Exploration

2017/03/20

John Schulman

What is the **exploration** problem?

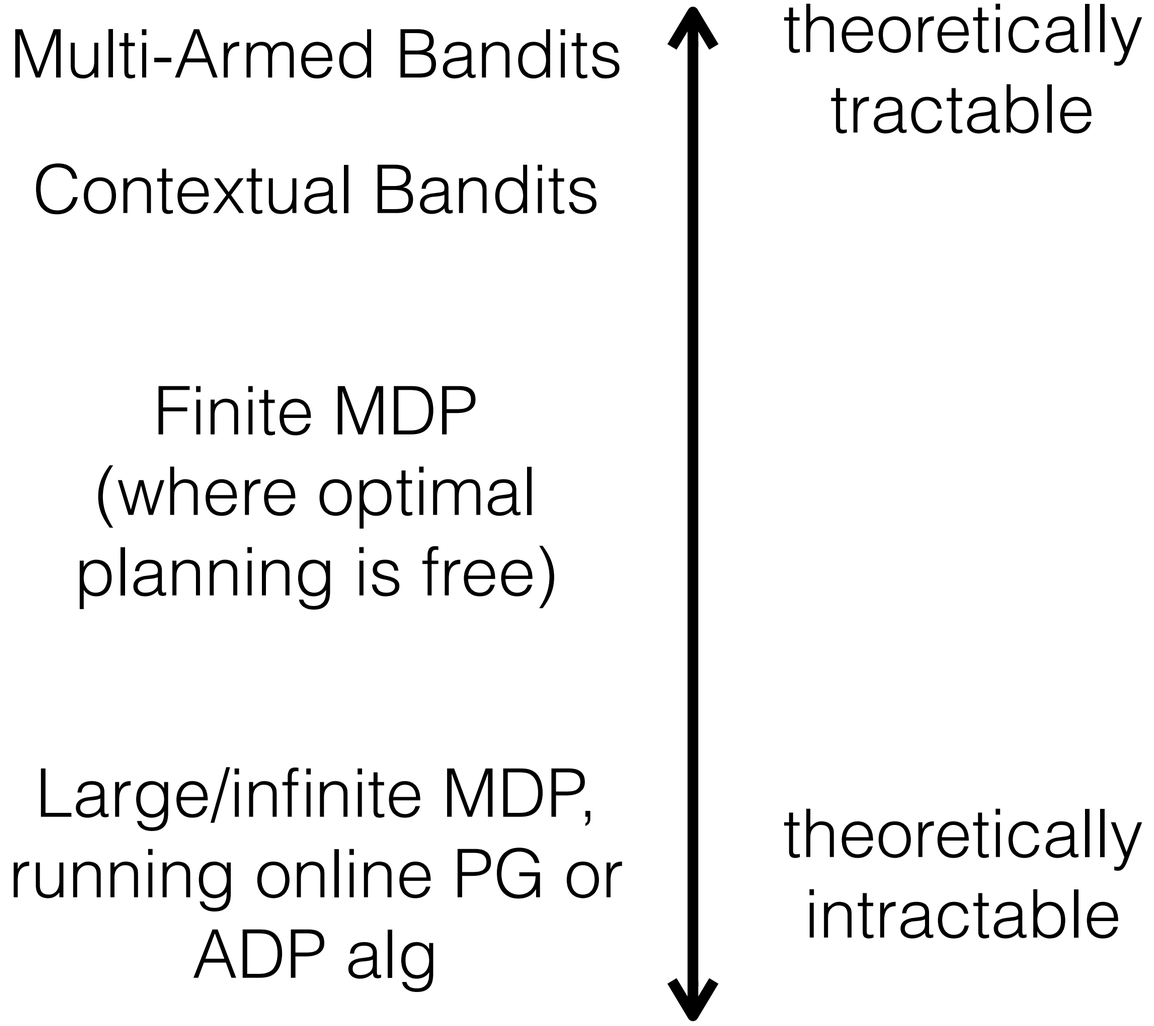
- Given a long-lived agent (or long-running learning algorithm), how to balance exploration and exploitation to maximize long-term rewards
- (Informal) How to search through the space of possible strategies of the agent to avoid getting stuck in local optima of behavior

Exploration vs Exploitation

- Restaurant Selection
  - Exploitation Go to your favourite restaurant
  - Exploration Try a new restaurant
- Online Banner Advertisements
  - Exploitation Show the most successful advert
  - Exploration Show a different advert
- Oil Drilling
  - Exploitation Drill at the best known location
  - Exploration Drill at a new location
- Game Playing
  - Exploitation Play the move you believe is best
  - Exploration Play an experimental move

