00A0   2203∃ 2200∀ 2286⊆ 2713x 27FA⟺ 221A$\surd$ 221B$\sqrt[3]{}$ 2295⊕ 2297⊗

0.0.1

Richard S. Sutton Andrew G. Barto

2021 04 13

Contents

Reinforcement Learning: An Introduction   Reinforcement Learning

CHAPTER 1

OpenAI Spinning Up

CHAPTER 2

(Richard S. Sutton Andrew G. Barto)

-
-
-

-
-
-
-

CHAPTER 3

3.1

3.2

3.3

( )

$\doteq$	
$\approx$	
$\propto$	
$\Pr\{X = x\}$	X x
$X \sim p$	X $p(x) \doteq \Pr\{X = x\}$
$\mathbb{E}[X]$	X , $\mathbb{E}[X] = \sum_x p(x)x$
$\arg\max_a f(a)$	$f(a)$ a
$\ln(x)$	x
$e^x, exp(x)$	$e \approx 2.71828$ x $e^{\ln x} = x$
$\mathbb{R}$	
$f : \mathcal{X} \to \dagger$	f $\mathcal{X}$ $\dagger$
$\leftarrow$	
$(a, b]$	
ε	ε -
α, β	
γ	
λ	
$\nVdash_{predicate}$	($predicate$ $\nVdash_{predicate} \doteq 1,$ $0)$

:

k	( )
t	
$q_*(a)$	a ( )
$Q_t(a)$	$q_*(a)$ t
$N_t(a)$	t a
$H_t(a)$	t a
$\pi_t(a)$	t a
$\overline{R}_t$	π_t , t

:

s, s'	
a	
r	
$\mathcal{S}$	
$\mathcal{S}^+$	,
$\mathcal{A}(s)$	s
$\mathcal{R}$	, $\mathbb{R}$
$\subset$	, $\mathcal{R} \subset \mathbb{R}$
$\in$	, $s \in \mathcal{S}, r \in \mathcal{R}$
$\lvert\mathcal{S}\rvert$	$\mathcal{S}$
t	
$T, T(t)$	, t
A_t	t
S_t	t , S_{t-1} A_{t-1}
R_t	t , S_{t-1} A_{t-1}
π	( )
$\pi(s)$	π , s
$\pi(a\|s)$	π , s a
G_t	t
h	horizon, the time step one looks up to in a forward view
$G_{t:t+n}, G_{t:h}$	$t+1$ $t+n$ h n
$overline G_{t:h}$	$t+1$ h 5.8
G_t^λ	λ - 12.1
$G_{t:h}^\lambda$	:math:$lambda$ - 12.3
$G_t^{\lambda s}, G_t^{\lambda a}$	:math:$lambda$ - 12.8
$p(s', r\|s, a)$	s a , r s'
$p(s'\|s, a)$	s a, s'
$r(s, a)$	a s
$r(s, a, s')$	a s s'
$v_\pi(s)$	π s ( )
$v_*(s)$	s
$q_\pi(s, a)$	π , s a
$q_*(s, a)$	, s a

1 –

V, V_t	v_π v_*
Q, Q_t	q_π q_*
$\overline{V}_t(s)$	, $\overline{V}_t(s) \doteq \sum_a \pi(a\|s)Q_t(s,a)$
U_t	t
δ_t	t TD 6.1
δ_t^s, δ_t^a	TD 12.9
n	n n
d	$\mathbf{w}$
d'	θ
$\mathbf{w}, \mathbf{w}_t$	d
$w_i, w_{t,i}$	i
$\hat{v}(s,\mathbf{w})$	mathbf{w} s
$v_{\mathbf{w}}(s)$	$\hat{v}(s,\mathbf{w})$
$\hat{q}(s,a,\mathbf{w})$	- s, a $\mathbf{w}$
$\hat{\nabla}\hat{v}(s,\mathbf{w})$	$\mathbf{w}$ $\hat{v}(s,\mathbf{w})$
$\nabla\hat{q}(s,a,\mathbf{w})$	$\mathbf{w}$ $\hat{q}(s,a,\mathbf{w})$
$\mathbf{x}(s)$	s
$\mathbf{x}(s,a)$	s a
$x_i(s), x_i(s,a)$	$\mathbf{x}(s)$ $\mathbf{x}(s,a)$
$\mathbf{x}_t$	$\mathbf{x}(S_t)$ $\mathbf{x}(S_t, A_t)$
$\mathbf{W}^\top\mathbf{x}$	$\mathbf{w}^\top\mathbf{x} \doteq \sum_i w_i x_i$ $\hat{v}(s,\mathbf{w}) \doteq \mathbf{w}^\top\mathbf{x}(s)$
$\mathbf{V}, \mathbf{V}_t$	$\mathbf{w}$ d 11
$\mathbf{Z}_t$	t d 12
θ, θ_t	13
$\pi(a\|s,\theta)$	θ s a
π_θ	θ
$\nabla\pi(a\|s,\theta)$	θ $\pi(a\|s,\theta)$
$\mathbf{J}(\theta)$	
$\nabla\mathbf{J}(\theta)$	θ $\mathbf{J}(\theta)$
$h(s,a,\theta)$	θ s a
$b(a\|s)$	;math:pi
$b(s)$	$b : \mathcal{S} \mapsto \mathbb{R}$
b	MDP
$\rho_{t:h}$	t h 5.5
ρ_t	t $\rho_t \doteq \rho_{t:t}$
$r(\pi)$	π 10.3
$\overline{R}_t$	t $r(\pi)$
$\mu(s)$	9.2
μ	$s \in \mathcal{S}$ $\mu(s)$ $\|\mathcal{S}\|$
$\\|v\\|_\mu^2$	v μ $\\|v\\|_\mu^2 \doteq \sum_{s\in\mathcal{S}} \mu(s)v(s)^2$
$\eta(s)$	s 199
Π	268
B_π	Bellman 11.4

1 –

$\mathbf{A}$	$d \times d$ $\mathbf{A} \doteq \mathbb{E}\left[\mathbf{x}_t\left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right)^\top\right]$
$\mathbf{b}$	d $\mathbf{b} \doteq \mathbb{E}\left[R_{t+1}\mathbf{x}_t\right]$
$\mathbf{w}_{TD}$	TD $\mathbf{w}_{\mathrm{TD}} \doteq \mathbf{A}^{-1}\mathbf{b}$ 9.4
$\mathbf{I}$	
$\mathbf{P}$	π $\lvert\mathcal{S}\rvert \times \lvert\lvert mathcalS\rvert$
$\mathbf{D}$	μ $\lvert\mathcal{S}\rvert \times \lvert\lvert mathcalS\rvert$
$\mathbf{X}$	$\mathbf{x}(s)$ $\lvert\mathcal{S}\rvert \times d$
$\bar{\delta}_{\mathbf{w}}(s)$	s $v_{\mathbf{w}}$ Bellman TD 11.4
$\bar{\delta}_{\mathbf{w}}, \mathrm{BE}$	Bellman $\bar{\delta}_{\mathbf{w}}(s)$
$\overline{\mathrm{VE}}(\mathbf{w})$	$\overline{\mathrm{VE}}(\mathbf{w}) \doteq \lVert v_{\mathbf{w}} - v_\pi \rVert_\mu^2$ 9.2
$\overline{\mathrm{BE}}(\mathbf{w})$	Bellman $\overline{\mathrm{BE}}(\mathbf{w}) \doteq \lVert \bar{\delta}_{\mathbf{w}} \rVert_\mu^2$
$\overline{\mathrm{PBE}}(\mathbf{w})$	Bellman $\overline{\mathrm{PBE}}(\mathbf{w}) \doteq \left\lVert \Pi \bar{\delta}_{\mathbf{w}} \right\rVert_\mu^2$
$\overline{\mathrm{TDE}}(\mathbf{w})$	$\overline{\mathrm{TDE}}(\mathbf{w}) \doteq \mathbb{E}_b\left[\rho_t \delta_t^2\right]$ 11.5
$\overline{\mathrm{RE}}(\mathbf{w})$	11.6

3.4 1

[1]

3.4.1 1.1

- -

"ing"

3

Exploration Exploitation

goal-seeking

[1] 14 15

“ ”

14 15

20 60
“ ” “ ”

3.4.2 1.2

- - anticipating possible replies and counterreplies
- / /
- 20
-
- -

- -

3.4.3 1.3

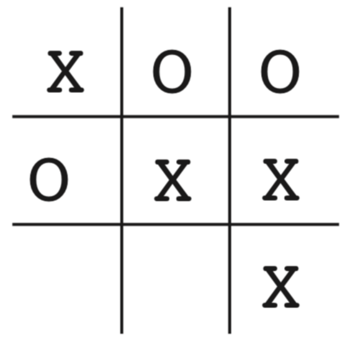
-

8

3.4.4 1.4

- “ ” 3
17.3

3.4.5 1.5



“ minimax ” “ ”

A B A B A B“ ”
1 0 0.5 50

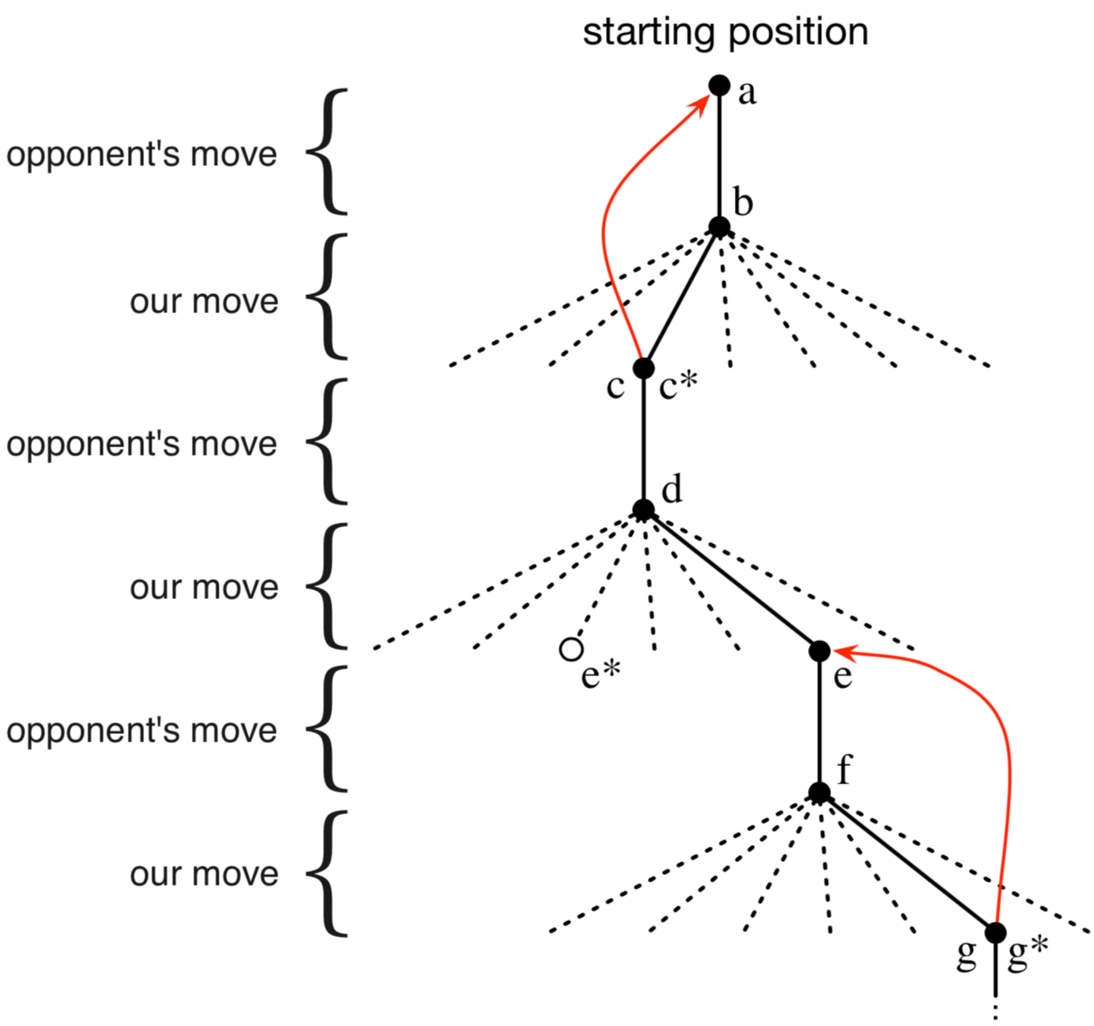
1.1

1.1 S_t
S_{t+1} S_t $V(S_t)$

$$V(S_t) \leftarrow V(S_t) + \alpha \left[V(S_{t+1}) - V(S_t) \right]$$

α $V(S_{t+1}) - V(S_t)$

“ ”


starting position
a
b
c
c*
d
e*
e
f
g
g*
opponent's move
our move
opponent's move
our move
opponent's move
our move

1: 1.1 e^* e^*

Gerry Tesauro 1992,1995 10^20
Tesauro 16.1
9.6
a tabula rasa view 9.5 17.4 13.1

8

“ ”

1.1

1.2

1.3

1.4

1.5

3.4.6 1.6

a sense of uncertainty and nondeterminism

3.4.7 1.7

20 80
20 80

“ ” 20 50 · Richard Bellman 20 50
19 Hamilton Jacobi “ ” Bellman Bellman 1957a
Bellman 1957b MDPs Ronald Howard 1960 MDP

“ ” 20 50 MDP Lovejoy 1991
White 1985,1988,1993 Rust 1996 Bertsekas 1982,1983
Bertsekas 2005 2012 Puterman 1994; Ross 1983 Whittle 1982 1983 Bryson 1996

Bellman
Bellman Dreyfus 1959 Witten 1977 Werbos 1987 1989 Chris Watkins MDP Dimitri Bertsekas John Tsitsiklis 1996 “ ”
“ ”

MDP

14.3 R. S. Woodworth 1938 19 50
· Alexander Bain “ ” Conway Lloyd Morgan 1894 Edward Thorndike

;
Thorndike 1911 244

Thorndike “ Law of Effect ” Thorndike Gallistel 2005 Herrnstein 1970 Kimble 1961,1967 Mazur 1994 Hilgard Bower 1975 Dennett 1978; Campbell 1960 Cziko 1995 Clark Hull 1943 1952 B. F. Skinner 1938

“ ” Thorndike 1927 - -

1948 · Alan Turing “ - ”
1948

Thomas Ross 1933 1951 W. Gray Walter “ ” Walter 1950 1952 Claude Shannon Theseus Shannon 1951 J. A. Deutsch 1954 Deutsch 1953 8 - Marvin Minsky 1954 SNARC - 15 cyberneticzoo.com

Farley Clark 1954 Clark and Farley 1955 Rosenblatt 1962 Widrow Hoff 1960 perceptual learning “ ”

20 60 70 20 60 “ ” “ ” Waltz Fu 1965 Mendel 1966 Fu 1970 Mendel and McClaren 1970 “ ” Minsky 1961 Minsky
20 60 70

John Andreae STeLLA “ ” Andreae 1963,1969a b Andreae 1977 “ ” Andreae 1998 Andreae 2017a b

Donald Michie 1961 1963 MENACE Matchbox Educable Naughts Crosses Engine naughts MENACE MENACE Michie Chambers 1968 GLEE BOXES BOXES Widrow Smith 1964 Michie Chambers Barto Sutton Anderson 1983 Sutton 1984 Michie Michie 1974

Widrow Gupta Maitra 1973 Widrow Hoff 1960 LMS “ ” “ ” “ ” Widrow “ ” Widrow Gupta Maitra Buchanan, Mitchell, Smith, and Johnson 1978 Dietterich Buchanan 1984

k k “ ” 2 20 60 M. L. Tsetlin 1973 Tsetlin Narendra Thathachar 1974,1989 Harth Tzanakou 1974 Alopex Barto Sutton Brouwer 1981 · William Estes 1950 Bush Mosteller 1955 Sternberg 1963

1973 Bush and Mosteller Cross 1973 Arthur 1991 Camerer 2011 Now ´e Vrancx De Hauwere 2012 Szita 2012

John Holland 1975 k 1976 1986 Holland “ - ” 6 Urbanowicz Moore 2009 Fogel Owens Walsh 1966 Koza 1992

· Harry Klopf 1972,1975,1982 Klopf K-lopf 15.9 Klopf Barto Sutton 1981a Barto Sutton Brouwer 1981 Barto Sutton 1981b Barto Anandan 1985 Barto Anderson Sutton 1982 Barto Anderson 1985 Barto 1985,1986; 1987 15.10 -

Minsky 1954 Arthur Samuel 1959 16.2

Samuel Claude Shannon 1950 Bellman Minsky 1961 “ ”

Minsky Samuel 1972 Klopf Klopf “ ” Samuel Klopf

Sutton 1978a b c Klopf Barto Sutton Barto 1981a Barto Sutton 1982 Klopf 1988 Moore 1986 Sutton Barto 1987,1990 Hawkins Kandel 1984 Byrne Gingrich Baxter 1990 Gelperin Hopfield Tank 1985 Tesauro 1986; Friston 1994

Klopf Minsky “ ” Samuel 1981 - Michie Chambers Barto Sutton Anderson 1983 Sutton 1984 Anderson 1986 Holland 1986 bucket-brigade Sutton 1988 TD(λ)

1981 - Ian Witten 1977,1976a TD(0) MDP 1974 Witten 1976 Witten Andreae STeLLA Witten 1977 - -

1989 Chris Watkins Q-learning Paul Werbos 1987 1977 Watkins 1992 Gerry Tesauro TD-Gammon

Friston 1994 Barto 1995a; Houk Adams Barto 1995 Montague Dayan Sejnowski 1996 Schultz Dayan Montague 1997 15

3.4.8

Szepesv ari 2010 Bertsekas Tsitsiklis 1996 Kaelbling 1993a Sugiyama Hachiya Morimura 2013 Si Barto Powell Wunsch 2004 Powell 2011 Lewis Liu 2012 Bertsekas 2012 Cao 2009 Sutton 1992a Kaelbling 1996 Singh 2002 Barto 1995b Kaelbling Littman Moore 1996 Keerthi Ravindran 1997 Weiring van Otterlo 2012

1.2 Agre 1988

1.5 6

3.5

- -

3.5.1 2

k

2.1 k

k 1000

k “ ” k
“ ”

k k t A_t R_t a $q_*(a)$ a

$$q_*(a) \doteq \mathbb{E}[R_t|A_t = a]$$

k t a $Q_t(a)$ $Q_t(a)$ $q_*(a)$

“ ”

k

k k

2.2

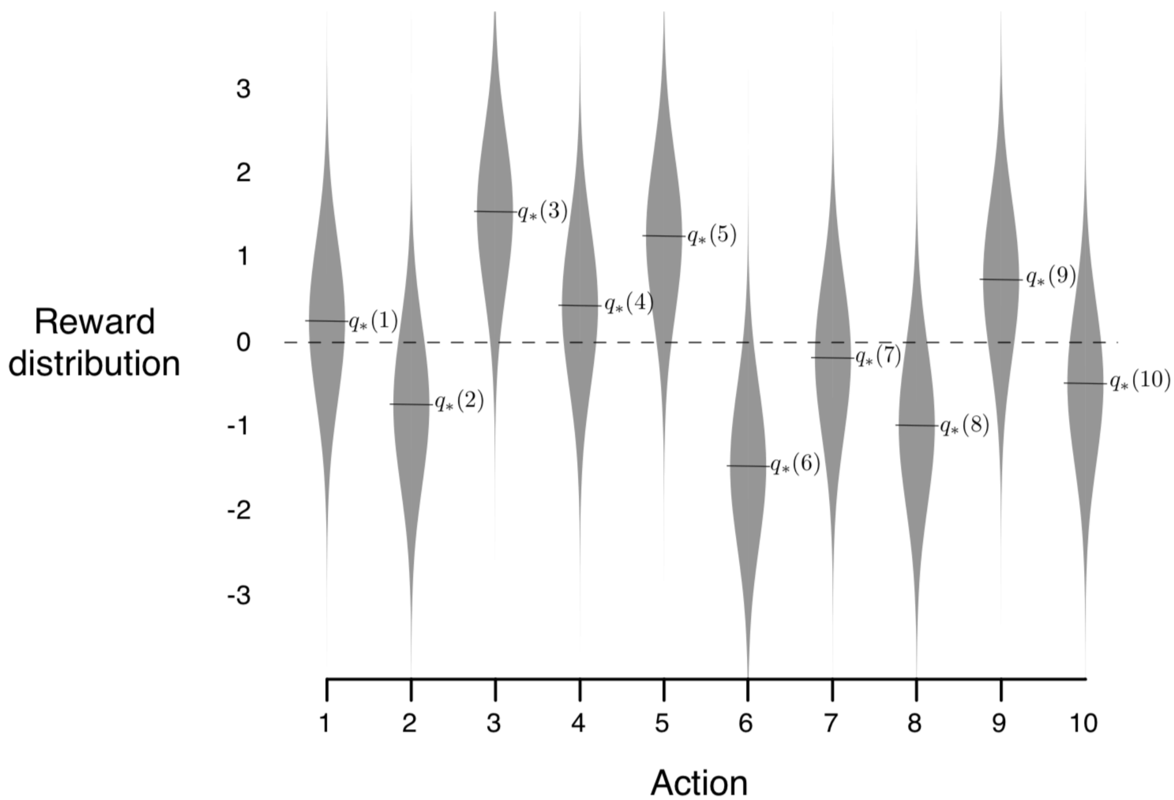
$$Q_t(a) \doteq \frac{t \quad a}{t \quad a} = \frac{\sum_{i=1}^{t-1} R_i \cdot ⊮_{A_i=a}}{\sum_{i=1}^{t-1} ⊮_{A_i=a}} \tag{3.1}$$

⊮ 1 0 $Q_t(a)$ 0 $Q_t(a)$ $q_*(a)$

$$A_t = \underset{a}{argmax}\, Q_t(a) \tag{3.2}$$

$argmax_a$ a ε
ε $Q_t(a)$ $q_*(a)$ $1 - \varepsilon$

2.1 ε $\varepsilon = 0.5$


Reward distribution
3
2
1
0
-1
-2
-3
$q_*(1)$
$q_*(2)$
$q_*(3)$
$q_*(4)$
$q_*(5)$
$q_*(6)$
$q_*(7)$
$q_*(8)$
$q_*(9)$
$q_*(10)$
1
2
3
4
5
6
7
8
9
10
Action

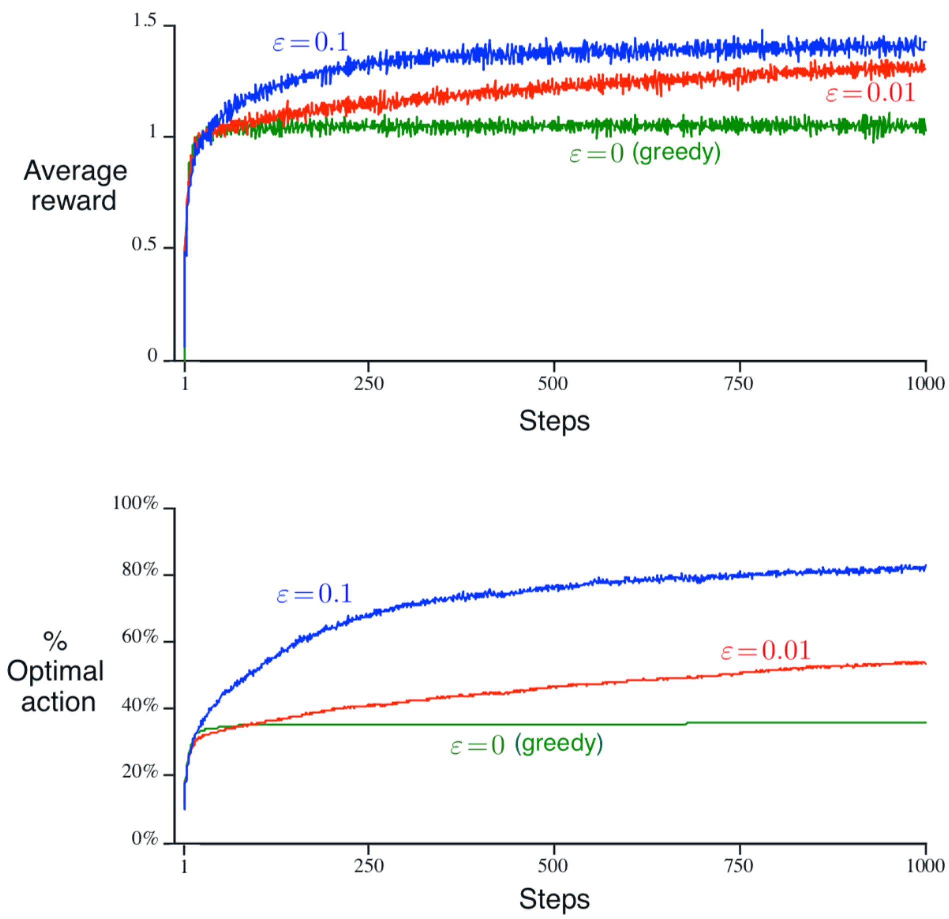
2: **2.1** 10 0 $q_*(a)$ $q_*(a)$

2.3 10

ε 2000 k $k = 10$ 2.1 $q_*(a), a = 1, \ldots, 10$ 0 1

A_t $q_*(A_t)$ 1 R_t 2.1 *10* 1000 2000

2.2 10 ε $\varepsilon = 0.01$ $\varepsilon = 0.1$ 1 1.55 ε $\varepsilon = 0.1$ 91 $\varepsilon = 0.01$ $\varepsilon = 0.1$ ε



3: **2.2** 10 ε 2000

ε 10 1 ε

2.2 $k = 4$ k 1 2 3 4 ε a $Q_1(a) = 0$ $A_1 = 1$ $R_1 = 1$ $A_2 = 2$ $R_2 = 1$ $A_3 = 2$ $R_3 = 2$ $A_4 = 2$ $R_4 = 2$ $A_5 = 3$ $R_5 = 0$ ε

2.3 2.2

2.4

R_i i Q_n

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

Q_n n R_n n

$$\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}(R_n + \sum_{i=1}^{n-1} R_i) \\
&= \frac{1}{n}(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i) \\
&= \frac{1}{n}(R_n + (n-1)Q_n) \\
&= \frac{1}{n}(R_n + nQ_n - Q_n) \\
&= Q_n + \frac{1}{n}(R_n - Q_n)
\end{aligned} \tag{3.3}$$

$n = 1$ Q_1 $Q_2 = R_1$ Q_n n 2.3 2.3

$$\leftarrow \quad + \quad [\quad - \quad] \tag{3.4}$$

[−] “ ” n

2.3 StepSize a n $\frac{1}{n}$ α $\alpha_t(a)$

ε $bandit(a)$

---

a 1 k

$$Q(a) \leftarrow 0$$
$$N(a) \leftarrow 0$$

$$A \leftarrow \begin{cases} argmax_a Q(a) & 1-\varepsilon \\ & \varepsilon \end{cases}$$
$$R \leftarrow bandit(a)$$
$$N(A) \leftarrow N(A) + 1$$
$$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$$

2.5

$n-1$ Q_n 2.3

$$Q_{n+1} \doteq Q_n + \alpha(R_n - Q_n) \tag{3.5}$$

$\alpha \in (0,1]$ Q_{n+1} Q_1

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha(R_n - Q_n) \\ &= \alpha R_n + (1-\alpha)Q_n \\ &= \alpha R_n + (1-\alpha)[\alpha R_{n-1} + (1-\alpha)Q_{n-1}] \\ &= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \\ &\qquad \cdots + (1-\alpha)^{n-1}\alpha R_1 + (1-\alpha)^n Q_1 \\ &= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i \end{aligned} \tag{3.6}$$

$(1-\alpha)^n + \sum_{i=1}^n \alpha(1-\alpha)^{n-i} = 1$ R_i $\alpha(1-\alpha)^{n-i}$ $n-i$ $1-\alpha$ 1 R_i $1-\alpha$ $1-\alpha=0$ R_n $0^0=1$

$\alpha_n(a)$ n a $\alpha_n(a) = \frac{1}{n}$ $\{\alpha_n(a)\}$ 1

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \quad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty \tag{3.7}$$

$\alpha_n(a) = \frac{1}{n}$ $\alpha_n(a) = n$ 2.7

2.4 $\alpha(a)$ Q_n 2.6 2.6

2.5 10 $q_*(a)$ $q_*(a)$ 0 0.01 $\alpha = 0.1$ 2.2 $\varepsilon = 0.1$ 10000

2.6

$Q_1(a)$ α 2.6

10 +5 $q_*(a)$ 0 1 +5 “ ”

2.3 a $Q_1(a) = +5$ 10 ε $Q_1(a) = 0$

2.6 2.3 2000 10

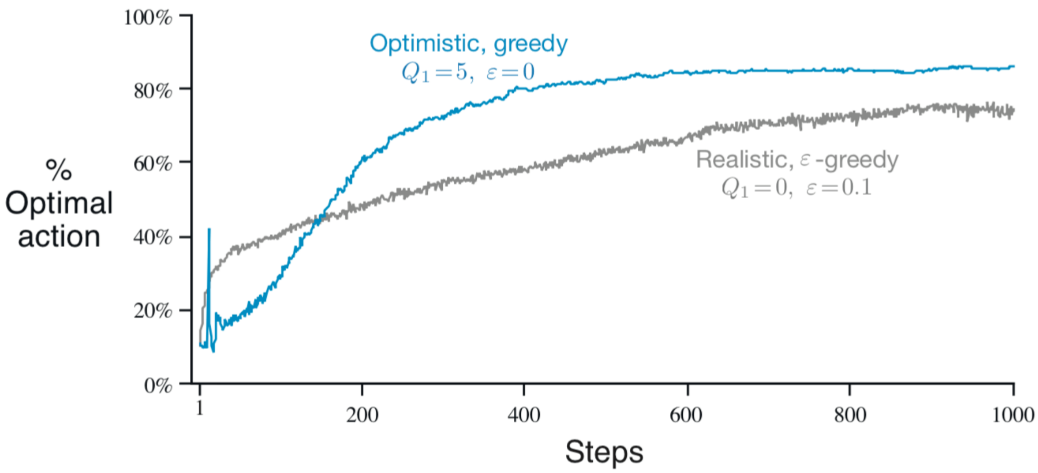
2.7 2.6

$$\beta_n \doteq \alpha / \bar{o}_n \tag{3.8}$$

n $\alpha > 0$ $\bar{o}_n$ 0

$$\bar{o}_n \doteq \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1}) for n \geq 0, with \bar{o}_0 \doteq 0 \tag{3.9}$$

2.6 Q_n



4: **2.3** 10 $alpha = 0.1$

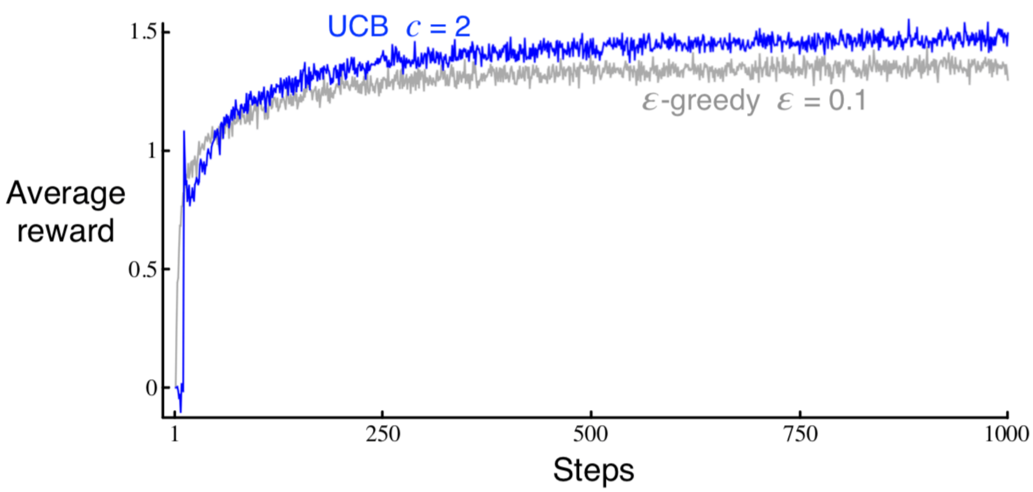
2.7

ε

$$A_t \doteq \underset{a}{argmax} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right] \tag{3.10}$$

$\ln t$ t $e \approx 2.71828$ t $N_t(a)$ t a 2.1 $c > 0$ $N_t(a) = 0$ a UCB a c a $N_t(a)$ a t $N_t(a)$ t

10 UCB 2.4 UCB ε UCB 2.5 UCB



5: **2.4** 10 UCB k UCB ε ”

2.8 USB 2.4 UCB 11 11 $c = 1$

2.8

a $H_t(a)$
1000 soft-max Gibbs Boltzmann

$$Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a) \tag{3.11}$$

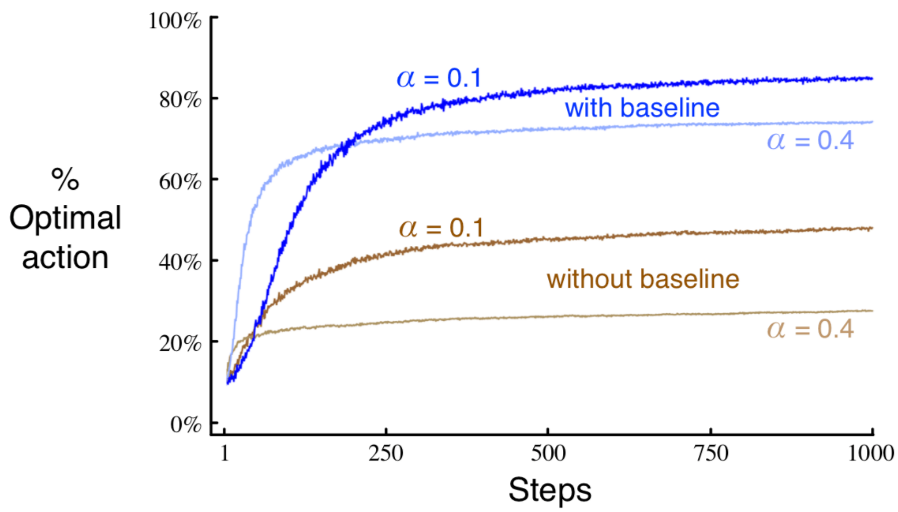
$\pi_t(a)$ t a $H_1(a) = 0$

2.9 soft-max sigmoid

A_t R_t

$$\begin{aligned} H_{t+1}(A_t) &\doteq H_t(A_t) + \alpha(R_t - \overline{R}_t)(1 - \pi_t(A_t)) \\ H_{t+1}(a) &\doteq H_t(a) - \alpha(R_t - \overline{R}_t)\pi_t(a) \qquad a \neq A_t \end{aligned} \tag{3.12}$$

$\alpha > 0$ $\overline{R}_t \in (R)$ t 2.4 2.5 $\overline{R}_t$ A_t ;
2.5 10 +4 $\overline{R}_t$
2.12



6: **2.5** $q_*(a)$ +4 10

$H_t(a)$

$$H_{t+1}(a) \doteq H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} \tag{3.13}$$

$$\mathbb{E}[R_t] = \sum_x \pi_t(x) q_*(x)$$

$q_*(x)$ 2.12 2.13

$$
\begin{aligned}
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)}\left[\sum_x \pi_t(x) q_*(x)\right] \\
&= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\
&= \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)}
\end{aligned}
$$

B_t x $\sum_x \frac{\partial \pi_t(x)}{\partial H_t(a)} = 0$ $H_t(a)$

$\pi_t(x)/\pi_t(x)$

$$
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \pi_t(x)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x)
$$

A_t x

$$
\begin{aligned}
&= \mathbb{E}\left[(q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t)\right] \\
&= \mathbb{E}\left[(R_t - \overline{R}_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t)\right]
\end{aligned}
$$

$B_t = \overline{R}_t$ R_t $q_*(A_t)$ $\mathbb{E}[R_t | A_t] = q_*(A_t)$ $\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x)(\nVdash_{a=A_t} - \pi_t(a))$ $a = x$ $\nVdash_{a=A_t}$ 1 0

$$
\begin{aligned}
&= \mathbb{E}\left[(R_t - \overline{R}_t)\pi_t(A_t)(\nVdash_{a=A_t} - \pi_t(a))/\pi_t(A_t)\right] \\
&= \mathbb{E}\left[(R_t - \overline{R}_t)(\nVdash_{a=A_t} - \pi_t(a))\right]
\end{aligned}
$$

2.13

$$
H_{t+1}(a) = H_t(a) + \alpha(R_t - \overline{R}_t)(\nVdash_{a=A_t} - \pi_t(a)) \quad a
$$

2.12

$\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x)(\nVdash_{a=A_t} - \pi_t(a))$

$$
\frac{\partial}{\partial x}\left[\frac{fx}{gx}\right] = \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x)\frac{\partial g(x)}{\partial x}}{g(x)^2}
$$

$$
\begin{aligned}
\frac{\partial \pi_t(x)}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \pi_t(x) \\
&= \frac{\partial}{\partial H_t(a)}\left[\frac{e^{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}}\right] \\
&= \frac{\frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_{y=1}^k e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_{y=1}^k e^{H_t(y)}}{\partial H_t(a)}}{(\sum_{y=1}^k e^{H_t(y)})^2} \\
&= \frac{\nVdash_{a=x} e_{H_t(x)} \sum_{y=1}^k e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{(\sum_{y=1}^k e^{H_t(y)})^2} \left(\quad \frac{\partial e^x}{\partial x} = e^x \right) \\
&= \frac{\nVdash_{a=x} e_{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}} - \frac{e^{H_t(x)} e^{H_t(a)}}{(\sum_{y=1}^k e^{H_t(y)})^2} \\
&= \nVdash_{a=x} \pi_t(x) - \pi_t(x)\pi_t(a) \\
&= \pi_t(x)(\nVdash_{a=x} - \pi_t(a)) \qquad\qquad Q.E.D.
\end{aligned}
$$

1000 2.5

2.9

k k

1 2

k k

2.10

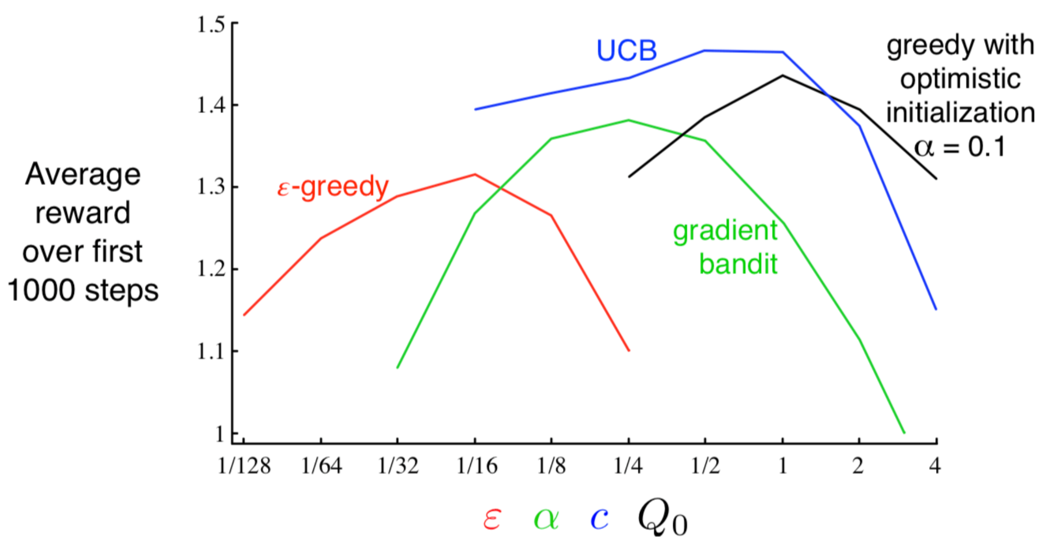
ε UCB favoring at soft-max

10

1000 2.6 x

2 U

UCB



7: **2.6** 1000

5

k *Gittins*

Gittins

1000 1000 2^2000

2.11 2.5 2.6 *varepsilon* $\alpha = 0.1$ 200,000 100,000

2.1 Thompson 1933,1934 Robbins 1952 Bellman 1956 “ ” Berry Fristedt 1985 Narendra Thathachar 1989 Bush Mosteller 1955 Estes 1950

Pearl 1984 Witten 1976b Feldbaum 1965 Holland 1975

2.2 k Thathachar Sastry 1985 Watkins 1989 “ε ” Watkins 1989 p.187

2.4-5 Bertsekas Tsitsiklis 1996

2.6 Sutton 1996

2.7 Lai Robbins 1985 Kaelbling 1993b Agrawal 1995 UCB UCB1 Auer Cesa-Bianchi Fischer 2002

2.8 Williams 1992 Balaraman Ravindran Greensmith Bartlett Baxter 2002,2004 Dick 2015 Sutton 1984 2.11 *soft-max* Bridle 1990 Luce 1959

2.9 Barto Sutton Brouwer 1981 Barto Anandan 1985 Sutton 1984 “ ” Thorndike 1 Skinner 1938

2.10 Bellman 1956 Gittins Gittins Jones 1974 Duff 1995 Gittins Kumar 1985 MDP Lovejoy 1991

Kakade 2003 Li 2012 Russo Van Roy Kazerouni Osband Wen 2018 Thompson

3.5.2 3

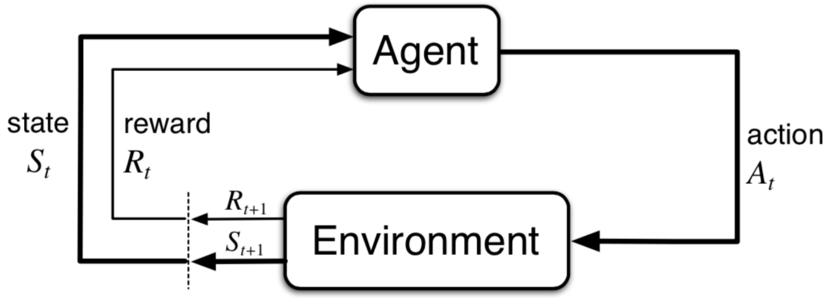
MDP MDP MDP a $q_*(a)$ MDP s a $q_*(s,a)$ $v_*(s)$

MDP Bellman MDP 17 MDP

3.1

MDP *agent* [1]

[1]



8: 3.1 -

$t = 0, 1, 2, 3, \ldots$[2] t $S_t \in \mathcal{S}$ $A_t \in \mathcal{S}(s)$[3] $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ S_{t+1}[4] MDP

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots \tag{3.14}$$

MDP $\mathcal{S}$ $\mathcal{A}$ $\mathcal{R}$ R_t S_t $s' \in \mathcal{S}$ $r \in \mathcal{R}$ t

$$p(s', r|s, a) \doteq Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\} \tag{3.15}$$

$s', s \in \mathcal{S}$ $\nabla \in \mathcal{R}$ $a \in \mathcal{A}(s)$ p MDP p $p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ "|" p s a

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) = 1 \quad s \in \mathcal{S} \; a \in \mathcal{A}(s) \tag{3.16}$$

p S_t R_t S_{t-1} A_{t-1} - 17

p $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$

$$p(s'|s, a) \doteq Pr\{S_t = s'|S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r|s, a) \tag{3.17}$$

- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$r(s, a) \doteq \mathbb{E}\left[R_t|S_{t-1} = s, A_{t-1} = a\right] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a) \tag{3.18}$$

- - $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

$$r(s, a, s') \doteq \mathbb{E}\left[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'\right] = \sum_{r \in \mathcal{R}} r \frac{p(s', r|s, a)}{p(s'|s, a)} \tag{3.19}$$

p 3.2

MDP

MDP

[2] Bertsekas Tsitsiklis 1996 Werbos 1992 Doya 1996
[3] R_{t+1} R_t A_t R_{t+1} S_{t+1}
[4] Lin 1992 Maclin Shavlik 1994 Clouse 1996

-

- -

MDP

3.1

3.2
+1 “ ”

3.1 MDP

3.2 MDP

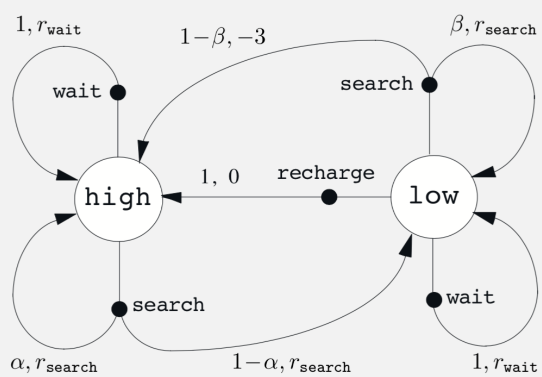
3.3

3.3

$\mathcal{S} = \{ \ \}$ 1 2 3
$\mathcal{A}(\ ) = \{\ ,\ \}$ $\mathcal{A}(\ ) = \{\ ,\ ,\ \}$

α $1-\alpha$ β $1-\beta$
-3 r r r $>$ r
MDP

s	a	s'	$p(s' \mid s,a)$	$r(s,a,s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-



s a $a \in \mathcal{A}(s)$ MDP
- s a s (s,a) (s,a) (s,s',a) s'
$p(s'|s,a)$ $r(s,a,s')$ 1

3.4 3.3 $p(s',r|s,a)$ s,a,s',r $p(s',r|s,a)$ $p(s',r|s,a) > 0$ 4

3.2

$R_t \in \mathbb{R}$

-1
+1 +1 -1 0
[5]
[6]

3.3

t $R_{t+1}, R_{t+2}, R_{t+3}, \dots$ G_t

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \tag{3.20}$$

T - [7]
$\mathcal{S}$
$\mathcal{S}^+$ T
- 3.7 $T =$
∞ +1
A_t

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{3.21}$$

γ $0 \leq \gamma \leq 1$
k γ^{k-1} $\gamma < 1$ $\{R_k\}$ 3.8 $\gamma = 0$ “ ”
A_t R_{t+1} 3.8 γ
1

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{3.22}$$

$t < T$ $t+1$ $G_T = 0$
3.8 $\gamma < 1$ +1

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma} \tag{3.23}$$

3.5 3.1 3.3

3.4

[5]
[6] 17.4
[7] “ ”

$+1$

$1 - \gamma^K$ [1] K K

3.5 -
1

3.7
$+1$
3.7

3.8 $\gamma = 0.5$ $R_1 =$
1 $R_2 = 2$ $R_3 = 6$ $R_4 = 3$ $R_5 = 2$ $T = 5$ G_0 G_1 $\dots$ G_5

3.9 $\gamma = 0.9$ $R_1 = 2$ 7s G_1 G_0

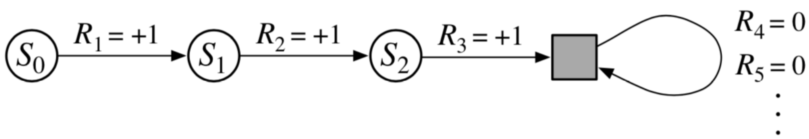
3.10 3.10

3.4

-

t S_t i t $S_{t,i}$ $A_{t,i}$ $R_{t,i}$ $\pi_{t,i}$
T_i S_t $S_{t,i}$

3.7 3.8



9: state transition diagram

S_0 +1 +1 +1 0 0 0 ... T $T = 3$
3.8 $\gamma = 1$

$$G_t \doteq \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \tag{3.24}$$

$T = \infty$ $\gamma = 1$ 10

3.5

- “ ”

t π $\pi(a|s)$ $S_t = s$ $A_t = a$ p π $\pi(a|s)$ “|” $s \in \mathcal{S}$ $a \in \mathcal{A}(s)$

3.11 S_t π π p 3.2 R_{t+1}

s π $v_\pi(s)$ s π **MDPs** $v_\pi(s)$

$$v_\pi(s) \doteq \mathbb{E}_\pi\left[G_t|S_t = s\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s\right] \quad s \in \mathbb{S} \tag{3.25}$$

$\mathbb{E}[\dot{]}$ π t 0 v_π π —

π s a $q_\pi(s,a)$ s ‘ a : *math* :pi‘

$$q_\pi(s,a) \doteq \mathbb{E}_\pi\left[G_t|S_t = s, A_t = a\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s, A_t = a\right] \tag{3.26}$$

q_π π

3.12 q_π π v_π *3.13* v_π p q_π

v_π q_π π $q_\pi(s,a)$
5 v_π q_π

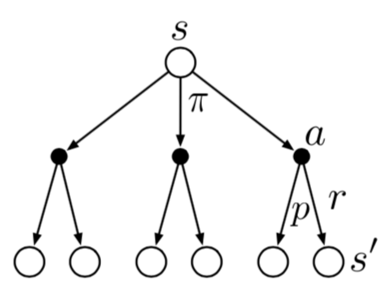
3.9 π s s

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] (\ (3.9)) \\
&= \sum_a \pi(a|s)\sum_{s'}\sum_r p(s',r|s,a)\left[r + \gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\right] \\
&= \sum_a \pi(a|s)\sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')], \quad s \in \mathcal{S}
\end{aligned} \tag{3.27}$$

a $\mathcal{A}(s)$ s' $\mathcal{S}$ $\mathcal{S}+$ r $\mathcal{R}$ s' r
a s' r $\pi(a|s)p(s',r|s,a)$

3.14 v_π - s
π s' r p
3.14

v_π v_π
— -



10: v_π

3.5 3.2 MDP
-1 A B
0 A +10 A' B +5 B'

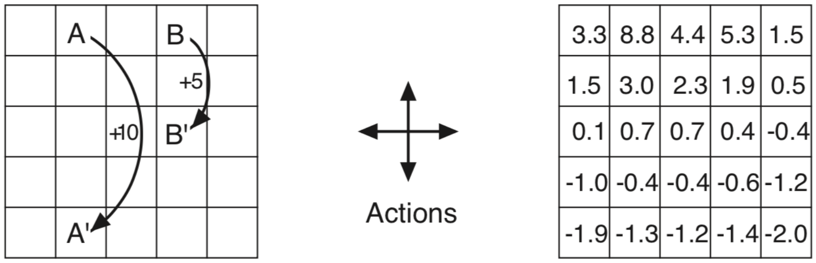
3.2 v_π $\gamma = 0.9$ 3.14
A 10 A A'
B 5 B B' B' A B

