Brain Inspired Artificial Intelligence 7: Introduction to generative adversarial imitation learning

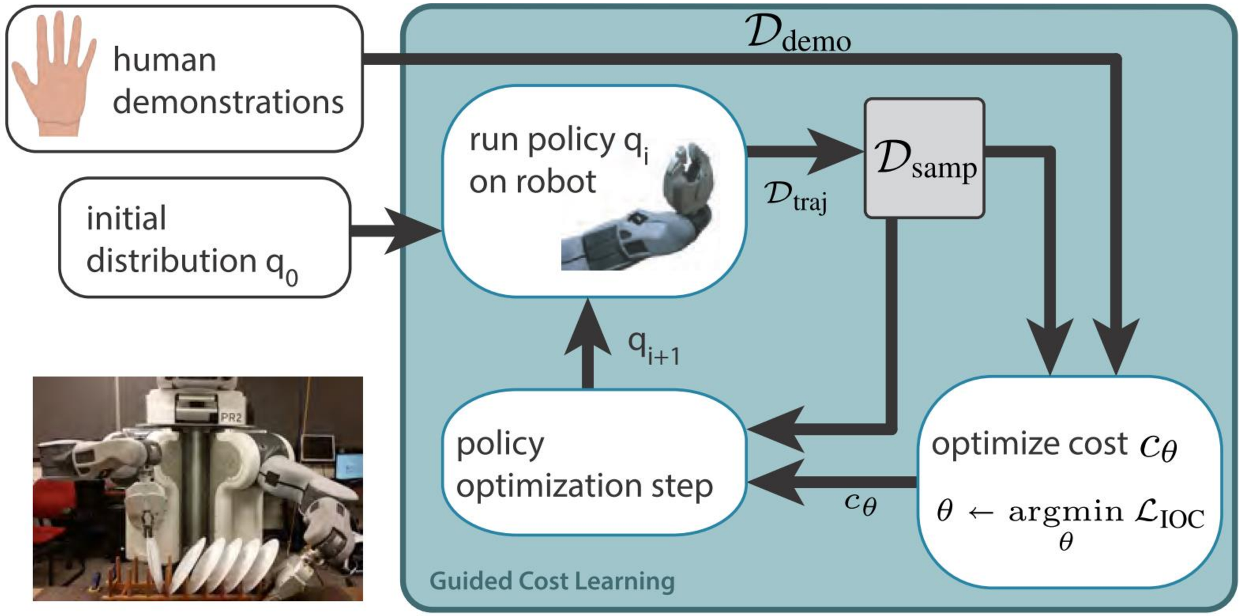
Eiji Uchibe

Dept. of Brain Robot Interface

ATR Computational Neuroscience Labs.

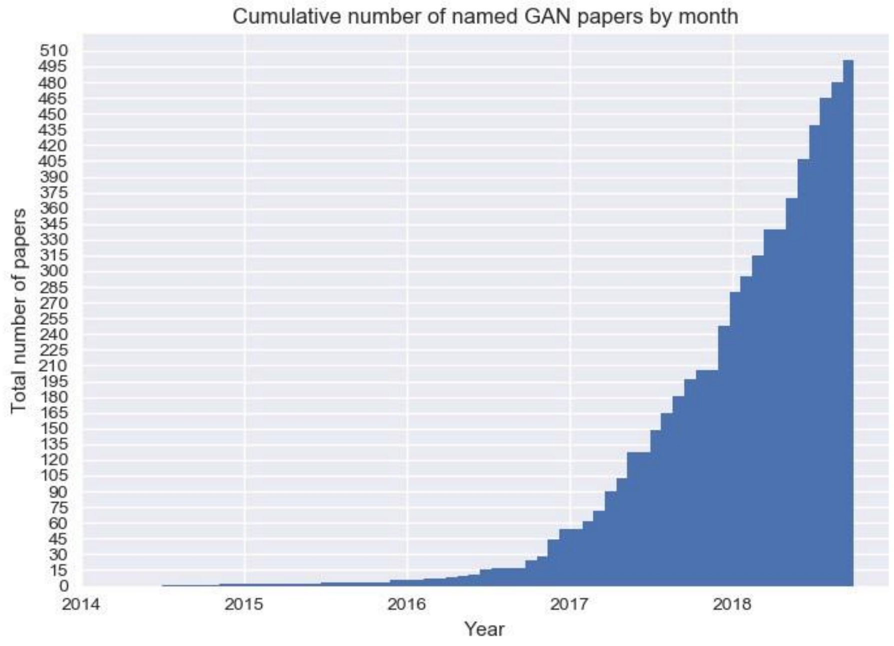
Reward estimation by forward and inverse reinforcement learning

- Recap: Guided Cost Learning (Finn et al., 2016a)
  - The normalizing constant Z and its derivative should be evaluated to estimate reward
  - Z is evaluated by importance sampling estimator
  - Forward reinforcement learning is used to improve a sampling distribution
- The entire architecture is very similar to that of Generative Adversarial Networks (GANs)



