

Convergence of TD Methods

tl;dr: Noisy, bootstrapped updates work

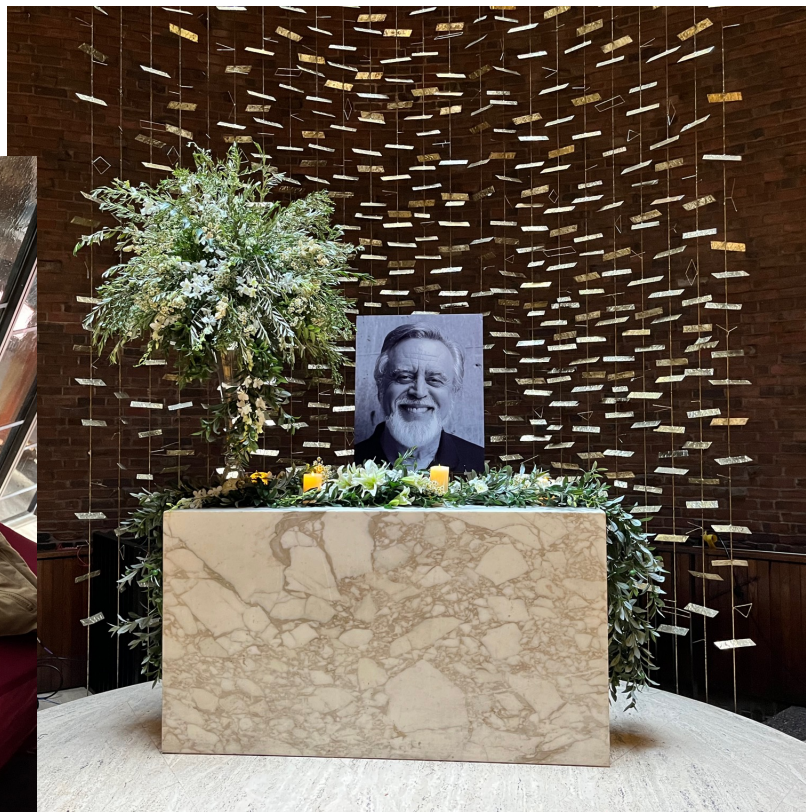
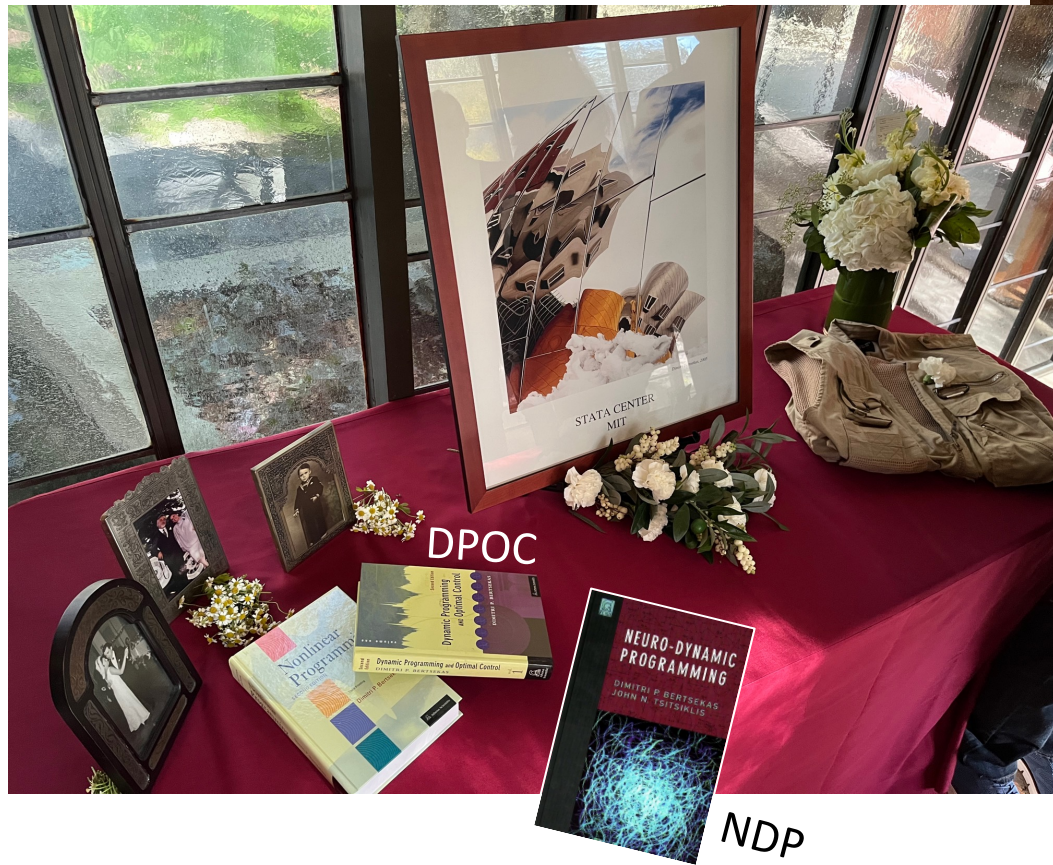
Cathy Wu

