CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Cooperative IRL

Motivation: values

- There's a difference between having a robot optimize for the human's reward function from it's own point of view (imitation) vs. optimize the reward that the human receives (assisting)
- Or a different framing: taking on the human's values itself vs. understanding the human's values to enable cooperation with them
    - We don't want the robot to make itself a cup of coffee!

Hadfield-Menell D, Russell SJ, Abbeel P, Dragan A. Cooperative inverse reinforcement learning. Advances in neural information processing systems. 2016.

Motivation: teaching behavior

- Humans aren’t optimal — and often purposely so!
    - Teaching using suboptimal trajectories
    - Gesturing, narrating, explaining branching logic of contingencies
- Teaching is often interactive and iterative
    - Learner might ask questions, try and make mistakes, etc.

Cooperative IRL: definition

- A two-player partial information game, in which the human (H) knows the reward function, while the robot (R) does not
- The robot's payoff is the human's reward, thus optimal solutions to this game maximize human reward
- Incentivizes active instructive behavior by the human and active learning by the robot, without directly encoding that objective

Hadfield-Menell D, Russell SJ, Abbeel P, Dragan A. Cooperative inverse reinforcement learning. Advances in neural information processing systems. 2016.

Formulation

Definition 1. *A* cooperative inverse reinforcement learning *(CIRL) game M is a two-player Markov game with identical payoffs between a human or principal, $\mathbf{H}$, and a robot or agent, $\mathbf{R}$. The game is described by a tuple, $M = \langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}, T(\cdot|\cdot,\cdot,\cdot), \{\Theta, R(\cdot,\cdot,\cdot;\cdot)\}, P_0(\cdot,\cdot), \gamma\rangle$, with the following definitions:*

$\mathcal{S}$ *a set of world states: $s \in \mathcal{S}$.*
$\mathcal{A}^{\mathbf{H}}$ *a set of actions for $\mathbf{H}$: $a^{\mathbf{H}} \in \mathcal{A}^{\mathbf{H}}$.*
$\mathcal{A}^{\mathbf{R}}$ *a set of actions for $\mathbf{R}$: $a^{\mathbf{R}} \in \mathcal{A}^{\mathbf{R}}$.*
$T(\cdot|\cdot,\cdot,\cdot)$ *a conditional distribution on the next world state, given previous state and action for both agents: $T(s'|s, a^{\mathbf{H}}, a^{\mathbf{R}})$.*
Θ *a set of possible static reward parameters, only observed by $\mathbf{H}$: $\theta \in \Theta$.*
$R(\cdot,\cdot,\cdot;\cdot)$ *a parameterized reward function that maps world states, joint actions, and reward parameters to real numbers. $R : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \Theta \to \mathbb{R}$.*
$P_0(\cdot,\cdot)$ *a distribution over the initial state, represented as tuples: $P_0(s_0, \theta)$*
γ *a discount factor: $\gamma \in [0,1]$.*

Complexity

- Naively as hard as a **Dec-POMDP** to solve for an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{R}})$, if posed as a general cooperative game.
    - NEXP-complete —> Doubly exponential in worst case!
- Instead, if both policies are generated by a *centralized* coordinator that observes all common observations, then the problem can be reduced to a **single-agent POMDP**.
- POMDPs are still very hard! PSPACE-complete: exponential time worst case.

Apprenticeship as a special case of CIRL

- Fixed H that gives only expert demonstrations
- R gives single-round best response
- In general, not an optimal joint policy!
- Example: manufacturing paperclips and staples

$$\mathcal{A}^{\mathbf{H}} = \{(0,2),(1,1),(2,0)\} \qquad \mathcal{A}^{\mathbf{R}} = \{(0,90),(50,50),(90,0)\}$$

Theorem 3. *There exist ACIRL games where the best-response for* $\mathbf{H}$ *to* $\pi^{\mathbf{R}}$ *violates the expert demonstrator assumption. In other words, if* $\mathbf{br}(\pi)$ *is the best response to* π*, then* $\mathbf{br}(\mathbf{br}(\pi^{\mathbf{E}})) \neq \pi^{\mathbf{E}}$.

The supplementary material proves this theorem by computing the optimal equilibrium for our example. In that equilibrium, $\mathbf{H}$ selects $(1,1)$ if $\theta \in [\frac{41}{92}, \frac{51}{92}]$. In contrast, $\pi^{\mathbf{E}}$ only chooses $(1,1)$ if $\theta = 0.5$. The change arises because there are situations (e.g., $\theta = 0.49$) where the immediate loss of reward to $\mathbf{H}$ is worth the improvement in $\mathbf{R}$'s estimate of θ.

