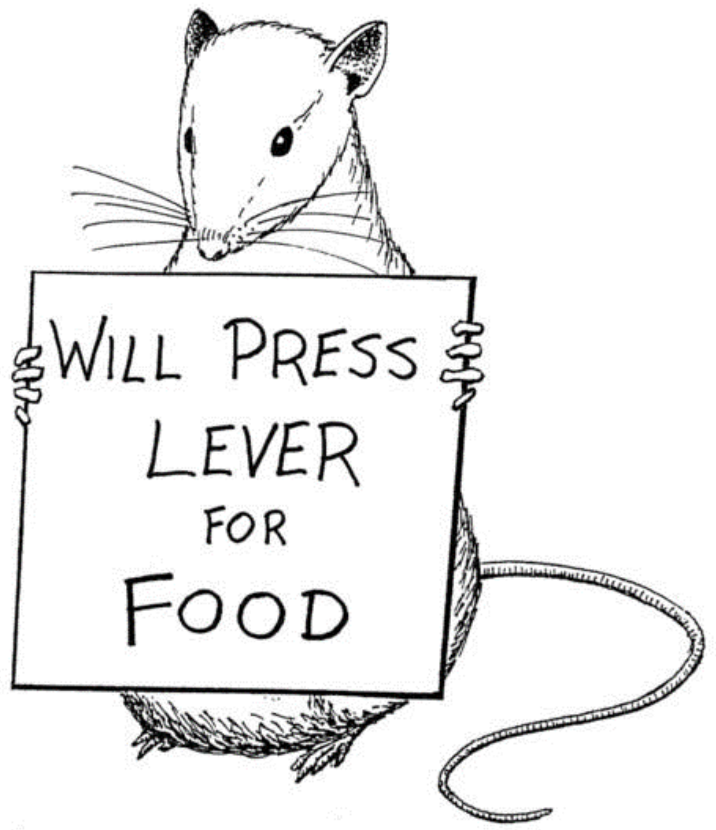
CS 277: Control and Reinforcement Learning

Winter 2026

Lecture 7: Exploration

Roy Fox
Department of Computer Science
School of Information and Computer Sciences
University of California, Irvine



Logistics

assignments

- Exercise 2 and Quiz 4 due next Monday

Today's lecture

Trust-region methods

Multi-Armed Bandits

Exploration in Deep RL

Importance Sampling

- Suppose you want to estimate $\mathbb{E}_{x\sim p}[f(x)]$
  - but only have samples $x \sim p'$

- Importance sampling:

$$\mathbb{E}_{x\sim p}[f(x)] = \mathbb{E}_{x\sim p'}\left[\frac{p(x)}{p'(x)} f(x)\right]$$

  - Importance (IS) weights: $\rho(x) = \frac{p(x)}{p'(x)}$
  - Estimate: $\rho(x) f(x)$ with $x \sim p'$

IS application 1: multi-step Q-Learning

- n-step Q-Learning: $Q(s_t, a_t) \to \sum_{\Delta t=0}^{n-1} \gamma^{\Delta t} r_{t+\Delta t} + \gamma^n \max_a Q(s_{t+n}, a)$

- Reminder: $Q^*(s_t, a_t)$ evaluates any a_t but optimal behavior afterward
  - We need data from $a_{t+\Delta t} = \arg\max_a Q(s_{t+\Delta t}, a)$ for RHS to estimate optimal target

- To be off-policy: update $Q(s_t, a_t) \to \sum_{\Delta t=0}^{n-1} \gamma^{\Delta t} \rho_t^{\Delta t} r_{t+\Delta t} + \gamma^n \max_a Q(s_{t+n}, a)$
  - with $\rho_t^{\Delta t} = \prod_{i=t+1}^{t+\Delta t} \frac{\pi(a_i | s_i)}{\pi'(a_i | s_i)}$ for data from π'