3.14 3.5 3.2 v_π 3.12
+2.3 +0.4 −0.4 +0.7 +0.7

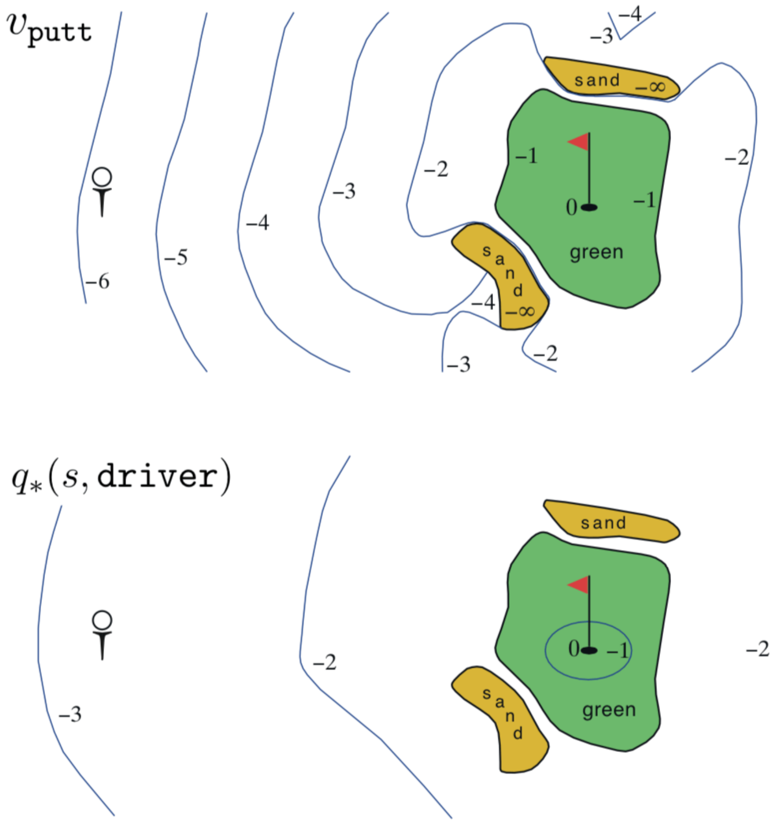
3.15 3.8 c v_c v_c c
γ

3.16 c

3.6 −1
3.3 $v_{putt}(s)$ 0 −1
−2 −2 −2
−3 $-\infty$

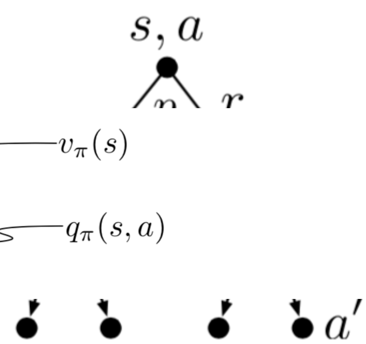

A
B
+5
+10
B'
A'
Actions
3.3 8.8 4.4 5.3 1.5
1.5 3.0 2.3 1.9 0.5
0.1 0.7 0.7 0.4 -0.4
-1.0 -0.4 -0.4 -0.6 -1.2
-1.9 -1.3 -1.2 -1.4 -2.0

11: **3.2**

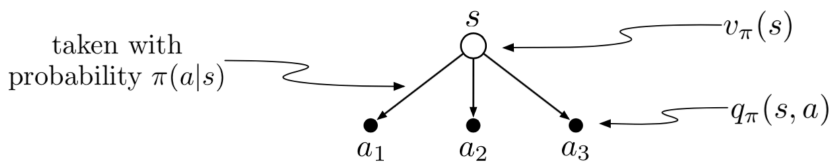

$v_{\texttt{putt}}$
sand
$-\infty$
green
-1
0
-2
-3
-4
-5
-6
$q_*(s, \texttt{driver})$
sand
green
0
-1
-2
-3

12: **3.3**

3.17 q_π — (s,a) $q_\pi(s',a')$ $q_\pi(s,a)$
3.14



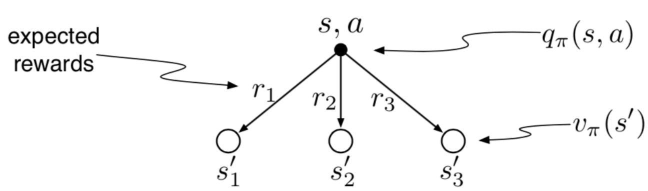
3.18



$S_t = s$ $q_\pi(s,a)$ $v_\pi(s)$ π
$\pi(a|s)$

13: q_π

3.19 $q_\pi(s,a)$ —



$St = s$ $A_t = a$ R_{t+1} $v_\pi(S_{t+1})$ $q_\pi(s,a)$
3.2 $p(s',r|s,a)$

3.6

MDP π
π' π π' $s \in \mathcal{S}$ $v_\pi(s) \geq v_{\pi'}(s)$ $\pi \geq \pi'$
π_* v_*

$$v_*(s) \doteq \max_\pi v_\pi(s) \tag{3.28}$$

$s \in \mathcal{S}$

q_*

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a) \tag{3.29}$$

$s \in \mathcal{S}$ $a \in \mathcal{A}(s)$ — (s,a) s a v_* q_*

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a\right] \tag{3.30}$$

3.17 3.3 $q_*(s, driver)$
$q_*(s, driver)$ -1 -2 -1
-3

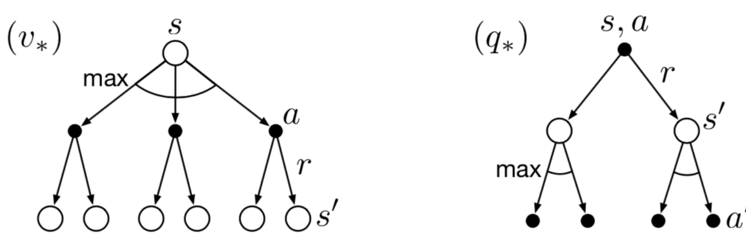
v_* 3.14 v_* v_*

$$
\begin{aligned}
v_*(s) &= \max_{a\in\mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a \mathbb{E}_{\pi_*}[G_t|S_t=s,A_t=a] \\
&= \max_a \mathbb{E}_{\pi_*}[R_{t+1}+\gamma G_{t+1}|S_t=s,A_t=a] \quad ((3.9)) \\
&= \max_a \mathbb{E}[R_{t+1}+\gamma v_*(S_{t+1})|S_t=s,A_t=a] \quad (3.18) \\
&= \max_{a\in\mathcal{A}(s)} \sum_{s',r} p(s',r|s,a)[r+\gamma v_*(s')] \quad (3.19)
\end{aligned}
$$

v_* q_*

$$
\begin{aligned}
q_*(s,a) &= \mathbb{E}\left[R_{t+1}+\gamma\sum_{a'} q_*(S_{t+1,a'})|S_t=s,A_t=a\right] \\
&= \sum_{s',r} p(s',r|s,a)[r+\gamma\max_{a'} q_*(s',a')]
\end{aligned} \quad (3.31)
$$

v_* q_* v_π q_π 3.4 3.19 3.20



14: **3.4** v_* q_*

MDP v_* 3.19 n n n p v_* q_*

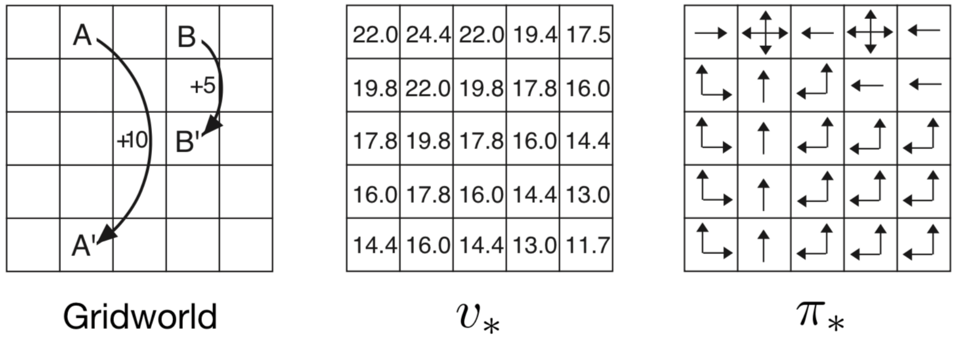
v_* v_* v_* v_* v_* v_*

q_* q_* s $q_*(s,a)$ — - -

3.8 3.5 v_* 3.5 A +10 A' B +5 B' 3.5 3.5

3.19 3.19 **h l s w re** $v_*(h)$

$$
\begin{aligned}
v_*(h) &= \max\left\{\begin{array}{l} p(h|h,s)[r(h,s,h)+\gamma v_*(h)]+p(l|h,s)[r(h,s,l)+\gamma v_*(l)], \\ p(h|h,w)[r(h,w,h)+\gamma v_*(h)]+p(l|h,w)[r(h,w,l)+\gamma v_*(l)] \end{array}\right\} \\
&= \max\left\{\begin{array}{l} \alpha[r_s+\gamma v_*(h)]+(1-\alpha)[r_s+\gamma v_*(l)], \\ l[r_w+\gamma v_*(h)]+0[r_w+\gamma v_*(l)] \end{array}\right\} \\
&= \max\left\{\begin{array}{l} r_s+\gamma[\alpha v_*(h)+(1-\alpha)v_*(l)], \\ r_w+\gamma v_*(h) \end{array}\right\}
\end{aligned}
$$



15: **3.5**

$v_*(l)$

$$v_*(l) = \max \begin{Bmatrix} \beta r_s - 3(1-\beta) + \gamma[(1-\beta)v_*(h) + \beta v_*(l)], \\ r_w + \gamma v_*(l), \\ \gamma v_*(h) \end{Bmatrix}$$

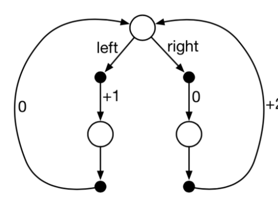
r_s r_w α β γ $0 \leq \gamma < 1$ $0 \leq \alpha, \beta \leq 1$ $v_*(h)$ $v_*(l)$

1 2 3 10^{20}
v_* q_*
3.19 “ ” “ ” v_* A^*

3.20

3.21 $q_*(s, putter)$

3.22 MDP π π $\gamma = 0$ $\gamma = 0.9$ $\gamma = 0.5$



3.23 q_*

3.24 3.5 24.4 3.8

3.25 q_* v_*

3.26 v_* p

3.7

-

Tesauro TD-Gammon *TD-Gammon*
MDP

3.8

MDP MDP MDP MDP

—

- - - - MDP

MDP p 3.2

MDP 1 MDP MDP
MDP MDP

MDP Bertsekas 2005 White 1969 Whittle 1982 1983 Puterman 1994 Ross 1983 MDP
Kumar 1985 Kumar Varaiya 1986

MDP Thompson 1933,1934 Robbins 1952 2

MDP Andreae 1969b Witten Corbin 1973 MDP Witten 1977,1976a
MDP Werbos 1977 Werbos 1982,1987,1988,1989,1992 Werbos MDP Watkins 1989

3.1 $p(s', r|s, a)$ MDP MDP $p(s'|s, a)$ $r(s, a)$
S_t R_t

Minsky 1967

Ungar 1990 Miller and Williams 1992 Jonathan Connell 1989
Kober Peters 2012

3.2 Michael Littman

3.3-4 MDP 1 2 3

Michie Chambers 1968 Barto Sutton Anderson 1983

3.5-6 20 50 19 Schultz Melsa 1967
Shannon 1950

Watkins 1989 q_* Q 6 - "Q " - Shannon 1950 $h(P, M)$
P M Michie 1961,1963 MENACE Michie and Chambers 1968 BOXES
Goldstein 1957 Denardo 1967 v_* Richard
Bellman 1957a " " Hamilton-Jacobi-Bellman Hamilton-Jacobiequation
Schultz Melsa 1967

3.5.3 4

DP MDP DP
DP DP

$\mathcal{S}$ $\mathcal{A}(\int)$ $\mathcal{R}$ $p(s', r|s, a)$ $\int \in \mathcal{S}$ $a \in \mathcal{A}(\int)$ $r \in \mathcal{R}$ $\mathcal{S}' \in \mathcal{S}^+$ ($\mathcal{S}^+$ $\mathcal{S}$) DP DP II

DP DP 3 3 v_* q_*

$$
\begin{aligned}
v_*(s) &= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1})|S_t = s, A_t = a] \\
&= \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')]
\end{aligned}
\tag{3.32}
$$

$$
\begin{aligned}
q_*(s, a) &= \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a')|S_t = s, A_t = a] \\
&= \sum_{s',r} p(s', r|s, a)[r + \gamma \max_{a'} q_*(s', a')],
\end{aligned}
\tag{3.33}
$$

$s \in \mathcal{S}$ $a \in \mathcal{A}(s)$ $s' \in \mathcal{S}^+$ DP

4.1

π v_π DP $s \in \mathcal{S}$

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] && \text{((3.9))} \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] && (4.3) \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')] && (4.4)
\end{aligned}
$$

$\pi(a|s)$ s π a π π $\gamma < 1$ $\pi(a|s)$ v_π

(4.4) $|\mathcal{S}|$ $|\mathcal{S}|$ $v_\pi(s), s \in \mathcal{S}$ $v_0, v_1, v_2, ...$ S^+ $\mathbb{R}$ v_0 0 v_π(3.12)

$$
\begin{aligned}
v_{k+1}(s) &\doteq \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1})|S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_k(s')],
\end{aligned}
\tag{3.34}
$$

$s \in \mathcal{S}$ $v_k = v_\pi$ v_π v_π $k \to \infty$ $\{v_k\}$ v_π

v_k v_{k+1} s s s v_{k+1} - DP " v_π "

(4.5) $v_k(s)$ $v_{k+1}(s)$ (4.5) v_π DP

$\max_{s \in \mathcal{S}} |v_{k+1}(s) - v_k(s)|$

$V \approx v_\pi$

π

$\theta > 0$

$V(s) = 0$ $s \in \mathcal{S}^+$ $V(\ ) = 0$

$\Delta \leftarrow 0$

$s \in \mathcal{S}$:

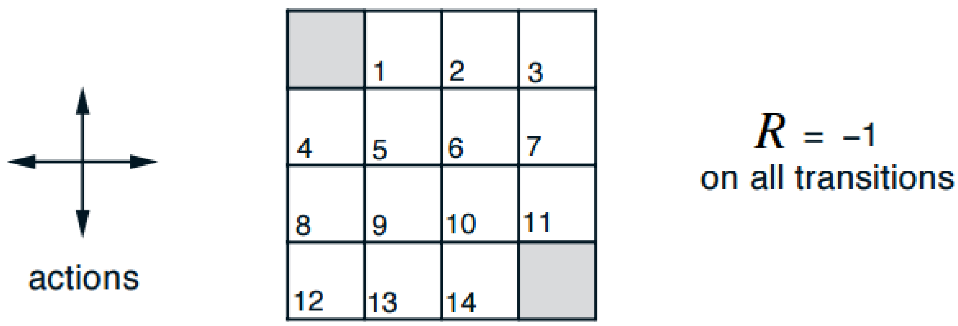
$v \leftarrow V(s)$

$V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

$\Delta < \theta$

4.1 4×4



$S = \{1, 2, ...14\}$ $\mathcal{A} =$, , , $r \in \mathcal{R}$ $p(6, -1|5,) = 1$ $p(7, -1|7,) = 1$ $p(10, r|5,) = 0$ -1 s s' a $r(s, a, s') = -1$ 4.1 $\{v_k\}$ v_π

4.1 4.1 π $q_\pi(11,)$ $q_\pi(7,)$

4.2 4.1 15 13 12 13 14 15 $v_\pi(15)$ 13 13 15 $v_\pi(15)$

4.3 - q_π $q_0, q_1, q_2, \ldots$ q_π (4.3) (4.4) (4.5)

4.2

π v_π s $a \neq \pi(s)$ s —— $v_\pi(s)$—— s a π

$$q_\pi(s,a) \doteq \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t = a] \\ = \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')] \tag{3.35}$$

v_π —— s a π π —— s a

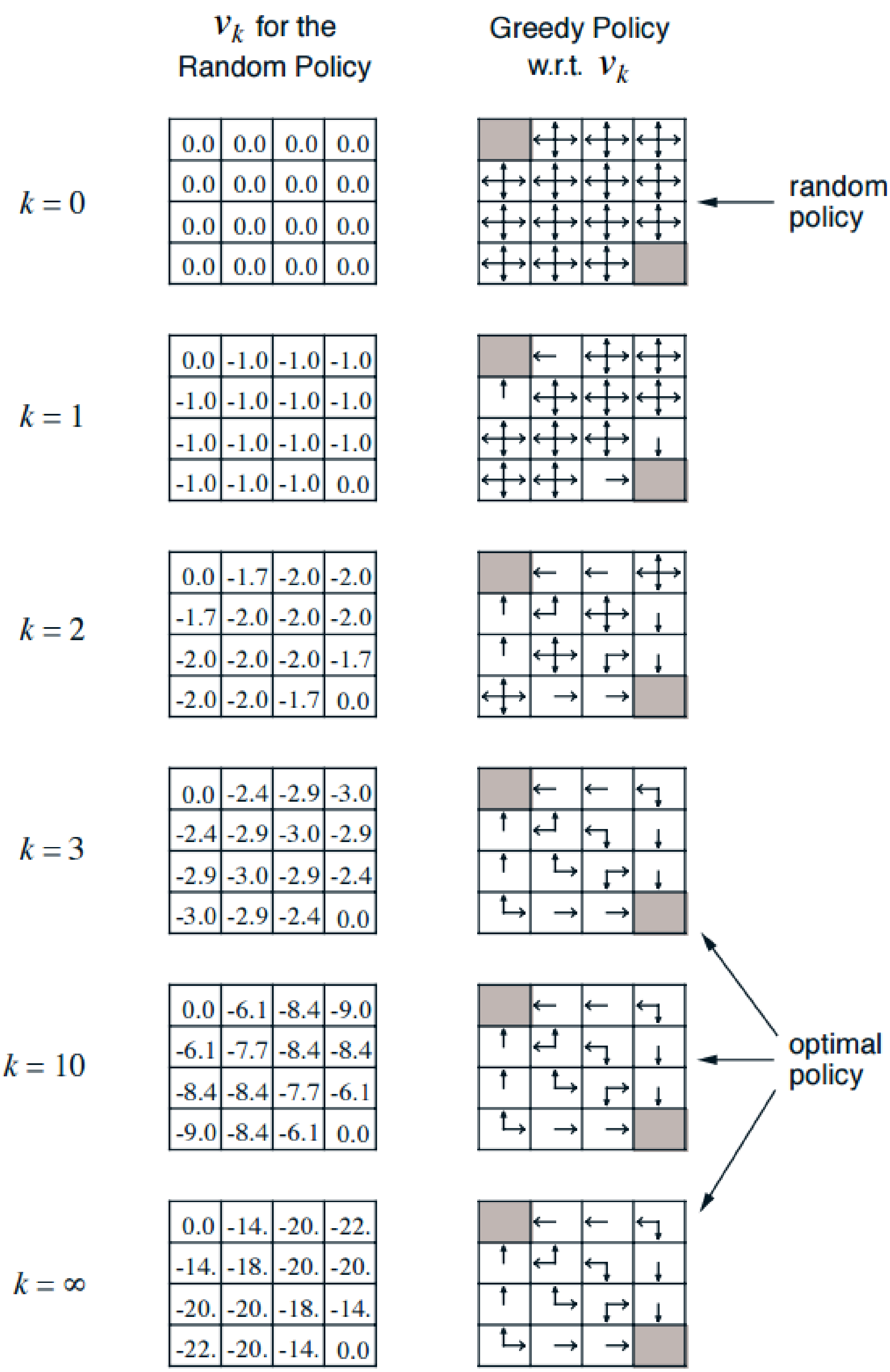
π π' $s \in \mathcal{S}$

$$q_\pi(s, \pi'(s)) \geq v_\pi(s) \tag{3.36}$$

π' π π $s \in \mathcal{S}$:

$$v'_\pi(s) \geq v_\pi(s) \tag{3.37}$$

(4.7) (4.8) π π' $\pi'(s) = a \neq \pi(s)$ π (4.7) s $q_\pi(s,a) > v_\pi(s)$ π


v_k for the Random Policy
Greedy Policy w.r.t. v_k
$k = 0$
0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0
random policy
$k = 1$
0.0 -1.0 -1.0 -1.0
-1.0 -1.0 -1.0 -1.0
-1.0 -1.0 -1.0 -1.0
-1.0 -1.0 -1.0 0.0
$k = 2$
0.0 -1.7 -2.0 -2.0
-1.7 -2.0 -2.0 -2.0
-2.0 -2.0 -2.0 -1.7
-2.0 -2.0 -1.7 0.0
$k = 3$
0.0 -2.4 -2.9 -3.0
-2.4 -2.9 -3.0 -2.9
-2.9 -3.0 -2.9 -2.4
-3.0 -2.9 -2.4 0.0
$k = 10$
0.0 -6.1 -8.4 -9.0
-6.1 -7.7 -8.4 -8.4
-8.4 -8.4 -7.7 -6.1
-9.0 -8.4 -6.1 0.0
optimal policy
$k = \infty$
0.0 -14. -20. -22.
-14. -18. -20. -20.
-20. -20. -18. -14.
-22. -20. -14. 0.0

: 4.1

(4.7) q_π (4.7) $v'_\pi(s)$

$$
\begin{aligned}
v_\pi(s) &\le q_\pi(s, \pi'(s)) \\
&= \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t = \pi'(s)] && (\ (4.6)) \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s] \\
&\le \mathbb{E}_{\pi'}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) | S_t = s] && (\ (4.7)) \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma \mathbb{E}_{\pi'}[R_{t+2} + \gamma v_\pi(S_{t+2}) | S_{t+1}, A_t = \pi'(s+1)] | S_t = s] \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) | S_t = s] \\
&\le \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+3}) | S_t = s] \\
&\vdots \\
&\le \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots | S_t = s] \\
&= v_{\pi'}(s)
\end{aligned}
$$

$q_\pi(s,a)$ π'

$$
\begin{aligned}
\pi'(s) &\doteq \arg\max_a q_\pi(s,a) \\
&= \arg\max_a \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t = a] \\
&= \arg\max_a \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')],
\end{aligned}
\tag{3.38}
$$

$\arg\max_a$ a —— —— v_π (4.7)

π' π $v_\pi = v_{\pi'}$ (4.9) $s \in \mathcal{S}$

$$
\begin{aligned}
v_{\pi'}(s) &= \max_a \mathbb{E}[R_{t+1} + \gamma v_{\pi'}(S_{t+1}) | S_t = s, A_t = a] \\
&= \max_a \sum_{s',r} p(s', r | s, a)[r + \gamma v_{\pi'}(s')].
\end{aligned}
$$

(4.1) $v_{\pi'}$ v_* π π'

π s a $\pi(a|s)$
(4.9) —— ——

4.1 π π' v_π v_π π' π' (4.9)
$v_{\pi'}(s)$ $s \in \mathcal{S}$ -1,-2 -3 $v_\pi(s)$ -14 $s \in \mathcal{S}$ $v_{\pi'}(s) \ge v_\pi(s)$
π'

4.3

π v_π π' $v_{\pi'}$ π''

$$
\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} v_*
$$

$\xrightarrow{E}$ $\xrightarrow{I}$ MDP

$\pi \approx \pi_*$

1.

$s \in \mathcal{S}$ $V(s) \in \mathbb{R}$ $\pi(s) \in \mathcal{A}(\int)$

2.

$\Delta \leftarrow 0$

$s \in S$:

$v \leftarrow V(s)$

$V(s) \leftarrow \sum_{s',r} p(s',r|s,\pi(s))[r+\gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

$\Delta < \theta$ ( )

3.

- $\leftarrow$

$s \in \mathcal{S}$:

$\leftarrow \pi(s)$

$\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s',r|s,a)[r+\gamma V(s')]$

$\neq \pi(s)$ - $\leftarrow$

- $V \approx v_*$ $\pi \approx \pi_*$ 2

4.2: 10 2 n $\frac{\lambda^n}{n!}e^{-\lambda}$ λ λ 3 4 3 2 20 $\gamma = 0.9$ MDP 4.2

4.1 4.1

4.4

4.7 2 10 4

4.5 q_* v_*

4.7 $\epsilon - soft$ s $\frac{\epsilon}{|\mathcal{A}(\int)|}$ v_* 3 2 1

4.4

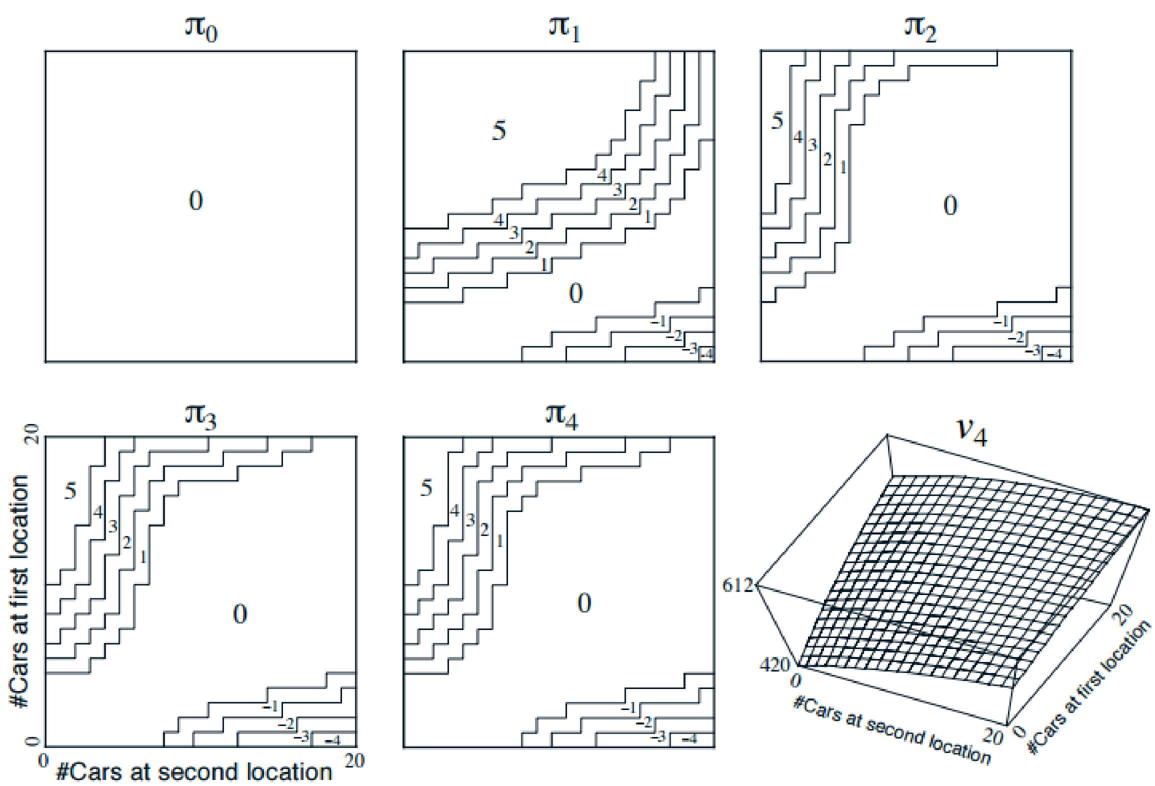
v_π 4.1

$$v_{k+1}(s) \doteq \max_a \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s, A_t = a]$$
$$= \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v_k(s')], \quad (3.39)$$

$s \in \mathcal{S}$ v_0 v_* :math:$\{v\_k\}$ v_*

(4.1) (4.5) 3.4 v_π v_*

v_*



π_0
0
π_1
5
4
3
2
1
0
−1
−2
−3
−4
π_2
5
4
3
2
1
0
−1
−2
−3
−4
π_3
20
#Cars at first location
0
5
4
3
2
1
0
−1
−2
−3
−4
0
#Cars at second location
20
π_4
5
4
3
2
1
0
−1
−2
−3
−4
v_4
612
420
0
#Cars at second location
20
0
#Cars at first location
20

17: 4.2:

$\pi \approx \pi_*$

$\theta > 0$

$s \in \mathcal{S}$ $V(s)$ $V(\ ) = 0$

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

$\Delta < \theta$ ( )

$\pi \approx \pi_*$ $\pi(s) = \arg\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$

4.10 MDP

4.3 100 MDP $s \in \{1, 2, \ldots, 99\}$ $a \in \{0, 1, \ldots, \min(s, 100 - s)\}$ +1 0 p_h p_h 4.3 $p_h = 0.4$ argmax

4.8 50 51

4.9 $p_h = 0.25$ $p_h = 0.55$ 0 100 0 1 4.3 $\theta \leftarrow 0$

4.10 4.10 $q_{k+1}(s, a)$

4.5

DP MDP 10^{20}

DP DP D-P

4.10 k s_k $0 \leq \gamma < 1$ $\{s_k\}$ v_* DP DP

MDP *MDP* DP DP DP DP

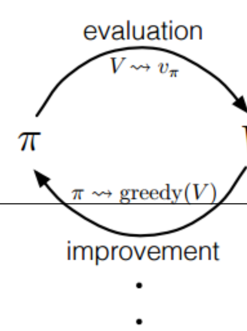
4.6

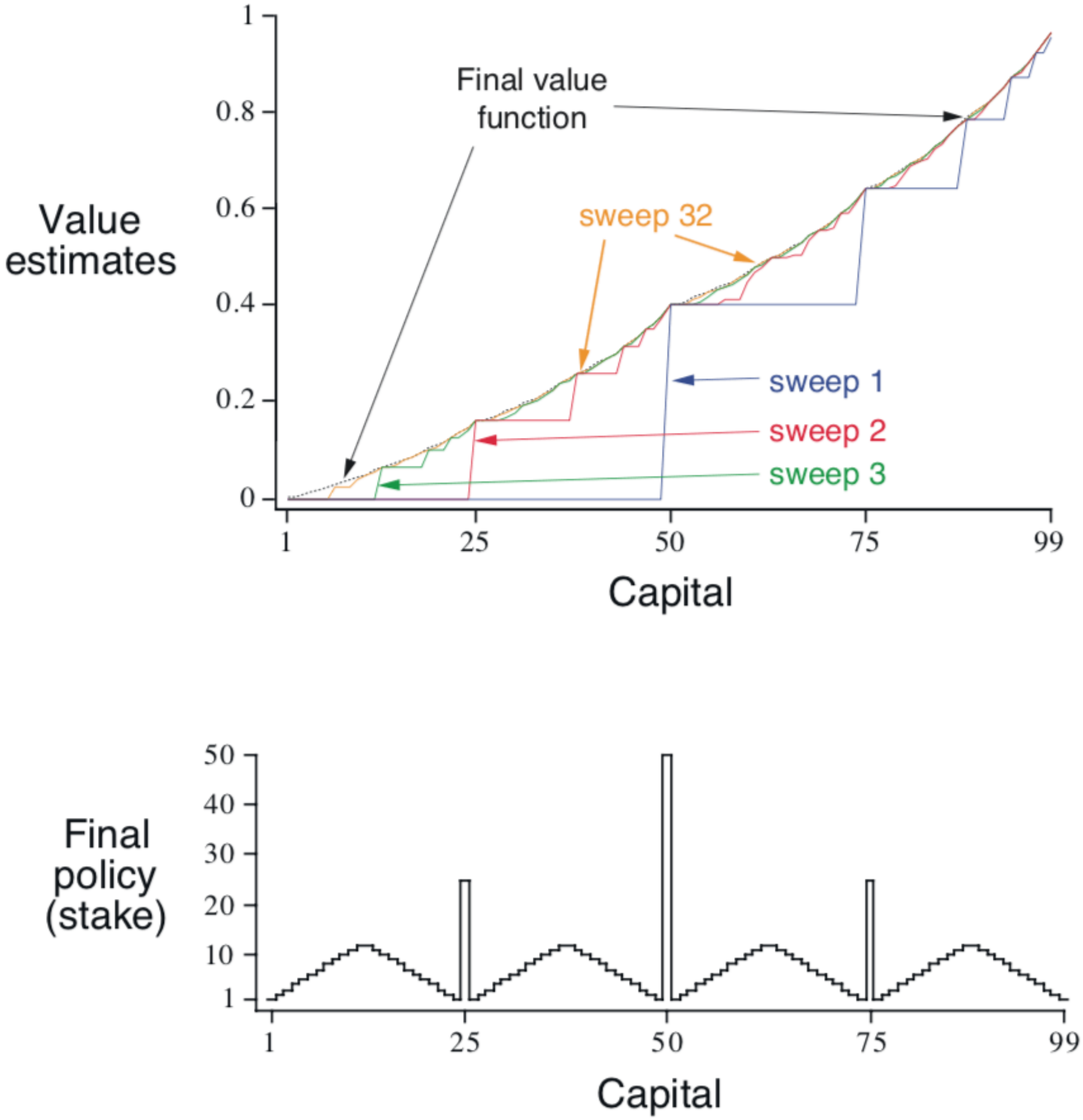
D-P

GPI GPI

4.1

GPI




Value estimates
Final value function
sweep 32
sweep 1
sweep 2
sweep 3
Capital
Final policy (stake)
Capital

18: **4.3** $p_h = 0.4$

GPI G-
PI

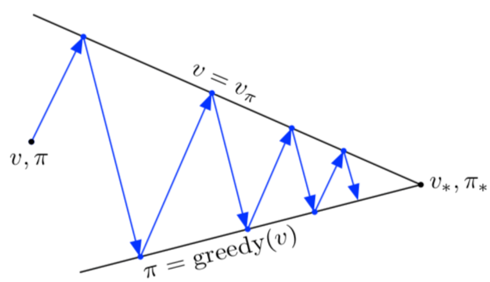
4.7

DP MDP DP DP n
k DP n k ( ) k^n DP
DP MDPs DP
100 DP DP

DP DP
DP

DP MDP

DP
G-
PI



4.8

MDP
DP MD-
P MDP
DP
v_π v_* q_π
q_* DP
DP GPI GPI

GPI DP GPI
DP DP GPI
DP DP

“ ” 1957a DP Bertsekas(2005,2012) Bertsekas Tsitsiklis(1996) Dreyfus Law(1977) Ross(1983) White(1969) Whittle(1982,1983) DP MDPs DP Kumar Kanal(1988)

DP Minsky(1961) Samuel Minsky DP Samuel
DP Andreae(1969b) DP DP
Werbos(1977) DP “ ” (Werbos,1982,1987,1988,1989,1992)
“ ”

4.1-4 DP Bellman(1957a) Howard(1960) Watkins(1989)
Puterman Shin(1978) Bertsekas(1987)

- - - DP Bertsekas Tsitsiklis(1989)

4.5 DP Bersekas(1982,1983) DP DP Bertsekas Tsitsiklis(1989) - - - DP Williams Baird(1990) DP

4.7 Michael Littman Littman Dean Kaelbling(1995) “ ” Bellman(1957)

3.5.4 5

* *

DP

“ ”

- -

DP GPI MDP MDP GPI DP π v_π q_π DP

5.1

-

$v_\pi(s)$ π s π s s s s s s $v_\pi(s)$ MC s M-C MC eligibility traces M-C MC MC S_t

---

MC $V \approx v_\pi$

π

$s \in \mathcal{S}$ $V(s) \in \mathbb{R}$

$Returns(s) \leftarrow$ $s \in \mathcal{S}$

π $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$t = T-1, T-2, \ldots, 0$

$G \leftarrow \gamma G + R_{t+1}$

S_t $S_0, S_1, \ldots, S_{t-1}$

G $Returns(s)$

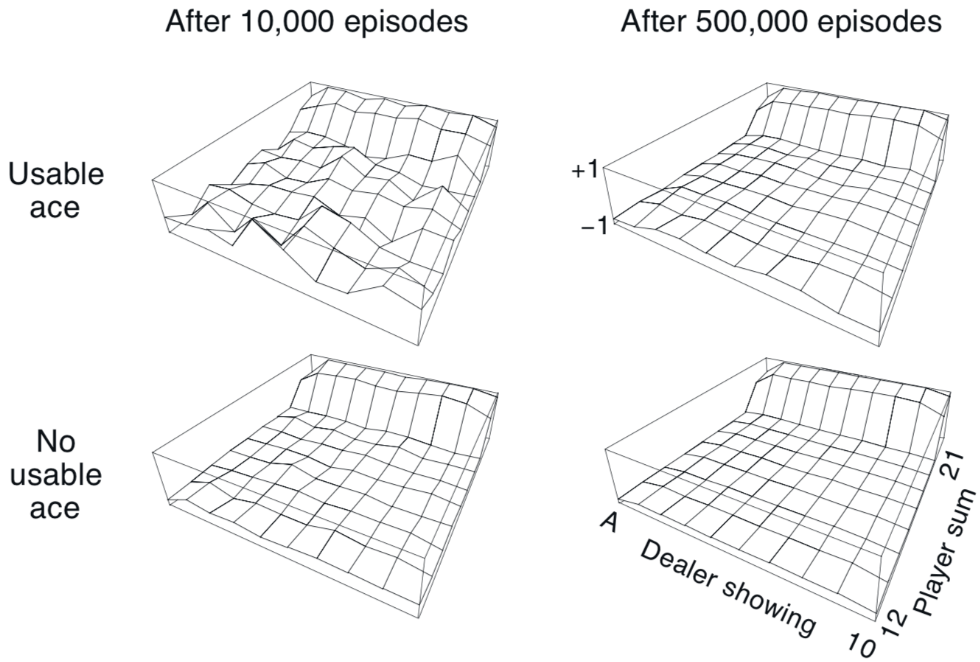
$V(s) \leftarrow average(Returns(s))$

MC MC s $v_\pi(s)$ MC $v_\pi(s)$ $1/\sqrt{n}$ n $v_\pi(s)$ Singh and Sutton 1996

5.1 Blackjack 21 J Q K 10 A 1 11 21 10 A *natural* natural draw 21 21 17 21

21 draw 1 -1 0 0 $\gamma = 1$ A 11 A 11 1 11 12-21 A-10 A 200

20 21 5.1 A A 500,000



19: 5.1 20 21

5.1 5.1

5.2 MC MC

DP DP —— $p(s^{'}, r|s, a)$ —— 14 $p(s^{'}, r|s, a)$ DP DP

v_π

DP v_π DP MC DP MC

MC DP MC *bootstrap*

MC

MC DP

5.2

DP

MC
MC

5.2

- DP
q_*

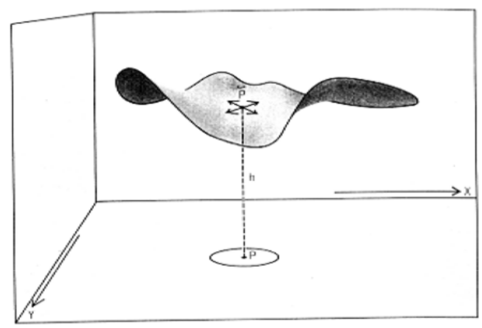
$q_\pi(s,a)$ $q_\pi(s,a)$ s a π return -
- s a s a MC - MC MC
MC

- π π

k-
-

-

5.2 q_π



20: Hersh Griego 1969 1969 Scientific American Nature America Inc

5.3

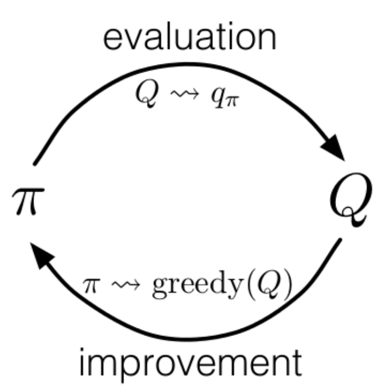
GPI
GPI

MC π_0 -

$$\pi_0 \xrightarrow{E} q_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} q_{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} q_*$$

$\xrightarrow{E}$ $\xrightarrow{I}$ - - π_k
q_{π_k}

- - q $s \in \mathcal{S}$ -



$$\pi(s) \doteq argmax_{a} q(s,a) \tag{3.40}$$

π_{k+1} q_{π_k} 4.2 π_k π_{k+1} $s \in \mathcal{S}$

$$\begin{aligned} q_{\pi_k}(s, \pi_{k+1}(s)) &= q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a)) \\ &= \max_a q_{\pi_k}(s,a) \\ &\geq q_{\pi_k}(s, \pi_k(s)) \\ &\geq v_{\pi_k}(s) \end{aligned}$$

π_{k+1} π_k π_k
MC

MC

DP D-
P MC q_{π_k}

q_{π_k} 4.6 GPI

Monte Carlo ES Monte Carlo with Exploring Starts

Monte Carlo ES $V \approx v_\pi$

$s \in \mathcal{S}, a \in \mathcal{A}(s)$:

$s \in \mathcal{S}$ $\pi(s) \in \mathcal{A}(s)$

$s \in \mathcal{S}, a \in \mathcal{A}(s)$ $Q(s,a) \in \mathbb{R}$

$s \in \mathcal{S}, a \in \mathcal{A}(s)$ $Returns(s,a) \leftarrow$

$S_0 \in \mathcal{S}$ $A_0 \in \mathcal{A}(S_0)$ - 0

S_0, A_0 π $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$t = T-1, T-2, \ldots, 0$

$G \leftarrow \gamma G + R_{t+1}$

S_t, A_t $S_0, A_0, R_1, \ldots, S_{t-1}, A_{t-1}$

G $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow average(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \underset{a}{argmax}\, Q(S_t, a)$

5.4 - 2.4 -
- -
Tsitsiklis 2002

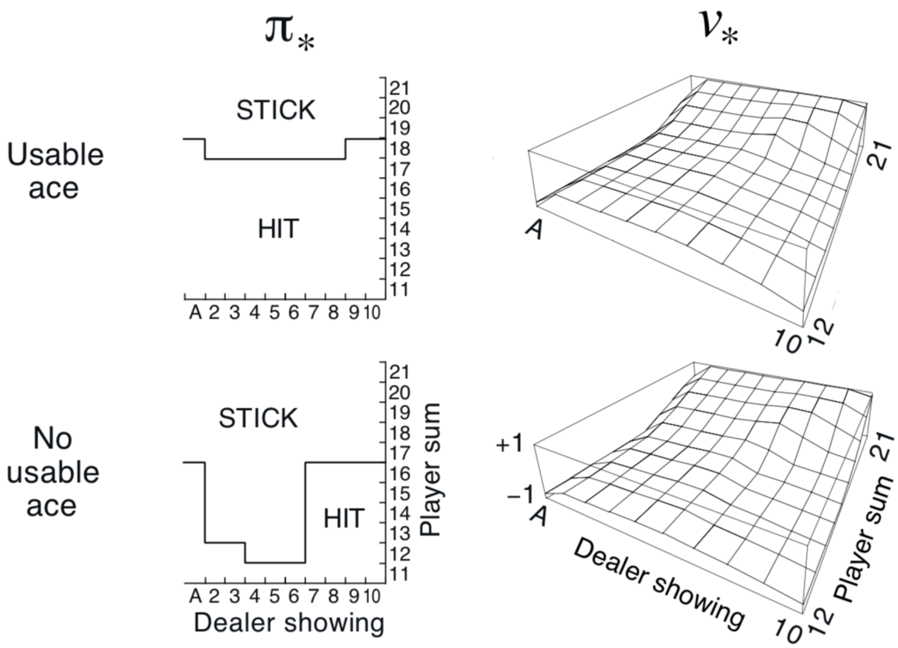
5.3 21 21 A
20 21 - 5.2 A Thorp 1966 “ ” Thorp
21

5.4

on-policy *off-policy*
on-policy Monte Carlo control

soft $s \in \mathcal{S}$ $a \in \mathcal{A}(s)$ $\pi(a|s) > 0$ $\epsilon-$ $\epsilon - greedy$
ϵ $\frac{\epsilon}{|\mathcal{A}(s)|}$ $1-\epsilon+\frac{\epsilon}{|\mathcal{A}(s)|}$ $\epsilon-$ $\epsilon - soft$ $\epsilon - soft$
$\pi(a|s) \geq \frac{\epsilon}{|\mathcal{A}(s)|}$ $\epsilon - soft$ $\epsilon-$

GPI -
GPI $\epsilon-$ $\epsilon - soft$ π q_π $\epsilon-$ π



21: **5.2** 21 - - -

$\epsilon - soft$ $V \approx v_\pi$

$\epsilon > 0$

:

$\pi \leftarrow$ $\epsilon - soft$

$s \in \mathcal{S}, a \in \mathcal{A}(s)$ $Q(s,a) \in \mathbb{R}$

$s \in \mathcal{S}, a \in \mathcal{A}(s)$ $Returns(s,a) \leftarrow$

π $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$t = T-1, T-2, \ldots, 0$

$G \leftarrow \gamma G + R_{t+1}$

S_t, A_t $S_0, A_0, R_1, \ldots, S_{t-1}, A_{t-1}$

G $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow average(Returns(S_t, A_t))$

$A^* \leftarrow \underset{a}{argmax}\, Q(S_t, a)$

$a \in \mathcal{A}(S_t)$:

$$\pi\left(a \mid S_t\right) \leftarrow \begin{cases} 1-\varepsilon+\varepsilon /\left|\mathcal{A}\left(S_t\right)\right| & \text {if } a=A^* \\ \varepsilon /\left|\mathcal{A}\left(S_t\right)\right| & \text {if } a \neq A^* \end{cases}$$

q_π $\epsilon-$ $\epsilon-soft$ π π' $\epsilon-$ $s \in \mathcal{S}$:

$$
\begin{aligned}
q_\pi\left(s, \pi'(s)\right) &= \sum_a \pi'(a|s) q_\pi(s,a) \\
&= \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \max_a q_\pi(s,a) (5.2) \\
&\geq \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \sum_a \frac{\pi(a|s) - \frac{\varepsilon}{|\mathcal{A}(s)|}}{1-\varepsilon} q_\pi(s,a)
\end{aligned}
\tag{3.41}
$$

1

$$
\begin{aligned}
&= \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) - \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + \sum_a \pi(a|s) q_\pi(s,a) \\
&= v_\pi(s)
\end{aligned}
$$

$\pi' \geq \pi$ $s \in \mathcal{S}$ $v_{\pi'}(s) \geq v_\pi(s)$ π' π $\epsilon-soft$

$\epsilon-soft$ s a $1-\epsilon$ ϵ

$\epsilon-soft$ $\tilde{v}_*$ $\tilde{q}_*$ π $v_\pi = \tilde{v}_*$ $\tilde{v}_*$

$$
\begin{aligned}
\widetilde{v}_*(s) =& (1-\varepsilon) \max_a \widetilde{q}_*(s,a) + \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a \widetilde{q}_*(s,a) \\
=& (1-\varepsilon) \max_a \sum_{s',r} p\left(s', r|s,a\right)\left[r + \gamma \widetilde{v}_*\left(s'\right)\right] \\
&+ \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a \sum_{s',r} p\left(s', r|s,a\right)\left[r + \gamma \widetilde{v}_*\left(s'\right)\right]
\end{aligned}
$$

$\epsilon-soft$ π 5.2

$$
\begin{aligned}
v_\pi(s) =& (1-\varepsilon) \max_a q_\pi(s,a) + \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) \\
=& (1-\varepsilon) \max_a \sum_{s',r} p\left(s', r|s,a\right)\left[r + \gamma v_\pi\left(s'\right)\right] \\
+& \quad \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a \sum_{s',r} p\left(s', r|s,a\right)\left[r + \gamma v_\pi\left(s'\right)\right]
\end{aligned}
$$

$\tilde{v}_*$ v_π $\tilde{v}_*$ $v_\pi = \tilde{v}_*$

$\epsilon-soft$ $\epsilon-soft$ $\epsilon-soft$ -
$\epsilon-soft$

5.5

—— -
" off "

17.2 ; Sutton, 2009; Sutton et al., 2011

v_π q_π b $\neq \pi$ π b

b π π b $\pi(a|s) > 0$ $b(a|s) > 0$ *coverage* b π
π $\epsilon-$

S_t π

$A_t, S_{t+1}, A_{t+1}, \dots, S_T$

$$
\begin{aligned}
&Pr\{A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t, A_{t:T-1} \sim \pi\} \\
&= \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \cdots p(S_T|S_{T-1}, A_{T-1}) \\
&= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k),
\end{aligned}
$$

p 3.4

$$\rho_{t:T-1} \doteq \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)} \tag{3.42}$$

MDP MDP

G_t $\mathbb{E}[G_t|S_t = s] = v_b(s)$ v_π $rho_{t:T-1}$

$$\mathbb{E}[\rho_{t:T-1} G_t | S_t = s] = v_\pi(s) \tag{3.43}$$

b $v_\pi(s)$ 100 $t = 101$ s $\mathcal{T}(\int)$ $\mathcal{T}(\int)$ s $T(t)$ G_t t $T(t)$ $\{G_t\}_{t \in \mathcal{T}(\int)}$ s $\{\rho_{t:T(t)-1}\}_{t \in \mathcal{T}(\int)}$ $v_\pi(s)$

