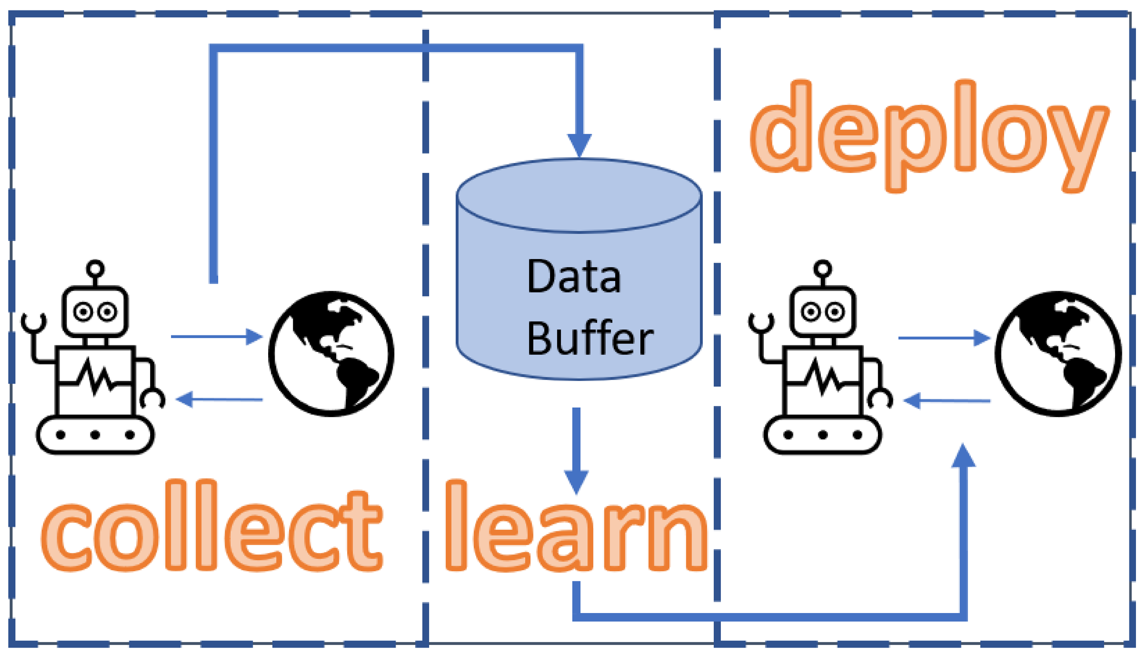
Value Regularization Using Model Uncertainty in Offline Reinforcement Learning

Alexie Pogue

Goal: Develop methods in offline RL that are robust to errors in model uncertainty estimation

