Bar-Ilan University
אוניברסיטת בר־אילן

Turning Sand to Gold: Recycling Data to Bridge On-Policy and Off-Policy Learning via Causal Bound

Tal Fiskus, Uri Shaham

NeurIPS 2025

The 39th Conference on Neural Information Processing Systems

DRL Limitations

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- **Deep reinforcement learning (DRL) agents** excel in solving complex decision-making tasks across various domains.
- **Main challenges:**
  - Substantial number of training steps.
  - Very Large experience replay buffer.
  - High computational demands.
  - High resource demands.

Contribution

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

SUFT, a causal upper-bound loss optimization method for DRL.

- **Significantly improve sample efficiency at a negligible cost.**
- **Reduce computational costs.**
- **Reduce resource demands.**
- **Bridge on-policy and off-policy learning methods.**
- **Strong theoretical result in the causal framework.**
- **Seamlessly applicable to any DRL agent with a V or Q-value network.**
- **Transform overlooked data into valuable causal insights.**

Theoretical Causal Bound

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- To the best of our knowledge, all other works bound the counterfactual loss using the factual loss.

Theoretical Causal Bound

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- To the best of our knowledge, all other works bound the counterfactual loss using the factual loss.
- **Our work bounds the factual loss using the counterfactual loss.**

Theoretical Causal Bound

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- To the best of our knowledge, all other works bound the counterfactual loss using the factual loss.
- **Our work bounds the factual loss using the counterfactual loss.**
- **Why?**

Theoretical Causal Bound

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- To the best of our knowledge, all other works bound the counterfactual loss using the factual loss.
- **Our work bounds the factual loss using the counterfactual loss.**
- **Why?**
- **How?**

Theoretical Causal Bound

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- To the best of our knowledge, all other works bound the counterfactual loss using the factual loss.
- **Our work bounds the factual loss using the counterfactual loss.**
- **Why?**
- **How?**
- **In DRL,** off-policy agents observe and store **counterfactual** data while **factual** data remains unobserved.

SUFT OPE Term

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- The **estimated treatment effect** loss in the causal framework is interpreted as our **SUFT off-policy evaluation (OPE) term** in the DRL framework.
- This term aligns with the principles of OPE, focusing on quantifying the discrepancy between the target policy and the behavior policy using off-policy data.

$$\psi_{\mathrm{SUFT}_Q} := \mathbb{E}_{(s,a)\sim D_0}[\mathbf{L}\left(Q(s,a;\theta_{\mathrm{behavior}}), Q(s,a;\theta_{\mathrm{target}})\right)].$$

SUFT Causal Bound

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- We conduct a **theoretical result** in the causal framework.
- Seamlessly **adapting it** to the DRL framework.

Causal Framework

- $\epsilon_{F_\phi} \leq \epsilon_{CF_\phi} + \psi_\phi + \delta.$
- Upper bound the **factual** loss.
- Using the **counterfactual** loss.
- The **estimated treatment effect** loss.
- And a constant term independent of ϕ.

DRL Framework

- $\epsilon_{\text{On-Policy}_Q} \leq \epsilon_{\text{Off-Policy}_Q} + \psi_{\text{SUFT}_Q} + \delta.$
- Upper bound the **on-policy** loss.
- Using the **off-policy** loss.
- The **SUFT OPE term.**
- And a constant term independent of Q.

Turning Sand to Gold

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- Recycling Value Network Outputs
- The Q values are **already being calculated** in the action selection when generating the experience.
- Our method, metaphorically: **Turning Sand to Gold.**
- By storing the old Q-value in the replay buffer:

$$(s, a, r, s', Q(s, a; \theta_{\text{behavior}}))$$

- The SUFT OPE term:

$$\psi_{\text{SUFT}_Q} := \mathbb{E}_{(s,a)\sim D_0}[\mathbf{L}\left(Q(s, a; \theta_{\text{behavior}}), Q(s, a; \theta_{\text{target}})\right)].$$

SUFT Universal Implementation

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- **Optimizing the SUFT causal bound** instead of the standard DRL loss.
- The agent's loss objective function: $\epsilon_{\text{Off-Policy}_Q} + \lambda_{\text{TF}} \cdot \psi_{\text{SUFT}_Q}$.
- Adding a coefficient to the SUFT OPE term improves the flexibility and accuracy of the loss optimization.
- **Our method is applicable to any DRL agent** with a V or Q-value network.

