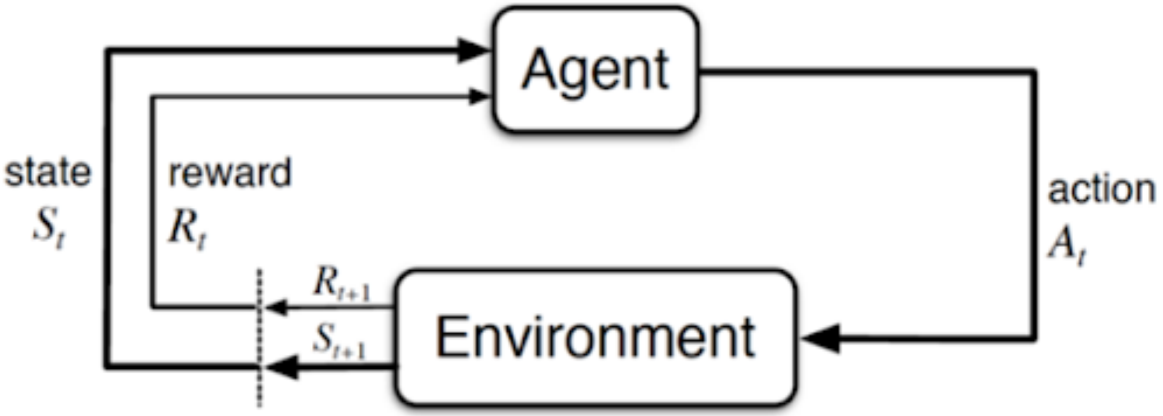
Unified Approach – Bayesian Adaptive MDPs

Markov Decision Processes (MDP)

- MDP: framework for sequential decision making
- State – $S_t \in \mathcal{S}$
- Action – $A_t \in \mathcal{A}$
- Dynamics – $P(\cdot\,|S, A) \in \mathcal{P}(\mathcal{S})$
- Reward – $R(S, A) \sim q(\cdot\,|S, A) \in \mathcal{P}(\mathcal{R})$
- MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, q)$ with objective: $\text{Maximize } \mathbb{E}(\sum_{t=0}^{T-1} \gamma^t R\,(S_t, A_t)|S_0 = s)$



Partially Observable MDP

- Complete states are not observable, instead observation O_t is observed at time step t
- State – $S_t \in \boldsymbol{\mathcal{S}}$, Action – $A_t \in \boldsymbol{\mathcal{A}}$, Dynamics – $P(\cdot\,|S, A) \in \boldsymbol{\mathcal{P}}(\boldsymbol{\mathcal{S}})$
- Observation $\mathrm{O} \in \boldsymbol{\mathcal{O}}$, where $\mathrm{O_t} \sim \Omega(\cdot\,|S_{t+1}, A_t)$
- POMDP $\boldsymbol{\mathcal{M}} = (\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{O}}, P, \Omega, q)$

MDP with unknown dynamics

- Dynamics – $\mathrm{P}_\theta(\cdot\,|S, A)$ with $\theta \sim \mu_0$
- Reformulation as Partially observable MDP
  - ✓ Define a new hyper-state $S' = (\theta, S)$ with observations $\mathrm{O}'_\mathrm{t} = S_t$
  - ✓ Dynamics over hybrid state

$$P'(S'_{t+1}|A_t, S'_t\,) = P'(S_{t+1}, \theta|A_t, S_t, \theta\,) = \; P_\theta(S_{t+1}|A_t, S_t)$$

  - ✓ $\Omega'(O_{t+1}|S'_{t+1}, A_t) = \Omega(O_{t+1}|S_{t+1}, \theta, A_t) = (\mathrm{S}_{\mathrm{t+1}}\ \mathrm{w.p.}\,1)$
  - ✓ POMDP $\boldsymbol{\mathcal{M}}' = (\boldsymbol{\mathcal{S}}', \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{O}}', P', \Omega', q)$

