Announcement

ASEN 6519 Advanced Topics in Sequential Decision Making

Projects due 5pm.

DMU
- Prob. Models
- MDPs
- RL
- POMDPs
- Games

$P(A)$
$P(A, B)$
$P(A|B)$

1) $0 \leq P(x|Y) \leq 1$
$\sum_{x \in X} P(x|Y) = 1$

2) $P(X) = \sum_{Y} P(X, Y)$

3) $P(x|Y) = \frac{P(x,Y)}{P(Y)}$

Bayes Rule

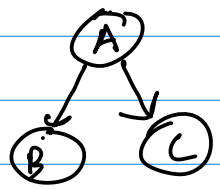
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Independence

$A \perp B \iff P(A,B) = P(A)\,P(B)$

$A \perp B|C \iff P(A,B|C) = P(A|C)\,P(B|C)$

Bayes Nets



$P(x|all) = P(x|Pa(x))$

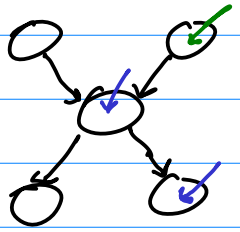
$X \perp Y|G$ if all paths between X and Y are d-separated by G

Sampling

Topological Sort then sample from each node

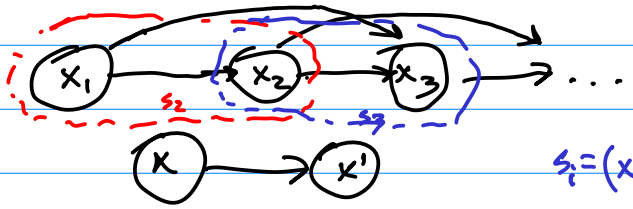
Inference

Given: BN, values of some variables
Output: Distributions of target Variables



Exact: NP-hard
Approximate: - Direct - Sampling ← Unweighted PF
- Likelihood-weighted Sampling ← Weighted PF
- Gibb's method

Stochastic Process

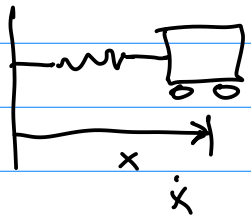


$s_i = (x_i, x_{i-1})$

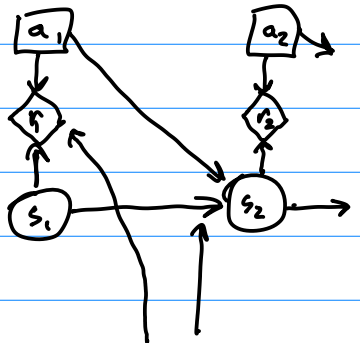
Markov Property

$P(X_t | X_{t-1}, X_{t-2} \dots X_1) = P(X_t | X_{t-1})$

Dyn. Bayes Net



Decision Network



$s_3 \perp a_1$? No

$s_3 \perp a_1 | s_2$? Yes

$r_k \perp a_1 | s_2$ $\forall k \geq 2$

$V(s)$

MDP

(S, A, R, T, γ)

$S = \{1, 2, 3\}$

$s = (x, \dot{x}) \in S = \mathbb{R}^2$

$R(s,a) = \mathop{E}_{s'}[R(s,a,s')]$

$T(s'|s,a)$

$$\underset{\pi: S \to A}{\text{maximize}} \; E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$$ Hard

$$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_1 = s\right]$$

Bellman's Equation

$$V^*(s) = \max_a \left(R(s,a) + \gamma E[V^*(s') \mid s,a]\right)$$

$$E[V^*(s') \mid s,a] = \sum_{s' \in S} T(s'|s,a) V^*(s')$$

Value Iteration

loop

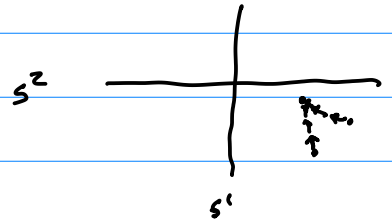
for $s \in S$

$$V^{k+1}(s) = \max_a \left(R(s,a) + \gamma E[V^*(s') \mid s,a]\right)$$

Converges because Bellman Operator

$$B[V](s) = \max_a \left(R(s,a) + \gamma E[V(s')]\right)$$

is a contraction in infinity norm



Policy Iteration

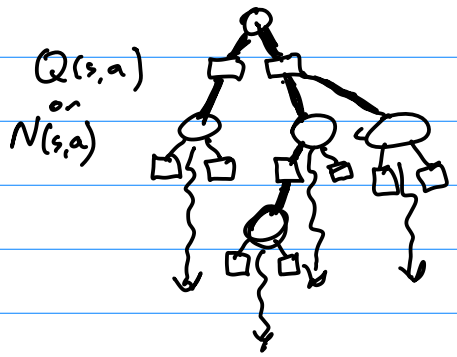
loop

Evaluate Policy

Improve Policy

Online Planning MCTS

Search Expand Rollout Backup



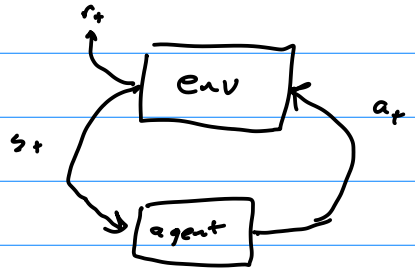
$$Q(s,a) + c\sqrt{\frac{\log N(s)}{N(s,a)}}$$

simulate(...)

$$q \leftarrow r_t + \gamma \text{ simulate}(\dots)$$

$$Q(s,a) \mathrel{+}= \frac{(q - Q(s,a))}{N(s,a)}$$

Reinforcement Learning



$\theta_1, \theta_2, \dots \theta_n$

Exploration vs. Exploitation: Multi-Armed Bandit

- ε-greedy
- softmax
- UCB

Thompson Sampling

Optimal DP Solution

↑ POMDP!

