2022-10-06

Value-based reinforcement learning

All about “Q”

Cathy Wu

6.7950 Reinforcement Learning: Foundations and 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).
3. Daniela Pucci De Farias. MIT 2.997 Decision-Making in Large-Scale Systems. Spring 2004, Lecture 8.

Outline

1. Policy learning
2. Stochastic approximation of a fixed point

From exact DP to approximate DP

Types of approximation

- Model-free updates for policy evaluation (today)
  - Techniques: Monte Carlo approximation, temporal differencing
- Model-free updates for optimal value functions ["RL"]
  - e.g., Q-learning; technique: stochastic approximation
- Approximating value functions
  - E.g., Approximate VI / PI
- Finite sample approximation ["RL"]
  - E.g., Fitted Q iteration, DQN
- Approximating policies ["RL"]
  - E.g., Policy gradient methods

} Tabular methods

Outline

1. **Policy learning**
   a. State-action value function
   b. SARSA
   c. Q-Learning
   d. Preview of stochastic approximation of a fixed point
2. Stochastic approximation of a fixed point

Policy Learning

Learn optimal policy π^*

For $i = 1, \ldots, n$

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

Definition

In discounted infinite horizon problems, for any policy π, 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

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

State-Action Value Function Operators*

- $T^{\pi}Q(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) Q(s', \pi(s))$
- $TQ(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} Q(s', a')$

- *Abuse of notation for the operators

State-Action and State Value Function

- $Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^{\pi}(s')$
- $V^{\pi}(s) = Q^{\pi}\big(s, \pi(s)\big)$

- $Q^{*}(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^{*}(s')$
- $V^{*}(s) = Q^{*}\big(s, \pi^{*}(s')\big) = \max_{a \in A} Q^{*}(s,a)$

Q-value Iteration

Q-iteration:

1. Let Q_0 be any Q-function
2. At each iteration $k = 1, 2, \ldots, K$
   - Compute $Q_{k+1} = \mathrm{T}Q_k$
3. Return the greedy policy

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

Discuss: Why is it desirable to work with Q-value function, rather than state value function, when designing a model-free method?

Comparison with value iteration

- Bonus: computing the greedy policy from the Q-function does not require the MDP
- Increased space to O(SA), same time complexity at O(S2A)
- Reduced time complexity to compute the greedy policy O(SA)

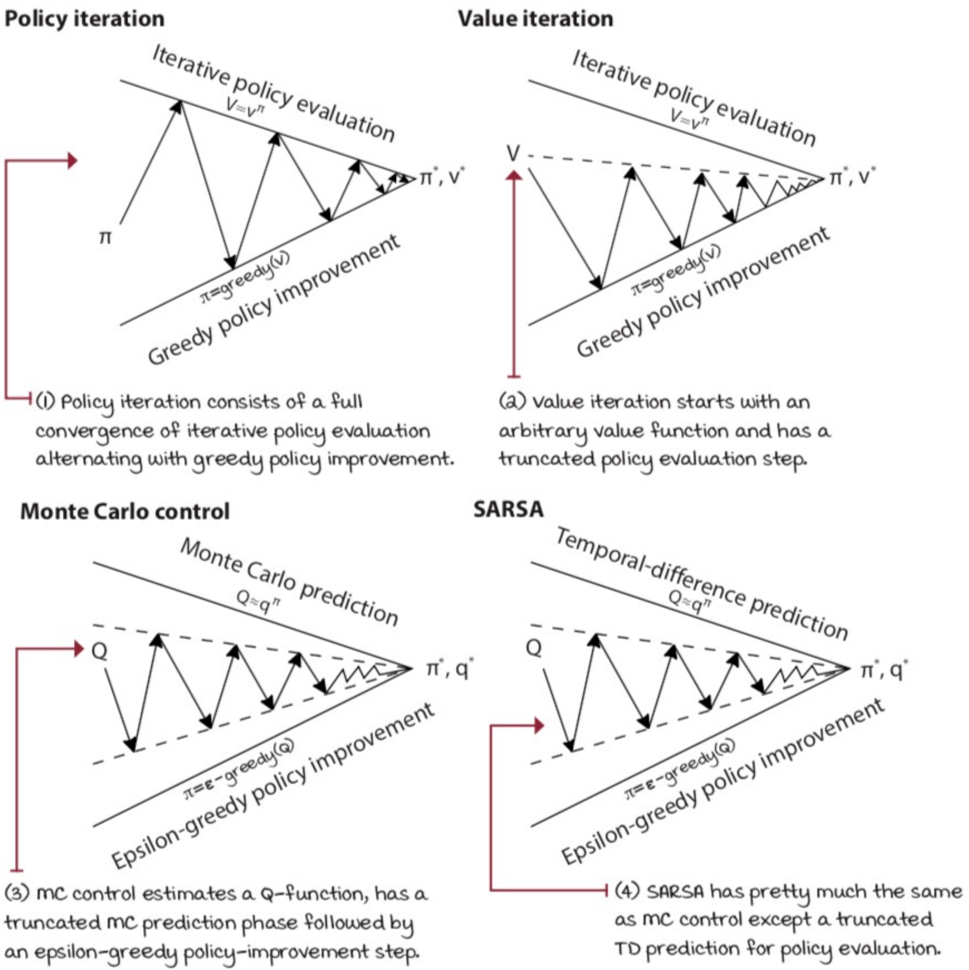
Policy Iteration (w/ Q-value function)

