Imitation learning

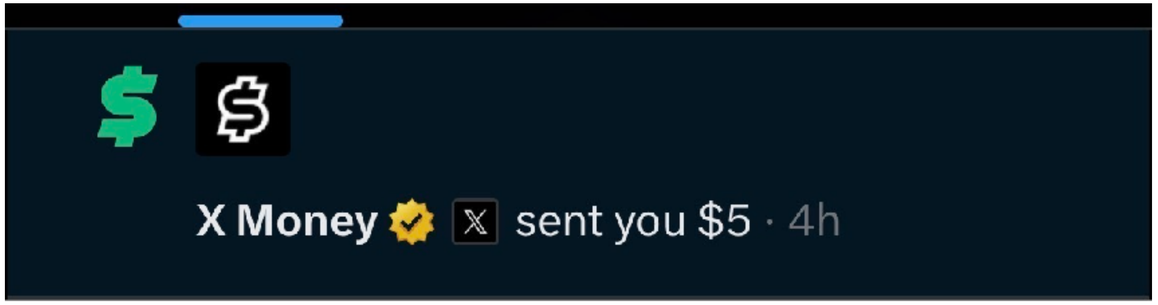
Preston Fu
2026/11/03






FrestonPu
FrestonPu




X Money
sent you $5 · 4h



Imitation learning

Let π_E be an (unknown) expert policy. We observe $\mathscr{D} = \{\tau^{(i)}\}_{1 \leq i \leq N}$, $\tau^{(i)} \sim p_{\pi_E}(\tau)$.

Pick some "distance" $\Delta(\pi, \pi_E)$, and our goal is to find some

$$\hat{\pi}_\Delta \in \arg\min_{\pi \in \Pi} \widehat{\Delta}(\pi, \pi_E)$$

Imitation learning

Let π_E be an (unknown) expert policy. We observe $\mathscr{D} = \{\tau^{(i)}\}_{1 \le i \le N}$, $\tau^{(i)} \sim p_{\pi_E}(\tau)$.

Pick some "distance" $\Delta(\pi, \pi_E)$, and our goal is to find some

$$\hat{\pi}_\Delta \in \arg\min_{\pi \in \Pi} \widehat{\Delta}(\pi, \pi_E)$$

Example: If the desired reward r^* is known: $\Delta_{r^*}(\pi, \pi_E) = \mathbb{E}_{\pi_E}[r^*] - \mathbb{E}_{\pi}[r^*]$

Imitation learning

Let π_E be an (unknown) expert policy. We observe $\mathscr{D} = \{\tau^{(i)}\}_{1 \le i \le N}$, $\tau^{(i)} \sim p_{\pi_E}(\tau)$.

Pick some "distance" $\Delta(\pi, \pi_E)$, and our goal is to find some

$$\hat{\pi}_\Delta \in \arg\min_{\pi \in \Pi} \widehat{\Delta}(\pi, \pi_E)$$

Example: If the desired reward r^* is known: $\Delta_{r^*}(\pi, \pi_E) = \mathbb{E}_{\pi_E}[r^*] - \mathbb{E}_{\pi}[r^*]$

Example: $\Delta_{\mathrm{BC}}(\pi, \pi_E) = \mathbb{E}_{s \sim d_E(s)}\left[D_{\mathrm{KL}}\big(\pi_E(\cdot \mid s) \,\|\, \pi(\cdot \mid s)\big)\right]$

BC in trajectory form

Assume finite horizon T, and that the learner and expert share the same p_0 and P:

$$
\begin{aligned}
D_{\mathrm{KL}}\left(p_{\pi_E} \| p_\pi\right) &= \mathbb{E}_{\tau \sim p_{\pi_E}}\left[\log \frac{p_{\pi_E}(\tau)}{p_\pi(\tau)}\right] \\
&= \mathbb{E}_{\tau \sim p_{\pi_E}}\left[\log \frac{p_0(s_0)\prod_{t=0}^{T-1} \pi_E(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t)}{p_0(s_0)\prod_{t=0}^{T-1} \pi(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t)}\right] \\
&= \mathbb{E}_{\tau \sim p_{\pi_E}}\left[\log \prod_{t=0}^{T-1} \frac{\pi_E(a_t \mid s_t)}{\pi(a_t \mid s_t)}\right] \\
&= \mathbb{E}_{\tau \sim p_{\pi_E}}\left[\sum_{t=0}^{T-1} \log \frac{\pi_E(a_t \mid s_t)}{\pi(a_t \mid s_t)}\right]
\end{aligned}
$$

BC in trajectory form

Assuming the learner and expert share the same p_0 and P:

$$D_{\mathrm{KL}}\left(p_{\pi_E} \| p_\pi\right) = \mathbb{E}_{\tau \sim p_{\pi_E}}\left[\sum_{t=0}^{T-1} \log \frac{\pi_E(a_t \mid s_t)}{\pi(a_t \mid s_t)}\right] = \text{const} + \mathbb{E}_{\tau \sim p_{\pi_E}}\left[\sum_{t=0}^{T-1} -\log \pi(a_t \mid s_t)\right]$$

$-H(\pi_E)$ $\qquad$ Δ_{BC}

So we may equivalently consider the "distance" $\Delta_\tau(\pi, \pi_E) = D_{\mathrm{KL}}\left(p_{\pi_E} \| p_\pi\right)$.

