Bayesian Reinforcement Learning

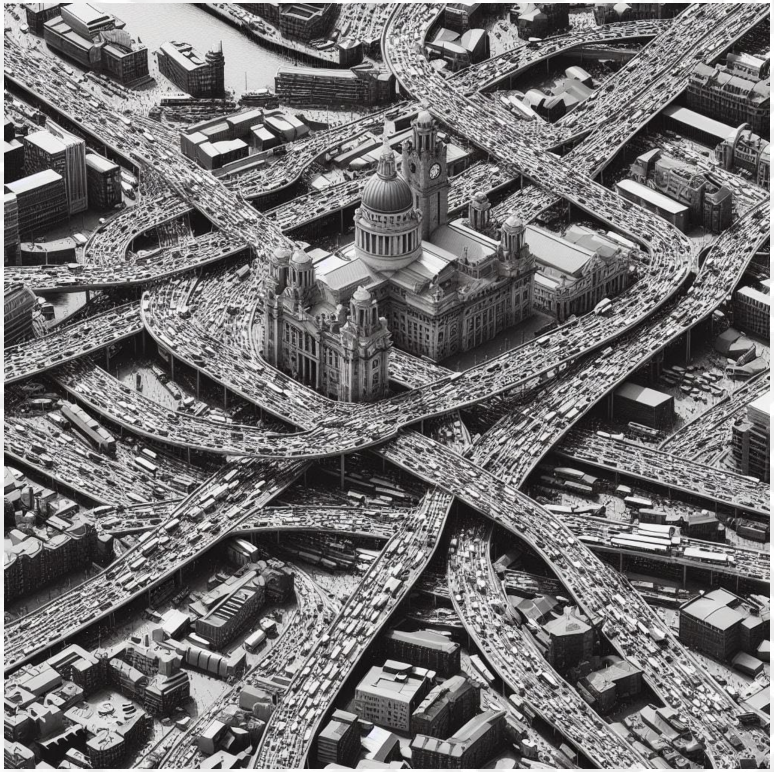
(or Posterior Sampling for RL)

University of Liverpool

Distributed Algorithms CDT Showcase 2023

Malcolm Strens

AI is moving towards *multi-step decision- making* under uncertainty…



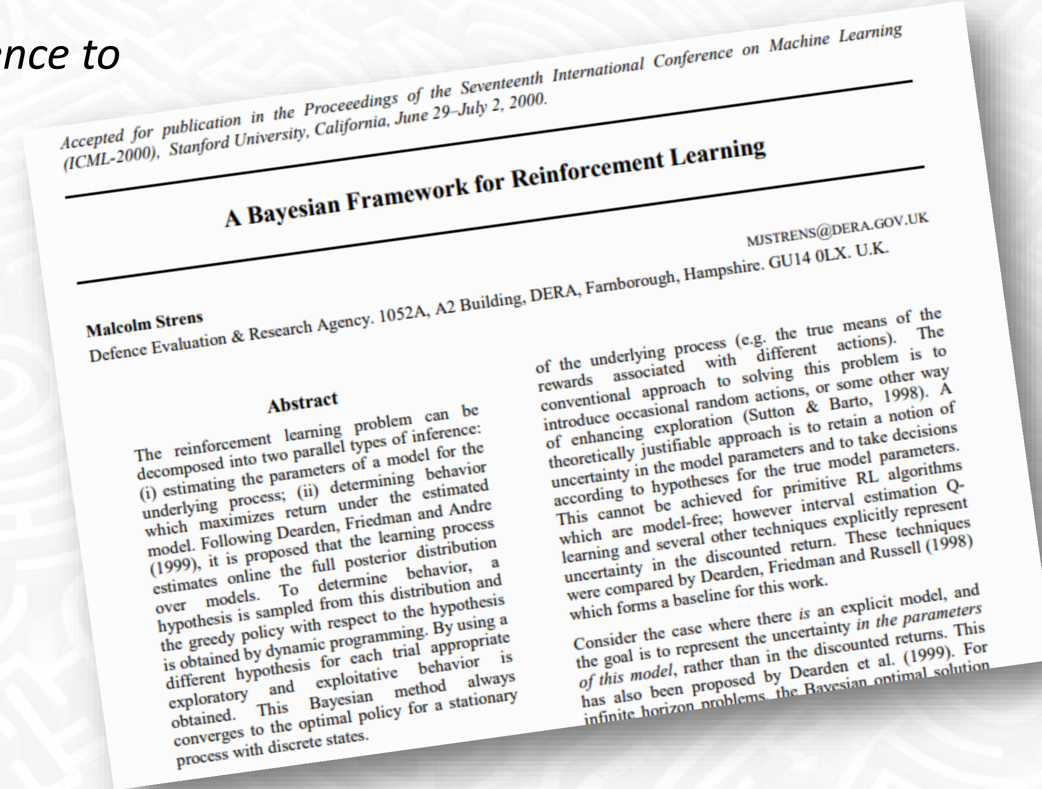
"Sequential Decision Problems"

