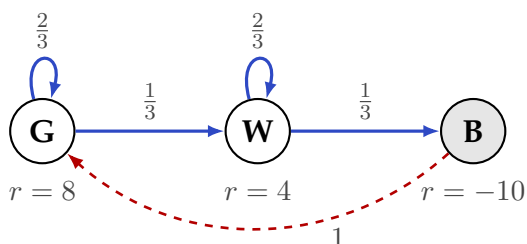


Lecture 9: Temporal-Difference Learning

How to use these notes. Board content in board order: figures, equations, tables, one-line statements. Today's task is *policy evaluation without a model*: a fixed policy π runs a three-state machine, the learner sees the stream of weeks and never P or r , and must estimate the value V^π . The note stands on its own: every symbol is defined where it is first used. Reading: [Sutton and Barto \(2018, Sections 6.1–6.3\)](#); optional, [Szepesvári \(2010, Section 3.1.1\)](#). Throughout the course, the notes draw on the public materials of Stanford CS 234 ([Brunskill, 2026](#)), UIUC CS 443 ([Jiang, 2026](#)), and Cornell CS 4789 ([Sun, 2025](#)); some examples are adapted from them.

The Task: Policy Evaluation Without a Model

A workshop runs one machine. Each week it is good (G), worn (W), or broken (B), and the manager runs it or has it repaired. The manager's rule is fixed: $\pi = \text{"repair when broken"}$. *Policy evaluation* asks how good π is from each state: three numbers, $V^\pi(G)$, $V^\pi(W)$, $V^\pi(B)$. The machine's numbers below are the truth, for checking; the learner never sees them.



The machine under $\pi = \text{"repair when broken"}$.

Run at G and W (solid): a good machine earns 8 and stays good w.p. $\frac{2}{3}$, else wears; a worn one earns 4 and stays worn w.p. $\frac{2}{3}$, else breaks. Repair at B (dashed): costs 10, leaves it good. Gray: hidden from the learner.

Known to the learner	Hidden
$\mathcal{S} = \{G, W, B\}$	P : $3 \cdot 2 \cdot 3 = 18$ numbers
$\mathcal{A} = \{\text{run}, \text{repair}\}$	r : $3 \cdot 2 = 6$ numbers
$\gamma = 0.9$; the policy π	the stream uses 12 of the 24

Under π only three rows of P and three rewards are ever exercised.

Policy evaluation: what is wanted (Sutton and Barto, 2018, Section 3.5). The machine is a Markov decision process $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$: action a in state s pays a reward with mean $r(s, a)$ (on the machine, exactly $r(s, a)$) and moves to s' with probability $P(s' | s, a)$; $\gamma \in [0, 1)$ discounts. A policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ picks the action. Fix π ; write $r^\pi(s) = r(s, \pi(s))$ and $P^\pi(s' | s) = P(s' | s, \pi(s))$.

$$G_t := R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots, \quad V^\pi(s) := \mathbb{E}[G_t | S_t = s] \text{ under } \pi$$

the return from week t , and its mean: the value of π at s

$$V^\pi(s) = r^\pi(s) + \gamma \sum_{s'} P^\pi(s' | s) V^\pi(s') \quad \text{for every } s$$

Bellman's equation: this week's reward, then the value of next week's state

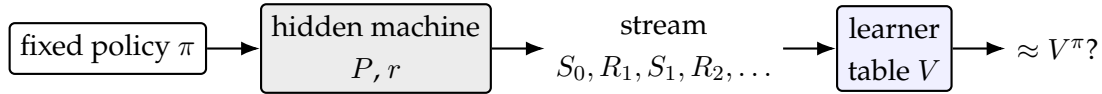
$$(\mathcal{T}^\pi V)(s) := r^\pi(s) + \gamma \sum_{s'} P^\pi(s' | s) V(s'); \quad V^\pi = \mathcal{T}^\pi V^\pi$$

the backup of any table V ; Bellman's equation says V^π is a fixed point

$$\|\mathcal{T}^\pi V - \mathcal{T}^\pi V'\|_\infty \leq \gamma \|V - V'\|_\infty, \quad \|V\|_\infty := \max_s |V(s)|$$

each entry of $\mathcal{T}^\pi V - \mathcal{T}^\pi V'$ is γ times a P^π -average of entries of $V - V'$

So V^π is the only fixed point (Banach, 1922), and the exact iteration $V \leftarrow \mathcal{T}^\pi V$ converges to it from any start, the error shrinking by at least the factor γ per iteration. Rewards on the machine lie in $[-10, 8]$, so $|G_t| \leq 10/(1 - \gamma) = 100$. On the machine: $V(G) = 8 + 0.9(\frac{2}{3}V(G) + \frac{1}{3}V(W))$, $V(W) = 4 + 0.9(\frac{2}{3}V(W) + \frac{1}{3}V(B))$, $V(B) = -10 + 0.9V(G)$; the solution, to two decimals, is $V^\pi = (44.30, 32.41, 29.87)$. With P and r in hand, policy evaluation is these three equations. Today P and r are hidden.