$R(s,a) = \theta_a$

$s = (\theta_1, \dots \theta_n)$

$O = \{0,1\}$

$Z(1 \mid a, s') = \theta_a$

Montezuma's Revenge

Advanced Exploration

- Pseudo Counts $N(s,a)$
- Curiosity : bonus reward if dynamics hard to predict
- Random Network Distillation

RL Algorithms

(S, A, R, γ)

POMDP
$\tilde{S} = (s, T)$

Model Based ⟷ Model-Free

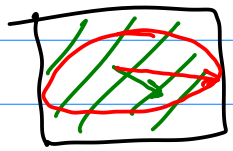
MLMBRL
BAMDP

Model-Free
- Learn Q
  - SARSA — On-Policy
  - Q-Learning — Off-Policy
- Actor Critic — π
- Learn π — Policy Gradient

Policy Gradient

- Likelihood Ratio Trick
- Causality
- Baseline Subtraction
- Natural Gradient

$$\nabla_\theta p_\theta(\tau) = p_\theta(\tau) \log p_\theta(\tau)$$



KL-divergence

Q-learning

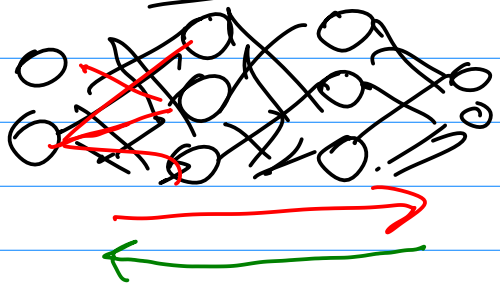
$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r_t + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$$

Double Q

Sarsa

Eligibility Traces

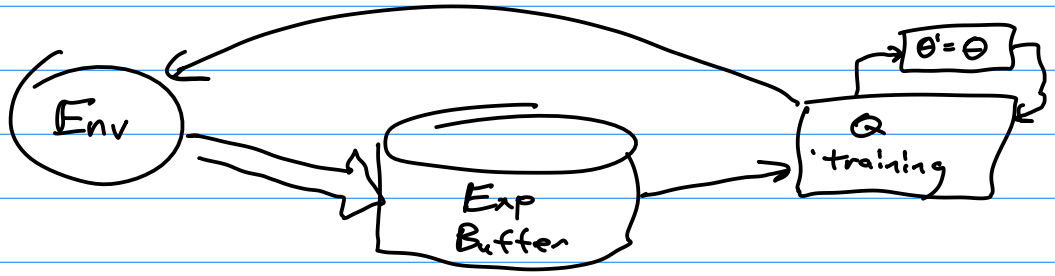
Neural Network



$$f_\theta(x) = W_3\sigma\left(W_2\,\sigma\left(W_1 x + b_1\right) + b_2\right) + b_3$$

Back prop

DQN



$Q(s,a) \rightarrow \mathbb{R}$

1. Experience Buffer
2. Periodically freeze target
3. Q-network outputs values for all actions

Rainbow

POMDPs

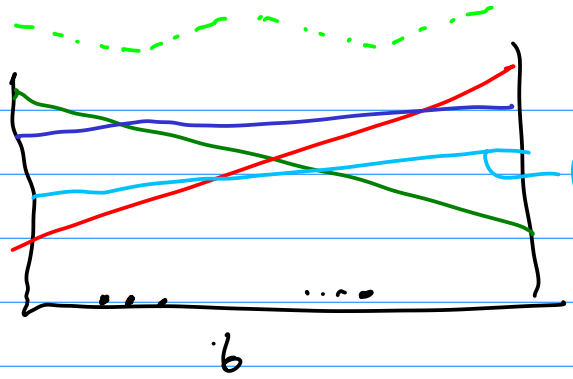
$(S, A, T, R, O, Z, \gamma)$

Belief-Space MDP

Belief-Updates

- Discrete Bayes Filter
- Particle Filter

Particle Depletion



PBVI
SARSOP ←

POMDP is PSPACE Complete

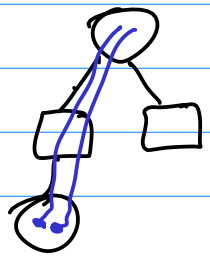
Formulation Approximation

- Certainty Equivalent ← Optimal LQG
- QMDP

Online POMDP Methods

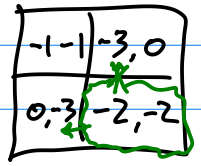
POMCP/PO-UCT/MCTS

DESPOT
↑determinized scenarios



Games

X Optimal Solution
✓ Nash Equilibria Pure, Mixed



Markov Games
POMG / POSG / Extensive Form Games

Imitation
IRL
Transfer Learning
Meta Learning (POMDP)

Successor Features
MAML θ so that 1 step get to optimal

Big Problems

	Unc (Outcome/Action)	Tool
1. Immediate + Future Rewards	Aleatoric	Value
2. Unknown Models	Model/Static Epistemic	RL
3. Partial Observability	State/Dyn. Epistemic	POMDP
4. Other Agents	Interaction Unc.	Games