"Optimal Control"

"Reinforcement Learning"

Q. Can we apply Bayesian Inference to sequential decision-making?

A. Yes, if we can express uncertainty over RL models (of interaction and delayed reward).



Accepted for publication in the Proceeedings of the Seventeenth International Conference on Machine Learning (ICML-2000), Stanford University, California, June 29–July 2, 2000.

A Bayesian Framework for Reinforcement Learning

Malcolm Strens MJSTRENS@DERA.GOV.UK
Defence Evaluation & Research Agency. 1052A, A2 Building, DERA, Farnborough, Hampshire. GU14 0LX. U.K.

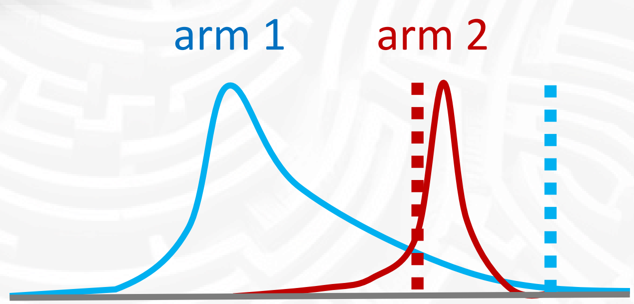
Abstract

The reinforcement learning problem can be decomposed into two parallel types of inference: (i) estimating the parameters of a model for the underlying process; (ii) determining behavior which maximizes return under the estimated model. Following Dearden, Friedman and Andre (1999), it is proposed that the learning process estimates online the full posterior distribution over models. To determine behavior, a hypothesis is sampled from this distribution and the greedy policy with respect to the hypothesis is obtained by dynamic programming. By using a different hypothesis for each trial appropriate exploratory and exploitative behavior is obtained. This Bayesian method always converges to the optimal policy for a stationary process with discrete states.

of the underlying process (e.g. the true means of the rewards associated with different actions). The conventional approach to solving this problem is to introduce occasional random actions, or some other way of enhancing exploration (Sutton & Barto, 1998). A theoretically justifiable approach is to retain a notion of uncertainty in the model parameters and to take decisions according to hypotheses for the true model parameters. This cannot be achieved for primitive RL algorithms which are model-free; however interval estimation Q-learning and several other techniques explicitly represent uncertainty in the discounted return. These techniques were compared by Dearden, Friedman and Russell (1998) which forms a baseline for this work.

