STA 4273: Minimizing Expectations

Lecture 11 - Inference and control

Chris J. Maddison

University of Toronto

Announcements

- Extension on the final project report. Now due April 14.

Today

- Bayesian RL, distinct from RL as inference.
- Thompson Sampling.

MDPs today

For todays lecture an MDP $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, T, \rho \rangle$ will be defined by

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Transition matrix $\mathcal{P}$
- Initial state distribution ρ
- Reward function r
- Horizon T

Exploration vs. exploitation

- Discussed exploration vs. exploitation tradeoff in the context of bandits ($T = 1$).
- Can we formalize exactly what we mean by this in a general MDP setting?
- Consider the following distinction. Suppose that either
  1. Observed M: we know the full description of the MDP M, in which case we can implement planning or optimal control.
  2. Unobserved $M \in \mathcal{M}$: we know that $M \in \mathcal{M}$ is in a family of Markov decision processes, but we must explore to figure out which M we're in.

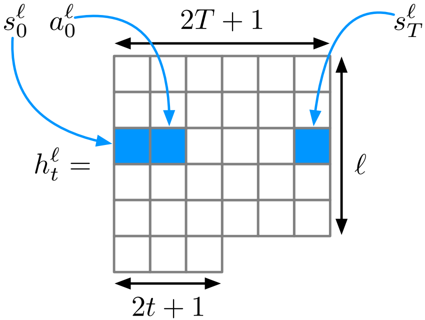
Exploration vs. exploitation

- By considering unobserved $M \in \mathcal{M}$, we can formalize what we mean by exploration vs. exploitation. We will focus on this setting.
- This discussion is based on the following references:
  - Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar. (2015). Bayesian Reinforcement Learning: A Survey.
  - Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen. (2020). A Tutorial on Thompson Sampling.
  - Arthur Guez, David Silver, Peter Dayan. (2013). Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search.
  - Brendan O'Donoghue, Ian Osband, Catalin Ionescu. (2020). Making Sense of Reinforcement Learning and Probabilistic Inference.

Exploration vs. exploitation

Suppose we have interacted with an MDP M for ℓ episodes and t timesteps on the $\ell + 1$ episode.

- We observe histories h_t^ℓ,



- An RL algorithm alg maps histories $h_t^\ell \to \pi_{\ell,t}$ to policies.
- Given M, alg we can define the sequence of histories h_t^ℓ as those produced by iteratively interacting with M via $\pi_{\ell,t}$.

Exploration vs. exploitation

If we have a budget of L episodes, we can evaluate algs according to

- Worst-case regret

$$\max_{M \in \mathcal{M}} \mathbb{E}\left[\sum_{\ell=1}^{L} V_0^{M,*}(s_0^\ell) - \sum_{t=1}^{T} r(s_t^\ell, a_t^\ell) \,\middle|\, M, \text{alg} \right]$$

- Bayesian regret for some prior p over $\mathcal{M}$,

$$\mathbb{E}_{M \sim p}\left[\mathbb{E}\left[\sum_{\ell=1}^{L} V_0^{M,*}(s_0^\ell) - \sum_{t=1}^{T} r(s_t^\ell, a_t^\ell) \,\middle|\, M, \text{alg} \right] \right]$$

These are the same for Dirac priors. Let's focus on Bayesian regret.

Exploration vs. exploitation

- To do well on Bayesian regret, an agent needs to be statistically efficient and consider the value of information.
- This means maintaining an estimate of M, so that it can direct its action to states that reveal more information about M.
- Yet, not sacrificing too much in terms of accumulated returns.

Example

Consider a bandits example.

- $\mathcal{S} = \{1\}, T = 1, \mathcal{M} = \{M^+, M^-\}, \mathcal{A} = \{1, 2, 3, \ldots, N\}$.
- Only difference is rewards (color = optimal arm):

$$r^+(1,1) = 1, r^+(1,2) = +2, r^+(1,a) = 1 - \epsilon \text{ for } a \geq 3$$
$$r^-(1,1) = 1, r^-(1,2) = -2, r^-(1,a) = 1 - \epsilon \text{ for } a \geq 3$$

Now let's consider different settings.

Example–M known

$$r^+(1,1) = 1, r^+(1,2) = +2, r^+(1,a) = 1 - \epsilon \text{ for } a \geq 3$$
$$r^-(1,1) = 1, r^-(1,2) = -2, r^-(1,a) = 1 - \epsilon \text{ for } a \geq 3$$

- If M is known, then the optimal policy is trivially $a^\ell = 2$ in M^+ and $a^\ell = 1$ in M^-.