1. Let π_0 be any stationary policy
2. At each iteration $k = 1, 2, \dots, K$
   - Policy evaluation: given π_k, compute Q^{π_k}
   - Policy improvement: compute the greedy policy

$$\pi_{k+1}(s) \in \operatorname*{argmax}_{a \in A} Q_k^{\pi}(s, a)$$

3. Return the last policy π_K

Recall:


Policy iteration
Iterative policy evaluation
V=v^π
π
π*, v*
π=greedy(V)
Greedy policy improvement
(1) Policy iteration consists of a full convergence of iterative policy evaluation alternating with greedy policy improvement.
Value iteration
Iterative policy evaluation
V=v^π
V
π*, v*
π=greedy(V)
Greedy policy improvement
(2) Value iteration starts with an arbitrary value function and has a truncated policy evaluation step.
Monte Carlo control
Monte Carlo prediction
Q≈q^π
Q
π*, q*
π=ε-greedy(Q)
Epsilon-greedy policy improvement
(3) MC control estimates a Q-function, has a truncated MC prediction phase followed by an epsilon-greedy policy-improvement step.
SARSA
Temporal-difference prediction
Q≈q^π
Q
π*, q*
π=ε-greedy(Q)
Epsilon-greedy policy improvement
(4) SARSA has pretty much the same as MC control except a truncated TD prediction for policy evaluation.

SARSA

Idea: Alternate policy evaluation and policy improvement (both model-free!)

- Issue: greedy policy might not visit states needed to improve Q-value function
- Approach: Define a soft-max (random) exploratory policy with temperature τ

$$\pi_Q(a|s) = \frac{\exp\left(\frac{Q(s,a)}{\tau}\right)}{\sum_{a'} \exp\left(\frac{Q(s,a')}{\tau}\right)}$$

  The higher $Q(x, a)$, the more probability to take action a in state s.

- Compute the temporal difference on the trajectory $\langle s_t, a_t, r_t, s_{t+1}, a_{t+1} \rangle$ (with actions chosen according to $\pi_Q(a|s)$)

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

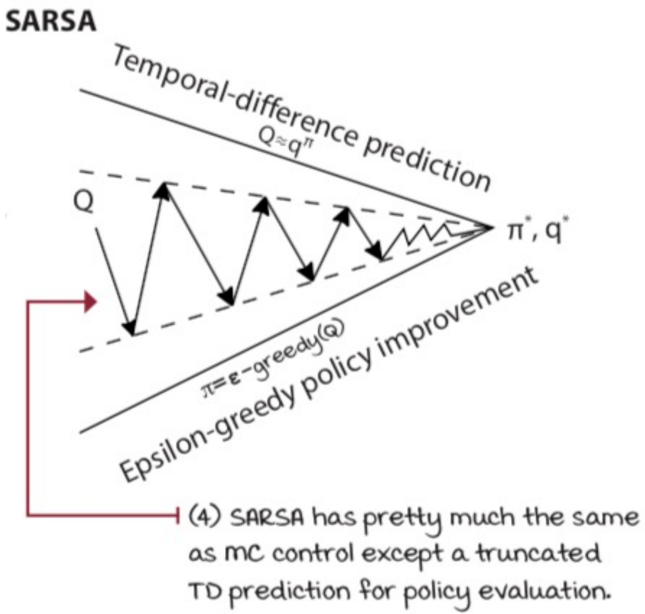
- Update the estimate of Q as

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

SARSA: Properties (Informal)

