Exploration

2015/10/12

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

- How to search through the space of possible strategies of the agent to avoid getting stuck in local optima of behavior

Problem Settings

Multi-Armed Bandits

theoretically tractable

Contextual Bandits

Finite MDP
(where optimal
planning is free)

Large/infinite MDP,
running online PG or
ADP alg

theoretically intractable

Problem Settings

Multi-Armed Bandits

Contextual Bandits

Themes:
- Use optimistic value estimates
- Thompson sampling

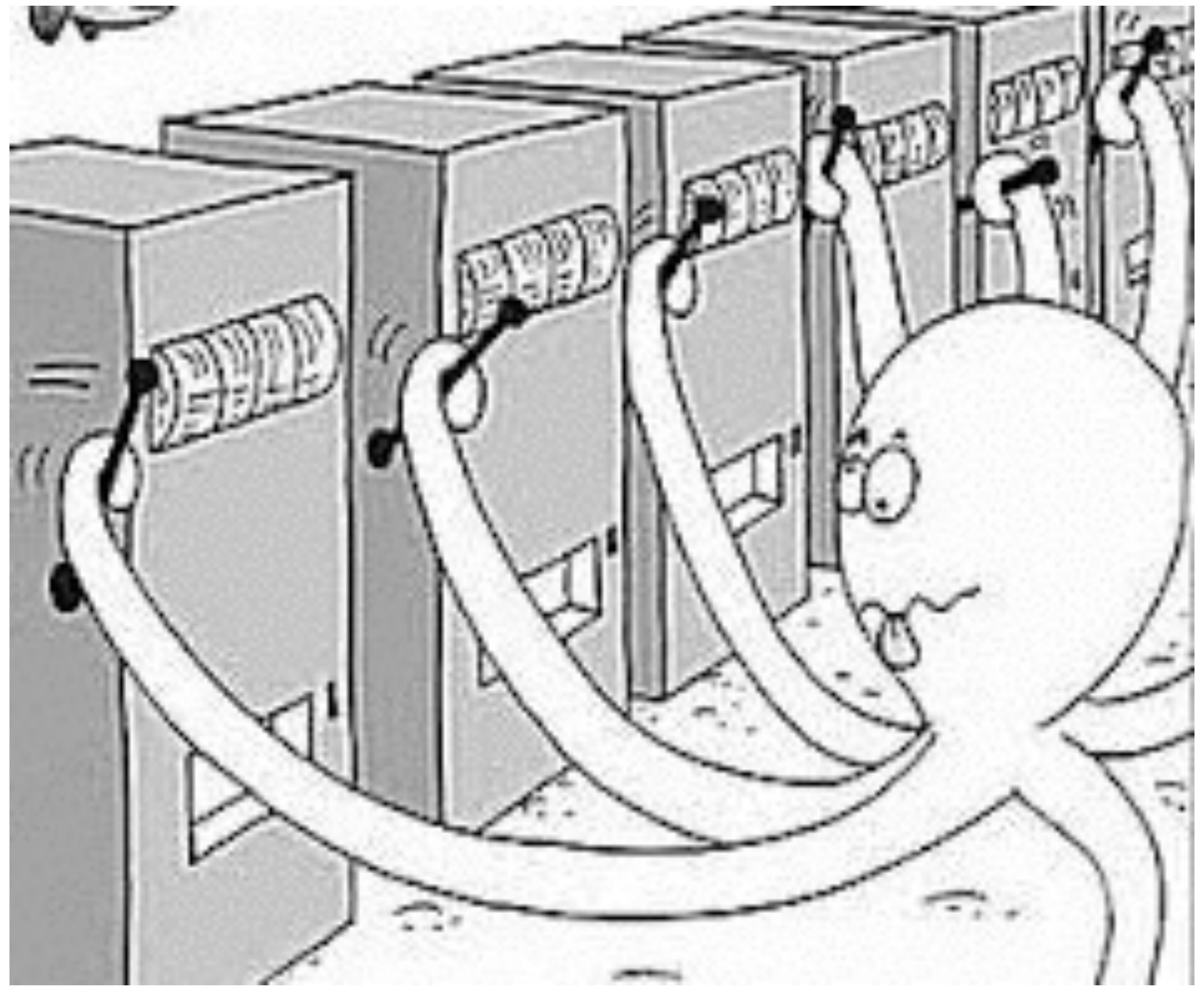
Finite MDP
(where optimal
planning is free)

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*

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 $t = 1,2, \ldots$
    - 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).

