Robot Learning

GAN, GAIL, MCTEIL

Prof. Songhwai Oh
ECE, SNU

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, Yoshua Bengio, "**Generative Adversarial Nets**," Advances in Neural Information Processing Systems (NIPS), Dec, 2014.

GENERATIVE ADVERSARIAL NETWORKS (GAN)

Adversarial Networks

- $D(x;\theta_d)$: discriminator (NN with parameters θ_d)
  - Probability that x is from the data, not from the generator
- $G(z;\theta_g)$: generator (NN with parameters θ_g)
  - $p_g(x)$: generated sample distribution
- $z \sim p_z(z)$: prior

- Cost function

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Algorithm

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

Analysis

Proposition 1. *For G fixed, the optimal discriminator D is*

$$D_G^*(\boldsymbol{x}) = \frac{p_{data}(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \tag{2}$$

Proof. The training criterion for the discriminator D, given any generator G, is to maximize the quantity $V(G, D)$

$$\begin{aligned} V(G,D) &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x}))dx + \int_z p_{\boldsymbol{z}}(z) \log(1 - D(g(z)))dz \\ &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_g(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))dx \end{aligned} \tag{3}$$

For any $(a, b) \in \mathbb{R}^2 \setminus \{0, 0\}$, the function $y \to a \log(y) + b \log(1 - y)$ achieves its maximum in $[0, 1]$ at $\frac{a}{a+b}$. The discriminator does not need to be defined outside of $Supp(p_{\text{data}}) \cup Supp(p_g)$, concluding the proof. □

Analysis

Theorem 1. *The global minimum of the virtual training criterion $C(G)$ is achieved if and only if $p_g = p_{data}$. At that point, $C(G)$ achieves the value $-\log 4$.*

$$\begin{aligned} C(G) &= \max_D V(G,D) \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\log(1 - D_G^*(G(\boldsymbol{z})))] \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log(1 - D_G^*(\boldsymbol{x}))] \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \end{aligned} \tag{4}$$

Proof. For $p_g = p_{\text{data}}$, $D_G^*(\boldsymbol{x}) = \frac{1}{2}$, (consider Eq. 2). Hence, by inspecting Eq. 4 at $D_G^*(\boldsymbol{x}) = \frac{1}{2}$, we find $C(G) = \log \frac{1}{2} + \log \frac{1}{2} = -\log 4$. To see that this is the best possible value of $C(G)$, reached only for $p_g = p_{\text{data}}$, observe that

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[-\log 2] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[-\log 2] = -\log 4$$

and that by subtracting this expression from $C(G) = V(D_G^*, G)$, we obtain:

$$C(G) = -\log(4) + KL\left(p_{\text{data}} \,\middle\|\, \frac{p_{\text{data}} + p_g}{2}\right) + KL\left(p_g \,\middle\|\, \frac{p_{\text{data}} + p_g}{2}\right) \tag{5}$$

where KL is the Kullback–Leibler divergence. We recognize in the previous expression the Jensen–Shannon divergence between the model's distribution and the data generating process:

$$C(G) = -\log(4) + 2 \cdot JSD\left(p_{\text{data}} \,\|\, p_g\right) \tag{6}$$

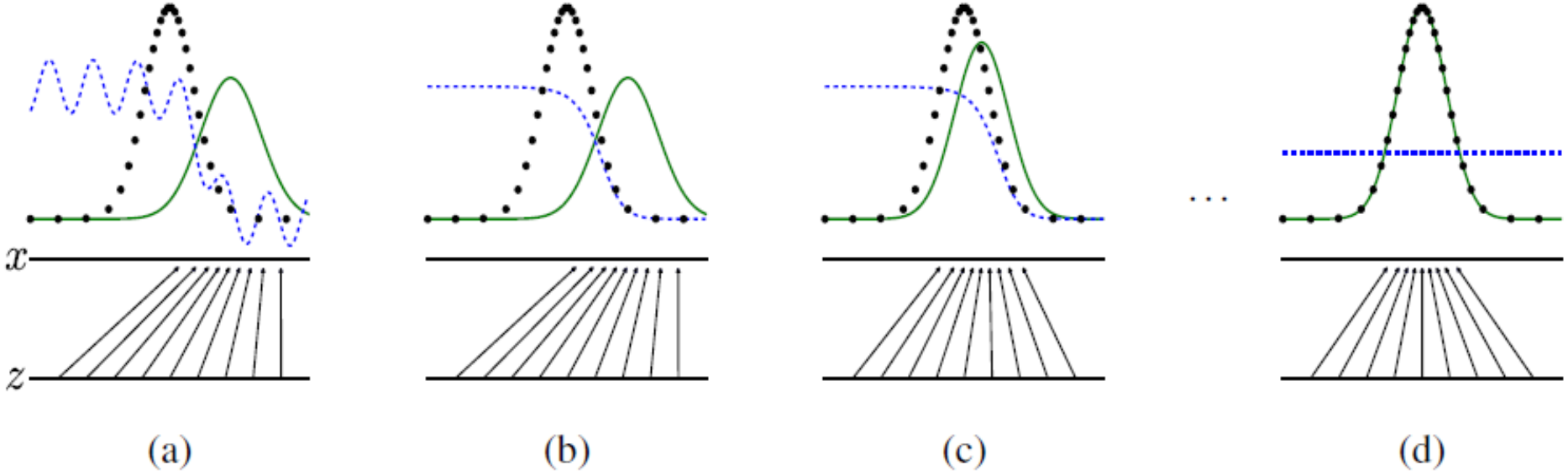
Since the Jensen–Shannon divergence between two distributions is always non-negative, and zero iff they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ and that the only solution is $p_g = p_{\text{data}}$, i.e., the generative model perfectly replicating the data distribution. □