Policy evaluation without a model. π acts (run at G and W, repair at B), the machine answers, the learner watches and updates three numbers. The learner never chooses an action and never sees P or r .

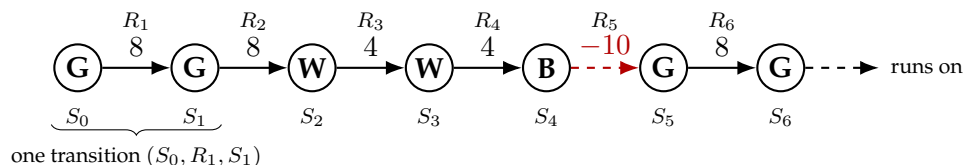
The data: one stream. Fix π . Week after week,

$$S_0, R_1, S_1, R_2, S_2, R_3, S_3, \dots \quad A_t = \pi(S_t) \text{ is fixed and not written}$$

$$S_{t+1} \sim P(\cdot | S_t, A_t), \quad R_{t+1} = r(S_t, A_t) \text{ on the machine}$$

one step of the hidden machine; in general R_{t+1} has mean $r(S_t, A_t)$

- One transition: (S_t, R_{t+1}, S_{t+1}) ; the action $A_t = \pi(S_t)$ is left out.
- No end: the machine is repaired, never retired. No return G_t is ever complete.
- The learner watches; it never acts. Wanted: V^π , from the stream alone.



Six weeks of the stream, used by hand below: the machine starts good, wears, breaks, is repaired, runs again.

The return $G_0 = 8 + 0.9 \cdot 8 + 0.9^2 \cdot 4 + \dots$ has no last term.

Access	The data	Method
Sampling oracle (a simulator)	name any (s, a) , receive one (R, S') : n samples at every pair, the learner's choice	certainty equivalence, next box
Online interaction	one stream; the counts fall where π goes	today
Logged data	another policy's stream	not today

Three ways to hold data, one hidden machine behind each. Today: the second row.

Certainty equivalence: policy evaluation by way of an estimated model (Sutton and Barto, 2018, Section 6.3). A record of transitions (s_i, r_i, s'_i) produced by π , with $n(s) > 0$ of them leaving each s ; V any table:

$$\hat{P}(s' | s) = \frac{\#\{i : s_i = s, s'_i = s'\}}{n(s)}, \quad \hat{r}(s) = \frac{1}{n(s)} \sum_{i: s_i=s} r_i$$

count and average: the estimated chain $\hat{M} = (\hat{P}, \hat{r})$

$$(\mathcal{T}_{\hat{M}}^{\pi} V)(s) := \hat{r}(s) + \gamma \sum_{s'} \hat{P}(s' | s) V(s'); \quad \hat{V}^{\pi} := \text{the solution of } \hat{V}^{\pi} = \mathcal{T}_{\hat{M}}^{\pi} \hat{V}^{\pi}$$

Bellman's equation solved in $\hat{M}$; one solution, since $\mathcal{T}_{\hat{M}}^{\pi}$ is a γ -contraction too

$$(\mathcal{T}_{\hat{M}}^{\pi} V)(s) = \frac{1}{n(s)} \sum_{i: s_i=s} (r_i + \gamma V(s'_i))$$

the identity: $\hat{M}$'s backup at s is the average of the sampled backups $r_i + \gamma V(s'_i)$

With a sampling oracle, n samples at every pair. Cost: a table of $|\mathcal{S}|^2 + |\mathcal{S}|$ numbers and a linear solve.

Policy evaluation from one stream, with no P , no r , and no end in sight: what is moved toward each week, why does it land on V^{π} , and where does it land on a finite record?

Today's targets.

