CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Adversarial imitation learning

Final project proposals

- Can work with up to 2 other people
- Example projects could include: extending an algorithm in a novel way; comparing several algorithms on an interesting problem; designing a new approach to attack a problem relevant to the class.
- In all cases, there should be a novel intellectual contribution, as well as empirical results on a problem of interest.
- Writeup should include:
  - A clear description of the problem you are investigating, both abstractly and in context of a particular experimental domain
  - References to a few papers that are relevant to the subject of interest
  - A proposed plan to address your problem, which should outline what method(s) you plan to develop, implement, compare, or extend (and how)
  - A testable hypothesis
  - An experiment to test your hypothesis and a clear evaluation criteria to determine the outcome of your experiment / hypothesis

IRL problems so far

- RL in the inner loop
- Overfitting to noisy estimates of expert feature counts or (s,a) occupancies
- Doesn't scale to large problems: restricted to linear rewards with carefully designed features
- Indirect - if we want to match expert, why can't we just learn a policy directly?

A generalized view of IRL

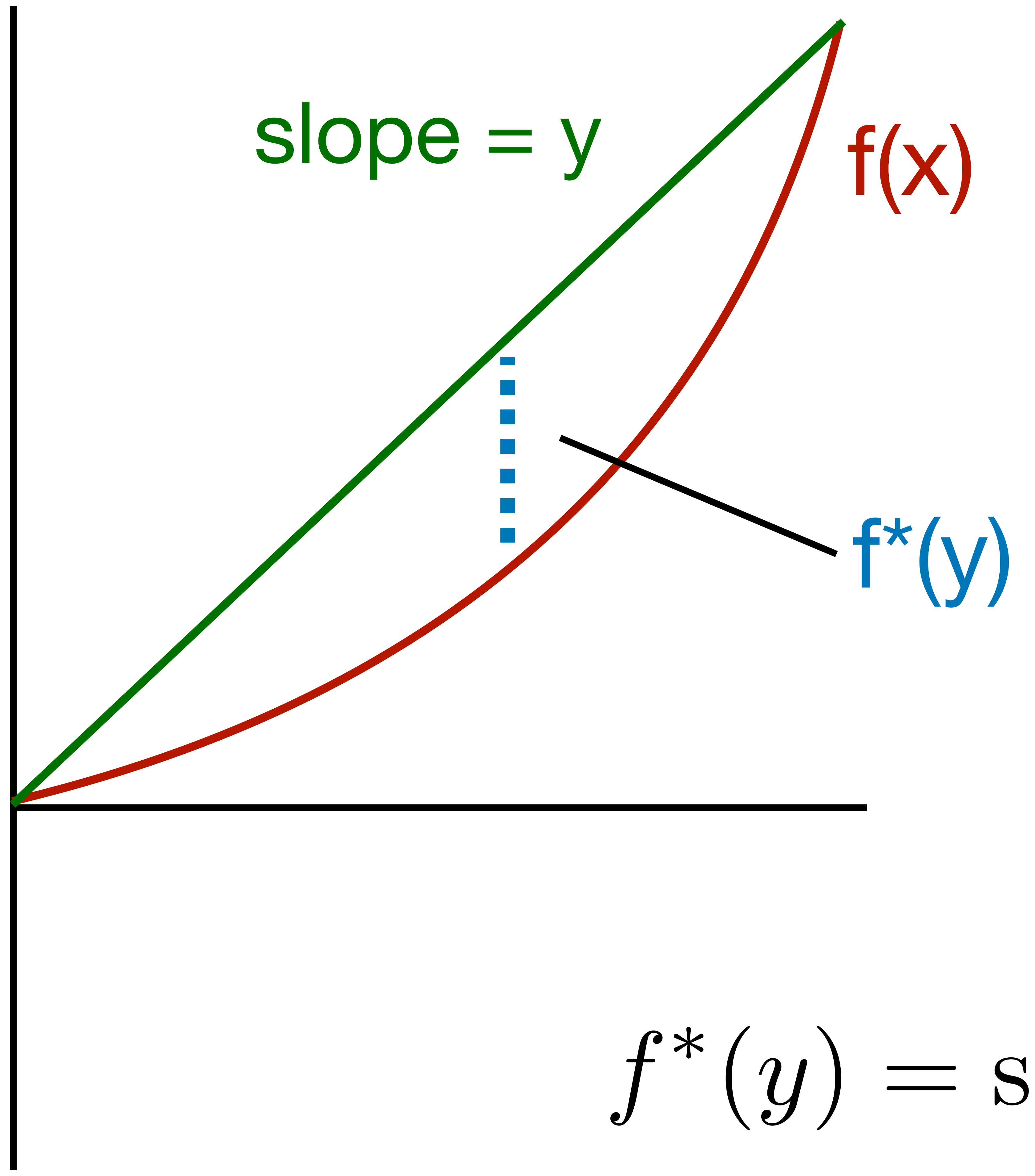
$$\mathrm{IRL}_\psi(\pi_E) = \underset{c \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max} \underbrace{-\psi(c)}_{\text{Regularizer}} + \underbrace{\left(\min_{\pi\in\Pi} -H(\pi) + \mathbb{E}_\pi[c(s,a)]\right)}_{\text{Entropy-regularized RL}} - \underbrace{\mathbb{E}_{\pi_E}[c(s,a)]}_{\text{Expert performance}}$$

Proposition 3.2. $\mathrm{RL} \circ \mathrm{IRL}_\psi(\pi_E) = \arg\min_{\pi\in\Pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E})$

$$\psi^*(y) = \sup_{x\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} x^\top y - \psi(x)$$

Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." *Advances in neural information processing systems* 29 (2016).