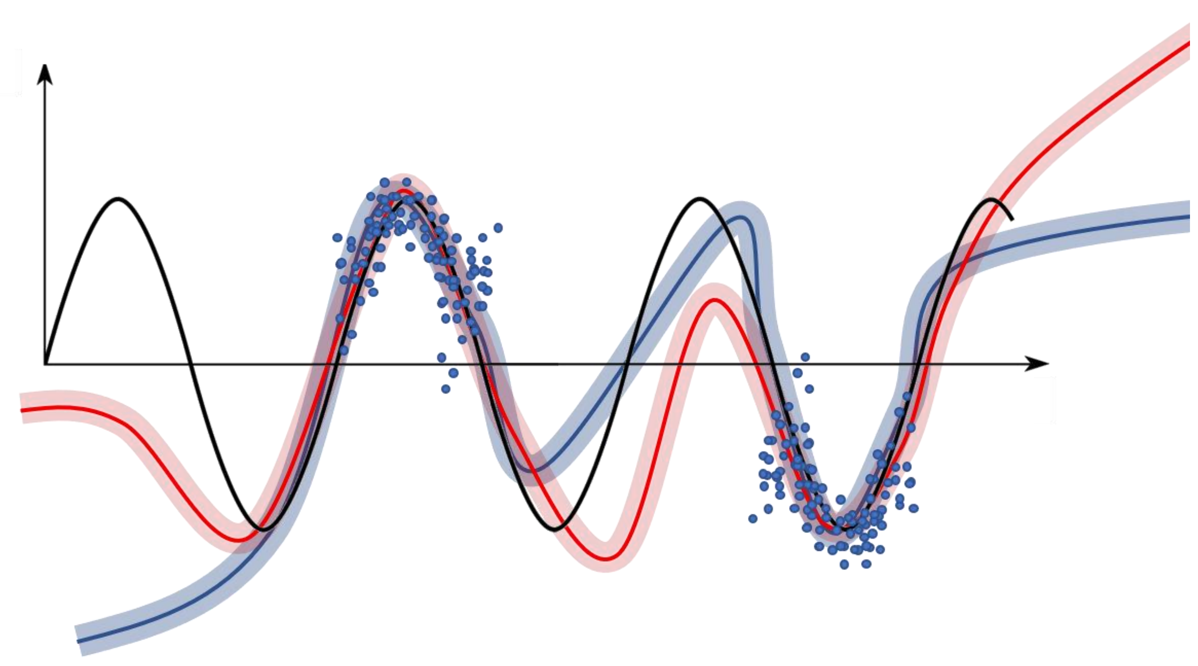


Summary

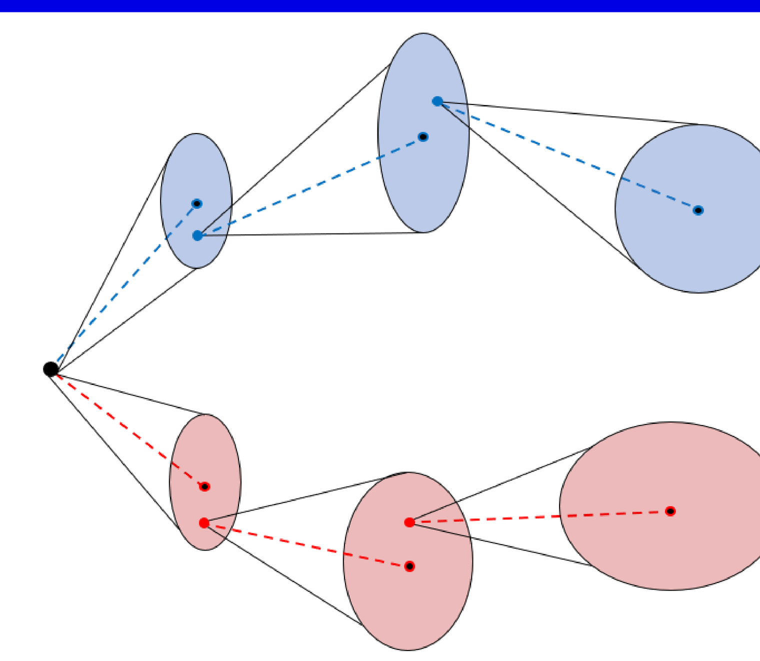
- Model-based approaches to offline reinforcement are prone to error due to incomplete supervision over state-action tuples
- Investigate outcomes when least aggressive error measures and reward penalties plus generous regularization towards the batch policy are used; including cases when the error estimator is inaccurate.

