



DeepMind

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

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



Meta Paris, 2023

@ Virtual





Dong et al. 2021

Interplay Between Planning Horizon & Meta-Reinforcement Learning

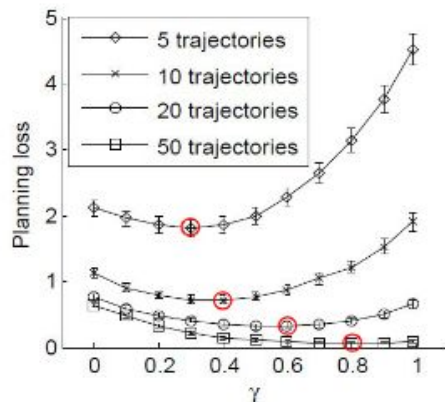


Motivation – choice of planning horizon

Private & Confidential

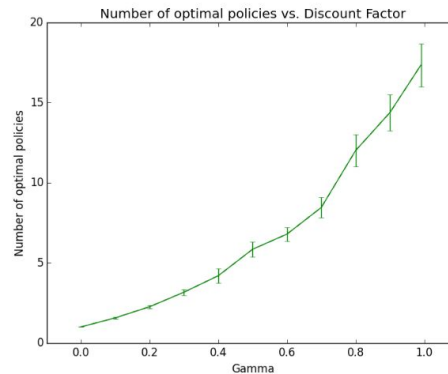
- A key component in the lifetime of an RL agent is the **planning horizon** $H = \frac{1}{1 - \gamma}$
- The choice of the planning horizon plays an important role, for e.g.

Optimality



Jiang et al. 2015

Complexity of the Policy Class



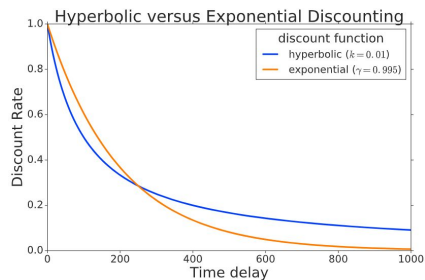
Arumugam et al. 2020



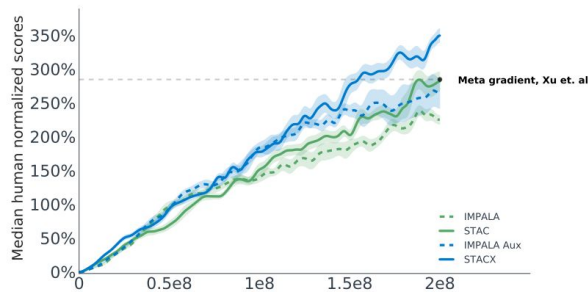
Motivation- empirical impact

Private & Confidential

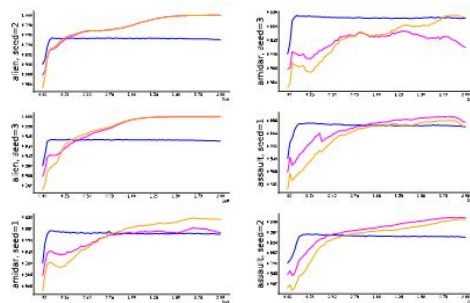
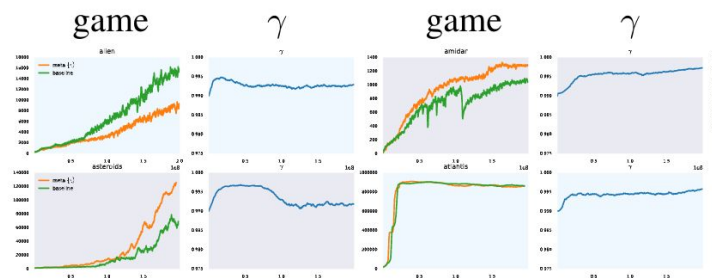
Fedus et al. 2019



Zahavy et al. 2021



Xu et al. 2018



Adapting discount factors has proven to be successful for many Deep RL algorithms



