
# AIL 722: Reinforcement Learning

## Lecture 29: Double Q-Learning

Raunak Bhattacharyya

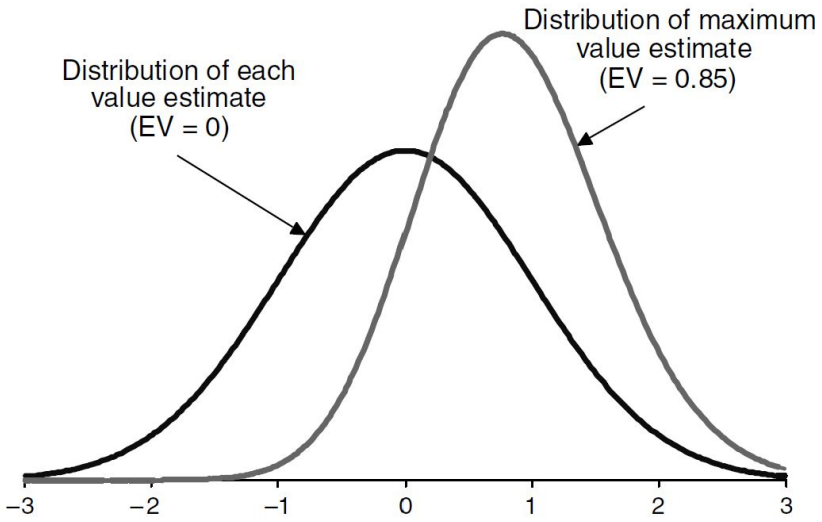
# Recap & Today's Outline

- Overestimation Bias
- Roulette and ALE
- Deep Q Network case study
- Single and double estimator
- Double Q Learning
- Double DQN

# The Optimizer’s Curse

- Decision science literature
- n alternatives
- Computing true values might cost many crores
- Computing estimates of the values might be lakhs of consulting effort
- Consulting firm tells us to pick the maximum expected estimated value

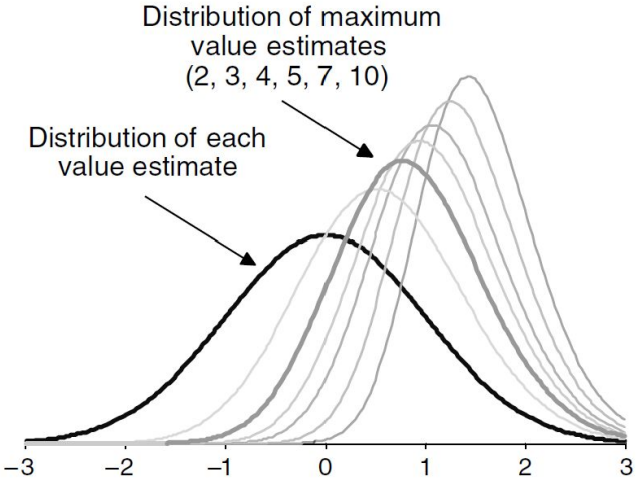
- 3 alternatives example






The Optimizer’s Curse, Smith & Winkler, Management Science 2006

# Increasing Number of Alternatives

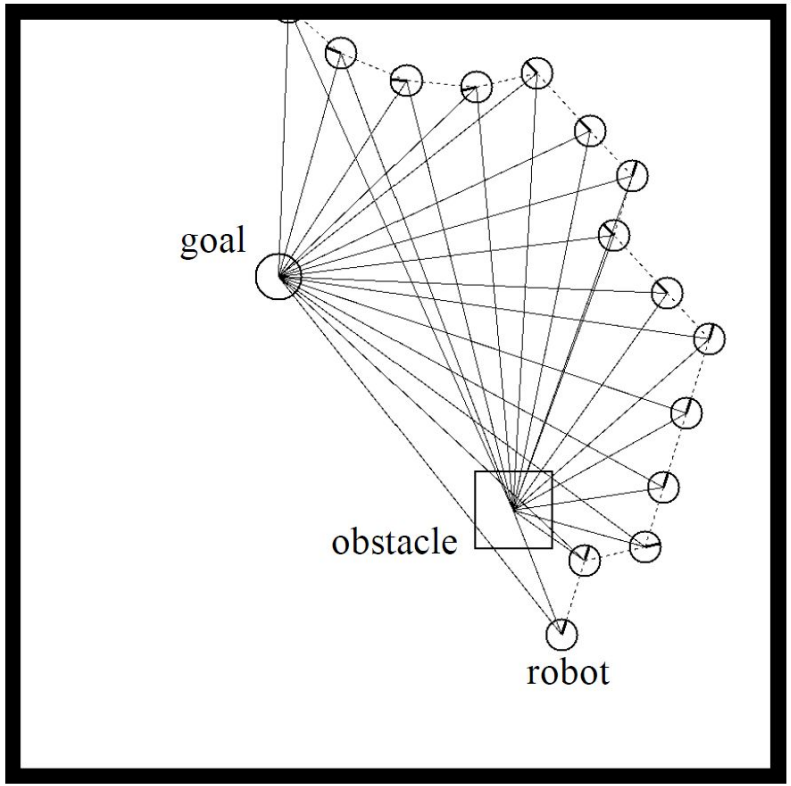


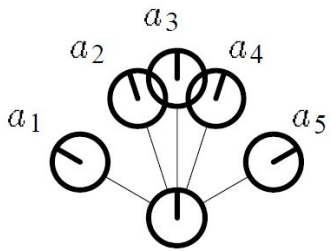

| Number of alternatives | Expected disappointment |
|---|---|
| 1 | 0.00 |
| 2 | 0.56 |
| 3 | 0.85 |
| 4 | 1.03 |
| 5 | 1.16 |
| 6 | 1.27 |
| 7 | 1.35 |
| 8 | 1.43 |
| 9 | 1.48 |
| 10 | 1.54 |

