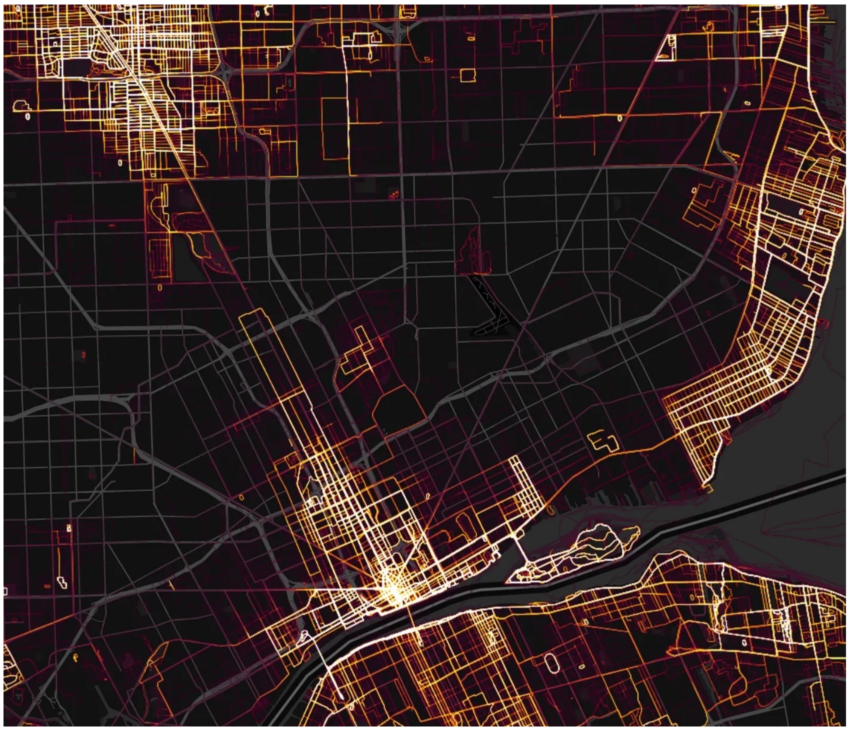
INVERSE REINFORCEMENT LEARNING:

FROM MAXIMUM MARGIN

TO MAXIMUM ENTROPY



– SANJIBAN CHOUDHURY

Formalizing Inverse Reinforcement Learning (IRL)

MDP: $\langle S, A, T, C \rangle$

known, known, known → UNKNOWN!

$$C_\theta(s,a) := w^1\, \mathbb{1}(s \in \text{WATER}) + w^2\, \mathbb{1}(s \in \text{GRASS}) + w^3\, \mathbb{1}(s \in \text{ROCKS})$$

$$:= w^1 f^1(s,a) + w^2 f^2(s,a) + w^3 f^3(s,a)$$

(feature map)

$$:= w^T f(s,a)$$

$C_\theta(s,a) \longmapsto$ $\boxed{\text{PLANNER} \quad \arg\min_{a_1,\dots,a_T} \sum_{t=1}^{T} C_\theta(s_t,a_t)}$ → TRAJECTORY $\xi = s_1, a_1, \dots s_T, a_T$