6.7920 Reinforcement Learning: Foundations and Methods

Readings

1. Neuro-dynamic Programming (NDP). §4.1-4.3, §6.1-6.2. Skim Ch 3-5 as needed.
2. DPOC2 §6.3

Dimitri P. Bertsekas (1942-2026)



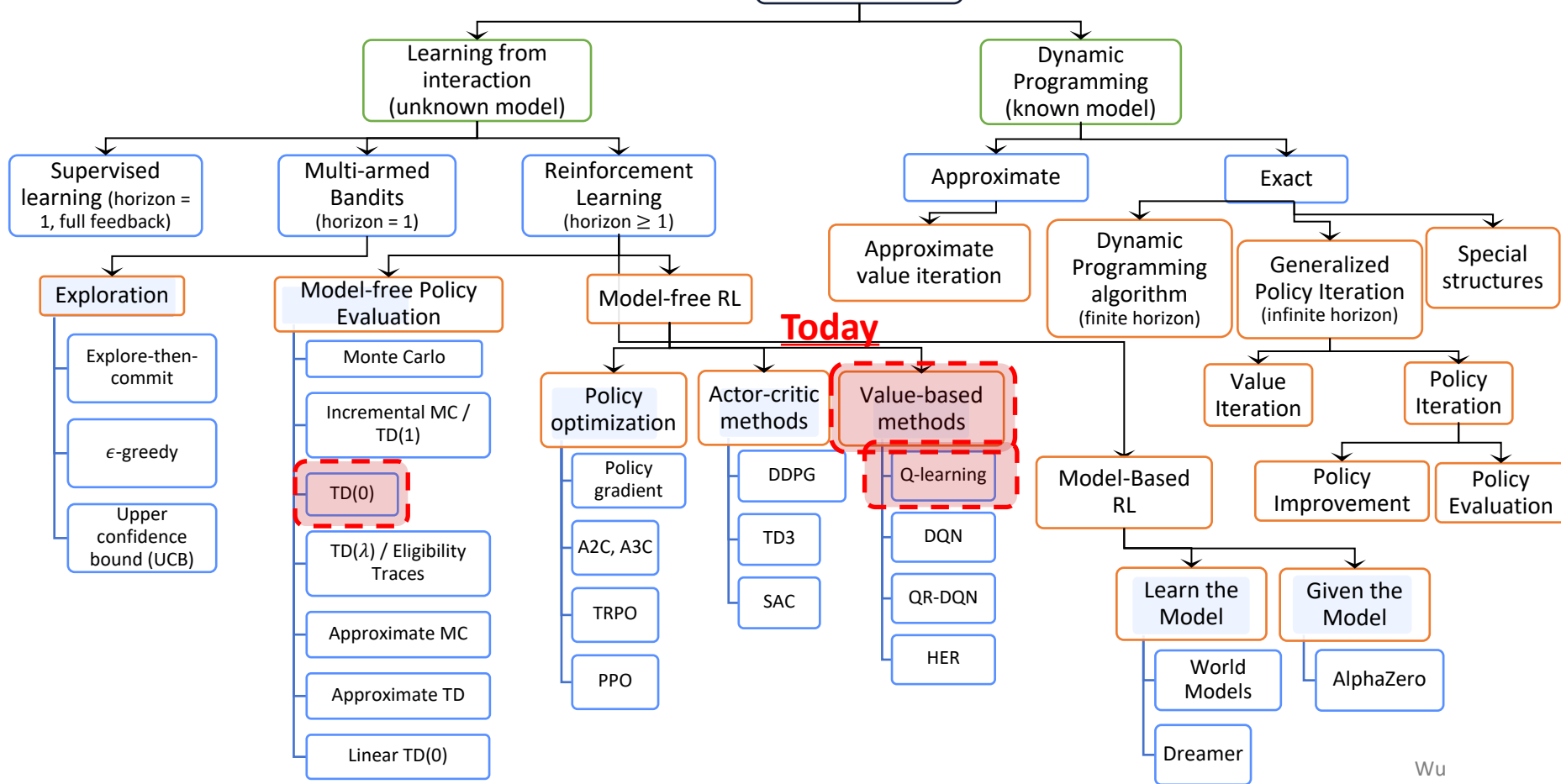
A brief history of 6.7920

- 6.231: Programming Linguistics (late 1960s)
 - Evolved into *Structure and Interpretation of Computer Programs* (SICP)
 - → 6.001 Notorious Intro to CS with LISP (sunsetted in 2008)
 - → 6.01 Intro to EECS w/ Robotics → 6.0001 → 6.100
- 6.231: Dynamic Programming and Stochastic Control (1980s–2010s)
 - Taught by Dimitri P. Bertekas and others: John Tsitsiklis, Munther Dahleh
 - Textbooks: DPOC volumes I and II
 - Renamed to Dynamic Programming and *Reinforcement Learning* in 2017 but not offered
 - Cathy Wu is hired in 2018, joins in 2019, and is asked to teach this course
- 6.7920: Reinforcement Learning (2019–Current)
 - Taught by Cathy Wu and Munther Dahleh (starting in 2024)
 - Textbooks: DPOC vol 1, NDP, S&B

Taxonomy of RL Methods

Methods for
sequential
decision making

Reinforcement learning (RL) is a framework for
sequential decision making under uncertainty.



Outline

1. Fixed points
2. Stochastic approximation
3. Application to TD(0) & Q-learning
4. Max norm convergence analysis
5. Handling non-i.i.d. noise

Fixed Point

