# Value-based reinforcement learning

Policy learning without knowing how the world works

**Cathy Wu**

6.7920 Reinforcement Learning: Foundations and Methods

# Readings

1. Neuro-dynamic Programming (NDP). §5.6, §4.1-4.3, §6.1-6.2. Skim Ch 3-5 as needed.

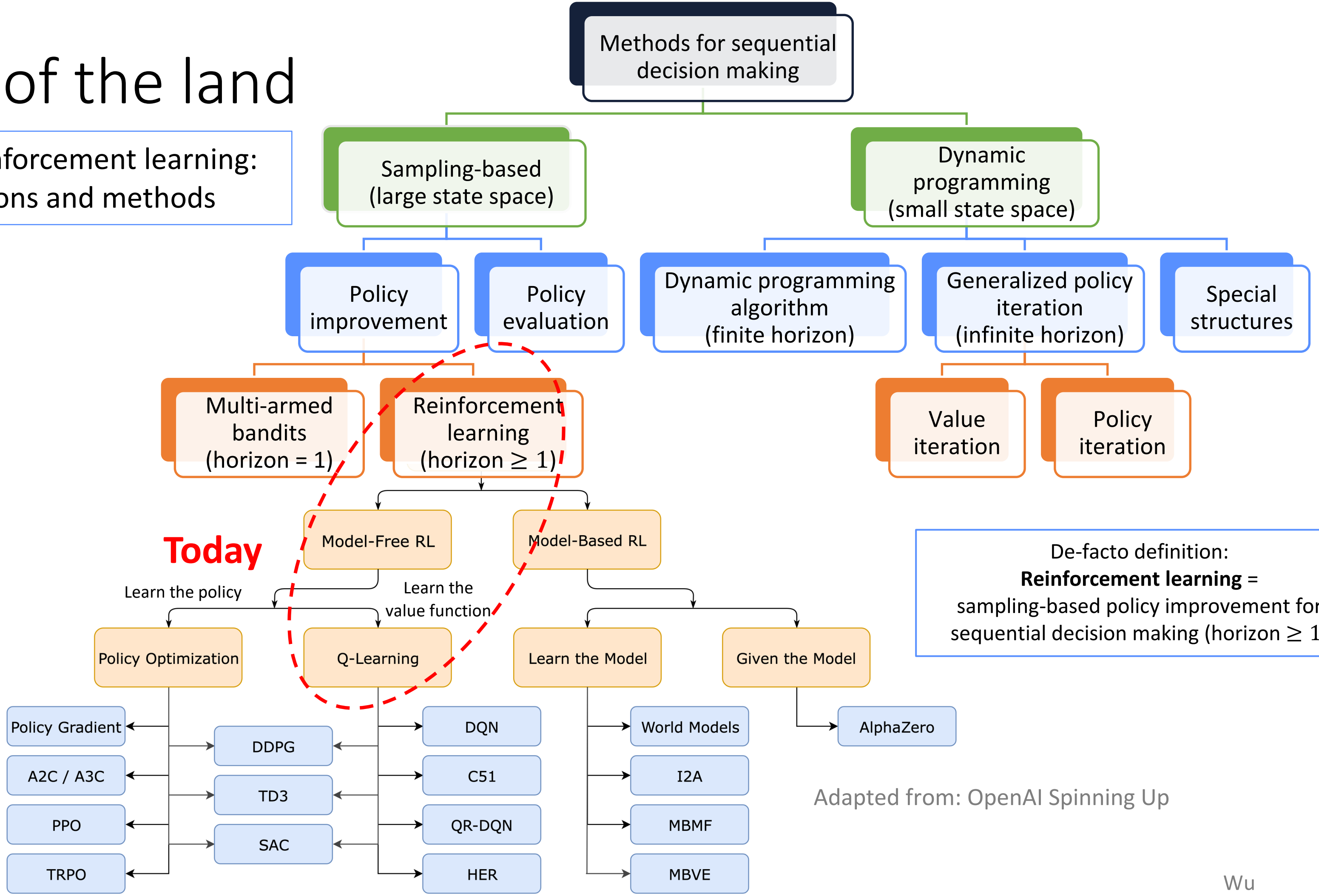
2. DPOC2 §6.3

# Outline

1. Policy learning
2. Convergence analysis – stochastic approximation of a fixed point

# Lay of the land

6.7920: Reinforcement learning: foundations and methods




De-facto definition:
**Reinforcement learning** = sampling-based policy improvement for sequential decision making (horizon ≥ 1)

Adapted from: OpenAI Spinning Up

# Outline

1. **Policy learning**
    a. State-action value function
    b. Q-iteration
    c. Q-learning
    d. On-policy vs off-policy learning
2. Convergence analysis – stochastic approximation of a fixed point

# Policy Learning

**Learn optimal policy** $\pi^*$

**For** $i = 1, \ldots, n$ [each of $n$ episodes]

1. Set $t = 0$
2. Set initial state $s_0$
3. **While** ($s_{t,i}$ not terminal) [execute one trajectory]
    1. Take action $a_{t,i}$ [Compare Policy Evaluation: Take action $a_{t,i} = \pi(s_{t,i})$]
    2. Observe next state $s_{t+1,i}$ and reward $r_{t,i} = r(s_{t,i}, a_{t,i})$
    3. Set $t = t + 1$

    **EndWhile**

**Endfor**

**Return**: Estimate of the value function $\hat{\pi}^*$

# State-Action Value Function ("Q")

## Definition

In discounted infinite horizon problems, for any policy $\pi$, the state-action value function (or Q-function) $Q^\pi : S \times A \mapsto \mathbb{R}$ is

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a, a_t = \pi(s_t), \forall t \geq 1\right]$$

The optimal Q-function is

$$Q^*(s, a) = \max_\pi Q^\pi(s, a)$$

and the optimal policy can be obtained as

$$\pi^*(s) = \arg\max_a Q^*(s, a)$$

- Recall: definition of value function, $V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r\left(s_t, \pi(s_t)\right) | s_0 = s; \pi\right]$

# State-Action Value Function Operators*

- $\mathcal{T}^\pi Q(s,a) \coloneqq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[Q(s',\pi(s))]$
  - Compare: $\mathcal{T}^\pi V(s) \coloneqq r\big(s,\pi(s)\big) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))}[V(s')]$
- $\mathcal{T} Q(s,a) \coloneqq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} Q(s',a')\right]$
  - Compare: $\mathcal{T} V(s) \coloneqq \max_{a \in A} r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s')$

Still true:

- $Q^* = \mathcal{T} Q^*$
- $Q^\pi = \mathcal{T}^\pi Q^\pi$

Note: Abuse of notation for the operators

# State-Action and State Value Function

- $Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^{\pi}(s')]$
- $V^{\pi}(s) = Q^{\pi}\big(s, \pi(s)\big)$

- $Q^{*}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^{*}(s')]$
- $V^{*}(s) = Q^{*}\big(s, \pi^{*}(s')\big) = \max_{a \in A} Q^{*}(s,a)$

# Q-value Iteration

1. Let $Q_0(s,a)$ be any Q-function $Q_0 : S \times A \to \mathbb{R}$
2. At each iteration $k = 1, 2, \dots, K$
   - Compute $Q_{k+1} = \mathcal{T} Q_k$
3. Terminate when $Q_k$ stops improving
   - e.g. when $\max_s |Q_{k+1}(s) - Q_k(s)|$ is small.
4. Return the greedy policy

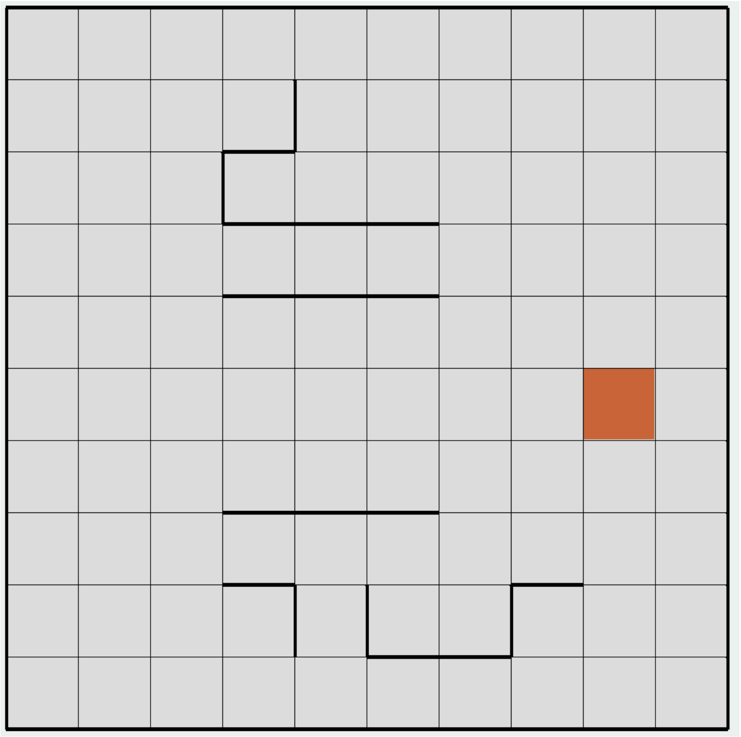
$$\pi_K(s) \in \arg \max_{a \in A} Q_K(s,a)$$

# Compare: Value iteration algorithm

1. Let $V_0(s)$ be any function $V_0 : S \to \mathbb{R}$.
2. At each iteration $k = 1, 2, \ldots, K$
   - Compute $V_{k+1} = \mathcal{T} V_k$
3. Terminate when $V_k$ stops improving
   - e.g. when $\max_s |V_{k+1}(s) - V_k(s)|$ is small.
4. Return the greedy policy

$$\pi_K(s) \in \arg\max_{a \in A} r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V_K(s')$$

# The Grid-World Problem

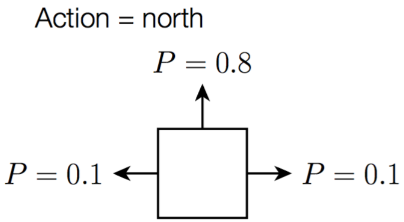


State: agent location

# Example: Winter parking (with ice and potholes)

- Simple grid world with a *goal state* (green, desired parking spot) with reward (+1), a *"bad state"* (red, pothole) with reward (-100), and all other states neural (+0).
- *Omnidirectional vehicle (agent)* can head in any direction. Actions move in the desired direction with probably 0.8, in one of the perpendicular directions with.
- Taking an action that would bump into a wall leaves agent where it is.

| | | | |
|---|---|---|---|
| 0 | 0 | 0 | 1 |
| 0 | | 0 | -100 |
| 0 | 0 | 0 | 0 |




*[Source: adapted from Kolter, 2016]*

# Example: value iteration

Running value iteration with $\gamma = 0.9$

| | | | |
|---|---|---|---|
| 0 | 0 | 0 | 1 |
| 0 | | 0 | -100 |
| 0 | 0 | 0 | 0 |

Original reward function

(a)

Running value iteration with $\gamma = 0.9$

| | | | |
|---|---|---|---|
| 0 | 0 | 0.72 | 1.81 |
| 0 | | 0 | -99.91 |
| 0 | 0 | 0 | 0 |

$\hat{V}$ at one iteration

(b)

Running value iteration with $\gamma = 0.9$

| | | | |
|---|---|---|---|
| 0.809 | 1.598 | 2.475 | 3.745 |
| 0.268 | | 0.302 | -99.59 |
| 0 | 0.034 | 0.122 | 0.004 |

$\hat{V}$ at five iterations

(c)

Running value iteration with $\gamma = 0.9$

| | | | |
|---|---|---|---|
| 2.686 | 3.527 | 4.402 | 5.812 |
| 2.021 | | 1.095 | -98.82 |
| 1.390 | 0.903 | 0.738 | 0.123 |

$\hat{V}$ at 10 iterations

(d)

Running value iteration with $\gamma = 0.9$

| | | | |
|---|---|---|---|
| 5.470 | 6.313 | 7.190 | 8.669 |
| 4.802 | | 3.347 | -96.67 |
| 4.161 | 3.654 | 3.222 | 1.526 |

$\hat{V}$ at 1000 iterations

(e)

Running value iteration with $\gamma = 0.9$

| | | | |
|---|---|---|---|
| $\rightarrow$ | $\rightarrow$ | $\rightarrow$ | $\uparrow$ |
| $\uparrow$ | | $\leftarrow$ | $\leftarrow$ |
| $\uparrow$ | $\leftarrow$ | $\leftarrow$ | $\downarrow$ |

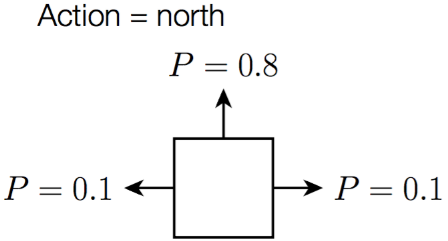
Resulting policy after 1000 iterations

(f)

# State-Action Value Function ("Q table")

- Example: Winter parking (with ice and potholes)

| | | | |
|---|---|---|---|
| 0 | 0 | 0 | 1 |
| 0 | | 0 | -100 |
| 0 | 0 | 0 | 0 |




Running value iteration with $\gamma = 0.9$

$V(s)$

| | | | |
|---|---|---|---|
| 5.470 | 6.313 | 7.190 | 8.669 |
| 4.802 | | 3.347 | -96.67 |
| 4.161 | 3.654 | 3.222 | 1.526 |

$\hat{V}$ at 1000 iterations

It is convenient to keep track of not only the long term value of a state, but also the state, jointly with the next action.

