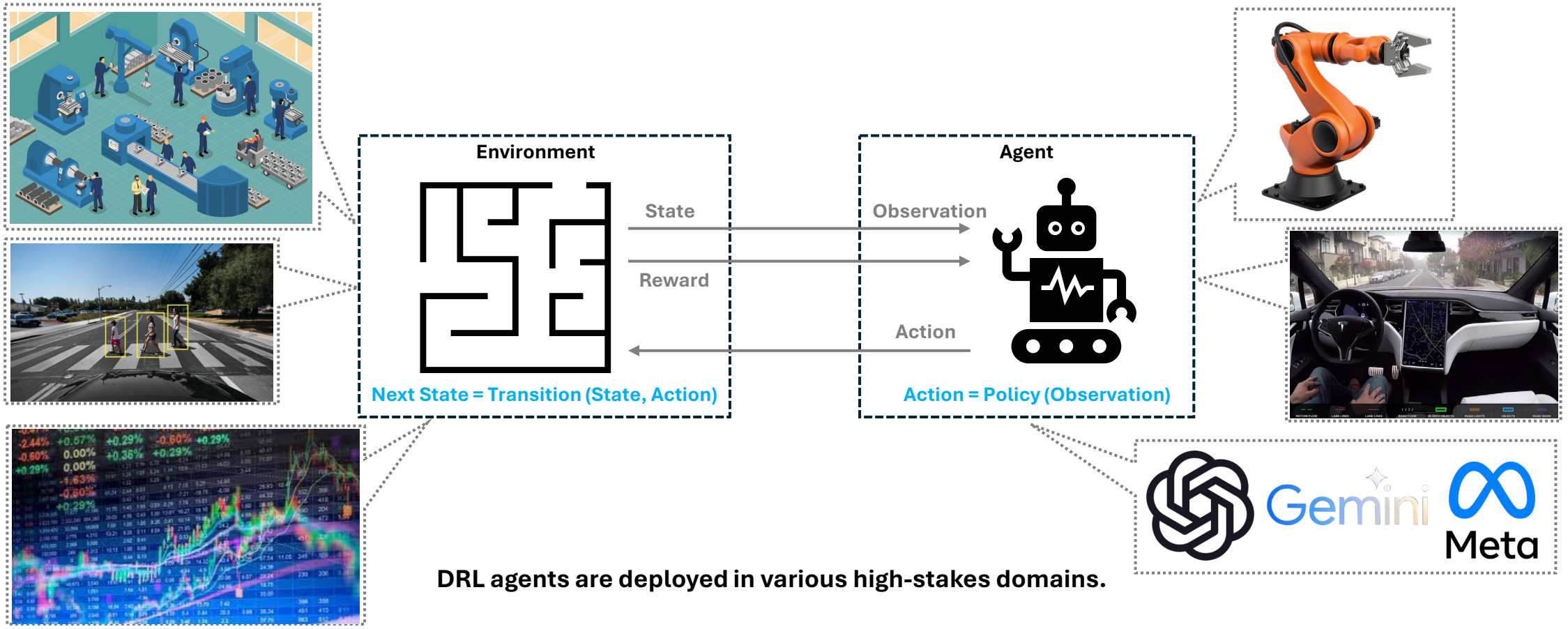
{CAMP} in the Odyssey: Provably Robust Reinforcement Learning with Certified Radius Maximization

Derui Wang* (*Presenter*), Kristen Moore, Diksha Goel, Minjune Kim, Gang Li, Yang Li, Robin Doss, Minhui Xue, Bo Li, Seyit Camtepe, Liming Zhu

- Background and Motivation
- Certified Robustness Enhancement via CAMP
- Agent Training by Policy Imitation
- Experiments
- Conclusion

Background and Motivation

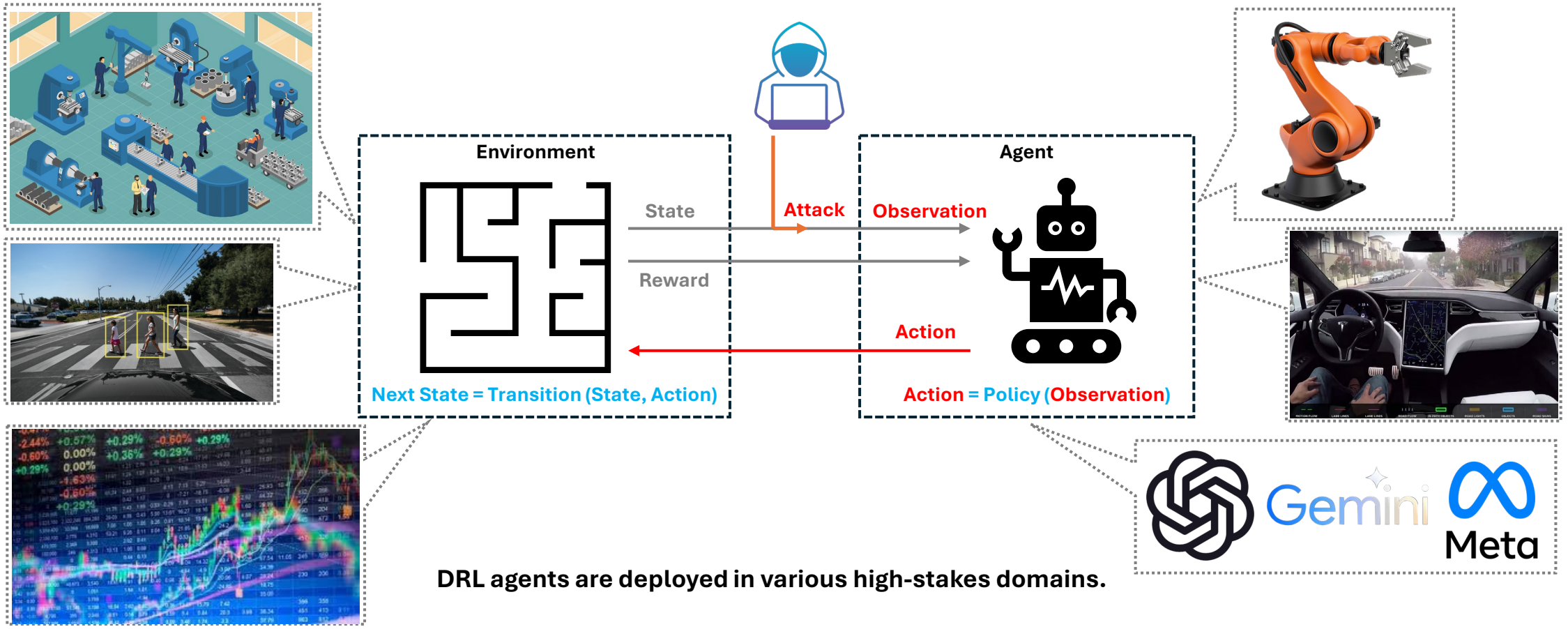
- **Deep Reinforcement Learning Agents are Susceptible to Observation Manipulation**



DRL agents are deployed in various high-stakes domains.