- Let H be a mapping from $\mathbb{R}^n \rightarrow \mathbb{R}^n$ (into itself).
- We are interested in solving a system of (possibly nonlinear) equations
$$H(x) = x$$
- A solution $x^* \in \mathbb{R}^n$ which satisfies $H(x^*) = x^*$ is called a **fixed point** of H .

Example: Simple fixed point equations

- **Mean.** Consider $H(x) := \mu$, where μ can be treated as simply some constant.
- **Gradient descent.** Consider $H(x) := x - \nabla f(x)$ for some cost function f .

Possible algorithms

$H(x)$ is known precisely

- $x \leftarrow H(x)$
- $x \leftarrow (1 - \eta)x + \eta H(x)$ (small steps version)

$H(x)$ is not precisely known $\rightarrow$ stochastic approximation algorithm

- $x \leftarrow (1 - \eta)x + \eta(H(x) + w)$
- E.g., stochastic gradient descent

Example: Fixed points in dynamic programming

- H is some operator that returns an object in the same space!
 - Example (Linear, Bellman operator): $H(V) := \mathcal{T}^\pi(V)$
 - Example (Nonlinear, Optimal Bellman operator): $H(V) := \mathcal{T}(V)$
 - Both take in value functions and return value functions.
- A solution $x^* \in \mathbb{R}^n$ which satisfies $H(x^*) = x^*$ is called a **fixed point** of H .
 - Example (Linear, Bellman operator): $V^\pi = \mathcal{T}^\pi V^\pi$
 - Example (Nonlinear, Optimal Bellman operator): $V^* = \mathcal{T}V^*$