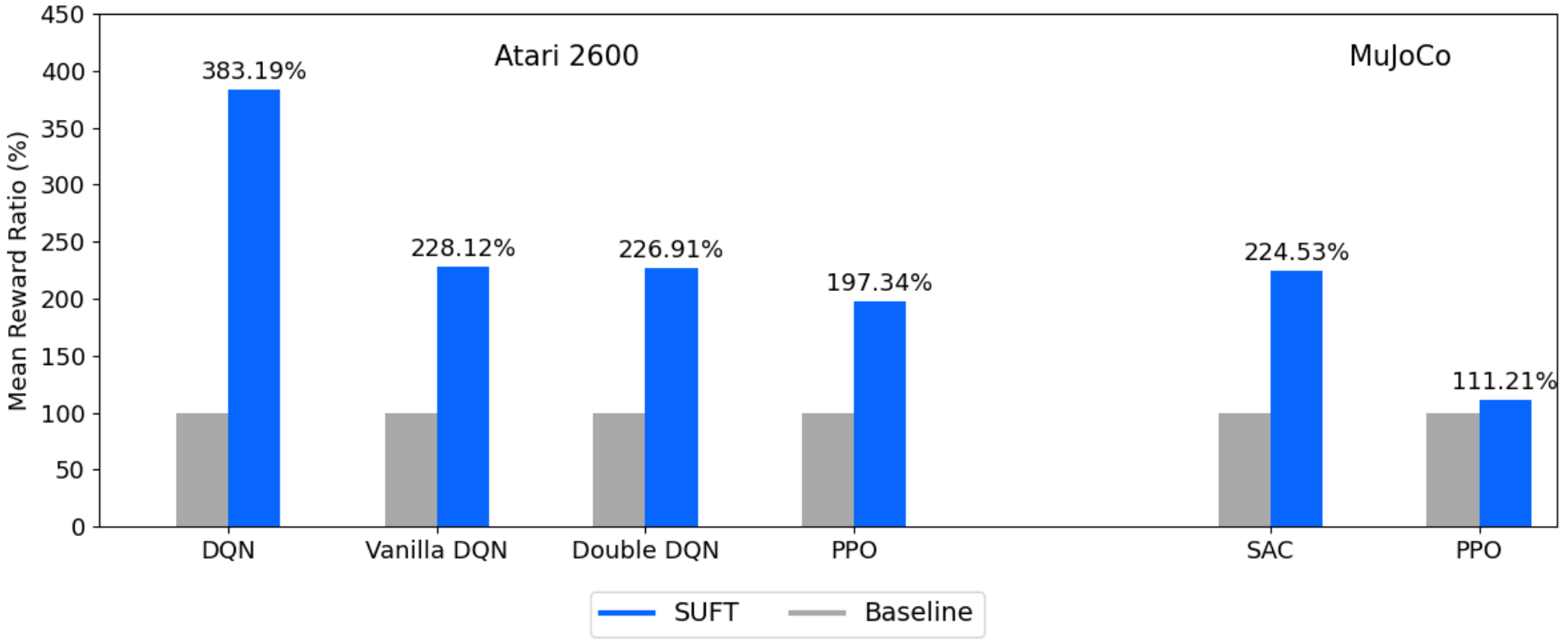
- The SUFT causal bound:

$$\epsilon_{\text{On-Policy}_Q} \leq \epsilon_{\text{Off-Policy}_Q} + \psi_{\text{SUFT}_Q} + \delta.$$

Experiments

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- **Mean reward ratio** comparison between agents using SUFT and baselines.
- Highlighting the **profound reward gains** across diverse agents and domains.



