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.

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

Monte Carlo (MC) methods estimate based on sample averages.
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.

<table>
<thead>
<tr>
<th>Episode 1</th>
<th>Episode 2</th>
<th>Episode 3</th>
<th>Episode 4</th>
<th>Episode 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T )</td>
<td>( s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T )</td>
<td>( s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T )</td>
<td>( s_3, 1, s_T )</td>
<td>( s_2, 3, s_2, 3, s_1, 1, s_T )</td>
</tr>
</tbody>
</table>

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

Monte Carlo (MC) methods estimate based on sample averages.
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:** \( s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T \).
- **Episode 2:** \( s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T \).
- **Episode 3:** \( s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T \).
- **Episode 4:** \( s_3, 1, s_T \).
- **Episode 5:** \( s_2, 3, s_2, 3, s_1, 1, s_T \).

What is your estimate of \( V^\pi \) (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: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T$.
  Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$.
  Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T$.
  Episode 4: $s_3, 1, s_T$.
  Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_T$.

- What is your estimate of $V^\pi$ (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

- \( 1(s, i, j) \) be 1 if \( s \) is visited at least \( j \) times on episode \( i \) (else \( 1(s, i, j) = 0 \)), 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 \( 1(s, i, j) = 0 \); also \( 0/0 = 0 \).

Episode 1: \( s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T \).
Episode 2: \( s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T \).
Episode 3: \( s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T \).
Episode 4: \( s_3, 1, s_T \).
Episode 5: \( s_2, 3, s_2, 3, s_1, 1, s_T \).
Defining Relevant Quantities

For $s \in S$, $i \geq 1, j \geq 1$, let

- $1(s, i, j)$ be 1 if $s$ is visited at least $j$ times on episode $i$ (else $1(s, i, j) = 0$), 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 $1(s, i, j) = 0$; also $0/0 = 0$.

\[
\begin{align*}
\text{Episode 1: } & s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T. \\
\text{Episode 2: } & s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T. \\
\text{Episode 3: } & s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T. \\
\text{Episode 4: } & s_3, 1, s_T. \\
\text{Episode 5: } & s_2, 3, s_2, 3, s_1, 1, s_T.
\end{align*}
\]

- $1(s_1, 1, 1) = 1$, $G(s_1, 1, 1) = 5 + \gamma \cdot 2 + \gamma^2 \cdot 3 + \gamma^3 \cdot 1 = 11$.
- $1(s_1, 1, 3) = 0$.
- $1(s_2, 5, 1) = 1$, $G(s_2, 5, 1) = 3 + \gamma \cdot 3 + \gamma^2 \cdot 1 = 7$.
- $1(s_2, 5, 2) = 1$, $G(s_2, 5, 2) = 3 + \gamma \cdot 1 = 4$. 
Some Standard Estimates of $V^\pi(s)$

<table>
<thead>
<tr>
<th>Episode 1:</th>
<th>$s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Episode 2:</td>
<td>$s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$.</td>
</tr>
<tr>
<td>Episode 3:</td>
<td>$s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T$.</td>
</tr>
<tr>
<td>Episode 4:</td>
<td>$s_3, 1, s_T$.</td>
</tr>
<tr>
<td>Episode 5:</td>
<td>$s_2, 3, s_2, 3, s_1, 1, s_T$</td>
</tr>
</tbody>
</table>

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.

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

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: | $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T$. |
| Episode 2: | $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$. |
| Episode 3: | $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T$. |
| Episode 4: | $s_3, 1, s_T$. |
| Episode 5: | $s_2, 3, s_2, 3, s_1, 1, s_T$. |

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

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

$$\hat{V}^N_{\text{Every-visit}}(s) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{\infty} G(s, i, j)}{\sum_{i=1}^{N} \sum_{j=1}^{\infty} 1(s, i, j)}.$$
Some Standard Estimates of $V_\pi(s)$

| Episode 1: | $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T$ |
| Episode 2: | $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$ |
| Episode 3: | $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T$ |
| Episode 4: | $s_3, 1, s_T$ |
| Episode 5: | $s_2, 3, s_2, 3, s_1, 1, s_T$ |

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

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

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

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: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T$.
Episode 4: $s_3, 1, s_T$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_T$.