Stochastic Approximation

■ Stochastic approximation of a mean

- Desired: $\mu_t \rightarrow \mu = \mathbb{E}[X]$
- Data we get is noisy, $\mu + w_t$
- Applications: TD(1)

■ Stochastic approximation of a fixed point

- Desired: $x_t \rightarrow x^*$, where x^* is a solution to $H(x) = x$
- Data we get is noisy, $H(x_t) + w_t$
- Applications: TD(0), TD(λ), Q-learning

Stochastic Approximation

- Hope (and actuality):

$$\mu_{t+1} = (1 - \eta_t)\mu_t + \eta_t(\mu + w_t)$$

$$x_{t+1} = (1 - \eta_t)x_t + \eta_t(H(x_t) + w_t)$$

converge to the desired quantity, under appropriate conditions.

- Generalization to component-wise updates:

$$x_{t+1}(s) = (1 - \eta_t)x_t(s) + \eta_t(H(x_t)(s) + w_t(s)) \quad \forall s \in \mathcal{S}$$

Stochastic Approximation of a Fixed Point

Summary of results: two kinds of norms, two kinds of analysis

- H is contraction w.r.t. max norm ($\|\cdot\|_\infty$)
- H is a contraction w.r.t. Euclidean norm ($\|\cdot\|_2$)

Under these contractive norms, with some additional assumptions,
 $x_t \rightarrow x^*$ a.s.

Max Norm Convergence Result (Prop 4.4, NDP)

Proposition

Let x_t be the sequence generated by the iteration

$$x_{t+1}(s) = (1 - \eta_t)x_t(s) + \eta_t(H(x_t)(s) + w_t(s)) \quad t = 0, 1, \dots$$

If:

- a) [Robbins-Monro stepsize] The step sizes $\eta_t \geq 0$ and are such that

$$\sum_{t \geq 0} \eta_t = \infty; \quad \sum_{t \geq 0} \eta_t^2 < \infty$$

- b) [Unbiasedness] For every s, t we have zero-mean noise $\mathbb{E}[w_t(s) | \mathcal{F}_t] = 0$.

- c) [Bounded variance] Given any norm $\|\cdot\|$ on $\mathbb{R}^n$, there exist constants A and B such that the variance of the noise is bounded as

$$\mathbb{E}[w_t^2(s) | \mathcal{F}_t] \leq A + B \|x_t\|^2, \quad \forall s, t$$

- d) [Contraction] The mapping H is a max norm contraction.

Then, x_t converges to x^* with probability 1.

Terminology: Filtration $\mathcal{F}_t$ (probability theory) can be thought of as history up to time t .
 $\mathcal{F}_t = \{x_0, \dots, x_t, s_0, \dots, s_{t-1}, \eta_0, \dots, \eta_t\}$

Application of Max Norm Convergence to Q-Learning

Recall:

- Compute the (optimal) temporal difference on the trajectory $\langle s_t, a_t, r_{t+1}, s_{t+1} \rangle$

$$\delta_t = r_{t+1} + \gamma \max_{a'} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t)$$

- Then, update the estimate of Q as

$$\hat{Q}(x_t, a_t) = \hat{Q}(s_t, a_t) + \eta(s_t, a_t) \delta_t$$

Proposition

If the learning rate satisfies the Robbins-Monro conditions in all states $s, a \in S \times A$

$$\sum_{i=0}^{\infty} \eta_t(s, a) = \infty \quad \sum_{i=0}^{\infty} \eta_t^2(s, a) < \infty$$

And all state-action pairs are tried **infinitely often**, then for all $s, a \in S \times A$

$$\hat{Q}(s, a) \xrightarrow{a.s.} \hat{Q}^*(s, a)$$

Terminology: Referred to as first-visit TD(0) in S&B.