Generating instructive demonstrations

- How to compute H's best response if R uses IRL as an estimator of theta?
- Can be reduced to a POMDP where the state is a tuple of the world state, reward parameters (since H knows them), and R's belief about theta
- With linear reward features, H tries to give demo such that if R matches features as closely as possible under its action space, true reward will be maximized:

$$\tau^{\mathbf{H}} \leftarrow \operatorname{argmax} \phi(\tau)^\top \theta - \eta||\phi_\theta - \phi(\tau)||^2.$$

Experiments

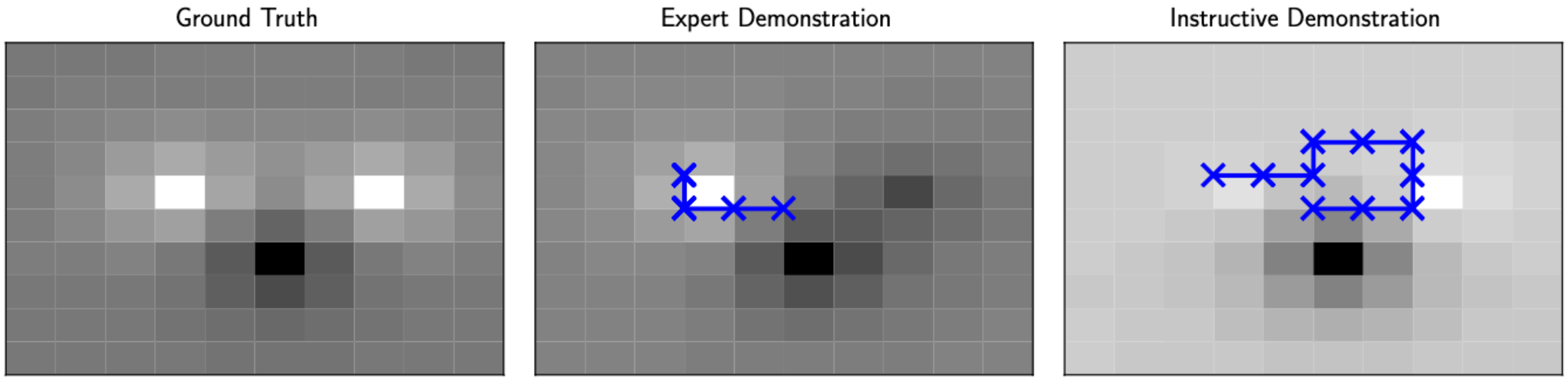
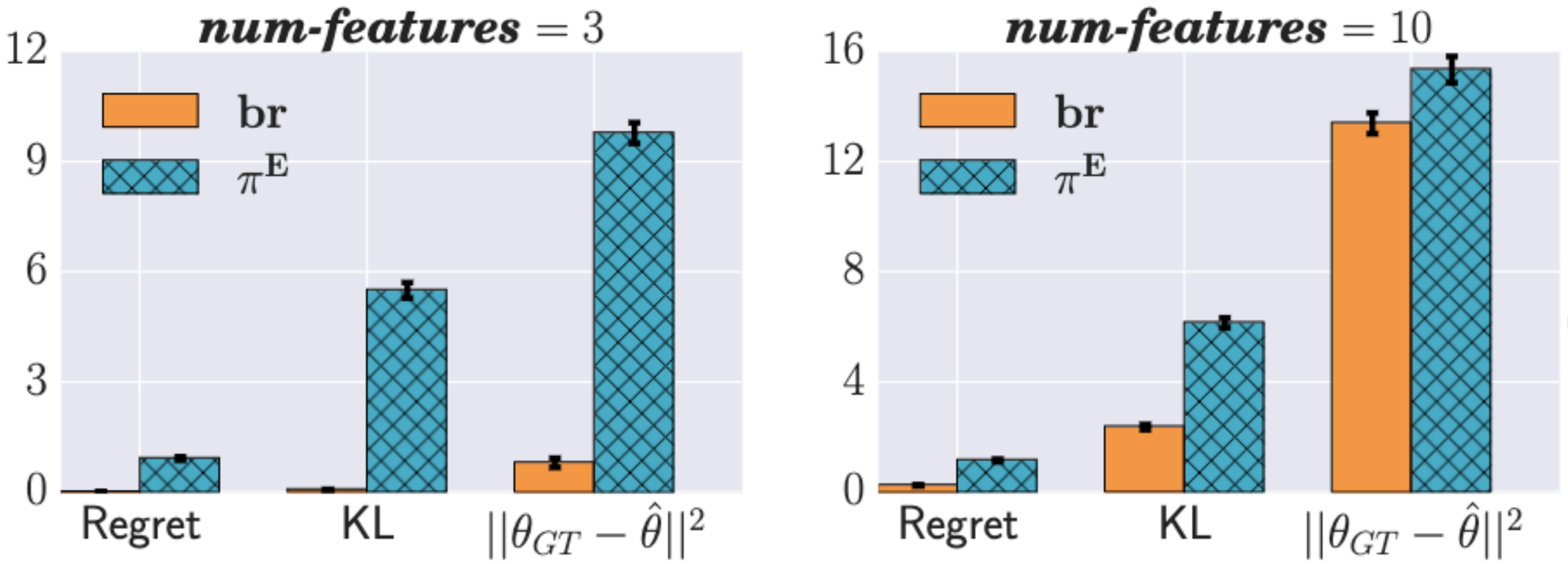


