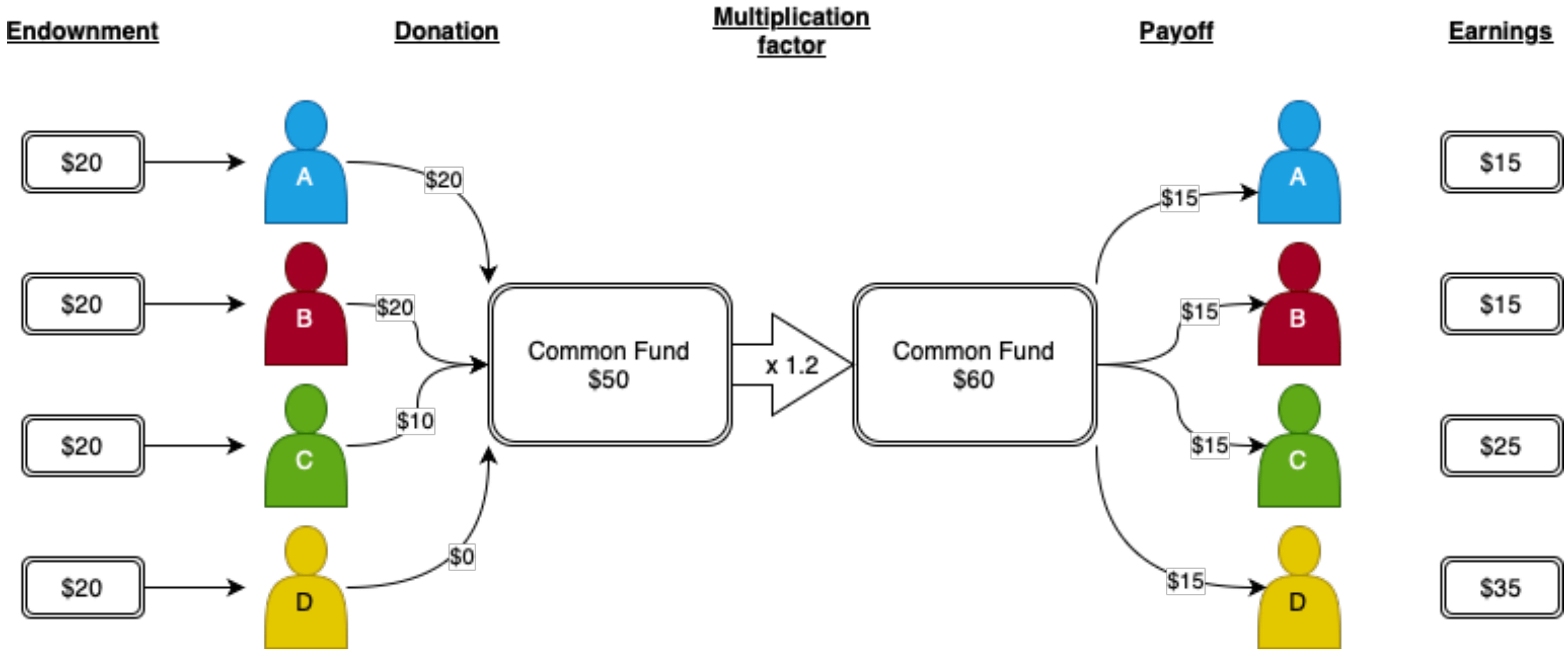
Empirical Evaluation of Overestimation Bias in Q-learning and Double Q-learning

Lai Jien Weng (ID: 2104438 - AM)
Supervised by: Dr. Tan Wei Lun

Introduction



“Freedom in a commons brings ruins to all.”

Hardin, G. (1968) ‘The Tragedy of the Commons’, 1, pp. 243–253. Available at: https://doi.org/10.1126/science.162.3859.1243.

Problem Statement

1. How does the multiplication factor affect cooperation in the Public Goods Game when modelled with Q-learning and Double Q-learning?
2. How do different endowments influence contribution strategies in the Public Goods Game using Q-learning and Double Q-learning?
3. How does a large action space impact fairness in the Public Goods Game when analysed with Q-learning and Double Q-learning?

Objectives

1. To study how the multiplication factor affects cooperation in the Public Goods Game using Q-learning and Double Q-learning.
2. To explore the impact of different endowments on contribution strategies in the Public Goods Game with Q-learning and Double Q-learning.
3. To assess how a large action space influences fairness in the Public Goods Game using Q-learning and Double Q-learning.

Significance

1. Create strategy to encourage cooperation.
2. Address inequities in resource allocation.
3. Promotes fairness in cooperative systems.

Literature Review

Summary of Key Prior Works

Study	Approach	Focus	Limitations
ManChon U and Zhen Li (2010)	TD-learning	Homogeneous endowments, cooperation rates	Binary actions, no fairness metrics
Fehr and Gächter (2002)	Experimental PGG	Variable contributions, punishment	Non-RL, human subjects only
Fischbacher, Gächter and Fehr (2001)	Experimental PGG	Conditional cooperation	Non-RL, no MARL framework
Isaac, Walker and Thomas (1984)	Experimental PGG	Heterogeneous endowments	Non-RL, limited to small groups
Rashid *et al.* (2018)	Deep RL (QMIX)	Scalable cooperation in dilemmas	Limited interpretability, no fairness focus
Jaques *et al.* (2019)	Deep RL with communication	Coordination via social influence	Homogeneous agents, no endowment variation