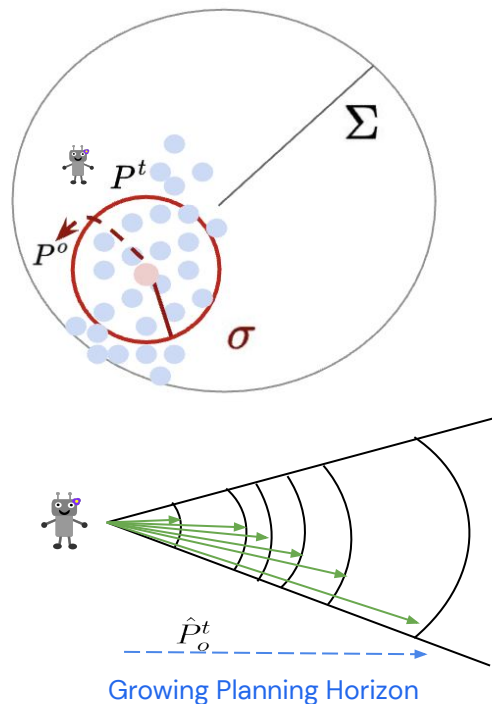
There is a direct correlation between the knowledge acquired by the agent and the effective planning horizon: the **more knowledgeable the agent, the longer its planning horizon.**

Research Question

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



- The agent is presented with a sequence of **T** RL tasks
 - ❑ In each task, the agent observes **m transitions from each state action pair** (generative model)
 - ❑ The task is sampled from a task distribution centered at P^o
- The agent **estimates** a model for the current task \hat{P}^t
 - ❑ The model is used for **planning** to find a policy
 - ❑ **Better model → better policy**
- The **estimate** is based on
 - ❑ The **m samples** from the current task
 - ❑ A **meta-learned** initialization/prior \hat{P}_o^t (from all the tasks)
 - ❑ **Better prior → better model → better policy**
- **Adapt the discount factor** as we meta-learn
 - ❑ Estimate σ
 - ❑ **Better prior → better model → increase the discount → better policy**



Background: *Discount factor as a regularizer (planning)*

Private & Confidential

Planning Loss:

$$\left\| V_{P^t, \gamma_{\text{eval}}}^{\pi_{P^t}^*} - V_{P^t, \gamma_{\text{eval}}}^{\pi_{\hat{P}^t}^*} \right\|_{\infty}$$

A policy is found by planning in an estimated model with a guidance discount factor

The value of optimal policy wrt to the true model and evaluation discount factor (evaluated there).

The value of the policy when evaluated in true model with an evaluation discount factor.



Background: *Discount factor as a regularizer (planning)*

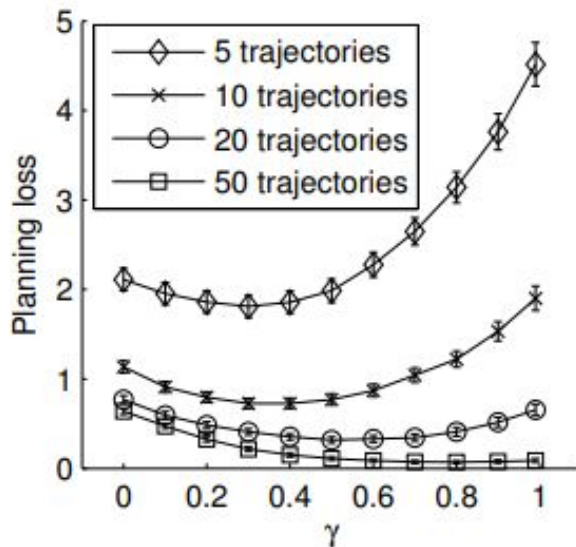
Private & Confidential

High-probability bound:

$$\left\| V_{P^t, \gamma_{\text{eval}}}^{\pi_{P^t, \gamma_{\text{eval}}}^*} - V_{P^t, \gamma_{\text{eval}}}^{\pi_{\hat{P}^t, \gamma}^*} \right\|_{\infty} \leq \frac{\gamma_{\text{eval}} - \gamma}{(1 - \gamma_{\text{eval}})(1 - \gamma)} + \frac{2\gamma R_{\max}}{(1 - \gamma)^2} \left(\sqrt{\frac{1}{2m} \log \frac{2|\mathcal{S}||\mathcal{A}||\Pi_{R, \gamma}|}{\delta}} \right)$$

