Model-based Bayesian Reinforcement Learning with Adaptive State Aggregation

Cosmin Paduraru
joint work with Arthur Guez, Doina Precup and Joelle Pineau

Reasoning and Learning Lab
McGill University

Starting point

- Learning MDP models can be useful
- Learning a distribution over models can be more useful than learning a single model (helps exploration)
- We want to handle continuous state spaces

Bayesian learning for MDP models

Given a history $h_t = s_0, a_0, r_1, s_1, ...s_t, a_t,$ we have

$$P(M|h_t, r_{t+1}, s_{t+1}) = \frac{P(s_{t+1}, r_{t+1}|M, h_t)P(M|h_t)}{P(s_{t+1}, r_{t+1}|h_t)}$$

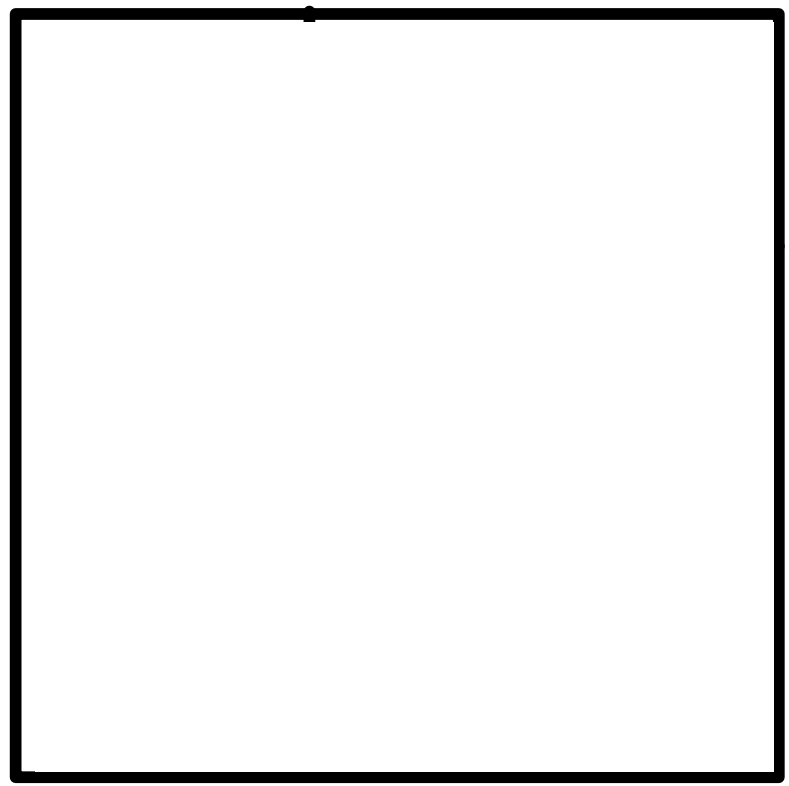
Posterior

Likelihood

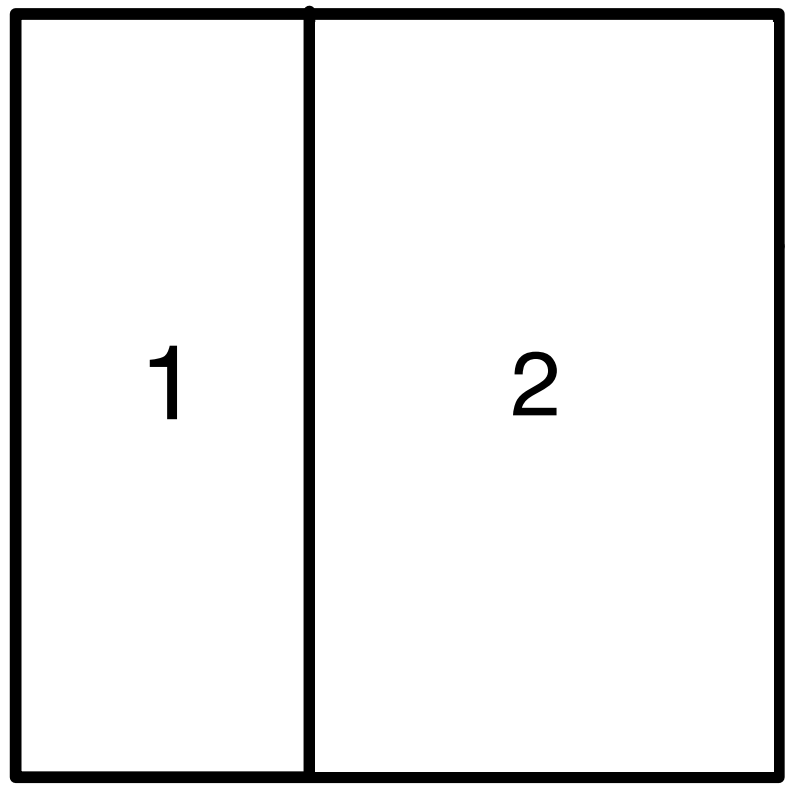
Normalization factor
(requires integration over all models)