Summary of Key Prior Works

Methodology

PGG: Nash vs. Pareto Outcomes

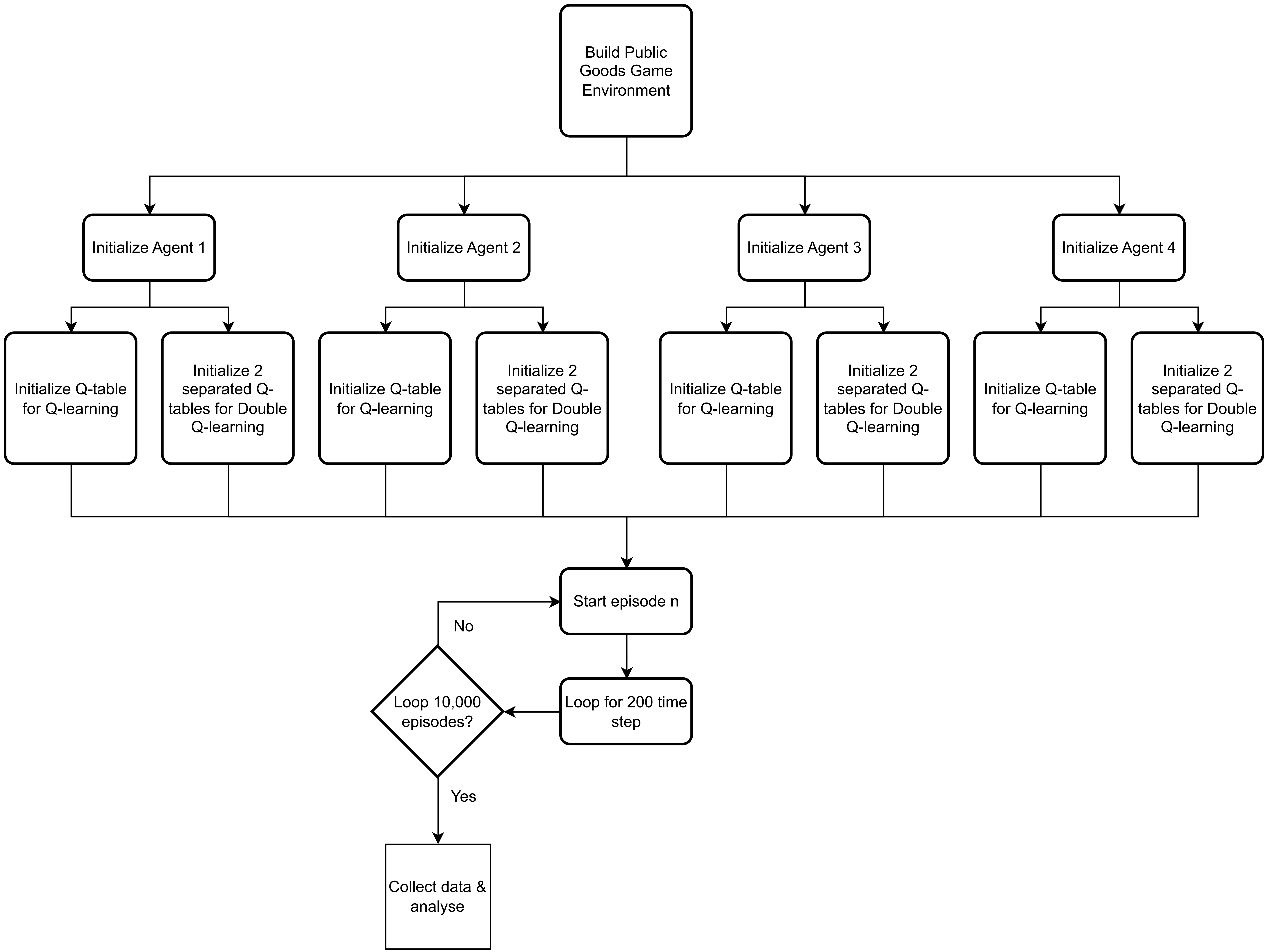
		Agent 2	
		Free-ride	**Contribute**
Agent 1	**Free-ride**	(1, 1)	(1.75, 0.75)
	Contribute	(0.75, 1.75)	(1.5, 1.5)

Note: Example shown with $r = 1.5$**,** $n = 2$**,** $\mathscr{A}_i = \{0, e_i\}$**,** $e_i = \{1,1\}$

PGG: Nash vs. Pareto Outcomes

		Agent 2	
		Free-ride	**Contribute**
Agent 1	**Free-ride**	Nash Equilibrium (1, 1)	(1.75, 0.75)
	Contribute	(0.75, 1.75)	Pareto Optimality (1.5, 1.5)

Note: Example shown with $r = 1.5, n = 2, \mathcal{A}_i = \{0, e_i\}, e_i = \{1, 1\}$


Build Public Goods Game Environment
Initialize Agent 1
Initialize Agent 2
Initialize Agent 3
Initialize Agent 4
Initialize Q-table for Q-learning
Initialize 2 separated Q-tables for Double Q-learning
Initialize Q-table for Q-learning
Initialize 2 separated Q-tables for Double Q-learning
Initialize Q-table for Q-learning
Initialize 2 separated Q-tables for Double Q-learning
Initialize Q-table for Q-learning
Initialize 2 separated Q-tables for Double Q-learning
Start episode n
No
Loop 10,000 episodes?
Loop for 200 time step
Yes
Collect data & analyse

Q-learning (Watkins and Dayan, 1992)

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(s, a)\left[u_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$

where,

- $s_t \in S$ is the state at time t,
- $a_t \in A$ is the action taken at time t,
- $s_{t+1} \in S$ is the next state at time t,
- u_t is the reward at time t,
- $\gamma \in [0,1]$ is the discount factor,
- $\alpha(s, a) \in (0,1]$ is the learning rate, and
- $\max_a Q(s_{t+1}, a)$ is the highest expected future rewards.

Double Q-learning (Hasselt, 2010)

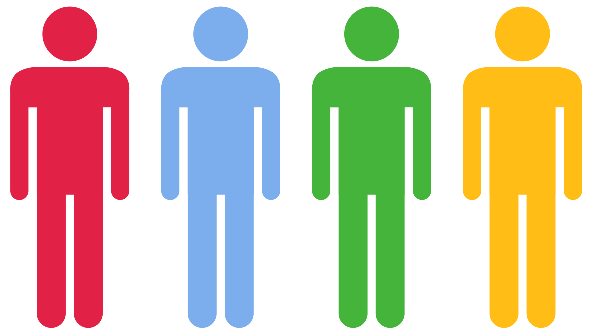
$$Q^A(s_t, a_t) \leftarrow Q^A(s_t, a_t) + \alpha \left[u_t + \gamma Q^B[s_{t+1}, \arg\max_a Q^A(s_{t+1}, a)] \right]$$

$$Q^B(s_t, a_t) \leftarrow Q^B(s_t, a_t) + \alpha \left[u_t + \gamma Q^A[s_{t+1}, \arg\max_a Q^B(s_{t+1}, a)] \right]$$

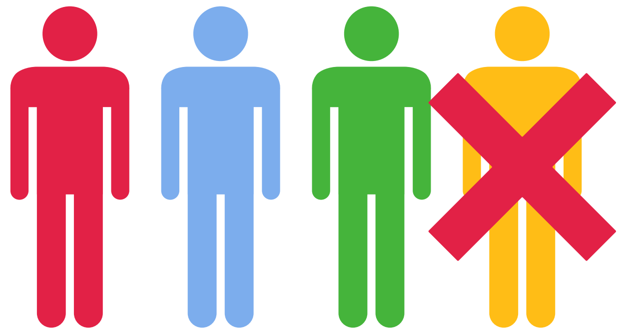
where all variables follow the definition in previous slide.

Uncertainty in PGG

- Random exclusion: 75% all agent participates, 25% one random agent excluded.

