Bayesian RL

Bayesian Decision Making

- In most part of this course we've taken a frequentist view of decision-making under uncertainty
  - e.g., the sample complexity guarantees we give in the exploration section are also worst-case bounds
  - that is, regardless of how nature picks the problem instance from a predetermined family (e.g., all MDPs whose state space is S)—possibly in an adversarial manner—the guarantee always holds
- The alternative: Bayesian RL
  - assume some *prior* over problem instances
  - use data to update the *posterior* according to Bayes rule

Review: Bayesian estimation of the bias of a coin

- Suppose we have a coin with unknown bias θ
- Want to estimate θ from i.i.d. coin tosses $X_1, \ldots, X_n$
- Frequentist approach/analysis: $\hat{\theta} = \frac{1}{n}\sum_i X_i$; can bound $|\theta - \hat{\theta}|$ by Hoeffding's regardless of what value θ takes
  - worst-case over all Bernoulli distributions with $\theta \in [0,1]$
  - Fix θ, the distribution of X_i is well-defined, but there is no such thing as "distribution of θ"

Review: Bayesian estimation of the bias of a coin

- Suppose we have a coin with unknown bias θ
- Want to estimate θ from i.i.d. coin tosses $X_1, \ldots, X_n$
- Bayesian approach
  - First, pick a prior, which is a distribution over θ
  - Often pick beta distribution (conjugate to Bernoulli) $\theta \sim p = beta(a, b)$, where a and b represents belief in prior
  - Use data to compute posterior: $q(\theta) \propto p(\theta) \Pr[X_{1:n} | \theta]$
  - In the special case here, the update is easy: q is still a beta, but you add #heads to a and #tails to b

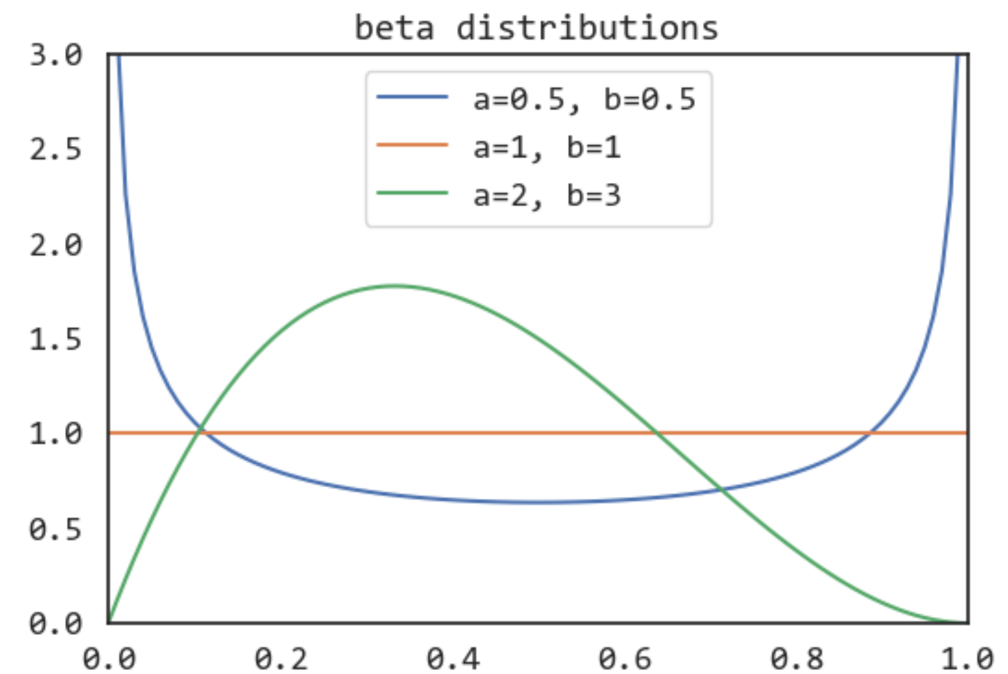

fig from: https://towardsdatascience.com/dirichlet-distribution-a82ab942a879

From Bayesian Prediction to Decision-making

- The bayesian stuffs you learn from a standard ML class is about prediction
- You get a posterior over the true world, which is often not what you want (e.g., we may want point estimates or confidence intervals)
- You are told to induce the quantities of interest from the posterior in anyway you want—there is no unique answer to how you do this
- In Bayesian decision-making, there is always a well-defined notion of optimal decision-making
  - e.g., in the exploration-exploitation setting, we will see that Bayesian optimality is well-defined with an interesting connection to POMDPs