From Dave Silver's lecture notes
http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/XX.pdf

Problem Settings



Problem Settings

Multi-Armed Bandits

Contextual Bandits

Themes:
- Use optimistic value estimates
- Posterior (Thompson) sampling

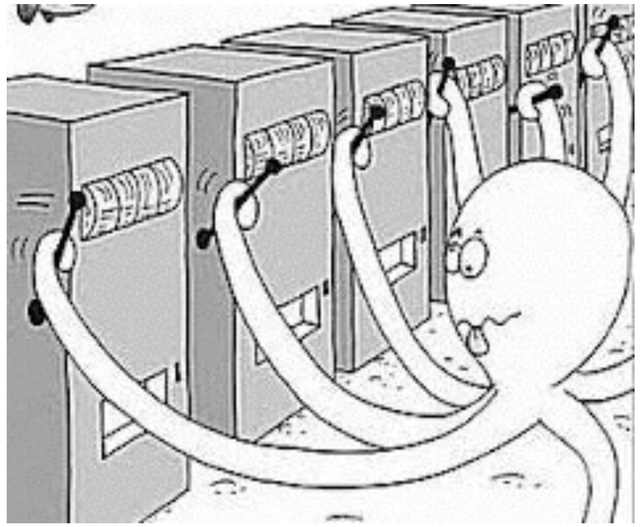
Finite MDP
(where optimal
planning is free)

Themes:
- Optimistic dynamics model
- Exploration bonuses

Large/infinite MDP,
running online PG or
ADP alg

Themes:
- Optimistic dynamics model
- Optimistic values
- Thompson sampling
- *Intrinsic rewards / intrinsic motivation*

