

# CS885 Reinforcement Learning

## Lecture 7a: May 23, 2018

Policy Gradient Methods

[SutBar] Sec. 13.1-13.3, 13.7

[SigBuf] Sec. 5.1-5.2, [RusNor] Sec. 21.5

# Outline

- Stochastic policy gradient
  - REINFORCE algorithm
- AlphaGo

# Model-free Policy-based Methods

- Q-learning
  - **Model-free value-based method**
  - **No explicit policy representation**
- Policy gradient
  - **Model-free policy-based method**
  - **No explicit value function representation**

# Stochastic Policy

- Consider stochastic policy  $\pi_{\theta}(a|s) = \Pr(a|s; \theta)$  parametrized by  $\theta$ .
- Finitely many discrete actions

$$\text{Softmax: } \pi_{\theta}(a|s) = \frac{\exp(h(s,a;\theta))}{\sum_{a'} \exp(h(s,a';\theta))}$$

where  $h(s, a; \theta)$  might be

$$\text{linear in } \theta: h(s, a; \theta) = \sum_i \theta_i f_i(s, a)$$

$$\text{or non-linear in } \theta: h(s, a; \theta) = \text{neuralNet}(s, a; \theta)$$

- Continuous actions:

$$\text{Gaussian: } \pi_{\theta}(a|s) = N(a|\mu(s; \theta), \Sigma(s; \theta))$$

# Supervised Learning

- Consider a stochastic policy  $\pi_{\theta}(a|s)$
- Data: state-action pairs  $\{(s_1, a_1^*), (s_2, a_2^*), \dots\}$
- Maximize log likelihood of the data

$$\theta^* = \operatorname{argmax}_{\theta} \sum_n \log \pi_{\theta}(a_n^* | s_n)$$

- Gradient update

$$\theta_{n+1} \leftarrow \theta_n + \alpha_n \nabla_{\theta} \log \pi_{\theta}(a_n^* | s_n)$$

# Reinforcement Learning

- Consider a stochastic policy  $\pi_{\theta}(a|s)$

- Data: state-action-reward triples  
 $\{(s_1, a_1, r_1), (s_2, a_2, r_2), \dots\}$

- Maximize discounted sum of rewards

$$\theta^* = \operatorname{argmax}_{\theta} \sum_n \gamma^n E_{\theta}[r_n | s_n, a_n]$$

- Gradient update

$$\theta_{n+1} \leftarrow \theta_n + \alpha_n \gamma^n G_n \nabla_{\theta} \log \pi_{\theta}(a_n | s_n)$$

where  $G_n = \sum_{t=0}^{\infty} \gamma^t r_{n+t}$

# Stochastic Gradient Policy Theorem

- Stochastic Gradient Policy Theorem

$$\nabla V_{\theta}(s_0) \propto \sum_s \mu_{\theta}(s) \sum_a \nabla \pi_{\theta}(a|s) Q_{\theta}(s, a)$$

$\mu_{\theta}(s)$ : stationary state distribution when executing policy parametrized by  $\theta$

$Q_{\theta}(s, a)$ : discounted sum of rewards when starting in  $s$ , executing  $a$  and following the policy parametrized by  $\theta$  thereafter.