MF

θ

DP

π'

max

IS application 2: off-policy policy evaluation

MF
θ
DP
π'
max

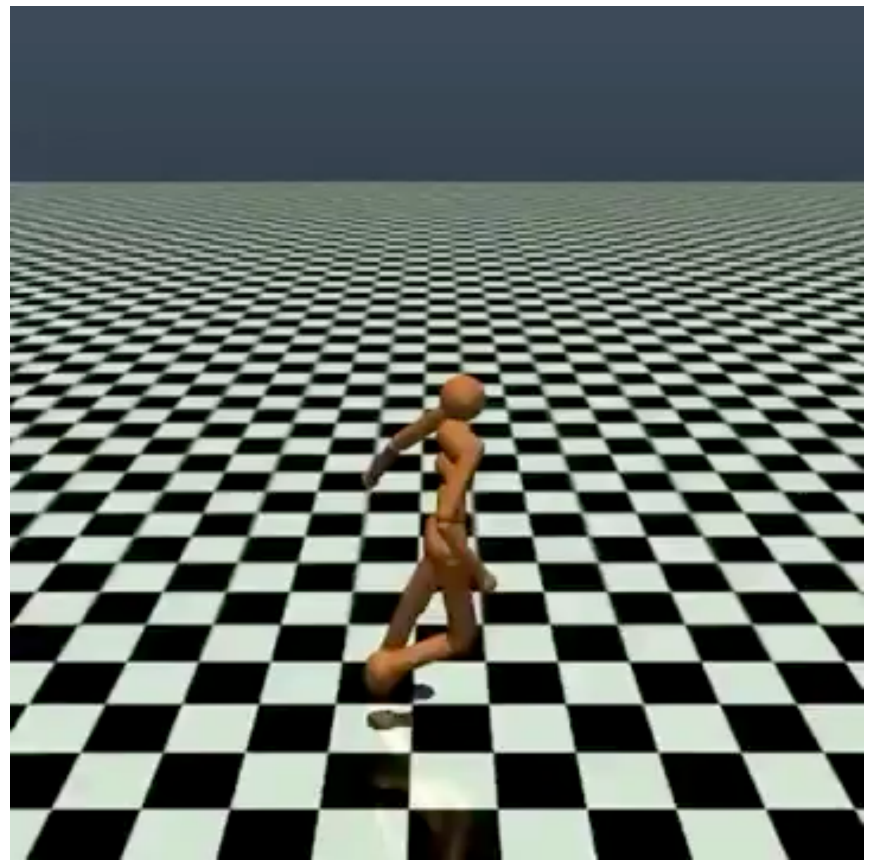
- Estimate $J_\pi = \mathbb{E}_{\xi \sim p_\pi}[R(\xi)]$ off-policy: $J_\pi = \mathbb{E}_{\xi \sim p_{\pi'}}[\rho^\pi_{\pi'}(\xi) R(\xi)]$
  - with $\rho^\pi_{\pi'}(\xi) = \frac{p_\pi(\xi)}{p_{\pi'}(\xi)} = \prod_t \frac{\pi(a_t | s_t)}{\pi'(a_t | s_t)}$ ← $p(s' | s, a)$ **cancels out**
- $\rho(\xi)$ can be very large or small ⇒ high variance
- Some reduction: r_t is not affected by future actions

$$J_\pi = \sum_t \mathbb{E}_{\xi_{\leq t} \sim p_{\pi'}}[\gamma^t \rho^\pi_{\pi'}(\xi_{\leq t}) r_t] = \sum_t \mathbb{E}_{\xi_{\leq t} \sim p_{\pi'}}\left[\gamma^t r_t \prod_{t' \leq t} \frac{\pi(a_{t'} | s_{t'})}{\pi'(a_{t'} | s_{t'})}\right]$$

[Precup et al., 2000]

IS application 3: Off-policy Policy Gradient

- Policy Gradient: $\nabla_\theta J_\theta = \sum_t \gamma^t \mathbb{E}_{\xi \sim p_\theta}[R_{\geq t}(\xi) \nabla_\theta \log \pi_\theta(a_t | s_t)]$
- Off-Policy PG: $\nabla_\theta J_\theta = \sum_t \gamma^t \mathbb{E}_{\xi \sim p_{\theta'}}[\rho_{\theta'}^{\theta}(\xi_{\leq t}) R_{\geq t}(\xi) \nabla_\theta \log \pi_\theta(a_t | s_t)]$
  - $R_{\geq t}(\xi)$ = future discounted rewards affected by $\pi_\theta(a_t | s_t)$
  - $\rho_{\theta'}^{\theta}(\xi_{\leq t})$ = past probability ratios that affect $\pi_\theta(a_t | s_t)$



- Should we discount by γ^t? Not if we care about evidence from later states
- $\rho_{\theta'}^{\theta}(\xi_{\leq t})$ has high variance, some methods just use $\rho_{\theta'}^{\theta}(a_t | s_t) = \dfrac{\pi_\theta(a_t | s_t)}{\pi_{\theta'}(a_t | s_t)}$

