Learning Policies with External Memory

Leonid Peshkin, Nicolas Meuleau, and Leslie Pack Kaelbling

*Computer Science Department, Brown University (1999)**

Simplified VAPS algorithm for online stigmergic policies

Robert Moss

January 22, 2020

* In *Proceedings of the 16th International Conference on Machine Learning*

Motivation: Goals and Strategy

Goal: Reduce the curse of history in partially observable domains

- Incorporate external memory in learning online policies
- While continuing to learn a memoryless mapping from observations to actions
- Learn a reactive policy (i.e. online) in a highly non-Markovian domain

Strategy: Explore a *stigmergic* approach (i.e. indirect coordination)

- *Def*: "The process of an insect's activity acting as a stimulus to further activity"
- An agent's actions include the ability to set and clear external memory bits
- Describes an agent's environmental changes that affect future behavior
    - Cited in the mobile robotics literature

Motivation: Importance and Problems

Importance:

- Incorporate history when solving for policies using a POMDP framework
- Preserve the notion of memorylessness in traditionally non-Markovian domains

Why is it difficult?

- Basic RL (i.e. Q-learning) can perform poorly in partially observable domains
  - Due to strong Markov assumptions
- Finding the optimal memoryless policy is NP-Hard [10]

[10] Michael L. Littman. **Markov games as a framework for multi-agent reinforcement learning**. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163, San Francisco, CA, 1994.

Motivation: Prior Work

Three classes of learning in partially observable domains:

1. ***Optimal memoryless***
   - Many policies can perform well in partially observable domains (e.g. $\textsc{Sarsa}(\lambda)$)
   - *Drawbacks*: Finding the optimal memoryless policy in partially observable domains is NP-Hard [10]
2. ***Finite memory***
   - Chooses actions based on a finite window of previous observations
   - *Drawbacks*: Relies on a finite-sized memory allocation of past policies
3. ***Model-based***
   - Assumes complete knowledge of the underlying process (modeled as a POMDP)
   - Can find optimal solution (or search for approximations)
   - *Drawbacks*: State-space dimensionality limitations & the model may not be known

[10] Michael L. Littman. **Markov games as a framework for multi-agent reinforcement learning**. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163, San Francisco, CA, 1994.

Contributions

- Applied stigmergic approaches to learning policies in partially observable domains
    - Employed online in highly non-Markovian domains
- Derived a simplified version of the VAPS algorithm for stigmergic policies
    - VAPS: Value and Policy Search [1]
        - General method for gradient descent in reinforcement learning
- Calculated the same gradient as VAPS with less computational effort

Novelty:

- This work incorporated external memory into learning memoryless online policies
- Assigns credit to (s, a) pairs proportional to the deviation from expected behavior

[1] Leemon C. Baird and Andrew W. Moore. **Gradient descent for general reinforcement learning**. In *Advances in Neural Information Processing Systems 11*. The MIT Press, 1999.

Problem Setting: Environment

The agent has two components (Figure 1):

1. Reinforcement-learning agent
2. Set of external memory bits

Input:

- The environmental observation o, augmented by the memory

Output:

- Action a and modifications to the state of the memory

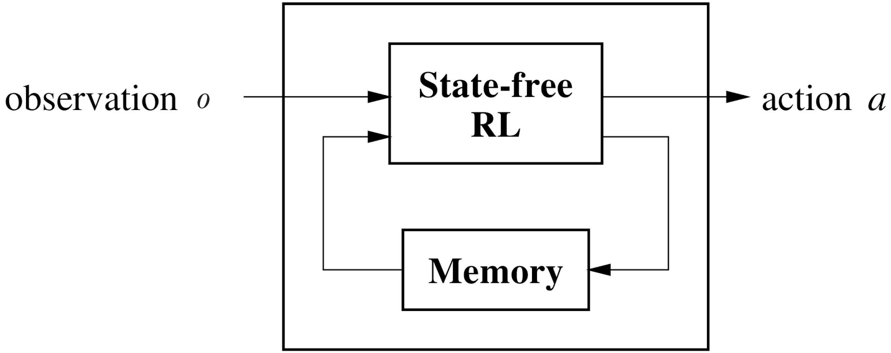


Figure 1: The architecture of a stigmergic policy.

* *See last slide for notation changes*

Example Problem: Load-Unload

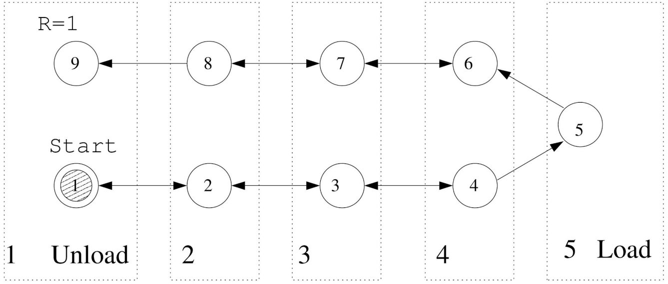


Figure 2: The state-transition diagram of the load-unload problem; aliased states are grouped by dashed boxes.

- Drive from *unload* to *load* then back to *unload*
  - Simple MDP with one-bit hidden variable (making it non-Markovian)
- Can be solved using a one-bit external memory
  - Set the bit when *unload* is observed (then go right)
  - Clear the bit when *load* is observed (then go left)

Architectural design with external memory:

- *Augment* the action space with memory-changing actions
  - *Adds* a new action for each memory bit
  - Changing the state may require additional steps (fixed by not discounting these actions)

Algorithm 1: $\textsc{Sarsa}(\lambda)$

"On-policy temporal-difference control learning algorithm":

- Q-values are updated via the rule (recall from AA228):

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma Q(s',a') - Q(s,a)\right]$$

- $\textsc{Sarsa}$ uses the actual action taken from state s'
  - In other words, $a' = \pi(s')$ or based on some exploration strategy (hence "on-policy")
- $\textsc{Sarsa}(\lambda)$ augments the $\textsc{Sarsa}$ algorithm via eligibility traces (where $0 \leq \lambda \leq 1$)
- For non-Markovian domains, $\textsc{Sarsa}(\lambda)$ has been shown to be appropriate ($\lambda \approx 1$)

eligibility traces [0]

$$\left\{\begin{array}{l} \textbf{for } s \in S \\ \quad \textbf{for } a \in A \\ \qquad Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a) \\ \qquad N(s,a) \leftarrow \gamma\lambda N(s,a) \end{array}\right.$$

[0] Mykel J. Kochenderfer. **Decision Making Under Uncertainty: Theory and Application**. The MIT Press, 2015.

Algorithm 2: VAPS

Value And Policy Search (VAPS):

- A stochastic gradient descent algorithm for reinforcement learning
- Seeks to minimize the expected cost of the policy: $B = \sum_{T=0}^{\infty} \sum_{\tilde{z} \in \tilde{Z}_T} P(\tilde{z}) \varepsilon(\tilde{z})$
- Where $\tilde{Z}_T$ is the set of all possible *experience sequences* that terminate at time T

$$\tilde{z} = \langle o_0, a_0, r_0, \ldots, o_t, a_t, r_t, \ldots, o_T, a_T, r_T \rangle$$

- The loss accumulated by a sequence $\tilde{z}$ is $\varepsilon(\tilde{z})$, where

$$\varepsilon(\tilde{z}) = \sum_{t=0}^{T} e(\mathrm{trunc}(\tilde{z}, t)), \qquad \text{for all } \tilde{z} \in \tilde{Z}_T$$

- Where $e(z)$ is the instantaneous error function associate with each sequence z
- Where $\mathrm{trunc}(\tilde{z}, t)$ represents the sequence $\tilde{z}$ truncated after time t

Algorithm 2: VAPS

(continued 1)

- The $\textsc{Sarsa}$ error measurement is given by

$$\underline{e_{\textsc{sarsa}}(z)} = \frac{1}{2}\sum_{o} P(o_t = o \mid o_{t-1}, a_{t-1}) \sum_{a} P(a_t = a \mid o_t)\left[r_{t-1} + \gamma Q(o,a) - Q(o_{t-1}, a_{t-1})\right]^2$$

- Where the policy search error measurement is given by $\underline{e_{\text{policy}}(z)} = b - \gamma^t r_t$
  - Where b is any constant (always set to 0 in their experiments)
  - The immediate error e is summed over all time t, summing all discounted immediate rewards $\gamma^t r_t$

(Recall): $$\varepsilon(\tilde{z}) = \sum_{t=0}^{T} \underline{e}(\text{trunc}(\tilde{z}, t)), \qquad \text{for all } \tilde{z} \in \tilde{Z}_T$$

- A linear combination of the error functions is used to balance both criteria

$$\underline{e} = (1-\beta)\underline{e_{\textsc{sarsa}}} + \underline{\beta e_{\text{policy}}}$$

- Combines criterion for Value And Policy Search, hence VAPS (or VAPS(β))

Algorithm 2: VAPS

(continued 2)

The ***problem*** with traditional VAPS:

- Although the immediate error gradient of e_{SARSA} is "easy" to compute

$$\frac{\partial}{\partial w_k} e_{\mathrm{SARSA}}(z) = \sum_{o} P(o_t = o \mid o_{t-1}, a_{t-1}) \sum_{a} P(a_t = a \mid o_t) \left[r_{t-1} + \gamma Q(o, a) - Q(o_{t-1}, a_{t-1}) \right] \left[\gamma \frac{\partial}{\partial w_k} Q(o, a) - \frac{\partial}{\partial w_k} Q(o_{t-1}, a_{t-1}) \right]$$

- The o_t and a_t quantities are sampled twice to avoid bias in gradient estimation
  - Yet, the only way to get a new observation is to *perform* the action!

This is **unrealistic** in the *online* case!

Algorithm 3: $\mathrm{VAPS}(1)$

$$e = (1-\beta)e_{\text{SARSA}} + \beta e_{\text{policy}}$$

Author's **novel simplifications** to VAPS:

- When $\beta = 1$, i.e. $\mathrm{VAPS}(1)$, the second sample is not needed (e_{SARSA} is not used)
  - Effective in the *online* case!
- The gradient for policy search, $\frac{\partial}{\partial w_k} e_{\text{policy}}(z) = 0$, for all w_k
  - Importance of policy search are state visits, which enter in the weight update through the trace
- Q-values are stored in a lookup table, where $w_k = Q(s,a)$
- Assuming an achievement task, each $Q(s,a)$ has an associated exploration trace:

$$\Phi_{s,a,t} = \frac{1}{c}\left[N_{s,a}^t - N_s^t P(a_t = a \mid s_t = s)\right] = \frac{1}{c}\left[N_{s,a}^t - E\left[N_{s,a}^t\right]\right]$$

$$\underbrace{P(a_t = a \mid s_t = s)}_{\substack{\text{Boltzmann}\\ \text{distribution}}} = \frac{e^{Q(s,a)/c}}{\sum_{a'} e^{Q(s,a')/c}}$$

Simplified VAPS(1): Key Contribution

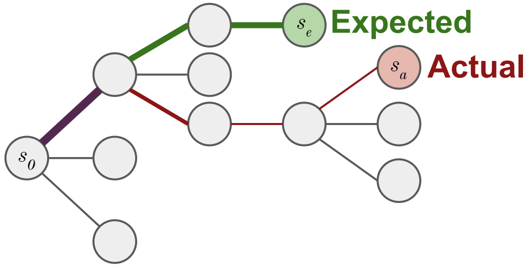
- "At each time-step where the current trial does not complete, we just **increment the counter** $N^t_{s,a}$ of the current state-action pair. When the trial completes, this **trace is used to update all the Q-values** [via $\Phi_{s,a,t}$]."

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

$$\underbrace{\delta c}_{\text{temperature decay}} = \left(\frac{c_{min}}{c_{max}}\right)^{1/(N-1)}$$

Discussion: Simplified Properties

1. The algorithm adjusts the unlikely actions when something surprising happens
   - "...assigns **credit to state-action pairs** proportional to the **deviation from the expected behavior**"
   - Sarsa(λ) is not capable of this discrimination

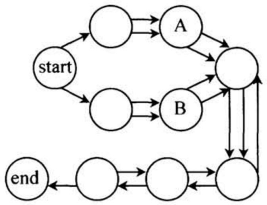


$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

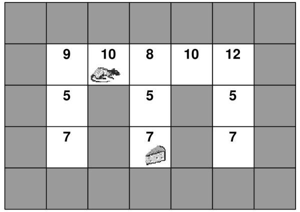
2. The Q-value updates approach 0 as the length of the trial approaches infinity
   - "...when **too many actions** have been performed, there is no reason to **attribute the final result** more to one of them than to others."

Experiments Setup

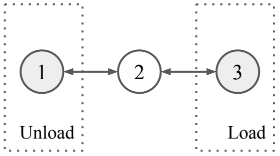
- **Evaluated SARSA(λ) and VAPS(1) on five simple problems:**
  1. Baird and Moore's problem, designed to illustrate the behavior of VAPS [1]
  2. McCallum's 11-state maze, with only six observations [12]
  3. The *load-unload* problem, with three locations (*loading*, *unloading*, and an *intermediate*)
  4. A five-location *load-unload* problem (as seen before)
  5. A *load-unload* problem specifically designed to highlight the benefit of VAPS(1)



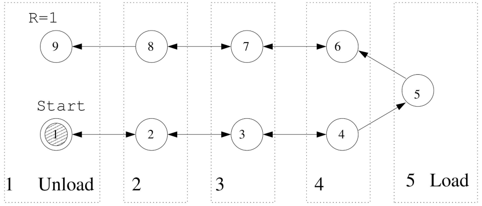
1



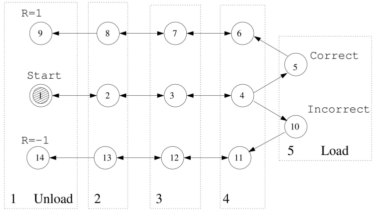
2



3



4

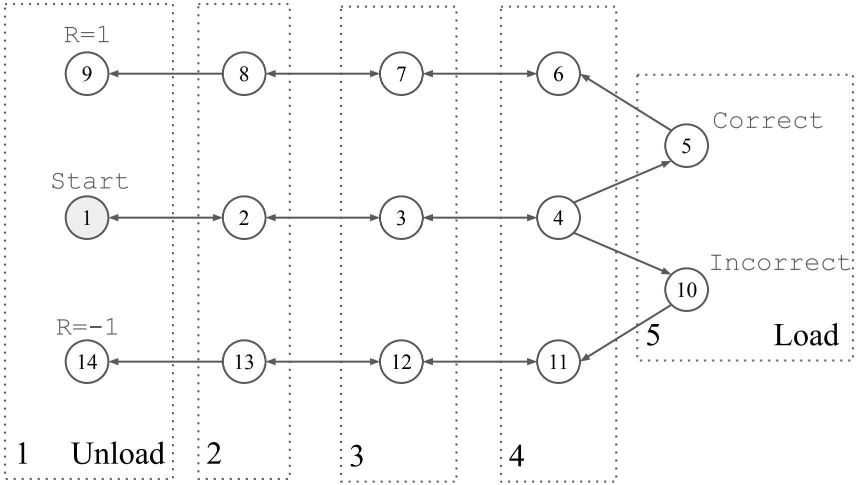


5

[1] Leemon C. Baird and Andrew W. Moore. **Gradient descent for general reinforcement learning**. In *Advances in Neural Information Processing Systems 11*. The MIT Press, 1999.

[12] Andrew Kachites McCallum. **Reinforcement Learning with Selective Perception and Hidden State**. *PhD thesis*, University of Rochester, Rochester, New York, 1995.

Experiments Setup: Problem 5



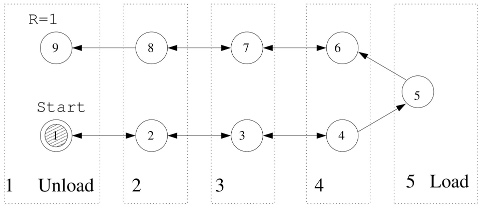
- Add additional *incorrect* load location
- Agent is punished for loading at wrong location ($R = -1$)
- Dashed-boxed states are observationally indistinguishable

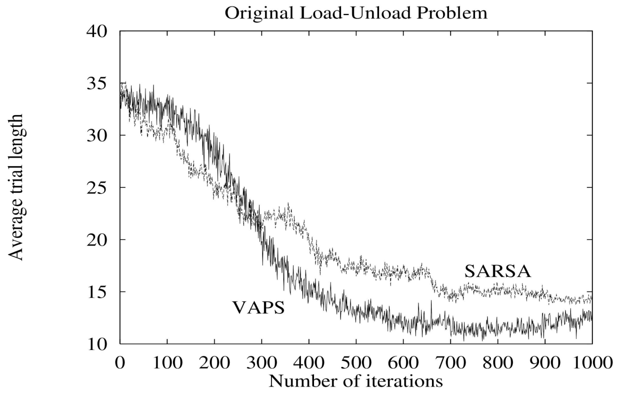
- **Idea**: A single action can ruin the agents long-term plan
  - $\textsc{Sarsa}(\lambda)$ will punish all action choices along that chain
  - $\mathrm{VAPS}(1)$ will only punish that specific action

Experimental Results: Discussion

Original Load-Unload Problem:

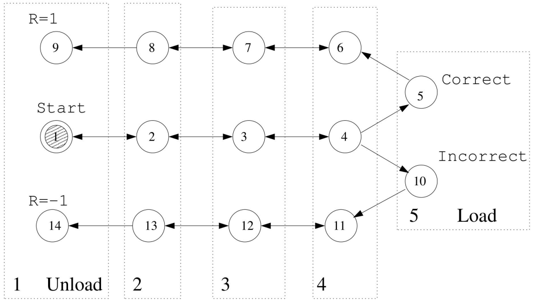
- Algorithms perform "essentially equivalently" on the original *load-unload* problem
  - Most runs converge to the optimal trial length of 9

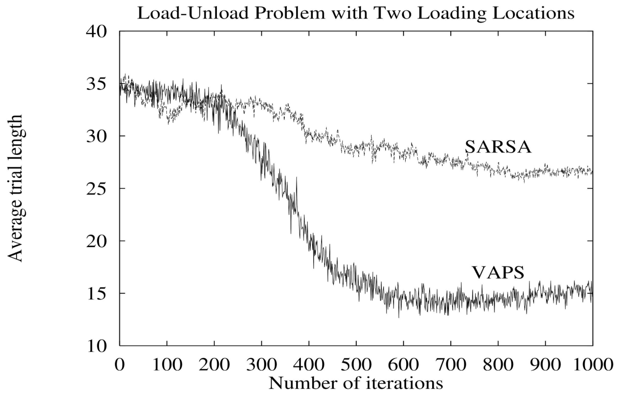


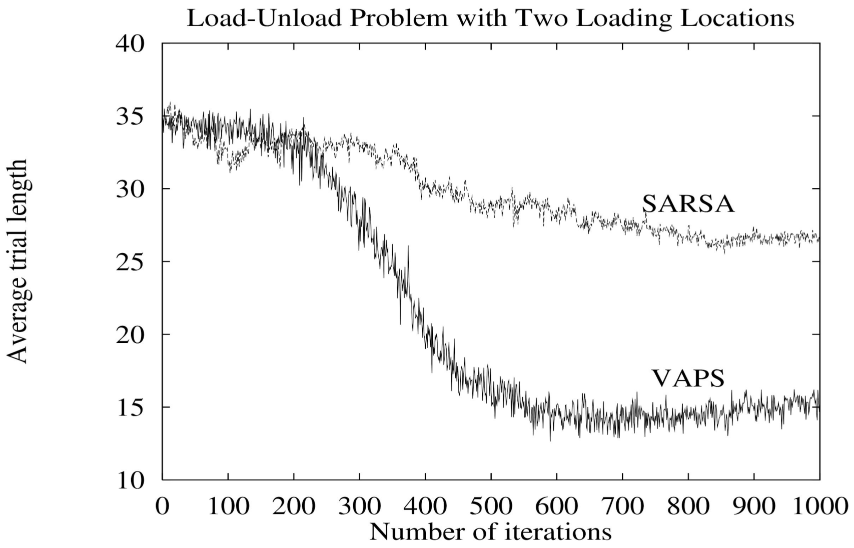


Load-Unload Problem with Two Loading Locations:

- $\mathrm{VAPS}(1)$ converge to a near-optimal policy in extended *load-unload* problem
  - $\textsc{Sarsa}(1)$ does not: exploring may punished the agent for picking the wrong load