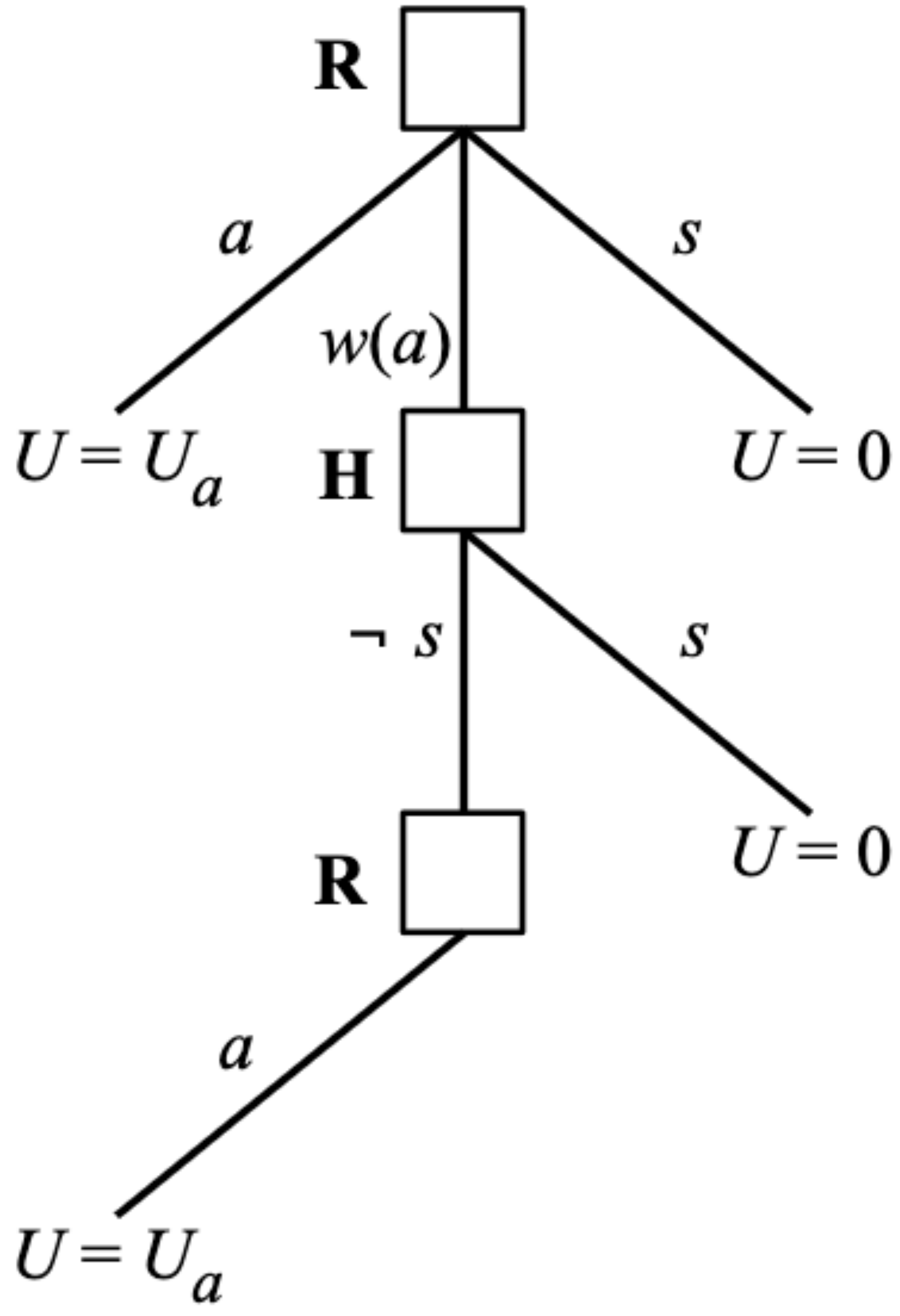
Figure 1: The difference between demonstration-by-expert and instructive demonstration in the mobile robot navigation problem from Section 4. Left: The ground truth reward function. Lighter grid cells indicates areas of higher reward. Middle: The demonstration trajectory generated by the expert policy, superimposed on the maximum a-posteriori reward function the robot infers. The robot successfully learns where the maximum reward is, but little else. Right: An instructive demonstration generated by the algorithm in Section 3.4 superimposed on the maximum a-posteriori reward function that the robot infers. This demonstration highlights both points of high reward and so the robot learns a better estimate of the reward.

