Robot Learning

Maximum Margin Planning
Maximum Entropy Inverse Reinforcement Learning

Prof. Songhwai Oh
ECE, SNU

Bagnell, J. Andrew, Nathan Ratliff, and Martin Zinkevich. "**Maximum margin planning**." In Proceedings of the International Conference on Machine Learning (ICML). 2006.

Maximum Margin Planning

Notation

- $x \in \mathcal{X}$: state
- $a \in \mathcal{A}$: action
- $p(y|x,a)$: transition probability
- s: initial state distribution
- $v \in V$: value function
- $\mu \in \mathcal{G}$: state-action frequency count
- $\mathcal{D} = \{(\mathcal{X}_i, \mathcal{A}_i, p_i, F_i, y_i, \mathcal{L}_i)\}_{i=1}^n$, training set
  - y_i: desired trajectory (or full policy)
  - $\mathcal{L}_i$ (or l_i): loss function
  - F_i: feature matrix ($d \times |\mathcal{X}| \cdot |\mathcal{A}|$)
  - $\mu_i \in \mathcal{G}$: state-action frequency count
  - $F_i\mu$: expected feature count
  - $w^T F_i \mu$: reward (w: parameter)
  - $l_i^T \mu$: loss (difference from demonstrated policy μ_i)

Maximum Margin Planning

- Maximum margin planning ($q \in \{1, 2\}$)

$$\min_{w,\zeta_i} \quad \frac{1}{2}\|w\|^2 + \frac{\gamma}{n}\sum_i \beta_i \zeta_i^q \tag{1}$$

subject to

$$\forall i \qquad w^T F_i \mu_i + \zeta_i \geq \max_{\mu \in \mathcal{G}_i} \; w^T F_i \mu + l_i^T \mu \tag{2}$$

- Demonstrated policy has a higher reward than any other policies by a margin
- ζ_i: slack variable
- $\mu \in \mathcal{G}_i$: Bellman-flow constraint ($\mu \geq 0$):

$$\sum_{x,a} \mu^{x,a} p_i(x'|x,a) + s_i^{x'} = \sum_a \mu^{x',a}$$

Duality

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \iff \begin{aligned} \min \quad & v^T b \\ \text{s.t.} \quad & v^T A \geq c^T \end{aligned}$$

- Maximum margin planning ($q \in \{1, 2\}$)

$$\min_{w,\zeta_i} \quad \frac{1}{2}\|w\|^2 + \frac{\gamma}{n}\sum_i \beta_i \zeta_i^q \tag{1}$$

$$\text{subject to}$$

$$\forall i \qquad w^T F_i \mu_i + \zeta_i \geq \max_{\mu \in \mathcal{G}_i} \; w^T F_i \mu + l_i^T \mu \tag{2}$$

- $\mu \in \mathcal{G}_i$: Bellman-flow constraint ($\mu \geq 0$):

$$\sum_{x,a} \mu^{x,a} p_i(x'|x,a) + s_i^{x'} = \sum_a \mu^{x',a}$$

- RHS of (2) becomes

$$\max \quad w^T F_i \mu + l_i^T \mu$$

$$\text{s.t.} \sum_x \sum_a \mu(x,a) p_i(x'|x,a) + s_i(x') = \sum_a \mu(x',a)$$

$$\mu(x,a) \geq 0 \qquad \forall x, a$$

- RHS of (2)

$$\begin{aligned} \max \quad & w^T F_i \mu + l_i^T \mu \\ \text{s.t.} \quad & \sum_x \sum_a \mu(x,a) p_i(x'|x,a) + s_i(x') = \sum_a \mu(x',a) \\ & \mu(x,a) \geq 0 \qquad \forall x, a \end{aligned}$$