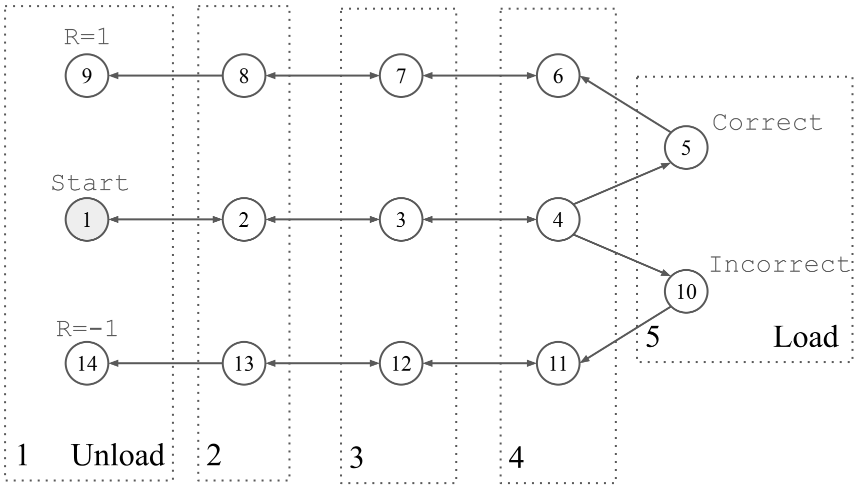


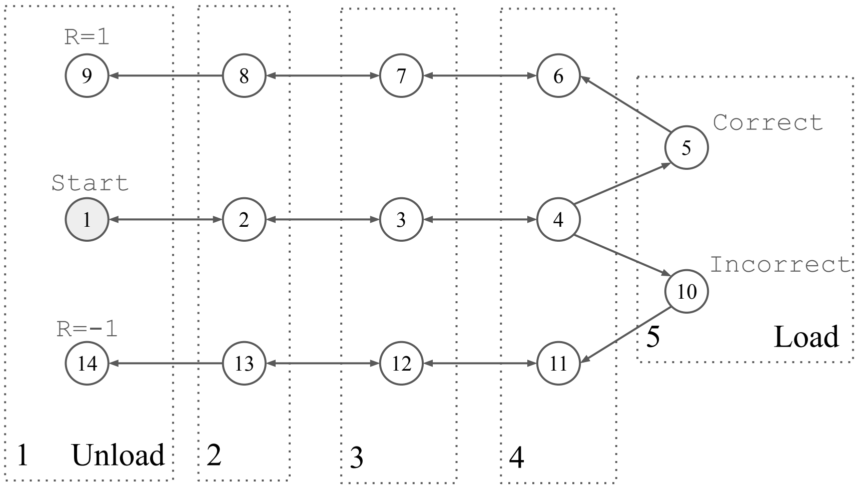


"SARSA will punish all state-action pairs **equally**"

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

Intuition Visualization





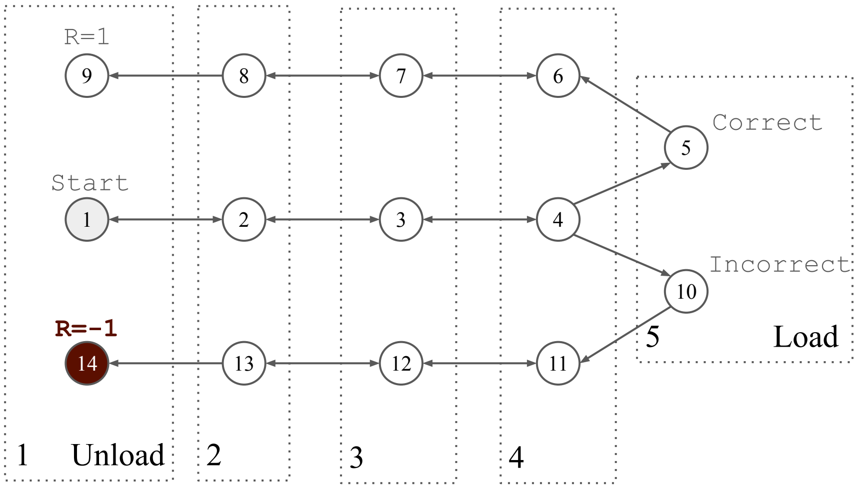
"SARSA will punish all state-action pairs **equally**"

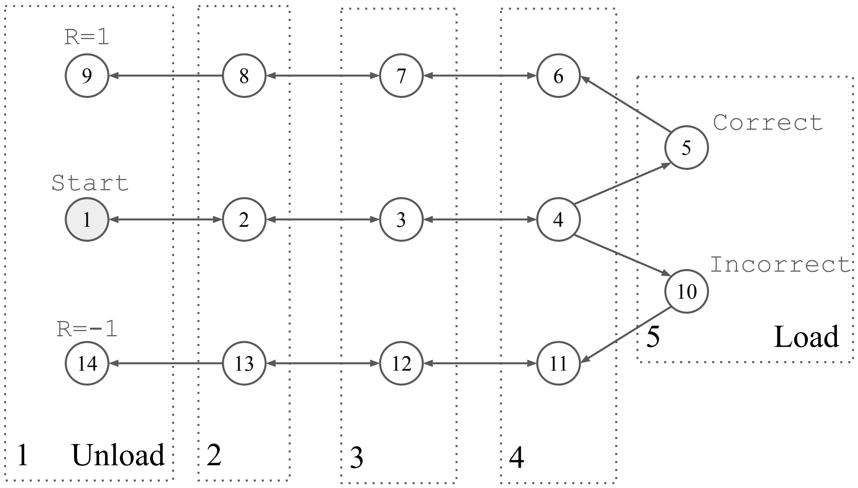
"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





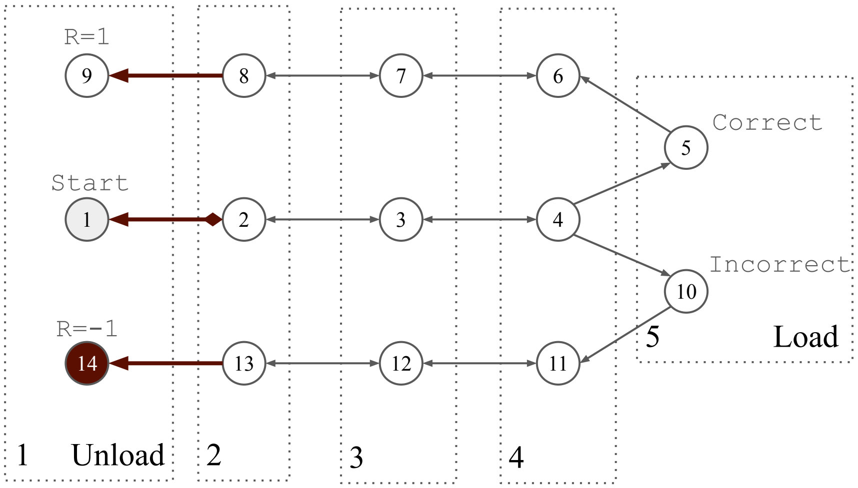
"SARSA will punish all state-action pairs **equally**"

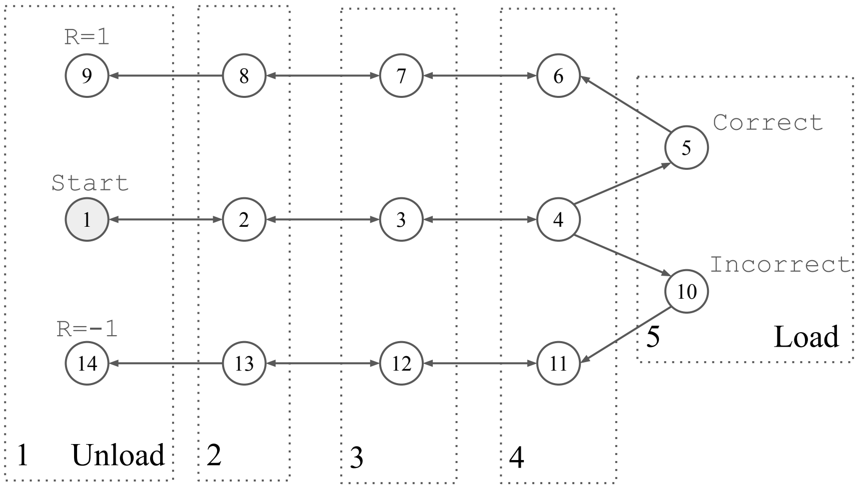
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





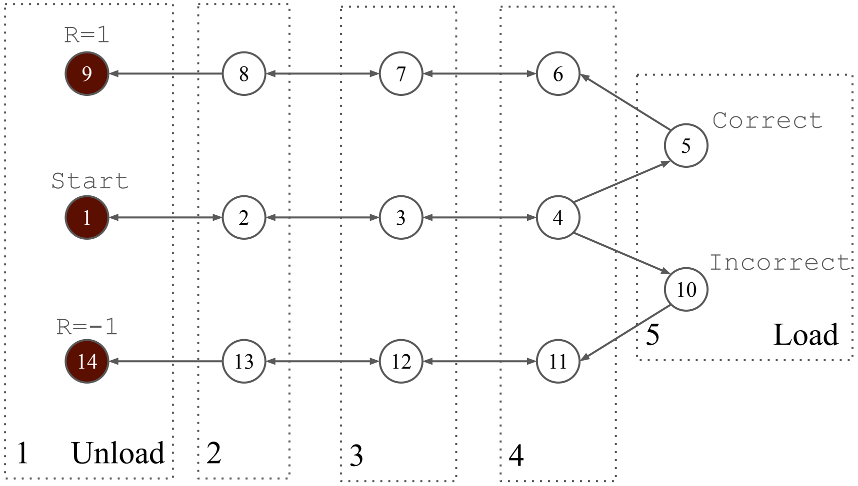
"SARSA will punish all state-action pairs **equally**"

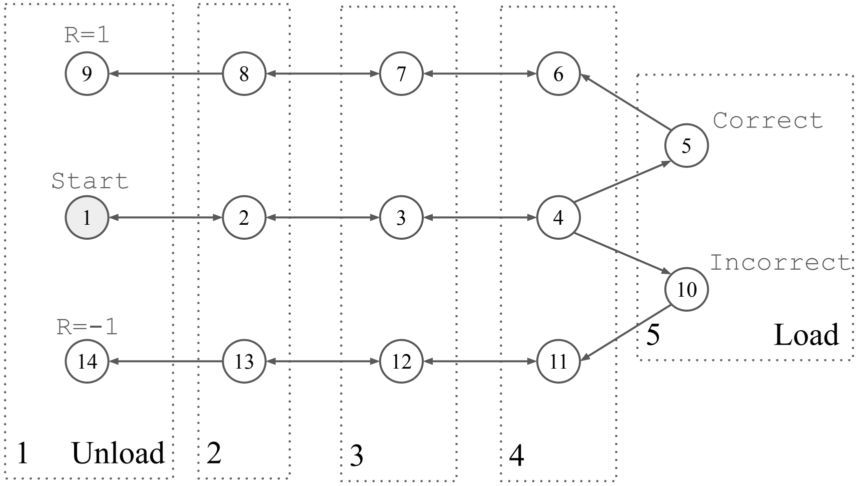
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





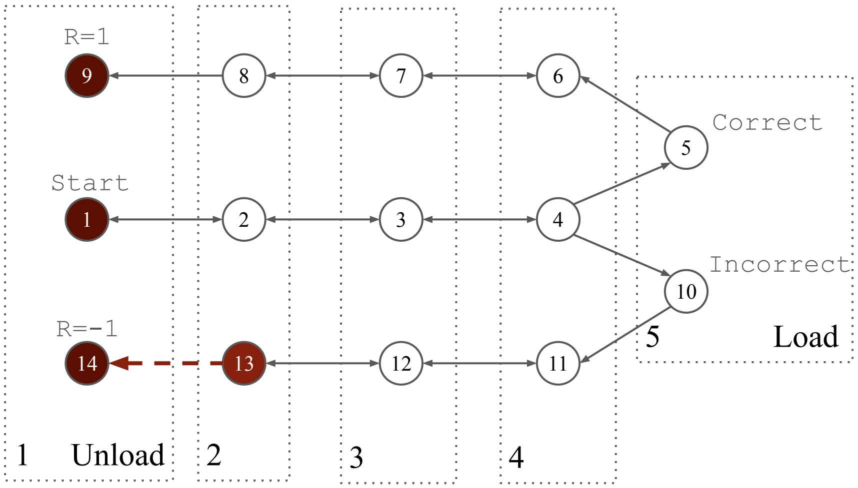
"SARSA will punish all state-action pairs **equally**"

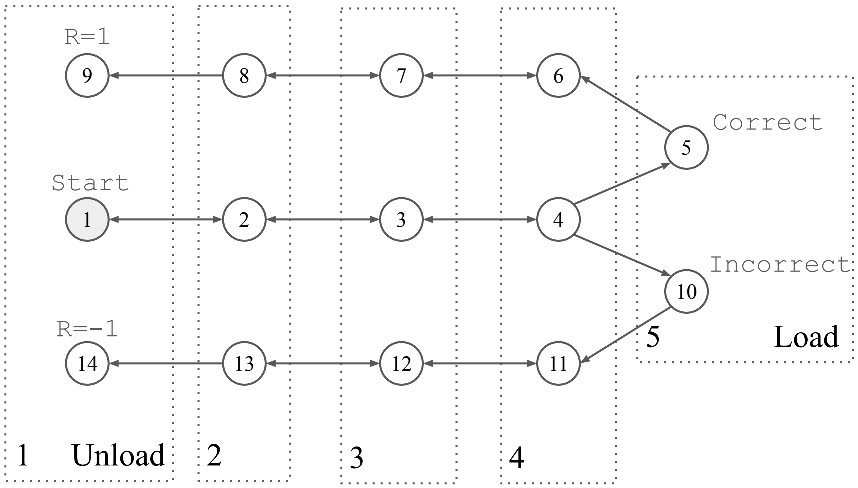
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
$N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





"SARSA will punish all state-action pairs **equally**"

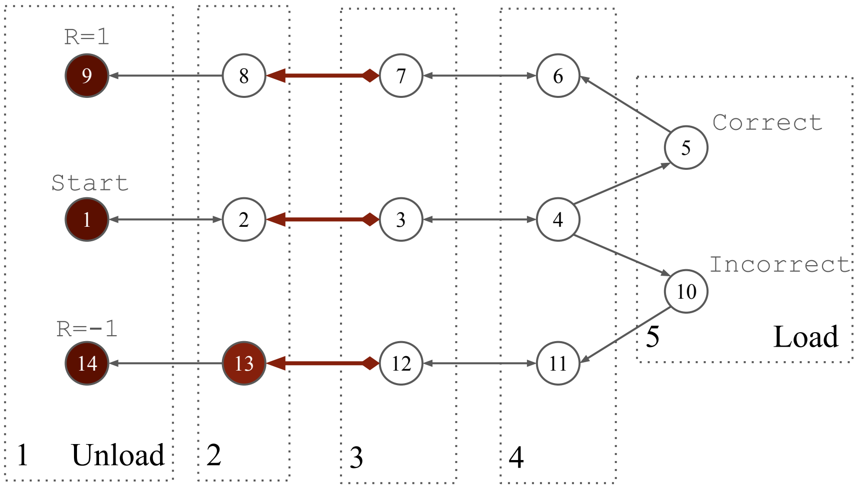
for $s \in S$
for $a \in A$

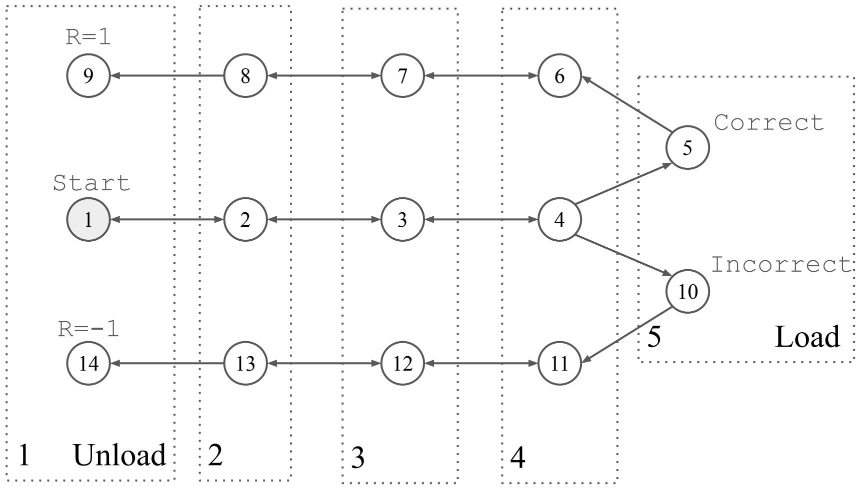
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





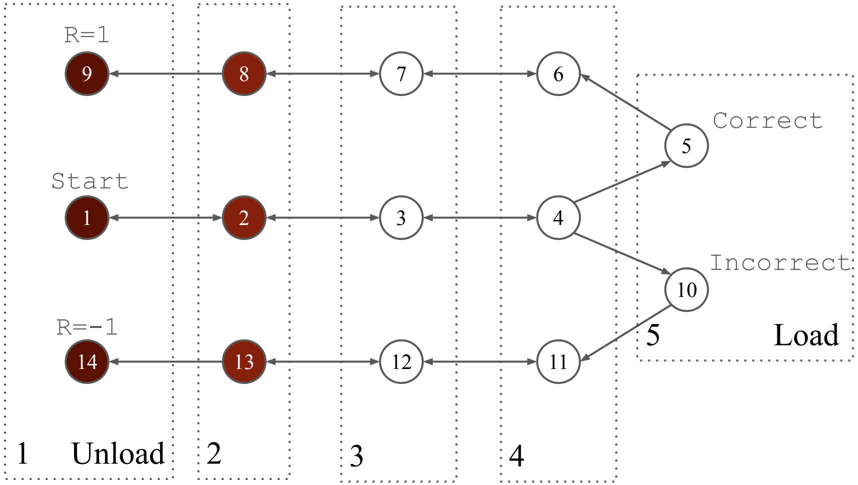
"SARSA will punish all state-action pairs **equally**"

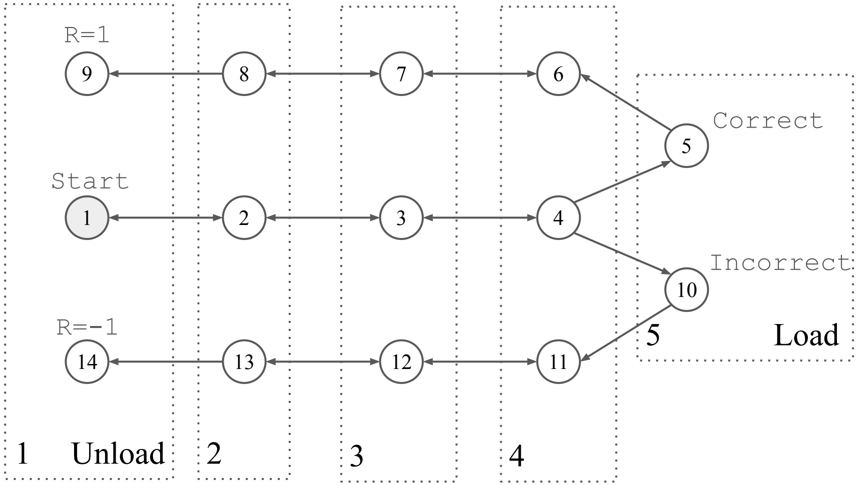
for $s \in S$
for $a \in A$
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





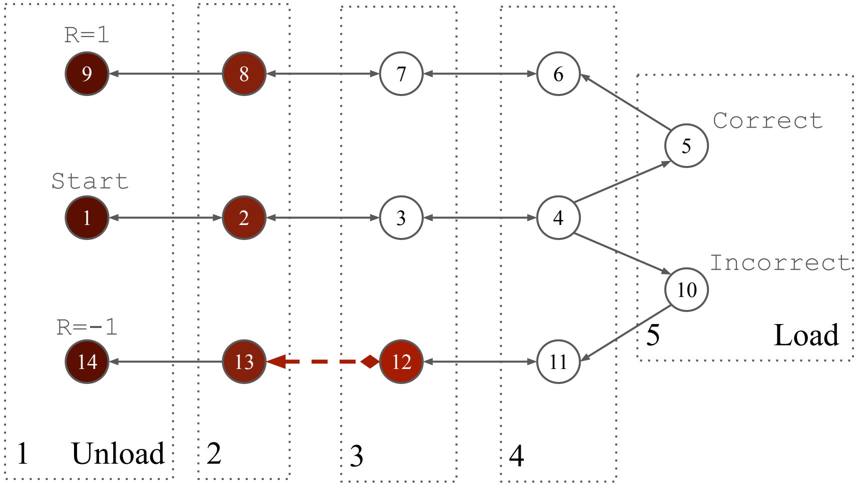
"SARSA will punish all state-action pairs **equally**"