# Derivation

$$\begin{aligned}
 \nabla V_\theta(s_0) &= \nabla \left[ \sum_{a_0} \pi_\theta(a_0|s_0) Q_\theta(s_0, a_0) \right] \quad \forall s_0 \in \mathcal{S} \\
 &= \sum_{a_0} \left[ \nabla \pi_\theta(a_0|s_0) Q_\theta(s_0, a_0) + \pi_\theta(a_0|s_0) \nabla Q_\theta(s_0, a_0) \right] \\
 &= \sum_{a_0} \left[ \nabla \pi_\theta(a_0|s_0) Q_\theta(s_0, a_0) + \pi_\theta(a_0|s_0) \nabla \sum_{s_1, r_0} \Pr(s_1, r_0|s_0, a_0) (r_0 + \gamma V_\theta(s_1)) \right] \\
 &= \sum_{a_0} \left[ \nabla \pi_\theta(a_0|s_0) Q_\theta(s_0, a_0) + \pi_\theta(a_0|s_0) \sum_{s_1} \gamma \Pr(s_1|s_0, a_0) \nabla V_\theta(s_1) \right] \\
 &= \sum_{a_0} \left[ \nabla \pi_\theta(a_0|s_0) Q_\theta(s_0, a_0) + \pi_\theta(a_0|s_0) \sum_{s_1} \gamma \Pr(s_1|s_0, a_0) \right. \\
 &\quad \left. \sum_{a_1} \left[ \nabla \pi_\theta(a_1|s_1) Q_\theta(s_1, a_1) + \pi_\theta(a_1|s_1) \sum_{s_2} \gamma \Pr(s_2|s_1, a_1) \nabla V_\theta(s_2) \right] \right] \\
 &= \sum_{s \in \mathcal{S}} \underbrace{\sum_{n=0}^{\infty} \gamma^n \Pr(s_0 \rightarrow s; n, \theta)}_{\text{Probability of reaching } s \text{ from } s_0 \text{ at time step } n} \sum_a \nabla \pi_\theta(a|s) Q_\theta(s, a)
 \end{aligned}$$

Probability of reaching  $s$  from  $s_0$  at time step  $n$

Since  $\mu_\theta(s) \propto \sum_{n=0}^{\infty} \gamma^n \Pr(s_0 \rightarrow s; n, \theta)$  then

$$\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s) Q_\theta(s, a)$$



# REINFORCE: Monte Carlo Policy Gradient

- $$\begin{aligned}\nabla V_\theta(s_0) &= \sum_{s \in \mathcal{S}} \sum_{n=0}^{\infty} \gamma^n \Pr(s_0 \rightarrow s; n, \theta) \sum_a \nabla \pi_\theta(a|s) Q_\theta(s, a) \\ &= E_\theta \left[ \sum_{n=0}^{\infty} \gamma^n \sum_a Q_\theta(S_n, a) \nabla \pi_\theta(a|S_n) \right] \\ &= E_\theta \left[ \sum_{n=0}^{\infty} \gamma^n \sum_a \pi_\theta(a|S_n) Q_\theta(S_n, a) \frac{\nabla \pi_\theta(a|S_n)}{\pi_\theta(a|S_n)} \right] \\ &= E_\theta \left[ \sum_{n=0}^{\infty} \gamma^n Q_\theta(S_n, A_n) \frac{\nabla \pi_\theta(A_n|S_n)}{\pi_\theta(A_n|S_n)} \right] \\ &= E_\theta \left[ \sum_{n=0}^{\infty} \gamma^n G_n \frac{\nabla \pi_\theta(A_n|S_n)}{\pi_\theta(A_n|S_n)} \right] \\ &= E_\theta \left[ \sum_{n=0}^{\infty} \gamma^n G_n \nabla \log \pi_\theta(A_n|S_n) \right]\end{aligned}$$
- Stochastic gradient at time step  $n$ 
$$\nabla V_\theta \approx \gamma^n G_n \nabla \log \pi_\theta(a_n|s_n)$$

# REINFORCE Algorithm (stochastic policy)

**REINFORCE**( $s_0, \pi_\theta$ )

Initialize  $\pi_\theta$  to anything

Loop forever (for each episode)

Generate episode  $s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T, a_T, r_T$  with  $\pi_\theta$