$Q(s, a)$

$s$

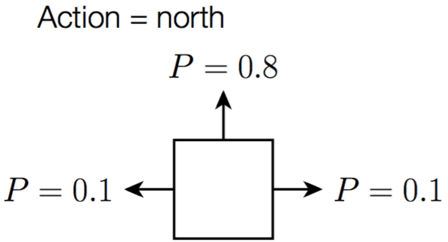
| ↑ | ↓ | ← | → |
|---|---|---|---|
| 2.5 | 1.4 | 3.2 | 5.4 |
| 1.0 | 3.2 | 5.1 | 6.3 |
| 5.2 | 4.2 | 5.5 | 7.2 |
| 8.7 | 3.4 | 2.0 | 8.0 |
| 4.8 | 2.5 | 3.5 | 4.2 |
| 1.0 | 3.0 | 3.3 | 1.2 |
| -180 | -172 | -99.7 | -150 |
| 4.2 | 2.1 | 3.2 | 3.7 |
| 2.1 | 2.0 | 3.7 | 3.1 |
| 3.0 | 1.2 | 3.2 | 2.7 |
| 0.1 | 1.5 | 0.1 | 1.0 |

$a$

# Convenient for selecting next action!

- Winter parking (with ice and potholes)

| | | | |
|---|---|---|---|
| 0 | 0 | 0 | 1 |
| 0 | | 0 | -100 |
| 0 | 0 | 0 | 0 |




Before

Running value iteration with $\gamma = 0.9$

| | | | |
|---|---|---|---|
| 5.470 | 6.313 | 7.190 | 8.669 |
| 4.802 | | 3.347 | -96.67 |
| 4.161 | 3.654 | 3.222 | 1.526 |

$\hat{V}$ at 1000 iterations

Running value iteration with $\gamma = 0.9$

| | | | |
|---|---|---|---|
| → | → | → | ↑ |
| ↑ | | ← | ← |
| ↑ | ← | ← | ↓ |

Resulting policy after 1000 iterations

$$\pi_K(s) = \arg\max_{a \in A} r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V_K(s')$$

$Q(s,a)$

$s$

| ↑ | ↓ | ← | → |
|---|---|---|---|
| 2.5 | 1.4 | 3.2 | 5.4 |
| 1.0 | 3.2 | 5.1 | 6.3 |
| 5.2 | 4.2 | 5.5 | 7.2 |
| 8.7 | 3.4 | 2.0 | 8.0 |
| 4.8 | 2.5 | 3.5 | 4.2 |
| 1.0 | 3.0 | 3.3 | 1.2 |
| -180 | -172 | -99.7 | -150 |
| 4.2 | 2.1 | 3.2 | 3.7 |
| 2.1 | 2.0 | 3.7 | 3.1 |
| 3.0 | 1.2 | 3.2 | 2.7 |
| 0.1 | 1.5 | 0.1 | 1.0 |

$a$

# Policy Iteration (w/ Q-value function)

1. Let $\pi_0$ be any stationary policy
2. At each iteration $k = 1, 2, \ldots, K$
   - Policy evaluation: given $\pi_k$, compute $Q^{\pi_k}$
   - Policy improvement: compute the greedy policy

$$\pi_{k+1}(s) \in \arg\max_{a \in A} Q_k^{\pi}(s, a)$$

3. Stop if $Q^{\pi_k} = Q^{\pi_{k-1}}$
4. Return the last policy $\pi_K$

# Compare: Policy Iteration

1. Let $\pi_0$ be any stationary policy
2. At each iteration $k = 1, 2, \ldots., K$
   - Policy evaluation: given $\pi_k$, compute $V^{\pi_k}$
   - Policy improvement: compute the greedy policy

$$\pi_{k+1}(s) \in \arg\max_{a \in A} \left[ r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^{\pi_k}(s') \right]$$

1. Stop if $V^{\pi_k} = V^{\pi_{k-1}}$
2. Return the last policy $\pi_K$

# Q-Learning (Watkins, 1992)

- Model-free algorithm for learning the optimal policy
- Stochastic approximation lens
  - Model-free Q-function improvement via **incremental updates**
    - Compute $TD$ error for the **optimal** Bellman operator (compare: Bellman operator)
  - Use $\epsilon$-greedy policy to collect data, to ensure that all state-actions are visited enough (for convergence)
    - With probability $1-\epsilon$, choose the best predicted action $\underset{a'}{\operatorname{argmax}}\, \hat{Q}(s_{t+1}, a')$
    - With probability $\epsilon$, choose an action uniformly at random.
- Intuition
  - Use $\epsilon$-greedy policy for data collection (exploration)
  - But use greedy policy for learning (exploitation)

# Recall: Temporal Difference $TD(0)$

**For** $i = 1, \dots, n$ [each of $n$ episodes]

1. Set $t = 0$
2. Set initial state $s_0$
3. **While** ($s_t$ not terminal) [execute one trajectory]
    1. Take action $a_{t,i} = \pi(s_{t,i})$
    2. Observe next state $s_{t+1,i}$ and reward $r_{t,i} = r(s_{t,i}, a_{t,i})$
    3. Set $t = t + 1$
    4. Update $\hat{V}^{\pi}(s_{t,i})$ using $TD(0)$ estimation

    **EndWhile**
4. Update $\hat{V}_i^{\pi}(s_0)$ using incremental Monte-Carlo estimation

**Endfor**

# Learning the Optimal Policy

**For** $i = 1, \ldots, n$

1. Set $t = 0$; Set initial state $s_0$
2. **While** ($s_t$ not terminal)
    1. Take action $a_t$ according to a suitable exploration policy

$$\pi_{\hat{Q}}(a|s) = \begin{cases} \underset{a'}{\operatorname{argmax}}\, \hat{Q}(s_{t+1}, a') & w.p.\ 1-\epsilon \\ Unif(A) & w.p.\ \epsilon \end{cases} \quad \text{($\epsilon$-greedy policy)}$$

$$\pi_{\hat{Q}}(a|s) = \frac{\exp\left(\frac{\hat{Q}(s,a)}{\tau}\right)}{\sum_{a'} \exp\left(\frac{\hat{Q}(s,a')}{\tau}\right)} \quad \text{(soft-max policy)}$$