Convergence

Proposition 2. *If G and D have enough capacity, and at each step of Algorithm 1, the discriminator is allowed to reach its optimum given G, and p_g is updated so as to improve the criterion*

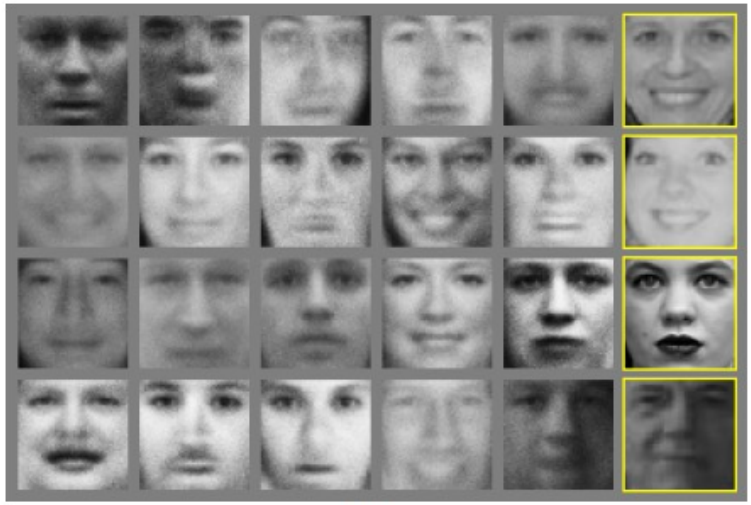
$$\mathbb{E}_{\boldsymbol{x}\sim p_{data}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log(1 - D_G^*(\boldsymbol{x}))]$$

then p_g converges to p_{data}



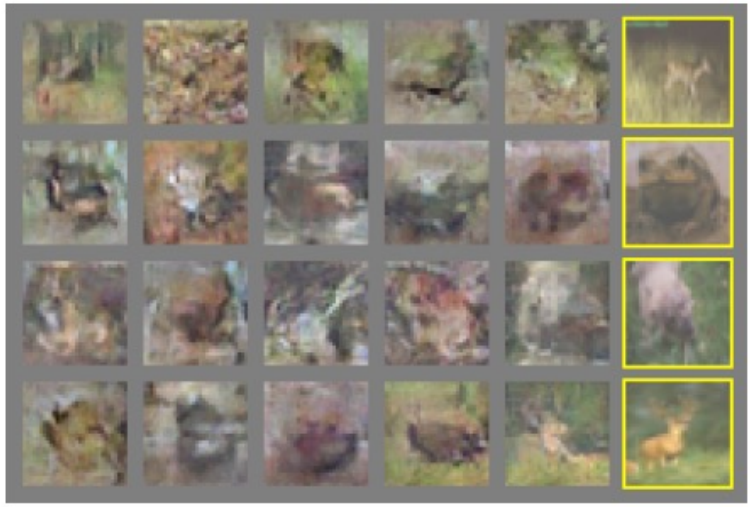
Experiments

a)



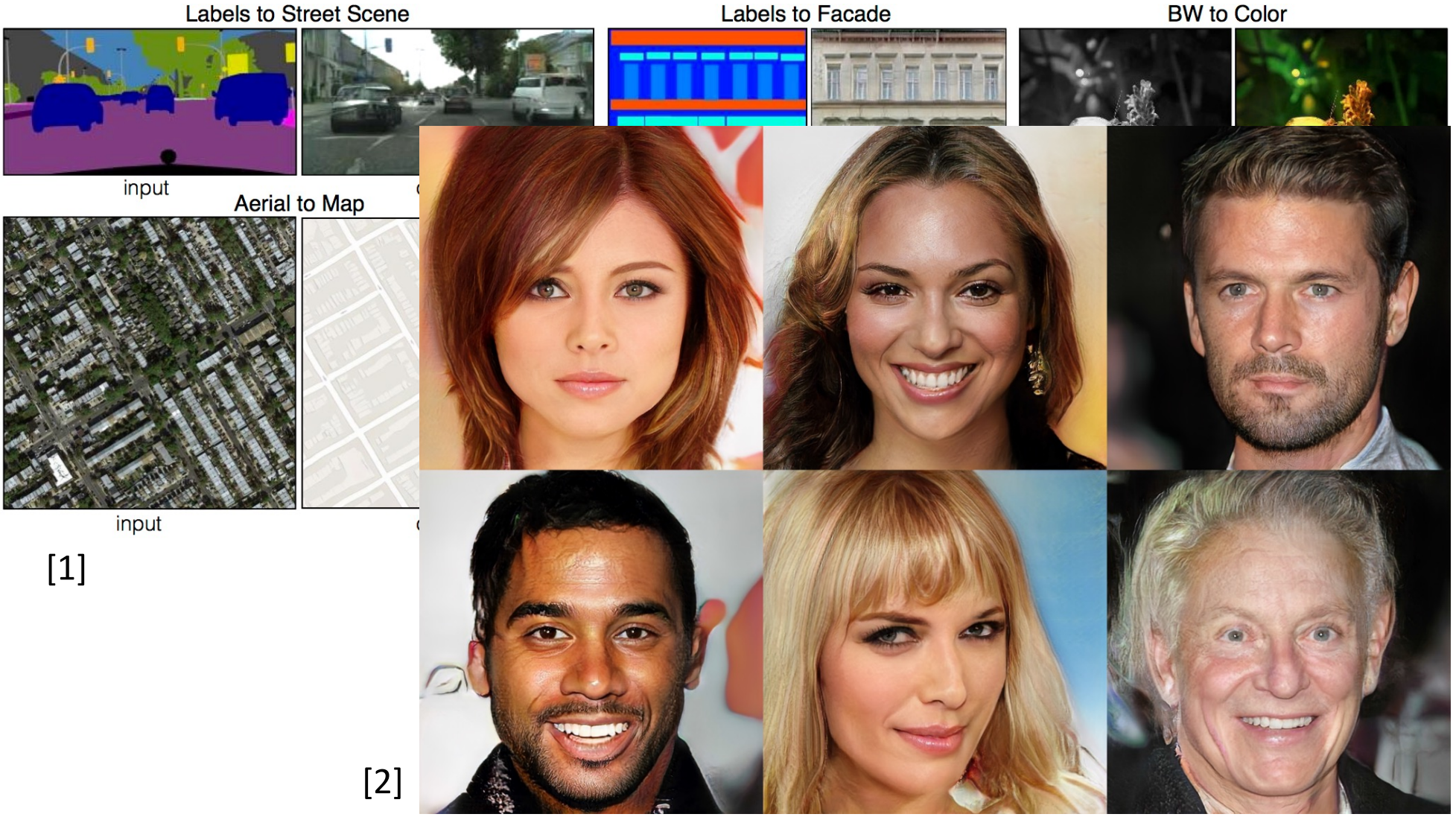
b)