Detour: convex conjugates

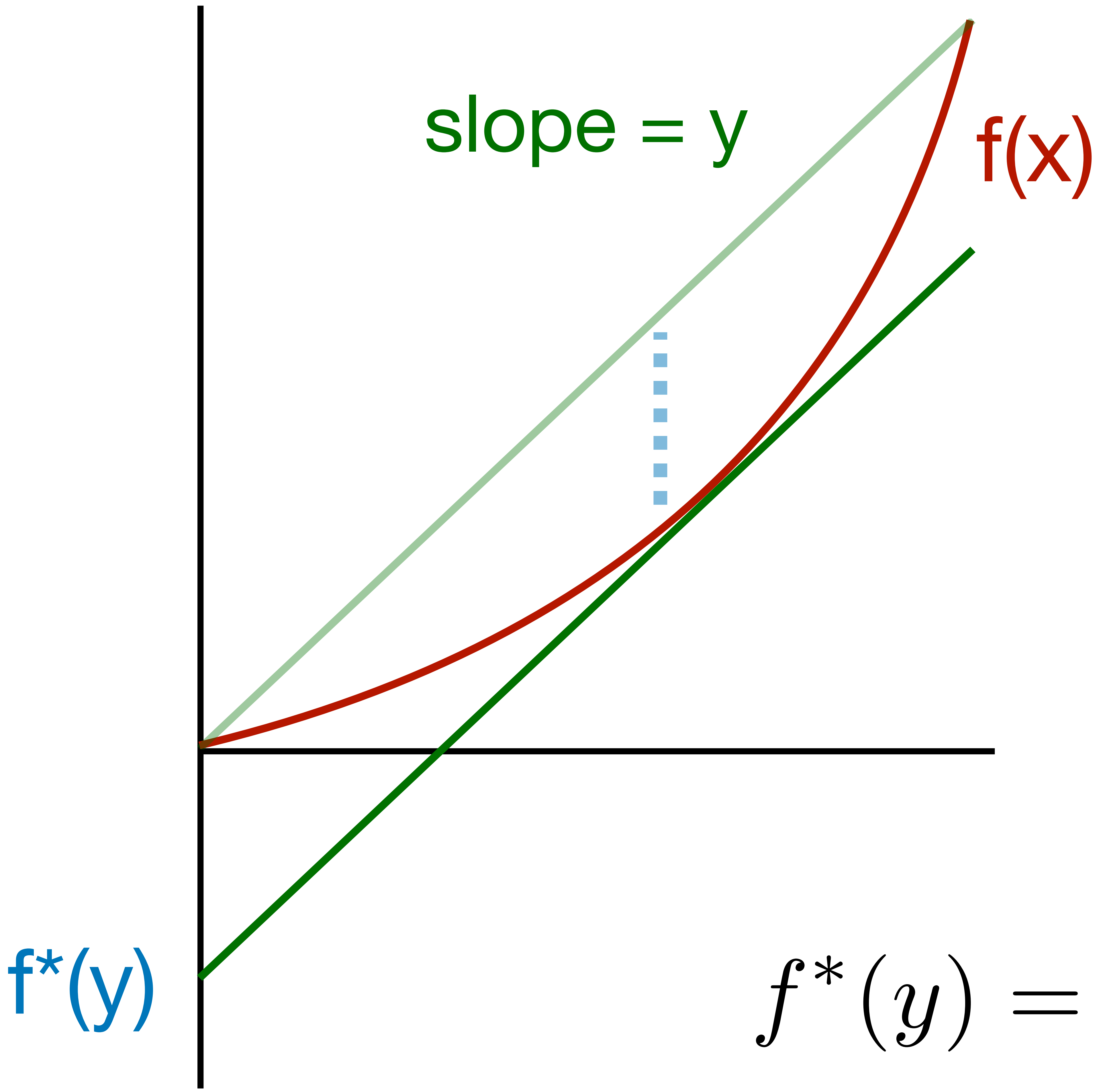


Economic view:
If it costs f(x) to make x widgets and I can sell them for y each, then f*(y) is the max profit I can make

ML view:
If I have a regularizer f(x) and utility function $u_y(x) = xy$, then f*(y) tells me the value of the best regularized utility I can achieve by choosing the best “weight” x for a fixed y: $f^*(y) = \sup_x u_y(x) - f(x)$