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(\int)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(\int)|}. \tag{3.44}$$

ordinary importance sampling

weighted importance sampling

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(\int)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(\int)} \rho_{t:T(t)-1}}, \tag{3.45}$$

s $\rho_{t:T(t)-1}$ $v_b(s)$ $v_\pi(s)$ 5.5 $v_\pi(s)$

1

Precup, Sutton, and Dasgupta 2001

MC

5.5 MDP p $1-p$ $+1$ $\gamma = 1$ 10 10

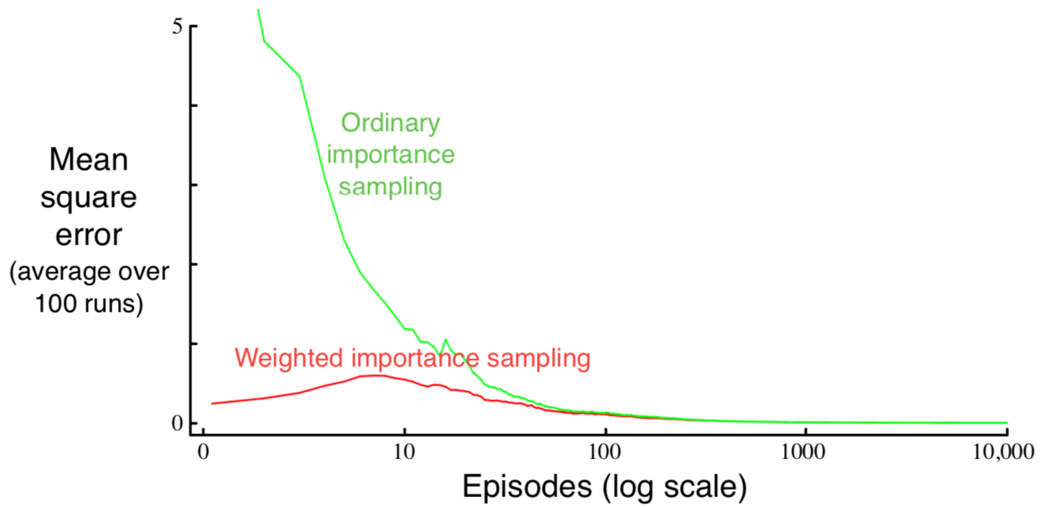
5.4 **21** 21 5.1 13 A A 2 20 21 5.1 -0.27726 1000 100 10000 5.3 —— 100

5.5 —— 5.4 s 0.9 s 0.1 $+1$ s s $+1$ s 1 $\gamma = 1$

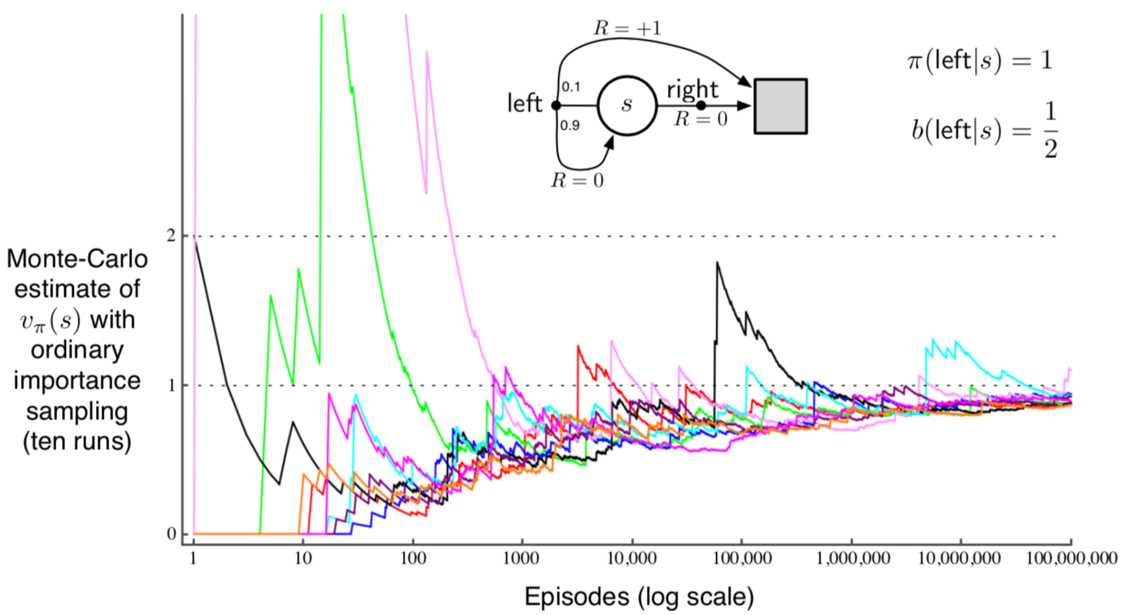
5.4 MC 1 1 1 $\rho_{t:T(t)}$ 5.6 1

X

$$Var[X] \doteq \mathbb{E}[(X - \overline{X})^2] = \mathbb{E}[X^2 - 2X\overline{X}^2 + \overline{X}^2] = \mathbb{E}[X^2] - \overline{X}^2.$$



22: **5.3** 21



23: **5.4** 5.5 MDP 1 $\gamma = 1$ MC

X

$$\mathbb{E}_b\left[\left(\prod_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}G_0\right)^2\right].$$

1 G_0

$$
\begin{aligned}
&= \frac{1}{2}\cdot 0.1\left(\frac{1}{0.5}\right)^2 \ (\quad 1 \quad) \\
&+ \frac{1}{2}\cdot 0.9\cdot\frac{1}{2}\cdot 0.1\left(\frac{1}{0.5}\frac{1}{0.5}\right)^2 \ (\quad 2 \quad) \\
&+ \frac{1}{2}\cdot 0.9\cdot\frac{1}{2}\cdot 0.9\cdot\frac{1}{2}\cdot 0.1\left(\frac{1}{0.5}\frac{1}{0.5}\frac{1}{0.5}\right)^2 \ (\quad 3 \quad) \\
&+ \cdots \\
&= 0.1\sum_{k=0}^{\infty}0.9^k\cdot 2^k\cdot 2 = 0.2\sum_{k=0}^{\infty}1.8^k = \infty
\end{aligned}
$$

5.6 b 5.6 $V(s)$ $Q(s,a)$

5.7 5.3

5.8 5.5 5.4 MC MC

5.6

2.4

$\rho_{t:T(t)-1}$ 5.3 5.5

$G_1, G_2, \ldots, G_{n-1}$ W_i $W_i = \rho_{t_i:T(t_i)-1}$

G_n V_n n C_n V_n

$$V_{n+1} \doteq V_n + \frac{W_n}{C_n}\left[G_n - V_n\right], \quad n \geq 1, \tag{3.46}$$

$$C_{n+1} \doteq C_n + W_{n+1},$$

$C_0 \doteq 0$ V_1 $\pi = b$ W 1
Q q_π - b

5.9 5.1 MC 2.4

5.10 5.7 5.8 2.3

MC $Q \approx q_\pi$

π

$s \in \mathcal{S}$ $a \in \mathcal{A}(s)$

$Q(s,a) \in \mathbb{R}$

$C(s,a) \leftarrow 0$

$b \leftarrow$ π

b $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$W \leftarrow 1$

$t = T-1, T-2, \ldots, 0$ $W \neq 0$

$G \leftarrow \gamma G + R_{t+1}$

$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$

$W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

5.7

soft

π_* q_* GPI $\pi \approx \pi_*$ Q Q q_π b π
b $\epsilon - soft$ b b π

MC $\pi \approx \pi_*$

$s \in \mathcal{S}$ $a \in \mathcal{A}(s)$

$Q(s, a) \in \mathbb{R}$

$C(s, a) \leftarrow 0$

$\pi(s) \leftarrow argmax_a Q(s, a)$

$b \leftarrow$

b $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$W \leftarrow 1$

$t = T-1, T-2, \ldots, 0$

$G \leftarrow \gamma G + R_{t+1}$

$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$

$\pi(S_t) \leftarrow argmax_a Q(S_t, a)$

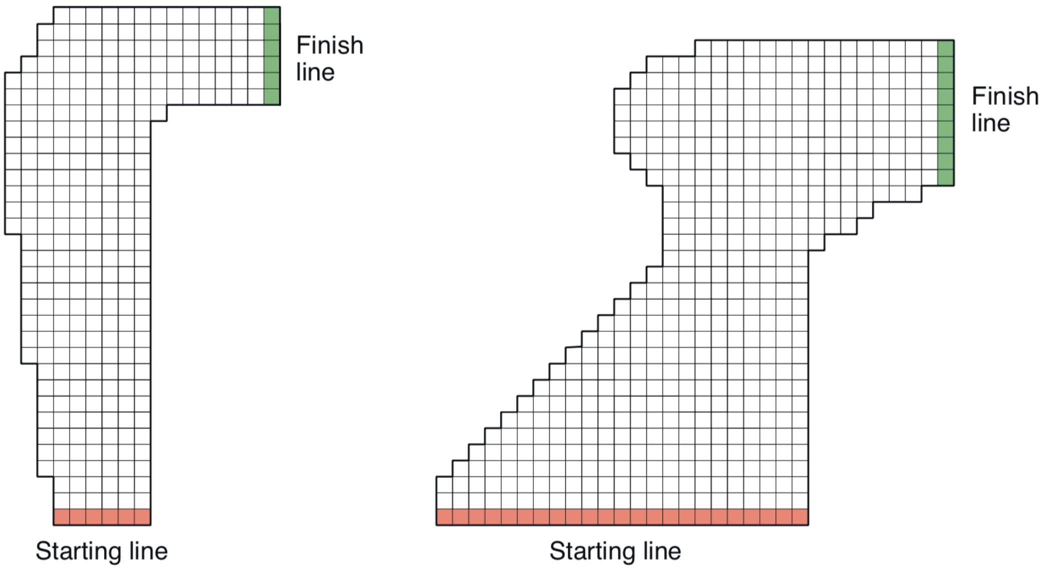
$A_t \neq \pi(S_t)$

$W \leftarrow W \frac{1}{b(A_t|S_t)}$

temporal-difference γ 1

5.11 MC W $\frac{\pi(A_t|S_t)}{b(A_t|S_t)}$ $\frac{1}{b(A_t|S_t)}$

5.12 5.5



24: 5.5

$+1, -1, 0$ 3×3 5 -1 0.1

5.8 *

γ 1 100 $\gamma = 0$ 0 $G_0 = R_1$ $\frac{\pi(A_0|S_0)}{b(A_0|S_0)}\frac{\pi(A_1|S_1)}{b(A_1|S_1)}\cdots\frac{\pi(A_{99}|S_{99})}{b(A_{99}|S_{99})}$ $\frac{\pi(A_0|S_0)}{b(A_0|S_0)}$ 99 $\frac{\pi(A_1|S_1)}{b(A_1|S_1)}\cdots\frac{\pi(A_{99}|S_{99})}{b(A_{99}|S_{99})}$ 1

degree $\gamma \in [0, 1)$ G_0 $1-\gamma$ R_1 $(1-\gamma)\gamma$ $R_1 + R_2$ $(1-\gamma)\gamma$ γ $1-\gamma$ $(1-\gamma)\gamma^2$ γ^2 *flat partial returns*

$$\overline{G}_{t:h} \doteq R_{t+1} + R_{t+2} + \cdots + R_h, \quad 0 \le t < h \le T$$

“ ” “ ” h h *horizon* T G_t

$$
\begin{aligned}
G_t \doteq & R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T \\
= & (1-\gamma) R_{t+1} \\
& + (1-\gamma)\gamma \left(R_{t+1} + R_{t+2}\right) \\
& + (1-\gamma)\gamma^2 \left(R_{t+1} + R_{t+2} + R_{t+3}\right) \\
& \vdots \\
& + (1-\gamma)\gamma^{T-t-2} \left(R_{t+1} + R_{t+2} + \cdots + R_T\right) \\
& + \gamma^{T-t-1} \left(R_{t+1} + R_{t+2} + \cdots + R_T\right) \\
& = (1-\gamma) \sum_{h=t+1}^{T-1} \gamma^{h-t-1} \overline{G}_{t:h} + \gamma^{T-t-1} \overline{G}_{t:T}
\end{aligned}
$$

$G_{t:h}$ h h 5.5

$$
V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t:h-1} \overline{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t:T(t)-1} \overline{G}_{t:T(t)} \right)}{|\mathcal{T}(s)|} \tag{3.47}
$$

5.6

$$
V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t:h-1} \overline{G}_{t:h} + \gamma^{T(t)-t-1} \rho_{t:T(t)-1} \overline{G}_{t:T(t)} \right)}{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_{t:h-1} + \gamma^{T(t)-t-1} \rho_{t:T(t)-1} \right)} \tag{3.48}
$$

discounting-aware $\gamma = 1$ 5.5

5.9 *

$\gamma = 1$ 5.5 5.6

$$
\begin{aligned}
\rho_{t:T-1} G_t &= \rho_{t:T-1} \left(R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_T\right) \\
&= \rho_{t:T-1} R_{t+1} + \gamma \rho_{t:T-1} R_{t+2} + \cdots + \gamma^{T-t-1} \rho_{t:T-1} R_T (5.11)
\end{aligned}
$$

5.11 5.3

$$
\rho_{t:T-1} R_{t+1} = \frac{\pi(A_t|S_t)}{b(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})} \frac{\pi(A_{t+2}|S_{t+2})}{b(A_{t+2}|S_{t+2})} \cdots \frac{\pi(A_{T-1}|S_{T-1})}{b(A_{T-1}|S_{T-1})} R_{t+1}. \tag{3.49}
$$

$: label : 5.13$

$$
\mathbb{E}\left[\frac{\pi(A_k|S_k)}{b(A_k|S_k)}\right] = \sum_a b(a|S_k) \frac{\pi(a|S_k)}{b(a|S_k)} = \sum_a \pi(a|S_k) = 1.
$$

$$
\mathbb{E}[\rho_{t:T-1} R_{t+1}] = \mathbb{E}[\rho_{t:t} R_{t+1}]. \tag{3.50}
$$

5.11 k

$$
\mathbb{E}[\rho_{t:T-1} R_{t+k}] = \mathbb{E}[\rho_{t:t+k-1} R_{t+k}].
$$

5.11

$$
\mathbb{E}[\rho_{t:T-1} G_t] = \mathbb{E}[\tilde{G}_t],
$$

$$\tilde{G}_t = \rho_{t:t}R_{t+1} + \gamma\rho_{t:t+1}R_{t+2} + \gamma^2\rho_{t:t+2}R_{t+3} + \cdots + \gamma^{T-t-1}\rho_{t:T-1}R_T$$

per-decision 5.5 $\tilde{G}_t$

$$V(s) \doteq \frac{\sum_{t\in\mathcal{T}(s)} \tilde{G}_t}{|\mathcal{T}(s)|}, \tag{3.51}$$

** 5.13* 5.12 5.14 ** 5.14* 5.10 5.7

5.10

DP D-P

bootstrap

GPI GPI

-

-

bootstrap
bootstrap

" " 1940s Los Alamos Kalos and Whitlock, 1986 Rubinstein, 1981

5.1-2 Singh Sutton 1996 MC MC 21 Widrow, Gupta Maitra 1973 Dirichlet

Kakutani 1945 Hersh Griego 1969 Doyle Snell 1984 Barto Duff 1994 Curtiss 1954

5.3-4 1998 1968 Michie Chambers 3.4 " " MC Narendra Wheeler 1986

5.5 Pearl 1995 Balke Pearl 1994 " interventions " " counterfactuals " normalized Koller Friedman 2009 Rubinstein 1981 Hesterberg 1988 Shelton 2001 Liu 2001

" "

5.7 Barto Bradtke Singh 1995 Gardner 1973

5.8 Sutton Mahmood Precup van Hasselt 2014 Mahmood Mahmood van Hasselt Sutton 2014

5.9 Precup Sutton Singh 2000

3.5.5 6

TD TD DP TD D-P TD TD DP 7 n TD 12 TD(λ)

π v_π DP TD GPI

6.1 TD

TD π S_t v_π V $V(S_t)$

$$V(S_t) \leftarrow V(S_t) + \alpha\left[G_t - V(S_t)\right] \tag{3.52}$$

G_t t α 2.4 α MC $V(S_t)$ G_t TD $t+1$ R_{t+1} TD

$$V(S_t) \leftarrow V(S_t) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right] \tag{3.53}$$

S_{t+1} R_{t+1} G_t TD $R_{t+1} + \gamma V(S_{t+1})$ TD *TD(0)* *TD* 12 7 TD(λ) n TD TD(0)

TD(0) v_π

π

$\alpha in (0, 1]$

$s \in \mathbb{S}^+$ $V(\ ) = 0$ $V(s)$

S

$A \leftarrow$ π S

A R S'

$V(S) \leftarrow V(S) + \alpha\left[R + \gamma V(S') - V(S)\right]$

$S \leftarrow S'$

S

TD(0) *bootstrapping* DP 3

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi\left[G_t | S_t = s\right] (6.3) \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} | S_t = s\right] \qquad (\ (3.9)) \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma v_\pi\left(S_{t+1}\right) | S_t = s\right] (6.4) \end{aligned}$$

6.3 DP 6.4 6.3 DP $v_\pi(S_{t+1})$ $V(S_{t+1})$ TD 6.4 V v_π TD DP DP

TD(0) TD - ? - DP

TD(0)

TD(0) S_t $R_{t+1} + \gamma V(S_{t+1})$ TD

$$\delta_t \doteq R_{t+1} + \gamma V\left(S_{t+1}\right) - V\left(S_t\right) \tag{3.54}$$

TD TD δ_t $V(S_{t+1})$ $t + 1$ V (?) TD

$$\begin{aligned} G_t - V\left(S_t\right) &= R_{t+1} + \gamma G_{t+1} - V\left(S_t\right) + \gamma V\left(S_{t+1}\right) - \gamma V\left(S_{t+1}\right) && \text{((3.9))} \\ &= \delta_t + \gamma\left(G_{t+1} - V\left(S_{t+1}\right)\right) \\ &= \delta_t + \gamma\delta_{t+1} + \gamma^2\left(G_{t+2} - V\left(S_{t+2}\right)\right) \\ &= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}\left(G_T - V\left(S_T\right)\right) \\ &= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}(0-0) \\ &= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k (6.6) \end{aligned}$$

V TD(0)

6.1 V 6.6 V_t TD 6.5 TD 6.2 t TD

6.1 6 30 6:05 35 40 35 6:40

6	0	30	30
	5	35	40
	20	15	35
	30	10	40
	40	3	43
	43	0	43

[1] $\gamma = 1$

6.1 α MC 6.1 $\alpha = 1$ 15 23 6.1 $G_t - V(S_t)$ 8 α 1/2

30 25 50 30

TD 30 50 6.1 TD 6.2 $\alpha = 1$

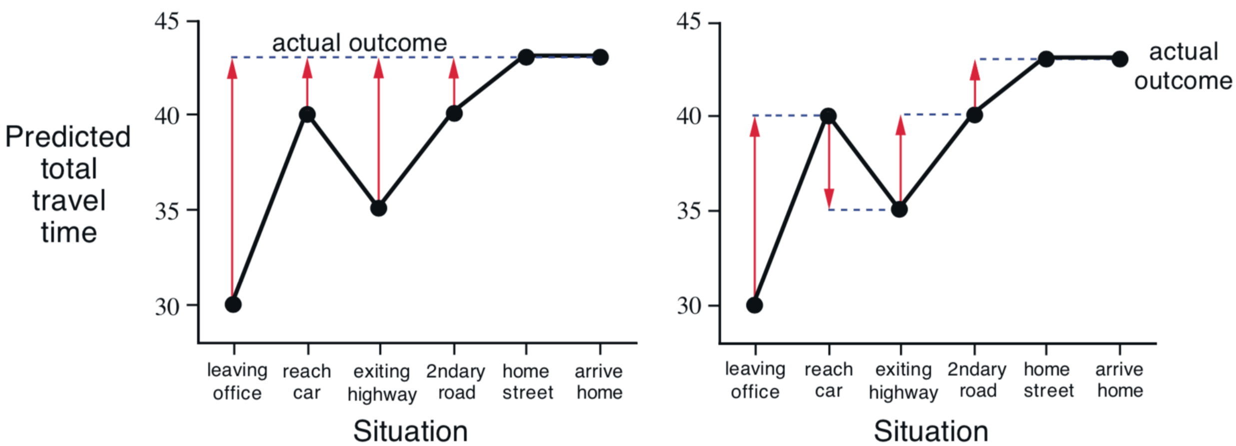
6.2 TD TD TD - - T-D TD

6.2 TD

TD - TD DP

TD DP

[1]



25: **6.1** TD

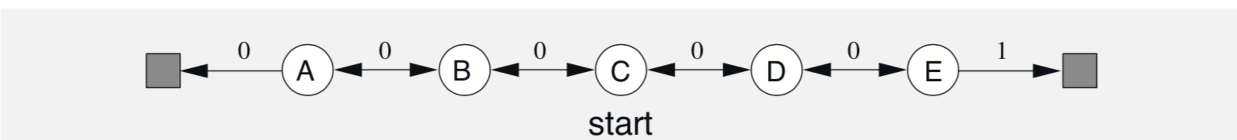
TD TD
TD

TD π TD(0) v_π
2.7 1 6.2 9.4

TD “ ”
TD - α MC 6.2

6.2

TD(0) - *alpha* MC



MRP MRP MRP C
+1 $C, 0, B, 0, C, 0, D, 0, E, 1$ $v_\pi(C) = 0.5$ A E
$\frac{1}{6}$ $\frac{2}{6}$ $\frac{3}{6}$ $\frac{4}{6}$ $\frac{5}{6}$
TD(0) 100 - $\alpha = 0.1$ α
RMS 100 s $V(s) = 0.5$ TD MC

6.3 $V(A)$

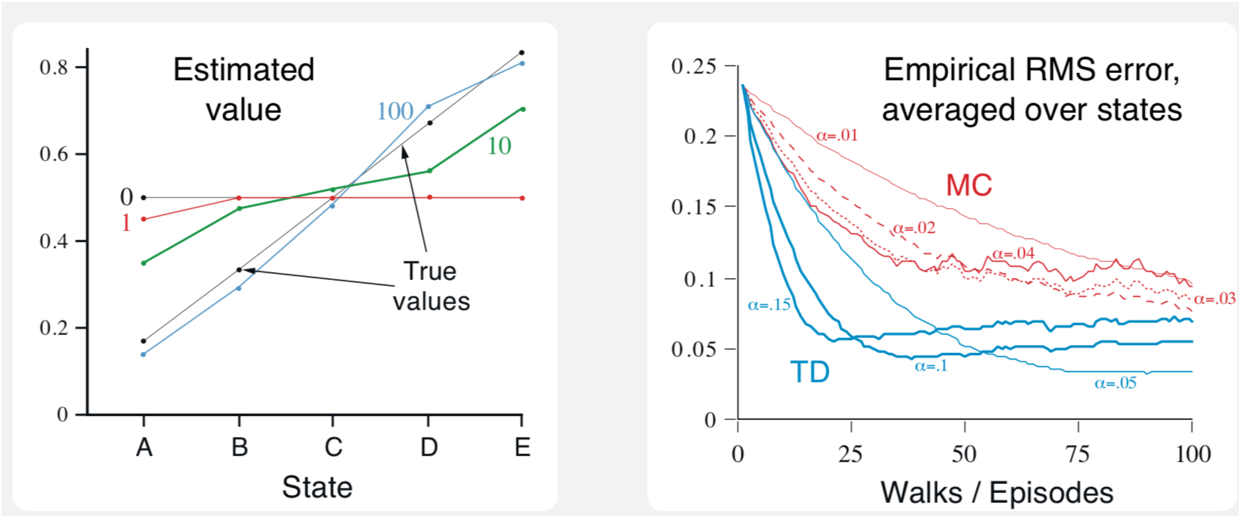
6.4 α α α

* *6.5* TD RMS α

6.6 6.2 A E $\frac{1}{6}$ $\frac{2}{6}$ $\frac{3}{6}$ $\frac{4}{6}$ $\frac{5}{6}$

6.3 TD(0)

10 100 V t 6.1 6.2



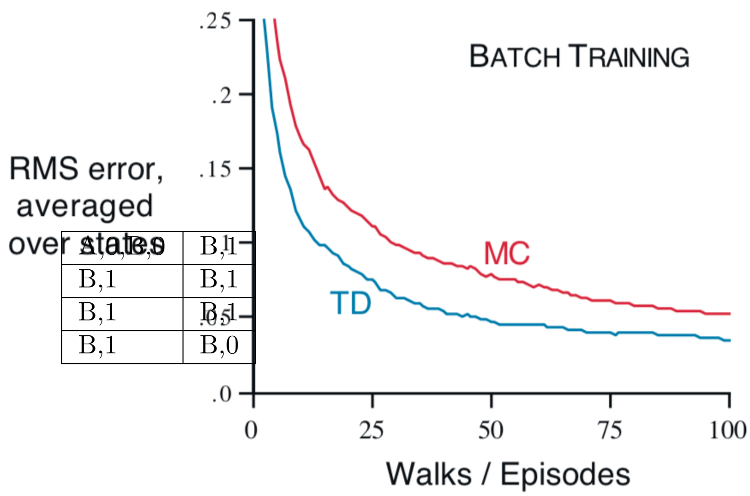
TD(0) α α - α MC

6.3 TD(0) - α MC 6.2 TD(0) - α MC α v_π 100 6.2 TD

- α MC $V(s)$

TD TD TD

6.4



A,0,B,0	B,1
B,1	B,1
B,1	B,1
B,1	B,0

A B 0 B 0 B $V(A)$ $V(B)$ $V(B)$ $\frac{3}{4}$ B 8 6 1 B 0

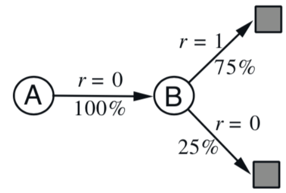
$V(A)$ A 100 B 0 B $\frac{3}{4}$ A $\frac{3}{4}$

$V(A) = \frac{3}{4}$ TD(0)

A 0 $V(A)$ 0

26: **6.2** TD(0) - α MC

6.4 TD(0) TD(0) i j i j



TD(0)

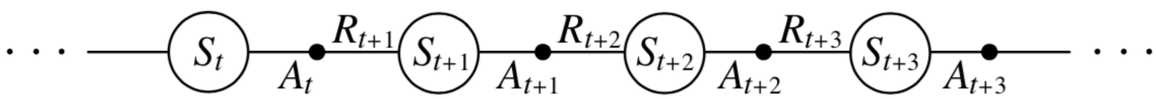
TD TD(0) TD(0) 6.2
TD(0) 6.2 125
TD(0) - α MC TD

$n = |\mathcal{S}|$ n^2
n^3 TD n TD

* *6.7* TD 0 b t 5.3

6.4 Sarsa TD

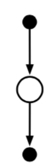
TD GPI TD TD
π s a $q(s,a)$ v_π TD -



- - - TD(0)

$$Q\left(S_t, A_t\right) \leftarrow Q\left(S_t, A_t\right) + \alpha\left[R_{t+1} + \gamma Q\left(S_{t+1}, A_{t+1}\right) - Q\left(S_t, A_t\right)\right]. \tag{3.55}$$

S_t S_{t+1} $Q(S_{t+1}, A_{t+1})$ $(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$ - -
Sarsa Sarsa



Sarsa

Sarsa π q_π π q_π Sarsa

Sarsa Q ε - ε -soft - ε - $\varepsilon = 1/t$
Sarsa 1 -

6.8 6.6 TD $\delta_t = R_{t+1} + \gamma Q\left(S_{t+1}, A_{t+1}\right) - Q\left(S_t, A_t\right)$

Sarsa TD $Q \approx q_*$

$\alpha \in (0,1]$ $\varepsilon > 0$

$s \in (S)^+$ $a \in (A)(s)$ $Q(s,a)$ $Q(\ ,\cdot) = 0$

S

Q S Avarepsilon‘ -

A R, S'

Q S' A'varepsilon‘ -

$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma Q\left(S', A'\right) - Q(S,A)\right]$

$S \leftarrow S'$ $A \leftarrow A'$

S

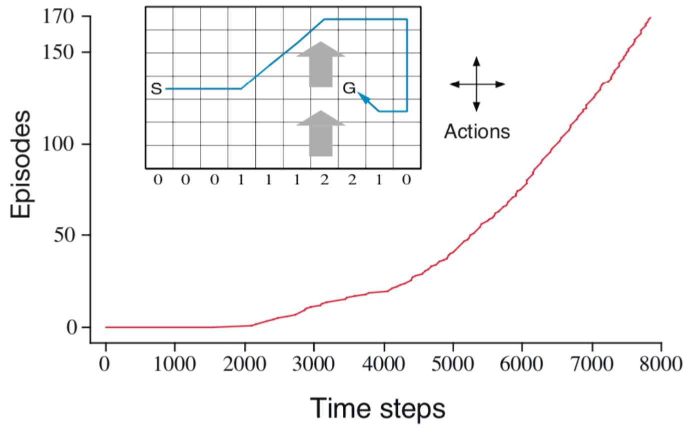
6.5 “ ” -1

ε - Sarsa $\varepsilon = 0.1$ $\alpha = 0.5$ s, a $Q(s, a) = 0$ 8000 ε - 17 15

Sarsa

6.9

6.10 1



6.5 Q-learning TD

Q-learning Watkins 1989 TD

$$Q\left(S_t, A_t\right) \leftarrow Q\left(S_t, A_t\right) + \alpha\left[R_{t+1} + \gamma \max_a Q\left(S_{t+1}, a\right) - Q\left(S_t, A_t\right)\right]. \tag{3.56}$$

- Q q_* - - - 5 Q 1 q_* Q-learning

Q-learning TD $\pi \approx \pi_*$

$\alpha \in (0, 1]$ $\varepsilon > 0$

$s \in (S)^+$ $a \in (A)(s)$ $Q(s, a)$ $Q(\ , \cdot) = 0$

S

Q S *Avarepsilon‘* -

A R, S'

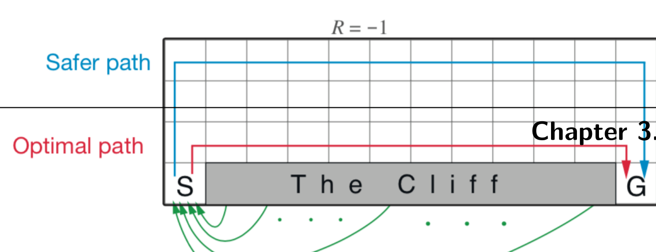
$Q(S, A) \leftarrow Q(S, A) + \alpha\left[R + \gamma \max_a Q\left(S', a\right) - Q(S, A)\right]$

$S \leftarrow S'$

S

Q-learning 6.8 - “ ” 3.4- 6.4

6.6 Sarsa Q-learning Sarsa Q-learning



1 “ ” −100

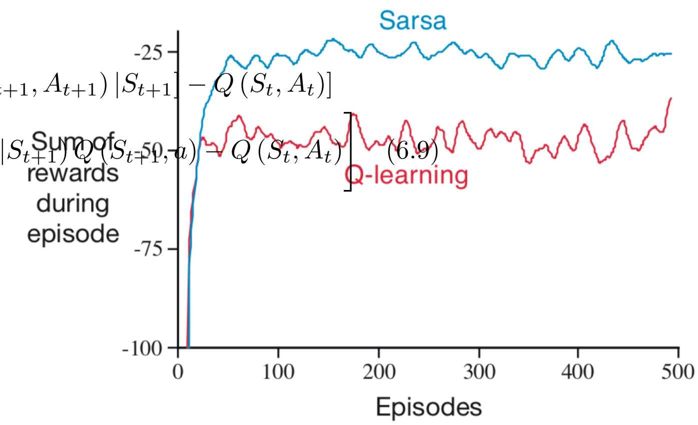
ε - Sarsa Q-learning $\alpha = 0.1$ Q-learning “ε - ” Sarsa Q-learning Sarsa ε

6.11 Q-learning

6.12 Q-learning Sarsa

6.6 Sarsa

Q-learning -



$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \mathbb{E}_\pi\left[Q(S_{t+1}, A_{t+1}) \mid S_{t+1}\right] - Q(S_t, A_t)\right]$$
$$\leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t)\right] \quad (6.9)$$

Q-learning S_{t+1} Sarsa *Sarsa* 6.4

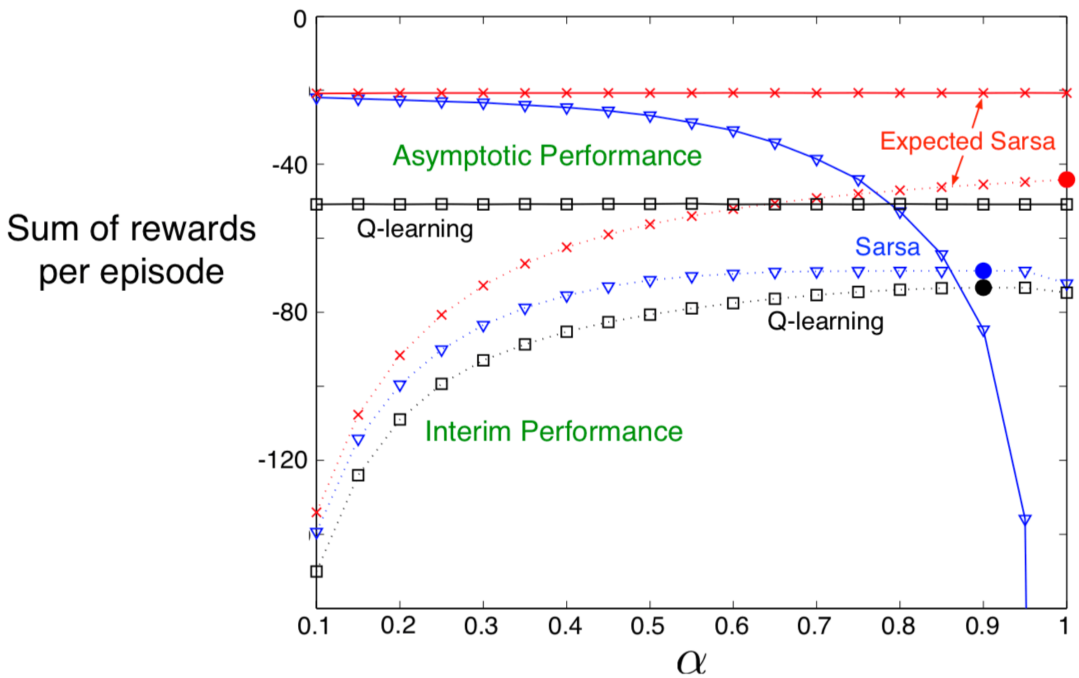
Sarsa Sarsa A_{t+1} Sarsa 6.3 Sarsa Sarsa Q-learning Sarsa Sarsa Q-learning αalpha=1‘ Sarsa α Sarsa Sarsa

Sarsa π π Sarsa Q-learning Sarsa Q-learning Sarsa Sarsa TD

6.8

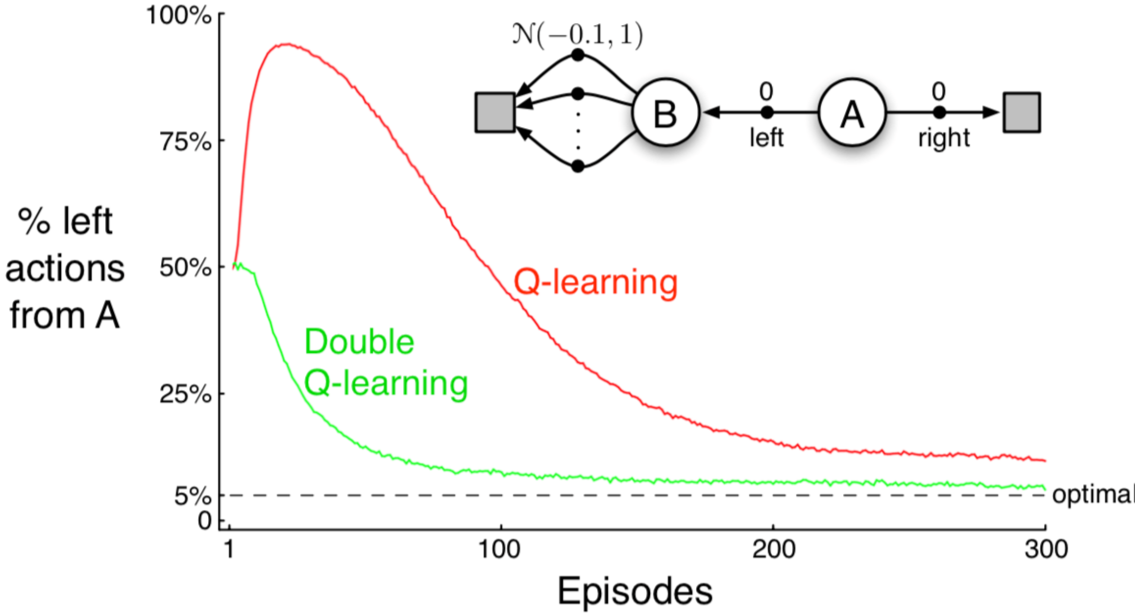
Q-learning max Sarsa ε - s a $q(s,a)$ $Q(s,a)$

6.7 6.5 MDP TD MDP **A** **B** **A** **B** −0.1 1.0 −0.1 **A** **B** 6.5 ε - Q-learning Q-learning $\varepsilon = 0.1$ $\alpha = 0.1$ $\gamma = 0.1$ 5



27: 6.3 TD α ε - $\varepsilon = 0.1$ 100,000 100 50,000 10
van Seijen et al.(2009)

28: 6.4 Q-learning Sarsa



29: **6.5** MDP Q-learning Q-learning Q-learning $\varepsilon = 0.1$ ε - 5 Q-learning 10,000 ε -

$Q_1(a)$ $Q_2(a)$
$a \in \mathcal{A}$ $q(a)$
Q_1 $A^* = \arg\max_a Q_1(a)$ Q_2
$Q_2(A^*) = Q_2(\arg\max_a Q_1(a))$
$\mathcal{E}[Q_2(A^*)] = q(A^*)$
$Q_1(\arg\max_a Q_2(a))$
Q-learning

MDP Q-learning Q-learning

$$Q_1(S_t, A_t) \leftarrow Q_1(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q_2\left(S_{t+1}, \arg\max_a Q_1(S_{t+1}, a)\right) - Q_1(S_t, A_t)\right] \tag{3.57}$$

Q_1 Q_2 Q_2
Q-learning
ε - Q-learning 6.5
Sarsa Sarsa

Q-learning $Q_1 \approx Q_2 \approx q_*$

$\alpha \in (0, 1]$ $\varepsilon > 0$

$s \in (S)^+$ $a \in (A)(s)$ $Q_1(s, a)$ $Q_2(s, a)$

$Q(\ ,\cdot) = 0$

S

$Q_1 + Q_2$ ε - S A

A R, S'

0.5

$Q_1(S,A) \leftarrow Q_1(S,A) + \alpha\left(R + \gamma Q_2\left(S', \arg\max_a Q_1\left(S', a\right)\right) - Q_1(S,A)\right)$

$Q_2(S,A) \leftarrow Q_2(S,A) + \alpha\left(R + \gamma Q_1\left(S', \arg\max_a Q_2\left(S', a\right)\right) - Q_2(S,A)\right)$

$S \leftarrow S'$

S

---

* *6.13* ε - Sarsa

6.9 Afterstates

- TD -

afterstates *afterstate* afterstates af-terstate

afterstates
-

- “ ” -
afterstate -
-

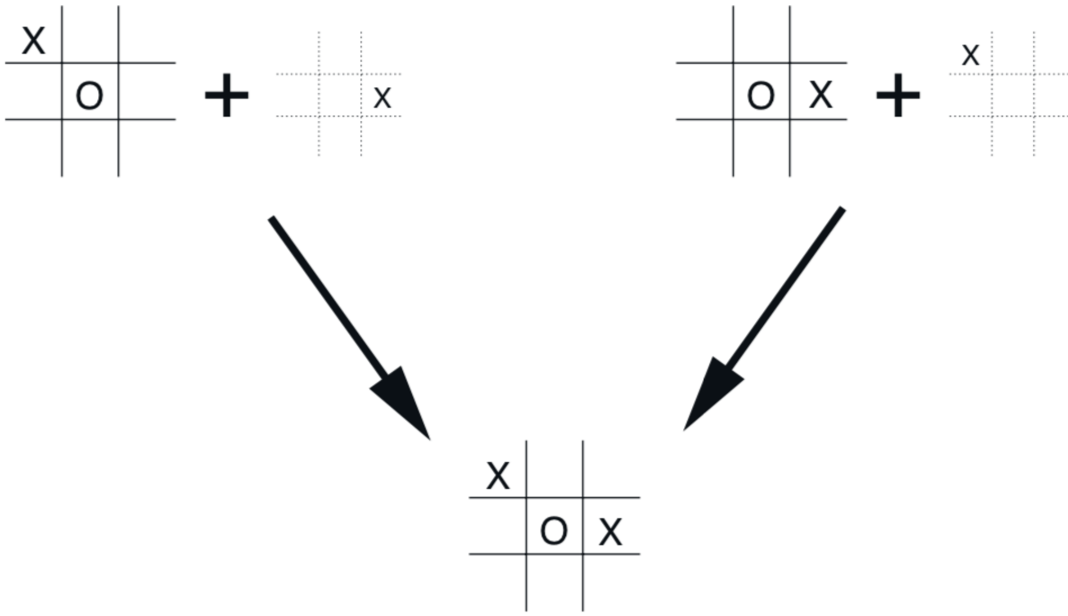
Afterstates

afterstate afterstate

6.14 4.2 afterstates

6.10

TD
TD



GPI

GPI ε
- TD
Sarsa Q-
learning Sarsa
TD - 13

TD
TD TD
n

TD TD
TD
TD
TD

1 TD
Samuel 1959 Klopf 1972
Samuel 16.2 TD Holland 1975,1976
Barto 1970 1975 Holland
Holland TD -
Booker 1982 Holland bucket
brigade 1986 Sarsa

6.1-2 Sut-

ton 1988 TD(0) “ ”
Watkins 1989 Werbos 1987

Watkins Dayan 1992 TD(0) Sutton 1988 Dayan 1992 1
Jaakkola Jordan Singh 1994 Tsitsiklis 1994

6.3 Sutton 1988 TD
Barnard 1993 TD

Goodwin Sin 1984

6.4 Sarsa Rummery Niranjan 1994 “ Q-learning” Sutton 1996 “Sarsa”
Singh Jaakkola Littman Szepesvari 2000 Sarsa Tom
Kalt “ ”

Holland 1986 Sarsa
T-D
Sarsa Wilson 1994

6.5 Watkins 1989 Q-learning Watkins Dayan 1992
Jaakkola Jordan Singh 1994 Tsitsiklis 1994

6.6 Sarsa George John 1994 “$\overline{Q}$-learning” Q-learning Sarsa van Seijen van Hasselt Whiteson Weiring 2009
Sarsa Sarsa Q-learning John 6.3 Van Seijen “ Sarsa”
van Hasselt 2011 Sarsa “ Q ”

6.7 van Hasselt 2010 2011
6.5 MDP 4.1 van Hasselt 2011

6.8 Afterstate “ ” Van Roy Bertsekas Lee Tsitsiklis 1997; Powell 2011

3.5.6 7 n Bootstrapping

MC TD MC TD
n TD
n MC TD

n T-D

TD
n

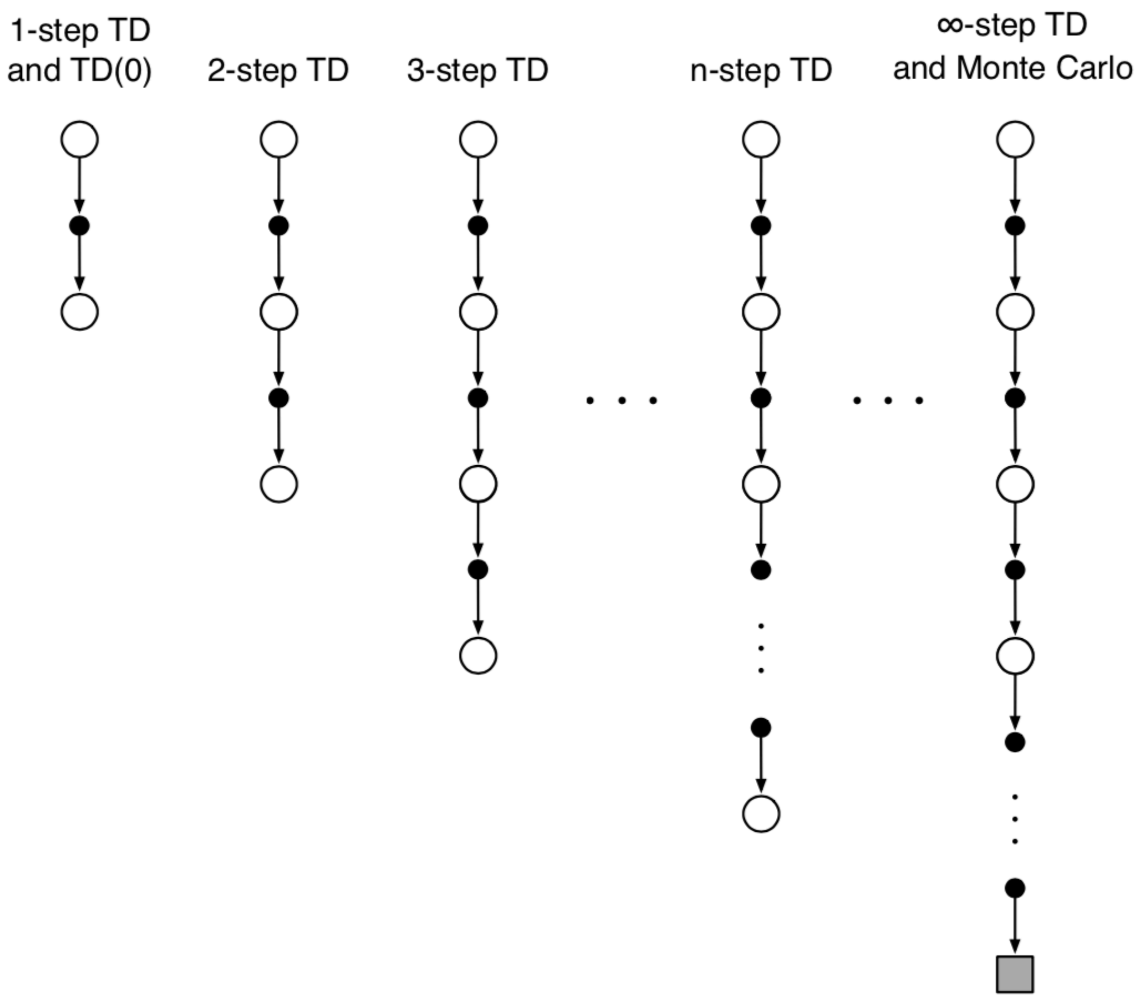
n 12
n
n

n
v_π

7.1 n TD

TD π
v_π
TD

7.1
v_π n TD



30: **7.1** n TD

n TD

n n n TD T-
D TD

S_t
$S_t, R_{t+1}, S_{t+1}, R_{t+2}, \ldots, R_T, S_T$
$v_\pi(S_t)$

$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$

T

$$G_{t:t+1} \doteq R_{t+1} + \gamma V_t\left(S_{t+1}\right)$$

$V_t : \mathcal{S} \to \mathbb{R}$ v_π t $G_{t:t+1}$
t $t+1$ $\gamma V_t\left(S_{t+1}\right)$
$\gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$

$$G_{t:t+2} \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 V_{t+1}\left(S_{t+2}\right)$$

$\gamma^2 V_{t+1}\left(S_{t+2}\right)$ $\gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots + \gamma^{T-t-1} R_T$ n n

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}\left(S_{t+n}\right) \tag{3.58}$$

$s \neq S_t$ $V_{t+n}(s) = V_{t+n-1}(s)$ n TD n-1

---

n **TD(0)** $V \approx v_\pi$

π

$\alpha \in (0,1]$ n

$s \in \mathcal{S}$ $V(s)$

S_t R_t $mod n+1$

$S_0 \neq$

$T \leftarrow \infty$

$t = 0, 1, 2, \ldots$

$t < T$

$\pi(\cdot|S_t)$

R_{t+1} S_{t+1}

S_{t+1} $T \leftarrow t+1$

$\tau \leftarrow t - n + 1$ τ

$\tau \geq 0$

$G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$

$\tau + n < T$ $G \leftarrow G + \gamma^n V\left(S_{\tau+n}\right)$

$V\left(S_\tau\right) \leftarrow V\left(S_\tau\right) + \alpha\left[G - V\left(S_\tau\right)\right]$
$\left(G_{\tau:\tau+n}\right)$

$\tau = T - 1$

7.1 6 TD 6.6
7.2 n TD

7.2 n
TD 7.2

n V_{t+n-1} R_{t+n}
n V_{t+n-1} v_π
$n \geq 1$ n V_{t+n-1} γ^n

$$\max_s \left|\mathbb{E}_\pi \left[G_{t:t+n} | S_t = s\right] - v_\pi(s)\right| \leq \gamma^n \max_s \left|V_{t+n-1}(s) - v_\pi(s)\right| \tag{3.59}$$

n n TD
n TD sound TD

7.1 **n TD** 6.2 5 n TD
C **D** **E** 1
$V(s) = 0.5$
$V(E)$ 1 $V(D)$
$V(E)$ 1 n $n \geq 2$ 1

n 7.2 19 5
-1 0 n α n T-
D 19
10 100 n
TD n

7.3 19 5
n 0 -1 n

7.2 n Sarsa

n n Sarsa TD
Sarsa n n Sarsa *Sarsa*
Sarsa(0)

- ε -
n Sarsa 7.3 n TD 7.1
Sarsa n

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}\left(S_{t+n}, A_{t+n}\right), \quad n \geq 1, 0 \leq t < T - n \tag{3.60}$$