c)



d)

State-of-the-Art GANs



[1]

[2]

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros , "**Image-to-Image Translation with Conditional Adversarial Nets**," CVPR 2017.
[2] Karras, T., Aila, T., Laine, S., & Lehtinen, J. **Progressive growing of GANs for improved quality, stability, and variation.** ICLR 2018.

J. Ho, and S. Ermon, "**Generative adversarial imitation learning**," Advances in Neural Information Processing Systems (NIPS), Dec, 2016

GENERATIVE ADVERSARIAL IMITATION LEARNING (GAIL)

Inverse Reinforcement Learning

- $E_\pi[c(s,a)] = E[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)]$: expected discounted cost
- $H(\pi) = E_\pi[-\log \pi(a|s)]$: discounted causal entropy

- Inverse Reinforcement Learning (**MaxEnt IRL**, dual form)

$$\underset{c \in \mathcal{C}}{\text{maximize}} \left(\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[c(s,a)] \right) - \mathbb{E}_{\pi_E}[c(s,a)]$$

- Reinforcement Learning

$$\text{RL}(c) = \arg\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[c(s,a)]$$

- GAIL: Solves IRL without the RL step

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

regularizer

Dual for MDPs

- $\rho_\pi(s,a) = \pi(a|s)\sum_{t=0}^{\infty}\gamma^t P(s_t = s|\pi)$: occupancy measure wrt π
  - $E_\pi[c(s,a)] = E[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)] = \sum_{s,a}\rho_\pi(s,a)c(s,a)$

Proposition 3.1 (Theorem 2 of Syed et al. [29]). *If $\rho \in \mathcal{D}$, then ρ is the occupancy measure for $\pi_\rho(a|s) \triangleq \rho(s,a)/\sum_{a'}\rho(s,a')$, and π_ρ is the only policy whose occupancy measure is ρ.*

Bellman flow constraint: $\mathcal{D} = \left\{\rho : \rho \geq 0 \quad \text{and} \quad \sum_a \rho(s,a) = p_0(s) + \gamma\sum_{s',a} P(s|s',a)\rho(s',a) \;\; \forall\, s \in \mathcal{S}\right\}$

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

- ψ^* is a convex conjugate of ψ, i.e., $\psi^*(x) = \sup_y [x^T y - \psi(y)]$.
- This IRL formulation finds a policy which matches its occupancy measure to the occupancy measure of an expert.

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

Let $\tilde{c} \in \mathrm{IRL}_\psi(\pi_E)$, $\tilde{\pi} \in \mathrm{RL}(\tilde{c}) = \mathrm{RL} \circ \mathrm{IRL}_\psi(\pi_E)$, and

$$\pi_A \in \arg\min_\pi -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E}) \tag{31}$$

(convex conjugate)

$$= \arg\min_\pi \max_c -H(\pi) - \psi(c) + \sum_{s,a} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a))c(s,a) \tag{32}$$

We wish to show that $\pi_A = \tilde{\pi}$. To do this, let ρ_A be the occupancy measure of π_A, let $\tilde{\rho}$ be the occupancy measure of $\tilde{\pi}$, and define $\bar{L} : \mathcal{D} \times \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$ by

$$\bar{L}(\rho, c) = -\bar{H}(\rho) - \psi(c) + \sum_{s,a} \rho(s,a)c(s,a) - \sum_{s,a} \rho_{\pi_E}(s,a)c(s,a). \tag{33}$$

$$\rho_A \in \arg\min_{\rho \in \mathcal{D}} \max_c \bar{L}(\rho, c), \tag{34}$$

$$\tilde{c} \in \arg\max_c \min_{\rho \in \mathcal{D}} \bar{L}(\rho, c), \tag{35}$$

$$\tilde{\rho} \in \arg\min_{\rho \in \mathcal{D}} \bar{L}(\rho, \tilde{c}). \tag{36}$$

$$\min_{\rho \in \mathcal{D}} \max_{c \in \mathcal{C}} \bar{L}(\rho, c) = \max_{c \in \mathcal{C}} \min_{\rho \in \mathcal{D}} \bar{L}(\rho, c) \tag{37}$$

(minimax duality)

Hence, from Eqs. (34) and (35), $(\rho_A, \tilde{c})$ is a saddle point of $\bar{L}$, which implies that

$$\rho_A \in \arg\min_{\rho \in \mathcal{D}} \bar{L}(\rho, \tilde{c}). \tag{38}$$

Because $\bar{L}(\cdot, c)$ is strictly convex for all c (Lemma 3.1), Eqs. (36) and (38) imply $\rho_A = \tilde{\rho}$. Since policies corresponding to occupancy measures are unique (Proposition 3.1), we get $\pi_A = \tilde{\pi}$. □