Experiments



Pros/Cons of CIRL?

The off-switch game



R	H	
	s	$\neg s$
$w(a)$	0	U_a
a	U_a	U_a
s	0	0

In general, **R**'s actions will fall into one of three categories: some prevent **H** from switching **R** off, by whatever means; some allow **H** to switch **R** off; and, for completeness, some lead to **R** switching *itself* off. In the off-switch game, **R** moves first and has three choices:

1. action a simply bypasses human oversight (disabling the off switch is one way to do this) and acts directly on the world, achieving utility $U = U_a$ for **H**.
2. action $w(a)$ informs **H** that **R** would like to do a, and waits for **H**'s response.
3. action s switches **R** off; without loss of generality, we assign this outcome $U = 0$.

If **R** chooses $w(a)$, then **H** can choose action s to switch **R** off, or $\neg s$ to allow **R** to go ahead (in which case **R** does a as promised). Figure 1 shows the basic structure of the game.

Hadfield-Menell D, Dragan A, Abbeel P, Russell S. The off-switch game. In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence. 2017.

The off-switch game

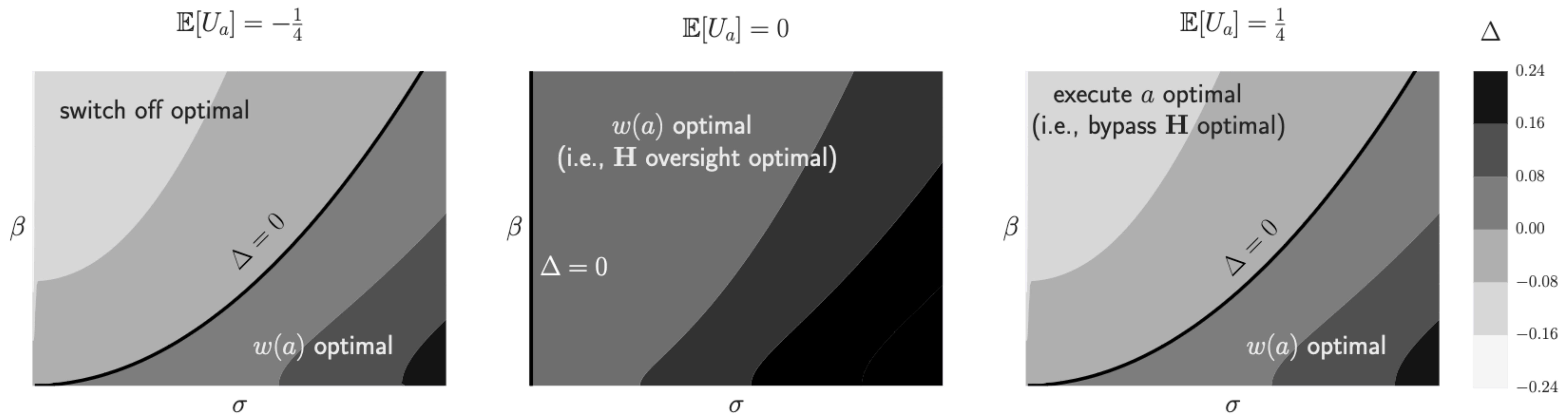
