Double Q-Learning

Weiwei Zhang

Q-Learning

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

- Disadvantages
  - **Overestimations** of the action values resulting from using the maximum value as approximation for the maximum expected value.

Double Q-Learning, Hasselt [2011]

- Randomly pick Q1 or Q2
  - Q1 updating

$$Q_1(S_t, A_t) \leftarrow Q_1(S_t, A_t) + \alpha\Big(R_{t+1} + Q_2(S_{t+1}, \arg\max_a Q_1(S_{t+1}, a)) - Q_1(S_t, A_t)\Big)$$

Double Q-Learning, Hasselt [2011]

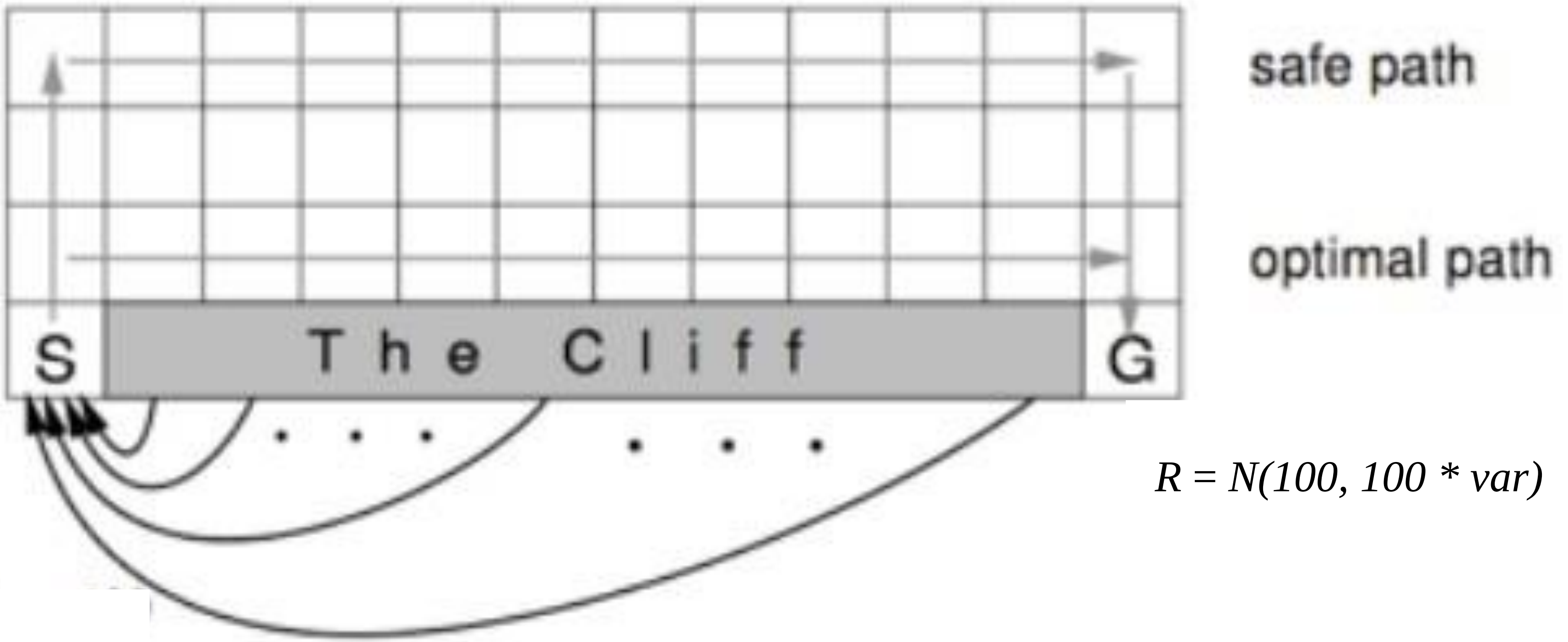
Lemma 1. *Let $X = \{X_1, \ldots, X_M\}$ be a set of random variables and let $\mu^A = \{\mu_1^A, \ldots, \mu_M^A\}$ and $\mu^B = \{\mu_1^B, \ldots, \mu_M^B\}$ be two sets of unbiased estimators such that $E\{\mu_i^A\} = E\{\mu_i^B\} = E\{X_i\}$, for all i. Let $\mathcal{M} \stackrel{\text{def}}{=} \{j \mid E\{X_j\} = \max_i E\{X_i\}\}$ be the set of elements that maximize the expected values. Let a^* be an element that maximizes μ^A: $\mu_{a^*}^A = \max_i \mu_i^A$. Then $\boxed{E\{\mu_{a^*}^B\} = E\{X_{a^*}\} \leq \max_i E\{X_i\}}$. Furthermore, the inequality is strict if and only if $P(a^* \notin \mathcal{M}) > 0$.*