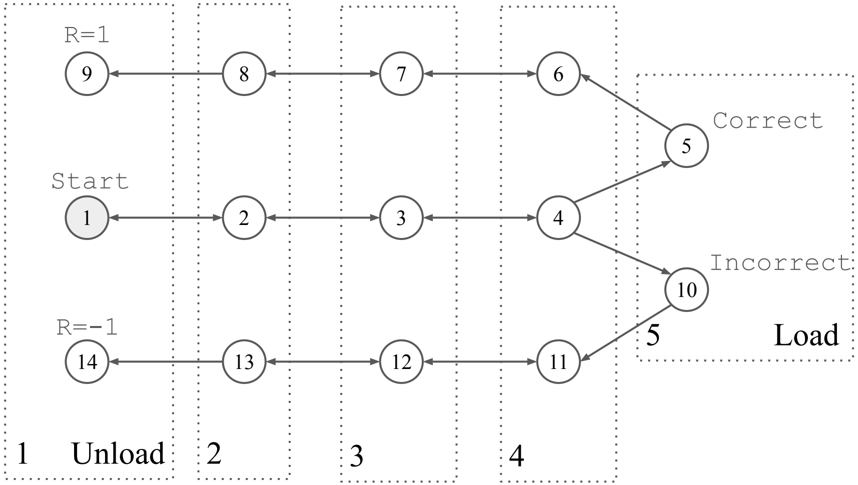
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





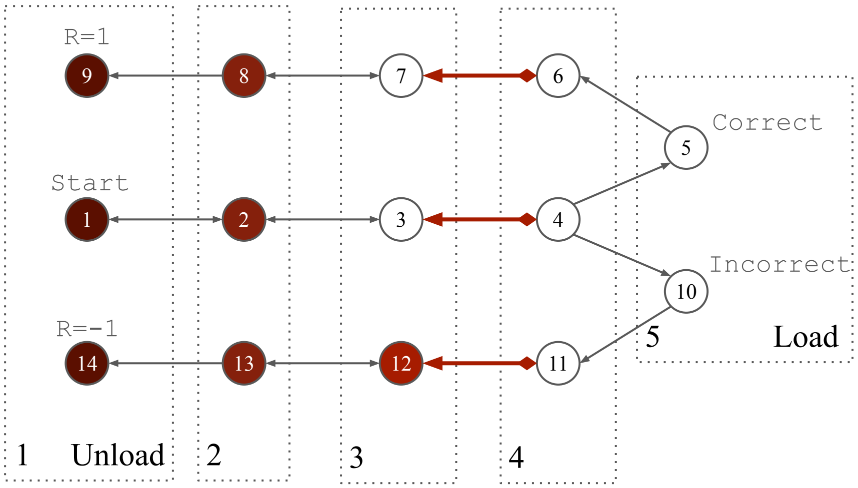
"SARSA will punish all state-action pairs **equally**"

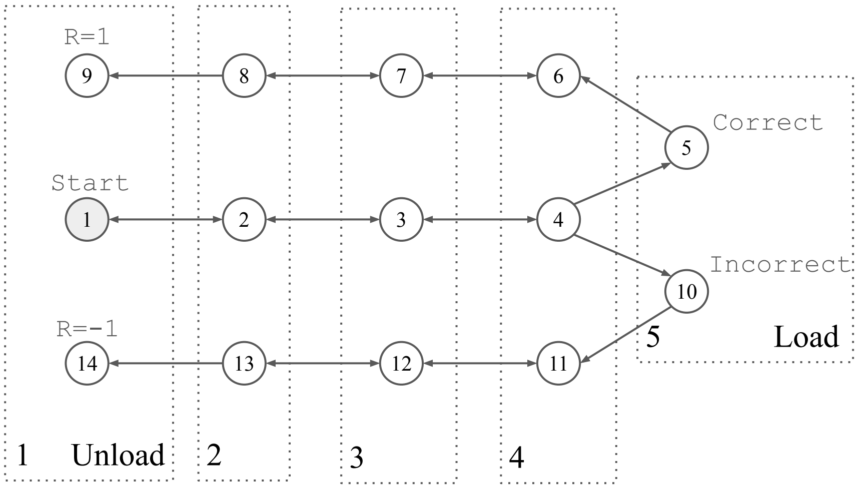
"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





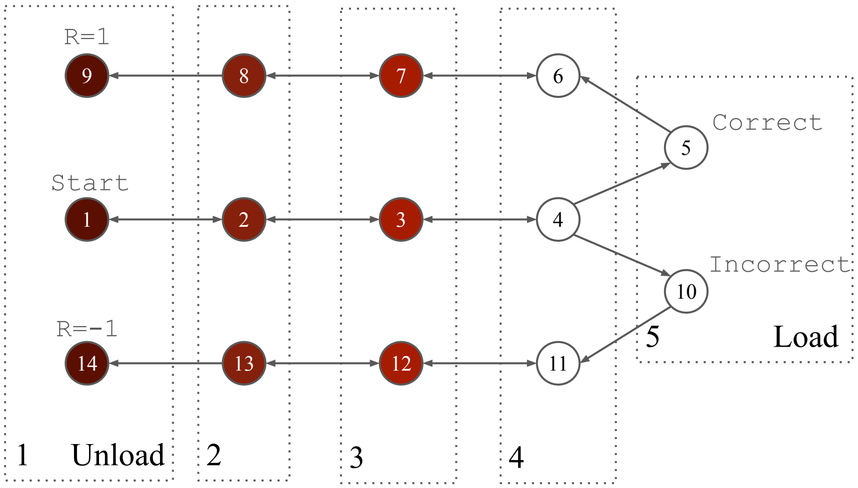
"SARSA will punish all state-action pairs **equally**"

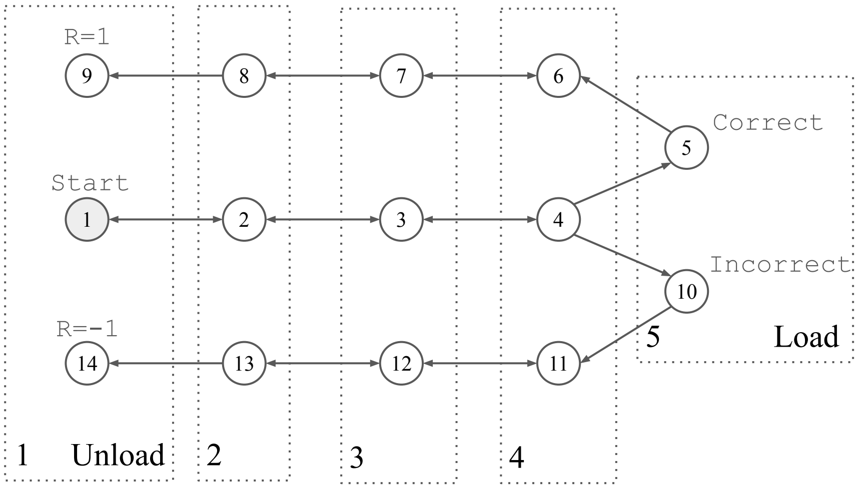
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





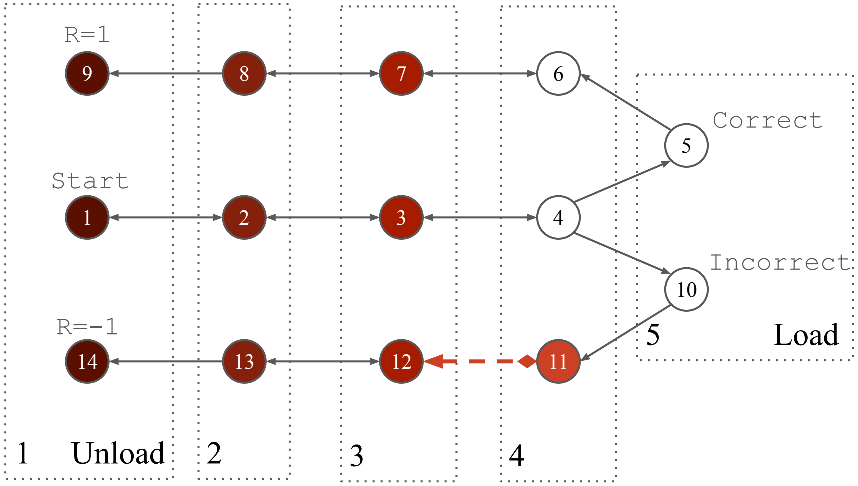
"SARSA will punish all state-action pairs **equally**"

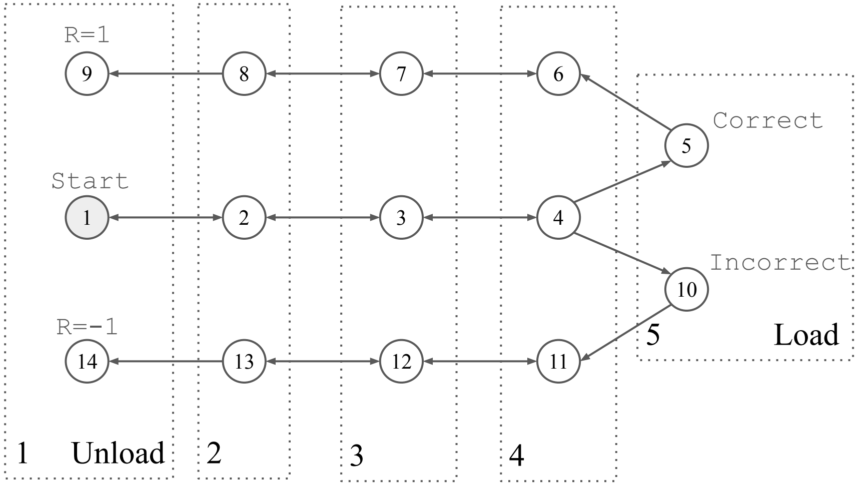
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
$N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





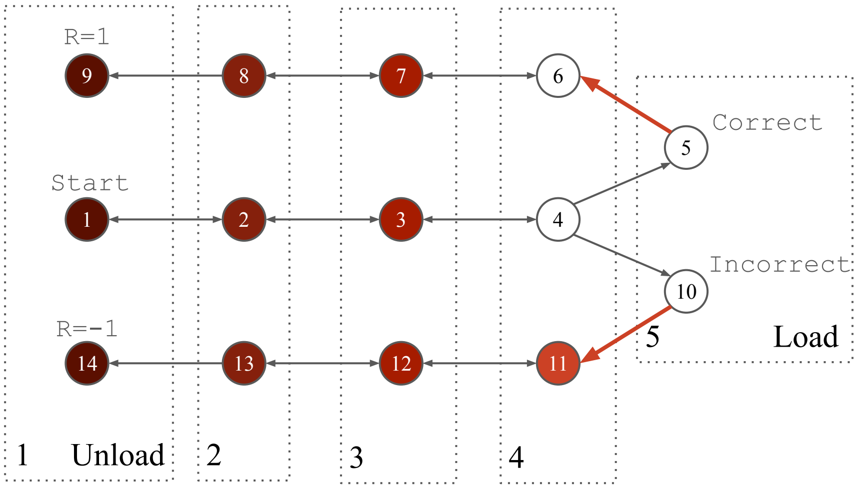
"SARSA will punish all state-action pairs **equally**"

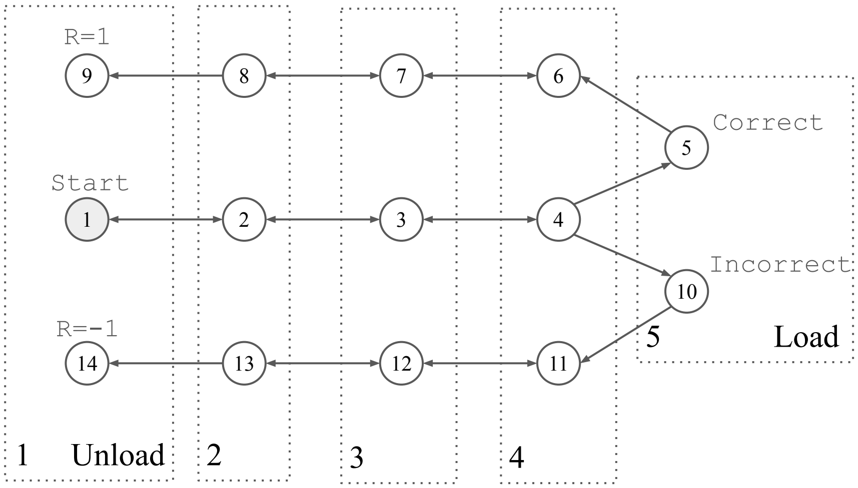
for $s \in S$
    for $a \in A$
        $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
        $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





"SARSA will punish all state-action pairs **equally**"

for $s \in S$

for $a \in A$

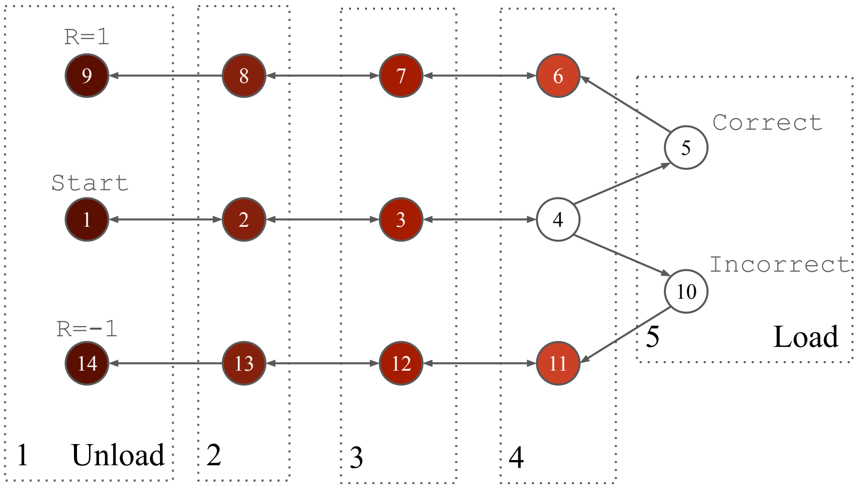
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$

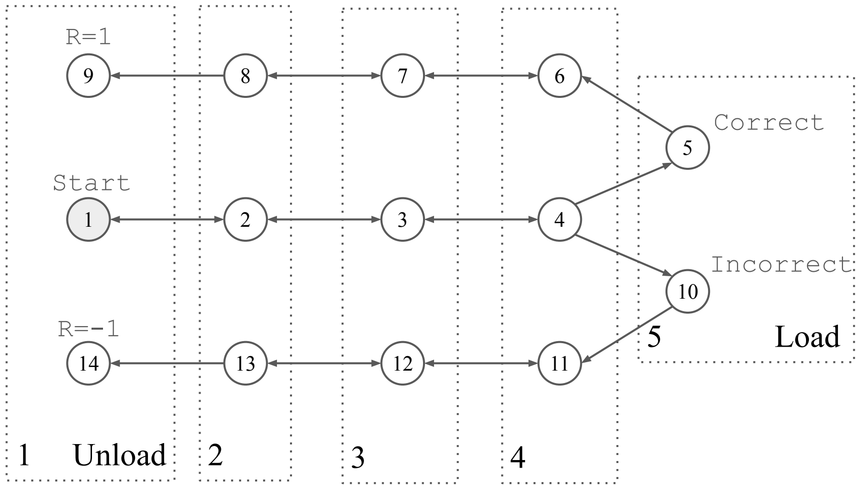
$N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





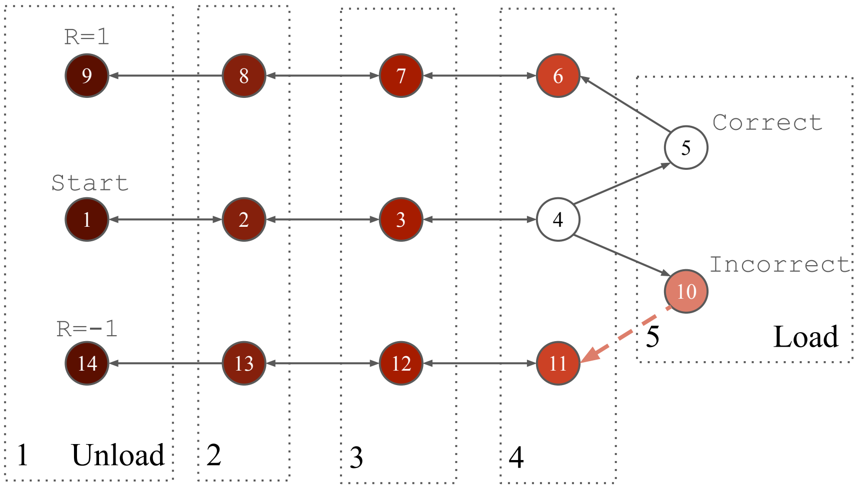
"SARSA will punish all state-action pairs **equally**"

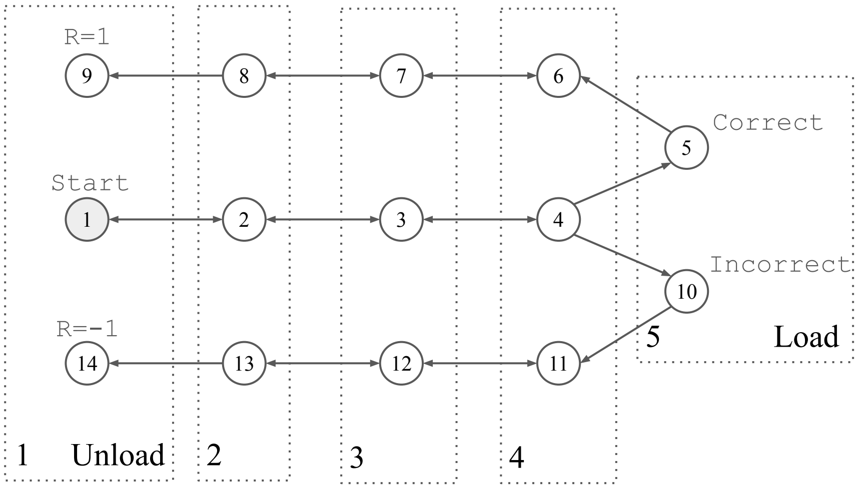
for $s \in S$
for $a \in A$
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





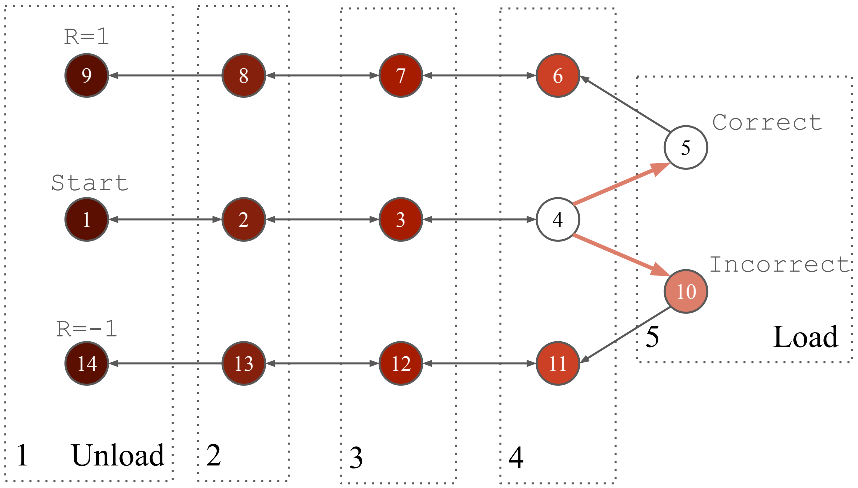
"SARSA will punish all state-action pairs **equally**"

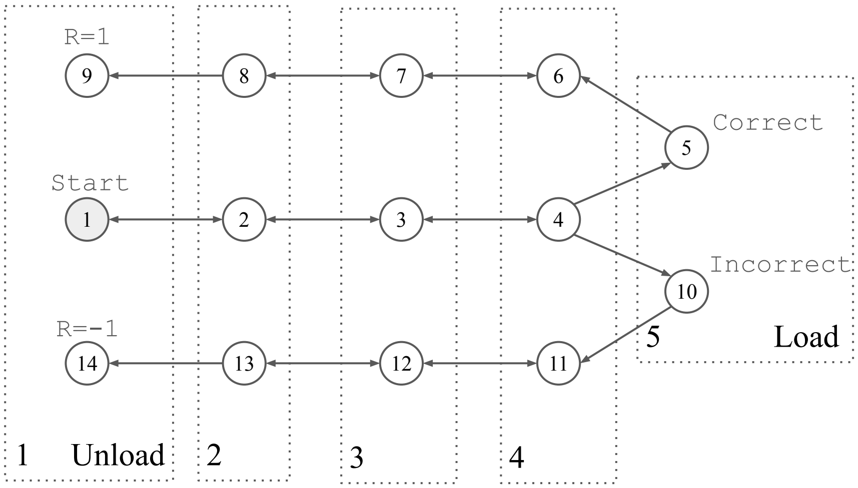
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
$N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