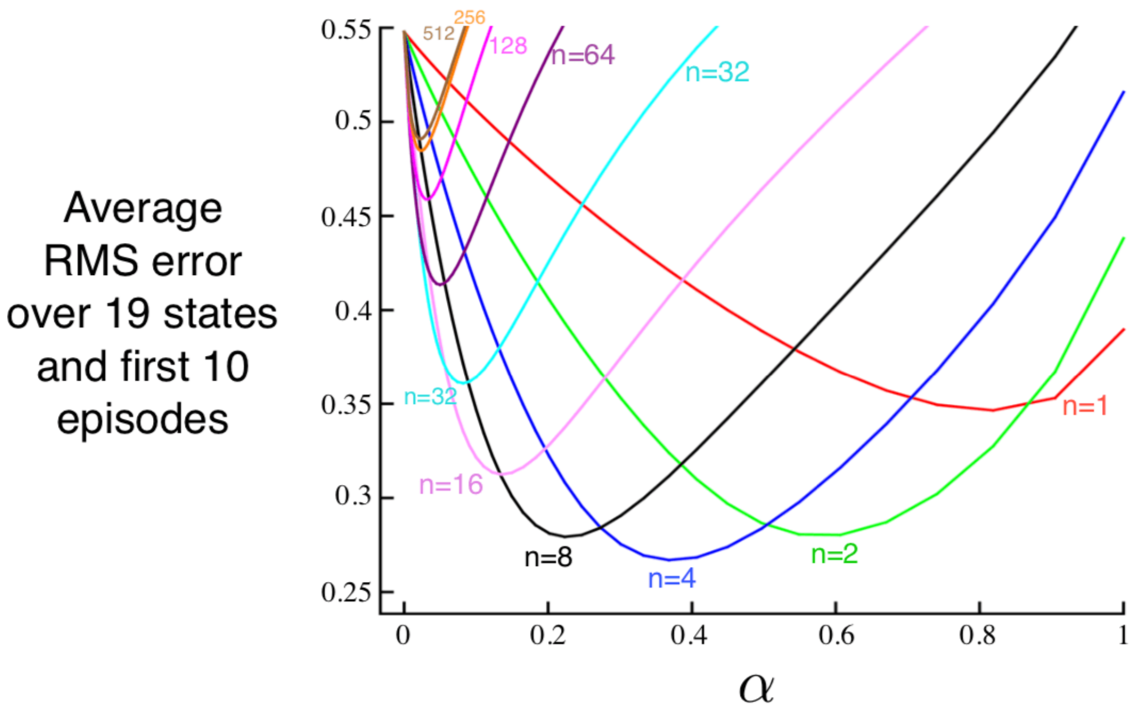
$t + n \geq T$ $G_{t:t+n} \doteq G_t$

$$Q_{t+n}\left(S_t, A_t\right) \doteq Q_{t+n-1}\left(S_t, A_t\right) + \alpha\left[G_{t:t+n} - Q_{t+n-1}\left(S_t, A_t\right)\right], \quad 0 \leq t < T \tag{3.61}$$

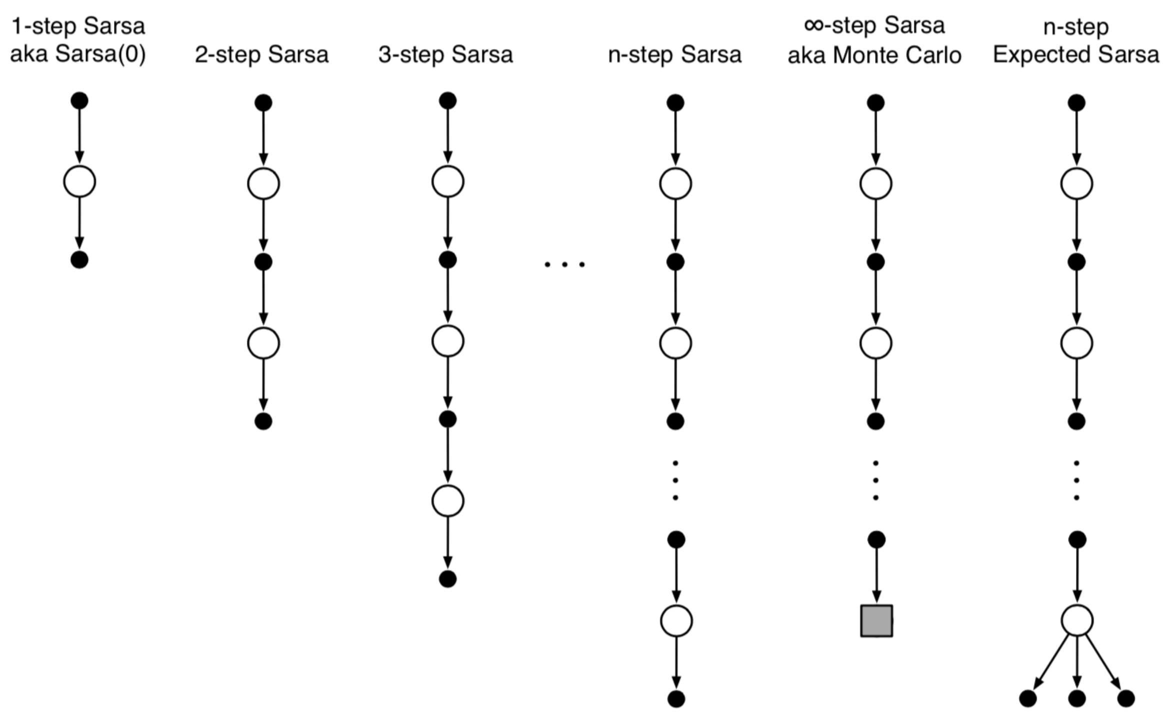
$Q_{t+n}(s,a) = Q_{t+n-1}(s,a)$ s, a
$s \neq S_t$ $a \neq A_t$ *n Sarsa*
7.4

---

n **Sarsa** $Q \approx q_*$ q_π



31: **7.2** 19 n n TD α 7.1



32: **7.3** - n Sarsa(0) n n n - n Sarsa

$s \in (S)$ $a \in (A)$ $Q(s,a)$

π Q ε -

$\alpha \in (0,1]$ $\varepsilon > 0$ n

S_t A_t R_t $mod n+1$

$S_0 \neq$

$A_0 \sim \pi(\cdot|S_0)$

$T \leftarrow \infty$

$t = 0, 1, 2, \ldots$

$t < T$

A_t

R_{t+1} S_{t+1}

S_{t+1}

$T \leftarrow t+1$

$A_{t+1} \sim \pi(\cdot|S_{t=1})$

$\tau \leftarrow t-n+1$ τ

$\tau \geq 0$

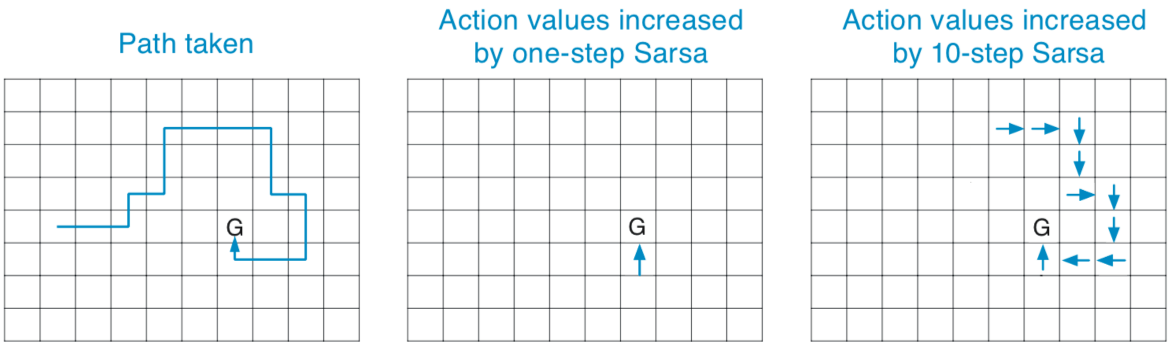
$G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$

$\tau+n < T$ $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$ $(G_{\tau:\tau+n})$

$Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha[G - Q(S_\tau, A_\tau)]$

π $\pi(\cdot|S_\tau)$ Q ε -

$\tau = T-1$



33: **7.4** n **G** 0 **G** Sarsa n Sarsa Sarsa n n

7.4 Sarsa 7.4 n TD

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{\min(t+n,T)-1} \gamma^{k-t}\left[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)\right] \tag{3.62}$$

Sarsa Sarsa n 7.3
n Sarsa π
n Sarsa n

$$G_{t:t+n} \doteq R_{t+1} + \cdots + \gamma^{n-1}R_{t+n} + \gamma^n \overline{V}_{t+n-1}(S_{t+n}), \quad t+n < T \tag{3.63}$$

$t + n \geq T$ $G_{t:t+n} \doteq G_t$ $\overline{V}_t(s)$ s
t

$$\overline{V}_t(s) \doteq \sum_a \pi(a|s)Q_t(s,a), \qquad s \in \mathcal{S} \tag{3.64}$$

s 0

7.3 n

π b π -
- b ε -
b 5.5
n n n n TD
t $t+n$ $\rho_{t:t}+n-1$

$$V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \alpha\rho_{t:t+n-1}\left[G_{t:t+n} - V_{t+n-1}(S_t)\right], \quad 0 \leq t < T \tag{3.65}$$

$\rho_{t:t}+n-1$ A_t A_{t+n-1} n
5.3

$$\rho_{t:h} \doteq \prod_{k=t}^{\min(h,T-1)} \frac{\pi(A_k|S_k)}{b(A_k|S_k)} \tag{3.66}$$

π $\pi(A_k|S_k) = 0$
n π _
π
_ over-
weight 1
7.9 n TD n Sarsa

$$Q_{t+n}(S_t, A_t) \doteq Q_{t+n-1}(S_t, A_t) + \alpha\rho_{t+1:t+n}\left[G_{t:t+n} - Q_{t+n-1}(S_t, A_t)\right] \tag{3.67}$$

$0 \leq t < T$ n TD 7.9 -

---

n **Sarsa** $Q \approx q_*$ q_π

$s \in (S)$ b $b(a|s) > 0$

$s \in (S)$ $a \in (A)$ $Q(s,a)$

π Q

$\alpha \in (0,1]$ n

S_t A_t R_t $modn+1$

$S_0 \neq$

$A_0 \sim \pi(\cdot|S_0)$

$T \leftarrow \infty$

$t = 0,1,2,\ldots$

$t < T$

A_t

R_{t+1} S_{t+1}

S_{t+1}

$T \leftarrow t+1$

$A_{t+1} \sim \pi(\cdot|S_{t=1})$

$\tau \leftarrow t-n+1$ τ

$\tau \geq 0$

$\rho \leftarrow \prod_{i=\tau+1}^{\min(\tau+n-1,T-1)} \frac{\pi(A_i|S_i)}{b(A_i|S_i)}$ $(\rho_\tau + 1 : t+n-1)$

$G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$

$\tau + n < T$ $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$ $(G_{\tau:\tau+n})$

$Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha\rho[G - Q(S_\tau, A_\tau)]$

π $\pi(\cdot|S_\tau)$ Q

$\tau = T-1$

---

n Sarsa n Sarsa $\rho_t + 1 : t+n-1$ $\rho_t + 1 : t+n$ n 7.7 Sarsa Sarsa

7.4 * per-decision

5.9 n 7.1 h n n

$$G_{t:h} = R_{t+1} + \gamma G_{t+1:h}, \quad t < h < T \quad (3.68)$$

$G_{h:h} \doteq V_{h-1}(S_h)$ h
$t+n$ b R_{t+1}
S_{t+1} t $\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$
t π ρ_t
n h
n

$$G_{t:h} \doteq \rho_t \left(R_{t+1} + \gamma G_{t+1:h}\right) + (1-\rho_t) V_{h-1}(S_t), \quad t < h < T \tag{3.69}$$

$G_{h:h} \doteq V_{h-1}(S_h)$
ρ_t
7.13 1 5.9
7.13 n 7.1 ρ_t 1

n 7.13 n TD 7.2

7.5

n

n h 7.7 7.12
$G_{h:h} \doteq \overline{V}_{h-1}(S_h)$ 7.8

$$\begin{aligned} G_{t:h} &\doteq R_{t+1} + \gamma \left(\rho_{t+1} G_{t+1:h} + \overline{V}_{h-1}(S_{t+1}) - \rho_{t+1} Q_{h-1}(S_{t+1}, A_{t+1})\right) \\ &= R_{t+1} + \gamma \rho_{t+1} \left(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\right) + \gamma \overline{V}_{h-1}(S_{t+1}), t < h \le T \end{aligned} \tag{7.14}$$

$h < T$ $G_{h:h} \doteq Q_{h-1}(S_h, A_h)$ $t \geq$
T $G_{T-1:h} \doteq R_T$ 7.5 Sarsa

7.6

** 7.7*

7.8 n 7.13 TD 6.5

7.9 n 7.14 Sarsa
TD 6.9

7.10 7.13 7.2 7.1 7.9

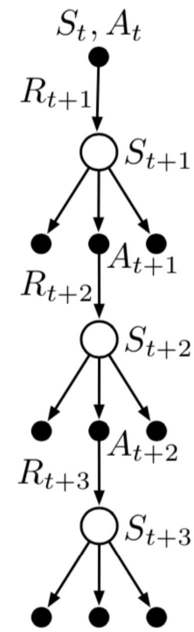
5 sound
-

Autostep Mahmood Sutton Degris Pilarski 2012
Tian TD Karampatziakis Langford 2010
Mahmood 2017 Mahmood Sutton 2015

7.5 n

6 Q-
learning Sarsa
n

3 - S_t A_t



the 3-step tree-backup update

π a $\pi\left(a|S_{t+1}\right)$
A_{t+1} $\pi\left(A_{t+1}|S_{t+1}\right)$ a' $\pi\left(A_{t+1}|S_{t+1}\right)\pi\left(a'|S_{t+2}\right)$
$\pi\left(A_{t+1}|S_{t+1}\right)\pi\left(A_{t+2}|S_{t+2}\right)\pi\left(a''|S_{t+3}\right)$

3 6

n Sarsa

$$G_{t:t+1} \doteq R_{t+1} + \gamma \sum_a \pi\left(a|S_{t+1}\right) Q_t\left(S_{t+1}, a\right) \tag{3.70}$$

$t < T-1$

$$\begin{aligned} G_{t:t+2} &\doteq R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi\left(a|S_{t+1}\right) Q_{t+1}\left(S_{t+1}, a\right) \\ &\quad + \gamma\pi\left(A_{t+1}|S_{t+1}\right)\left(R_{t+2} + \gamma \sum_a \pi\left(a|S_{t+2}\right) Q_{t+1}\left(S_{t+2}, a\right)\right) \\ &= R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi\left(a|S_{t+1}\right) Q_{t+1}\left(S_{t+1}, a\right) + \gamma\pi\left(A_{t+1}|S_{t+1}\right) G_{t+1:t+2} \end{aligned}$$

$t < T-2$ n

$$G_{t:t+n} \doteq R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi\left(a|S_{t+1}\right) Q_{t+n-1}\left(S_{t+1}, a\right) + \gamma\pi\left(A_{t+1}|S_{t+1}\right) G_{t+1:t+n} \tag{3.71}$$

$0 \leq t < T$ - $Q_{t+n}(s,a) = Q_{t+n-1}(s,a)$ s, a $s \neq S_t$ $a \neq A_t$

---

n $Q \approx q_*$ q_π

$s \in (S)$ $a \in (A)$ $Q(s,a)$

π Q

$\alpha \in (0,1]$ n

$mod n+1$

$S_0 \neq$

$A_0 \sim \pi\left(\cdot|S_0\right)$

A_0 S_0 A_0

$T \leftarrow \infty$

$t = 0, 1, 2, \ldots$

$t < T$

A_t R_{t+1} S_{t+1}

S_{t+1}

$T \leftarrow t+1$

A_{t+1} S_{t+1} A_{t+1}

$\tau \leftarrow t+1-n$ τ

$\tau \geq 0$

$t+1 \geq T$:

$G \leftarrow R_T$

$G \leftarrow R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a)$

$k = \min(t, T-1)$ $\tau + 1$

$G \leftarrow R_k + \gamma \sum_{a \neq A_k} \pi(a|S_k) Q(S_k, a) + \gamma\pi(A_k|S_k) G$

π $\pi(\cdot|S_\tau)$ Q

$\tau = T-1$

7.11 7.16 TD

$$G_{t:t+n} = Q(S_t, A_t) + \sum_{k=t}^{\min(t+n-1, T-1)} \delta_k \prod_{i=t+1}^{k} \gamma\pi(A_i|S_i)$$

$\delta_t \doteq R_{t+1} + \gamma\overline{V}_t(S_{t+1}) - Q(S_t, A_t)$ $\overline{V}_t$ 7.8

7.6 * n $Q(\sigma)$

7.5 n Sarsa - n Sarsa

7.5 Sarsa Sarsa Sarsa

$\sigma_t \in [0,1]$ t $\sigma = 1$ $\sigma = 0$ σ_t t - n $Q(\sigma)$

n $Q(\sigma)$ $h+n$ n 7.16 $\overline{V}$ 7.8

$$\begin{aligned} G_{t:h} &= R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q_{h-1}(S_{t+1}, a) + \gamma\pi(A_{t+1}|S_{t+1}) G_{t+1:h} \\ &= R_{t+1} + \gamma\overline{V}_{h-1}(S_{t+1}) - \gamma\pi(A_{t+1}|S_{t+1}) Q_{h-1}(S_{t+1}, A_{t+1}) + \gamma\pi(A_{t+1}|S_{t+1}) G_{t+1:h} \\ &= R_{t+1} + \gamma\pi(A_{t+1}|S_{t+1})(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})) + \gamma\overline{V}_{h-1}(S_{t+1}) \end{aligned}$$

7.14 Sarsa n $\pi(A_{t+1}|S_{t+1})$ ρ_{t+1} $Q(\sigma)$

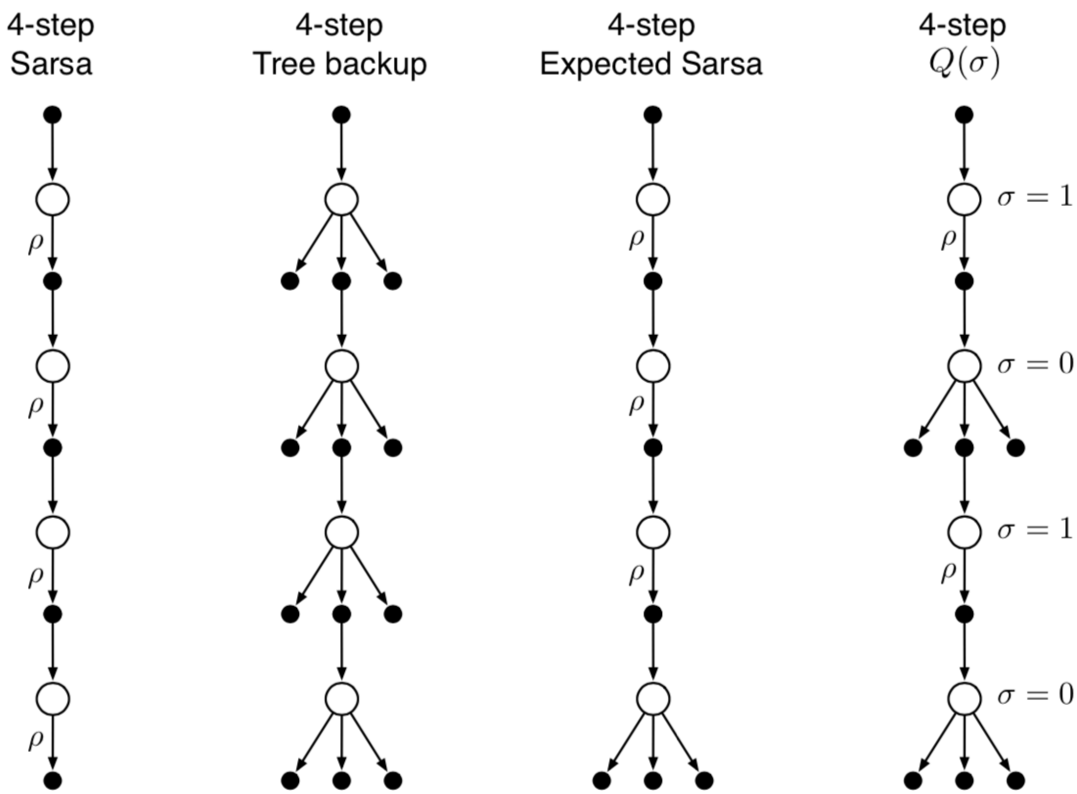
$$\begin{aligned} G_{t:h} \doteq R_{t+1} &+ \gamma(\sigma_{t+1}\rho_{t+1} + (1-\sigma_{t+1})\pi(A_{t+1}|S_{t+1}))(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})) \\ &+ \gamma\overline{V}_{h-1}(S_{t+1}) \end{aligned} \tag{7.17}$$

$t < h \leq T$ $h < T$ $G_{h:h} \doteq Q_{h-1}(S_h, A_h)$ $h = T$ $G_{T-1:T} \doteq R_T$ n Sarsa 7.11

n **Sarsa** $Q \approx q_*$ q_π

$s \in (S)$ b $b(a|s) > 0$

$s \in (S)$ $a \in (A)$ $Q(s,a)$



34: **7.5** n 4 ρ $\sigma_t = 1$ $\sigma_t = 0$

π Q ε -

$\alpha \in (0, 1]$ $\varepsilon > 0$ n

$mod n + 1$

$S_0 \neq$

$A_0 \sim b(\cdot|S_0)$

$T \leftarrow \infty$

$t = 0, 1, 2, \ldots$

$t < T$

A_t R_{t+1} S_{t+1}

S_{t+1}

$T \leftarrow t + 1$

$A_{t+1} \sim b(\cdot|S_{t+1})$

σ_{t+1}

$\frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})}$ ρ_{t+1}

$\tau \leftarrow t - n + 1$ τ

$\tau \geq 0$

$G \leftarrow 0$

$k = \min(t + 1, T)$ $\tau + 1$

$k = T$

$G \leftarrow R_T$

$\overline{V} \leftarrow \sum_a \pi(a|S_k) Q(S_k, a)$

$G \leftarrow R_k + \gamma(\sigma_k \rho_k + (1 - \sigma_k) \pi(A_k|S_k))(G - Q(S_k, A_k)) + \gamma \overline{V}$

$Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha [G - Q(S_\tau, A_\tau)]$

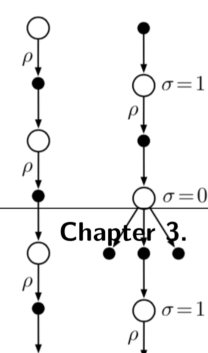
π $\pi(\cdot|S_\tau)$ Q

$\tau = T - 1$

7.7

TD

n n 4 n TD n $Q(\sigma)$ Sarsa Q-learning n n n n 12 TD



n n
Q-learning n

n Watkins 1989 n Cichosz 1995 van Seijen 2016 12

7.1-2 Sutton 1988 Singh Sutton 1996

7.3-5 Precup Sutton Singh 2000 Precup Sutton Dasgupta 2001 Sutton Mahmood Precup van Hasselt 2014

Precup Sutton Singh 2000

7.6 $Q(\sigma)$ De Asis Hernandez-Garcia Holland Sutton 2017

3.5.7 8

model-based *model-free* * * *planning* *learning*
n

8.1

MDP $p\left(s', r|s, a\right)$
5

$$\longrightarrow$$

“ ” Russell Norvig 2010
1 2

$$\longrightarrow \quad longrightarrow \quad \longrightarrow$$

Q-learning
Q-planning Q-learning 1 - α

Q-planning

1. $S \in \mathcal{S}$ $A \in \mathcal{A}$
2. S, A R S'
3. S, A, R, S' Q-learning

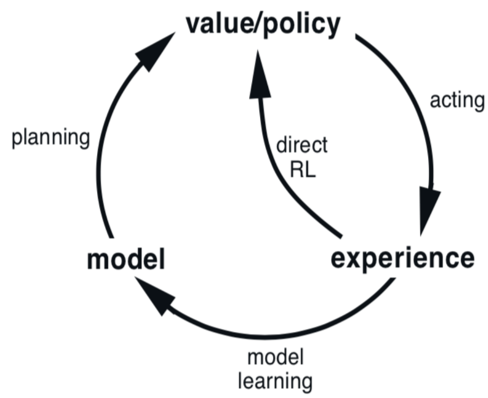
$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$

8.2 Dyna

Dyna-Q Dyna-Q

RL

14



Dyna-Q RL Q-planning RL Q-learning $S_t, A_t \rightarrow R_{t+1}, S_{t+1}$ S_t, A_t S_{t+1}, A_{t+1} - Q-planning - 1 -

Dyna Dyna-Q 8.1

Dyna-Q " "

RL Dyna Dyna-Q RL RL Q-plainning n 1-3 Dyna-Q $Model(s,a)$ - (s,a) d e f e f Q-learning

Dyna-Q

$s \in \mathcal{S}$ $a \in \mathcal{A}$ $Q(s,a)$ $Model(s,a)$

a $S \leftarrow$

b $A \leftarrow \varepsilon$- (S,Q)

c A R S'

d $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$

e $Model(S,A) \leftarrow R, S'$

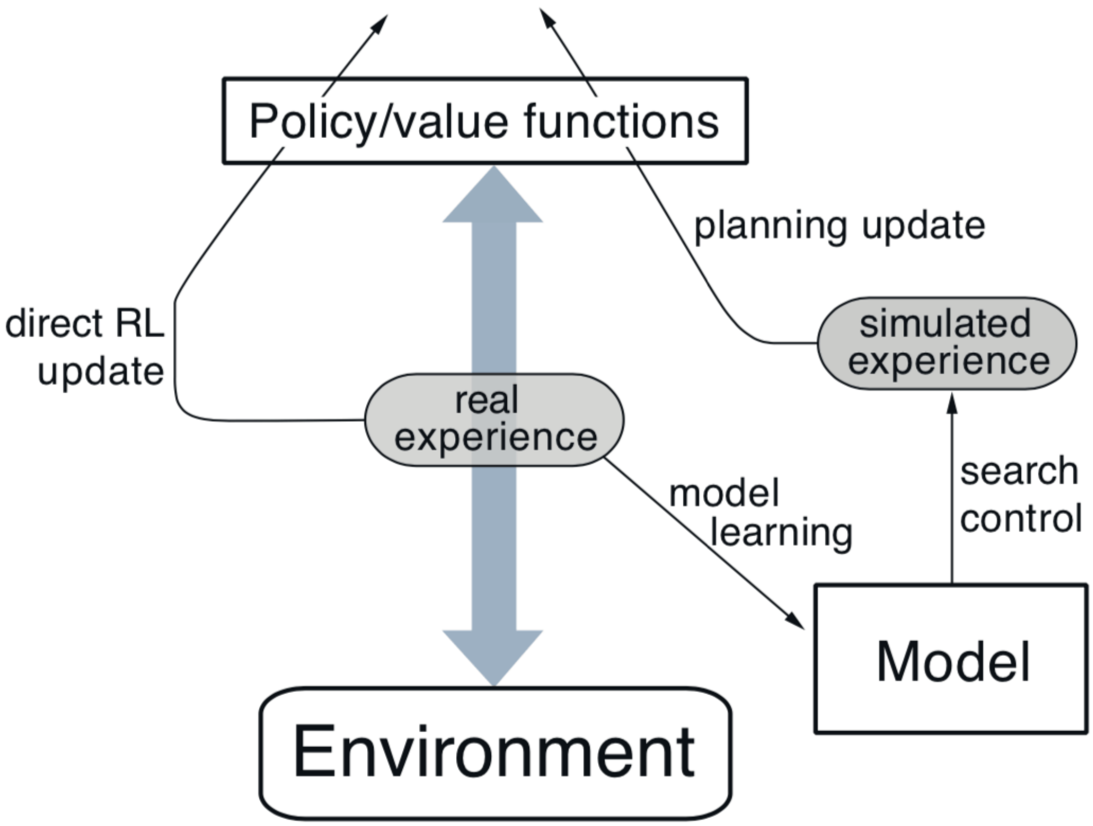
f n

$S \leftarrow$

$A \leftarrow$ S

$R, S' \leftarrow Model(S,A)$

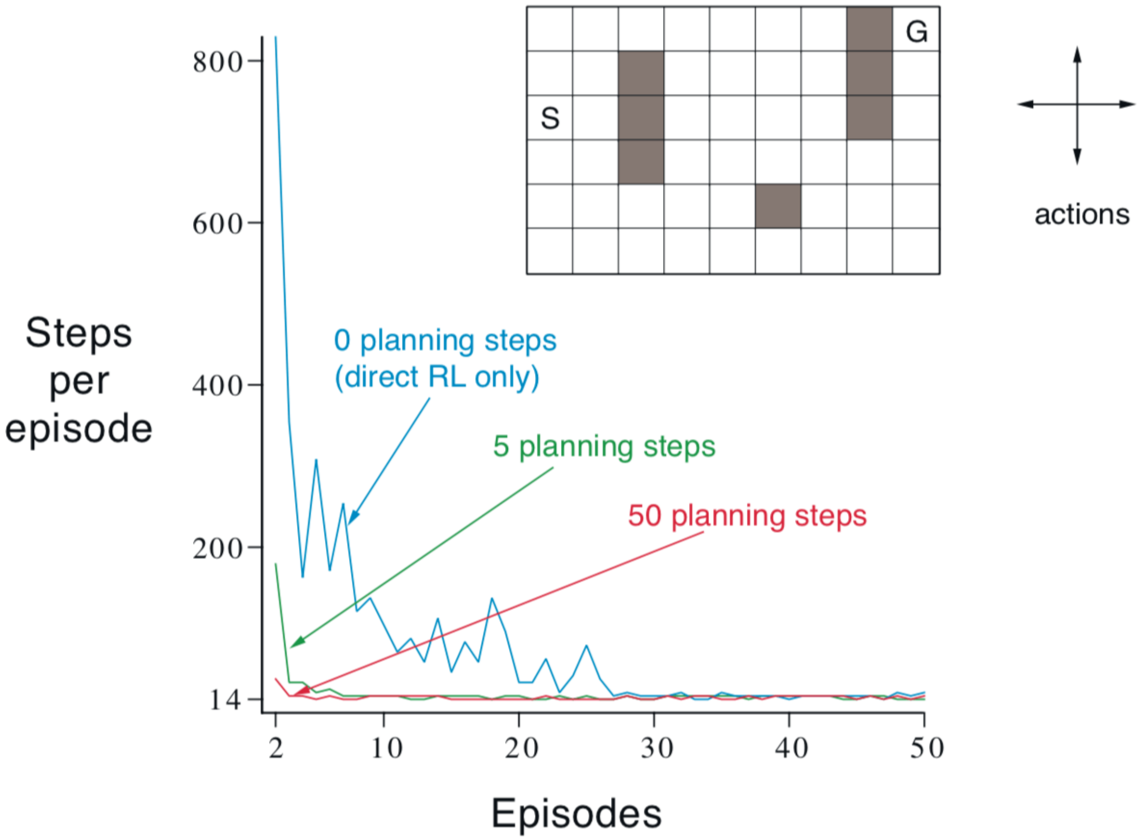
$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$


Policy/value functions
planning update
direct RL
update
simulated
experience
real
experience
model
learning
search
control
Model
Environment

35: **8.1** Dyna

8.1 Dyna 8.2 47 +1 **G** **S** $\gamma = 0.95$

8.2 Dyna-Q $\alpha = 0.1$ $\varepsilon = 0.1$ ties n n 30 n 1700 n $n = 0$ Q-learning α ε 25 ε - $n = 5$ 5 $n = 50$ 3



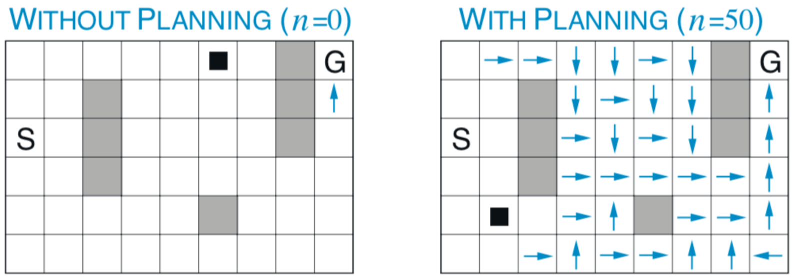
36: **8.2** Dyna-Q n **S** **G**

8.3 $n = 0$ $n = 50$ $n = 0$

Dyna-Q

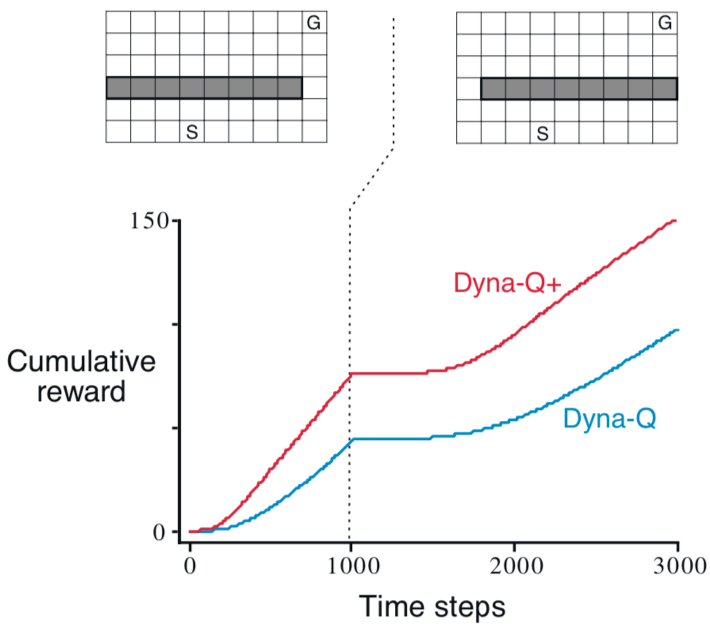
8.1 8.3 7 Dyna

8.3



37: **8.3** Dyna-Q

8.2 8.4 1000 “ ” Dyna-Q Dyna-Q+ Dyna 1000



38: **8.4** Dyna 1000 Dyna-Q+ Dyna-Q

8.3 8.5 3000 Dyna-Q ε -

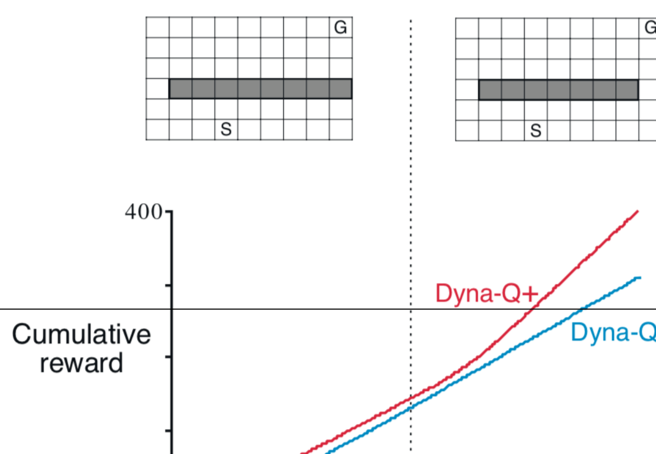
/

Dyna-Q+

-

“ ”

r τ $r + \kappa\sqrt{\tau}$



κ [1]

8.2 Dyna-Q+ Dyna

8.3 8.5 Dyna-Q+ Dyna-Q

8.4 $\kappa\sqrt{\tau}$ $Q(S_t, a)$ + $\kappa\sqrt{\tau(S_t, a)}$

8.5 8.2 Dyna-Q

8.4

Dyna - - - 8.3 -

“ ” g *backward focusing* - - - - - -

---

s, a $Q(s, a)$ $Model(s, a)$ $PQueue$ 0

a $S \leftarrow$

b $A \leftarrow$ (S, Q)

c A R S'

d $Model(S, A) \leftarrow R, S'$

e $P \leftarrow |R + \gamma \max_a Q(S', a) - Q(S, A)|$

f $P > \theta$ S, A $PQueue$ P

g n $PQueue$

$S, A \leftarrow$ first $(PQueue)$

$R, S' \leftarrow Model(S, A)$

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$

S $\overline{S}, \overline{A}$

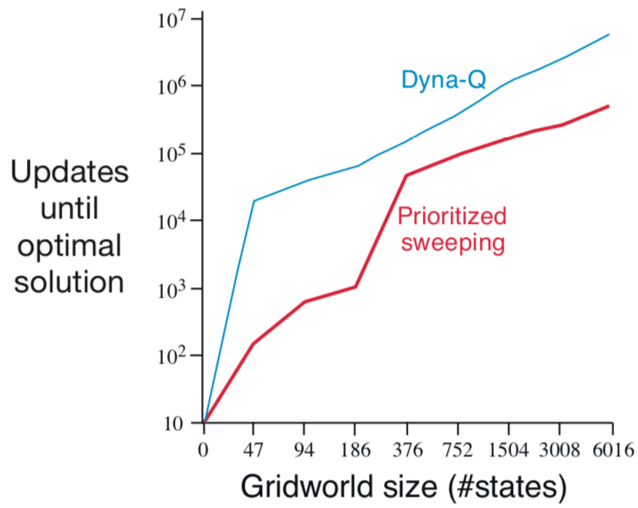
[1] Dyna-Q+ Dyna-Q f

$\overline{R} \leftarrow \overline{S}, \overline{A}, S$

$P \leftarrow \left|\overline{R} + \gamma \max_a Q(S,a) - Q(\overline{S}, \overline{A})\right|$

$P > \theta$ $\overline{S}, \overline{A}$ $PQueue$ P

8.4 5 10
8.2
Dyna-Q $n = 5$
Peng Williams 1993
-

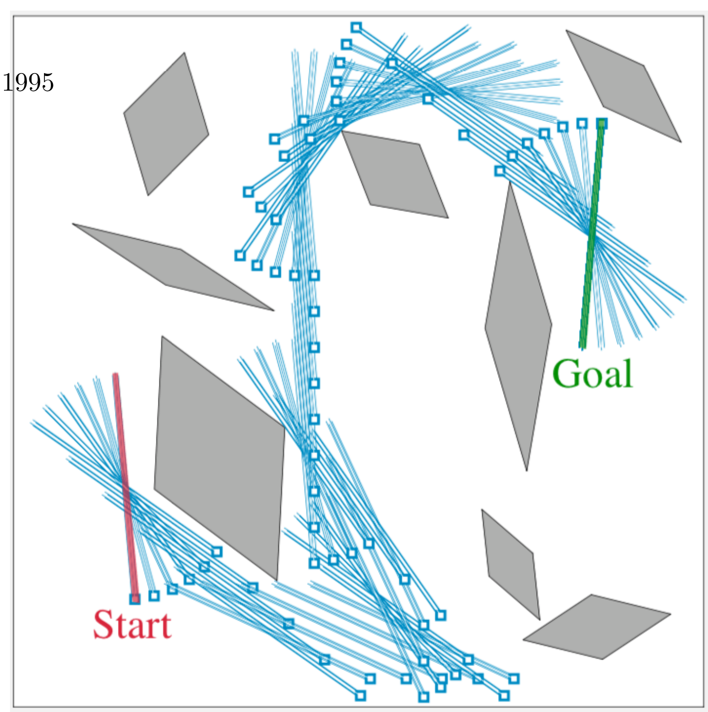


8.5

1/20 10
Translations 20 $\times$
20
14,400
Moore Atkeson 1993

van Seijen Sutton 2013 “ ”

Peng Williams 1993 Barto Bradtke Singh 1995



8.5

q_* v_* q_π v_π

Dyna-Q q_* q_*
q_π Dyna-AC v_π 13

6

q_* :math‘Q‘
$\hat{p}\left(s^{\prime}, r \mid s, a\right)$ - s, a

$$Q(s, a) \leftarrow \sum_{s^{\prime}, r} \hat{p}\left(s^{\prime}, r \mid s, a\right)\left[r+\gamma \max _{a^{\prime}} Q\left(s^{\prime}, a^{\prime}\right)\right] \tag{3.72}$$

s, a S_0 R Q-learning

$$Q(s, a) \leftarrow Q(s, a)+\alpha\left[R+\gamma \max _{a^{\prime}} Q\left(S^{\prime}, a^{\prime}\right)-Q(s, a)\right] \tag{3.73}$$

α

$\alpha = 1$
In favor of
$Q(s, a)$ $Q(s\prime, a_0)$
Q
- s, a b *branching factor*
$\hat{p}(s'|s, a) > 0$ $s\prime$ -
b

b
b
- - -
- b

8.7 b
b 1
$\sqrt{\frac{b-1}{bt}}$ t
$\alpha = \frac{1}{t}$ b b -
-
8.7

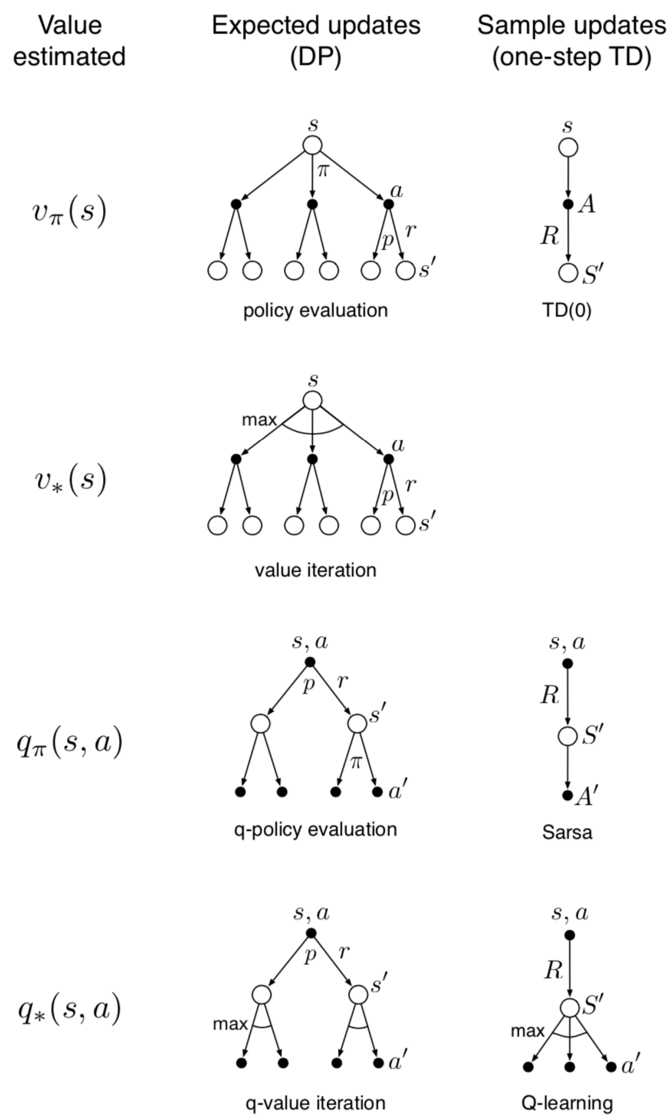
8.6 b b

8.6

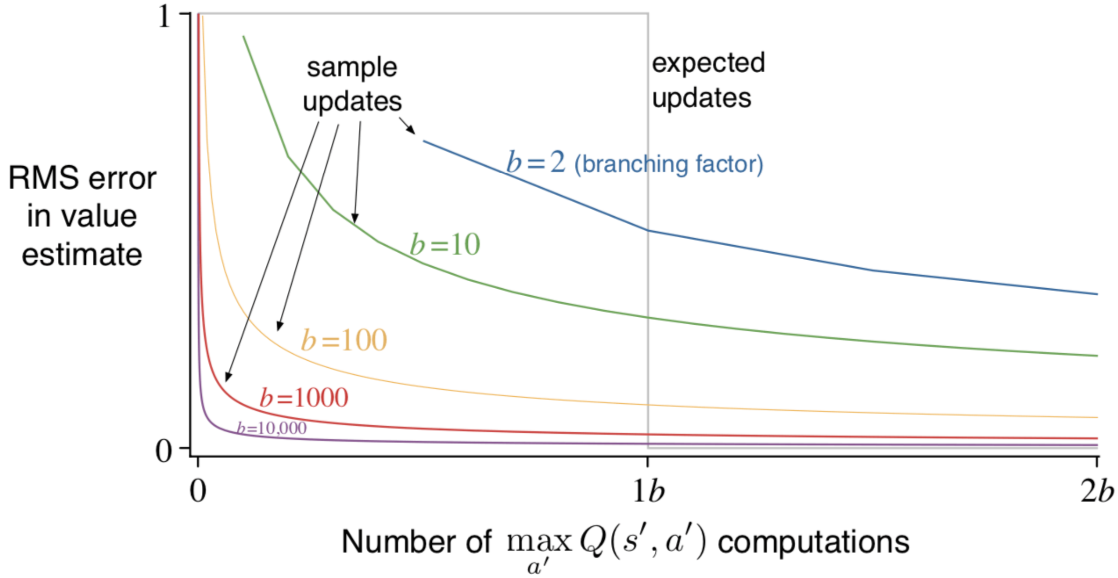
-
-

4
-
8.7

- Dyna-Q



40: **8.6**



41: **8.7**

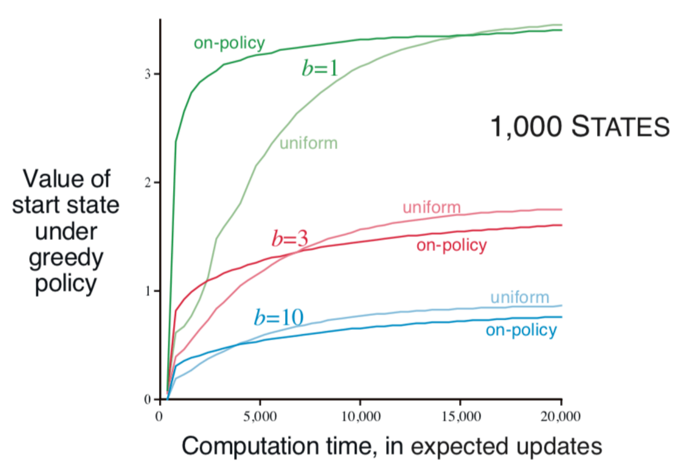
-

-

8.1
- ε - ε = 0.1 - $|\mathcal{S}|$
b - b - b 0.1
0 1 $v_{\tilde{\pi}}(\mathcal{S}_0)$ $\tilde{\pi}$ - Q

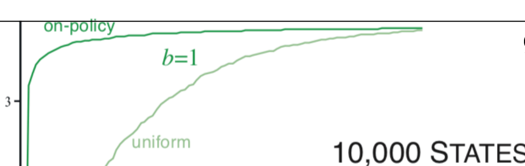
200 1000 1,3 10

10,000 1



8.7 8.8 $b = 1$

8.8 8.8 $b = 3$



8.7

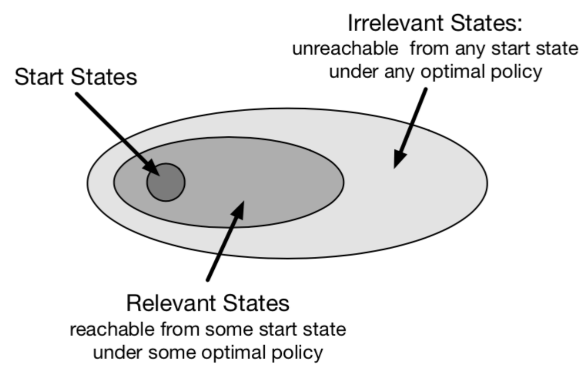
RTDP DP
RTDP
RTDP 4.10
8.8

RTDP DP
RTDP 4.5 DP D-
P
RTDP

Sarsa 6.4
-
5.3
RTDP RTDP MDP
-

RTDP RTDP

MDP 3.4
RTDP



RTDP 1
1 2 1
3 4
Barto Bradtke Singh 1995 DP
A^* Korf 1990

-1
3.5

8.6 RTDP 5.12 RTDP D-
P

5.5

5.12 -1

5.5 9,115 599 10^7

DP RTDP 25 DP Gauss-
Seidel Jacobi 4.8 10^{-4} DP
20 RTDP RTDP

–	DP	RTDP
	28	4000
	252,784	127,600
	–	31.9
≤ 100	–	98.45
≤ 10	–	80.51
0	–	3.18

14 15 RTDP DP RTDP DP RTDP RTDP 98.45 100 80.51 10 290

RTDP v_* v_* 4 v_* DP

DP DP 15 136,725 DP v_* 252,784 RTDP 127,600

RTDP RTDP RTD-P RTDP RTDP 50

8.8

Dyna S_t S_t *background planning*

S_t A_t S_{t+1} A_{t+1} 1 *decision-time planning*

8.1

8.9

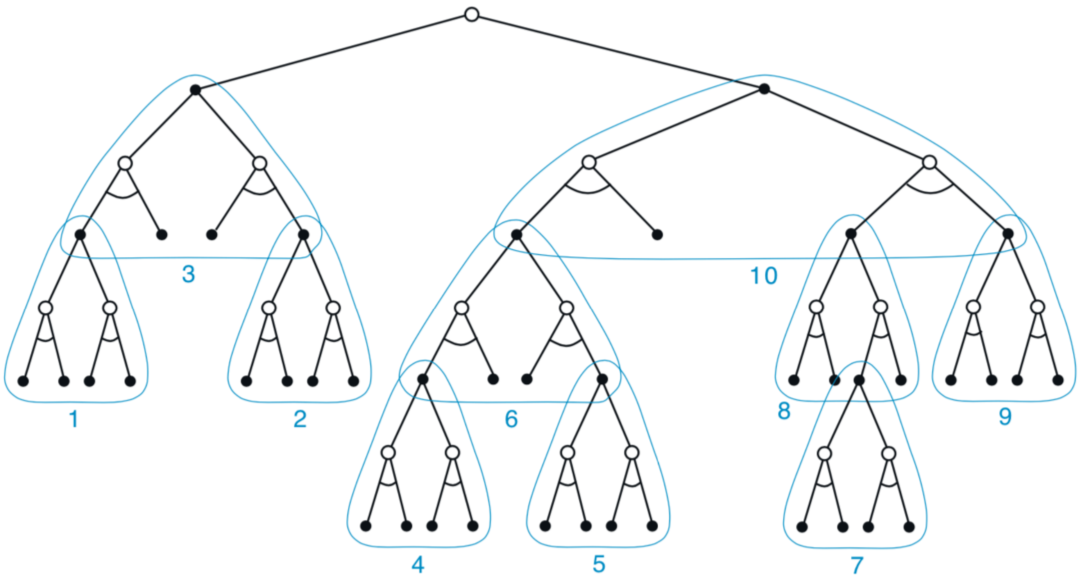
maxes v_* q_*

ε - UCB 2.7

- [2] k γ^k Tesauro TD-Gammon 16.1 TD afterstate TD-Gammon TD-Gammon Tesauro TD-Gammon

8.9

[2] Pearl 1984



43: **8.9**

8.10 Rollout

Rollout
Tesauro Galperin 1997 Rollout "Rollout" "Rollout"

5 Rollout q_* q_π *rollout rollout policy*
Rollout Rollout -

Rollout 4.2 π π' s $\pi'(s) = a \neq \pi(s)$ $q_\pi(s,a) \geq v_\pi(s)$ π' π π π' π Rollout s π rollout $a' \in \mathcal{A}(s)$ $q_\pi(s,a')$ s π π 4.3 4.5

Rollout rollout Rollout Tesauro Galperin 1997 rollout rollout Rollout rollout rollout Rollout Gelly Silver 2007

rollout Rollout Rollout roll-out

rollout Te-sauro Galperin 1997 Tesauro Galperin

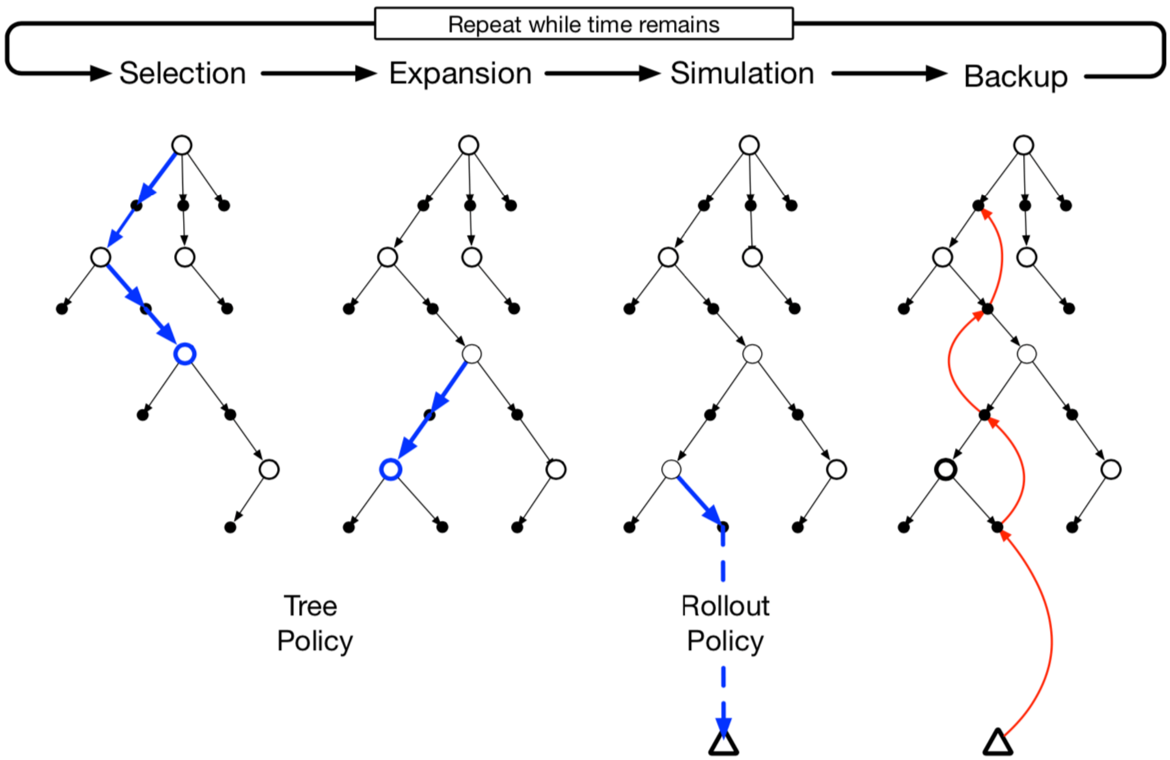
Rollout
Rollout

8.11

MCTS MCTS rollout MCTS 2005 2015 6 16.6 2016 AlphaGo 18 MCTS Finnsson Bjrnsson 2008 Genesereth Thi

MCTS rollout MCT-S MCTS

rollout rollout rollout -
- 8.10 MCTS
rollout ε - UCB 2



44: 8.10 MCTS
Chaslot Bakkes Szita Spronck 2008

MCTS 8.10

1.