$$C_\theta(\xi) = w^T f(\xi) = \sum_{(s,a)\in\xi} w^T f(s,a)$$

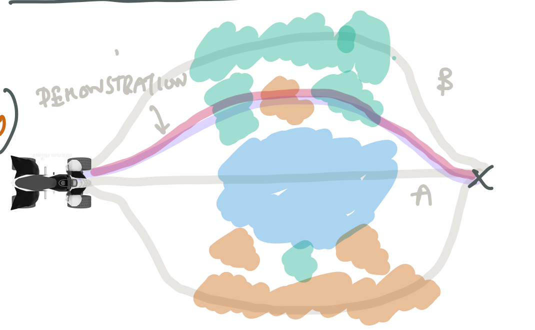
Objective

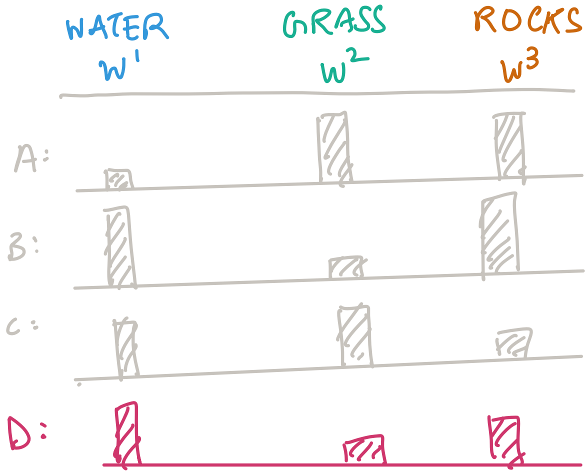
Given (optimal) demonstration $\xi^* = (s_1, a_1, \dots s_T, a_T)$

find $C_\theta(s,a)$ that generates it (find w)

Known as Inverse Optimal Control (IOC), IRL, ...

Off-Terrain Navigation

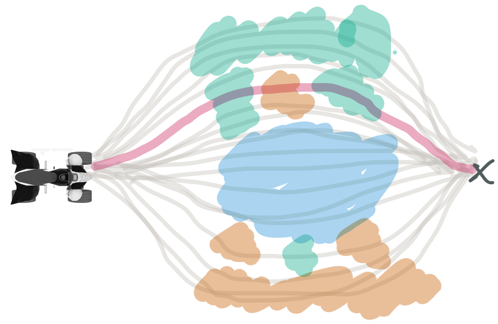




JUST A (VERY VERY LARGE) MULTI-CLASS CLASSIFICATION PROBLEM?

Given N datapoints $\{(\Phi_i, \xi_i^h)\}_{i=1}^N$

MDP HUMAN TRAJ



FIND θ

s.t. $C_\theta(\xi_i^h, \Phi_i) \leq C_\theta(\xi, \Phi_i) \;\forall \xi, \forall i$

COST OF HUMAN TRAJ IN MDP — COST OF ALL TRAJ IN MDP

(i^{th} datapoints)

EVERY TRAJ IS A CLASS $\rightarrow$ A VERY LARGE CLASSIFICATION PROB

LINEAR SETTING: EVERY MDP Φ_i INDUCES A FEATURE MAP f_i

$$C_\theta(\xi, \Phi_i) = w^T f_i(\xi) = \sum_{(s,a)\in\xi} \left(w^1 f_i^1(s,a) + w^2 f_i^2(s,a) + \dots \right)$$

FIND $\mathbf{w}$

s.t. $w^T f_i(\xi_i^h) \leq w^T f_i(\xi) \;\forall \xi \;\forall i$

$$w^T f_i(\xi_i^h) \le w^T f_i(\xi) \quad \forall \xi$$

FAIL

$$\begin{cases} 0 & \text{if } \xi = \xi^h \\ 1 & \text{else} \end{cases}$$

USE MARGIN!

$$w^T f_i(\xi_i^h) \le w^T f_i(\xi) - \underbrace{\gamma_i(\xi)}_{\text{MARGIN}} \quad \forall \xi$$

$$\le \min_{\xi} \left[w^T f_i(\xi) - \gamma_i(\xi) \right]$$

FAIL

MAXIMIZE MARGIN!

$$\min_{w} \quad \|w\|^2 + \frac{1}{N} \sum_{i=1}^{N} \eta_i$$

s.t. $$w^T f_i(\xi_i^h) \le \min_{\xi} \left[w^T f_i(\xi) - \gamma_i(\xi) \right] + \eta_i$$ (SLACK)

? $W = 0$ satisfies this opt

TRIVIAL SOLUTIONS!

⚠ Exponential # of constraints ...

?? let w' be a weight s.t.

$$w' f(\xi^h) \le w' f(\xi) - \epsilon$$ (TINY)

Setting $w = w' \times 10^{1000}$ ... solves the problem!

(MULTIPLE SOLUTIONS!)