Similarly for $TD(0)$

- $TD(0)$ update (for t^{th} trajectory τ_t):

$$V_{t+1}(s) = V_t(s) + \eta_t \delta_t(s), \quad \forall s \in \mathcal{S}$$

With temporal difference $\delta_t(s)$

$$\delta_t(s) = r(s, s') + \gamma V_t(s') - V_t(s) \quad \text{when } s \in \tau_t, \text{ otherwise } 0$$

- Exercise: Apply Prop 4.4 to show that TD(0) converges to V^π

Peeling back the onion for Q-learning analysis



Max Norm Convergence Result (Prop 4.4, NDP)

Proposition

Let x_t be the sequence generated by the iteration

$$x_{t+1}(s) = (1 - \eta_t(s))x_t(s) + \eta_t(s)(H(x_t)(s) + w_t(s)) \quad t = 0, 1, \dots$$

If for each coordinate s :

- a) [Robbins-Monro stepsize] The step sizes $\eta_t \geq 0$ and are such that

$$\sum_{t \geq 0} \eta_t(s) = \infty; \quad \sum_{t \geq 0} \eta_t^2(s) < \infty$$

- b) [Unbiasedness] For every s, t we have zero-mean noise $\mathbb{E}[w_t(s) | \mathcal{F}_t] = 0$.

- c) [Bounded variance] Given any norm $\|\cdot\|$ on $\mathbb{R}^n$, there exist constants A and B such that the variance of the noise is bounded as

$$\mathbb{E}[w_t^2(s) | \mathcal{F}_t] \leq A + B \|x_t\|^2, \quad \forall s, t$$

- d) [Contraction] The mapping H is a max norm contraction.

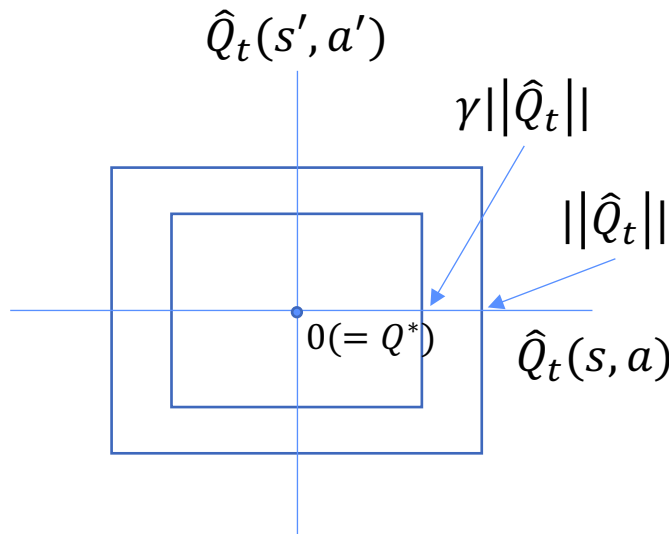
Then, x_t converges to x^* with probability 1.

Terminology: Filtration $\mathcal{F}_t$ (probability theory) can be thought of as history up to time t .
 $\mathcal{F}_t = \{x_0, \dots, x_t, s_0, \dots, s_{t-1}, \eta_0, \dots, \eta_t\}$

Sketch: Max Norm Contraction Analysis (Prop 4.4)

- Overall proof strategy: show that **an upper bound of the iterates $\|x_t\|$ contracts**.
Therefore, $\|x_t\|$ contracts.
- Note: w.l.o.g. assume that $x^* = 0$
 - Can translate the origin of the coordinate system.
- Assume that x_t is bounded.
 - This can be shown precisely (see NDP Prop 4.7).
- The upper bound can be decomposed into a **deterministic** and a **stochastic (noise)** component (induction argument).
- The deterministic component **contracts as expected in due time** (induction argument, Bellman operators).
- The noise component **goes to 0 w.p. 1** (**Supermartingale Convergence Theorem**).
- Therefore, the overall x_t contracts.

For Q-learning, let $x_t := \hat{Q}_t$



Supermartingale Convergence Theorem

Generalization to a **probabilistic context** of the fact that **a bounded monotonic sequence converges**.

