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.

`$\defeq$`	$\doteq$	equality relationship that is true by definition
`$\approx$`	$\approx$	approximately equal
`$\propto$`	$\propto$	proportional to
`$\Pr{X\!=\!x}$`	$\Pr\{X\!=\!x\}$	probability that a random variable X takes on the value x
`$X\sim p$`	$X \sim p$	random variable X selected from distribution $p(x) \doteq \Pr\{X$
`$\E{X}$`	$\mathbb{E}[X]$	expectation of a random variable X, i.e., $\mathbb{E}[X] \doteq \sum_x p(x)x$
`$\arg\max_a f(a)$`	$\arg\max_a f(a)$	a value of a at which $f(a)$ takes its maximal value
`$\ln x$`	$\ln x$	natural logarithm of x
`$e^x$`	e^x	the base of the natural logarithm, $e \approx 2.71828$, carried to p
`$\Re$`	$\mathbb{R}$	set of real numbers
`$f:\X\rightarrow\Y$`	$f: \mathcal{X} \rightarrow \mathcal{Y}$	function f from elements of set $\mathcal{X}$ to elements of set $\mathcal{Y}$
`$\leftarrow$`	$\leftarrow$	assignment
`$(a,b]$`	$(a, b]$	the real interval between a and b including b but not includ

`$\e$`	ε	probability of taking a random action in an ε-greedy policy
`$\alpha, \beta$`	α, β	step-size parameters
`$\gamma$`	γ	discount-rate parameter
`$\lambda$`	λ	decay-rate parameter for eligibility traces
`$\ind{\text{\emph{predicate}}}$`	$\mathbb{1}_{predicate}$	indicator function ($\mathbb{1}_{predicate} \doteq 1$ if the *predicate* is true, els

In a multi-arm bandit problem:

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

In a Markov Decision Process:

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

`$\in$`	$\in$	is an element of; e.g., $s \in \mathcal{S}$, $r \in \mathcal{R}$
`$\|\S\|$`	$\vert\mathcal{S}\vert$	number of elements in set $\mathcal{S}$
`$t$`	t	discrete time step
`$T, T(t)$`	$T, T(t)$	final time step of an episode, or of the episode including tin
`$A_t$`	A_t	action at time t
`$S_t$`	S_t	state at time t, typically due, stochastically, to S_{t-1} and A_t
`$R_t$`	R_t	reward at time t, typically due, stochastically, to S_{t-1} and
`$\pi$`	π	policy (decision-making rule)
`$\pi(s)$`	$\pi(s)$	action taken in state s under *deterministic* policy π
`$\pi(a\|s)$`	$\pi(a \vert s)$	probability of taking action a in state s under *stochastic* po
`$G_t$`	G_t	return following time t
`$h$`	h	horizon, the time step one looks up to in a forward view
`$G_{t:t+n}, G_{t:h}$`	$G_{t:t+n}, G_{t:h}$	n-step return from $t+1$ to $t+n$, or to h (discounted and c
`$\bar G_{t:h}$`	$\bar{G}_{t:h}$	flat return (undiscounted and uncorrected) from $t+1$ to h
`$G^\lambda_t$`	G^λ_t	λ-return
`$G^\lambda_{t:h}$`	$G^\lambda_{t:h}$	truncated, corrected λ-return
`$G^{\lambda s}_t$, $G^{\lambda a}_t$`	$G^{\lambda s}_t, G^{\lambda a}_t$	λ-return, corrected by estimated state, or action, values
`$\p(s',r\|s,a)$`	$p(s', r \mid s, a)$	probability of transition to state s' with reward r, from sta
`$\p(s'\|s,a)$`	$p(s' \mid s, a)$	probability of transition to state s', from state s taking acti
`$r(s,a)$`	$r(s, a)$	expected immediate reward from state s after action a
`$r(s,a,s')$`	$r(s, a, s')$	expected immediate reward on transition from s to s' under
`$\vpi(s)$`	$v_\pi(s)$	value of state s under policy π (expected return)
`$\vstar(s)$`	$v_*(s)$	value of state s under the optimal policy
`$\qpi(s,a)$`	$q_\pi(s, a)$	value of taking action a in state s under policy π
`$\qstar(s,a)$`	$q_*(s, a)$	value of taking action a in state s under the optimal policy
`$V, V_t$`	V, V_t	array estimates of state-value function v_π or v_*
`$Q, Q_t$`	Q, Q_t	array estimates of action-value function q_π or q_*
`$\bar V_t(s)$`	$\bar{V}_t(s)$	expected approximate action value, e.g., $\bar{V}_t(s) \doteq \sum_a \pi(a \vert s)$(
`$U_t$`	U_t	target for estimate at time t
`$\delta_t$`	δ_t	temporal-difference (TD) error at t (a random variable)
`$\delta^s_t, \delta^a_t$`	δ^s_t, δ^a_t	state- and action-specific forms of the TD error
`$n$`	n	in n-step methods, n is the number of steps of bootstrappin
`$d$`	d	dimensionality—the number of components of $\mathbf{w}$
`$d'$`	d'	alternate dimensionality—the number of components of $\boldsymbol{\theta}$
`$\w,\w_t$`	$\mathbf{w}, \mathbf{w}_t$	d-vector of weights underlying an approximate value functio
`\>$w_i,w_{t,i}$`	$w_i, w_{t,i}$	ith component of learnable weight vector
`$\hat v(s,\w)$`	$\hat{v}(s, \mathbf{w})$	approximate value of state s given weight vector $\mathbf{w}$
`$v_\w(s)$`	$v_{\mathbf{w}}(s)$	alternate notation for $\hat{v}(s, \mathbf{w})$

