CS 747, Autumn 2023: Lecture 14

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2023

Reinforcement Learning

1. Prediction with Monte Carlo methods

2. On-line implementation

Reinforcement Learning

1. Prediction with Monte Carlo methods

2. On-line implementation

Prediction

- Assume we have an episodic task. $S = \{s_1, s_2, s_3\}$, $\gamma = 1$.
  On each episode, start state picked uniformly at random.

Prediction

- Assume we have an episodic task. $S = \{s_1, s_2, s_3\}$, $\gamma = 1$.
  On each episode, start state picked uniformly at random.

- Here are the first 5 episodes.

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Prediction

- Assume we have an episodic task. $S = \{s_1, s_2, s_3\}$, $\gamma = 1$.
  On each episode, start state picked uniformly at random.

- Here are the first 5 episodes.

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

- What is your estimate of V^π (call it $\hat{V}^5$)?

Prediction

- Assume we have an episodic task. $S = \{s_1, s_2, s_3\}$, $\gamma = 1$.
  On each episode, start state picked uniformly at random.

- Here are the first 5 episodes.

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

- What is your estimate of V^π (call it $\hat{V}^5$)?
  Monte Carlo (MC) methods estimate based on sample averages.

Defining Relevant Quantities

- For $s \in S$, $i \geq 1, j \geq 1$, let
  - $\mathbf{1}(s, i, j)$ be 1 if s is visited at least j times on episode i (else $\mathbf{1}(s, i, j) = 0$), and
  - $G(s, i, j)$ be the discounted long-term reward starting from the j-th visit of s on episode i,
  - Taking $G(s, i, j) = 0$ if $\mathbf{1}(s, i, j) = 0$; also $0/0 = 0$.

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Defining Relevant Quantities

- For $s \in S$, $i \geq 1, j \geq 1$, let
  - $\mathbf{1}(s, i, j)$ be 1 if s is visited at least j times on episode i (else $\mathbf{1}(s, i, j) = 0$), and
  - $G(s, i, j)$ be the discounted long-term reward starting from the j-th visit of s on episode i,
  - Taking $G(s, i, j) = 0$ if $\mathbf{1}(s, i, j) = 0$; also $0/0 = 0$.

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

- $\mathbf{1}(s_1, 1, 1) = 1$, $G(s_1, 1, 1) = 5 + \gamma \cdot 2 + \gamma^2 \cdot 3 + \gamma^3 \cdot 1 = 11$.
- $\mathbf{1}(s_1, 1, 3) = 0$.
- $\mathbf{1}(s_2, 5, 1) = 1$, $G(s_2, 5, 1) = 3 + \gamma \cdot 3 + \gamma^2 \cdot 1 = 7$.
- $\mathbf{1}(s_2, 5, 2) = 1$, $G(s_2, 5, 2) = 3 + \gamma \cdot 1 = 4$.

Some Standard Estimates of $V^\pi(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Let $\hat{V}^N$ denote estimate after N *episodes*.

First-visit MC: Average the G's of every first occurrence of s in an episode.

$$\hat{V}^N_{\text{First-visit}}(s) = \frac{\sum_{i=1}^N G(s, i, 1)}{\sum_{i=1}^N \mathbf{1}(s, i, 1)}.$$

Some Standard Estimates of $V^{\pi}(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Let $\hat{V}^N$ denote estimate after N *episodes*.

First-visit MC: Average the G's of every first occurrence of s in an episode.

$$\hat{V}^N_{\text{First-visit}}(s) = \frac{\sum_{i=1}^{N} G(s, i, 1)}{\sum_{i=1}^{N} \mathbf{1}(s, i, 1)}.$$

$$\text{Hence } \hat{V}^5_{\text{First-visit}}(s_2) = \frac{4 + 7 + 8 + 7}{4} = 6.5.$$

Some Standard Estimates of $V^{\pi}(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Let $\hat{V}^N$ denote estimate after N *episodes*.

Every-visit MC: Average the G's of every occurrence of s in an episode.

$$\hat{V}^N_{\text{Every-visit}}(s) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{\infty} G(s, i, j)}{\sum_{i=1}^{N} \sum_{j=1}^{\infty} \mathbf{1}(s, i, j)}.$$

Some Standard Estimates of $V^\pi(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Let $\hat{V}^N$ denote estimate after N *episodes*.