Generative Adversarial Imitation Learning (GAIL)

$$\psi_{\mathrm{GA}}(c) \triangleq \begin{cases} \mathbb{E}_{\pi_E}[g(c(s,a))] & \text{if } c < 0 \\ +\infty & \text{otherwise} \end{cases} \quad \text{where } g(x) = \begin{cases} -x - \log(1-e^x) & \text{if } x < 0 \\ +\infty & \text{otherwise} \end{cases} \tag{13}$$

$$\psi^*_{\mathrm{GA}}(\rho_\pi - \rho_{\pi_E}) = \max_{D \in (0,1)^{\mathcal{S}\times\mathcal{A}}} \mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1-D(s,a))] \tag{14}$$

connection to GAN !!!

GAIL: $$\min_\pi \max_D \; \mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1-D(s,a))] - \lambda H(\pi)$$

Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

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

5: 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

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

6: **end for**

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

$$R_\phi(\pi, \pi_E) = \sum_{s,a} \min_{\gamma \in \mathbb{R}} \rho_\pi(s,a)\phi(\gamma) + \rho_{\pi_E}(s,a)\phi(-\gamma)$$

ϕ: strictly decreasing convex loss function

$$g_\phi(x) = \begin{cases} -x + \phi(-\phi^{-1}(-x)) & \text{if } x \in T \\ +\infty & \text{otherwise} \end{cases}$$

$$\psi_\phi(c) = \begin{cases} \sum_{s,a} \rho_{\pi_E}(s,a) g_\phi(c(s,a)) & \text{if } c(s,a) \in T \text{ for all } s,a \\ +\infty & \text{otherwise} \end{cases}$$

$$\begin{aligned}
\psi^*_\phi(\rho_\pi - \rho_{\pi_E}) &= \max_{c \in \mathcal{C}} \sum_{s,a} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a))c(s,a) - \sum_{s,a} \rho_{\pi_E}(s,a) g_\phi(c(s,a)) \\
&= \sum_{s,a} \max_{c \in T} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a))c - \rho_{\pi_E}(s,a)[-c + \phi(-\phi^{-1}(-c))] \\
&= \sum_{s,a} \max_{c \in T} \rho_\pi(s,a)c - \rho_{\pi_E}(s,a)\phi(-\phi^{-1}(-c)) \\
&= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a)(-\phi(\gamma)) - \rho_{\pi_E}(s,a)\phi(-\phi^{-1}(\phi(\gamma))) \qquad c \to -\phi(\gamma) \\
&= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a)(-\phi(\gamma)) - \rho_{\pi_E}(s,a)\phi(-\gamma) \\
&= -R_\phi(\rho_\pi, \rho_{\pi_E})
\end{aligned}$$

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

Corollary A.1.1. *The cost regularizer* (13)

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

satisfies

$$\psi^*_{GA}(\rho_\pi - \rho_{\pi_E}) = \max_{D \in (0,1)^{\mathcal{S}\times\mathcal{A}}} \mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))]. \tag{47}$$

Proof. Using the logistic loss $\phi(x) = \log(1 + e^{-x})$, we see that Eq. (40) reduces to the claimed ψ_{GA}. Applying Proposition A.1, we get

$$\psi^*_{\text{GA}}(\rho_\pi - \rho_{\pi_E}) = -R_\phi(\rho_\pi, \rho_{\pi_E}) \tag{48}$$

$$= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a) \log\left(\frac{1}{1 + e^{-\gamma}}\right) + \rho_{\pi_E}(s,a) \log\left(\frac{1}{1 + e^{\gamma}}\right) \tag{49}$$

$$= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a) \log\left(\frac{1}{1 + e^{-\gamma}}\right) + \rho_{\pi_E}(s,a) \log\left(1 - \frac{1}{1 + e^{-\gamma}}\right) \tag{50}$$

$$= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a) \log(\sigma(\gamma)) + \rho_{\pi_E}(s,a) \log(1 - \sigma(\gamma)), \tag{51}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function. Because the range of σ is $(0, 1)$, we can write

$$\psi^*_{\text{GA}}(\rho_\pi - \rho_{\pi_E}) = \sum_{s,a} \max_{d \in (0,1)} \rho_\pi(s,a) \log d + \rho_{\pi_E}(s,a) \log(1 - d) \tag{52}$$

$$= \max_{D \in (0,1)^{\mathcal{S}\times\mathcal{A}}} \sum_{s,a} \rho_\pi(s,a) \log(D(s,a)) + \rho_{\pi_E}(s,a) \log(1 - D(s,a)), \tag{53}$$

which is the desired expression. □

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh, "**Maximum Causal Tsallis Entropy Imitation Learning**", in Proc. of Neural Information Processing Systems (NIPS), Dec. 2018.