- The TD updates make $\hat{Q}$ converge to Q^{π}
- The update of π_Q allows improvement of the policy
- A decreasing temperature allows us to become more and more greedy

$\Longrightarrow$ If $\tau \to 0$ with a proper rate, then $\hat{Q} \to Q^*$ and $\pi_Q \to \pi^*$



SARSA: Limitations

The actions a_t need to be selected according to the current Q

$\Longrightarrow$ On-policy learning

The Optimal Bellman Equation

Proposition

The optimal value function Q^* (i.e. $Q^* = \max_{\pi} Q^{\pi}$) is the solution to the optimal Bellman equation:

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a' \in A} Q^*(s',a')$$

Q-Learning

Idea: Use TD for the optimal Bellman operator.

- Compute the (optimal) temporal difference on the trajectory $\langle s_t, a_t, r_t, s_{t+1} \rangle$ (with actions chosen arbitrarily!)

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

- Update the estimate of Q as

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

Q-Learning: Properties

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.

Learning the Optimal Policy

For $i = 1, \dots, n$

1. Set $t = 0$
2. Set initial state s_0
3. **While** (s_t not terminal)
    1. Take action a_t according to a suitable exploration policy
    2. Observe next state s_{t+1} and reward r_t
    3. Compute the temporal difference
       $$\delta_t = r_t + \gamma \hat{Q}(s_{t+1}, a_{t+1}) - \hat{Q}(s_t, a_t) \quad \text{(SARSA)}$$
       $$\delta_t = r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t) \quad \text{(Q−learning)}$$
    4. Update the Q-function
       $$\hat{Q}(s_t, a_t) = \hat{Q}(s_t, a_t) + \eta(s_t, a_t)\delta_t$$
    5. Set $t = t + 1$

   EndWhile

Endfor

How to avoid enumerating S?

Should you use Q-learning?

For: small state space problems

For: when you don't have P, r

Worth a try

Each iteration is extremely fast

✅ Step 1: Avoid this

Fitted Q iteration

(Ernst, 2005)

Step 2: Avoid this

Q-learning

(Watkins, 1992)

$$V_{i+1}(s) = \max_{a \in \mathcal{A}} r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [V_i(s')] \quad \text{for all } s$$

$$\mathcal{O}(KS^2A)$$

Wu

Stochastic Approximation of a Fixed Point

Definition

Let $\mathcal{T}: \mathbb{R}^N \to \mathbb{R}^N$ be a contraction in some norm $||\cdot||$ with fixed point V. For any function W and state s, a noisy observation $\hat{\mathcal{T}}V'(s) = \mathcal{T}V'(s) + w(s)$ is available. For any s $\in S = \{1, \ldots, N\}$, we defined the stochastic approximation:

$$V_{n+1}(s) = (1 - \eta_n(s))V_n(s) + \eta_n(s)\left(\hat{\mathcal{T}}V_n(s)\right)$$
$$= (1 - \eta_n(s))V_n(s) + \eta_n(s)(\mathcal{T}V_n(s) + w_n)$$

Where η_n is a sequence of learning steps.

Stochastic Approximation of a Fixed Point

Proposition

Let $\mathcal{F}_n = \{V_0, \dots, V_n, w_0, \dots, w_{n-1}, \eta_0, \dots, \eta_n\}$ the filtration of the algorithm and assume that:

$$\mathbb{E}[w_n(s) | \mathcal{F}_n] = 0 \quad and \quad \mathbb{E}[w_n^2(s) | \mathcal{F}_n] \leq A + B\|V_n\|^2$$

For constants A, B.

