Policy Shaping: Integrating Human Feedback with RL

Kushagra Chandak

Paper by: Griffith et al, NIPS'13

22nd Feb, 2021

Introduction: Policy Shaping

- Integrate human feedback with interactive RL.

Introduction: Policy Shaping

- Integrate human feedback with interactive RL.
- Previously: Treat human feedback as a shaping reward.

Introduction: Policy Shaping

- Integrate human feedback with interactive RL.
- Previously: Treat human feedback as a shaping reward.
- This paper: Human feedback as policy advice.

Introduction: Policy Shaping

- Integrate human feedback with interactive RL.
- Previously: Treat human feedback as a shaping reward.
- This paper: Human feedback as policy advice.
- **Advise**: An algorithm for

Introduction: Policy Shaping

- Integrate human feedback with interactive RL.
- Previously: Treat human feedback as a shaping reward.
- This paper: Human feedback as policy advice.
- **Advise**: An algorithm for
  - estimating human's Bayes optimal feedback policy;

Introduction: Policy Shaping

- Integrate human feedback with interactive RL.
- Previously: Treat human feedback as a shaping reward.
- This paper: Human feedback as policy advice.
- **Advise**: An algorithm for
  - estimating human's Bayes optimal feedback policy;
  - combining human policy with RL policy (agent's direct experience with the environment)

Bayesian Q-Learning

- Q-Learning: $Q(s,a)$ represent point estimate of the long term expected discounted reward for taking action a in state s
- BQL maintains parameters that specify a normal distribution with unknown mean and precision for each Q-value
- Mean and the precision are estimated using a NG distribution with hyperparameters $< \mu_0^{s,a}, \lambda^{s,a}, \alpha^{s,a}, \beta^{s,a} >$
- Parameters updated at each time step of RL
- RL policy π_R:
  - Optimal action estimate: $\arg\max_a \hat{Q}(s,a)$.
  - Estimate of probability that an action is optimal: Sample a large number of times and count the number of times an action has the highest Q-value.

Policy Shaping

- Agent has access to communication from a human during its learning process.

Policy Shaping

- Agent has access to communication from a human during its learning process.
- Receive "right"/"wrong" label after performing *a*: comment on optimality of the action just performed. (in addition to environment reward).

Policy Shaping

- Agent has access to communication from a human during its learning process.
- Receive "right"/"wrong" label after performing *a*: comment on optimality of the action just performed. (in addition to environment reward).
- Reward shaping: "right" becomes +1 and "wrong" becomes -1 which is used to modify Q-values or to bias action selection.

Policy Shaping

- Agent has access to communication from a human during its learning process.
- Receive "right"/"wrong" label after performing *a*: comment on optimality of the action just performed. (in addition to environment reward).
- Reward shaping: "right" becomes +1 and "wrong" becomes -1 which is used to modify Q-values or to bias action selection.
- Policy shaping: Use the label directly to infer human optimal policy.

Policy Shaping

- Agent has access to communication from a human during its learning process.
- Receive "right"/"wrong" label after performing *a*: comment on optimality of the action just performed. (in addition to environment reward).
- Reward shaping: "right" becomes $+1$ and "wrong" becomes -1 which is used to modify Q-values or to bias action selection.
- Policy shaping: Use the label directly to infer human optimal policy.
- Noise in the feedback channel introduces inconsistencies between what the human intends to communicate and what the agent observes. Feedback consistent with the optimal policy with probability $\mathcal{C}$.

Policy Shaping

- Agent has access to communication from a human during its learning process.
- Receive "right"/"wrong" label after performing *a*: comment on optimality of the action just performed. (in addition to environment reward).
- Reward shaping: "right" becomes +1 and "wrong" becomes -1 which is used to modify Q-values or to bias action selection.
- Policy shaping: Use the label directly to infer human optimal policy.
- Noise in the feedback channel introduces inconsistencies between what the human intends to communicate and what the agent observes. Feedback consistent with the optimal policy with probability $\mathcal{C}$.
- Likelihood of receiving feedback has probability $\mathcal{L}$.