Learning rewards via maximum likelihood

Inverse RL (Ziebart08)**:** parametrize trajectories via rewards $R_\theta(\tau) = \sum_{t=0}^{T-1} r_{\theta,t}(s_t, a_t)$!

Assume access to reference $q_0(\tau)$.

$$q_\theta \in \arg\max_q \left\{ \mathbb{E}_{\tau\sim q}[R_\theta(\tau)] - \alpha D_{\mathrm{KL}}(q\|q_0) \right\}$$

Learning rewards via maximum likelihood

Inverse RL (Ziebart08)**:** parametrize trajectories via rewards $R_\theta(\tau) = \sum_{t=0}^{T-1} r_{\theta,t}(s_t, a_t)$!

Assume access to reference $q_0(\tau)$.

$$q_\theta \in \arg\max_q \left\{ \mathbb{E}_{\tau\sim q}[R_\theta(\tau)] - \alpha D_{\mathrm{KL}}(q\|q_0) \right\}$$

+ Reward specification
+ Reward transfer
+ Preference learning / alignment
+ Improve beyond BC

Learning rewards via maximum likelihood

Inverse RL (Ziebart08)**:** parametrize trajectories via rewards $R_\theta(\tau) = \sum_{t=0}^{T-1} r_{\theta,t}(s_t, a_t)$!

Assume access to reference $q_0(\tau)$.

$$q_\theta \in \arg\max_q \left\{ \mathbb{E}_{\tau\sim q}[R_\theta(\tau)] - \alpha D_{\mathrm{KL}}(q\|q_0) \right\}$$

$$= \frac{q_0(\tau)\exp(R_\theta(\tau)/\alpha)}{Z_\theta(s_0)}$$

$$Z_\theta(s_0) = \int q_0(\tau)\exp(R_\theta(\tau)/\alpha)\, d\tau$$

Learning rewards via maximum likelihood

Inverse RL (Ziebart08)**:** parametrize trajectories via rewards $R_\theta(\tau) = \sum_{t=0}^{T-1} r_{\theta,t}(s_t, a_t)$!

Assume access to reference $q_0(\tau)$.

$$q_\theta \in \arg\max_q \left\{ \mathbb{E}_{\tau\sim q}[R_\theta(\tau)] - \alpha D_{\mathrm{KL}}(q\|q_0) \right\}$$

$$= \frac{q_0(\tau)\exp(R_\theta(\tau)/\alpha)}{Z_\theta(s_0)}$$

$$Z_\theta(s_0) = \int q_0(\tau)\exp(R_\theta(\tau)/\alpha)\, d\tau$$

Aside: if q_0 is uniform, this reduces to $\mathbb{E}_{\tau\sim q}[R_\theta(\tau)] + \alpha H(q)$. If additionally dynamics are deterministic, then $H(q) = H(\pi)$, hence max entropy IRL.

Learning rewards via maximum likelihood

Inverse RL (Ziebart08)**:** parametrize trajectories via rewards $R_\theta(\tau) = \sum_{t=0}^{T-1} r_{\theta,t}(s_t, a_t)$!

Assume access to reference $q_0(\tau)$.

$$q_\theta \in \arg\max_q \left\{ \mathbb{E}_{\tau\sim q}[R_\theta(\tau)] - \alpha D_{\mathrm{KL}}(q\|q_0) \right\}$$

$$= \frac{q_0(\tau)\exp(R_\theta(\tau)/\alpha)}{Z_\theta(s_0)}$$

$$Z_\theta(s_0) = \int q_0(\tau)\exp(R_\theta(\tau)/\alpha)\, d\tau$$

$$\begin{aligned}
\theta^* &\in \arg\min_{\theta\in\Theta} \Delta_\tau(q_\theta, p_{\pi_E}) \\
&= \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}\left(p_{\pi_E}\|q_\theta\right) \\
&= \arg\min_{\theta\in\Theta} \text{const} - \mathbb{E}_{\tau\sim p_{\pi_E}}[\log q_\theta(\tau)] \\
&\approx \arg\max_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log q_\theta(\tau_i)
\end{aligned}$$

Learning rewards via maximum likelihood

How can we extract a policy?

Assume:

- fixed initial state s_0
- finite time horizon T
- deterministic dynamics $s_{t+1} = f(s_t, a_t)$

Learning rewards via maximum likelihood

How can we extract a policy?

Assume:

- fixed initial state s_0
- finite time horizon T
- deterministic dynamics $s_{t+1} = f(s_t, a_t)$

So $q_0(\tau) = p_{\pi_0}(\tau \mid s_0) = \prod_{t=0}^{T-1} \pi_0(a_t \mid s_t)$.