2.

3. rollout rollout

4. MCTS rollout 8.10 rollout -
-

MCTS
MCTS MCTS MCTS

MCTS Go rollout 16.6 AlphaGo MCTS MCT-
S

MCTS MCTS rollout
MCTS -
MCTS -

MCTS

8.12

8.2

Dyna -

-

8.13

GPI

GPI

8.11

8.11 bootstrapping TD TD n

12 n λ -

8.11

afterstate -

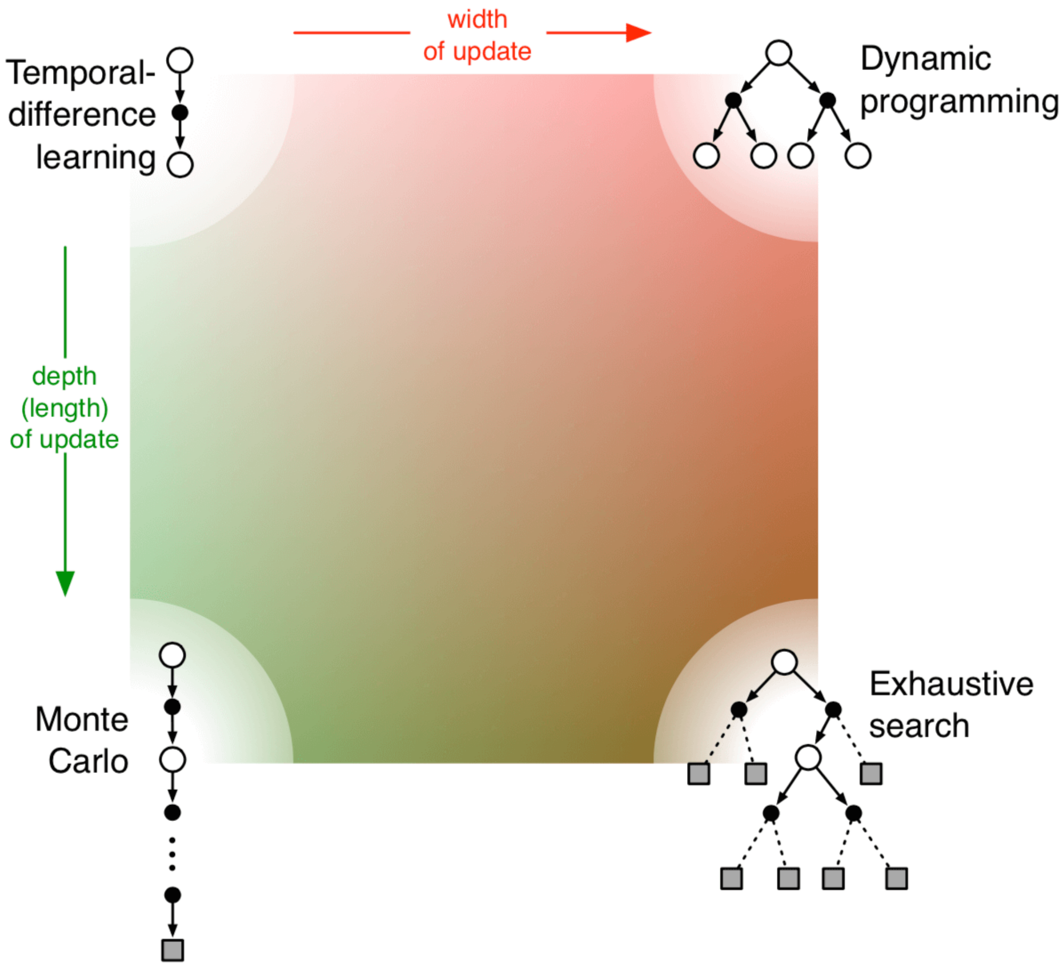
/ ε - soft-max

- -

Dyna

8.1 Sutton 1990,1991a 1991b Barto Bradtke Singh 1991,1995 Sutton Pinette 1985 Sutton Barto 1981b Agre Chapman 1990 Agre 1988 Bertsekas Tsitsiklis 1989 Singh 1993 Tolman 1932 Galanter Gerstenhaber 1956; Craik 1943 Campbell 1960; Dennett 1978 14.6 15.11

8.2 Goodwin Sin 1984 Goodwin Sin 1984; Ljung Soderstrom 1983; Young 1984 Dyna Sutton 1990 Barto and


width
of update
Temporal-
difference
learning
Dynamic
programming
depth
(length)
of update
Monte
Carlo
Exhaustive
search

45:   8.11

Singh 1990 Dyna 9.4 Sutton Szepesv́ari Geramifard Bowling 2008 Parr Li Taylor Painter-Wakefield Littman 2008

8.3 Kearns Singh E^3 2002 Brafman Tennenholtz 2003 R-max

8.4 Moore Atkeson 1993 Peng Williams 1993 170 Peng and Williams 1993 171 Moore Atkeson McMahan Gordon 2005 van Seijen Sutton 2013

8.5 Singh 1993

8.6-7 Barto Bradtke Singh 1995 RTDP Korf 1990 A^* LRTA* DP Korf LRTA* RTDP Barto 1995 Korf 1990 LRTA* Bertsekas 1982 Bertsekas Tsitsiklis 1989 DP RTDP RTDP Barto 1995 Barto 2011

8.9 Russell Norvig 2009 Korf 1988 Peng Williams 1993

8.10 Abramson 1990 Rollout " " " " Tesauro Galperin 1997 Rollout "Rollout" Bertsekas Tsitsiklis Wu 1997 Rollout Bertsekas 2013 " "

8.11 MCT Coulom 2006 Kocsis Szepesv́ari 2006 Browne Powley Whitehouse Lucas Cowling Rohlfshagen Tavener Perez Samothrakis Colton 2012 MCTS David Silver

3.6

9 10 11 12 -

3.6.1 9

- π v_π $\mathbf{w} \in \mathbb{R}^d$ $\mathbf{w}$ s $\hat{v}(s,\mathbf{w}) \approx v_\pi(s)$ $\hat{v}$ $\mathbf{w}$ $\hat{v}$ $\mathbf{w}$ $\hat{v}$ $\mathbf{w}$ $\mathbf{w}$ $d \ll |\mathcal{S}|$

$\hat{v}$ 17.3

9.1

" " $s \mapsto u$ s u s $S_t \mapsto G_t$ TD(0) $S_t \mapsto R_{t+1} + \gamma\hat{v}(S_{t+1},\mathbf{w}_t)$ n TD $S_t \mapsto G_{t:t+n}$ DP $s \mapsto \mathbb{E}_\pi\left[R_{t+1} + \gamma\hat{v}(S_{t+1},\mathbf{w}_t)|S_t = s\right]$ s S_t

- $s \mapsto u$ s u s u
s u - $s \mapsto u$

GPI π
q_π DP TD

9.2 $\overline{\mathrm{VE}}$

$\mu(s) \geq 0, \sum_s \mu(s) = 1$ s s
$\hat{v}(s, \mathbf{w})$ $v_\pi(s)$ μ $\overline{\mathrm{VE}}$

$$\overline{\mathrm{VE}}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w})\right]^2 \tag{3.74}$$

$\overline{\mathrm{VE}}$ $\mu(s)$ s math:pi

;math:$h(s)$ s $\eta(s)$ s s $\bar{s}$ s s

$$\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_a \pi(a|\bar{s}) p(s|\bar{s}, a), \qquad s \in \mathcal{S} \tag{3.75}$$

$\eta(s)$

$$\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}, \qquad s \in \mathcal{S} \tag{3.76}$$

$\gamma < 1$ 9.2 γ

$\overline{\mathrm{VE}}$ $\overline{\mathrm{VE}}$ $\overline{\mathrm{VE}}$
$\overline{\mathrm{VE}}$ $\mathbf{w}^*$ $\mathbf{w}$ $\overline{\mathrm{VE}}(\mathbf{w}^*) \leq \overline{\mathrm{VE}}(\mathbf{w})$
$\mathbf{w}$ $\mathbf{w}^*$ $\mathbf{w}$ $\overline{\mathrm{VE}}(\mathbf{w}^*) \leq \overline{\mathrm{VE}}(\mathbf{w})$
$\overline{\mathrm{VE}}$
$\overline{\mathrm{VE}}$

9.3

SGD SGD
$\mathbf{w} \doteq (w_1, w_2, \ldots, w_d)^\top$[1] $\hat{v}(s, \mathbf{w})$ $s \in \mathcal{S}$ $\mathbf{w}$ $t = 0, 1, 2, 3, \ldots,$ $\mathbf{w}$
$\mathbf{w}_t$ $S_t \mapsto v_\pi(S_t)$ S_t S_t
$v_\pi(S_t)$ resolution $\mathbf{w}$
μ 9.1 $\overline{\mathrm{VE}}$ SGD

$$\begin{aligned}
\mathbf{w}_{t+1} &\doteq \mathbf{w}_t - \frac{1}{2}\alpha \nabla \left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right]^2 && (9.4)\\
&= \mathbf{w}_t + \alpha \left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)\right] \nabla \hat{v}(S_t, \mathbf{w}_t) && (9.5)
\end{aligned}$$

[1] $\top$

α $\mathbf{w}$ $f(\mathbf{w})$ $\nabla f(\mathbf{w})$

$$\nabla f(\mathbf{w}) \doteq \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \frac{\partial f(\mathbf{w})}{\partial w_2}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_d}\right)^\top \tag{3.77}$$

math:f $\mathbf{w}$ SGD " " $\mathbf{w}_t$ 9.4 " " VE

SGD

SGD α 2.7 SGD 9.5

t $S_t \mapsto U_t$ $U_t \in \mathbb{R}$ $v_\pi(S_t)$ U_t $v_\pi(S_t)$ $\hat{v}$ 9.5 $v_\pi(S_t)$ U_t $v_\pi(S_t)$ SGD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[U_t - \hat{v}\left(S_t, \mathbf{w}_t\right)\right] \nabla \hat{v}\left(S_t, \mathbf{w}_t\right) \tag{3.78}$$

U_t t $\mathbb{E}\left[U_t | S_t = s\right] = v_\pi(S_t)$ $\mathbf{w}_t$ α 2.7

π $U_t \doteq G_t$ $v_\pi(S_t)$ SGD 9.7 $v_\pi(S_t)$

$\hat{v} \approx v_\pi$

π

$\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$

$\alpha > 0$

$\mathbf{w} \in \mathbb{R}^d$ $\mathbf{w} = \mathbf{0}$

π $S_0, A_0, R_1, S_1, A_1, \ldots, R_T, S_T$

$t = 0, 1, \ldots, T-1$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[G_t - \hat{v}(S_t, \mathbf{w})\right] \nabla \hat{v}(S_t, \mathbf{w})$

$v_\pi(S_t)$ 9.7 U_t n $G_{t:t+n}$ DP $\sum_{a,s',r} \pi(a|S_t) p(s', r|S_t, a) \left[r + \gamma \hat{v}(s', \mathbf{w}_t)\right]$ $\mathbf{w}_t$ 9.4 9.5 $\mathbf{w}_t$ $v_\pi(S_t)$ Barnard 1993 $\mathbf{w}_t$

6 7

TD(0) $U_t \doteq R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w})$

TD(0) $\hat{v} \approx v_\pi$

π

$\hat{v} : \mathcal{S}^+ \times \mathbb{R}^d \to \mathbb{R}$ $\hat{v}(\ , \cdot) = 0$

$\alpha > 0$

$\mathbf{w} \in \mathbb{R}^d$ $\mathbf{w} = \mathbf{0}$

S

$A \sim \pi(\cdot|S)$

A R S'

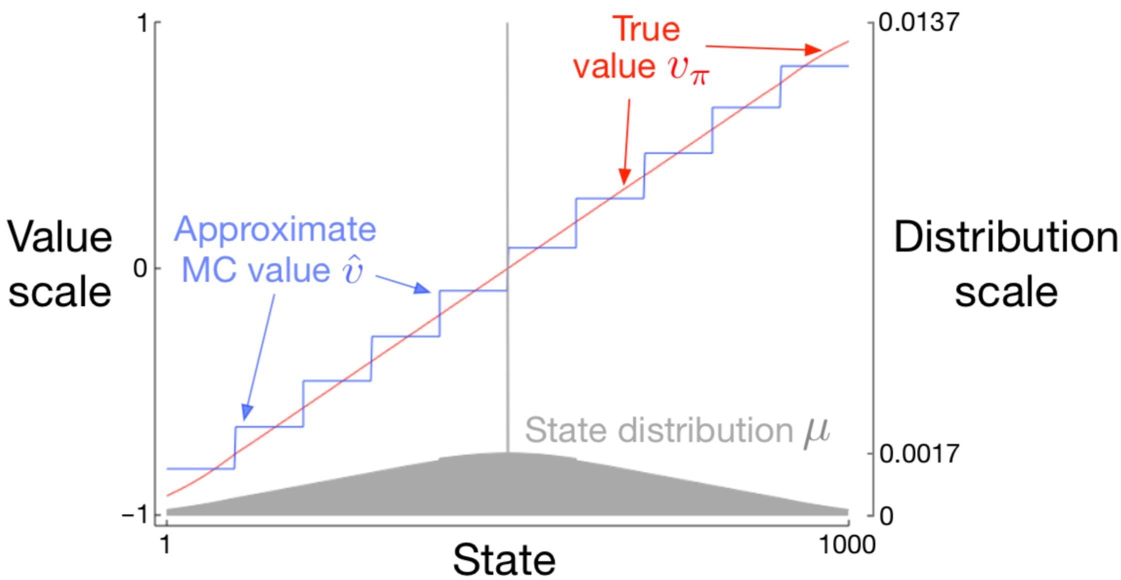
$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})\right] \nabla \hat{v}(S, \mathbf{w})$

$S \leftarrow S'$

S

$\mathbf{w}$ SGD 9.7 $\nabla \hat{v}(S_t, \mathbf{w}_t)$ S_t 1 0

9.1 1000 1000 6.2 7.1 1 1000 500 100 100 100 1 0.5 950 0.25 -1 $+1$



46: **9.1** 1000

9.1 v_π 100 α $=$ 20 $-$ 5 100,000 1000 10 100 1-100 101-200 $\overline{\text{VE}}$ 9.1

μ 500 1.37 0.17 μ 1 1000 0.0147 μ 100 1 3 100 1

9.4

$\hat{v}(\cdot, \mathbf{w})$ $\mathbf{w}$ s $\mathbf{x}(s) \doteq (x_1(s), x_2(s), \ldots, x_d(s))^\top$ $\mathbf{w}$ $\mathbf{w}$ $\mathbf{x}(s)$

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^\top \mathbf{x}(s) \doteq \sum_{i=1}^{d} w_i x_i(s) \tag{3.79}$$

linear in the weights

$\mathbf{x}(s)$ s $\mathbf{x}(s)$ $x_i(s)$ $x_i : \mathcal{S} \to \mathbb{R}$ s d
d

SGD $\mathbf{w}$

$$\nabla \hat{v}(s, \mathbf{w}) = \mathbf{x}(s)$$

SGD 9.7

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[U_t - \hat{v}\left(S_t, \mathbf{w}_t\right)\right] \mathbf{x}\left(S_t\right)$$

SGD

α $\overline{\mathrm{VE}}$

TD(0) SGD t

$$\begin{aligned} \mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^{\top} \mathbf{x}_{t+1} - \mathbf{w}_t^{\top} \mathbf{x}_t\right) \mathbf{x}_t \qquad (9.9) \\ &= \mathbf{w}_t + \alpha \Big(R_{t+1} \mathbf{x}_t - \mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right)^{\top} \mathbf{w}_t\Big) \end{aligned}$$

$\mathbf{x}_t = \mathbf{x}\left(S_t\right)$ $\mathbf{w}_t$

$$\mathbb{E}\left[\mathbf{w}_{t+1} | \mathbf{w}_t\right] = \mathbf{w}_t + \alpha\left(\mathbf{b} - \mathbf{A} \mathbf{w}_t\right) \tag{3.80}$$

$$\mathbf{b} \doteq \mathbb{E}\left[R_{t+1} \mathbf{x}_t\right] \in \mathbb{R}^d \qquad \mathbf{A} \doteq \mathbb{E}\left[\mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right)^{\top}\right] \in \mathbb{R}^d \times \mathbb{R}^d \tag{3.81}$$

9.10 $\mathbf{W}_{\mathrm{TD}}$

$$\begin{aligned} \mathbf{b} - \mathbf{A} \mathbf{w}_{\mathrm{TD}} &= \mathbf{0} \\ \Rightarrow \mathbf{b} &= \mathbf{A} \mathbf{w}_{\mathrm{TD}} \\ \Rightarrow \mathbf{w}_{\mathrm{TD}} &\doteq \mathbf{A}^{-1} \mathbf{b} \qquad (9.12) \end{aligned}$$

TD TD(0)

TD(0)

TD(0) 9.9 9.10

$$\mathbb{E}\left[\mathbf{w}_{t+1} | \mathbf{w}_t\right] = (\mathbf{I} - \alpha \mathbf{A}) \mathbf{w}_t + \alpha \mathbf{b} \tag{3.82}$$

$\mathbf{A}$ $\mathbf{w}_t$ $\mathbf{b}$ $\mathbf{A}$ $\mathbf{A}$ $\mathbf{I} - \alpha \mathbf{A}$ 1 $\mathbf{w}_t$ $\mathbf{A}$
α $\mathbf{I} - \alpha \mathbf{A}$ 0 1 $\mathbf{w}_t$ $\mathbf{A}$ $y \neq 0$ $y^{\top} \mathbf{A} y > 0$ $\mathbf{w}_t$
$\mathbf{A}^{-1}$

TD(0) $\gamma < 1$ $\mathbf{A}$ 9.11

$$\begin{aligned} \mathbf{A} &= \sum_s \mu(s) \sum_a \pi(a|s) \sum_{r, s'} p\left(r, s' | s, a\right) \mathbf{x}(s) \left(\mathbf{x}(s) - \gamma \mathbf{x}\left(s'\right)\right)^{\top} \\ &= \sum_s \mu(s) \sum_{s'} p\left(s' | s\right) \mathbf{x}(s) \left(\mathbf{x}(s) - \gamma \mathbf{x}\left(s'\right)\right)^{\top} \\ &= \sum_s \mu(s) \mathbf{x}(s) \left(\mathbf{x}(s) - \gamma \sum_{s'} p\left(s' | s\right) \mathbf{x}\left(s'\right)\right)^{\top} \\ &= \mathbf{X}^{\top} \mathbf{D} (\mathbf{I} - \gamma \mathbf{P}) \mathbf{X} \end{aligned}$$

$\mu(s)$ π $p(s'|s)$ π s s' $\mathbf{P}$ $|\mathcal{S}| \times |\mathcal{S}|$ $\mathbf{D}$ $|\mathcal{S}| \times |\mathcal{S}|$ $\mu(s)$ $\mathbf{X}$ $|\mathcal{S}| \times d$ $\mathbf{x}(s)$ $\mathbf{D}(\mathbf{I} - \gamma\mathbf{P})$ $\mathbf{A}$

key matrix Sutton 1988 p.27 $\mathbf{S} = \mathbf{M} + \mathbf{M}^\top$ $\mathbf{M}$ Sutton 1988 $\mathbf{S}$ Varga 1962 23 $\mathbf{D}(\mathbf{I} - \gamma\mathbf{P})$ $\mathbf{P}$ $\gamma < 1$ $\mathbf{M}$ $\mathbf{1}^\top\mathbf{M}$ $\mathbf{1}$ 1 μ $\mu(s)$ $|\mathcal{S}|$ - $\mu = \mathbf{P}^\top\mu$ μ

$$\begin{aligned}
\mathbf{1}^\top\mathbf{D}(\mathbf{I} - \gamma\mathbf{P}) &= \boldsymbol{\mu}^\top(\mathbf{I} - \gamma\mathbf{P}) \\
&= \boldsymbol{\mu}^\top - \gamma\boldsymbol{\mu}^\top\mathbf{P} \\
&= \boldsymbol{\mu}^\top - \gamma\boldsymbol{\mu}^\top \quad \mu \\
&= (1 - \gamma)\boldsymbol{\mu}^\top
\end{aligned}$$

$\mathbf{A}$ TD(0) α 1

TD $\overline{\mathrm{VE}}$

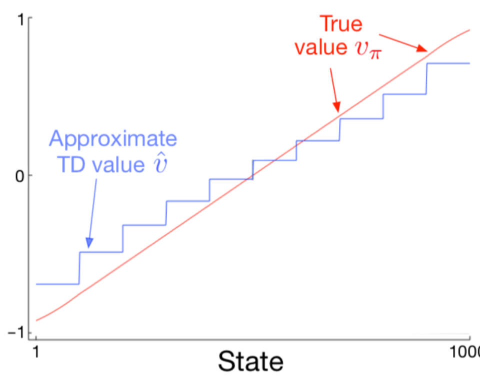
$$\overline{\mathrm{VE}}\,(\mathbf{w}_{\mathrm{TD}}) \leq \frac{1}{1-\gamma} \min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) \tag{3.83}$$

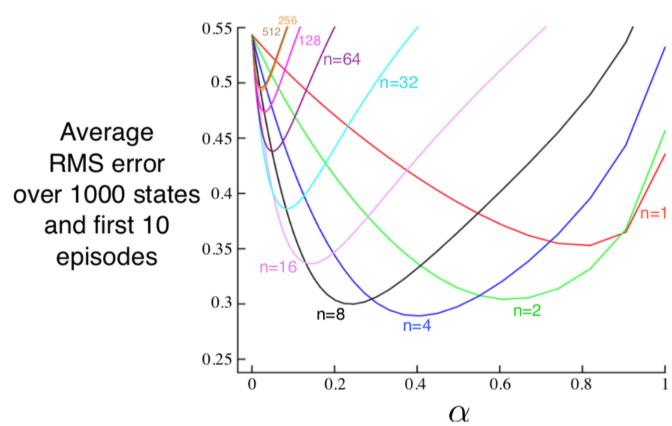
TD $\frac{1}{1-\gamma}$ γ 1 TD TD 6 7

9.14 DP 9.7 $U_t \doteq \sum_a \pi(a|S_t) \sum_{s',r} p(s',r|S_t,a)[r + \gamma\hat{v}(s',\mathbf{w}_t)]$ T-D Sarsa(0) Bertsekas Tsitsiklis 1996 Tsitsiklis Van Roy 1997

11

9.2 1000 1000 9.2 9.1 TD(0) 9.3 T-D 9.1





47: **9.2** 1000 TD 9.1 n 7.2 100

TD 7 n TD 9.2 n TD 1000 19 7.2 20 50 20 19 100 50 19 10 RMS $\overline{\mathrm{VE}}$

n TD 7 n TD

---

n TD $\hat{v} \approx v_\pi$

π

$\hat{v} : \mathcal{S}^{+} \times \mathbb{R}^{d} \to \mathbb{R}$ $\hat{v}(\ , \cdot) = 0$

$\alpha > 0$ n

$\mathbf{w}$ $\mathbf{w} = \mathbf{0}$

S_{t+1} R_{t+1} $mod n + 1$

$S_0 \neq$

$T \leftarrow \infty$

$t = 0, 1, 2, \ldots$

$t < T$

$\pi(\cdot|S_t)$

R_{t+1} S_{t+1}

S_{t+1} $T \leftarrow t + 1$

$\tau \leftarrow t - n + 1$ τ

$\tau \geq 0$

$G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$

$\tau + n < T$ $G \leftarrow G + \gamma^{n} \hat{v}\left(S_{\tau+n}, \mathbf{w}\right)$ $(G_{\tau:\tau+n})$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[G - \hat{v}\left(S_{\tau}, \mathbf{w}\right)\right] \nabla \hat{v}\left(S_{\tau}, \mathbf{w}\right)$

$\tau = T - 1$

7.2

$$\mathbf{w}_{t+n} \doteq \mathbf{w}_{t+n-1} + \alpha \left[G_{t:t+n} - \hat{v}\left(S_t, \mathbf{w}_{t+n-1}\right)\right] \nabla \hat{v}\left(S_t, \mathbf{w}_{t+n-1}\right), \quad 0 \leq t < T \tag{3.84}$$

n 7.1

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^{n} \hat{v}\left(S_{t+n}, \mathbf{w}_{t+n-1}\right), \quad 0 \leq t \leq T - n \tag{3.85}$$

9.1

9.5

i j 3.4

9.5.1

3.4 4.2 4.3

s $s_1 \in \mathbb{R}$ $s_2 \in \mathbb{R}$ s $\mathbf{x}(s) = (s_1, s_2)^\top$ s_1 s_2 $\mathbf{x}(s) = (1, s_1, s_2, s_1 s_2)^\top$ 1 $s_1 s_2$ $\mathbf{x}(s) = (1, s_1, s_2, s_1 s_2, s_1^2, s_2^2, s_1 s_2^2, s_1^2 s_2, s_1^2 s_2^2)^\top$ 2 k

---

: s k $s_1, s_2, \ldots, s_k$ $s_i \in \mathbb{R}$ k -n order-n x_i

$$x_i(s) = \Pi_{j=1}^{k} s_j^{c_{i,j}} \tag{3.86}$$

$c_{i,j}$ $\{0, 1, \ldots, n\}$ $n > 0$ k -n $(n+1)^k$

-n k $n > 0$

9.2 9.17 k $(n+1)^k$

9.2 n $c_{i,j}$ $\mathbf{x}(s) = (1, s_1, s_2, s_1 s_2, s_1^2, s_2^2, s_1 s_2^2, s_1^2 s_2, s_1^2 s_2^2)^\top$

9.5.2

time-honored $f(x) = f(x + \tau)$ x τ f

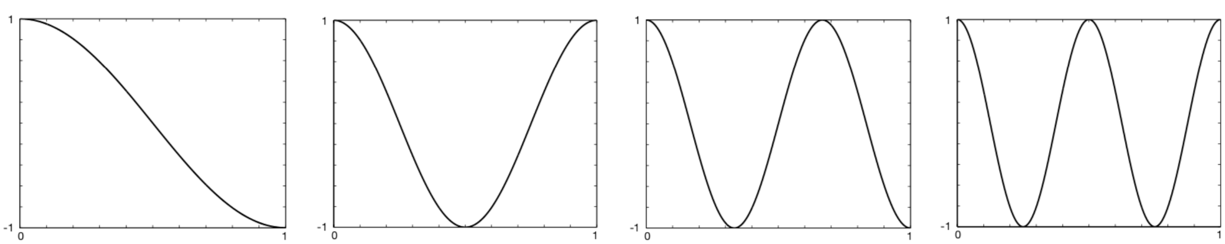
τ τ $1/\tau$ τ

τ $[0, \tau/2]$ $[0, \tau/2]$
“ ” “ ” “ ”
$[0, \tau/2]$

$\tau = 2$ τ $[0, 1]$ n $n+1$

$$x_i(s) = \cos(i\pi s), \quad s \in [0, 1]$$

$i = 0, \ldots, n$ 9.3 math:*x_{i}, i=1,2,3,4* x_0



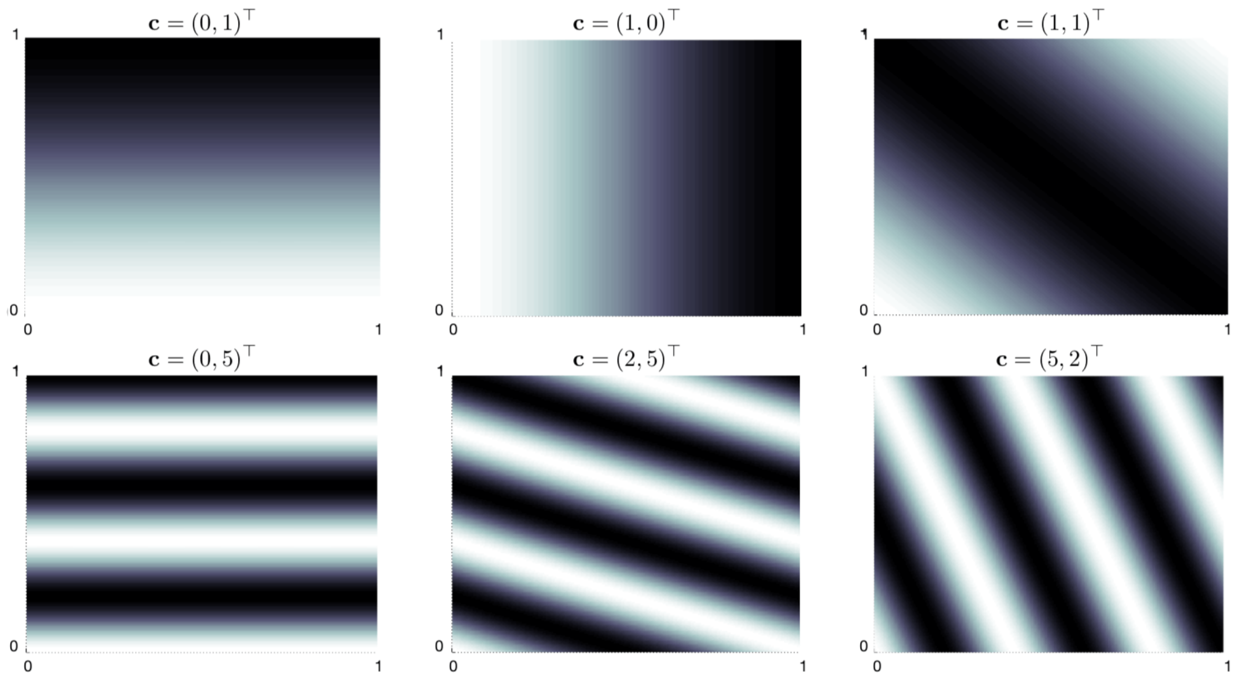
48: **9.3** math:*x_{i}, i=1,2,3,4* $[0, 1]$ Konidaris 2011

: s k $\mathbf{s} = (s_1, s_2, \ldots, s_k)^\top$ $s_i \in [0, 1]$ n i

$$x_i(s) = \cos(\pi \mathbf{s}^\top \mathbf{c}^i) \tag{3.87}$$

$\mathbf{c}^i = (c_1^i, \ldots, c_k^i)^\top$ $j = 1, \ldots, k$ $i = 1, \ldots, (n+1)^k$ $c_j^i \in \{0, \ldots, n\}$ $(n+1)^k$ $\mathbf{c}^i$ $\mathbf{s}^\top \mathbf{c}^i$ $\{0, \ldots, n\}$ $\mathbf{s}$

$k=2$ $\mathbf{s}=(s_1,s_2)^\top$ $\mathbf{c}^i=(c_1^i,c_2^i)^\top$ 9.4 $\mathbf{c}^i$ s_1 $\mathbf{c}^i$ i $\mathbf{c}$ $\mathbf{c}=(0,0)^\top$ $\mathbf{c}=(c_1,0)^\top$ c_1 $\mathbf{c}=(0,c_2)^\top$ $\mathbf{c}=(c_1,c_2)^\top$ $c_j=0$ c_1 c_2



49: **9.4** $\mathbf{c}$ s_1 $\mathbf{c}^i$ i Konidaris 2011

9.7 TD(0) Sarsa α Konidaris Osentoski Thomas 2011 x_i $\alpha_i=\alpha/\sqrt{(c_1^i)^2+\cdots+(c_k^i)^2}$ $c_j^i=0$ $\alpha_i=\alpha$ Sarsa “ ”

n k $\leq$ 5 n n $\mathbf{c}^i$ $\mathbf{c}^i$

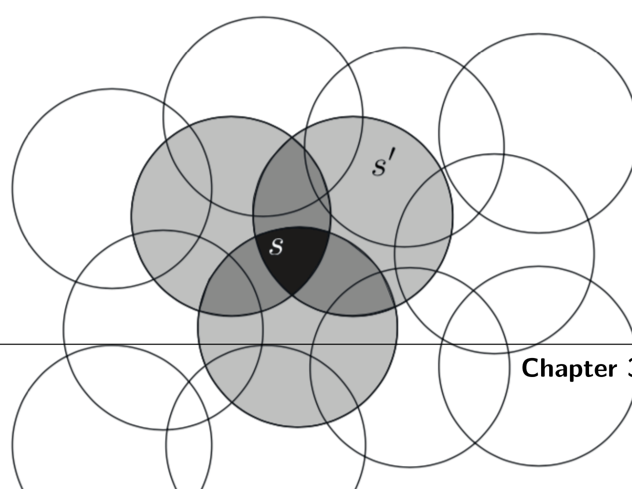
9.5 1000 [2]

9.5.3 Coarse Coding

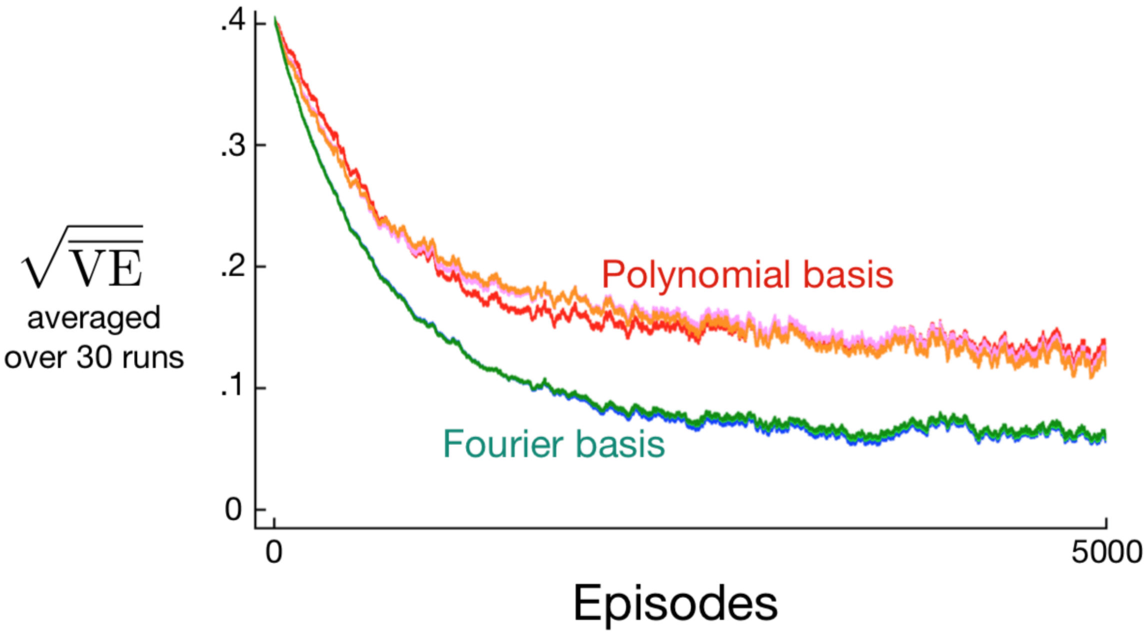
1 *present* 0 *absent* 1-0

$\mathbf{w}$

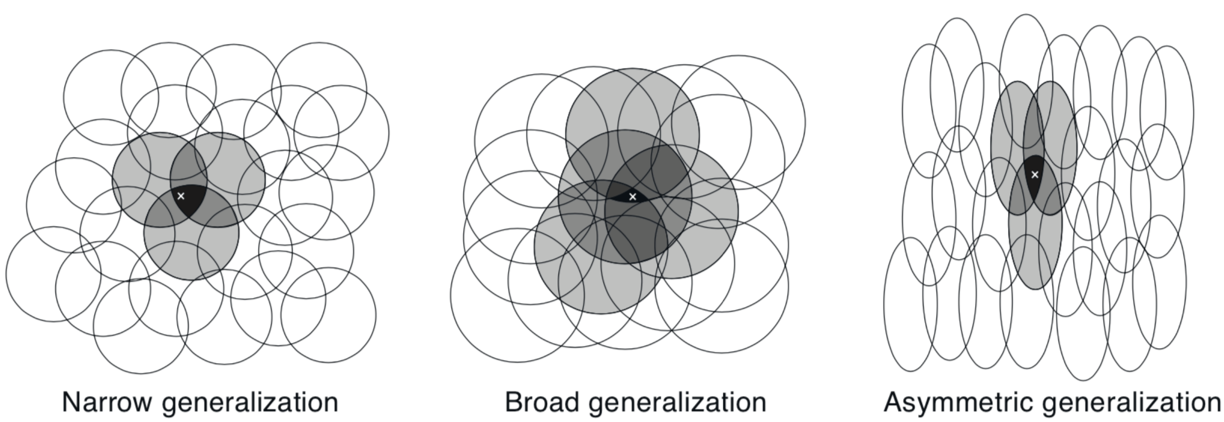
9.8 “ ” 9.6 9.7 9.7 9.7



[2]

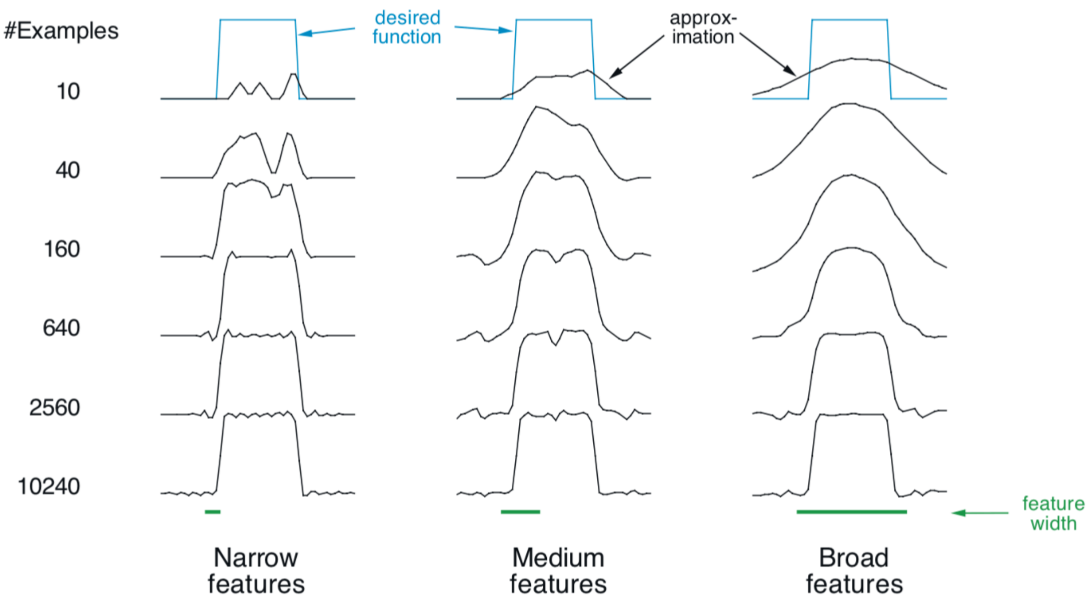


50: **9.5** 1000 5,10 20 $\alpha = 0.0001$ $\alpha = 0.00005$
y 9.1



52: **9.7**

9.3
9.7 9.8
U_t
50
α = $\frac{0.2}{n}$ n
9.8



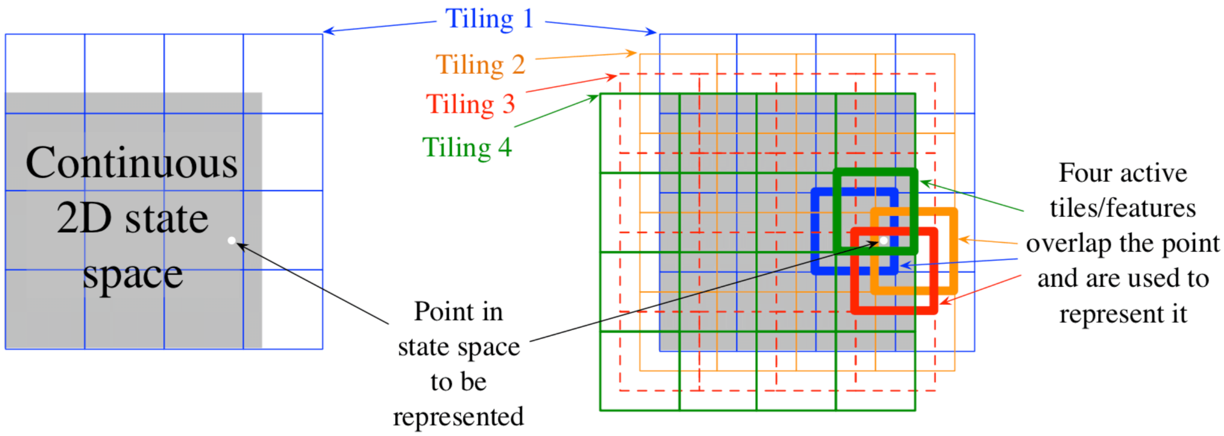
53: **9.8**

9.5.4 Tile Coding

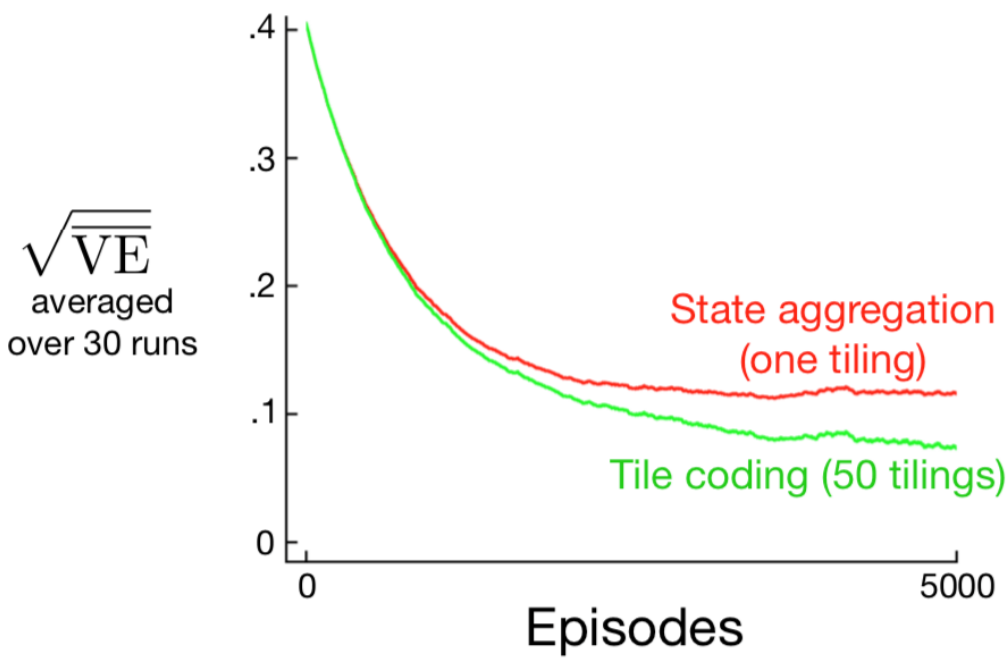
tiling *tile*
9.9 9.6

9.9
$\mathbf{x}(s)$ $4 \times 4 \times 4 = 64$ s 0 9.10 1000
α $\alpha = \frac{1}{n}$ n one-trial
$s \mapsto v$ $\hat{v}(s, \mathbf{w}_t)$ $\hat{v}(s, \mathbf{w}_{t+1}) = v$ $\alpha = \frac{1}{10n}$

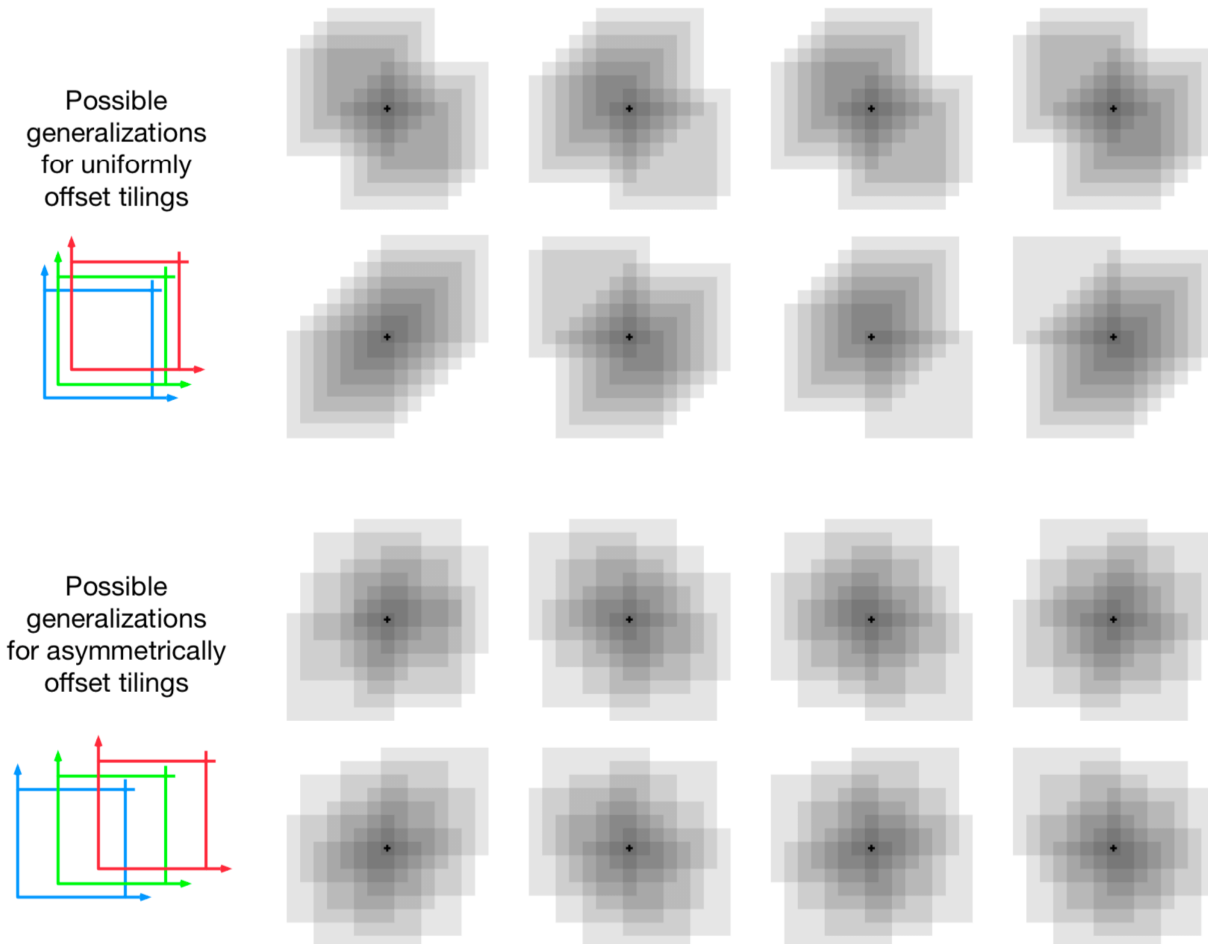
0 1 9.8 d $n \ll d$ indices n
9.9 9.11
64 artifacts
w n $\frac{w}{n}$ $\frac{w}{n}$ $\frac{w}{n}$ /
$(1, 1)$ $\frac{w}{n}$ 9.11 $(1, 3)$



