

## Summary of Notation

Capital letters are used for random variables, whereas lower case letters are used for the values of random variables and for scalar functions. Quantities that are required to be real-valued vectors are written in bold and in lower case (even if random variables). Matrices are bold capitals.

$\doteq$	equality relationship that is true by definition
$\approx$	approximately equal
$\propto$	proportional to
$\Pr\{X=x\}$	probability that a random variable $X$ takes on the value $x$
$X \sim p$	random variable $X$ selected from distribution $p(x) \doteq \Pr\{X=x\}$
$\mathbb{E}[X]$	expectation of a random variable $X$ , i.e., $\mathbb{E}[X] \doteq \sum_x p(x)x$
$\arg \max_a f(a)$	a value of $a$ at which $f(a)$ takes its maximal value
$\ln x$	natural logarithm of $x$
$e^x$	the base of the natural logarithm, $e \approx 2.71828$ , carried to power $x$ ; $e^{\ln x} = x$
$\mathbb{R}$	set of real numbers
$f: \mathcal{X} \rightarrow \mathcal{Y}$	function $f$ from elements of set $\mathcal{X}$ to elements of set $\mathcal{Y}$
$\leftarrow$	assignment
$(a, b]$	the real interval between $a$ and $b$ including $b$ but not including $a$
$\varepsilon$	probability of taking a random action in an $\varepsilon$ -greedy policy
$\alpha, \beta$	step-size parameters
$\gamma$	discount-rate parameter
$\lambda$	decay-rate parameter for eligibility traces
$\mathbb{1}_{predicate}$	indicator function ( $\mathbb{1}_{predicate} \doteq 1$ if the <i>predicate</i> 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 of 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 $\pi_t$

In a Markov Decision Process:

$s, s'$	states
$a$	an action
$r$	a reward
$\mathcal{S}$	set of all nonterminal states
$\mathcal{S}^+$	set of all states, including the terminal state
$\mathcal{A}(s)$	set of all actions available in state $s$
$\mathcal{R}$	set of all possible rewards, a finite subset of $\mathbb{R}$
$\subset$	subset of; e.g., $\mathcal{R} \subset \mathbb{R}$