Generative Adversarial Network (GAN)

- Minimax game between Generator and Discriminator
- Generator wants to minimize the objective function J while Discriminator wants to maximize it.



https://deephunt.in/the-gan-zoo-79597dc8c347

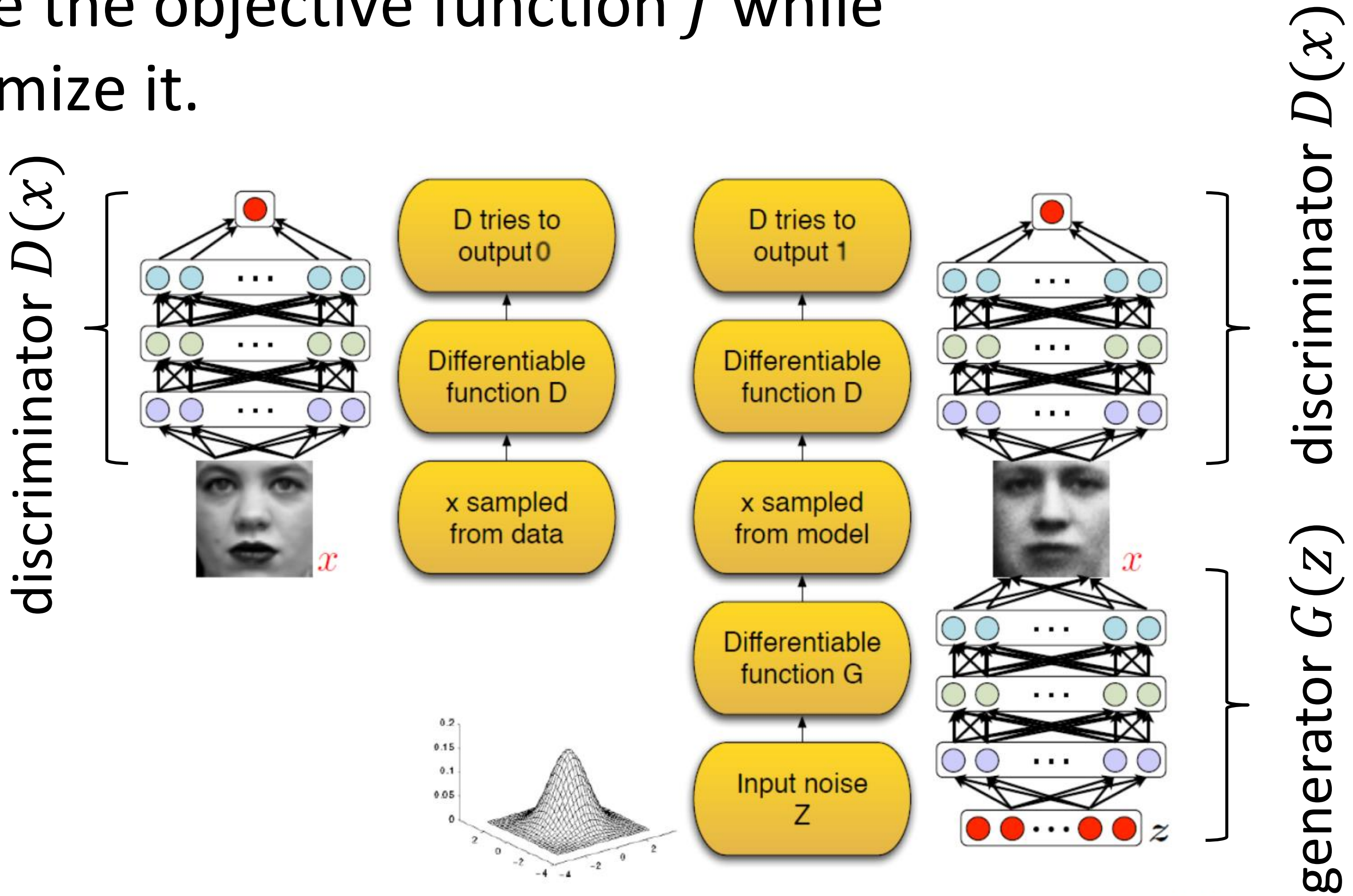


Figure from Goodfellow et al, 2014

Goal of GAN

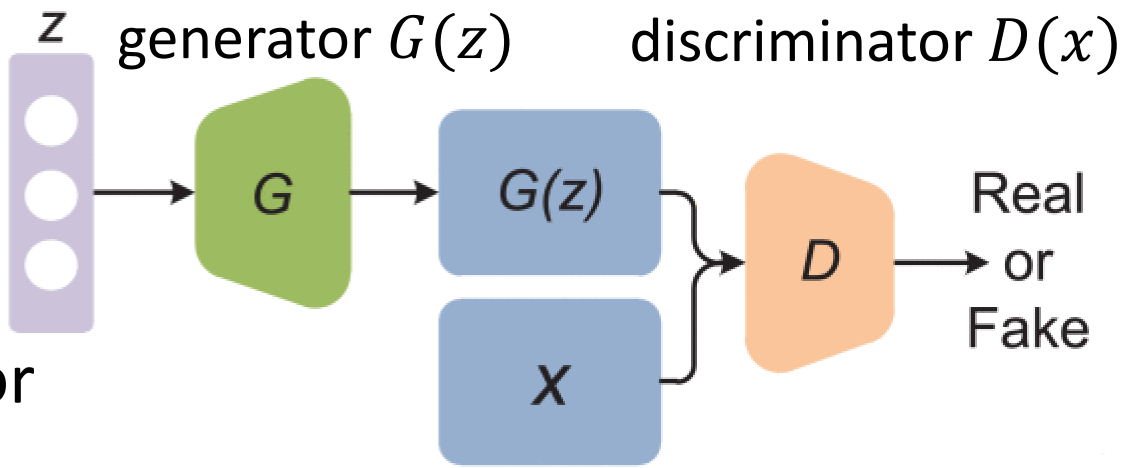


figure from [Pan et al., 2019]

- $D(x)$ discriminates real data from the generated ones
- Generator tries to fool the discriminator
  - $x = G(z), z \sim P_z$
- Objective function

$$\min_G \max_D \mathbb{E}_{x \sim p_r}[\ln(D(x))] + \mathbb{E}_{z \sim P_Z}\left[\ln\left(1 - D(G(z))\right)\right]$$