MF | θ | DP | π' | max

[Liu et al., 2018]

Performance Difference Lemma

- Policy gradient = small changes in policy; can we make large changes?

telescopic cancelation

- For any π, ξ: $\sum_t \gamma^t A_\pi(s_t, a_t) = \sum_t \gamma^t (r_t + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)) = R(\xi) - V_\pi(s_0)$

advantage of entire trajectory

- Expectation by different policy: Performance Difference Lemma

$$\sum_t \gamma^t \mathbb{E}_{(s_t,a_t)\sim p_\pi}[A_{\bar{\pi}}(s_t, a_t)] = \mathbb{E}_{\xi\sim p_\pi}[R(\xi) - V_{\bar{\pi}}(s_0)] = J_\pi - J_{\bar{\pi}}$$

$s_0 \sim p$ in both π and π'

  - We want to maximize over π, with $\bar{\pi}$ fixed

- Compare: PG Theorem $\nabla_\theta J_\theta = \sum_t \gamma^t \mathbb{E}_{(s_t,a_t)\sim p_\theta}[A_{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t \mid s_t)]$

[Kakade and Langford, 2002]

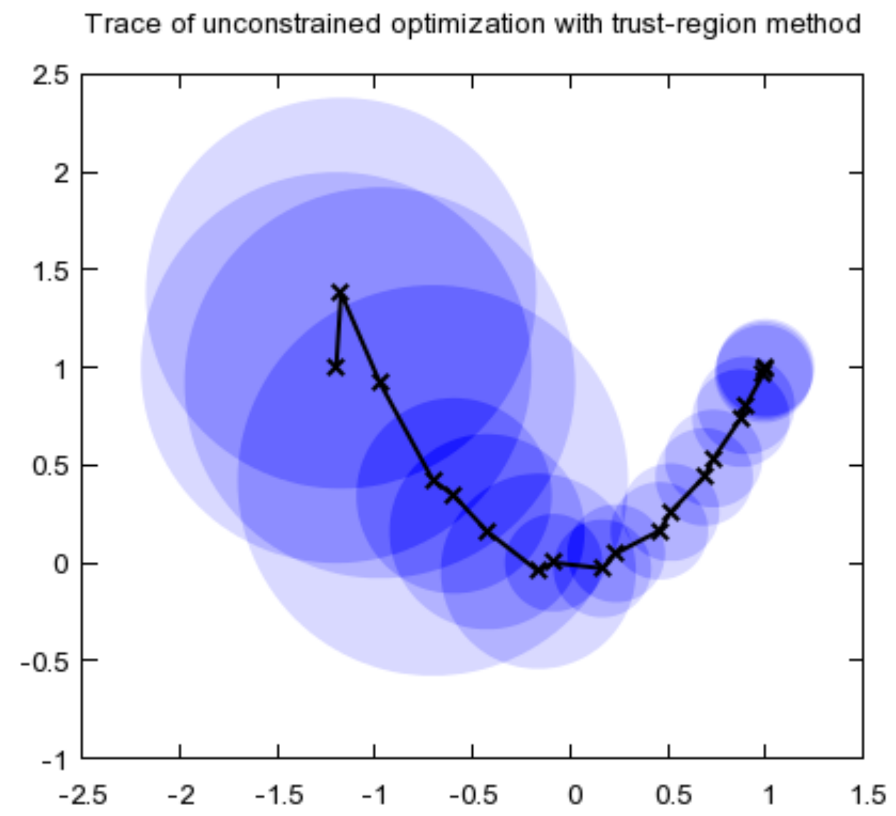
Finding best next policy

- With current policy $\bar{\pi}$: find $\max_\pi J_\pi - J_{\bar{\pi}} = \max_\pi \sum_t \gamma^t \mathbb{E}_{(s_t,a_t)\sim p_\pi}[A_{\bar{\pi}}(s_t, a_t)]$
  - Can use $\bar{\pi}$ to evaluate $A_{\bar{\pi}}$
- But we don't have data $(s_t, a_t) \sim p_\pi$; idea: sample from $\bar{\pi}$
  - Trick question: is this on-policy or off-policy? On-policy data, but needs IS weight

$$\max_\pi \sum_t \gamma^t \mathbb{E}_{\xi_{\le t}\sim p_{\bar{\pi}}}[\rho^{\pi}_{\bar{\pi}}(\xi_{\le t}) A_{\bar{\pi}}(s_t, a_t)]$$