Learning rewards via maximum likelihood

How can we extract a policy?

Assume:

- fixed initial state s_0
- finite time horizon T
- deterministic dynamics $s_{t+1} = f(s_t, a_t)$

So $q_0(\tau) = p_{\pi_0}(\tau \mid s_0) = \prod_{t=0}^{T-1} \pi_0(a_t \mid s_t)$.

$$q_\theta(\tau) = \frac{q_0(\tau)\exp(R_\theta(\tau)/\alpha)}{Z_\theta(s_0)}$$

$$\begin{aligned} Z_\theta(s_0) &= \int q_0(\tau)\exp(R_\theta(\tau)/\alpha)\, d\tau \\ &= \sum_{a_{0:T-1}} \prod_{t=0}^{T-1} \pi_0(a_t \mid s_t)\exp\left(\frac{r_{\theta,t}(s_t, a_t))}{\alpha}\right) \end{aligned}$$

$$\begin{aligned} \pi_{\theta,t}(a \mid s) &= q_\theta(a_t = a \mid s_t = s) \\ &= \frac{q_\theta(s_t = s, a_t = a)}{q_\theta(s_t = s)} \end{aligned}$$

Learning rewards via maximum likelihood

Decompose into “forward” and “backward” components!

$$Z_{\theta,t}(s) = \sum_{a_{t:T-1}} \prod_{k=t}^{T-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right] \qquad B_{\theta,t}(s) = \sum_{a_{0:t-1}} \prod_{k=0}^{t-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right]$$

Learning rewards via maximum likelihood

Decompose into “forward” and “backward” components!

$$Z_{\theta,t}(s) = \sum_{a_{t:T-1}} \prod_{k=t}^{T-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right] \qquad B_{\theta,t}(s) = \sum_{a_{0:t-1}} \prod_{k=0}^{t-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right]$$

$$Z_{\theta,t}(s, a) = \sum_{a_{t+1:T-1}} \prod_{k=t}^{T-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right]$$

Learning rewards via maximum likelihood

Decompose into “forward” and “backward” components!

$$Z_{\theta,t}(s) = \sum_{a_{t:T-1}} \prod_{k=t}^{T-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right] \qquad B_{\theta,t}(s) = \sum_{a_{0:t-1}} \prod_{k=0}^{t-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right]$$

$$Z_{\theta,t}(s,a) = \sum_{a_{t+1:T-1}} \prod_{k=t}^{T-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right] = \pi_0(a \mid s) \exp\left(\frac{r_{\theta,t}(s,a)}{\alpha} \right) Z_{\theta,t+1}(f(s,a))$$

Dynamic programming!

Learning rewards via maximum likelihood

Decompose into “forward” and “backward” components!

$$Z_{\theta,t}(s) = \sum_{a_{t:T-1}} \prod_{k=t}^{T-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right] \qquad B_{\theta,t}(s) = \sum_{a_{0:t-1}} \prod_{k=0}^{t-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right]$$

$$Z_{\theta,t}(s,a) = \sum_{a_{t+1:T-1}} \prod_{k=t}^{T-1} \left[\pi_0(a_k \mid s_k) \exp\left(\frac{r_{\theta,k}(s_k, a_k)}{\alpha} \right) \right] = \pi_0(a \mid s) \exp\left(\frac{r_{\theta,t}(s,a)}{\alpha} \right) Z_{\theta,t+1}(f(s,a))$$

Dynamic programming!

$$q_\theta(s_t = s, a_t = a) = \frac{B_{\theta,t}(s) Z_{\theta,t}(s,a)}{Z_\theta(s_0)} \qquad q_\theta(s_t = s) = \frac{B_{\theta,t}(s) Z_{\theta,t}(s)}{Z_\theta(s_0)}$$

Learning rewards via maximum likelihood

$$Z_{\theta,t}(s,a) = \pi_0(a \mid s)\exp\left(\frac{r_{\theta,t}(s,a)}{\alpha}\right) Z_{\theta,t+1}(f(s,a))$$

$$q_\theta(s_t = s, a_t = a) = \frac{B_{\theta,t}(s)Z_{\theta,t}(s,a)}{Z_\theta(s_0)} \qquad q_\theta(s_t = s) = \frac{B_{\theta,t}(s)Z_{\theta,t}(s)}{Z_\theta(s_0)}$$

$$\pi_{\theta,t}(a \mid s) = \frac{q_\theta(s_t = s, a_t = a)}{q_\theta(s_t = s)} = \frac{Z_{\theta,t}(s,a)}{Z_{\theta,t}(s)} = \frac{\pi_0(a \mid s)\exp(r_{\theta,t}(s,a)/\alpha)Z_{\theta,t+1}(f(s,a))}{Z_{\theta,t}(s)}$$

So now we have a way to implicitly learn policies from just rewards!