UCB-style algorithms

- “Upper Confidence Bound”, not UC Berkeley unfortunately
- Pick the arm that maximizes *mean + const * stdev*
- I.e., 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

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 $t = 1,2, \ldots$
  - environment provides context s_t
  - agent chooses $a_t \in \{1,2,\ldots,k\}$
  - environment provides reward R_t according to $p(R \mid a_t)$

Applications of 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

Finite MDPs, PAC Exploration

Definition 1 *(Kakade, 2003) Let* $c = (s_1, a_1, r_1, s_2, a_2, r_2, \ldots)$ *be a path generated by executing an algorithm* $\mathcal{A}$ *in an MDP* M*. For any fixed* $\epsilon > 0$*, the* **sample complexity of exploration** *(***sample complexity***, for short) of* $\mathcal{A}$ *with respect to* c *is the number of timesteps* t *such that the policy at time* t*,* $\mathcal{A}_t$*, is not* ϵ*-optimal from the current state,* s_t *at time* t *(formally,* $V^{\mathcal{A}_t}(s_t) < V^*(s_t) - \epsilon$*).*

Definition 2 *An algorithm* $\mathcal{A}$ *is said to be an* **efficient PAC-MDP** *(Probably Approximately Correct in Markov Decision Processes) algorithm if, for any* ϵ *and* δ*, the per-step computational complexity and the sample complexity of* $\mathcal{A}$ *are less than some polynomial in the relevant quantities* $(|S|, |A|, 1/\epsilon, 1/\delta, 1/(1-\gamma))$*, with probability at least* $1 - \delta$*. For convenience, we may also say that* $\mathcal{A}$ *is* **PAC-MDP**.

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

Finite MDPs, PAC Exploration

Delayed Q-Learning
no epsilon greedy!
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

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.

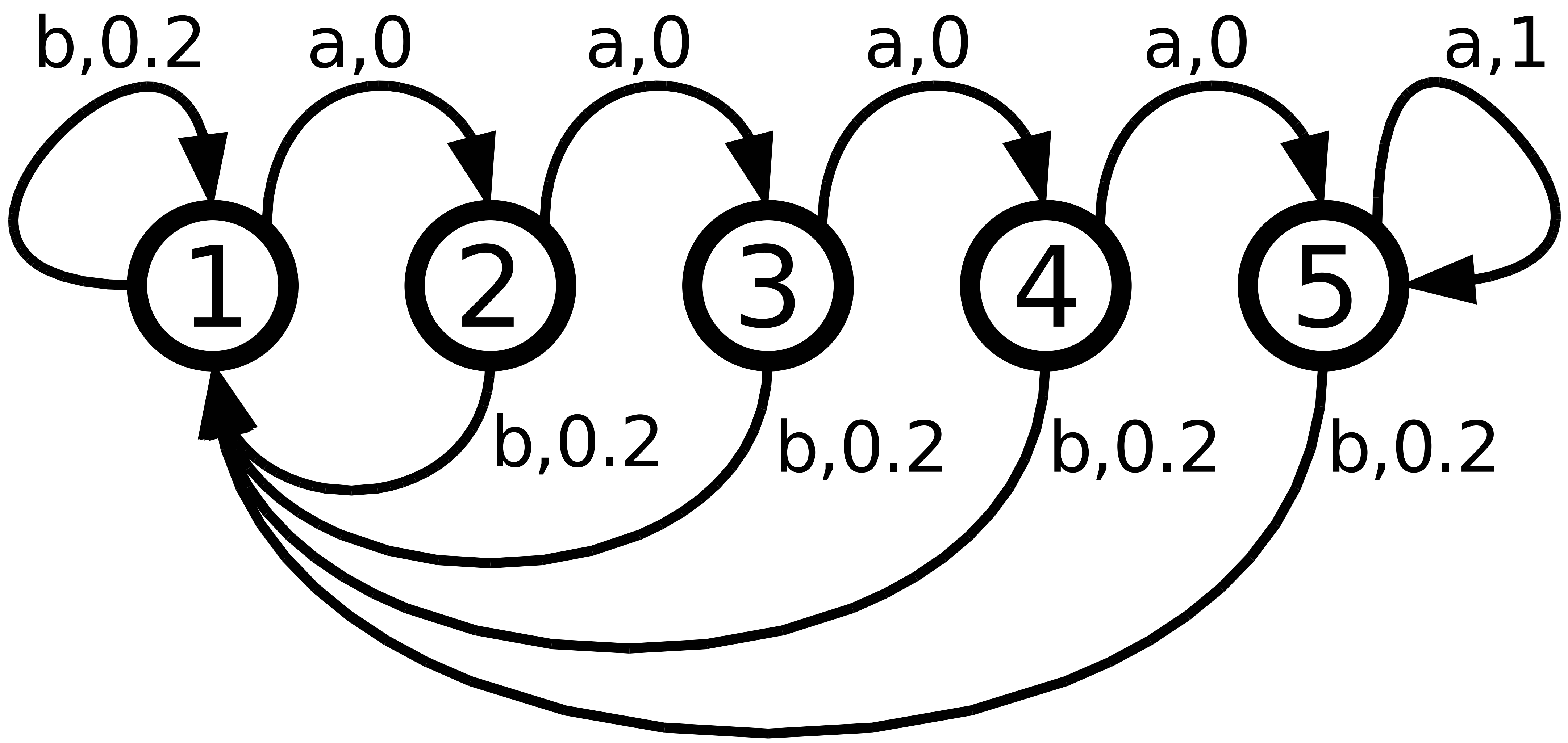
Szita, István, and András Lőrincz. "The many faces of optimism: a unifying approach." ICML 2008.