Maximum Margin Planning

Nathan D. Ratliff
J. Andrew Bagnell
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA. 15213 USA

Martin A. Zinkevich
Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E

EXTEND TO NON-LINEAR COST FUNCTIONS [LEARCH]

$$\min_{C_\theta} \frac{1}{N} \sum_{i=1}^{N} \Big(\underbrace{C_\theta(\xi_i^h, \phi_i)}_{\text{(HUMAN COST)}} - \underbrace{\min_{\xi} \big[C_\theta(\xi, \phi_i) - \gamma_i(\xi) \big]}_{\text{(PLANNER COST)}} \Big) + \underbrace{R(C_\theta)}_{\text{(REG)}}$$

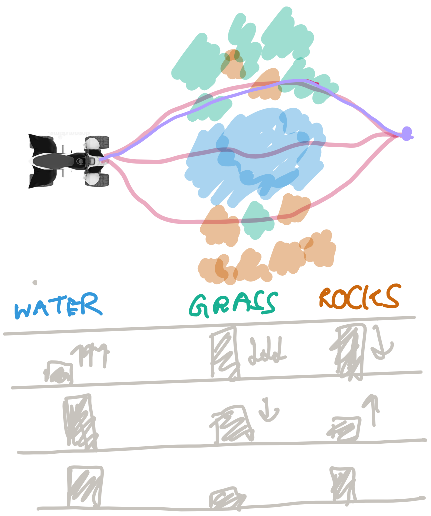
Learning to Search:
Functional Gradient Techniques for Imitation Learning

Nathan D. Ratliff
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
ndr@ri.cmu.edu

David Silver
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
dsilver@ri.cmu.edu

J. Andrew Bagnell
Robotics Institute and Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213
dbagnell@ri.cmu.edu

ASIDE: 3 ways to do this (1) Boosting (2) Kernel gradient descent (3) Parametric functional gradient descent (e.g deep learning)



Alg

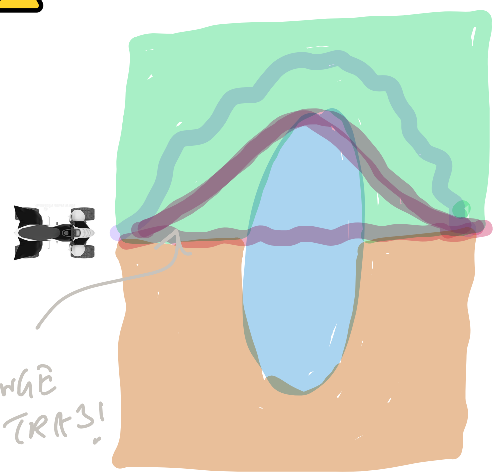
for $i = 1 \ldots N$ # loop over datapoints (MDP ϕ_i , HUMAN DEM ξ_i^h)

$$\xi_i^* = \min_{\xi} \big[C_\theta(\xi, \phi_i) - \gamma_i(\xi) \big]$$ # Call planner to find optimal path

Iter 0:

$$\theta^+ = \theta - \alpha \big[\nabla_\theta C_\theta(\xi_i^h, \phi_i) - \nabla_\theta C_\theta(\xi_i^*(\theta)) \big] + \nabla_\theta R(C_\theta)$$

PROBLEM: SUBOPTIMAL DEMONSTRATIONS



NO CHANGE IN TRAJ!

UNREALIZABLE EXPERT:

$\nexists$ ANY C_θ that results in expert traj being optimal

↳ GRADIENT keeps running away!

↳ No convergence!

INSTEAD OF:

FIND COST s.t.

EXPERT COST $\leq$ OPTIMAL TRAJ COST

PROBABILISTIC

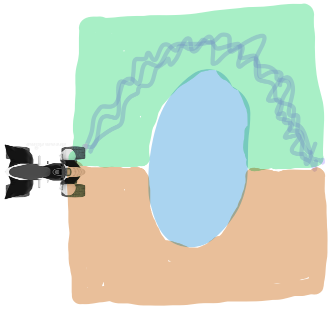
FIND DISTRIBUTION over traj $P(\xi)$ s.t.

EXPECTED EXPERT COST = EXPECTED TRAJ COST | FOR ALL COST FUNCTIONS!