1. One template, $V(s) \leftarrow V(s) + \alpha(\text{target} - V(s))$, and two targets for $V^\pi(s)$: the return G_t (Monte Carlo) and the sampled backup $R_{t+1} + \gamma V(S_{t+1})$. The second is TD(0).
2. By hand, six weeks from $V = 0$ with $\alpha = \frac{1}{2}$: $V = (9.7, 3, -2.3)$ against the truth $(44.30, 32.41, 29.87)$.
3. Derivation: the expected target is $\mathcal{T}^\pi V$; the expected update is $(1 - \alpha)V + \alpha \mathcal{T}^\pi V$, a contraction with modulus $1 - \alpha(1 - \gamma)$ and fixed point V^π . Convergence under Robbins–Monro step sizes, stated.
4. Theorem, two lines: on a fixed record TD(0) lands on certainty equivalence, Monte Carlo on least squares. Nine transitions: $(50, 40, 35)$; AB: $\frac{3}{4}$ against 0.

One Template, Two Targets

The template: move the estimate toward a target (Sutton and Barto, 2018, Section 2.4).

$$V(s) \leftarrow V(s) + \alpha(\text{target} - V(s)) = (1 - \alpha)V(s) + \alpha \cdot \text{target}$$

move a fraction α of the way from the estimate to the target

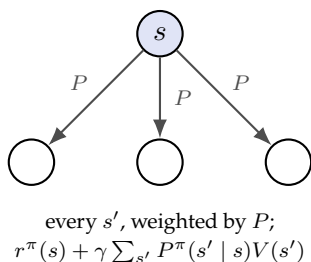
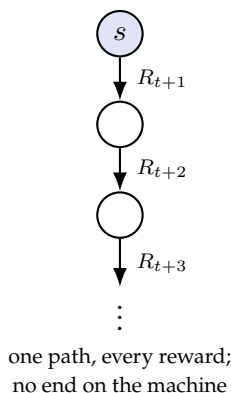
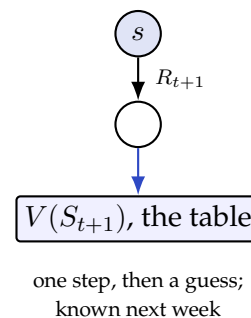
$$\alpha = \frac{1}{n} : \quad \bar{x}_n = \bar{x}_{n-1} + \frac{1}{n}(x_n - \bar{x}_{n-1})$$

x_n the n -th target seen at s , $\bar{x}_n$ the average of the first n : the running average

$$\alpha \text{ constant} : \quad V(s) = \sum_{k=0}^{n-1} \alpha(1 - \alpha)^k x_{n-k} + (1 - \alpha)^n V_0(s)$$

after n targets at s : recent ones weigh more, the start is forgotten at rate $1 - \alpha$

Step sizes for a noisy target: $\sum_n \alpha_n = \infty$ and $\sum_n \alpha_n^2 < \infty$ (Robbins and Monro, 1951), the conditions of the convergence theorem below. $\alpha_n = \frac{1}{n}$ satisfies both; a constant fails the second.

With the model: a backup**Target 1: the return G_t** **Target 2: $R_{t+1} + \gamma V(S_{t+1})$** 

Three ways to say what $V^\pi(s)$ is. The backup averages over P . Monte Carlo follows one path to the end and weights nothing. TD(0) takes one sampled step and reads the rest off its own table: bootstrapping, a guess from a guess.

	Target 1: the return (Monte Carlo)	Target 2: the sampled backup (TD(0))
formula	$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$	$R_{t+1} + \gamma V(S_{t+1})$
known at	when the sum ends: the end of an <i>episode</i> , a run that stops in a terminal state. On the machine never; a cut at $H = 44$ weeks drops a tail of size at most $0.9^{44} \cdot 100 = 0.97$	week $t + 1$
mean of the target, $\mathbb{E}[\cdot S_t = s]$	$V^\pi(s)$: unbiased	$(\mathcal{T}^\pi V)(s)$: biased toward the table, by at most $\gamma \ V - V^\pi\ _\infty$
noise in one target	a whole path of rewards	one reward, one next state
data needed	complete returns: episodes, or cut returns	single transitions (s, r, s')

The same template, two targets: Monte Carlo prediction (Sutton and Barto, 2018, Chapter 5) and temporal-difference learning (Sutton, 1988). Bias against noise: the mean of the target is where the template is pulled, the noise is what it is pulled through. The sampled backup's bias shrinks with the table's error; the derivation below pays it back.

- Monte Carlo on an episodic task: average the returns that followed visits to s . Unbiased: $\mathbb{E}[G_t | S_t = s] = V^\pi(s)$.
- On the machine: wait 44 weeks per return and lose at most 0.97 to the cut. The second target is known a week later, with no cut.
- Certainty equivalence averages the sampled backups at each state, then solves. TD(0) moves $V(s)$ toward each sampled backup as it arrives.

