Reinforcement Learning for Business, Economics, and Social Sciences

Unit 4-4: Policy Gradient Methods

Davud Rostam-Afschar (Uni Mannheim)

If I nudge my policy a little, can I win more often?

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 a stochastic policy

$$\pi_\theta(a \mid s) = \mathbb{P}(a \mid s; \theta),$$

parameterized by θ.

Stochastic Policy

Consider a stochastic policy

$$\pi_\theta(a \mid s) = \mathbb{P}(a \mid s; \theta),$$

parameterized by θ.

Discrete actions: SoftMax exploration

$$\pi_\theta(a \mid s) = \frac{\exp\big(h(s, a; \theta)\big)}{\sum_{a'} \exp\big(h(s, a'; \theta)\big)},$$

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

- *linear* in θ: $h(s, a; \theta) = \sum_i \theta_i f_i(s, a)$
- *nonlinear* in θ: $h(s, a; \theta) = \mathtt{NeuralNet}(s, a; \theta)$

Stochastic Policy

Consider a stochastic policy

$$\pi_\theta(a \mid s) = \mathbb{P}(a \mid s; \theta),$$

parameterized by θ.

Discrete actions: SoftMax exploration

$$\pi_\theta(a \mid s) = \frac{\exp\big(h(s, a; \theta)\big)}{\sum_{a'} \exp\big(h(s, a'; \theta)\big)},$$

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

- *linear* in θ: $h(s, a; \theta) = \sum_i \theta_i f_i(s, a)$
- *nonlinear* in θ: $h(s, a; \theta) = \mathrm{NeuralNet}(s, a; \theta)$

Continuous actions: Gaussian

$$\pi_\theta(a \mid s) = \mathcal{N}\big(a \mid \mu(s; \theta), \Sigma(s; \theta)\big).$$

Stochastic Gradient Policy

- Consider a stochastic policy

$$\pi_\theta(a \mid s)$$

- Data: state–action pairs $\{(s_1, a_1^*), (s_2, a_2^*), \dots\}$
- Maximize log-likelihood of the data:

$$\theta^* = \arg\max_\theta \sum_n \log \pi_\theta(a_n^* \mid s_n).$$

- Policy gradient update:

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

Reinforcement Learning

- Consider a stochastic policy

$$\pi_\theta(a \mid s)$$

- Data: state–action–reward triples $\{(s_1, a_1, r_1), (s_2, a_2, r_2), \ldots\}$
- Maximize expected discounted return

$$\theta^* = \arg\max_\theta \sum_n \gamma^n \mathbb{E}_\theta[r_n \mid s_n, a_n].$$

- Stochastic policy gradient update:

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

- Consider a stochastic policy

$$\pi_\theta(a \mid s)$$

- Data: state–action–reward triples $\{(s_1, a_1, r_1), (s_2, a_2, r_2), \ldots\}$
- Maximize expected discounted return

$$\theta^* = \arg\max_\theta \sum_n \gamma^n \mathbb{E}_\theta[r_n \mid s_n, a_n].$$

- Stochastic policy gradient update:

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

- Consider a stochastic policy

$$\pi_\theta(a \mid s)$$

- Data: state–action–reward triples $\{(s_1, a_1, r_1), (s_2, a_2, r_2), \ldots\}$
- Maximize expected discounted return

$$\theta^* = \arg\max_\theta \sum_n \gamma^n \mathbb{E}_\theta[r_n \mid s_n, a_n].$$

- Stochastic policy gradient update:

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

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

- **Stochastic Gradient Policy Theorem**

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

- $\mu_\theta(s)$: stationary state distribution when executing policy parametrized by θ
- $Q_\theta(s, a)$: discounted sum of rewards when starting in s, executing a and following the policy parametrized by θ thereafter.