Background and Motivation

- **Deep Reinforcement Learning Agents are Susceptible to Observation Manipulation**



DRL agents are deployed in various high-stakes domains.

Background and Motivation

- **Policy Smoothing (ICLR 2022): Expected Return of Randomized State-Action Trajectories can be Lower Bounded**

Lower Bound on Probability

$$P_{\pi}^{\mathbf{z}'}(C) \geq \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(C)\right) - \frac{\tau}{\sigma}\right), \forall \|\Delta\|_2 \leq \tau$$

"The probability that the return over randomized trajectories z' exceeds C is lower bounded."

Gaussian CDF-Expectation Conversion

Lower Bound on Expected Return

$$\mathbb{E}[F_{\pi}(\mathbf{z}')] \geq \mathbf{r}_1 \cdot \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(\mathbf{r}_1)\right) - \frac{\tau}{\sigma}\right) + \sum_{i=2}^{m} (\mathbf{r}_i - \mathbf{r}_{i-1}) \cdot \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(\mathbf{r}_i)\right) - \frac{\tau}{\sigma}\right) \quad \forall \|\Delta\|_2 \leq \tau.$$

"The expected return is lower bounded."

Background and Motivation

- **Policy Smoothing (ICLR 2022): Expected Return of Randomized State-Action Trajectories can be Lower Bounded**

Lower Bound on Probability

$$P_{\pi}^{\mathbf{z}'}(C) \geq \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(C)\right) - \frac{\tau}{\sigma}\right), \forall \|\Delta\|_2 \leq \tau$$

"The probability that the return over randomized trajectories z' exceeds C is lower bounded."

Gaussian CDF-Expectation Conversion

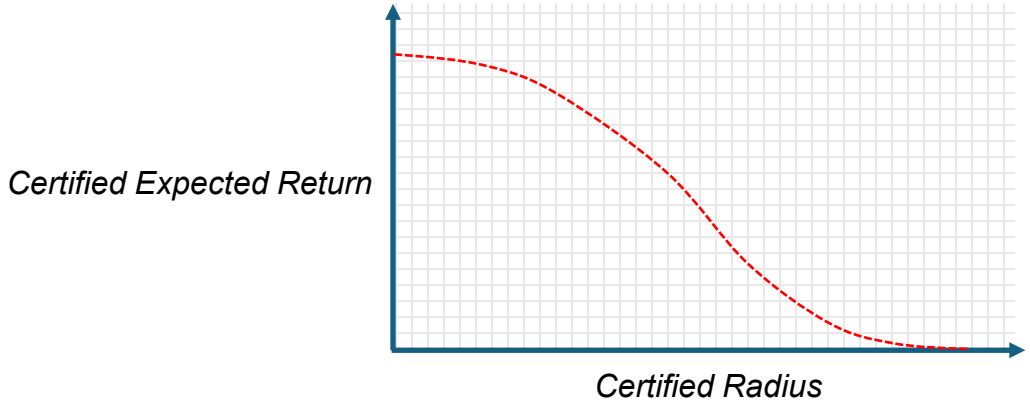
Lower Bound on Expected Return

$$\mathbb{E}[F_{\pi}(\mathbf{z}')] \geq \mathbf{r}_1 \cdot \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(\mathbf{r}_1)\right) - \frac{\tau}{\sigma}\right) + \sum_{i=2}^{m} (\mathbf{r}_i - \mathbf{r}_{i-1}) \cdot \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(\mathbf{r}_i)\right) - \frac{\tau}{\sigma}\right) \quad \forall \|\Delta\|_2 \leq \tau.$$

"The expected return is lower bounded."

- **There is A Trade-Off Between Certified Expected Return and Certified Radius τ**

❑ This trade-off prohibits the deployment of the provably robust agent!



Background and Motivation

- **Policy Smoothing (ICLR 2022): Expected Return of Randomized State-Action Trajectories can be Lower Bounded**

Lower Bound on Probability

$$P_{\pi}^{\mathbf{z}'}(C) \geq \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(C)\right) - \frac{\tau}{\sigma}\right), \forall \|\Delta\|_2 \leq \tau$$

"The probability that the return over randomized trajectories z' exceeds C is lower bounded."

Gaussian CDF-Expectation Conversion

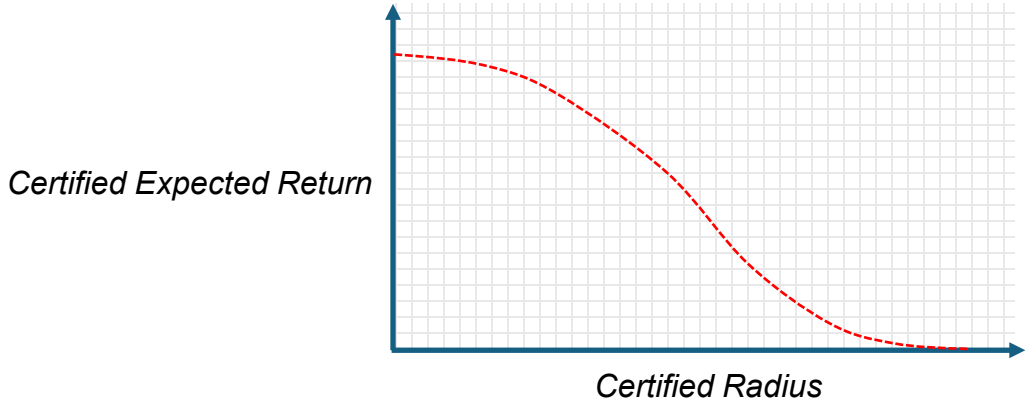
Lower Bound on Expected Return

$$\mathbb{E}[F_{\pi}(\mathbf{z}')] \geq \mathbf{r}_1 \cdot \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(\mathbf{r}_1)\right) - \frac{\tau}{\sigma}\right) + \sum_{i=2}^{m} (\mathbf{r}_i - \mathbf{r}_{i-1}) \cdot \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(\mathbf{r}_i)\right) - \frac{\tau}{\sigma}\right) \quad \forall \|\Delta\|_2 \leq \tau.$$

"The expected return is lower bounded."

- **There is A Trade-Off Between Certified Expected Return and Certified Radius τ**

- ❑ This trade-off prohibits the deployment of the provably robust agent!
- ❑ Existing certification method determines the certified radius τ via binary search



Background and Motivation

- **Policy Smoothing (ICLR 2022): Expected Return of Randomized State-Action Trajectories can be Lower Bounded**

Lower Bound on Probability

$$P_{\pi}^{\mathbf{z}'}(C) \geq \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(C)\right) - \frac{\tau}{\sigma}\right), \forall \|\Delta\|_2 \leq \tau$$

"The probability that the return over randomized trajectories z' exceeds C is lower bounded."

