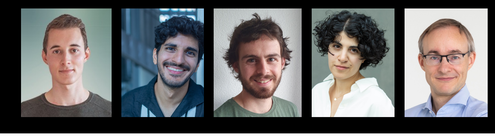
Hallucinated Adversarial Control for Conservative Offline Policy Evaluation

Jonas Rothfuss*, Bhavya Sukhija*, Tobias Birchler*, Parnian Kassraie, Andreas Krause

ETH zürich — Learning & Adaptive Systems (LAS)



Summary

How do we ensure that a policy fulfils the criteria to safely interact with an environment, without ever testing it on the environment? Our solution is HAMBO, a practical method for conservative off-policy evaluation.

- Given an **offline dataset** of transitions and an **evaluation policy**, how to **reliably estimate the performance** of the policy in **safety-critical** domains?
- HAMBO uses an **uncertainty-aware model** to estimate a tight and provably consistent **lower bound** on the evaluation **policy's performance**.
- We show that empirically HAMBO reliably gives a lower bound on the performance and analyze its behavior with varying dataset sizes and horizons for both Gaussian Process and Bayesian Neural Network models.
- Applications: Safe Offline RL, Uncertainty-aware offline RL

Problem Setting

- Continuous state and action space finite-horizon MDP

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, p_0, p, r, T)$$

$$\mathcal{A} \subseteq \mathbb{R}^{d_a} \quad \mathcal{S} \subseteq \mathbb{R}^{d_s} \quad r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

- With noisy, unknown transitions

$$s_{t+1} = f_s(s_t, a_t) + \epsilon_t \qquad \epsilon_t \sim p_\epsilon(\epsilon_t | s_t, a_t)$$

Transition distribution:

$$p(s_{t+1}|s_t, a_t) = p_\epsilon(s_{t+1} - f(s_t, a_t) | s_t, a_t)$$

- Transition data collected by behavior policy π_b:

$$\mathcal{D}_b = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$$

- Expected return of evaluation policy $\pi_e(a_t|s_t)$:

$$J(\pi_e) = \mathbb{E}_{\pi_e, p_\epsilon} \sum_{t=0}^{T} r(s_t, a_t)$$

Goal

Find largest possible lower bound $\tilde{J} \in \mathbb{R}$ that satisfies

$$\tilde{J} \leq J(\pi_e)$$

with probability at least $1 - \delta$.

The HAMBO Framework

- Use offline data $\mathcal{D}_b$ to train calibrated model $(\hat{\mu}_n, \hat{\sigma}_n)$

$$\mathbb{P}\left(\forall (s,a) : |\hat{\mu}_n(s,a) - f(s,a)| \leq \beta_n \hat{\sigma}_n(s,a)\right) \geq 1 - \delta \quad \text{(coordinate-wise)}$$

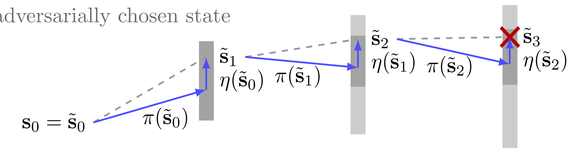
e.g.: GPs, BNNs

- Hallucinate pessimistic trajectories

$$p_\eta(s_{t+1}|s_t, a_t) \leftarrow p_\epsilon(s_{t+1} - \hat{\mu}_n(s_t, a_t) - \beta_n \eta(s_t, a_t) \hat{\sigma}_n(s_t, a_t))$$

η : hallucinated controls

$\tilde{s}_t$: adversarially chosen state



- Obtain lower bound by choosing adversarial transitions

$$\tilde{J}(\pi_e) \leftarrow \min_\eta \mathbb{E}_{p_\eta, \pi_e}\left[\sum_{t=0}^{T} r(s_t, a_t)\right]$$

→ treat η as a vector or parametrize with a network

→ optimize for η using any trajectory or policy optimizer, respectively.