Proposition (Supermartingale convergence theorem (Neveu, 1975, p33))

Let X_t, Y_t , and $Z_t, t = 0, 1, 2, \dots$, be three sequences of random variables. Furthermore, let $\mathcal{F}_t, t = 0, 1, 2, \dots$, be sets of random variables such that $\mathcal{F}_t \subset \mathcal{F}_{t+1}, \forall t$. Suppose that:

- a) **[Nonnegative]** The random variables X_t, Y_t , and Z_t are nonnegative, and are functions of the random variables in $\mathcal{F}_t$.
- b) **[Almost Supermartingale energy]** For each t , we have

$$\mathbb{E}[Y_{t+1} | \mathcal{F}_t] \leq Y_t - X_t + Z_t$$

- a) **[Energy injection budget]** There holds $\sum_{t=0}^{\infty} Z_t < \infty$.

Then,

1. **[Energy]** Y_t converges to a limit with probability 1,
2. **[Total energy leak]** $\sum_{t=1}^{\infty} X_t < \infty$ with probability 1.

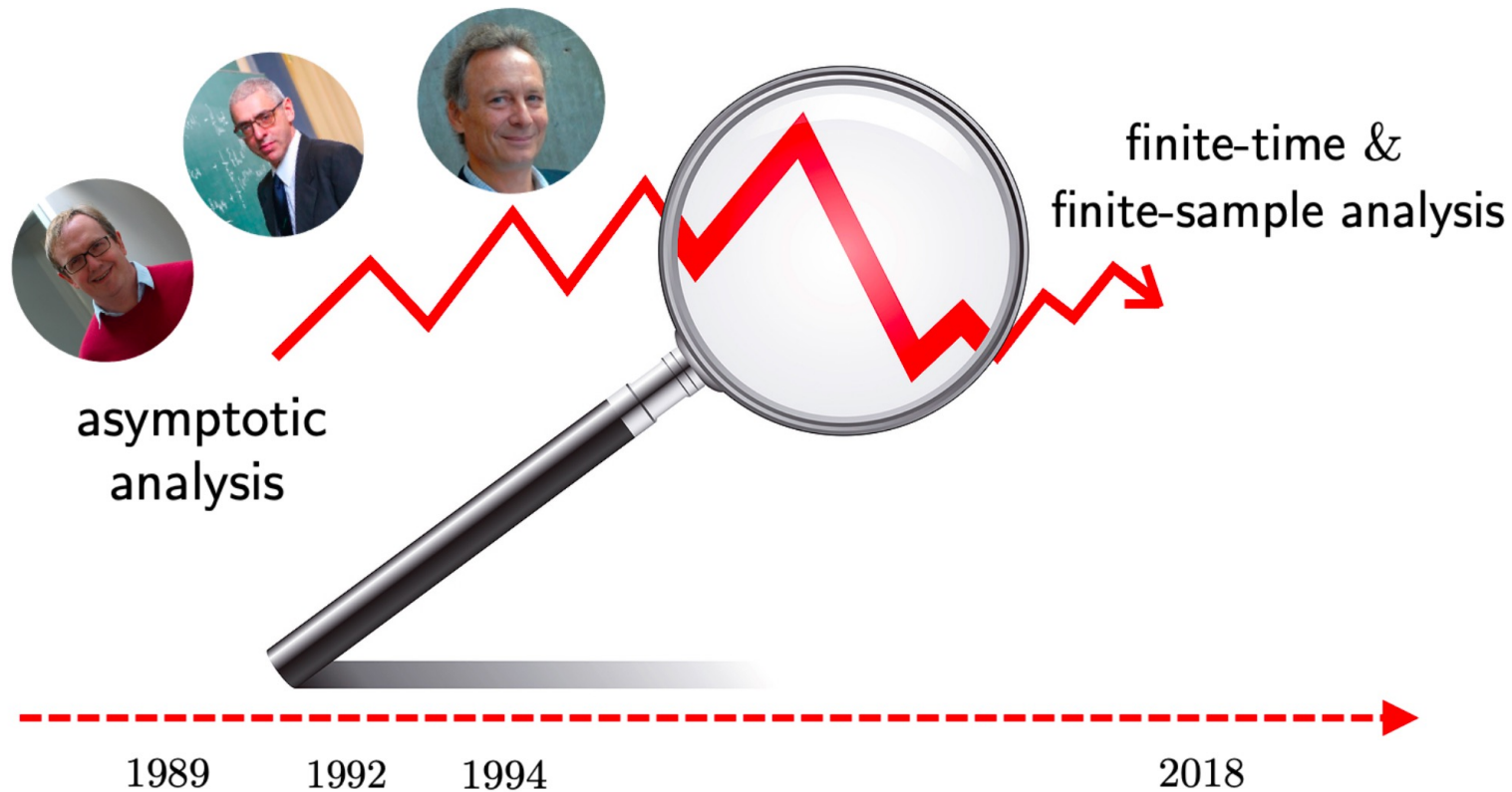
Correspondence to noise
upper bound (intuition)