Part I: Bandit Problems

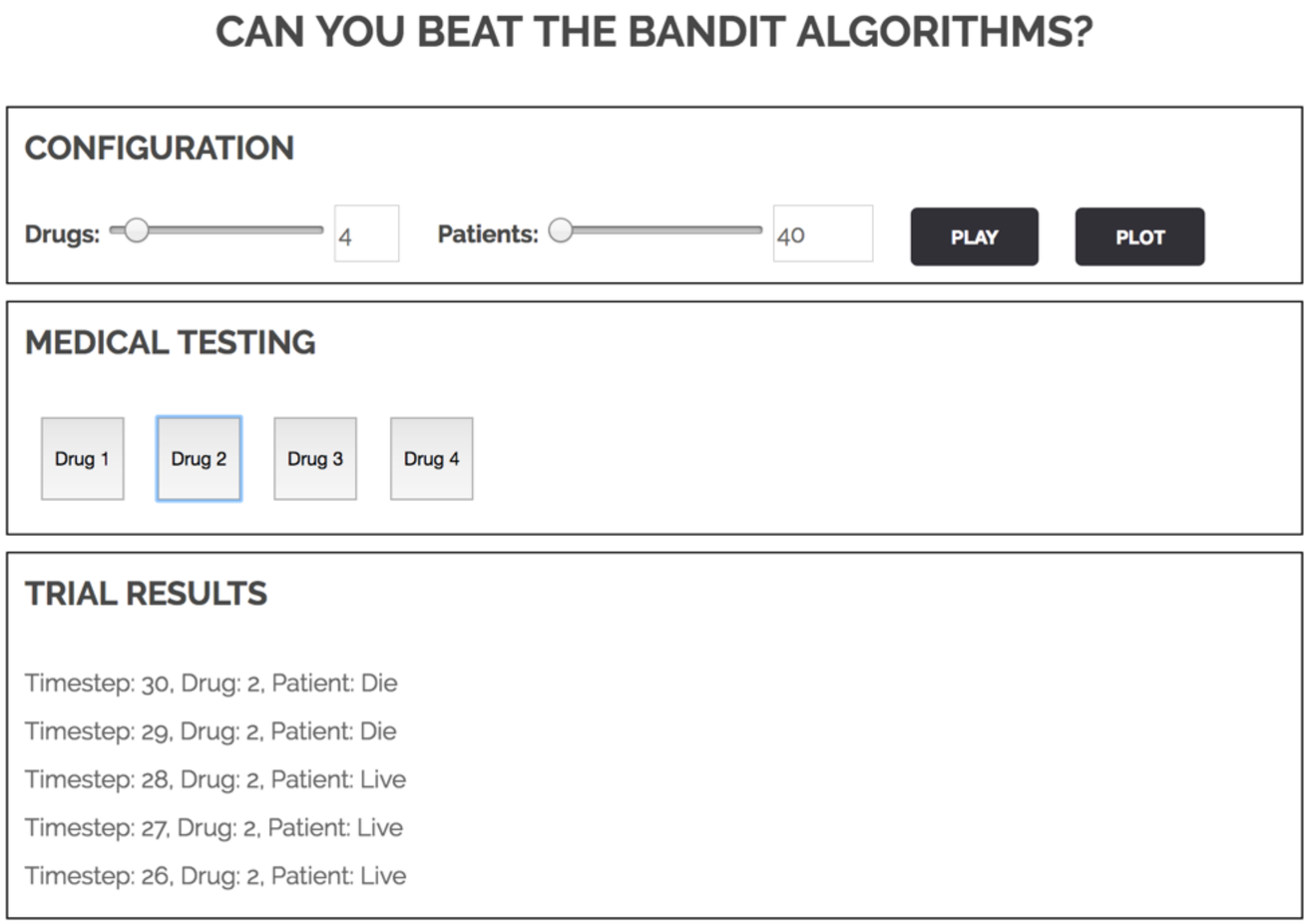


“bandit” = slot machine
pick the best one

Bandit Problems

- k arms, n rounds, $n \geq k$
- Unknown: probability distributions $p(R \mid a)$ for each action
- For t = 1,2, …
  - agent chooses $a_t \in \{1,2,\ldots,k\}$
  - environment provides reward R_t according to $p(R \mid a)$
  - Let $Q(a) = E[R \mid a]$
- Goal: maximize cumulative reward, equivalently, minimize regret
  - $\text{Regret}_n := \Sigma_t \ (Q^* - Q(a_t))$

Review: Bubeck, Sébastien, and Nicolo Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems." arXiv preprint arXiv:1204.5721 (2012).



http://iosband.github.io/2015/07/28/Beat-the-bandit.html

Bandit Problem As a POMDP

- Let's say arm parameters are randomly sampled, then bandit problem is a POMDP
  - State = history of all actions and rewards so far
  - Bandit isn't stateful, but our belief state is
- Gittins index theorem: if arm parameters are sampled independently, optimal strategy for maximizing discounted return is to always maximize Gittins index, which is computed independently for each arm
  - Limitations: no guarantees for misspecified priors, doesn't extend to other settings (contextual bandits, MDPs)

Optimism: UCB-style algorithms

- “Upper Confidence Bound”, not UC Berkeley unfortunately
- Pick the arm that maximizes *mean + stdev * very_slow_growing_factor*
- I.e., arm with best return *if we’re a bit optimistic*
- Favor high expected return and high variance
- Logarithmic regret (which is optimal)

Peter Auer, Nicolò Cesa-Bianchi and Paul Fischer, Finite-Time Analysis of the Multi-Armed Bandit Problem, Mach. Learn., 47 (2002), 235–256

Probability Matching / Posterior Sampling

- Probability matching - pull lever with probability that it's the optimal one
- Posterior (Thompson) sampling - sample from posterior distribution over model, then choose optimal action according to that sample

Daniel Russo, Benjamin Van Roy (2014) Learning to Optimize via Posterior Sampling. Mathematics of Operations Research

Chapelle O. and Li, L. "An Empirical Evaluation of Thompson Sampling". NIPS, 2011

Information Gain

- “Information gain” (e.g., in Bayesian experiment design)
  - $\mathrm{I}(y_t$ [=next observation], θ [latent var. of interest])
    - $I(\theta, y_t) := E_{yt}[H(\theta) - H(\theta \mid y_t)] \qquad = E_{\theta}[H(y_t) - H(y_t \mid \theta)]$
    - Depends on action a_t, so we want $\max_{at} I(\theta, y_t \mid a_t)$
- Bandit alg: information-directed sampling [1]
  - $\min_{\pi}$ E[suboptimality]^2 / E[information gain]
  - sparse linear bandit example

[1] Daniel Russo, Benjamin Van Roy (2014) Learning to Optimize via Information-Directed Sampling. NIPS

Contextual Bandits

- Each timestep, we also get a "context" s_t and reward follows distribution $P(R \mid s_t, a_t)$
  - unlike in MDP, s_t does not depend on history
- For t = 1,2, …
  - environment provides context s_t
  - agent chooses $a_t \in \{1,2,\ldots,k\}$
  - environment provides reward R_t according to $p(R \mid a_t, s_t)$