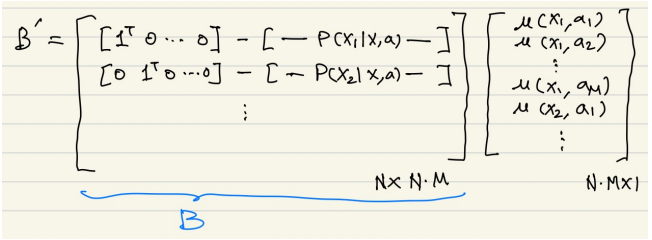
- Equivalent to

$$\begin{aligned} \max \quad & (w^T F_i + l_i^T)\mu \\ \text{s.t.} \quad & \mu(x,a) \geq 0 \qquad \forall x, a \end{aligned}$$

$$s_i = \begin{bmatrix} s_i(x_1) \\ \vdots \\ s_i(x_N) \end{bmatrix} = \begin{bmatrix} \sum_a \mu(x_1,a) - \sum_x \sum_a \mu(x,a) p_i(x_1|x,a) \\ \vdots \\ \sum_a \mu(x_N,a) - \sum_x \sum_a \mu(x,a) p_i(x_N|x,a) \end{bmatrix} = B'$$

- Equivalent to

$$\begin{aligned} \max \quad & (w^T F_i + l_i^T)\mu \\ \text{s.t.} \quad & B\mu = s_i \\ & \mu \geq 0 \end{aligned}$$



- RHS of (2)

$$\begin{aligned} \max \quad & (w^T F_i + l_i^T)\mu \\ \text{s.t.} \quad & B\mu = s_i \\ & \mu \geq 0 \end{aligned}$$

- Dual problem

$$\begin{aligned} \min \quad & v^T s_i \\ \text{s.t.} \quad & v^T B \geq (w^T F_i + l_i^T) \end{aligned}$$

- Since $(v^T B)^{x,a} = v^x - \sum_{x'} v(x') p_i(x'|x,a)$, the constraint becomes, for all x and a,

$$v^x \geq (w_i^T F_i + l_i)^{x,a} + \sum_{x'} p_i(x'|x,a) v^{x'}$$

Maximum Margin Planning

- Maximum margin planning ($q \in \{1, 2\}$)

$$\min_{w,\zeta_i} \quad \frac{1}{2}\|w\|^2 + \frac{\gamma}{n}\sum_i \beta_i \zeta_i^q \tag{1}$$

subject to

$$\forall i \qquad w^T F_i \mu_i + \zeta_i \geq \max_{\mu \in \mathcal{G}_i} \; w^T F_i \mu + l_i^T \mu \tag{2}$$

- Using duality, (2) becomes

$$\forall i \qquad w^T F_i \mu_i + \zeta_i \geq \min_{v \in V_i} \; s_i^T v$$

$$\forall x, a \qquad v^x \geq (w_i^T F_i + l_i)^{x,a} + \sum_{x'} p_i(x'|x,a) v^{x'}$$

- Putting all together

$$\min_{w,\zeta_i,v_i} \quad \frac{1}{2}\|w\|^2 + \frac{\gamma}{n}\sum_i \beta_i \zeta_i^q$$

subject to

$$\forall i \qquad w^T F_i \mu_i + \zeta_i \geq s_i^T v_i$$

$$\forall i, x, a \qquad v_i^x \geq (w_i^T F_i + l_i)^{x,a} + \sum_{x'} p_i(x'|x,a) v_i^{x'}$$

Ziebart, Brian D., Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. "**Maximum Entropy Inverse Reinforcement Learning**." In AAAI, vol. 8, pp. 1433-1438. 2008.

Maximum Entropy Inverse Reinforcement Learning

Entropy

- X: random variable ($X \sim P(X)$)
- Entropy: $H(X) = \mathbb{E}(-\log P(X)) = -\sum_x P(x) \log P(x)$
  - measures the uncertainty of the distribution $P(X)$.
  - the number of bits needed to encode samples from $P(X)$ on average
  - If X is a continuous random variable, we use differential entropy, $h(X) = -\int f(x) \log f(x) dx$
- Cross entropy: $H(P, Q) = \mathbb{E}_P(-\log Q(X)) = -\sum_x P(x) \log Q(x)$
  - expected number of bits needed to encode samples from P if coding is based on another distribution Q
- Kullback-Liebler divergence (relative entropy)

$$H(P\|Q) = \mathbb{E}_P\left(\log \frac{Q}{P}\right) = \mathbb{E}_P\left(\log \frac{P}{Q}\right) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$
$$= -\sum_x P(x) \log Q(x) + \sum_x P(x) \log P(x) = H(P, Q) - H(P)$$

