Maximum Entropy Stochastic Control and Reinforcement Learning

Insoon Yang

Department of Electrical and Computer Engineering
Seoul National University

CORE
Control + Optimization Research Lab

So far...

- DQN, Double-DQN
- Policy gradient
- Actor-critic
- DDPG
- TRPO

Common issue: exploration

- Is there a value of exploring unknown regions of the environment?
- Which action should we try to explore unknown regions of the environment?

Tradeoff between exploration and exploitation

- Exploitation: Make the best decision given current information
- Exploration: Gather more information

Q) Why dilemma?

Tradeoff between exploration and exploitation

- Exploitation: Make the best decision given current information
- Exploration: Gather more information

Q) Why dilemma?

- The best long-term strategy may involve short-term sacrifices
- Need to gather enough information to make the best overall decisions

Examples

1. Restaurant Selection
    - Exploitation: Go to your favorite restaurant
    - Exploration: Try a new restaurant

Examples

1. Restaurant Selection
   - Exploitation: Go to your favorite restaurant
   - Exploration: Try a new restaurant

2. Online Banner Advertisements
   - Exploitation: Show the most successful advertisement
   - Exploration: Show a different advertisement

Examples

1. Restaurant Selection
   - Exploitation: Go to your favorite restaurant
   - Exploration: Try a new restaurant
2. Online Banner Advertisements
   - Exploitation: Show the most successful advertisement
   - Exploration: Show a different advertisement
3. Oil Drilling
   - Exploitation: Drill at the best known location
   - Exploration: Drill at a new location

Examples

1. Restaurant Selection
   - Exploitation: Go to your favorite restaurant
   - Exploration: Try a new restaurant
2. Online Banner Advertisements
   - Exploitation: Show the most successful advertisement
   - Exploration: Show a different advertisement
3. Oil Drilling
   - Exploitation: Drill at the best known location
   - Exploration: Drill at a new location
4. Game Playing
   - Exploitation: Play the move you believe is best
   - Exploration: Play an experimental move

It's not just an issue in RL

- It happens in almost all our daily living activities involving decision-making

It's not just an issue in RL

- It happens in almost all our daily living activities involving decision-making
- How can we algorithmically solve this issue in exploration vs exploitation?

It's not just an issue in RL

- It happens in almost all our daily living activities involving decision-making
- How can we algorithmically solve this issue in exploration vs exploitation?
  1. ϵ-greey

It's not just an issue in RL

- It happens in almost all our daily living activities involving decision-making

- How can we algorithmically solve this issue in exploration vs exploitation?
  1. ϵ-greey
  2. Optimism in the face of uncertainty

It's not just an issue in RL

- It happens in almost all our daily living activities involving decision-making
- How can we algorithmically solve this issue in exploration vs exploitation?
  1. ϵ-greey
  2. Optimism in the face of uncertainty
  3. Thompson (posterior) sampling

It's not just an issue in RL

- It happens in almost all our daily living activities involving decision-making
- How can we algorithmically solve this issue in exploration vs exploitation?
  1. ϵ-greey
  2. Optimism in the face of uncertainty
  3. Thompson (posterior) sampling
  4. Noisy networks

It's not just an issue in RL

- It happens in almost all our daily living activities involving decision-making

- How can we algorithmically solve this issue in exploration vs exploitation?
  1. ϵ-greey
  2. Optimism in the face of uncertainty
  3. Thompson (posterior) sampling
  4. Noisy networks
  5. Information theoretic exploration

Method I: ϵ-greedy

Idea:

Method I: ϵ-greedy

Idea:

- Choose a greedy action $(\arg\max Q(s, a))$ with probability $(1 - \epsilon)$
- Choose a random action with probability ϵ

Method I: ϵ-greedy

Idea:

- Choose a greedy action ($\arg\max Q(s, a)$) with probability $(1 - \epsilon)$
- Choose a random action with probability ϵ

Advantages:

1. Very simple
2. Light computation

Method I: ϵ-greedy

Idea:

- Choose a greedy action $(\arg\max Q(s, a))$ with probability $(1 - \epsilon)$
- Choose a random action with probability ϵ

Advantages:

1. Very simple
2. Light computation

Disadvantages:

1. Fine tuning ϵ
2. Not quite systematic: No focus on unexplored regions
3. Inefficient

Method II: Optimism in the face of uncertainty (OFU)

Idea:

Method II: Optimism in the face of uncertainty (OFU)

Idea:

1. Construct a confidence set for MDP parameters using collected samples

Method II: Optimism in the face of uncertainty (OFU)

Idea:

1. Construct a confidence set for MDP parameters using collected samples
2. Choose MDP parameters that give the highest rewards

Method II: Optimism in the face of uncertainty (OFU)

Idea:

1. Construct a confidence set for MDP parameters using collected samples
2. Choose MDP parameters that give the highest rewards
3. Construct an optimal policy of the optimistically chosen MDP

Method II: Optimism in the face of uncertainty (OFU)

Idea:

1. Construct a confidence set for MDP parameters using collected samples
2. Choose MDP parameters that give the highest rewards
3. Construct an optimal policy of the optimistically chosen MDP