- Is it reasonable to use $\rho^{\pi}_{\bar{\pi}}(a_t \mid s_t) = \dfrac{\pi(a_t \mid s_t)}{\bar{\pi}(a_t \mid s_t)}$ instead? i.e. drop $\rho^{\pi}_{\bar{\pi}}(\xi_{<t})$

Trust-Region Policy Optimization (TRPO)



- Trust region = space around $\bar{\pi}$ where $\rho(\xi_{<t}) \approx 1$
  - Easier to consider $\mathbb{E}_{\xi_{<t} \sim p_{\bar{\pi}}}[\log \rho(\xi_{<t})] \approx 0$

$$-\mathbb{E}_{\xi_{<t} \sim p_{\bar{\pi}}}[\log \rho(\xi_{<t})] = \mathbb{D}[\bar{\pi}(\xi_{<t}) \| \pi(\xi_{<t})] = \sum_{t'<t} \mathbb{E}_{\xi_{<t'} \sim p_{\bar{\pi}}}[\mathbb{D}[\bar{\pi}(a_{t'} | s_{t'}) \| \pi(a_{t'} | s_{t'})]]$$

- TRPO: $\max_{\theta} \mathbb{E}_{(s,a) \sim p_{\bar{\theta}}}[\rho_{\bar{\theta}}^{\theta}(a | s) A_{\bar{\theta}}(s, a)]$ s.t. $\mathbb{E}_{s \sim p_{\bar{\theta}}}[\mathbb{D}[\pi_{\bar{\theta}}(a | s) \| \pi_{\theta}(a | s)]] \leq \epsilon$
  - $A_{\bar{\theta}}$ estimated with critic A_{ϕ}
  - Computational tricks for gradient-based optimization

MF
θ
DP
π'
max

[Schulman et al., 2015]

Proximal Policy Optimization (PPO)

MF
θ
DP
π'
max

- Same motivation: ascend $\mathbb{E}_{(s,a)\sim p_{\bar\theta}}[\rho_{\bar\theta}^{\theta}(a|s)A_{\bar\theta}(s,a)]$ with π_θ staying near $\pi_{\bar\theta}$
  - PPO-Penalty: add a penalty term for $\mathbb{E}_{s\sim p_{\bar\theta}}[\mathbb{D}[\pi_{\bar\theta}(a|s)\|\pi_\theta(a|s)]]$
  - PPO-Clip: ascend $\mathbb{E}_{(s,a)\sim p_{\bar\theta}}[L_{\bar\theta}^{\theta}(s,a)]$ with

$$L_{\bar\theta}^{\theta}(s,a) = \min(\rho_{\bar\theta}^{\theta}(a|s)A_{\bar\theta}(s,a), A_{\bar\theta}(s,a) + |\epsilon A_{\bar\theta}(s,a)|)$$

- Positive / negative advantage ⇒ increase / decrease $\rho_{\bar\theta}^{\theta}(a|s) = \dfrac{\pi_\theta(a|s)}{\pi_{\bar\theta}(a|s)}$
  - But no incentive beyond $\rho_{\bar\theta}^{\theta}(a|s) = 1 \pm \epsilon$

- **no incentive ≠ doesn't happen**
- **PPO has lots more tricks to limit divergence**

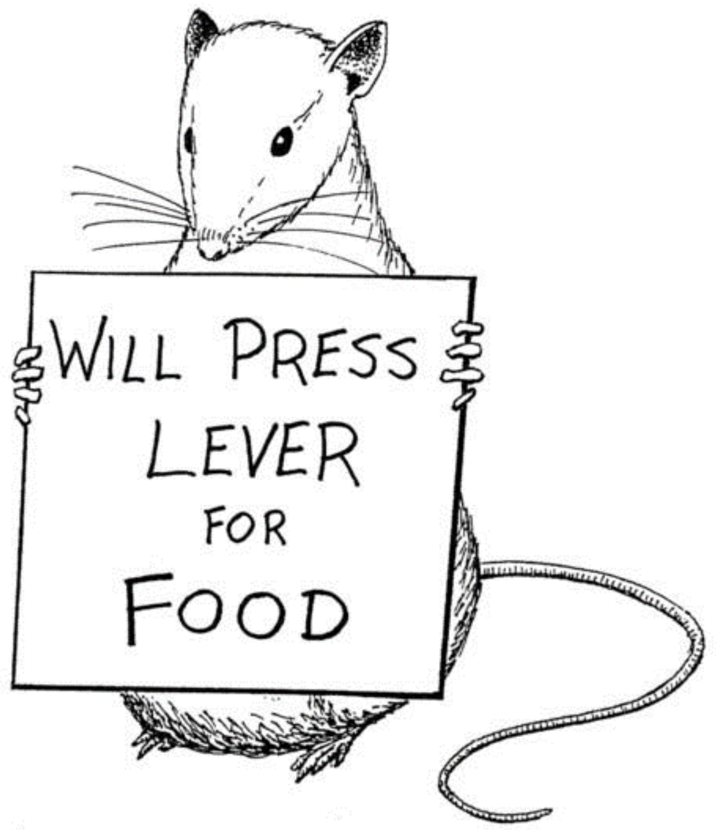
[Schulman et al., 2017]

