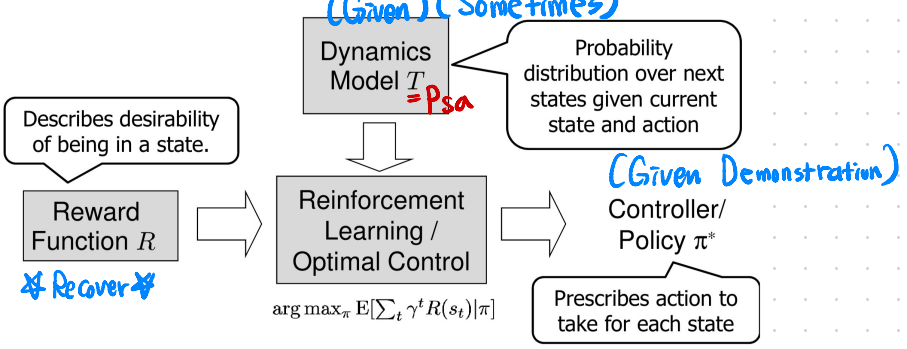
2024.09.11 (Reference : Slide from Abbeel, Levine, Sadigh, & Salakhutdinov)

- BC vs IRL → Learn r s.t. $\pi^* = \arg\max_\theta \mathbb{E}_{s\sim p(s|\theta)} r(s, \pi_\theta(s))$
  - ↳ $\arg\min_\theta \mathbb{E}_{(s,a^*)\sim p^*} L(a^*, \pi_\theta(s))$

<pros>
- No reasoning about outcomes or dynamics
- No notion of intentions
- Expert can be suboptimal
- ⋮

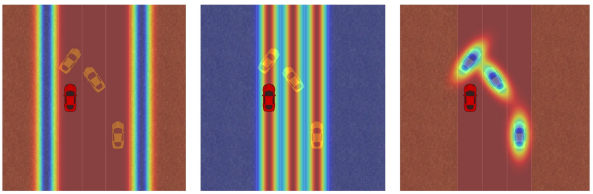
- Problem Setting : MDP/r (same as BC)
  - → State Space : S
  - Action Space : A
  - An expert policy : $\pi^* : S \mapsto A$
  - Transition model : $P_{sa}(S_{t+1} | S_t, a_t)$ (Given) (Sometimes)

} Goal : Learn a reward function $r(S,a)$ or $r(S)$ assuming π^* == optimal



- Big Assumption (1) : Reward func. is a linear combination of features.

→ $R(s) = w^T \varphi(s)$
- ↓ weights $w \in \mathbb{R}^n$
- ↳ features of state ⇒ $\varphi : S \mapsto \mathbb{R}^n$



(a) Features for the boundaries of the road (b) Feature for staying inside the lanes. (c) Features for avoiding other vehicles.

- Big Assumption (2) : Expert is optimal, R^* explains π^*.

→ Find R^* s.t. $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R^*(s_t)|\pi^*] \geq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R^*(s_t)|\pi] \quad \forall \pi$.

- **Combine two assumptions.**

⇒ $R(s) = w^T \varphi(s)$

$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^*(s_t) \mid \pi^*\right] \geq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^*(s_t) \mid \pi\right] \quad \forall \pi$

$V^{\pi}(s)$ in RL!

$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t w^T \varphi(s_t) \mid \pi\right]$

$= w^T \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \varphi(s_t) \mid \pi\right]$

$= w^T \mu(\pi)$ — Depend on state visitation distribution.

feature expectations.

- how to compute for π^*? "Requires Scale"

⇒ Find w^* s.t. $w^{*T} \mu(\pi^*) \geq w^{*T} \mu(\pi), \forall \pi$. (If π^* is suboptimal, Infeasible!)

($V^{\pi^*}(s) \geq V^{\pi}(s) \quad \forall \pi$)

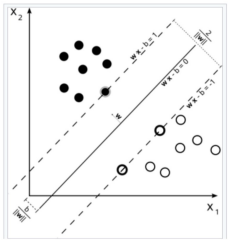
- $R=0$ can be solution <ambiguity>

Reward Ambiguity: Many R exist under π^*'s demonstrations.

- How to pick R?
  - (1) Maximum Margin Planning
  - (2) Maximum Entropy IRL

(MMP)

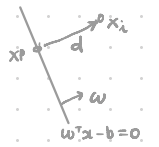
- **Maximum Margin Planning**: Find R that best separates π^* from $\forall \pi$.



"Minimize $\|\vec{w}\|$ subject to $y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1$, for $i = 1, \ldots, n$"

Given a training dataset of $(x_1, y_1), \ldots, (x_n, y_n)$, where y_i is either 1 or -1 identifying the class x_i is in. We want to find the maximum margin hyperplane that divides the points so the distance between the hyperplane and the nearest point from each class is maximized.