PROPOSITION 1. *Let* $V_1, \ldots, V_n$ *be estimates of* $\mu_1, \ldots, \mu_n$ *that are conditionally unbiased in that* $\mathrm{E}[V_i \mid \mu_1, \ldots, \mu_n] = \mu_i$ *for all* $i$. *Let* $i^*$ *denote the alternative with the maximal estimated value* $V_{i^*} = \max\{V_1, \ldots, V_n\}$. *Then,*

$$\mathrm{E}[\mu_{i^*} - V_{i^*}] \leq 0. \qquad (1)$$

# Example from Robotics



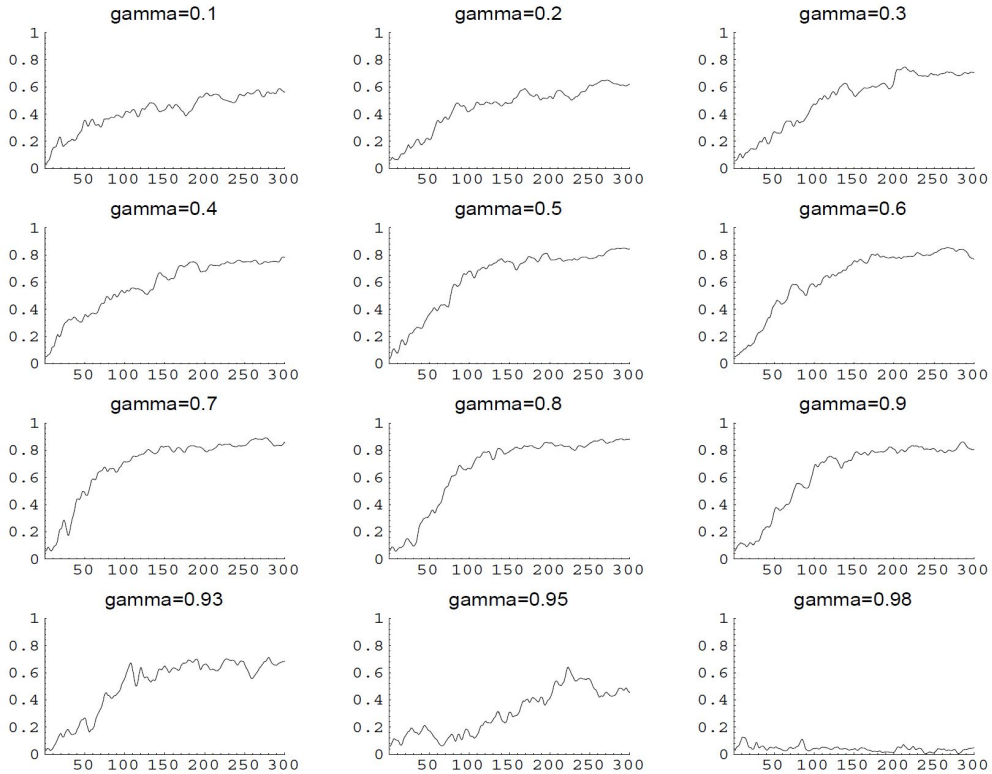




**Derives an upper bound on the overestimation**



Issues in using func. Approx. for RL, Thrun and Schwartz, 1993

# Overestimation




$$0 \leq E[Z_s] \leq \gamma c \; with \; c = \varepsilon \frac{n-1}{n+1}$$

# Towards Double Q-Learning

$M$ random variables $X = \{X_1, \ldots, X_M\}$

$$\max_i E\{X_i\}$$

**Notation alert**

$$S = \bigcup_{i=1}^{M} S_i$$

**Si contains samples of Xi**

$$\mu_i(S) \overset{\text{def}}{=} \frac{1}{|S_i|} \sum_{s \in S_i} s$$

$$\max_i E\{X_i\} = \max_i E\{\mu_i\} \approx \max_i \mu_i(S)$$

**Single estimator**



Double Q-Learning, van Hasselt, NeurIPS 2010

# Double Estimator

$$S^A \cap S^B = \emptyset$$

$$\mu_i^A(S) = \frac{1}{|S_i^A|} \sum_{s \in S_i^A} s \qquad \mu_i^B(S) = \frac{1}{|S_i^B|} \sum_{s \in S_i^B} s$$

$$\mu^A = \{\mu_1^A, \ldots, \mu_M^A\} \text{ and } \mu^B = \{\mu_1^B, \ldots, \mu_M^B\}$$

$\mu_i^A$ and $\mu_i^B$ are unbiased

**What does this mean?**

# Double Estimator

$$Max^A(S) \stackrel{\text{def}}{=} \{j \mid \mu_j^A(S) = \max_i \mu_i^A(S)\}$$

$$E\{\mu_j^B\} = E\{X_j\} \text{ for all } j, \text{ including all } j \in Max^A$$

$$\mu_{a^*}^A(S) \stackrel{\text{def}}{=} \max_i \mu_i^A(S)$$

$$\max_i E\{X_i\} = \max_i E\{\mu_i^B\} \approx \mu_{a^*}^B$$