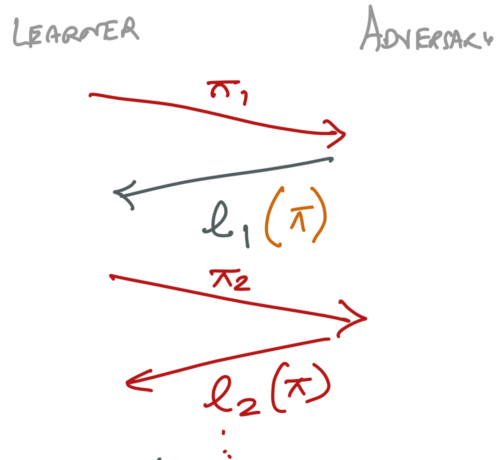
Analyzing DAgger

① Imitation Learning can be viewed as interactive online learning

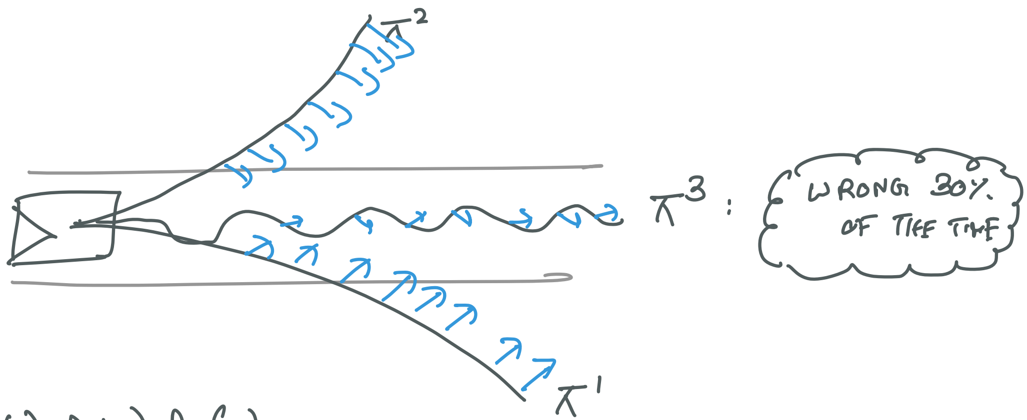
Policies (Π)	$\ell_1(\pi)$	$\ell_2(\pi)$	$\ell_N(\pi)$
π^1	0.1	1.0	
π^2	0.3	0.1	
π^3	1.0	0.3	- - - - - -
⋮	⋮	⋮	
π^k	.	⋮	



Goal: Minimize Avg Regret

$$\text{AvgReg}(\pi_{1:N}) = \frac{1}{N}\Big(\underbrace{\sum_{i=1}^{N} \ell_i(\pi)}_{\text{loss of learner}} - \min_{\pi\in\Pi} \sum_{i=1}^{N} \ell_i(\pi) \Big)$$

No Regret Algorithm $:= \lim_{N\to\infty} \text{AvgReg}(\pi_{1:N}) \to 0$



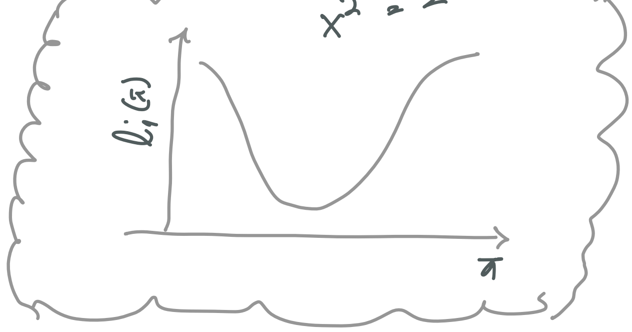
	$\ell_1(\pi)$	$\ell_2(\pi)$	$\ell_3(\pi)$			
π^1	1.0	0.1	0.7	0.7		
π^2	0.1	1.0	0.7	0.7		
π^3	0.3	0.3	0.3	0.3	0	. . . - - -

② Follow The Leader

$$\pi^i = \arg\min_{\pi\in\Pi} \sum_{j=0}^{i-1} \ell_j(\pi)$$

$x = 0$

If $l_i(\pi)$ is strongly convex ∞
then FTL is no-regret



③ DAGGER IS FTL.

$$\pi_{i+1} = \text{TRAIN}(D)$$

$$= \arg\min_{\pi\in\Pi} \sum_{j=0}^{i} l_j(\pi)$$

④ DAGGER RETURNS ATLEAST ONE POLICY π_i
THAT DOES WELL ON IT'S OWN INDUCED DISTRIBUTION

$$\pi_i \quad \text{s.t} \quad \boxed{\sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{s_t\sim d_t^{\pi}} l(s_t,\pi_i)} \le \underline{O(\dots T)}$$

ASSUMPTION: [RICH POLICY CLASS] ∞

$$l(\pi) = \sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{s_t\sim d_t^{\pi}} l(s_t,\pi)$$

FOR ANY LOSS FUNCTION $l(\pi)$, $\exists\, \pi\in\Pi$ THAT IS GOOD

$$\min_{\pi\in\Pi} l(\pi) \le O(\varepsilon T H)$$

recoverability of MDP

PROOF:

LOOK AT THE BEST POLICY THAT DAGGER RETURNS.

$$\min_{i=1,\dots N} l_i(\pi_i)$$

$$\le \frac{1}{N}\sum_{i=1}^{N} l_i(\pi_i)$$

$$\le \frac{1}{N}\left(\sum_{i=1}^{N} l_i(\pi) - \min_{\pi\in\Pi}\sum_{i=1}^{N} l_i(\pi)\right) + \min_{\pi\in\Pi}\frac{1}{N}\sum_{i=1}^{N} l_i(\pi)$$

$\lesssim \underbrace{\qquad\qquad}_{\text{Avg Reg}(\pi_{1:N})}$

By our assumption

$\lesssim O(\varepsilon H T)$

As $N \to \infty$, Av Reg $\to 0$

$$\frac{\log N}{N}$$