54: **9.9**



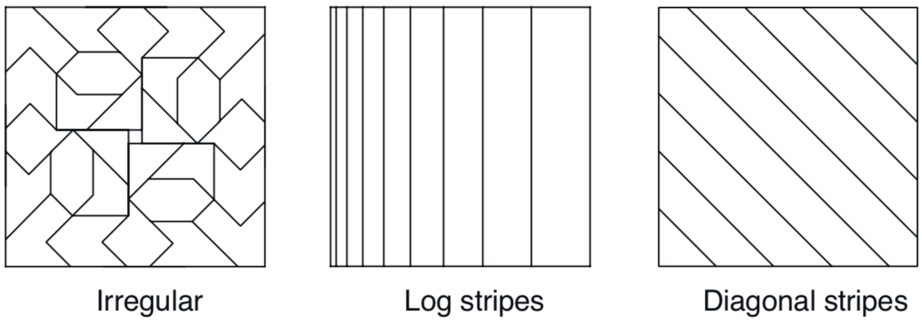
55: **9.10** 1000 1000 200 4 $\alpha =$ 0.0001 50 $\alpha = 0.0001/50$


Possible generalizations for uniformly offset tilings
Possible generalizations for asymmetrically offset tilings

56: **9.11**

Parks Militzer 1991 An 1991 An Miller Parks 1991 Miller An Glanz Carter 1990 $(1,1)$ Miller Glanz 1996 k $1,3,5,7,\dots,2k-1$ n $4k$ 2 9.11 $k=2$ $n=2^3\geq 4k$ $(1,3)$ $(0,0,0)$ $(1,3,5)$ $(2,6,10)$ $(3,9,15)$ k

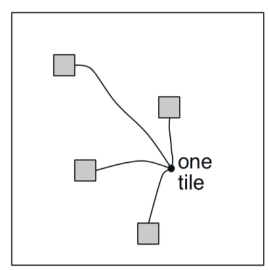
9.8 9.7 9.11 9.12 9.12 discrimination 9.12 9.12



57: **9.12**

9.9 Sutton 1996

hashing



9.4

9.5.5

RBF 0 1 $[0,1]$ RBF x_i $x_i(s)$ s c_i σ_i

$$x_i(s) \doteq \exp\left(-\frac{\|s-c_i\|^2}{2\sigma_i^2}\right)$$

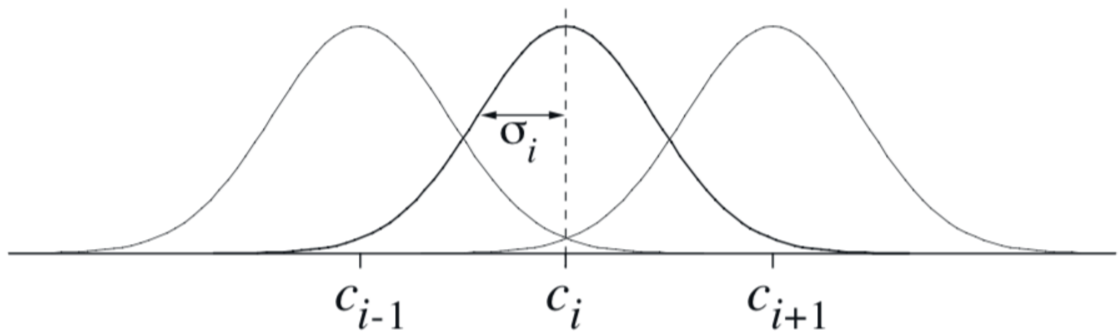
RBF RBF An 1991 Miller 1991 An 1991 Lane Handelman Gelfand

RBF RBF 9.7 9.8 RBF RBF RBF

9.6

SGD α

2.7 MC $\alpha_t = 1/t$ TD 9.8 LSTD $O\left(d^2\right)$ d



58: **9.13**

$\alpha = 1$ 2.4 1 9.3 $\alpha = \frac{1}{10}$ 10
100 $\alpha = \frac{1}{100}$ $\alpha = \frac{1}{\tau}$ τ

τ SGD

$$\alpha \doteq \left(\tau \mathbb{E}\left[\mathbf{x}^\top \mathbf{x}\right]\right)^{-1} \tag{3.88}$$

$\mathbf{x}$ SGD $\mathbf{x}^\top \mathbf{x}$

9.5 $\hat{v}(s, \mathbf{w}) \approx v_\pi(s)$ $7 \times 8 = 56$
$\begin{pmatrix} 7 \\ 2 \end{pmatrix} = 21$ $21 \times 2 + 56 = 98$ 10

9.7

ANN ANN
16 ANN

9.14 ANN “ ”
15.1 ANN ANN

9.14 S sigmoid $f(x) = 1/\left(1 + e^{-x}\right)$
$f(x) = \max(0, x)$ $x \geq \theta$ $f(x) = 1$ 0 θ

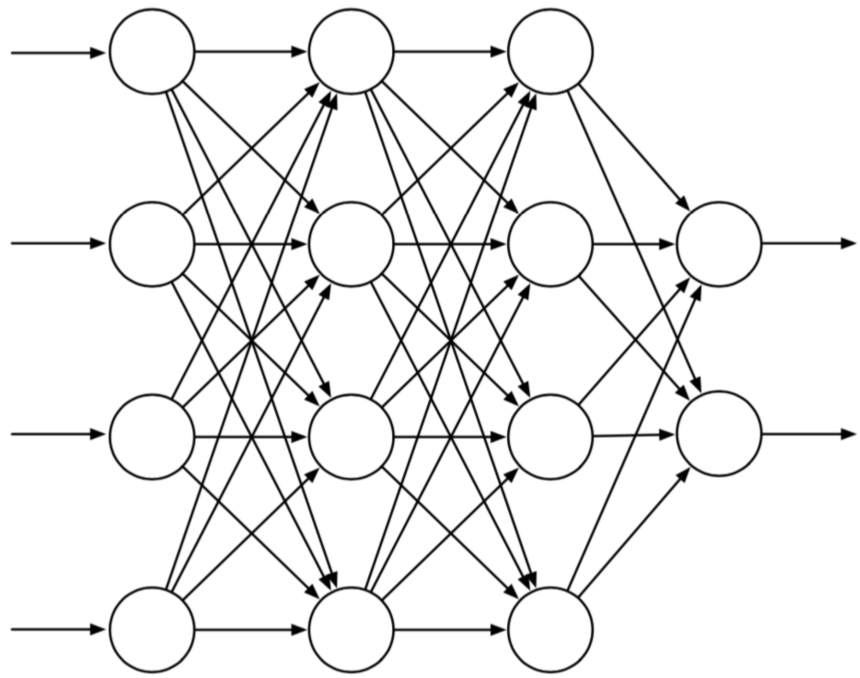
ANN ANN ANN S
Cybenko 1989 ANN

“ ” Bengio 2009
ANN “ ” -

ANN
9.3 TD 2.8 13

15.10 ANN

1 2 ANN k + 1 k
Bengio 2009 ANN



59: **9.14** ANN

Srivastava Hinton Krizhevsky Sutskever Salakhutdinov 2014 dropout
“ ” dropout Srivas-
tava

Hinton Osindero Teh 2006

Ioffe Szegedy 2015 ANN ANN Iof-
fe Szegedy 2015 “ ” ANN

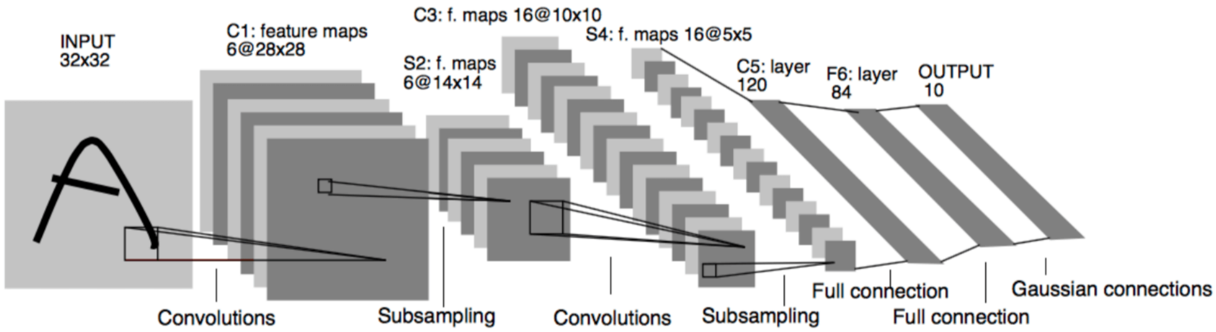
ANN He Zhang Ren Sun 2016
He 2016 16

ANN 16 LeCun Bottou Bengio Haffner 1998

9.15 LeCun 1998
“ ”

9.15 6 28×28 5×5 6 25
9.15 6 2×2 14×14

16



60: **9.15** IEEE LeCun Bottou Bengio Haffner 86 1998 Copyright Clearance Center, Inc

9.8 TD

9.4 TD(0) TD

$$\mathbf{w}_{\mathrm{TD}} = \mathbf{A}^{-1}\mathbf{b}$$

$$\mathbf{A} \doteq \mathbb{E}\left[\mathbf{x}_t\left(\mathbf{x}_t - \gamma\mathbf{x}_{t+1}\right)^\top\right] \qquad \mathbf{b} \doteq \mathbb{E}\left[R_{t+1}\mathbf{x}_t\right]$$

$\mathbf{A}$ $\mathbf{b}$ TD *TD* LSTD

$$\widehat{\mathbf{A}}_t \doteq \sum_{k=0}^{t-1}\mathbf{x}_k\left(\mathbf{x}_k - \gamma\mathbf{x}_{k+1}\right)^\top + \varepsilon\mathbf{I} \qquad \widehat{\mathbf{b}}_t \doteq \sum_{k=0}^{t-1} R_{k+1}\mathbf{x}_k \tag{3.89}$$

$\mathbf{I}$ $\varepsilon > 0$ $\varepsilon\mathbf{I}$ $\widehat{\mathbf{A}}_t$ t ; t $\mathbf{A}$ t $\mathbf{b}$ LSTD TD t

$$\mathbf{w}_t \doteq \widehat{\mathbf{A}}_t^{-1}\widehat{\mathbf{b}}_t \tag{3.90}$$

TD(0) TD(0) $O(d)$

LSTD t 9.20 2 $\widehat{\mathbf{A}}_t$ $O(d^2)$ $\widehat{\mathbf{A}}_t$ $O(d^2)$

9.21 $\widehat{\mathbf{A}}_t$ $O(d^3)$ $O(d^2)$

$$\begin{aligned}
\widehat{\mathbf{A}}_t^{-1} &= \left(\widehat{\mathbf{A}}_{t-1} + \mathbf{x}_{t-1}\left(\mathbf{x}_{t-1} - \gamma\mathbf{x}_t\right)^\top\right)^{-1} && (9.20)\\
&= \widehat{\mathbf{A}}_{t-1}^{-1} - \frac{\widehat{\mathbf{A}}_{t-1}^{-1}\mathbf{x}_{t-1}\left(\mathbf{x}_{t-1} - \gamma\mathbf{x}_t\right)^\top\widehat{\mathbf{A}}_{t-1}^{-1}}{1 + \left(\mathbf{x}_{t-1} - \gamma\mathbf{x}_t\right)^\top\widehat{\mathbf{A}}_{t-1}^{-1}\mathbf{x}_{t-1}} && (9.22)
\end{aligned}$$

$t > 0$ $\widehat{\mathbf{A}}_0 \doteq \varepsilon\mathbf{I}$ *Sherman-Morrison* 9.22 - - $O(d^2)$ $\widehat{\mathbf{A}}_t^{-1}$ 9.22 9.21 $O(d^2)$

LSTD $\hat{v} = \mathbf{w}^\top\mathbf{x}(\cdot) \approx v_\pi$ $O(d^2)$

$\mathbf{x} : \mathcal{S}^+ \to \mathbb{R}^d$ $\mathbf{x}(\ ) = 0$

$\varepsilon > 0$

$\widehat{\mathbf{A}^{-1}} \leftarrow \varepsilon^{-1}\mathbf{I}$ $d \times d$

$\widehat{\mathbf{b}} \leftarrow \mathbf{0}$ d

S $\mathbf{x} \leftarrow \mathbf{x}(S)$

$A \sim \pi(\cdot|S)$ R, S' $\mathbf{x}' \leftarrow \mathbf{x}(S')$

$\mathbf{v} \leftarrow \widehat{\mathbf{A}^{-1}}^\top (\mathbf{x} - \gamma\mathbf{x}')$

$\widehat{\mathbf{A}^{-1}} \leftarrow \widehat{\mathbf{A}^{-1}} - \left(\widehat{\mathbf{A}^{-1}}\mathbf{x}\right) \mathbf{v}^\top / \left(1 + \mathbf{v}^\top \mathbf{x}\right)$

$\widehat{\mathbf{b}} \leftarrow \widehat{\mathbf{b}} + R\mathbf{x}$

$\mathbf{w} \leftarrow \widehat{\mathbf{A}^{-1}}\widehat{\mathbf{b}}$

$S \leftarrow S'$ $\mathbf{x} \leftarrow \mathbf{x}'$

S'

$O(d^2)$ TD $O(d)$ LSTD d LSTD LSTD ε ε ε LSTD π GPI LSTD

9.9

$s \mapsto g$

local-learning

nearest neighbor s s' $\mapsto$ g s' s g s 9.1

8.6 -

k k k n n k n

$k-d$ k k $k-d$ naive

9.10

$s' \mapsto g$ s' s $k : \mathbb{R} \to \mathbb{R}$ $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ $k(s, s')$ s' s

$k(s, s')$ s' s 9.11

$\mathcal{D}$ $g(s')$ s' $\mathcal{D}$

$$\hat{v}(s, \mathcal{D}) = \sum_{s' \in \mathcal{D}} k(s, s') g(s') \tag{3.91}$$

$k(s, s')$ s s' $\mathcal{D}$

9.5.5 RBF RBF RBF RBF RBF RBF RBF 9.23 9.4 $\mathbf{x}(s) = (x_1(s), x_2(s), \ldots, x_d(s))^\top$ $k(s, s')$ s s'

$$k(s, s') = \mathbf{x}(s)^\top \mathbf{x}(s') \tag{3.92}$$

Bishop 2006 9.24 9.24 d “ ”

9.11

MDP MDP

I_t t - t t $\overline{\mathrm{VE}}$ 9.1 μ M_t t 9.15 n

$$\mathbf{w}_{t+n} \doteq \mathbf{w}_{t+n-1} + \alpha M_t \left[G_{t:t+n} - \hat{v}(S_t, \mathbf{w}_{t+n-1})\right] \nabla \hat{v}(S_t, \mathbf{w}_{t+n-1}), \quad 0 \le t < T \tag{3.93}$$

9.16 n

$$M_t = I_t + \gamma^n M_{t-n}, \quad 0 \le t < T \tag{3.94}$$

$t < 0$ $M_t \doteq 0$ $G_{t:t+n} = G_t$ $n = T - t$ $M_t = I_t$

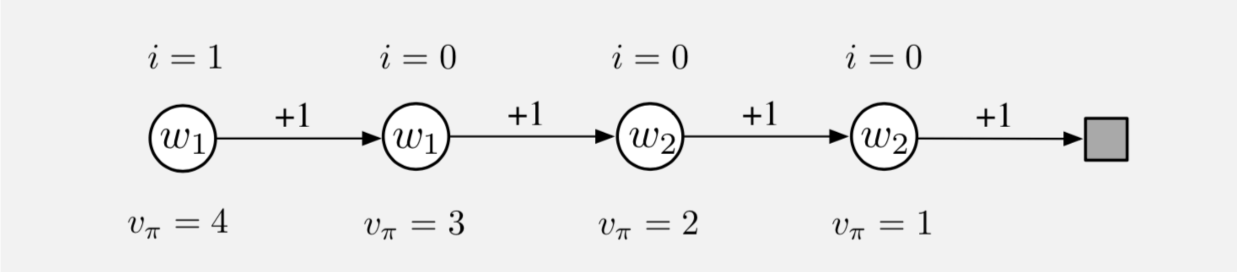
9.4

9.4

+1 4 3 $\mathbf{w} = (w_1, w_2)^\top$ w_1 w_2 1 0

9.7 9.3 $\mathbf{w}_\infty = (3.5, 1.5)$ _ 3.5 w_1 4 w_2

TD (9.15 9.16 9.4 $\mathbf{w}_\infty = (3.5, 1.5)$ $\mathbf{w}_\infty = (4\ 2)$



9.12

w $\overline{VE}(\mathbf{w})$ μ **w** $v_{\pi_\mathbf{w}}(s)$ $\overline{VE}$

SGD n *TD* TD(0) $n = \infty$ $n = 1$ TD DP SGD

LSTD TD SGD

n n TD $\overline{VE}$ n $n \to \infty$ n $n < \infty$ 7 n 5 TD

Bertsekas Tsitsiklis 1996 Bertsekas 2012 Sugiyama 2013

9.3 Widrow Hoff 1960 LMS Widrow Stearns 1985 Bishop 1995 Duda Hart 1973

Sutton 1984,1988 TD(0) 12 TD(λ) “ ”

Michie Chambers BOXES 1968 Singh Jaakkola Jordan 1995 Tsitsiklis Van Roy 1996 Bellman 1957a

9.4 Sutton 1988 $\{\mathbf{x}(s) : s \in \delta\}$ TD(0) $\overline{\text{VE}}$ 1 Peng 1993 Dayan Sejnowski 1994 Tsitsiklis 1994 Gurvits Lin Hanson 1994 Jaakkola Jordan Singh 1994 $\mathbf{w}_t$ Dayan 1992 Tsitsiklis Van Roy 1997 Dayan

9.5 Barto 1990

9.5.2 Konidaris Osentoski Thomas 2011

9.5.3 Hinton 1984 9.6 Waltz Fu 1965

9.5.4 Albus 1971,1981 “ ” CMAC “ ” CMAC Watkins 1989 Shewchuk Dean 1990 Lin Kim 1991 Miller Scalera Kim 1994 Sofge White 1992 Tham 1994 Sutton 1996 Watkins Kraft Campagna 1990; Kraft Miller Dietz 1992 Miller Glanz 1996 http://incompleteideas.net/tiles/tiles3.html

9.5.5 Broomhead Lowe 1988 Powell 1987 RBF Poggio Girosi 1989,1990

9.6 RMSprop Tiele-man Hinton 2012 Adam Kingma Ba 2015 Delta-Bar-Delta Jacobs 1988 Sutton 1992b c; Mahmood 2012 Schraudolph 1999,2002 AlphaBound Dabney Barto 2 SID NOSID Dabney 2014 TIDBD Kearney Schraudolph Yu Aberdeen 2006

9.6 McCulloch Pitts 1943 Perceptron Rosenblatt 1962 ADALINE ADAptive LINear Element Widrow Hoff 1960 LeCun 1985; Rumelhart Hinton Williams 1986 Bengio Courville Vincent 2012; Goodfellow Bengio Courville 2016 Haykin 1994 Bishop 1995 Ripley 2007

Farley Clark 1954 Widrow Gupta Maitra 1973 ADALINE Werbos 1987,1994 ANN TD Barto Sutton Brouwer 1981 Barto Sutton 1981b Kohonen 1977; Anderson Silverstein Ritz Jones 1977 Barto Anderson Sutton 1982 ANN Hampson 1983,1989 Barto Sutton Anderson 1983 15.7 15.8 Barto Anandan 1985 Widrow 1973 A_{R-P} Barto 1985,1986 Barto Jordan 1987 A_{R-P} ANN A_{R-P} Barto 1985 15.10 Anderson 1986,1987,1989 Hanoi Williams 1988 Gullapalli 1990 Williams 1992 Barto Sutton Watkins 1990 Williams 1992 REINFORCE 13.3 Tesauro TD-Gammon Tesauro 1992,1994; 16.1 TD(λ) ANN Silver AlphaGo AlphaGo Zero AlphaZero 2016,2017a b; 16.6 Schmidhuber 2015

9.8 LSTD Bradtke Barto Bradtke 1993,1994 Bradtke Barto 1996 Bradtke Ydstie Barto 1994 Boyan 1999,2002 Nedi c Bertsekas 2003 Yu 2010 1949 Sherman Morrison 1949 Lagoudakis Parr 2003 Buşoniu Lazaric Ghavamzadeh Munos Babŭska De Schutter 2012

9.9 Atkeson Moore Schaal 1997 Atkeson 1992 Stanfill Waltz 1986 Baird Klopf 1993 Q-learning Schaal Atkeson 1994 Peng 1995 Tadepalli Ok 1996 Bottou Vapnik 1992

Bentley 1975 k-d n $O(\log n)$ Friedman Bentley Finkel 1977 k-d Omohundro 1987 k-d Moore Schneider Deng 1997 k-d

9.10 Aizerman Braverman Rozonoer 1964 * * Connell Utgoff 1987 "Shepard " Shepard 1968 Ormoneit Sen 2002 Dietterich Wang 2002 Xu Xie Hu Lu 2005 Taylor Parr 2009 Barreto Precup Pineau 2011 Bhat Farias Moallemi 2012

9.11 Emphatic-TD 11.8

Samuel 1959,1967 Samuel Shannon 1950 Samuel Grith 1966,1974; Page 1977; Biermann Fairfield and Beres 1982

Samuel Bellman Dreyfus 1959 DP Bellman Samuel DP Bellman Dreyfus 1959 Bellman Kalaba Kotkin 1973 Daniel 1976 Whitt 1978 Reetz 1977; Schweitzer Seidmann 1985 Chow Tsitsiklis 1991 Kushner Dupuis 1992 Rust 1996

Holland 1986 - " " Holland - 1 Holland Holland

Christensen Korf 1986 Chapman Kaelbling 1991 Tan 1991 Yee Saxena Utgoff Barto 1990 Dietterich Flann 1995

3.6.2 10

- $\hat{q}(s, a, \mathbf{w}) \approx q_*(s, a)$ $\mathbf{w} \in \mathbb{R}^d$ 11 Sarsa TD(0) " " " "

GPI ε n Sarsa

10.1

9 $\hat{q} \approx q_\pi$ $\mathbf{w}$ $S_t \mapsto U_t$ $S_t, A_t \mapsto U_t$ U_t
$q_\pi(S_t, A_t)$ G_t n Sarsa 7.4

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[U_t - \hat{q}\left(S_t, A_t, \mathbf{w}_t\right)\right] \nabla \hat{q}\left(S_t, A_t, \mathbf{w}_t\right) \tag{3.95}$$

Sarsa

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[R_{t+1} + \gamma \hat{q}\left(S_{t+1}, A_{t+1}, \mathbf{w}_t\right) - \hat{q}\left(S_t, A_t, \mathbf{w}_t\right)\right] \nabla \hat{q}\left(S_t, A_t, \mathbf{w}_t\right) \tag{3.96}$$

Sarsa TD(0) 9.14

S_t
$\hat{q}(S_t, a, \mathbf{w}_t)$ $A_t^* = \arg\max_a \hat{q}(S_t, a, \mathbf{w}_{t-1})$ ε -

---

Sarsa $\hat{q} \approx q_*$

$\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$

$\alpha > 0$ $\varepsilon > 0$

$\mathbf{w} \in \mathbb{R}^d$ $\mathbf{w} = \mathbf{0}$

$S, A \leftarrow$ ε -

A R, S'

S'

$\mathbf{w} \leftarrow \mathbf{w} + \alpha[R - \hat{q}(S, A, \mathbf{w})]\nabla\hat{q}(S, A, \mathbf{w})$

A' $\hat{q}(S', \cdot, \mathbf{w})$ ε -

$\mathbf{w} \leftarrow \mathbf{w} + \alpha\left[R + \gamma\hat{q}\left(S', A', \mathbf{w}\right) - \hat{q}(S, A, \mathbf{w})\right]\nabla\hat{q}(S, A, \mathbf{w})$

$S \leftarrow S'$

$A \leftarrow A'$

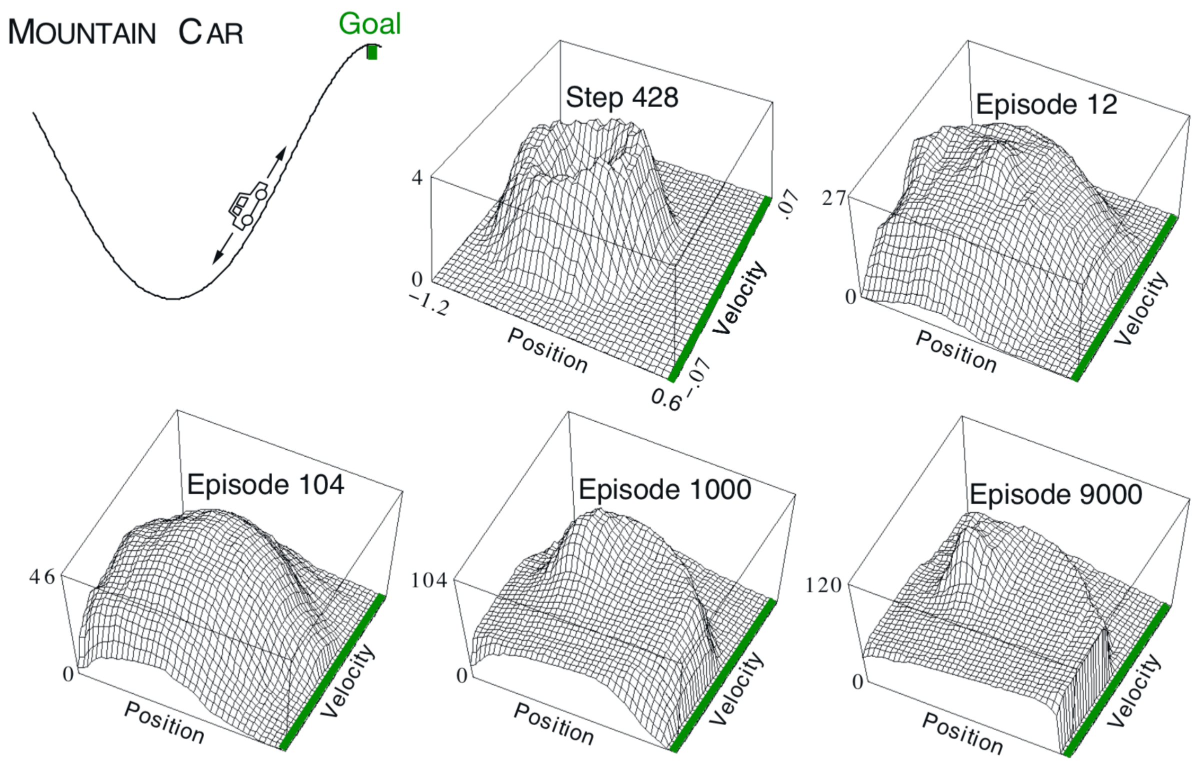
10.1 10.1

-1 $+1$ -1 0 x_t $\dot{x}_t$

$$x_{t+1} \doteq \text{bound}\left[x_t + \dot{x}_{t+1}\right]$$
$$\dot{x}_{t+1} \doteq \text{bound}\left[\dot{x}_t + 0.001 A_t - 0.0025\cos\left(3x_t\right)\right]$$

bound $-1.2 \le x_{t+1} \le 0.5$ $-0.07 \le \dot{x}_{t+1} \le 0.07$ $\dot{x}_{t+1}$ $\dot{x}_{t+1}$
$x_t \in [-0.6, -0.4)$ 9.9 8 1/8 9.5.4 [1] $\mathbf{x}(s, a)$

[1] http://incompleteideas.net/tiles/tiles3.html int = IHT(4096) tiles(iht, 8, $[8 * x/(0.5 + 1.2), 8 * xdot/(0.07 + 0.07)], A)$ $(x\ xdot)$ A



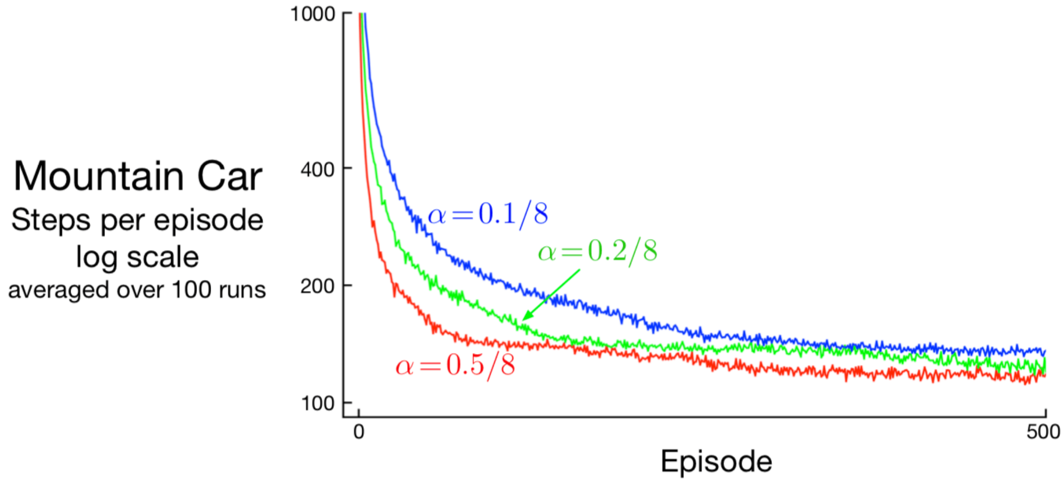
61: **10.1** $-\max_a \hat{q}(s, a, \mathbf{w})$

$$\hat{q}(s,a,\mathbf{w}) \doteq \mathbf{w}^{\top}\mathbf{x}(s,a) = \sum_{i=1}^{d} w_i \cdot x_i(s,a) \tag{3.97}$$

s a

10.1 [2] *cost-* *to-go* ε 0 "Step 428"

10.2 Sarsa



62: **10.2** ε - Sarsa

10.2 n Sarsa

Sarsa 10.1 n Sarsa n n 7.4

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^{n} \hat{q}\left(S_{t+n}, A_{t+n}, \mathbf{w}_{t+n-1}\right), \quad t+n < T \tag{3.98}$$

$G_{t:t+n} = G_t$ $t+n \geq T$ n

$$\mathbf{w}_{t+n} \doteq \mathbf{w}_{t+n-1} + \alpha \left[G_{t:t+n} - \hat{q}\left(S_t, A_t, \mathbf{w}_{t+n-1}\right)\right] \nabla \hat{q}\left(S_t, A_t, \mathbf{w}_{t+n-1}\right), \quad 0 \leq t < T \tag{3.99}$$

---

n Sarsa $\hat{q} \approx q_*$ q_π

$\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$

π q_π

$\alpha > 0$ $\varepsilon > 0$ n

$\mathbf{w} \in \mathbb{R}^d$ $\mathbf{w} = \mathbf{0}$

S_t A_t R_t $mod n + 1$

---

[2] " Sarsa(λ)" 12 Sarsa

$S_0 \neq$

$A_0 \sim \pi(\cdot|S_0)$ $\hat{q}(S_0, \cdot, \mathbf{w})$ ε -

$S, A \leftarrow$ ε -

$T \leftarrow \infty$

$t = 0, 1, 2, cdots$

$t < T$

A_t

R_{t+1} S_{t+1}

S_{t+1}

$T = t + 1$

$A_{t+1} \sim \pi(\cdot|S_{t+1})$ $\hat{q}(S_{t+1}, \cdot, \mathbf{w})$ ε -

$\tau \leftarrow t - n + 1$ τ

$\tau \geq 0$

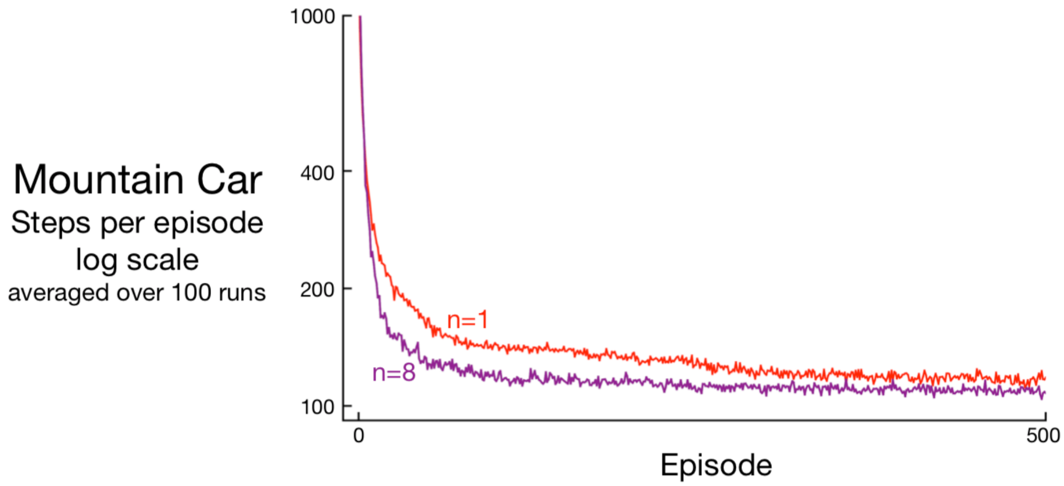
$G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$

$\tau + n < T$ $G \leftarrow G + \gamma^n \hat{q}\left(S_{\tau+n}, A_{\tau+n}, \mathbf{w}\right)$ $G_{\tau:\tau+n}$

$\mathbf{w} \doteq \mathbf{w} + \alpha\left[G - \hat{q}\left(S_t, A_t, \mathbf{w}\right)\right] \nabla \hat{q}\left(S_\tau, A_\tau, \mathbf{w}\right)$

$\tau = T - 1$

1 n 10.3 $n = 8$ $n = \infty$ 10.4 α n

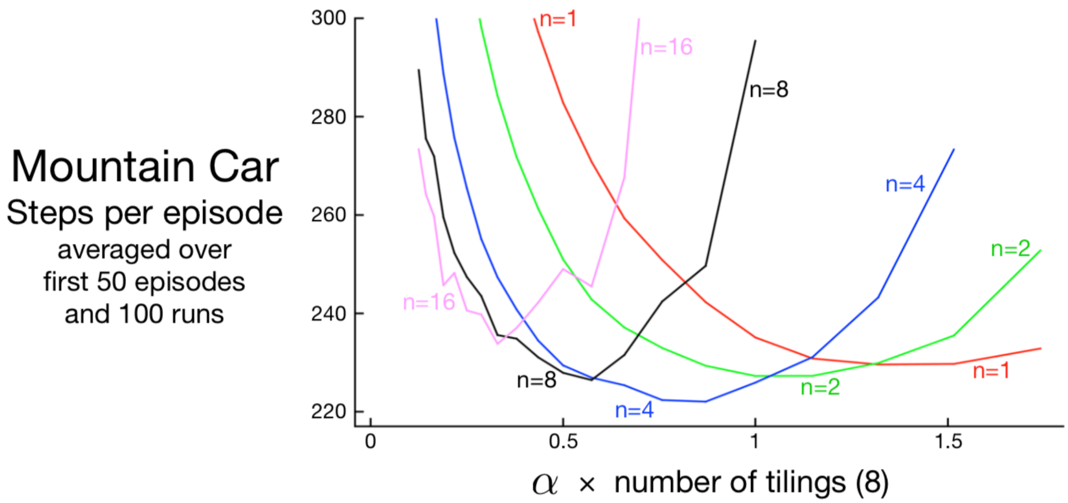


63: **10.3** Sarsa $n = 1$ $\alpha = 0.5/8$ $n = 8$ $\alpha = 0.3/8$

10.1

10.2 Sarsa

10.3 10.4 n n



64: **10.4** α n n Sarsa $n = 4$ α $n = 16$ $n = 1$ 0.5 $n = 16$ 4

10.3

MDP

π $r(\pi)$

$$r(\pi) \doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}\left[R_t | S_0, A_{0:t-1} \sim \pi\right] \quad (10.6)$$
$$= \lim_{t\to\infty} \mathbb{E}\left[R_t | S_0, A_{0:t-1} \sim \pi\right] \quad (10.7)$$
$$= \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_{s',r} p\left(s', r|s, a\right) r$$

S_0 π $A_0, A_1, \ldots, A_{t-1}$ μ_π $\mu_\pi(s) \doteq \lim_{t\to\infty} Pr\left\{S_t = s | A_{0:t-1} \sim \pi\right\}$ π S_0 MDP *ergodicity* MDP MDP $r(\pi)$ 10.7 π $r(\pi)$

π

$$\sum_s \mu_\pi(s) \sum_a \pi(a|s) p\left(s'|s, a\right) = \mu_\pi\left(s'\right) \quad (3.100)$$

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots \quad (3.101)$$

$v_\pi(s) \doteq \mathbb{E}_\pi\left[G_t | S_t = s\right]$ $q_\pi(s, a) \doteq \mathbb{E}_\pi\left[G_t | S_t = s, A_t = a\right]$ v_* q_* Bellman γs

$$v_\pi(s) = \textstyle\sum_a \pi(a|s) \sum_{r,s'} p\left(s', r|s, a\right)\left[r - r(\pi) + v_\pi\left(s'\right)\right]$$
$$q_\pi(s, a) = \textstyle\sum_{r,s'} p\left(s', r|s, a\right)\left[r - r(\pi) + \sum_{a'} \pi\left(a'|s'\right) q_\pi\left(s', a'\right)\right]$$
$$v_*(s) = \max_a \textstyle\sum_{r,s'} p\left(s', r|s, a\right)\left[r - \max_\pi r(\pi) + v_*\left(s'\right)\right],$$
$$q_*(s, a) = \textstyle\sum_{r,s'} p\left(s', r|s, a\right)\left[r - \max_\pi r(\pi) + \max_{a'} q_*\left(s', a'\right)\right]$$

3.14 3.17 3.19 3.20

TD

$$\delta_t \doteq R_{t+1} - \overline{R}_t + \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t) \tag{3.102}$$

$$\delta_t \doteq R_{t+1} - \overline{R}_t + \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_t) - \hat{q}(S_t, A_t, \mathbf{w}_t) \tag{3.103}$$

$\overline{R}_t$ $r(\pi)$ t

Sarsa 10.2 TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \delta_t \nabla \hat{q}(S_t, A_t, \mathbf{w}_t) \tag{3.104}$$

10.11 t

Sarsa $\hat{q} \approx q_*$

$\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$

$\alpha, \beta > 0$

$\mathbf{w} \in \mathbb{R}^d$ $\mathbf{w} = \mathbf{0}$

$\overline{R} \in \mathbb{R}$ $\overline{R} = 0$

S A

A R, S'

A' $\hat{q}(S', \cdot, \mathbf{w})$ ε -

$\delta \leftarrow R - \overline{R} + \hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})$

$\overline{R} \leftarrow \overline{R} + \beta\delta$

$\overline{R} \leftarrow \overline{R} + \beta\delta$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\hat{q}(S, A, \mathbf{w})$

$S \leftarrow S'$

$A \leftarrow A'$

10.4 Q-learning

10.5 10.10 TD(0)

10.6 MDP $+1, 0, +1, 0, +1, 0, \ldots$ μ_π 10.7 10.6 MDP **A** $+1$ **B** 0 $+1, 0, +1, 0, \ldots$ 10.9

$$v_\pi(s) \doteq \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \left(\mathbb{E}_\pi \left[R_{t+1} | S_0 = s \right] - r(\pi) \right) \tag{3.105}$$

A **B**

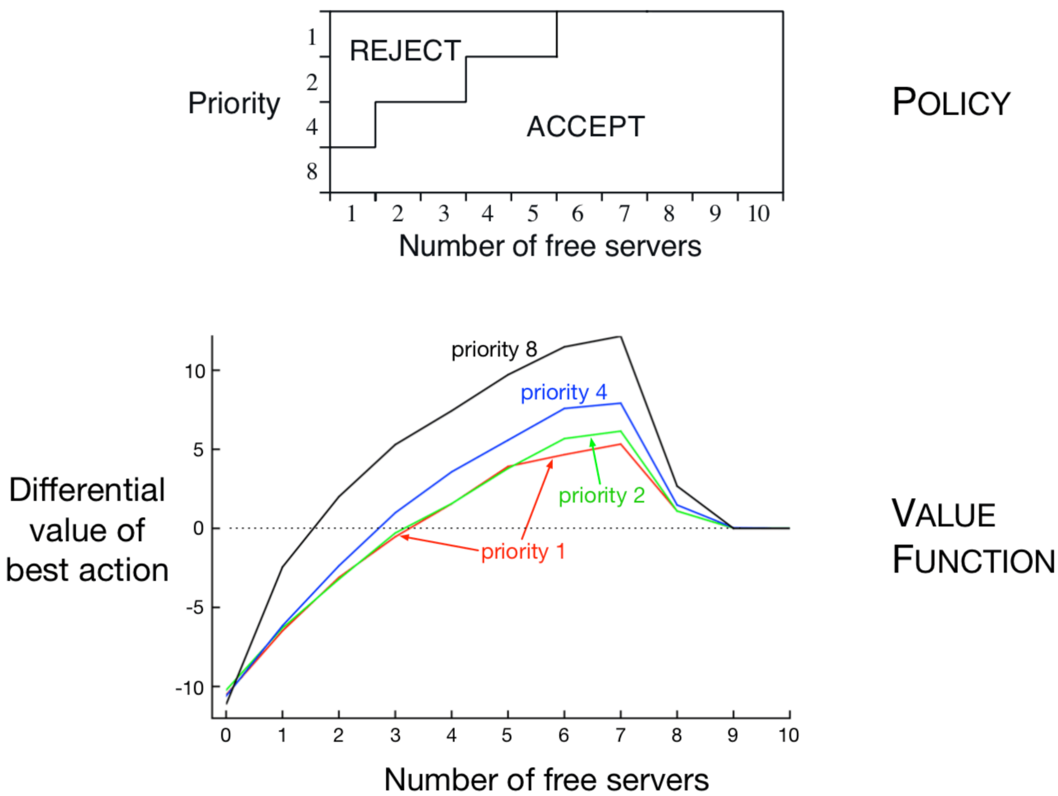
10.7 **A B** **C** **A** $+1$ 0 10.13

10.8 δ_t $R_{t+1} - \overline{R}_t$ $\overline{R}_t$ δ_t 10.7 MRP $\frac{1}{3}$ $R_{t+1} - \overline{R}_t$ δ_t 10.10

10.2 10 1,2,4 8

$p = 0.06$

- 10.5 Sarsa $\alpha = 0.01$ $\beta = 0.01$ $\varepsilon = 0.1$ $\overline{R}$



65: 10.5 Sarsa 200 $\overline{R}$ 2.31

10.4

π

$r(\pi)/(1-\gamma)$ $r(\pi)$ γ

t $t-1$ $t-2$ $t-1000$ 999

t $1 + \gamma + \gamma^2 + \gamma^3 + \cdots = 1/(1-\gamma)$ $r(\pi)/(1-\gamma)$

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) v_\pi^\gamma(s) && (\quad v_\pi^\gamma \quad) \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_{s'} \sum_r p\left(s', r|s, a\right)\left[r + \gamma v_\pi^\gamma\left(s'\right)\right] && (\text{Bellman} \quad) \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_{s'} \sum_r p\left(s', r|s, a\right) \gamma v_\pi^\gamma\left(s'\right) && (\ 10.7) \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma\left(s'\right) \sum_s \mu_\pi(s) \sum_a \pi(a|s) p\left(s'|s, a\right) && (\ 3.4) \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma\left(s'\right) \mu_\pi\left(s'\right) && (\ 10.8) \\
&= r(\pi) + \gamma J(\pi) \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 J(\pi) \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 r(\pi) + \gamma^3 r(\pi) + \cdots \\
&= \frac{1}{1-\gamma} r(\pi)
\end{aligned}
$$

γ

γ γ

4.2

13 “ ”

Perkins Precup 2003 ε - Gordon 1996a

10.5 n Sarsa

n TD n n 7.4

$$G_{t:t+n} \doteq R_{t+1} - \overline{R}_{t+n-1} + \cdots + R_{t+n} - \overline{R}_{t+n-1} + \hat{q}\left(S_{t+n}, A_{t+n}, \mathbf{w}_{t+n-1}\right) \tag{3.106}$$

$\overline{R}$ $r(\pi)$ $n \geq 1$ $t+n<T$ $t+n \geq T$ $G_{t:t+n} \doteq G_t$ n TD

$$\delta_t \doteq G_{t:t+n} - \hat{q}\left(S_t, A_t, \mathbf{w}\right) \tag{3.107}$$

Sarsa 10.12

---

n Sarsa $\hat{q} \approx q_\pi$ q_*

$\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \rightarrow \mathbb{R}$ π

$\mathbf{w} \in \mathbb{R}^d$ $\mathbf{w} = \mathbf{0}$

$\overline{R} \in \mathbb{R}$ $\overline{R} = 0$

$\alpha, \beta > 0$ n

S_t A_t R_t $mod n+1$

S_0 A_0

$t = 0, 1, 2, \ldots$

A_t

R_{t+1} S_{t+1}

$A_{t+1} \sim \pi(\cdot|S_{t+1})$ $\hat{q}(S_{t+1}, \cdot, \mathbf{w})$ ε -

$\tau \leftarrow t - n + 1$ τ

$\tau > 0$

$\delta \leftarrow \sum_{i=\tau+1}^{\tau+n} \left(R_i - \overline{R}\right) + \hat{q}\left(S_{\tau+n}, A_{\tau+n}, \mathbf{w}\right) - \hat{q}\left(S_\tau, A_\tau, \mathbf{w}\right)$

$\overline{R} \leftarrow \overline{R} + \beta\delta$

$\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\hat{q}\left(S_\tau, A_\tau, \mathbf{w}\right)$

10.9 n Sarsa β $\overline{R}$ $\overline{R}$ $\overline{R}$ $\overline{R}$ 2.7 n Sarsa

10.6

$r(\pi)$

Bellman TD

10.1 Rummery Niranjan 1994 Sarsa ε - Sarsa Gordon 1996a 2001 Pupcup Perkins 2003 Perkins Pendrith 2002 Melo Meyn Ribeiro 2008 - Moore 1990 Sutton 1996

10.2 n Sarsa van Seijen 2016 Sarsa(λ)

10.3 Puterman 1994 Mahadevan 1996 Tadepalli Ok 1994 Bertsekas Tsitsiklis 1996 Tsitsiklis Van Roy 1999 Schwartz 1993 “R-learning” R-learning Q-learning Carlstřom Nordstřom 1997

10.4 Singh Jaakkola Jordan 1994

3.6.3 11 *

5
6 7

π b $\hat{v} \approx v_\pi$ $\hat{q} \approx q_\pi$ π $\hat{q}$ b $\hat{q}$ ε -

5 7

11.1

7 $\hat{v}$ $\hat{q}$ V Q $\mathbf{w}$

$$\rho_t \doteq \rho_{t:t} = \frac{\pi(A_t|S_t)}{b(A_t|S_t)} \tag{3.108}$$

TD(0) 203 9.3 ρ_t

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\rho_t\delta_t\nabla\hat{v}(S_t, \mathbf{w}_t) \tag{3.109}$$

ρ_t

$$\delta_t \doteq R_{t+1} + \gamma\hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t), \tag{3.110}$$

$$\delta_t \doteq R_{t+1} - \overline{R}_t + \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t) \tag{3.111}$$

Sarsa

$$\begin{aligned}
&\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\delta_t\nabla\hat{q}(S_t, A_t, \mathbf{w}_t), \\
&\delta_t \doteq R_{t+1} + \gamma\textstyle\sum_a \pi(a|S_{t+1})\,\hat{q}(S_{t+1}, a, \mathbf{w}_t) - \hat{q}(S_t, A_t, \mathbf{w}_t), \text{ or} \qquad (\quad) \\
&\delta_t \doteq R_{t+1} - \overline{R}_t + \textstyle\sum_a \pi(a|S_{t+1})\,\hat{q}(S_{t+1}, a, \mathbf{w}_t) - \hat{q}(S_t, A_t, \mathbf{w}_t). \qquad (\quad)
\end{aligned}$$

