Constrained Offline Policy Optimization

Nicholas Polosky[1] Bruno C. da Silva[2 †] Madalina Fiterau[2 †] Jithin Jagannath[1 †]

Contact: npolosky@androcs.com

[1]ANDRO Computational Solutions, LLC

[2]University of Massachusetts Amherst

[†]Shared Senior Authorship

MOTIVATION & BACKGROUND

Trained via
Offline RL

Issues:

1. Unreliable policy evaluation (distributional shift)
2. Does the data accurately represent the underlying MDP?

MOTIVATION & BACKGROUND

Trained via
Offline RL

Issues:

1. Unreliable policy evaluation (distributional shift)
2. Does the data accurately represent the underlying MDP?

Question: Is offline RL suitable for safety-critical environments?

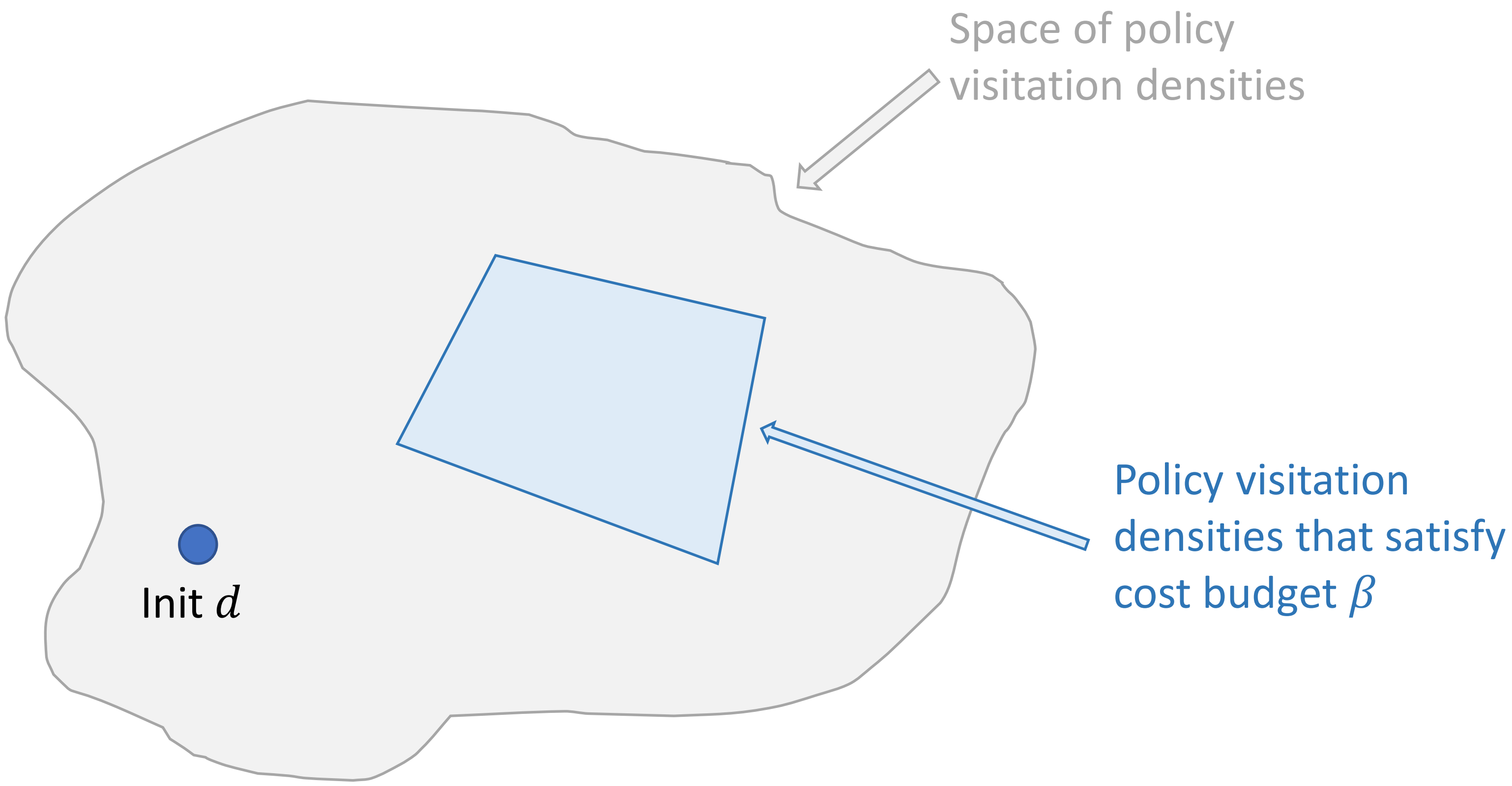
COPO: Offline RL with Safety Guarantees

Model safety-critical environments using Constrained Markov Decision Processes (CMDPs)

COPO: Offline RL with Safety Guarantees

Model safety-critical environments using Constrained Markov Decision Processes (CMDPs)

