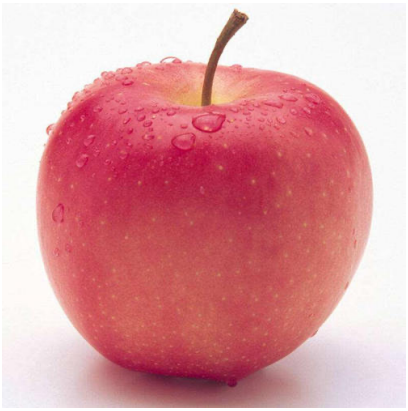
南京大学
人工智能学院
School of Artificial Intelligence, Nanjing University

Lecture 2
RL by Behavior Cloning

Supervised learning

Instance

Label



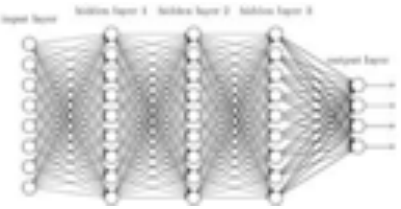
apple

x

y

Learn a model to fit the data $f(x) = y$

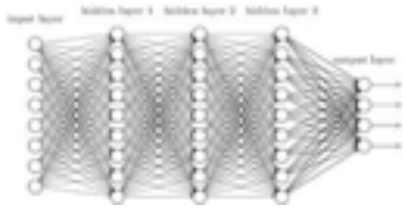
function model



Supervised learning

Learn a model to fit the data $f(x) = y$

function model



Find the model parameters:

$$\theta^* = \arg\min_{\theta} \sum_{i} \|f(x_i|\theta) - y_i\| + \|\theta\|$$

training data distribution

test data distribution

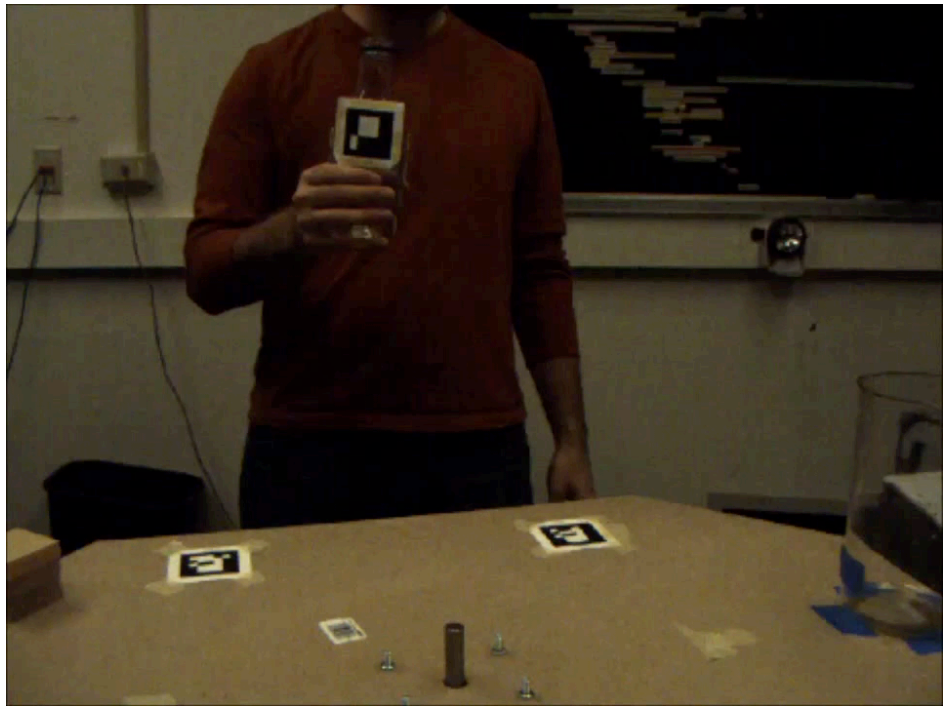

嬰兒的成長
"Imitation"
Andrew Meltzoff's Lab
Center for Mind Brain & Learning
University of Washington, USA
我要給你看這個
準備好了嗎?注意看


iQIYI 爱奇艺
幼崽们之前从不明白它这样做的原因

Imitation learning

Expert/teacher provide demonstrations

$$s_0 \rightarrow a_1 \rightarrow s_1 \rightarrow a_2 \rightarrow s_2 \rightarrow \cdots \rightarrow a_m \rightarrow s_m$$



https://www.youtube.com/watch?v=ydnjS__8Ooc

agent learns from demonstrations to imitate the expert

Copy actions: behavior cloning

demonstration data

$$s_0 \rightarrow a_1 \rightarrow s_1 \rightarrow a_2 \rightarrow s_2 \rightarrow \cdots \rightarrow a_m \rightarrow s_m$$