Policy Shaping

- How to estimate human feedback policy?

Policy Shaping

- How to estimate human feedback policy?
- $\Delta_{s,a}$ = No. of "right" and "wrong" labels associated with it.

Policy Shaping

- How to estimate human feedback policy?
- $\Delta_{s,a}$ = No. of "right" and "wrong" labels associated with it.
- Probability s, a is optimal: $Pr(\pi | d_{s,a}) = \frac{\mathcal{C}^{\Delta_{s,a}}}{\mathcal{C}^{\Delta_{s,a}} + (1-\mathcal{C})^{\Delta_{s,a}}}$

Policy Shaping

- How to estimate human feedback policy?
- $\Delta_{s,a}$ = No. of "right" and "wrong" labels associated with it.
- Probability s, a is optimal: $Pr(\pi | d_{s,a}) = \frac{\mathcal{C}^{\Delta_{s,a}}}{\mathcal{C}^{\Delta_{s,a}} + (1-\mathcal{C})^{\Delta_{s,a}}}$
- If only one optimal action in each state:
  $Pr(\pi^k | D_s) = \mathcal{C}^{\Delta_{s,a}} (1 - \mathcal{C})^{\sum_{j \neq a}^{n} \Delta_{s,j}} = \pi_F(s, a)$

Policy Shaping

- How to estimate human feedback policy?
- $\Delta_{s,a}$ = No. of "right" and "wrong" labels associated with it.
- Probability s, a is optimal: $Pr(\pi | d_{s,a}) = \frac{\mathcal{C}^{\Delta_{s,a}}}{\mathcal{C}^{\Delta_{s,a}} + (1-\mathcal{C})^{\Delta_{s,a}}}$
- If only one optimal action in each state:
  $Pr(\pi^k | D_s) = \mathcal{C}^{\Delta_{s,a}}(1 - \mathcal{C})^{\sum_{j \neq a}^{n} \Delta_{s,j}} = \pi_F(s, a)$
- It is the Bayes optimal feedback policy given "right" and "wrong" labels seen, value of $\mathcal{C}$, and only one optimal action per state.

Policy Shaping

- How to reconcile policy information from multiple sources? (human and environment)

Policy Shaping

- How to reconcile policy information from multiple sources? (human and environment)
- Although, the human can make a mistake ($\mathcal{C}$), given sufficient time, $\mathcal{L} > 0$, and $\mathcal{C} \neq 0.5$, the total amt of information from the human should be enough for the agent to choose the optimal policy with probability 1.

Policy Shaping

- How to reconcile policy information from multiple sources? (human and environment)
- Although, the human can make a mistake ($\mathcal{C}$), given sufficient time, $\mathcal{L} > 0$, and $\mathcal{C} \neq 0.5$, the total amt of information from the human should be enough for the agent to choose the optimal policy with probability 1.
- What action to perform using the policy information each source provide?

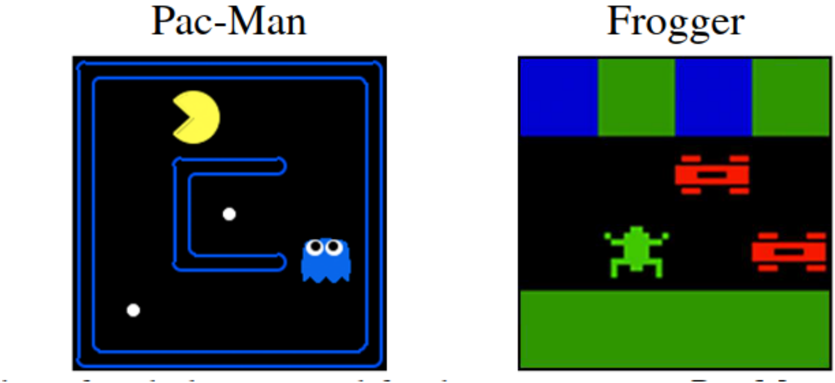
Policy Shaping

- How to reconcile policy information from multiple sources? (human and environment)
- Although, the human can make a mistake ($\mathcal{C}$), given sufficient time, $\mathcal{L} > 0$, and $\mathcal{C} \neq 0.5$, the total amt of information from the human should be enough for the agent to choose the optimal policy with probability 1.
- What action to perform using the policy information each source provide?
- $\pi \propto \pi_R \times \pi_F$.

