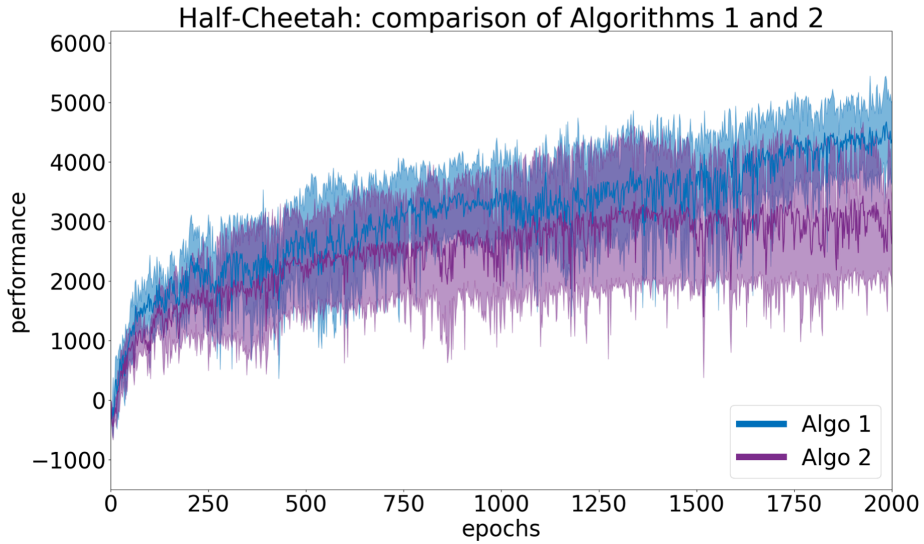
Institute of
Artificial Intelligence

Leibniz
Universität
Hannover

Statistical Tests in RL

Artur Ganzha 10019651

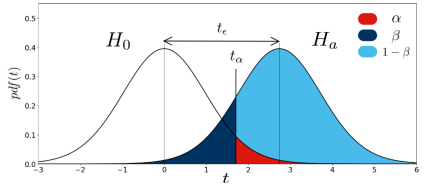
Reproducibility Crisis [CSO18]



- In RL & ML reproducibility is a challenge
  - Codebase not available
  - Paper does not explain enough
  - Versions are not mentioned
- Running several experiments show different results
- After interpreting the figure you might assume that the blue line outperforms the purple line
  - Actually the performances are the same

Welch's t-test & Bootstrapped confidence interval [CSO18]

- Welch's t-test was applied as the implications and constraint appeared suitable for RL
  - continuous and ordinal data
  - independent measurements
  - sampling from population
  - normal distributed
  - Overlap in Hypotheses are considered as errors
- Bootstrap is already widely used in ML/RL/DS
- Stick to Welch's t-test



Welch's t-test	bootstrapped confidence interval
normal data	distribution-free
compares two means	estimates parameter confidence
Is robust on smaller samples	large sample size increases accuracy
easy implementation	Rather complex due to resampling

Sources

[CSO18] https://arxiv.org/pdf/1806.08295