Advantages:

1. Regret optimal
2. Systematic: More focus on unexplored regions

Method II: Optimism in the face of uncertainty (OFU)

Idea:

1. Construct a confidence set for MDP parameters using collected samples
2. Choose MDP parameters that give the highest rewards
3. Construct an optimal policy of the optimistically chosen MDP

Advantages:

1. Regret optimal
2. Systematic: More focus on unexplored regions

Disadvantages:

1. Complicated
2. Computation intensive

Method III: Thompson (posterior) sampling

Idea:

Method III: Thompson (posterior) sampling

Idea:

1. Sample MDP parameters from posterior distribution

Method III: Thompson (posterior) sampling

Idea:

1. Sample MDP parameters from posterior distribution
2. Construct an optimal policy of the sampled MDP

Method III: Thompson (posterior) sampling

Idea:

1. Sample MDP parameters from posterior distribution
2. Construct an optimal policy of the sampled MDP
3. Update the posterior distribution

Method III: Thompson (posterior) sampling

Idea:

1. Sample MDP parameters from posterior distribution
2. Construct an optimal policy of the sampled MDP
3. Update the posterior distribution

Advantages:

1. Regret optimal
2. Systematic: More focus on unexplored regions

Method III: Thompson (posterior) sampling

Idea:

1. Sample MDP parameters from posterior distribution
2. Construct an optimal policy of the sampled MDP
3. Update the posterior distribution

Advantages:

1. Regret optimal
2. Systematic: More focus on unexplored regions

Disadvantages:

1. Somewhat complicated

Method IV: Noisy networks

Idea:

- Inject noise to the weights of NN

Method IV: Noisy networks

Idea:

- Inject noise to the weights of NN

Advantages:

1. Simple
2. Good empirical performance

Method IV: Noisy networks

Idea:

- Inject noise to the weights of NN

Advantages:

1. Simple
2. Good empirical performance

Disadvantages:

1. Not systematic: No focus on unexplored regions

Method V: Information theoretic exploration

Idea:

- Use high entropy policy policy to explore more

Method V: Information theoretic exploration

Idea:

- Use high entropy policy policy to explore more

Advantages:

1. Simple
2. Good empirical performance

Method V: Information theoretic exploration

Idea:

- Use high entropy policy policy to explore more

Advantages:

1. Simple
2. Good empirical performance

Disadvantages:

1. More theoretic analyses needed

What is entropy?

- Average rate at which information is produced by a stochastic source of data:

What is entropy?

- Average rate at which information is produced by a stochastic source of data:

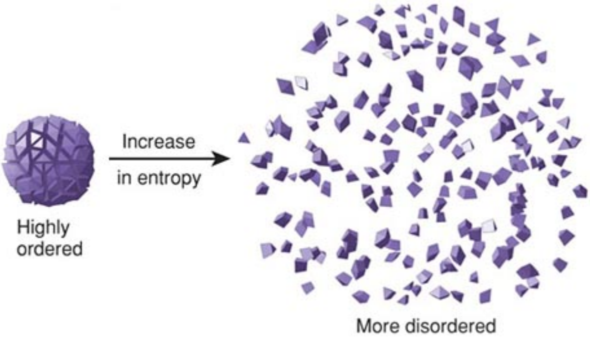
$$H(p) := -\sum_i p_i \log p_i = \mathbb{E}[-\log p_i]$$

What is entropy?

- Average rate at which information is produced by a stochastic source of data:

$$H(p) := -\sum_i p_i \log p_i = \mathbb{E}[-\log p_i]$$

- More generally, entropy refers to *disorder*



What's the benefit of high entropy policy?

Entropy of policy:

$$H(\pi(\cdot|s)) := -\sum_{a} \pi(a|s) \log \pi(a|s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[-\log \pi(a|s)]$$

- Higher disorder in π
- Try new risky behaviors: Potentially explore unexplored regions

Maximum entropy stochastic control

- Standard MDP problem:

$$\max_{\pi} \ \mathbb{E}^{\pi}\left[\sum_t r(s_t, u_t)\right]$$

Maximum entropy stochastic control

- Standard MDP problem:

$$\max_{\pi} \; \mathbb{E}^{\pi}\left[\sum_t r(s_t, u_t)\right]$$

- Maximum entropy MDP problem:

$$\max_{\pi} \; \mathbb{E}^{\pi}\left[\sum_t [r(s_t, u_t) + \alpha H(\pi_t(\cdot|s_t))]\right],$$

where α is called the temperature

Soft value functions

- Soft Q-function of policy π:

$$Q^\pi(s_t, a_t) := r(s_t, a_t) + \mathbb{E}^\pi\left[\sum_{l=1}^{\infty} \gamma^l [r(s_{t+l}, a_{t+l}) + \alpha H(\pi(\cdot|s_{t+l}))]\right]$$

Soft value functions

- Soft Q-function of policy π:

$$Q^\pi(s_t, a_t) := r(s_t, a_t) + \mathbb{E}^\pi\left[\sum_{l=1}^{\infty} \gamma^l [r(s_{t+l}, a_{t+l}) + \alpha H(\pi(\cdot|s_{t+l}))]\right]$$