Recap

- Model-based policy evaluation can be solved linearly
- Deep RL isn't just SGD
  - Exception: policy gradient on offline (batch) data
- Value-based methods struggle to $\max$ in continuous action spaces
  - DDPG: π_θ learns to maximize Q_ϕ (actor–critic method)
- Importance Sampling decouples expectation and sampling distributions
  - Optimize on-policy objectives with off-policy data
  - TRPO and PPO: sample from current policy to evaluate next policy, if it's close

State of the Course

- Model-Free RL: done!
- Up next:
  - Model-Based RL (related: Optimal Control)
  - Twists and turns!
    - Exploration, partial observability, non-reward feedback, structure
  - Advanced settings!
    - Inverse RL, Bounded RL, Offline RL, Multi-Agent RL & more



Today's lecture

Trust-region methods

Multi-Armed Bandits

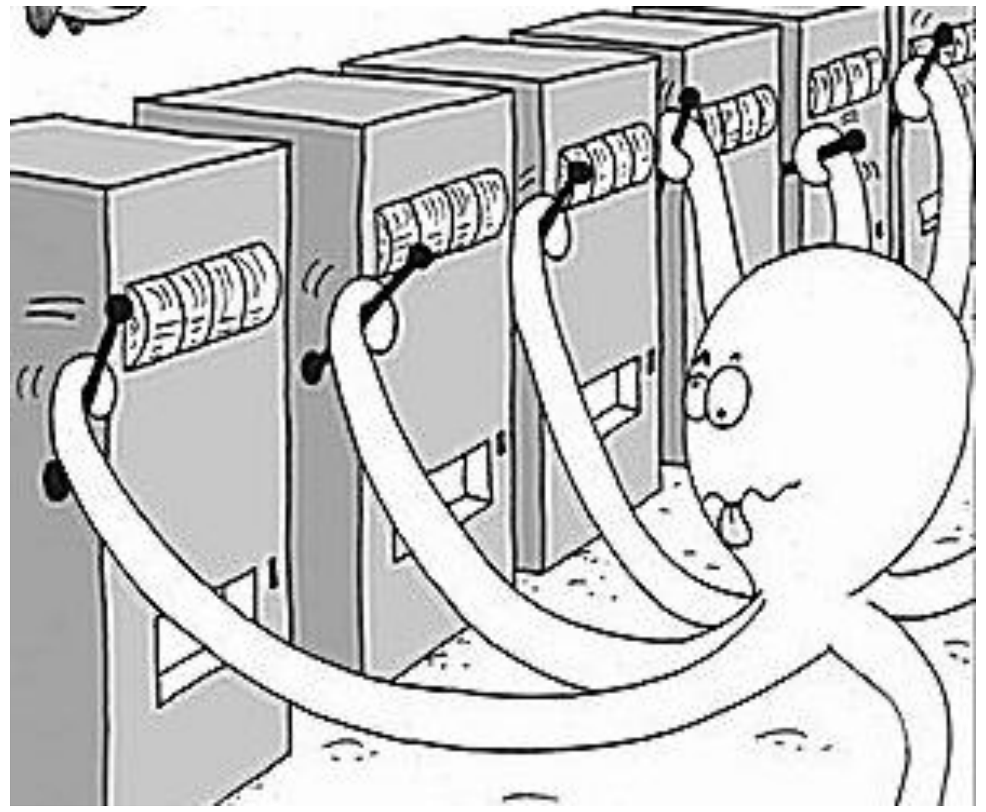
Exploration in Deep RL

Multi-Armed Bandits (MABs)

- Basic setting: single instance x, multiple actions $a_1, \ldots, a_k$
  - Each time we take action a_i we see a noisy reward $r_t \sim p_i$
- Can we maximize the expected reward $\max_i \mathbb{E}_{r \sim p_i}[r]$?
  - We can use the mean as an estimate $\mu_i = \mathbb{E}_{r \sim p_i}[r] \approx \frac{1}{n(i)} \sum_{t \in \mathcal{T}_i} r_t$
- Challenge: is the best mean so far the best action?
  - Or is there another that's better than it appeared so far?

