Safe Policy Improvement by Minimizing Robust Baseline Regret

Marek Petrik, Mohammad Ghavamzadeh, Yinlam Chow
University of New Hampshire, Adobe Research, Stanford University

Contribution: Model-based policy improvement

- Often available baseline policy (e.g., deployed, a simple heuristic)
- Data-driven MDP models are imprecise
- Safe policy: Guarantee that the solution to the imprecise model is better than baseline policy
- We use robust MDPs and show why to minimize baseline regret
- Regret minimization improves on baseline policy with only few samples

General problem setting

- Discounted infinite horizon MDP: Compute π : states $\rightarrow$ actions
- Transition probability P is unknown, available limited samples of state-to-state transitions
- Return for discount factor $\gamma \in [0,1]$:

$$\mathrm{return}(\mathrm{policy}, \mathrm{model}) = \rho(\pi, P) = \mathbf{E}_{p_0}\left[\sum_{t=0}^{\infty} \gamma^t \, \mathrm{reward}_t\right]$$

- Baseline policy π_{B}: best known solution

Method 1: Solve average model (standard approach)

1. Estimate an average transition model $\bar{P}$ from samples:

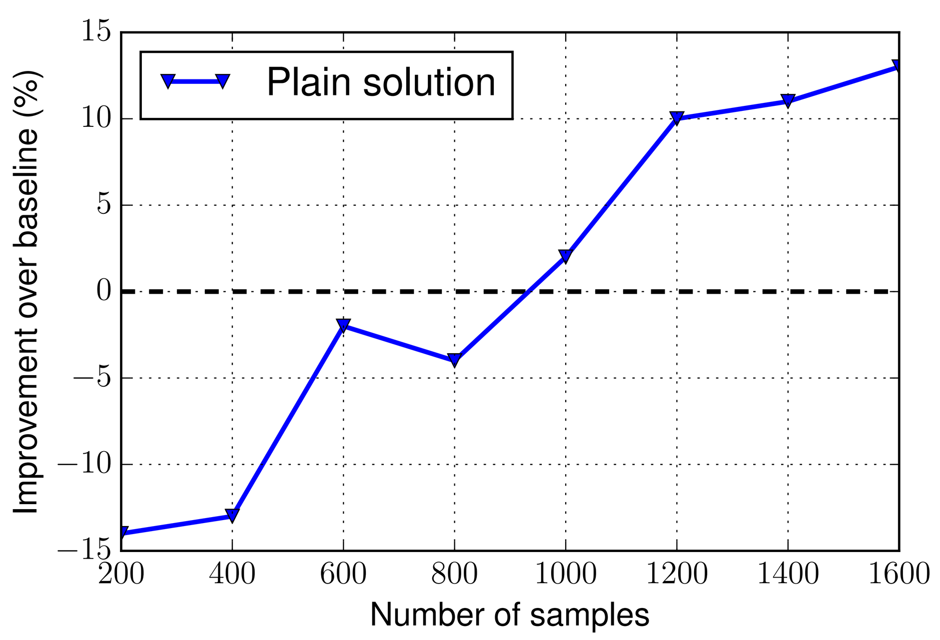
$$\bar{P}(s_1, a, s_2) = \frac{|\,\mathrm{samples\_from}(s_1, a) \cap \mathrm{samples\_to}(s_2)|}{|\,\mathrm{samples\_from}(s_1, a)|}$$

2. Solve a regular MDP with average model P

$$\pi_{\mathsf{A}} \in \arg\max_{\pi} \; \mathrm{return}(\pi, \bar{P}) \tag{1}$$

Benchmark problem results

- Simulate samples from an assumed true model
- Evaluate with respect to true model



- Optimized policy is significantly worse than baseline policy with few samples (large uncertainty)
- Decreased performance is inapparent from the solution or average model alone

Method 2: Simple robust solution

- Represent uncertainty due to limited samples:

$$\bar{P}(s_1, a, s_2) = \frac{|\,\mathrm{samples\_from}(s_1, a) \cap \mathrm{samples\_to}(s_2)|}{|\,\mathrm{samples\_from}(s_1, a)|} \pm \sqrt{\frac{\mathrm{constants}}{|\,\mathrm{samples\_from}(s_1, a)|}}$$

