CAST: Certifiable Aggregation of Smoothed Teachers for Robust Policy Adaptation

Zhongrui Zhao[1,2,*] Yanan Cai[1,2,*] Zhigang Lu[2] Longkun Guo[3,†]
Ickjai Lee[1] Shuchao Pang[4] Minhui Xue[5]

[1]James Cook University, Australia [2]Western Sydney University, Australia [3]Fuzhou University, China
[4]Nanjing University of Science and Technology, China [5]CSIRO, Australia

zhongrui.zhao@yahoo.com z.lu@westernsydney.edu.au longkun.guo@gmail.com







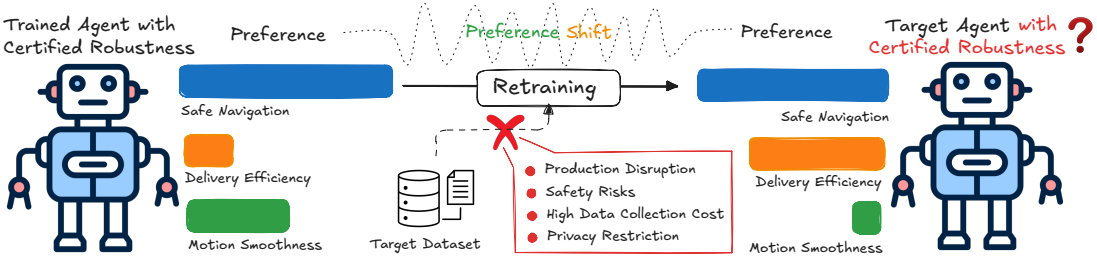




*Equal contribution †Corresponding author

Background

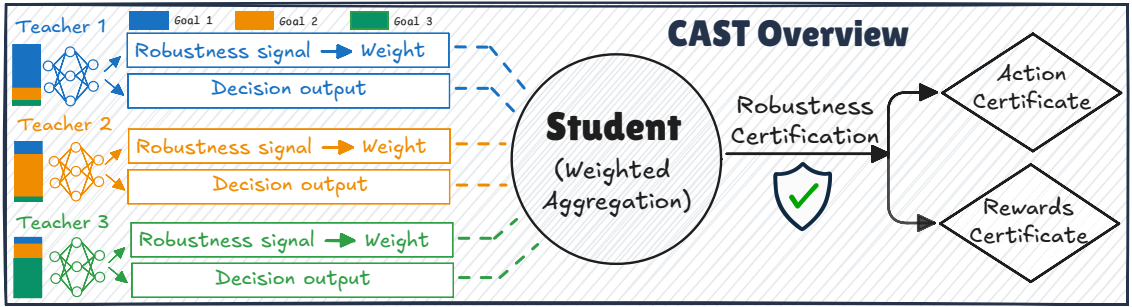
Background: Agents face preference shifts, target-data unavailability, and robustness requirements.



Question: How can policies adapt to preference shifts without target-task data while ensuring certified robustness?

CAST: Certifiable Aggregation of Smoothed Teachers

CAST: an MoE-inspired approach to policy adaptation without target-task data. Instead of training a router, CAST uses **teacher robustness signals** to determine aggregation weights.

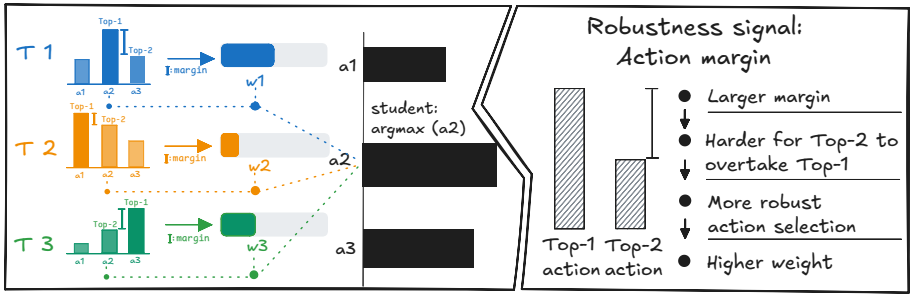


Discrete actions in CAST

Aggregation in discrete scenarios: weighted log-sum-exp of teacher Q-values.

$$Q_{\text{agg}}(s,a) := \frac{1}{\lambda} \log\left(\sum_{i=1}^{N} w_i(s) \exp\left(\lambda \widetilde{Q}_i(s,a)\right)\right)$$

Smoothed teacher action certificate depends on the **top-two action-value margin**[1].