Model Uncertainty

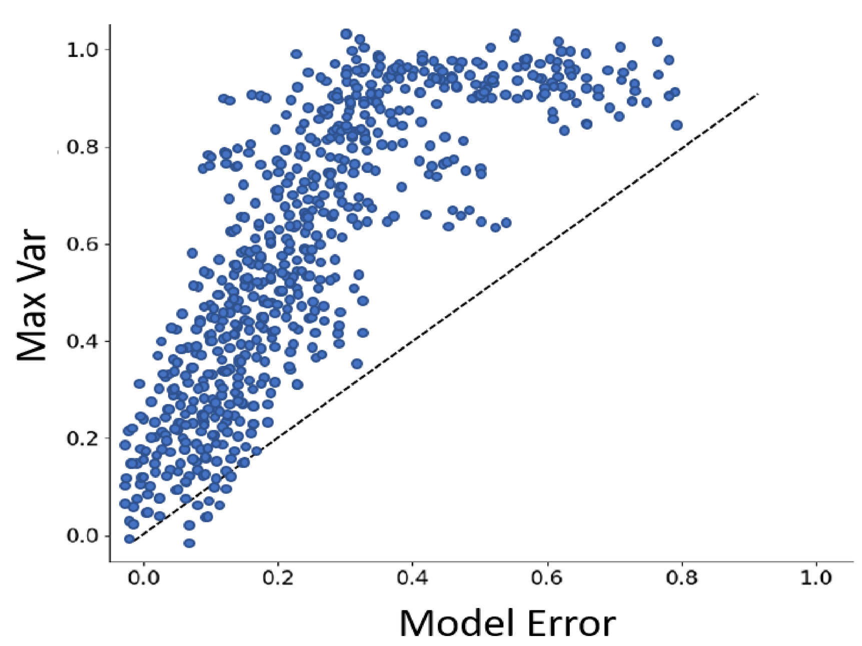
- Probabilistic neural networks estimate aleatoric uncertainty
- Bootstrap ensembles of networks estimate epistemic uncertainty
- Estimate the upper bound on sampling error



Models are accurate close to data samples



Ensemble agreement alludes to correct model generalization



Narrow distribution dataset

- Maximum variance is overly conservative, ensemble variance can create bias leading to model exploitation
- High error when the true error is zero motivates analysis when the state-action tuples are within the batch distribution, is behavior cloning achievable?
- When the error is high, encouragement towards the support of the batch policy via increased rewards leads to value regularization

Uncertainty-based Value Regularization

$$\hat{Q}^{k+1} \leftarrow \arg\min_{Q} \beta(\mathbb{E}_{s,a\sim\rho(s,a)}[Q(s,a)] - \mathbb{E}_{s,a\sim\mathcal{D}}[Q(s,a)]) + \frac{1}{2}\mathbb{E}_{s,a,s'\sim h}\left[(Q(s,a) - \hat{\mathcal{B}}^{\pi}\hat{Q}^{k}(s,a))^2\right]$$

$$\hat{Q}^{k+1}(s,a) = \hat{\mathcal{B}}^{\pi}\hat{Q}^{k}(s,a) - \beta\frac{\rho(s,a)-d(s,a)}{h(s,a)}$$

f-interp mix of distributions: Dyna (Sutton, 1991)

$$\hat{Q}^{k+1} = \hat{\mathcal{B}}^{\pi}_{\mathcal{M}}\hat{Q}^{k}(s,a) - \beta\frac{\rho(s,a)-d(s,a)}{h(s,a)} + (1-f)\left[\hat{\mathcal{B}}^{\pi}_{\hat{\mathcal{M}}}\hat{Q}^{k}(s,a) - \hat{\mathcal{B}}^{\pi}_{\mathcal{M}}\hat{Q}^{k}(s,a)\right]$$

Error in Bellman backup

$$\leq \hat{\mathcal{B}}^{\pi}_{\mathcal{M}}\hat{Q}^{k}(s,a) - \underbrace{\beta\frac{\rho(s,a)-d(s,a)}{h(s,a)}} + \underbrace{(1-f)\left|\hat{\mathcal{B}}^{\pi}_{\hat{\mathcal{M}}}\hat{Q}^{k}(s,a) - \hat{\mathcal{B}}^{\pi}_{\mathcal{M}}\hat{Q}^{k}(s,a)\right|}$$