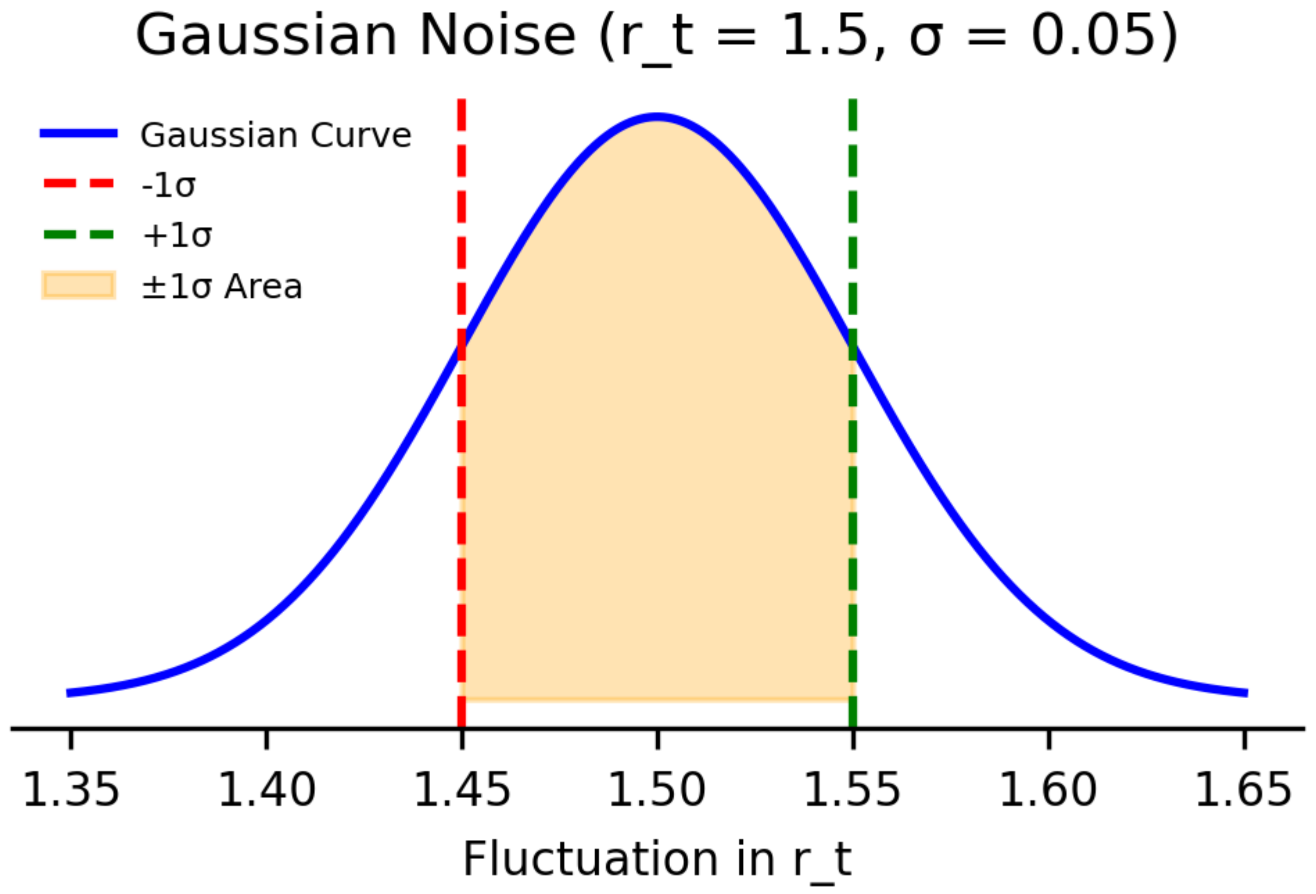


75%



25%

- Gaussian noise: Multiplication factor fluctuate within $\sigma_r = 0.05$.