Every-visit MC: Average the G's of every occurrence of s in an episode.

$$\hat{V}^N_{\text{Every-visit}}(s) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{\infty} G(s,i,j)}{\sum_{i=1}^{N} \sum_{j=1}^{\infty} \mathbf{1}(s,i,j)}.$$

$$\text{Hence } \hat{V}^5_{\text{Every-visit}}(s_2) = \frac{(4+1)+(7+1)+8+(7+4)}{7} \approx 4.57.$$

Some Not-so-standard Estimates of $V^\pi(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Let $\hat{V}^N$ denote estimate after N *episodes*.

Second-visit MC: Average the G's of every second occurrence of s in an episode.

$$\hat{V}^N_{\text{Second-visit}}(s) = \frac{\sum_{i=1}^{N} G(s, i, 2)}{\sum_{i=1}^{N} \mathbf{1}(s, i, 2)}.$$

Some Not-so-standard Estimates of $V^{\pi}(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Let $\hat{V}^N$ denote estimate after N *episodes*.

Second-visit MC: Average the G's of every second occurrence of s in an episode.

$$\hat{V}^N_{\text{Second-visit}}(s) = \frac{\sum_{i=1}^{N} G(s, i, 2)}{\sum_{i=1}^{N} \mathbf{1}(s, i, 2)}.$$

$$\text{Hence } \hat{V}^5_{\text{Second-visit}}(s_2) = \frac{1+1+4}{3} = 2.$$

Some Not-so-standard Estimates of $V^\pi(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Let $\hat{V}^N$ denote estimate after N *episodes*.

Last-visit MC: Average the G's of every last occurrence of s in episode i (assume $times(s, i)$ visits).

$$\hat{V}^N_{\text{Last-visit}}(s) = \frac{\sum_{i=1}^{N} G(s, i, times(s, i))}{\sum_{i=1}^{N} \mathbf{1}(s, i, times(s, i))}.$$

Some Not-so-standard Estimates of $V^\pi(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_3, 1, s_\top$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_\top$

Let $\hat{V}^N$ denote estimate after N *episodes*.

Last-visit MC: Average the G's of every last occurrence of s in episode i (assume $times(s, i)$ visits).

$$\hat{V}^N_{\text{Last-visit}}(s) = \frac{\sum_{i=1}^{N} G(s, i, times(s, i))}{\sum_{i=1}^{N} \mathbf{1}(s, i, times(s, i))}.$$

Hence $\hat{V}^5_{\text{Last-visit}}(s_2) = \dfrac{1+1+8+4}{4} = 3.5.$

Question

- Recall that we generate N episodes.
- Which claims below are true?

$$\lim_{N \to \infty} \hat{V}^N_{\text{First-visit}} = V^\pi.$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Every-visit}} = V^\pi.$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Second-visit}} = V^\pi.$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Last-visit}} = V^\pi.$$

Question

- Recall that we generate N episodes.
- Which claims below are true?

$$\lim_{N \to \infty} \hat{V}^N_{\text{First-visit}} = V^\pi. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Every-visit}} = V^\pi.$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Second-visit}} = V^\pi.$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Last-visit}} = V^\pi.$$

Question

- Recall that we generate N episodes.
- Which claims below are true?

$$\lim_{N \to \infty} \hat{V}^{N}_{\text{First-visit}} = V^{\pi}. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^{N}_{\text{Every-visit}} = V^{\pi}. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^{N}_{\text{Second-visit}} = V^{\pi}.$$

$$\lim_{N \to \infty} \hat{V}^{N}_{\text{Last-visit}} = V^{\pi}.$$

Question

- Recall that we generate N episodes.
- Which claims below are true?

$$\lim_{N \to \infty} \hat{V}^N_{\text{First-visit}} = V^\pi. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Every-visit}} = V^\pi. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Second-visit}} = V^\pi. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Last-visit}} = V^\pi.$$

Question

- Recall that we generate N episodes.
- Which claims below are true?

$$\lim_{N \to \infty} \hat{V}^N_{\text{First-visit}} = V^\pi. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Every-visit}} = V^\pi. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Second-visit}} = V^\pi. \text{ True.}$$

$$\lim_{N \to \infty} \hat{V}^N_{\text{Last-visit}} = V^\pi. \text{ False.}$$

Reinforcement Learning

1. Prediction with Monte Carlo methods
2. On-line implementation

First-visit MC Again

- Assume episodic task with $S = \{s_1, s_2, s_3\}$; following π.
- Say we start each episode with state s (for illustration s_2).

Episode 1: $s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_2, 3, s_2, 3, s_1, 1, s_\top$

First-visit MC Again

- Assume episodic task with S = $\{s_1, s_2, s_3\}$; following π.
- Say we start each episode with state s (for illustration s_2).

Episode 1: $s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_2, 3, s_2, 3, s_1, 1, s_\top$

- $\hat{V}^1 = G(s_2, 1, 1) = 4$.