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

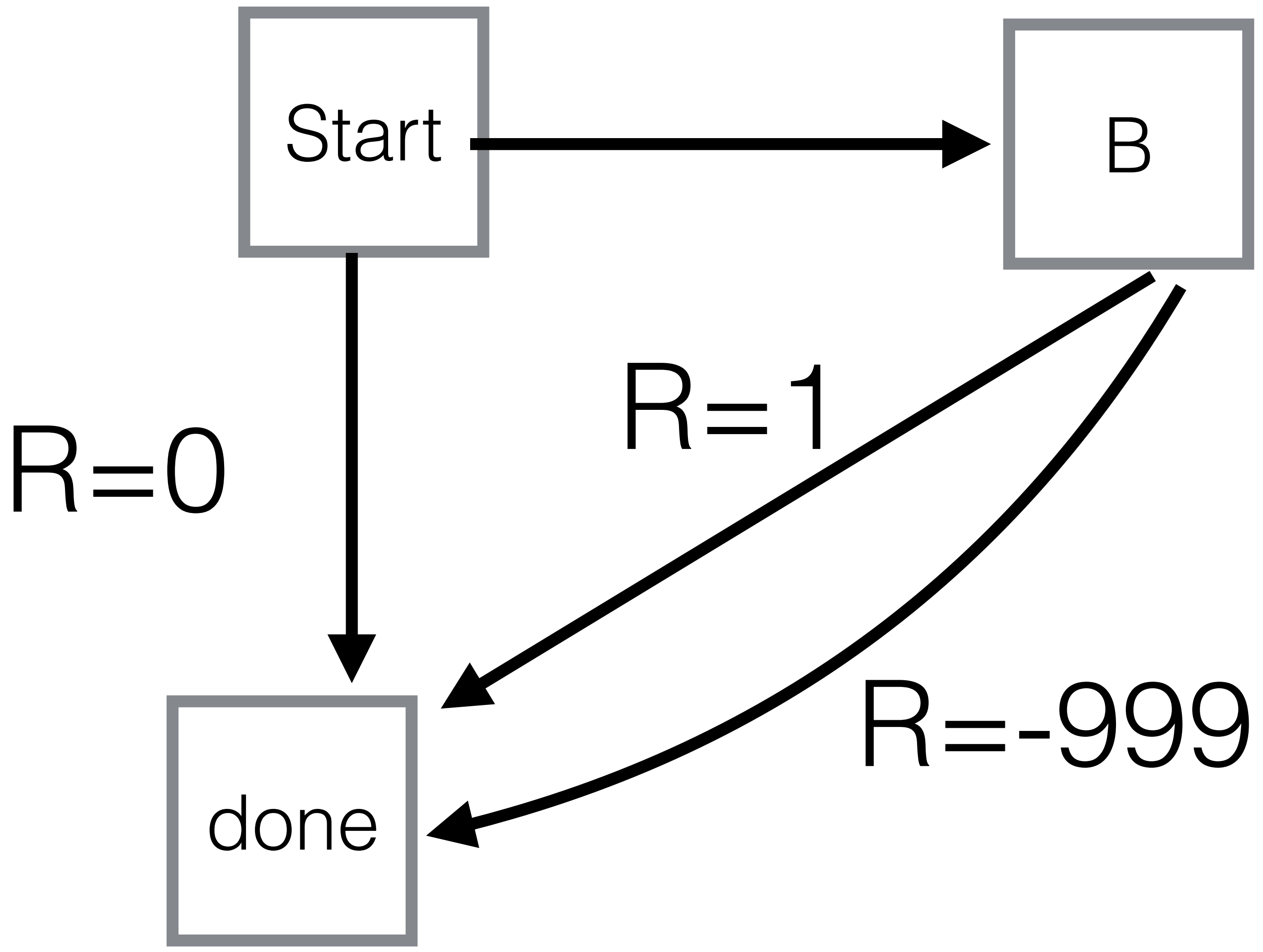
MDPs — examples



