DeepMind

POMRL: No-Regret Learning-to-Plan with Increasing Horizons

Khimya Khetarpal*, Claire Vernade*, Brendan O' Donoghue, Satinder Singh & Tom Zahavy

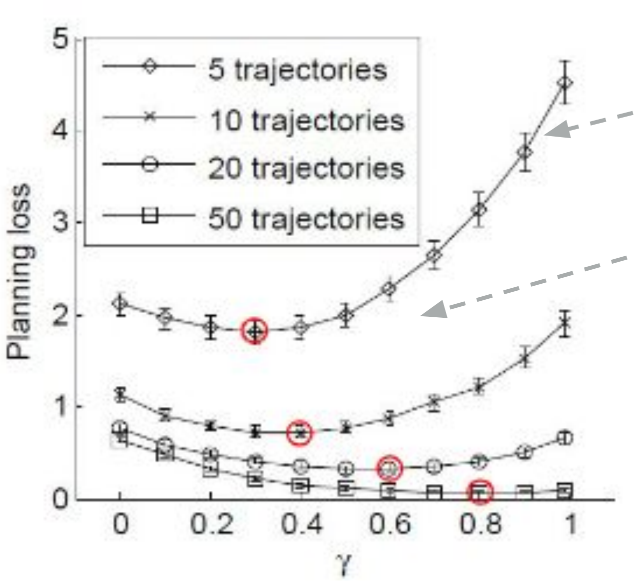




GenPlan Workshop, 2023
@ NeurIPS

Motivation - *Choice of planning horizon*

- ➢ A key component in the lifetime of an RL agent is the planning horizon $H = \frac{1}{1-\gamma}$
- ➢ The choice of the planning horizon plays an important role

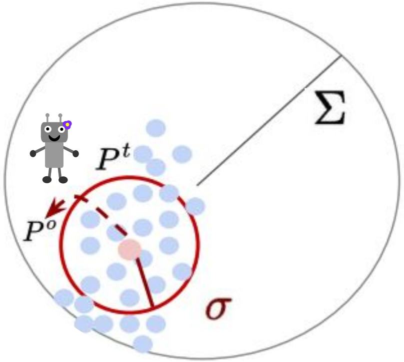


Jiang et al. 2015

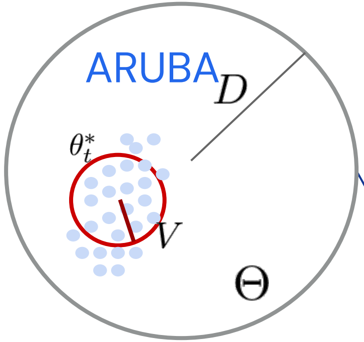
In the low-data regime, it might be optimal to not plan with the true model (i.e. gamma eval)

One could instead learn and adapt the planning horizon accordingly.

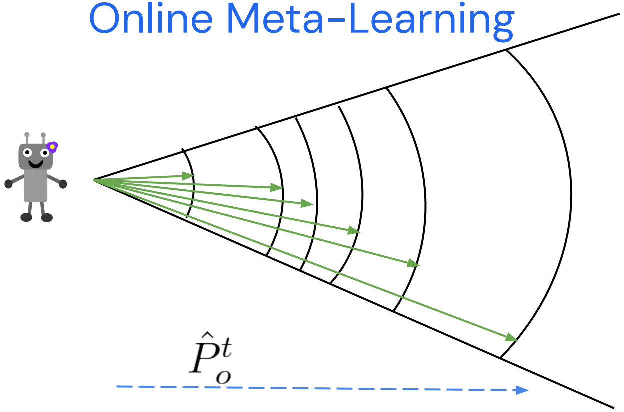
The Bigger Picture - *Problem Setting and Overview*