Loop for each step of the episode  $n = 0, 1, \dots, T$

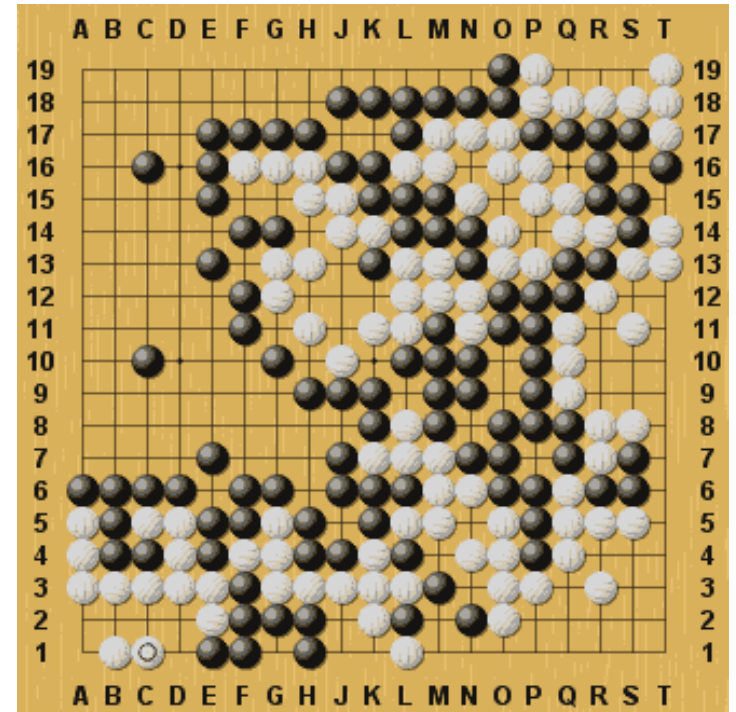
$$G_n \leftarrow \sum_{t=0}^{T-n} \gamma^t r_{n+t}$$

Update policy:  $\theta \leftarrow \theta + \alpha \gamma^n G_n \nabla \log \pi_\theta(a_n | s_n)$