samples needed ~ 2^{Length}

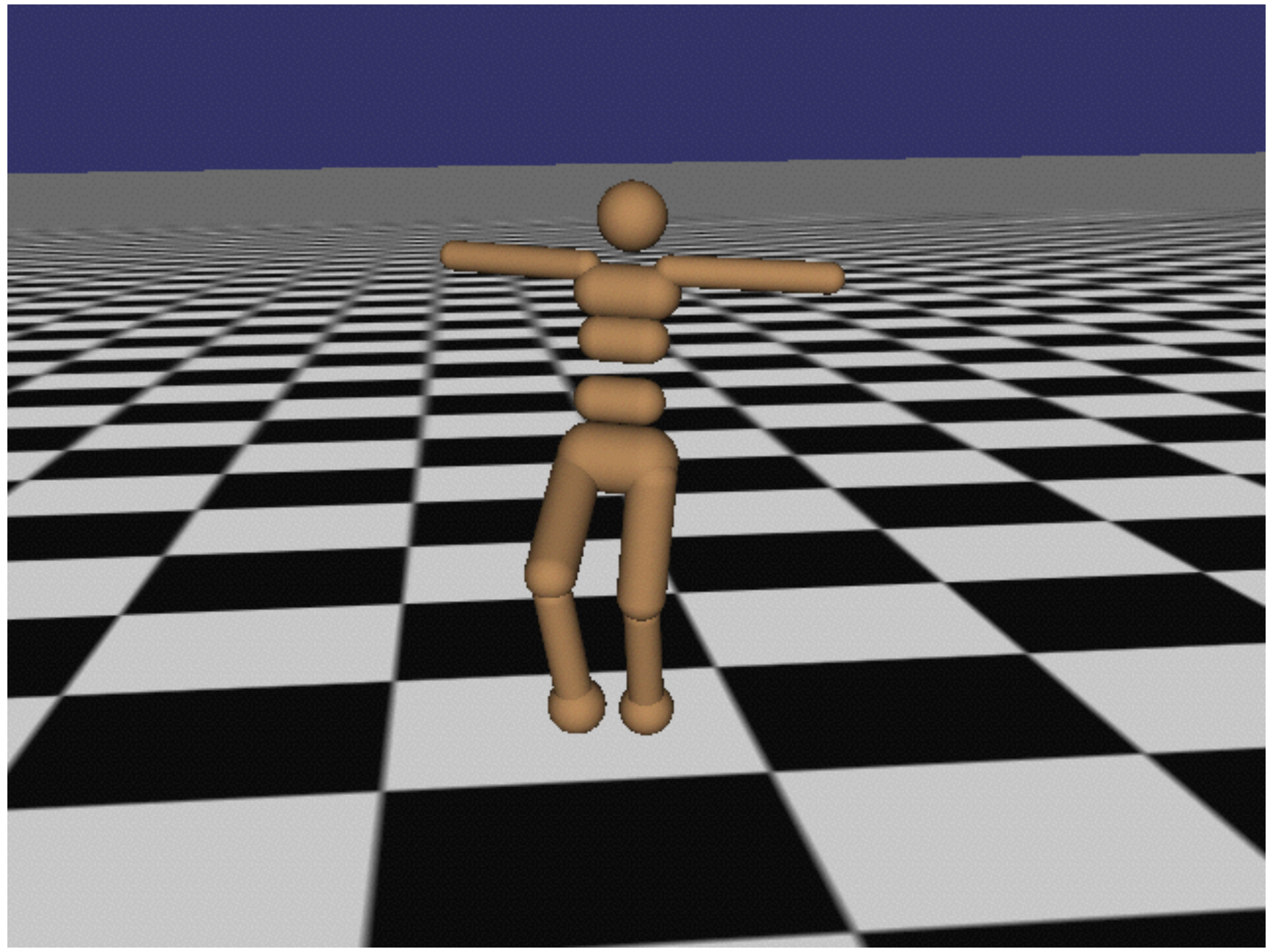
from Kolter & Ng, Near-Bayesian Exploration in Polynomial Time