Bayesian Multi-armed Bandits

- Consider a multi-armed bandit, where the reward of each arm follows a Bernoulli distribution with unknown parameter θ_i (for $i=1, \dots, K$, where K is the number of arms)
- In the Bayesian setting, we need to pick a prior p for $\{\theta_i\}_{i=1,\dots,K}$
  - For simplicity, let's say each θ_i follows an i.i.d. beta
  - Here i.i.d.ness of $\{\theta_i\}$ implies that data from one arm will not be used to update the posterior of any other arm (i.e., no generalization)
- (Bayesian) Metric for the algorithm's performance
  - Suppose algorithm interacts with the env for T rounds
  - In round t, the algorithm gets reward r_t
  - Metric: $\mathbb{E}_{\{\theta_i\}\sim p}\left[\sum_{t=1}^{T} r_t \,\middle|\, \textit{exec alg in problem instance } \{\theta_i\}\right]$
- What is the optimal value and what is an algorithm that achieves it?

Define Bayesian Optimality

- Key result: The Bayesian optimal value and algorithm are defined by the optimal value and policy in a belief MDP (sometimes also called Bayesian Adaptive MDP/POMDP)
- Defining the belief MDP
    - State space: the space of possible posteriors q over $\{\theta_i\}$ (sometimes also called an *information state*)
    - Action space: same as the original problem (K arms)
    - Reward function: $R(q, a) = \mathbb{E}_{\{\theta_i\} \sim q}[\theta_a]$
    - Transition function: (defined via a generative process) when we take action a in state q, we transition to q' as: $\{\theta_i\} \sim q, \quad r \sim Ber(\theta_a), \quad q' = BeliefUpdate(q, a, r)$
    - Horizon is T (finite-horizon, undiscounted)
- Claim: optimal policy in this MDP (which maps (belief, time-step) to actions) is an algorithm that achieves Bayes optimality

Compare the original vs the Bayesian problems

- Learning vs planning
  - Original: learning under uncertainty (model unknown)
  - Bayesian RL: *planning* with fully known transition model
- Horizon
  - Original: one-shot decision making (bandits)
  - Bayesian RL: sequential decision-making with (extremely) long horizon T
- Algorithm style for exploration-exploitation
  - Original: the metric requires to balance exploration and exploitation
  - Bayesian RL: no need to explicit balance exp-exp. The optimal policy balances exp-exp optimally (by definition)!

Challenges in Bayesian RL & Practical Algorithms

- Solving the belief MDP is computationally very challenging
  - State space is too large and complex (all posteriors)
  - Horizon is extremely long
- Practical heuristic algorithm: Thompson sampling
  - Extremely simple: given posterior q, sample a problem instance from q, and make decisions greedily w.r.t. the sampled instance!
  - Automatically balance exp-exp
  - No hyperparameters (apart from the prior)
- Practical meta-level algorithm: MCTS
  - Simplified case: use Monte-Carlo control for one-step policy improvement (over a heuristic algorithm)
  - Computation: $O(T) \Rightarrow O(nT^2)$ where n is the number of simulations run in each real time step

Further comments

- We consider a MAB here, but the way to handle an MDP or even POMDP (say with finite horizon H) is very similar
    - The corresponding Bayesian-Adaptive MDP (BAMDP) has a horizon of $HT,$ where $T,$ is the total number of episodes
    - The state in the BAMDP is (original state, posterior over MDP family)
    - Exercise: define the BAMDP yourself
- Besides computation, another issue is the choice of prior
    - (Die-hard Bayesians will tell you prior is “never wrong”)
    - Similar to the choice of function approximation in the frequentist approach (bias-variance trade-off)
    - For MDPs, a popular choice is i.i.d. Dirichlet for each P(.|s,a) => Bayesian version of “tabular RL”
    - Another limitation of Bayesian RL: must be model-based almost by definition