Reinforcement Learning 2

Kintan Saha

January 8, 2026

Outline

Markov Decision Process (MDP)

- States: $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$
- Actions: $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$
- State transition probabilities: $P(s'|s, a)$
- Rewards: $r(s, a)$
- Reward discounting factor: $0 \leq \gamma < 1$

Reinforcement Learning (II)

Value function:

$$V^{\mu}(s) = \mathbb{E}^{\mu}\left[\sum_{t=0}^{\infty} \gamma^t R_t \;\middle|\; S_0 = s\right]$$

Bellman evaluation operator (policy μ):

$$(T_{\mu} V)(s) \;:=\; \sum_{a \in \mathcal{A}} \mu(a \mid s)\Big(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V(s')\Big).$$

Bellman optimality operator:

$$(TV)(s) \;:=\; \max_{a \in \mathcal{A}} \Big(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V(s')\Big).$$

Operators and Contractions

Contraction in $\|\cdot\|_\infty$: for all bounded $V, W : \mathcal{S} \to \mathbb{R}$,

$$\|T_\mu V - T_\mu W\|_\infty \le \gamma \|V - W\|_\infty, \qquad \|TV - TW\|_\infty \le \gamma \|V - W\|_\infty.$$

Iterates converge to fixed points: for any V_0,

$$T_\mu^k V_0 \to V^\mu, \qquad T^k V_0 \to V^* \quad (k \to \infty),$$

where V^μ is the unique fixed point of T_μ and V^* is the unique fixed point of T.

$$\|x\|_\infty \to 0 \quad \Rightarrow \quad \max_{i \in [d]} |x_i| \to 0$$

$$\Rightarrow \quad |x_i| \to 0 \quad \forall\, i \in [d]$$

$$\|x\|_2 \to 0 \quad \Rightarrow \quad \sqrt{\sum x_i^2} \to 0$$

Outline

Policy Iteration

Input: an initial stationary policy μ_0.
For $k = 0, 1, 2, \ldots$ **iterate:**

1. **Policy evaluation:** compute V^{μ_k} as the unique solution of

$$V^{\mu_k} = T_{\mu_k} V^{\mu_k}.$$

Approximate policy evaluation

2. **Policy improvement:** choose a greedy policy μ_{k+1} w.r.t. V^{μ_k}:

$$\mu_{k+1}(s) \in \arg\max_{a \in \mathcal{A}} \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^{\mu_k}(s') \right).$$

Stop when $\mu_{k+1} = \mu_k$.

Policy Iteration: Guarantees

Assume finite $\mathcal{S}, \mathcal{A}$, bounded rewards, and $0 \leq \gamma < 1$.

- **Monotonic improvement:** $V^{\mu_{k+1}}(s) \geq V^{\mu_k}(s)$ for all $s \in \mathcal{S}$.
- **Finite termination:** PI terminates in finitely many iterations (cannot cycle).
- **Optimality at termination:** if $\mu_{k+1} = \mu_k$, then $V^{\mu_k} = V^*$ and μ_k is optimal.

Modified PI

ADP

Proof I: Policy Improvement Lemma

Lemma (Policy Improvement). Let μ' be greedy w.r.t. V^μ. Then $V^{\mu'} \geq V^\mu$ componentwise.

Proof. Greediness implies for every $s \in \mathcal{S}$,

$$(T_{\mu'} V^\mu)(s) = (TV^\mu)(s) \;\geq\; (T_\mu V^\mu)(s) = V^\mu(s).$$

Define the iterates

$$W_0 := V^\mu, \qquad W_{n+1} := T_{\mu'} W_n \quad (n \geq 0).$$

Since $T_{\mu'}$ is monotone (order-preserving), $W_{n+1} \geq W_n$ for all n.
Because $T_{\mu'}$ is a γ-contraction in $\|\cdot\|_\infty$, we have

$$W_n = T_{\mu'}^n V^\mu \xrightarrow[n\to\infty]{} V^{\mu'},$$

the unique fixed point of $T_{\mu'}$. Taking limits in $W_n \geq W_0$ yields $V^{\mu'} \geq V^\mu$. $\square$