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

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.

- Papers:

  - Osband, Ian, and Benjamin Van Roy. "Bootstrapped Thompson Sampling and Deep Exploration." arXiv preprint arXiv:1507.00300 (2015).

  - Yarin Gail, and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning." arXiv preprint arXiv:1506.02142 (2015).

Exploration Bonus via State Novelty

- Stadie, Bradly C., Sergey Levine, and Pieter Abbeel. "Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models." arXiv preprint arXiv: 1507.00814 (2015).

- Pazis, Jason, and Ronald Parr. "PAC Optimal Exploration in Continuous Space Markov Decision Processes." AAAI. 2013.

- Curiosity papers of Schmidhuber et al.

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

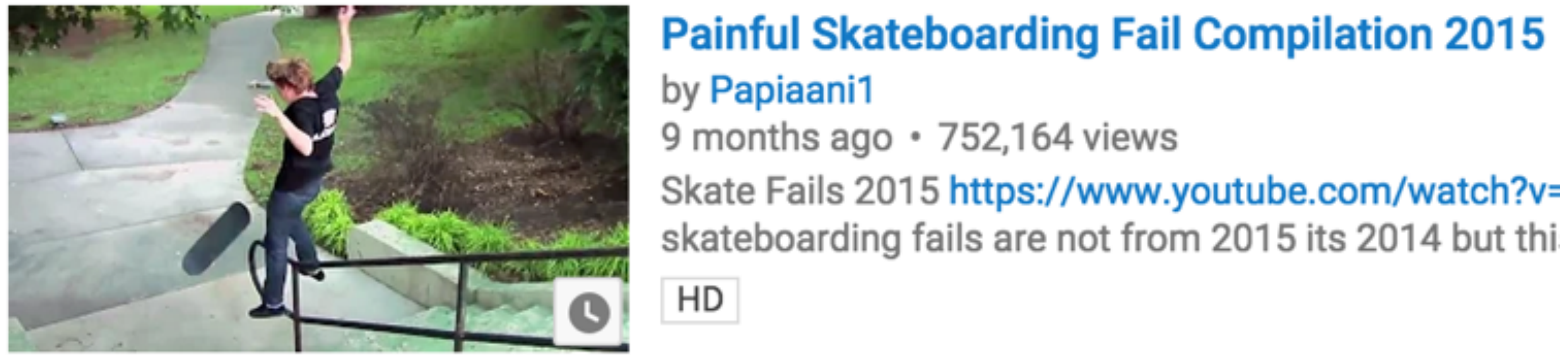
Intrinsic Motivation

- 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
- Competence based models
    - maximize learning
    - tasks should be hard but not too hard

The End

Learning dynamics models

Curiosity



Optimistic dynamics models