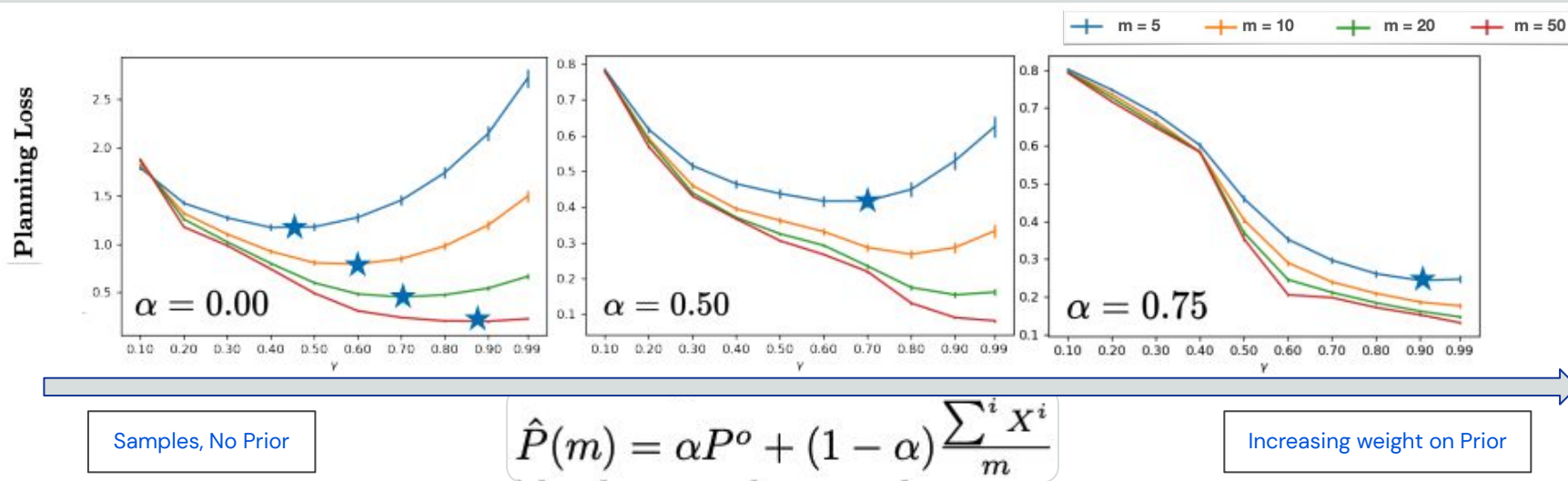
Jiang et al. 2015

There is a discount factor that minimizes the RHS, that is **not discount eval**



Incorporating a fixed prior

Private & Confidential



- The optimal discount is lower than 0.99 (eval)
- The optimal discount is lower when there are less samples
- Quality of the prior \sim samples



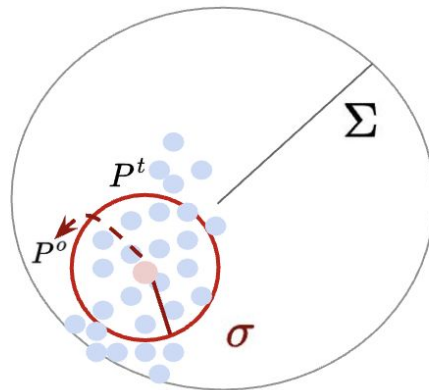
Planning with Online Meta-Learning



In this work, we assume that for all $t \in [T]$,

$P^t \sim \mathcal{P}$ centered at some fixed but unknown $P^o \in \Delta_S^{S \times A}$ and such that for any (s, a) ,

$$\|P_{s,a}^t - P_{s,a}^o\|_\infty \leq \sigma \quad \text{a.s.}$$



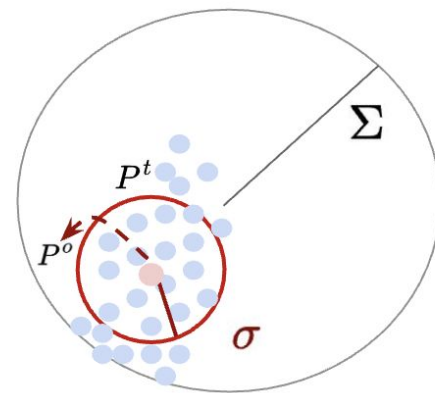
For simplicity we shall assume throughout that the rewards are known and focus on learning an approximate dynamics model.