Theorem (HAMBO lower-bounds the expected return)

Suppose r, f, and π_e are Lipschitz-continuous, and the model $(\hat{\mu}_n, \hat{\sigma}_n)$ is calibrated. Then HAMBO lower and upper bounds the expected risk, w.h.p.

$$\tilde{J}_n(\pi_e) \leq J(\pi_e) \leq \tilde{J}_n(\pi_e) + C_n \mathbb{E}_{(s,a)} \|\hat{\sigma}_n(s,a)\|_2$$

$$C_n = \mathcal{O}\left(\bar{L}_r T^2 \left(\bar{L}_f + \sqrt{d_s}\beta_n(\delta)\bar{L}_\sigma\right)^T\right)$$

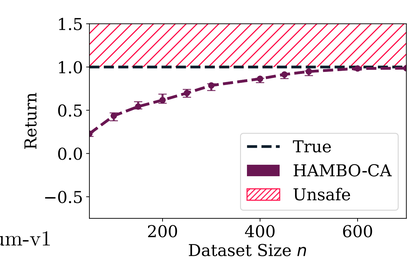
Modeling with Gaussian Processes (GPs)

Theorem (HAMBO converges to the expected return)

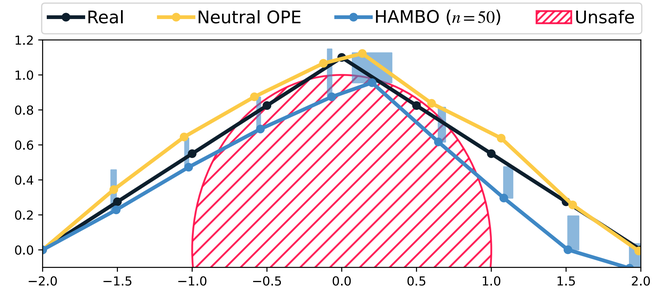
Under mild assumptions on the MDP, and if $\text{supp}(\pi_e) \subseteq \text{supp}(\pi_b)$, *HAMBO converges almost surely*

$$\tilde{J}_n(\pi_e) \xrightarrow{n \to \infty} J(\pi_e)$$

- HAMBO lower bound convergences to the true return



Env: Pendulum-v1



Modeling with Bayesian Neural Networks (BNNs)

- Using $\mathcal{D}_b$, train a BNN with ensemble of K neural networks $\{\theta_1, \ldots, \theta_K\}$ using Stein-Variational Gradient Descent (SVGD).
- Estimate BNN posterior mean and variance from ensemble:

$$\hat{\mu}_n(s,a) = \frac{1}{K}\sum_{k=1}^{K} f_{\theta_k}(s,a) \quad \hat{\sigma}_n^2(s,a) = \frac{1}{K}\sum_{k=1}^{K}\left(f_{\theta_k}(s,a) - \mu_n(s,a)\right)^2$$

→ use $\hat{\mu}_n$ and $\hat{\sigma}_n$ to hallucinate pessimistic trajectories

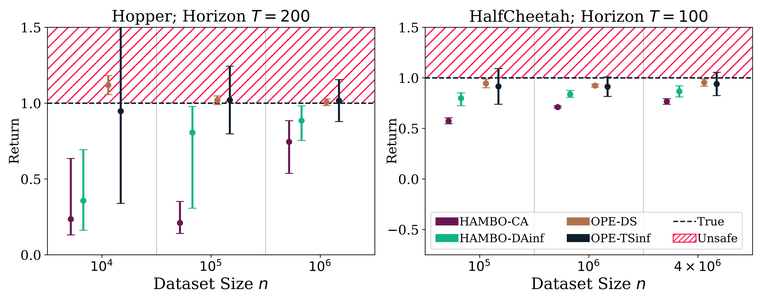
- (Even more) practical variant: HAMBO-DAINF :

→ Estimate $J(\pi_e)$ for each NN model, then pick worst

$$\tilde{J}_{\text{DAINF}}(\pi_e) = \min_{k \in \{1,\ldots,K\}} \tilde{J}_{f_{\theta_k}}(\pi_e)$$

- HAMBO estimate for varying dataset sizes:

→ Lower bound gets closer to the true return with more data.



- HAMBO estimate for increasing horizons:

→ Lower bound becomes tighter for smaller horizons.