  - $D(x) = \begin{cases} 1 & x \text{ is real} \\ 0 & x \text{ is created by Generator } G(x) \end{cases}$
  - Note: $D'(x) \triangleq 1 - D(x)$ is sometimes used as the definition of discriminator

Discriminator's objective

- $D(x)$ **maximizes** J^D

$$J^D = \mathbb{E}_{x\sim p_r}[\ln(D(x))]$$
$$+\mathbb{E}_{z\sim p_Z}\left[\ln\left(1 - D(G(z))\right)\right]$$

  – Binary classification

- Optimal discriminator

$$D^*(x) = \frac{p_r(x)}{p_r(x) + p_g(x)}$$

  – p_g is the data distribution generated by G

  – $J^D = \mathbb{E}_{x\sim p_r}[\ln(D(x))]$
$$+\mathbb{E}_{x\sim p_g}[\ln(1 - D(x))]$$

generator $G(z)$ discriminator $D(x)$

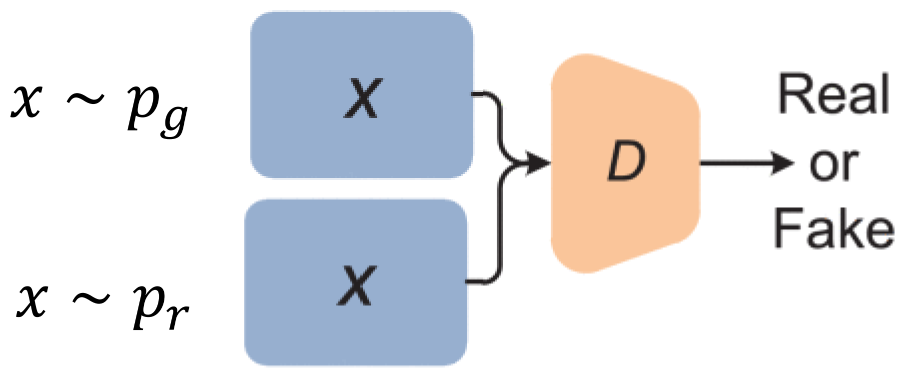
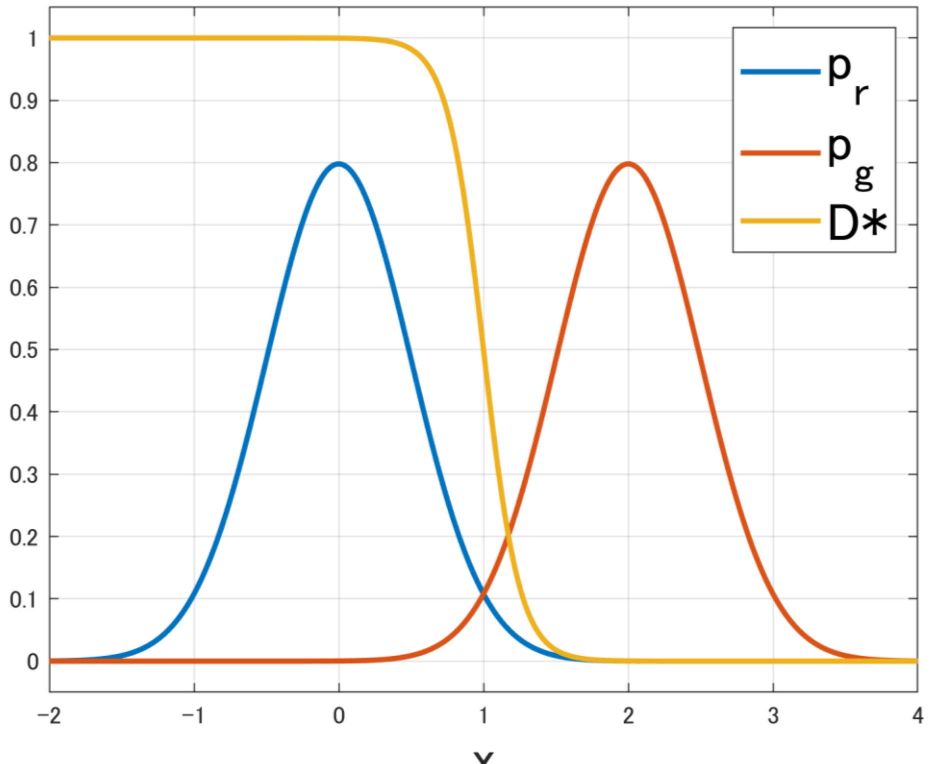