Example—M unknown, worst-case regret optimal

$$r^+(1,1) = 1, r^+(1,2) = +2, r^+(1,a) = 1 - \epsilon \text{ for } a \geq 3$$
$$r^-(1,1) = 1, r^-(1,2) = -2, r^-(1,a) = 1 - \epsilon \text{ for } a \geq 3$$

1. Choose $a^0 = 2$, observe r^0.
2. If $r^0 = +2$, then pick $a^\ell = 2$ for all $\ell \geq 1$.
3. If $r^0 = -2$, then pick $a^\ell = 1$ for all $\ell \geq 1$.

This achieves a regret of 3, and is worst-case optimal (also Bayes optimal as long as $p(M^+)L > 3$).

Bayes-optimal strategies

- In general, policies that optimize the Bayesian regret are still poorly understood. (As I understand it, I am not an expert in this area.)
- Instead, let us consider so-called Bayes-optimal strategies that directly maximize the expected return over a single episode with unobserved M:

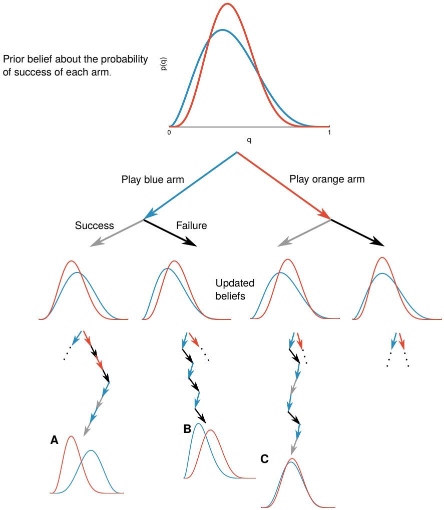
$$\arg\max_{\pi} \mathbb{E}_{M \sim p} \left[\mathbb{E} \left[\sum_{t=1}^{T} r(s_t, a_t) \,\middle|\, M \right] \right]$$

- This objective is not the expected return of an MDP. Technically, it's a POMDP, where M is an unobserved variable.
- Algorithms that approximate Bayes-optimal strategies often good in practice.

Bayes-optimal strategies

- An agent can receive higher reward, if it performs Bayes-rationally about the information it's received.
- So, Bayes-optimal strategy gives value to exploration moves.
- To see what I mean, let's consider a T round bandit problem as an unobserved MDP.
    - Can think of this as a single state MDP.

Bayes-optimal strategies



(Guez, 2015)

Bayes-optimal strategies

Interaction can with an unobserved M can be formulated as a fully-observed MDP by expanding the state-space.

- Expanded state space $\mathcal{S}^+ = \mathcal{S} \times \mathcal{H}$ where $\mathcal{H}$ is the set of all histories $h_t = s_0 a_0 s_1 a_1 \ldots a_{t-1} s_t$ for $t \leq T$.
- Expanded transition probability

$$\mathcal{P}^+(s', h' | a, s, h) = \mathbb{1}(h' = has') \int \mathcal{P}(s' | s, a) p(\mathcal{P} | h) d\mathcal{P}$$

  where $p(\mathcal{P}|h) \propto p(\mathcal{P}) p(h|\mathcal{P})$ is the posterior under prior p.
- Expanded reward function

$$r^+(s, h, a) = r(s, a)$$

This expanded MDP is called the Bayes-Adaptive MDP (BAMDP).

Bayes-optimal strategies

- The optimal policy of the BAMDP is also the optimal policy,

$$\arg\max_{\pi} \mathbb{E}_{M\sim p}\left[\mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t) \,\middle|\, M\right]\right]$$