If students less know linear algebra...



$x_p = x_i - d$

$d = \alpha w$

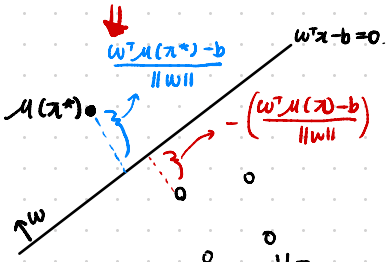
$w^T x_p - b = 0$

$\rightarrow w^T(x_i - d) - b = 0$

$= w^T(x_i - \alpha w) - b = 0$

$\alpha = \frac{w^T x_i - b}{w^T w}$

$\rightarrow \|d\|_2 = \sqrt{d^T d} = \sqrt{\alpha^2 w^T w} = \frac{(w^T x_i - b)}{\|w\|_2}$



Our goal

$\rightarrow \max_w \min_{\pi} \frac{\left(\frac{w^T \mu(\pi^*) - b}{\|w\|}\right) + \left(-\frac{w^T \mu(\pi) - b}{\|w\|}\right)}{2}$

$= \max_w \min_{\pi} \frac{w^T \mu(\pi^*) - w^T \mu(\pi)}{\|w\|}$

$= \max_w \frac{1}{\|w\|} \left(\min_{\pi} w^T \mu(\pi^*) - w^T \mu(\pi)\right)$

↳ Scale is invariant, Let's be clever, choose margin

$w^T \mu(\pi^*) \geq w^T \mu(\pi)$

$\Rightarrow w^T \mu(\pi^*) - b \geq 0$ & $w^T \mu(\pi) - b \leq 0$

$\Rightarrow \min_w \|w\|$

$s.t. \quad w^T \mu(\pi^*) - w^T \mu(\pi) \geq 1$

- MMP recap

→ $\min_{w} \|w\|_2^2$

s.t. $w^T \mu(\pi^*) \geq w^T \mu(\pi) + 1 \quad \forall \pi.$ → can be $m(\pi^*, \pi)$ (adaptive margin)

- Max Entropy IRL.

→ $\xi = \{(s_1, a_1), \dots, (s_T, a_T)\}$: trajectory , $\xi \in \Xi$

→ $D = \{\xi_1, \dots \xi_{|D|}\}$: set of π^*'s demonstration.

→ $f : \Xi \mapsto \mathbb{R}^n$: traj. feature. , $f_D = \frac{1}{|D|} \sum_{\xi \in D} f(\xi)$ is empirical feature expectation.

Then, $\mathbb{E}_{\xi \sim P(\xi)} \left[\sum_{t=1}^{T} \gamma^t R(s_t) \right] = \mathbb{E}_{\xi \sim P(\xi)} \left[\sum_{t=1}^{T} \gamma^t w^T \phi(s_t) \right] = w^T \mathbb{E}_{\xi \sim P(\xi)} [f(\xi)]$ ("weighted" ~ feature expectation)

└ Let's make noisy assumption (yet rational)

$P(\xi) \propto \exp(R(\xi))$

Goal : $\max_P \int -P(\xi) \log P(\xi) \, d\xi$

s.t. $\mathbb{E}_{\xi \sim P(\xi)} [f(\xi)] = \int P(\xi) f(\xi) \, d\xi = f_D$

$\int P(\xi) \, d\xi = 1$

$P(\xi) \geq 0, \forall \xi \in \Xi$

Find P from D that matches f_D, & maximize Entropy.

Strong preference for low-cost paths, equal cost paths be equally probable.

⇒ Lagrangian : $L(P, \lambda, v) \Rightarrow \int -P(\xi) \log P(\xi) \, d\xi + \lambda^T \left(\int P(\xi) f(\xi) \, d\xi - f_D \right) + v \left(\int P(\xi) \, d\xi - 1 \right)$

$\Rightarrow \int \underbrace{\left(-P(\xi) \log P(\xi) + \lambda^T P(\xi) f(\xi) + v P(\xi) \right)}_{F(P(\xi))} d\xi - v - \lambda^T f_D$

$\Rightarrow \int F(P(\xi)) \, d\xi - v - \lambda^T f_D$, $\nabla_P L(P, \lambda, v) = 0$

set gradient of each term in the integral.