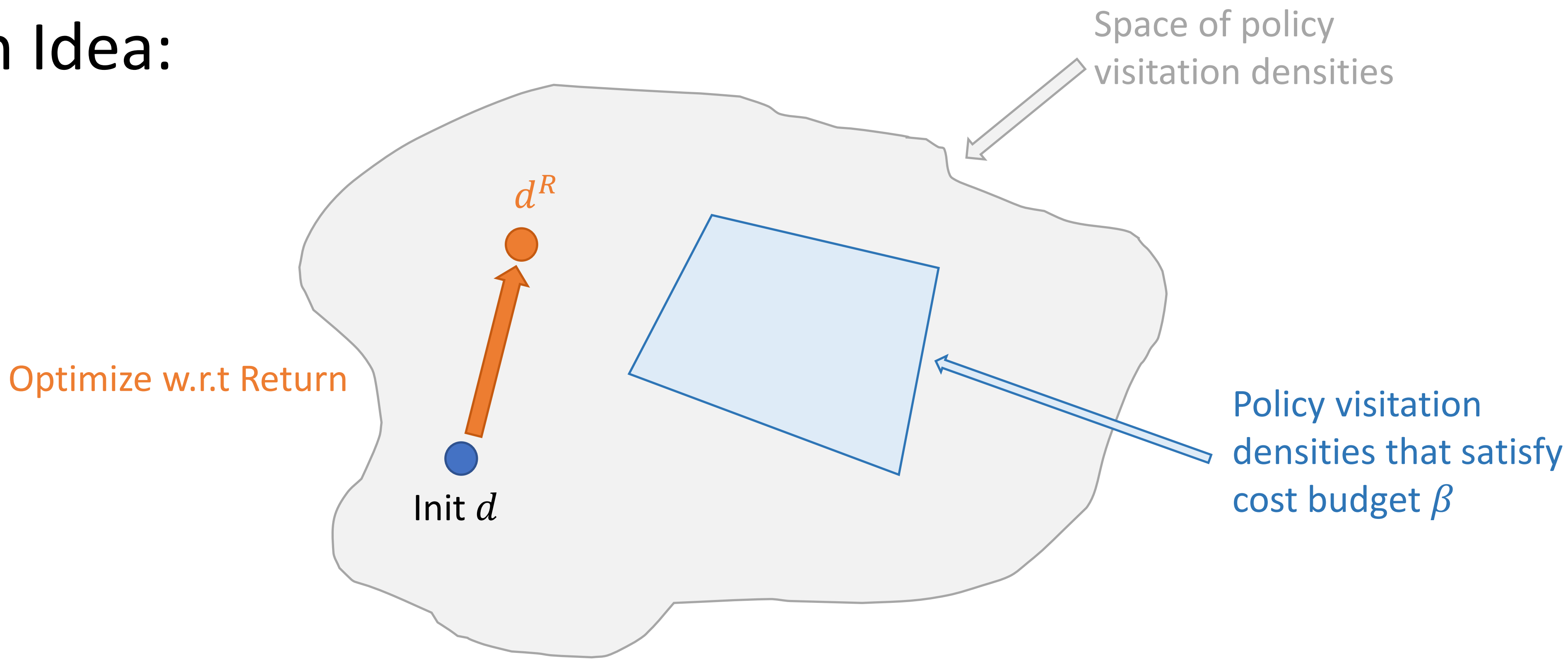
Main Idea:



COPO: Offline RL with Safety Guarantees

Model safety-critical environments using Constrained Markov Decision Processes (CMDPs)

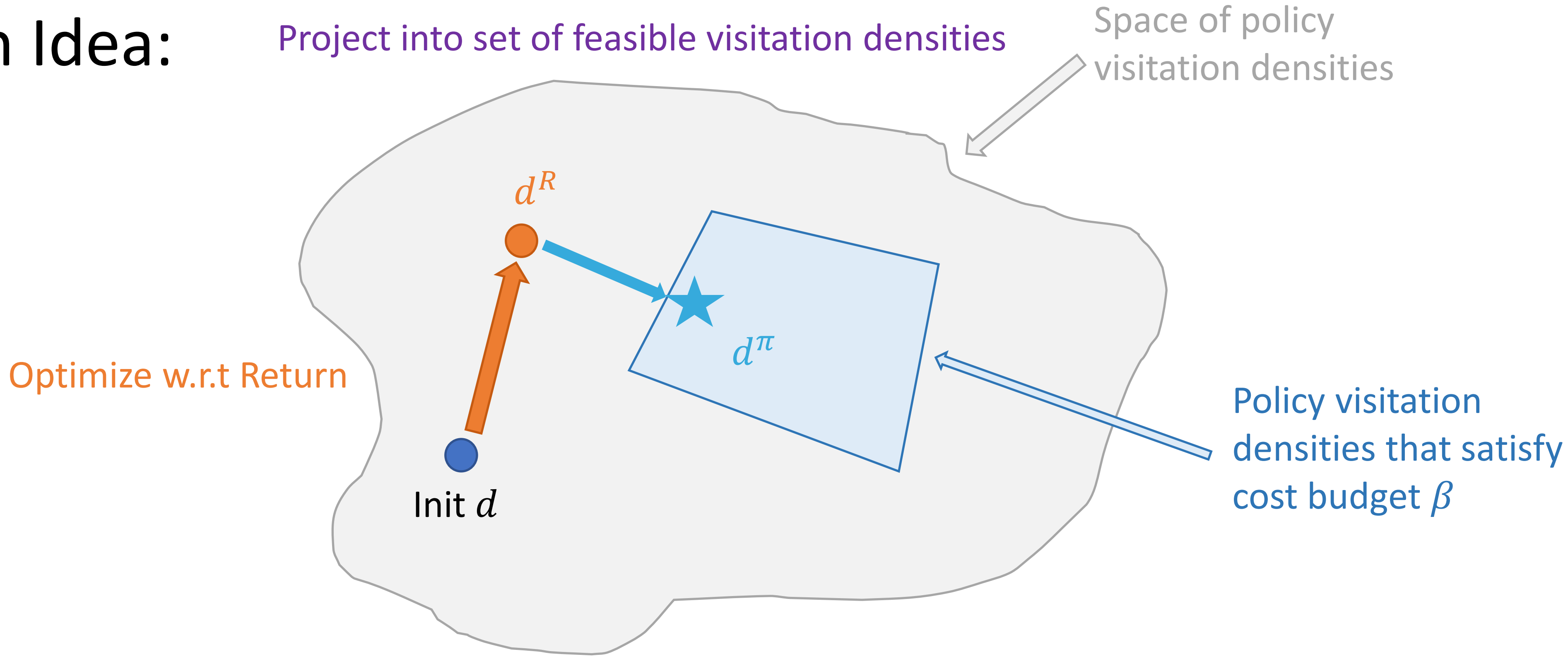
Main Idea:



COPO: Offline RL with Safety Guarantees

Model safety-critical environments using Constrained Markov Decision Processes (CMDPs)

Main Idea:



Novel Constrained Projection

Assume we are given the visitation density, d^R, of a policy that maximizes return

Novel Constrained Projection

Assume we are given the visitation density, d^R, of a policy that maximizes return

We want to find the policy with the *closest* visitation to d^R that satisfies the cost budget β

Problem:

$$\min_{\pi} D(d^{\pi}, d^R) \quad \text{s.t.} \quad \rho_C(\pi) \leq \beta$$

where $\rho_C(\pi)$ is the cost-value of the policy, β is the cost budget, and D is a metric or pseudo-metric on policy space.

Novel Constrained Projection

Assume we are given the visitation density, d^R, of a policy that maximizes return

We want to find the policy with the *closest* visitation to d^R that satisfies the cost budget β

Problem:

$$\min_{\pi} D(d^{\pi}, d^{R}) \quad \text{s.t.} \quad \rho_C(\pi) \leq \beta$$

where $\rho_C(\pi)$ is the cost-value of the policy, β is the cost budget, and D is a metric or pseudo-metric on policy space.

Take the Lagrangian ➡ Expand the OPE term ➡ Rearrange, we get:

$$\min_{\pi} \max_{\lambda \geq 0, \nu} \min_{d} D(d, d^R) + \sum d(s,a)(\lambda C(s,a) + \nu) - (\lambda\beta + \nu) \quad s.t. \; d(s,a) = P_*^{\pi} d(s,a)$$

Novel Constrained Projection

Assume we are given the visitation density, d^R, of a policy that maximizes return

We want to find the policy with the *closest* visitation to d^R that satisfies the cost budget β

Problem:

$$\min_{\pi} D(d^{\pi}, d^R) \quad \text{s.t.} \quad \rho_C(\pi) \le \beta$$

where $\rho_C(\pi)$ is the cost-value of the policy, β is the cost budget, and D is a metric or pseudo-metric on policy space.

Take the Lagrangian ➡ Expand the OPE term ➡ Rearrange, we get:

$$\min_{\pi} \max_{\lambda \ge 0, \nu} \min_{d} D(d, d^R) + \sum d(s,a)(\lambda C(s,a) + \nu) - (\lambda\beta + \nu) \quad s.t.\, d(s,a) = P_*^{\pi} d(s,a)$$

We will leverage ideas from the DIstribution Correction Estimation (DICE) Offline RL framework

Novel Constrained Projection

Assume we are given the visitation density, d^R, of a policy that maximizes return

We want to find the policy with the *closest* visitation to d^R that satisfies the cost budget β

Problem:

$$\min_{\pi} D(d^{\pi}, d^R) \quad \text{s.t.} \;\; \rho_C(\pi) \leq \beta$$

where $\rho_C(\pi)$ is the cost-value of the policy, β is the cost budget, and D is a metric or pseudo-metric on policy space.

Take the Lagrangian ➡ Expand the OPE term ➡ Rearrange, we get:

$$\min_{\pi} \max_{\lambda \geq 0, \nu} \min_{d} D(d, d^R) + \sum d(s,a)(\lambda C(s,a) + \nu) - (\lambda\beta + \nu) \quad s.t. \; d(s,a) = P_*^{\pi} d(s,a)$$

We will leverage ideas from the DIstribution Correction Estimation (DICE) Offline RL framework

From DICE and Fenchel-Rockafeller duality we have:

Primal $\quad \min_{x} f(x) + g(Ax)$

Dual $\quad \max_{y} -f_*(A_* y) - g_*(y)$

Novel Constrained Projection

Assume we are given the visitation density, d^R, of a policy that maximizes return

We want to find the policy with the *closest* visitation to d^R that satisfies the cost budget β

Problem:

$$\min_{\pi} D(d^{\pi}, d^{R}) \quad \text{s.t.} \quad \rho_C(\pi) \leq \beta$$

where $\rho_C(\pi)$ is the cost-value of the policy, β is the cost budget, and D is a metric or pseudo-metric on policy space.