MAXIMUM CAUSAL TSALLIS ENTROPY IMITATION LEARNING

Inverse Reinforcement Learning (IRL)

Maximum Entropy IRL

$$\max_{\pi} \alpha H(\pi)$$

$$\text{subject to} \quad \forall\, s, a \quad \sum_{a} \pi(a|s) = 1, \qquad \pi(a|s) \geq 0$$

$$\forall\, s, a \quad \mathbb{E}_{\pi}[\phi(s,a)] = \mathbb{E}_{\pi_E}[\phi(s,a)]$$

- Causal entropy regularization: $H(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} -\gamma^t \, log(\pi(a_t|s_t))]$
- Feature expectation matching: behave similarly to expert's policy
- Dual problem of maximum causal entropy problem

$$\max_{\theta} \min_{\pi} -\alpha H(\pi) - \mathbb{E}_{\pi}[\theta^T \phi(s,a)] + \mathbb{E}_{\pi_E}[\theta^T \phi(s,a)]$$

Policy Learning: Soft MDP under cost $-\theta^T\phi$ or reward $\theta^T\phi$

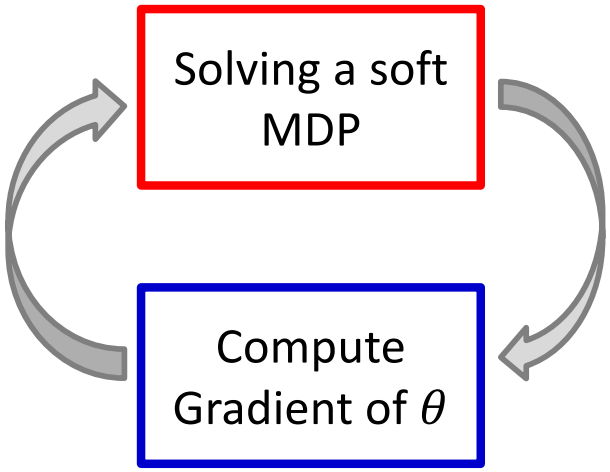
Reward Learning: maximizing the performance gap between the learner and expert

Ziebart et al. "**Maximum Entropy Inverse Reinforcement Learning**." *AAAI*. Vol. 8. 2008.
Michael Bloem and Nicholas Bambos. "**Infinite time horizon maximum causal entropy inverse reinforcement learning**." *CDC*, 2014.

Maximum Entropy IRL

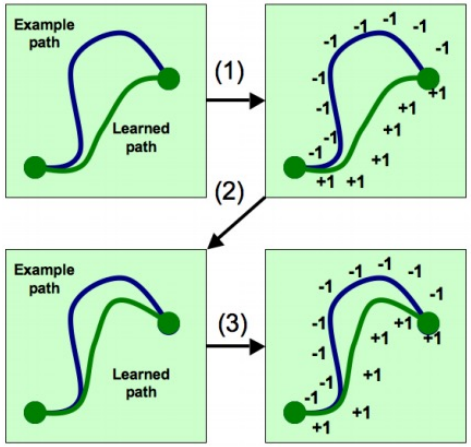
MaxEnt IRL

$$\max_{\theta} \min_{\pi} -\alpha H(\pi) - \mathbb{E}_{\pi}[\theta^T \phi(s,a)] + \mathbb{E}_{\pi_E}[\theta^T \phi(s,a)]$$



Optimal Solution of MaxEnt IRL

- $r(s,a) = \theta^{*T} \phi(s,a)$
- $q(s,a) = r(s,a) + \sum_{s'} v(s') T(s'|s,a)$
- $v(s) = \log \sum_{a'} q(s,a')$
- $\pi(a|s) = \frac{1}{Z} \exp \frac{q(s,a')}{\alpha}$ ← **Softmax**



- Drawbacks
  - Not scalable
  - Transition model is required
  - Solving RL as a subroutine is computationally expensive

Brian D. Ziebart, "**Modeling purposeful adaptive behavior with the principle of maximum causal entropy**." Ph.D. Thesis, CMU, 2010.

Generative Adversarial Imitation Learning (GAIL)

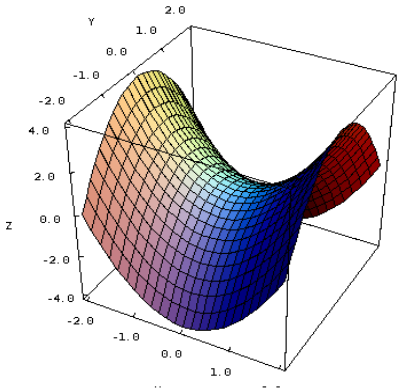
Unifying View of IRL

$$\max_{c} \min_{\pi} -\alpha H(\pi) + \mathbb{E}_{\pi}[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)] - \psi(c)$$

- $\psi(c)$ is reward regularization, where $c = -\theta^T \phi$
- Many existing IRL methods can be interpreted under this framework
  - Apprenticeship Learning: l2-ball regularization
  - MaxEnt IRL: constant
  - Gaussian process (GP) IRL: Hilbert norm regularization

Generative Adversarial Setting

$$\min_{\pi} \max_{D} -\alpha H(\pi) + \mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))]$$



- It does not require to solve MDPs or RL at every iteration
- Sample efficient: after sampling from policy, both discriminator and policy are updated
- Note the similarity to generative adversarial networks (GAN)

Jonathan Ho and Stefano Ermon. "**Generative adversarial imitation learning**." NIPS. 2016.

Maximum Causal Tsallis Entropy Imitation Learning

Maximum Causal Tsallis Entropy Imitation Learning (MCTEIL)

$$\min_{\pi} \max_{D} -\alpha W(\pi) + \mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log 1 - D(s,a)]$$

Reinforcement Learning Problem

Real Fake Classification Problem

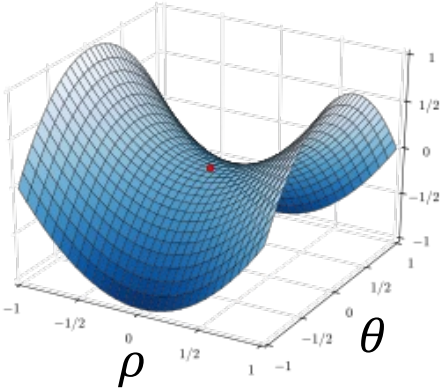
- By setting regularization to $\psi_{GA}(\theta)$

$$\psi_{GA}(\theta) \coloneqq \mathbb{E}_{\pi}[g(\theta^T \phi(s,a))]$$

$$\text{where } g(x) = \begin{cases} -x - \log(1 - e^x), & x < 0 \\ \infty, & x \geq 0 \end{cases}$$