figure from [Pan et al., 2019]



Discriminator's objective

- When $D(x)$ is optimal $\quad J^D(D^*) = JS\left(p_r \parallel p_g\right) - 2\ln 2$
  - JS represents Jensen-Shannon divergence defined by

$$JS\left(p_r \parallel p_g\right) = KL\left(p_r \parallel \frac{p_r + p_g}{2}\right) + KL\left(p_g \parallel \frac{p_r + p_g}{2}\right)$$

  - Note: KL is Kullback-Leibler divergence

$$KL\left(p_r \parallel \frac{p_r + p_g}{2}\right) = \mathbb{E}_{p_r}\left[\ln \frac{p_r(x)}{\left(p_r(x) + p_g(x)\right)/2}\right]$$

- The role of the discriminator estimates JS divergence between p_r and p_g to measure a gap

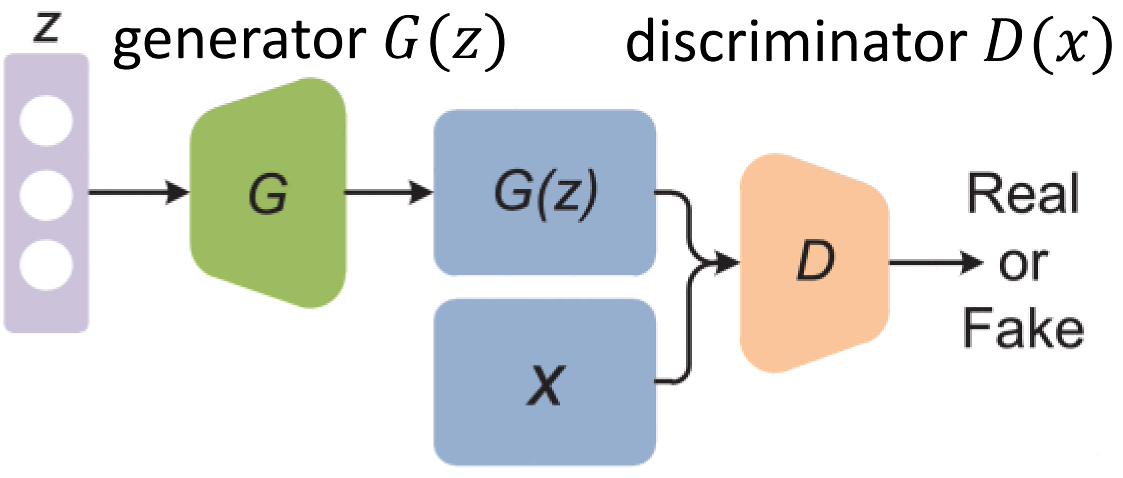
Generator's objective

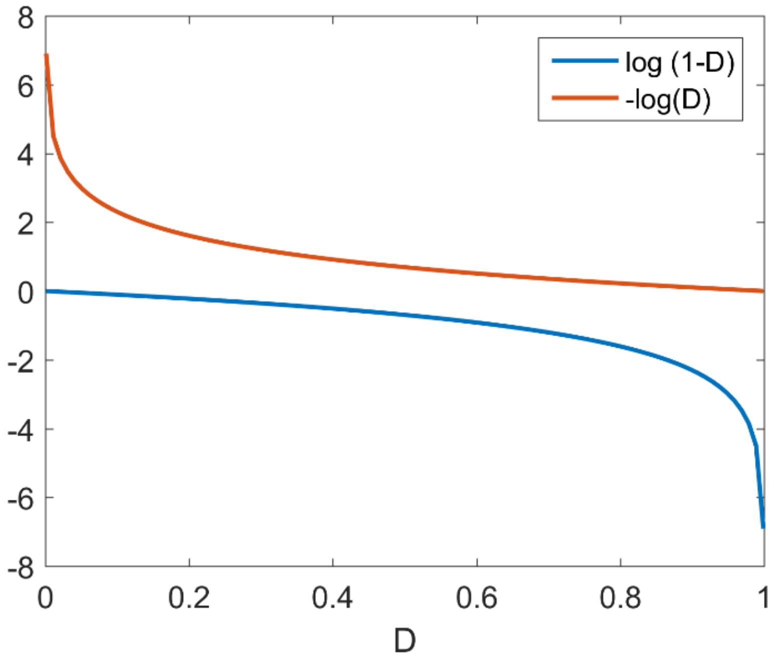
- $G(z)$ **minimizes** J^G

$$J^G = \cancel{\mathbb{E}_{x\sim p_r}[\ln(D(x))]} + \mathbb{E}_{z\sim P_Z}\left[\ln\left(1 - D\left(G(z)\right)\right)\right]$$