TD(0): Policy Evaluation From the Stream

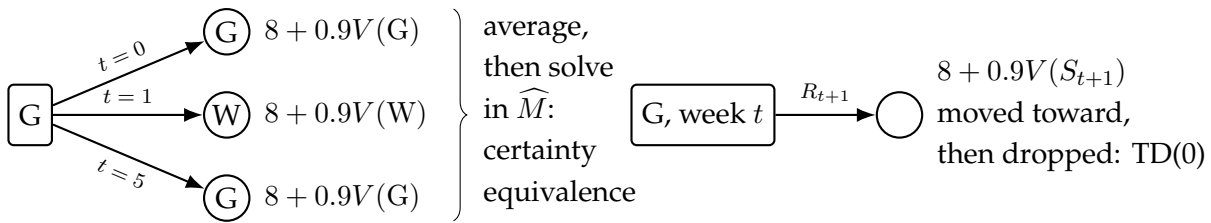
TD(0) (Sutton, 1988), (Sutton and Barto, 2018, Section 6.1), (Szepesvári, 2010, Algorithm 1).

- Input: the stream of π on the machine; a step size $\alpha \in (0, 1]$. Output: a table V , meant to approach V^π .
- Start: $V_0(s) = 0$ for every s . V_t is the table before week t 's update.
- Each week t : observe (S_t, R_{t+1}, S_{t+1}) , then

$$\delta_t = R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t) \quad \text{the TD error}$$

$$V_{t+1}(S_t) = V_t(S_t) + \alpha \delta_t \quad \text{the template at one entry; other entries unchanged}$$

- Nothing stored, nothing waited for. Memory $|S|$; work per week $O(1)$.



The six-week stream leaves G three times, at weeks 0, 1, 5. Certainty equivalence averages the three sampled backups and solves; TD(0) moves toward each one the week it arrives, then forgets it.

Example: TD(0) by hand, six weeks of evaluating π . $\alpha = \frac{1}{2}$, $V = (0, 0, 0)$, the stream of the figure. Fill the table one week at a time; the target and the error first, then the one entry that moves.

week t	(S_t, R_{t+1}, S_{t+1})	target $R_{t+1} + 0.9 V(S_{t+1})$	δ_t	V after: (G, W, B)
0	(G, 8, G)	$8 + 0.9 \cdot 0 = 8$	$8 - 0 = 8$	(4 , 0, 0)
1	(G, 8, W)	$8 + 0.9 \cdot 0 = 8$	$8 - 4 = 4$	(6 , 0, 0)
2	(W, 4, W)	$4 + 0.9 \cdot 0 = 4$	$4 - 0 = 4$	(6, 2 , 0)
3	(W, 4, B)	$4 + 0.9 \cdot 0 = 4$	$4 - 2 = 2$	(6, 3 , 0)
4	(B, -10, G)	$-10 + 0.9 \cdot 6 = -4.6$	$-4.6 - 0 = -4.6$	(6, 3, -2.3)
5	(G, 8, G)	$8 + 0.9 \cdot 6 = 13.4$	$13.4 - 6 = 7.4$	(9.7 , 3, -2.3)
truth V^π , for comparison				(44.30, 32.41, 29.87)

Six updates, one entry each (bold). Week 4 reads a guess, $V(G) = 6$, into the target for B; week 5 reads it again for G. Bootstrapping from zero: the targets are near zero too.

Check: where is the table pulled next? With P , which the learner lacks: $\mathcal{T}^\pi V$ for $V = (9.7, 3, -2.3)$.

G: $8 + 0.9(\frac{2}{3} \cdot 9.7 + \frac{1}{3} \cdot 3) = 14.72$; W: $4 + 0.9(\frac{2}{3} \cdot 3 + \frac{1}{3} \cdot (-2.3)) = 5.11$; B: $-10 + 0.9 \cdot 9.7 = -1.27$.

At W the target is 6.7 (w.p. $\frac{2}{3}$) or 1.93 (w.p. $\frac{1}{3}$), mean 5.11. Every expected target sits above the entry; the truth sits far above both.

- $\alpha = 1$ copies the target: $V(W)$ becomes $4 + 0.9V(W)$ or $4 + 0.9V(B)$, whichever the week brings, and never settles. The step must shrink or stay small.
- Six weeks moved six entries by half an error each. How far the table is pulled per week, and whether it arrives at V^π : the derivation.