Proof of the Policy Gradient Theorem (episodic case)

$$\nabla V_\theta(s) \;=\; \nabla \left[\sum_a \pi_\theta(a \mid s)\, Q_\theta(s, a) \right], \text{for all } s \in \mathcal{S}$$

Proof of the Policy Gradient Theorem (episodic case)

$$
\begin{aligned}
\nabla V_\theta(s) &= \nabla\left[\sum_a \pi_\theta(a \mid s)\, Q_\theta(s,a)\right], \text{for all } s \in \mathcal{S} \\
&= \sum_a \left[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s)\, \nabla Q_\theta(s,a)\right] \text{ (product rule)}
\end{aligned}
$$

Proof of the Policy Gradient Theorem (episodic case)

$$
\begin{aligned}
\nabla V_\theta(s) \;&=\; \nabla\left[\sum_a \pi_\theta(a \mid s)\, Q_\theta(s,a)\right], \text{for all } s \in \mathcal{S} \\
&= \sum_a \left[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s)\, \nabla Q_\theta(s,a)\right] \text{ (product rule)} \\
&= \sum_a \left[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \nabla \sum_{s',\, r} \mathbb{P}(s', r \mid s, a) \left(r + \gamma V_\theta(s')\right)\right]
\end{aligned}
$$

Proof of the Policy Gradient Theorem (episodic case)

$$
\begin{aligned}
\nabla V_\theta(s) &= \nabla \Big[\sum_a \pi_\theta(a \mid s)\, Q_\theta(s,a)\Big], \text{ for all } s \in \mathcal{S} \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s)\, \nabla Q_\theta(s,a)\Big] \text{ (product rule)} \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \nabla \sum_{s',\,r} \mathbb{P}(s', r \mid s, a)\,\big(r + \gamma V_\theta(s')\big)\Big] \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \sum_{s'} \gamma \mathbb{P}(s' \mid s, a)\, \nabla V_\theta(s')\Big]
\end{aligned}
$$

Proof of the Policy Gradient Theorem (episodic case)

$$
\begin{aligned}
\nabla V_\theta(s) &= \nabla\left[\sum_a \pi_\theta(a \mid s)\, Q_\theta(s,a)\right], \text{for all } s \in \mathcal{S} \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s)\, \nabla Q_\theta(s,a)\Big] \text{ (product rule)} \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \nabla \sum_{s',\,r} \mathbb{P}(s', r \mid s, a)\, \big(r + \gamma V_\theta(s')\big)\Big] \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \sum_{s'} \gamma \mathbb{P}(s' \mid s, a)\, \nabla V_\theta(s')\Big] \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \sum_{s'} \gamma \mathbb{P}(s' \mid s, a) \\
&\qquad \sum_{a'} \Big[\nabla \pi_\theta(a' \mid s')\, Q_\theta(s',a') \;+\; \pi_\theta(a' \mid s') \sum_{s''} \gamma \mathbb{P}(s'' \mid s', a')\, \nabla V_\theta(s'')\Big]\Big] \text{ (unrolling)}
\end{aligned}
$$

Proof of the Policy Gradient Theorem (episodic case)

$$
\begin{aligned}
\nabla V_\theta(s) &= \nabla\left[\sum_a \pi_\theta(a \mid s)\, Q_\theta(s,a)\right], \text{for all } s \in \mathcal{S} \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s)\, \nabla Q_\theta(s,a)\Big] \text{ (product rule)} \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \nabla \sum_{s',\, r} \mathbb{P}(s', r \mid s, a)\, \big(r + \gamma V_\theta(s')\big)\Big] \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \sum_{s'} \gamma \mathbb{P}(s' \mid s, a)\, \nabla V_\theta(s')\Big] \\
&= \sum_a \Big[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \sum_{s'} \gamma \mathbb{P}(s' \mid s, a) \\
&\qquad \sum_{a'} \Big[\nabla \pi_\theta(a' \mid s')\, Q_\theta(s',a') \;+\; \pi_\theta(a' \mid s') \sum_{s''} \gamma \mathbb{P}(s'' \mid s', a')\, \nabla V_\theta(s'')\Big]\Big] \text{ (unrolling)} \\
&= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \gamma^k \mathbb{P}\big(s \to x, k, \pi\big) \; \sum_a \nabla \pi(a \mid x)\, Q_\theta(x,a),
\end{aligned}
$$