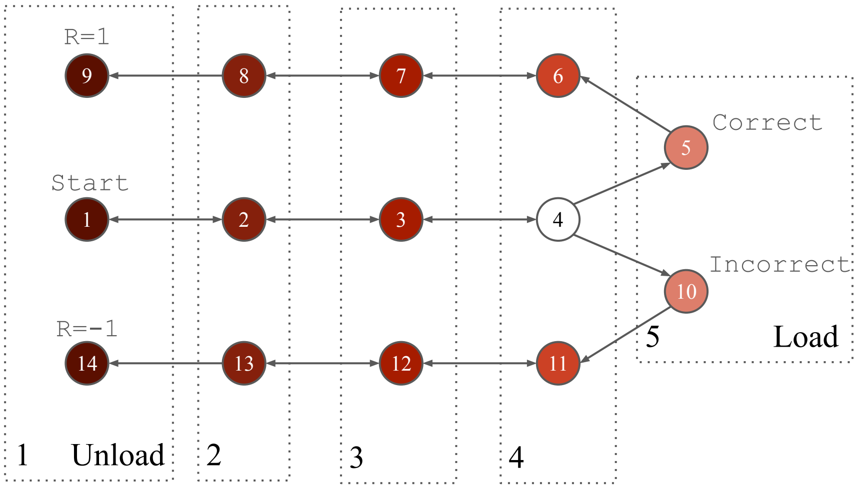
"SARSA will punish all state-action pairs **equally**"

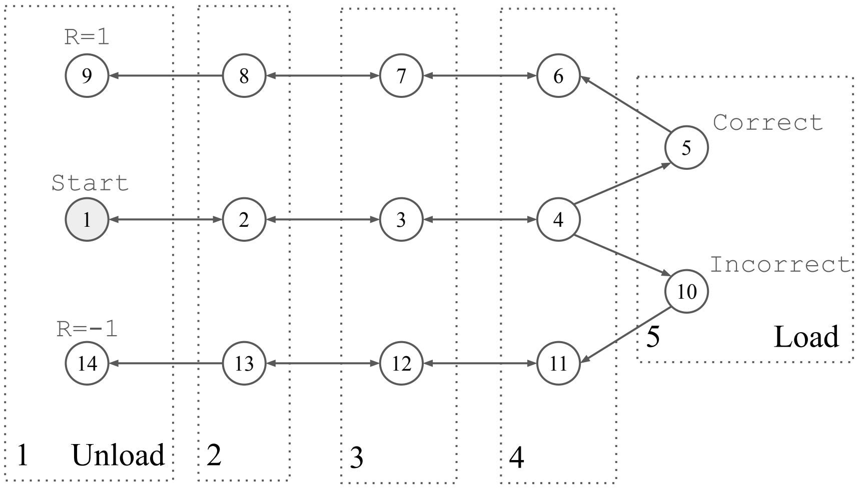
for $s \in S$
 for $a \in A$
 $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
 $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





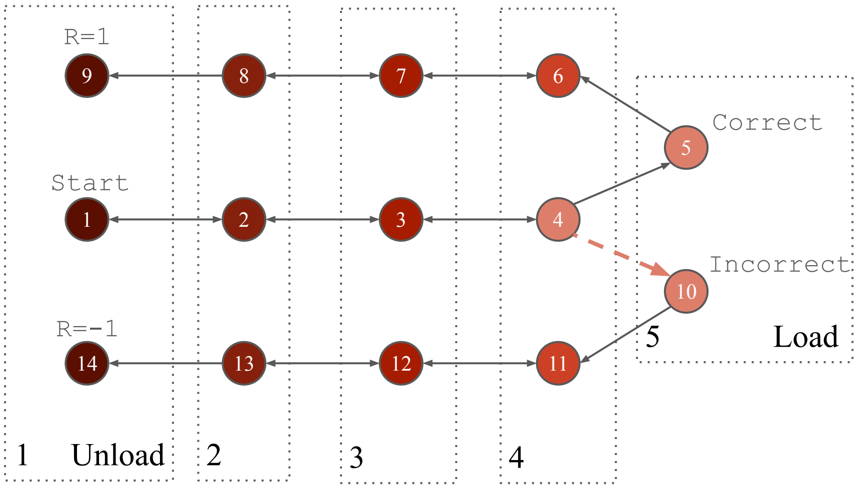
"SARSA will punish all state-action pairs **equally**"

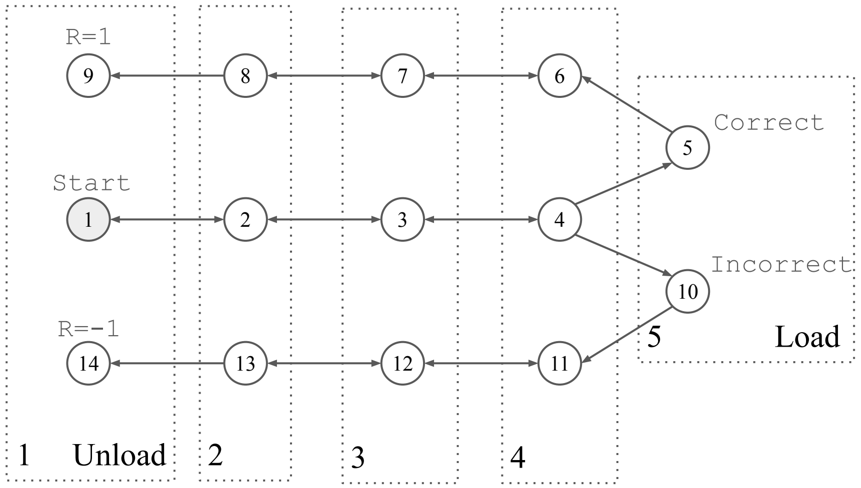
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
$N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





"SARSA will punish all state-action pairs **equally**"

for $s \in S$

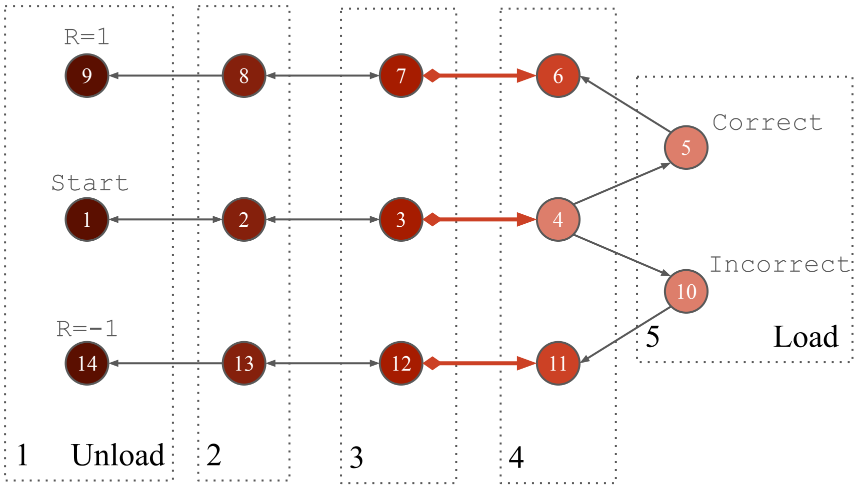
for $a \in A$

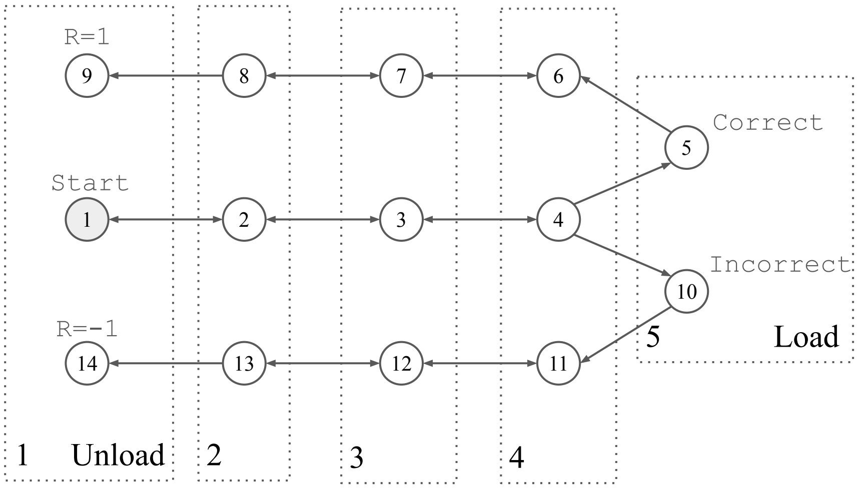
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





"SARSA will punish all state-action pairs **equally**"

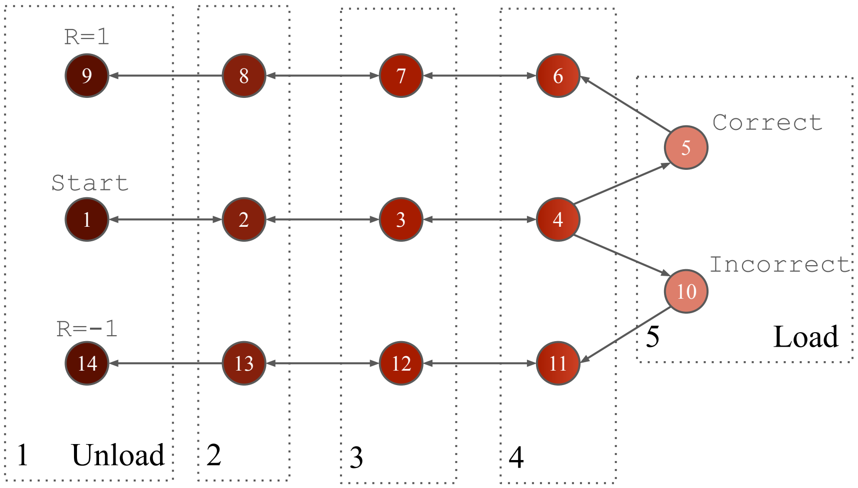
for $s \in S$
for $a \in A$

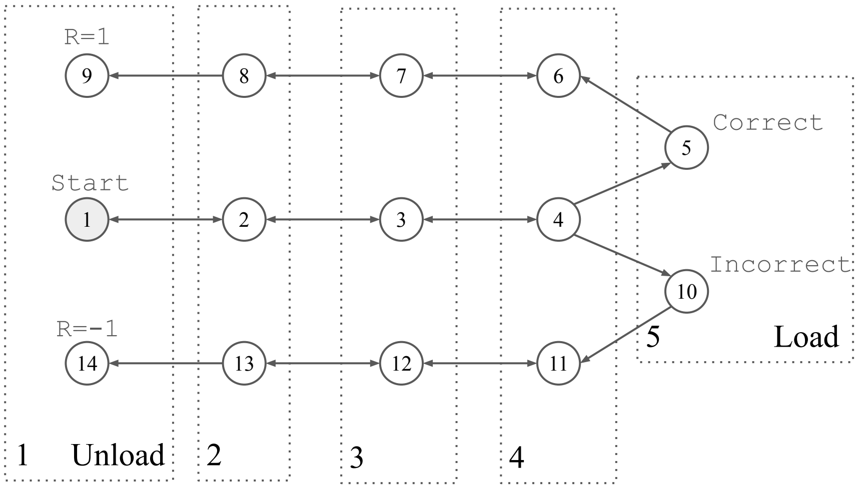
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





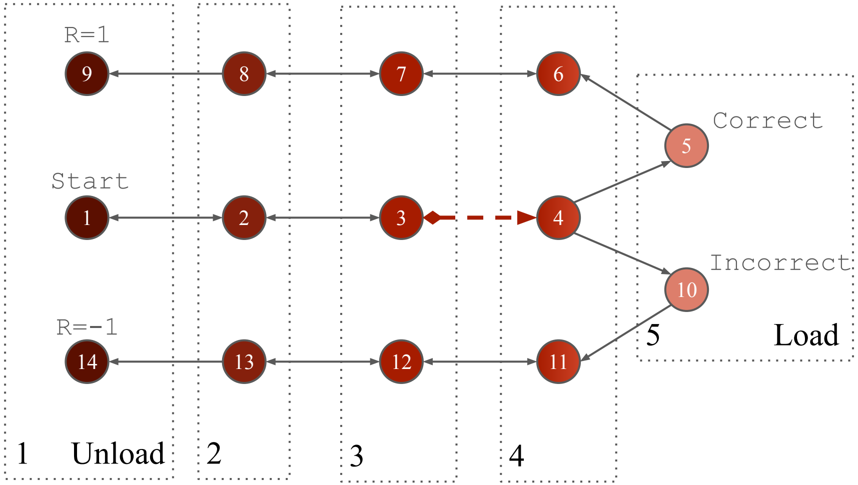
"SARSA will punish all state-action pairs **equally**"

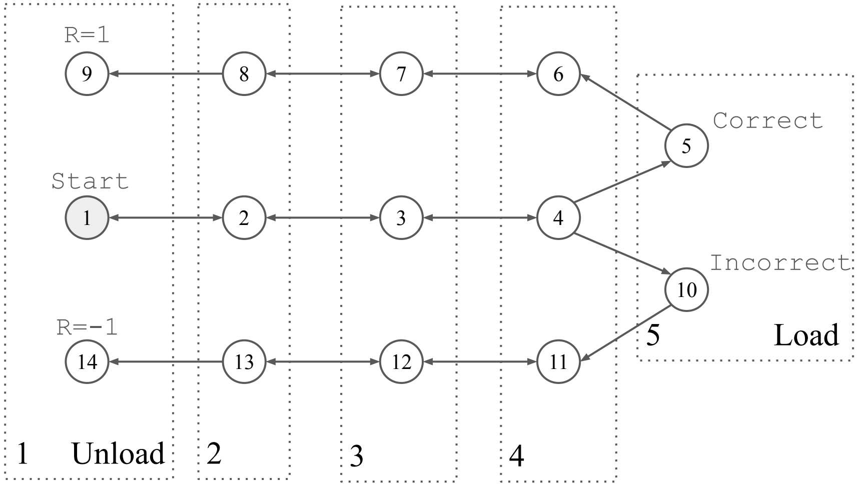
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
$N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





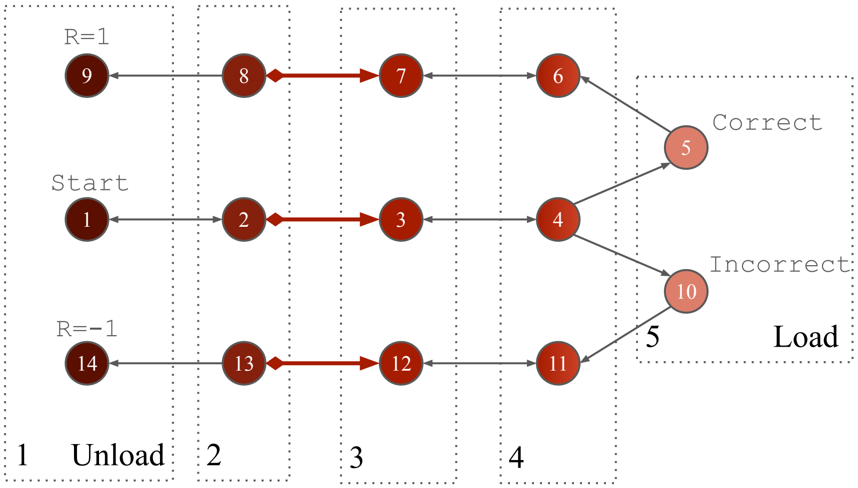
"SARSA will punish all state-action pairs **equally**"

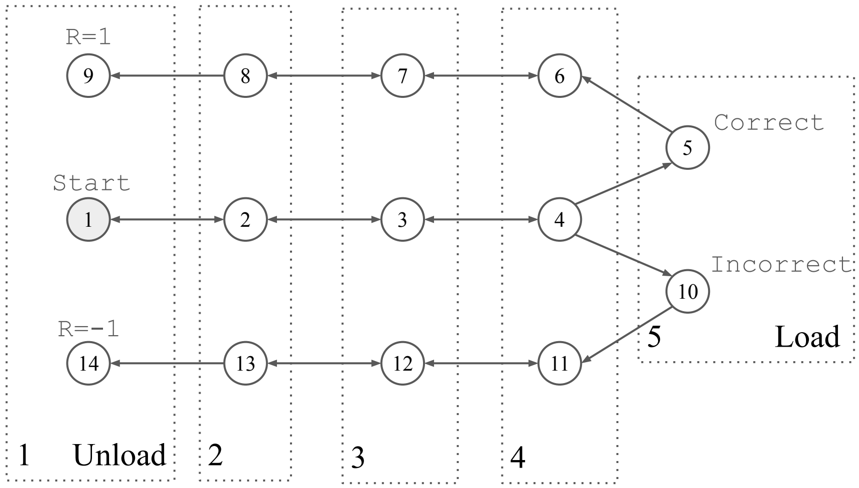
for $s \in S$
for $a \in A$
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





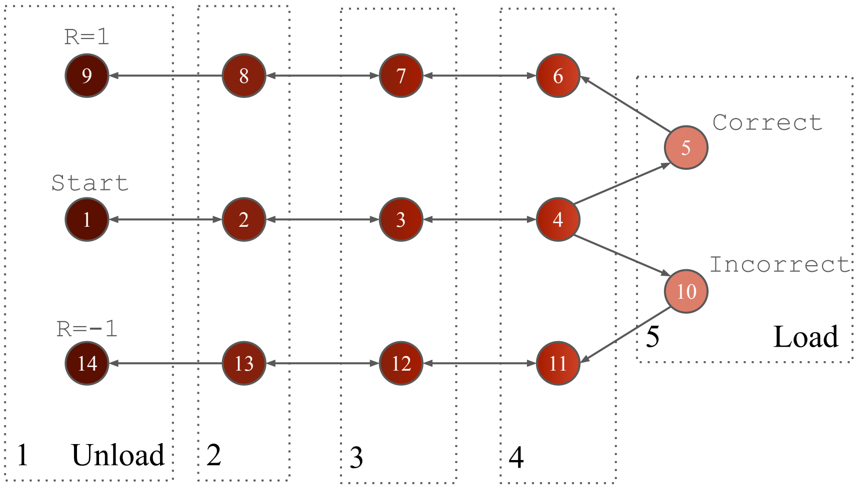
"SARSA will punish all state-action pairs **equally**"

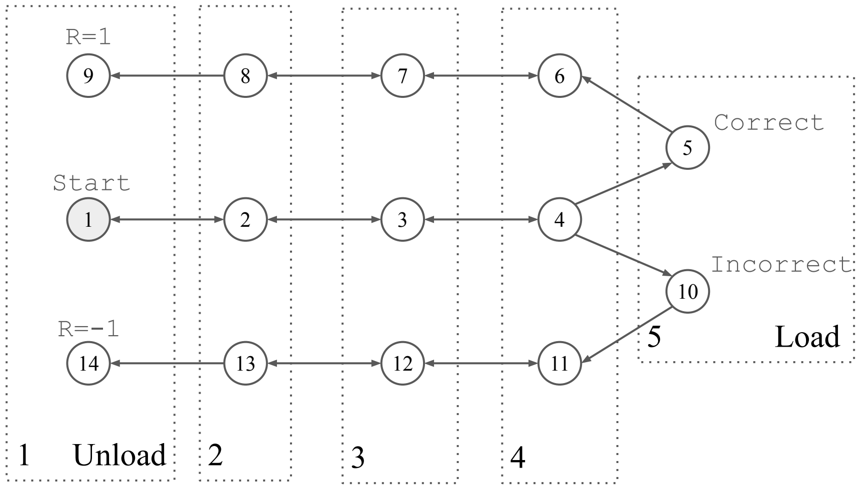
for $s \in S$
for $a \in A$
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





