Last Time: Utility Theory, Games

- Instead of optimal soln., Nash Equillibrium

Today: Alt optimization obj.

- Weighted sum
- Pareto Frontiers
- Constrained MDPs + POMDPs
- Coherent Risk Measures
- Pure Info Gathering
- Penalties for uncertainty

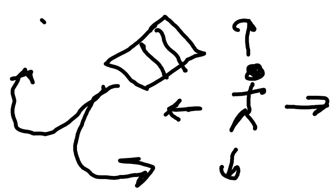
Differential Games

$$\dot{x} = f(x, u_1, \dots u_N)$$

$$J_i = \int_0^T g_i(x, u_1 \dots u_N)\, dt$$

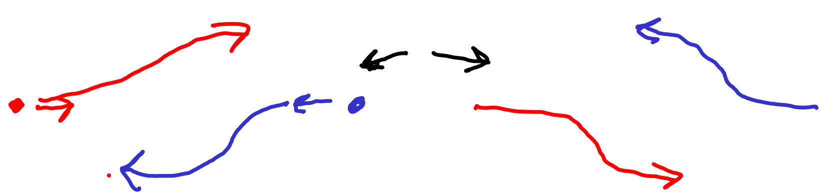
$$u_i(t) = \gamma_i(t, x)$$ strategy for player i

Zero-sum Example: Homicidal Chauffer



Solved with Hamilton-Jacobi-Isaacs PDE, Level Set Methods

General-Sum Differential Game Example: Hallway



iLQ Games

MDP: Expectimax DP

$$\rightarrow \quad V^*(s) = \max_a \left(E\left[R(s,a,s') + \gamma V^*(s') \mid s,a \right] \right)$$

Zero-Sum Game against nature

Minimax

$$\rightarrow \quad V^*(s) = \max_a \left(\min_{s'} \left(R(s,a,s') + \gamma V^*(s') \mid s,a \right] \right)$$