  - It is easy for Discriminator to distinguish at the early stage of learning because Generator is poor.
  - $\ln(1 - D(G(z)))$ is saturated and its gradient is close to 0

- Alternative $\quad \tilde{J}^G = \mathbb{E}_{z\sim P_z}\left[-\ln D\left(G(z)\right)\right]$
  - $\tilde{J}^G$ has the same fixed point of J^G





More advanced objective function

- Sum of the previous functions

$$\bar{J}^G = J^G + \tilde{J}^G = \mathbb{E}_{z\sim P_Z}\left[\ln\left(1 - D\big(G(z)\big)\right)\right] - \mathbb{E}_{z\sim P_Z}\left[\ln D\big(G(z)\big)\right]$$

$$= \mathbb{E}_{z\sim P_Z}\left[\ln\frac{1 - D\big(G(z)\big)}{D\big(G(z)\big)}\right]$$

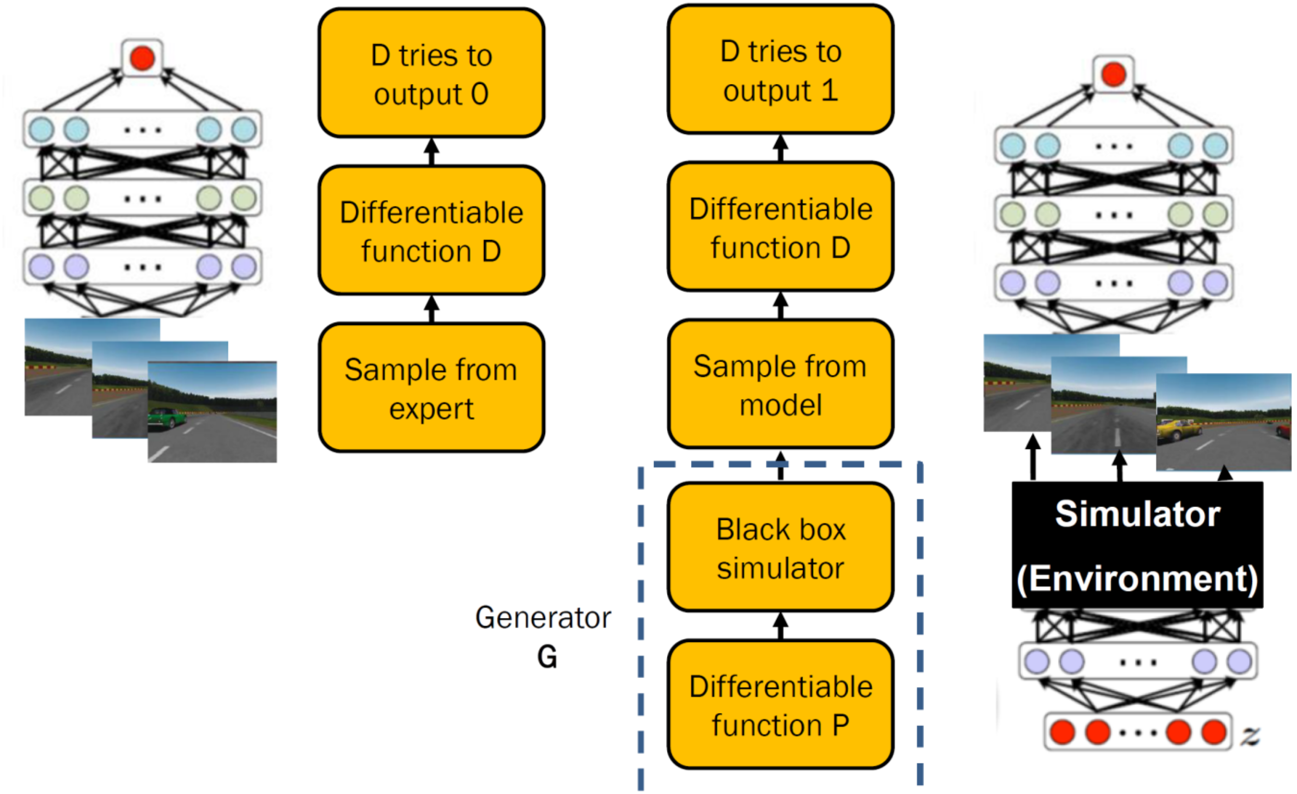
- Connection to Reverse KL divergence

$$\mathrm{KL}\big(p_g \parallel p_r\big) = \mathbb{E}_{x\sim p_g}\left[\ln\frac{p_g}{p_r}\right] = \mathbb{E}_{x\sim p_g}\left[\ln\frac{1 - D^*(x)}{D^*(x)}\right] \approx \mathbb{E}_{x\sim p_g}\left[\ln\frac{1 - D(x)}{D(x)}\right]$$