Return  $\pi_\theta$

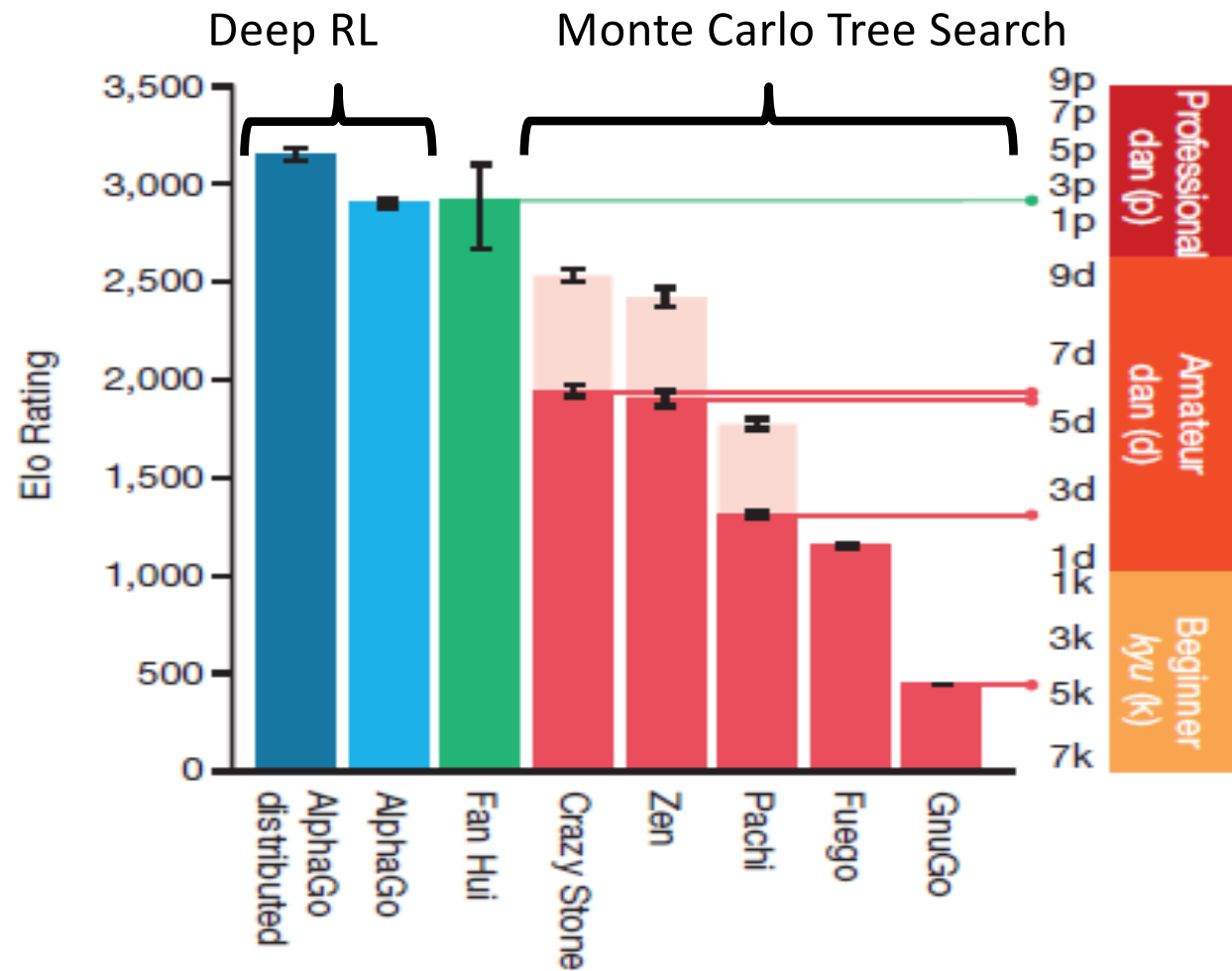
# Example: Game of Go

- (simplified) rules:
  - Two players (black and white)
  - Players alternate to place a stone of their color on a vacant intersection.
  - Connected stones without any liberty (i.e., no adjacent vacant intersection) are captured and removed from the board
  - Winner: player that controls the largest number of intersections at the end of the game



# Computer Go

- Oct 2015:



# Computer Go

- March 2016: AlphaGo defeats Lee Sedol (9-dan)

*“[AlphaGo] can’t beat me”* Ke Jie (world champion)

- May 2017: AlphaGo defeats Ke Jie (world champion)

*“Last year, [AlphaGo] was still quite humanlike when it played. But this year, it became like a god of Go”* Ke Jie (world champion)

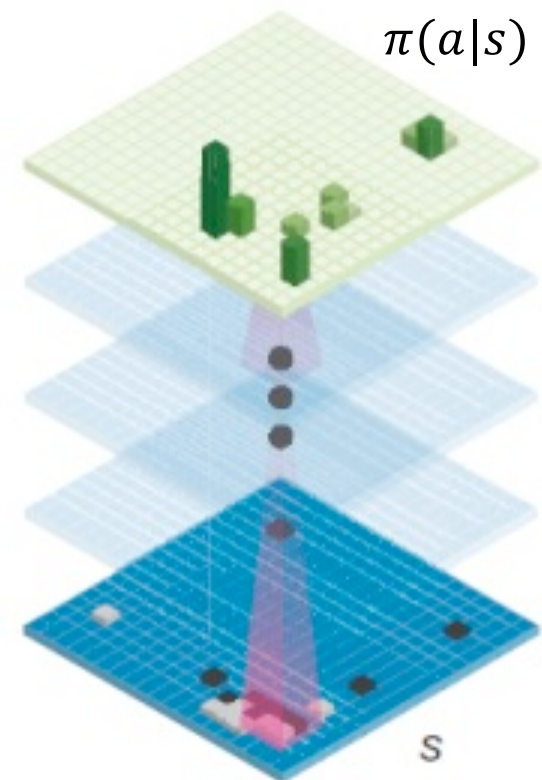
# Winning Strategy

Four steps:

1. Supervised Learning of Policy Networks
2. Policy gradient with Policy Networks
3. Value gradient with Value Networks
4. Searching with Policy and Value Networks

# Policy Network

- Train policy network to imitate Go experts based on a database of 30 million board configurations from the KGS Go Server.
- Policy network:  $\pi(a|s)$ 
  - Input: state  $s$   
(board configuration)
  - Output: distribution over actions  $a$   
(intersection on which the next stone will be placed)



# Supervised Learning of the Policy Network

- Let  $\theta$  be the weights of the policy network
- Training:
  - Data: suppose  $a$  is optimal in  $s$
  - Objective: maximize  $\log \pi_{\theta}(a|s)$
  - Gradient:  $\nabla \theta = \frac{\partial \log \pi_{\theta}(a|s)}{\partial \theta}$
  - Weight update:  $\theta \leftarrow \theta + \alpha \nabla \theta$