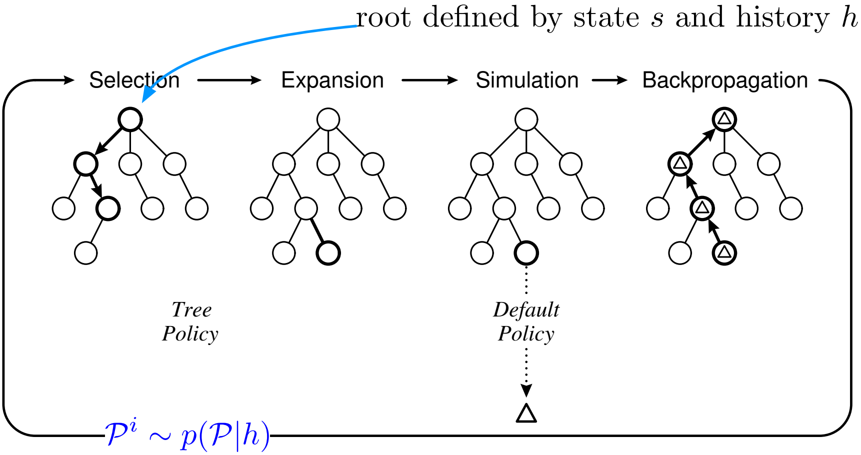
- The BAMDP construction is an application of a classical technique in partially observed MDPs.
- Could always do planning in the BAMDP to get the optimal policy, but this requires Bayesian inference at every node of the search tree.
- Guez et al. (2013) provide a more efficient MCTS method for approximating this Bayes-optimal policy.

BA-UCT (Guez et al., 2013)

BA-UCT is a MCTS method for approximating the Bayes-optimal policy.

1. Starting from the root in state s with history h.
2. For simulation $i = 1, \ldots,$
   - Sample $\mathcal{P}^i \sim p(\mathcal{P}|h)$.
   - Run one simulation of UCT with $\mathcal{P}^i$.
   - Share estimates of $Q^*(s, h, a)$ between simulations.
3. Return best action a according to current UCT estimates of $Q^*(s, h, a)$.
4. Get next state s' and update history h'.

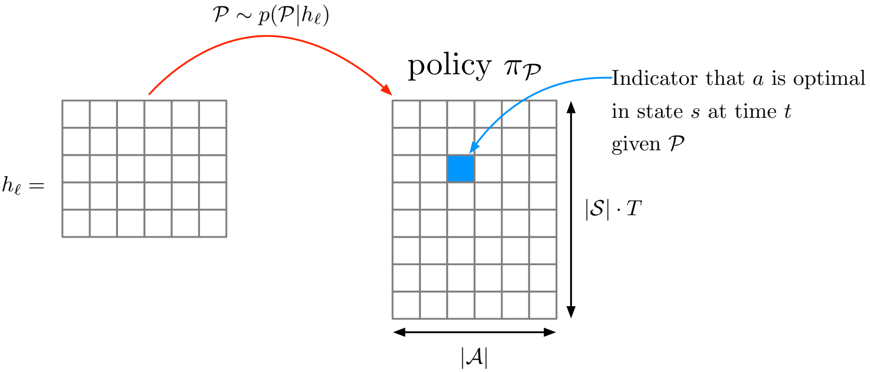
BA-UCT (Guez et al., 2013)



Thompson Sampling

- Convergence can be slow for this algorithm.
- Thompson sampling is a heuristic approximation to Bayes-optimal strategies that converges faster.
  - Given a history h_ℓ (from ℓ episodes).
  - For episode $\ell + 1$, use policy $\pi_{\mathcal{P}}$, which is defined as following optimal actions under a random MDP $\mathcal{P} \sim p(\mathcal{P}|h_\ell)$.
- This can be implemented using Bayesian posterior updating to keep track of $p(\mathcal{P}|h_\ell)$

Thompson sampling



Thompson sampling—example

Let's consider a detailed Beta-Bernoulli Bandit example.

- M is a K arms Bernoulli bandit problem.
  - Pulling arm k on round ℓ returns a Bernoulli reward $r^\ell(k) \sim \mathrm{Bern}(\theta_k)$ where $r^\ell(k) \in \{0,1\}, \theta_k \in (0,1)$.
  - M is fully defined by θ_k values.
- Agent has a Beta prior over $\theta_k \sim \mathrm{Beta}(\alpha_k, \beta_k)$ where $\alpha_k, \beta_k > 0$.

$$p(\theta_k) \propto \theta_k^{\alpha_k - 1}(1-\theta_k)^{\beta_k - 1}$$

- In this model, Bayesian updating has a simple form.