Proof II: If PI Stops, It Is Optimal

Step 1 (Stopping condition $\Rightarrow$ Bellman optimality). If policy iteration stops at μ (i.e., the greedy improvement returns the same policy), then μ is greedy w.r.t. V^μ, hence for every s,

$$(TV^\mu)(s) = (T_\mu V^\mu)(s).$$

But V^μ satisfies the evaluation equation $V^\mu = T_\mu V^\mu$, so

$$TV^\mu = V^\mu.$$

Step 2 (Uniqueness of the fixed point). Since T is a γ-contraction, it has a unique fixed point V^*. Therefore $TV^\mu = V^\mu$ implies

$$V^\mu = V^*,$$

and thus μ is an optimal policy. $\square$

Proof III: Finite Termination (No Cycles)

Key fact: each non-terminating PI step strictly improves the value somewhere.
If $\mu_{k+1} \neq \mu_k$, the improvement lemma gives

$$V^{\mu_{k+1}} \geq V^{\mu_k}.$$

If μ_k is not optimal, then $TV^{\mu_k} \neq V^{\mu_k}$, so $\exists\, s$ such that

$$(TV^{\mu_k})(s) > V^{\mu_k}(s).$$

For a greedy μ_{k+1},

$$(T_{\mu_{k+1}} V^{\mu_k})(s) = (TV^{\mu_k})(s) > V^{\mu_k}(s),$$

and applying $T^n_{\mu_{k+1}}$ and taking $n \to \infty$ yields

$$V^{\mu_{k+1}}(s) > V^{\mu_k}(s) \quad \text{for some } s.$$

Conclusion (finite termination). PI cannot revisit a previous policy (values would have to strictly increase yet repeat). Since the number of stationary deterministic policies is finite ($|\mathcal{A}|^{|\mathcal{S}|}$), PI must terminate.

Outline

Model-based algorithm

Recall the operators:

$$(T_\mu V)(s) = \sum_{a \in \mathcal{A}} \mu(a \mid s) \Big(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V(s') \Big),$$

$$(TV)(s) = \max_{a \in \mathcal{A}} \Big(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V(s') \Big).$$

Model dependence. Computing $T_\mu V$ or TV requires knowing:

- the reward function $r(s,a)$, and
- the transition kernel $P(\cdot \mid s, a)$,

which are typically unknown and only accessible via samples. This motivates **model-free** algorithms.

2 Fundamental Questions of RL

Policy Evaluation

Given a policy μ, estimate its value function

$$V^{\mu}(s) = \mathbb{E}^{\mu}\left[\sum_{t=0}^{\infty} \gamma^t R_t \;\middle|\; S_0 = s\right]$$

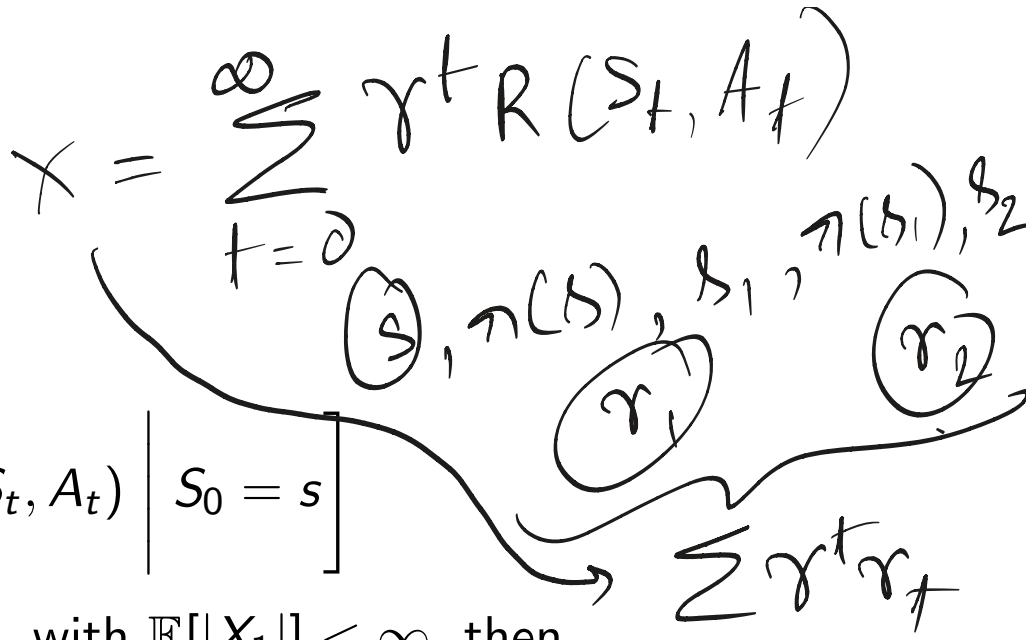
to quantify how good μ is.