  - To train G, minimizing $\bar{J}^G$ is identical to minimizing reverse KL divergence

Generative Adversarial Imitation Learning (GAIL)

- Imitation learning formulated as GAN
- The most fundamental imitation learning framework
- Generator: stochastic policy and environmental dynamics
- Discriminator: pseudo reward from the difference between expert's behavior and learner's behavior



Ho and Ermon, *Generative Adversarial Imitation Learning*

Ho, J. & Ermon, S. (2016). Generative adversarial imitation learning. NIPS29.

Objective function of GAIL

- $D(s, a)$ discriminates generated state-action pair (s, a) from the real ones
- $\min_{\pi} \max_{D} \mathbb{E}_{(s,a)\sim p_r}\left[\ln\left(1 - D(s, a)\right)\right] + \mathbb{E}_{(s,a)\sim \pi}\left[\ln\left(D(s, a)\right)\right] - \lambda \mathcal{H}(\pi)$
  - $D(s, a) = \begin{cases} 1 & (s, a) \text{ is created by Generator} \\ 0 & (s, a) \text{ is real} \end{cases}$
- $\mathbb{E}_{(s,a)\sim p_r}[\cdot]$ is the expectation operator under p_r caused by expert's policy $\pi_E(a \mid s)$, which is defined by

$$p_r(s, a) = \pi_E(a \mid s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi_E)$$

  - $\mathbb{E}_{(s,a)\sim \pi}[\cdot]$ is defined in the same manner

Ho, J. & Ermon, S. (2016). Generative adversarial imitation learning. NIPS29.

Objective functions of discriminator and generator

- The discriminator's goal is binary classification task

$$\max_D J^D(D)\,, J^D(D) = \mathbb{E}_{(s,a)\sim p_r}\left[\ln\left(1 - D(s,a)\right)\right] + +\mathbb{E}_{(s,a)\sim \pi}\left[\ln\left(D(s,a)\right)\right]$$

- Objective function of the generator

$$\min_\pi J^\pi(\pi),\ J^\pi(\pi) = \mathbb{E}_{(s,a)\sim \pi}\left[\ln\left(D(s,a)\right)\right]$$

$$\Rightarrow \quad \max_\pi \mathbb{E}_{(s,a)\sim \pi}[r(s,a)] \qquad r(s,a) = -\ln\left(D(s,a)\right)$$

  - Generator's objective function is interpreted as that of forward reinforcement learning by defining a pseudo-reward
  - The original GAIL uses Trust Region Policy Optimization (Schulman, et al., 2015), which is on-policy

Another interpretation

- Objective function of apprenticeship learning (Abbeel & Ng, 2004)

$$\min_{\pi} d_{\psi}(\rho_{\pi}, p_r) - \lambda \mathcal{H}(\pi)$$

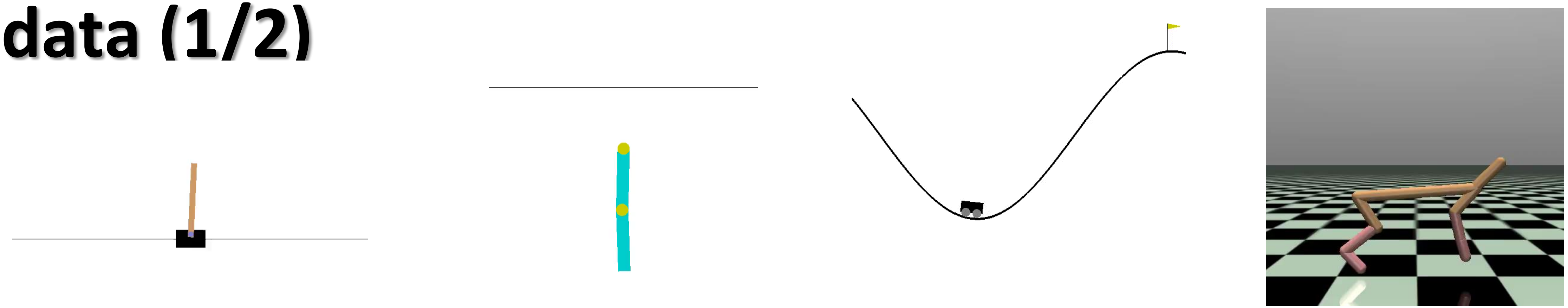
  - $d_{\psi}(\rho_{\pi}, p_r)$: distance between two joint state-action distributions
  - Minimize the gap between distributions while maximizing the entropy