Gaussian CDF-Expectation Conversion

Lower Bound on Expected Return

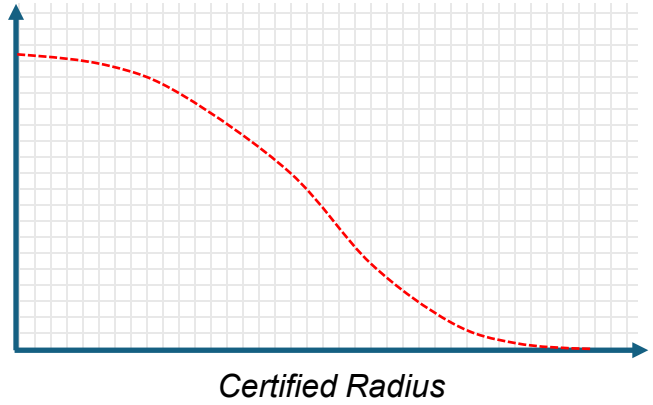
$$\mathbb{E}[F_{\pi}(\mathbf{z}')] \geq \mathbf{r}_1 \cdot \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(\mathbf{r}_1)\right) - \frac{\tau}{\sigma}\right) + \sum_{i=2}^{m} (\mathbf{r}_i - \mathbf{r}_{i-1}) \cdot \Phi\left(\Phi^{-1}\left(\underline{P_{\pi}^{\mathbf{z}}}(\mathbf{r}_i)\right) - \frac{\tau}{\sigma}\right) \quad \forall \|\Delta\|_2 \leq \tau.$$

"The expected return is lower bounded."

- **There is A Trade-Off Between Certified Expected Return and Certified Radius τ**

❑ This trade-off prohibits the deployment of the provably robust agent!

❑ Existing certification method determines the certified radius τ via binary search



Can τ be optimized through other variables?

- Background and Motivation
- Certified Robustness Enhancement via CAMP
- Agent Training by Policy Imitation
- Experiments
- Conclusion

Certified Robustness Enhancement via CAMP

- **How CAMP Works**

Reformulating Certified Radius

Theorem 1 (*Change of variable*). *Given a target expected return threshold* ξ *for the randomized policy, let the perturbed trajectory be* $\mathbf{z}' = \mathbf{z} + \Delta$ *and define* $P_\pi^{\mathbf{z}}(C) := \Pr[F_\pi(\mathbf{z}) \geq C]$. *Let* $\mathbf{R} = \{\mathbf{r}_1, ..., \mathbf{r}_m\}$ *represent a set of sampled and sorted values of* $F_\pi(\mathbf{z})$ *such that* $\mathbf{r}_1 \leq \mathbf{r}_2 \leq ... \leq \mathbf{r}_m$. *If* $\xi/\mathbf{r}_1 \leq \underline{P_\pi^{\mathbf{z}}}(\mathbf{r}_1))$ *and*

$$\|\Delta\|_2 \leq \sigma \left[\Phi^{-1}(\underline{P_\pi^{\mathbf{z}}}(\mathbf{r}_1)) - \Phi^{-1}(\xi/\mathbf{r}_1) \right], \tag{8}$$

then $\mathbb{E}[F_\pi(\mathbf{z}')] \geq \xi$.

τ is Changed to a Function of Variables

Certified Robustness Enhancement via CAMP

- **How CAMP Works**

Reformulating Certified Radius

Theorem 1 (*Change of variable*). *Given a target expected return threshold ξ for the randomized policy, let the perturbed trajectory be $\mathbf{z}' = \mathbf{z} + \Delta$ and define $P^{\mathbf{z}}_{\pi}(C) := \Pr[F_\pi(\mathbf{z}) \geq C]$. Let $\mathbf{R} = \{\mathbf{r}_1, ..., \mathbf{r}_m\}$ represent a set of sampled and sorted values of $F_\pi(\mathbf{z})$ such that $\mathbf{r}_1 \leq \mathbf{r}_2 \leq ... \leq \mathbf{r}_m$. If $\xi/\mathbf{r}_1 \leq \underline{P^{\mathbf{z}}_{\pi}}(\mathbf{r}_1))$ and*

$$\|\Delta\|_2 \leq \sigma \left[\Phi^{-1}(\underline{P^{\mathbf{z}}_{\pi}}(\mathbf{r}_1)) - \Phi^{-1}(\xi/\mathbf{r}_1)\right], \quad (8)$$

then $\mathbb{E}[F_\pi(\mathbf{z}')] \geq \xi$.

Surrogate Radius

Theorem 3 (*Soft certified radius*). *Given a target expected return $\xi \leq \mathbb{E}[F_\pi(\mathbf{z})]$ where $F_\pi : \mathbf{z} \in (\mathbb{S} \times \mathbb{A} \times \mathbb{R}^d)^T \to [A, B]$, suppose an adversary perturbs the observed states by applying $\Delta = (\delta_0, \delta_1, ..., \delta_{T-1})$. The target expected return will not drop below ξ if the perturbations satisfy:*

$$\|\Delta\|_2 \leq \sigma \left[\Phi^{-1}\left(\frac{\mathbb{E}[F_\pi(\mathbf{z})] - A}{B - A}\right) - \Phi^{-1}\left(\frac{\xi - A}{B - A}\right)\right]. \quad (12)$$

τ is Changed to a Function of Variables

An Optimizable Surrogate Function of τ

Certified Robustness Enhancement via CAMP

How CAMP Works

Reformulating Certified Radius

Theorem 1 (*Change of variable*). *Given a target expected return threshold ξ for the randomized policy, let the perturbed trajectory be $\mathbf{z}' = \mathbf{z} + \Delta$ and define $P_\pi^{\mathbf{z}}(C) := \Pr[F_\pi(\mathbf{z}) \geq C]$. Let $\mathbf{R} = \{\mathbf{r}_1, ..., \mathbf{r}_m\}$ represent a set of sampled and sorted values of $F_\pi(\mathbf{z})$ such that $\mathbf{r}_1 \leq \mathbf{r}_2 \leq ... \leq \mathbf{r}_m$. If $\xi/\mathbf{r}_1 \leq \underline{P_\pi^{\mathbf{z}}}(\mathbf{r}_1))$ and*

$$\|\Delta\|_2 \leq \sigma \left[\Phi^{-1}(\underline{P_\pi^{\mathbf{z}}}(\mathbf{r}_1)) - \Phi^{-1}(\xi/\mathbf{r}_1)\right], \quad (8)$$

then $\mathbb{E}[F_\pi(\mathbf{z}')] \geq \xi$.