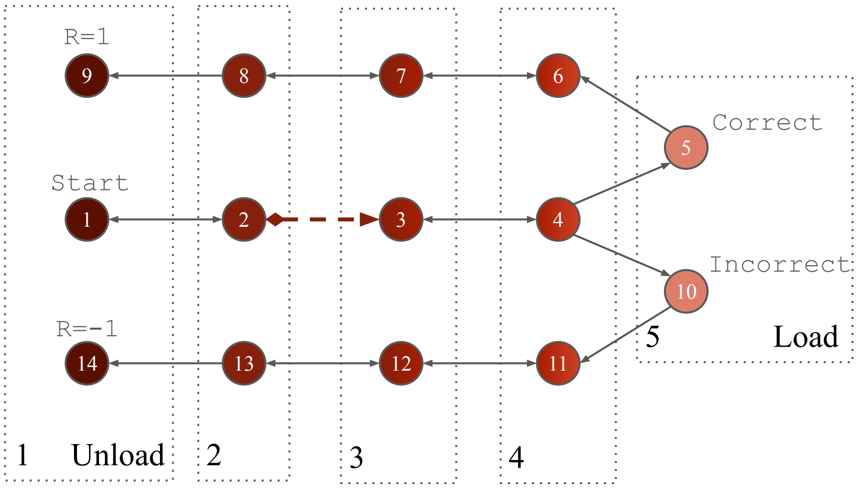
"SARSA will punish all state-action pairs **equally**"

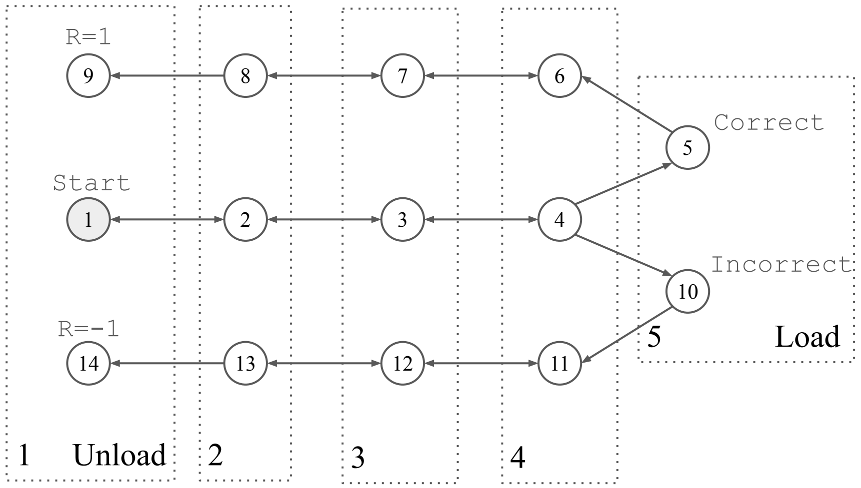
$$
\begin{aligned}
&\textbf{for } s \in S \\
&\quad \textbf{for } a \in A \\
&\qquad Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a) \\
&\qquad N(s,a) \leftarrow \gamma \lambda N(s,a)
\end{aligned}
$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





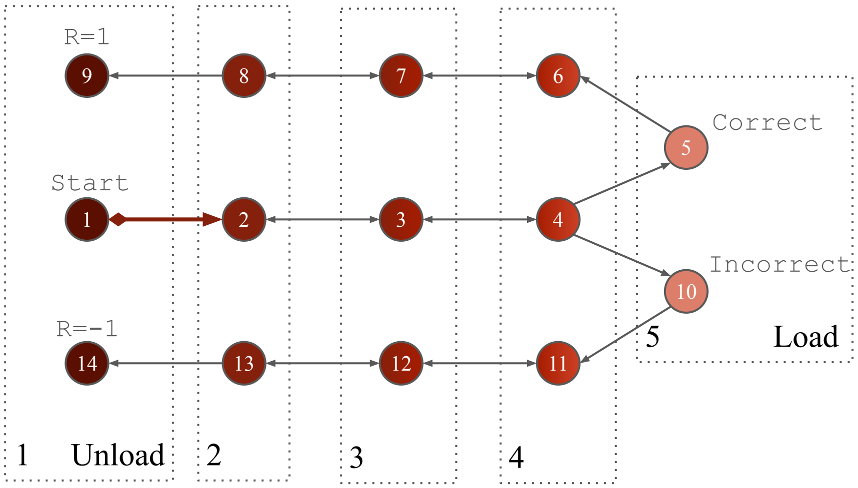
"SARSA will punish all state-action pairs **equally**"

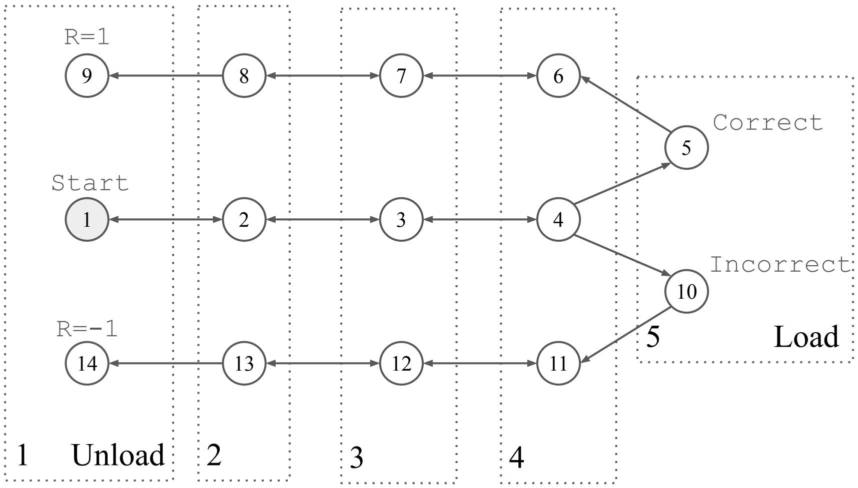
"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





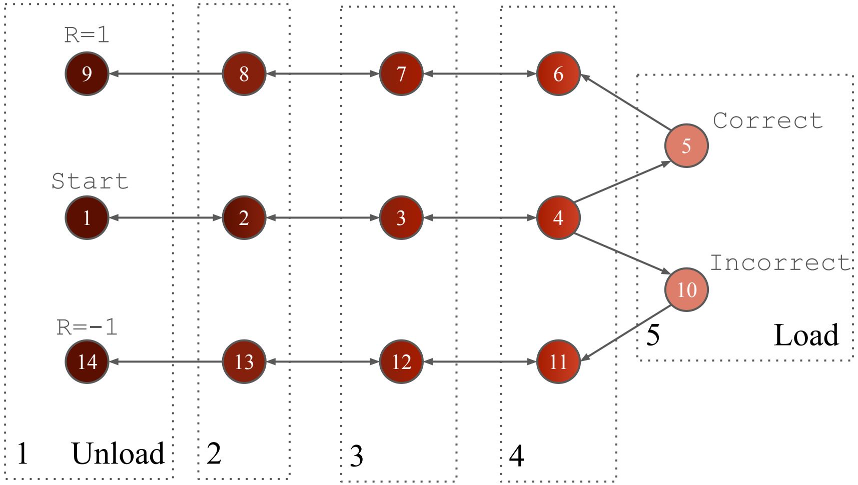
"SARSA will punish all state-action pairs **equally**"

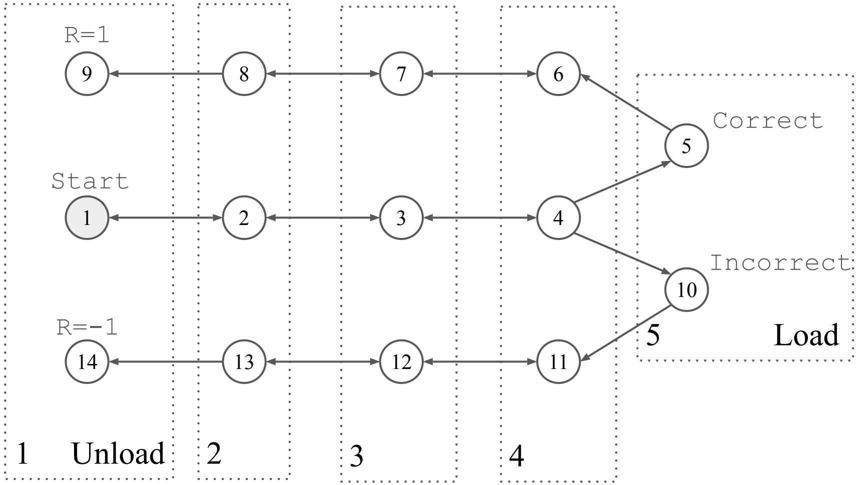
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





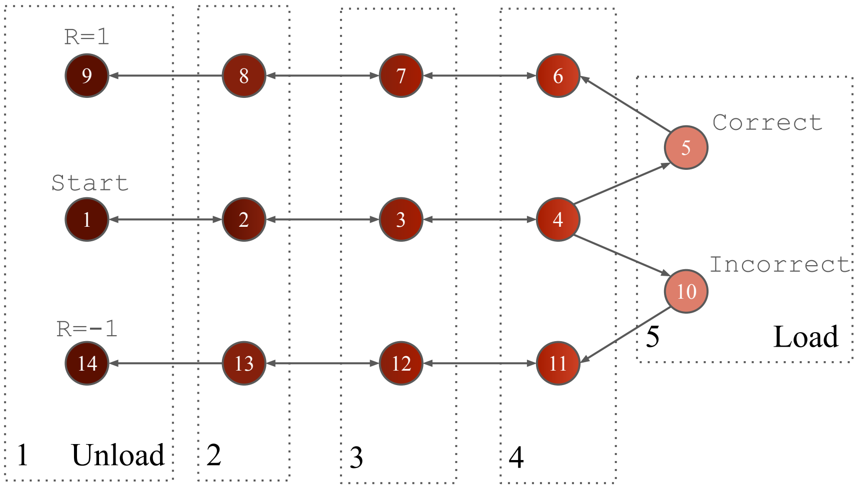
"SARSA will punish all state-action pairs **equally**"

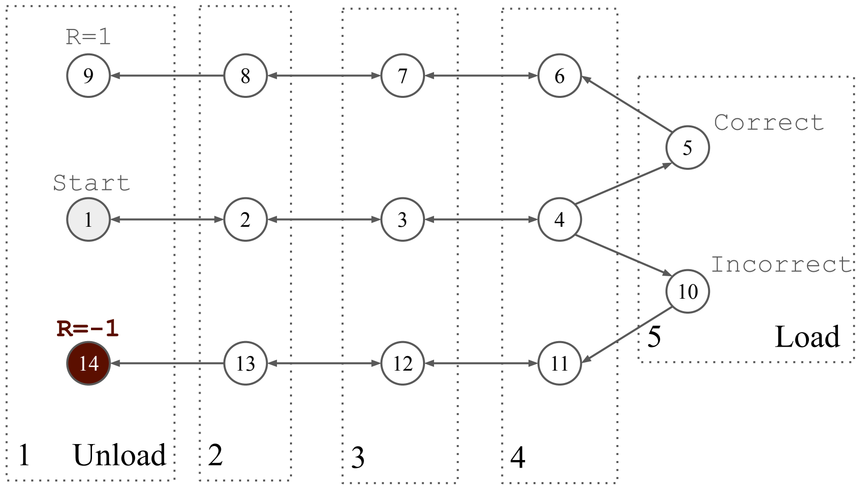
for $s \in S$
    for $a \in A$
        $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
        $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





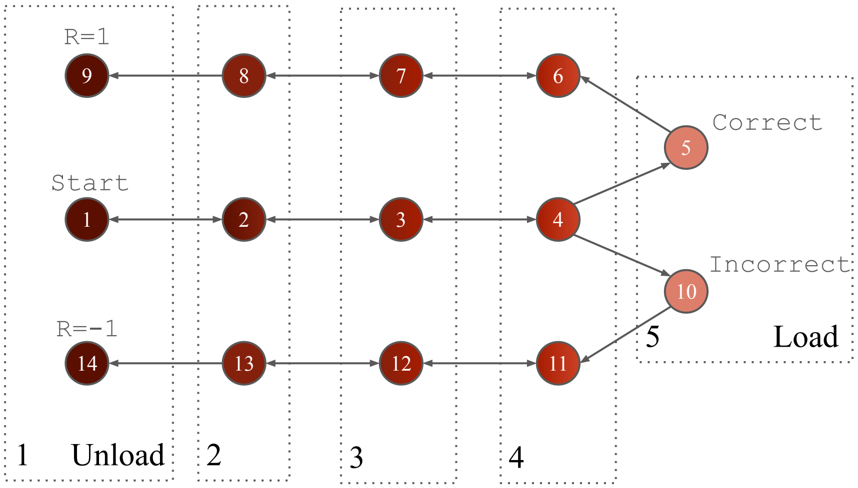
"SARSA will punish all state-action pairs **equally**"

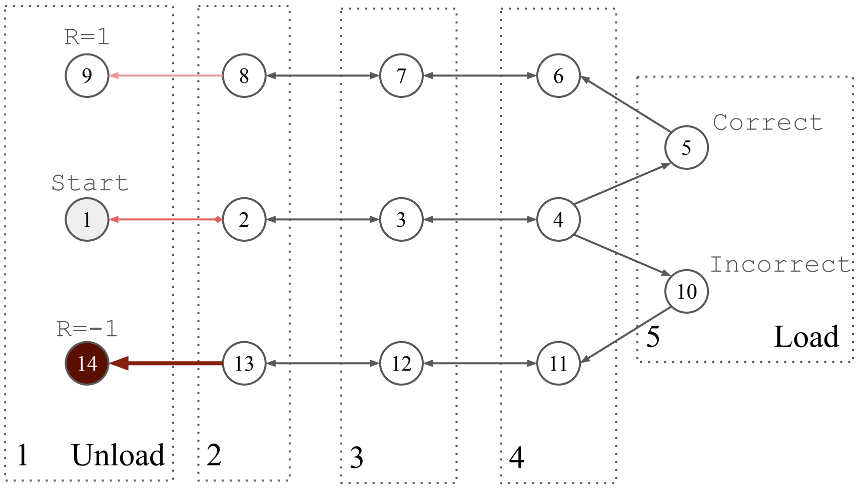
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





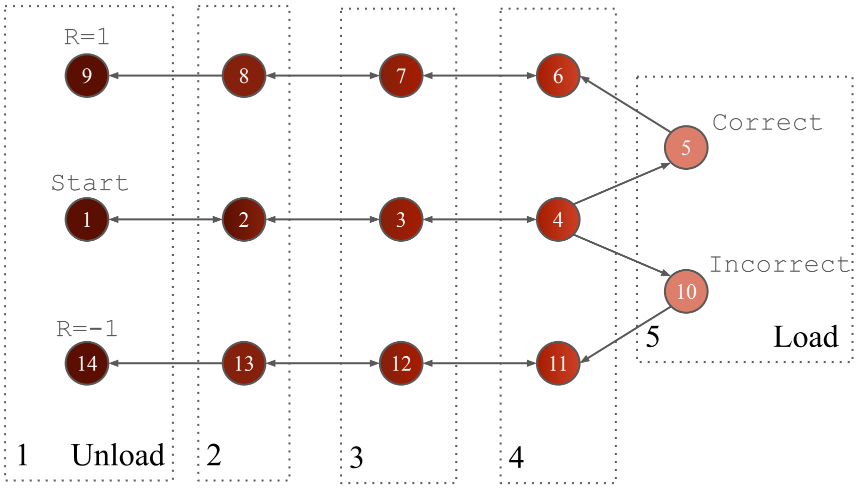
"SARSA will punish all state-action pairs **equally**"

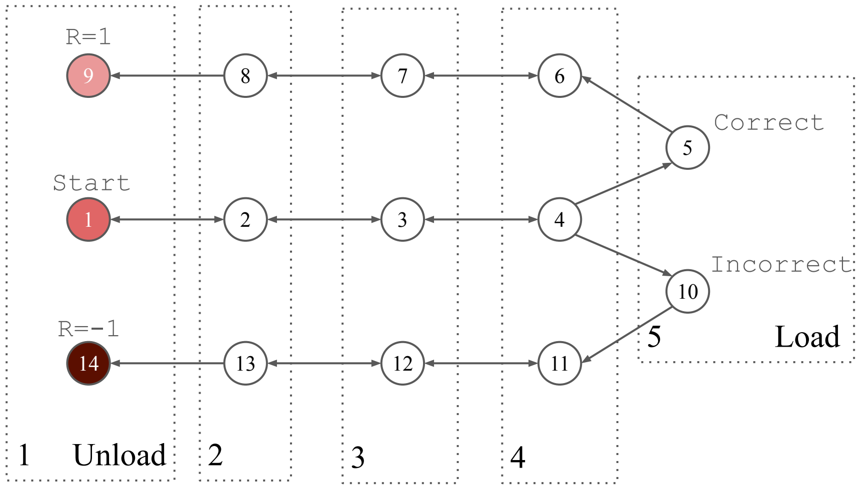
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





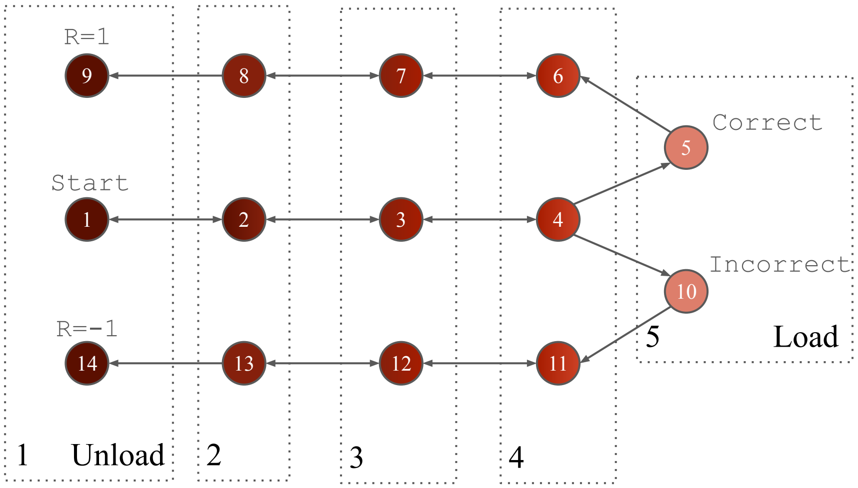
"SARSA will punish all state-action pairs **equally**"

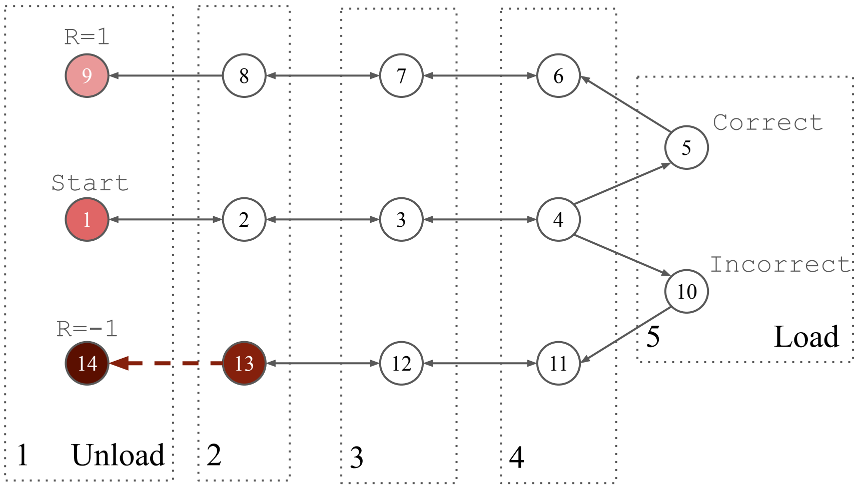
for $s \in S$
    for $a \in A$
        $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
        $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





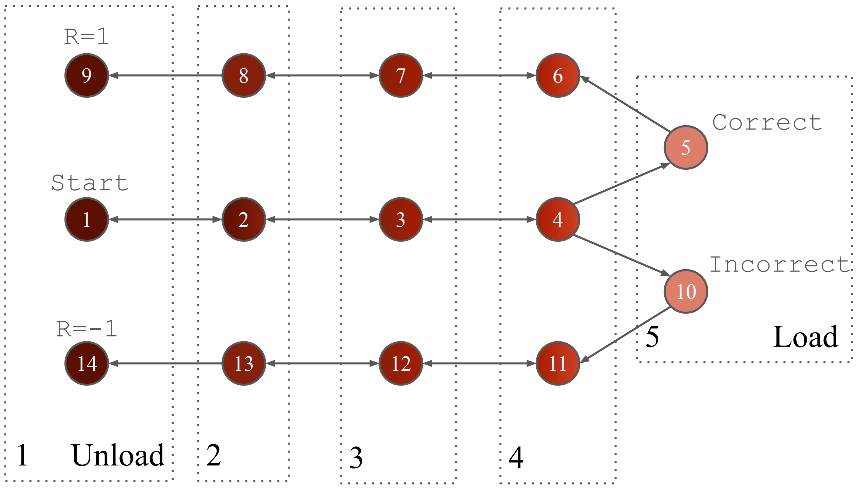
"SARSA will punish all state-action pairs **equally**"