Why TD(0) Evaluates π : the Expected Update

Derivation: the expected target is the backup (Sutton and Barto, 2018, Section 6.1). For any table V and any state s :

$$\mathbb{E}[R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s] = r^\pi(s) + \gamma \sum_{s'} P^\pi(s' \mid s) V(s') = (\mathcal{T}^\pi V)(s)$$

R_{t+1} has mean $r^\pi(s)$ and $S_{t+1} \sim P^\pi(\cdot \mid s)$

$$\mathbb{E}[V_{t+1}(s) \mid V_t, S_t = s] = (1 - \alpha)V_t(s) + \alpha(\mathcal{T}^\pi V_t)(s) =: (\mathcal{U}_\alpha V_t)(s)$$

the template in expectation: a damped backup

$$\|\mathcal{U}_\alpha V - \mathcal{U}_\alpha V'\|_\infty \leq (1 - \alpha)\|V - V'\|_\infty + \alpha\gamma\|V - V'\|_\infty = (1 - \alpha(1 - \gamma))\|V - V'\|_\infty$$

the triangle inequality, then the contraction of $\mathcal{T}^\pi$

$$\mathcal{U}_\alpha V = V \iff \mathcal{T}^\pi V = V \iff V = V^\pi$$

a contraction has one fixed point, and V^π is one $\square$

TD(0) is the exact iteration $V \leftarrow \mathcal{T}^\pi V$ run at speed α , at one entry per week, with one sample in place of the expectation.

α	1 (the exact iteration)	0.5	0.3	0.1
modulus $1 - \alpha(1 - \gamma)$ at $\gamma = 0.9$	0.90	0.95	0.97	0.99
expected updates to shrink the bound by γ	1	2.1	3.5	10.5

The third row is $\ln \gamma / \ln(1 - \alpha(1 - \gamma))$: at $\alpha = 0.1$ the bound shrinks by 0.99 per expected update of an entry, and $0.99^{10.5} \approx 0.90$, one exact iteration's worth. The exact iteration $V \leftarrow \mathcal{T}^\pi V$ from $V = 0$ needs 24 iterations for a max-norm error under 3 and 35 for under 1.

Why nothing but V^π (informal) (Szepesvári, 2010, Section 3.1.1). Suppose the table settles at some $\hat{V}$ while every state keeps being visited.

$$\mathbb{E}[\delta_t \mid S_t = s, V_t = \hat{V}] = (\mathcal{T}^\pi \hat{V})(s) - \hat{V}(s) \text{ must be 0 at every } s$$

a nonzero expected increment, repeated forever with $\sum \alpha = \infty$, would move the entry

$$\hat{V} = \mathcal{T}^\pi \hat{V} \implies \hat{V} = V^\pi \quad \text{only } V^\pi \text{ solves Bellman's equation}$$

Theorem (stated): tabular TD(0) converges (Sutton, 1988; Dayan, 1992; Jaakkola et al., 1994; Tsitsiklis and Van Roy, 1997). If every state is visited infinitely often and the step sizes used at each state satisfy $\sum \alpha = \infty$, $\sum \alpha^2 < \infty$, then $V_t \rightarrow V^\pi$ with probability 1. With a constant α the table does not converge: it keeps fluctuating around V^π , the less the smaller α (Sutton and Barto, 2018, Section 6.2), (Szepesvári, 2010, Section 3.1.1). Proof by stochastic approximation (Bertsekas and Tsitsiklis, 1996); not given.

On a Fixed Record: Batch TD(0) Is Certainty Equivalence

A finite record $D = \{(s_i, r_i, s'_i)\}_{i=1}^N$ produced by π , with $n(s) > 0$ transitions leaving each s : two ways to evaluate π from it. *Batch TD(0)* (Sutton and Barto, 2018, Section 6.3): hold V , apply all of D 's increments at once, repeat:

$$V(s) \leftarrow V(s) + \alpha \sum_{i: s_i=s} (r_i + \gamma V(s'_i) - V(s)) \quad \text{for every } s.$$

Replay, TD(0) on transitions drawn from D at random, is this update in expectation, scaled by $1/N$.