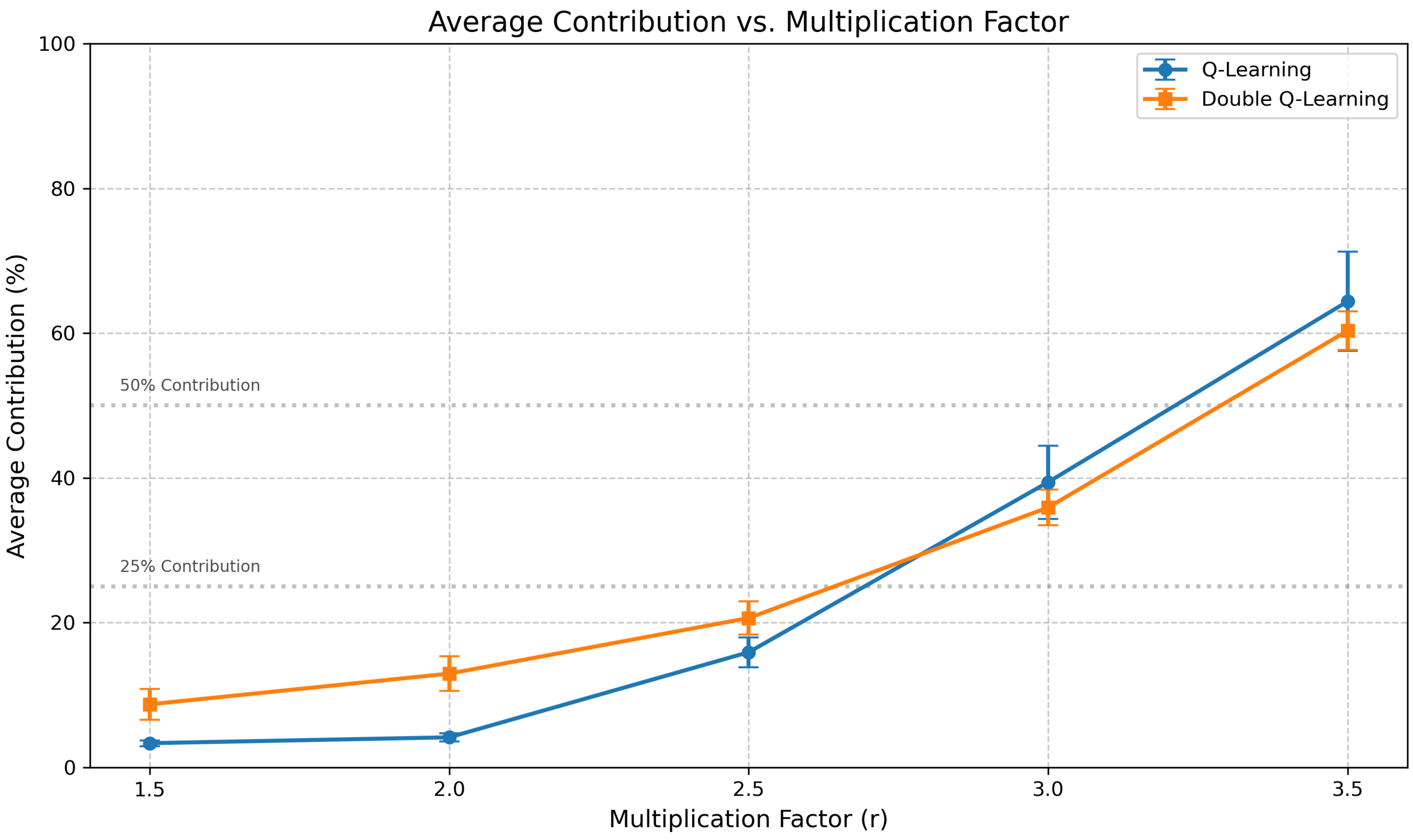


Experiment 1: Multiplication Factor

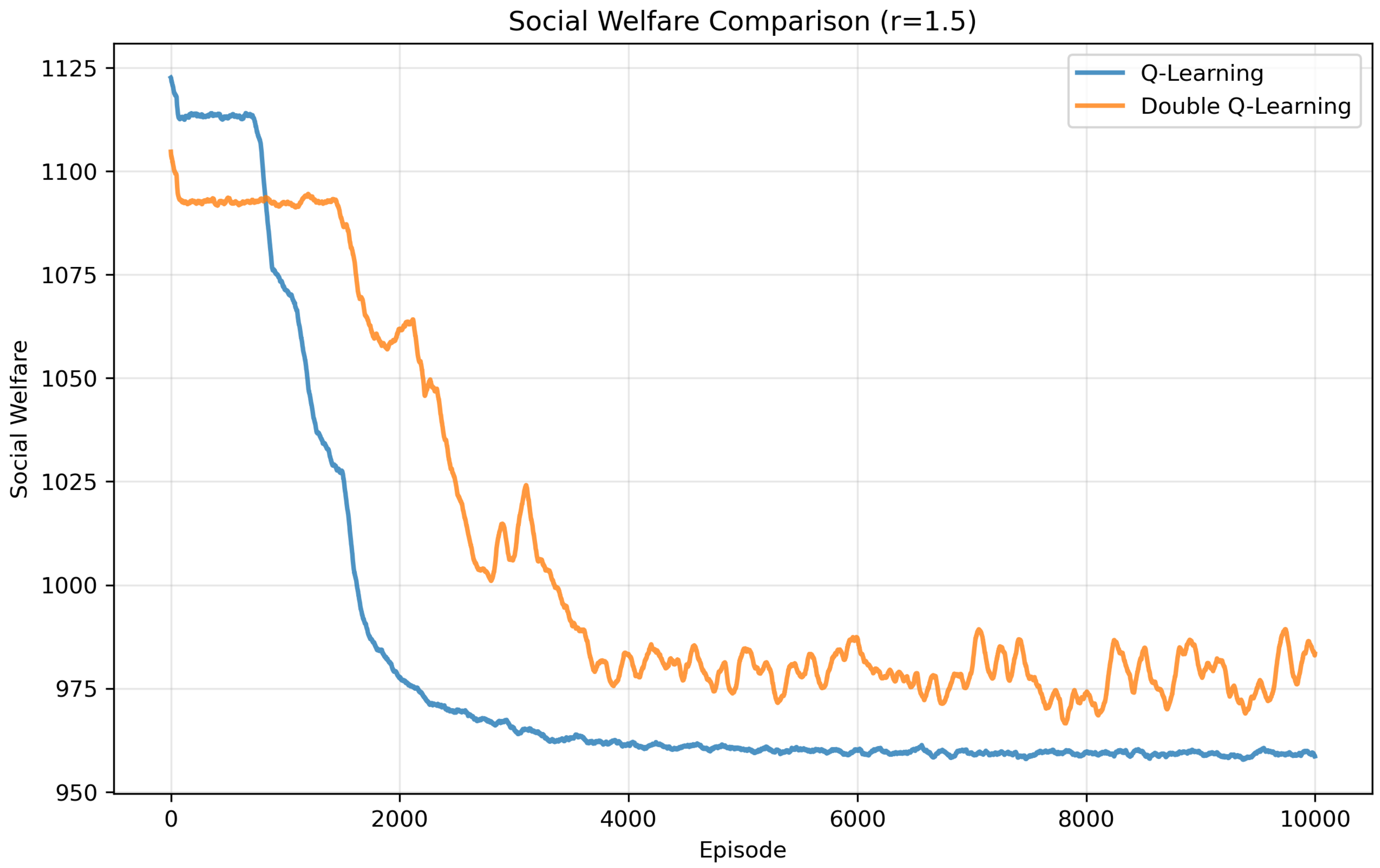
- Run experiment for $r_t = \{1.5, 2.0, 2.5, 3.0, 3.5\}$.