Policy Shaping

- How to reconcile policy information from multiple sources? (human and environment)
- Although, the human can make a mistake ($\mathcal{C}$), given sufficient time, $\mathcal{L} > 0$, and $\mathcal{C} \neq 0.5$, the total amt of information from the human should be enough for the agent to choose the optimal policy with probability 1.
- What action to perform using the policy information each source provide?
- $\pi \propto \pi_R \times \pi_F$.
- Bayes optimal method for combining probabilities from (conditionally) independent sources.

Experiments



Experiments

- BQL with Advise compared against
  - BQL + Action Biasing
  - BQL + Control Sharing
  - BQL + Reward Shaping
- In all the algorithms, positive feedback gets a reward $+r_h$ and negative feedback gets a reward of $-r_h$.
- Parameter $B[s, a]$ controls the influence of feedback on learning.
- $B[s, a]$ is incremented by b when feedback is received for s, a and decayed by d at all other time steps.

Other algorithms for integrating feedback (baselines)

- Action Biasing.

Other algorithms for integrating feedback (baselines)

- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.

Other algorithms for integrating feedback (baselines)

- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.
  - Action selection mechanism: $\arg\max_a \hat{Q}(s,a) + B[s,a] * H[s,a]$.

Other algorithms for integrating feedback (baselines)

- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.
  - Action selection mechanism: $\arg\max_a \hat{Q}(s, a) + B[s, a] * H[s, a]$.
  - Guiding exploration toward human favored state-action pairs.

Other algorithms for integrating feedback (baselines)

- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.
  - Action selection mechanism: $\arg\max_a \hat{Q}(s,a) + B[s,a] * H[s,a]$.
  - Guiding exploration toward human favored state-action pairs.
- Control Sharing.

Other algorithms for integrating feedback (baselines)

- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.
  - Action selection mechanism: $\arg\max_a \hat{Q}(s,a) + B[s,a] * H[s,a]$.
  - Guiding exploration toward human favored state-action pairs.
- Control Sharing.
  - $P(a = \arg\max_a H[s,a]) = \min(B[s,a], 1)$

Other algorithms for integrating feedback (baselines)

- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.
  - Action selection mechanism: $\arg\max_a \hat{Q}(s, a) + B[s, a] * H[s, a]$.
  - Guiding exploration toward human favored state-action pairs.
- Control Sharing.
  - $P(a = \arg\max_a H[s, a]) = \min(B[s, a], 1)$
  - An agent transfers control to a feedback policy as feedback is received, and begins to switch control to the underlying RL algorithm as $B[s, a]$ decays.

Other algorithms for integrating feedback (baselines)

- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.
  - Action selection mechanism: $\arg\max_a \hat{Q}(s,a) + B[s,a] * H[s,a]$.
  - Guiding exploration toward human favored state-action pairs.
- Control Sharing.
  - $P(a = \arg\max_a H[s,a]) = \min(B[s,a], 1)$
  - An agent transfers control to a feedback policy as feedback is received, and begins to switch control to the underlying RL algorithm as $B[s,a]$ decays.
- Reward shaping.