    1. Observe next state $s_{t+1}$ and reward $r_t$, take action $a_{t+1}$ according to a suitable exploration policy (if needed)
    2. Compute the temporal difference $\delta_t$

$$\delta_t = r_t + \gamma \hat{Q}(s_{t+1}, a_{t+1}) - \hat{Q}(s_t, a_t) \qquad \text{(SARSA)}$$

$$\delta_t = r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t) \qquad \text{(Q–learning)}$$

    1. Update the Q-function

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

    2. Set $t = t + 1$

**EndWhile**

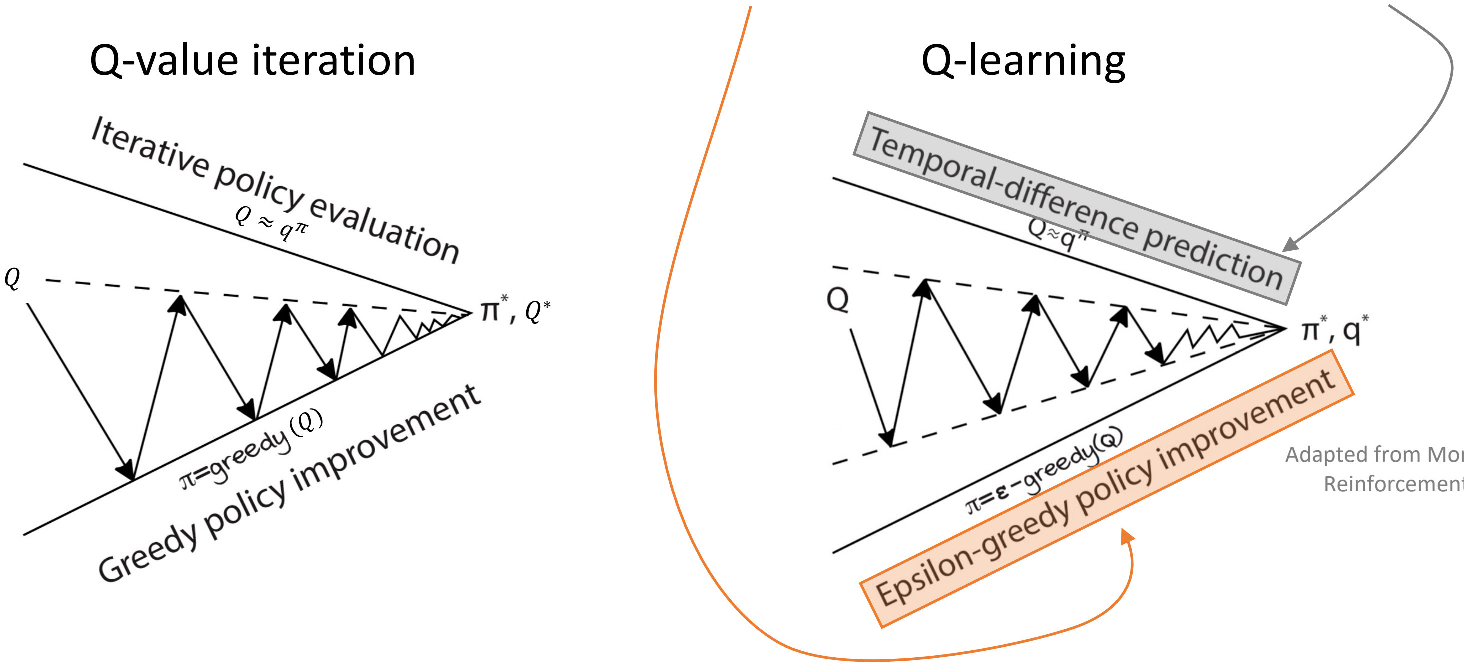
**Endfor**

**Terminology**: on-policy vs off-policy learning

- Two uses of policies
    - Behavior policy: Policy used for interacting (collecting data)
    - Target policy: Policy used for learning
- Q-learning
    - Interacting policy: $\epsilon$-greedy
    - Learning policy: greedy
    - Different → off-policy
- SARSA
    - Interacting policy: $\epsilon$-greedy
    - Learning policy: $\epsilon$-greedy
    - Same → on-policy
- Off-policy = “learning from others”
- On-policy = “learning from oneself”

# Q-learning

- Key idea: incrementally obtain new data and update Q function using the optimal Bellman equation (greedy)




Adapted from Morales, Grokking Deep Reinforcement Learning, 2020.

# Q-Learning: Properties

Understanding this Proposition is the main subject of today + next time.

## Proposition

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

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

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

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

- **Remark**: “infinitely often” requires a steady exploration policy.

# Outline

1. Policy learning

2. **Convergence analysis – stochastic approximation of a fixed point**
   a. Fixed points
   b. Stochastic approximation
   c. Examples: TD(0) & Q-learning
   d. Max norm convergence result & analysis
   e. Handling non-i.i.d. noise

# Fixed Point

- We are interested in solving a system of (possibly nonlinear) equations
$$H(x) = x$$
where $H$ is a mapping from $\mathbb{R}^n \to \mathbb{R}^n$ (into itself).

- 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 $H(x) \coloneqq \mu$, where $\mu$ can be treated as simply some constant.