Consider the case where there *is* an explicit model, and the goal is to represent the uncertainty *in the parameters of this model*, rather than in the discounted returns. This has also been proposed by Dearden et al. (1999). For infinite horizon problems, the Bayesian optimal solution

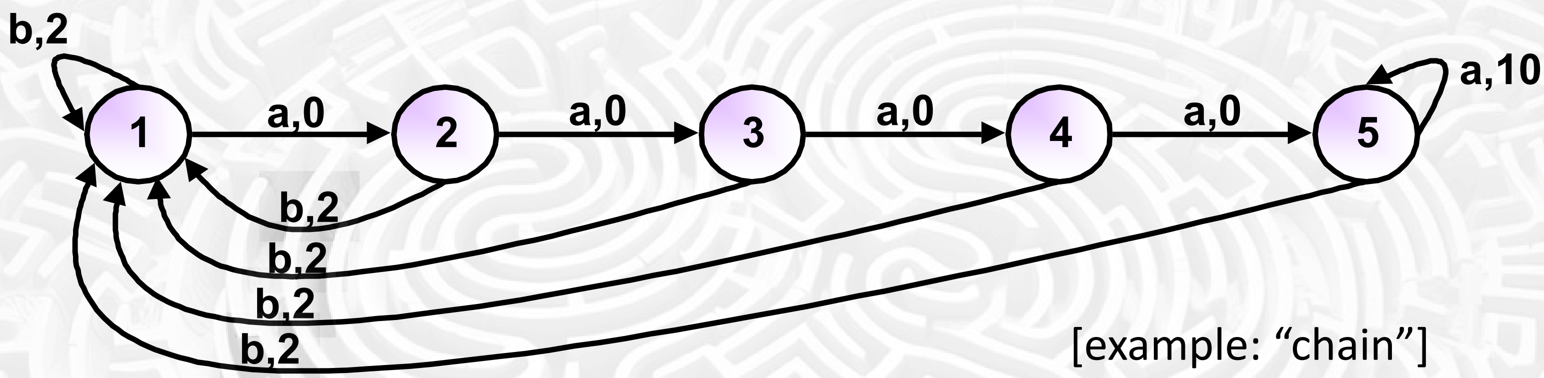
Builds on Thompson Sampling (1933):
for repeated state-free decision
e.g. 2-armed bandit



posteriors on expected payoffs

Optimal: "select the best action under a single sample from the Bayesian posterior"
i.e. P(E arm 1 payoff > E arm 2 payoff | trials so far)

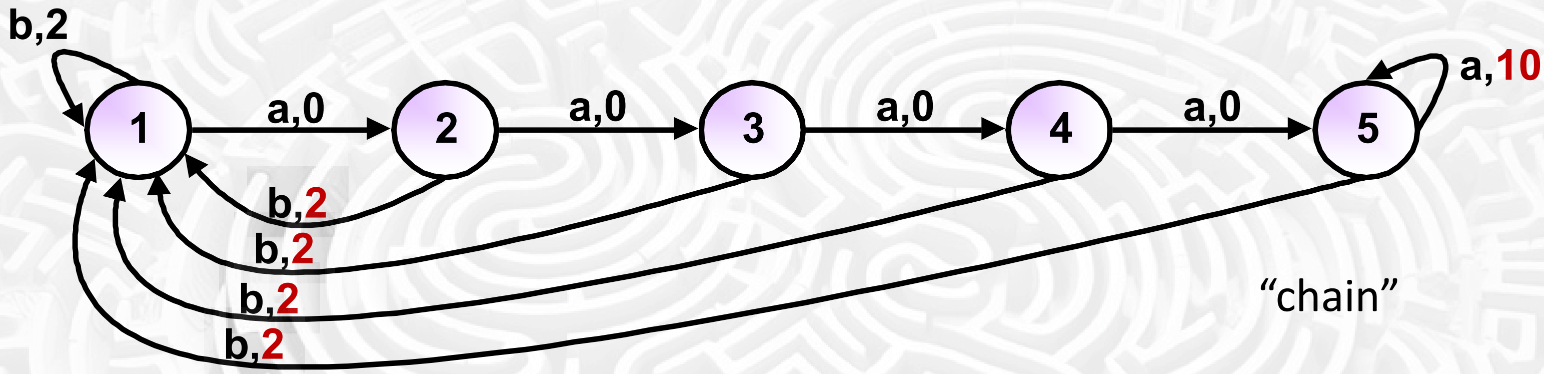
RL target for inference: Markov Decision Process