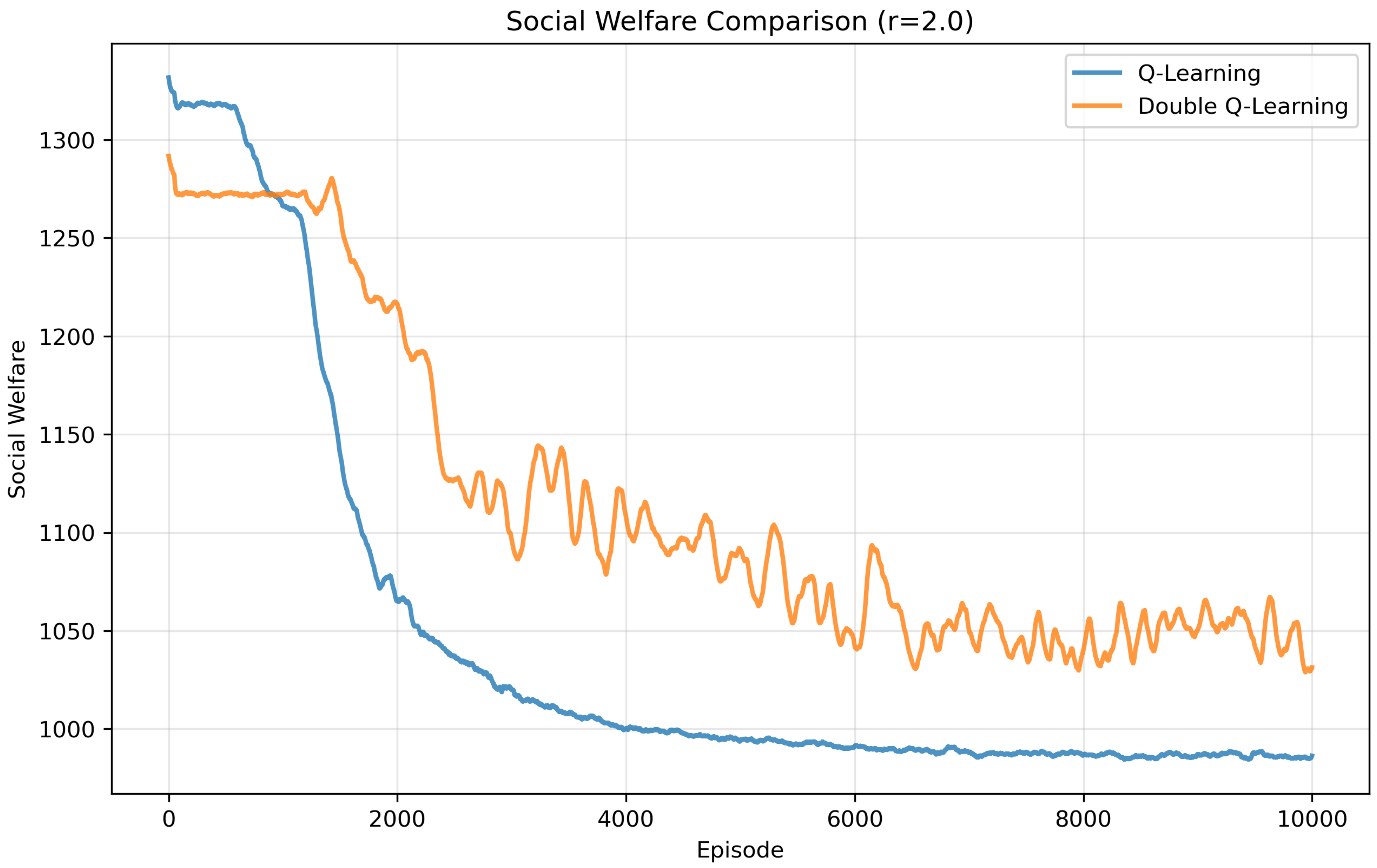
- Measure with:
  - Contribution rate, $\bar{a}_i$ against r_t,
  - Social welfare, W_t,
  - $\sigma_{contrib}$.

Experiment 1: Multplication Factor (cont.)



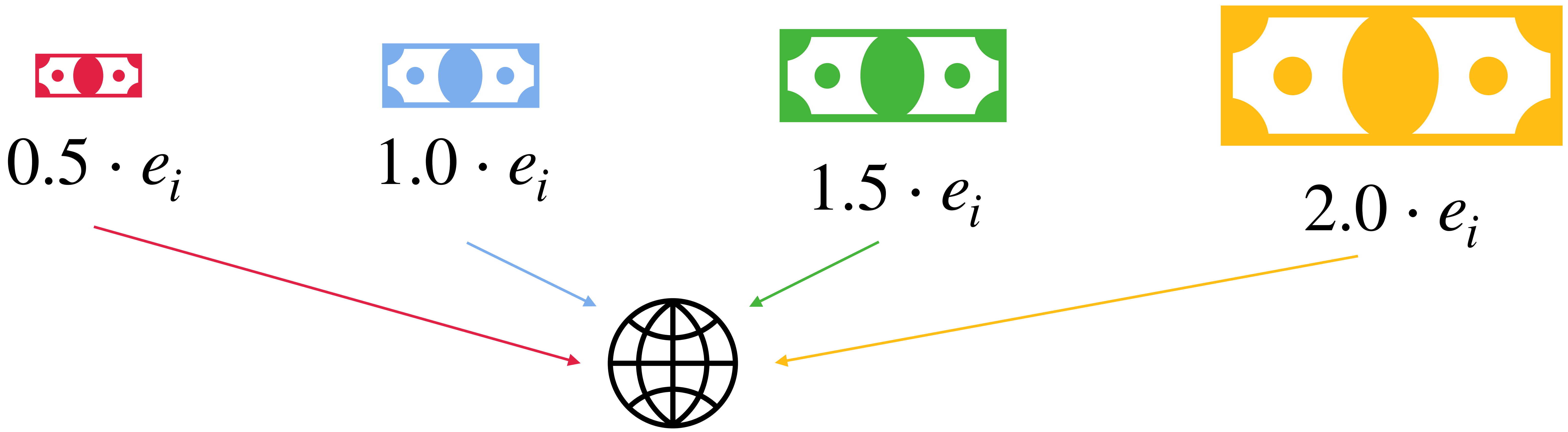
Experiment 1: Multplication Factor (cont.)