- Stochastic gradient descent. Consider $H(x) \coloneqq x - \nabla f(x)$ 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 → 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): $H(V) \coloneqq \mathcal{T}^{\pi}(V)$
  - Example (Nonlinear, Optimal Bellman operator): $H(V) \coloneqq \mathcal{T}(V)$
  - 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($\lambda$), 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:

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

# 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, $x_t \to x^*$ a.s.

# Max Norm Convergence Result (Prop 4.4, NDP)

## Proposition

Let $x_t$ be the sequence generated by the iteration

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

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

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

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 $t$.
$\mathcal{F}_t = \{x_0, \dots, x_t, s_0, \dots, s_{t-1}, \eta_0, \dots \eta_t\}$

# Example for max norm: $TD(0)$

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

- $TD(0)$ update (for $t^{\text{th}}$ trajectory $\tau_t$):

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

With temporal difference $\delta_t(s)$

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

- Exercise: Apply Prop 4.4 to show that TD(0) converges to $V^{\pi}$

# Similarly for Q-Learning (see HW)

Recall:

- Compute the (optimal) temporal difference on the trajectory $\langle s_t, a_t, r_t, s_{t+1} \rangle$
$$\delta_t = r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t)$$
- 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 $s, a \in S \times A$

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

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

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

# Summary of Q-learning analysis

Max Norm Convergence $\overset{\text{(HW)}}{\Longrightarrow}$ Q-learning

# Peeling back the onion for Q-learning

# Max Norm Convergence Result (Prop 4.4, NDP)

## Proposition

Let $x_t$ be the sequence generated by the iteration

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

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

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

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

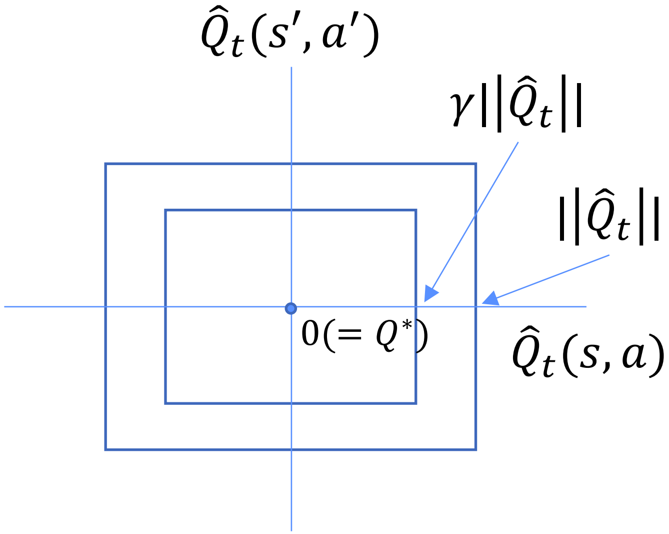
Then, $x_t$ converges to $x^*$ with probability 1.

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

# 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 $x_t \coloneqq \hat{Q}_t$



# Remark

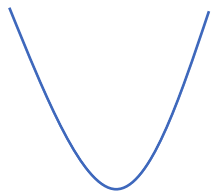
- Deterministic-only upper bound
  - Corresponds to convergence analysis for asynchronous value iteration!

- Q-learning as noisy extension of value iteration.

# Sketch: Handling the noise component in Prop 4.4

$$W_{t+1}(s) = (1-\eta_t)W_t(s) + \eta_t w_t(s) \qquad (1)$$

- Interpretation: $\{W_t(s)\}$ as stochastic gradient descent along a quadratic (Lyapunov) function



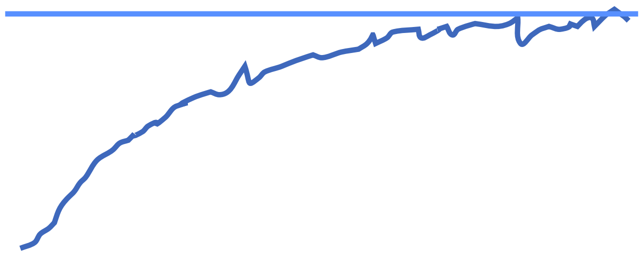
- Descent direction interpretation (take $H(x) \coloneqq x - \nabla f(x)$):

$$\begin{aligned} x_{t+1} &= (1-\eta_t)x_t + \eta_t(x_t - \nabla f(x_t) + w_t) \\ &= x_t + \eta_t(x_t - \nabla f(x_t) - x_t + w_t) \\ &= x_t + \eta_t(-\nabla f(x_t) + w_t) \end{aligned}$$

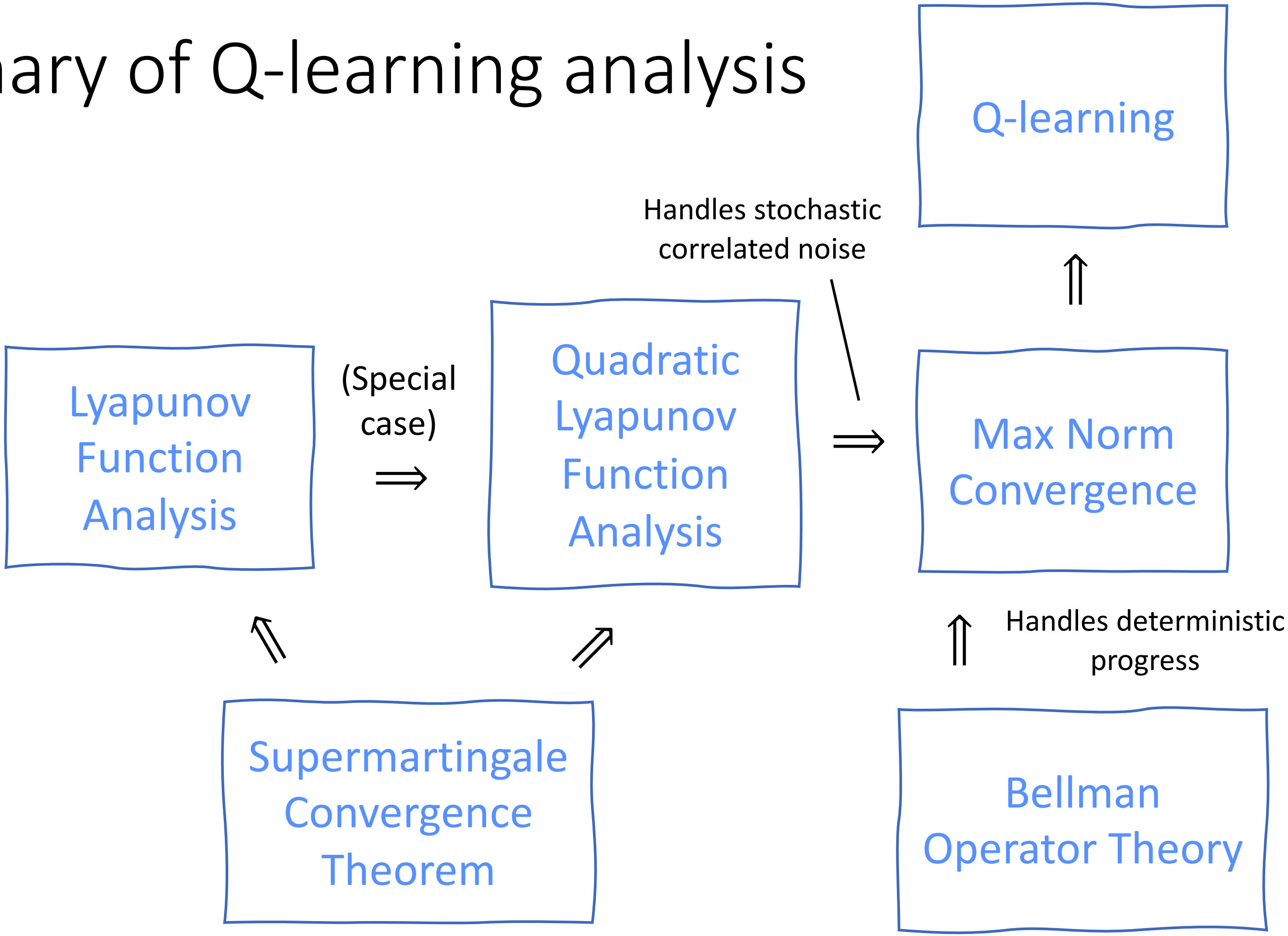
- Corresponds to taking Lyapunov function $f(x) = \frac{1}{2}x^2$
  - Take $x_t \coloneqq W_t(s)$ to recover stochastic approximation update for $W_{t+1}(s)$
  - That is, $-\nabla f(x_t) = x_t = W_t(s)$ recovers (1)
- To show that $W_t(s) \to 0$, sufficient to show that $f(x_t) \to 0$.

