Last Time Model-Based
Model-Free

~~Max~~ ML MB RL

loop
- choose action
- update $N, \rho \leftarrow$
- solve for optimal Q in most likely MDP ← expensive

Dyna

replace solve with

$$Q(s,a) \leftarrow \underbrace{R(s,a)}_{r} + \gamma \sum_{s'} \underbrace{T(s'|s,a)}_{\text{based } N} \max_{a'} Q(s',a')$$

- use extra time to update Q for random s

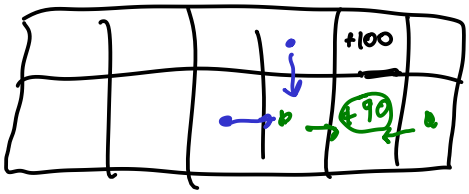
Prioritized Sweeping

Maintain a priority queue of states

$$\begin{bmatrix} s^1 \\ s^4 \\ \vdots \\ s^{10} \end{bmatrix}$$

Prioritized Sweeping (s)
- Increase $p(s)$ to ∞
- while pq not empty
  - $s \leftarrow$ highest priority state
  - Update (s)

Update (s)
- $v \leftarrow V(s) = \max_a Q(s,a)$
- $V(s) \leftarrow \max_a \left(R(s,a) + \gamma \sum_{s'} T(s'|s,a) V(s')\right)$
- for $\tilde{s}, \tilde{a} \in$ pred (s)
  - $p(\tilde{s}) \leftarrow T(s|\tilde{s},\tilde{a}) \cdot |V(s) - v|$



Bayesian RL

before : Most Likely MDP
now : Belief over all MDPs

θ model params

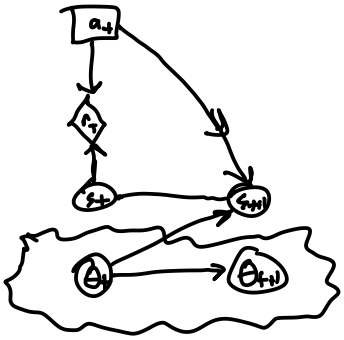
$T_\theta(s'|s,a)$

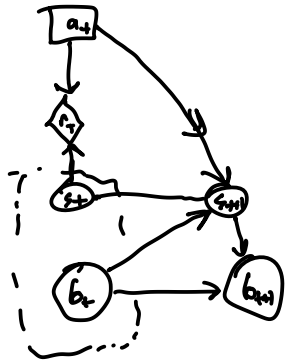
$R_\theta(s,a)$

$b_t = P(\theta \mid s_1, a_1, s_2, a_2 \dots s_t)$

Dirichlet

Bayes Adaptive MDP BAMDP



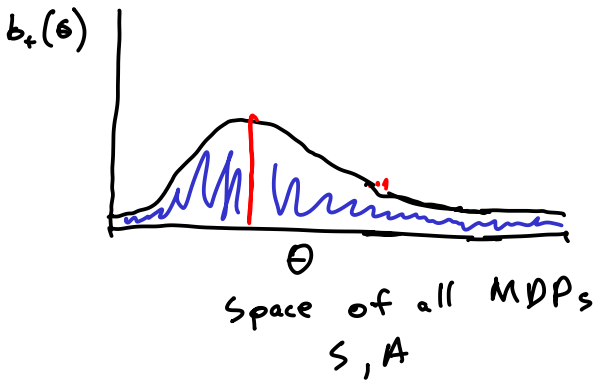


"belief state" (s_t, b_t)

Transition

$T(s', b' \mid s, b, a)$

hard to solve
can use e.g. MCTS

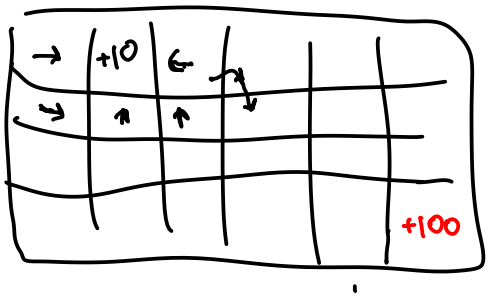


Space of all MDPs
S, A

$\pi(s) = \text{argmax}\ Q_{\hat{\theta}}(s,a)$

$\rightarrow \pi(s) = \text{argmax}\ \underset{\theta \sim b}{E}[Q_\theta(s,a)]$

↖ No exploration



Thompson Sampling

sample $\hat{\theta}$ from b_t
solve $Q_{\hat{\theta}}$
take best w.r.t. $Q_{\hat{\theta}}$