Theorem: batch TD(0) converges to the certainty-equivalence value (Sutton and Barto, 2018, Section 6.3). Let $\hat{M} = (\hat{P}, \hat{r})$ be D 's counts and averages and $\hat{V}^\pi$ the solution of Bellman's equation in $\hat{M}$, as in the certainty-equivalence box. A table $\hat{V}$ is a fixed point of batch TD(0) if and only if $\hat{V} = \hat{V}^\pi$; and for $\alpha \leq 1/\max_s n(s)$ the batch update reaches

it from any start.

$$\text{fixed point: } \sum_{i: s_i=s} (r_i + \gamma \hat{V}(s'_i) - \hat{V}(s)) = 0 \quad \text{for every } s$$

the summed increment at s is zero

$$\text{divide by } n(s) : \quad \hat{V}(s) = \hat{r}(s) + \gamma \sum_{s'} \hat{P}(s' | s) \hat{V}(s')$$

the identity: the average of the sampled backups is $\hat{M}$'s backup

$$\hat{V} = \mathcal{T}_{\hat{M}}^{\pi} \hat{V} \iff \hat{V} = \hat{V}^{\pi} \quad \text{Bellman's equation in } \hat{M} \text{ has one solution} \quad \square$$

$$\text{the update reads } V(s) \leftarrow (1 - \alpha n(s)) V(s) + \alpha n(s) (\mathcal{T}_{\hat{M}}^{\pi} V)(s)$$

by the identity; for $\alpha n(s) \leq 1$ a contraction with modulus $1 - \alpha \min_s n(s) (1 - \gamma)$

Batch Monte Carlo on a record of complete returns: the fixed point is the average of the returns after visits to s , which is $\text{argmin}_V \sum_t (G_t - V(S_t))^2$, least squares on the observed returns.

Example: a record of nine transitions, replayed. The record D : three transitions (s_i, r_i, s'_i) out of each state.

out of G	(G, 8, G), (G, 8, W), (G, 8, G)
out of W	(W, 4, W), (W, 4, W), (W, 4, W)
out of B	(B, -10, G), (B, -10, G), (B, -10, G)

Counts and averages: $\hat{P}(G | G) = \frac{2}{3}$, $\hat{P}(W | G) = \frac{1}{3}$, $\hat{P}(W | W) = 1$, $\hat{P}(G | B) = 1$; $\hat{r} = (8, 4, -10)$. Batch TD(0) on D converges to its fixed point: which table?

W: $\hat{V}(W) = 4 + 0.9 \hat{V}(W)$, so 40. G: $\hat{V}(G) = 8 + 0.9(\frac{2}{3} \hat{V}(G) + \frac{1}{3} \cdot 40)$, so 50. B: $-10 + 0.9 \cdot 50 = 35$.

(50, 40, 35): the certainty-equivalence value of D , which never saw W break; reached by batch TD(0) with $\hat{M}$ never written down.

Example: you are the predictor (Sutton and Barto, 2018, Example 6.4). Two states, A and B; $\gamma = 1$; every episode ends right after B. Eight episodes were seen:

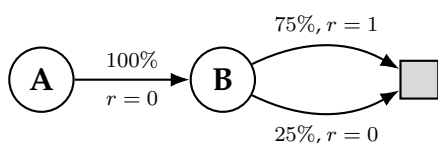
episode	what was seen	return
1	start in A, reward 0; move to B, reward 0; end	from A: $0 + 0 = 0$; from B: 0
2–7 (six times)	start in B, reward 1; end	from B: 1
8	start in B, reward 0; end	from B: 0

From the eight episodes alone: $V(B)$ and $V(A)$?

B: eight visits, returns 1, 1, 1, 1, 1, 1, 0, 0, average $\frac{3}{4}$; both targets agree, as B is always followed by the end.

A, TD: $0 + V(B) = \frac{3}{4}$. A, Monte Carlo: the only return seen after A is 0.

	$V(A)$	$V(B)$	squared error on the eight returns
batch TD(0) = certainty equivalence	$\frac{3}{4}$	$\frac{3}{4}$	$6 \cdot \frac{1}{16} + 2 \cdot \frac{9}{16} + \frac{9}{16} = 2.06$
batch Monte Carlo = least squares	0	$\frac{3}{4}$	$6 \cdot \frac{1}{16} + 2 \cdot \frac{9}{16} = 1.50$



$\widehat{M}$ counted from the eight episodes; the square is the terminal state. TD's $\frac{3}{4}$ is this chain's value of A. Monte Carlo's 0 minimizes the squared error on the record (Sutton and Barto, 2018, Section 6.3).