Aside: soft Q-learning

$$Z_{\theta,t}(s,a) = \pi_0(a \mid s)\exp\left(\frac{r_{\theta,t}(s,a)}{\alpha}\right) Z_{\theta,t+1}(f(s,a))$$

$$\pi_{\theta,t}(a \mid s) = \frac{\pi_0(a \mid s)\exp(r_{\theta,t}(s,a)/\alpha) Z_{\theta,t+1}(f(s,a))}{Z_{\theta,t}(s)}$$

Define:

$$V_{\theta,t}(s) = \alpha \log Z_{\theta,t}(s)$$

$$Q_{\theta,t}(s,a) = \alpha \log \frac{Z_{\theta,t}(s,a)}{\pi_0(a \mid s)}$$

Aside: soft Q-learning

$$Z_{\theta,t}(s,a) = \pi_0(a \mid s)\exp\left(\frac{r_{\theta,t}(s,a)}{\alpha}\right) Z_{\theta,t+1}(f(s,a))$$

$$\pi_{\theta,t}(a \mid s) = \frac{\pi_0(a \mid s)\exp(r_{\theta,t}(s,a)/\alpha) Z_{\theta,t+1}(f(s,a))}{Z_{\theta,t}(s)}$$

Define:

$$V_{\theta,t}(s) = \alpha \log Z_{\theta,t}(s)$$

$$Q_{\theta,t}(s,a) = \alpha \log \frac{Z_{\theta,t}(s,a)}{\pi_0(a \mid s)}$$

Then:

$$Q_{\theta,t}(s,a) = r_{\theta,t}(s,a) + V_{\theta,t+1}(f(s,a))$$

$$V_{\theta,t}(s) = \alpha \log \left(\sum_a \pi_0(a \mid s)\exp\left(\frac{Q_{\theta,t}(s,a)}{\alpha}\right)\right)$$

$$\pi_{\theta,t}(a \mid s) = \pi_0(a \mid s)\exp\left(\frac{1}{\alpha}(Q_{\theta,t}(s,a) - V_{\theta,t}(s))\right)$$

Aside: soft Q-learning

$$Z_{\theta,t}(s,a) = \pi_0(a \mid s)\exp\left(\frac{r_{\theta,t}(s,a)}{\alpha}\right) Z_{\theta,t+1}(f(s,a))$$

$$\pi_{\theta,t}(a \mid s) = \frac{\pi_0(a \mid s)\exp(r_{\theta,t}(s,a)/\alpha) Z_{\theta,t+1}(f(s,a))}{Z_{\theta,t}(s)}$$

In the stochastic dynamics case, instead we get $\mathbb{E}_{s' \sim P(s'|s,a)}[V_{\theta,t+1}(s')]$

Define:

$$V_{\theta,t}(s) = \alpha \log Z_{\theta,t}(s)$$

$$Q_{\theta,t}(s,a) = \alpha \log \frac{Z_{\theta,t}(s,a)}{\pi_0(a \mid s)}$$

Then:

$$Q_{\theta,t}(s,a) = r_{\theta,t}(s,a) + V_{\theta,t+1}(f(s,a))$$

$$V_{\theta,t}(s) = \alpha \log \left(\sum_a \pi_0(a \mid s)\exp\left(\frac{Q_{\theta,t}(s,a)}{\alpha}\right)\right)$$

$$\pi_{\theta,t}(a \mid s) = \pi_0(a \mid s)\exp\left(\frac{1}{\alpha}(Q_{\theta,t}(s,a) - V_{\theta,t}(s))\right)$$

Reward learning is underspecified

Is there a reward that rationalizes the demonstrations?

$$\mathrm{IRL}_{\psi}(\pi_E) \in \arg\max_{r} \left\{ -\psi(r) + \mathbb{E}_{\pi_E}[r] - \max_{\pi \in \Pi} \left[H(\pi) + \mathbb{E}_{\pi}[r] \right] \right\} \qquad \mathrm{RL}(r) \in \arg\max_{\pi \in \Pi} \left\{ H(\pi) + \mathbb{E}_{\pi}[r] \right\}$$

Return $\mathrm{RL} \circ \mathrm{IRL}_{\psi}(\pi_E)$.

Reward learning is underspecified

Is there a reward that rationalizes the demonstrations?

$$\mathrm{IRL}_\psi(\pi_E) \in \arg\max_r \left\{ -\psi(r) + \mathbb{E}_{\pi_E}[r] - \max_{\pi \in \Pi} \left[H(\pi) + \mathbb{E}_\pi[r] \right] \right\} \qquad \mathrm{RL}(r) \in \arg\max_{\pi \in \Pi} \left\{ H(\pi) + \mathbb{E}_\pi[r] \right\}$$

Return $\mathrm{RL} \circ \mathrm{IRL}_\psi(\pi_E)$.

Problems:

- $\widehat{r} = 0$ rationalizes every policy.
- For any potential $\Phi : \mathcal{S} \to \mathbb{R}$, the potential-shaped reward $\widehat{r}(s, a, s') = r(s, a, s') + \gamma\Phi(s') - \Phi(s)$ also preserves optimal policies (Ng99).
- IRL contains RL in the inner loop!