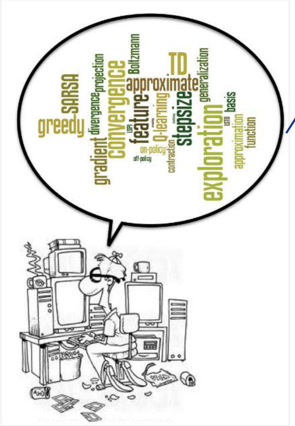
Online Meta-Learning



Khodak et al. 2019



Growing Planning Horizon



Dong et al. 2021

Interplay Between Planning Horizon & Meta-Reinforcement Learning

Research Question

There is a direct correlation between the knowledge acquired by the agent and the effective planning horizon: the **more knowledgeable the agent, the longer its planning horizon**.

Research Question

Can we meta-learn a good initialization of the model across tasks and adapt the effective planning horizon better?

Planning with Online Meta-learning: *Our Approach*

for *task* $t \in [T]$ **do**

for t^{th} *batch of* m *samples* **do**

$\hat{P}^t(m) = (1-\alpha_t)\frac{1}{m}\sum_{i=1}^{m} X_i + \alpha_t \hat{P}^{o,t}$ // regularized least squares minimizer.

$\gamma^\star \leftarrow \gamma$-Selection-Procedure$(m, \alpha_t, \sigma_t, T, S, A)$

$\pi^\star_{\hat{P}^t,\gamma} \leftarrow$ Planning$(\hat{P}^t(m))$ // $\forall \gamma \leq \gamma_{\texttt{eval}}$

Output: $\pi^\star_{\hat{P}^t,\gamma}$

A batch within-task RLS Loss

Update $\hat{P}^{o,t+1}$, $\hat{\sigma}_{t+1} \leftarrow$ Welford's online algorithm$\left((\hat{\sigma_o})_t, \hat{P}^{o,t+1}, \hat{P}^{o,t}\right)$ // meta-update AoM (Eq. 5) and task-similarity parameter.

Update $\alpha_{t+1} = \frac{1}{\hat{\sigma}_{t+1}{}^2(1+1/t)m+1}$ // meta-update mixing rate, plug $\max(\sigma_{S\times A})$

Meta-learn the task similarity and a universal dynamics model

Planning with Online Meta-learning: *Theory Result*

- After T tasks, the agent is evaluated via the average planning loss

$$\bar{\mathcal{L}} = \frac{1}{T}\sum_{t=1}^{T} \left\| V_{P^t,\gamma_{\mathtt{eval}}}^{\pi^*_{P^t,\gamma_{\mathtt{eval}}}} - V_{P^t,\gamma_{\mathtt{eval}}}^{\pi^*_{\hat{P}^t,\gamma}} \right\|_\infty$$

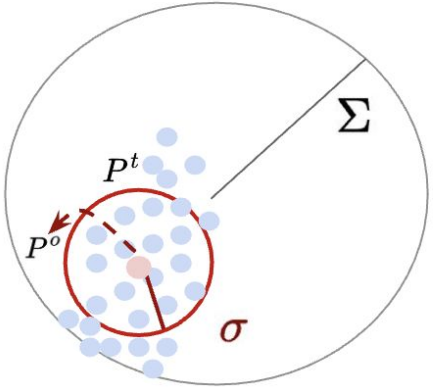
- **Average Regret Upper Bound** for Planning with Online Meta-Learning **(POMRL)**

Task Similarity

Our result: $\bar{\mathcal{L}} \leq \tilde{O}\left(\frac{\sigma}{\sqrt{T}} + \frac{\Sigma}{\sqrt{mT}}\right)$

#Tasks

Samples per task



Without meta-learning:

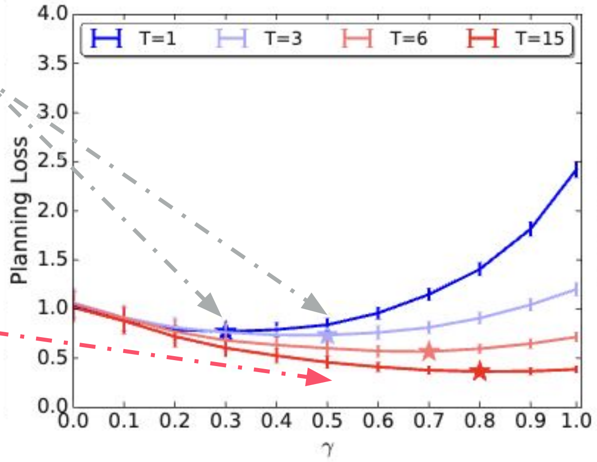
$$\bar{\mathcal{L}} \leq \tilde{O}\left(\frac{\Sigma}{\sqrt{m}}\right)$$

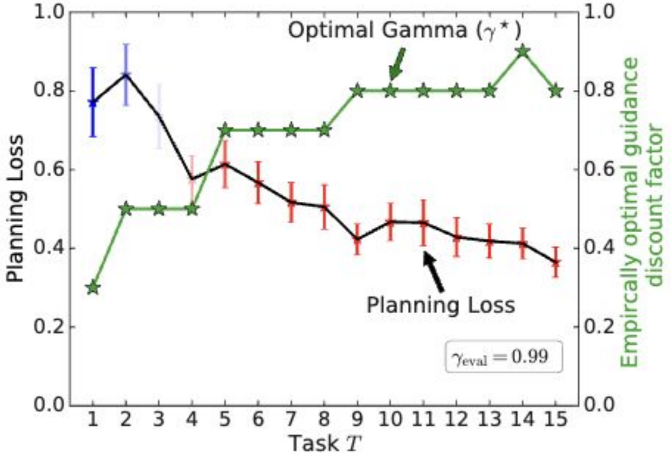
Planning with Online Meta-learning: *Experiments*

Does meta-learning a good initialization of dynamics model enables longer planning horizons and improved planning accuracy?

For initial tasks, an intermediate value of gamma is optimal

A better meta-learned initialization of the task dynamics, led to longer effective planning horizon.



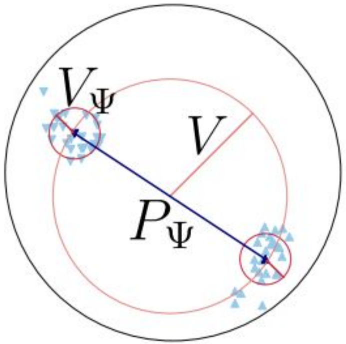


➔ **Meta-reinforcement learning leads to improved planning accuracy.**

➔ **The more knowledgeable the agent, the longer its planning horizon.**

Open Research Questions

- ➢ Non-stationary or shifts in underlying task distribution
- ➢ Scaling up with meta-gradients.
- ➢ More tractable algorithm with a proxy to planning loss



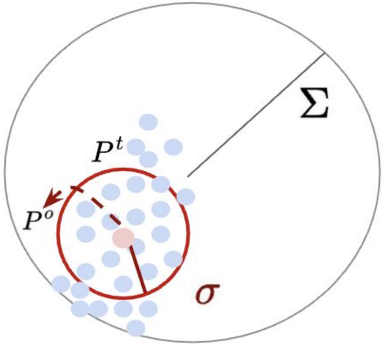
(ARUBA by Khodak et al. 2019)

tl;dr Adaptive Planning Horizon and Meta-Reinforcement Learning

Meta-learning a *good* initialization of the transition model across *similar* tasks allows to *plan longer ahead*.

Our result: $$\bar{\mathcal{L}} \leq \tilde{O}\left(\frac{\sigma}{\sqrt{T}} + \frac{\Sigma}{\sqrt{mT}}\right)$$

Without meta-learning: $$\bar{\mathcal{L}} \leq \tilde{O}\left(\frac{\Sigma}{\sqrt{m}}\right)$$



Come to our poster for more details!