- Soft value function of policy π:

$$V^\pi(s_t) := \alpha \log \int \exp\left(\frac{1}{\alpha} Q^\pi(s_t, a)\right) da$$

Soft value functions

- Soft Q-function of policy π:

$$Q^\pi(s_t, a_t) := r(s_t, a_t) + \mathbb{E}^\pi\left[\sum_{l=1}^{\infty} \gamma^l [r(s_{t+l}, a_{t+l}) + \alpha H(\pi(\cdot|s_{t+l}))]\right]$$

- Soft value function of policy π:

$$V^\pi(s_t) := \alpha \log \int \exp\left(\frac{1}{\alpha} Q^\pi(s_t, a)\right) da$$

- Optimal value functions:

$$Q^*(s_t, a_t) := \max_\pi Q^\pi(s_t, a_t)$$
$$V^*(s_t) := \alpha \log \int \exp\left(\frac{1}{\alpha} Q^*(s_t, a)\right) da$$

Soft policy improvement theorem

Suppose we are given a policy π_{old}.

Soft policy improvement theorem

Suppose we are given a policy π_{old}.

Q) How can we use the soft value functions to improve the policy?

Soft policy improvement theorem

Suppose we are given a policy π_{old}.

Q) How can we use the soft value functions to improve the policy?

- Define a new policy by

$$\pi_{new}(a|s) := \exp\left(\frac{1}{\alpha}[Q^{\pi_{old}}(s,a) - V^{\pi_{old}}(s)]\right)$$

Soft policy improvement theorem

Suppose we are given a policy π_{old}.

Q) How can we use the soft value functions to improve the policy?

- Define a new policy by

$$\pi_{new}(a|s) := \exp\left(\frac{1}{\alpha}[Q^{\pi_{old}}(s,a) - V^{\pi_{old}}(s)]\right)$$

- Then, we have

$$Q^{\pi_{old}}(s,a) \leq Q^{\pi_{new}}(s,a) \quad \forall (s,a)$$

$\implies$ **Monotonic policy improvement**

What is the optimal policy?

Idea: Keep updating using the soft policy improvement theorem

$$\pi_{old}(a|s) \to \pi_{new}(a|s) := \exp\left(\frac{1}{\alpha}[Q^{\pi_{old}}(s,a) - V^{\pi_{old}}(s)]\right)$$

Result:

- Optimal policy

$$\pi^*(a|s) = \exp\left(\frac{1}{\alpha}[Q^*(s,a) - V^*(s)]\right)$$

Soft Bellman equation

$$
\begin{aligned}
Q^*(s,a) &= r(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')] \\
&= r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^*(s')
\end{aligned}
$$

Q) Looks familiar?

Comparison to standard MDP: Bellman equation

- Standard:

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')]$$
$$V^*(s) = \max_a Q^*(s,a)$$

- Max-Entropy:

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')]$$
$$V^*(s) = \alpha \log \int \exp\left(\frac{1}{\alpha} Q^*(s,a)\right) da$$

Comparison to standard MDP: Bellman equation

- Standard:

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')]$$
$$V^*(s) = \max_a Q^*(s,a)$$

- Max-Entropy:

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')]$$
$$V^*(s) = \alpha \log \int \exp\left(\frac{1}{\alpha} Q^*(s,a)\right) da$$

- Connection:
  Note that as $\alpha \to 0$, $\exp(\frac{1}{\alpha} Q^*(s,a))$ emphasizes $\max_a Q^*(s,a)$:

Comparison to standard MDP: Bellman equation

- Standard:

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')]$$
$$V^*(s) = \max_a Q^*(s,a)$$

- Max-Entropy:

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')]$$
$$V^*(s) = \alpha \log \int \exp\left(\frac{1}{\alpha} Q^*(s,a)\right) da$$

- Connection:
  Note that as $\alpha \to 0$, $\exp(\frac{1}{\alpha} Q^*(s,a))$ emphasizes $\max_a Q^*(s,a)$:

$$\alpha \log \int \exp\left(\frac{1}{\alpha} Q^*(s,a)\right) da \to \max_a Q^*(s,a) \quad \text{as } \alpha \to 0$$

Comparison to standard MDP: Optimal policy

- Standard:

$$\pi^*(s) \in \arg\max_a Q^*(a, s)$$

$\Longrightarrow$ deterministic policy

Comparison to standard MDP: Optimal policy

- Standard:

$$\pi^*(s) \in \arg\max_a Q^*(a, s)$$

$\Longrightarrow$ deterministic policy

- Max-Entropy:

$$\pi^*(a|s) = \exp\left(\frac{1}{\alpha}[Q^*(s, a) - V^*(s)]\right)$$

$\Longrightarrow$ stochastic policy

So, what's the advantage of max-entropy?

- Computation:
  No maximization involved

So, what's the advantage of max-entropy?

- Computation:
  No maximization involved
- Exploration:
  High entropy policy

So, what's the advantage of max-entropy?

- Computation:
  No maximization involved
- Exploration:
  High entropy policy
- Structural similarity:
  Can combine it with many RL methods for standard MDP