Greater than zero under expectation

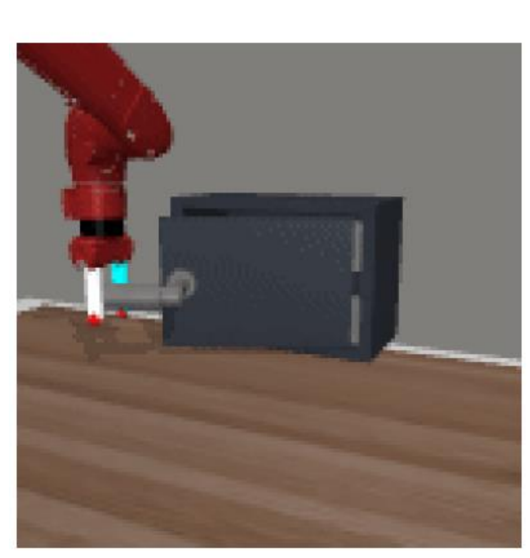
$$\left|r_{\hat{\mathcal{M}}}(s,a) - r_{\mathcal{M}}(s,a)\right| + \gamma\frac{R_{\max}}{1-\gamma}D_{\mathrm{TV}}(T_{\hat{\mathcal{M}}}, T_{\mathcal{M}})$$

Bound on Bellman backup error

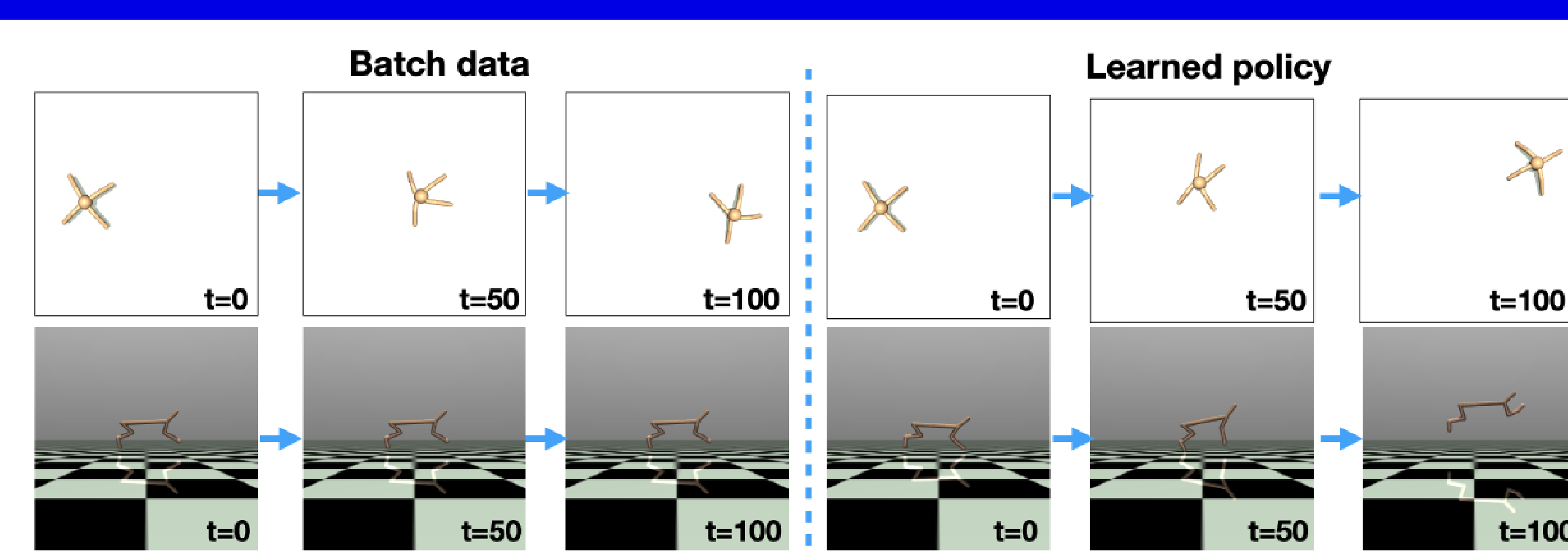
Solve β such that the sum of extra terms are negative

Validation and Further Study

- Validation within narrow data domains
- Analysis via OpenAI Gym subsets of the D4RL benchmark (Brockman et al. 2016)
- Test outcomes on datasets requiring generalization to different tasks
- Determine outcomes on high-dimensional observations such as vision experiments



Sawyer vision door close



Ant and cheetah run forward → Change direction, jump

Yu et al. 2021