Prior
(I know..)

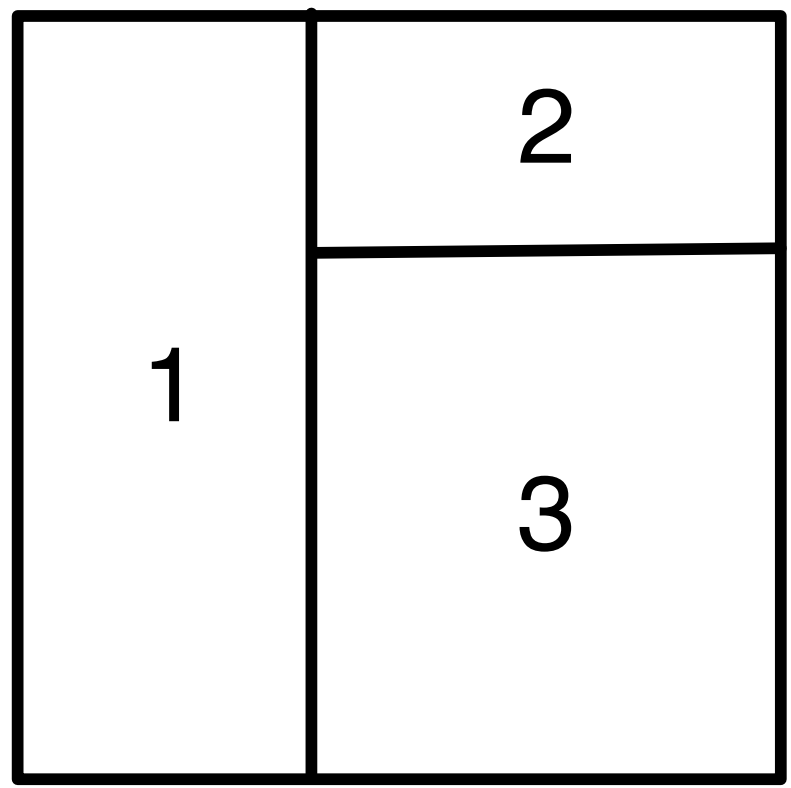
Example: state aggregation



Example: state aggregation



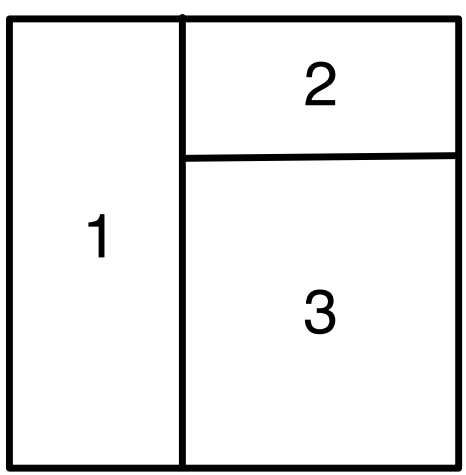
Example: state aggregation



can easily construct a partition-to-partition transition matrix

Idea #1

- Separate out everything that can be analytically updated
- Given a discretization of the state space, the probabilities of transitioning between partitions can be updated analytically



$P(x_{t+1} = \cdot | x_t = 1, a)$ is a multinomial

- can use a Dirichlet prior for closed form updates

What about the distribution over discretizations?