We assume that for each task $t \in [T]$ we have access to a simulator of transitions providing m i.i.d. samples $(X_{s,a}^{t,i})_{i=1..m} \in \mathcal{S}^m \sim P^t(\cdot|s, a)$

For each (s, a) , we can compute an empirical estimator

$$\bar{P}_{s,a}^t(s') = \sum_{i=1}^m \mathbb{1}\{X_{s,a}^{t,i} = s'\} / m,$$

$$\text{where } \sum_{s'} \bar{P}_{s,a}^t(s') = 1.$$



- We perform Meta-RL by alternating between minimizing a batch ***within-task*** regularized least-squares loss (RLS), and **an outer-loop** step where we optimize the regularization to optimally balance bias and variance of the next estimator.
- **Estimating the dynamics model:** At each round, the current model is estimated by minimizing the RLS loss for a given regularizer:

$$\hat{P}_{(s,a)}^t = \arg \min_{P_{(s,a)} \in \Delta_S} \left\| \underbrace{\frac{1}{m} \sum_{i=1}^m \mathbb{1}\{X_{s,a}^{t,i}\}}_{\text{empirical transition prob.}} - P_{(s,a)} \right\|_2^2 + \lambda_t \|P_{(s,a)} - h_t\|_2^2,$$

- **Outer-loop: Meta learning the regularization:** At the beginning of each task t , the learner has already observed $t-1$ related but different tasks. We use Average of Means

$$h_t \leftarrow \hat{P}^{o,t} = \frac{1}{t-1} \sum_{j=1}^{t-1} \frac{\sum_i \mathbb{1}\{X^{i,j}\}}{m} := \frac{1}{t-1} \sum_{j=1}^{t-1} \bar{P}^j.$$



Planning with Online Meta-learning: *Our Approach*

Private & Confidential

$$\hat{P}_{(s,a)}^t = \arg \min_{P_{(s,a)} \in \Delta_S} \left\| \underbrace{\frac{1}{m} \sum_{i=1}^m \mathbb{1}\{X_{s,a}^{t,i}\}}_{\text{empirical transition prob.}} - P_{(s,a)} \right\|_2^2 + \lambda_t \|P_{(s,a)} - h_t\|_2^2,$$

- The solution of this can be computed in closed form as a convex combination of the empirical average (count based) and the prior, with α_t as the mixing parameter.

$$\hat{P}^t = \alpha_t h_t + (1 - \alpha_t) \bar{P}^t \quad \text{where } \alpha_t = \frac{\lambda_t}{1 + \lambda_t}$$

- **Outer Loop-Deriving the mixing rate:** To set mixing rate, we compute the Mean Squared Error (MSE) of the estimator and minimize an upper bound,

$$\text{MSE}(\hat{P}^t) \leq \alpha_t^2 \sigma^2 \left(1 + \frac{1}{t}\right) + (1 - \alpha_t)^2 \frac{1}{m} \quad \Rightarrow \quad \alpha_t = \frac{1}{\sigma^2(1 + 1/t)m + 1}$$



Algorithm 1: POMRL (σ) – Planning with Online Meta-Reinforcement Learning

Input: Set meta-initialization $\hat{P}^{o,1}$ to uniform, task-similarity ($\sigma(s, a)$) a matrix of size $S \times A$, mixing rate $\alpha_1 = 0$, and γ_{eval}

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.

$\gamma^* \leftarrow \gamma\text{-Selection-Procedure}(m, \alpha_t, \sigma, T, S, A)$

$\pi_{\hat{P}^t, \gamma^*}^* \leftarrow \text{Planning}(\hat{P}^t(m))$ //

Output: $\pi_{\hat{P}^t, \gamma^*}^*$

 Update $\hat{P}^{o,t+1}, \alpha_{t+1} = \frac{1}{\sigma^2(1+1/t)m+1}$ // meta-update AoM (Eq. 5) and mixing rate



Planning with Online Meta-learning: *Theory Result*

Private & Confidential

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

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

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

$$\bar{\mathcal{L}} \leq \frac{\gamma_{\text{eval}} - \gamma}{(1 - \gamma_{\text{eval}})(1 - \gamma)} + \frac{2\gamma S}{(1 - \gamma)^2} \tilde{O} \left(\frac{\sigma + \sqrt{\frac{1}{T}} \left(\sigma + \sqrt{\sigma^2 + \frac{\Sigma}{m}} \right)}{\sigma^2 m + 1} + \frac{\sigma^2 m \sqrt{\frac{\Sigma}{m}}}{\sigma^2 m + 1} \right)$$