Policy Improvement / Control

Given a policy (or value estimates), produce an improved policy, ultimately targeting an optimal policy.

In what follows, we focus on **policy evaluation**.

Policy Evaluation



We will focus on policy evaluation for now.

Value Function

$$V^\pi(s) := \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s \right]$$

Law of Large Numbers. If $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}[|X_1|] < \infty$, then

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[X_1].$$

Motivation. The return $\sum_{t=0}^{\infty} \gamma^t R_t$ is random and its expectation is $V^\mu(s)$. Since P and r are unknown, we estimate expectations using samples from interaction.

Policy Evaluation - Monte Carlo Methods

Monte Carlo idea: for a fixed start state s, generate N independent rollouts under μ and average returns:

$$\widehat{V}_N^{\mu}(s) = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=0}^{\infty} \gamma^t R_t^{(i)} \right) \xrightarrow[N \to \infty]{a.s.} V^{\mu}(s).$$

Issue: MC needs (approximately) the full discounted sum, so it is most natural in episodic / finite-horizon settings.

Motivation for TD: learn incrementally from one-step samples (S_n, A_n, R_n, S_{n+1}).

Temporal Difference Learning: Optimization View

Goal: learn the value function V^μ from samples (S_n, A_n, R_n, S_{n+1}) under a fixed policy μ.

We begin with an **idealized optimization problem**:

$$f(x) \;:=\; \frac{1}{2}\sum_{s\in\mathcal{S}} d(s)\big(V^\mu(s) - x(s)\big)^2 \;=\; \frac{1}{2}\|V^\mu - x\|_D^2,$$

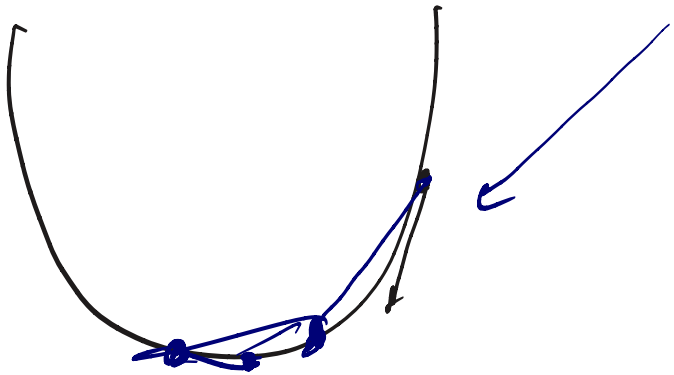
$d(s) \in [0,1]$

where d is a weighting distribution on $\mathcal{S}$ (e.g. the stationary distribution under μ), and $x : \mathcal{S} \to \mathbb{R}$ is our estimate.

Observation: the unique minimizer of $f(x)$ is $x = V^\mu$.

$\min_{x \in \mathbb{R}} f(x)$

$x_{k+1} = x_k - \alpha f'(x_k)$ → GD update



$x_{k+1} = x_k$

Gradient Descent on the Ideal Objective

For

$$f(x) = \frac{1}{2}\sum_{s\in\mathcal{S}} d(s)\big(V^{\mu}(s) - x(s)\big)^2,$$

the gradient (tabular form) is

$$-\nabla f(x)(s) = d(s)\big(x(s) - V^{\mu}(s)\big).$$

A formal gradient descent step is

$$x_{n+1} = x_n - \alpha_n \nabla f(x_n) = x_n + \alpha_n \sum_{s\in\mathcal{S}} d(s)\big(V^{\mu}(s) - x_n(s)\big)e_s$$

Issue: this update is not implementable since V^{μ} is unknown.