Example: $R'(\tau) = R_\theta(\tau) + C$ gives $Z'(s_0) = e^{C/\alpha} Z_\theta(s_0)$ and $q'(\tau) = q_\theta(\tau)$.

Perspective 1: skip reward learning

Use *convexity* of objective vs $p_\pi(s,a) = \sum_{t \geq 0} \gamma^t \mathbb{P}_\pi(s_t = s, a_t = a)$ to skip reward learning! (Ho16)

- $\mathbb{E}_\pi[r] = \sum_{s,a} p_\pi(s,a) r(s,a) = \langle p_\pi, r \rangle.$

Perspective 1: skip reward learning

Use *convexity* of objective vs $p_\pi(s,a) = \sum_{t\geq 0} \gamma^t \mathbb{P}_\pi(s_t = s, a_t = a)$ to skip reward learning! (Ho16)

- $\mathbb{E}_\pi[r] = \sum_{s,a} p_\pi(s,a) r(s,a) = \langle p_\pi, r \rangle.$
- Each policy induces a point in the set of occupancy measures

$$\mathscr{P} = \left\{ p > 0 : \sum_a p(s,a) = p_0(s) + \gamma \sum_{\bar{s},\bar{a}} P(s \mid \bar{s},\bar{a}) p(\bar{s},\bar{a}) \ \forall s \in \mathcal{S} \right\}.$$

$\mathscr{P}$ is convex.

- Each such p corresponds to a policy, namely $\pi_p(a \mid s) = p(s,a) \Big/ \sum_{\bar{a}} p(s,\bar{a})$.

Perspective 1: skip reward learning

Use *convexity* of objective vs $p_\pi(s,a) = \sum_{t\geq 0} \gamma^t \mathbb{P}_\pi(s_t = s, a_t = a)$ to skip reward learning! (Ho16)

- $\mathbb{E}_\pi[r] = \sum_{s,a} p_\pi(s,a) r(s,a) = \langle p_\pi, r \rangle.$
- Each policy induces a point in the set of occupancy measures

$$\mathcal{P} = \left\{ p > 0 : \sum_a p(s,a) = p_0(s) + \gamma \sum_{\bar{s},\bar{a}} P(s \mid \bar{s},\bar{a}) p(\bar{s},\bar{a}) \ \forall s \in \mathcal{S} \right\}.$$

- Each such p corresponds to a policy, namely $\pi_p(a \mid s) = p(s,a) \Big/ \sum_{\bar{a}} p(s,\bar{a})$.
- For each π, $\mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t \big(-\log \pi(a_t \mid s_t) \big) \right] =: H(\pi) = \bar{H}(p_\pi) := - \sum_{s,a} p_\pi(s,a) \log \frac{p_\pi(s,a)}{\sum_{\bar{a}} p_\pi(s,\bar{a})}.$

Perspective 1: skip reward learning

Use *convexity* of objective vs $p_\pi(s,a) = \sum_{t\geq 0} \gamma^t \mathbb{P}_\pi(s_t = s, a_t = a)$ to skip reward learning! (Ho16)

- $\mathbb{E}_\pi[r] = \sum_{s,a} p_\pi(s,a) r(s,a) = \langle p_\pi, r \rangle.$
- Each policy induces a point in the set of occupancy measures

$$\mathscr{P} = \left\{ p > 0 : \sum_a p(s,a) = p_0(s) + \gamma \sum_{\bar{s},\bar{a}} P(s \mid \bar{s}, \bar{a}) p(\bar{s}, \bar{a}) \ \forall s \in \mathcal{S} \right\}.$$

- Each such p corresponds to a policy, namely $\pi_p(a \mid s) = p(s,a) \Big/ \sum_{\bar{a}} p(s,\bar{a})$.
- For each π, $\mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t \big(-\log \pi(a_t \mid s_t) \big) \right] =: H(\pi) = \bar{H}(p_\pi) := - \sum_{s,a} p_\pi(s,a) \log \frac{p_\pi(s,a)}{\sum_{\bar{a}} p_\pi(s,\bar{a})}.$

$\bar{H}$ is strictly concave on $\mathscr{P}$.