Take the Lagrangian ➡ Expand the OPE term ➡ Rearrange, we get:

$$\min_{\pi} \max_{\lambda \geq 0, \nu} \min_{d} \underbrace{D(d, d^R) + \sum d(s,a)(\lambda C(s,a) + \nu) - (\lambda\beta + \nu)}_{f(d)} \; s.t.\, d(s,a) = P_*^{\pi} d(s,a)$$

We will leverage ideas from the DIstribution Correction Estimation (DICE) Offline RL framework

From DICE and Fenchel-Rockafeller duality we have:

Primal $\quad \min_{x} f(x) + g(Ax)$

Dual $\quad \max_{y} -f_*(A_* y) - g_*(y)$

Novel Constrained Projection

Assume we are given the visitation density, d^R, of a policy that maximizes return

We want to find the policy with the *closest* visitation to d^R that satisfies the cost budget β

Problem:

$$\min_{\pi} D(d^{\pi}, d^{R}) \ \text{ s.t. } \ \rho_C(\pi) \leq \beta$$

where $\rho_C(\pi)$ is the cost-value of the policy, β is the cost budget, and D is a metric or pseudo-metric on policy space.

Take the Lagrangian ➡ Expand the OPE term ➡ Rearrange, we get:

$$\min_{\pi} \max_{\lambda \geq 0, \nu} \min_{d} \underbrace{D(d, d^R) + \sum d(s,a)(\lambda C(s,a) + \nu) - (\lambda\beta + \nu)}_{f(d)} \ s.t. \underbrace{d(s,a) = P_*^{\pi} d(s,a)}_{g(d)}$$

We will leverage ideas from the DIstribution Correction Estimation (DICE) Offline RL framework

From DICE and Fenchel-Rockafeller duality we have:

Primal $\quad \min_{x} f(x) + g(Ax)$

Dual $\quad \max_{y} -f_*(A_* y) - g_*(y)$

Novel Constrained Projection

Assume we are given the visitation density, d^R, of a policy that maximizes return

We want to find the policy with the *closest* visitation to d^R that satisfies the cost budget β

Problem:

$$\min_{\pi} D(d^{\pi}, d^R) \quad \text{s.t.} \quad \rho_C(\pi) \leq \beta$$

where $\rho_C(\pi)$ is the cost-value of the policy, β is the cost budget, and D is a metric or pseudo-metric on policy space.

Take the Lagrangian ➡ Expand the OPE term ➡ Rearrange, we get:

$$\min_{\pi} \max_{\lambda \geq 0, \nu} \min_{d} \underbrace{D(d, d^R) + \sum d(s,a)(\lambda C(s,a) + \nu) - (\lambda\beta + \nu)}_{f(d)} \; s.t. \underbrace{d(s,a) = P_*^{\pi} d(s,a)}_{g(d)}$$

We will leverage ideas from the DIstribution Correction Estimation (DICE) Offline RL framework

From DICE and Fenchel-Rockafeller duality we have:

Primal $\quad \min_{x} f(x) + g(Ax)$

Dual $\quad \max_{y} -f_*(A_* y) - g_*(y)$

Pick a distance metric D and transform

Transforming allows to optimize π directly, rather than through visitation d.

Algorithm

Recall: we assumed we had d^R

Algorithm

Recall: we assumed we had d^R

If we have d^R:

Run constrained projection to get safe policy

Algorithm

Recall: we assumed we had d^R

If we have d^R:

Run constrained projection to get safe policy

If we don't have d^R:

Run offline RL w.r.t reward to find d^R, then do constrained projection

Algorithm

Recall: we assumed we had d^R

If we have d^R:

Run constrained projection to get safe policy

If we don't have d^R:

Run offline RL w.r.t reward to find d^R, then do constrained projection

Issue 2: What if the data set doesn't capture MDP dynamics?

Algorithm

Recall: we assumed we had d^R

If we have d^R:

Run constrained projection to get safe policy

If we don't have d^R:

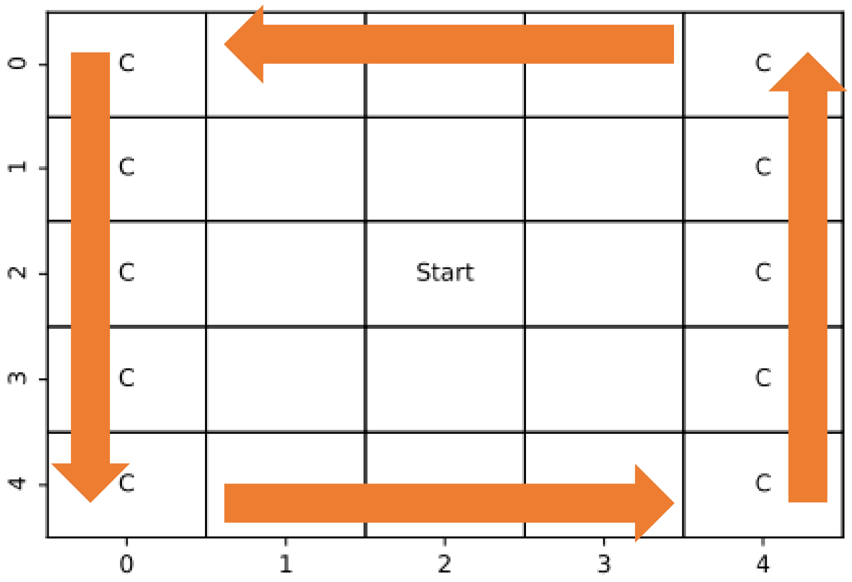
Run offline RL w.r.t reward to find d^R, then do constrained projection

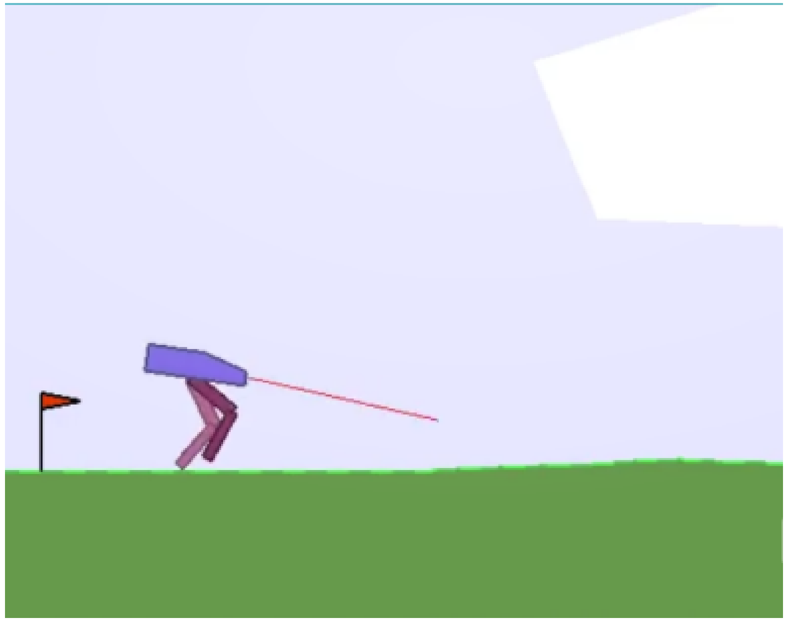
Issue 2: What if the data set doesn't capture MDP dynamics?

Answer: Novel finite sample upper confidence interval on policy cost

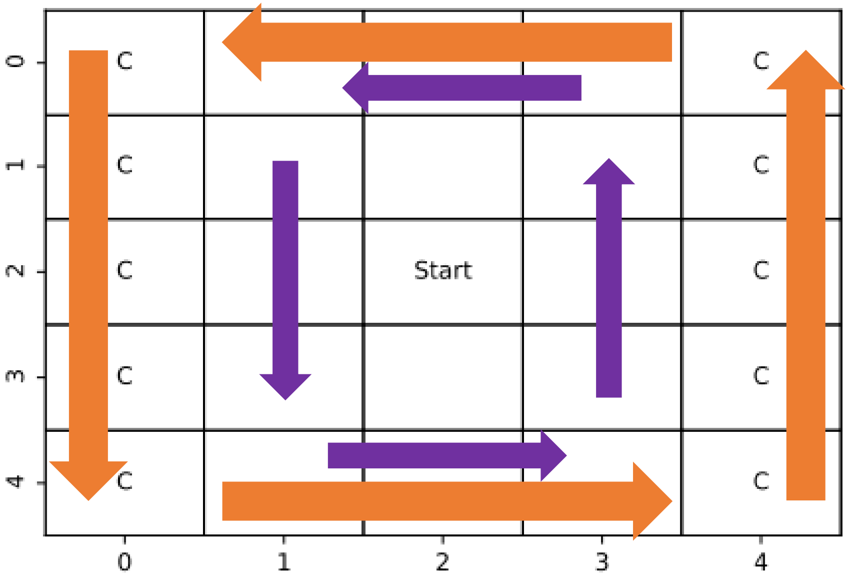
Details in the paper!