Without meta-learning:
(Jiang et al, 2015)

$$\bar{\mathcal{L}} \leq \frac{\gamma_{\text{eval}} - \gamma}{(1 - \gamma_{\text{eval}})(1 - \gamma)} + \frac{2\gamma}{(1 - \gamma)^2} \times \tilde{O} \left(\sqrt{\frac{1}{m}} \right)$$



Planning with Online Meta-learning: *Theory Result*

Private & Confidential

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

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

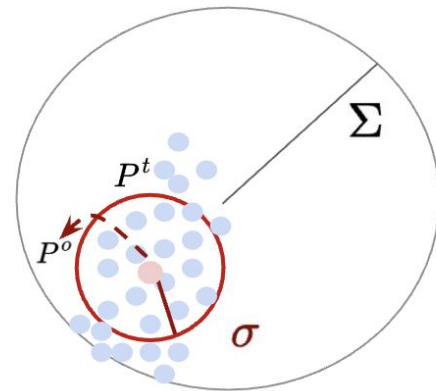
- **Average Regret Upper Bound** for Planning with Online Meta-Learning (POMRL)

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

Task Similarity (points to σ)

#Tasks (points to T)

Samples per task (points to m)



Without meta-learning:
(Jiang et al, 2015)

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



Planning with Online Meta-learning: *Implications for extreme cases*

Private & Confidential

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

$$\bar{\mathcal{L}} \leq \frac{\gamma_{\text{eval}} - \gamma}{(1 - \gamma_{\text{eval}})(1 - \gamma)} + \frac{2\gamma S}{(1 - \gamma)^2} \tilde{O} \left(\frac{\sigma + \sqrt{\frac{1}{T}} \left(\sigma + \sqrt{\sigma^2 + \frac{\Sigma}{m}} \right)}{\sigma^2 m + 1} + \frac{\sigma^2 m \sqrt{\frac{\Sigma}{m}}}{\sigma^2 m + 1} \right)$$

➤ When tasks are all exactly the same, estimate only one model in a space of set size with mT samples

when $\sigma = 0$

$$\bar{\mathcal{L}} \leq \frac{\gamma_{\text{eval}} - \gamma}{(1 - \gamma_{\text{eval}})(1 - \gamma)} + \frac{2\gamma S}{(1 - \gamma)^2} \tilde{O} \left(\sqrt{\frac{\Sigma}{mT}} \right)$$

➤ When tasks are all very different, meta-learning is not relevant, estimate T models with m samples.

when $\sigma = 1$, then $\sigma = \Sigma$

$$\bar{\mathcal{L}} \leq \frac{\gamma_{\text{eval}} - \gamma}{(1 - \gamma_{\text{eval}})(1 - \gamma)} + \frac{2\gamma S}{(1 - \gamma)^2} \tilde{O} \left(\frac{1}{m} \left(1 + \frac{1}{\sqrt{T}} \left(1 + \sqrt{1 + \frac{1}{m}} \right) \right) + \frac{1}{\sqrt{m}} \right)$$

Added bias due to regularization but
second order in $1/m$

Usual estimation error for each task



Planning with Online Meta-learning: *Practical Considerations on Algorithm*

Algorithm 2: ada-POMRL – Planning with Online Meta-Reinforcement Learning

Input: Set meta-initialization $\hat{P}^{o,1}$ to uniform, initialize $(\hat{\sigma})_1$ as a matrix of size $S \times A$, mixing rate

$\alpha_1 = 0$, and γ_{eval}

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.

