UNIVERSITY OF TWENTE.

Convergence proof for Q-learning

Anne Zander

Value updates in RL algorithms

(Tabular, value-based) RL algorithms update value estimates of states/state-action pairs:

$$\textit{NewEstimate} \leftarrow \textit{OldEstimate} + \textit{Stepsize}[\textit{Target} - \textit{OldEstimate}].$$

Target is some noisy value estimate.

We want to show: $\textit{Estimates} \rightarrow \textit{Optimal Value Function}$

Example Q-learning:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_t [R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

Overview

A stochastic approximation scheme

An asynchronous stochastic approximation scheme

Stochastic fixed point iterations for contractions

Example Q-learning

Summary and outlook

References

Back-up

A stochastic approximation scheme

Stochastic approximation scheme in $\mathbb{R}^d$:

$$x_{n+1} = x_n + a(n)\left[h\left(x_n\right) + M_{n+1}\right], n \geq 0 \tag{1}$$

with prescribed $x_0 \in \mathbb{R}^d$, (small) positive stepsizes $a(n) \in \mathbb{R}_+$, $h : \mathbb{R}^d \to \mathbb{R}^d$, ”zero-mean” random vectors M_n.

Motivation

Euler method: First-order numerical procedure for solving ordinary differential equations (ODEs).

$$\dot{x}(t) = h(x(t)), x(0) = x_0.$$

Set $t_n = a \cdot n$ with step size a.

Then approximate $x(t_{n+1})$ with $x_{n+1} = x_n + ah(x_n)$.

ODE approach

Limiting ODE which (1) might track asymptotically:

$$\dot{x}(t) = h(x(t)), t \geq 0. \tag{2}$$

Idea: Construct continuous interpolated trajectory of $\{x_n\}$, and show that it asymptotically approaches a solution of (2).

Then, e.g., showing that (2) has a globally asymptotically stable point (a root of h) shows the convergence of the iterates, too.

Assumptions

(A1) The map $h : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz:

$$\|h(x) - h(y)\| \leq L\|x - y\| \text{ for some } 0 < L < \infty.$$

(A2) Step sizes $\{a(n)\}$ are positive scalars satisfying

$$\sum_n a(n) = \infty, \sum_n a(n)^2 < \infty.$$

Assumptions continued

(A3) $\{M_n\}$ is a martingale difference sequence with respect to the increasing family of σ-fields $\mathcal{F}_n = \sigma(x_m, M_m, m \leq n)$:

$$E[M_{n+1} \mid \mathcal{F}_n] = 0 \text{ almost surely (a.s.)}, n \geq 0.$$

Furthermore:

$$E\left[\|M_{n+1}\|^2 \mid \mathcal{F}_n\right] \leq K\left(1 + \|x_n\|^2\right) \text{ a.s., } n \geq 0 \text{ for some } K > 0 \quad (3)$$

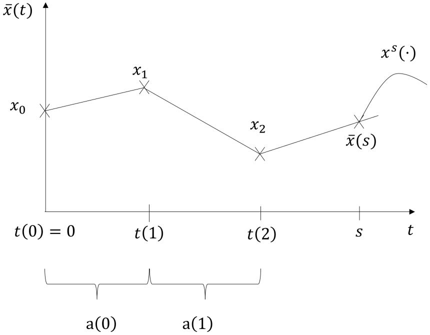
Assumptions continued

(A4) The iterates of (1) remain bounded a.s., i.e.,

$$\sup_n \|x_n\| < \infty, \text{ a.s.}$$

This assumption is in general not easy to establish. Later, we will come back to it.