# Sketch: Handling the noise component in Prop 4.4

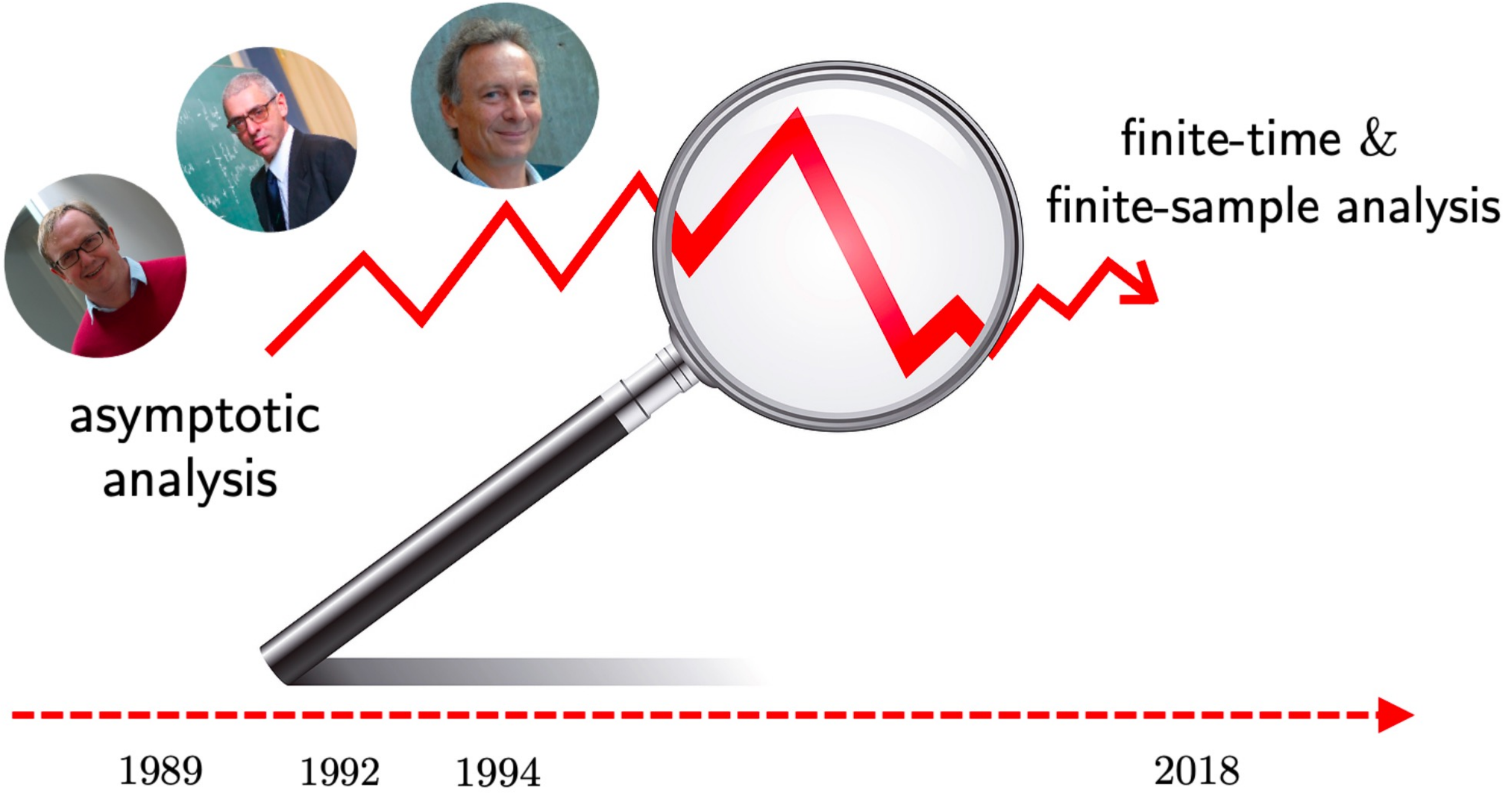
- **Key fact**: $f(x_t)$ turns out to be martingale noise.
  - Martingale noise corresponds to a stochastic Lyapunov function.
- Consequently, martingale noise averages out over time to zero.
- Uses Supermartingale Convergence Theorem
  - Generalization to a probabilistic context of the fact that a bounded monotonic sequence converges.