- Construct set of plausible transition probabilities: (e.g. concentration inequalities)

$$\mathcal{P} = \left\{ P \; : \; \|\bar{P}(s, a, \cdot) - P(s, a, \cdot)\|_1 \le e(s, a) \right\} \quad e(s, a) \sim \sqrt{\frac{\mathrm{constants}}{|\,\mathrm{samples\_from}(s_1, a)|}}$$

- Solve for a robust solution (lower bound):

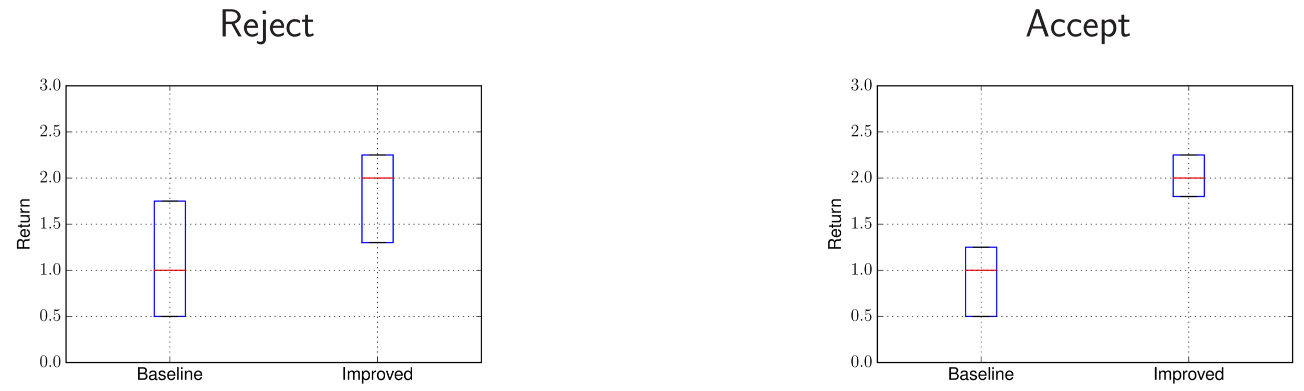
$$\pi_{\mathsf{R}} \leftarrow \arg\max_{\pi} \min_{P \in \mathcal{P}} \mathrm{return}(\pi, P) \tag{2}$$

- Accept only if surely outperforms baseline policy π_{B}:

$$\min_{P \in \mathcal{P}} \mathrm{return}(\pi_{\mathsf{R}}, P) \ge \max_{P \in \mathcal{P}} \mathrm{return}(\pi_{\mathsf{B}}, P)$$

Accepting the robust solution

- Accept the robust solution only if it is guaranteed to be better than the baseline policy
- Otherwise, use the baseline policy

