state s taking action s
$x_i(s)$, $x_i(s, a)$	the ith component of $\mathbf{x}(s)$ or $\mathbf{x}(s, a)$
$\mathbf{x}_t$	short-hand for $\mathbf{x}(S_t)$ or $\mathbf{x}(S_t, A_t)$
$\mathbf{w}^\top x$	inner product of vectors, $\mathbf{w}^\top x \doteq= \sum_i w_i x_i$; for example, $\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^\top x(s)$
$\mathbf{v}$, $\mathbf{v}_t$	secondary d-vector of weights, used to learn $\mathbf{w}$
$\mathbf{z}_t$	d-vector of eligibility traces at tim t
Θ, Θ	parameter vector of target policy
$\pi(a\|s, \Theta)$	probability of taking action a in state s given parameter vector Θ
π_Θ	policy corresponding to parameter vector Θ
$\nabla\pi(a\|s, \Theta)$	columb vector of partial derivatives of $\pi(a\|s, \Theta)$ with respect to Θ
$J(\Theta)$	performance measure for th epolicy π_Θ
$\nabla J(\Theta)$	column vector of partial derivatives of $J(\Theta)$ with respect to Θ
$h(s, a, \Theta)$	preference for selecting action a in state s based on Θ
$b(a\|s)$	behavior of policy used to select actions while learning about target policy π
$b(s)$	a baseline function $g : \mathcal{S} \mapsto \mathbb{R}$
b	brnahcing factor for an MDP or search tree
$\rho_{t:h}$	importance sapling ratio for time t through time h
ρ_t	importance sampling ratio for time t alone, $\rho \doteq \rho_{t:t}$
$r(\pi)$	average reward (reward rate) for policy π
$\bar{R}_t$	estimate of $r(\pi)$ at time t
$\mu(s)$	on-policy distribution over states
μ	$\|\mathcal{S}\|$-vector of the $\mu(s)$ for all $s \in \mathcal{S}$
$\|\|v\|\|^2_\mu$	μ-weighted squared norm of value function v,i.e. $\|\|v\|\|^2_\mu \doteq \sum_{s\in\mathcal{S}} \mu(s)v(s)^2$
$\eta(s)$	expected number of visits to state s per episode
Π	projection operator for value functions
B_π	Bellman operator for value functions
$\mathbf{A}$	$d \times d$ matrix $\mathbf{A} \doteq \mathbb{E}\left[\mathbf{x}(\mathbf{x}_t - \gamma\mathbf{x}_{t+1})^\top\right]$
$\mathbf{b}$	d-dimensional vector $\mathbf{b} \doteq \mathbb{E}\left[R_{t+1}\mathbf{x}_t\right]$
$\mathbf{w}_{TD}$	TD fixed point $\mathbf{w}_{TD} \doteq \mathbf{A}^{-1}\mathbf{b}$ (a d-vector)
$\mathbf{I}$	identity matrix
$\mathbf{P}$	$\|\mathcal{S}\| \times \|\mathcal{S}\|$ matrix of state transition probabilities under π
$\mathbf{D}$	$\|\mathcal{S}\| \times \|\mathcal{S}\|$ diagonal matrix with μ on its diagonal
$\mathbf{X}$	$\|\mathbf{S}\| \times d$ matrix with the $\mathbf{x}(s)$ as its rows
$\bar{\delta}(s)$	Bellmann error (expected TD error) for $v_{\mathbf{w}}$ at state s
$\bar{\delta}(s)$, BE	Bellmann error vector, with components $\delta_{\mathbf{w}}(s)$
$\bar{V}E(\mathbf{w})$	mean square value error $\bar{V}E(\mathbf{w}) \doteq \|\|v_{vbw} - v_\pi\|\|^2_\mu$
$\bar{B}E(\mathbf{w})$	mean square Bellman error $\bar{B}E(\mathbf{w}) \doteq \|\|\bar{\delta}_{\mathbf{w}}\|\|^2_\mu$
$P\bar{B}E(\mathbf{w})$	mean square projected Bellman error $P\bar{B}E(\mathbf{w}) \doteq \|\|\bar{\delta}_{\mathbf{w}}\|\|^2_\mu$
$T\bar{D}E(\mathbf{w})$	mean square temporal difference error $T\bar{D}E(\mathbf{w}) \doteq \|\|\bar{\delta}_{\mathbf{w}}\|\|^2_\mu$
$\bar{R}E(\mathbf{w})$	mean square return error $\bar{R}E(\mathbf{w}) \doteq \|\|\bar{\delta}_{\mathbf{w}}\|\|^2_\mu$

1 Introduction

1.1 Recap

foo:

$$\mathbb{E}[X] = \sum_{x_i} x_i \cdot Pr\{X = x_i\} \tag{1}$$

foo:

$$\mathbb{E}[X|Y = y_j] = \sum_{x_i} x_i \cdot Pr\{X = x_i|Y = y_j\} \tag{2}$$

foo:

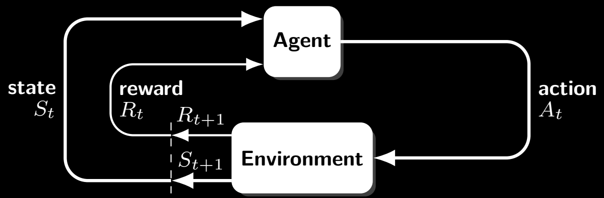
$$\mathbb{E}[X|Y = y_j] = \sum_{z_k} Pr\{Z = z_k|Y = y_j\} \cdot \mathbb{E}[X|Y = y_j, Z = z_k] \tag{3}$$

2 Multi-armed Bandits

$q_*(a) \quad \doteq \quad$ value of action
$\mathbb{E}[T_t|A_t = a]$

3 Finite Markov Decision Process

3.1 Agent-Environment Interaction



4 Dynamic Programming

5 Monte Carlo Methods

6 Temporal-Difference Learning

7 n-step Bootstrapping

8 Planning and Learning

9 On Policy Prediction with Approximation

10 On Policy Control with Approximation

11 Off Policy Prediction with Approximation

12 Eligibility Traces

13 Policy Gradient Methods

14 Psychology

15 Neuroscience

15.1 Neuroscience Basics

TODO

15.2 Reward Signals, Reinforcement Signals, Values and Prediction Errors

15.2.1 Reward Signals

In reinforcement learning reward **defines the problem a reinforcement learning agent is trying to solve**.

15.2.2 Reinforcement Signals

Reinforcement signals in reinforcement learning are are different from reward signals. **The function of a reinforcement signal is to direct the changes a learning algorithm makes in an agent's policy**, value estimates, or envirnment models. e.g.: TD error (TD method): $\delta_{t-1} = R_t + \gamma V(S_t) - V(S_{t-1})$

16 Application and case studies

17 Frontiers