Thompson sampling—example

- In the Beta-Bernoulli Bandit model, Bayesian updating has a simple form.
- The posterior over θ_ks is itself a Beta, after observing rewards.
- Let $\alpha_k^\ell, \beta_k^\ell$ be the parameters of the posterior Beta after observing $r^1, \ldots r^\ell$, then if we pull arm k on round $\ell + 1$, we get the following update

$$\alpha_k^{\ell+1}, \beta_k^{\ell+1}) \leftarrow (\alpha_k^\ell + r^\ell, \beta_k^\ell + 1 - r^\ell)$$
$$\alpha_j^{\ell+1}, \beta_j^{\ell+1}) \leftarrow (\alpha_j^\ell, \beta_j^\ell) \text{ for } j \neq k$$

Thompson sampling—example

- In this case, Thompson sampling is: on round ℓ
  - Sample $\theta_k \sim \mathrm{Beta}(\alpha_k^{\ell-1}, \beta_k^{\ell-1})$ from the current posterior.
  - Pull arm $k^* = \arg\max_k \theta_k$ and observe reward $r^\ell(k^*)$.
  - Apply Bayesian posterior updates to get $\alpha_k^\ell, \beta_k^\ell$.
- Can compare this to a greedy approach that has all the same updating, but uses $\hat{\theta}_k = \mathbb{E}[\theta_k]$ to decide which arm to pull.

Thompson sampling—example

Algorithm 3.1 BernGreedy(K, α, β)

for $t = 1, 2, \ldots$ **do**
  #estimate model:
  for $k = 1, \ldots, K$ **do**
    $\hat{\theta}_k \leftarrow \alpha_k / (\alpha_k + \beta_k)$
  end for

  #select and apply action:
  $x_t \leftarrow \text{argmax}_k \, \hat{\theta}_k$
  Apply x_t and observe r_t

  #update distribution:
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t} + r_t, \beta_{x_t} + 1 - r_t)$
end for

Algorithm 3.2 BernTS(K, α, β)

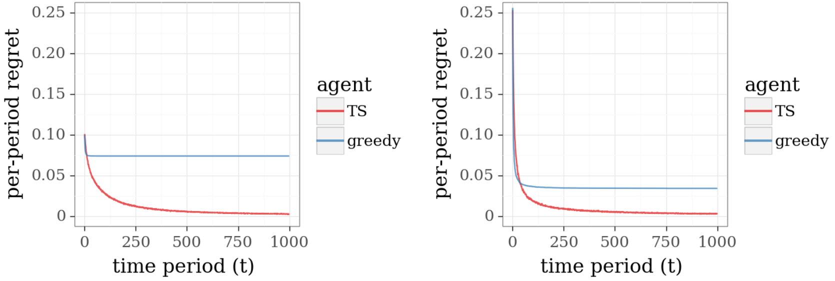
for $t = 1, 2, \ldots$ **do**
  #sample model:
  for $k = 1, \ldots, K$ **do**
    Sample $\hat{\theta}_k \sim \text{beta}(\alpha_k, \beta_k)$
  end for

  #select and apply action:
  $x_t \leftarrow \text{argmax}_k \, \hat{\theta}_k$
  Apply x_t and observe r_t

  #update distribution:
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t} + r_t, \beta_{x_t} + 1 - r_t)$
end for

(Russo et al., 2020)

Thompson sampling—example



(a) $\theta = (0.9, 0.8, 0.7)$ **(b)** average over random θ

Figure 3.2: Regret from applying greedy and Thompson sampling algorithms to the three-armed Bernoulli bandit.

(Russo et al., 2020)

Thompson sampling—example

- Why is Thompson sampling better than greedy?
- Greedy can get stuck, but even if we allow it to take a uniform random action w.p. ϵ, it will still ignore uncertainty.

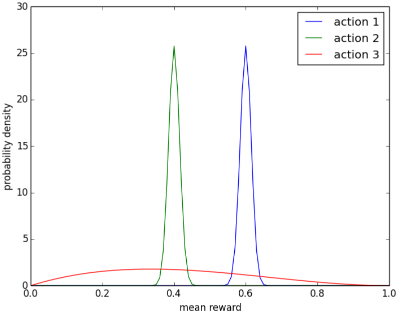


Figure 2.2: Probability density functions over mean rewards.

(Russo et al., 2020)

Recap

- We can formalize the notion of exploration by considering uncertainty over the MDP M.
- This gives us a natural class of algorithms that update their posterior beliefs about the MDP specification after observing state-action histories.
- Despite having a posterior, the RL as inference that we've been seeing a lot of (VIREL, MPO, SAC, etc.) is very distinct from this view on Bayesian RL.

Next week

- Next week Brendan O'Donoghue will talk about K-learning and a variational inference perspective on Bayesian RL.