Applications of (Contextual) Bandits

- *Originally considered by Allied scientists in World War II, it proved so intractable that, according to Peter Whittle, the problem was proposed to be dropped over Germany so that German scientists "could also waste their time on it” [1]*

- Ads and recommendation engines

[1] wikipedia, Multi-Arm Bandits

Part II: Finite MDPs, PAC Exploration

- How to define sample complexity / near optimal learning in a previously unknown MDP?
  - Kearns and Singh [1]: measure amount of time it takes until final policy is epsilon-suboptimal
  - Kakade [2]: number of timesteps where policy is epsilon-suboptimal

[1] Kearns & Singh, “Near-Optimal Reinforcement Learning in Polynomial Time” (1999)
[2] Kakade, “On the sample complexity of reinforcement learning” (thesis) (2003)

Optimistic Initial Model

- Make optimistic assumption about dynamics model of MDP and plan according to it
- Szita & Lorincz alg: Initially assume that every state-action pair has deterministic transition to “Garden of Eden State” with maximal reward. Also see R-MAX.

Szita, István, and András Lőrincz. "The many faces of optimism: a unifying approach." ICML 2008.

Moldovan, Teodor Mihai, and Pieter Abbeel. "Safe exploration in markov decision processes." arXiv preprint arXiv:1205.4810 (2012).

Optimistic Initial Value

- Initialize Q-values with large positive value
- Heuristic method inspired by OIM methods

Finite MDPs, PAC Exploration

Delayed Q-Learning with Interval Estimation

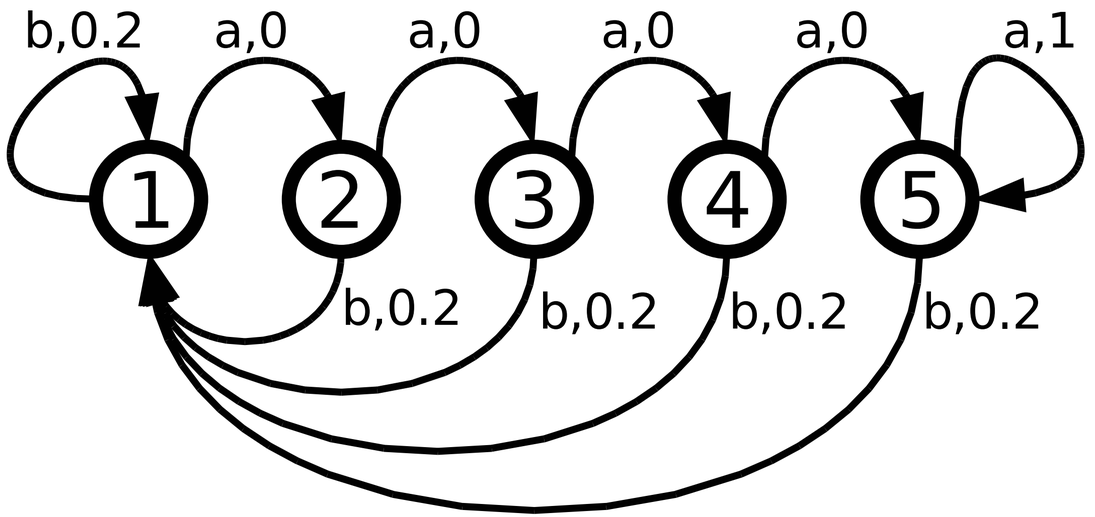
add exploration bonus to Q-values

All insufficiently visited states are highly rewarding

Summary Table			
Algorithm	Comp. Complexity	Space Complexity	Sample Complexity
Q-Learning	$O(\ln(A))$	$O(SA)$	Unknown, Possibly EXP
DQL	$O(\ln(A))$	$O(SA)$	$\tilde{O}\left(\frac{SA}{\epsilon^4(1-\gamma)^8}\right)$
DQL-IE	$O(\ln(A))$	$O(SA)$	$\tilde{O}\left(\frac{SA}{\epsilon^4(1-\gamma)^8}\right)$
RTDP-RMAX	$O(S+\ln(A))$	$O(S^2A)$	$\tilde{O}\left(\frac{S^2A}{\epsilon^3(1-\gamma)^6}\right)$
RTDP-IE	$O(S+\ln(A))$	$O(S^2A)$	$\tilde{O}\left(\frac{S^2A}{\epsilon^3(1-\gamma)^6}\right)$
RMAX	$O\left(\frac{SA(S+\ln(A))\ln\frac{1}{\epsilon(1-\gamma)}}{1-\gamma}\right)$	$O(S^2A)$	$\tilde{O}\left(\frac{S^2A}{\epsilon^3(1-\gamma)^6}\right)$
MBIE-EB	$O\left(\frac{SA(S+\ln(A))\ln\frac{1}{\epsilon(1-\gamma)}}{1-\gamma}\right)$	$O(S^2A)$	$\tilde{O}\left(\frac{S^2A}{\epsilon^3(1-\gamma)^6}\right)$