Idea #2

- Maintain an approximate distribution over structures by sampling from the posterior
- Sampling can be performed using Markov Chain Monte Carlo (MCMC) methods
  - propose changes to the discretization (e.g. splits, merges)
  - accept the changes with prob. proportional to likelihood of stored data
- If re-sampling is too expensive, re-weight existing set of discretizations

Idea #3

- Using the likelihood of stored data for model selection does not mean overfitting!

Given a partition D whose quality we want to evaluate:

Overfitting

1. use stored data to estimate transition matrix
2. compute likelihood of estimated matrix on the same data

Incremental (Bayesian) likelihood

1. use data up to s_t to estimate transition matrix
2. compute likelihood of estimated matrix on s_{t+1}
3. do this for all t and multiply results

Idea #4

- The posterior distribution over models can be used for optimally trading off between exploration and exploitation in any environment

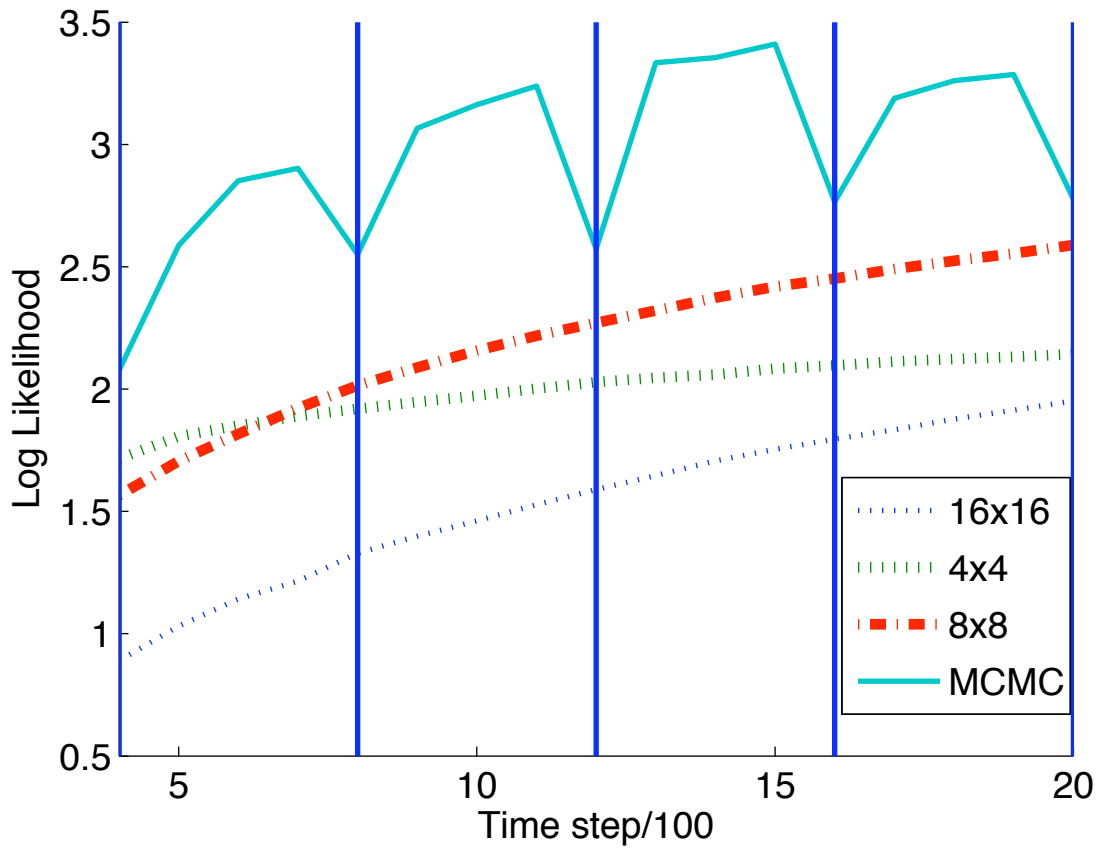
Idea #4

- The posterior distribution over models can be used for optimally trading off between exploration and exploitation in any environment.. NOT!