<u>POSG</u> - Reason about other player's beliefs

Humans don't behave according to game theory

- Unclear which Equilibrium
- Difficult to compute
- Doubt opponent's ability

Logit-level k model ← good for modeling humans in games

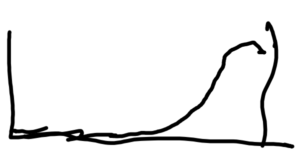
$\lambda \geq 0$ precision

$k > 0$ depth

Level 0: selects uniformly

Level k: assumes others adopt k-1

$$P(a_i) \propto e^{\lambda U_i(a_i,\, s_{-i})}$$



Multiple Objectives

$R(s,a)$

$R_1(s,a)$ $\qquad R_2(s,a)$ $\qquad$ Efficiency

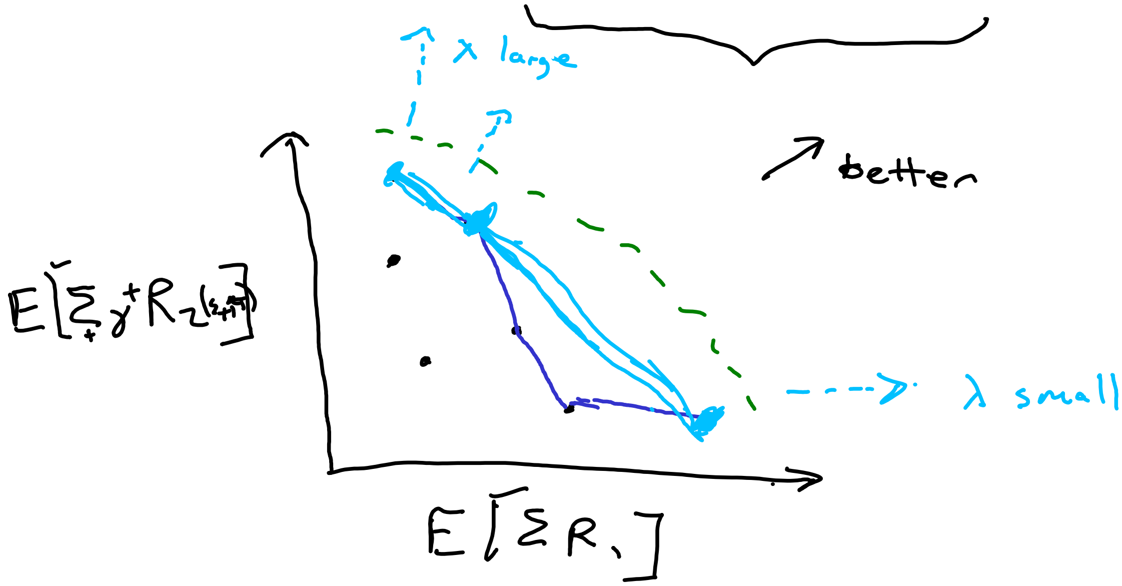
Safety

Coronavirus $\qquad$ Economy

Deaths

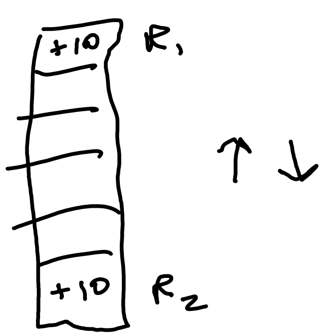
Weighted Sum

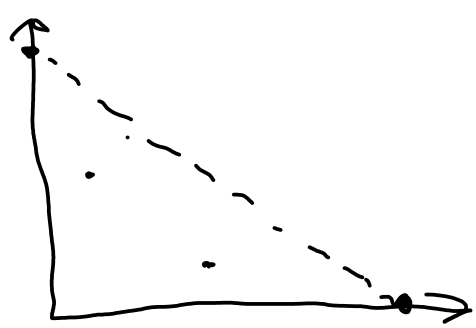
$$R(s,a) = \underbrace{R_1(s,a) + \lambda_2 R_2(s,a)} + \dots \lambda_N R_N(s,a)$$

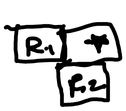


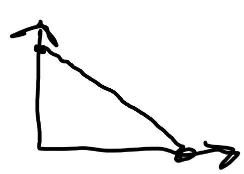
Is there a policy that achieves every point on the convex pareto frontier?

No









Option 2

Constrained (PO) MDPs

maximize $E\left[\sum \gamma^t R(s_t, a_t)\right]$

subject to $E\left[\sum \gamma^t C_1(s_t, a_t)\right] \leq D_1$

$E\left[\sum \gamma^t C_2(\quad)\right] \leq D_2$

Safe 99.999%

Solution: Altman CMDPs LP, Lagrange multipliers

1. In C(PO)MDPs, stochastic policies may outperform deterministic ones.

MDP: suppose that stochastic policy π^* is optimal

$$V^*(s) = \underset{a \sim \pi^*}{E}\left[R(s,a) + \gamma E\left[V^*(s') \mid s,a\right]\right]$$

$$= \underset{a \sim \pi^*}{E}\left[Q^*(s,a)\right]$$

Claim: $Q^*(s,a_1) = Q^*(s,a_2) \;\forall a$ s.t. $\pi^*(a|s) > 0$

suppose claim false

then $\pi'(s) = \underset{a \in A}{\operatorname{argmax}}\, Q^*(s,a)$ is a better policy than π^* $\times$

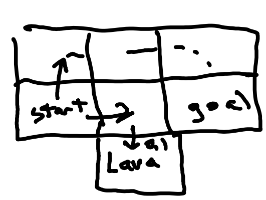
$\therefore$ claim is true

if we choose any action in π^* deterministically, that action is optimal

(1) in any MDP, $\exists$ a deterministic policy that is at least as good as any stochastic policy

CMDP, (1) is not true

Counterexample:



2. Limit to deterministic policies

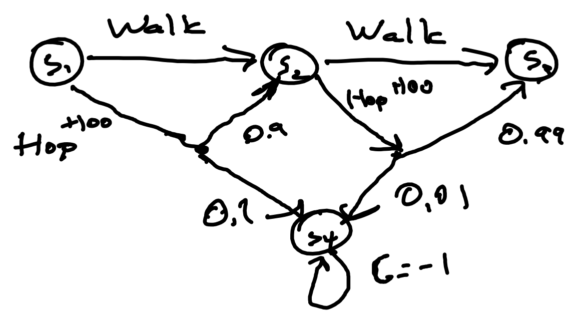
Is there a weighted reward function that achieves the best deterministic policy for a CMDP