τ is Changed to a Function of Variables

Surrogate Radius

Theorem 3 (*Soft certified radius*). *Given a target expected return $\xi \leq \mathbb{E}[F_\pi(\mathbf{z})]$ where $F_\pi : \mathbf{z} \in (\mathbb{S} \times \mathbb{A} \times \mathbb{R}^d)^T \to [A, B]$, suppose an adversary perturbs the observed states by applying $\Delta = (\delta_0, \delta_1, ..., \delta_{T-1})$. The target expected return will not drop below ξ if the perturbations satisfy:*

$$\|\Delta\|_2 \leq \sigma \left[\Phi^{-1}\left(\frac{\mathbb{E}[F_\pi(\mathbf{z})] - A}{B - A}\right) - \Phi^{-1}\left(\frac{\xi - A}{B - A}\right)\right]. \quad (12)$$

An Optimizable Surrogate Function of τ

Local Radius Maximization

Theorem 4 (*Local certified radius*). *Let l_i and u_i denote the lower and upper bounds of the expected action-value function at step i, respectively. Let the perturbations from step t to step $T-1$ be $\{\delta_i\}_{i=t}^{T-1}$. Under a greedy policy, the optimal reward at step t is obtained by taking action $a_i^{(1)} = \arg\max_{a_i} \bar{Q}_{\pi^*}(s_i, a_i)$. The optimal expected return from step t to $T-1$ will not be reduced if the local perturbation δ_i at each step i satisfies:*

$$\|\delta_i\|_2 \leq \frac{\sigma}{2}\left[\Phi^{-1}\left(\frac{\bar{Q}_{\pi^*}(s_i, a_i^{(1)}) - l_i}{u_i - l_i}\right) - \Phi^{-1}\left(\frac{\bar{Q}_{\pi^*}(s_i, a_i^{(2)}) - l_t}{u_i - l_i}\right)\right], \quad (16)$$

where $a_i^{(2)}$ is the runner-up action $a_i^{(2)} = \arg\max_{a_i : a_i \neq a_i^{(1)}} \bar{Q}_{\pi^*}(s_i, a_i)$.

Break Down the Optimizable Surrogate Function into Per-step Q-Value Gaps

Certified Robustness Enhancement via CAMP

- **How CAMP Works**

Local Radius Maximization

Theorem 4 (*Local certified radius*). *Let l_i and u_i denote the lower and upper bounds of the expected action-value function at step i, respectively. Let the perturbations from step t to step $T-1$ be $\{\delta_i\}_{i=t}^{T-1}$. Under a greedy policy, the optimal reward at step t is obtained by taking action $a_i^{(1)} = \arg\max_{a_i} \bar{Q}_{\pi^*}(s_i, a_i)$. The optimal expected return from step t to $T-1$ will not be reduced if the local perturbation δ_i at each step i satisfies:*

$$\|\delta_i\|_2 \leq \frac{\sigma}{2}\left[\Phi^{-1}\left(\frac{\bar{Q}_{\pi^*}(s_i, a_i^{(1)}) - l_i}{u_i - l_i}\right) - \Phi^{-1}\left(\frac{\bar{Q}_{\pi^*}(s_i, a_i^{(2)}) - l_t}{u_i - l_i}\right)\right], \quad (16)$$

where $a_i^{(2)}$ is the runner-up action $a_i^{(2)} = \arg\max_{a_i : a_i \neq a_i^{(1)}} \bar{Q}_{\pi^*}(s_i, a_i)$.

The Certified Radius is Positively Correlated to the Q-Value Gap between Top-1 and Runner-up Actions

CAMP Loss

When the oracle policy suggests a top-1 action and a runner up action

$$\ell_{ro}(\pi, \tilde{\pi}; s, \varepsilon) = \lambda \cdot \mathbb{1}_{\{Q_{\tilde{\pi}}(s+\varepsilon, a^{(1)}) \geq Q_{\tilde{\pi}}(s+\varepsilon, a^{(2)})\}} \max\{0, \eta - [Q_\pi(s+\varepsilon, a^{(1)}) - Q_\pi(s+\varepsilon, a^{(2)})]\},$$

$$a^{(1)} = \arg\max_a \pi(s+\varepsilon)_a,$$

$$a^{(2)} = \arg\max_{a : a \neq a^{(1)}} \pi(s+\varepsilon)_a.$$

CAMP accordingly maximizes the Q-Value gap between the two actions

- Background and Motivation
- Certified Robustness Enhancement via CAMP
- Agent Training by Policy Imitation
- Experiments
- Conclusion

Agent Training by Policy Imitation

- **Why CAMP Loss Cannot Be Directly Applied to the Training Loss**

Time-Difference Loss in Q-Learning

$$\min_{\pi} \mathop{\mathbb{E}}_{(s,a,s')\sim \mathbf{Z}} \left[\underbrace{Q_{\pi}(s,a)}_{\text{Current Q Value}} - \underbrace{\left(\mathcal{R}(s,a) + \gamma \max_{a'} Q_{\pi^*}(s',a') \right)}_{\text{Target Q Value}} \right]^2$$

Agent Training by Policy Imitation

- **Why CAMP Loss Cannot Be Directly Applied to the Training Loss**

Time-Difference Loss in Q-Learning

$$\min_{\pi} \mathop{\mathbb{E}}_{(s,a,s')\sim \mathbf{Z}} \left[\underbrace{Q_\pi(s,a)}_{\text{Current Q Value}} - \underbrace{\left(\mathcal{R}(s,a) + \gamma \max_{a'} Q_{\pi^*}(s',a') \right)}_{\text{Target Q Value}} \right]^2$$

Overestimating Q Values Leads to Sub-Optimal Convergence
(From NIPS'10 to NeurIPS'22)

[NIPS'10] *Hado V. Hasselt, "Double Q-learning."*
[NeurIPS'22] *Rui Yang, et al. "RORL: Robust Offline Reinforcement Learning via Conservative Smoothing."*

Agent Training by Policy Imitation

- **Why CAMP Loss Cannot Be Directly Applied to the Training Loss**

Time-Difference Loss in Q-Learning

$$\min_{\pi} \mathop{\mathbb{E}}_{(s,a,s')\sim \mathbf{Z}} \left[Q_{\pi}(s,a) - \left(\mathcal{R}(s,a) + \gamma \max_{a'} Q_{\pi^*}(s',a') \right) \right]^2$$

Current Q Value — Target Q Value

Overestimating Q Values Leads to Sub-Optimal Convergence
(From NIPS'10 to NeurIPS'22)

$$\ell_{ro}(\pi, \tilde{\pi}; s, \varepsilon) = \lambda \cdot \mathbb{1}_{\{Q_{\tilde{\pi}}(s+\varepsilon, a^{(1)}) \geq Q_{\tilde{\pi}}(s+\varepsilon, a^{(2)})\}}$$
$$\max\{0, \eta - [Q_{\pi}(s+\varepsilon, a^{(1)}) - Q_{\pi}(s+\varepsilon, a^{(2)})]\},$$

CAMP Loss Could Increase the Maximal (Top-1) Q Value

[NIPS'10] *Hado V. Hasselt, "Double Q-learning."*
[NeurIPS'22] *Rui Yang, et al. "RORL: Robust Offline Reinforcement Learning via Conservative Smoothing."*

Agent Training by Policy Imitation

- **Policy Imitation Stabilizes the Q-Learning Process**

  Intuition:

  ❑ Use a reference policy to model the oracle policy

  ❑ The primary policy imitates the actions, but not the Q-values, of the reference policy

Agent Training by Policy Imitation

- **Policy Imitation Stabilizes the Q-Learning Process**

  Intuition:

  ❑ Use a reference policy to model the oracle policy

  ❑ The primary policy imitates the actions, but not the Q-values, of the reference policy

Imitation Loss for the Primary Policy

$$\ell_{im}(\pi, \tilde{\pi}; s, \varepsilon) = -\sum_{i=1}^{|\mathbb{A}|} Softmax(\tilde{\pi}(s+\varepsilon))_i \log Softmax(\pi(s+\varepsilon))_i$$