$$Y_t \leftarrow W_t^2; \mathcal{F}_t \leftarrow \tau_t;$$

$$X_t \leftarrow \eta_t W_t^2;$$

$$Z_t \leftarrow \eta_t^2 \mathbb{V}(w_t)$$

Developments on Q-learning



Developments on Q-learning (an incomplete list!)

■ Asynchronous Q-learning

Asymptotic
analysis

- Watkins, Dayan '92
- Tsitsiklis '94
- Jaakkola, Jordan, Singh '94
- Szepesvári '98
- Borkar, Meyn '00

Finite-time and
finite-sample
analysis

- Even-Dar, Mansour '03
- Beck, Srikant '12
- Chi, Zhu, Bubeck, Jordan '18
- Lee, He '18
- Chen, Zhang, Doan, Maguluri, Clarke '19
- Du, Lee, Mahajan, Wang '20
- Chen, Maguluri, Shakkottai, Shanmugam '20
- Qu, Wierman '20
- Devraj, Meyn '20
- Weng, Gupta, He, Ying, Srikant '20
- Li, Wei, Chi, Gu, Chen '20
- Li, Cai, Chen, Wei, Chi '21
- Chen, Maguluri, Shakkottai, Shanmugam '21
- ...

Question: how many samples are needed to ensure $\|\hat{Q} - Q^*\|_\infty \leq \varepsilon$?

other papers	sample complexity
Even-Dar, Mansour '03	$\frac{(t_{\text{cover}})^{\frac{1}{1-\gamma}}}{(1-\gamma)^4 \varepsilon^2}$
Even-Dar, Mansour '03	$\left(\frac{t_{\text{cover}}^{1+3\omega}}{(1-\gamma)^4 \varepsilon^2}\right)^{\frac{1}{\omega}} + \left(\frac{t_{\text{cover}}}{1-\gamma}\right)^{\frac{1}{1-\omega}}, \omega \in (\frac{1}{2}, 1)$
Beck & Srikant '12	$\frac{t_{\text{cover}}^3 \mathcal{S} \mathcal{A} }{(1-\gamma)^5 \varepsilon^2}$
Qu & Wierman '20	$\frac{t_{\text{mix}}}{\mu_{\min}^2 (1-\gamma)^5 \varepsilon^2}$
Li, Wei, Chi, Gu, Chen '20	$\frac{1}{\mu_{\min} (1-\gamma)^5 \varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min} (1-\gamma)}$
Chen, Maguluri, Shakkottai, Shanmugam '21	$\frac{1}{\mu_{\min}^3 (1-\gamma)^5 \varepsilon^2} + \text{other-term}(t_{\text{mix}})$

— cover time: $t_{\text{cover}} \asymp \frac{t_{\text{mix}}}{\mu_{\min}}$

Summary

- **Stochastic approximation of fixed points** (results, contractive norms, analyses)
 - Supermartingale convergence theorem: Helps handle non-i.i.d. noise
- TD and Q-learning as stochastic approximation methods

References

1. Alessandro Lazaric. INRIA Lille. Reinforcement Learning. 2017, Lectures 2-3.
2. Neuro-dynamic Programming (NDP). Ch 3-5 (esp. §5.6, §4.1-4.3, §6.1-6.2).
3. DPOC2 §6.3
4. Daniela Pucci De Farias. MIT 2.997 Decision-Making in Large-Scale Systems. Spring 2004, Lecture 8.