- TD(0) needs single transitions; Monte Carlo needs complete returns, so an end or a cut. The machine has no end.
- Certainty equivalence stores $|\mathcal{S}|^2 + |\mathcal{S}|$ numbers and solves a system; one pass of TD(0) stores $|\mathcal{S}|$ and does $O(1)$ a week; batch TD(0) stores the record and lands on certainty equivalence's answer, $\widehat{M}$ never written down.
- One pass over a stream stops short of the fixed point; batch TD(0) on the stored stream reaches it. By computer, next.

By Computer: Evaluating π on the Machine

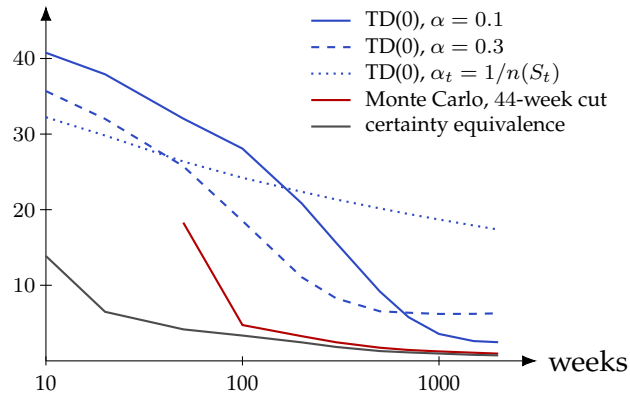
The experiment.

1. Start the machine at G and run π for 2,000 weeks: one *stream* of 2,000 transitions.
2. Feed the stream to each method below. After t weeks each method holds a table V_t ; its error is $\|V_t - V^\pi\|_\infty = \max_s |V_t(s) - V^\pi(s)|$.
3. Repeat with 200 fresh streams; report the average error at $t = 50, 100, 200, 500, 1,000, 2,000$.

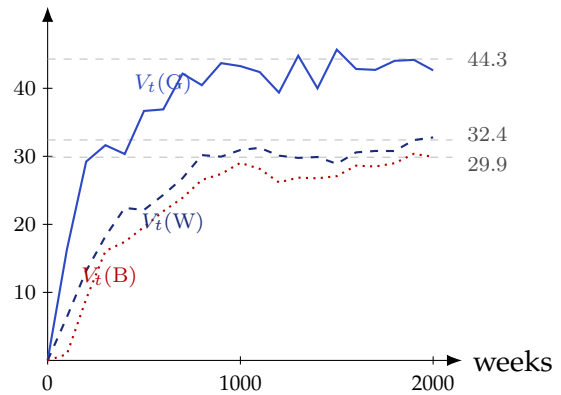
Method	What V_t is
TD(0), $\alpha = 0.1$ or 0.3	the box above, one pass over the first t weeks, $V_0 = 0$
TD(0), $\alpha_t = 1/n(S_t)$	the same with $n(S_t)$ the number of visits to S_t so far: at each state, the running average of the sampled backups
Monte Carlo, 44-week cut	$V_t(s)$ is the average of the cut returns $\sum_{k=0}^{43} \gamma^k R_{u+k+1}$ over the visits u to s with $u + 44 \leq t$; 0 before the first one
certainty equivalence	$\hat{V}^\pi$ of the record of the first t transitions; every state has been left by week 50 in all 200 streams
batch TD(0) on the first t transitions	its fixed point: the row above, by the theorem

mean error over 200 streams at week $t =$	50	100	200	500	1,000	2,000
TD(0), $\alpha = 0.1$	32.05	28.08	20.85	9.18	3.56	2.47
TD(0), $\alpha = 0.3$	25.75	18.52	11.08	6.56	6.21	6.29
TD(0), $\alpha_t = 1/n(S_t)$	26.37	24.22	22.37	20.17	18.70	17.37
Monte Carlo, 44-week cut	18.29	4.74	3.27	1.74	1.25	0.96
certainty equivalence	4.17	3.35	2.45	1.30	0.97	0.73
batch TD(0) on the first t transitions	= certainty equivalence, by the theorem					

Error against weeks. One pass of TD(0) from zero is slow here; a constant step leaves a floor; the running-average step keeps the weight of the early, near-zero targets; batch TD(0) on the stored stream matches certainty equivalence exactly.



The table as curves: average error against weeks, the week axis logarithmic. The first cut return is complete at week 44.



