数学符号

本书尽可能地减少了和数学相关的内容，以帮助读者更加直观地理解深度强化学习。本书的数学符号约定如下。

基础符号

x	scalar，标量
$\boldsymbol{x}$	vector，向量
$\boldsymbol{X}$	matrix，矩阵
$\mathbb{R}$	the set of real numbers，实数集
$\frac{\mathrm{d}y}{\mathrm{d}x}$	derivative of y with respect to x，标量的导数
$\frac{\partial y}{\partial x}$	partial derivative of y with respect to x，标量的偏导数
$\nabla_{\boldsymbol{x}} y$	gradient of y with respect to $\boldsymbol{x}$，向量的梯度
$\nabla_{\boldsymbol{X}} y$	matrix derivatives of y with respect to $\boldsymbol{X}$，矩阵的导数
$P(X)$	a probability distribution over a discrete variable，离散变量的概率分布
$p(X)$	a probability distribution over a continuous variable, or over a variable whose type has not been specified，连续变量（或者未定义连续或者离散的变量）的概率分布
$X \sim p$	the random variable X has distribution，随机变量 X 满足概率分布 p
$\mathbb{E}[X]$	expectation of a random variable，随机变量的期望
$\mathrm{Var}[X]$	variance of a random variable，随机变量的方差
$\mathrm{Cov}(X, Y)$	covariance of two random variables，两个随机变量的协方差

$D_{\mathrm{KL}}(P\|Q)$	Kullback-Leibler divergence of P and Q，两个概率分布的 KL 散度
$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$，平均值为 $\boldsymbol{\mu}$ 且协方差为 $\boldsymbol{\Sigma}$ 的多元高斯分布

强化学习符号

s, s'	states，状态
a	action，动作
r	reward，奖励
R	reward function，奖励函数
$\mathcal{S}$	set of all non-terminal states，非终结状态
$\mathcal{S}^+$	set of all states, including the terminal state，全部状态，包括终结状态
$\mathcal{A}$	set of actions，动作集合
$\mathcal{R}$	set of all possible rewards，奖励集合
$\boldsymbol{P}$	transition matrix，转移矩阵
t	discrete time step，离散时间步
T	final time step of an episode，回合内最终时间步
S_t	state at time t，时间 t 的状态
A_t	action at time t，时间 t 的动作
R_t	reward at time t, typically due, stochastically, to A_t and S_t，时间 t 的奖励，通常为随机量，且由 A_t 和 S_t 决定
G_t	return following time t，回报
$G_t^{(n)}$	n-step return following time t，n 步回报
G_t^{λ}	λ-return following time t，λ-回报
π	policy, decision-making rule，策略
$\pi(s)$	action taken in state s under *deterministic* policy π，根据确定性策略 π，状态 s 时的动作

$\pi(a\|s)$	probability of taking action a in state s under *stochastic* policy π，根据随机性策略 π，状态 s 时执行动作 a 的概率
$p(s', r\|s, a)$	probability of transitioning to state s', with reward r, from state s and action a，根据状态 s 和动作 a，使得状态转移成 s' 且获得奖励 r 的概率
$p(s'\|s, a)$	probability of transitioning to state s', from state s taking action a，根据状态 s 和动作 a，使得状态转移成 s' 的概率
$v_\pi(s)$	value of state s under policy π (expected return)，根据策略 π，状态 s 的价值（回报期望）
$v_*(s)$	value of state s under the optimal policy，根据最优策略，状态 s 的价值
$q_\pi(s, a)$	value of taking action a in state s under policy π，根据策略 π，在状态 s 时执行动作 a 的价值
$q_*(s, a)$	value of taking action a in state s under the optimal policy，根据最优策略，在状态 s 时执行动作 a 的价值
V, V_t	estimates of state-value function $v_\pi(s)$ or $v_*(s)$，状态价值函数的估计
Q, Q_t	estimates of action-value function $q_\pi(s, a)$ or $q_*(s, a)$，动作价值函数的估计
τ	trajectory, which is a sequence of states, actions and rewards, $\tau = (S_0, A_0, R_0, S_1, A_1, R_1, \cdots)$，状态、动作、奖励的轨迹
γ	reward discount factor, $\gamma \in [0, 1]$，奖励折扣因子
ϵ	probability of taking a random action in ϵ-greedy policy，根据 ϵ-贪婪策略，执行随机动作的概率
α, β	step-size parameters，步长
λ	decay-rate parameter for eligibility traces，资格迹的衰减速率