(1D) Continuous interpolation



Let $x^s(t), t \geq s$ be the solution to (2) starting at s:

$$\dot{x}^s(t) = h\left(x^s(t)\right), t \geq s, x^s(s) = \bar{x}(s), s \in \mathbb{R}.$$

Interpolated trajectory converges to ODE solution

Lemma 1 (Lemma 2.1 in [Borkar, 2023])

Given (A1)- (A4), for any $T > 0$,

$$\lim_{s\to\infty} \sup_{t\in[s,s+T]} \left\|\bar{x}(t) - x^s(t)\right\| = 0, \ \textit{a.s.}$$

Convergence illustrated

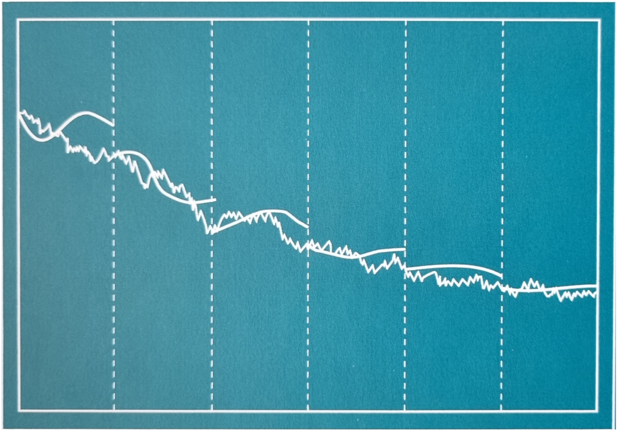


Figure: Taken from [Borkar, 2008]

Stability of the iterates

(A5) $h_c(x) = h(cx)/c, c \geq 1, x \in \mathcal{R}^d$, satisfy $h_c(x) \to h_\infty(x)$ as $c \to \infty$, uniformly on compacts for some $h_\infty \in C\left(\mathcal{R}^d\right)$.

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

has origin as unique globally asymptotically stable equilibrium.

Theorem 2 (Theorem 4.1 in [Borkar, 2023])

Under (A1) - (A3) and (A5), we have (A4): $\sup_n \|x_n\| < \infty,$ *a.s.*

Overview

A stochastic approximation scheme

An asynchronous stochastic approximation scheme

Stochastic fixed point iterations for contractions

Example Q-learning

Summary and outlook

References

Back-up

An asynchronous scheme

Only one component i of x_n is updated each iteration:

$$x_{n+1}(i) = x_n(i) + a(\nu(i,n))\mathbb{1}_{Y_n=i}\left[h_i\left(x_n\right) + M_{n+1}(i)\right], n \geq 0, \tag{4}$$

where Y_n is a RV on $\{1, \ldots, d\}$ and $\nu(i,n) = \sum_{m=0}^{n} \mathbb{1}_{Y_m=i}$.

We need $\nu(i,n) \to \infty$ for $n \to \infty$ for all components i to obtain (random) stepsizes that fulfill (A2)

Asynchronous limiting ODE

Define the continuous interpolation $\bar{x}(\cdot)$ as before.

Stepsizes $a(n)$ become very small and the ODE only sees averaged values of how often a component is updated.

Theorem 3 (Theorem 6.1 in [Borkar, 2023])

If (A1) - (A4) hold and $\{Y_n\}$ is an ergodic Markov chain with stationary distribution π, $\bar{x}$ tracks the ODE

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

where $\Lambda = diag(\pi_1, \ldots, \pi_d)$ with $\pi_i > 0$ for all i. Hence, $I \geq \Lambda \geq \epsilon I$ for all $t > 0$ for some positive ϵ.

Overview

A stochastic approximation scheme

An asynchronous stochastic approximation scheme

Stochastic fixed point iterations for contractions

Example Q-learning

Summary and outlook

References

Back-up

Contractions

Let $h(x) = F(x) - x$ for some contraction $F(\cdot)$ in the max-norm, i.e.,

$$\|F(x) - F(y)\|_\infty \leq \beta \|x - y\|_\infty \quad \forall x, y$$

and for some $\beta \in [0, 1) \Rightarrow (A1)$.

Due to the contraction mapping theorem there is a unique root x^* of h which is equal to the fixed point of F.