$$f^*(y) = \sup_x xy - f(x)$$

Detour: convex conjugates

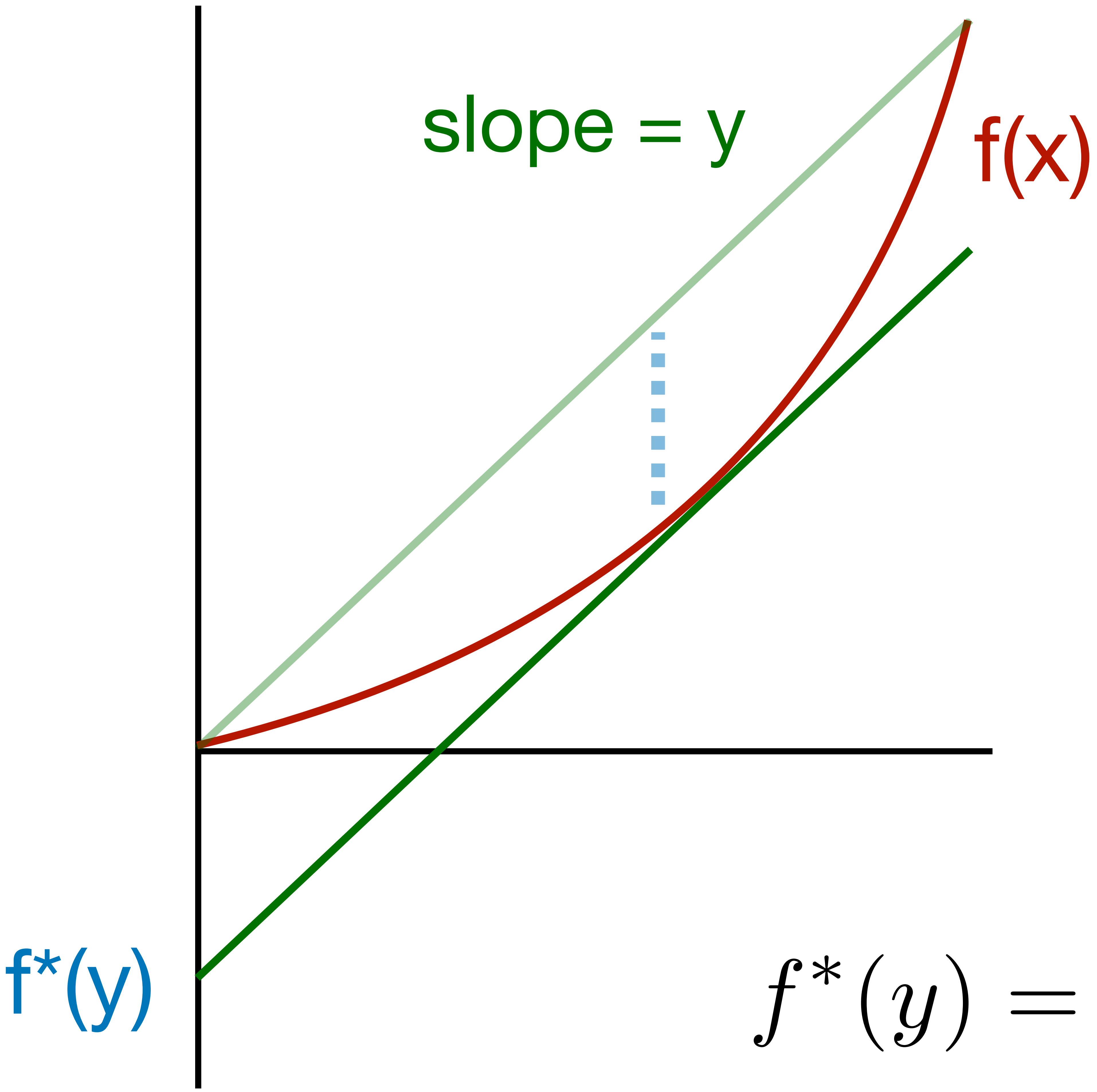


$$f^*(y) = \sup_x xy - f(x)$$

Economic view:
If it costs f(x) to make x widgets and I can sell them for y each, then f*(y) is the max profit I can make

ML view:
If I have a regularizer f(x) and utility function $u_y(x) = xy$, then f*(y) tells me the value of the best regularized utility I can achieve by choosing the best “weight” x for a fixed y:
$f^*(y) = \sup_x u_y(x) - f(x)$

Detour: convex conjugates



$$f^*(y) = \sup_x xy - f(x)$$

Economic view:
If it costs f(x) to make x widgets and I can sell them for y each, then f*(y) is the max profit I can make

(Negative) ML view:
If I have a regularizer f(x) and cost function $c_y(x) = xy$, then f*(y) tells me the value of the largest regularized cost an adversary can force by choosing the worst "weight" x for a fixed y: $f^*(y) = \sup_x c_y(x) - f(x)$

A generalized view of IRL

$$\mathrm{IRL}_\psi(\pi_E) = \underset{c \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max} \, -\psi(c) + \underbrace{\left(\min_{\pi\in\Pi} -H(\pi) + \mathbb{E}_\pi[c(s,a)]\right)}_{\text{Entropy-regularized RL}} - \mathbb{E}_{\pi_E}[c(s,a)]$$

Regularizer

Entropy-regularized RL

Expert performance

Proposition 3.2. $\mathrm{RL} \circ \mathrm{IRL}_\psi(\pi_E) = \arg\min_{\pi\in\Pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E})$

$$\psi^*(y) = \sup_{x\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} x^\top y - \psi(x)$$

Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." *Advances in neural information processing systems* 29 (2016).

A dual optimization view of non-regularized RL+IRL

Corollary 3.2.1. *If ψ is a constant function, $\tilde{c} \in \mathrm{IRL}_\psi(\pi_E)$, and $\tilde{\pi} \in \mathrm{RL}(\tilde{c})$, then $\rho_{\tilde{\pi}} = \rho_{\pi_E}$.*

Proof of Corollary 3.2.1. Define $\bar{L}(\rho, c) = -\bar{H}(\rho) + \sum_{s,a} c(s,a)(\rho(s,a) - \rho_E(s,a))$. Given that ψ is a constant function, we have the following, due to Lemma 3.2:

$$\tilde{c} \in \mathrm{IRL}_\psi(\pi_E) = \underset{c \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max} \min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)] + \text{const.} \tag{5}$$

$$= \underset{c \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max} \min_{\rho \in \mathcal{D}} -\bar{H}(\rho) + \sum_{s,a} \rho(s,a)c(s,a) - \sum_{s,a} \rho_E(s,a)c(s,a) = \underset{c \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max} \min_{\rho \in \mathcal{D}} \bar{L}(\rho, c). \tag{6}$$

This is the dual of the optimization problem

$$\underset{\rho \in \mathcal{D}}{\text{minimize}} -\bar{H}(\rho) \quad \text{subject to} \quad \rho(s,a) = \rho_E(s,a) \quad \forall\, s \in \mathcal{S}, a \in \mathcal{A}$$

with Lagrangian $\bar{L}$, for which the costs $c(s,a)$ serve as dual variables for equality constraints.