Idea #4

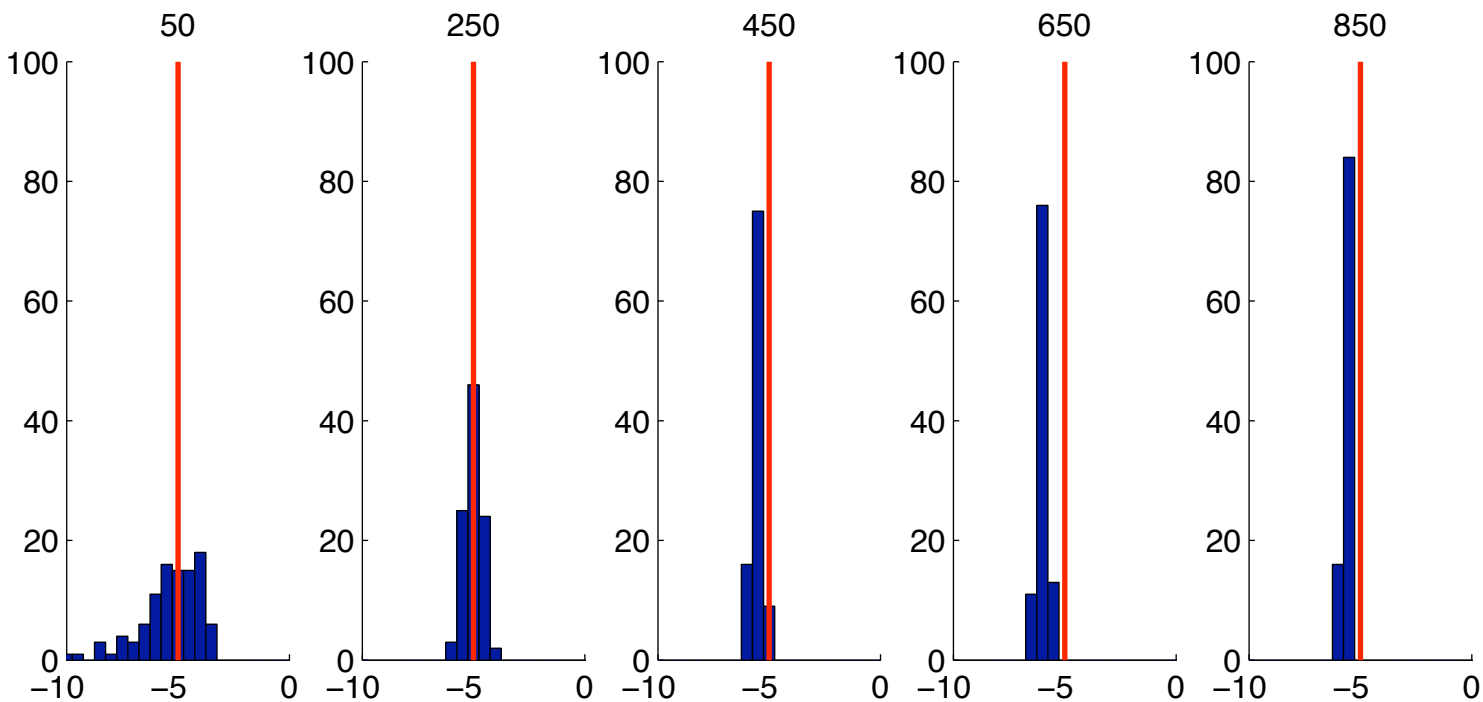
- The posterior distribution over models can be used for optimally trading off between exploration and exploitation in any environment.. NOT!
- Even given the correct posterior, one would need to solve continuous high dimensional POMDPs to find the optimal trade-off
- The hope is that approximate POMDP solutions are still good enough
- Myopic exploration heuristics could still be used (e.g., value of perfect information)

Average model likelihood



Continuous stochastic navigation domain, random policy
No obstacles, some varying dynamics

Value function distribution



estimated distribution of $V^{\pi^*}(s_0)$
after various amounts of experience

Summary

- Separate the posterior over models into
  - posterior over discretizations (updated implicitly by re-sampling)
  - posterior over transition matrices given a discretization (updated analytically)
- Cross-validation style measure for likelihood
- Need to explore possibilities for using the resulting distribution for exploration