  - $H(P\|Q) \geq 0$
  - $H(P\|Q) = 0$ if $P \overset{d}{=} Q$

Maximum Entropy Probability Distribution

- Maximum entropy probability distribution problem ($X \sim P(X)$)

$$\begin{aligned} \max \quad & H(X) \\ \text{s.t.} \quad & \mathbb{E}(f_k(X)) = a_k \qquad k = 1, \ldots, n \\ & \sum_x P(x) = 1 \\ & P(x) \geq 0 \qquad \forall x \end{aligned}$$

- Solution

$$P(x) = \frac{1}{Z} \exp\left(\sum_k \alpha_k f_k(x) \right),$$

where

$$Z = \sum_x \exp\left(\sum_k \alpha_k f_k(x) \right)$$

Inverse Reinforcement Learning (IRL)

- Expert's demonstration: $\zeta = (s_t, a_t)$
  - ζ: trajectory
  - s_t: state
  - a_t: action
- $f_{s_t} \in \mathbb{R}^k$, feature of state s_t
- $f_\zeta = \sum_{s_t \in \zeta} f_{s_t}$, feature count
- $\text{reward}(f_\zeta) = \theta^T f_\zeta = \sum_{s_t \in \zeta} \theta^T f_{s_t}$ (θ, parameter)
- From m demonstrations, the empirical average feature count is $\tilde{f} = \frac{1}{m} \sum_{i=1}^{m} f_{\zeta_i}$
- (Abbeel & Ng, 2004) aims to match feature expectations

$$\mathbb{E}(f_\zeta) = \tilde{f}$$

  - Too many solutions
  - Need to use some criteria or regularization

Maximum Entropy IRL

- Maximum entropy principle
  - we are making the least assumptions
  - minimizes the worst-case error
- MaxEnt IRL

$$\begin{aligned} \arg \max_{\pi(a_t|s_t)} \quad & H(\pi) \\ \text{subject to } & \mathbb{E}(f_\zeta) = \tilde{f} \\ & \sum_{a_t} \pi(a_t|s_t) = 1 \\ & \pi(a_t|s_t) \geq 0 \qquad \forall a_t \end{aligned}$$

- Solution

$$\pi(a_t|s_t) = \frac{Z(s_t, a_t)}{Z(s_t)}$$

$$\log Z(s_t) = \log \sum_{a_t} Z(s_t, a_t)$$

$$Z(s_t, z_t) = \exp\left(\sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \log Z(s_{t+1}) + \theta^T f_{s_t, a_t} \right)$$

Maximum Entropy IRL

- MaxEnt IRL solution

$$\pi(a_t|s_t) = \frac{Z(s_t, a_t)}{Z(s_t)}$$

$$\log Z(s_t) = \log \sum_{a_t} Z(s_t, a_t)$$

$$Z(s_t, a_t) = \exp\left(\sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \log Z(s_{t+1}) + \theta^T f_{s_t, a_t}\right)$$

- Interpretation

$$Q(s_t, a_t) = \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \log Z(s_{t+1}) + \theta^T f_{s_t, a_t}$$

$$= \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \log Z(s_{t+1}) + r(s_t, a_t)$$

$$V(s_t) = \log Z(s_t) = \log \sum_{a_t} \exp(Q(s_t, a_t))$$

$$\pi(a_t|s_t) = \frac{Z(s_t, a_t)}{Z(s_t)} = \frac{e^{Q(s_t, a_t)}}{e^{V(s_t)}} = \exp\left(Q(s_t, a_t) - V(s_t)\right)$$