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.

$$\hat{V}^N_{\text{Second-visit}}(s) = \frac{\sum_{i=1}^{N} G(s, i, 2)}{\sum_{i=1}^{N} 1(s, i, 2)}.$$
Some Not-so-standard Estimates of $V^\pi(s)$

| Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T$. |
| Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$. |
| Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T$. |
| Episode 4: $s_3, 1, s_T$. |
| Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_T$. |

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.

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

Hence $\hat{V}_{\text{Second-visit}}^5(s_2) = \frac{1 + 1 + 4}{3} = 2$. 
Some Not-so-standard Estimates of $V^\pi(s)$

Episode 1: $s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T$.
Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$.
Episode 3: $s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T$.
Episode 4: $s_3, 1, s_T$.
Episode 5: $s_2, 3, s_2, 3, s_1, 1, s_T$.

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).

$$\hat{V}^N_{\text{Last-visit}}(s) = \frac{\sum_{i=1}^{N} G(s, i, times(s, i))}{\sum_{i=1}^{N} 1(s, i, times(s, i))}.$$
Some Not-so-standard Estimates of $V^\pi(s)$

<table>
<thead>
<tr>
<th>Episode</th>
<th>States</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$s_1, 5, s_1, 2, s_2, 3, s_2, 1, s_T$</td>
</tr>
<tr>
<td>2</td>
<td>$s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$</td>
</tr>
<tr>
<td>3</td>
<td>$s_1, 2, s_2, 2, s_1, 5, s_1, 1, s_T$</td>
</tr>
<tr>
<td>4</td>
<td>$s_3, 1, s_T$</td>
</tr>
<tr>
<td>5</td>
<td>$s_2, 3, s_2, 3, s_1, 1, s_T$</td>
</tr>
</tbody>
</table>

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 $\text{times}(s, i)$ visits).

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

Hence $\hat{V}_5^\text{Last-visit}(s_2) = \frac{1 + 1 + 8 + 4}{4} = 3.5.$
Question

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

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

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

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

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

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

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

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

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

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

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

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

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

\[
\lim_{N \to \infty} \hat{V}^N_{\text{Last-visit}} = V^\pi.
\]
Recall that we generate $N$ episodes.
Which claims below are true?

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

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

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

\[
\lim_{N \to \infty} \hat{V}^N_{\text{Last-visit}} = V^\pi. \text{ False.}
\]
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 $\pi$.
- Say we start each episode with state $s$ (for illustration $s_2$).

<table>
<thead>
<tr>
<th>Episode</th>
<th>States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Episode 1</td>
<td>$s_2, 3, s_2, 1, s_T$</td>
</tr>
<tr>
<td>Episode 2</td>
<td>$s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$</td>
</tr>
<tr>
<td>Episode 3</td>
<td>$s_2, 2, s_1, 5, s_1, 1, s_T$</td>
</tr>
<tr>
<td>Episode 4</td>
<td>$s_2, 3, s_2, 3, s_1, 1, s_T$</td>
</tr>
</tbody>
</table>
First-visit MC Again

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

| Episode 1:    | $s_2, 3, s_2, 1, s_T$. |
| Episode 2:    | $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$. |
| Episode 3:    | $s_2, 2, s_1, 5, s_1, 1, s_T$. |
| Episode 4:    | $s_2, 3, s_2, 3, s_1, 1, s_T$ |

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

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

<table>
<thead>
<tr>
<th>Episode 1: $s_2, 3, s_2, 1, s_\top$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_\top$.</td>
</tr>
<tr>
<td>Episode 3: $s_2, 2, s_1, 5, s_1, 1, s_\top$.</td>
</tr>
<tr>
<td>Episode 4: $s_2, 3, s_2, 3, s_1, 1, s_\top$.</td>
</tr>
</tbody>
</table>