# Policy gradient for the Policy Network

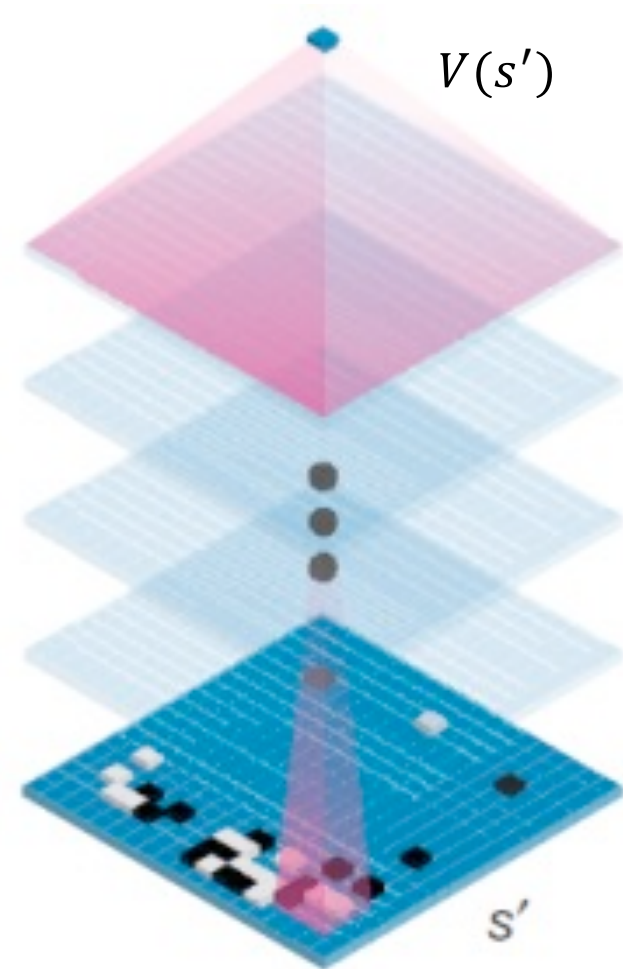
- How can we update a policy network based on reinforcements instead of the optimal action?
- Let  $G_n = \sum_t \gamma^t r_{n+t}$  be the discounted sum of rewards in a trajectory that starts in  $s$  at time  $n$  by executing  $a$ .
- Gradient:  $\nabla \theta = \frac{\partial \log \pi_\theta(a|s)}{\partial \theta} \gamma^n G_n$ 
  - Intuition rescale supervised learning gradient by  $G_n$
- Policy update:  $\theta \leftarrow \theta + \alpha \nabla \theta$

# Policy gradient for the Policy Network

- In computer Go, program repeatedly plays games against its former self.
- For each game  $G_n = \begin{cases} 1 & \text{win} \\ -1 & \text{lose} \end{cases}$
- For each  $(s_n, a_n)$  at turn  $n$  of the game, assume  $\gamma = 1$  and compute
  - Gradient:  $\nabla \theta = \frac{\partial \log \pi_{\theta}(a|s)}{\partial \theta} \gamma^n G_n$
  - Policy update:  $\theta \leftarrow \theta + \alpha \nabla \theta$

# Value Network

- Predict  $V(s')$  (i.e., who will win game) in each state  $s'$  with a value network
  - Input: state  $s$  (board configuration)
  - Output: expected discounted sum of rewards  $V(s')$