Bayesian Adaptive MDP

- Let $\mu_t = \mu(\theta | H_t, \mu_0)$ be the posterior over the parameter θ given the history $H_t = (S_0, A_0, \cdots, A_{t-1}, S_t)$
- Define an MDP as follows
  - ✓ Hyper-state with posteriors $\tilde{S}_t = (S_t, \mu_t)$
  - ✓ Transition dynamics over the hybrid state space $\tilde{P}\left(\tilde{S}_{t+1} \middle| \tilde{S}_t, A_t\right)$

    $$\tilde{P}(S_{t+1}, \mu_{t+1} | S_t, \mu_t, A_t) = \mathbb{1}(\mu_{t+1} | S_{t+1}, S_t, A_t, \mu_t)\left(\int P_\theta(S_{t+1} | S_t, A_t)\, d\mu_t(\theta)\right)$$
  - ✓ Reward function $\tilde{q}\left(\cdot \middle| \tilde{S}, A\right) = \tilde{q}(\cdot | S, \mu, A) = q(\cdot | S, A)$
  - ✓ Bayesian Adaptive MDP $\widetilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, \tilde{q})$

Generalizing to unknown reward functions

- Reward $R(S_t, A_t) \sim q_\alpha(\cdot\,|S_t, A_t)$ with $\alpha \sim \nu_0$
- Define $\mu_t = \mu(\theta|H_t, \mu_0)$ and $\nu_t = \nu(\alpha|H_t, \nu_0)$ be the posterior over the parameter θ and α given the history H_t
- Define an MDP as follows
  - ✓ Hyper-state with posteriors $\tilde{S}_t = (S_t, \mu_t, \nu_t)$
  - ✓ Transition dynamics over the hybrid state space and rewards

$$\tilde{P}\left(\tilde{S}_{t+1}, R_t \middle| \tilde{S}_t, A_t\right) = \tilde{P}(S_{t+1}, \mu_{t+1}, \nu_{t+1}, R_t|S_t, \mu_t, \nu_t, A_t)$$

$$= \mathbb{1}(\mu_{t+1}|S_{t+1}, S_t, A_t, \mu_t)\left(\int P_\theta(S_{t+1}|S_t, A_t)\, d\mu_t(\theta)\right) \mathbb{1}(\nu_{t+1}|R_t, S_t, A_t, \nu_t)\left(\int q_\alpha(R_t|S_t, A_t)\, d\nu_t(\theta)\right)$$

  - ✓ Bayesian Adaptive MDP $\widetilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P})$

Connecting back…

- Bayesian bandits as an MDP with unknown reward function
  - ✓ K arms with means $\alpha = (\alpha_1, \cdots, \alpha_K)$ with $\alpha \sim \nu_0$ (Prior)
  - ✓ Actions A_t – which arm to pull
  - ✓ States $S_t = \phi$, with $P(S_{t+1} = \phi | S_t, A_t) = 1$
  - ✓ Reward $R_t = q_\alpha(\cdot\,|A_t, S_t) = q_\alpha(\cdot\,|A_t)$