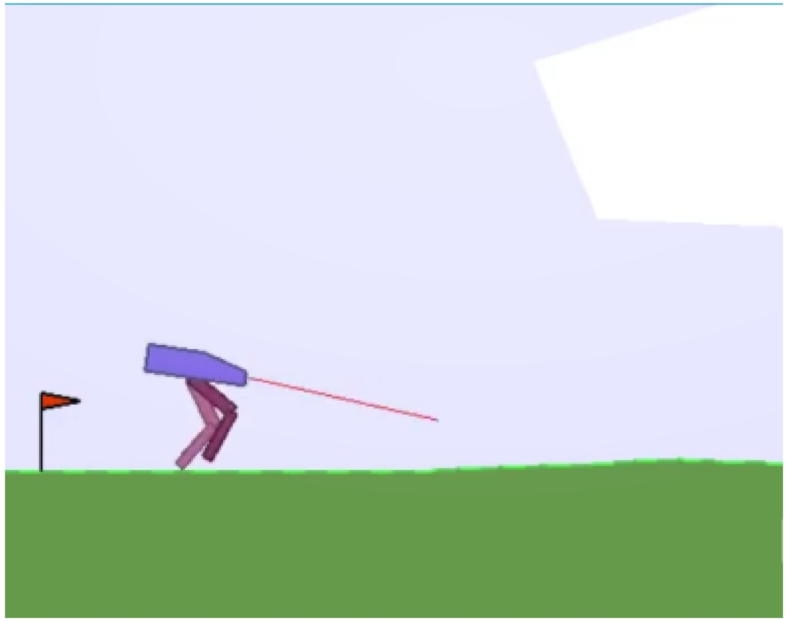
Experiments



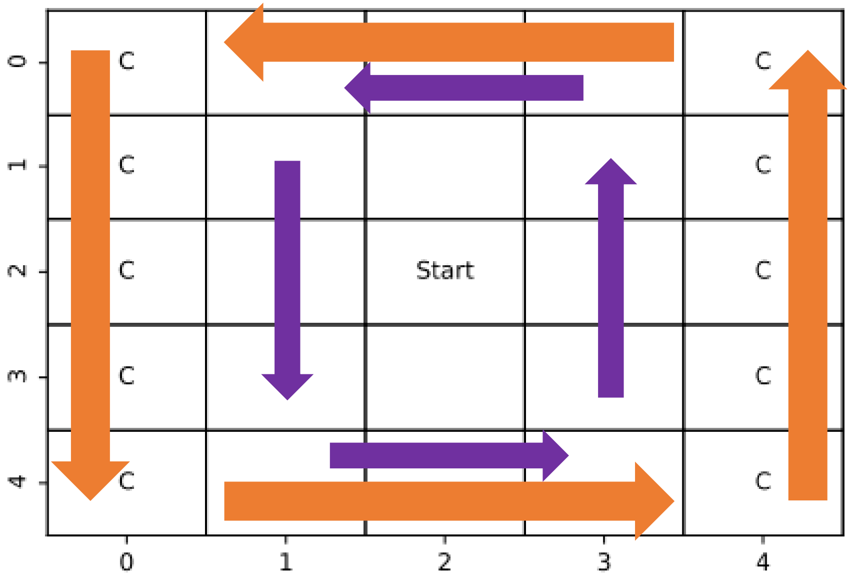


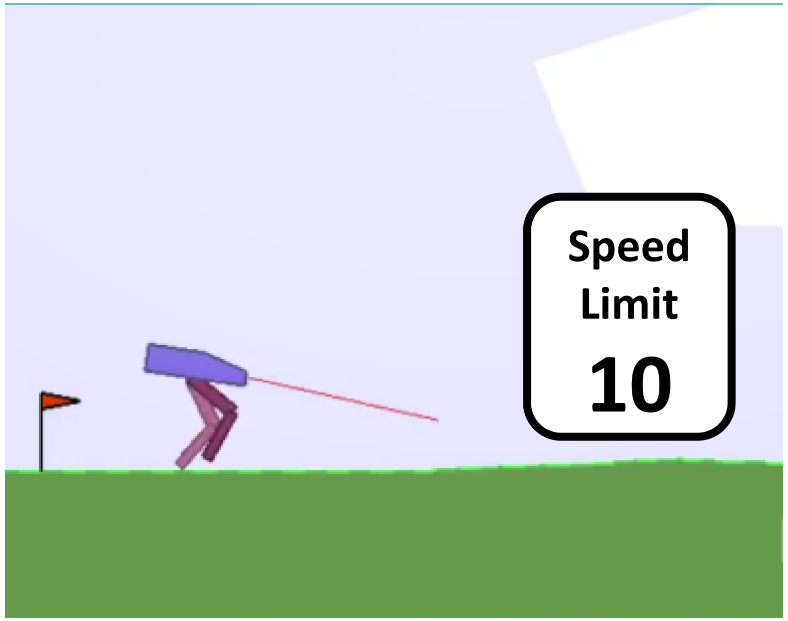
Experiments



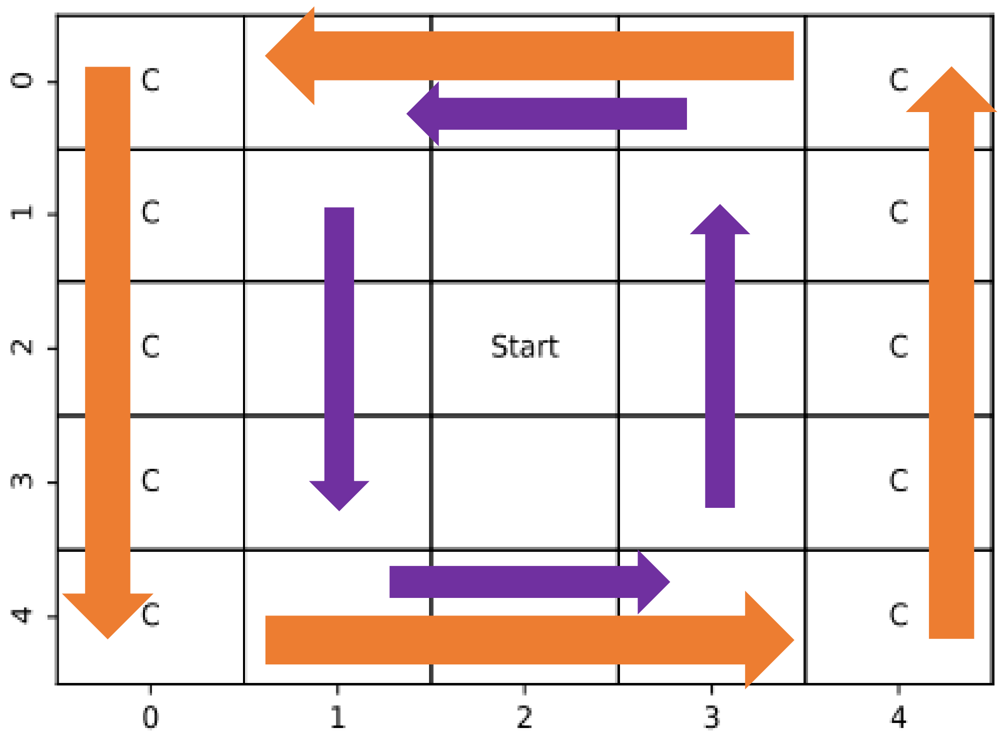


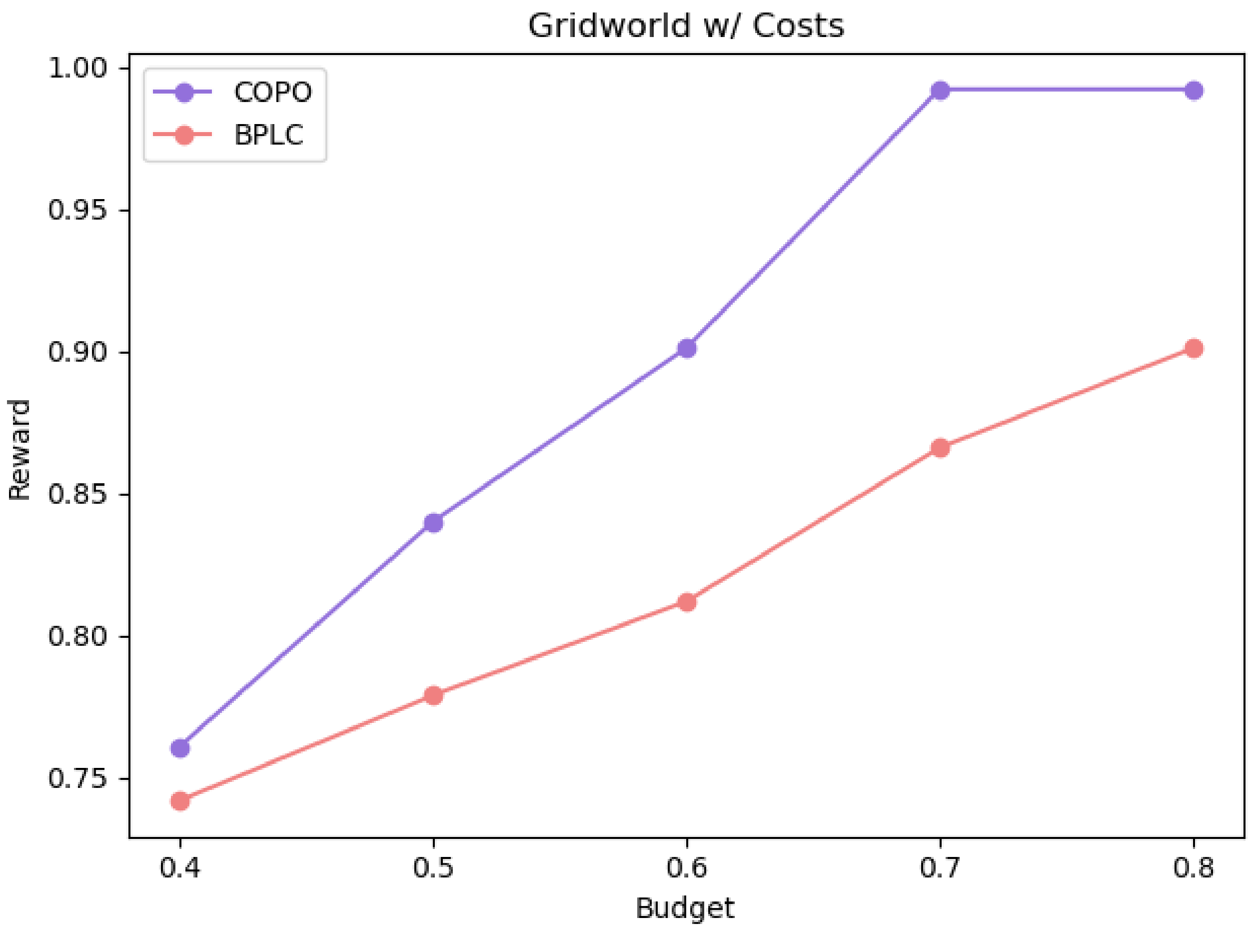
Experiments



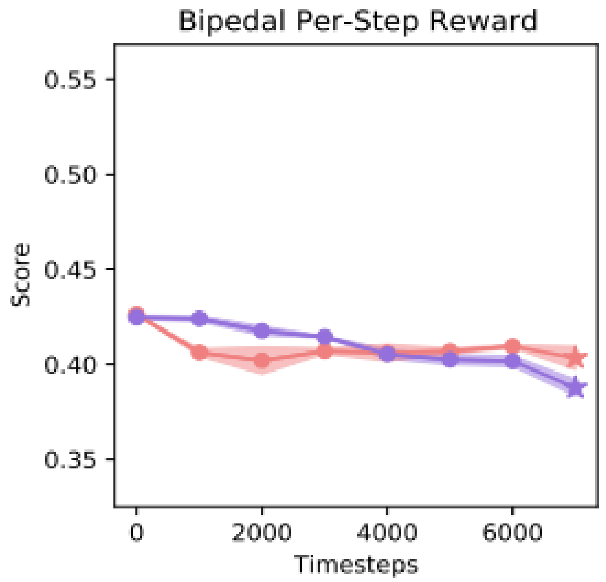


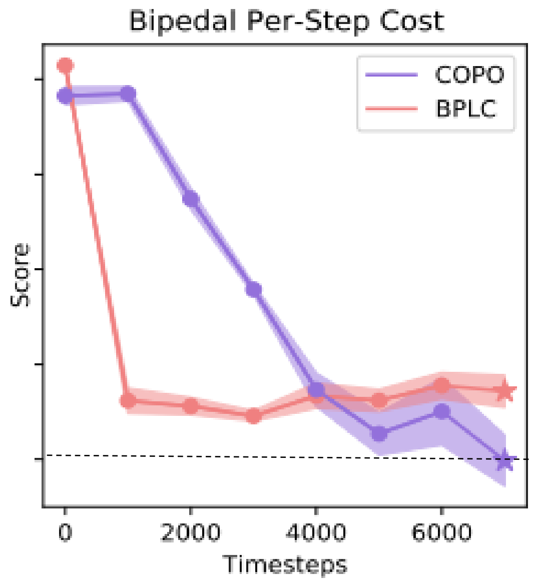