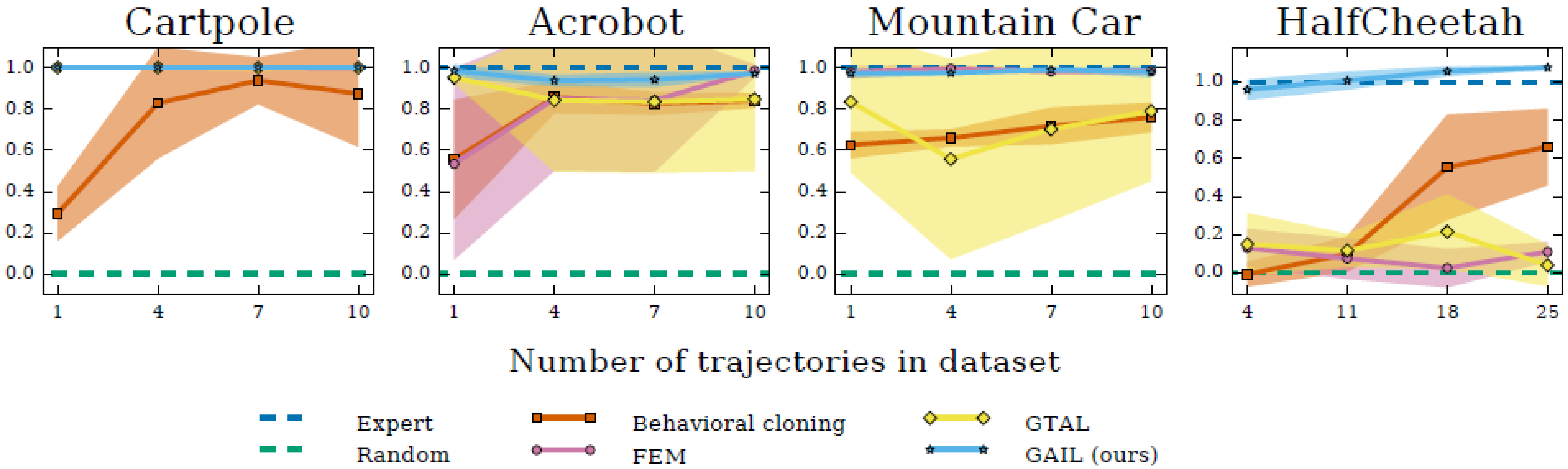
- The role of the GAN’s discriminator is to estimate JS divergence

$$d_{\psi}\left(\rho_{\pi}, \rho_{\pi_E}\right) = \max_{D} \mathbb{E}_{\pi}[\ln D(s,a)] + \mathbb{E}_{\pi_E}\left[\ln\left(1 - D(s,a)\right)\right]$$

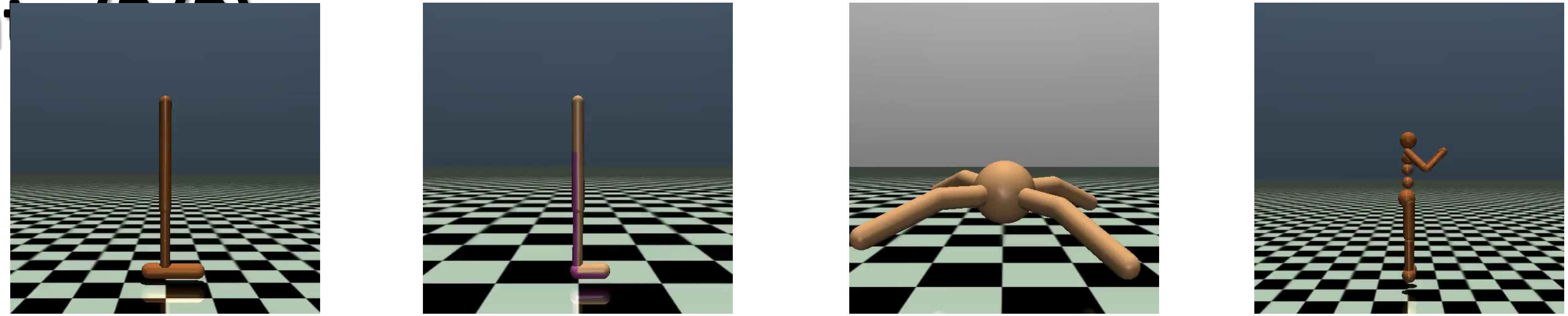
- According to choices of d_{ψ}, different algorithms such as AL (Abbeel & Ng, 2004) and MWAL (Syed & Schapire, 2008) can be derived