If the learning rates $\eta_n(s)$ are positive and satisfy the stochastic approximation conditions:

$$\sum_{n \geq 0} \eta_n = \infty \qquad \sum_{n \geq 0} \eta_n^2 = \infty$$

Then for any s $\in S$:

$$V_n(s) \xrightarrow{a.s.} V(s)$$

Outline

1. Policy learning
2. **Stochastic approximation of a fixed point**
   a. Overview
   b. Examples: TD(0), Q-learning
   c. Max norm contraction analysis
   d. (Quadratic) Lyapunov function analysis

Stochastic Approximation

- Stochastic approximation of a mean. Earlier: Wanted iterates μ_t to get closer and closer to some $\mu = \mathbb{E}[X]$, so that we could evaluate a policy using Monte Carlo samples.
- Stochastic approximation of a fixed point. Now, more generally: Want iterates x_t to get closer and closer to some fixed point x^* that is a solution to $H(x) = x$.
  - Application: Exploit the Bellman equation to evaluate a policy as soon as new information is available.
  - Application: Exploit optimal Bellman equation to improve a policy as soon as new information is available.
- 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:
$$x_{t+1}(s) = (1 - \eta_t)x_t(s) + \eta_t\big(H(x_t)(s) + w_t(s)\big) \quad \forall s \in \mathcal{S}$$

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 (Linear): $H(V) \coloneqq \mathcal{T}^\pi(V), \text{i.e.}\, V^\pi = \mathcal{T}^\pi V^\pi$

Example (Nonlinear): $H(V) \coloneqq \mathcal{T}(V), \text{i.e.}\, V^* = \mathcal{T} V^*$

Examples

Mean

$$H(x) \coloneqq \mathbb{E}[X] = \sum_{x'} p(x')x'$$

Stochastic gradient descent. Consider

$$H(x) \coloneqq x - \nabla f(x)$$

for some cost function f.

In this case, the system $H(x) = x$ is of the form $\nabla f(x) = 0$, which is closely related to finding the minimum of a convex function.

Possible algorithms:

- $x \coloneqq H(x)$
- $x \coloneqq (1-\eta)x + \eta H(x)$ (small steps version)
- $x \coloneqq (1-\eta)x + \eta(H(x) + w)$ (since $H(x)$ is not precisely known; this is a stochastic approximation algorithm)

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 ($\|\cdot\|_\infty$)

- As discussed earlier.
- Enables analysis of TD, Q-learning.

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

- Use where the expected update directions at each iteration are descent directions corresponding to a smooth potential (or Lyapunov) function.
- Enables analysis of the mean, stochastic gradient descent, and $TD(\lambda)$ with linear approximation.
- Above analysis uses this analysis as a sub-routine!

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

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

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.

- Related result for contractions w.r.t. the Euclidean norm (later)

Discuss

Why do we need these extra assumptions on noise?

Why not just apply the law of large numbers for the noise term $w_t(s)$?

Example for max norm: First Visit $TD(0)$

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

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

With temporal difference $\delta_t(s)$

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

Need to show assumptions for Prop. 4.4 are met.

(Condition b) Equivalently (construct w_t s.t. it is zero mean):

$$V_{t+1}(s) = (1 - \eta_t) V_t(s) + \eta_t \Big(\underbrace{\mathbb{E}[\delta_t(s)] + V_t(s)}_{H(V_t)(s)} \Big) + \eta_t \big(\underbrace{\delta_t(s) - \mathbb{E}[\delta_t(s)]}_{w_t(s)} \big)$$

Thus,

$$\mathbb{E}[w_t(s) | \mathcal{F}_t] = 0, \qquad \forall s, t$$

Example for max norm: First Visit $TD(0)$

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

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

With temporal difference $\delta_t(s)$

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

Need to show assumptions for Prop. 4.4 are met.

(Condition c) Need to confirm that $TD(0)$ has bounded variance.
Recall: $TD(0)$ is low variance (but high bias).

$$\mathbb{V}(\delta_t(s) - \mathbb{E}[\delta_t(s)] | \mathcal{F}_t) = \mathbb{V}(\delta_t(s) | \mathcal{F}_t)$$

$$\begin{aligned}\mathbb{V}(\delta_t(s) | \mathcal{F}_t) &\leq \mathbb{E}\left[\left(r(s, s') + \gamma V_t(s') - V_t(s)\right)^2 \middle| \mathcal{F}_t\right] \\ &\leq (r_{\max} + 2\|V_t\|_\infty)^2 \\ &\leq 3r_{\max}^2 + 6\|V_t\|_\infty^2\end{aligned}$$

Since $2xy \leq x^2 + y^2$.