$\in$	is an element of; e.g., $s \in \mathcal{S}$ , $r \in \mathcal{R}$
$ \mathcal{S} $	number of elements in set $\mathcal{S}$
$t$	discrete time step
$T, T(t)$	final time step of an episode, or of the episode including time step $t$
$A_t$	action at time $t$
$S_t$	state at time $t$ , typically due, stochastically, to $S_{t-1}$ and $A_{t-1}$
$R_t$	reward at time $t$ , typically due, stochastically, to $S_{t-1}$ and $A_{t-1}$
$\pi$	policy (decision-making rule)
$\pi(s)$	action taken in state $s$ under <i>deterministic</i> policy $\pi$
$\pi(a s)$	probability of taking action $a$ in state $s$ under <i>stochastic</i> policy $\pi$
$G_t$	return following time $t$
$h$	horizon, the time step one looks up to in a forward view
$G_{t:t+n}, G_{t:h}$	$n$ -step 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 $h$
$G_t^\lambda$	$\lambda$ -return
$G_{t:h}^\lambda$	truncated, corrected $\lambda$ -return
$G_t^{\lambda s}, G_t^{\lambda a}$	$\lambda$ -return, corrected by estimated state, or action, values
$p(s', r   s, a)$	probability of transition to state $s'$ with reward $r$ , from state $s$ and action $a$
$p(s'   s, a)$	probability of transition to state $s'$ , from state $s$ taking action $a$
$r(s, a)$	expected immediate reward from state $s$ after action $a$
$r(s, a, s')$	expected immediate reward on transition from $s$ to $s'$ under action $a$
$v_\pi(s)$	value of state $s$ under policy $\pi$ (expected return)
$v_*(s)$	value of state $s$ under the optimal policy
$q_\pi(s, a)$	value of taking action $a$ in state $s$ under policy $\pi$
$q_*(s, a)$	value of taking action $a$ in state $s$ under the optimal policy
$V, V_t$	array estimates of state-value function $v_\pi$ or $v_*$
$Q, Q_t$	array estimates of action-value function $q_\pi$ or $q_*$
$\bar{V}_t(s)$	expected approximate action value, e.g., $\bar{V}_t(s) \doteq \sum_a \pi(a s) Q_t(s, a)$
$U_t$	target for estimate at time $t$
$\delta_t$	temporal-difference (TD) error at $t$ (a random variable)
$\delta_t^s, \delta_t^a$	state- and action-specific forms of the TD error
$n$	in $n$ -step methods, $n$ is the number of steps of bootstrapping
$d$	dimensionality—the number of components of $\mathbf{w}$
$d'$	alternate dimensionality—the number of components of $\boldsymbol{\theta}$
$\mathbf{w}, \mathbf{w}_t$	$d$ -vector of weights underlying an approximate value function
$w_i, w_{t,i}$	$i$ th component of learnable weight vector
$\hat{v}(s, \mathbf{w})$	approximate value of state $s$ given weight 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 $\mathbf{w}$
$\nabla \hat{v}(s, \mathbf{w})$	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 when in state $s$
$\mathbf{x}(s, a)$	vector of features visible when in state $s$ taking action $a$
$x_i(s), x_i(s, a)$	$i$ th component of vector $\mathbf{x}(s)$ or $\mathbf{x}(s, a)$
$\mathbf{x}_t$	shorthand for $\mathbf{x}(S_t)$ or $\mathbf{x}(S_t, A_t)$
$\mathbf{w}^\top \mathbf{x}$	inner product of vectors, $\mathbf{w}^\top \mathbf{x} \doteq \sum_i w_i x_i$ ; e.g., $\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^\top \mathbf{x}(s)$
$\mathbf{v}, \mathbf{v}_t$	secondary $d$ -vector of weights, used to learn $\mathbf{w}$
$\mathbf{z}_t$	$d$ -vector of eligibility traces at time $t$
$\boldsymbol{\theta}, \boldsymbol{\theta}_t$	parameter vector of target policy
$\pi(a s, \boldsymbol{\theta})$	probability of taking action $a$ in state $s$ given parameter vector $\boldsymbol{\theta}$
$\pi_{\boldsymbol{\theta}}$	policy corresponding to parameter $\boldsymbol{\theta}$
$\nabla \pi(a s, \boldsymbol{\theta})$	column vector of partial derivatives of $\pi(a s, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$
$J(\boldsymbol{\theta})$	performance measure for the policy $\pi_{\boldsymbol{\theta}}$
$\nabla J(\boldsymbol{\theta})$	column vector of partial derivatives of $J(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$
$h(s, a, \boldsymbol{\theta})$	preference for selecting action $a$ in state $s$ based on $\boldsymbol{\theta}$
$b(a s)$	behavior policy used to select actions while learning about target policy $\pi$
$b(s)$	a baseline function $b : \mathcal{S} \mapsto \mathbb{R}$ for policy-gradient methods
$b$	branching factor for an MDP or search tree
$\rho_{t:h}$	importance sampling ratio for time $t$ through time $h$
$\rho_t$	importance sampling ratio for time $t$ alone, $\rho_t \doteq \rho_{t:t}$
$r(\pi)$	average reward (reward rate) for policy $\pi$
$\bar{R}_t$	estimate of $r(\pi)$ at time $t$
$\mu(s)$	on-policy distribution over states
$\boldsymbol{\mu}$	$ \mathcal{S} $ -vector of the $\mu(s)$ for all $s \in \mathcal{S}$
$\ v\ _\mu^2$	$\mu$ -weighted squared norm of value function $v$ , i.e., $\ v\ _\mu^2 \doteq \sum_{s \in \mathcal{S}} \mu(s) v(s)^2$
$\eta(s)$	expected number of visits to state $s$ per episode
$\Pi$	projection operator for value functions
$B_\pi$	Bellman operator for value functions
$\mathbf{A}$	$d \times d$ matrix $\mathbf{A} \doteq \mathbb{E} \left[ \mathbf{x}_t (\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^\top \right]$
$\mathbf{b}$	$d$ -dimensional vector $\mathbf{b} \doteq \mathbb{E} [R_{t+1} \mathbf{x}_t]$
$\mathbf{w}_{\text{TD}}$	TD fixed point $\mathbf{w}_{\text{TD}} \doteq \mathbf{A}^{-1} \mathbf{b}$ (a $d$ -vector)
$\mathbf{I}$	identity matrix
$\mathbf{P}$	$ \mathcal{S}  \times  \mathcal{S} $ matrix of state-transition probabilities under $\pi$
$\mathbf{D}$	$ \mathcal{S}  \times  \mathcal{S} $ diagonal matrix with $\boldsymbol{\mu}$ on its diagonal
$\mathbf{X}$	$ \mathcal{S}  \times d$ matrix with the $\mathbf{x}(s)$ as its rows
$\bar{\delta}_{\mathbf{w}}(s)$	Bellman error (expected TD error) for $v_{\mathbf{w}}$ at state $s$
$\bar{\delta}_{\mathbf{w}}, \text{BE}$	Bellman error vector, with components $\bar{\delta}_{\mathbf{w}}(s)$

$\overline{\text{VE}}(\mathbf{w})$	mean square value error $\overline{\text{VE}}(\mathbf{w}) \doteq \ v_{\mathbf{w}} - v_{\pi}\ _{\mu}^2$
$\overline{\text{BE}}(\mathbf{w})$	mean square Bellman error $\overline{\text{BE}}(\mathbf{w}) \doteq \ \bar{\delta}_{\mathbf{w}}\ _{\mu}^2$
$\overline{\text{PBE}}(\mathbf{w})$	mean square projected Bellman error $\overline{\text{PBE}}(\mathbf{w}) \doteq \ \Pi \bar{\delta}_{\mathbf{w}}\ _{\mu}^2$
$\overline{\text{TDE}}(\mathbf{w})$	mean square temporal-difference error $\overline{\text{TDE}}(\mathbf{w}) \doteq \mathbb{E}_b[\rho_t \delta_t^2]$
$\overline{\text{RE}}(\mathbf{w})$	mean square return error