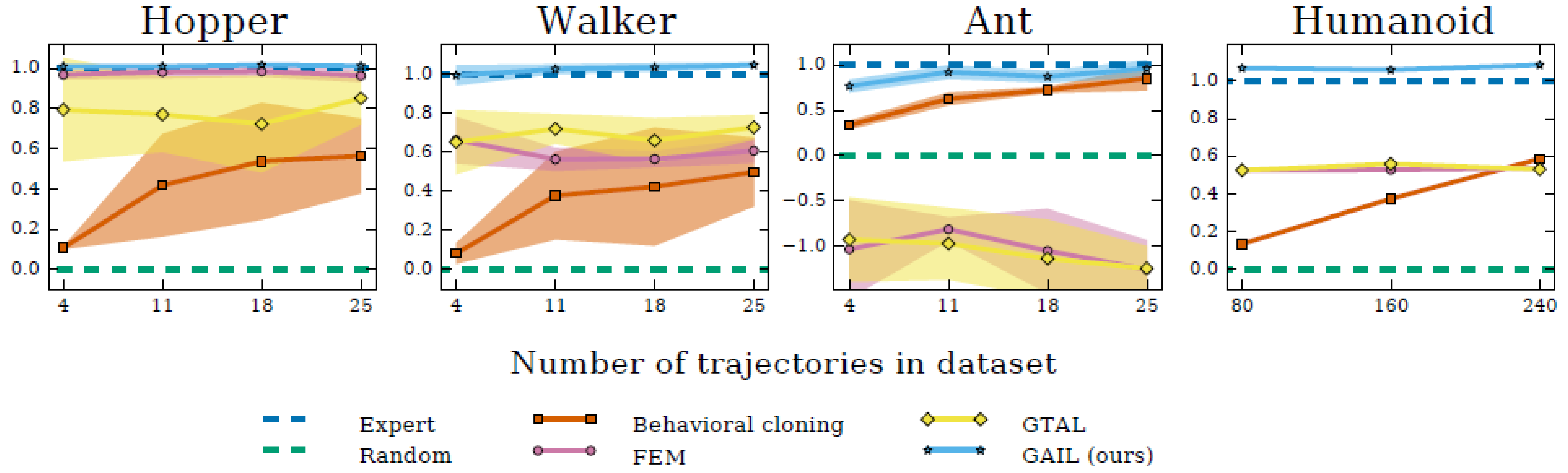
Sample efficiency w.r.t. the number of expert's data (1/2)





Sample efficiency w.r.t. the number of expert's da





References

- Abbeel, P. & Ng, A.Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proc. of ICML*.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. In *Proc. of ICML,* 214–23.
- Baram, N., Anschel, O., Caspi, I., & Mannor, S. (2017). End-to-End Differentiable Adversarial Imitation Learning. Proc. of the 34th International Conference on Machine Learning, 390–399.
- Finn, C., Christiano, P., Abbeel, P., and Levine, S. (2016). A Connection Between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models. *NIPS 2016 Workshop on Adversarial Training*.
- Fu, J., Luo, K., and Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. *Proc. of ICLR*.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *NIPS 27*, 2672–2680.
- Heess, N., Wayne, G., Silver, D., Lillicrap, T., Tassa, Y., & Erez, T. (2015). Learning Continuous Control Policies by Stochastic Value Gradients. *NIPS28*.

References

- Henderson, P., Chang, W.-D., Bacon, P.-L., Meger, D., Pineau, J., & Precup, D. (2018). OptionGAN: Learning Joint Reward-Policy Options using Generative Adversarial Inverse Reinforcement Learning. In *Proc. of AAAI*.
- Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. *NIPS29*.
- Huszár, F. (2016). An Alternative Update Rule for Generative Adversarial Networks.
- 小林, 堀井, 岩城, 長井, 浅田. (2019). 複数の報酬関数を推定可能なタスク条件付き敵対的模倣学習. 第33回人工知能学会全国大会予稿集.
- Li, Y., Song, J., & Ermon, S. (2017). InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations. *NIPS30*.
- Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F. & Zheng, Y. (2019). Recent Progress on Generative Adversarial Networks (GANs): A Survey. *IEEE Access* 7: 36322–33.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust Region Policy Optimization. *Proc. of ICML*, 1889–1897.
- Sun, M., & Ma, X. (2019). Adversarial Imitation Learning from Incomplete Demonstrations. In *Proc. of IJCAI*, 2019.

References

- Syed, U. & Schapire, R.E. (2008). A Game-Theoretic Approach to Apprenticeship Learning. *NIPS 20*.
- Torabi, F., Warnell, G., & Stone, P. (2018). Behavioral Cloning from Observation. In *Proc. of IJCAI-ECAI*, 4950–57.
- Torabi, F., Warnell, G., & Stone, P. (2019). Generative Adversarial Imitation from Observation. *ICML 2019 Workshop on Imitation, Intent, and Interaction*.

- Torabi, F., Warnell, G., & Stone, P. (2019). Generative Adversarial Imitation from Observation. ICML 2019 Workshop on Imitation, Intent, and Interaction.