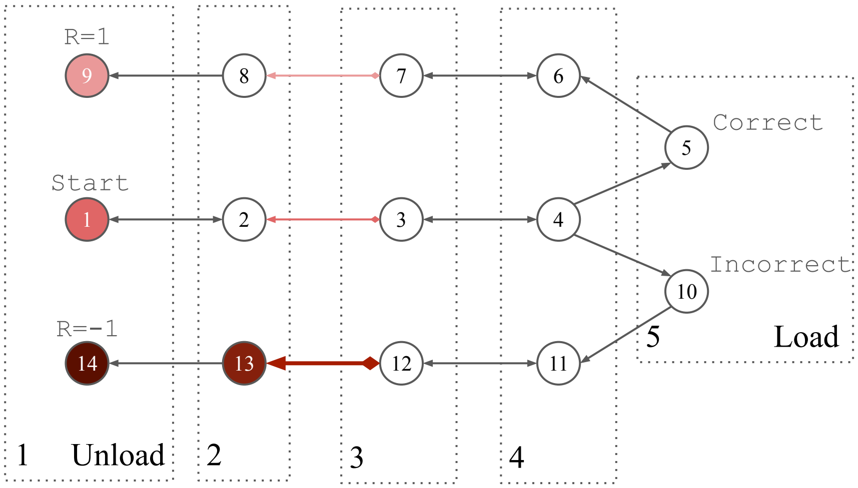
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
$N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





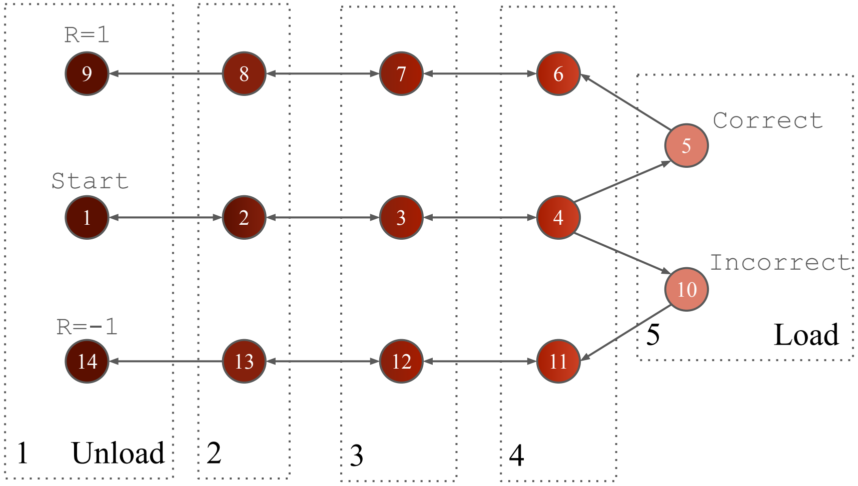
"SARSA will punish all state-action pairs **equally**"

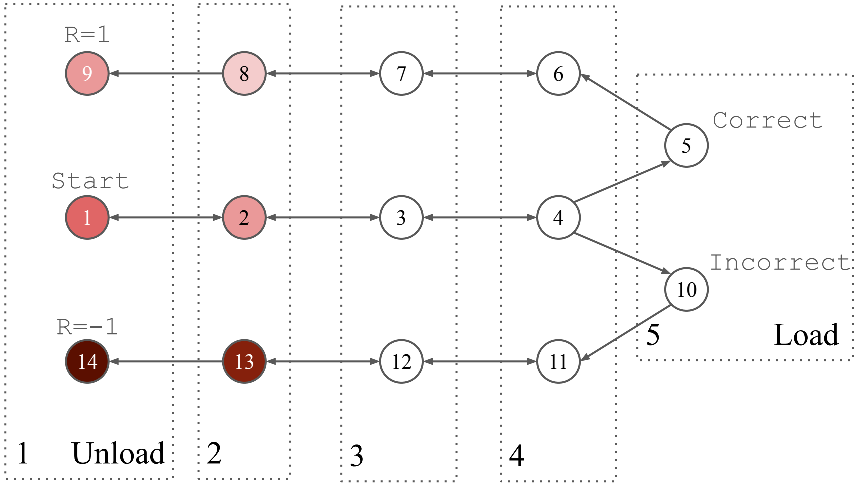
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





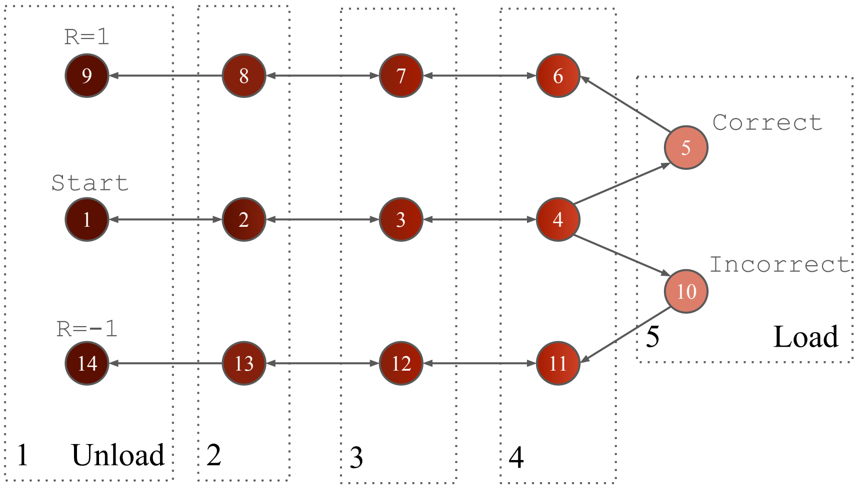
"SARSA will punish all state-action pairs **equally**"

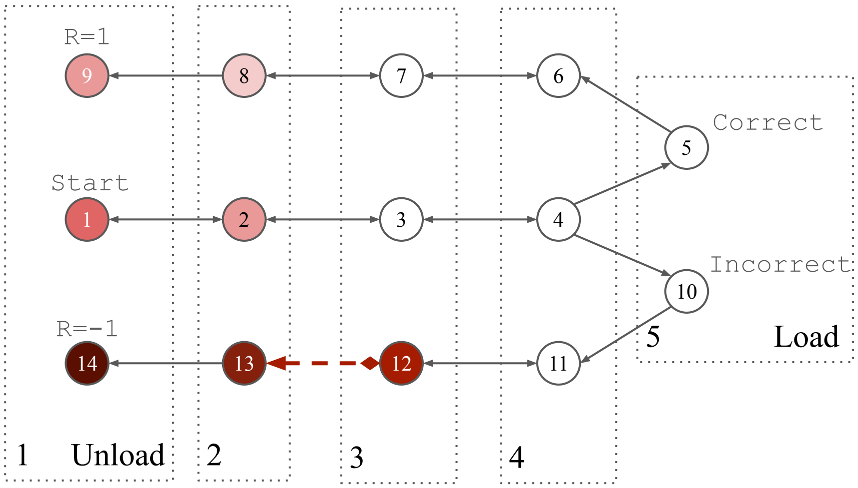
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





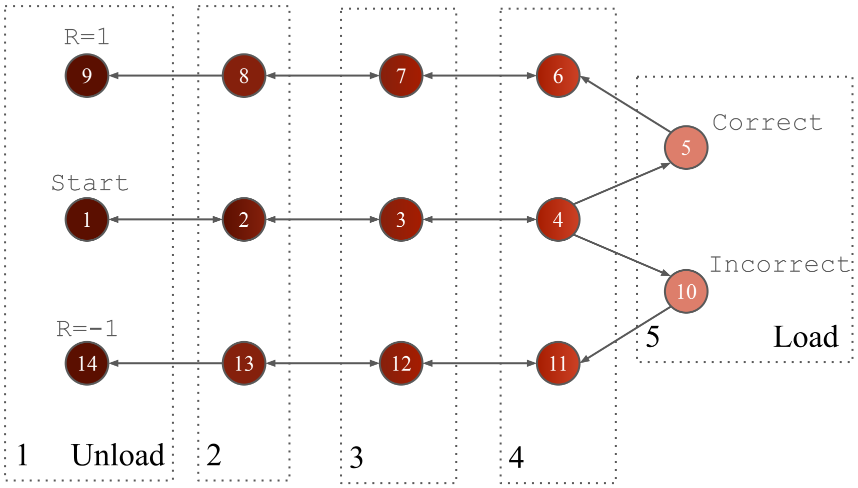
"SARSA will punish all state-action pairs **equally**"

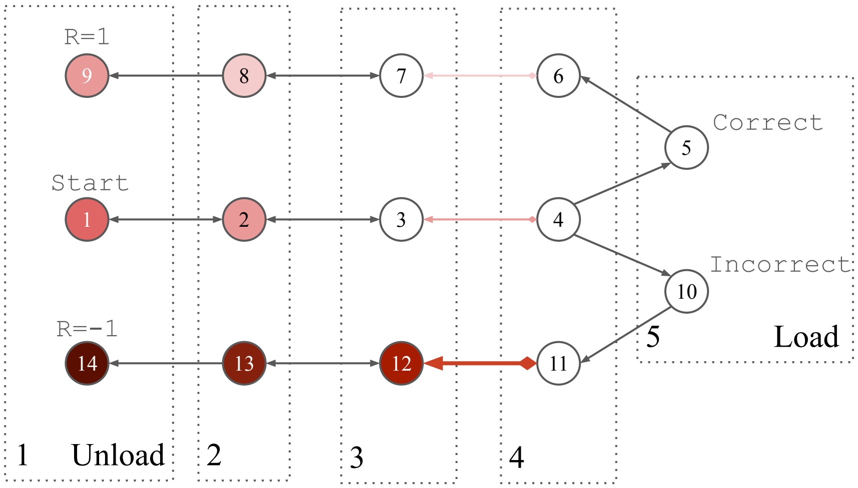
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





"SARSA will punish all state-action pairs **equally**"

for $s \in S$

for $a \in A$

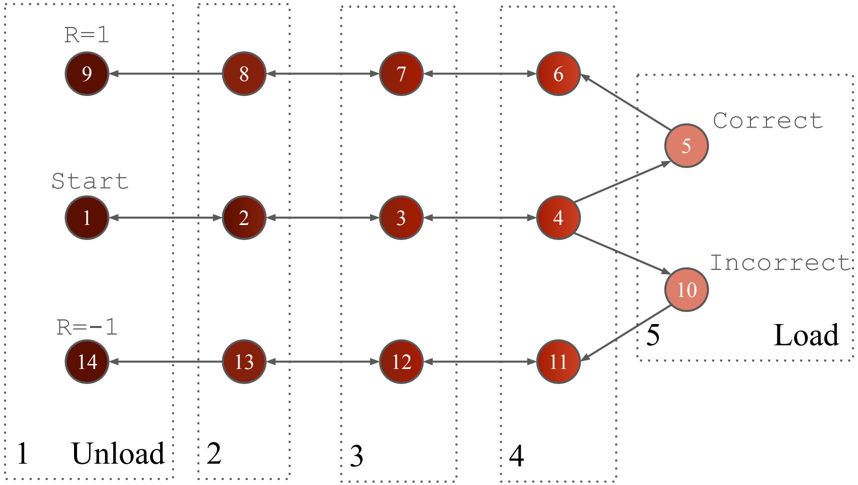
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$

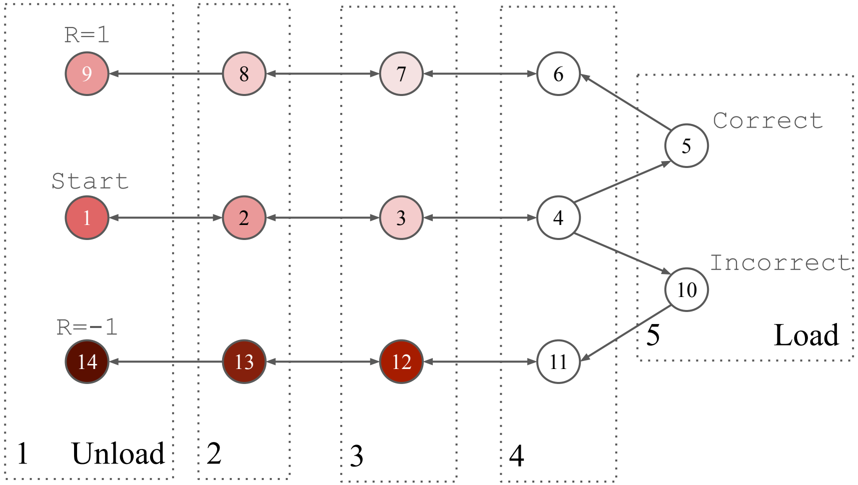
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





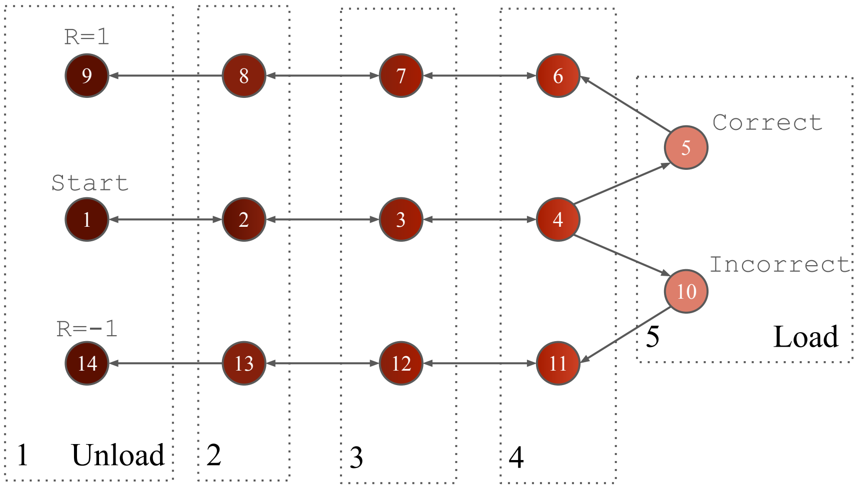
"SARSA will punish all state-action pairs **equally**"

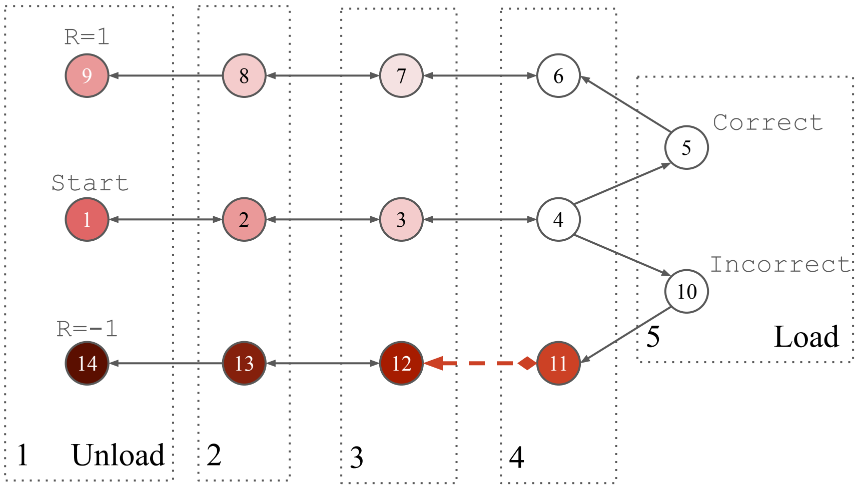
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
$N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





"SARSA will punish all state-action pairs **equally**"

for $s \in S$

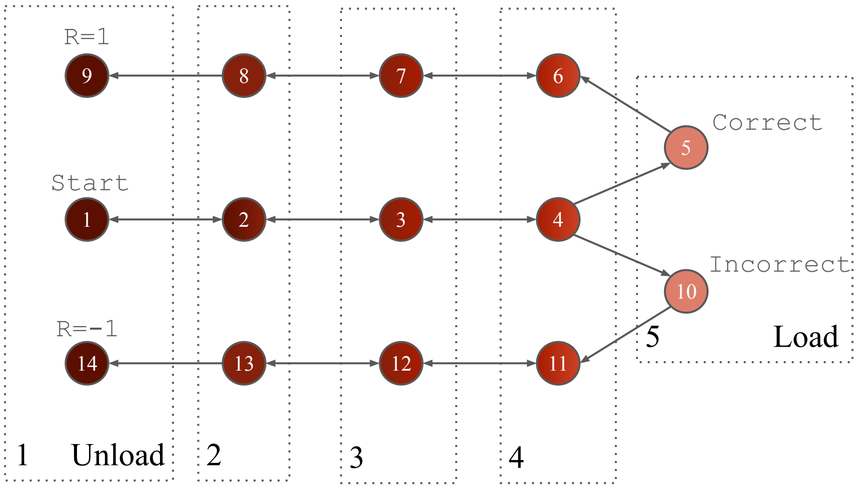
for $a \in A$

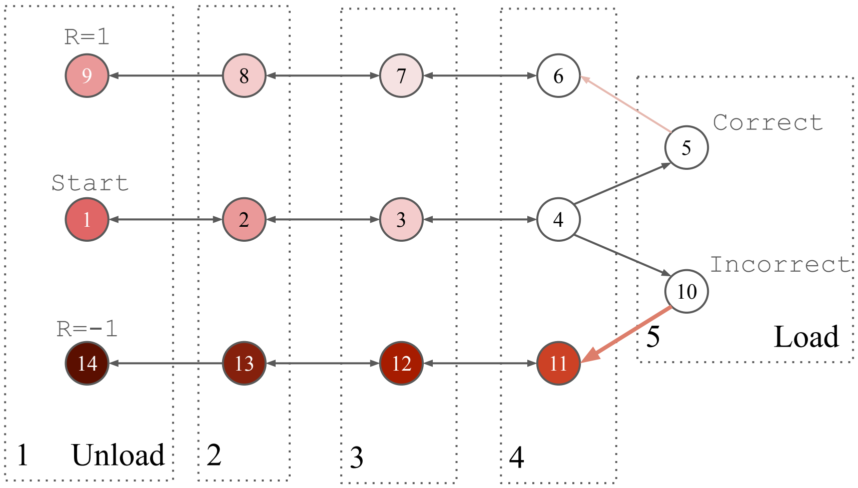
$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$$
$$N(s,a) \leftarrow \gamma \lambda N(s,a)$$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





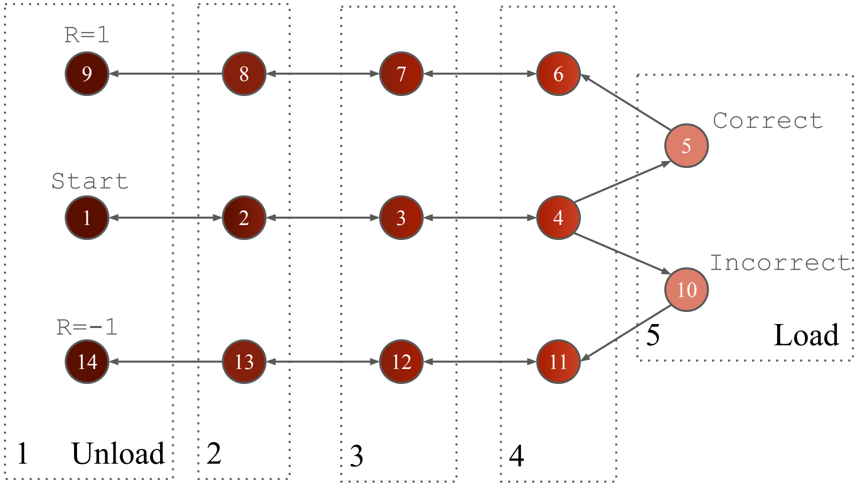
"SARSA will punish all state-action pairs **equally**"