Using the Bellman Equation

Recall the Bellman evaluation equation:

$$V^{\mu}(s) = \sum_{a \in \mathcal{A}} \mu(a \mid s) \Big(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^{\mu}(s') \Big).$$

Substituting into the gradient update yields

$$x_{n+1} = x_n + \alpha_n \sum_{s,a,s'} d(s) \mu(a \mid s) P(s' \mid s, a) \Big(r(s,a) + \gamma V^{\mu}(s') - x_n(s) \Big) e_s.$$

Still not implementable: $V^{\mu}(s')$ is an infinite-horizon return.

$$d(s)\left[V^{\mu}(s) - x_n(s)\right]$$

$$V^{\mu}(s) = \sum_a \mu(a|s)\left[r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^{\mu}(s')\right]$$

$$\frac{}{\sum_a \mu(a|s)\, x_n(s) \underbrace{\sum_{s'} P(s'|s,a)}_{1}\; x_n(s)\;\Big\}}$$

$$= \sum_a \mu(a|s)\, x_n(s) = x_n(s) \underbrace{\sum_a \mu(a|s)}_{1} = x_n(s)$$

Bootstrapping and Loss of a True Objective

To obtain a computable recursion, we replace $V^{\mu}(s')$ with it's current estimate $x_n(s')$:

$$V^{\mu}(s') \longrightarrow x_n(s')$$

This gives

$$x_{n+1} = x_n + \alpha_n \sum_{s,a,s'} d(s)\mu(a \mid s)P(s' \mid s,a)\Big(r(s,a) + \gamma x_n(s') - x_n(s)\Big)e_s.$$

Key conceptual point:

- The update is *not* the gradient of the original objective $f(x)$.
- Temporal-Difference learning is therefore **not** true gradient descent.
- It is a noisy gradient descent (?) scheme
- We still don't know $P(s' \mid s,a)$ so how are we even supposed to implement this?

Replacing Expectations with Samples: TD(0)

The quantity

$$d(s)\mu(a \mid s)P(s' \mid s, a)$$

defines the joint distribution of (S, A, S') under policy μ.

Assume we can sample i.i.d. tuples

$$(S_n, A_n, R_n, S_{n+1}) \sim \text{stationary one-step distribution under } \mu.$$

Replacing expectations by a single sample yields the TD(0) update:

$$\boxed{x_{n+1}(S_n) = x_n(S_n) + \alpha_n\Big(R_n + \gamma x_n(S_{n+1}) - x_n(S_n)\Big),}$$

with $x_{n+1}(s) = x_n(s)$ for all $s \neq S_n$.

TD Learning - Remarks

- **Not classical GD:** expectations are replaced by random samples, so the update is a stochastic recursion.
- **Interpretation:** TD(0) fits the stochastic approximation / stochastic gradient paradigm.
- **Sampling mechanism (i.i.d. version):** each update uses an independent tuple $S_n \sim d_\mu$ (stationary distribution), $A_n \sim \mu(\cdot \mid S_n)$, $S_{n+1} \sim P(\cdot \mid S_n, A_n)$, and $R_n = r(S_n, A_n)$

Outline

Stochastic Approximation

TD(0):

$$x_{n+1}(S_n) = x_n(S_n) + \alpha_n\Big(R_n + \gamma x_n(S_{n+1}) - x_n(S_n)\Big).$$

Stochastic Approximation(in $\mathbb{R}^d$):

$$\boxed{x_{n+1} = x_n + a(n)\big(h(x_n) + M_{n+1}\big)}$$

where:

- $a(n) > 0$ is the stepsize,
- $h : \mathbb{R}^d \to \mathbb{R}^d$ is the mean drift (typically $h(x) = \mathbb{E}[H(x, \xi)]$),
- M_{n+1} is the noise term with $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = 0$.

What do we want?

Goal of stochastic approximation: find conditions on

- stepsizes $\{a(n)\}$,
- regularity of $h(\cdot)$,
- noise sequence $\{M_n\}$,
- stability of iterates $\{x_n\}$,

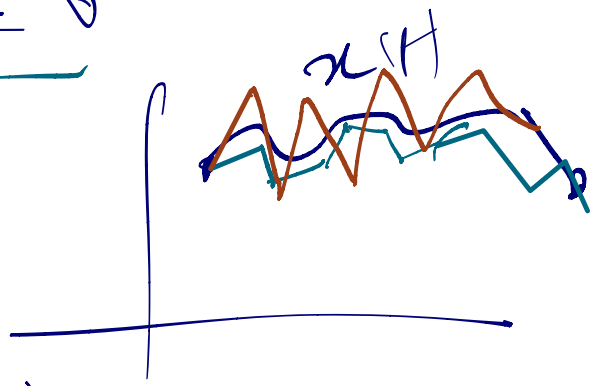
so that

$$x_n \to x^\star \quad \text{a.s. (or to a limit set).}$$

For TD policy evaluation : $x^\star = V^\mu$.