Experiment 2: Heterogeneous Endowments

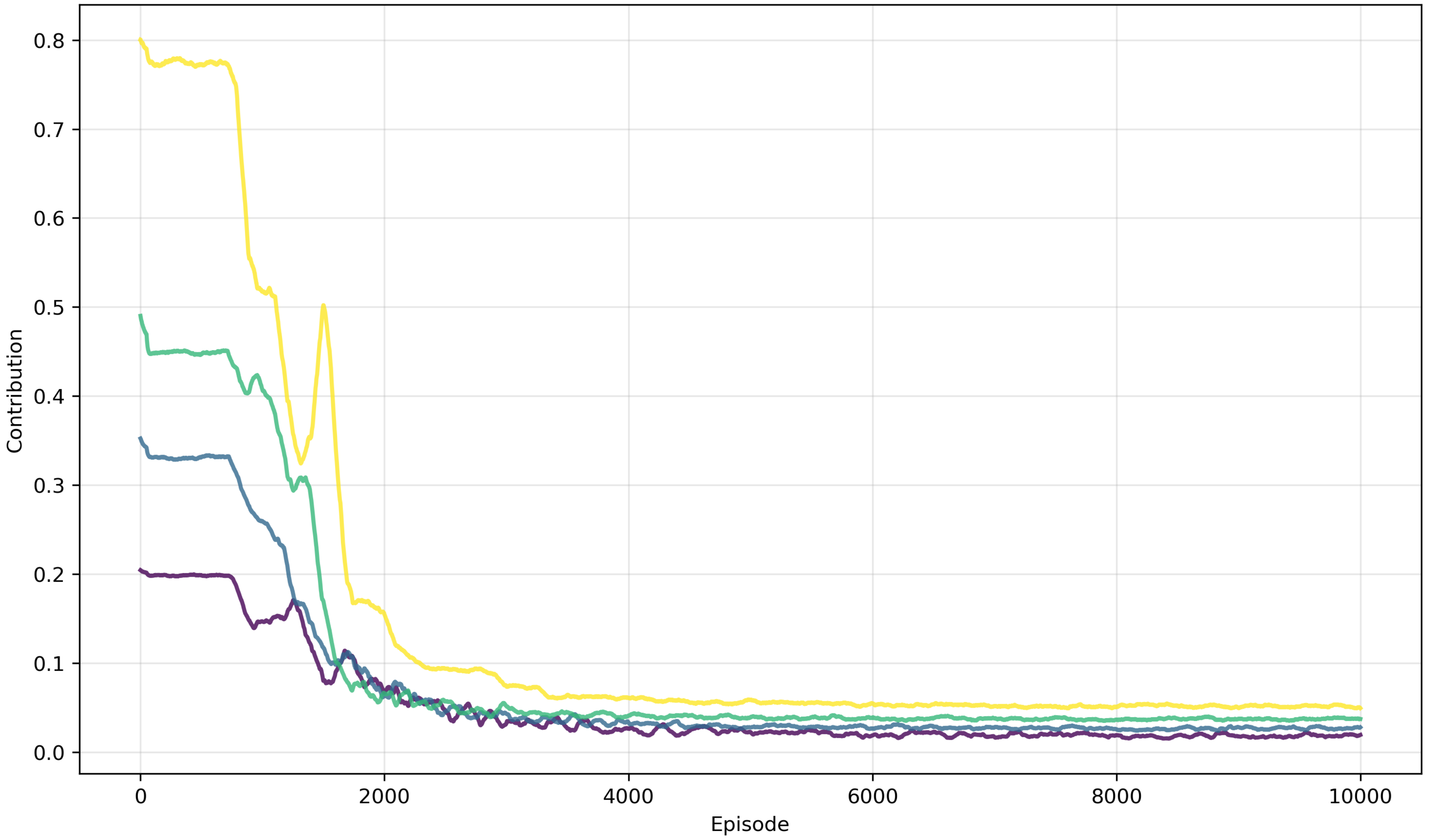
- Run experiment for $e_i = \{0.5, 1.0, 1.5, 2.0\}$.



- Measure with:
  - Individual contribution per epsiode, $\tilde{a}_{i,k}$.

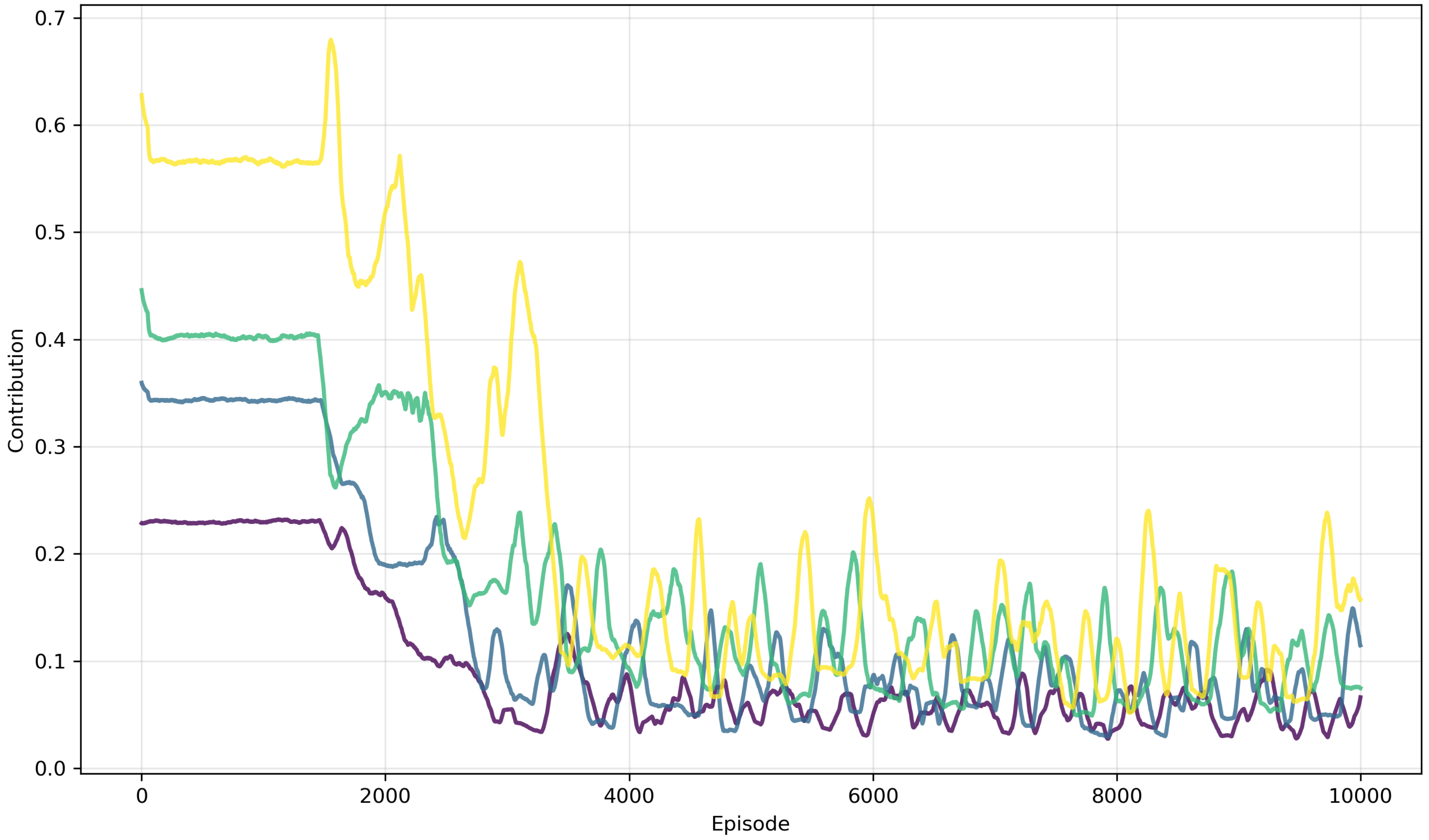
Experiment 2: Heterogeneous Endowment (cont.)

Individual Contribution of Q-learning



With $r_t = 1.5$

Individual Contribution of Double Q-learning



With $r_t = 1.5$

Legends

2.0	1.5	1.0	0.5

Experiment 3: Fairness

- Run experiment for 25-level discrete contribution options, $\mathscr{A}_i = \{0, 0.04e_i, \ldots, e_i\}$.
- Test with a control set with 4-level discrete contribution options.