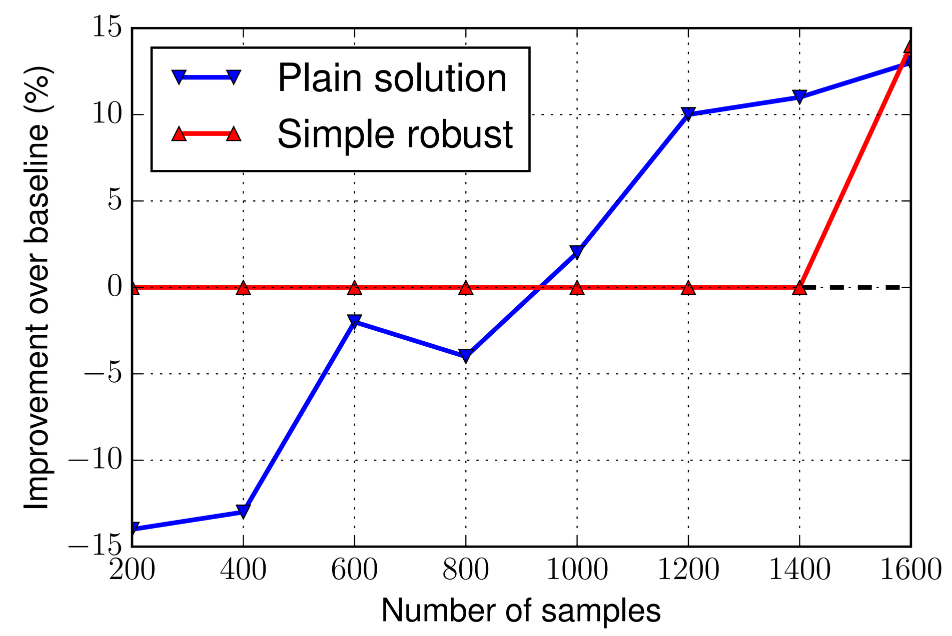


Solution using robust MDPs

- Problem (2) is non-convex: Tractably solve using robust Markov decision process (a game with nature)
- Rectangular uncertainty sets (independent uncertainty sets between states and actions)
- Similar properties as regular MDPs (Markov policies optimal), easy to solve
- Robust Bellman optimality:

$$v^\star(s) = \max_{a} \min_{P \in \mathcal{P}} \left(\mathrm{reward}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{S}} P(s, a, s') \, v^\star(s') \right)$$

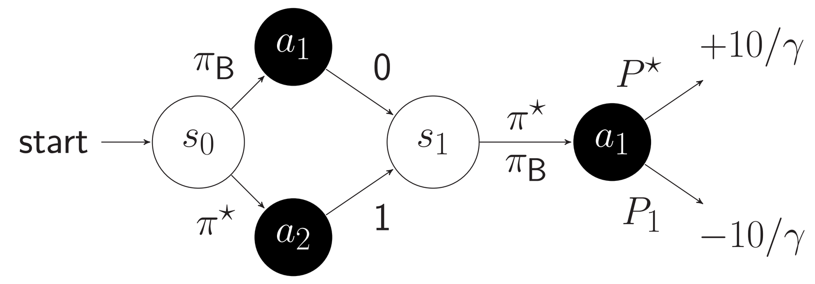
Benchmark problem results



- Guarantees solution is no worse than baseline
- All or nothing behavior even when some states are better known than others
- Can we better leverage the model to get improvement with few samples?

How to do better with a model

- State s_0 transition probabilities are certain
- State s_1 transition probabilities are uncertain

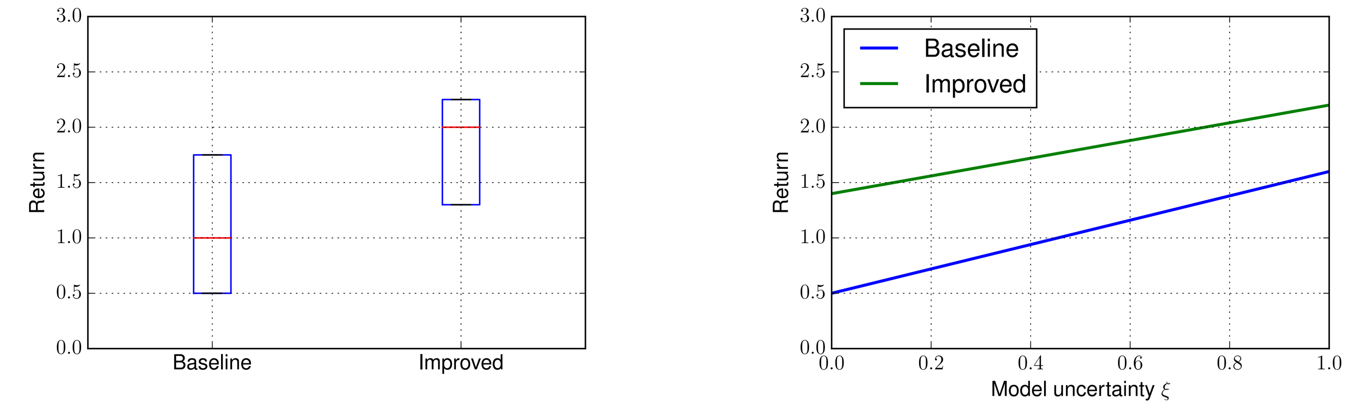


- Policy $\pi^\star$ always better than baseline π_{B}
- Method 2 does not improve on baseline in this example:

$$\min_{P \in \mathcal{P}} \mathrm{return}(\pi^\star, P) = -9 \qquad \max_{P \in \mathcal{P}} \mathrm{return}(\pi_{\mathsf{B}}, P) = +10$$