- $-\log D(s,a)$ is the reward function

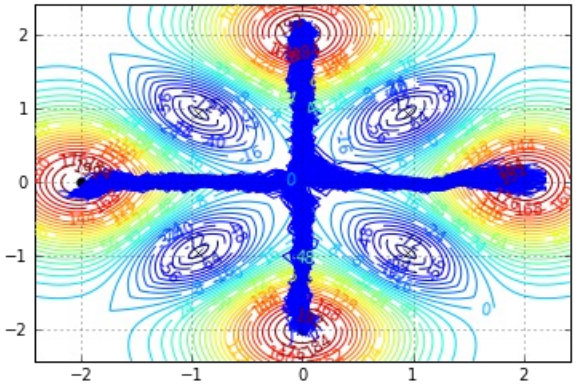
- Since the objective function has a unique saddle point, we can update policy and reward simultaneously



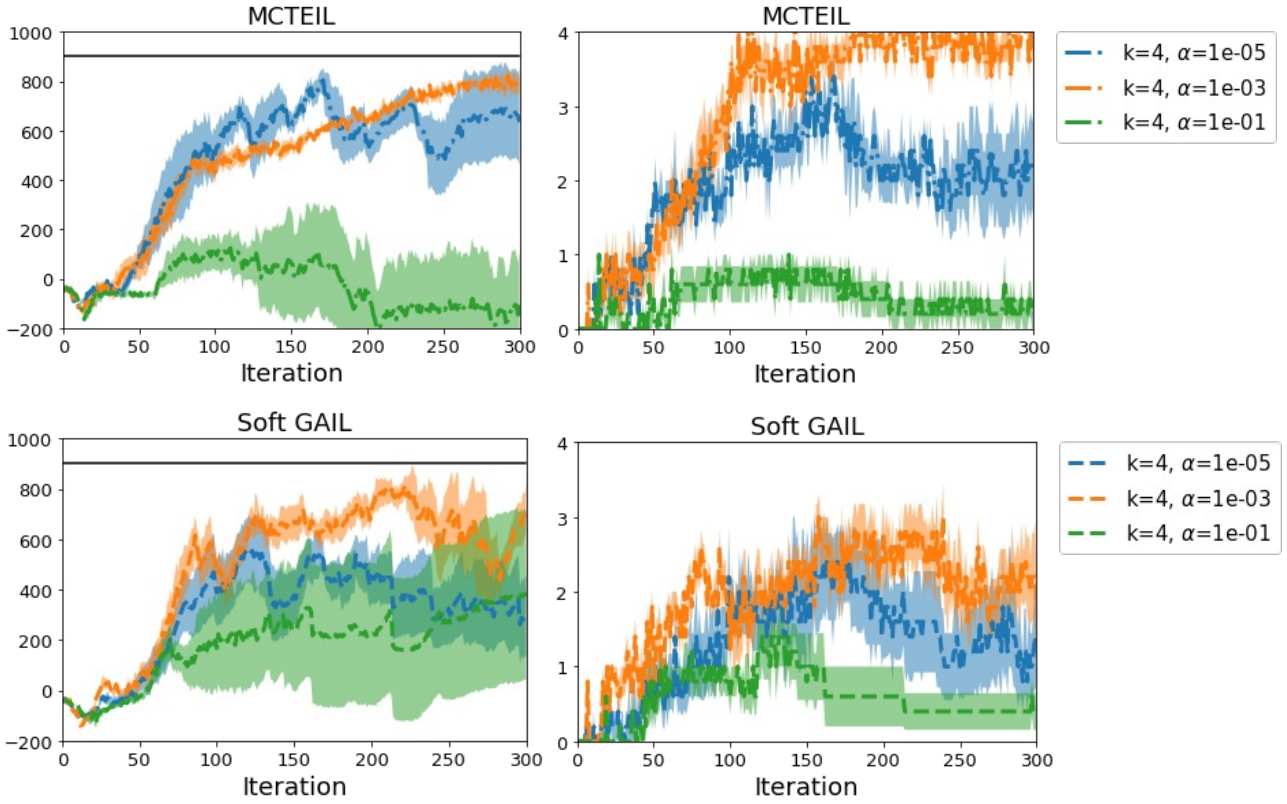
Experiments

- **Multi-Goal Environment**
  - Simple 2D environment with multiple optima in the rewards
  - Agent follows pointmass dynamics
  - Verifying that the proposed method can learn multi-modal behavior
- **MuJoCo (Continuous Control)**
  - Four robotic control problems: Reacher, Half Cheetah, Ant, Walker2d
  - Comparison with GAIL
- **Object World and Highway Driving (Discrete Control)**
  - Control agents with discrete action space
  - Omitted in paper