A_t -

Sarsa n

$$\mathbf{w}_{t+n} \doteq \mathbf{w}_{t+n-1} + \alpha\rho_{t+1}\cdots\rho_{t+n-1}\left[G_{t:t+n} - \hat{q}(S_t, A_t, \mathbf{w}_{t+n-1})\right]\nabla\hat{q}(S_t, A_t, \mathbf{w}_{t+n-1}) \tag{3.112}$$

$$\begin{aligned}
&G_{t:t+n} \doteq R_{t+1} + \cdots + \gamma^{n-1}R_{t+n} + \gamma^n\hat{q}(S_{t+n}, A_{t+n}, \mathbf{w}_{t+n-1}), \qquad (\quad) \\
&G_{t:t+n} \doteq R_{t+1} - \overline{R}_t + \cdots + R_{t+n} - \overline{R}_{t+n-1} + \hat{q}(S_{t+n}, A_{t+n}, \mathbf{w}_{t+n-1}) \qquad (\quad)
\end{aligned}$$

$k \geq T$ $\rho_k\mathbf{S}$ 1 $t+n \geq T$ $G_{t:n}$ G_t

7 n

$$\mathbf{w}_{t+n} \doteq \mathbf{w}_{t+n-1} + \alpha\left[G_{t:t+n} - \hat{q}(S_t, A_t, \mathbf{w}_{t+n-1})\right]\nabla\hat{q}(S_t, A_t, \mathbf{w}_{t+n-1}) \tag{3.113}$$

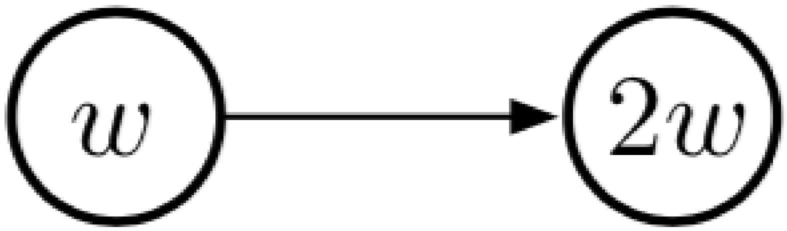
$$G_{t:t+n} \doteq \hat{q}(S_t, A_t, \mathbf{w}_{t-1}) + \sum_{k=t}^{t+n-1}\delta_k\prod_{i=t+1}^{k}\gamma\pi(A_i|S_i) \tag{3.114}$$

ρ_t Sarsa 7 n Q(σ) n

11.1 n TD 7.9 *11.2* n Q(σ) 7.11 7.17

11.2

MDP w $2w$ $\mathbf{w}$ w 1 2
0



$w = 10$ 10 20 w γ 1 TD 10 $\alpha = 0.1$ TD w 11 22 ≈ 1 ≈ 22 TD ≈ 11 ≈ 12.1 w

TD

$$\delta_t = R_{t+1} + \gamma \hat{v}\left(S_{t+1}, \mathbf{w}_t\right) - \hat{v}\left(S_t, \mathbf{w}_t\right) = 0 + \gamma 2 w_t - w_t = (2\gamma - 1) w_t$$

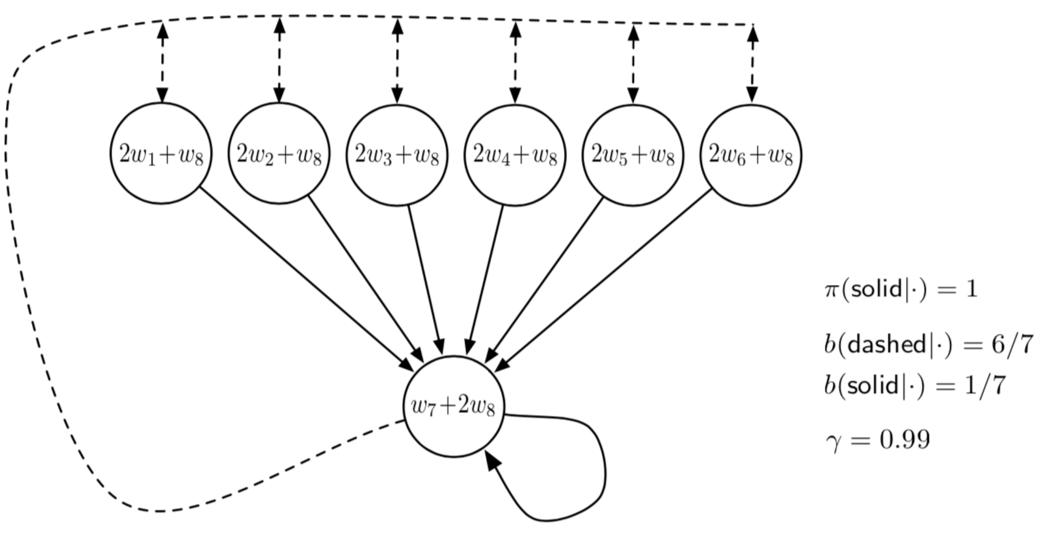
TD(0) 11.2

$$w_{t+1} = w_t + \alpha \rho_t \delta_t \nabla \hat{v}\left(S_t, w_t\right) = w_t + \alpha \cdot 1 \cdot (2\gamma - 1) w_t \cdot 1 = (1 + \alpha(2\gamma - 1)) w_t$$

ρ_t 1 1 $1 + \alpha(2\gamma - 1)$ 1 w $\gamma > 0.5$ 1 $\alpha > 0$ w

w ρ_t ρ_t w $2w$ w $2w$ w $\gamma = 1$ $2w$ w Eventually the piper must be paid

MDP *Baird* 11.1 MDP b $\frac{6}{7}$ $\frac{1}{7}$ π π $\gamma = 0.99$



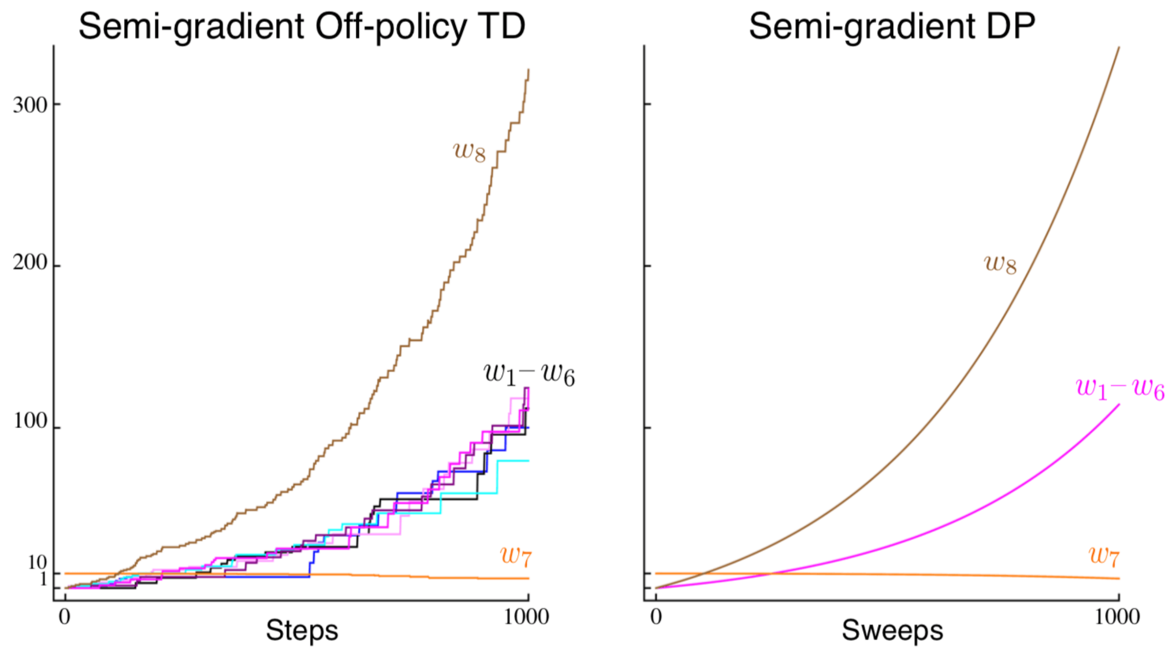
66: **11.1** Baird

$2w_1 + w_8$ $\mathbf{w} \in \mathbb{R}^8$ $\mathbf{x}(1) = (2, 0, 0, 0, 0, 0, 0, 1)^\top$
s $v_\pi(s) = 0$ $\mathbf{w} = \mathbf{0}$ 8 7 $\{\mathbf{x}(s) : s \in \mathcal{S}\}$

TD(0) 11.2 11.2 DP 11.2
DP $\mathbf{w}_k$

$$\mathbf{w}_{k+1} \doteq \mathbf{w}_k + \frac{\alpha}{|\mathcal{S}|} \sum_s \left(\mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}\left(S_{t+1}, \mathbf{w}_k\right) | S_t = s\right] - \hat{v}\left(s, \mathbf{w}_k\right)\right) \nabla \hat{v}\left(s, \mathbf{w}_k\right) \tag{3.115}$$

DP



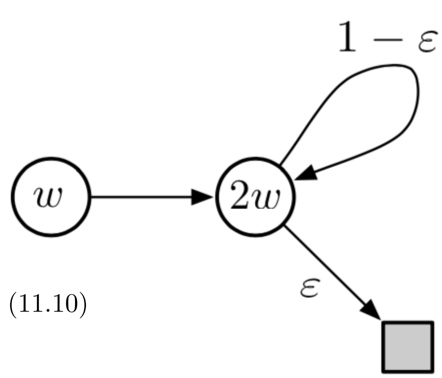
67: **11.2** Baird $\mathbf{w}$ $\alpha = 0.01$ $\mathbf{w} = (1, 1, 1, 1, 1, 1, 10, 1)^\top$

Baird DP 9.14 TD DP

Baird Q-learning Q-learning ε -
Q-learning Q-learning

Baird $\{\mathbf{x}(s) : s \in \mathcal{S}\}$
Baird DP

11.1 Tsitsiklis Van Roy DP
w-to-$2w$ w $2w$
$w = 0$ w_{k+1}
$\overline{\text{VE}}$



$$\begin{aligned} w_{k+1} &= \arg\min_{w \in \mathbb{R}} \sum_{s \in \mathcal{S}} \left(\hat{v}(s, w) - \mathbb{E}_\pi \left[R_{t+1} + \gamma \hat{v}\left(S_{t+1}, w_k\right) | S_t = s\right]\right)^2 \\ &= \arg\min_{w \in \mathbb{R}} \left(w - \gamma 2 w_k\right)^2 + \left(2w - (1-\varepsilon)\gamma 2 w_k\right)^2 \\ &= \frac{6 - 4\varepsilon}{5} \gamma w_k \end{aligned} \tag{11.10}$$

$\gamma > \frac{5}{6-4\varepsilon}$ $w_0 \neq 0$ $\{w_k\}$

averagers ANN

11.3 Q-learning Baird

11.3

ANN

TD MC

LSTD

.

12

7 7.2 9 9.2 10 10.4 12.14

Şimsek Alg orta Kothiyal 2016

9.14 n $\lambda \approx 1$ 12

Sarsa Q-learning

11.4

$v : \mathcal{S} \to \mathbb{R}$

$\mathcal{S}$ = $\{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ v $\left[v(s_1), v(s_2), \ldots, v(s_{|\mathcal{S}|})\right]^\top$

$\mathcal{S} = \{s_1, s_2, s_3\}$ $\mathbf{w} = (w_1, w_2)^\top$ / $\mathbf{w} = (w_1, w_2)^\top$

$v_{\mathbf{w}}$ - 11.3

π v_π v_π

v_π v_1 v_2 $v = v_1 - v_2$ v

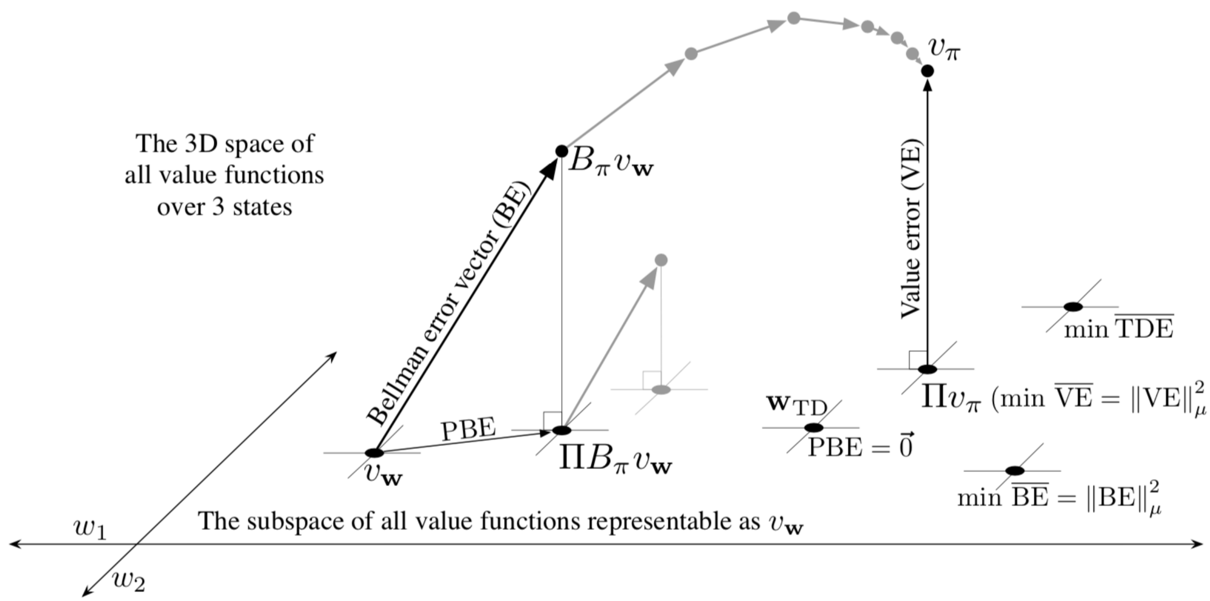
9.2 9.11 9.2 $\mu : \mathcal{S} \to [0, 1]$

$$\|v\|_\mu^2 \doteq \sum_{s \in \mathcal{S}} \mu(s) v(s)^2 \tag{3.116}$$

9.2 $\overline{\mathrm{VE}}$ $\overline{\mathrm{VE}}(\mathbf{w}) = \|v_{\mathbf{w}} - v_\pi\|_\mu^2$ v Π

$$\Pi v \doteq v_{\mathbf{w}} \quad \mathbf{w} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|v - v_{\mathbf{w}}\|_\mu^2 \tag{3.117}$$

v_π Π_{v_π} 11.3



68: **11.3** - $\mathbf{w} = (w_1, w_2)^\top$ v_π Π VE Bellman BE Bellman PBE TDE VE BE PBE Bellman Bellman PBE

$|\mathcal{S}| \times |\mathcal{S}|$

$$\Pi \doteq \mathbf{X}\left(\mathbf{X}^\top \mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{D} \tag{3.118}$$

9.4 $\mathbf{D}$ $|\mathcal{S}| \times |\mathcal{S}|$ $\mu(s)$ $\mathbf{X}$ $|\mathcal{S}| \times d$ $\mathbf{x}(s)^\top$ s 11.14

$$\|v\|_\mu^2 = v^\top \mathbf{D} v \tag{3.119}$$

$$v_\mathbf{w} = \mathbf{X}\mathbf{w} \tag{3.120}$$

TD v_π Bellman

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p\left(s', r|s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right], \qquad s \in \mathcal{S} \tag{3.121}$$

v_π 11.13 $v_\mathbf{w}$ v_π $v_\mathbf{w}$ v_π s *Bellman*

$$\begin{aligned} \bar{\delta}_\mathbf{w}(s) &\doteq \left(\sum_a \pi(a|s) \sum_{s',r} p\left(s', r|s, a\right)\left[r + \gamma v_\mathbf{w}\left(s'\right)\right]\right) - v_\mathbf{w}(s) && (11.17) \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma v_\mathbf{w}\left(S_{t+1}\right) - v_\mathbf{w}\left(S_t\right) | S_t = s, A_t \sim \pi\right] && (11.18) \end{aligned}$$

Bellman TD 11.3 Bellman TD

Bellman $\overline{\delta}_{\mathbf{w}} \in \mathbb{R}^{|\mathcal{S}|}$ *Bellman* 11.3 BE *Bellman*

$$\overline{\mathrm{BE}}(\mathbf{w}) = \left\|\overline{\delta}_{\mathbf{w}}\right\|_{\mu}^{2} \tag{3.122}$$

$\overline{\mathrm{BE}}$ $v_{\mathbf{w}} = v_{\pi}$ $\mathbf{w}$ $\overline{\mathrm{BE}}$ 11.3 $\min \overline{\mathrm{BE}}$ $\overline{\mathrm{VE}}$ Πv_{π} $\overline{\mathrm{BE}}$ Bellman 11.3 *Bellman* $B_{\pi} : \mathbb{R}^{|\mathcal{S}|} \rightarrow \mathbb{R}^{|\mathcal{S}|}$ Bellman

$$\left(B_{\pi} v\right)(s) \doteq \sum_{a} \pi(a|s) \sum_{s', r} p\left(s', r|s, a\right)\left[r+\gamma v\left(s'\right)\right] \tag{3.123}$$

$s \in \mathcal{S}$ $v : \mathcal{S} \rightarrow \mathbb{R}$ v Bellman $\overline{\delta}_{\mathbf{w}} = B_{\pi} v_{\mathbf{w}} - v_{\mathbf{w}}$

Bellman 11.3 v_{π} Bellman

$$v_{\pi} = B_{\pi} v_{\pi} \tag{3.124}$$

π 11.13 Bellman

11.3 Bellman DP

Bellman Bellman $\Pi\overline{\delta}_{v_{\mathbf{w}}}$ 11.3 PBE v *Bellman* $\overline{\mathrm{PBE}}$

$$\overline{\mathrm{PBE}}(\mathbf{w}) = \left\|\Pi \overline{\delta}_{\mathbf{w}}\right\|_{\mu}^{2} \tag{3.125}$$

$\overline{\mathrm{PBE}}$ 9.4 TD $\mathbf{w}_{TD}$ TD $\overline{\mathrm{VE}}$ $\overline{\mathrm{BE}}$ 11.7 11.8

11.5 Bellman

SGD 9.3 SGD SGD Tsitsiklis Van Roy 1997 S-GD

SGD *Bellman*

Bellman TD TD TD

$$\delta_{t} = R_{t+1} + \gamma \hat{v}\left(S_{t+1}, \mathbf{w}_{t}\right) - \hat{v}\left(S_{t}, \mathbf{w}_{t}\right)$$

TD

$$\begin{aligned} \overline{\mathrm{TDE}}(\mathbf{w}) &= \sum_{s \in \mathcal{S}} \mu(s) \mathbb{E}\left[\delta_{t}^{2} | S_{t} = s, A_{t} \sim \pi\right] \\ &= \sum_{s \in \mathcal{S}} \mu(s) \mathbb{E}\left[\rho_{t} \delta_{t}^{2} | S_{t} = s, A_{t} \sim b\right] \\ &= \mathbb{E}_{b}\left[\rho_{t} \delta_{t}^{2}\right] && (\ \mu\ b\ ) \end{aligned}$$

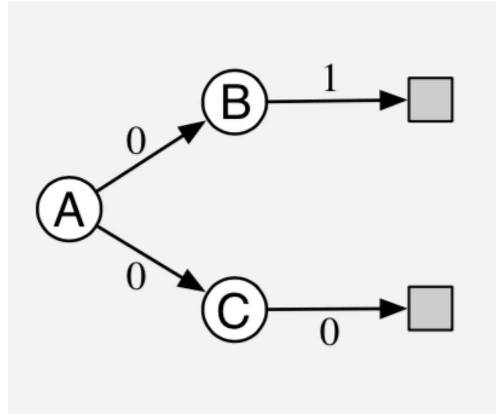
SGD b SGD

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_{t} - \frac{1}{2} \alpha \nabla\left(\rho_{t} \delta_{t}^{2}\right) \\ &= \mathbf{w}_{t} - \alpha \rho_{t} \delta_{t} \nabla \delta_{t} \\ &= \mathbf{w}_{t} + \alpha \rho_{t} \delta_{t}\left(\nabla \hat{v}\left(S_{t}, \mathbf{w}_{t}\right) - \gamma \nabla \hat{v}\left(S_{t+1}, \mathbf{w}_{t}\right)\right) \end{aligned}$$

TD 11.2 SGD Baird 1995

11.2 A

MRP **A** “ ” **B** 1 **C**
A γ 1 ρ_t 1



A 1 0 **A** $\frac{1}{2}$ **B** 1 1
C 0 0

B **C** $\frac{3}{4}$ B $\frac{1}{4}$ C A $\frac{1}{2}$ $\overline{\mathrm{TDE}}$

$\overline{\mathrm{TDE}}$ **A** $\frac{1}{2}$ **B** $\frac{3}{4}$ $\frac{1}{4}$ **A** $\frac{1}{2}$ **C** $\frac{1}{4}$
$-\frac{1}{4}$ $\gamma = 1$ TD TD $\frac{1}{16}$ **B**
$\frac{3}{4}$ 1 0 **C** $\frac{1}{4}$ 0 0 TD $\pm\frac{1}{4}$ $\frac{1}{16}$
TDE $\frac{1}{16}$

$\overline{\mathrm{TDE}}$ **B** 1 **C** 0 **A** 1 $\frac{1}{2}$ **B** 1 $\frac{1}{2}$ **C**
0 $\frac{1}{2}$ $\frac{1}{4}$ 1 0 **B** **C**
TD $\frac{1}{4}$ 0 $\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{16}$ $\overline{\mathrm{TDE}}$
$\overline{\mathrm{TDE}}$

A $\overline{\mathrm{TDE}}$ $\overline{\mathrm{TDE}}$ TD

Bellman Bellman Bellman A Bellman
Bellman TD Bellman TD TD S_t

$$
\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{1}{2}\alpha\nabla\left(\mathbb{E}_\pi\left[\delta_t\right]^2\right) \\
&= \mathbf{w}_t - \frac{1}{2}\alpha\nabla\left(\mathbb{E}_b\left[\rho_t\delta_t\right]^2\right) \\
&= \mathbf{w}_t - \alpha\mathbb{E}_b\left[\rho_t\delta_t\right]\nabla\mathbb{E}_b\left[\rho_t\delta_t\right] \\
&= \mathbf{w}_t - \alpha\mathbb{E}_b\left[\rho_t\left(R_{t+1}+\gamma\hat{v}\left(S_{t+1},\mathbf{w}\right)-\hat{v}\left(S_t,\mathbf{w}\right)\right)\right]\mathbb{E}_b\left[\rho_t\nabla\delta_t\right] \\
&= \mathbf{w}_t + \alpha\left[\mathbb{E}_b\left[\rho_t\left(R_{t+1}+\gamma\hat{v}\left(S_{t+1},\mathbf{w}\right)\right)\right]-\hat{v}\left(S_t,\mathbf{w}\right)\right]\left[\nabla\hat{v}\left(S_t,\mathbf{w}\right)-\gamma\mathbb{E}_b\left[\rho_t\nabla\hat{v}\left(S_{t+1},\mathbf{w}\right)\right]\right]
\end{aligned}
$$

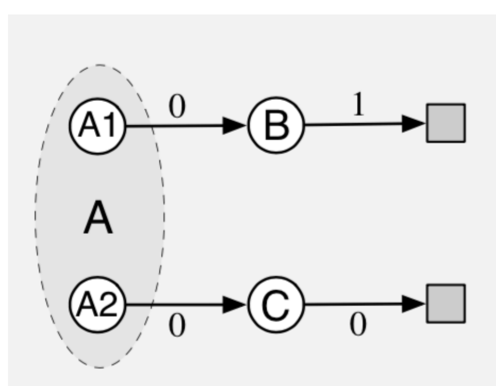
11.23 [1] S_{t+1}

S_t S_{t+1}
$\overline{\mathrm{BE}}$ SGD $\overline{\mathrm{BE}}$ $\mathbf{w}$

Baird Moore 1999
A Bellman $\overline{\mathrm{BE}}$
A A $\overline{\mathrm{BE}}$

11.3 A $\overline{\mathrm{BE}}$

MRP **A1** **A2** **A**
B **C** **B** **C** **A1**
A2 **A1** **B** 0 1 **A2**
C



A **A** **B** **C** 1 0
A **B** **C** 1 0 **A1** **A2** $\frac{1}{2}$

A TD **B** **C**
$\frac{3}{4}$ $\frac{1}{4}$ “ ” $\overline{\mathrm{BE}}$

[1] ρ_t

Bellman TD $\overline{\text{BE}}$ $\overline{\text{TDE}}$ $\overline{\text{BE}}$
A

$\overline{\text{BE}}$

11.6 Bellman

effciently " "[2] Bellman $\overline{\text{BE}}$
Bellman
[3] MRP

$x = 1$ w MRP 0 2 0.5
MRP MRP 0 2 0 2 MRP MR-
P MRP MRP

MRP $\overline{\text{VE}}$ 9.1 $\gamma = 0$ MRP 1,0 2 $w = 1$ MRP $\overline{\text{VE}}$ 0 MRP VE 1 $\overline{\text{VE}}$
$\overline{\text{VE}}$ $\overline{\text{VE}}$ $\overline{\text{VE}}$

$\overline{\text{VE}}$ MRP $w = 1$ MRP μ MDP
$\overline{\text{VE}}$ $\overline{\text{VE}}$

$\overline{\text{RE}}$ μ $\overline{\text{RE}}$

$$\begin{aligned}\overline{\text{RE}}(\mathbf{w}) &= \mathbb{E}\left[\left(G_t - \hat{v}\left(S_t, \mathbf{w}\right)\right)^2\right] \\ &= \overline{\text{VE}}(\mathbf{w}) + \mathbb{E}\left[\left(G_t - v_\pi\left(S_t\right)\right)^2\right] \qquad (11.24)\end{aligned}$$

$\mathbf{w}^*$ 11.4

11.4 11.24 $\overline{\text{RE}}$ $S_t = s$ s s 11.24
$\overline{\text{BE}}$ $\overline{\text{BE}}$ $\overline{\text{VE}}$ MDP $\overline{\text{RE}}$ MRP
$\overline{\text{PBE}}$ $\overline{\text{TDE}}$ $\overline{\text{BE}}$ 11.4

11.4 Bellman

MRP MRP

MRP MRP **B** **B'** $\mathbf{w}$ **A** **B** **B'**
MRP s $\mu(s) = 1$

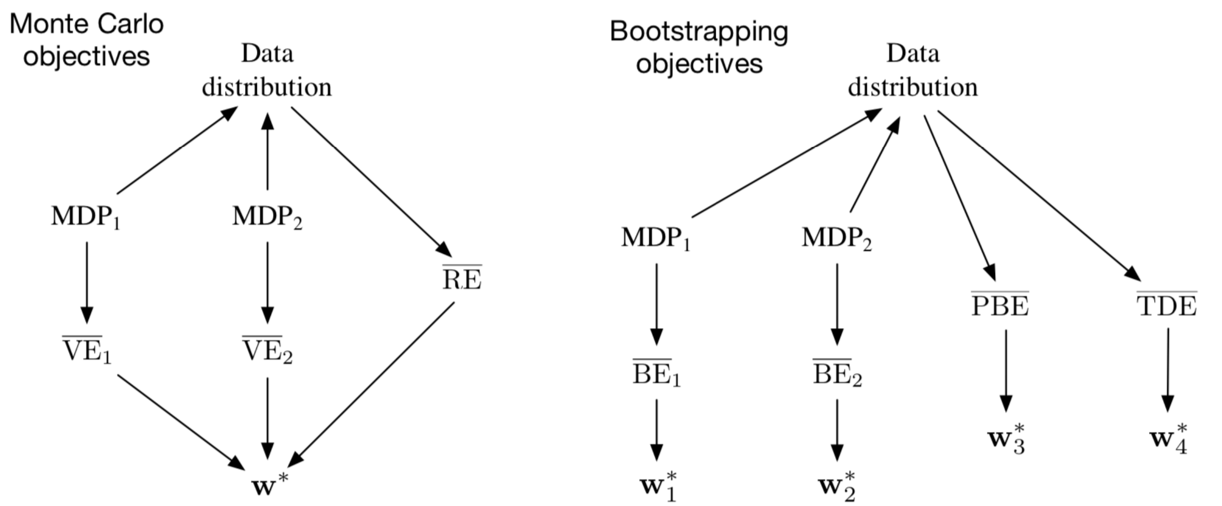
MRP **A** 0 B -1 1 **A** 0 MRP k B 2^{-k}

$\mathbf{w} = \mathbf{0}$ MRP $\overline{\text{BE}}$ MRP **B** **B'** 1 $\overline{\text{BE}} = \mu(\text{B})1 + \mu\left(\text{B}'\right)1 = \frac{2}{3}$ MRP
$\overline{\text{BE}}$ $\overline{\text{BE}}$

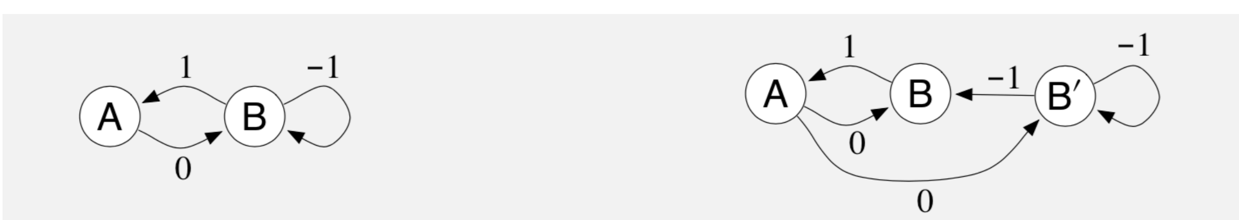
$\overline{\text{VE}}$ MRP $\mathbf{w}$ MRP $\mathbf{w} = \mathbf{0}$ $\overline{\text{BE}}$ MRP $\mathbf{w}$ γ $\gamma \to 1$ $\left(-\frac{1}{2}, 0\right)^\top$
$\overline{\text{BE}}$ MRP BE

---

[2]
[3] MRP MDP MRP MDP



69: **11.4** MDP MDP $\overline{\mathrm{VE}}$ $\overline{\mathrm{VE}}$ $\overline{\mathrm{VE}}$ $\mathbf{w}^*$ $\mathbf{w}^*$ $\overline{\mathrm{RE}}$ $\overline{\mathrm{VE}}$ $\mathbf{w}^*$ $\overline{\mathrm{RE}}$ MDP $\overline{\mathrm{BE}}$ $\overline{\mathrm{PBE}}$ $\overline{\mathrm{TDE}}$



MRP **A** $\overline{\mathrm{BE}}$ **A** **A** 0 0 $v_{\mathbf{w}}(A)$ 0 0 **B** **A** $v_{\mathbf{w}}(A)$ 1 **B** **A** 1 **B** **A** -1 $\overline{\mathrm{BE}}$ **A** $\approx -\frac{1}{2}$ **A**

$\overline{\mathrm{BE}}$ $\overline{\mathrm{BE}}$ $\overline{\mathrm{BE}}$ MDP $\overline{\mathrm{BE}}$ $\overline{\mathrm{BE}}$ MDP 11.3 A $\overline{\mathrm{BE}}$ $\overline{\mathrm{PBE}}$

11.7 TD

$\overline{\mathrm{PBE}}$ SGD SGD -*TD* TD $\mathbf{W}_{\mathrm{TD}}$ $\overline{\mathrm{PBE}}$ 9.8 $O\left(d^2\right)$ SGD $O(d)$ -TD

$\overline{\mathrm{PBE}}$ SGD 11.22

$$
\begin{aligned}
\overline{\mathrm{PBE}}(\mathbf{w}) &= \left\|\Pi\bar{\delta}_{\mathbf{w}}\right\|_{\mu}^{2} \\
&= \left(\Pi\bar{\delta}_{\mathbf{w}}\right)^{\top}\mathbf{D}\Pi\bar{\delta}_{\mathbf{w}} && \text{((11.15))} \\
&= \bar{\delta}_{\mathbf{w}}^{\top}\Pi^{\top}\mathbf{D}\Pi\bar{\delta}_{\mathbf{w}} \\
&= \bar{\delta}_{\mathbf{w}}^{\top}\mathbf{D}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{D}\bar{\delta}_{\mathbf{w}} && (11.25) \\
&\left(\text{ (11.14) } \Pi^{\top}\mathbf{D}\Pi = \mathbf{D}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{D}\right) \\
&= \left(\mathbf{X}^{\top}\mathbf{D}\bar{\delta}_{\mathbf{w}}\right)^{\top}\left(\mathbf{X}^{\top}\mathbf{D}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{\top}\mathbf{D}\bar{\delta}_{\mathbf{w}}\right) && (11.26)
\end{aligned}
$$

$\mathbf{w}$

$$
\nabla\overline{\mathrm{PBE}}(\mathbf{w}) = 2\nabla\left[\mathbf{X}^{\top}\mathbf{D}\bar{\delta}_{\mathbf{w}}\right]^{\top}\left(\mathbf{X}^{\top}\mathbf{D}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{\top}\mathbf{D}\bar{\delta}_{\mathbf{w}}\right)
$$

SGD μ

$$
\mathbf{X}^{\top}\mathbf{D}\bar{\delta}_{\mathbf{w}} = \sum_{s}\mu(s)\mathbf{x}(s)\bar{\delta}_{\mathbf{w}}(s) = \mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right]
$$

TD(0) 11.2

$$
\begin{aligned}
\nabla\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right]^{\top} &= \mathbb{E}\left[\rho_t\nabla\delta_t^{\top}\mathbf{x}_t^{\top}\right] \\
&= \mathbb{E}\left[\rho_t\nabla\left(R_{t+1} + \gamma\mathbf{w}^{\top}\mathbf{x}_{t+1} - \mathbf{w}^{\top}\mathbf{x}_t\right)^{\top}\mathbf{x}_t^{\top}\right] && (\ \delta_t) \\
&= \mathbb{E}\left[\rho_t\left(\gamma\mathbf{x}_{t+1} - \mathbf{x}_t\right)\mathbf{x}_t^{\top}\right]
\end{aligned}
$$

$$
\mathbf{X}^{\top}\mathbf{D}\mathbf{X} = \sum_{s}\mu(s)\mathbf{x}_s\mathbf{x}_s^{\top} = \mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^{\top}\right]
$$

$\overline{\mathrm{PBE}}$

$$
\nabla\overline{\mathrm{PBE}}(\mathbf{w}) = 2\mathbb{E}\left[\rho_t\left(\gamma\mathbf{x}_{t+1} - \mathbf{x}_t\right)\mathbf{x}_t^{\top}\right]\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^{\top}\right]^{-1}\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right] \tag{3.126}
$$

$\mathbf{x}_{t+1}$

$d \times d$

9.8 $O\left(d^2\right)$

-TD -TD 11.27 $d \times d$ d d $\mathbf{w}$ $\mathbf{v}$

$$
\mathbf{v} \approx \mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^{\top}\right]^{-1}\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right] \tag{3.127}
$$

$\rho_t\delta_t$ $(\mathbf{v}^\top\mathbf{x}_t - \rho_t\delta_t)^2$ $\mathbf{v}$ SGD LMS

$$\mathbf{v}_{t+1} \doteq \mathbf{v}_t + \beta\rho_t\left(\delta_t - \mathbf{v}_t^\top\mathbf{x}_t\right)\mathbf{x}_t$$

$\beta > 0$ 11.28 $O(d)$

$\mathbf{v}_t$ 11.28 11.27 SGD $\mathbf{w}_t$

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{1}{2}\alpha\nabla\overline{\mathrm{PBE}}(\mathbf{w}_t) && (\quad \text{SGD} \quad)\\
&= \mathbf{w}_t - \frac{1}{2}\alpha 2\mathbb{E}\left[\rho_t(\gamma\mathbf{x}_{t+1} - \mathbf{x}_t)\mathbf{x}_t^\top\right]\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\right]^{-1}\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right] && ((11.27))\\
&= \mathbf{w}_t - \frac{1}{2}\alpha 2\mathbb{E}\left[\rho_t(\gamma_t - \gamma\mathbf{x}_{t+1})\mathbf{x}_t^\top\right]\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\right]^{-1}\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right] && (11.29)\\
&\approx \mathbf{w}_t + \alpha\mathbb{E}\left[\rho_t(\mathbf{x}_t - \gamma\mathbf{x}_{t+1})\mathbf{x}_t^\top\right]\mathbf{V}_t && (\quad (11.28))\\
&\approx \mathbf{w}_t + \alpha\rho_t(\mathbf{x}_t - \gamma\mathbf{x}_{t+1})\mathbf{x}_t^\top\mathbf{v}_t && (\quad)
\end{aligned}$$

GTD2 $\mathbf{x}_t^\top\mathbf{v}_t$ $O(d)$

$\mathbf{v}_t$ 11.29

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t + \alpha\mathbb{E}\left[\rho_t(\mathbf{x}_t - \gamma\mathbf{x}_{t+1})\mathbf{x}_t^\top\right]\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\right]^{-1}\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right]\\
&= \mathbf{w}_t + \alpha\left(\mathbb{E}\left[\rho_t\mathbf{x}_t\mathbf{x}_t^\top\right] - \gamma\mathbb{E}\left[\rho_t\mathbf{x}_{t+1}\mathbf{x}_t^\top\right]\right)\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\right]^{-1}\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right]\\
&= \mathbf{w}_t + \alpha\left(\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\right] - \gamma\mathbb{E}\left[\rho_t\mathbf{x}_{t+1}\mathbf{x}_t^\top\right]\right)\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\right]^{-1}\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right]\\
&= \mathbf{w}_t + \alpha\left(\mathbb{E}\left[\mathbf{x}_t\rho_t\delta_t\right] - \gamma\mathbb{E}\left[\rho_t\mathbf{x}_{t+1}\mathbf{x}_t^\top\right]\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\right]^{-1}\mathbb{E}\left[\rho_t\delta_t\mathbf{x}_t\right]\right)\\
&\approx \mathbf{w}_t + \alpha\left(\mathbb{E}\left[\mathbf{x}_t\rho_t\delta_t\right] - \gamma\mathbb{E}\left[\rho_t\mathbf{x}_{t+1}\mathbf{x}_t^\top\right]\mathbf{v}_t\right) && (\quad (11.28))\\
&\approx \mathbf{w}_t + \alpha\rho_t\left(\delta_t\mathbf{x}_t - \gamma\mathbf{x}_{t+1}\mathbf{x}_t^\top\mathbf{v}_t\right) && (\quad)
\end{aligned}$$

$\mathbf{x}_t^\top\mathbf{v}_t$ $O(d)$ *TDC* *TD(0)* *GTD(0)*

11.5 Baird TDC $\overline{\mathrm{PBE}}$ s $\hat{v}(s) = 0$ $\mathbf{w}$ $(1,1,1,1,1,1,4,-2)^\top$ 1000 $\overline{\mathrm{VE}}$ 2 $\overline{\mathrm{PBE}}$

GTD2 TDC $\mathbf{w}$ $\mathbf{v}$ *two-time-scale* α β $\beta \to 0$ $\frac{\alpha}{\beta} \to 0$

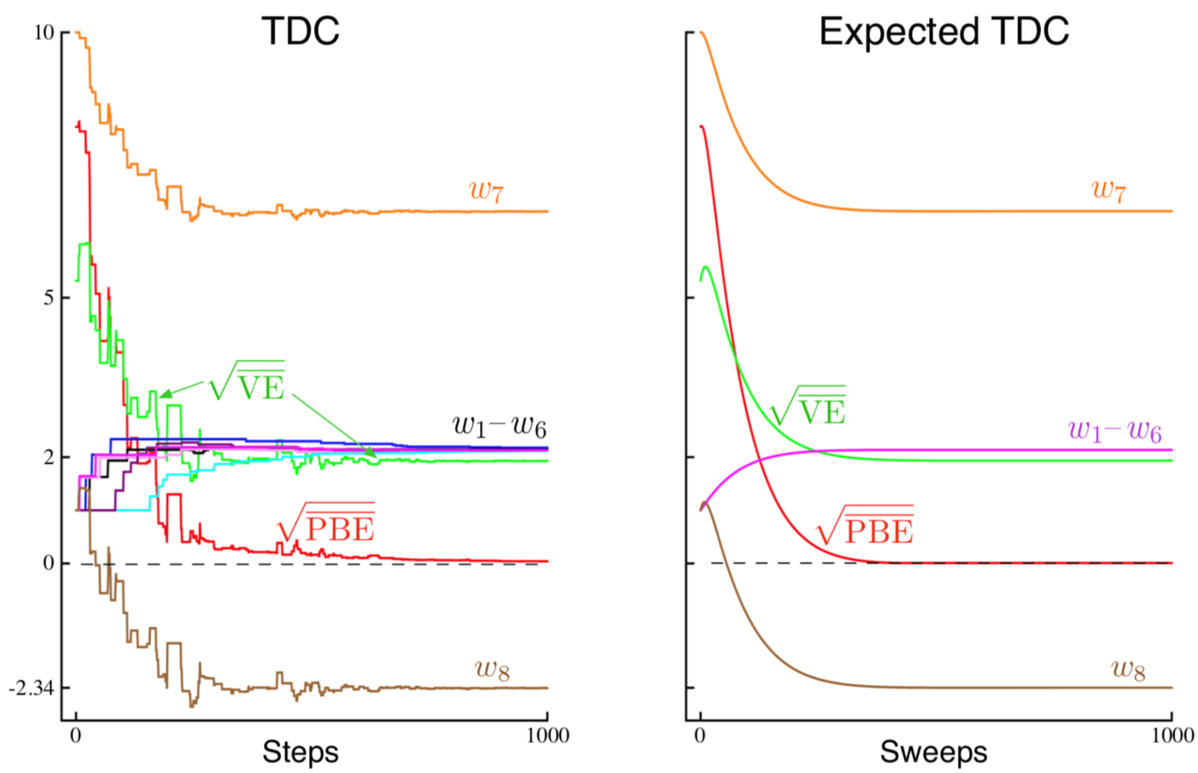
-TD GQ Maei 2010 GTD(λ) GQ(λ) Maei 2011 Maei Sutton 2010 Maei 2009 TD TD Hackman 2012 White and White 2016 TD - TD TD Mahadevan 2014 Du 2017

11.8 TD Emphatic-TD

TD 9.4 **A** 9.11 [4] μ_π $p(s|s,a)$ - TD 9.11 " "

$\gamma = 0.9$ 0.1 $1 - \gamma$

[4] **A** $\mathbb{E}_{s\sim b}\left[\mathbf{x}(s)\mathbb{E}\left[\mathbf{x}(S_{t+1})^\top|S_t = s, A_t \sim \pi\right]\right]$

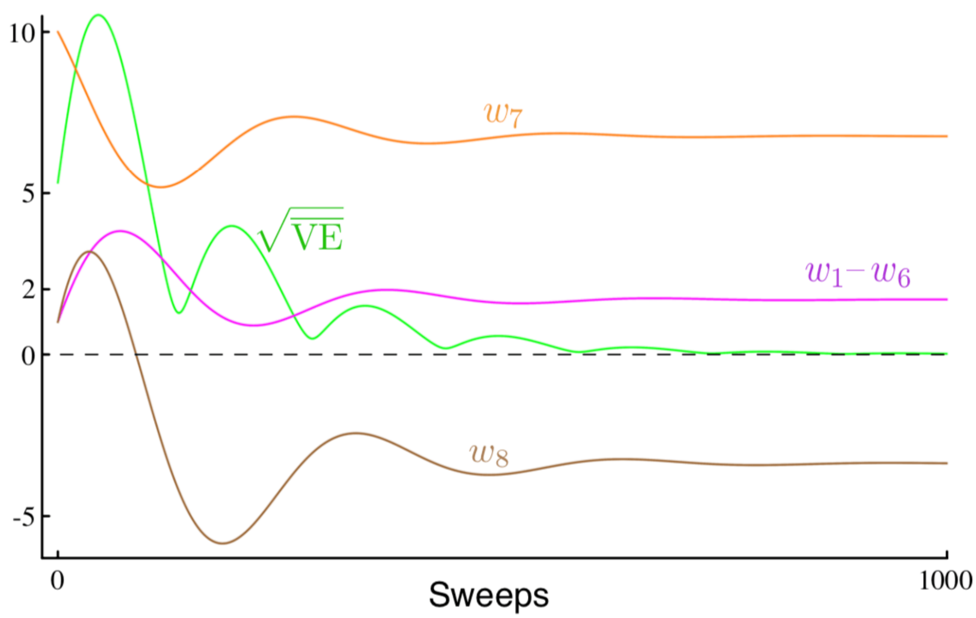


70: **11.5** Baird TDC 11.9 TDC $\alpha = 0.005$ $\beta = 0.05$

TD

$$\delta_t = R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)$$
$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha M_t \rho_t \delta_t \nabla \hat{v}(S_t, \mathbf{w}_t)$$
$$M_t = \gamma \rho_{t-1} M_{t-1} + I_t$$

I_t M_t $M_{t-1} = 0$ Baird 11.6 t $I_t = 1$ $\overline{\text{VE}}$ TD Baird



71: **11.6** TD Baird $\alpha = 0.03$

11.9

5.13
SGD SGD SGD
Polyak-Ruppert Polyak 1990 Ruppert 1988 Polyak Juditsky 1992 Derthick 1984 Jacobs 1988 Sutton 1992b c Karampatziakis Langford 2010 “ ”

5 $O(d)$ Mahmood Sutton 2015

7.5 Munos Stepleton Harutyunyan Bellemare 2016 Mahmood Yu Sutton 2017

Precup 2006 “ ”

11.10

Q-learning Sarsa

TD

TD TD
Bellman SGD Bellman $\overline{\text{BE}}$
TD Bellman SGD $\overline{\text{PBE}}$ $O(d)$ TD

11.1 TD(λ) Sutton 1988 “ ” Sutton 2015a Sutton Mahmood White 2016 TD(0) Precup Sutton Singh 2000 12 n

11.2 w – to – $2w$ Tsitsiklis Van Roy 1996 263 Baird Baird 1995 Gordon 1995,1996b Boyan Moore 1995 DP Bradtke 1993 Q-learning

11.3 Sutton 1995b Tsitsiklis Van Roy 1997 “ ” Sutton 2015a

11.4 Tsitsiklis Van Roy 1996; 1997 11.3 Lagoudakis Parr 2003 Bellman B_π I^π “ ” $T^{(\lambda)}$ “TD(λ) ” Tsitsiklis 1996 1997

11.5 $\overline{\text{BE}}$ Schweitzer Seidmann 1985 Baird 1995,1999 TD $\overline{\text{BE}}$ Bellman

A Dayan 1992 Sutton 2009 a

11.6

11.7 Sutton Szepesvari Maei 2009b -TD Sutton 2009a Mahmood 2014 Mahadevan 2014 TD -TD Geist Scherrer 2014 Dann Neumann Peters 2014 White 2015 Ghiassian Patterson White Sutton White 2018 Yu 2017 -TD

11.8 Sutton Mahmood White 2016 -TD Yu 2015; 2016; Yu Mahmood Sutton 2017 Hallak Tamar Mannor 2015 Hallak Tamar Munos Mannor 2016

3.6.4 12 Eligibility Traces

TD(λ) λ TD Q-learning Sarsa

TD TD $\lambda = 1$ $\lambda = 0$ TD

TD 7 n TD $\mathbf{z}_t \in \mathbb{R}^d$ $\mathbf{w}_t \in \mathbb{R}^d$ $\mathbf{w}_t$ $\mathbf{z}_t$ TD $\mathbf{w}_t$ $\lambda \in [0, 1]$

n n n

5 n TD 7 n n
TD
2014

12.1 λ

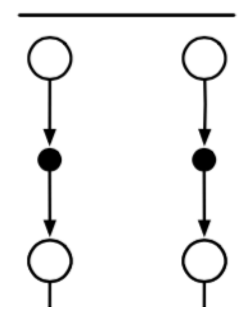
7 n n n 7.1

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{v}\left(S_{t+n}, \mathbf{w}_{t+n-1}\right), \quad 0 \leq t \leq T - n \tag{3.128}$$