Recall: Q-Learning

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

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

- 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 $s, a \in S \times A$

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

Example for max norm: Q-learning

- $\hat{Q}(x_t, a_t) = \hat{Q}(s_t, a_t) + \eta(s_t, a_t)\delta_t$
- $\hat{Q}(x_t, a_t) = \hat{Q}(s_t, a_t) + \eta(s_t, a_t)(r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t))$
- Let $M_{(s,a)}(Q) = \mathbb{E}\left[r(s,a) + \gamma \max_{a'} Q(s_{t+1}, a') \middle| \mathcal{F}_t\right]$
  - $\hat{Q}(x_t, a_t) = \hat{Q}(s_t, a_t) + \eta(s_t, a_t)\underbrace{(M_{(s_t,a_t)}(\hat{Q}) - \hat{Q}(s_t, a_t)}_{H(Q_t)(s)} + \underbrace{r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a') - M_{(s_t,a_t)}(\hat{Q})}_{w_t(s)})$
- (condition b) $\mathbb{E}[w_t(s) | \mathcal{F}_t] = 0, \quad \forall s, t$
- (condition c) Similar to previous example
- (condition d) Know how to show this
- (condition a) We choose this

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.

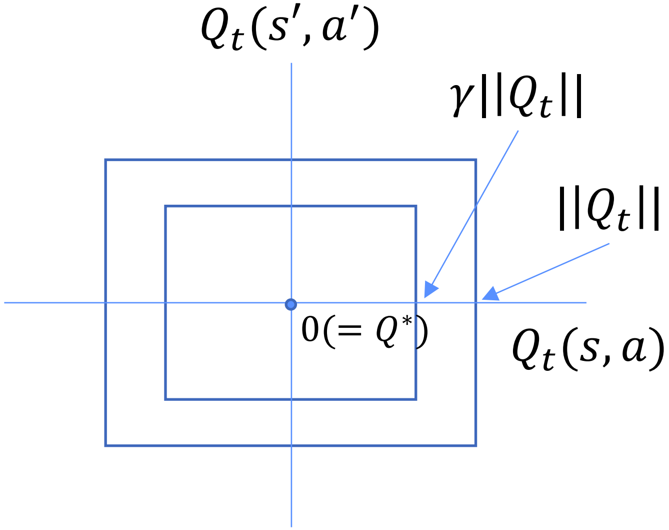
Then, x_t converges to x^* with probability 1.

- Related result for contractions w.r.t. the Euclidean norm (later)

Proof: Max Norm Contraction Analysis (Prop 4.4)

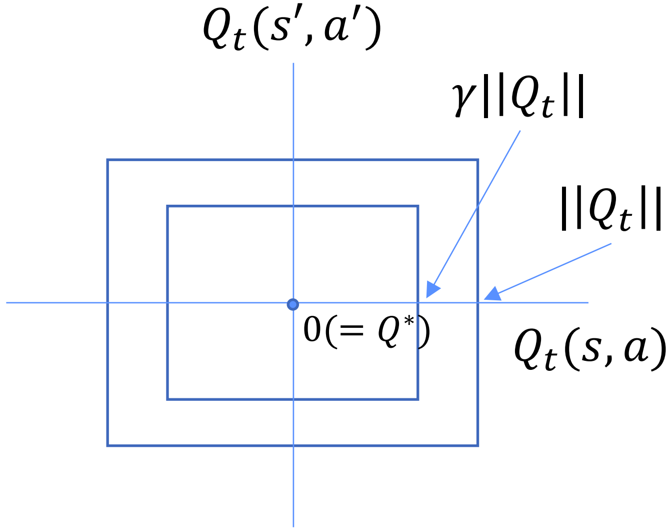
Sketch:

- 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.
- The deterministic component contracts as expected in due time.
- The noise component goes to 0 w.p. 1.
- Therefore, the overall x_t contracts.