Duality

IRL dual, policy form: $\max_{r} \min_{\pi \in \Pi} -H(\pi) + \langle p_E - p_\pi, r \rangle - \psi(r)$

Duality

IRL dual, policy form: $\max_{r} \min_{\pi \in \Pi} -H(\pi) + \langle p_E - p_\pi, r \rangle - \psi(r)$

IRL dual, occupancy form: $\max_{r} \min_{p \in \mathscr{P}} -\bar{H}(p) + \langle p_E - p, r \rangle - \psi(r)$

Duality

IRL dual, policy form: $\max_r \min_{\pi \in \Pi} -H(\pi) + \langle p_E - p_\pi, r \rangle - \psi(r)$

IRL dual, occupancy form: $\max_r \min_{p \in \mathscr{P}} -\bar{H}(p) + \langle p_E - p, r \rangle - \psi(r)$

Occupancy-matching primal: $\min_{p \in \mathscr{P}} -\bar{H}(p) + \psi^*(p_E - p)$

Convex conjugate:
$f^*(y) = \sup_x (\langle y, x \rangle - f(x))$

Duality

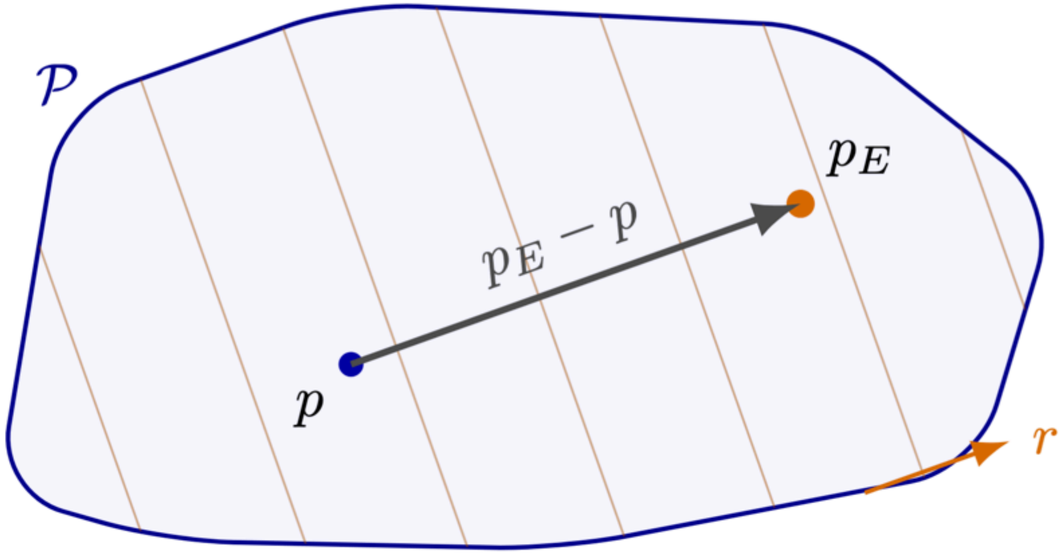
IRL dual, occupancy form: $\max_{r} \min_{p \in \mathscr{P}} -\bar{H}(p) + \langle p_E - p, r \rangle - \psi(r)$

Occupancy-matching primal: $\min_{p \in \mathscr{P}} -\bar{H}(p) + \psi^*(p_E - p)$

Example: If $\psi(r) \equiv 0$, then

$$\psi^*(p_E - p) = \begin{cases} 0 & p_E - p = 0 \\ +\infty & \text{else} \end{cases}$$

becomes a hard constraint!



Duality

IRL dual, occupancy form: $\max_{r} \min_{p \in \mathscr{P}} -\bar{H}(p) + \langle p_E - p, r \rangle - \psi(r)$

Occupancy-matching primal: $\min_{p \in \mathscr{P}} -\bar{H}(p) + \psi^*(p_E - p)$

Example (GAIL): If $\psi_{\mathrm{GAIL}}(r) = \begin{cases} \mathbb{E}_{d_E}\left[r - \log(1 - e^{-r})\right] & r > 0 \\ +\infty & \text{else} \end{cases}$

then $\psi^*_{\mathrm{GAIL}}(p_E - p) = \sum_{s,a} p(s,a) \log \frac{p(s,a)}{p(s,a) + p_E(s,a)} + p_E(s,a) \log \frac{p_E(s,a)}{p(s,a) + p_E(s,a)}$

is the standard discriminator objective.

Reward learning is underspecified

Is there a reward that rationalizes the demonstrations?

$$\mathrm{IRL}_\psi(\pi_E) \in \arg\max_r \left\{ -\psi(r) + \mathbb{E}_{\pi_E}[r] - \max_{\pi \in \Pi} \left[H(\pi) + \mathbb{E}_\pi[r] \right] \right\} \qquad \mathrm{RL}(r) \in \arg\max_{\pi \in \Pi} \left\{ H(\pi) + \mathbb{E}_\pi[r] \right\}$$

Return $\mathrm{RL} \circ \mathrm{IRL}_\psi(\pi_E)$.

Problems:

- $\widehat{r} = 0$ rationalizes every policy.
- For any potential $\Phi : \mathcal{S} \to \mathbb{R}$, the potential-shaped reward $\widehat{r}(s, a, s') = r(s, a, s') + \gamma\Phi(s') - \Phi(s)$ also preserves optimal policies (Ng99).
- IRL contains RL in the inner loop!

Example: $R'(\tau) = R_\theta(\tau) + C$ gives $Z'(s_0) = e^{C/\alpha} Z_\theta(s_0)$ and $q'(\tau) = q_\theta(\tau)$.

