Exploring-PILCO
Data-Efficient Reinforcement Learning using PILCO and Directed-Exploration
Rowan McAllister, Mark van der Wilk, Carl Rasmussen {rtm26, mv310, cer54}@cam.ac.uk

PILCO achieved unprecedented data-efficiency

PILCO Algorithm:
1. Execute random actions for initial data.
2. for episode $e = 1$ to $E$ do
3. Learn dynamics model $f$.
4. Predict system trajectories $p(x_t)$ to $p(X_{T'})$.
5. Evaluate policy: $J(\pi) = E_p[\mathcal{L}]$, $\mathcal{L} = \sum_{t \in \tau}[\text{cost}(x_t)|\pi]$.
6. Optimise policy: $\pi \leftarrow \arg\min\{J(\pi)\}$.
7. Execute system, record data.
8. end for

- Evaluating loss requires simulation for policy $\pi$ under current belief of dynamics.
- Gaussian approximation is possible with a GP model and exact moment matching.
- Loss is smooth and differentiable $\Rightarrow$ optimise policy with gradients.

PILCO discovers by accident

Exploration is not explicitly encouraged in the loss function!

Bayes-optimal solution is complex:
- Requires deep lookahead tree with continuous transition dynamics and actions.
- Information gained from a policy informs all future policies (losses are correlated).
- Evaluating the expected loss and its variance is already complex.

Can we improve PILCO by treating policy selection as Bayesian optimisation
as long as we have good estimates of the loss uncertainty?

PILCO with Directed Exploration

We test three Bayesian optimisation methods BO($\mu, \sigma$), where $\mathcal{L} \sim \mathcal{N}(\mu, \sigma)$:
1. Probability of improvement: $PI(\mu, \sigma) = -Pr(\mathcal{L} < \mathcal{L}^* \equiv \Phi(\frac{\mathcal{L}^* - \mathcal{L}}{\sigma}))$.
2. Expected improvement: $EI(\mu, \sigma) = E_p[\min(\mathcal{L} - \mathcal{L}^*, 0)]$.
3. Gittins index: $GI(\mu, \sigma) = \lambda : \frac{1}{\lambda} = E_p\left[\mathcal{L} + \lambda \cdot \min(\mathcal{L}, \lambda)\right]$.

Exploring using Total Loss Uncertainty

<table>
<thead>
<tr>
<th>$\mathcal{L}$</th>
<th>$\mu$</th>
<th>$PI$</th>
<th>$EI$</th>
<th>$GI$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Empirical}\mathcal{L}$</td>
<td>0.759</td>
<td>0.947</td>
<td>0.905</td>
<td>1.15</td>
</tr>
</tbody>
</table>

Exploring using Reducible Loss Uncertainty

Total loss uncertainty is not the right metric
Which bandit-type would you play next with $N$ plays remaining?

- Deterministic returns $r_1 = \$1$.
- Aleatorically uncertain returns $r_2 \sim \mathcal{N}(\mu, 1)$, $h_{\text{alea}}(\pi) = \mathcal{N}(0, 0)$.
- Epistemically uncertain returns $r_3 \sim \mathcal{N}(\mu, 0)$, $h_{\text{epis}}(\pi) = \mathcal{N}(0, 1)$.

We only care about uncertainty that can be reduced!

Variance reduction in Gaussian processes

For a tractable analytic approximate estimate we assume:
1. observation of $y$ in $D$ only reduces the variance for $x_{t+1}$ in the next simulation,
2. costs for consecutive times are independent, and
3. the locations of the observations are at the mean of $x$.

- Reduces problem to expected variance of a single output given a random input, with the expectation over a hypothetical observation.
- Since hypothetical observation is uncertain, we marginalise it out.

$\mathbb{E}_y[V_{x, t} | y|x, y] = \mathbb{E}_y \left[\mathbb{E}_y[V_{x, t} | y|x, y] + V_{x} | \mathbb{E}_y[y|x, y]\right]$ (1)

Issues with the current approximation:
- Variance reduction is currently swamped by initial variance
- Propagation of reduced uncertainty seems important
- Variance of the fantasy output currently disregards variance of the input

Discussion

- There is evidence that PILCO can be improved further with exploration strategies
- Gaussian processes allow for sophisticated analytic approximations
- More work is needed in GP approximations
- What is the benefit of analytic approximations vs sampling?