Reference: Proof of Prop 4.4 upper bound

Proof: Max Norm Contraction Analysis (Prop 4.4)

- **Deterministic part of upper bound:** Since x_t is bounded, there exists some D_0 s.t. $\|x_t\|_\infty \leq D_0, \forall t$. We define:

$$D_{k+1} = \gamma D_k, \quad k \geq 0$$

- Clearly, D_k converges to zero.
 - For TD(0), can think of D_k as upper bound on $H(V_t)(s) = \mathbb{E}[r(s, s') + \gamma V_t(s')]$.

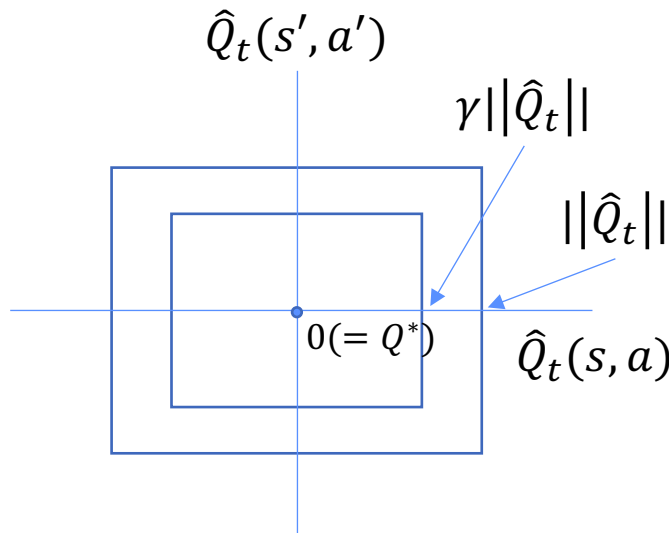
- Proof idea (by induction): suppose there exists some t_k s.t.

$$\|x_t\|_\infty \leq D_k, \forall t \geq t_k$$

Then, there exists some later time t_{k+1} s.t.

$$\|x_t\|_\infty \leq D_{k+1}, \forall t \geq t_{k+1}$$

For Q-learning, let $x_t := \hat{Q}_t$



Proof: Max Norm Contraction Analysis (Prop 4.4)

- Define combined upper bound (**need to confirm**) (for all $t \geq t_k$):

$$Y_{t_k}(s) = D_k + W_{t_k}(s); \quad Y_{t+1}(s) = (1 - \eta_t)Y_t(s) + \eta_t \gamma D_k + \eta_t w_t(s)$$
- Confirm combined upper bound via induction:
 - Suppose $|x_t(s)| \leq Y_t(s), \forall s$, for some $t \geq t_k$. We then have:

$$\begin{aligned} x_{t+1}(s) &= (1 - \eta_t)x_t(s) + \eta_t(H(x_t)(s) + w_t(s)) \\ &\leq (1 - \eta_t)Y_t(s) + \eta_t(H(x_t)(s) + w_t(s)) \\ &\leq (1 - \eta_t)Y_t(s) + \eta_t(\gamma D_k + w_t(s)) \\ &= Y_{t+1}(s) \end{aligned}$$

Where the last inequality is due to $|H(x_t)(s)| \leq \gamma \|x_t\| \leq \gamma D_k$.

- Since $\sum_t^\infty \eta_t = \infty$ and $\lim_{t \rightarrow \infty} W_t(s) = 0$, Y_t converges to γD_k as $t \rightarrow \infty$ a.s.
 This yields:

$$\limsup_{t \rightarrow \infty} \|x_t\| \leq \gamma D_k =: D_{k+1}$$

- Therefore, there exists some time t_{k+1} s.t. $\|x_t\| \leq D_{k+1}, \forall t \geq t_{k+1}$. ■