Perspective 2: optimize reward as a fixed point

Recall the occupancy-matching primal problem: $\min_{p \in \mathcal{P}} -\bar{H}(p) + \psi^*(p_E - p)$

We can equivalently optimize over $d_\pi(s,a) = (1-\gamma)p_\pi(s,a)$ (which is just the normalized version).

For the choice of ψ^* as KL divergence (Diwan26):

$$\max_{\pi \in \Pi,\, d \in \Delta(\mathcal{S}\times\mathcal{A})} \left\{ H(d_\pi) - \beta D_{\mathrm{KL}}(d \| d_E) \right\} \quad \text{s.t.} \quad d = d_\pi$$

Here, rewards are now Lagrange multipliers:

$$\mathcal{L}(\pi, d, r) = H(d_\pi) - \beta D_{\mathrm{KL}}(d \| d_E) + \langle d_\pi - d, r \rangle$$

Duality

$$\min_{r} \left(\max_{\pi \in \Pi} \left\{ H_d(\pi) + \langle d_\pi, r \rangle \right\} + \max_{d \in \Delta(\mathcal{S} \times \mathcal{A})} \left\{ -\beta D_{\mathrm{KL}}(d \| d_E) - \langle d, r \rangle \right\} \right)$$

Duality

$$\min_r \left(\max_{\pi \in \Pi} \left\{ H_d(\pi) + \langle d_\pi, r \rangle \right\} + \max_{d \in \Delta(\mathcal{S} \times \mathcal{A})} \left\{ -\beta D_{\mathrm{KL}}(d \| d_E) - \langle d, r \rangle \right\} \right)$$

$$= \max_{\pi \in \Pi} \mathbb{E}_{(s,a) \sim d_\pi} [r(s,a) - \log \pi(a \mid s)]$$
$$= (1-\gamma) \max_{\pi \in \Pi} (\mathbb{E}_\pi[r] + H(\pi))$$

The optimum satisfies the Bellman equations:

$$Q_r(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_r(s')]$$
$$V_r(s) = \log \sum_a \exp Q_r(s,a)$$
$$\pi_r(a \mid s) = \exp\big(Q_r(s,a) - V_r(s)\big)$$

Duality

$$\min_r \left(\max_{\pi \in \Pi} \left\{ H_d(\pi) + \langle d_\pi, r \rangle \right\} + \max_{d \in \Delta(\mathcal{S} \times \mathcal{A})} \left\{ -\beta D_{\mathrm{KL}}(d \| d_E) - \langle d, r \rangle \right\} \right)$$

$$= \max_{\pi \in \Pi} \mathbb{E}_{(s,a) \sim d_\pi}[r(s,a) - \log \pi(a \mid s)]$$
$$= (1-\gamma) \max_{\pi \in \Pi} (\mathbb{E}_\pi[r] + H(\pi))$$

The optimum satisfies the Bellman equations:

$$Q_r(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_r(s')]$$
$$V_r(s) = \log \sum_a \exp Q_r(s,a)$$
$$\pi_r(a \mid s) = \exp\big(Q_r(s,a) - V_r(s)\big)$$

Looks familiar?

$$\widetilde{d}_r(s,a) = \frac{d_E(s,a) \exp(-r(s,a)/\beta)}{\mathbb{E}_{(\bar{s},\bar{a}) \sim d_E}[\exp(-r(\bar{s},\bar{a})/\beta)]}$$

Duality

$$\min_r \left(\max_{\pi \in \Pi} \left\{ H_d(\pi) + \langle d_\pi, r \rangle \right\} + \max_{d \in \Delta(\mathcal{S} \times \mathcal{A})} \left\{ -\beta D_{\mathrm{KL}}(d \| d_E) - \langle d, r \rangle \right\} \right)$$

$$= \max_{\pi \in \Pi} \mathbb{E}_{(s,a) \sim d_\pi} [r(s,a) - \log \pi(a \mid s)]$$
$$= (1-\gamma) \max_{\pi \in \Pi} (\mathbb{E}_\pi[r] + H(\pi))$$

The optimum satisfies the Bellman equations:

$$Q_r(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_r(s')]$$
$$V_r(s) = \log \sum_a \exp Q_r(s,a)$$
$$\pi_r(a \mid s) = \exp\big(Q_r(s,a) - V_r(s)\big)$$

Looks familiar?

$$\widetilde{d}_r(s,a) = \frac{d_E(s,a) \exp(-r(s,a)/\beta)}{\mathbb{E}_{(\bar{s},\bar{a}) \sim d_E}[\exp(-r(\bar{s},\bar{a})/\beta)]}$$

Differentiate the objective vs r:

$$d_{\pi_r}(s,a) - \widetilde{d}_r(s,a) = 0$$

Perspective 2: optimize reward as a fixed point

Solve: $r^*(s,a) = \beta \log \frac{d_E(s,a)}{d_{\pi_{r^*}}(s,a)} \underbrace{-\beta \log Z_{r^*}}_{\text{constant in } (s,a)}$

Perspective 2: optimize reward as a fixed point

Solve: $r^*(s,a) = \beta \log \frac{d_E(s,a)}{d_{\pi_{r^*}}(s,a)} \underbrace{-\beta \log Z_{r^*}}_{\text{constant in } (s,a)}$

We can ignore this due to reward shaping!