Proof of the Policy Gradient Theorem (episodic case)

$$
\begin{aligned}
\nabla V_\theta(s) &= \nabla\left[\sum_a \pi_\theta(a \mid s)\, Q_\theta(s,a)\right], \text{for all } s \in \mathcal{S} \\
&= \sum_a \left[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s)\, \nabla Q_\theta(s,a)\right] \text{ (product rule)} \\
&= \sum_a \left[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \nabla \sum_{s',\, r} \mathbb{P}(s', r \mid s, a)\, \Big(r + \gamma V_\theta(s')\Big)\right] \\
&= \sum_a \left[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \sum_{s'} \gamma \mathbb{P}(s' \mid s, a)\, \nabla V_\theta(s')\right] \\
&= \sum_a \Bigg[\nabla \pi_\theta(a \mid s)\, Q_\theta(s,a) \;+\; \pi_\theta(a \mid s) \sum_{s'} \gamma \mathbb{P}(s' \mid s, a) \\
&\qquad \sum_{a'} \left[\nabla \pi_\theta(a' \mid s')\, Q_\theta(s', a') \;+\; \pi_\theta(a' \mid s') \sum_{s''} \gamma \mathbb{P}(s'' \mid s', a')\, \nabla V_\theta(s'')\right]\Bigg] \text{ (unrolling)} \\
&= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \gamma^k \mathbb{P}\Big(s \to x, k, \pi\Big) \;\sum_a \nabla \pi(a \mid x)\, Q_\theta(x,a),
\end{aligned}
$$

with $\mathbb{P}(s \to x, k, \pi)$ of transitioning from state s to state x in k steps under policy π.

Proof of the Policy Gradient Theorem (episodic case)

$$\nabla V_\theta(s_0) = \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \gamma^k \mathbb{P}(s_0 \to x, k, \theta) \sum_a \nabla \pi_\theta(a \mid x) \, Q_\theta(x, a)$$

Proof of the Policy Gradient Theorem (episodic case)

$$
\begin{aligned}
\nabla V_\theta(s_0) &= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \gamma^k \mathbb{P}(s_0 \to x, k, \theta) \sum_a \nabla \pi_\theta(a \mid x)\, Q_\theta(x, a) \\
&\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a \mid s)\, Q_\theta(s, a).
\end{aligned}
$$

Stochastic gradient

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

Stochastic gradient

$$
\begin{aligned}
\nabla V_\theta &\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s)\, Q_\theta(s,a) \\
&= \mathbb{E}_\theta\left[\gamma^n \sum_a Q_\theta(S_n, a)\, \nabla \pi_\theta(a|S_n)\right]
\end{aligned}
$$

Stochastic gradient

$$\begin{aligned}\nabla V_\theta &\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s)\, Q_\theta(s,a) \\ &= \mathbb{E}_\theta\Big[\gamma^n \sum_a Q_\theta(S_n,a)\, \nabla \pi_\theta(a|S_n)\Big]\end{aligned}$$

Stochastic gradient

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

$$= \mathbb{E}_\theta\Big[\gamma^n \sum_a Q_\theta(S_n,a)\, \nabla \pi_\theta(a|S_n)\Big]$$

$$= \mathbb{E}_\theta\Big[\gamma^n \sum_a \pi_\theta(a|S_n)\, Q_\theta(S_n,a)\, \frac{\nabla \pi_\theta(a|S_n)}{\pi_\theta(a|S_n)}\Big]$$