Regularized occupancy matching

$$\underset{\pi}{\text{minimize}} \;\; d_\psi(\rho_\pi, \rho_E) - H(\pi) \tag{8}$$

by modifying the IRL regularizer ψ so that $d_\psi(\rho_\pi, \rho_E) \triangleq \psi^*(\rho_\pi - \rho_E)$ smoothly penalizes violations in difference between the occupancy measures.

…but apprenticeship learning (Abbeel and Ng 2004) surprisingly already regularizes:

$$\underset{\pi}{\text{minimize}} \;\; \max_{c\in\mathcal{C}} \mathbb{E}_\pi[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)]$$ for function class $\mathcal{C}_{\text{linear}} = \{\sum_i w_i f_i \;:\; \|w\|_2 \le 1\}$

Define: $\delta_\mathcal{C}(c) = 0$ if $c \in \mathcal{C}$ and $+\infty$ otherwise, then:

$$\max_{c\in\mathcal{C}} \mathbb{E}_\pi[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)] = \max_{c\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} -\delta_\mathcal{C}(c) + \sum_{s,a}(\rho_\pi(s,a) - \rho_{\pi_E}(s,a))c(s,a) = \delta^*_\mathcal{C}(\rho_\pi - \rho_{\pi_E})$$

So what's the problem?

The problem with apprenticeship learning

- If expert's true reward function isn't in the representable class, then we can get poor performance!
- Just because learned policy looks as good as expert on a restricted set of cost functions, doesn't mean we've learned the expert policy. Not smooth regularization — very sharp, in fact.
- Thus, requires very careful feature design
- Can we do better?

GAIL

$$\psi_{\mathrm{GA}}(c) \triangleq \begin{cases} \mathbb{E}_{\pi_E}[g(c(s,a))] & \text{if } c < 0 \\ +\infty & \text{otherwise} \end{cases} \quad \text{where } g(x) = \begin{cases} -x - \log(1-e^x) & \text{if } x < 0 \\ +\infty & \text{otherwise} \end{cases}$$

$$\psi^*_{\mathrm{GA}}(\rho_\pi - \rho_{\pi_E}) = \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} \mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))]$$

GAIL objective:

$$\underset{\pi}{\text{minimize}} \ \psi^*_{\mathrm{GA}}(\rho_\pi - \rho_{\pi_E}) - \lambda H(\pi) = D_{\mathrm{JS}}(\rho_\pi, \rho_{\pi_E}) - \lambda H(\pi)$$

Where: $D_{\mathrm{JS}}(\rho_\pi, \rho_{\pi_E}) \triangleq D_{\mathrm{KL}}\left(\rho_\pi \| (\rho_\pi + \rho_E)/2\right) + D_{\mathrm{KL}}\left(\rho_E \| (\rho_\pi + \rho_E)/2\right)$

GAIL

Algorithm 1 Generative adversarial imitation learning

Input: Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0

for $i = 0, 1, 2, \ldots$ **do**

  Sample trajectories $\tau_i \sim \pi_{\theta_i}$

  Update the discriminator parameters from w_i to w_{i+1} with the gradient

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \tag{17}$$

  Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\begin{aligned} &\hat{\mathbb{E}}_{\tau_i}\left[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)\right] - \lambda \nabla_\theta H(\pi_\theta), \\ &\text{where } Q(\bar{s}, \bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \,|\, s_0 = \bar{s}, a_0 = \bar{a}] \end{aligned} \tag{18}$$

end for

Reward/dynamics entanglement

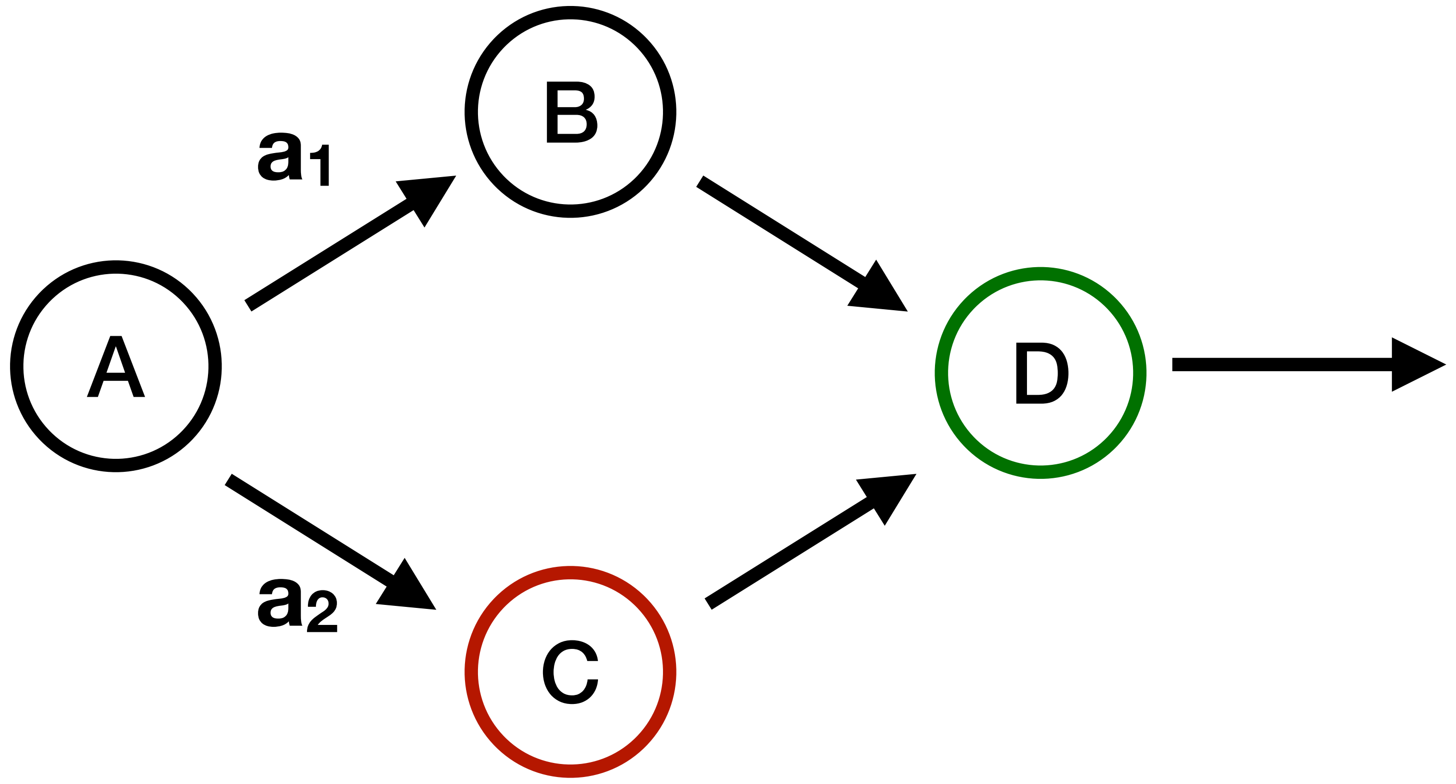
Two types of ambiguity in IRL:

(1) Many different policies explain demonstration data (MaxEnt rectifies this)

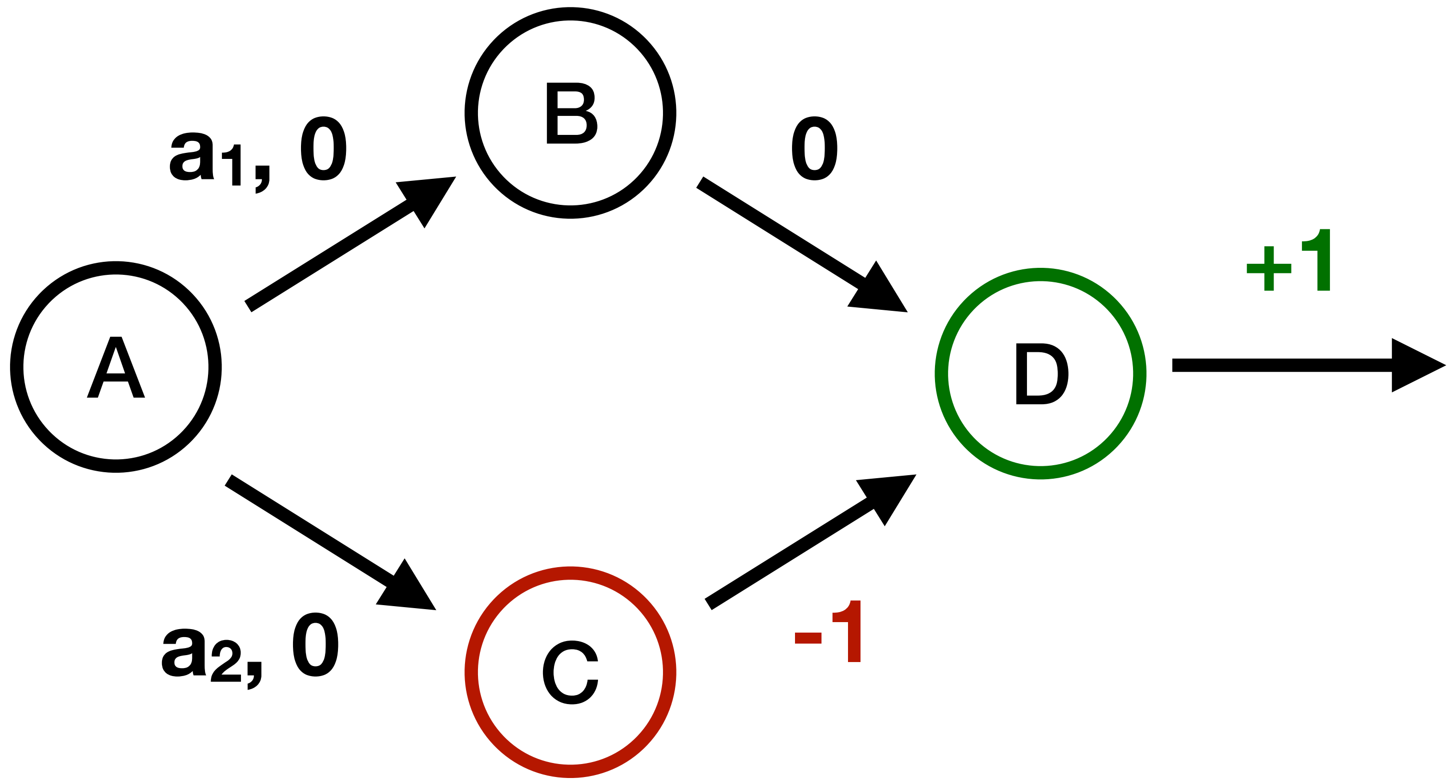
(2) Many different reward functions explain any given policy

- Some of those reward functions may be sparse; some may be heavily shaped
- Of the shaped reward functions, some may have shaping entangled with **dynamics**
- Bad if dynamics change at test time!

