Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

Policy learning in 10 minutes w/o reward via human-in-the-loop!

Proxy Value Propagation (PVP)

(1) Human oversees agent’s exploration
(2) Human intervenes and provides demonstration
(3) Update & propagate proxy values

Proxy Value Objective: $J^{PV}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - 1^2 + \max_{a'} Q_d(s',a') \right]$

Temporal Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Total Objective for Q network: $J(\theta) = J^{PV}(\theta) + J^{TD}(\theta)$

Objective for Policy: $J(\phi) = - \mathbb{E}_{s,a} \left[ Q_d(s,a) \right]$

Temporally Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Drop the reward term!

Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

• Reward engineering is hard to encapsulate human intentions.
• Human-in-the-loop methods are promising to achieve alignment.
• To ensure safety, active human involvement enhances training-time safety.

Learning from Active Human Involvement through Proxy Value Propagation
Zhenghao (Mark) Peng1, Wenjie Mo1, Chenda Duan1, Quanyi Li2, Bolei Zhou1
1UCLA, 2University of Edinburgh

Policy learning in 10 minutes w/o reward via human-in-the-loop!

Proxy Value Propagation (PVP)

(1) Human oversees agent’s exploration
(2) Human intervenes and provides demonstration
(3) Update & propagate proxy values

Proxy Value Objective: $J^{PV}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - 1^2 + \max_{a'} Q_d(s',a') \right]$

Temporal Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Total Objective for Q network: $J(\theta) = J^{PV}(\theta) + J^{TD}(\theta)$

Objective for Policy: $J(\phi) = - \mathbb{E}_{s,a} \left[ Q_d(s,a) \right]$

Temporally Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Drop the reward term!

Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

• Reward engineering is hard to encapsulate human intentions.
• Human-in-the-loop methods are promising to achieve alignment.
• To ensure safety, active human involvement enhances training-time safety.

Learning from Active Human Involvement through Proxy Value Propagation
Zhenghao (Mark) Peng1, Wenjie Mo1, Chenda Duan1, Quanyi Li2, Bolei Zhou1
1UCLA, 2University of Edinburgh

Policy learning in 10 minutes w/o reward via human-in-the-loop!

Proxy Value Propagation (PVP)

(1) Human oversees agent’s exploration
(2) Human intervenes and provides demonstration
(3) Update & propagate proxy values

Proxy Value Objective: $J^{PV}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - 1^2 + \max_{a'} Q_d(s',a') \right]$

Temporal Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Total Objective for Q network: $J(\theta) = J^{PV}(\theta) + J^{TD}(\theta)$

Objective for Policy: $J(\phi) = - \mathbb{E}_{s,a} \left[ Q_d(s,a) \right]$

Temporally Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Drop the reward term!

Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

• Reward engineering is hard to encapsulate human intentions.
• Human-in-the-loop methods are promising to achieve alignment.
• To ensure safety, active human involvement enhances training-time safety.

Learning from Active Human Involvement through Proxy Value Propagation
Zhenghao (Mark) Peng1, Wenjie Mo1, Chenda Duan1, Quanyi Li2, Bolei Zhou1
1UCLA, 2University of Edinburgh

Policy learning in 10 minutes w/o reward via human-in-the-loop!

Proxy Value Propagation (PVP)

(1) Human oversees agent’s exploration
(2) Human intervenes and provides demonstration
(3) Update & propagate proxy values

Proxy Value Objective: $J^{PV}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - 1^2 + \max_{a'} Q_d(s',a') \right]$

Temporal Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Total Objective for Q network: $J(\theta) = J^{PV}(\theta) + J^{TD}(\theta)$

Objective for Policy: $J(\phi) = - \mathbb{E}_{s,a} \left[ Q_d(s,a) \right]$

Temporally Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Drop the reward term!

Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

• Reward engineering is hard to encapsulate human intentions.
• Human-in-the-loop methods are promising to achieve alignment.
• To ensure safety, active human involvement enhances training-time safety.

Learning from Active Human Involvement through Proxy Value Propagation
Zhenghao (Mark) Peng1, Wenjie Mo1, Chenda Duan1, Quanyi Li2, Bolei Zhou1
1UCLA, 2University of Edinburgh

Policy learning in 10 minutes w/o reward via human-in-the-loop!

Proxy Value Propagation (PVP)

(1) Human oversees agent’s exploration
(2) Human intervenes and provides demonstration
(3) Update & propagate proxy values

Proxy Value Objective: $J^{PV}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - 1^2 + \max_{a'} Q_d(s',a') \right]$

Temporal Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Total Objective for Q network: $J(\theta) = J^{PV}(\theta) + J^{TD}(\theta)$

Objective for Policy: $J(\phi) = - \mathbb{E}_{s,a} \left[ Q_d(s,a) \right]$

Temporally Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Drop the reward term!

Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

• Reward engineering is hard to encapsulate human intentions.
• Human-in-the-loop methods are promising to achieve alignment.
• To ensure safety, active human involvement enhances training-time safety.

Learning from Active Human Involvement through Proxy Value Propagation
Zhenghao (Mark) Peng1, Wenjie Mo1, Chenda Duan1, Quanyi Li2, Bolei Zhou1
1UCLA, 2University of Edinburgh

Policy learning in 10 minutes w/o reward via human-in-the-loop!

Proxy Value Propagation (PVP)

(1) Human oversees agent’s exploration
(2) Human intervenes and provides demonstration
(3) Update & propagate proxy values

Proxy Value Objective: $J^{PV}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - 1^2 + \max_{a'} Q_d(s',a') \right]$

Temporal Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Total Objective for Q network: $J(\theta) = J^{PV}(\theta) + J^{TD}(\theta)$

Objective for Policy: $J(\phi) = - \mathbb{E}_{s,a} \left[ Q_d(s,a) \right]$

Temporally Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Drop the reward term!

Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

• Reward engineering is hard to encapsulate human intentions.
• Human-in-the-loop methods are promising to achieve alignment.
• To ensure safety, active human involvement enhances training-time safety.

Learning from Active Human Involvement through Proxy Value Propagation
Zhenghao (Mark) Peng1, Wenjie Mo1, Chenda Duan1, Quanyi Li2, Bolei Zhou1
1UCLA, 2University of Edinburgh

Policy learning in 10 minutes w/o reward via human-in-the-loop!

Proxy Value Propagation (PVP)

(1) Human oversees agent’s exploration
(2) Human intervenes and provides demonstration
(3) Update & propagate proxy values

Proxy Value Objective: $J^{PV}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - 1^2 + \max_{a'} Q_d(s',a') \right]$

Temporal Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Total Objective for Q network: $J(\theta) = J^{PV}(\theta) + J^{TD}(\theta)$

Objective for Policy: $J(\phi) = - \mathbb{E}_{s,a} \left[ Q_d(s,a) \right]$

Temporally Difference Objective: $J^{TD}(\theta) = \mathbb{E}_{(s,a)} \left[ Q_d(s,a) - \gamma \max_{a'} Q_d(s',a') \right]^2$

Drop the reward term!

Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

• Reward engineering is hard to encapsulate human intentions.
• Human-in-the-loop methods are promising to achieve alignment.
• To ensure safety, active human involvement enhances training-time safety.