ϵ-greedy

Exponential samples to find sparse reward

→ Ben Van Roy
→ "Deep RL"
"Deep Exploration"
ϵ-greedy : bad
UCB : good ?
Thompson : good, but expensive
randomized value functions

Model-Free RL

No estimation of T, R; Learn Q or V directly

Temporal Difference (TD) — SARSA, Q-Learning

~~X~~ estimate mean from samples

$$\hat{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$Q(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

$$\hat{x}_n = \hat{x}_{n-1} + \frac{1}{n}(x_n - \hat{x}_{n-1})$$

$$\approx \hat{x}_{n-1} + \alpha(x_n - \hat{x}_{n-1})$$

α: learning rate; $(x_n - \hat{x}_{n-1})$: temporal diff

$$Q(s,a) = R(s,a) + \gamma E\left[\max_{a'} Q(s',a')\right]$$

$$0 = R(s,a) + \gamma E\left[\max_{a'} Q(s',a')\right] - Q(s,a)$$

On-policy TD learning: Sarsa

loop
- observe s', choose a'
- $Q(s,a) \leftarrow Q(s,a) + \alpha\left[R(s,a) + \gamma Q(s',a') - Q(s,a)\right]$
- $a \leftarrow a'$
- $s \leftarrow s'$

Off-policy TD learning: Q-learning

loop
- choose a
- observe s'
- $Q(s,a) \leftarrow Q(s,a) + \alpha\left[R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$
- $s \leftarrow s'$

$R(s,a)$: sample; $\max_{a'} Q(s',a')$: π

Big problem with Q-learning: maximization bias

after n steps

$Q(s,a^1)$ +0.4
$Q(s,a^2)$ -0.1 max
$Q(s,a^3)$ +0.2

Double Q learning

Q_1, Q_2

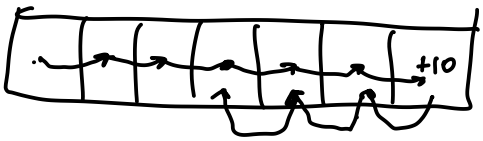
$$Q_1(s,a) \leftarrow Q_1(s,a) + \alpha\left[R(s,a) + \gamma Q_2\left(s', \arg\max_{a'} Q_1(s',a')\right) - Q_1(s,a)\right]$$

Q-learning: guaranteed to converge to Q^* w.p. 1

Sarsa : " " " given that exploration policy converges to the greedy policy

Big problem : slow

2. Credit Assignment

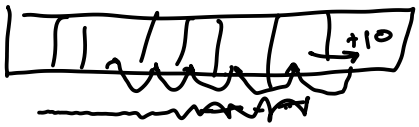


Eligibility traces

Exponentially-decaying visit count

$Q(\lambda)$ or $TD(\lambda)$

$Sarsa(\lambda)$



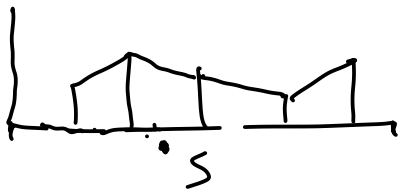
loop
  choose a
  observe s'
  $N(s,a)$ ++
  $\delta \leftarrow R(s,a) - \gamma \max_{a'} Q(s',a') - Q(s,a)$
  for $\tilde{s} \in S$, $\tilde{a} \in A$
    $Q(\tilde{s},\tilde{a}) \leftarrow Q(\tilde{s},\tilde{a}) + \alpha \delta N(\tilde{s},\tilde{a})$
    $N(\tilde{s},\tilde{a}) \leftarrow \gamma \lambda N(\tilde{s},\tilde{a})$

3. Generalization

$$Q(s,a) = \theta^T \beta(s,a) \qquad \frac{\partial Q}{\partial \theta} = \beta(s,a)$$

$$\theta \leftarrow \theta + \alpha \left(R(s,a) + \gamma \max_{a'} \theta^T \beta(s',a') - \theta^T \beta(s,a) \right) \beta(s,a)$$

Q



$$Q(s,a) = \theta^T \beta(s,a)$$