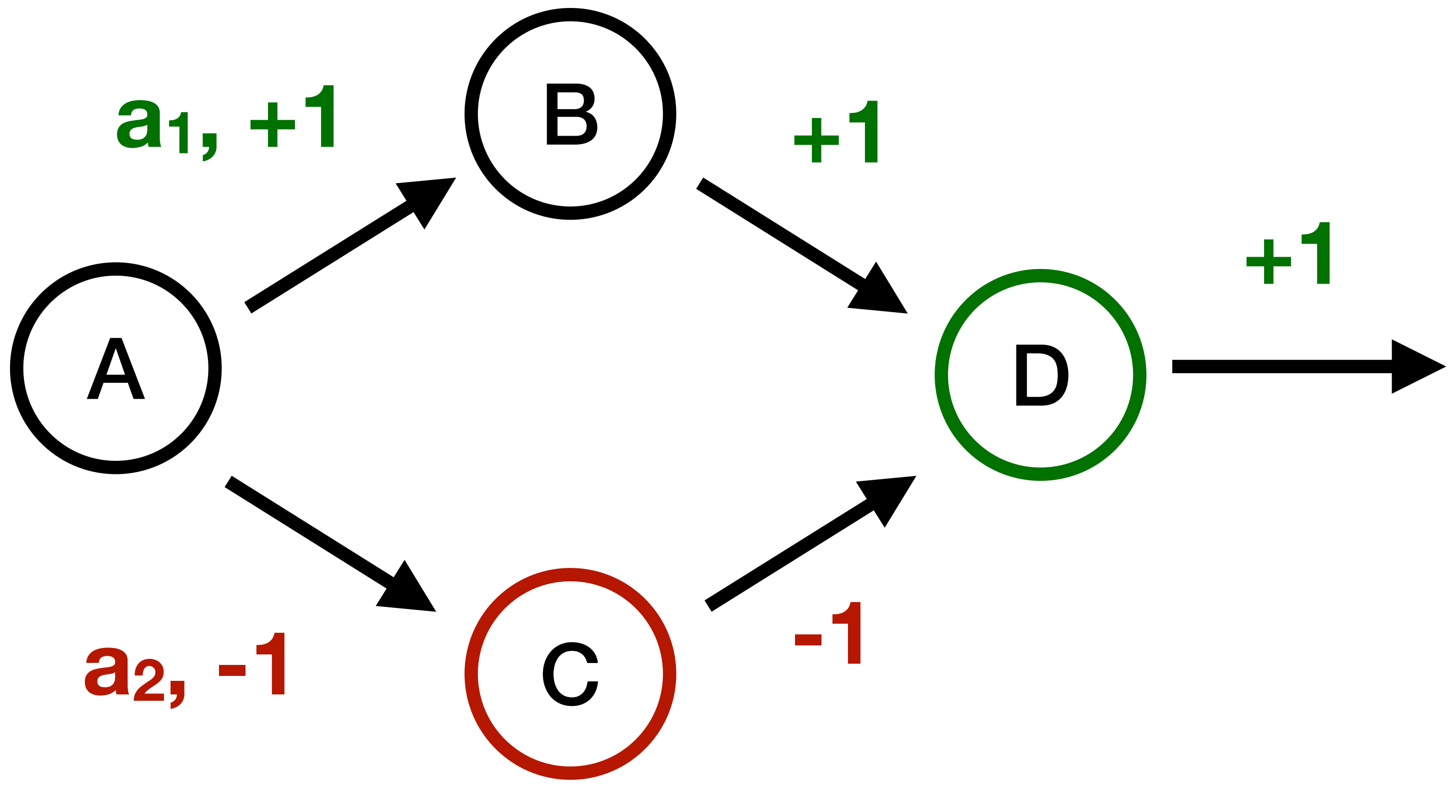
Reward/dynamics entanglement



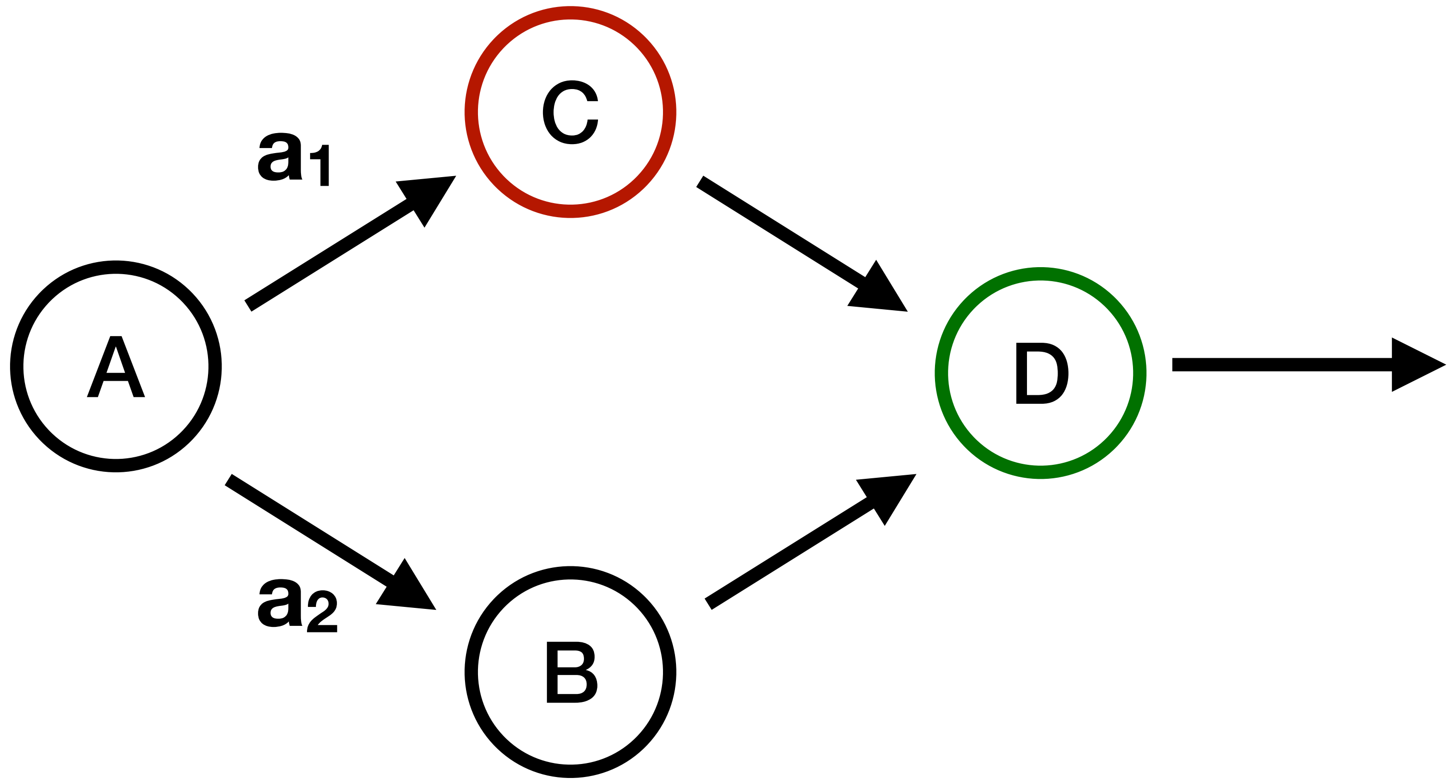
Dynamics



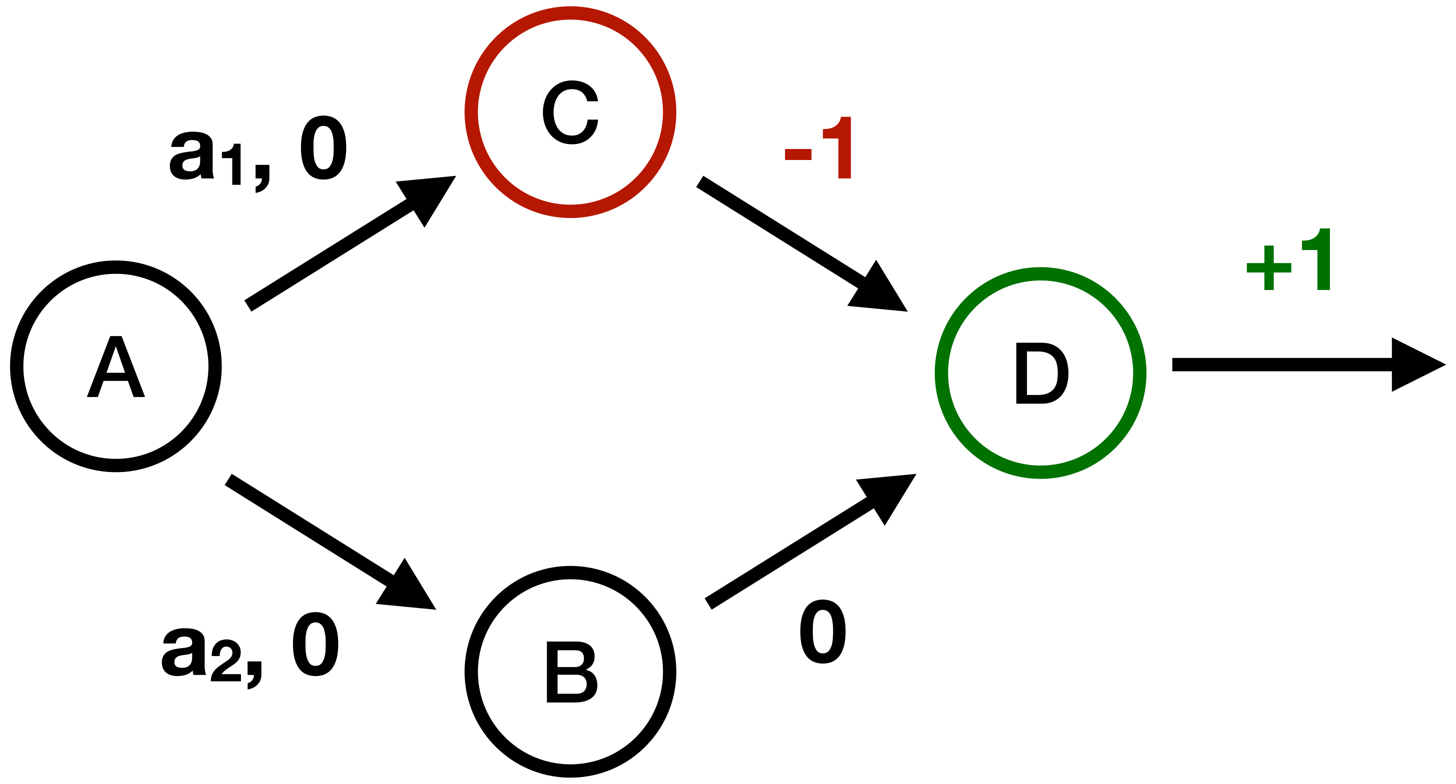
Sparse / ground truth



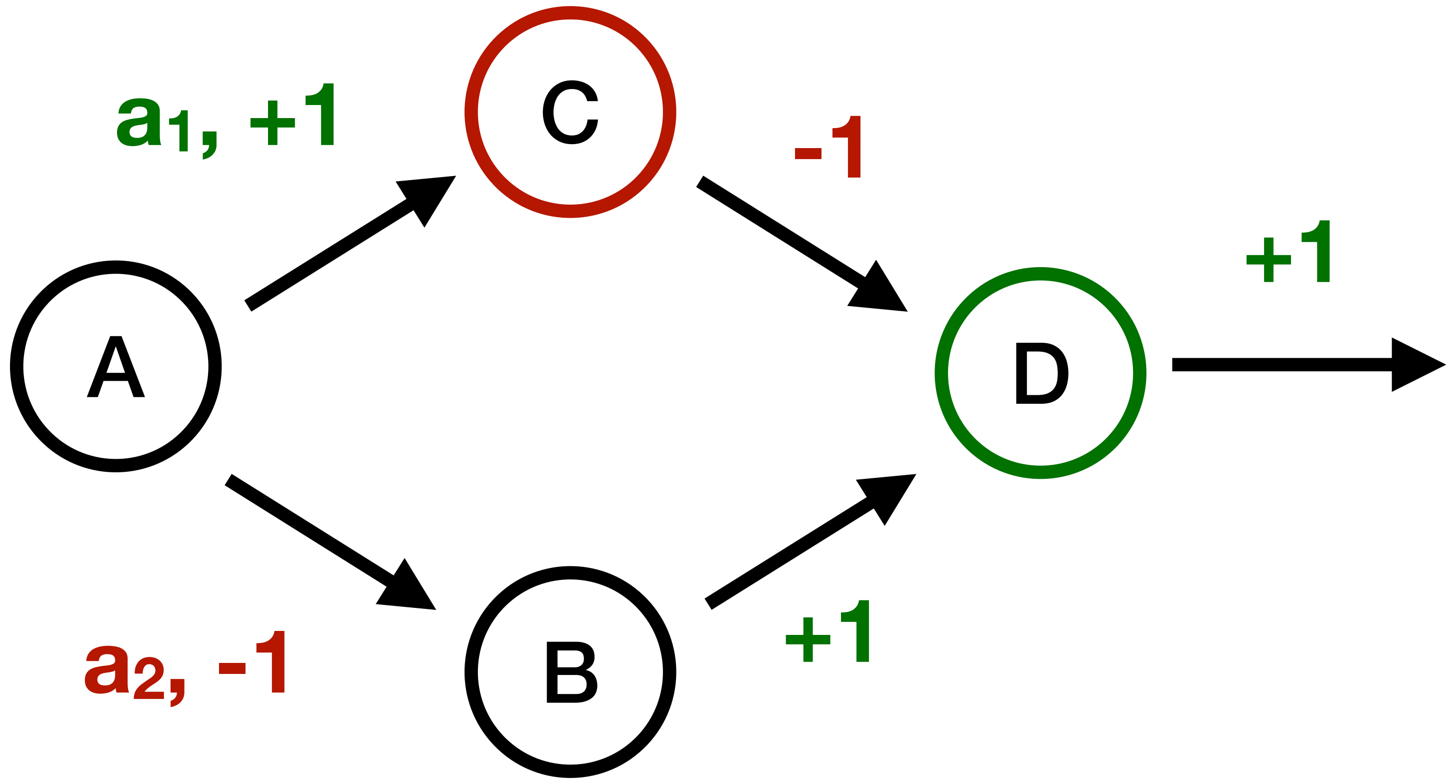
Shaped + dynamics entangled



Dynamics shift!



Sparse / ground truth



Shaped + dynamics entangled

Reward/dynamics entanglement

$$\hat{r}(s, a, s') = r(s, a, s') + \gamma\Phi(s') - \Phi(s)$$

Potential-based reward shaping (Ng et al. 1999)

Assume we have a reward function of the following form for MDP M with deterministic dynamics T:

$$\hat{r}(s, a) = r(s, a) + \gamma\Phi(T(s, a)) - \Phi(s).$$

But then then the MDP changes to M' with dynamics T', where T'(s,a) =/= T(s,a)

$\hat{r}(s, a)$ no longer guaranteed to lead to optimal policies in M' (as judged under r(s,a))