PARADIGM 2: PROBABILISTIC MODELING VIA MOMENT MATCHING

Let's return to linear case, i.e $c_\theta(s,a) = w^T f(s,a) = [w^1\ w^2 \dots] \begin{bmatrix} f^1(s,a) \\ \vdots \\ f^k(s,a) \end{bmatrix}$

Assume: Human traj distribution $P(\xi^h)$



Find $\boxed{P(\xi)}$

$$\boxed{\max \quad -\int P(\xi) \log P(\xi)\, d\xi} \quad \text{(ENTROPY)}$$

$$\text{s.t.} \quad \mathop{E}_{\xi^h \sim P(\xi^h)} f^k(\xi^h) = \mathop{E}_{\xi \sim \boxed{P(\xi)}} f^k(\xi) \quad \forall k \quad \text{(MOMENT MATCHING)}$$

E.g. Avg time expert spends on grass f^1 = Avg time learners spend on grass

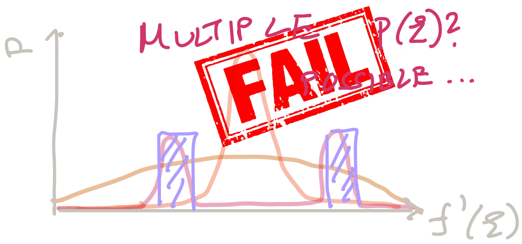
'' on water = '' on water

'' on rocks = '' on rocks

ANY LINEAR COMBINATION!

(If all moments matched, EXPERT COST = LEARNER COST)

... PROBLEM ....



PRINCIPLE OF MAXIMUM ENTROPY

$$\max_{\xi} \; -\sum P(\xi) \log P(\xi)$$

$$\text{s.t} \quad \sum_{\xi} P(\xi) f^k(\xi) = F^k$$

(EXPECTED FEATURE COUNT) (EXPERT FEATURE COUNT)

Eg Avg time on water $F^1 = 0.0$, Avg time on grass $F^2 = 0.8$, Avg time on rocks $F^3 = 0.2$

Information Theory and Statistical Mechanics

E. T. JAYNES
Department of Physics, Stanford University, Stanford, California
(Received September 4, 1956; revised manuscript received March 4, 1957)

(Pick a distribution that is least committal)

Sketch of solution: First write out LAGRANGIAN

$$\mathcal{L}(\cdots) = -\sum P(\xi) \log P(\xi) - \sum_k \lambda^k \left(\sum_{\xi} P(\xi) f^k(\xi) - F^k \right)$$

Take gradient & set to 0 and solve! Solution:

$$P_\lambda(\xi) = \frac{1}{Z} \exp\left(-\sum \lambda^k f^k(\xi)\right)$$

(THINK OF LAGRANGE MULTIPLIER λ^k as being weights w^k!

$P(\xi) \propto \exp(-\text{COST}) \Rightarrow$ LOW COST mean HIGH PROB!

FINAL step: Plug in solution to optimization to solve λ^k:

$$\max_{\lambda^k} \frac{1}{N} \sum_{i=1}^{N} \log P_\lambda(\xi_i^h)$$

↳ MAXIMIZE LOG LIK OF EXPERT TRAJ. → GENERATIVE MODELLING!

MAXIMUM ENTROPY INVERSE OPTIMAL CONTROL

$$\min_{C_\theta} \frac{1}{N}\sum_{i=1}^{N} -\log P_\theta(\xi_i^h \mid \phi) \quad \text{where} \quad \boxed{P_\theta(\xi \mid \phi) = \frac{1}{Z(\theta,\phi)} \exp(-C_\theta(\xi,\phi))}$$

(MODEL)