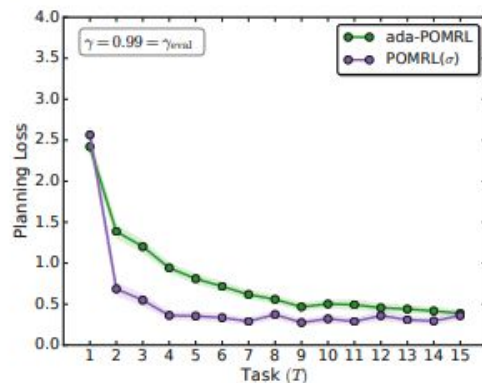
$\gamma^* \leftarrow \gamma\text{-Selection-Procedure}(m, \alpha_t, \sigma_t, T, S, A)$

$\pi_{\hat{P}^t, \gamma}^* \leftarrow \text{Planning}(\hat{P}^t(m))$ // $\forall \gamma \leq \gamma_{\text{eval}}$

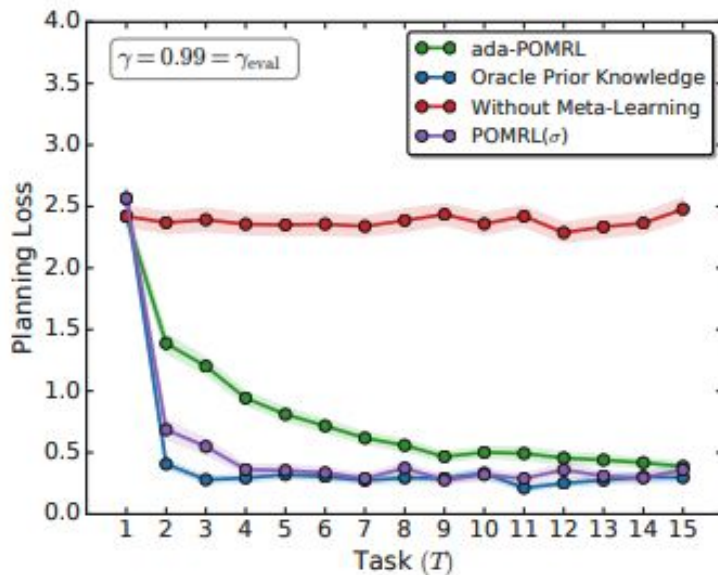
Output: $\pi_{\hat{P}^t, \gamma}^*$

Update $\hat{P}^{o,t+1}, \hat{\sigma}_{t+1} \leftarrow \text{Welford's online algorithm}((\hat{\sigma}_o)_t, \hat{P}^{o,t+1}, \hat{P}^{o,t})$ // 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 $\max(\sigma_{S \times A})$



Q1. Does meta-learning a good initialization of dynamics model facilitate improved planning accuracy for the choice of evaluation discount factor?



➤ Tl;dr: Meta-reinforcement learning leads to improved planning accuracy.

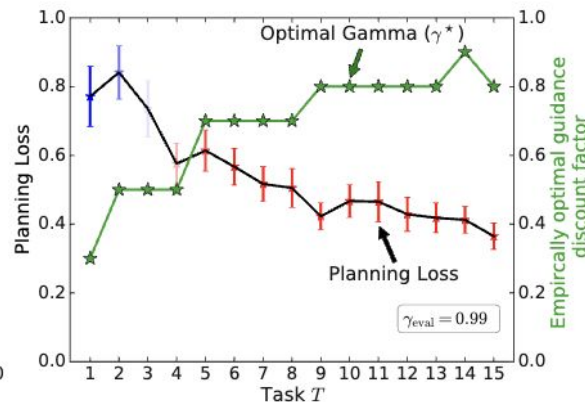
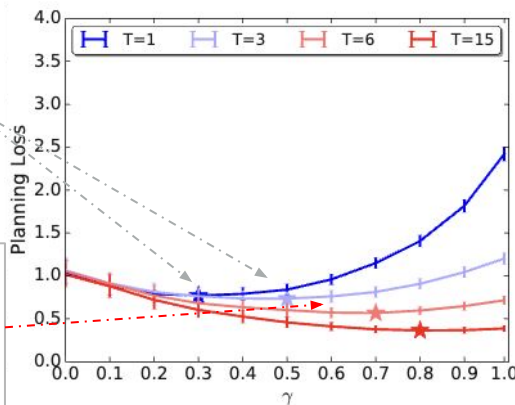


Q2. Does meta-learning a good initialization of dynamics model enables longer planning horizons?

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.

