Challenges for Using Impact Regularizers to Avoid Negative Side Effects

David Lindner, Kyle Matoba, Alexander Meulemans
Agents and side effects

- Agents fulfill tasks by maximizing cumulative reward
Agents and side effects

- Agents fulfill tasks by maximizing cumulative reward
- Reward specification problem:
  - What should the agent do?
  - What should the agent not do?
- Side effects
Agents and side effects

- Agents fulfill tasks by maximizing cumulative reward
- Reward specification problem:
  - What should the agent do?
  - What should the agent not do?
- Side effects
- Impact regularizers

\[ R(s) = R_{\text{spec}}(s) + R_{\text{IR}}(s) \]
Impact Regularizers

\[ R(s_t) = R_{\text{spec}}(s_t) - \lambda \cdot d(s_t, b(s_0, s_{t-1}, t)) \]

- magnitude
- baseline
- deviation
Impact Regularizers

\[ R(s_t) = R_{\text{spec}}(s_t) - \lambda \cdot d(s_t, b(s_0, s_{t-1}, t)) \]

- \( R(s_t) \): Regularizer value at time step \( t \)
- \( R_{\text{spec}}(s_t) \): Specified regularizer value at time step \( t \)
- \( \lambda \): Magnitude parameter
- \( d \): Deviation function
- \( b(s_0, s_{t-1}, t) \): Baseline value

Actual trajectory

\( S_0 \) or \( S_{t-1} \)

Compare using \( d \)

Baseline policy

\( S_t \)

Baseline
Baseline
Baseline

Problems with inaction baseline

- Unsafe inaction baseline
- Chaotic environment dynamics
- Offsetting
Offsetting

Undesirable offsetting

Source: Designing agent incentives to avoid side effects
Offsetting

Undesirable offsetting

Desirable offsetting

Source: Designing agent incentives to avoid side effects
Offsetting

Undesirable offsetting

Desirable offsetting

Consequence of completing the task

Instrumental towards achieving the task

Source: Designing agent incentives to avoid side effects
Deviation measure

How much should a deviation from the baseline be penalized?
Deviation measure

How much should a deviation from the baseline be penalized?

Problems with current deviation measures
▷ Positive, neutral and negative side effects
▷ Rollout policy
Positive, neutral and negative side effects

Not all impact is equally negative!

▷ Suboptimal solutions if notion of ‘value’ is omitted

Optimize reaction path 

Waste products 

Production rate
Tuning regularization magnitude

\[ R(s_t) = R_{\text{spec}}(s_t) - \lambda \cdot d(s_t, b(s_0, s_{t-1}, t)) \]

Dangerous

Safe action

Does nothing

Source: Armstrong & Levenstein 2017
Tuning regularization magnitude

\[ R(s_t) = R_{\text{spec}}(s_t) - \lambda \cdot d(s_t, b(s_0, s_{t-1}, t)) \]

Source: Armstrong & Levenstein 2017
Ways forward

▷ Causal framing of offsetting
▷ Probabilities instead of counterfactuals
▷ Improved Human-Computer interaction