$\dot{x} = -\nabla f(x)$ $\qquad \dot{x} = 0 \Rightarrow \nabla f(x) = 0$

$\frac{dx}{dt} = \lim_{h \to 0} \frac{x_{t+h} - x_t}{h} = -\nabla f(x_t)$



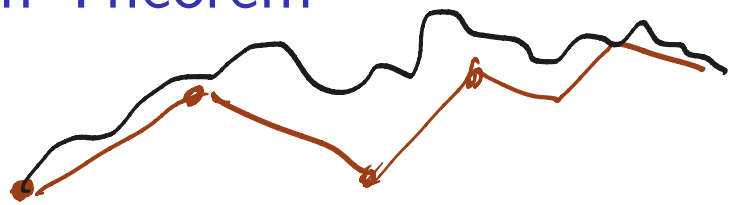
$$x(t+h) = x(t) - h \nabla f(x_t) + M$$

x_{t+1} $\qquad x_t$

$+100$ w.p. $1/2$

-100 w.p. $1/2$

Borkar–Meyn Theorem



Consider

$$x_{n+1} = x_n + a(n)\big(h(x_n) + M_{n+1}\big), \qquad x_n \in \mathbb{R}^d,$$

with filtration $\{\mathcal{F}_n\}$ such that $\mathcal{F}_n = \sigma(x_0, M_1, \ldots, M_n)$.

Conclusion (ODE method): Under some assumptions, the interpolated trajectory tracks the ODE

$$\dot{x}(t) = h(x(t)),$$

and x_n converges a.s. to the internally chain transitive invariant set of the ODE. If the ODE has a unique globally asymptotically stable equilibrium $x^\star$, then $x_n \to x^\star$ a.s.

Borkar–Meyn Theorem (Assumptions I)

(A1) Stepsizes (Robbins–Monro):

$$a(n) > 0, \qquad \sum_{n=0}^{\infty} a(n) = \infty, \qquad \sum_{n=0}^{\infty} a(n)^2 < \infty.$$

(A2) Drift regularity: The function $h : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\|h(x) - h(y)\| \le L\|x - y\| \quad \forall\, x, y \in \mathbb{R}^d$$

for some constant $L > 0$.

(A3) Martingale-difference noise with bounded conditional second moment:

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = 0, \qquad \mathbb{E}\big[\|M_{n+1}\|^2 \mid \mathcal{F}_n\big] \le C\big(1 + \|x_n\|^2\big) \quad \text{a.s.}$$

$\{M_n\}$ → martingale

① $\mathbb{E}[M_{n+1} \mid M_0, M_1, \dots, M_n] = M_n$

$\mathbb{E} M_{n+1} = \mathbb{E} M_n$

martingale difference → $X_n = M_n - M_{n-1}$

$\mathbb{E}[X_n \mid M_0, \dots M_{n-1}]$

$= 0$

Borkar–Meyn Theorem (Assumptions II)

(A4) Stability (assumed here):

$$\sup_{n \geq 0} \|x_n\| < \infty \quad \text{a.s.}$$

(A5 gives us stability)

(A5) Scaled drift / ODE at infinity: Assume the limit exists uniformly on compacts:

$$h_\infty(x) := \lim_{c \to \infty} \frac{h(cx)}{c}.$$

Assume the limiting ODE

$$\dot{x}(t) = h_\infty(x(t))$$

has the origin as a globally asymptotically stable equilibrium.

$$\dot{x} = f(u)$$

Outline

Remarks and Outlook

$$x_{n+1}(s) = x_n(s) + a(n)\left[r_n + \gamma x_n(s_{n+1}) - x_n(s_n)\right]$$

$$x_{n+1} \longrightarrow v_\mu$$

- We will formulate the TD learning update rule as a *stochastic approximation* recursion.
- We will then verify the required assumptions and use stochastic approximation theory to establish convergence of the iterates to the desired limit.

$$x_{n+1} = x_n + a(n)\left[h(x_n) + M_{n+1}\right]$$

$h(\cdot)$ M_{n+1}

Thank You

Thank you for your attention!

Questions, comments, or clarifications?