`$\hat q(s,a,\w)$`	$\hat{q}(s,a,\mathbf{w})$	approximate value of state–action pair s,a given weight vec
`$\grad \hat v(s,\w)$`	$\nabla\hat{v}(s,\mathbf{w})$	column vector of partial derivatives of $\hat{v}(s,\mathbf{w})$ with respect
`$\grad \hat q(s,a,\w)$`	$\nabla\hat{q}(s,a,\mathbf{w})$	column vector of partial derivatives of $\hat{q}(s,a,\mathbf{w})$ with respec
`$\x(s)$`	$\mathbf{x}(s)$	vector of features visible when in state s
`$\x(s,a)$`	$\mathbf{x}(s,a)$	vector of features visible when in state s taking action a
`$x_i(s), x_i(s,a)$`	$x_i(s), x_i(s,a)$	ith component of vector $\mathbf{x}(s)$ or $\mathbf{x}(s,a)$
`$\x_t$`	$\mathbf{x}_t$	shorthand for $\mathbf{x}(S_t)$ or $\mathbf{x}(S_t,A_t)$
`$\w\tr\x$`	$\mathbf{w}^\top\mathbf{x}$	inner product of vectors, $\mathbf{w}^\top\mathbf{x} \doteq \sum_i w_i x_i$; e.g., $\hat{v}(s,\mathbf{w}) \doteq \mathbf{w}$
`$\v,\v_t$`	$\mathbf{v}, \mathbf{v}_t$	secondary d-vector of weights, used to learn $\mathbf{w}$
`$\z_t$`	$\mathbf{z}_t$	d-vector of eligibility traces at time t
`$\th, \th_t$`	$\boldsymbol{\theta}, \boldsymbol{\theta}_t$	parameter vector of target policy
`$\pi(a\|s,\th)$`	$\pi(a \mid s,\boldsymbol{\theta})$	probability of taking action a in state s given parameter ve
`$\pi_\th$`	$\pi_{\boldsymbol{\theta}}$	policy corresponding to parameter $\boldsymbol{\theta}$
`$\grad\pi(a\|s,\th)$`	$\nabla\pi(a \mid s,\boldsymbol{\theta})$	column vector of partial derivatives of $\pi(a \mid s,\boldsymbol{\theta})$ with respec
`$J(\th)$`	$J(\boldsymbol{\theta})$	performance measure for the policy $\pi_{\boldsymbol{\theta}}$
`$\grad J(\th)$`	$\nabla J(\boldsymbol{\theta})$	column vector of partial derivatives of $J(\boldsymbol{\theta})$ with respect to
`$h(s,a,\th)$`	$h(s,a,\boldsymbol{\theta})$	preference for selecting action a in state s based on $\boldsymbol{\theta}$
`$b(a\|s)$`	$b(a \mid s)$	behavior policy used to select actions while learning about
`$b(s)$`	$b(s)$	a baseline function $b : \mathcal{S} \mapsto \mathbb{R}$ for policy-gradient methods
`$b$`	b	branching factor for an MDP or search tree
`$\rho_{t:h}$`	$\rho_{t:h}$	importance sampling ratio for time t through time h
`$\rho_{t}$`	ρ_t	importance sampling ratio for time t alone, $\rho_t \doteq \rho_{t:t}$
`$r(\pi)$`	$r(\pi)$	average reward (reward rate) for policy π