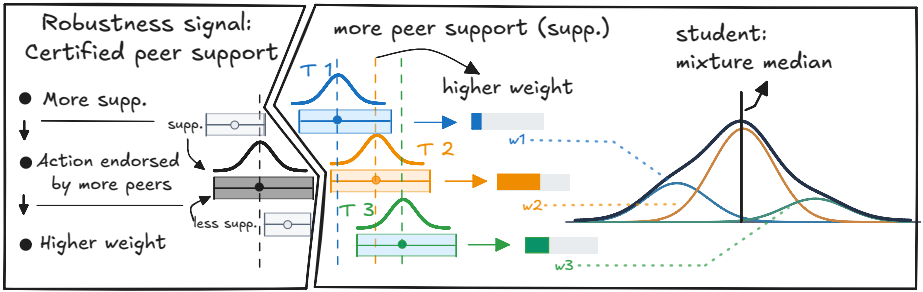
[1] Wu et al., 2022. CROP: Certifying Robust Policies for Reinforcement Learning through Functional Smoothing. ICLR-2022.

Continuous actions in CAST

Aggregation in continuous scenarios: a weighted Gaussian mixture of teachers' policies.

$$F(a \mid s) := \sum_{i=1}^{N} w_i(s)\Phi\left(\frac{a - a_i(s)}{\sigma_a}\right).$$

Smoothed teacher provides a **certified action interval**[2]. A teacher's action is considered acceptable by a peer if it lies within that peer's interval. **More peer support** yields a higher aggregation weight.



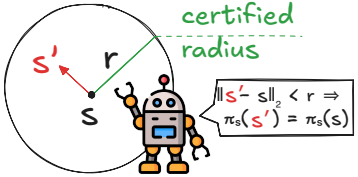
[2] Sun et al., 2024. Breaking the Barrier: Enhanced Utility and Robustness in Smoothed DRL Agents. ICML-2024.

Theoretical results

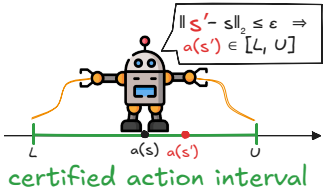
Three robustness guarantees

1. Discrete action robustness
2. Continuous action robustness
3. Cumulative reward robustness

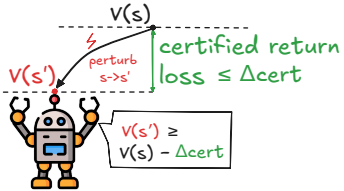
① Action unchanged within the certified radius.



② Action coordinates stay within certified intervals.



③ Certified lower bound on cumulative reward.



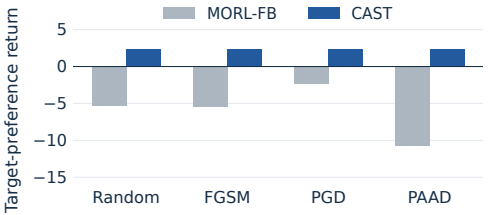
Experimental results

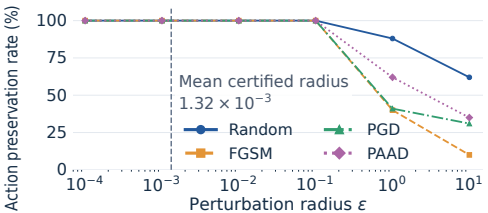
Evaluate utility under attack and validate action certificates.

Utility

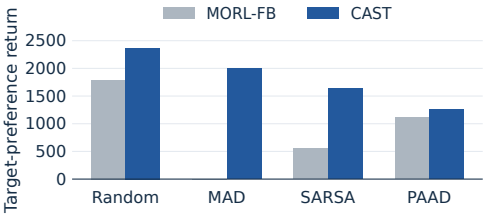
Action certificates

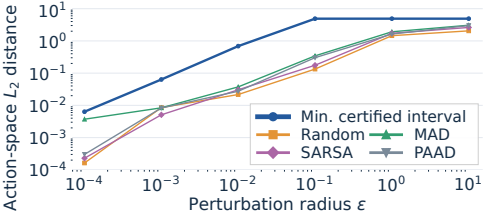
Discrete actions





Continuous actions





Conclusion

01

Adaptation without retraining

CAST aggregates smoothed teachers using robustness-aware weights, without retraining.

02

Action and reward certification

CAST provides discrete and continuous action certificates, and a cumulative reward lower bound.

03

Empirical validation

CAST achieves competitive utility under state perturbations across discrete and continuous benchmarks.

Thank you!

Questions & Discussion

Zhongrui Zhao
zhongrui.zhao@yahoo.com

Zhigang Lu
z.lu@westernsydney.edu.au

Longkun Guo
longkun.guo@gmail.com