Stochastic gradient

$$\begin{aligned}
\nabla V_\theta &\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s)\, Q_\theta(s,a) \\
&= \mathbb{E}_\theta\Big[\gamma^n \sum_a Q_\theta(S_n,a)\, \nabla \pi_\theta(a|S_n)\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n \sum_a \pi_\theta(a|S_n)\, Q_\theta(S_n,a)\, \frac{\nabla \pi_\theta(a|S_n)}{\pi_\theta(a|S_n)}\Big]
\end{aligned}$$

Stochastic gradient

$$
\begin{aligned}
\nabla V_\theta &\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s)\, Q_\theta(s, a) \\
&= \mathbb{E}_\theta\left[\gamma^n \sum_a Q_\theta(S_n, a)\, \nabla \pi_\theta(a|S_n)\right] \\
&= \mathbb{E}_\theta\left[\gamma^n \sum_a \pi_\theta(a|S_n)\, Q_\theta(S_n, a)\, \frac{\nabla \pi_\theta(a|S_n)}{\pi_\theta(a|S_n)}\right] \\
&= \mathbb{E}_\theta\left[\gamma^n\, Q_\theta(S_n, A_n)\, \frac{\nabla \pi_\theta(A_n|S_n)}{\pi_\theta(A_n|S_n)}\right]
\end{aligned}
$$

Stochastic gradient

$$
\begin{aligned}
\nabla V_\theta &\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s)\, Q_\theta(s,a) \\
&= \mathbb{E}_\theta\Big[\gamma^n \sum_a Q_\theta(S_n,a)\, \nabla \pi_\theta(a|S_n)\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n \sum_a \pi_\theta(a|S_n)\, Q_\theta(S_n,a)\, \frac{\nabla \pi_\theta(a|S_n)}{\pi_\theta(a|S_n)}\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n\, Q_\theta(S_n,A_n)\, \frac{\nabla \pi_\theta(A_n|S_n)}{\pi_\theta(A_n|S_n)}\Big]
\end{aligned}
$$

Stochastic gradient

$$
\begin{aligned}
\nabla V_\theta &\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s)\, Q_\theta(s,a) \\
&= \mathbb{E}_\theta\Big[\gamma^n \sum_a Q_\theta(S_n, a)\, \nabla \pi_\theta(a|S_n)\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n \sum_a \pi_\theta(a|S_n)\, Q_\theta(S_n, a)\, \frac{\nabla \pi_\theta(a|S_n)}{\pi_\theta(a|S_n)}\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n\, Q_\theta(S_n, A_n)\, \frac{\nabla \pi_\theta(A_n|S_n)}{\pi_\theta(A_n|S_n)}\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n\, G_n\, \nabla \log \pi_\theta(A_n|S_n)\Big]
\end{aligned}
$$

note that $Q_\theta(S_n, A_n) = \mathbb{E}_\theta[G_n|S_n, A_n]$.

Stochastic gradient

$$\begin{aligned}
\nabla V_\theta &\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s)\, Q_\theta(s,a) \\
&= \mathbb{E}_\theta\left[\gamma^n \sum_a Q_\theta(S_n,a)\, \nabla \pi_\theta(a|S_n)\right] \\
&= \mathbb{E}_\theta\left[\gamma^n \sum_a \pi_\theta(a|S_n)\, Q_\theta(S_n,a)\, \frac{\nabla \pi_\theta(a|S_n)}{\pi_\theta(a|S_n)}\right] \\
&= \mathbb{E}_\theta\left[\gamma^n\, Q_\theta(S_n,A_n)\, \frac{\nabla \pi_\theta(A_n|S_n)}{\pi_\theta(A_n|S_n)}\right] \\
&= \mathbb{E}_\theta\left[\gamma^n\, G_n\, \nabla \log \pi_\theta(A_n|S_n)\right]
\end{aligned}$$