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

Shivaram Kalyanakrishnan (2022)
First-visit MC Again

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

<table>
<thead>
<tr>
<th>Episode 1:</th>
<th>$s_2, 3, s_2, 1, s_T$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Episode 2:</td>
<td>$s_2, 2, s_3, 1, s_3, 1, s_2, 1, s_T$.</td>
</tr>
<tr>
<td>Episode 3:</td>
<td>$s_2, 2, s_1, 5, s_1, 1, s_T$.</td>
</tr>
<tr>
<td>Episode 4:</td>
<td>$s_2, 3, s_2, 3, s_1, 1, s_T$.</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Episode 1: $s_2, 3, s_2, 1, s_T$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Episode 2: $s_2, 2, s_3, 1, s_3, 1, s_3, 2, s_2, 1, s_T$.</td>
</tr>
<tr>
<td>Episode 3: $s_2, 2, s_1, 5, s_1, 1, s_T$.</td>
</tr>
<tr>
<td>Episode 4: $s_2, 3, s_2, 3, s_1, 1, s_T$.</td>
</tr>
</tbody>
</table>

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

- 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

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

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 \( \alpha_t \), \( \hat{V}_0(s) = 0 \)?
An On-line Implementation

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

\[
= \frac{1}{t} \left( \sum_{i=1}^{t-1} G(s, i, 1) + G(s, t, 1) \right)
\]

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 \(\alpha_t\), \(\hat{V}_0(s) = 0\)?
An On-line Implementation

\[ \hat{V}^t(s) = \frac{1}{t} \sum_{i=1}^{t} G(s, t, 1) \]
\[ = \frac{1}{t} \left( \sum_{i=1}^{t-1} G(s, i, 1) + G(s, t, 1) \right) \]
\[ = \frac{1}{t} \left( (t - 1) \hat{V}^{t-1}(s) + G(s, t, 1) \right) \]

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 \( \alpha^t \), \( \hat{V}^0(s) = 0 \)?
An On-line Implementation

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

\[
= \frac{1}{t} \left( \sum_{i=1}^{t-1} G(s, i, 1) + G(s, t, 1) \right)
\]

\[
= \frac{1}{t} \left( (t - 1) \hat{V}^{t-1}(s) + G(s, t, 1) \right)
\]

\[
= (1 - \alpha_t) \hat{V}^{t-1}(s) + \alpha_t G(s, t, 1) \text{ for } \alpha_t = \frac{1}{t}, \hat{V}^0(s) = 0.
\]
An On-line Implementation

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

\[ = \frac{1}{t} \left( \sum_{i=1}^{t-1} G(s, i, 1) + G(s, t, 1) \right) \]

\[ = \frac{1}{t} \left( (t - 1) \hat{V}^{t-1}(s) + G(s, t, 1) \right) \]

\[ = (1 - \alpha_t) \hat{V}^{t-1}(s) + \alpha_t G(s, t, 1) \text{ for } \alpha_t = \frac{1}{t}, \hat{V}^0(s) = 0. \]

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

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

\[ = \frac{1}{t} \left( \sum_{i=1}^{t-1} G(s, i, 1) + G(s, t, 1) \right) \]

\[ = \frac{1}{t} \left( (t - 1) \hat{V}^{t-1}(s) + G(s, t, 1) \right) \]

\[ = (1 - \alpha_t) \hat{V}^{t-1}(s) + \alpha_t G(s, t, 1) \text{ for } \alpha_t = \frac{1}{t}, \hat{V}^0(s) = 0. \]

- 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 \( \alpha_t, \hat{V}^0(s) \)?
Stochastic Approximation

- Result due to Robbins and Monro (1951).

Let the sequence $(\alpha_t)_{t \geq 1}$ satisfy

\[
P\lim_{t \to \infty} \alpha_t = 1 \quad \text{and} \quad \sum_{t=1}^{\infty} \alpha_t^2 < \infty.
\]

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.
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.
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.