Bayesian Policy Search with Probabilistic Programs

arXiv:2010.00284

David Tolpin
Yuan Zhou
Hongseok Yang

Motivation

- Policy search ≡ inference (via stochastic control)
- Wingate et al. 2011: Bayesian policy search in deterministic domains
- van de Meent et al. 2016: Policy optimization in stochastic domains
- **Bayesian** policy search in **stochastic** domains?

1. David Wingate, Noah D Goodman, Daniel M Roy, Leslie P Kaelbling, and Joshua B Tenenbaum. Bayesian policy search with policy priors. IJCAI 2011
2. Jan-Willem van de Meent, Brooks Paige, David Tolpin, and Frank Wood. Black-box policy search with probabilistic programs. AISTATS 2016

Stochastic Metropolis–Hastings

$$p_{common}(\mathbf{1}|\mathbb{E}(r|\theta)) = \exp(\mathbb{E}(r|\theta) - U_r) = \prod_r (\exp(r - U_r))^{p_S(r|\theta)} = \prod_r p_{common}(\mathbf{1}|r)^{p_S(r|\theta)}$$

$$p_{our}(\mathbf{1}|\mathbb{E}(r|\theta)) = \frac{\mathbb{E}(r|\theta) - L_r}{U_r - L_r} = \sum_r p_S(r|\theta)\left(\frac{r - L_r}{U_r - L_r}\right) = \sum_r p_S(r|\theta) p_{our}(\mathbf{1}|r)$$ – mixture!

Model

$\tau \sim D_\tau$ — simulator trace

$$\theta \sim D_\theta$$

$$\mathbf{1} \sim \mathrm{Bernoulli}\left(\frac{\mathbb{E}_\tau[\mathcal{P}(\theta, \tau)] - L_r}{U_r - L_r}\right)$$

Algorithm

```
loop
    τ ~ D_τ   (* always accept *)
    θ' ~ D_θ
    u ~ Uniform(0, 1)
    if {u < min(1, (P(θ',τ) − L_r)/(P(θ,τ) − L_r))} then
        θ ← θ'
    end if
    Output θ
end loop
```

Probabilistic programs for policy search

- $S(\theta)$ — simulator, returns reward r
- θ — parameters of interest

Common conditioning

$$p_{common}(\mathbf{1}|r) = \exp(r - U_r)$$

U_r — an upper bound on r

Generative model

$$\theta \sim D_\theta$$
$$\mathbf{1} \sim \mathrm{Bernoulli}(\exp(\mathbb{E}[\mathcal{S}(\theta)] - U_r))$$

Our conditioning

$$p_{our}(\mathbf{1}|r) = \frac{r - L_r}{U_r - L_r}$$

U_r, L_r — upper and lower bounds on r

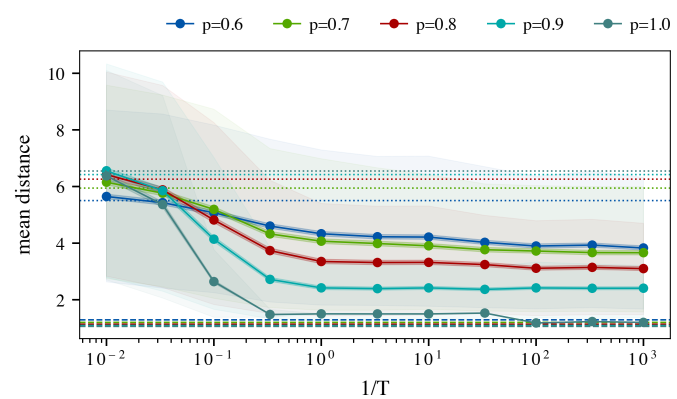
Generative model

$$\theta \sim D_\theta$$
$$\mathbf{1} \sim \mathrm{Bernoulli}\left(\frac{\mathbb{E}[\mathcal{S}(\theta)] - L_r}{U_r - L_r}\right)$$

Allows flattening of nested inference!

Case studies

Canadian traveller



RockSample