Avoid Direct Comparison to a Target Q-Value

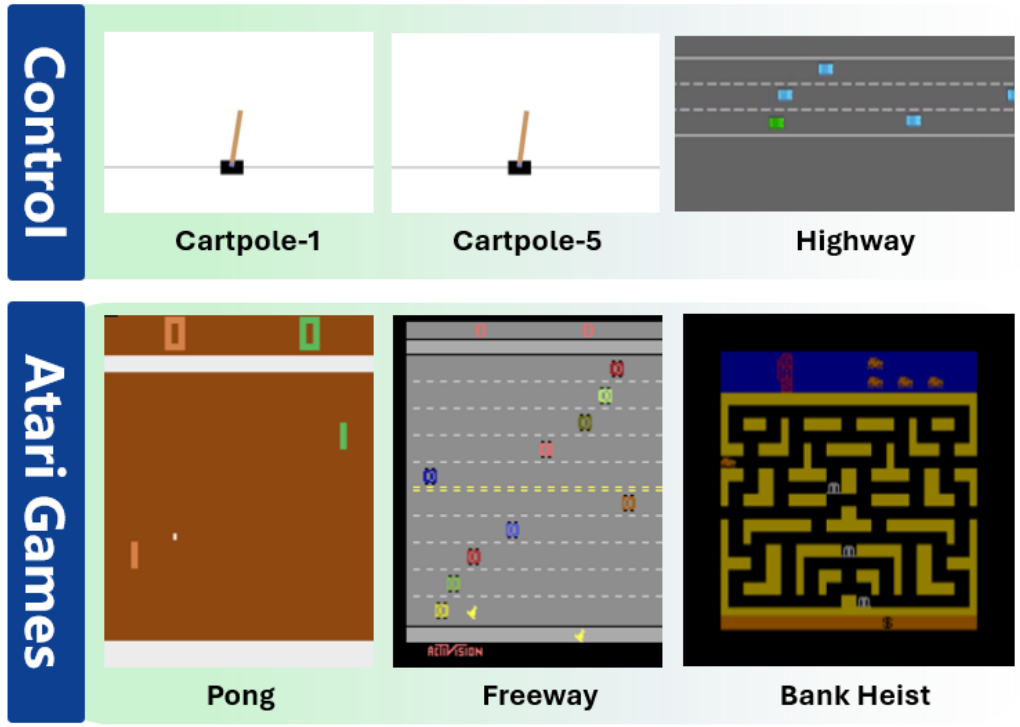
CAMP Loss can be Appended to the Imitation Loss without Affecting the Q-Value

- Background and Motivation
- Certified Robustness Enhancement via CAMP
- Agent Training by Policy Imitation
- Experiments
- Conclusion

Experiments

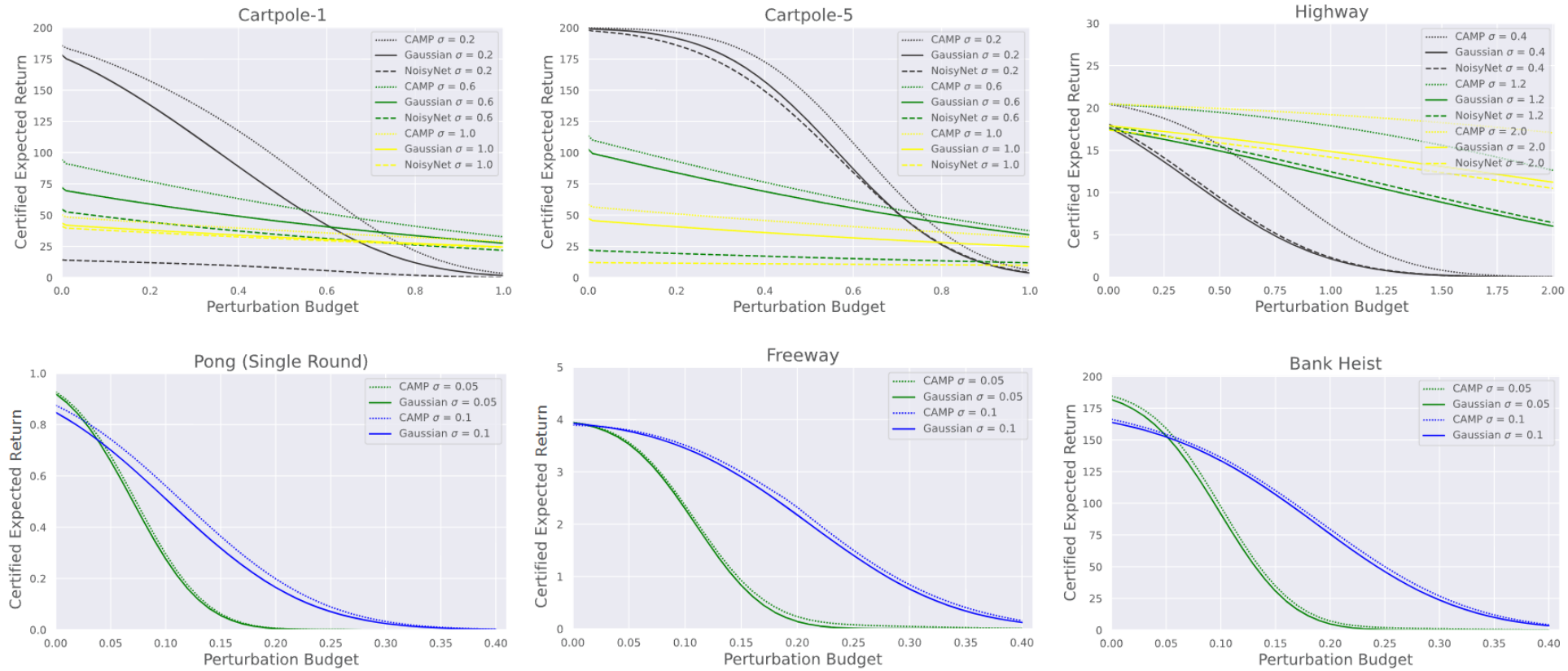
- **CAMP is Benchmarked Across 6 Environments**
  - ✓ Cartpole-1 (1 frame)
  - ✓ Cartpole-5 (5 consecutive frames)
  - ✓ Highway
  - ✓ Pong
  - ✓ Freeway
  - ✓ Bank Heist

- **Both Certified Robustness and Empirical Robustness**
  - ✓ Certified Expected Return – Certified Radius
  - ✓ Single-Episode Return – Perturbation Budget