Experiments

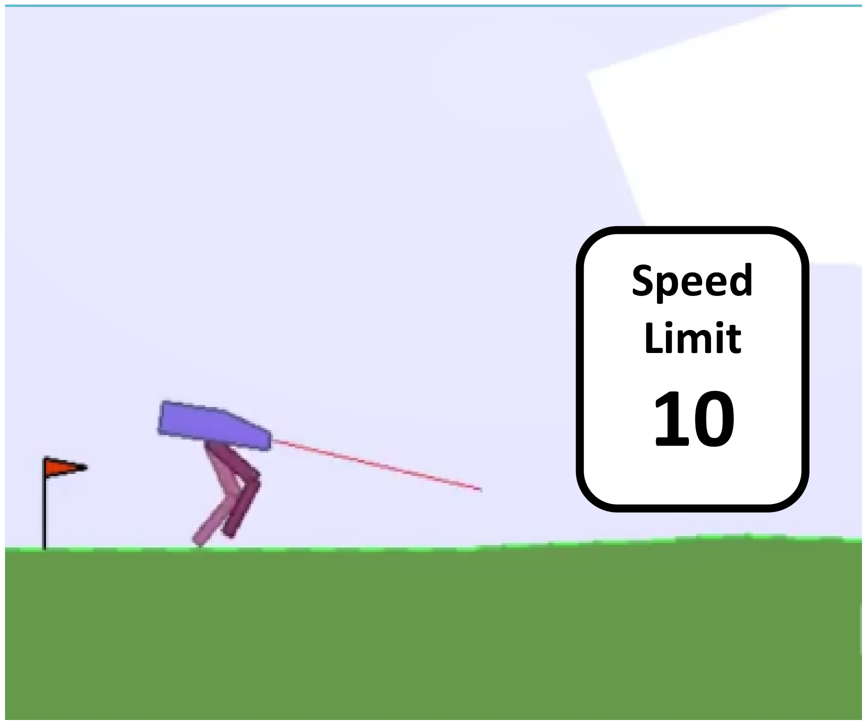




Experiments







Conclusion

COPO Offline RL Algorithm:

1. Suitable for safety-critical applications
2. Novel Constrained Projection
3. Finite Sample Confidence Interval

Conclusion

COPO Offline RL Algorithm:

1. Suitable for safety-critical applications
2. Novel Constrained Projection
3. Finite Sample Confidence Interval

COPO empirically outperforms the SOTA method BPLC

Conclusion

COPO Offline RL Algorithm:

1. Suitable for safety-critical applications
2. Novel Constrained Projection
3. Finite Sample Confidence Interval

COPO empirically outperforms the SOTA method BPLC

For more details visit **Poster #910**