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. Can think of D_k as upper bound on
  $$\mathbb{E}[r(s,s') + \gamma V_t(s')]$$
- 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 t_{k+1} s.t.
  $$\|x_t\|_\infty \leq D_{k+1}, \forall t \geq t_{k+1}$$



Proof: Max Norm Contraction Analysis (Prop 4.4)

- For the stochastic part of the upper bound, define:

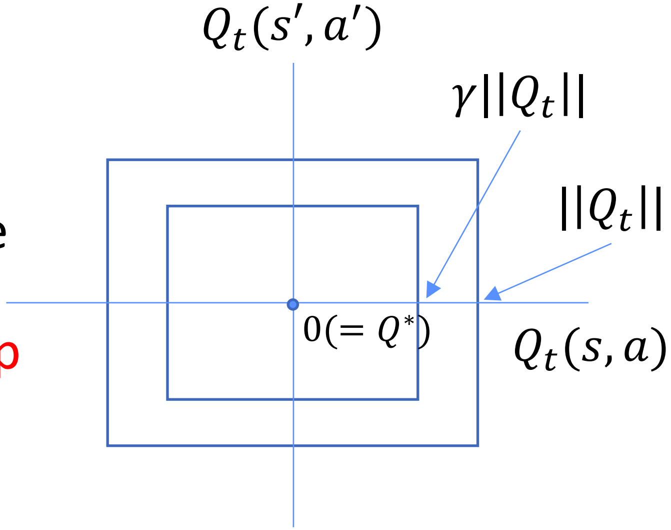
$$W_0(s) = 0;$$
$$W_{t+1}(s) = (1 - \eta_t)W_t(s) + \eta_t w_t(s)$$

- 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 next*),

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

a.s.

- That is, the noise averages out to zero.



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 γD_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. $\|x_t\| \leq D_{k+1}, \forall t \geq t_{k+1}$. ∎

Deterministic-only upper bound

Corresponds to convergence analysis for asynchronous value iteration!

Q-learning as noisy extension of value iteration.

Now for the noise

The remainder of the discussion is about noise.

We used two not-yet-justified tools:

1. Supermartingale Convergence Theorem
2. Lyapunov Function Analysis (NDP Prop 4.1)

Proof: Max Norm Contraction Analysis (Prop 4.4)

- For the stochastic part of the upper bound, define:

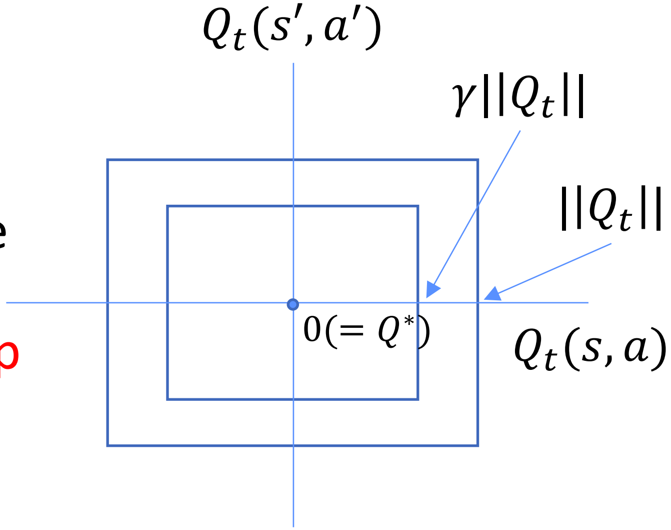
$$W_0(s) = 0;$$
$$W_{t+1}(s) = (1 - \eta_t)W_t(s) + \eta_t w_t(s)$$

- 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),

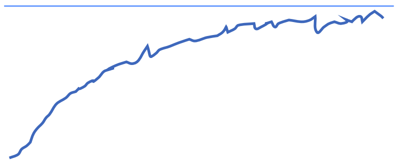
$$\lim_{t \to \infty} W_t(s) = 0$$

a.s.

- That is, the noise averages out to zero.

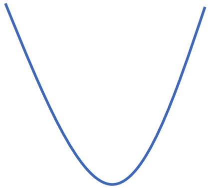


To Complete the Max Norm Analysis



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

- $W_t(s)$ turns out to be martingale noise.
- Martingale noise corresponds to a stochastic Lyapunov function.
- Consequently, martingale noise averages out over time to zero.
- Interpretation: $\{W_t(s)\}$ as stochastic gradient descent along a quadratic Lyapunov function



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

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

- Take Lyapunov function $f(x) = x^2$ (for noise terms W).
- Show that $W_t(s) \to 0$.