$\hat{v}(s, \mathbf{w})$ $\mathbf{w}$ 9 s T 7 $n \geq 1$ n SGD 9.7

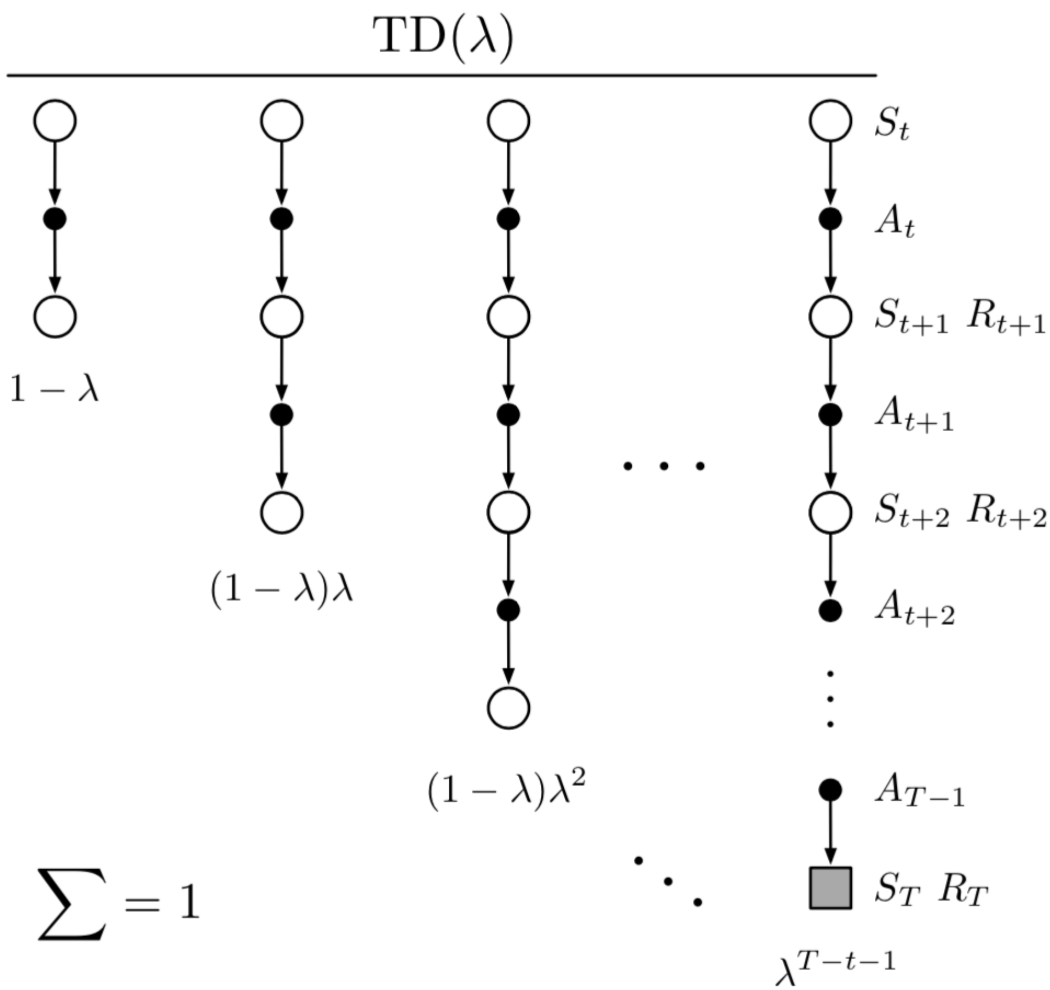
n n n $\frac{1}{2}G_{t:t+2} + \frac{1}{2}G_{t:t+4}$ n 1
n 7.3 TD DP 8

compound *update*
$t + 4$ t



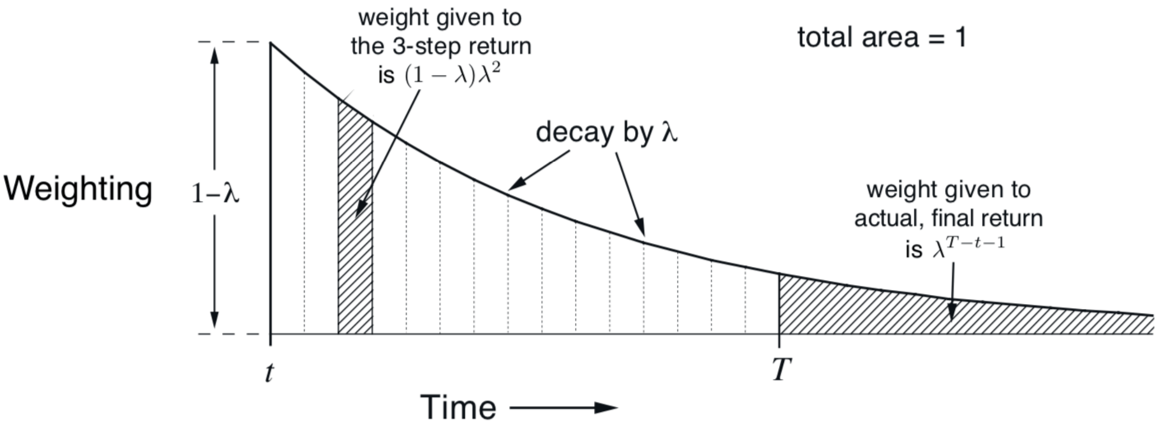
TD(λ) n n λ^{n-1} $\lambda \in [0, 1]$ $1 - \lambda$
1 12.1 λ

$$G_t^\lambda \doteq (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n} \tag{3.129}$$



72: **12.1** TD(λ) $\lambda = 0$ TD $\lambda = 0$

12.2 λ n $1 - \lambda$ $(1-\lambda)\lambda$ $(1-\lambda)\lambda^2$



73: **12.2** n λ

λ n G_t

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}G_{t+t+n} + \lambda^{T-t-1}G_t \qquad (3.130)$$

$\lambda = 1$ $\lambda = 1$ λ $\lambda = 0$
λ $G_{t:t+1}$ $\lambda = 0$ λ TD

12.1 3.9 λ 12.2 12.1

12.2 λ 12.2 λ λ
λ $\mathcal{T}_\lambda$

λ λ λ

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\left[G_t^\lambda - \hat{v}\left(S_t, \mathbf{w}_t\right)\right]\nabla\hat{v}\left(S_t, \mathbf{w}_t\right), \quad t = 0, \ldots, T-1 \qquad (3.131)$$

λ TD 7 n 19 7.1 144 12.3 λ
n 7.2 λ λ n 10 19
λ n n n λ λ

12.4

12.2 TD(λ)

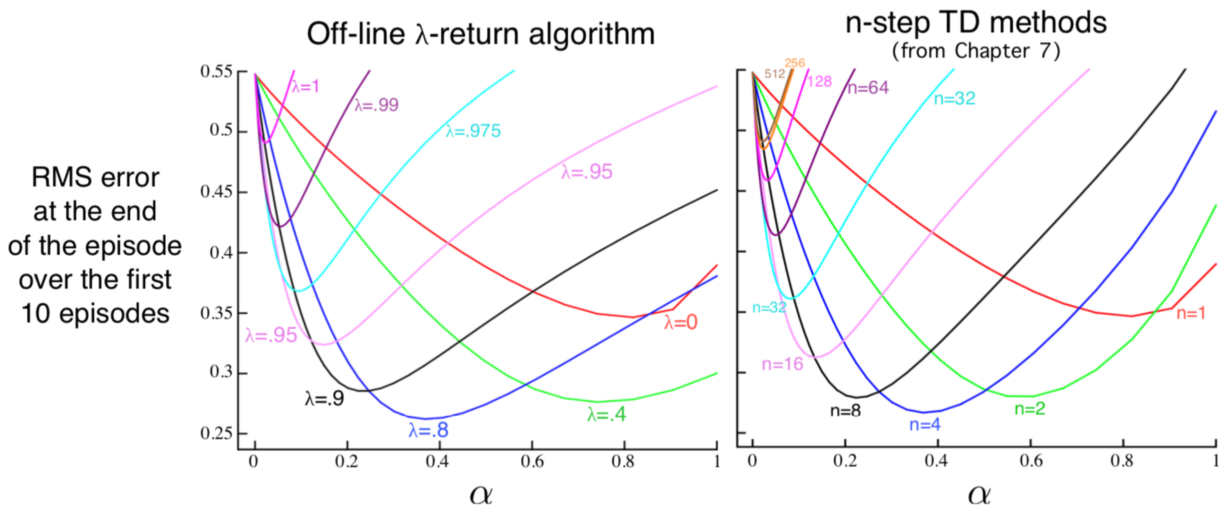
12.3 n λ

12.4 λ

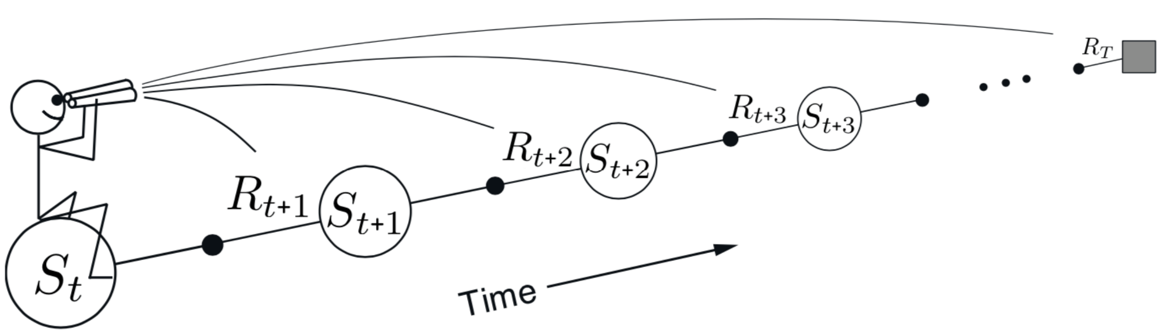
12.5 TD(λ)

12.6 Dutch

12.7 Sarsa(λ)



74: **12.3** 19 7.1 n TD λ n λ α λ α



75: **12.4**

12.8 λ γ

12.9

12.10 Watkins Q(λ) Tree-Backup(λ)

12.11

12.12

12.13

3.6.5 13

[1]

$\theta \in \mathbb{R}^{d'}$ t s θ t a $\pi(a|s,\theta) = Pr\{A_t = a|S_t = s, \theta_t = \theta\}$
$\mathbf{w} \in \mathbb{R}^d$ $\hat{v}(s,\mathbf{w})$

$J(\theta)$ J

$$\theta_{t+1} = \theta_t + \alpha \widehat{\nabla J(\theta_t)} \tag{3.132}$$

$\widehat{\nabla J(\theta_t)} \in \mathbb{R}^{d'}$ θ_t -
“ ” “ ” 10.3

13.1

$\pi(a|s,\boldsymbol{\theta})$ $\nabla\pi(a|s,\boldsymbol{\theta})$ ($\pi(a|s,\boldsymbol{\theta})$ $\boldsymbol{\theta}$) $s \in \mathcal{S}$ $a \in \mathcal{A}(s)$ $\boldsymbol{\theta} \in \mathbb{R}^{d'}$
s a $\boldsymbol{\theta}$ $\pi(a|s,\boldsymbol{\theta}) \in (0,1)$ 13.7

- preferences $h(s,a,\boldsymbol{\theta}) \in \mathbb{R}$ soft-max

$$\pi(a|s,\boldsymbol{\theta}) \doteq \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_b e^{h(s,b,\boldsymbol{\theta})}} \tag{3.133}$$

$e \approx 2.71828$ 1 *soft-max* soft-max in action preferences

ANN $\boldsymbol{\theta}$ 16.6 AlphaGo

$$h(s,a,\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}(s,a) \tag{3.134}$$

9.5 $\mathbf{x}(s,a) \in \mathbb{R}^{d'}$

soft-max ε ε soft-max
0 1 soft-max

soft-max (bluffing)
13.1

13.1

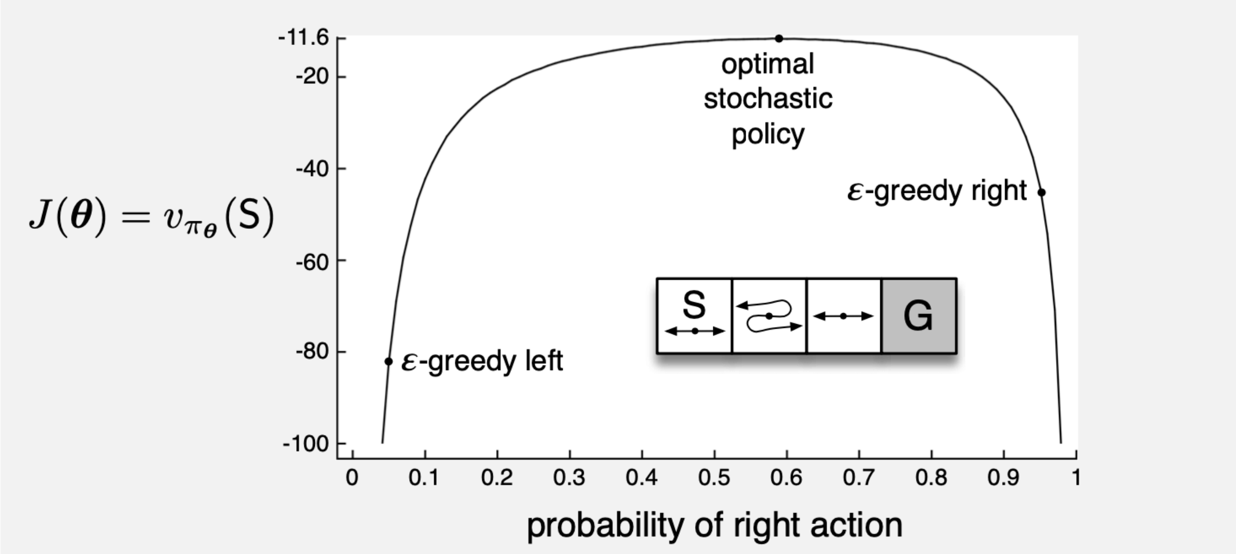
---

[1] 2.8 MDP

-1

s $\mathbf{x}(s,\text{ right }) = [1,0]^\top$ $\mathbf{x}(s,\text{ left }) = [0,1]^\top$ ε $1-\varepsilon/2$

$\varepsilon = 0.1$ -44 -82 0.59 -11.6



Şimşek, Algórta Kothiyal 2016

13.1 13.1

13.2

13.3

13.4

13.5 -

13.6

13.7

13.8

3.7

3.7.1 14

20

/

14.1

[1] 3

Modayil White Sutton 2014

1 Thorndike“ ” Law of Effect

/ /

“ ”

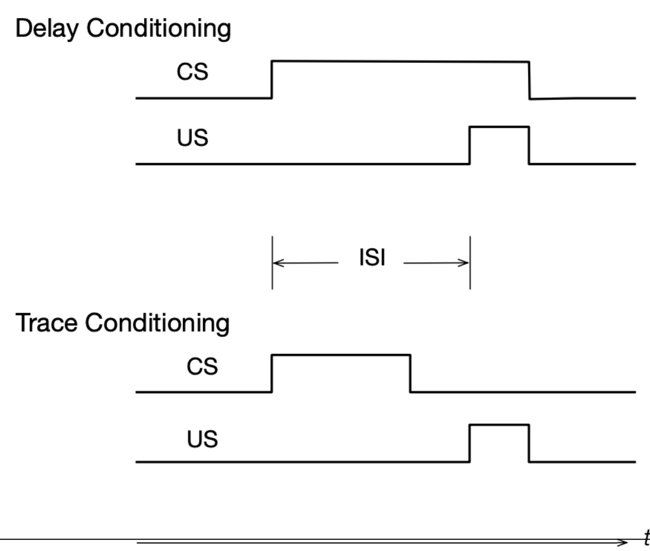
14.2

. 5 6 “ ” Pavlov 1927 22

Pavlov 1927 14

“ ” unconditioned response UR “ ” unconditioned stimuli US “ ” conditioned response CR “ ” conditioned stimulu CS “ ” “ ” US CR CS US

CS US interstimulus interval ISI CS US CS Us CS CS US



CS US C-s CR US CR UR

CS US UR

CR US

[1] “ ”

14.2.1

CR C-
S CR CS CR
CS US

/ CS
[2] US CS Rescorla–Wagner Rescorla and Wagner 1972

CS US
10 CS US CS
US CS US
TD CR Rescorla-Wagner

13.5 “ ” 15.7 15.8 TD
1.7 14.4

14.2.2 Rescorla-Wagner

Rescorla Wagner Rescorla-Wagner- Rescorla Wagner TD

Rescorla Wagner CS “ ” US CS
CS “ ”

Rescorla Wagners CS AX A X A X V_A V_X V_{AX} A X AX CS AX US US Y CS

$$\Delta V_A = \alpha_A \beta_Y \left(R_Y - V_{AX}\right)$$
$$\Delta V_X = \alpha_X \beta_Y \left(R_Y - V_{AX}\right)$$

$\alpha_A \beta_Y$ $\alpha_X \beta_Y$ US CS R_Y US Y Rescorla Wagner λ R R R US Hescorla-Wagner V_{AX} $V_A + V_X$ Δs

V CR Rescorla Wagner V CR V CR

Rescorla-Wagner CR V_{AX} US Y R_Y $R_Y - V_{AX}$ V_X V_{AX} R_Y US CS CS CS 0 CS CS US

Rescorla-Wagner TD TD 9.4 C-S “US ” US Y Rescorla-Wagner R_Y Rescorla-Wagner- TD CS

s $\mathbf{x}(s) = (x_1(s), x_2(s), \ldots, x_d(s))^\top$ CS i CS_i $x_i(s) = 1$ 0 d $mathbfw$ s

$$\hat{v}(s, \mathbf{w}) = \mathbf{w}^\top \mathbf{x}(s) \tag{3.135}$$

US

[2] Moore Schmajuk 2008

t TD t $S+t$ t t $\mathbf{w}_t$ $\mathbf{w}_{t+1}$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha\delta_t\mathbf{x}(S_t) \tag{3.136}$$

α Rescorla-Wagner δ_t

$$\delta_t = R_t - \hat{v}(S_t, \mathbf{w}_t) \tag{3.137}$$

R_t t US Rescorla Wagner US 14.2 $\mathbf{x}(S_t)$ CS US

Rescorla-Wagner LMS Widrow-Hoff Widrow Hoff, 1960 0 " " 9.4 [3]

Rescorla-Wagner " " US Rescorla-Wagner Moore and Schmajuk, 2008

Rescorla-Wagner T-D Rescorla-Wagner

14.2.3 TD

Rescorla-Wagner TD Rescorla-Wagner t 1 Rescorla-Wagner TD TD ——

Rescorla-Wagner TD t t $t+1$ 0.01 t t CS s $mathbfx(s)$

TD TD TD

TD s $\mathbf{x}(s) = (x_1(s), x_2(s), \ldots, x_n(s))^\top$ s Rescorla Wagner 14.1 TD $\mathbf{w}$ t TD

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha\delta_t\mathbf{z}_t \tag{3.138}$$

Rescorla-Wagner 14.2 $\mathbf{x}_t(S_t)$ $\mathbf{z}_t$ $\mathbf{z}_t$ δ_t 14.3 TD

$$\delta_t = R_{t+1} + \gamma\hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t) \tag{3.139}$$

γ 0 1 R_t t $\hat{v}(S_{t+1}, \mathbf{w}_t)$ $\hat{v}(S_t, \mathbf{w}_t)$ $t+1$ t 14.1 $\mathbf{z}_t$ i $x_i(S_t)$ $\mathbf{x}(S_t)$ $\gamma\lambda$

$$\mathbf{z}_{t+1} = \gamma\lambda\mathbf{z}_t + \mathbf{x}(S_t) \tag{3.140}$$

λ

$\gamma = 0$ TD Rescorla-Wagner t Rescorla-Wagner TD TD R TD 12 $TD(\lambda)$ TD R_t

[3] LMS Rescorl-Wagner LMS $\mathbf{x}_t$ α LMS

3.7.2 15

TD
TD

1 1.7
“ ”

15.1

“ ”

“ ” “ ”

“ ”

synaptic efficacies

15.2

3

14 R_t R_t R_t
R_t

t TD $\delta_{t-1} = R_t + \gamma V(S_t) - V(S_{t-1})$[1]
TD

V Q

[1] 6.1 δ_t $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ $t+1$ δ_t t TD $\delta_{t-1} = R_t + \gamma V(S_t) - V(S_{t-1})$

reward prediction errors RPE 6.5 TD RE RPE TD RPE TD TD Sarsa Q TD TD TD TD Modayil White Sutton 2014

RE TD Rescorla Wagner 14.3 TD “Q ” Sarsa RPE TD 15.1

15.3

Montague Dayan Sejnowski1996 TD 20 80 90 · 15.4 15.6 TD

Montague 1996 TD 14.2 TD λ Montague TD $\delta_{t-1} + b_t$ b_t TD [2]

14.2.4 Montague CSC 141 US TD TD

15.5 TD 15.5 TD 1 2 3 4

Schultz TD TD Montague CSC

15.4

SNpc VTA

· James Olds Peter Milner 1954 “…… ” olds Milner 1954

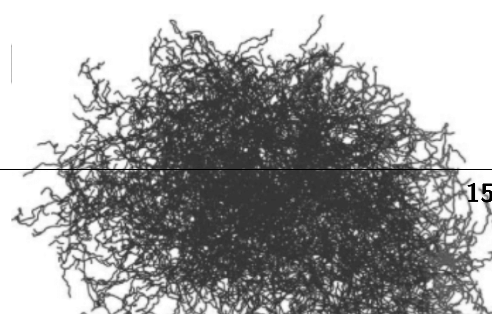
t $\delta_{t-1} = R_t + \gamma V(S_t) - V(S_{t-1})$ R_t

“ - - ” δ δ “ - ” 13.5 15.7 δ Q Sarsa R_t δ_{t-1} $\gamma V(S_t) - V(S_{t-1})$ δ_{t-1} $R_t \neq 0$ TD 15.6

Olds Milner1 1954

Tsai et al 2009 Steinberg 2013 14.2.1

[2] TD δ_t $\delta_{t-1} = R_t + \gamma V(S_t) - V(S_{t-1})$



“ ” Claridge Chang 2009
TD δ
$minus\delta$

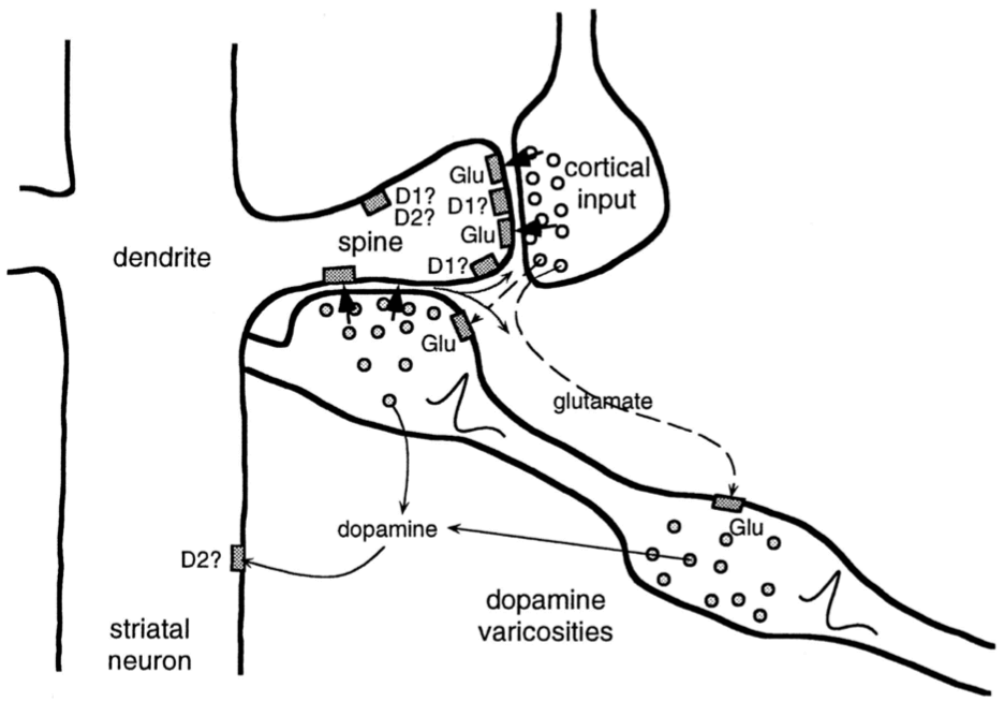
100~1000
SNpc spc VTA 500
000

δ SNpc V-
TA
R-
PE RPE RPE
RPE
15.10

“ ”

“ ” 15.1

15.1



77: **15.1** VTA SNpc
“ ” 500 000
D1 D2 Journal of Neurophysiology w. Schultz vol.80 1998 page10.

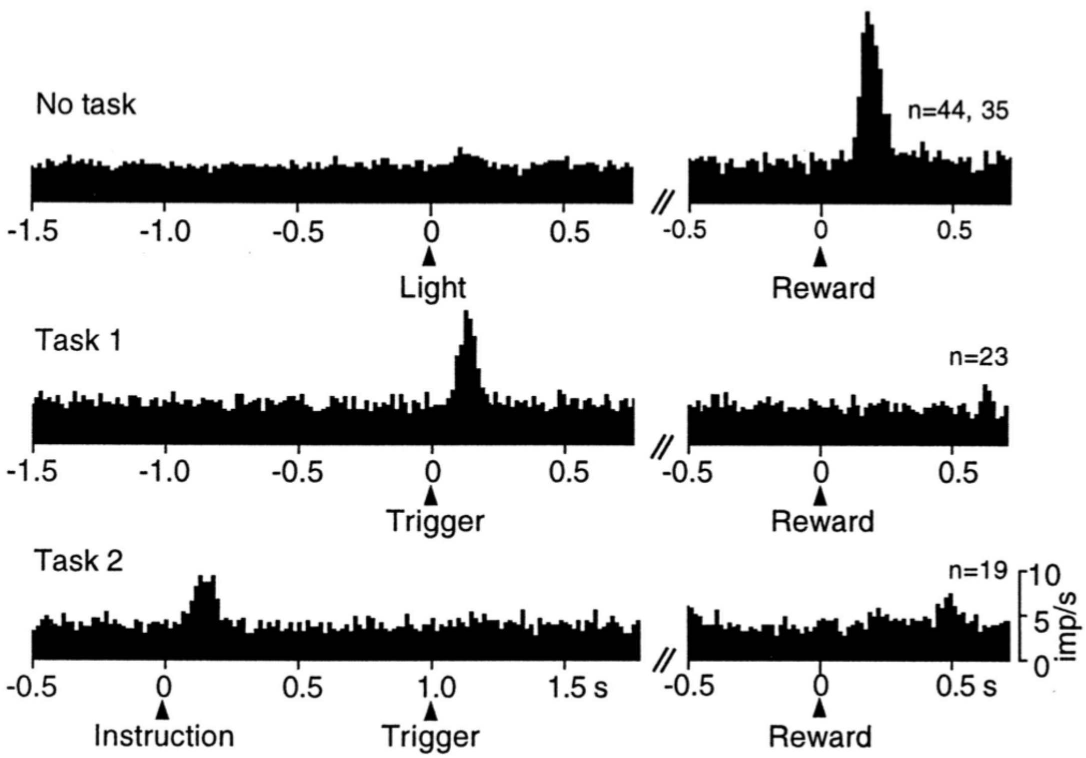
15.5

Romo
Schultz 1990 Schultz Romo 1990

Romo Schultz

Romo Schultz Romo Schultz
Romo Schultz

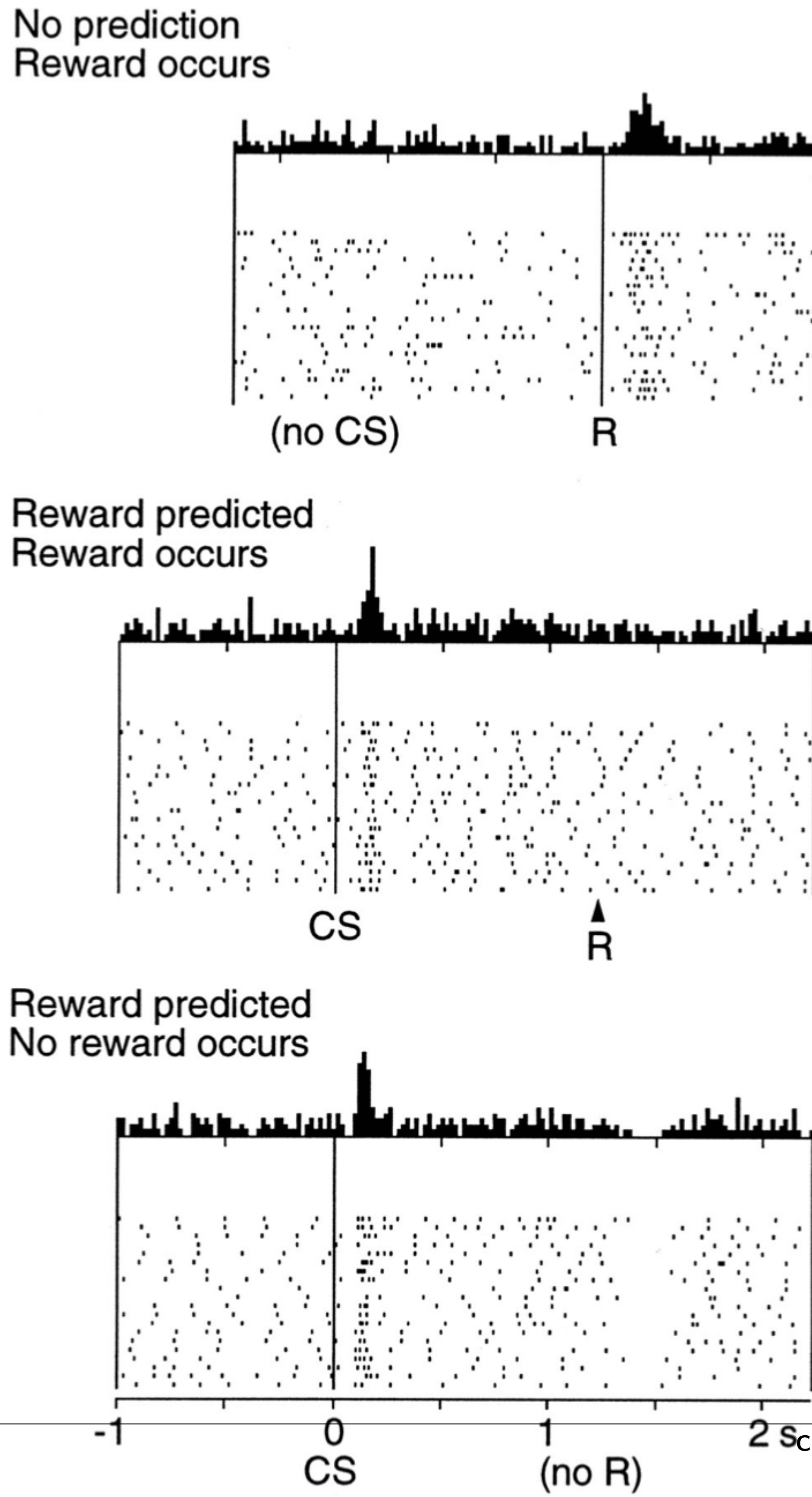
Schultz SNpc VT TD Rescorla Wagner 14.3 Ljungberg
Apicella Schultz 1992 “ ” Romo Schultz 15.2
15.2



78: **15.2** 23~4
1s Schultz et al. 1995 MIT Press.

Schultz Apicella Ljungberg 1993 “ ”
1 15.2
14.2
15.3

Schultz


No prediction
Reward occurs
(no CS)
R
Reward predicted
Reward occurs
CS
R
Reward predicted
No reward occurs
-1
0
1
2 s
CS
(no R)

15.6 TD

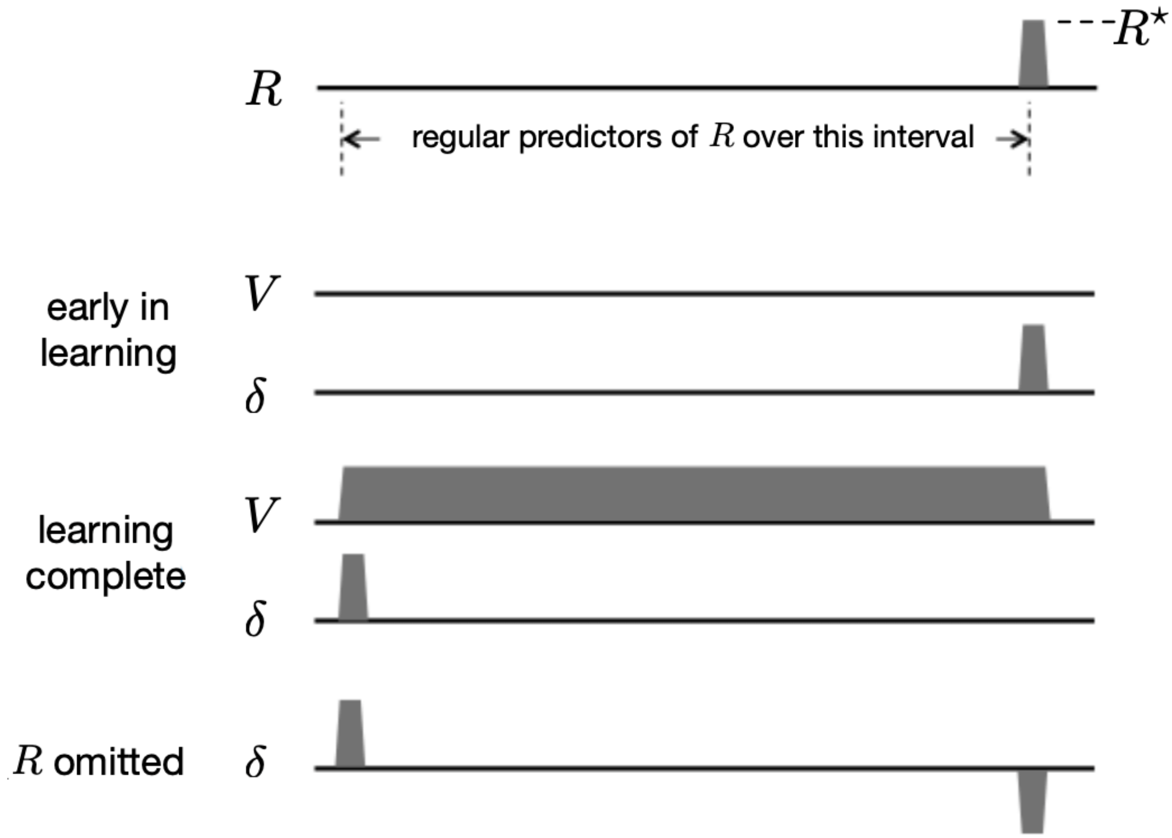
TD δ T-
D

4.1 6.1
“ -
”

Schultz “ ” “ ”

TD Montague CSC
I TD(0) V γ 1

15.4 R V δ $R^\star$
$R^\star$



80: **15.4** TD δ TD δ t δ_{t-1} R V δ $R^\star$
δ $\delta = 0$ R δ

Schultz
15.4

15.4 V δ “ ” V TD $R^\star$ $\delta_{t-1} = R_t + V_t - V_{t-1} = R_t + 0 - 0 = R_t$ $R^\star$ V_t $V_{t\,1}$ $t\,1$ TD

TD(0) 6 $R^\star$
15.4 V “ ” $R^\star$ 15.4 0

V TD $\delta_{t-1} = R_t + V_t - V_{t-1} = 0 + R^\star - R^\star = 0$
$\delta_{t-1} = R_t + V_t - V_{t-1} = R^\star + 0 - R^\star = 0$ TD
$\delta_{t-1} = R_t + V_t - V_{t-1} = 0 + R^\star - 0 = R^\star$ 15.4 “ ”

TD TD

TD $\delta_{t-1} = R_t + V_t - V_{t-1} = 0 + 0 - R^\star = -R^\star$ 15.4 “ R” δ
Schultz et al. 1993 15.3

TD

TD

TD δ
TD TD Montague et al. 1996 CSC TD TD TD Hollerman Schultz 198 CSC TD

TD CSC Suri Schultz 199 CSC
Daw Courville Touretzky 2006 TD Ludvig Sutton Kehoe2008 TD CSC 14.1 Pan Schmidt Wickens Hyland 205 CSC TD TD TD TD

TD
TD

TD TD
TD

TD
“ - ” TD “ - ”

15.7 “ - ”

“ - ” “ ” “ ” TD TD δ δ

“ - ” “ - ” 15.4 “ - ” TD

13.5 “ - ” Bart Sutton Anderson 1983 “ - ” Barto Takahashi Schoenbaum Niv 2008 “ - ” 15.8

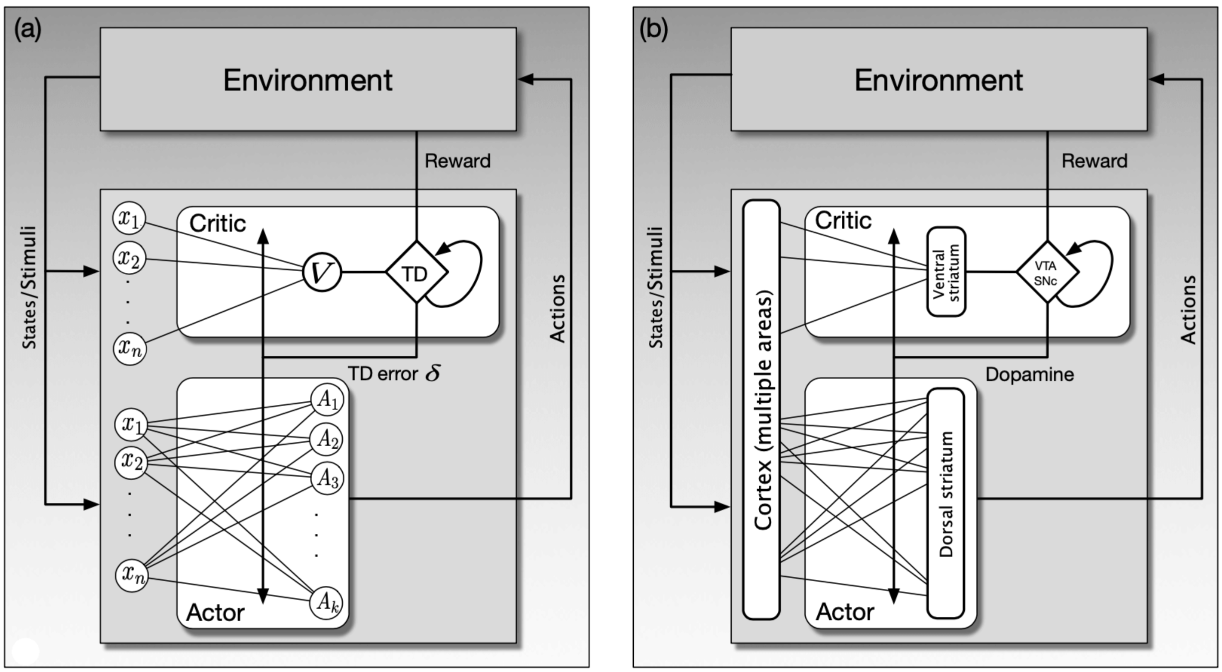
15.5a “ - ” V TD TD k $A_i, i = 1, \ldots, k$ k k A

1 “ ” $x_1, x_2, \ldots, x_n$
x_i V A_i “ ”

TD 15.5a “TD δ” “ - ”

15.5b Takahashi et al. 2008
15.4

“ - ” VTA SNpc TD 15.5a “TD δ” 15.5b “ ” VTA SNpc
15.1



81: **15.5** “ - ” a “ - ” TD δ R TD b “ - ” TD VTA SNp Frontiers in Neuroscience, vol. 2(1), 2008, Y. Takahashi, G. Schoenbaum, and Y. Niv, Silencing the critics: Understanding the e ects of cocaine sensitization on dorsolateral and ventral striatum in the context of an Actor/Critic model.

15.5b R_t R SNpc VTA 15.5b “ ” R_t

15.5b “ - ”

15.8

“ - 15.5b TD δ TD 1.7 δ TD δ 14.2 TD “ - ”

15.5b “ - ” 13.6 “ - ” S_t S_{t+1} A_t R_{t+1} TD $\mathbf{z}_t^{\mathbf{w}}$ $\mathbf{z}_t^{\theta}$ $\mathbf{w}$ θ

$$
\begin{aligned}
\delta_t &= R_{t+1} + \gamma \hat{v}\left(S_{t+1}, \mathbf{w}\right) - \hat{v}\left(S_t, \mathbf{w}\right) \\
\mathbf{z}_t^{\mathbf{w}} &= \lambda^{\mathbf{w}} \mathbf{z}_{t-1}^{\mathbf{w}} + \nabla \hat{v}\left(S_t, \mathbf{w}\right) \\
\mathbf{z}_t^{\theta} &= \lambda^{\theta} \mathbf{z}_{t-1}^{\theta} + \nabla \ln \pi\left(A_t | S_t, \boldsymbol{\theta}\right) \\
\mathbf{w} &\leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta_t \mathbf{z}_t^{\mathbf{w}} \\
\boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \delta \mathbf{z}_t^{\boldsymbol{\theta}}
\end{aligned}
$$

$\gamma \in [0,1)$ $\lambda^w c \in [0,1]$ $\lambda^w a \in [0,1]$ $\alpha^{\mathbf{w}} > 0$ $\alpha^{\theta} > 0$

$\hat{v}$ 15.5a V s $\mathbf{x}(s) = (x_1(s), \ldots, x_n(s))^\top$ $\mathbf{w} = (w_1, \ldots, w_n)^\top$

$$
\hat{v}(s, \mathbf{w}) = \mathbf{w}^\top \mathbf{x}(s) \tag{3.141}
$$

$x_i(s)$ w_i $\alpha^{\mathbf{w}} \delta_t \mathbf{z}_t^{\mathbf{w}}$ δ_t $\mathbf{z}_t^{\mathbf{w}}$ $\nabla \hat{v}(S_t, \mathbf{w})$ $\hat{v}(s, \mathbf{w})$ $\nabla \hat{v}(S_t, \mathbf{w}) = \mathbf{x}(S_t)$

$\mathbf{z}_t^{\mathbf{w}}$ $\mathbf{x}(S_t)$ $\lambda^{\mathbf{w}}$

14.2 TD 15.5a Barto et al. 1983 “ - ”

15.5a k t $\mathbf{x}(S_t)$ $j, j = 1, \ldots, k$ $\boldsymbol{\theta}_j$ “ - 0 1 A_t 1 $\boldsymbol{\theta}^\top \mathbf{x}(S_t)$ 13.2

$$
\pi(1|s, \boldsymbol{\theta}) = 1 - \pi(0|s, \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x}(s))} \tag{3.142}
$$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \delta_t \mathbf{z}_t^{\boldsymbol{\theta}}$ δ 15.5a δ_t 15.10 $\mathbf{z}_t^{\boldsymbol{\theta}}$ $\nabla \ln \pi(A_t | S_t, \boldsymbol{\theta})$ 13.5 a $\mathbf{x}(s)$ $\pi(a|s, \boldsymbol{\theta})$ $\nabla \ln \pi(a|s, \boldsymbol{\theta})$ t

$$
\nabla \ln \pi(A_t | S_t, \boldsymbol{\theta}) = (A_t - \pi(\mathbf{1}|S_t, \boldsymbol{\theta}))\mathbf{x}(S_t) \tag{3.143}
$$

$\mathbf{x}(S_t)$ 15.3 $\pi(\mathbf{1}|S_t, \boldsymbol{\theta})$ $A_t = 1$ δ δ δ δ

Hebb 1949 Hebbian δ δ

15.1 15.2 Heb-
bian

$(A_t - \pi(A_t|S_t, \boldsymbol{\theta}))\mathbf{x}(S_t)$ $(A_t - \pi(A_t|S_t, \boldsymbol{\theta}))$
$\mathbf{x}(S_t)$ $\mathbf{x}(S_t)$ $(A_t - \pi(A_t|S_t, \boldsymbol{\theta}))$

STDP STDP Hebbian
STDP Hebbian

STDP STDP STDP STD-
P STDP 15.5b “ -
STDP 10s etal.2014 ST-
D

Barto et al 1983 “ - ” A.H. Klopf 1972 1982 “ ”
Klopf STDP STDP Klopf Klopf

15.9

Klopf 1972 192
Klopf
Klopf “ ”
Klopf

Klopf

Klopf
Klopf λ γ Klopf

Klopf

. “ ”
“ ” Selfridge 1978 1984 “ ”

Klopf

15.10

Klopf “ - ”

15.5a k $(A_1, A_2, \dots, A_k)^\top$ δ δ δ

“ ”

15.8
15.8

“ - ”
“ - ”

15.8 REINFORCE 13
Williams 1992 REINFORCE

Williams 1992 REINFORCE
REINFORCE 15.8 STDP

15.11

14.6 “ - ” “ -
”

14.6 “ - ” DLS
DMS
OFC

Tolman 14.5 Hassabis
Maguire 20 Olafsdottir Barry Saleem Hassabis Spiers 2015

“ ” John-
son Redish 2007 Pfeiffer Foster 2013 “ ”

8.10 Dyna 8.2

Doll Simon Daw 2012 “ ”

15.12

15.7

TD Redish 2004 TD δ
15.6 δ “ ” TD

Redish
Redish

15.13

TD Wolfram Schultz

TD

15 R_t TD δ_t

" - " TD

13.5 " - "

" - " STDP STDP STDP 10s
STDP " - " " - "

Klopf " " Klopf 1972 1981
Klopf

TD

Niv 2009 Dayan Niv 2008 Gimcher 2011 Ludvig Bellemare Pearson 2011 Sah 2012

Glimcher 200 " " Glimcher Fehr 2013 Dayan Abbott 2001 Sterling Laughlin 015

15.1 Kandel Schwartz Jessell Siegelbaum Hudspeth 2013

15.2 Berridge Kringelbach 2008 Berridge Robinson 1998
" " " " Hare O Doherty Camerer Schultz Rangel 2008
Rangel Camerer Montague 2008 Rangel Hare 2010 Peters Buchel 2010

15.3 Schultz Montague Dayan 1997 Montague Dayan Sejnowski1996 " " RPE TD TD TD Montague Dayan Nowlan Pouget Sejnowski 1993 TD Hebbian Schultz Quartz Dayan Montague Sejnowski 1992 Montague Sejnowski 1994 TD Hebbian Friston Tononi Reeke Sporns Edelman 1994 TD Montague Dayan Person Sejnowski 1995 TD Hammer Menzel Hammer Menzel 1995 Hammer 1997 Montague 1995 Barto 1995 " - " TD Schultz Houk Adams Barto 1995 TD " - " DoyaSejnowski 1998 Doya Sejnowski 1994 TD O Reilly Fank 2006 Reilly Frank hazy Watz 2007 RPE TD TD TD Dayan Niv 2008 " " Glimcher 2011

15.4 Graybiel 000 Tsai Zhang Adamantidis Stuber Bonci de Lecea Deisseroth 2009 Steinberg Keiflin Boivin Witten Deisseroth Janak 203 Claridge Chang Roorda rontou Sjulson Li Hirsh Miesenbock 2009 Fiorillo yun Song 2013 Lammel im Malenka 2014 Saddoris Cacciapaglia Wightmman Carelli 2015 RPE Eshel Tian Bukwich Uchida 2016 VTA Gershman Pesaran - Daw 2009

15.5 Schultz 1998 Schultz 1998 Berns McClure Pagnoni Montague 2001 Breiter Aharon Kahneman Dale Shizgal 2001 Pagnoni Zink Montague Berns 2002 O Doherty Dayan Friston Critchley Dolan 2003 TD

15.6 Barto 1995a TD Schultz

15.7 Takahashi Schoenbaum Niv2008 Niv 2009 Barto 1995 Houk Adams Barto 1995 " - " O Doherty Dayan Schultz Deichmann Friston Dolan 2004 Gershman Moustafa Ludvig 2014

" - " Houk Adams Barto 1995 - Suri Schultz 1998 1999 Brown Bullock Grossberg 1999 Contreras Vidal Schultz 1999 Suri Bargas Arbib 2001 O Reilly Frank 2006 O Reilly Frank azy atz 2007 Joel Niv Ruppin 2002

15.8 Barto 1983 " - " At St At 1t x t Barto 3 Williams 1986 1992

Reynolds Wickens 2002 STDP Markram Lubke Frotscher Sakmann 1997 Levy Steward 1983 Rao Sejnowski 2001 STDP TD 10ms Dayan 2002 Sutton Barto 1981 TD Wickens 1990 Reynolds Wickens 2002 Calabresi Picconi Tozzi Di Filippo 2007 STDP Pawlak Kerr2008 STDP Pawlak Wickens Kirkwood Kerr 2010 Yagishita Hayashi Takagi llis Davies Urakubo Ishii Kasai 2014 STDP 0.3s~2s Izhikevich 2007 STDP Fremaux Sprekeler Gerstner 2010 TDP

15.9 Klopf Klof 1972 1982 " - " Barto Sutton Anerson 1983 Klopf Crow 1968

……
Crow 1968

" " Crow " " " "
Olds Milner 1954

Miller 1981

... B " "
... …… ... Miller 1981 81

Miller " " s " TD Miller Klopf " " STDP

Seung 2003 " " Minsky 1954 Minsky 1954 SNARC Unnikrishnan Venugopal 1994 Harth Tzanakou 1974

Frey Morris 1997 " " Klopf " " O Reilly rnk 2006 Reilly Frank hazy Watz 2007 Wickens Kotter 1995 He Huertas Hong Tie Hell Shouval Kirkwood 2015 Klopf

Barto 1989 Koshland Koshland 1980 Berg 1975 Shimansky 2009 Seung “ ” Montague Dayn Person Sejnowski 1995

15.10 M Tsetlin 1966 Tsetlin 73 1.7 4.8 Tsetlin Narendra Thathachar 1974 Viswanathan Narendra 1974 Lakshmivarahan Narendra 1982 Narendra Wheeler 1983 Narendra 1989 Thathachar Sastry 002 Thathachar Sastry 2011 2.9

Barto Sutton Brouwer 1981 Barto Sutton 1981b Harth Tzanakou 1974 Alopex ASE Barto Anandan 1985 A_{R-P} Barto 1985 1986 Barto Jordan 1987 A P XOR Barto 1985 Williams 1992 Williams 1988 Williams9 A_{R-P} REINFORCE A_{R-P} Barto 1985

STDP Bartlett Baxter 1999 2000 xie Seung 2004 Baras Meir 2007 farries fairhall 2007 Floian 2007 Izhikevich 2007 Pecevsk 2007 Pecevski Maass Legenstein (2008) Legenstein Pecevski Maass (2008) Kolodziejski Porr Worgotter 2009 Urbanczik Senn 2009 Vasilaki Fr ´emaux Urbanczik Senn Gerstner (2009) Now´e Vrancx De Hauwere 2012

15.11 Yin Knowlton 2006 L DMS Valentin Dickinson O Doherty 2007 OFC Padoa Schioppa Assad 2006 OFC Rangel Camerer Montague 2008 Rangel Hare 2010 Pezzulo van der Meer Lansink Pennartz 2014 Daw Shohamy 2008 Bromberg Martin Matsumoto Hong Hikosaka 2010 Doll Simon Daw 2012

15.12 Keifin Janak 2015 TD Nutt Lingford Hughes Erritzoe Stokes 201 Montague Dolan Friston Dayan 2012 Adams Huys Roiser 2015

3.7.3 16

3.7.4 17

3.8

- genindex
- modindex
- search