Experiments

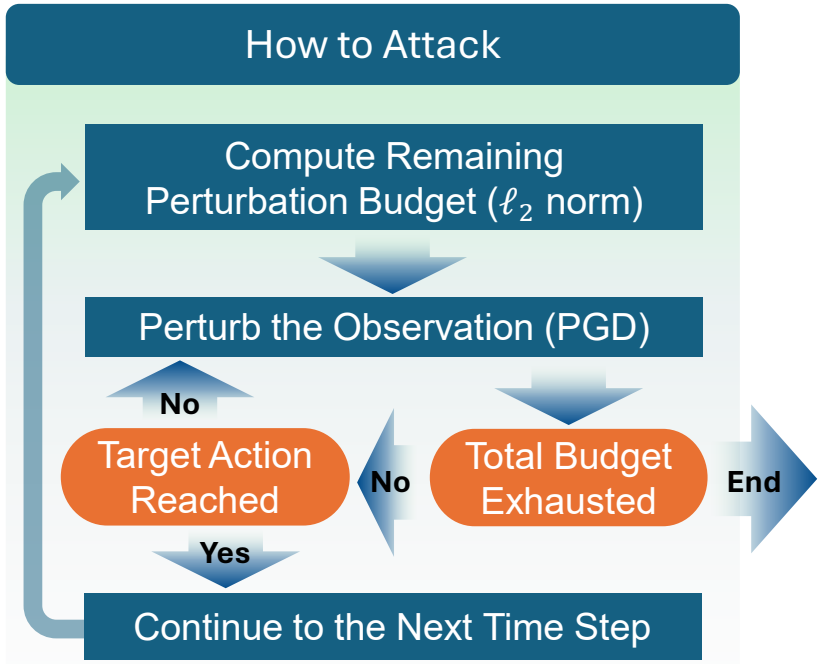
- **CAMP Significantly Enhances Certified Expected Return of Q-Learning Agents**



Agents show more significant improvements on classic control tasks.