Strehl, PROBABLY APPROXIMATELY CORRECT (PAC) EXPLORATION IN REINFORCEMENT LEARNING, 2007

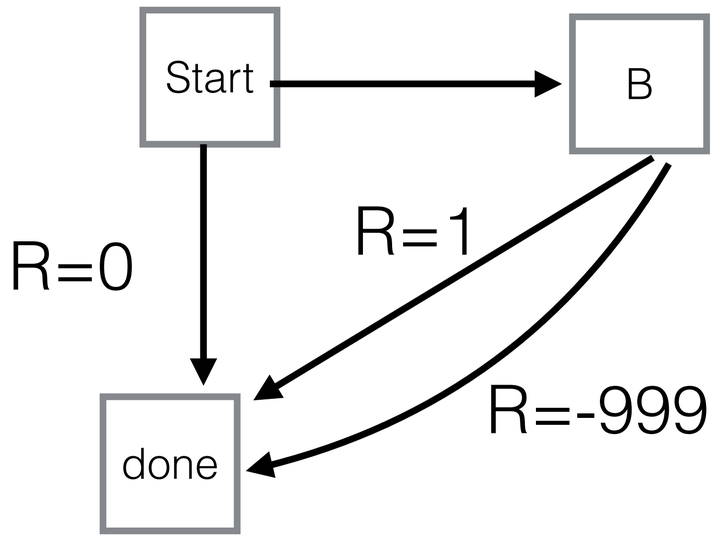
MDPs — examples



samples needed ~ 2^{Length}

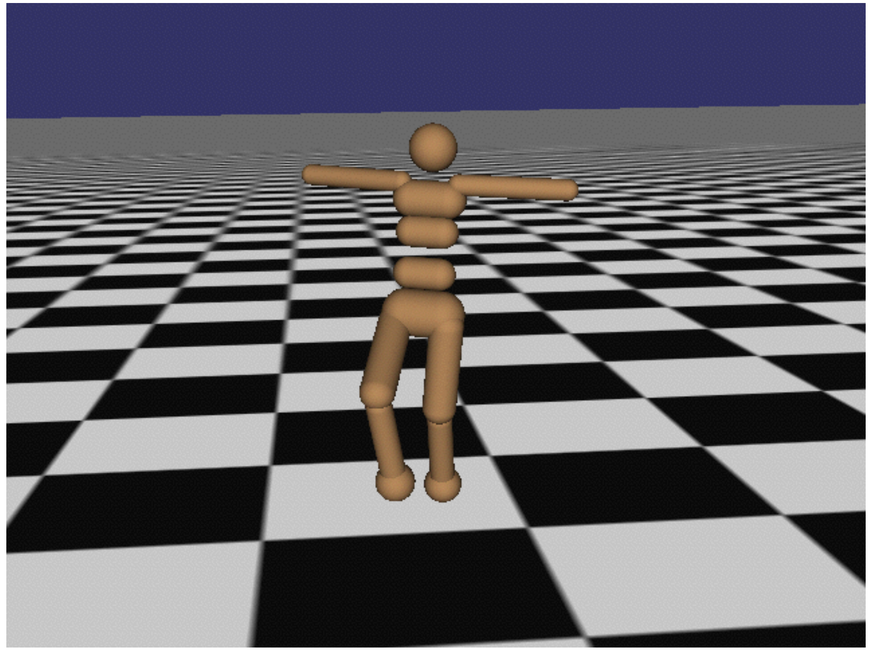
from Kolter & Ng, Near-Bayesian Exploration in Polynomial Time

MDPs — examples



problematic for
policy gradient methods

MDPs — examples



Local minima policies:

- lunge forward
- stand

MDPs — examples

Breakout

Local minima policies:

- Stay on one side

Part III: Exploration in Deep RL

Exploration in Deep RL

- Can’t optimally plan in the MDP, as was assumed by some prior algorithms
- Never reach the same state twice (need metric or some notion of “novelty”)