# Summary of Q-learning analysis



# Developments on Q-learning




Graphic courtesy Wei, Chen, Chi (2023)

# 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 $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$?

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

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

References courtesy Wei, Chen, Chi (2023)

Wu

# Summary

- **State-action value function** $(Q)$ vs state value function $(V)$
  - State-action value function permits model-free extraction of the policy
- Policy learning: SARSA and Q-learning (definition, guarantees)
- **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: Detailed proof of Prop 4.4*

# 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:

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

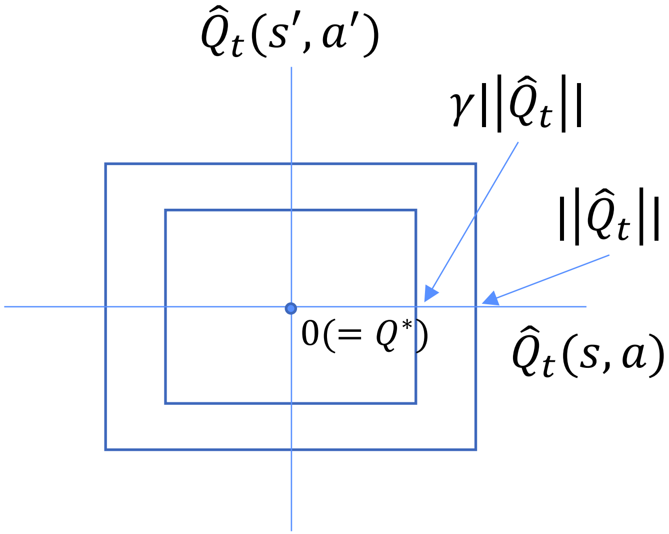
- Clearly, $D_k$ converges to zero.
  - For TD(0), can think of $D_k$ as upper bound on $H(V_t)(s) = \mathbb{E}[r(s,s') + \gamma V_t(s')]$.
- Proof idea (by induction): suppose there exists some $t_k$ s.t.

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

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 $x_t \coloneqq \hat{Q}_t$



# Proof: Max Norm Contraction Analysis (Prop 4.4)

- For the stochastic part of the upper bound, define (need to confirm):

$$W_0(s) = 0;$$
$$W_{t+1}(s) = (1 - \eta_t)W_t(s) + \eta_t w_t(s)$$

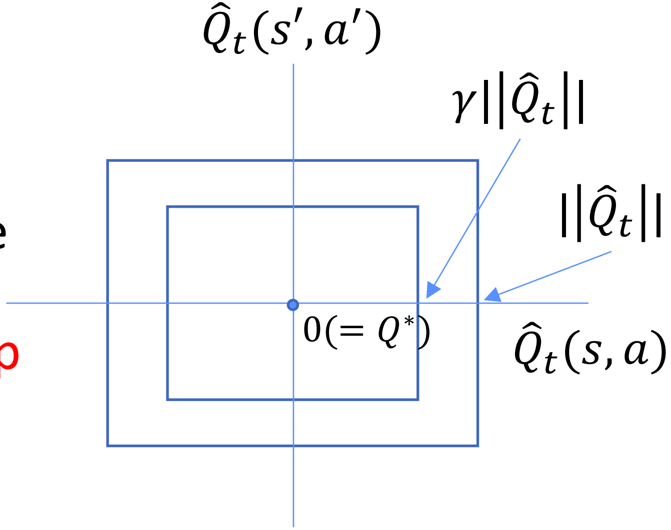
- Since $x_t$ is bounded, so is the conditional variance of $w_t(s)$. Then, as a result of the Supermartingale Convergence Theorem, and Lyapunov Function Analysis (NDP Prop 4.1) (*discussed later*),

$$\lim_{t\to\infty} W_t(s) = 0$$

a.s.

- That is, the noise averages out to zero.

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



# Proof: Max Norm Contraction Analysis (Prop 4.4)

- Define combined upper bound (need to confirm) (for all $t \geq t_k$):
  $Y_{t_k}(s) = D_k + W_{t_k}(s); \qquad Y_{t+1}(s) = (1-\eta_t)Y_t(s) + \eta_t \gamma D_k + \eta_t w_t(s)$
- 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:

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

  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 $\lim_{t\to\infty} W_t(s) = 0$, $Y_t$ converges to $\gamma D_k$ as $t \to \infty$ a.s. This yields:

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

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

*Reference: Detailed theorems and proofs for the noise (Prop 4.4)*

# Quadratic Lyapunov function (special case of NDP Prop 4.1)

## Proposition

Let $x_t$ be the sequence generated by the iteration

$$x_{t+1}(s) = x_t + \eta_t g_t \quad t = 0, 1, \ldots$$

Interpretation as noisy descent direction:
$g_t \coloneqq -\nabla f(x_t) + w_t = -\|r - r^*\| + w_t$

Suppose $f(r) = \frac{1}{2}\|r - r^*\|_2^2$ satisfies:

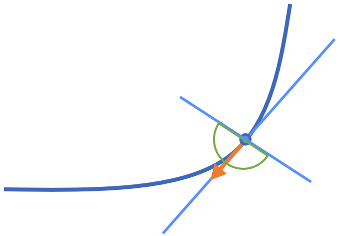
1. [Pseudogradient property] $\exists c$ such that $cf(x_t) \leq -\nabla f(x_t)^T \mathbb{E}[g_t | \mathcal{F}_t]$
2. [Bounded variance] $\exists K_1, K_2$ such that $\mathbb{E}[\|g_t\|_2^2 | \mathcal{F}_t] \leq K_1 + K_2 f(x_t)$

Then if $\eta_t > 0$ with $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$

$$x_t \to r^*, \qquad w.p.1$$

**Terminology:** *Filtration* $\mathcal{F}_t$ (probability theory) can be thought of as history up to time $t$.

- Consequence of conditions (1) and (2) is that $f(x_t)$ is a supermartingale.
- Note: Prop 4.1 will generalize $f(r)$ to general Lyapunov functions (conditions (a) and (b) in Prop 4.1).



# (General) Lyapunov Function Analysis Setup

Descent direction interpretation (take $H(x) := x - \nabla f(x)$):

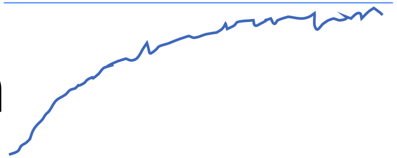
$$\begin{aligned} x_{t+1} &= (1-\eta_t)x_t + \eta_t(x_t - \nabla f(x_t) + w_t) \\ &= x_t + \eta_t(x_t - \nabla f(x_t) - x_t + w_t) \\ &= x_t + \eta_t \underbrace{(-\nabla f(x_t) + w_t)}_{g_t} \\ &= x_t + \eta_t g_t \end{aligned}$$

Slight re-write:

$$\begin{aligned} x_{t+1}(s) &= (1-\eta_t)x_t(s) + \eta_t\big(H(x_t)(s) + w_t(s)\big) \qquad t = 0, 1, \ldots \\ &= x_t(s) + \eta_t \underbrace{\big(H(x_t)(s) - x_t(s) + w_t(s)\big)}_{g_t(s)} \end{aligned}$$

$$\begin{aligned} x_{t+1} &= x_t + \eta_t \underbrace{(H(x_t) - x_t + w_t)}_{g_t} \\ &= x_t + \eta_t g_t \end{aligned}$$

# 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 $Z_t$, $t = 0, 1, 2, \ldots$, be three sequences of random variables. Furthermore, let $\mathcal{F}_t, t = 0, 1, 2, \ldots$, 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) [Non-increasing-ish] For each $t$, we have $\mathbb{E}[Y_{t+1} | \mathcal{F}_t] \le Y_t - X_t + Z_t$.

c) [Diminishing increase] There holds $\sum_{t=0}^{\infty} Z_t < \infty$.