Multi-Goal Experiments



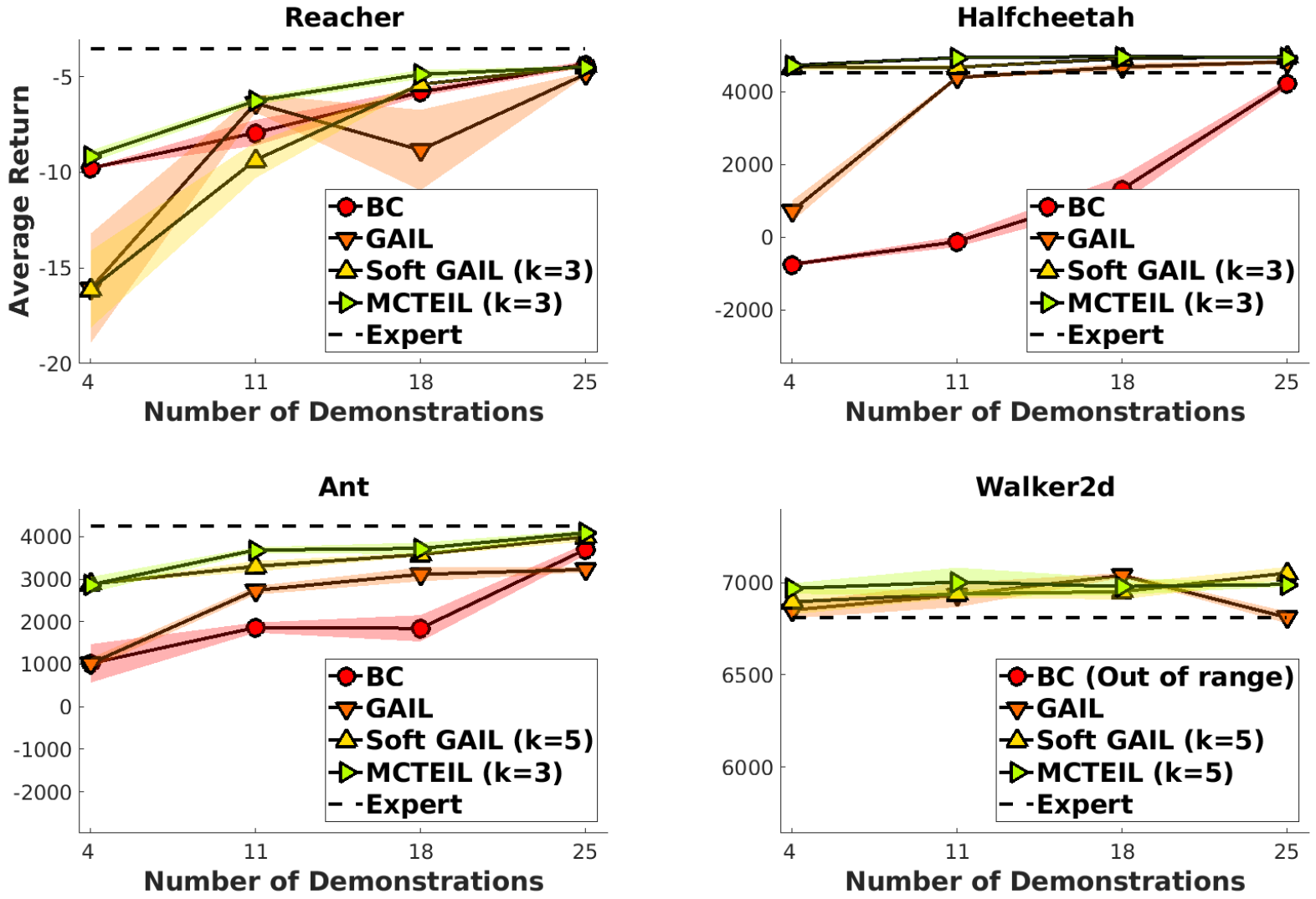
Rewards map and Demos



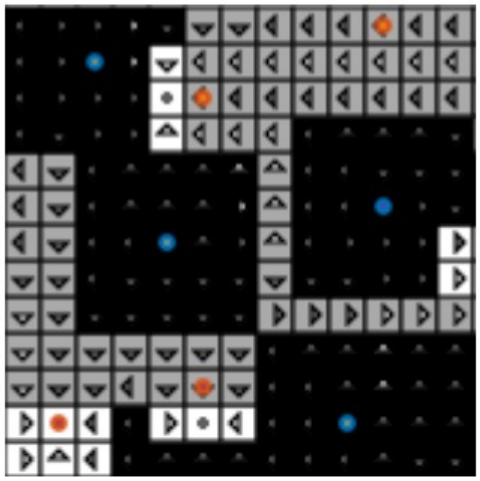
Average Return

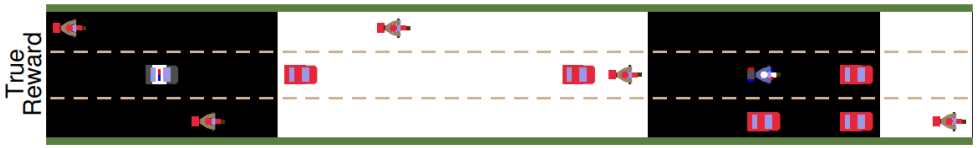
Reachability

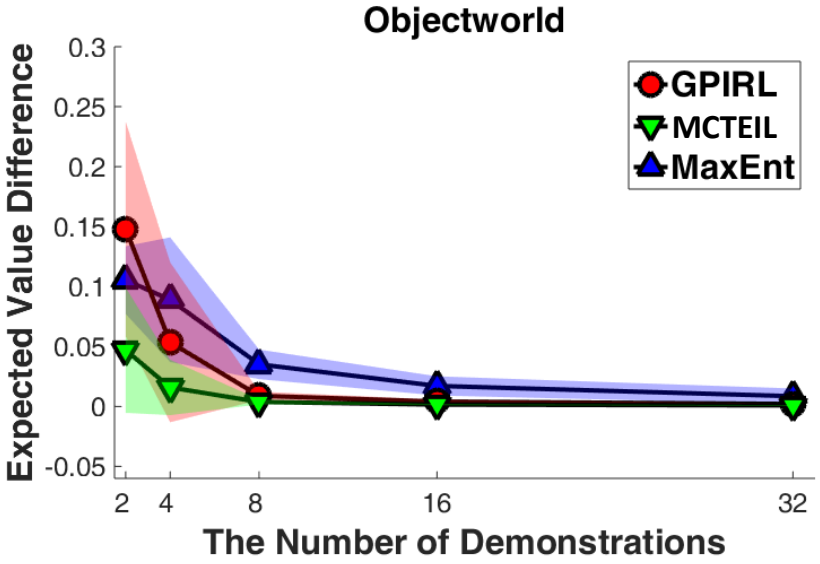
MuJoCo Experiments (Continuous Control)