Quadratic Lyapunov function (special case of Prop 4.1)

Proposition

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

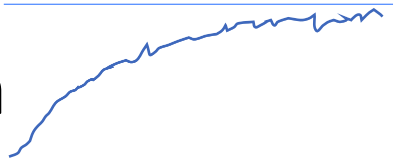
1. [Pseudogradient property] $\exists c$ such that $cf(x_t) \leq -\nabla f(x_t)^T \mathbb{E}[g_t | \mathcal{F}_t]$
2. [Bounded variance] $\exists K_1, K_2$ such that $\mathbb{E}[\|g_t\|_2^2 | \mathcal{F}_t] \leq K_1 + K_2 f(x_t)$

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

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

- 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)).

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

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

Then,

1. Y_t converges to a limit with probability 1,
2. $\sum_{t=1}^{\infty} X_t < \infty$

Correspondence to noise upper bound

(intuition)

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

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

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.

- $$E[f(x_{t+1})|\mathcal{F}_t] = E\left[\frac{1}{2}\|x_{t+1} - r^*\|_2^2 \middle| \mathcal{F}_t\right]$$
$$= E\left[\frac{1}{2}(x_t + \eta_t g_t - r^*)^T(x_t + \eta_t g_t - r^*)\middle|\mathcal{F}_t\right] \quad (g_t \triangleq g(x_t, w_t))$$
$$= \frac{1}{2}(x_t - r^*)^T(x_t - r^*) + \eta_t(x_t - r^*)^T E[g_t|\mathcal{F}_t] + \frac{\eta_t^2}{2}E[g_t^T g_t|\mathcal{F}_t]$$
- Since $f(x_t) = \frac{1}{2}\|x_t - r^*\|_2^2, \nabla f(x_t) = x_t - r^*$. Then:
- $$E[f(x_{t+1})|\mathcal{F}_t] = f(x_t) + \eta_t(x_t - r^*)^T E[g_t|\mathcal{F}_t] + \frac{\eta_t^2}{2}E[\|g_t\|_2^2|\mathcal{F}_t]$$
$$= f(x_t) + \eta_t \nabla f(x_t)^T E[g_t|\mathcal{F}_t] + \frac{\eta_t^2}{2}E[\|g_t\|_2^2|\mathcal{F}_t]$$
$$\le f(x_t) - \eta_t c f(x_t) + \frac{\eta_t^2}{2}(K_1 + K_2 f(x_t))$$
$$\le \underbrace{f(x_t)}_{Y_t} - \underbrace{\left(\eta_t c - \frac{\eta_t^2 K_2}{2}\right) f(x_t)}_{X_t} + \underbrace{\frac{\eta_t^2}{2}K_1}_{Z_t} \qquad \text{(condition b)}$$
- Since $\eta_t > 0$ and $\sum_{t=0}^{\infty}\eta_t^2 < \infty$, η_t must converge to zero, and $X_t \ge 0$ for all large enough t.

Proof: quadratic Lyapunov function

- Moreover: $\sum_{t=0}^{\infty} Z_t = \frac{K_1}{2} \sum_{t=0}^{\infty} \eta_t^2 < \infty$ (condition c)
- Therefore, by Supermartingale convergence theorem:

$$f(x_t) \text{ converges w.p. 1, and}$$

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

- 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:

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

■

(General) Lyapunov Function Analysis Setup

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

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

Slight re-write:

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

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

Lyapunov Function Analysis (Prop 4.1)

Proposition

Let x_t be the sequence generated by the iteration

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

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 ∇f] The function f is continuously differentiable and there exists some constant L such that

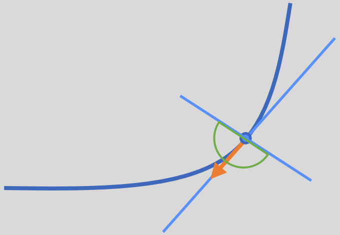
$$\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|, \qquad \forall x, x' \in \mathbb{R}^n$$

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 will prove the convergence for a special case where $f(r) = \frac{1}{2}\|r - r^*\|_2^2$ for some r^*.

Summary of Q-learning analysis

- Apply Max Norm Convergence
- Via
  - Supermartingale Convergence Theorem
  - Lyapunov Function Analysis (Prop 4.1)
    - Special case: Quadratic Lyapunov function

Summary

- Policy learning: SARSA and Q-learning (definition, guarantees)
- Stochastic approximation of fixed points (results, contractive norms, analyses)
- TD and Q-learning as stochastic approximation methods