Towards a practical Bayes-optimal agent

Arthur Guez[1], David Silver[2], and Peter Dayan[1]

[1] Gatsby Computational Neuroscience Unit, UCL [2] DeepMind Technologies

Introduction

Only rich statistical models are adequate for agents that must learn to navigate complex environments. However, it has not been clear how methods for planning can take advantage of these models. Myopic methods such as Thompson Sampling have shortcomings that we illustrate with formal counter-examples. **We show that Bayes-Adaptive planning can be combined in a principled way with approximate sampling, and demonstrate the power of the resulting method in a challenging task involving safe exploration.** This highlights the importance of propagating beliefs in realistic cases involving trade-offs between exploration and exploitation.

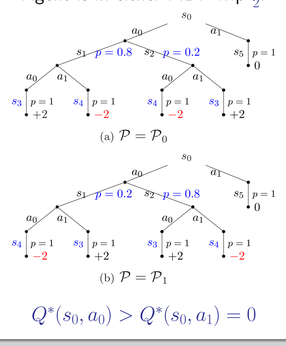
Model-based Bayesian RL

$\mathcal{P} :=$ Model of the dynamics.
$\mathcal{D} :=$ Interaction data (transitions, rewards).

- Start with prior $P(\mathcal{P})$.
- Receive data and update posterior: $P(\mathcal{P} \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mathcal{P})P(\mathcal{P})$.
- During interaction, choose actions to maximize $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$, with $\gamma < 1$.
- Requires balancing exploration and exploitation.
- **Bayes-optimal** policy integrates over how $P(\mathcal{P} \mid \mathcal{D})$ might change as the result of expected *future interactions*.

Bayes-Adaptive example

Agent is in either MDP w.p $\frac{1}{2}$.



$Q^*(s_0, a_0) > Q^*(s_0, a_1) = 0$

Can we make it work in practice?

Rich statistical models allow confident inferences from limited observations, but inference is generally hard and approximate. Existing work generally combines:

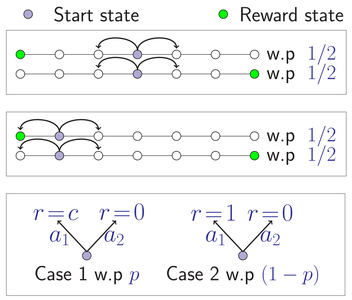
Rich statistical models + Myopic planning
or **Toy statistical models + Adaptive planning**.

- What's wrong with myopic planning? See **Box 1**.
- Is it feasible to combine rich stat. models+adaptive planning? Yes, see **Box 2**.
- Is there much to gain? Yes, see **Box 3-4**.
- What else is needed for a fully practical solution? See **Box 5**.

Box 1: Myopic planning

Myopic planning doesn't take into account how belief will evolve. For example, **Thompson Sampling** works by: (1) Sampling a single $\mathcal{P}$ according to $P(\mathcal{P} \mid \mathcal{D})$, (2) Solving MDP corresponding to $\mathcal{P}$, and (3) Selecting greedy action.

Understanding what can go wrong in examples:



E1: Thompson Sampling (TS) random walking. $O(x^2)$ steps to reach either end.

E2: Problem with myopic+commitment. Don't go right first!

E3: Unwarranted optimism about cost. Let $c \ll 0$, $V^*=0$ but $V_{TS}=(1-p)(pc+1-p)$. Can be worse for other alg. (e.g., BOSS). Chain instances together to obtain an MDP example.

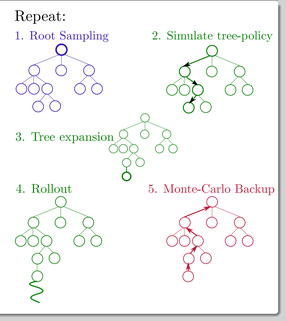
Box 2: Sample-based Forward-search with root sampling

Plan from current belief state only and repeat:

- Perform (approximate) inference to get samples from $P(\mathcal{P} \mid \mathcal{D})$ *at tree root only.*
- Run a simulation of a tree-based planner (Monte-Carlo Tree Search) for each sample.
- Perform Monte-Carlo backup.

→ **BAMCP** algorithm.

Th: Converges to Bayes-optimal policy *even when using MCMC inference.*

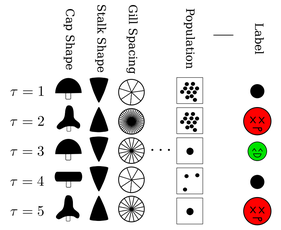


Box 3 - Case study: Mushroom exploration task

Task: Discrimination as generalization under safe exploration.

You can choose to eat or skip each mushroom. Some are poisonous.