→ $\frac{\partial F(P)}{\partial P} = -\log P(\xi) - 1 + \lambda^T f(\xi) + v = 0$

$\Rightarrow P^*(\xi) = \exp(\lambda^T f(\xi) + v - 1)$

→ $L(P^*, \lambda, v) = \int \left(-\exp(\lambda^T f(\xi) + v - 1) \cdot (\lambda^T f(\xi) + v - 1) + \lambda^T \cdot \exp(\lambda^T f(\xi) + v - 1) \cdot f(\xi) + v \cdot \exp(\lambda^T f(\xi) + v - 1) \right) d\xi - v - \lambda^T f_D$

$\Rightarrow \int \exp(\lambda^T f(\xi) + v - 1) \, d\xi - \lambda^T f_D - v$

→ v^* ? : $\frac{\partial L(P^*, \lambda, v)}{\partial v} = 0 \Rightarrow e^{v-1} \int \exp(\lambda^T f(\xi)) \, d\xi - 1 = 0$

$\Rightarrow e^{-v} = \int \exp(\lambda^T f(\xi) - 1) \, d\xi$

→ $v^* = -\log \left(\int \exp(\lambda^T f(\xi) - 1) \, d\xi \right)$

Recap

$\rightarrow p^*(\xi) = \exp(\lambda^T f(\xi) + v - 1)$

$v^* = -\log\left(\int \exp(\lambda^T f(\xi) - 1)\, d\xi\right)$

$\Rightarrow$ Put v^* to $p^* \Rightarrow p^* = \exp\left(\lambda^T f(\xi) - \log\left(\int \exp(\lambda^T f(\xi) - 1) d\xi\right) - 1\right)$

$$\Rightarrow p^* = \frac{\exp(\lambda^T f(\xi) - 1)}{\int \exp(\lambda^T f(\xi) - 1) d\xi} = \frac{\exp(\lambda^T f(\xi))}{\int \exp(\lambda^T f(\xi)) d\xi}.$$

match to $P(\xi) \propto \exp(R(\xi))$

if we set $R(\xi) = \lambda^T f(\xi)$

Reward weights are dual variable λ.

Finally, $\lambda^* = \arg\max_\lambda P(\xi_{1:D}; \lambda)$

$$= \arg\max_\lambda \underbrace{\lambda^T f_D - \log\left(\int \exp(\lambda^T f(\xi))\, d\xi\right)}_{M}$$

Then, we do gradient ascent $\nabla_\lambda M = f_D - \mathbb{E}_{\xi \sim P(\xi;\lambda)}[f(\xi)]$

$\lambda_{i+1} \leftarrow \lambda_i + \alpha \nabla_\lambda M$.

1) Initialize λ and collect expert demonstrations D.
2) Solve for the optimal policy $\pi_\lambda(a|s)$ with respect to λ. (RL Loop)
3) Solve for state visitation frequencies $P_\lambda(s)$.
4) Compute the gradient $\nabla_\lambda M$.
5) Update λ with one gradient step.

This assumes access to the dynamics (transition function) and having low dimensional systems to be able to solve for the policy using RL.

$\nabla_\lambda M$? $\Rightarrow \frac{\partial\left(\log \int \exp(\lambda^T f(\xi)) d\xi\right)}{\partial \lambda}$?

$\log g(\lambda)$, $g(\lambda) = \int \exp(\lambda^T f(\xi)) d\xi$

$$\frac{\partial \log g(\lambda)}{\partial \lambda} = \frac{1}{g(\lambda)} \cdot \frac{\partial g(\lambda)}{\partial \lambda}, \quad \frac{\partial g(\lambda)}{\partial \lambda} = \int \frac{d}{d\lambda} \exp(\lambda^T f(\xi)) d\xi$$

$$= \int f(\xi) \exp(\lambda^T f(\xi)) d\xi$$

$$\Rightarrow \frac{\int f(\xi) \exp(\lambda^T f(\xi)) d\xi}{\int \exp(\lambda^T f(\xi)) d\xi} = \mathbb{E}[f(\xi)]$$

$P(\xi)$

Lagrangian? : Combine obj. & const into one.

$\rightarrow$ $\min f(x)$ s.t. $g(x) = 0$ $\xrightarrow{\text{Lag. multiplier } \lambda}$ $L(x, \lambda) = f(x) + \lambda \cdot g(x)$

Find partial derivative w.r.t $x \& \lambda$ set them 0, find which $x \& \lambda$ make it "optimal"

ex) $\min f(x, y) = xy$ s.t. $x + y = 1$ $\rightarrow L(x, y, \lambda) = xy - \lambda(x + y - 1)$

$$\frac{\partial L}{\partial x} = y - \lambda = 0$$
$$\frac{\partial L}{\partial y} = x - \lambda = 0$$
$$\frac{\partial L}{\partial \lambda} = -(x + y - 1) = 0$$

$\Rightarrow \lambda = \frac{1}{2}$, $x = \lambda$, $y = \lambda$.