`$\bar R_t$`	$\bar{R}_t$	estimate of $r(\pi)$ at time t
`$\mu(s)$`	$\mu(s)$	on-policy distribution over states
`$\bm\mu$`	$\boldsymbol{\mu}$	$\lvert\mathcal{S}\rvert$-vector of the $\mu(s)$ for all $s \in \mathcal{S}$
`$\norm{v}$`	$\lVert v\rVert_\mu^2$	μ-weighted squared norm of value function v, i.e., $\lVert v\rVert_\mu^2 \doteq \sum$
`$\eta(s)$`	$\eta(s)$	expected number of visits to state s per episode
`$\Pi$`	Π	projection operator for value functions
`$B_\pi$`	B_π	Bellman operator for value functions
`${\bf A}$`	$\mathbf{A}$	$d \times d$ matrix $\mathbf{A} \doteq \mathbb{E}\Big[\mathbf{x}_t\big(\mathbf{x}_t - \gamma\mathbf{x}_{t+1}\big)^\top\Big]$
`${\bf b}$`	$\mathbf{b}$	d-dimensional vector $\mathbf{b} \doteq \mathbb{E}[R_{t+1}\mathbf{x}_t]$
`$\w_{\rm TD}$`	$\mathbf{w}_{\rm TD}$	TD fixed point $\mathbf{w}_{\rm TD} \doteq \mathbf{A}^{-1}\mathbf{b}$ (a d-vector
`${\bf I}$`	$\mathbf{I}$	identity matrix
`${\bf P}$`	$\mathbf{P}$	$\lvert\mathcal{S}\rvert \times \lvert\mathcal{S}\rvert$ matrix of state-transition probabilities under π
`${\bf D}$`	$\mathbf{D}$	$\lvert\mathcal{S}\rvert \times \lvert\mathcal{S}\rvert$ diagonal matrix with $\boldsymbol{\mu}$ on its diagonal
`${\bf X}$`	$\mathbf{X}$	$\lvert\mathcal{S}\rvert \times d$ matrix with the $\mathbf{x}(s)$ as its rows
`$\bar\delta_\w(s)$`	$\bar{\delta}_{\mathbf{w}}(s)$	Bellman error (expected TD error) for $v_{\mathbf{w}}$ at state s
`$\bar\delta_\w$, BE`	$\bar{\delta}_{\mathbf{w}}$, BE	Bellman error vector, with components $\bar{\delta}_{\mathbf{w}}(s)$

`$\MSVEm(\w)$`	$\overline{\mathrm{VE}}(\mathbf{w})$	mean square value error $\overline{\mathrm{VE}}(\mathbf{w}) \doteq \lVert v_{\mathbf{w}} - v_\pi \rVert_\mu^2$
`$\MSBEm(\w)$`	$\overline{\mathrm{BE}}(\mathbf{w})$	mean square Bellman error $\overline{\mathrm{BE}}(\mathbf{w}) \doteq \big\lVert \bar{\delta}_{\mathbf{w}} \big\rVert_\mu^2$
`$\MSPBEm(\w)$`	$\overline{\mathrm{PBE}}(\mathbf{w})$	mean square projected Bellman error $\overline{\mathrm{PBE}}(\mathbf{w}) \doteq \big\lVert \Pi\bar{\delta}_{\mathbf{w}} \big\rVert_\mu^2$
`$\MSTDEm(\w)$`	$\overline{\mathrm{TDE}}(\mathbf{w})$	mean square temporal-difference error $\overline{\mathrm{TDE}}(\mathbf{w}) \doteq \mathbb{E}_b\big[\rho_t\delta_t^2\big]$
`$\MSREm(\w)$`	$\overline{\mathrm{RE}}(\mathbf{w})$	mean square return error