$$R'(s,a) = R(s,a) - \lambda C(s,a)$$

sort of like a Lagrange Multiplier

No

Counterexample



$\gamma = 0.95$

Best deterministic policy for $D=0.1$

is $s_1 \to$ hop

$s_2 \to$ walk

$E[R] = 100$

$E[C] = 0.1$

$R' = R - \lambda C$

for $\lambda < 9.136\ldots$

hop on both

$E[R] = 185.5$

$E[C] = 0.1086\ldots$

Not Feasible

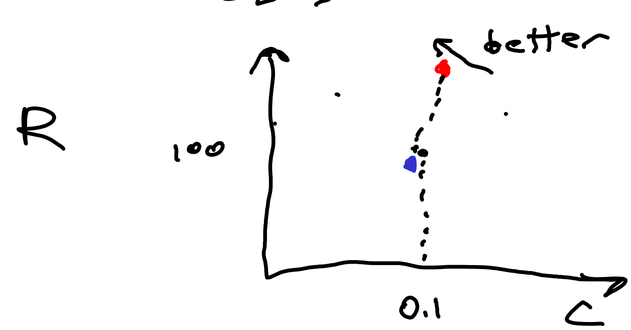
for $\lambda > 913.68$

$s_1 \to$ walk

$s_2 \to$ hop

$E[R] = 95$

$E[C] = 0.0095$

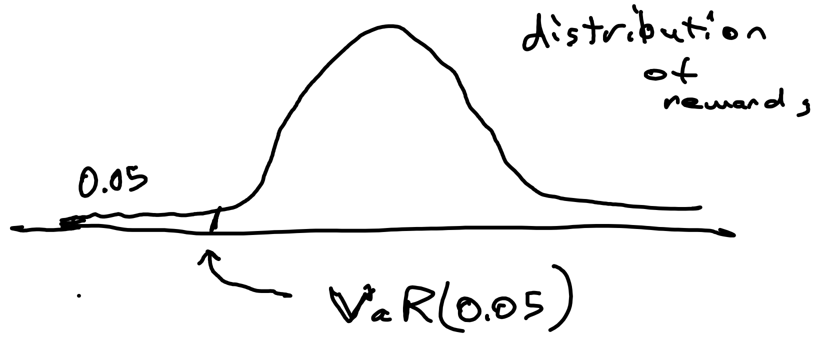


Weighted Sum CMDPs not the same

Coherent Risk Measures

$$E\left[\sum_t \gamma^t R(s_t, a_t)\right]$$

any function of distribution

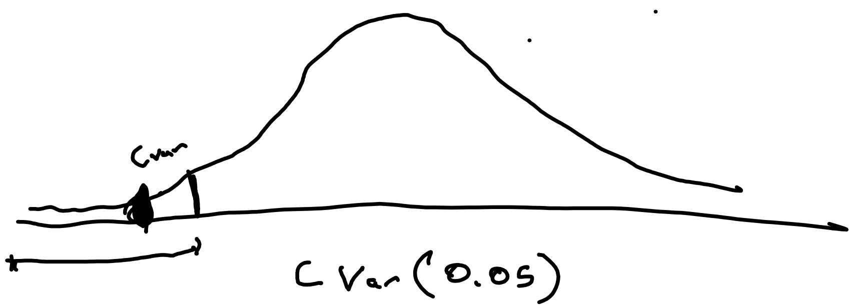


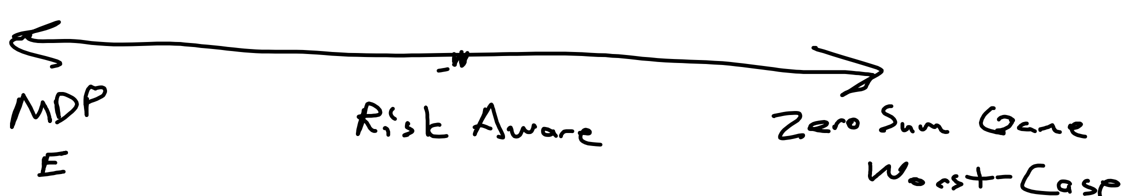
- To use dynamic programming, must use a "coherent" measure of the reward distribution

Convex, Monotonic Translation Invariant, Positively Homogeneous

Example not Coherent: VaR

Example Coherent: CVaR





	Objective	Thing to Remember
Weighted Sum	$R_1 + \lambda_2 R_2 + \lambda_3 R_3$	only finds policies on convex hull of Pareto Front
Constrained	R, $C \le D$	Stochastic Policies Dominate
RiskAware (CVaR)	R, avoid very bad	Between Expected and Worst-Case

In a POMDP

$$R(b,a) = \mathop{E}_{s\sim b}[R(s,a)]$$

What if you just want to gather info?

Options

1) $R(b,a) = -\overset{\swarrow \text{Entropy}}{H}(b)$ POMDP? No
Belief-space MDP? Yes

2) Final Action

$$R_T(s,a) = \mathbb{1}_s(a)$$ POMDP? Yes

3) ρ-POMDP

$$R(b,a) = \rho(b,a)$$

If ρ satisfies some convexity assumptions

V is still PWLC

Can slightly modify SARSOP

Incentivise Info Gain in a POMDP

$$R(b,a) = E[R(s,a)] - \boxed{\lambda H(b)}$$

Example: AMDP, iLQ method Van den Berg 2012
Belief space Planning ... max likelihood obs

Pros: Works well in practice

Cons: choose λ artificially