Experiments

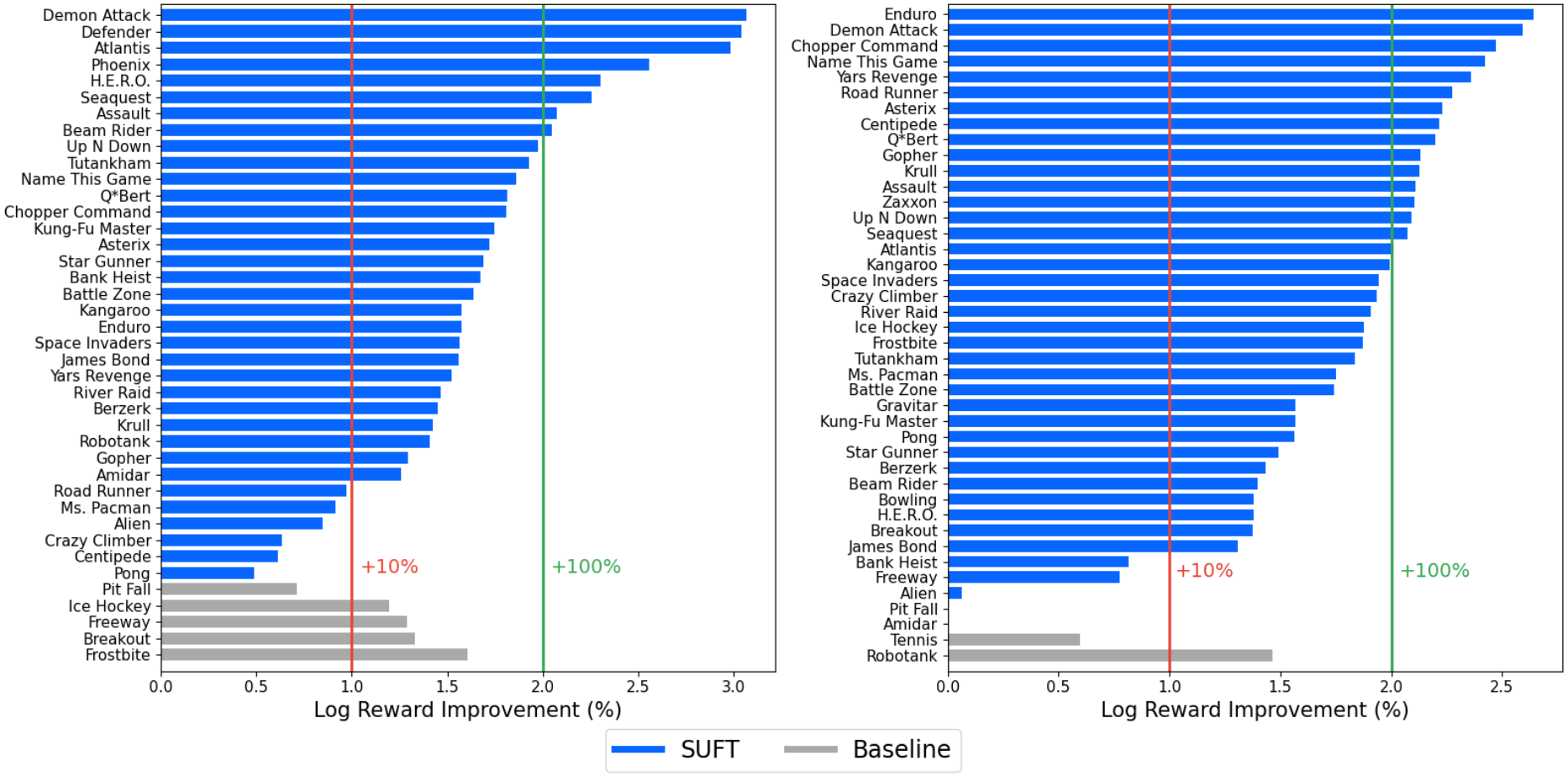
Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

Double DQN (Left):

- Over 100% improvement in **20%**
- Over 10% improvement in **72.5%**

PPO (Right):

- Over 100% improvement in **38.1%**
- Over 10% improvement in **83.3%**



Experiments

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- **SUFT achieves the highest Human Normalized Mean reward gain.**
- Better than all other tested methods that **require higher resources and higher computational costs.**
- **SUFT** surpasses the baseline that has a **x25** larger buffer, which is a **96%** buffer size reduction.

Table 1: An ablation study on the DQN agent comparing the SUFT additional term impact against enlarging the buffer size, adding a Vanilla DQN, and adding a Double DQN. The DQN, Vanilla DQN, Double DQN, and SUFT DQN use a 4K buffer size. All agents are using L2 loss.

Agent	Human Normalized Mean (%)
DQN	25.04%
DQN, buffer 100K	29.29%
Vanilla DQN	41.26%
Double DQN	42.31%
SUFT DQN	**63.45%**

Thank You For Watching!

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

Tal Fiskus

LinkedIn

Uri Shaham

LinkedIn

References

Department of Computer Science
Faculty of Exact Sciences
Bar-Ilan University

- Mnih et al., Human-level control through deep reinforcement learning. Nature, 2015.
- Haarnoja et al., Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of The 35th International Conference on Machine Learning. PMLR, 2018.
- Schulman et al., Proximal policy optimization algorithms. 2017.
- Van Hasselt et al., Deep reinforcement learning with double q-learning. In Proceedings of the AAAI conference on artificial intelligence, 2016.
- Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 2005.
- Shalit et al., Estimating individual treatment effect: generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning. PMLR, 2017.