Then,

*1.* $Y_t$ converges to a limit with probability 1,

*2.* $\sum_{t=1}^{\infty} X_t < \infty$ with probability 1.

Correspondence to noise upper bound

(intuition)

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

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

# Proof: quadratic Lyapunov function

Key idea: show that $f(x_t)$ is a supermartingale, so $f(x_t)$ converges. Then show converges to zero w.p. 1.

- $$E[f(x_{t+1})|\mathcal{F}_t] = E\left[\frac{1}{2}\|x_{t+1} - r^*\|_2^2 \middle| \mathcal{F}_t\right]$$
$$= E\left[\frac{1}{2}(x_t + \eta_t g_t - r^*)^T(x_t + \eta_t g_t - r^*)\middle|\mathcal{F}_t\right] \quad (g_t \triangleq g(x_t, w_t))$$
$$= \frac{1}{2}(x_t - r^*)^T(x_t - r^*) + \eta_t (x_t - r^*)^T E[g_t|\mathcal{F}_t] + \frac{\eta_t^2}{2} E[g_t^T g_t|\mathcal{F}_t]$$
$$= f(x_t) + \eta_t (x_t - r^*)^T E[g_t|\mathcal{F}_t] + \frac{\eta_t^2}{2} E[\|g_t\|_2^2|\mathcal{F}_t]$$
- Since $f(x_t) = \frac{1}{2}\|x_t - r^*\|_2^2, \nabla f(x_t) = x_t - r^*$. Then:
- $$E[f(x_{t+1})|\mathcal{F}_t] = f(x_t) + \eta_t \nabla f(x_t)^T E[g_t|\mathcal{F}_t] + \frac{\eta_t^2}{2} E[\|g_t\|_2^2|\mathcal{F}_t]$$
$$\le f(x_t) - \eta_t c f(x_t) + \frac{\eta_t^2}{2}(K_1 + K_2 f(x_t)) \quad \text{(P4.1 conditions 1 \& 2)}$$
$$\le \underbrace{f(x_t)}_{Y_t} - \underbrace{\left(\eta_t c - \frac{\eta_t^2 K_2}{2}\right) f(x_t)}_{X_t} + \underbrace{\frac{\eta_t^2}{2} K_1}_{Z_t} \quad \text{(SCT condition b)}$$

Correspondence to noise upper bound (intuition)

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

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

# Proof: quadratic Lyapunov function

$$E[f(x_{t+1})|\mathcal{F}_t] \leq \underbrace{f(x_t)}_{Y_t} - \underbrace{\left(\eta_t c - \frac{\eta_t^2 K_2}{2}\right) f(x_t)}_{X_t} + \underbrace{\frac{\eta_t^2}{2} K_1}_{Z_t}$$

- Since $\eta_t > 0$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$, then $X_t \geq 0$ for large enough $t$ (SCT condition a)
- Moreover: $\sum_{t=0}^{\infty} Z_t = \frac{K_1}{2} \sum_{t=0}^{\infty} \eta_t^2 < \infty$ (SCT condition c)
- Therefore, by **Supermartingale convergence theorem**:

$$f(x_t) \text{ converges w.p. 1}, \qquad \text{and} \qquad \sum_{t=0}^{\infty} \left(\eta_t c - \frac{\eta_t^2 K_2}{2}\right) f(x_t) < \infty, \text{ w.p. 1}$$

- Suppose that $f(x_t) \to \epsilon > 0$. Then, by hypothesis that $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$, we must have:

$$\sum_{t=0}^{\infty} \left(\eta_t c - \frac{\eta_t^2 K_2}{2}\right) f(x_t) = \infty$$

- Which is a contradiction. Therefore:

$$\lim_{t\to\infty} f(x_t) = \lim_{t\to\infty} \frac{1}{2} \|x_t - r^*\|_2^2 = 0 \quad \text{w.p. 1} \quad \Longrightarrow \quad x_t \to r^* \quad \text{w.p. 1}$$

$\blacksquare$

# Lyapunov Function Analysis (NDP Prop 4.1)

## Proposition

Let $x_t$ be the sequence generated by the iteration

$$x_{t+1}(s) = x_t + \eta_t g_t \quad t = 0, 1, \ldots$$

If the stepsizes $\eta_t \geq 0$ and are such that $\sum_{t\geq 0} \eta_t = \infty; \sum_{t\geq 0} \eta_t^2 < \infty$, and there exists a function $f: \mathbb{R}^n \to \mathbb{R}^n$, with:

Lyapunov function

a) [Non-negativity] $f(x) \geq 0, \forall x \in \mathbb{R}$.

b) [Lipschitz continuity of $\nabla f$] The function $f$ is continuously differentiable and there exists some constant $L$ such that

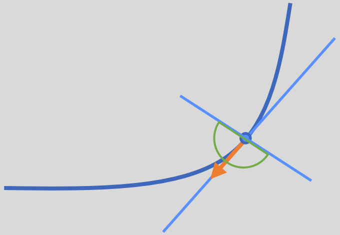
$$\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|, \qquad \forall x, x' \in \mathbb{R}^n$$

c) [Pseudogradient property] There exists a positive constant $c$ such that

$$c\|\nabla f(x_t)\|^2 \leq -\nabla f(x_t)^T \mathbb{E}[g_t | \mathcal{F}_t], \qquad \forall t$$

d) [Bounded variance] There exists positive constants $K_1, K_2$ s.t.

$$E[\|g_t\|^2 | \mathcal{F}_t] \leq K_1 + K_2\|\nabla f(x_t)\|^2, \qquad \forall t$$



Then, with probability 1, we have

1. The sequence $f(x_t)$ converges.
2. We have $\lim_{t\to\infty} \nabla f(x_t) = 0$.
3. Every limit point of $x_t$ is a stationary point of $f$.

Note: This holds for contractions w.r.t. the Euclidean norm.

We proved the convergence for the special case where $f(r) = \frac{1}{2}\|r - r^*\|_2^2$ for some $r^*$ (sufficient for Q-learning).