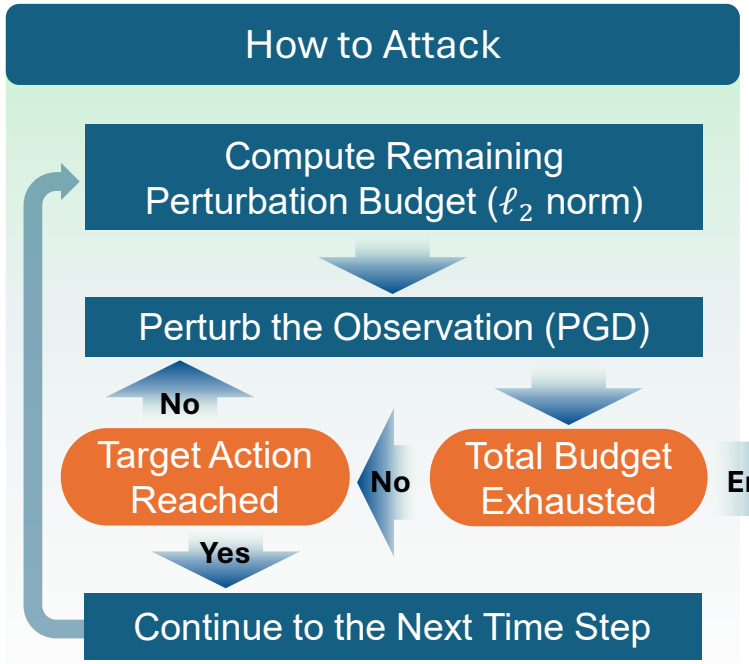
Experiments

- **Empirical Episodic Robustness: CAMP Agents are More Robust Against PGD**

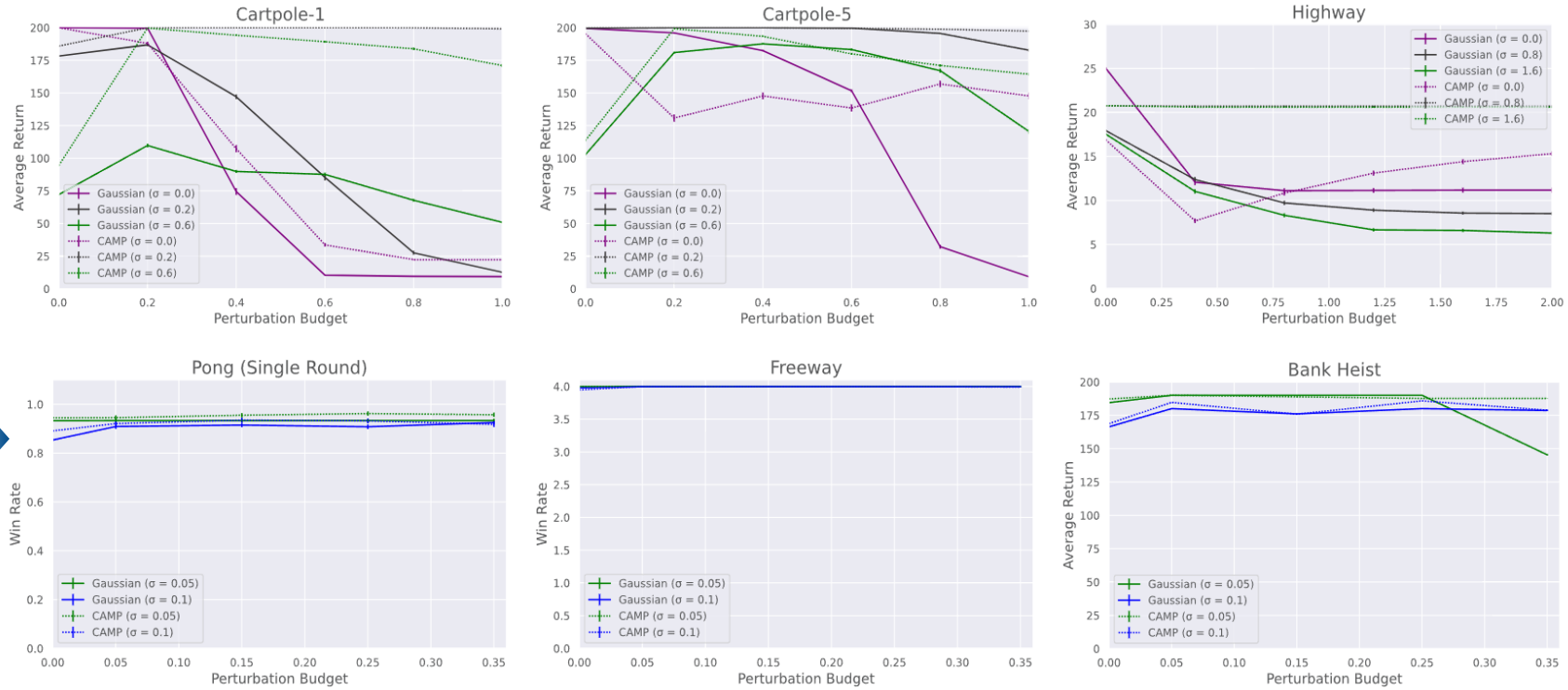


Experiments

- **Empirical Episodic Robustness: CAMP Agents are More Robust Against PGD**



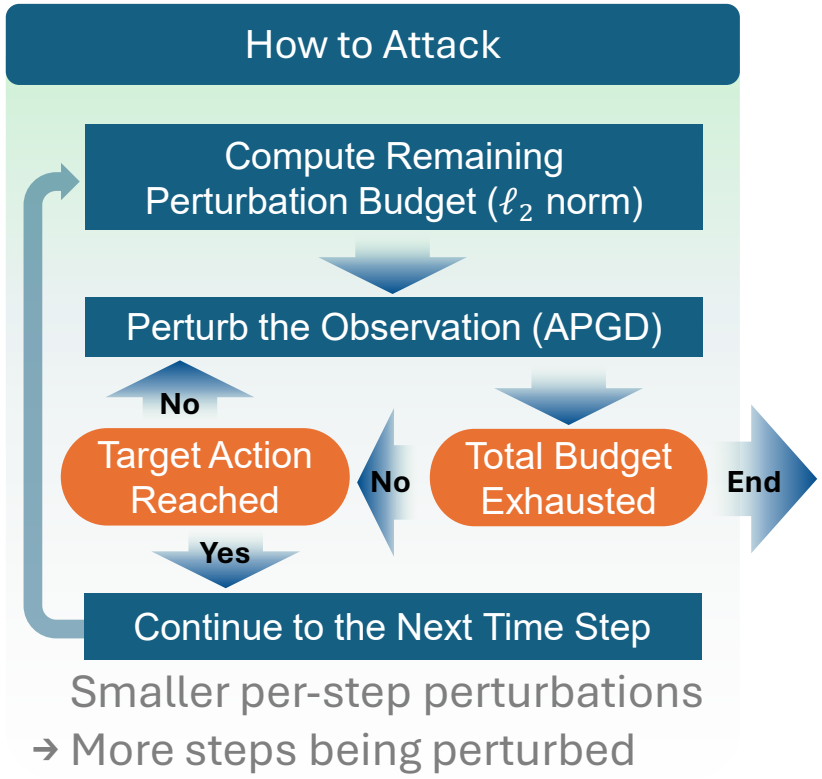
Average return over 1k independent episodes



Agents that were once vulnerable have been made robust.