Maximum Entropy Inverse Reinforcement Learning

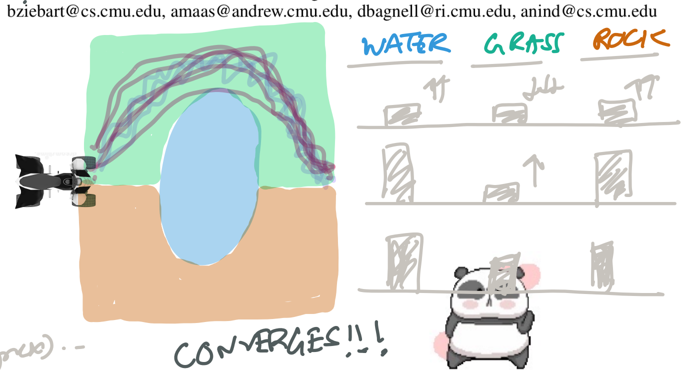
Brian D. Ziebart, Andrew Maas, J.Andrew Bagnell, and **Anind K. Dey**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
bziebart@cs.cmu.edu, amaas@andrew.cmu.edu, dbagnell@ri.cmu.edu, anind@cs.cmu.edu

Plugin MODEL in objective to get.

$$\min_{C_\theta} \underbrace{\frac{1}{N}\sum_{i=1}^{N} C_\theta(\xi_i^h, \phi)}_{\text{EXPERT COST}} + \underbrace{\log Z(\theta,\phi)}_{\text{SOFT MIN OF LEARNER COST}}$$

$$Z(\theta,\phi) = \int_\xi \exp(-C_\theta(\xi,\phi))\, d\xi$$

EXPERT COST − SOFT MIN OF LEARNER COST



ALGORITHM

for $i = 1 \ldots N$

- Sample $\xi^i \sim \exp(-C_\theta(\xi,\phi))$ # SOFT MIN (Sample low cost traj) (MCMC, Laplace approx)..

$$\theta^+ \leftarrow \theta - \alpha\left[\nabla_\theta C_\theta(\xi^h,\phi) - \mathop{E}_{\xi \sim \text{sample}} \nabla_\theta C_\theta(\xi,\phi)\right]$$

MAX-MARGIN PLANNER

$$\theta^+ \leftarrow \theta - \alpha \left[\nabla_\theta C_\theta(\xi_i^h, \phi_i) - \nabla_\theta C_\theta(\xi_i^*, \phi_i) \right]$$

HUMAN

OPTIMAL COST TRAJ

MAXIMUM ENTROPY INVERSE OPT CONTROL

$$\theta^+ \leftarrow \theta - \alpha \left[\nabla_\theta C_\theta(\xi_i^h, \phi) - \mathop{E}_{\xi \sim \text{sample}} \nabla_\theta C_\theta(\xi, \phi_i) \right]$$

(HUMAN)

SAMPLE LOW COST TRAJ

SOFT

TIE-BREAKING: MAX MARGIN

REGULARIZATION ON COST (L2)

OPTIMAL PLANNER (A^* e.g.)

TIE-BREAKING: MAX ENTROPY.

ENTROPIC REGULARIZATION ON PLANNER

"SOFT-" PLANNER (SAC e.g.)

BOTH SOLVING THE SAME GAME

NO REGRET + BEST RESPONSE

$$\boxed{\boxed{\max_{C_\theta}} \boxed{\min_{\xi}} \; C_\theta(\xi^h) - C_\theta(\xi)}$$

NO REGRET + ENTROPY REG BEST RESPONSE

KEY CHALLENGE: LEARN COST FUNCTION THAT EXPLAINS EXPERT DEMONSTRATIONS

* METHOD 1: Recover costs that make expert optimal
  - ↳ Maximum margin planning.
  - ↳ PROBLEM: No suboptimal expert

* METHOD 2: Learn a distribution over traj that matches all cost function moments
  - ↳ Maximum entropy
  - ↳ Can deal with any expert dist!

* Both method are solving the same game in different ways!

It's all a game between DISCRIMINATOR (COST FUNCTION / ADVANTAGE function) & GENERATOR (PLANNER / POLICY)

Hey everyone,

Welcome to sixth lecture in our series on imitation learning.

I am Sanjiban Choudhury, a research scientist at Aurora & soon to be assistant professor at Cornell

Today we'll talk about a class of imitation learning methods that try to recover the underlying reward or cost function that the human optimizes. We will look at the problem from 2 different perspectives — #1 by treating human as optimal planner #2 by treating human as a stochastic process. And in the end, we will see that they both converge at a unified game theoretic framework. Let's get started.