Perspective 2: optimize reward as a fixed point

Solve: $r^*(s,a) = \beta \log \frac{d_E(s,a)}{d_{\pi_{r^*}}(s,a)} \underbrace{-\beta \log Z_{r^*}}_{\text{constant in } (s,a)}$

Suppose iteration i has current reward r_i and occupancy $d_i = d_{\pi_{r_i}}$. The fixed point condition gives the target $\hat{r}_i(s,a) = \beta \log \frac{d_E(s,a)}{d_i(s,a)}$

Perspective 2: optimize reward as a fixed point

Solve: $r^*(s,a) = \beta \log \frac{d_E(s,a)}{d_{\pi_{r^*}}(s,a)} \underbrace{-\beta \log Z_{r^*}}_{\text{constant in } (s,a)}$

Suppose iteration i has current reward r_i and occupancy $d_i = d_{\pi_{r_i}}$. The fixed point condition gives the target $\hat{r}_i(s,a) = \beta \log \frac{d_E(s,a)}{d_i(s,a)}$

But we can optimize for this with gradient descent! $r_{i+1} \leftarrow r_i - \eta(r_i - \hat{r}_i)$

Perspective 2: optimize reward as a fixed point

Solve: $r^*(s,a) = \beta \log \frac{d_E(s,a)}{d_{\pi_{r^*}}(s,a)} \underbrace{-\beta \log Z_{r^*}}_{\text{constant in } (s,a)}$

Suppose iteration i has current reward r_i and occupancy $d_i = d_{\pi_{r_i}}$. The fixed point condition gives the target $\hat{r}_i(s,a) = \beta \log \frac{d_E(s,a)}{d_i(s,a)}$

But we can optimize for this with gradient descent! $r_{i+1} \leftarrow r_i - \eta(r_i - \hat{r}_i)$

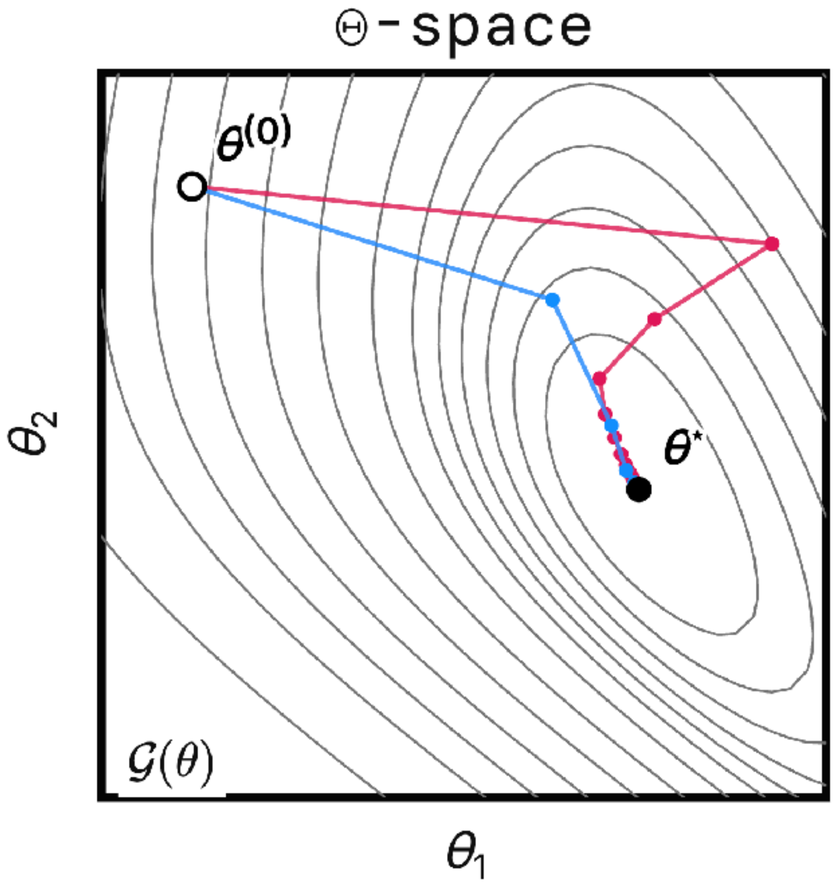


Figure from Diwan26

Conclusions

Perspective 1 (GAIL)

+ Cool framework for imitation learning with regularization.
- At convergence, rewards are 0.5, and so they cannot be transferred.
- Hard to scale adversarial training.
- Needs on-policy data.

Perspective 2 (fixed point)

+ Cool geometric view.
- Rewards are Lagrange multipliers and cannot be interpreted/transferred.
- Needs online or a lot of offline data, else reverse KL d_E/d_π estimation is hard.