One-armed bandit

Multi-armed bandit



Exploration vs. exploitation

- Exploitation = choose actions that seems good (so far)
- Exploration = see if we're missing out on even better ones
- Naïve solution: learn r by trying every action enough times
  - Suppose we can't wait that long: we care about rewards while we learn
- Regret = how much worse our return is than an optimal action

$$\rho(T) = T\mu_{a^*} - \sum_{t=0}^{T-1} r_t$$

  - Can we get the regret to grow sub-linearly with T? $\implies$ average goes to 0: $\frac{\rho(T)}{T} \to 0$

Let's play!

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

Simple exploration: ϵ-greedy

- With probability ϵ:
  - Select action uniformly at random
- Otherwise (w.p. $1 - \epsilon$):
  - Select best (on average) action so far
- Problem 1: all non-greedy actions selected with same probability
- Problem 2: must have $\epsilon \to 0$, or we keep accumulating regret
  - But at what rate should ϵ vanish?

Boltzmann exploration

- Keep an average of past rewards $\hat{\mu}_i = \frac{1}{n(i)} \sum_{t \in \mathcal{T}_i} r_t$

- Boltzmann (softmax) exploration: $\pi(a_i) = \text{softmax}_\beta \hat{\mu}_i = \frac{\exp(\beta \hat{\mu}_i)}{\sum_j \exp(\beta \hat{\mu}_j)}$

- Obviously bad actions $\hat{\mu}_i \ll \max_j \hat{\mu}_j$ are unlikely to be used (but can!)
  - Problem: still must have $\beta \to \infty$, or we keep accumulating regret
  - Some evidence that β should increase linearly

Optimism under uncertainty

- Tradeoff: explore less used actions, but don't be late to start exploiting what's known
  - Principle: optimism under uncertainty = explore to the extent you're uncertain, otherwise exploit
- By the central limit theorem, the mean reward $\hat{\mu}_i$ of arm i quickly $\to \mathcal{N}\left(\mu_i, O\left(\frac{1}{n(i)}\right)\right)$
- Be optimistic by slowly-growing number of standard deviations:

$$a = \arg\max_i \hat{\mu}_i + \sqrt{\frac{2\ln T}{n(i)}}$$

  - Upper confidence bound (UCB): likely $\mu_i \leq \hat{\mu}_i + c\sigma_i$; unknown variance $\implies$ let c grow
  - But not too fast, or we fail to exploit what we do know
- Regret: $\rho(T) = O(\log T)$, provably optimal

Thompson sampling

- Consider a model of the reward distribution $p_{\theta_i}(r \,|\, a_i)$
- Suppose we start with some prior $q(\theta)$
  - Taking action a_t, see reward $r_t \implies$ update posterior $q(\theta \,|\, \{(a_{\leq t}, r_{\leq t})\})$
- Thompson sampling:
  - Sample $\theta \sim q$ from the posterior
  - Take the optimal action $a^* = \max_i \mathbb{E}_{r \sim p_{\theta i}}[r]$
  - Update the belief (different methods for doing this)
  - Repeat

Other online learning settings

- What is the reward for action a_i?
  - MAB: random variable with distribution $p_i(r)$
  - Adversarial bandits: adversary selects r_i for every action
    - The adversary knows our algorithm! And past action selection! But not future actions
      - Learner must be stochastic (= unpredictable), but we can still have guarantees
  - Dueling bandits: just 1 bit of feedback, is a_i better or a_j?
- Contextual bandits: we also get instance $x \sim p$, make decision $\pi(a \mid x)$
  - Can we generalize to unseen instances?

Today's lecture

Trust-region methods

Multi-Armed Bandits

Exploration in Deep RL

Learning with sparse rewards

- Montezuma's Revenge
  - Key = 100 points
  - Door = 500 points
  - Skull = 0 points
    - Is it good? Bad? Affects something off-screen? Opens up an easter egg?
  - Humans have a head start with transfer from known objects

- Exploration before learning:
  - Random walk until you get some points — could take a while!

