Efficient Model-Based Reinforcement Learning
Through Optimistic Policy Search and Planning

tl;dr: H-UCRL, a practical algorithm for efficient optimistic exploration in deep RL

Problem Setting

Definition (Performance of policy $\pi$ on model $f$):
$$ J(\pi, f) = \mathbb{E}_{\mathbf{s}_0 \sim \rho_{\pi}} \left[ \sum_{t=0}^{N} r(\mathbf{s}_t, \pi(\mathbf{s}_t)) \right] $$

Objective (Maximize performance on true system $f^*$):
$$ \pi^* = \arg\max_{\pi} J(f^*, \pi) $$

Data Collection:

Exploration vs. Exploitation

We want to execute a policy that:

- Exploits the model to maximize the performance.
- Explores the environment to reduce the epistemic uncertainty.

<table>
<thead>
<tr>
<th>Name</th>
<th>Algorithm</th>
<th>Exploration</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy</td>
<td>$\max_{\pi} J(f^*, \pi)$</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>UCRL</td>
<td>$\max_{\pi} J(f^*, \pi)$</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>PSRL</td>
<td>$\max_{\pi} J(\pi, f^*)$</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

- Greedy exploitation does not explore enough. Many practical algorithms solve such problem (e.g., PILCO (Deisenroth & Rasmussen, 2011) and PETS (Chua et al., 2018)).
- UCRL (Jaksch et al. 2010) instantiates the Optimism in the Face of Uncertainty principle and explores efficiently. The optimization over policies and models is intractable.
- PSRL (Osband et al. 2013) instantiates OFU through stochasticity but requires samples from an exact posterior. When the model is sampled using approximate methods, the exploration guarantees are lost.

H-UCRL: A reduction from UCRL to greedy

- We reparameterize the set of plausible models $M_\theta$ introducing an auxiliary function $\eta$.
- H-UCRL is optimistic by hallucinating control over $\eta$.
- The resulting planning problem is a tractable greedy exploitation problem.

$$ \text{UCRL} = \max_{\theta} \max_{\pi} J(\pi; f) \quad \text{s.t.} \quad |f - \mu_\pi| \leq \beta |\sigma_\pi| $$

$$ \text{H-UCRL} = \max_{\theta} \max_{\pi} J(\pi; f) \quad \text{s.t.} \quad \mu_\pi + \beta |\sigma_\pi|_\eta : \mathcal{S} \rightarrow [-1, 1] $$

$$ \text{Greedy (augmented policies, structured model)} $$

H-UCRL plans by:

- Using the true policy to select the next-state distribution, alike to Greedy.
- Hallucinating control to select the next state optimistically from within the conditional next-state distribution.

H-UCRL scales to deep NN because it does not require calibrated multi-step ahead predictions, but only calibrated one-step ahead predictions.

Theoretical Results (Cumulative Regret)

$$ \text{Regret}_T = \sum_{t=1}^{T} J(f^*, \pi_T) - J(f^*, \pi^*) = O(K^2 \sqrt{T N^3 T_f}) $$

The regret is sublinear when the maximum information gain is also sublinear. This quantity depends on the model class we are trying to learn. For some GP kernels, this is sublinear. The regret also depends exponentially on the horizon. This is the price we pay for only requiring calibrated one-step ahead predictions.

Experimental Results

- H-UCRL outperforms Greedy and PSRL in hard exploration problems.
- H-UCRL also outperforms Greedy and PSRL in terms of learning speed.

References