- Measure with:
  - Shapley value, ϕ_i

Statistical Testing

All tests will be tested with paired t-test.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$$

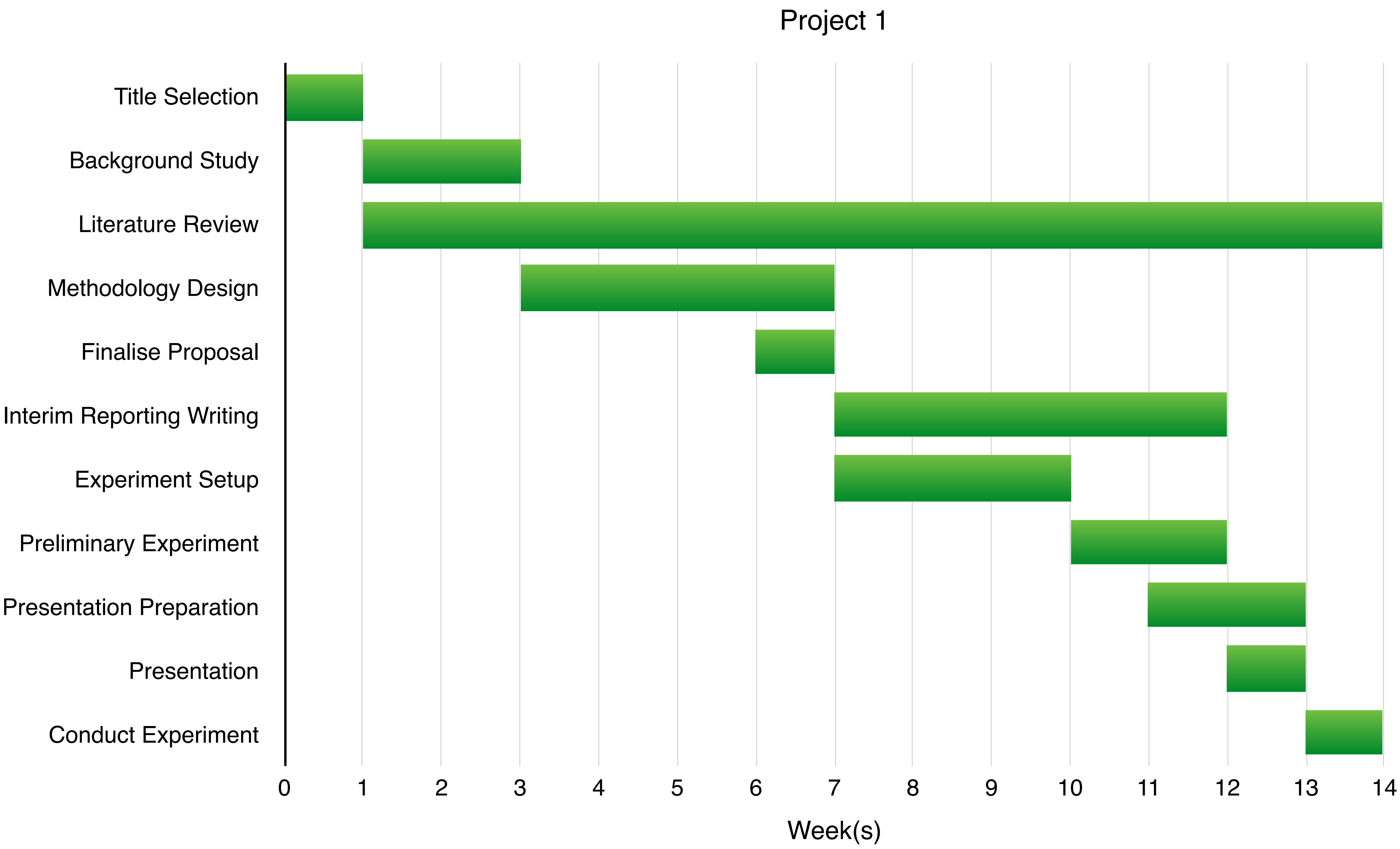
H_0 :There is no significant difference between the metric values of Q & DQ

$$H_0 : \mu_Q = \mu_{DQ}$$

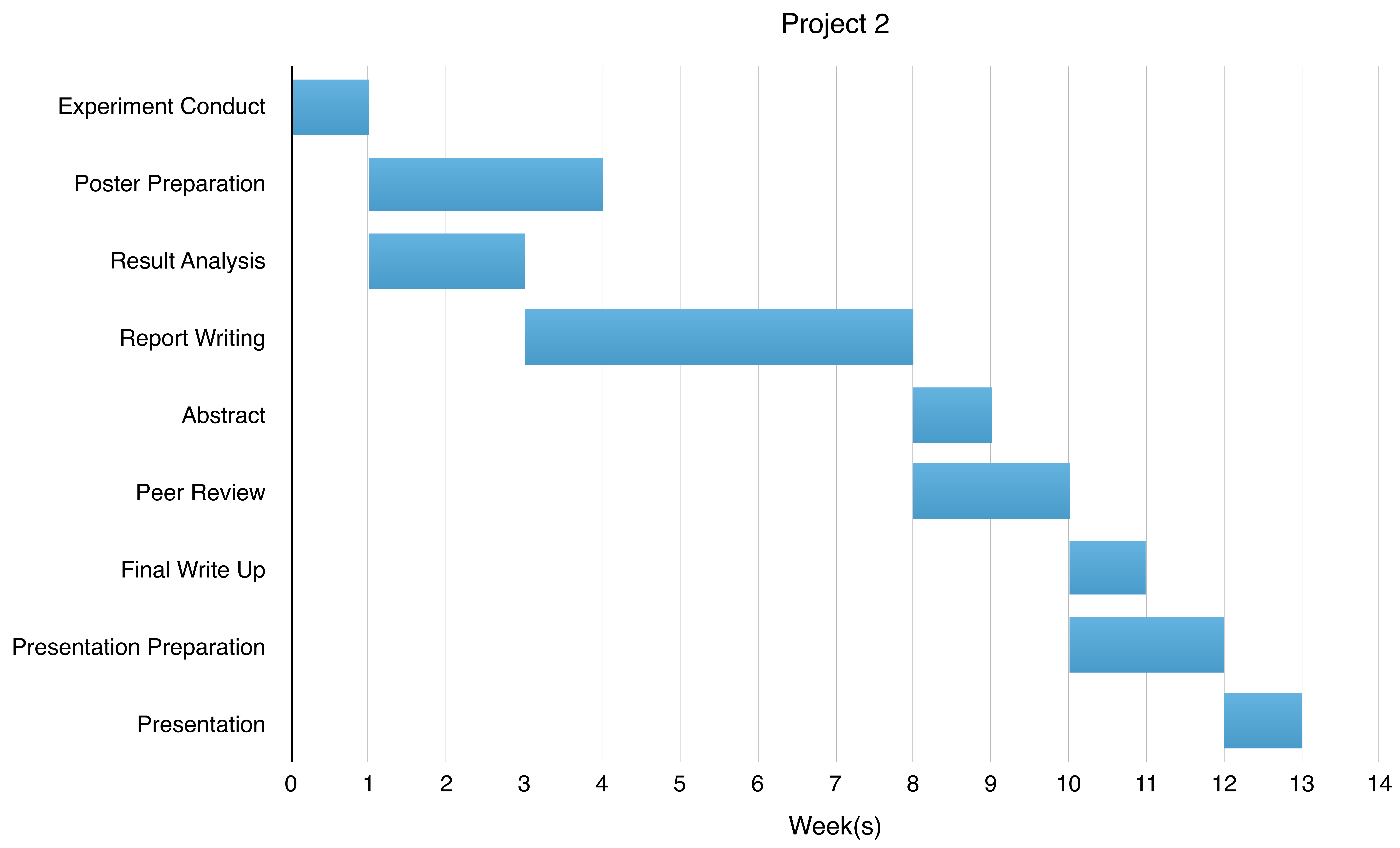
H_1 :There is a significant difference between the metric values of Q & DQ