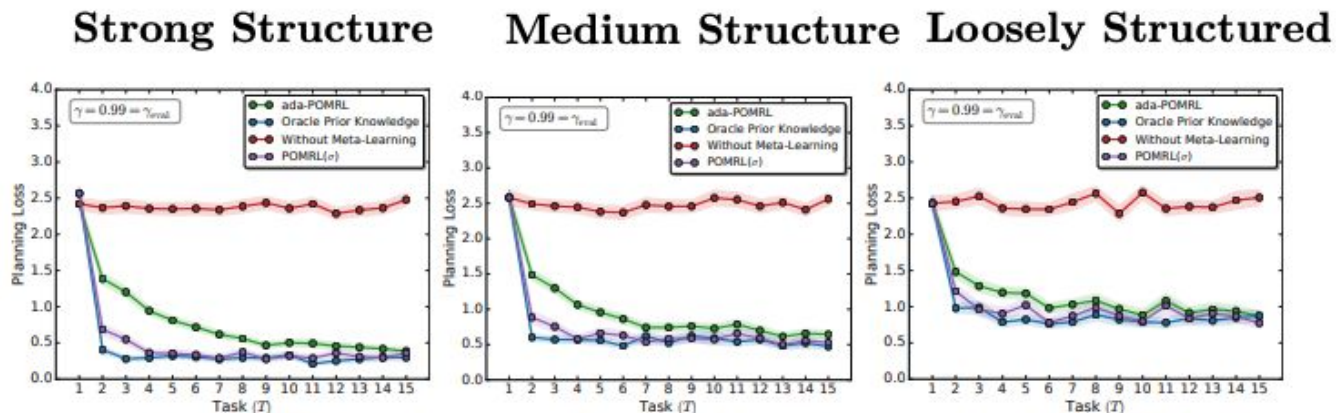
Recall



the more knowledgeable the agent, the longer its planning horizon.



Q3. How does performance depend on the amount of shared structure across tasks?



- **Tl;dr:** POMRL and ada-POMRL perform **consistently** well for varying task-similarity.
- An **intermediate value of task-similarity** still leads to gains, albeit at a lower speed of convergence.
- In contrast, a larger value indicates **little structure across tasks resulting in minimal gains from meta-learning**. The learner struggles to learn a good initialization of the model dynamics as there is no natural one. All planning loss curves remain U-shaped and overall higher with a smaller optimal guidance discount.



Adaptation of Planning Horizon



Adaptation of Discount Factor: *Intuition*

- Equivalence between effective planning horizon, tasks/samples, & meta-learned initialization

$$\gamma_t := \gamma_{\text{samples}} + \gamma_{\text{prior}}$$

Effective planning
horizon at time t

horizon gained from m
samples per task

horizon gained from meta-learned
initialization at round t

$m(t-1)$ samples equivalent to
meta-learned task dynamics at round t



Adaptation of Discount Factor: *Heuristic Based on Prior Work*

- We **propose two heuristics** to design an **adaptive schedule for discount factor**
- We adapt the schedule proposed by Dong et al. to our problem:

$$T_0 = m, \quad T_t = \frac{SA}{L} \underbrace{\left((1 - \alpha_t)m + \alpha_t m(t - 1) \right)}_{\text{efficient sample size}}, \quad \gamma_t = 1 - \frac{1}{T_t^{1/5}},$$

- Where L is the maximum trajectory length.
- **The size of samples in each task** is controlled by the **efficient sample size** which includes a combination of
 - ❑ the current task's samples and,
 - ❑ the sample observed so far, as used to construct our estimator in POMRL.



Adaptation of Discount Factor: *Theory driven schedule*

- Next, we **use the upper bound to guide the schedule**
- Having a second look at our main theory result, we see that the RHS is a function of the form

$$U : \gamma \mapsto \frac{1}{1 - \gamma_{\text{eval}}} + \frac{1}{\gamma - 1} + C_{m,T,S,A,\sigma,\delta} \frac{\gamma}{(1 - \gamma)^2},$$

Proposition 1. *The existence of a strict minimum in $(0, 1)$ is determined by $C = C_{m,T,S,A,\sigma,\delta}$ (which can be computed) as follows:*