First-visit MC Again

- Assume episodic task with S = $\{s_1, s_2, s_3\}$; following π.
- Say we start each episode with state s (for illustration s_2).

Episode 1: $s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_2, 3, s_2, 3, s_1, 1, s_\top$

- $\hat{V}^1 = G(s_2, 1, 1) = 4.$
- $\hat{V}^2 = \frac{1}{2}\{G(s_2, 1, 1) + G(s_2, 2, 1)\} = 5.5.$

First-visit MC Again

- Assume episodic task with S = $\{s_1, s_2, s_3\}$; following π.
- Say we start each episode with state s (for illustration s_2).

Episode 1: $s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_2, 3, s_2, 3, s_1, 1, s_\top$.

- $\hat{V}^1 = G(s_2, 1, 1) = 4$.
- $\hat{V}^2 = \frac{1}{2}\{G(s_2, 1, 1) + G(s_2, 2, 1)\} = 5.5$.
- $\hat{V}^3 = \frac{1}{3}\{G(s_2, 1, 1) + G(s_2, 2, 1) + G(s_2, 3, 1)\} \approx 6.33$.

First-visit MC Again

- Assume episodic task with S = $\{s_1, s_2, s_3\}$; following π.
- Say we start each episode with state s (for illustration s_2).

Episode 1: $s_2, 3, s_2, 1, s_\top$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.
Episode 3: $s_2, 2, s_1, 5, s_1, 1, s_\top$.
Episode 4: $s_2, 3, s_2, 3, s_1, 1, s_\top$.

- $\hat{V}^1 = G(s_2, 1, 1) = 4$.
- $\hat{V}^2 = \frac{1}{2}\{G(s_2, 1, 1) + G(s_2, 2, 1)\} = 5.5$.
- $\hat{V}^3 = \frac{1}{3}\{G(s_2, 1, 1) + G(s_2, 2, 1) + G(s_2, 3, 1)\} \approx 6.33$.
- In general, for $t \geq 1$:

$$\hat{V}^t(s) = \frac{1}{t}\sum_{i=1}^{t} G(s, i, 1).$$

An On-line Implementation

$$\hat{V}^t(s) = \frac{1}{t}\sum_{i=1}^{t} G(s, t, 1)$$

An On-line Implementation

$$
\begin{aligned}
\hat{V}^t(s) &= \frac{1}{t}\sum_{i=1}^{t} G(s, t, 1) \\
&= \frac{1}{t}\left(\sum_{i=1}^{t-1} G(s, i, 1) + G(s, t, 1)\right)
\end{aligned}
$$

An On-line Implementation

$$
\begin{aligned}
\hat{V}^t(s) &= \frac{1}{t}\sum_{i=1}^{t} G(s, t, 1) \\
&= \frac{1}{t}\left(\sum_{i=1}^{t-1} G(s, i, 1) + G(s, t, 1)\right) \\
&= \frac{1}{t}\Big((t-1)\hat{V}^{t-1}(s) + G(s, t, 1)\Big)
\end{aligned}
$$

An On-line Implementation

$$
\begin{aligned}
\hat{V}^t(s) &= \frac{1}{t}\sum_{i=1}^{t} G(s,t,1) \\
&= \frac{1}{t}\left(\sum_{i=1}^{t-1} G(s,i,1) + G(s,t,1)\right) \\
&= \frac{1}{t}\Big((t-1)\hat{V}^{t-1}(s) + G(s,t,1)\Big) \\
&= (1-\alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s,t,1) \text{ for } \alpha_t = \frac{1}{t}, \hat{V}^0(s) = 0.
\end{aligned}
$$

An On-line Implementation

$$
\begin{aligned}
\hat{V}^t(s) &= \frac{1}{t}\sum_{i=1}^{t} G(s, t, 1) \\
&= \frac{1}{t}\left(\sum_{i=1}^{t-1} G(s, i, 1) + G(s, t, 1)\right) \\
&= \frac{1}{t}\left((t-1)\hat{V}^{t-1}(s) + G(s, t, 1)\right) \\
&= (1-\alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s, t, 1) \text{ for } \alpha_t = \frac{1}{t}, \hat{V}^0(s) = 0.
\end{aligned}
$$

- We already know that $\lim_{t\to\infty} \hat{V}^t(s) = V^\pi(s)$.