Convergence

Theorem 4 (Theorem 12.1 in [Borkar, 2023])

x^* is a globally asymptotically stable point of the ODE.

Corollary 5

The iterates x_n converge to x^* if (A2), (A3) and (A4)/(A5) are satisfied.

These results also transfer to the asynchronous case.

Overview

A stochastic approximation scheme

An asynchronous stochastic approximation scheme

Stochastic fixed point iterations for contractions

Example Q-learning

Summary and outlook

References

Back-up

Q-learning

Asynchronous update for Q-learning:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t \mathbb{1}(S_t = s, A_t = a)\left[R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) - Q_t(s,a)\right]$$

$$\begin{aligned}
& R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) - Q_t(s,a) \\
= \quad & \mathbb{E}(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) \mid S_t = s, A_t = a) - Q_t(s,a) \\
& + R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) - \mathbb{E}(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) \mid S_t = s, A_t = a).
\end{aligned}$$

Q-learning

We define h and M componentwise as

$$
\begin{aligned}
&(h(Q_t))(s,a) \\
=&\mathbb{E}(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) - Q_t(s,a)) \mid S_t = s, A_t = a) - Q_t(s,a) \\
=&(B' Q_t)(s) - Q_t(s),
\end{aligned}
$$

and

$$
\begin{aligned}
&M_{t+1}(s,a) \\
=&R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) - \mathbb{E}(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) \mid S_t = s, A_t = a)
\end{aligned}
$$

with Bellman optimality operator B'. Hence, we have that $h(Q) = B'Q - Q$, where B' is a contraction in the max-norm with unique fixed point q_* (optimal Q-function). $\Rightarrow Q_t \to q_*$ if (A2), (A3), (A4)/(A5) hold and if we see state and action pairs according to a ergodic Markov Chain.

Q-learning

Assume (A2). Show (A3) and (A5).

(A3): We first need to show that the noise term has zero mean given the history.

$$
\begin{aligned}
&\mathbb{E}[M_{t+1}(s,a) \mid \mathcal{F}_t] = \mathbb{E}[M_{t+1}(s) \mid Q_t, S_t = s, A_t = a] \\
=&\mathbb{E}[R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) \\
&- \mathbb{E}(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) \mid S_t = s, A_t = a) \mid Q_t, S_t = s, A_t = a] \\
=&\mathbb{E}(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) \mid Q_t, S_t = s, A_t = a) \\
&- \mathbb{E}(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) \mid Q_t, S_t = s, A_t = a) \\
=&0
\end{aligned}
$$

Q-learning

Show $\mathbb{E}[\|M_{t+1}\|_\infty^2 \mid \mathcal{F}_t] \leq K(1 + \|Q_t\|_\infty^2)$ componentwise, assuming bounded rewards:

$$
\begin{aligned}
&\mathbb{E}[|M_{t+1}(s,a)|^2 \mid \mathcal{F}_t] = \mathbb{E}[|M_{t+1}(s,a)|^2 \mid Q_t, S_t = s, A_t = a] \\
=&\mathbb{E}[(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b) - \mathbb{E}(R_{t+1} + \gamma \max_b Q_t(S_{t+1}, b)))^2] \\
=&\mathbb{E}[(R_{t+1} - \mathbb{E}(R_{t+1}))^2 + 2\gamma(R_{t+1} - \mathbb{E}(R_{t+1}))(\max_b Q_t(S_{t+1}, b) \\
&\quad - \mathbb{E}(\max_b Q_t(S_{t+1}, b))) \\
&\quad + \gamma^2(\max_b Q_t(S_{t+1}, b) - \mathbb{E}(\max_b Q_t(S_{t+1}, b)))^2] \\
\leq& C_1 + C_2\|Q_t\|_\infty + C_3\|Q_t\|_\infty^2 \\
\leq& C_1 + C_2\left(1 + \|Q_t\|_\infty^2\right) + C_3\|Q_t\|_\infty^2 \\
=& C_4 + C_5\|Q_t\|_\infty^2 \\
\leq& \max(C_4, C_5)(1 + \|Q_t\|_\infty^2)
\end{aligned}
$$