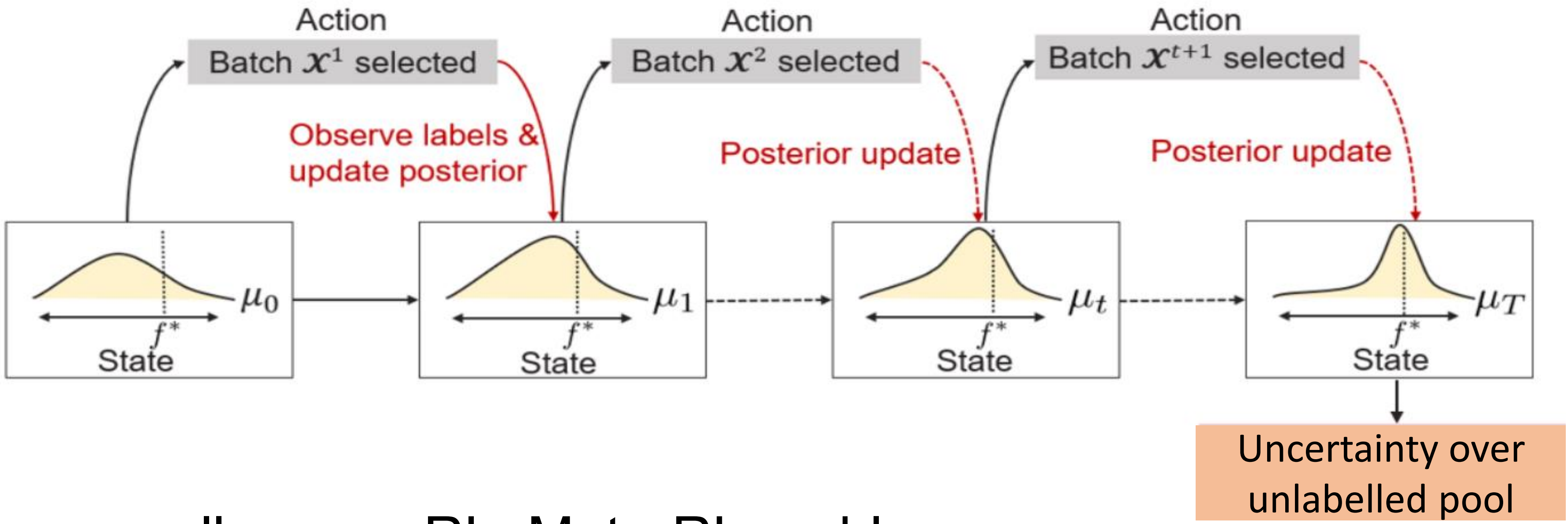
- Bayesian bandits as BAMDP
  - ✓ Maintain a posterior ν over α, and define hyper-state $\tilde{S}_t = (S_t, \nu_t) = \nu_t$
  - ✓ $\tilde{P}\left(\tilde{S}_{t+1}, R_t \middle| \tilde{S}_t, A_t\right) = \mathbb{1}(\nu_{t+1} | R_t, A_t, \nu_t)\left(\int q_\alpha(R_t | A_t)\, d\nu_t(\alpha)\right)$
  - ✓ Bayesian Adaptive MDP $\widetilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P})$

Connecting back…

- Similarly, we can express other decision-making problems as BAMDP

    - ✓ Bayesian Optimization

    - ✓ Adaptive data collection



    - ✓ More generally – any RL, Meta-RL problem

Conventional Solution Approaches

- Offline Value Approximation – (approximately solves BAMDP) – intractable for most domains
- Online near-myopic value approximation
  - ✓ Bayesian Dynamic Programming: extends Thompson sampling to BAMDP – sample a model θ from μ_t and take best action A_t according to θ (solve MDP θ).
  - ✓ Value of information heuristic: extends knowledge gradient ideas to BAMDP
- Online Tree search approximation
  - ✓ Perform a forward search in the space of hyper-states.

Two key challenges

- How to solve the BAMDP?
  - ✓ Continuous state space due to posteriors
  - ✓ Number of possible hyper-states increases exponentially with horizon

- How to get reliable posteriors?
  - ✓ Posterior updates are intractable for most domains
  - ✓ Further the current UQ methodologies (such as BNN etc.) suffer from following challenges
    - ✓ How much we should sharpen our belief as we see more data points – requires learning priors to effectively quantify uncertainty
  - ❖ Potential solution: UQ through meta learned sequence models

References

- Active Learning Literature Survey. Burr Settles (2010). https://burrsettles.com/pub/settles.activelearning.pdf
- A Survey on Deep Active Learning: Recent Advances and New Frontiers. Li et. al. (2024) https://arxiv.org/pdf/2405.00334
- Active Learning: A survey. Aggarwal et. al. https://charuaggarwal.net/active-survey.pdf
- Bayesian Reinforcement Learning: A Survey, Ghavamzadeh et. al. (2016) https://arxiv.org/pdf/1609.04436
- Deep Bayesian Active Learning with Image Data. Gal et. al. (2017) https://arxiv.org/abs/1703.02910