Method 3: Robust baseline regret (new approach)

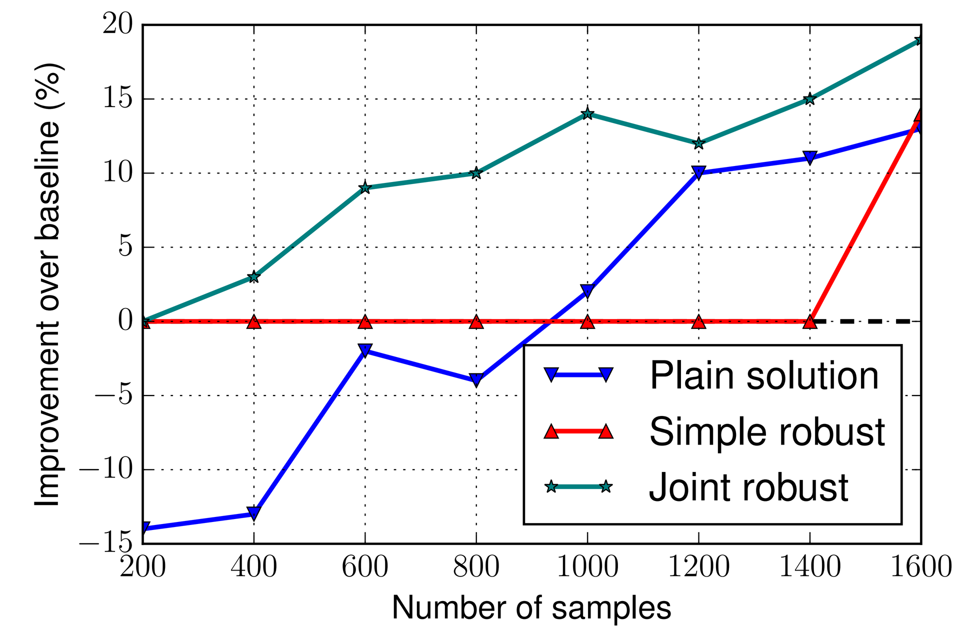
- Be more precise about the impact of model uncertainty on both improved and baseline policies
- Considering confidence intervals alone is insufficient, must consider the response of return with respect to model uncertainty ($\xi = P$)



- Minimize robust baseline regret:

$$\pi_{\mathsf{S}} = \arg\max_{\pi} \min_{P \in \mathcal{P}} \left(\mathrm{return}(\pi, P) - \mathrm{return}(\pi_{\mathsf{B}}, P) \right) \tag{3}$$

Benchmark problem results



- Improves baseline solution with very few samples

Guarantees on performance loss

- Performance loss: $\mathrm{loss}(\pi) = \mathrm{return}(\pi^\star, P^\star) - \mathrm{return}(\pi, P^\star)$; $P^\star$ is the *true unknown* model
- Method 1: Solve average model: π_{A}

$$\mathrm{loss}(\pi_{\mathsf{A}}) \le \frac{2\gamma}{(1-\gamma)^2} \max_{\pi} \left(\|r_\pi\|_\infty \|e_\pi\|_\infty \right).$$

- Method 2,3: Robust solution: π_{S}

$$\mathrm{loss}(\pi_{\mathsf{S}}) \le \min \left\{ \frac{2\gamma}{(1-\gamma)^2} \|r_{\pi^\star}\|_\infty \|e_{\pi^\star}\|_{1,u^\star}, \; \mathrm{loss}(\pi_{\mathsf{B}}) \right\}$$

Other notable results (see the paper)

1. Showed that it is NP hard to solve (3) (via an SAT reduction)
2. Proposed a simple approximate algorithm for solving (3)
3. Optimal policy in (3) may be randomized; arbitrarily better than the best deterministic policy
4. Case study using a realistic energy storage and arbitrage problem

Related work

- Most approaches based on model-free methods
- Off-policy learning and optimization (Perkins2002a; Thomas2015; Hallak2015)
- Robust/safe policy improvement (Pirotta2013)
- Conservative policy iteration (Kakade2002)
- Policy improvement with high confidence (Thomas2015a)
- Robust MDPs (Iyengar2005; Wiesemann2013)

NIPS 2016, Barcelona, Spain