$$H_1 : \mu_Q \neq \mu_{DQ}$$

Implementation Plan



Project 1
Title Selection
Background Study
Literature Review
Methodology Design
Finalise Proposal
Interim Reporting Writing
Experiment Setup
Preliminary Experiment
Presentation Preparation
Presentation
Conduct Experiment
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Week(s)


Project 2
Experiment Conduct
Poster Preparation
Result Analysis
Report Writing
Abstract
Peer Review
Final Write Up
Presentation Preparation
Presentation
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Week(s)

Summary

- This study investigated cooperative behaviour in PGG within a MARL framework of:
    1. Multiplication factor
    2. Heterogeneous endowments
    3. Fairness in reward distribution

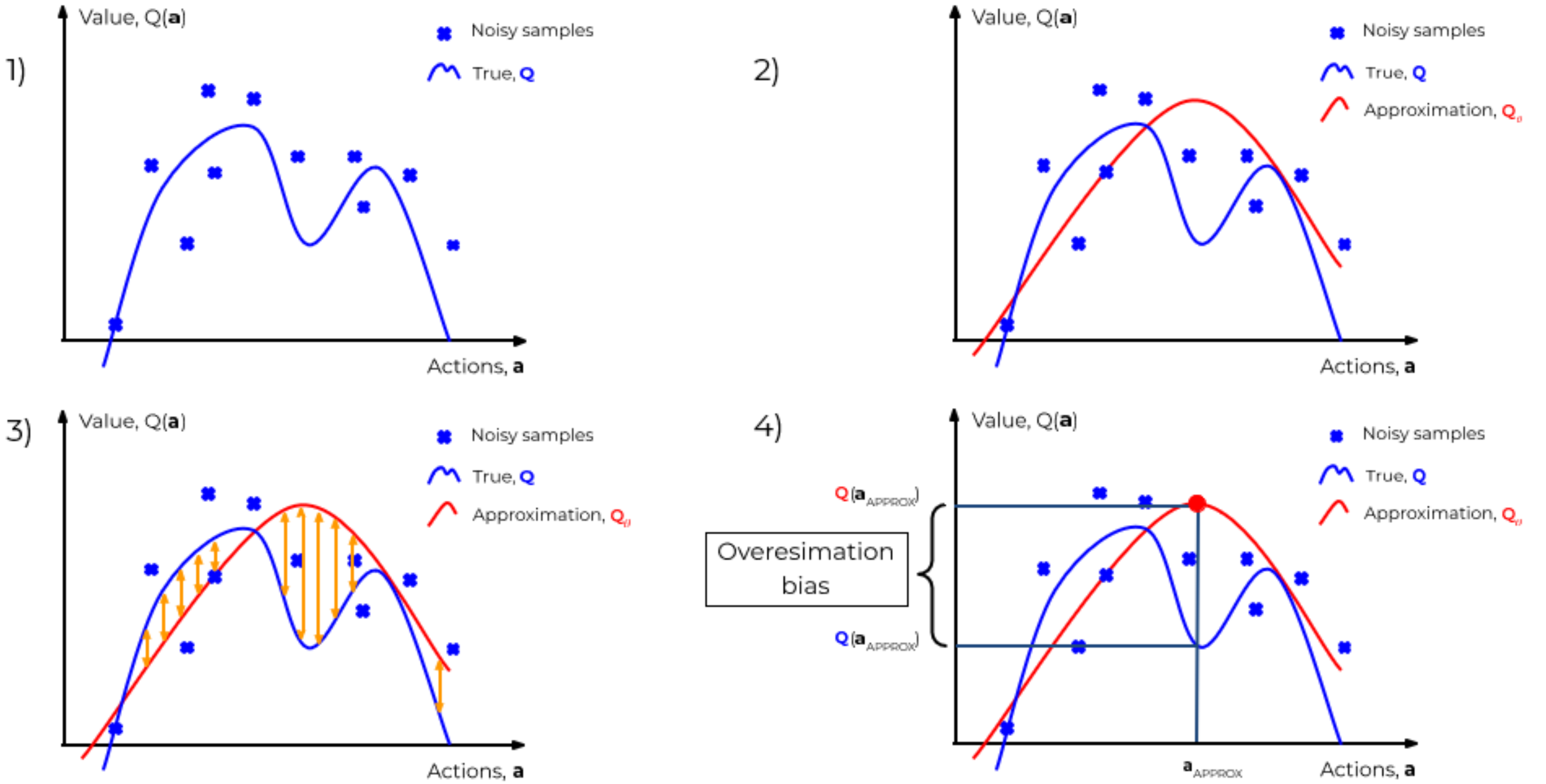
- The experiment employed:
    A. Tabular Q-learning
    B. Double Q-learning

Thank you

Appendices

Q-learning Overestimation Bias

$$\mathbb{E}\left[\max_{a'} Q(s', a')\right] \geq \max_{a'} Q^*(s', a')$$



Double Q-learning

Double Q-learning uses two Q-tables, Q^A and Q^B.

Instead of using the same values to both select and evaluate the best action (which causes overestimation), it splits the process:

$$Q^A(s,a) \leftarrow Q^A(s,a) + \alpha \left[r + \gamma Q^B(s', \arg\max_{a'} Q^A(s',a')) - Q^A(s,a) \right]$$

- Q^A chooses the best action.
- Q^B estimates its value (or vice versa).

This reduces the chance of both selecting and overestimating the same action, thus reducing bias.