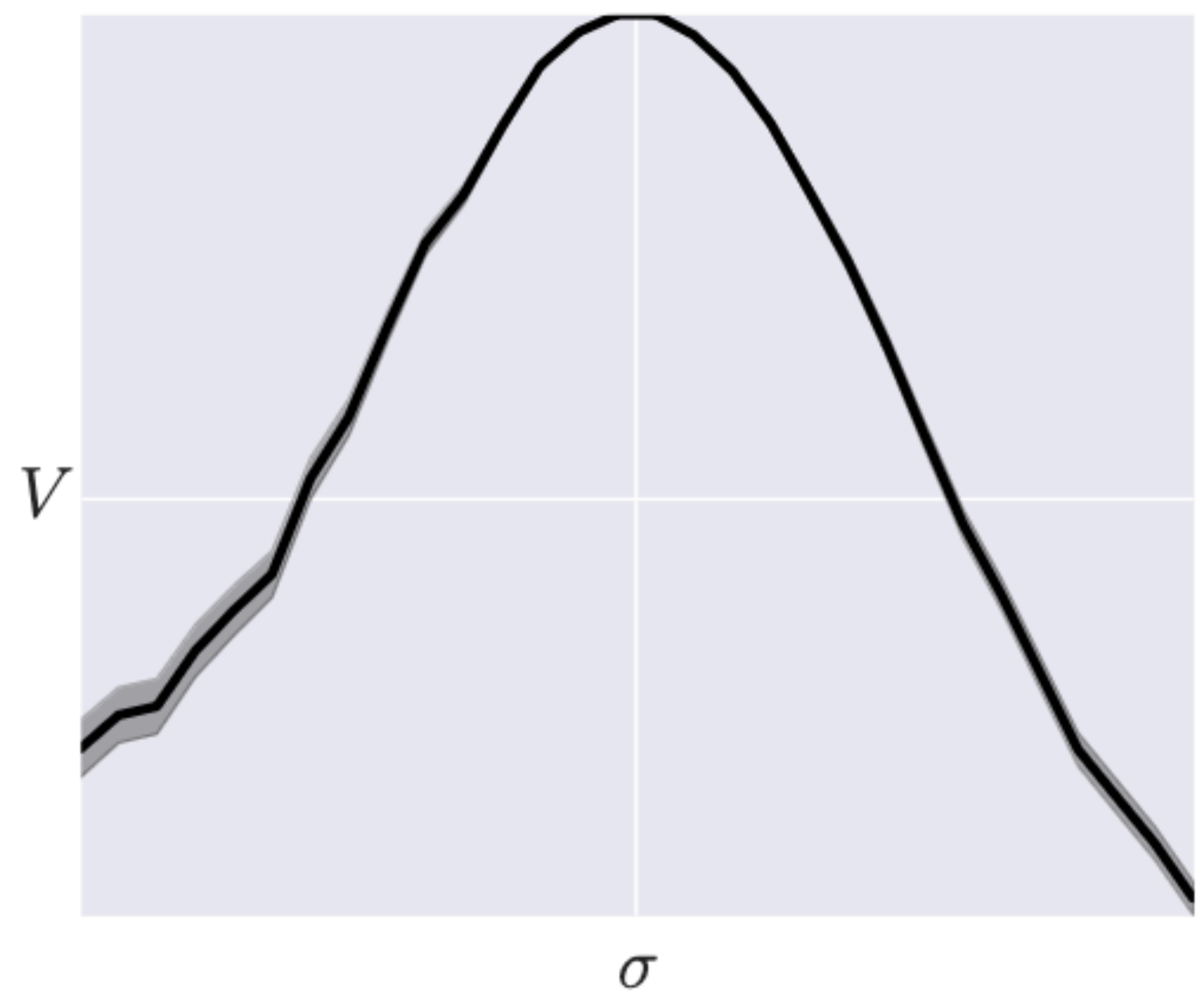


Figure 3: If $\mathbf{H}$ is an irrational actor, then $\mathbf{R}$ may prefer switching itself off or executing a immediately rather than handing over the choice to $\mathbf{H}$. $\mathbf{R}$'s belief $B^{\mathbf{R}}$ is a Gaussian with standard deviation σ and $\mathbf{H}$'s policy is a Boltzmann distribution (Equation 5). β measures $\mathbf{H}$'s suboptimality: $\beta = 0$ corresponds to a rational $\mathbf{H}$ and $\beta = \infty$ corresponds to a $\mathbf{H}$ that randomly switches $\mathbf{R}$ off (i.e., switching $\mathbf{R}$ off is independent of U_a). In all three plots Δ is lower in the top left, where $\mathbf{R}$ is certain (σ low) and $\mathbf{H}$ is very suboptimal (β high), and higher in the bottom right, where $\mathbf{R}$ is uncertain (σ high) and $\mathbf{H}$ is near-optimal (β low). The sign of $\mathbb{E}[U_a]$ controls $\mathbf{R}$'s behavior if $\Delta \leq 0$. **Left:** If it is negative, then $\mathbf{R}$ switches itself off. **Right:** If it is positive, $\mathbf{R}$ executes action a directly. **Middle:** If it is 0 , $\mathbf{R}$ is indifferent between $w(a)$, a, and s.

The off-switch game



The off-switch game