Fu, Justin, Katie Luo, and Sergey Levine. "Learning robust rewards with adversarial inverse reinforcement learning." *arXiv preprint arXiv:1710.11248* (2017).

Disentangled rewards

Definition 5.1 (Disentangled Rewards). *A reward function $r'(s, a, s')$ is (perfectly) disentangled with respect to a ground-truth reward $r(s, a, s')$ and a set of dynamics $\mathcal{T}$ such that under all dynamics $T \in \mathcal{T}$, the optimal policy is the same: $\pi^*_{r',T}(a|s) = \pi^*_{r,T}(a|s)$*

$$\downarrow$$

$$Q^*_{r',T}(s, a) = Q^*_{r,T}(s, a) - f(s)$$

Theorem 5.2. *If a reward function $r'(s, a, s')$ is disentangled for all dynamics functions, then it must be state-only. i.e. If for all dynamics T,*

$$Q^*_{r,T}(s, a) = Q^*_{r',T}(s, a) + f(s) \;\forall s, a$$

Then r' is only a function of state.

Disentangled rewards

$$D_\theta(s,a) = \frac{\exp\{f_\theta(s,a)\}}{\exp\{f_\theta(s,a)\} + \pi(a|s)}$$

$$\downarrow$$

$$D_{\theta,\phi}(s,a,s') = \frac{\exp\{f_{\theta,\phi}(s,a,s')\}}{\exp\{f_{\theta,\phi}(s,a,s')\} + \pi(a|s)}$$

where $f_{\theta,\phi}$ is restricted to a reward approximator g_θ and a shaping term h_ϕ as

$$f_{\theta,\phi}(s,a,s') = g_\theta(s,a) + \gamma h_\phi(s') - h_\phi(s).$$

To be consistent with Sec. $\boxed{4}$, an alternative way to interpret the form of Eqn. $\boxed{4}$ is to view $f_{\theta,\phi}$ as the advantage under deterministic dynamics

$$f^*(s,a,s') = \underbrace{r^*(s) + \gamma V^*(s')}_{Q(s,a)} - \underbrace{V^*(s)}_{V(s)} = A^*(s,a)$$