An On-line Implementation

$$
\begin{aligned}
\hat{V}^t(s) &= \frac{1}{t}\sum_{i=1}^{t} G(s,t,1) \\
&= \frac{1}{t}\left(\sum_{i=1}^{t-1} G(s,i,1) + G(s,t,1)\right) \\
&= \frac{1}{t}\left((t-1)\hat{V}^{t-1}(s) + G(s,t,1)\right) \\
&= (1-\alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s,t,1) \text{ for } \alpha_t = \frac{1}{t}, \hat{V}^0(s) = 0.
\end{aligned}
$$

- We already know that $\lim_{t\to\infty} \hat{V}^t(s) = V^\pi(s)$.
- Will we get convergence to $V^\pi(s)$ for other choices for α_t, $\hat{V}^0(s)$?

Stochastic Approximation

- Result due to Robbins and Monro (1951).

Stochastic Approximation

- Result due to Robbins and Monro (1951).
- Let the sequence $(\alpha_t)_{t \geq 1}$ satisfy
  - $\sum_{t=1}^{\infty} \alpha_t = \infty$.
  - $\sum_{t=1}^{\infty} (\alpha_t)^2 < \infty$.

Stochastic Approximation

- Result due to Robbins and Monro (1951).
- Let the sequence $(\alpha_t)_{t \geq 1}$ satisfy
  - $\sum_{t=1}^{\infty} \alpha_t = \infty$.
  - $\sum_{t=1}^{\infty} (\alpha_t)^2 < \infty$.
- For $t \geq 1$, set

$$\hat{V}^t(s) \leftarrow (1 - \alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s, t, 1),$$

where $\hat{V}^0$ is arbitrary (but bounded).

Stochastic Approximation

- Result due to Robbins and Monro (1951).
- Let the sequence $(\alpha_t)_{t \geq 1}$ satisfy
  - $\sum_{t=1}^{\infty} \alpha_t = \infty$.
  - $\sum_{t=1}^{\infty} (\alpha_t)^2 < \infty$.
- For $t \geq 1$, set

$$\hat{V}^t(s) \leftarrow (1 - \alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s, t, 1),$$

  where $\hat{V}^0$ is arbitrary (but bounded).
- Then $\lim_{t \to \infty} \hat{V}^t(s) = V^\pi(s)$.

Stochastic Approximation

- Result due to Robbins and Monro (1951).
- Let the sequence $(\alpha_t)_{t \geq 1}$ satisfy
  - $\sum_{t=1}^{\infty} \alpha_t = \infty$.
  - $\sum_{t=1}^{\infty} (\alpha_t)^2 < \infty$.
- For $t \geq 1$, set
$$\hat{V}^t(s) \leftarrow (1 - \alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s, t, 1),$$
  where $\hat{V}^0$ is arbitrary (but bounded).
- Then $\lim_{t \to \infty} \hat{V}^t(s) = V^{\pi}(s)$.

- $(\alpha_t)_{t \geq 1}$ is the "learning rate" or "step size".

Stochastic Approximation

- Result due to Robbins and Monro (1951).
- Let the sequence $(\alpha_t)_{t \geq 1}$ satisfy
  - $\sum_{t=1}^{\infty} \alpha_t = \infty$.
  - $\sum_{t=1}^{\infty} (\alpha_t)^2 < \infty$.
- For $t \geq 1$, set
$$\hat{V}^t(s) \leftarrow (1 - \alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s, t, 1),$$
  where $\hat{V}^0$ is arbitrary (but bounded).
- Then $\lim_{t \to \infty} \hat{V}^t(s) = V^{\pi}(s)$.

- $(\alpha_t)_{t \geq 1}$ is the "learning rate" or "step size".
- Must be large enough, as well as small enough!

Stochastic Approximation

- Result due to Robbins and Monro (1951).
- Let the sequence $(\alpha_t)_{t \geq 1}$ satisfy
  - $\sum_{t=1}^{\infty} \alpha_t = \infty$.
  - $\sum_{t=1}^{\infty} (\alpha_t)^2 < \infty$.
- For $t \geq 1$, set
$$\hat{V}^t(s) \leftarrow (1 - \alpha_t)\hat{V}^{t-1}(s) + \alpha_t G(s, t, 1),$$
  where $\hat{V}^0$ is arbitrary (but bounded).
- Then $\lim_{t \to \infty} \hat{V}^t(s) = V^\pi(s)$.

- $(\alpha_t)_{t \geq 1}$ is the "learning rate" or "step size".
- Must be large enough, as well as small enough!
- No need to store all previous episodes; t and $\hat{V}^t$ suffice.

Reinforcement Learning

1. Prediction with Monte Carlo methods

2. On-line implementation

Reinforcement Learning

1. Prediction with Monte Carlo methods
2. On-line implementation

Next class: Bootstrapping.