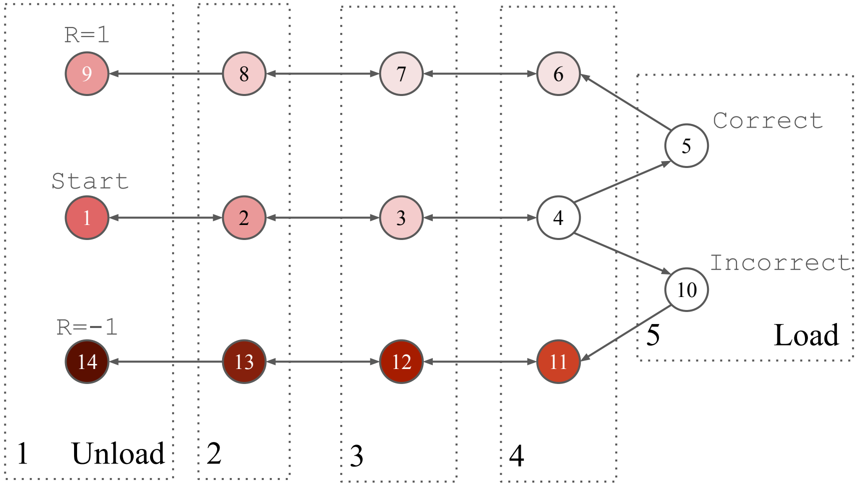
"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
$N(s,a) \leftarrow \gamma\lambda N(s,a)$

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





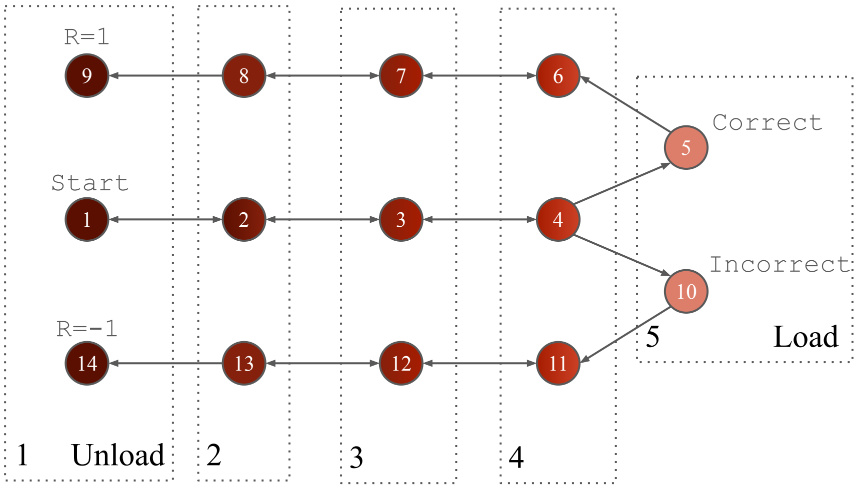
"SARSA will punish all state-action pairs **equally**"

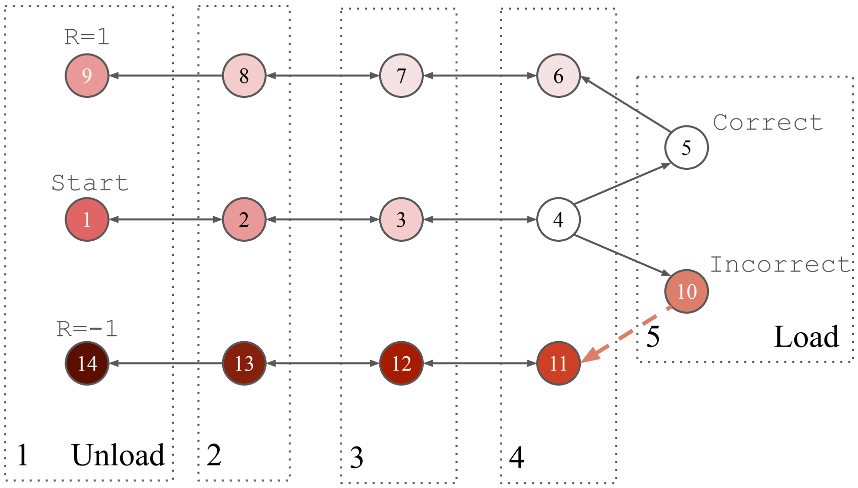
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





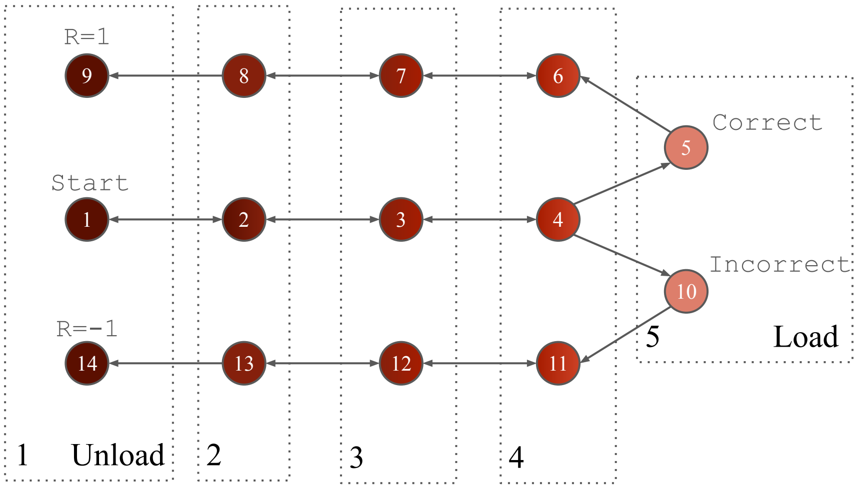
"SARSA will punish all state-action pairs **equally**"

for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

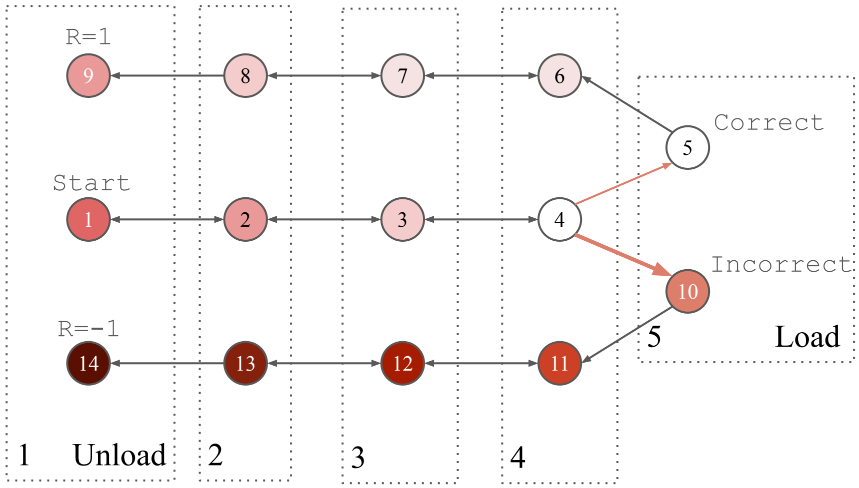
$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization



“Sarsa will punish all state-action pairs **equally**”

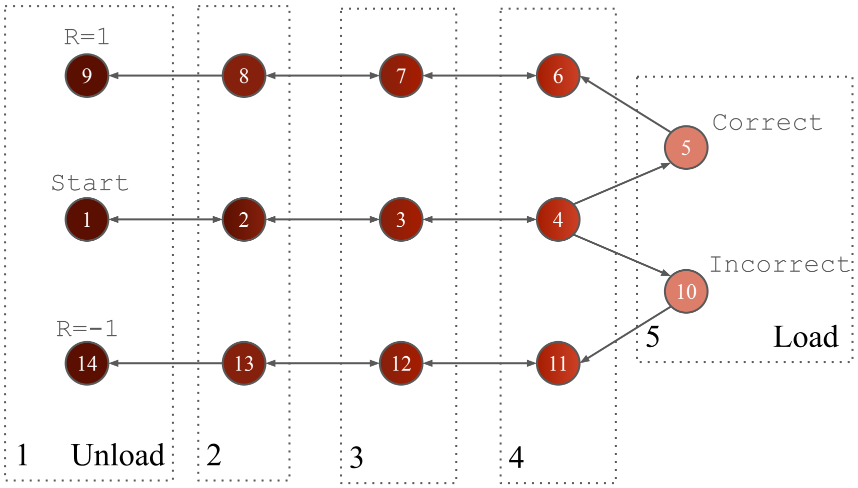
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
$N(s,a) \leftarrow \gamma \lambda N(s,a)$

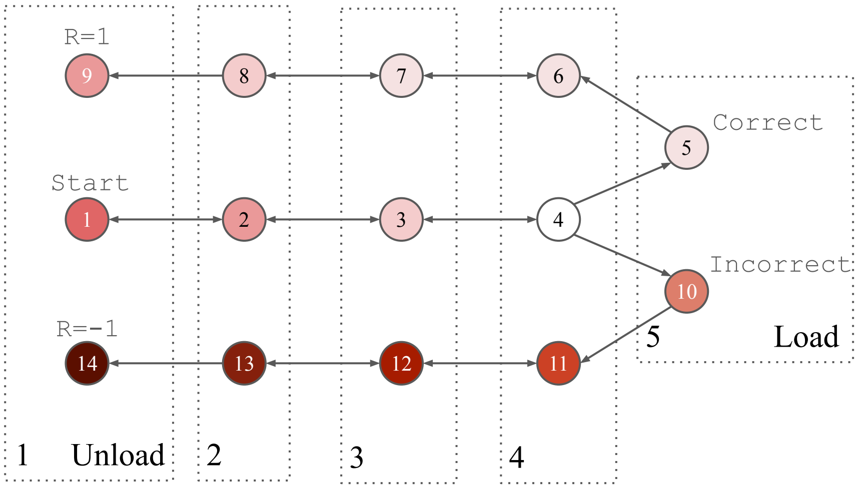


“VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**”

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





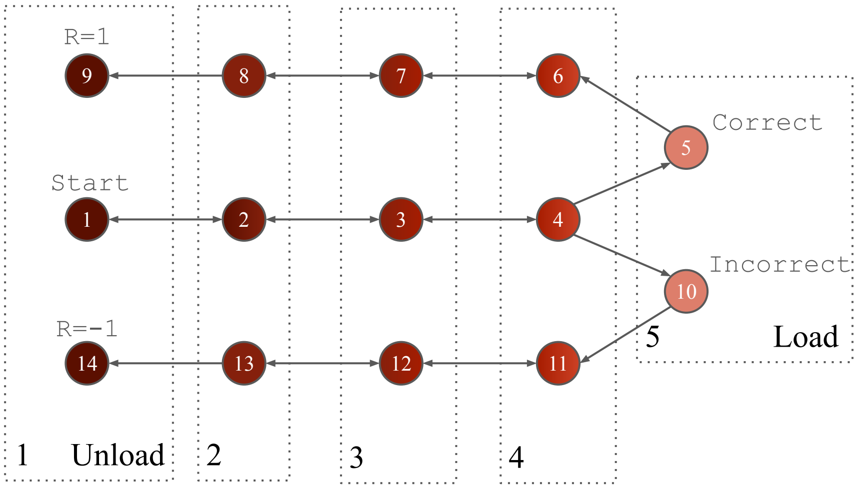
"Sarsa will punish all state-action pairs **equally**"

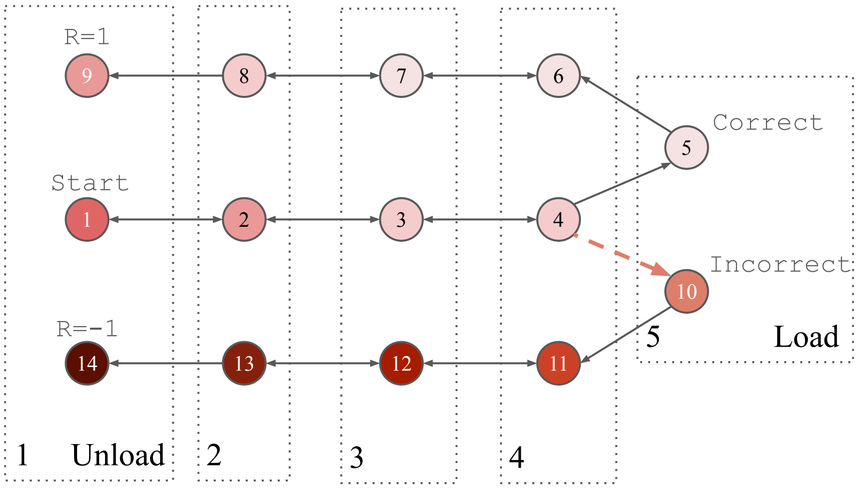
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





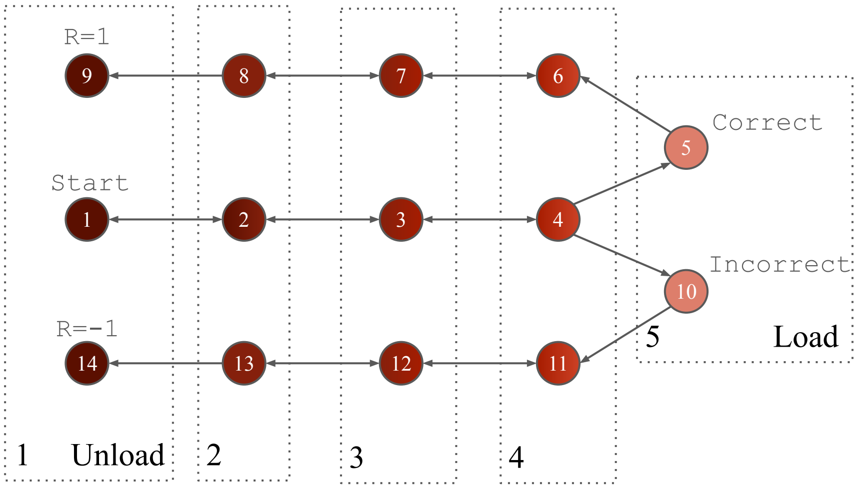
"SARSA will punish all state-action pairs **equally**"

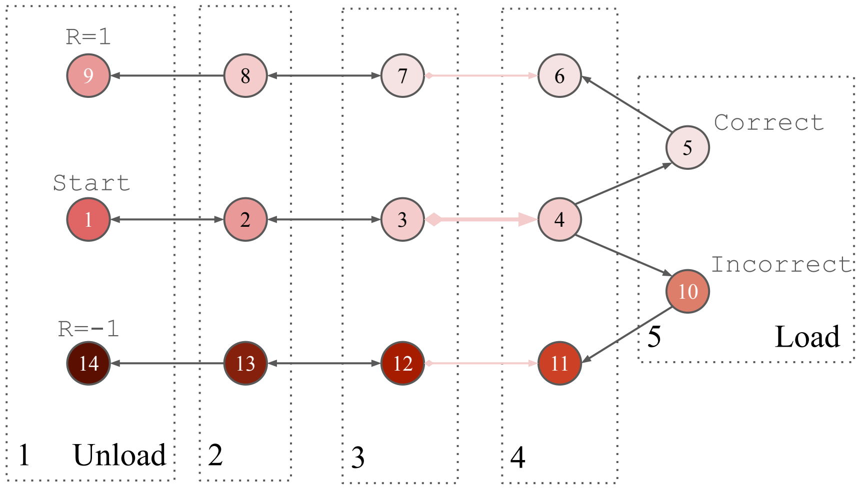
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





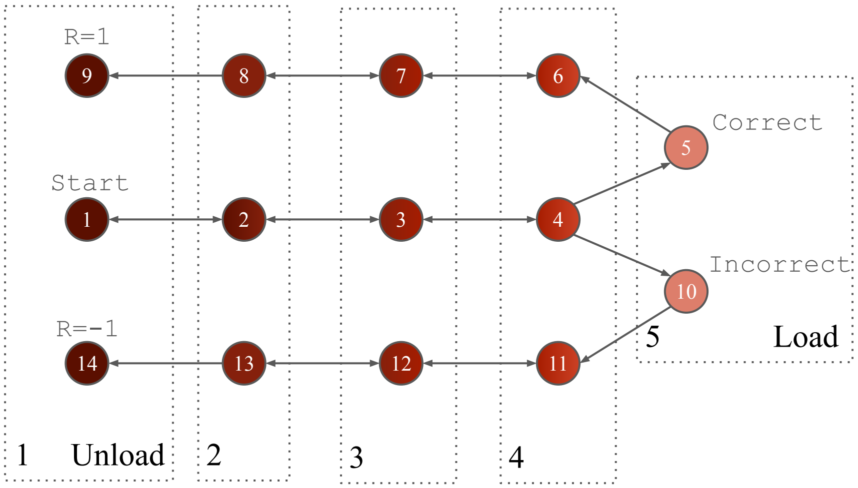
"SARSA will punish all state-action pairs **equally**"

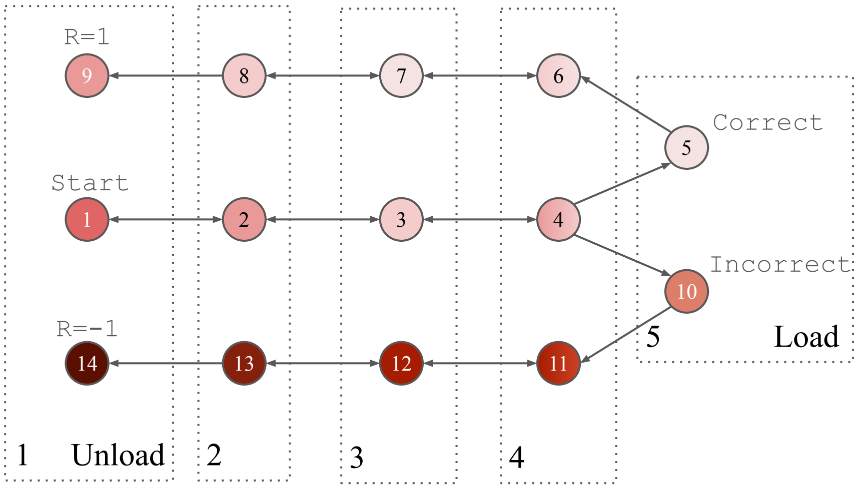
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





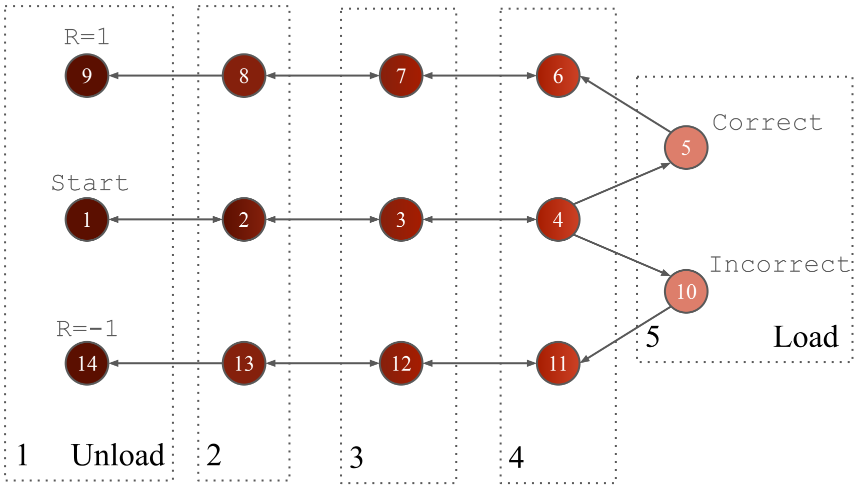
"SARSA will punish all state-action pairs **equally**"

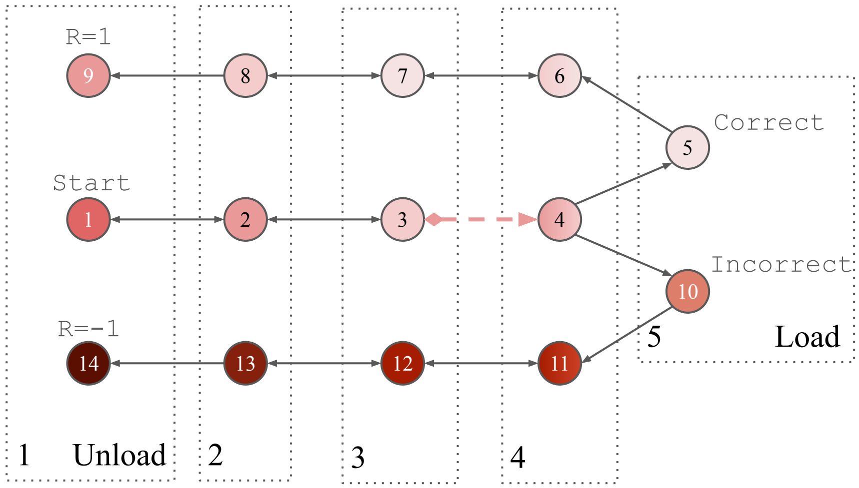
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