强化学习中术语总结

除了在本书开头的数学符号法则中定义的术语，强化学习中常见内容的相关术语总结如下：R 是奖励函数，$R_t = R(S_t)$ 是 MRP 中状态 S_t 的奖励，$R_t = R(S_t, A_t)$ 是 MDP 中的奖励，$S_t \in \mathcal{S}$。

$R(\tau)$ 是轨迹 τ 的 γ-折扣化回报，$R(\tau) = \sum_{t=0}^{\infty} \gamma^t R_t$。

$p(\tau)$ 是轨迹的概率：

- $p(\tau) = \rho_0(S_0)\prod_{t=0}^{T-1} p(S_{t+1}|S_t)$ 对于 MP 和 MRP，$\rho_0(S_0)$ 是起始状态分布（Start-State Distribution）。
- $p(\tau|\pi) = \rho_0(S_0)\prod_{t=0}^{T-1} p(S_{t+1}|S_t, A_t)\pi(A_t|S_t)$ 对于 MDP，$\rho_0(S_0)$ 是起始状态分布。

$J(\pi)$ 是策略 π 的期望回报，$J(\pi) = \int_\tau p(\tau|\pi)R(\tau) = \mathbb{E}_{\tau\sim\pi}[R(\tau)]$。

π^* 是最优策略：$\pi^* = \arg\max_\pi J(\pi)$。

$v_\pi(s)$ 是状态 s 在策略 π 下的价值（期望回报）。

$v_*(s)$ 是状态 s 在最优策略下的价值（期望回报）。

$q_\pi(s,a)$ 是状态 s 在策略 π 下采取动作 a 的价值（期望回报）。

$q_*(s,a)$ 是状态 s 在最优策略下采取动作 a 的价值（期望回报）。

$V(s)$ 是对 MRP 中从状态 s 开始的状态价值的估计。

$V^\pi(s)$ 是对 MDP 中在线状态价值函数的估计，给定策略 π，有期望回报：

- $V^\pi(s) \approx v_\pi(s) = \mathbb{E}_{\tau\sim\pi}[R(\tau)|S_0 = s]$

$Q^\pi(s,a)$ 是对 MDP 下在线动作价值函数的估计，给定策略 π，有期望回报：

- $Q^\pi(s,a) \approx q_\pi(s,a) = \mathbb{E}_{\tau\sim\pi}[R(\tau)|S_0 = s, A_0 = a]$

$V^*(s)$ 是对 MDP 下最优动作价值函数的估计，根据最优策略，有期望回报：

- $V^*(s) \approx v_*(s) = \max_\pi \mathbb{E}_{\tau\sim\pi}[R(\tau)|S_0 = s]$

$Q^*(s,a)$ 是对 MDP 下最优动作价值函数的估计，根据最优策略，有期望回报：

- $Q^*(s,a) \approx q_*(s,a) = \max_\pi \mathbb{E}_{\tau\sim\pi}[R(\tau)|S_0 = s, A_0 = a]$

$A^\pi(s,a)$ 是对状态 s 和动作 a 的优势估计函数：

- $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$

在线状态价值函数 $v_\pi(s)$ 和在线动作价值函数 $q_\pi(s,a)$ 的关系：

- $v_\pi(s) = \mathbb{E}_{a\sim\pi}[q_\pi(s,a)]$

最优状态价值函数 $v_*(s)$ 和最优动作价值函数 $q_*(s,a)$ 的关系：

- $v_*(s) = \max_a q_*(s,a)$

$a_*(s)$ 是状态 s 下根据最优动作价值函数得到的最优动作：

- $a_*(s) = \arg\max_a q_*(s,a)$

对于在线状态价值函数的贝尔曼方程：

- $v_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s),s'\sim p(\cdot|s,a)}[R(s,a) + \gamma v_\pi(s')]$

对于在线动作价值函数的贝尔曼方程：

- $q_\pi(s,a) = \mathbb{E}_{s'\sim p(\cdot|s,a)}[R(s,a) + \gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}[q_\pi(s',a')]]$

对于最优状态价值函数的贝尔曼方程：

- $v_*(s) = \max_a \mathbb{E}_{s' \sim p(\cdot|s,a)}[R(s,a) + \gamma v_*(s')]$

对于最优动作价值函数的贝尔曼方程：

- $q_*(s,a) = \mathbb{E}_{s' \sim p(\cdot|s,a)}[R(s,a) + \gamma \max_{a'} q_*(s',a')]$