RL exploration is more complicated...

- Need to consider states and dynamics
- Need coordinated behavior to get *anywhere*
  - E.g., cross a bridge to get the game started...
  - Random exploration will kill us with high probability
    - Structured exploration: noise over time has joint distribution, temporal structure
- How to define regret?
  - With respect to constant action? We can outperform it
  - With respect to optimal policy? May be too hard to learn $\implies$ linear regret
  - Most approaches are heuristic, no regret guarantees; often train-time rewards don't matter

Count-based exploration

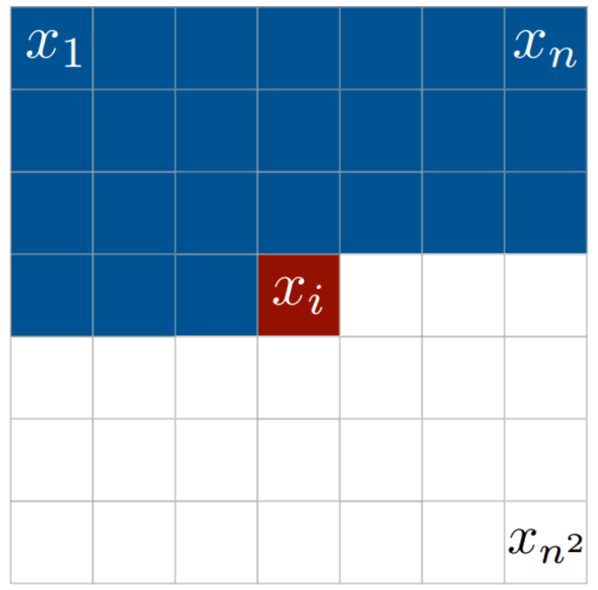
- Generalizing UCB exploration $a = \arg\max_i \hat{\mu}_i + \sqrt{\frac{2 \ln T}{n(i)}}$ from MAB to RL
- Count visitations to each state $n(s)$ (or state-action $n(s, a)$)
- Optimism under uncertainty: add exploration bonus to scarcely-visited states

$$\tilde{r} = r + r_e(n(s))$$

  - r_e should be monotonic decreasing in $n(s)$
  - Need to tune its weight

Density model for count-based exploration

- How to represent "counts" in large state spaces?
  - We may never see the same state twice
  - If a state is very similar to ones we've seen often, is it new?
- Train a density model $p_\phi(s)$ over past experience
- Unlike generative models, we care about getting the density correctly
  - But we don't care about the quality of samples



- Density models for images:
  - CTS, PixelRNN, PixelCNN, etc.

Pseudo-counts

- How to infer pseudo-counts from a density model?

$$p_\phi(s) = \frac{n(s)}{N}$$

- After another visit:

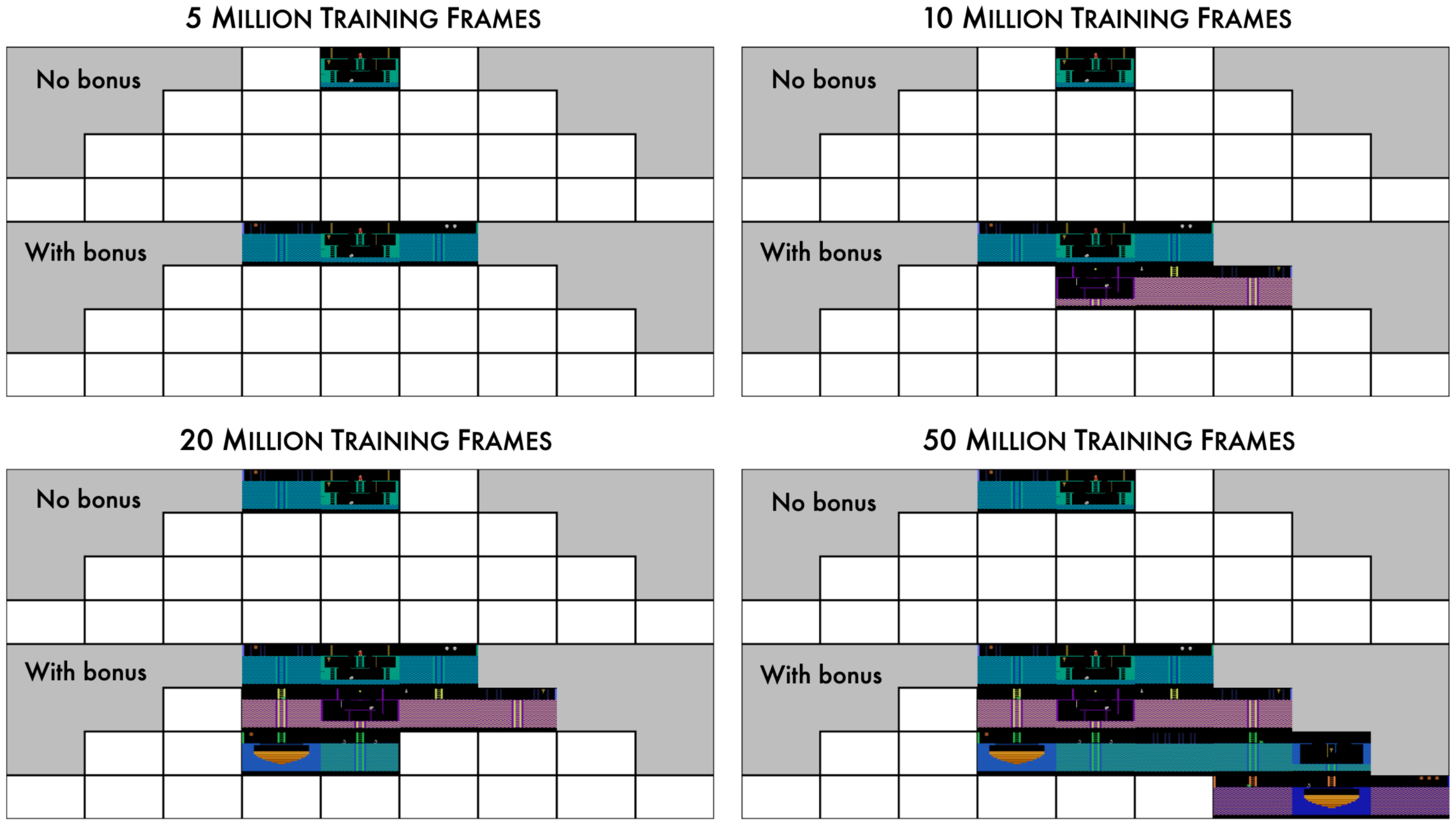
$$p_\phi(s) = \frac{n(s)+1}{N+1}$$

- To recover the pseudo-count:
  - $p_{\phi'} \leftarrow$ mock-update the density model with another visit of s
  - Compute

$$\hat{N} = \frac{1 - p_{\phi'}(s)}{p_{\phi'}(s) - p_\phi(s)} p_\phi(s) \qquad \hat{n}(s) = \hat{N} p_\phi(s)$$

Exploration bonus

- What's a good exploration bonus?
- In bandits: Upper Confidence Bound (UCB)
  - $$r_e(n(s)) = \sqrt{\frac{2 \ln N}{n(s)}}$$
- In RL, often:
  - $$r_e(n(s)) = \sqrt{\frac{1}{n(s)}}$$



[Bellemare et al., 2016]

Thompson sampling for RL

- Keep a distribution over models $p_\theta(\phi)$
- What's our “model”? Idea 1: MDP; Idea 2: Q-function
- Thompson sampling over Q-functions:
  - Sample $Q \sim p_\theta$
  - Roll out an episode with the greedy policy $\pi(s) = \arg\max_a Q(s, a)$
  - Update p_θ to be more likely for Q' that gives low empirical Bellman error
  - Repeat

Optimal exploration: simple settings

- Multi-Armed Bandits (MAB): single state, one-step horizon
  - Exploration–exploitation tradeoff very well understood
- Contextual bandits: random state, one-step horizon
  - Also has good theory (Online Learning)
- Tabular RL
  - Some good heuristics, recent theoretical guarantees
- Deep RL
  - Only few exploratory ideas and heuristics

Recap

- Online learning = getting good rewards while learning
  - In contrast: learn however, but deploy good policy
- Online learning requires trading off exploration–exploitation
  - Don't overfit to too little data
  - Don't be late to use what you've learned
- Optimism under uncertainty: exploration bonus for novelty
- Thompson sampling: coordinated exploration actions
- Same principles hold in RL