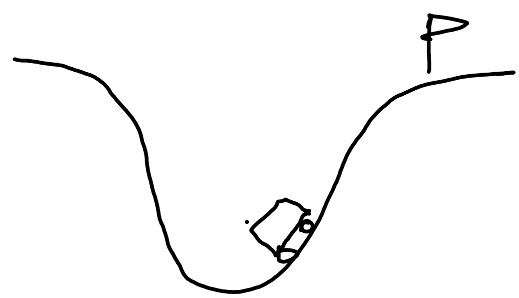
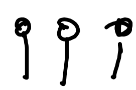
Last time: Inverse RL

Given	Find
$S, A, T, \{\tau_i\}$	R

This Time: Exploration ~~Hierarchical~~



Montezuma's Revenge Atari

Bandit 

ε-greedy

softmax e^{TQ}

$\rightarrow$ UCB $Q + c\sqrt{\frac{2 \ln T}{N}}$ $\leftarrow$ Optimistic

$\rightarrow$ Thompson Sampling distribution over models $\leftarrow$ Posterior Sampling

sample, take optimal w.r.t. sample

* Bayes Optimal BAMDP

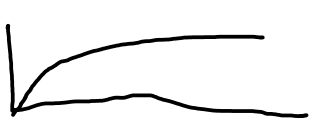
POMDP where unknown model params part of state

$O(\log n)$ Regret

$$\underset{\text{params}}{z},\ \underset{\text{obs}}{y},\ \underset{\text{action}}{a}: \quad IG(z, y \mid a) = E_y\left[H(\hat{p}(z)) - H(\hat{p}(z) \mid y) \mid a\right]$$

$$g(a) = IG(z, y \mid a) \qquad \Delta(a) = E\left[r(a^*) - r(a)\right]$$

$\rightarrow$ $\arg\min \frac{\Delta(a)^2}{g(a)}$ Russo + Van Roy "Learning to optimize with info directed sampling"



Optimistic

- new state = good state
- Exploration bonus

$$R^+(s,a) = R(s,a) + B(N(s,a))$$

$N(s)$

$N\uparrow \Rightarrow B\downarrow$

Large/ Continuous S

$\phi_\theta(s)$ "psedo-count"

Small MDP

$$P(s) = \frac{N(s)}{n}$$

after seeing a new state s

$$\rightarrow P'(s) = \frac{N(s)+1}{n+1}$$

model that fits these dynamics

fit $p_\theta(s)$ to all seen states D

fit $p_{\theta'}(s)$ to $D \cup \{s\}$

$p_\theta(s), p_{\theta'}(s) \longrightarrow \hat{N}(s)$

$R^+(s,a) = R(s,a) + B(\hat{N}(s))$

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

What B?

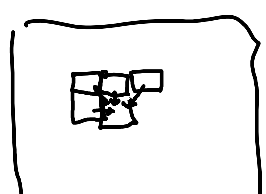
UCB

$\rightarrow \sqrt{\frac{1}{N(s)}}$

$\frac{1}{N(s)}$

Bellomare "Unifying Count-Based Exploration"

How to model $p_\theta(s)$? CTS



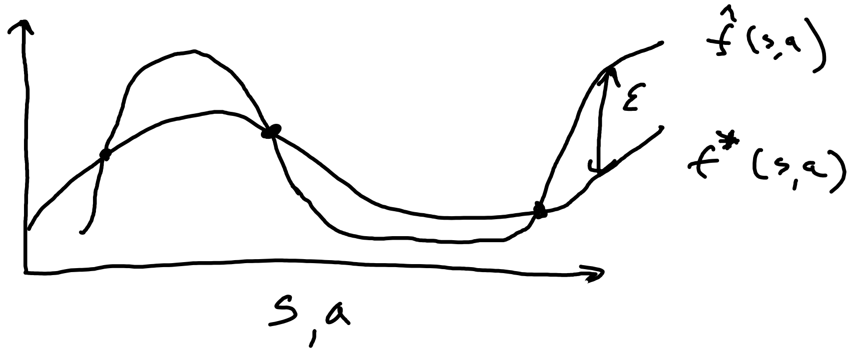
- #Exploration ← "hash"

compress s into k-bit code w/ $\phi(s)$, then count $N(\phi(s))$

- Use a classifier "Ex2"

$$\underline{p_\theta(s)} = \frac{1 - D_s(s)}{D_s(s)}$$ ← probability of positive

---



Use $\varepsilon(s,a) = \|\hat{f}(s,a) - f^*(s,a)\|$ as bonus

What is f^*

- $f^*(s,a) = s'$ "Curiosity"