Stochastic gradient

$$\begin{aligned}
\nabla V_\theta &\propto \sum_s \mu_\theta(s) \sum_a \nabla \pi_\theta(a|s)\, Q_\theta(s,a) \\
&= \mathbb{E}_\theta\Big[\gamma^n \sum_a Q_\theta(S_n,a)\, \nabla \pi_\theta(a|S_n)\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n \sum_a \pi_\theta(a|S_n)\, Q_\theta(S_n,a)\, \frac{\nabla \pi_\theta(a|S_n)}{\pi_\theta(a|S_n)}\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n\, Q_\theta(S_n,A_n)\, \frac{\nabla \pi_\theta(A_n|S_n)}{\pi_\theta(A_n|S_n)}\Big] \\
&= \mathbb{E}_\theta\Big[\gamma^n\, G_n\, \nabla \log \pi_\theta(A_n|S_n)\Big]
\end{aligned}$$

Stochastic gradient at time step n:

$$\nabla V_\theta \approx \gamma^n\, G_n\, \nabla \log \pi_\theta(A_n|S_n)$$

Stochastic gradient

Stochastic gradient at time step n:

$$\nabla V_\theta \approx \gamma^n \, G_n \, \nabla \log \pi_\theta(A_n | S_n)$$

Discounted returns times direction that most increases the probability of repeating the action A_n on future visits to state S_n.

REINFORCE Algorithm

REINFORCE: Monte Carlo Policy Gradient

REINFORCE(s_0, π_θ)

Initialize π_θ 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 π_θ

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

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

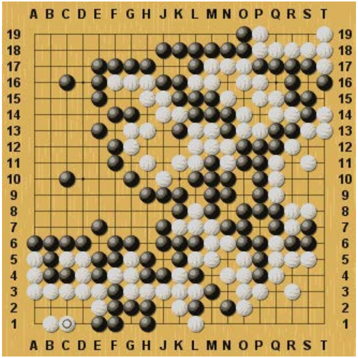
Update policy: $\theta \leftarrow \theta + \alpha \gamma^n G_n \nabla \log \pi_\theta(A_n|S_n)$

Return π_θ

Policy Gradient Methods in Practice

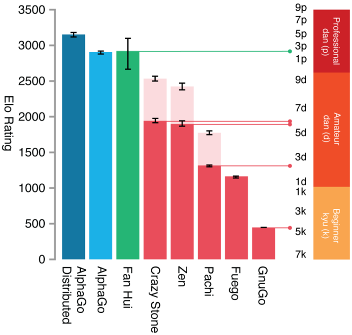
Example: Game of Go

- (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.



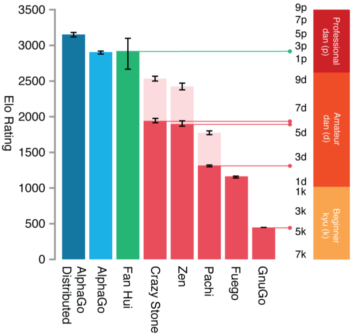
Computer Go

- Before: Monte Carlo Tree Search



Computer Go

- Before: Monte Carlo Tree Search
- Deep RL incl. policy gradient methods: AlphaGo



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)

Russell, S. J., and P. Norvig (2016): *Artificial intelligence: a modern approach*. Pearson.

Sigaud, O., and O. Buffet (2013): *Markov decision processes in artificial intelligence*. John Wiley & Sons, Available at https://zodml.org/sites/default/files/Markov_Decision_Processes_and_Artificial_Intelligence.pdf.

Sutton, R. S., and A. G. Barto (2018): "Reinforcement learning: An introduction," *A Bradford Book*, Available at http://incompleteideas.net/book/the-book-2nd.html.

Takeaways

Policy Gradient Methods

- Policy gradients directly optimize behavior to maximize rewards
- Stochastic policies explore actions with softmax or Gaussian distributions
- Good actions are reinforced by increasing their selection probability
- AlphaGo mastered Go by combining policy gradients, value networks, and search