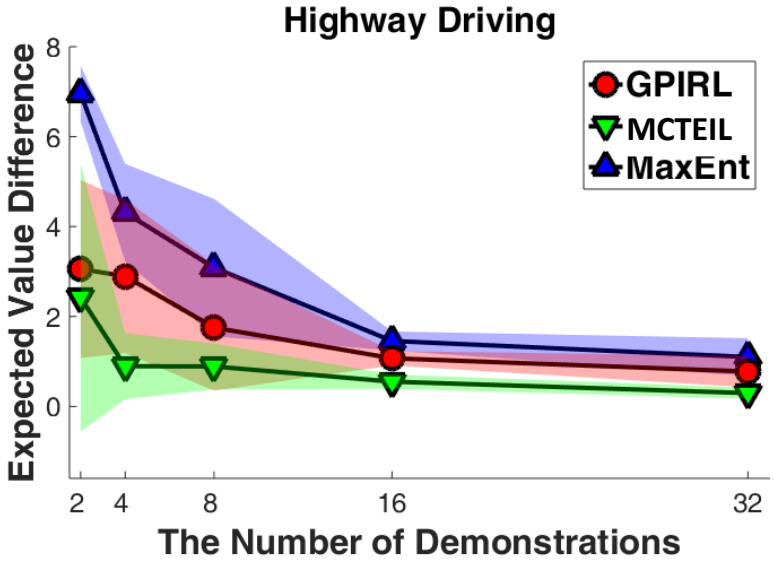


More Experiments



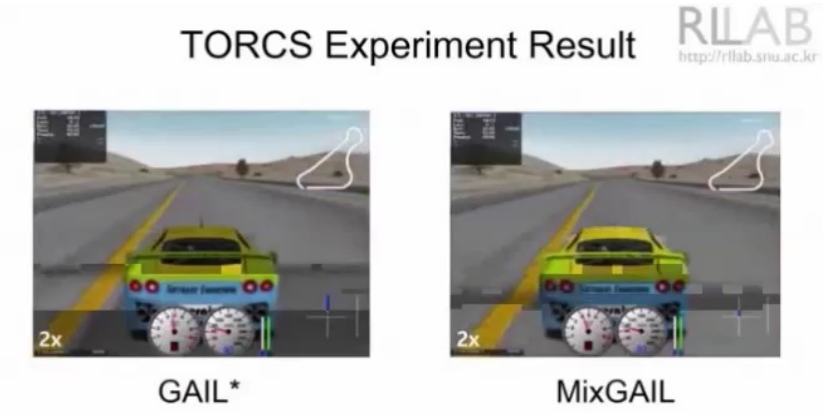






MixGAIL:
Autonomous Driving Using Demonstrations with Mixed Qualities

- Uses both expert demonstrations and negative demonstration (e.g., accidents)
- Experimental results
  - TORCS: Racing car simulator
  - RC Car with a lidar sensor
  - Learns faster and performs better



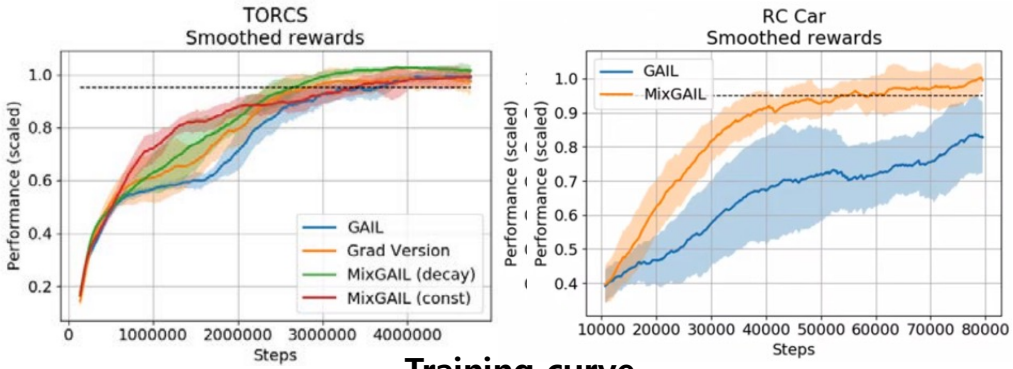
Optimization Problem.

$$\min_{\pi\in\Pi}\max_{D} E_{\pi}[\log(D(s,a)) - \eta \mathrm{CoR}(s, s_E, s_N)] + E_{\pi_E}[\log(1 - D(s,a))] - \lambda H(\pi),$$

$$\mathrm{CoR}(s, s_E, s_N) = \frac{(1+\frac{d_{s_E}}{\alpha})^{-\frac{\alpha+1}{2}}}{(1+\frac{d_{s_E}}{\alpha})^{-\frac{\alpha+1}{2}} + (1+\frac{d_{s_N}}{\alpha})^{-\frac{\alpha+1}{2}}},$$

$$d_{s_E} = \sqrt{\frac{1}{|s_E|}\sum_{s_e \in s_E} (\|s - s_e\|_2)^2},$$

$$d_{s_N} = \sqrt{\frac{1}{|s_N|}\sum_{s_n \in s_N} (\|s - s_n\|_2)^2},$$



Training curve.

Gunmin Lee, Dohyeong Kim, Wooseok Oh, Kyungjae Lee, and Songhwai Oh, "**MixGAIL: Autonomous Driving Using Demonstrations with Mixed Qualities**," in Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2020.