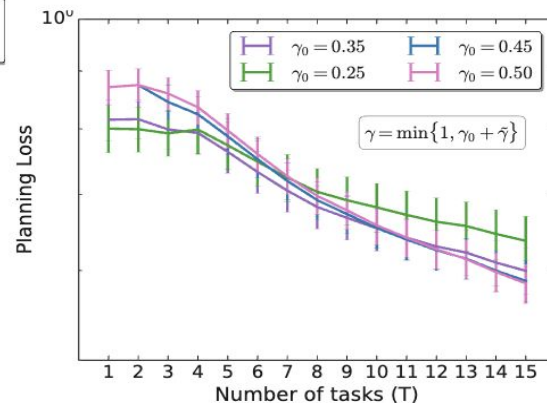
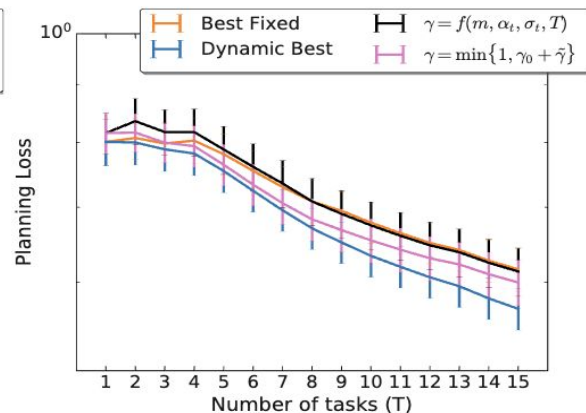
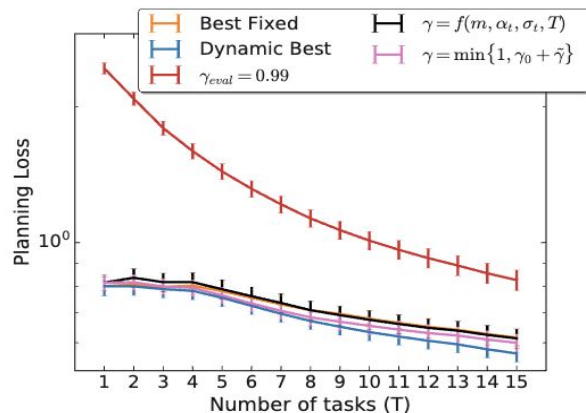
$$\tilde{\gamma} = \begin{cases} 0 & \text{if } C \geq 1 \\ 1 & \text{if } C < 1/2 \\ \frac{1-C}{1+C} & \text{otherwise, i.e if } 1/2 < C < 1 \end{cases}$$

$$[\gamma = \min\{1, \gamma_0 + \tilde{\gamma}\}]$$



Adaptation of Discount Factor: *Experiments*

Private & Confidential

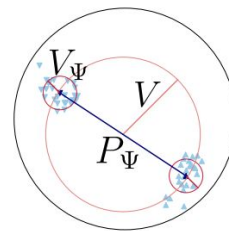


- Using the **evaluation discount factor** results in a **very high loss**, due to trying to plan too far ahead despite model uncertainty.
- The **proposed** $\gamma_t = 1 - \frac{1}{T_t^{1/5}}$, obtains **similar performance to best-fixed** and is within the significance range or the lower bound.
- The **upper bound guidance** for selection of gamma obtains **similarly good performance**.

TL;dr: Evidence suggests it is possible to adapt the planning horizon as a function of the problem structure (e.g. meta-learned task-similarity) and data per task.



- **Non-stationary** or **shifts** in underlying **task distribution** is an important problem to consider.
- Our analysis focused on **planning** and model based RL.
 - ❑ Learning in a model-free setting is a promising to explore
 - ❑ Preliminary investigation of Optimistic Q-learning [Dong et. al 2021] did not yield immediate results
- **Scaling up empirical** work to meta-gradients.
 - ❑ A better understanding of function approximation theory will provide further insight
 - ❑ Connections to DISTRAL – our work is groundwork for analysing similar meta-learning algorithms
- More tractable algorithm with a **proxy to planning loss** (doesn't require the true MDP)



([ARUBA by Khodak et al. 2019](#))



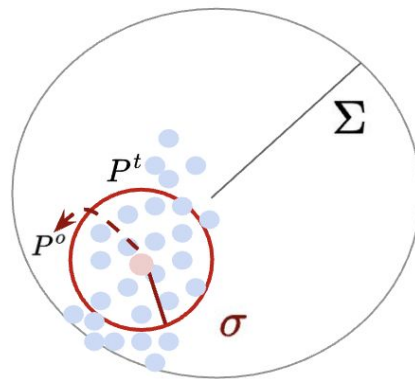
tl;dr Adaptive Planning Horizon and Meta-Reinforcement Learning

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

(Jiang et al, 2015)



Takeaway: meta-learning helps define longer planning horizons!