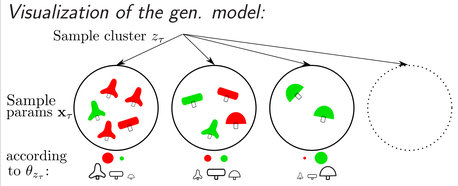
- Mushroom τ described by attributes $\mathbf{x}_\tau$. (Data from UCI mushroom dataset)
- $r << 0$ if poisonous, $r > 0$ if edible.
- Skip mushroom (•): still learn about distribution of mushrooms (unsupervised).
- *Can we safely explore and exploit without specific knowledge about the mushrooms?*



Prior

Assume general prior that only encodes 'mushrooms are likely to be like others'. The prior says that each τ belongs to a latent cluster z_τ, distributed according to a Chinese Restaurant Process (CRP).

Visualization of the gen. model:

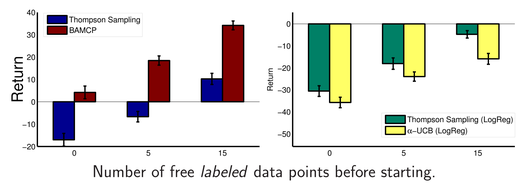


Formal generative model:

$$\alpha \sim \text{Gamma}(a, b),$$
$$z_\tau \sim \text{CRP}(\alpha),\ \forall \tau \in \mathbb{Z}^+,$$
$$\boldsymbol{\theta}_k^i \sim \text{Dirichlet}(\frac{\beta}{D_i}),\ \forall i \in \{1, \ldots, n\}, \forall k \in \mathbb{Z}^+,$$
$$x_\tau^i \sim \text{Categorical}(\boldsymbol{\theta}_{z_\tau}^i)\ \forall i \in \{1, \ldots, n\}, \forall \tau \in \mathbb{Z}^+,$$

Results

Combine with planning from Box 1-2.
Left: With CRP gen. model described above.
Right: Using logistic regression.



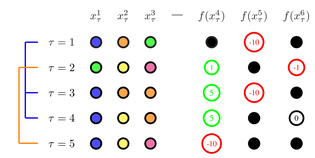
References

- *Asmuth et al. (2009) A Bayesian sampling approach to exploration in reinforcement learning.*
- *Doshi-Velez et al. (2010) Nonparametric Bayesian Policy Priors for Reinforcement Learning.*
- *Guez et al. (2012) Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search.*
- *Ross & Pineau (2008) Model-Based Bayesian Reinforcement Learning in Large Structured Domains.*
- *Silver & Veness (2010) Monte-Carlo Planning in Large POMDPs.*

Box 4 - Contextual CRP Bandits

Model

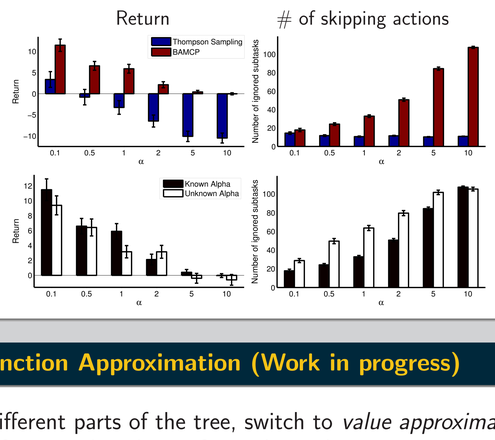
- Extend model in Box 3 with multiple actions/outcomes.
- Remove model mismatch by using data from generative model.
- $\mathbb{E}[r]$ for each action is negative!



Results

Vary concentration parameter α (Known). Bayes-Adaptive (BAMCP) vs myopic planning.

Comparison known/unknown α for BAMCP. ($\alpha \sim \text{Gam}(0.5, 0.5)$)



Box 5: Value Function Approximation (Work in progress)

Generalize between different parts of the tree, switch to *value approximation* to represent the belief-state value *during forward search*.

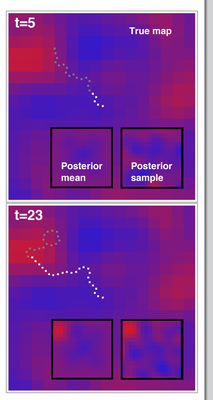
- For history h, represent $Q(h, a)$ as $\phi(h, a)^T w$.
- Should improve search efficiency further: enables longer search horizons.
- Compatible with continuous state space.
- Two main issues: Finding good $\phi(h, a)$ + Theoretical guarantees.

Example:

A search problem arising from a Gaussian Process with continuous states. A latent function $f \sim GP$ describes the reward signal (red: high values, blue: low values). Reward is only observed locally.

The agent needs to find the high reward regions. On the right, an agent implementing a forward-search with function approximation using history features. Root sampling is performed by sampling from the GP posterior at the current belief state for each forward simulation.

White: past trajectory of agent. Grey: forward simulation. Large map is the true function, inset is a posterior sample.



Conclusion

- Bayes-adaptive planning is conceivable in complex, infinite domains.
- Can exploit rich statistical models.
- Large gains vs myopic planning in a contextual-bandit task.
- For larger domains, need better models (capture more structure) and more efficient planning algorithms (see Box 5).

This work was funded by the Gatsby Charitable Foundation and the Natural Sciences and Engineering Research Council of Canada.

Contact: aguez@gatsby.ucl.ac.uk