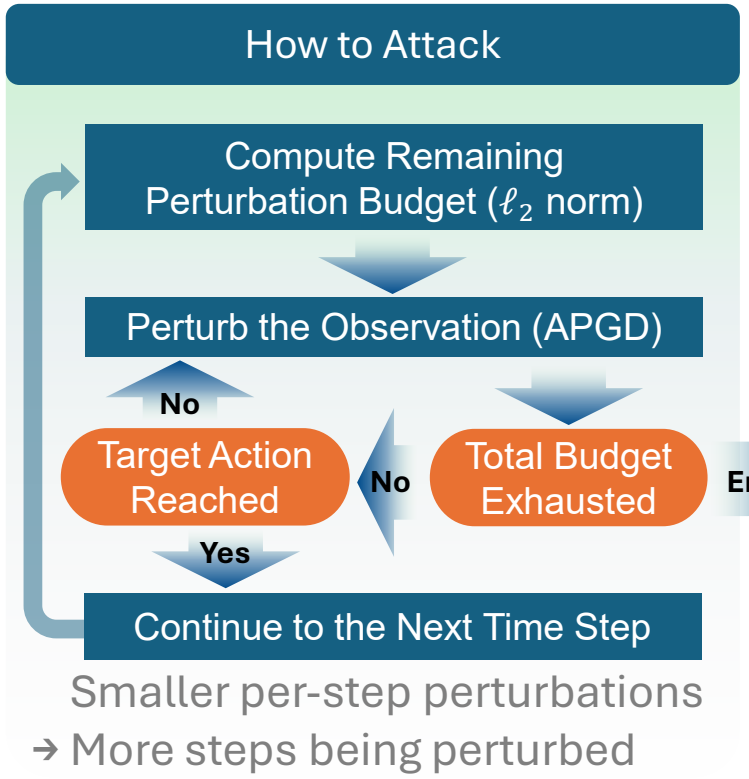
Experiments

- **Empirical Episodic Robustness: CAMP Agents are More Robust Against AutoPGD**

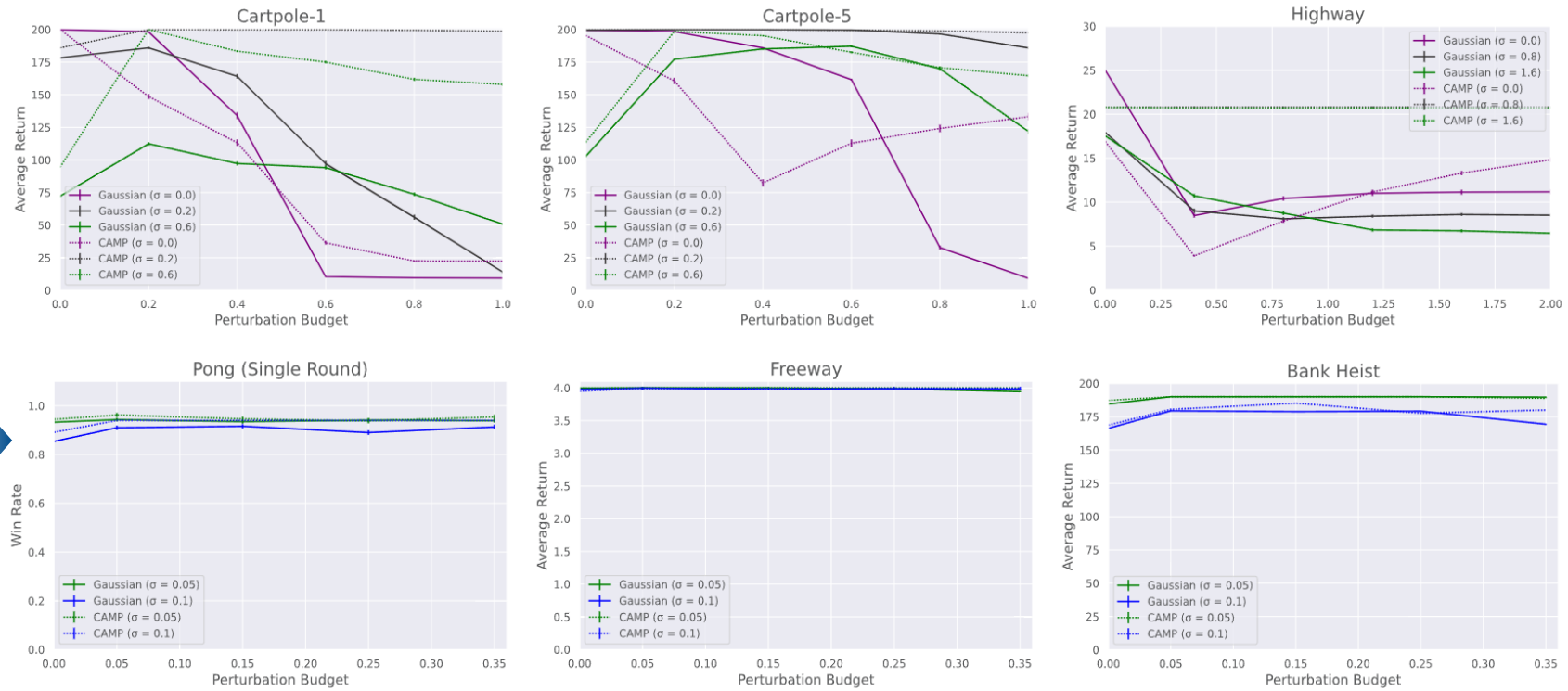


Experiments

- **Empirical Episodic Robustness: CAMP Agents are More Robust Against AutoPGD**



Average return over 1k independent episodes

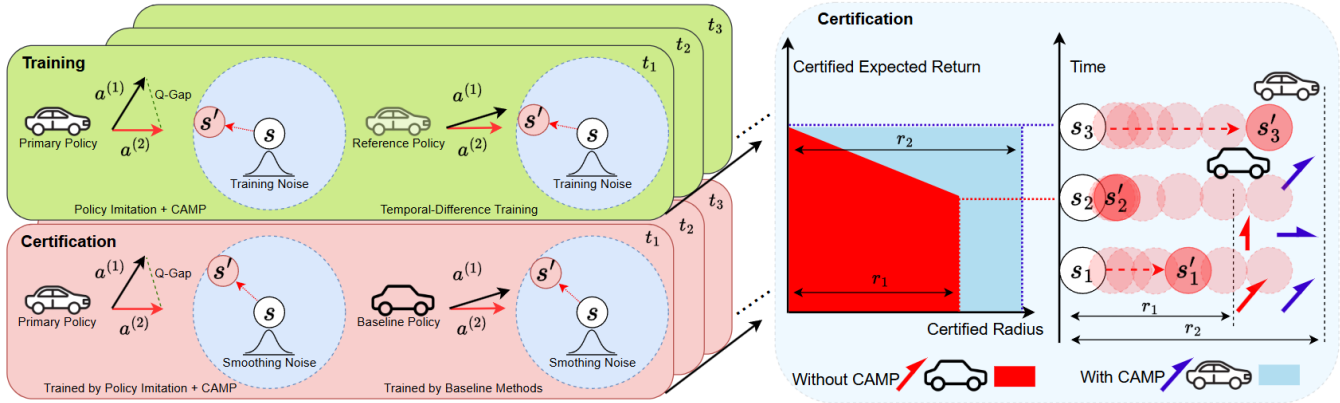


CAMP agents are robust to the AutoPGD attacks as well.

- Background and Motivation
- Certified Robustness Enhancement via CAMP
- Agent Training by Policy Imitation
- Experiments
- Conclusion

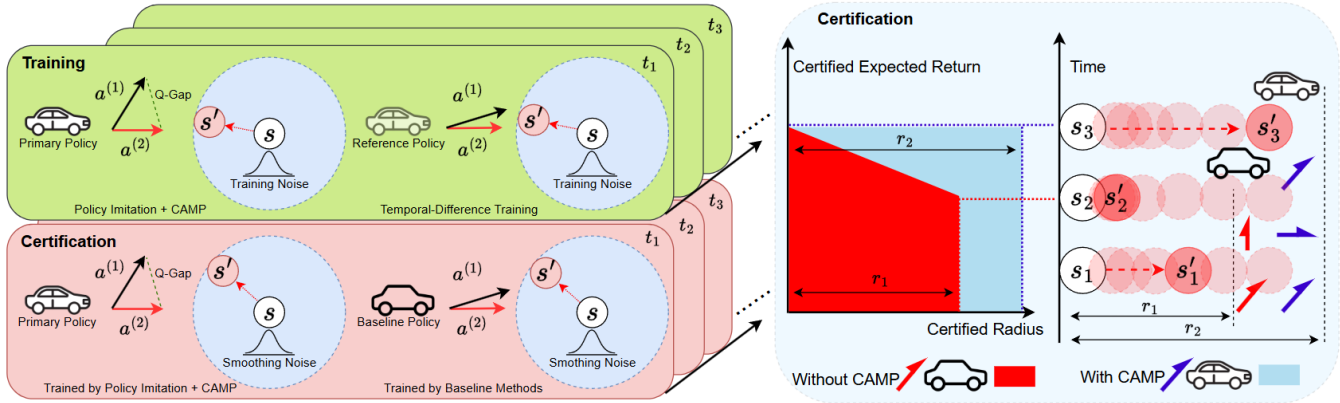
Conclusion

- **Advantages of CAMP + Policy Imitation**
  - ❑ Explainable improvements in certified robustness;
  - ❑ Significant gains in both certified and empirical robustness in classic control agents;
  - ❑ Scalable to environments with large discrete action spaces.



Conclusion

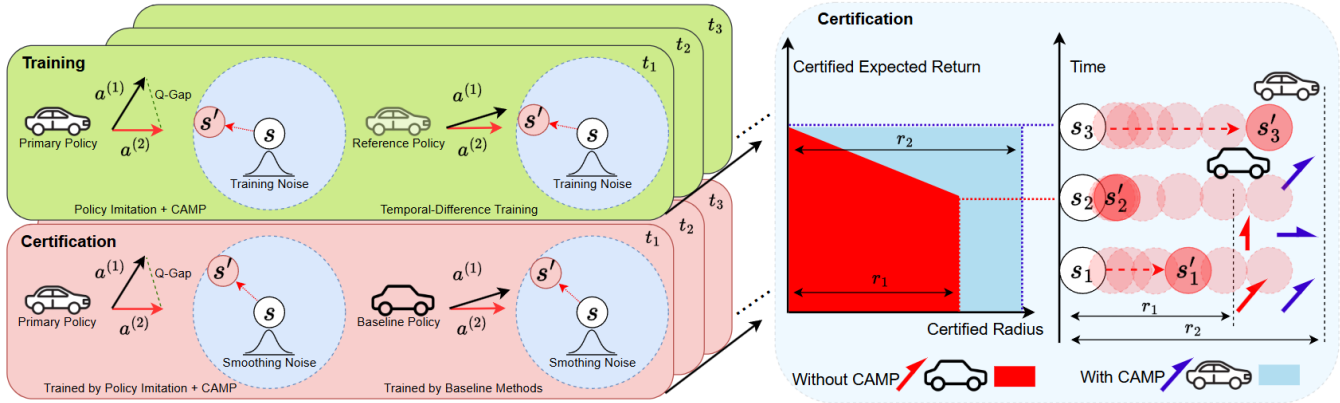
- **Advantages of CAMP + Policy Imitation**
  - ❑ Explainable improvements in certified robustness;
  - ❑ Significant gains in both certified and empirical robustness in classic control agents;
  - ❑ Scalable to environments with large discrete action spaces.



- **Limitations**
  - ❑ Limited to discrete action spaces;
  - ❑ Smaller improvements on Atari games (when baseline agents already exhibit higher robustness);
  - ❑ Increased training overhead.

Conclusion

- **Advantages of CAMP + Policy Imitation**
  - ❑ Explainable improvements in certified robustness;
  - ❑ Significant gains in both certified and empirical robustness in classic control agents;
  - ❑ Scalable to environments with large discrete action spaces.



- **Limitations**
  - ❑ Limited to discrete action spaces;
  - ❑ Smaller improvements on Atari games (when baseline agents already exhibit higher robustness);
  - ❑ Increased training overhead.

Thank you for your attention!

Derui (Derek) Wang
Research Scientist | **Data61, CSIRO** Australia's National Science Agency
derek.wang@data61.csiro.au