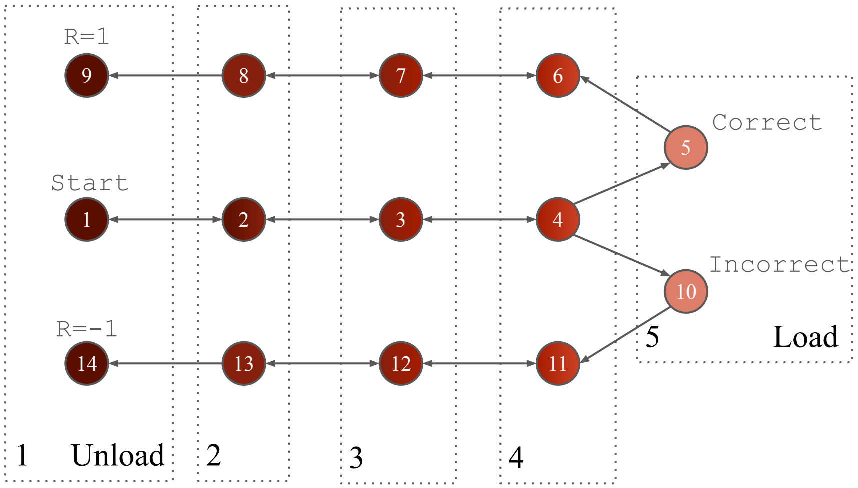
"SARSA will punish all state-action pairs **equally**"

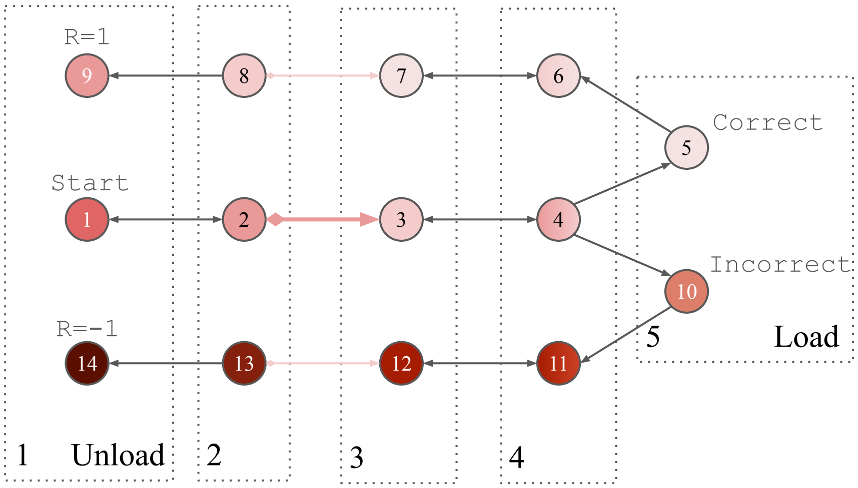
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
$N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





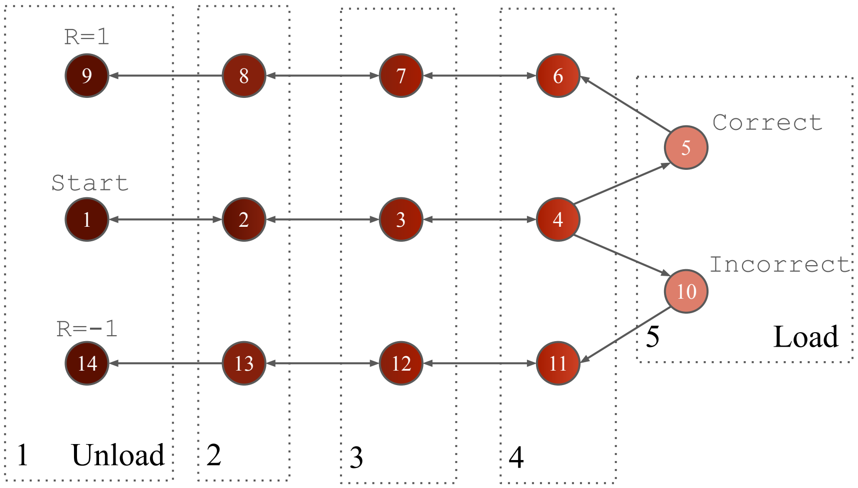
"SARSA will punish all state-action pairs **equally**"

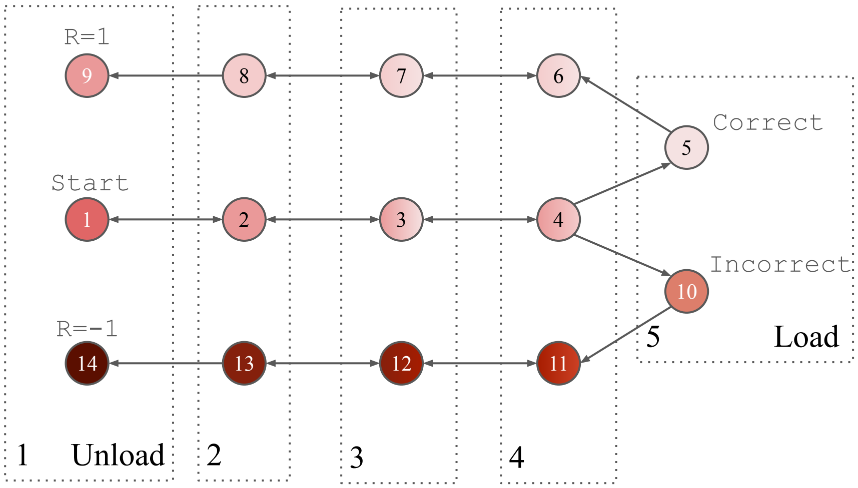
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





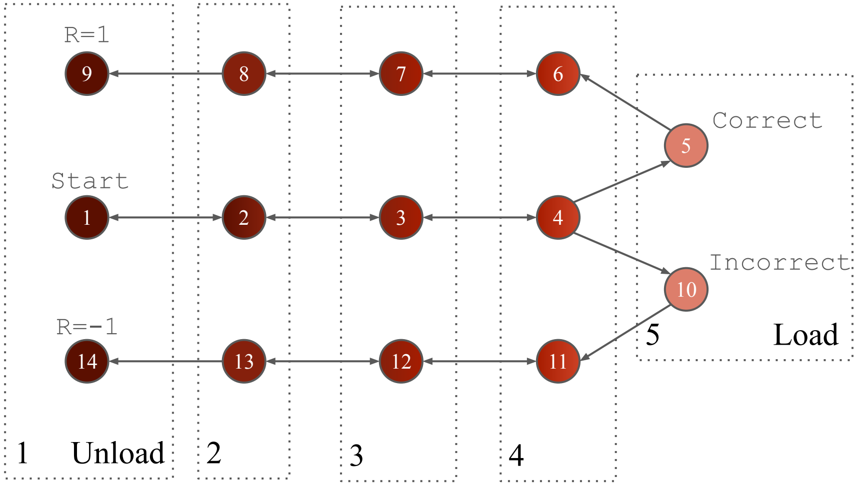
"SARSA will punish all state-action pairs **equally**"

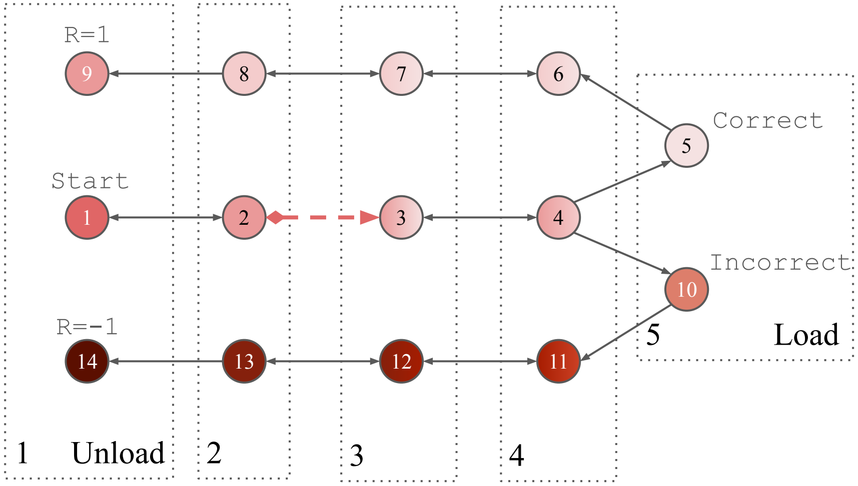
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
    $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





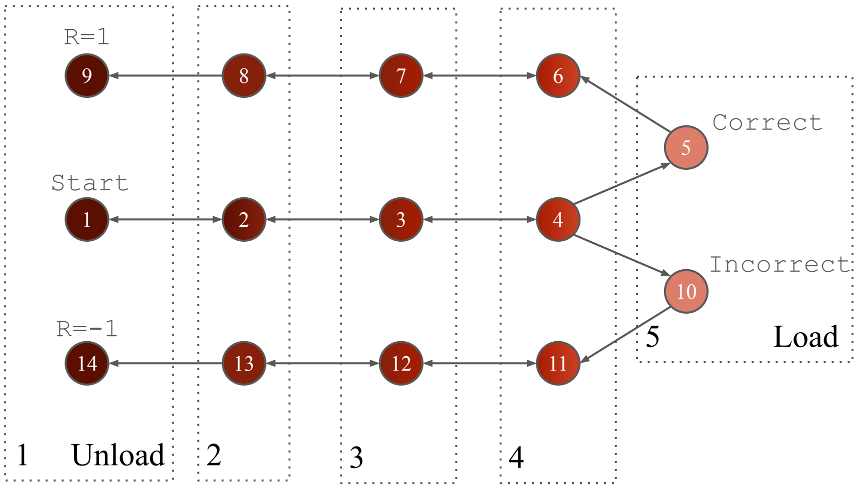
"Sarsa will punish all state-action pairs **equally**"

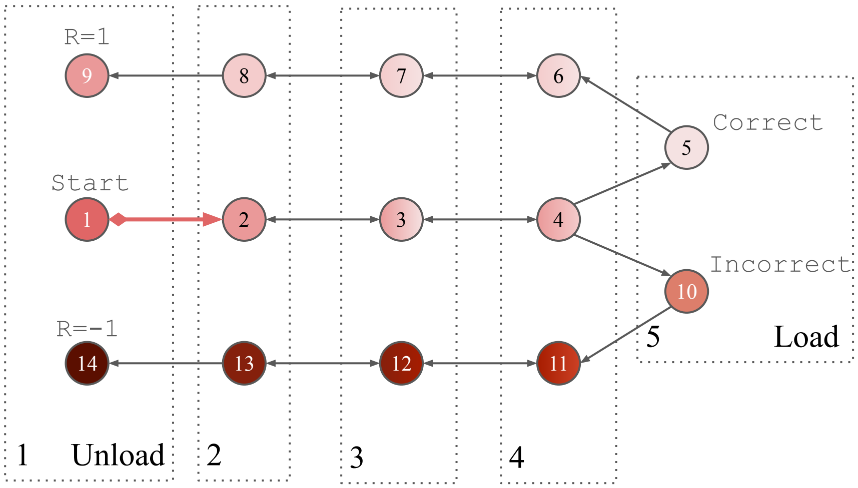
for $s \in S$
  for $a \in A$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
    $N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N_{s,a}^{t} - E\left[N_{s,a}^{t}\right]\right]$$

Intuition Visualization





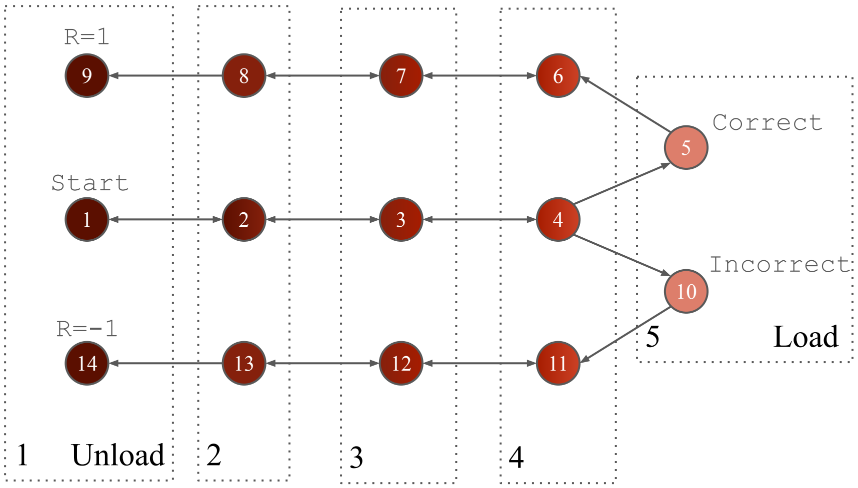
"SARSA will punish all state-action pairs **equally**"

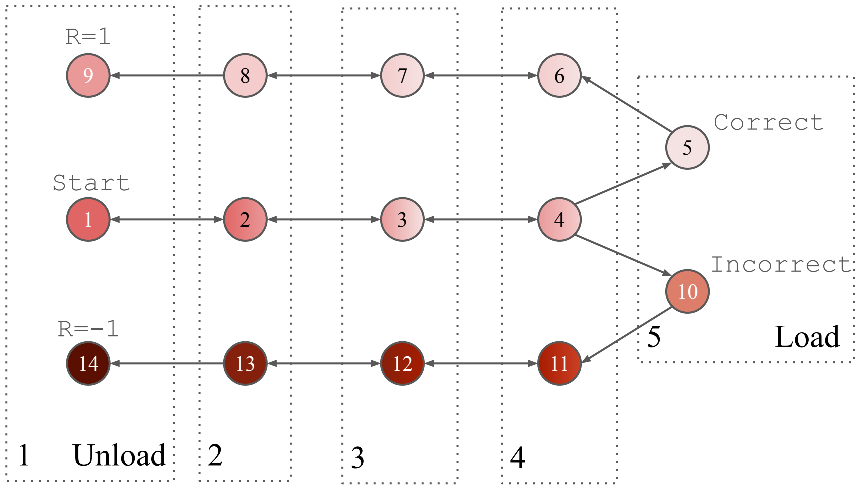
for $s \in S$
for $a \in A$
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a)$
$N(s,a) \leftarrow \gamma \lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Intuition Visualization





"SARSA will punish all state-action pairs **equally**"

for $s \in S$
    for $a \in A$
        $Q(s,a) \leftarrow Q(s,a) + \alpha\delta N(s,a)$
        $N(s,a) \leftarrow \gamma\lambda N(s,a)$

"VAPS(1) will punish the **bad state-action pair more** due to the **different principle of credit assignment**"

$$\underbrace{\Phi_{s,a,t}}_{\text{exploration trace}} = \frac{1}{c}\left[N^t_{s,a} - E\left[N^t_{s,a}\right]\right]$$

Critique / Limitations

Critiques:

- Shallow explanation of interesting results!
- Lack of more detailed aggregate results (no tables of results in the paper)
- Mentions a computational speed-up, but does not provide metrics to prove (lookup?)

Limitations:

- Authors do not mention specific limitations, but should explore scalability

Petty critiques:

- Why initially use a to represent an action, then switch to u inexplicably?
- Using x for both *state* and *observation*
- Minor math inconsistencies and typos (u^t instead of u_t, reuse of T, c_{min} vs $c_{\min}$)
- Lack of consistent `\mid` usage: $Pr(u_t = u | x_t = x)$ *(the absolute horror!)*

Future Work for Paper/Reading

- The authors did not provide any future work outline
- It would be interesting to see stigmergic approaches applied to larger problems
- Interested in $\mathrm{VAPS}(\beta)$ where β is swept; similar to the original VAPS paper [1]
    - Applied to the *load-unload* problem with two loading locations

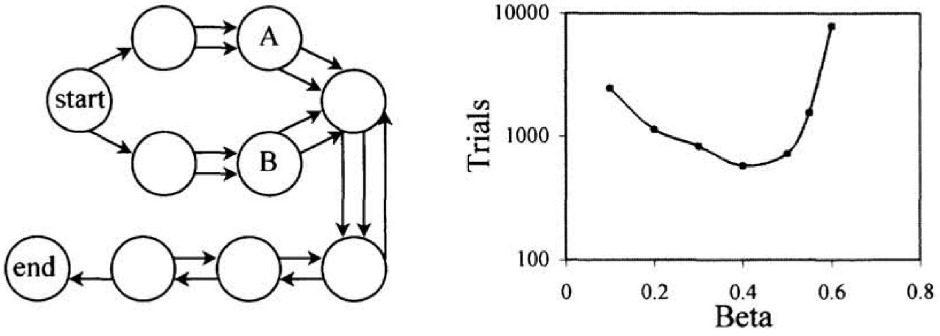


Figure 1. A POMDP and the number of trials needed to learn it vs. β.
A combination of policy-search and value-based RL outperforms either alone.

[1] Leemon C. Baird and Andrew W. Moore. **Gradient descent for general reinforcement learning**. In *Advances in Neural Information Processing Systems 11*. The MIT Press, 1999.

Closely Related Papers: Building on these ideas

[1+] Leonid Peshkin, Kee-Eung Kim, Nicolas Meuleau, and Leslie Pack Kaelbling. **Learning to cooperate via policy search**. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA, USA, 489–496. 2000

- From the same authors, extending this work to cooperative game playing

[2+] Leonid Peshkin. **Reinforcement Learning by Policy Search**. *PhD thesis*. Massachusetts Institute of Technology. 2003.

- PhD thesis of the first author, includes VAPS work and investigates other methods of controllers with memory

[3+] Bram Bakker. **Reinforcement learning with long short-term memory**. *Advances in Neural Information Processing Systems*. 2002.

- Solves non-Markovian tasks with long-term dependencies between relevant events (using RL-LSTMs)

[4+] Douglas Aberdeen and Jonathan Baxter. **Scaling internal-state policy-gradient methods for POMDPs**. *Machine Learning International Workshop Then Conference*. 2002.

- Extends this work of reinforcement with memory to the infinite-horizon case

[5+] Douglas Aberdeen. **Policy-Gradient Algorithms for Partially Observable Markov Decision Processes**. *PhD thesis*. Research School of Information Sciences and Engineering and The Australian National University. 2003.

- Uses high-order filters to replace eligibility traces of gradient estimators

[6+] Dustin J. Nowak and Gary B. Lamont. **Autonomous Self Organized UAV Swarm Systems**. In *IEEE National Aerospace and Electronics Conference*, pp. 183–189. 2008.

- Doesn’t cite this work, but uses the concept of stigmergy for autonomous UAV swarms

Contributions: Revisited

- Applied stigmergic approaches to learning policies in partially observable domains
    - Employed online in highly non-Markovian domains
- Derived a simplified version of the VAPS algorithm for stigmergic policies
    - Showed a problem where the eligibility traces of $\textsc{Sarsa}(\lambda)$ caused failures
- Calculated the same gradient as VAPS with less computational effort
    - Using a Q-value lookup table

Novelty:

- This work incorporated external memory into learning memoryless online policies
- Assigns credit to (s, a) pairs proportional to the deviation from expected behavior

Appendix: Notation Changes

The following notation changes were made to adhere to DMU [0]

Variable	**Old**	**New**
State	x	s
Observation	x	o
Action	u	a
Experience Sequence	$s \in S$	$z \in Z$
Probability	$Pr(...)$	$P(...)$
Trace	$T_{x,u,t}$	$\Phi_{s,a,t}$

[0] Mykel J. Kochenderfer. **Decision Making Under Uncertainty: Theory and Application**. The MIT Press, 2015.

Thank you!

Questions?

mossr@stanford.edu

Algorithm 2: VAPS

(gradient explanation)

- The gradient of the global error $B = \sum_{T=0}^{\infty} \sum_{\tilde{z} \in \tilde{Z}_T} P(\tilde{z})\varepsilon(\tilde{z})$ with respect to weight k

$$\frac{\partial}{\partial w_k} B = \sum_{t=0}^{\infty} \sum_{z \in Z_t} P(z) \left[\frac{\partial}{\partial w_k} e(z) + e(z) \sum_{j=1}^{t} \frac{\partial}{\partial w_k} \ln P(a_{j-1} \mid o_{j-1}) \right]$$

- Stochastic gradient descent can be performed for several trials
  - Producing a sample sequence z for a trial of length T
  - Per trial, weights are kept constant and the gradient in [brackets] is accumulated
  - Weights are updated at the end of each trial
- An incremental algorithm was used at every step t following these update rules

$$\underbrace{\Delta \Phi_{k,t}}_{\text{exploration trace}} = \frac{\partial}{\partial w_k} \ln P(a_{t-1} \mid o_{t-1})$$

$$\underbrace{\Delta w_k}_{\text{weights}} = -\alpha \left[\frac{\partial}{\partial w_k} e(z_t) + e(z_t) \Phi_{k,t} \right]$$