Other algorithms for integrating feedback (baselines)

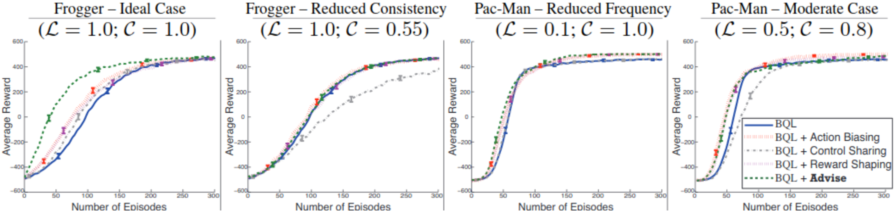
- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.
  - Action selection mechanism: $\arg\max_a \hat{Q}(s,a) + B[s,a] * H[s,a]$.
  - Guiding exploration toward human favored state-action pairs.
- Control Sharing.
  - $P(a = \arg\max_a H[s,a]) = \min(B[s,a], 1)$
  - An agent transfers control to a feedback policy as feedback is received, and begins to switch control to the underlying RL algorithm as $B[s,a]$ decays.
- Reward shaping.
  - Feedback first converted to a reward $+r_h$ or $-r_h$.

Other algorithms for integrating feedback (baselines)

- Action Biasing.
  - Human feedback to bias the action selection mechanism of the underlying RL algo.
  - Action selection mechanism: $\arg\max_a \hat{Q}(s,a) + B[s,a] * H[s,a]$.
  - Guiding exploration toward human favored state-action pairs.
- Control Sharing.
  - $P(a = \arg\max_a H[s,a]) = \min(B[s,a], 1)$
  - An agent transfers control to a feedback policy as feedback is received, and begins to switch control to the underlying RL algorithm as $B[s,a]$ decays.
- Reward shaping.
  - Feedback first converted to a reward $+r_h$ or $-r_h$.
  - $R'(s,a) = R(s,a) + B[s,a] * H[s,a]$

Results

	Ideal Case ($\mathcal{L}=1.0, \mathcal{C}=1.0$)		Reduced Consistency ($\mathcal{L}=0.1, \mathcal{C}=1.0$)		Reduced Frequency ($\mathcal{L}=1.0, \mathcal{C}=0.55$)		Moderate Case ($\mathcal{L}=0.5, \mathcal{C}=0.8$)	
	Pac-Man	Frogger	Pac-Man	Frogger	Pac-Man	Frogger	Pac-Man	Frogger
BQL + Action Biasing	0.58 ± 0.02	0.16 ± 0.05	-0.33 ± 0.17	0.05 ± 0.06	0.16 ± 0.04	0.04 ± 0.06	$\mathbf{0.25 \pm 0.04}$	0.09 ± 0.06
BQL + Control Sharing	0.34 ± 0.03	0.07 ± 0.06	-2.87 ± 0.12	-0.32 ± 0.13	0.01 ± 0.12	0.02 ± 0.07	-0.18 ± 0.19	0.01 ± 0.07
BQL + Reward Shaping	0.54 ± 0.02	0.11 ± 0.07	-0.47 ± 0.30	0 ± 0.08	0.14 ± 0.04	0.03 ± 0.07	0.17 ± 0.12	0.05 ± 0.07
BQL + **Advise**	$\mathbf{0.77 \pm 0.02}$	$\mathbf{0.45 \pm 0.04}$	$\mathbf{-0.01 \pm 0.11}$	0.02 ± 0.07	$\mathbf{0.21 \pm 0.05}$	$\mathbf{0.16 \pm 0.06}$	0.13 ± 0.08	$\mathbf{0.22 \pm 0.06}$



Other points

- Reward parameter affects action biasing: Large r_h is appropriate for more consistent feedback; smaller r_h for reduced consistency.
- Domain size affects learning: $B[s, a]$ function of domain size. Advise algorithm performs better.
- Inaccurate estimation of $\mathcal{C}$: Desirable to use $\hat{\mathcal{C}}$ as the closest overestimate to its true value.

Conclusion

- Advise performed on par or better.
- Robust to infrequent and inconsistent feedback.
- Future work:
  - Estimate $\hat{\mathcal{C}}$ during learning.
  - Errors in credit assignment by humans.
  - Knowledge transfer.