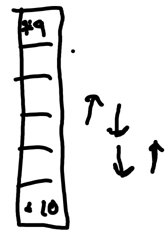
- $f^*(s,a) = f_\phi(s,a)$ where ϕ is random random Neural Network

RND - Random Network Distillation

Thompson Sampling Style

$p(Q)$

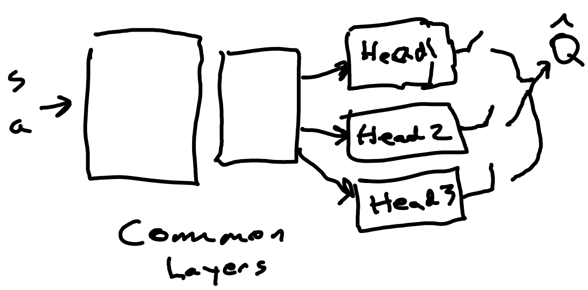
1. sample $\hat{Q}$ from $p(Q)$
2. act according to $\hat{Q}$ for 1 episode



Works with off-policy

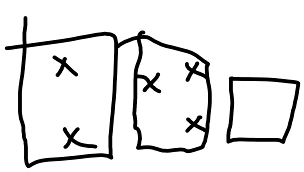
How to maintain $p(Q)$

- Bootstrapping
  - Resample D N times to get $D_1, \ldots D_N$
  - train f_{θ_i} on D_i
  - Sample from $p(\theta)$ by sampling $i \in 1..N$ using f_{θ_i}



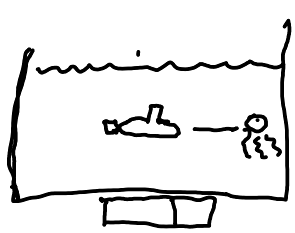
Osband et al. "Deep exploration via Bootstrapped DQN"

- Dropout



$p(Q)$

+ Don't have to modify Reward
- Doesn't work as well as optimistic



Information Gain

$IG(z, y \mid a)$ about what?

$R(s,a)$? bad for sparse

$p(s)$ state density

$p(s' \mid s,a)$ dynamics

Generally Intractible $\longrightarrow$ Approximations

Approximations

- prediction gain $\log p_{\theta'}(s) - \log p_\theta(s)$

  justification RND

- Variational Inference

  $q_\theta(s) \approx p(s \mid h)$ $\boxed{\text{VIME}}$

IG is like $D_{KL}(p(z \mid y) \,\|\, p(z))$

$p_\theta(s_{t+1} \mid s_t, a_t)$ $z = \theta$

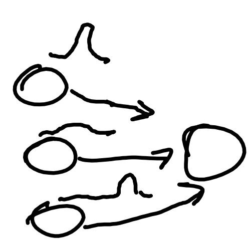
$q(\theta \mid \phi) \approx p(\theta \mid h)$ $y = (s_t, a_t, s_{t+1})$

$D_{KL}(p(\theta \mid h, s_t, a_t, s_{t+1}) \,\|\, p(\theta \mid h))$

specifically optimize variational lower bound

$D_{KL}(q(\theta \mid \phi) \,\|\, p(h \mid \theta)\, p(\theta))$

product of independent Gaussian param. distributions



Every step, update ϕ to ϕ'

Use $D_{KL}(q(\theta \mid \phi') \,\|\, q(\theta \mid \phi))$ as bonus

Review:

Optimistic: RND ← Best

theoretically justified → Thompson: Bootstrapping with many Q networks

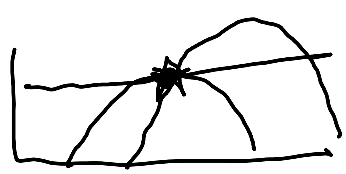
theoretically justified → IG: VIME

Using Expert Demonstration

Imitation Learning	RL
+ Simple, stable, supervised	+ Exceed human perf
− Demos	− Reward
− Unseen "Distributional shift"	− Exploration
− as good as expert	− Unstable

Simplest ↓

- pretrain + fine tune with RL
  - Flaws: − could bias
  - − could forget
- Use off-policy RL with data from expert
  - − Policy Gradient with importance sampling
  - $\boxed{\text{Guided Policy Search}}$
- Q-learning: Drop into Replay Buffer



$$Q(s,a) \leftarrow r(s,a) + \mathbb{E}_{a' \sim \pi_{new}}\left[Q(s',a')\right]$$

Hybrid objective

imitation $\sum_{(s,a)\in Demo} \log \pi_\theta(a|s)$

RL $E_{\pi_\theta}[R(s,a)$

Hybrid $E_{\pi_\theta}[R(s,a)] + \lambda \sum_{(s,a)\in Demo} \log \pi_\theta(a|s)$

Flaws: - choose weight
- domain-specific