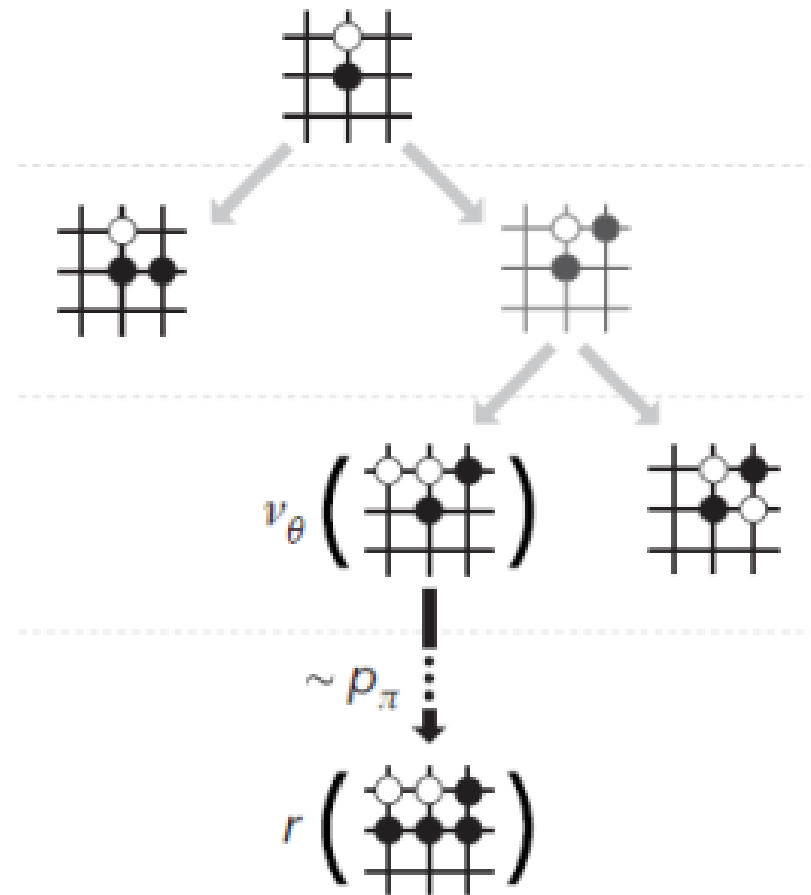


# Gradient Value Learning with Value Networks

- Let  $\mathbf{w}$  be the weights of the value network
- Training:
  - Data:  $(s, G)$  where  $G = \begin{cases} 1 & \text{win} \\ -1 & \text{lose} \end{cases}$
  - Objective: minimize  $\frac{1}{2}(V_{\mathbf{w}}(s) - G)^2$
  - Gradient:  $\nabla \mathbf{w} = \frac{\partial V_{\mathbf{w}}(s)}{\partial \mathbf{w}} (V_{\mathbf{w}}(s) - G)$
  - Weight update:  $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla \mathbf{w}$

# Searching with Policy and Value Networks

- AlphaGo combines policy and value networks into a **Monte Carlo Tree Search (MCTS)** algorithm
- Idea: construct a search tree
  - Node:  $s$
  - Edge:  $a$
- We will discuss MCTS in a few lectures



# Competition

**Extended Data Table 1 | Details of match between AlphaGo and Fan Hui**

Date	Black	White	Category	Result
5/10/15	Fan Hui	<i>AlphaGo</i>	Formal	<i>AlphaGo</i> wins by 2.5 points
5/10/15	Fan Hui	<i>AlphaGo</i>	Informal	Fan Hui wins by resignation
6/10/15	<i>AlphaGo</i>	Fan Hui	Formal	<i>AlphaGo</i> wins by resignation
6/10/15	<i>AlphaGo</i>	Fan Hui	Informal	<i>AlphaGo</i> wins by resignation
7/10/15	Fan Hui	<i>AlphaGo</i>	Formal	<i>AlphaGo</i> wins by resignation
7/10/15	Fan Hui	<i>AlphaGo</i>	Informal	<i>AlphaGo</i> wins by resignation
8/10/15	<i>AlphaGo</i>	Fan Hui	Formal	<i>AlphaGo</i> wins by resignation
8/10/15	<i>AlphaGo</i>	Fan Hui	Informal	<i>AlphaGo</i> wins by resignation
9/10/15	Fan Hui	<i>AlphaGo</i>	Formal	<i>AlphaGo</i> wins by resignation
9/10/15	<i>AlphaGo</i>	Fan Hui	Informal	Fan Hui wins by resignation

The match consisted of five formal games with longer time controls, and five informal games with shorter time controls. Time controls and playing conditions were chosen by Fan Hui in advance of the match.