Posterior (Thompson) Sampling

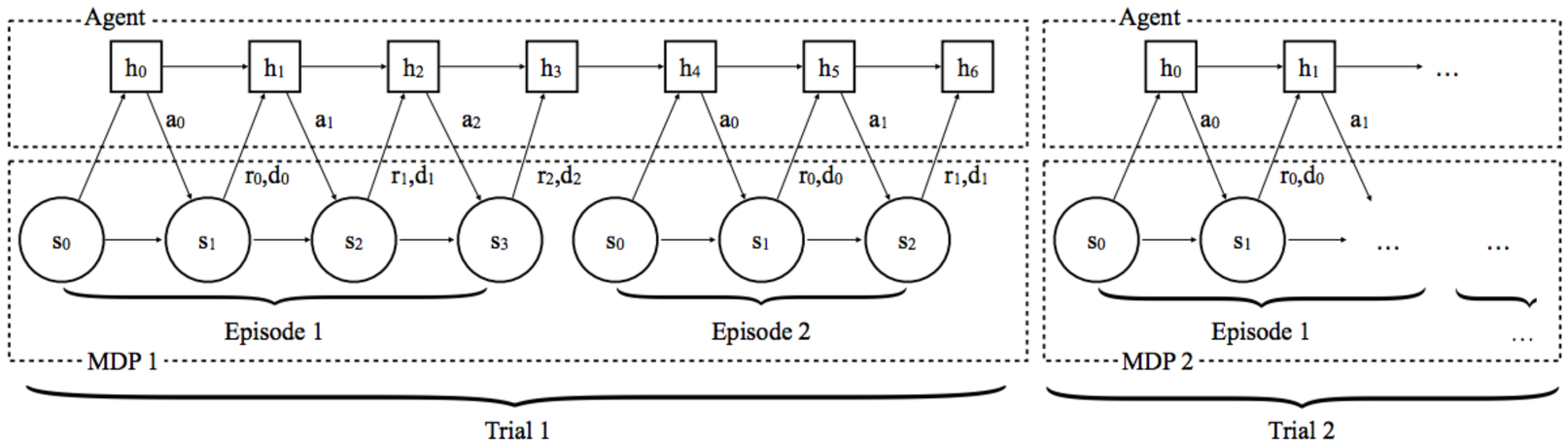
- Learn posterior distribution over Q functions. Sample Q function each episode.
- Recent Papers:
  - Osband, Ian, and Benjamin Van Roy. "Bootstrapped Thompson Sampling and Deep Exploration." (2015)
  - Yarin Gal, and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning." (2015).
  - Zachary Lipton et al., "Efficient Exploration for Dialogue Policy Learning with BBQ Networks & Replay Buffer Spiking" (2016)
    - Use Bayesian neural networks

Recent Papers: Exploration Bonus via State Novelty

- Stadie et al., "Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models." (2015)
  - exploration bonus := error in next-state prediction
- Bellemare et al., "Unifying Count-Based Exploration and Intrinsic Motivation" (2016)
  - Followup, Ostrovski et al., "Count-Based Exploration with Neural Density Models" (2017)
- Houthooft et al., "Variational information maximizing exploration" (2016).
  - Measure state novelty via information gain of dynamics model

Recent Papers: Learning to Explore, Unknown System as a POMDP

- Duan et al., "RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning." (2017)
- Wang et al., "Learning to Reinforcement Learn" (2017)
- Outer episodes (sample a new bandit problem / MDP) and inner episodes (of sampled MDP).
- Use RNN policy with no state reset between inner episodes



Intrinsic Motivation

- Reward functions that can be defined generically and lead to good long-term outcomes for agent
  - encourage visiting novel states
  - encourage safety
- Singh, S. P., Barto, A. G., and Chentanez, N. *Intrinsically motivated reinforcement learning*. In NIPS, 2005.
  - original ML paper on the topic
- Oudeyer, Pierre-Yves, and Frederic Kaplan. *How can we define intrinsic motivation?* 2008.
  - good extensive review
- Shakir Mohamed and Danilo J. Rezende, Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning, ArXiv 2015.
  - good short review & ideas on empowerment

Intrinsic Motivation

- Information theoretic intrinsic motivation signals listed by Oudeyer et al:
  - Uncertainty motivation: maximize prediction error / surprise of observations
  - Information gain about uncertain model
    - (see papers by Schmidhuber on “curiosity”, additional ideas on compression)
  - Empowerment — mutual information between action sequence and future state
  - Several other novelty measures

That’s All. Questions?