[example: "chain"]

- (S,A,T,R)
- S: Set of States
- A: Set of Actions
- T: Transition Probabilities $T(s,a,s')$ Multinomial posterior (Dirichilet conjugate prior)
- R: Reward Distributions $R(s,a)$ Gaussian posterior (Normal-Gamma conjugate prior)

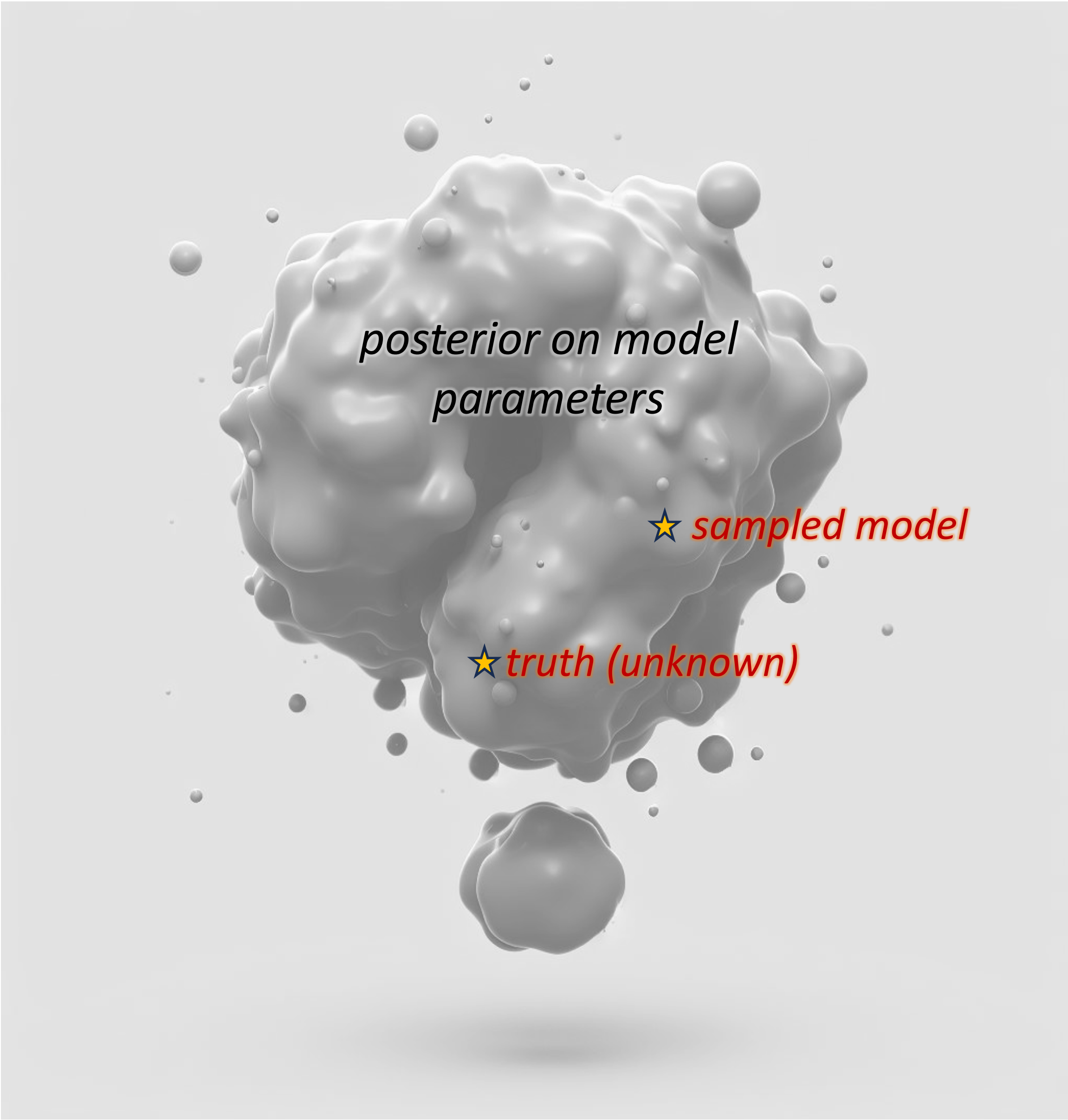
Exploration/exploitation dilemma



- Trade-off 'value of information' from exploration vs regret of not exploiting information already known.

- A generalisation of Thompson sampling to multi-step problems achieves this.

Bayesian RL on "models of interaction" like MDPs



Start a new interaction ("episode").

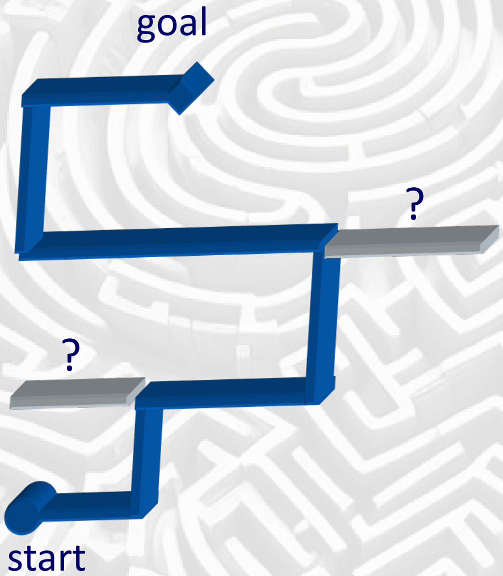
Sample a model for the world (e.g. MDP) from current posterior.

Solve the world model for optimal behaviour; *e.g.* Bellman backup to obtain control "policy".

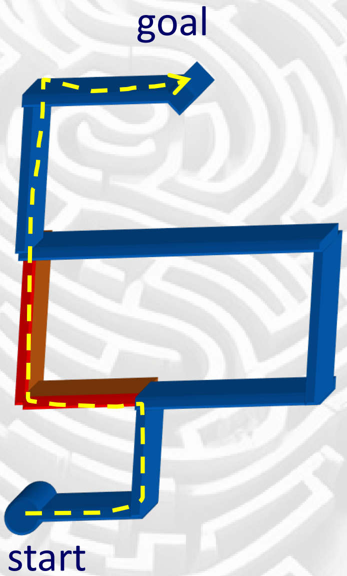
Take actions according to this policy until the end of the episode.

Use the collected transitions & rewards to update the posterior.

e.g. navigation while learning map



posterior represents uncertainty in map based on experience



drawing a hypothesis from the posterior, then solve for shortest path … yields *exploratory* behaviour!

(then update posterior)

Developments/applications in PSRL

- 2000: formulated for RL
- 2005: modelling cognition [Daw/Dayan]
- 2011+: medical applications
- 2012: MCTS for large models/games [Guez/Silver/Dayan]
- 2013 & 2017: regret bounds & outperformance … best for any RL algorithm [Osband/Russo/Van Roy]
- 2019: multi-agent
- 2021: partially observable tasks
- 2023: Langevin Thompson Sampling (MCMC)
- 2024: complex neural world models?

Questions?

Optimal Policy from Dynamic Programming

- Can apply to:
  - true MDP (not known during learning)
  - max likelihood MDP (changes during learning)
  - **hypothesised MDP**
- DP backup on an estimated MDP:

$$\hat{Q}(s,a) \leftarrow \hat{R}(s,a) + \gamma \sum_{s' \in S} \hat{T}(s,a,s') . \max_{a'} \hat{Q}(s',a')$$

Expected discounted reward for *a* in *s*

Expected immediate reward for *a* in *s*

Sum over possible transitions (discounted)

Best expected discounted reward in successor state