Proof. Assume $a^* \in \mathcal{M}$. Then $E\{\mu_{a^*}^B\} = E\{X_{a^*}\} \stackrel{\text{def}}{=} \max_i E\{X_i\}$. Now assume $a^* \notin \mathcal{M}$ and choose $j \in \mathcal{M}$. Then $E\{\mu_{a^*}^B\} = E\{X_{a^*}\} < E\{X_j\} \stackrel{\text{def}}{=} \max_i E\{X_i\}$. These two possibilities are mutually exclusive, so the combined expectation can be expressed as

$$
\begin{aligned}
E\{\mu_{a^*}^B\} &= P(a^* \in \mathcal{M}) E\{\mu_{a^*}^B | a^* \in \mathcal{M}\} + P(a^* \notin \mathcal{M}) E\{\mu_{a^*}^B | a^* \notin \mathcal{M}\} \\
&= P(a^* \in \mathcal{M}) \max_i E\{X_i\} + P(a^* \notin \mathcal{M}) E\{\mu_{a^*}^B | a^* \notin \mathcal{M}\} \\
&\leq P(a^* \in \mathcal{M}) \max_i E\{X_i\} + P(a^* \notin \mathcal{M}) \max_i E\{X_i\} \qquad = \max_i E\{X_i\} \ ,
\end{aligned}
$$

Randomized CliffWalking

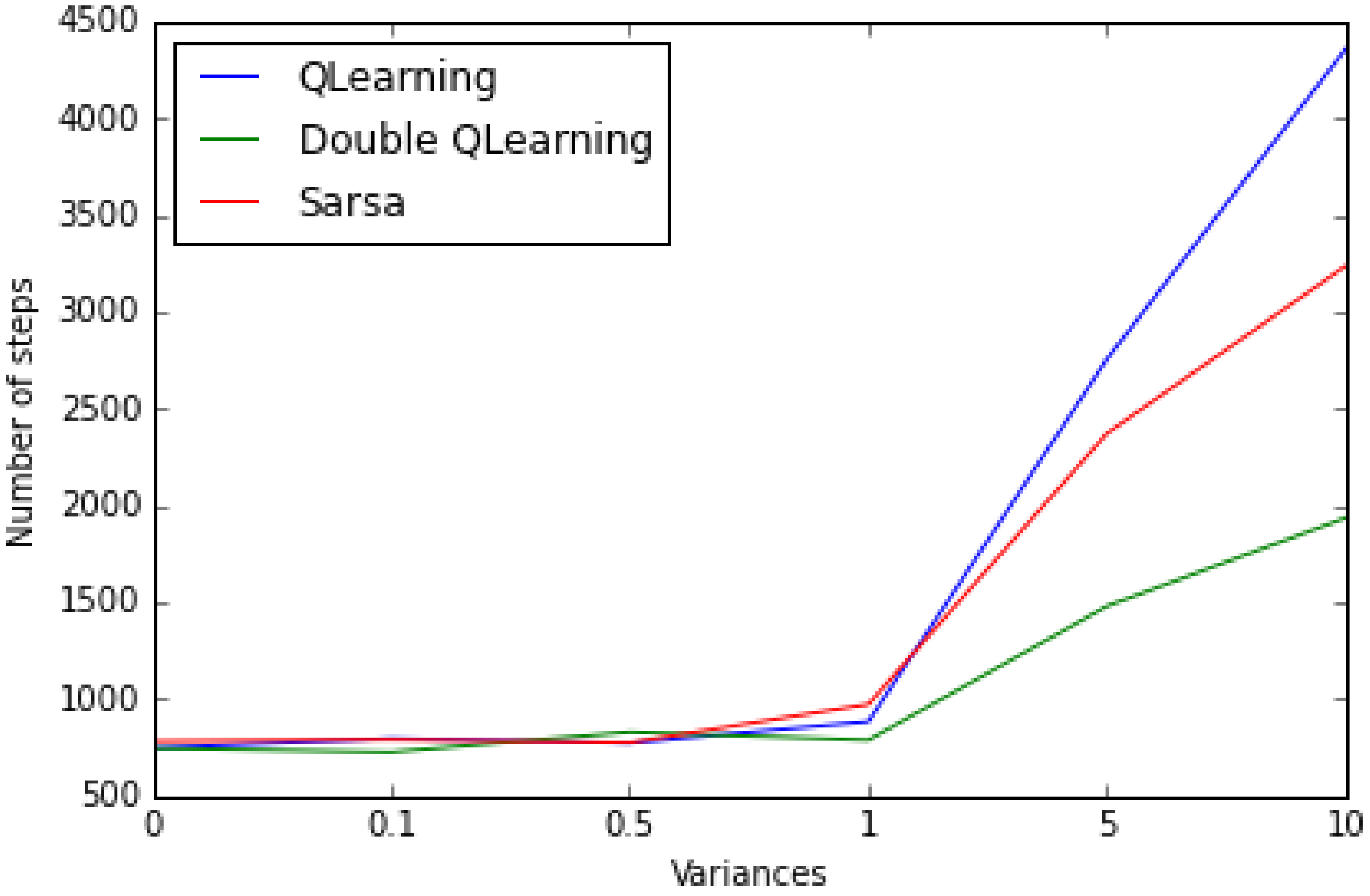
$R = N(-1, var)$



$R = N(100, 100 * var)$

$R = N(-100, 100 * var)$

Results



QLearning
Double QLearning
Sarsa
4500
4000
3500
3000
2500
2000
1500
1000
500
Number of steps
0
0.1
0.5
1
5
10
Variances