Q-learning

(A5):

$$\begin{aligned}(h_c(Q_t))(s,a) &= \left(\frac{h(cQ_t)}{c}\right)(s,a) = \left(\frac{B'(cQ_t)}{c}\right)(s,a) - \left(\frac{cQ_t}{c}\right)(s) \\ &= \mathbb{E}\left(\frac{R_{t+1}}{c} + \gamma \max_b Q_t(S_{t+1},b) \mid S_t = s, A_t = a\right) - Q_t(s,a) \\ &\longrightarrow \mathbb{E}\left(\gamma \max_b Q_t(S_{t+1},b) \mid S_t = s, A_t = a\right) - Q_t(s,a) \text{ for } c \to \infty\end{aligned}$$

$$\Longrightarrow (h_\infty(Q_t))(s,a) = \mathbb{E}\left(\max_b Q_t(S_{t+1},b) \mid S_t = s, A_t = a\right) - Q_t(s,a)$$

$h_\infty(Q) = B'Q - Q$ in an environment with zero rewards $\Rightarrow$ ODE $\dot{Q} = h_\infty(Q)$ has $Q = 0$ as its unique global asymptotical stable equilibrium.

Overview

A stochastic approximation scheme

An asynchronous stochastic approximation scheme

Stochastic fixed point iterations for contractions

Example Q-learning

Summary and outlook

References

Back-up

Summary

- Stochastic approximation scheme (SAS) $\approx$ noisy Euler method
- Synchronous and asynchronous schemes
- Convergence of a SAS
  - Tacking of ODE (fulfill assumptions)
  - Stability of the ODE (investigate h)
- Stochastic fixed point iterations for contractions
- Example Q-learning

Outlook

- Averaging over the natural time scale
- Stochastic gradient schemes
- Two time scales
- Projected schemes
- ...

Overview

A stochastic approximation scheme

An asynchronous stochastic approximation scheme

Stochastic fixed point iterations for contractions

Example Q-learning

Summary and outlook

References

Back-up

References

Borkar, V. S. (2008).
Stochastic Approximation A Dynamical Systems Viewpoint.
Hindustan Book Agency, Gurgaon, 1 edition.

Borkar, V. S. (2023).
Stochastic Approximation: A Dynamical Systems Viewpoint, volume 48.
Springer Nature Singapore.

Overview

A stochastic approximation scheme

An asynchronous stochastic approximation scheme

Stochastic fixed point iterations for contractions

Example Q-learning

Summary and outlook

References

Back-up

Martingale convergence

Let

$$\zeta_n = \sum_{m=0}^{n-1} a(m) M_{m+1}, n \geq 1.$$

By (A3) $(\zeta_n, \mathcal{F}_n)$, $n \geq 1$, is a zero-mean, square-integrable martingale. By (A2), (A3) and (A4),

$$\begin{aligned}\sum_{n\geq 0} E\left[\|\zeta_{n+1} - \zeta_n\|^2 \mid \mathcal{F}_n\right] &= \sum_{n\geq 0} a(n)^2 E\left[\|M_{n+1}\|^2 \mid \mathcal{F}_n\right] \\ &\leq \sum_{n\geq 0} a(n)^2 K\left(1 + \|x_n\|^2\right) < \infty, \quad \text{a.s.}\end{aligned}$$

Martingale convergence theorem: ζ_n converges a.s., $n \to \infty$.

Additional results

Chapter 7 in [Borkar, 2023]: Under some additional assumptions and if the ODE (2) has a unique globally asymptotically stable equilibrium point the asymptotic convergence rate of the iterates to that point is $O(\sqrt{a(n)})$.

Chapter 9 in [Borkar, 2023]: For a small constant stepsize $a \in (0, 1)$, we often can replace 'converges a.s. to' with 'concentrates with high probability in the neighbourhood of'.