One stream, TD(0) with $\alpha = 0.1$: the three entries $V_t(G)$, $V_t(W)$, $V_t(B)$ against weeks; dashed, V^π .

one stream, week $t =$	100	500	2,000
TD(0), $\alpha = 0.1$, one pass: V_t	(16.30, 6.38, 0.92)	(36.67, 22.12, 19.61)	(42.64, 32.78, 29.92)
batch TD(0), first t transitions: $\hat{V}^\pi$	(36.83, 27.23, 23.14)	(43.31, 31.51, 28.98)	(44.71, 32.55, 30.24)
its error $\ \hat{V}^\pi - V^\pi\ _\infty$	7.48	0.99	0.41

The same 100 transitions, two uses: one pass of TD(0) stands at (16.30, 6.38, 0.92); batch TD(0) on them, which is their certainty-equivalence value, stands at (36.83, 27.23, 23.14).

- One pass from zero, $\alpha = 0.1$: error 3.56 after 1,000 weeks. Each week moves one entry by a tenth of its error; the exact iteration moves all three fully, 24 times for an error under 3.
- A constant α leaves a floor that shrinks with α : 2.47 at 0.1 against 6.29 at 0.3. The running average weights the early, near-zero targets as much as the late ones: 17.37 after 2,000 weeks.
- Here Monte Carlo and certainty equivalence are ahead of one pass of TD(0): three states, a start at zero.
- Which converges faster is open in general (Sutton and Barto, 2018, Section 6.2); on the Random Walk of Example 6.2 there, TD(0) is ahead.
- Batch TD(0) on the stored stream: the theorem as numbers. Certainty equivalence's value, $\hat{M}$ never written down.

Summary

Policy evaluation without a model, on one board

- Task: evaluate a fixed π , that is, find $V^\pi = \mathcal{T}^\pi V^\pi = r^\pi + \gamma P^\pi V^\pi$, from one stream $S_0, R_1, S_1, R_2, \dots$ with no P , no r , no end.
- Template: $V(s) \leftarrow V(s) + \alpha(\text{target} - V(s))$; $\alpha = 1/n$ averages. Targets: G_t (Monte Carlo; on the machine, cut at 44, late by 44 weeks) and $R_{t+1} + \gamma V(S_{t+1})$ (TD(0)).
- TD(0): $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$, $V(S_t) \leftarrow V(S_t) + \alpha \delta_t$. By hand, $\alpha = \frac{1}{2}$: errors 8, 4, 4, 2, -4.6, 7.4; $V = (9.7, 3, -2.3)$; $\mathcal{T}^\pi V = (14.72, 5.11, -1.27)$; truth (44.30, 32.41, 29.87).
- Derivation: $\mathbb{E}[\text{target} \mid s] = (\mathcal{T}^\pi V)(s)$; expected update $\mathcal{U}_\alpha V = (1 - \alpha)V + \alpha \mathcal{T}^\pi V$; modulus $1 - \alpha(1 - \gamma)$ (0.99 at $\alpha = 0.1$); fixed point V^π . Theorem (stated): Robbins–Monro steps, every state visited forever, $V_t \rightarrow V^\pi$.
- Theorem (proved): batch TD(0)'s fixed point solves $\hat{V} = \hat{r} + \gamma \hat{P} \hat{V}$, the certainty-equivalence value; batch Monte Carlo is least squares on returns. Nine transitions: (50, 40, 35). AB: $\frac{3}{4}$ against 0.
- By computer, average error over 200 streams from week 50 to 2,000: TD(0) at $\alpha = 0.1$ from 32.05 to 2.47; floor 6.29 at $\alpha = 0.3$; 17.37 with $1/n$; Monte Carlo 0.96; certainty equivalence 0.73; batch TD(0) on the stream = certainty equivalence.

References

- Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3:133–181, 1922.
- Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- Emma Brunskill. CS 234: Reinforcement learning. Stanford University course materials, 2026. URL <https://web.stanford.edu/class/cs234/>. Winter 2026 offering.
- Peter Dayan. The convergence of TD(λ) for general λ . *Machine Learning*, 8(3–4):341–362, 1992.
- Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.
- Nan Jiang. CS 443: Reinforcement learning. University of Illinois Urbana-Champaign course materials, 2026. URL <https://nanjiang.cs.illinois.edu/cs443/>. Spring 2026 offering.
- Herbert Robbins and Sutton Monroe. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- Wen Sun. CS 4/5789: Introduction to reinforcement learning. Cornell University course materials, 2025. URL https://wensun.github.io/CS4789_spring_2025.html. Spring 2025 offering.
- Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. URL <http://incompleteideas.net/book/the-book-2nd.html>.
- Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2010. URL <https://sites.ualberta.ca/~szepesva/papers/RLAlgsInMDPs.pdf>.
- John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.