split into labeled data

$$D = \begin{array}{r} s_0 \rightarrow a_1 \\ s_1 \rightarrow a_2 \\ \cdots \\ s_{m-1} \rightarrow a_m \end{array}$$

learning objective

$$\theta^* = \arg\min_{\theta} E_{s,a\sim D}\, \mathbf{loss}(\pi(s|\theta), a)$$

Behavior cloning examples

used human player data to initialize the policy in AlphaGo and AlphaStar

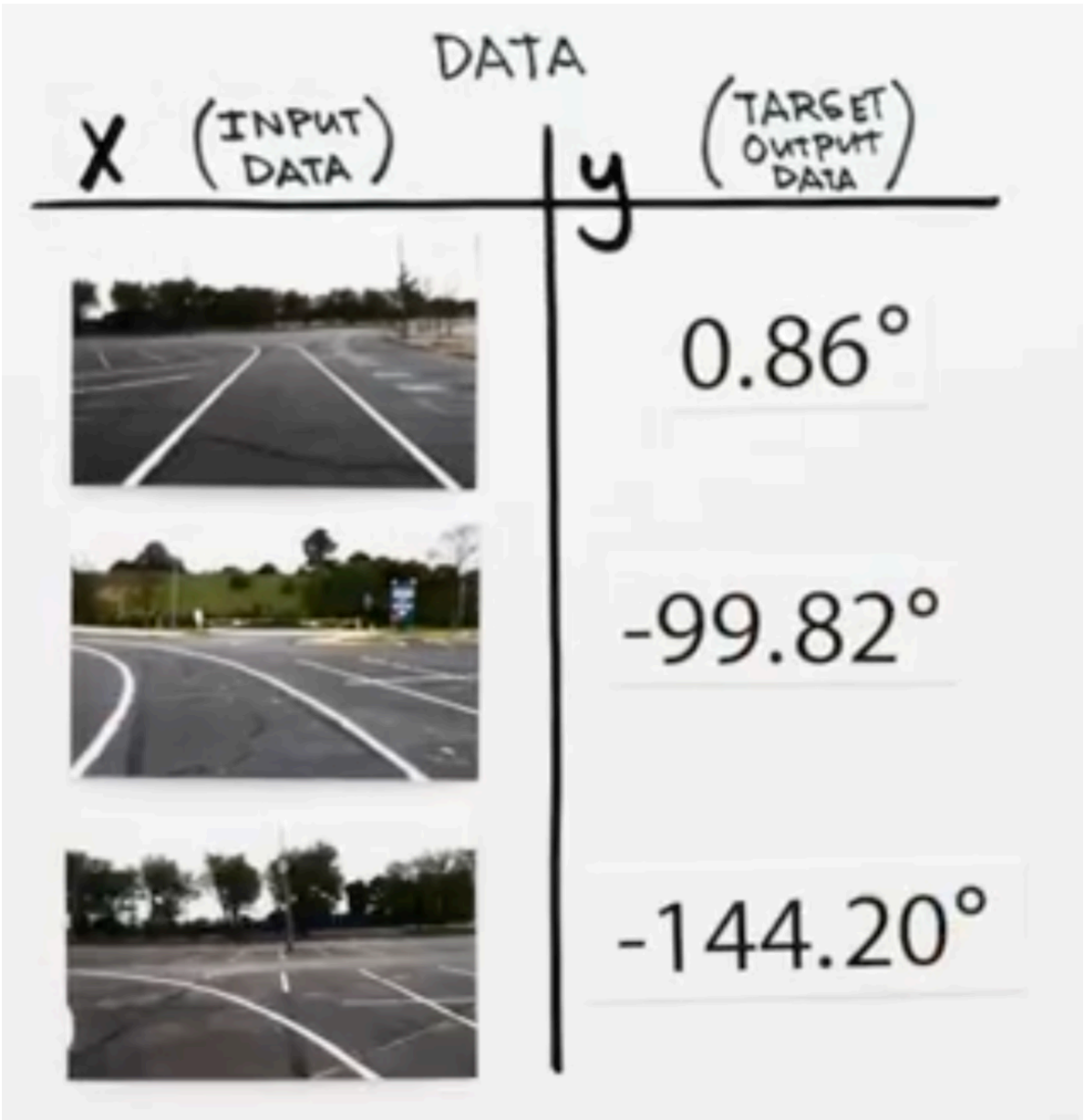
improvement in prepare the labeled data:
remove highly correlated data

$$D = \{ s_0 \to a_1,\ \cancel{s_1 \to a_2},\ \ldots,\ s_{m-1} \to a_m \}$$

quickly learn a rough policy, no trial-and-error cost
but with limited power

Behavior cloning examples

ALVINN Self-driving car



https://www.youtube.com/watch?v=H0igiP6Hg1k

Behavior cloning examples

ALVINN Self-driving car

https://www.youtube.com/watch?v=H0igiP6Hg1k

Behavior cloning examples

Behavior cloning limitation

Supervised learning objective

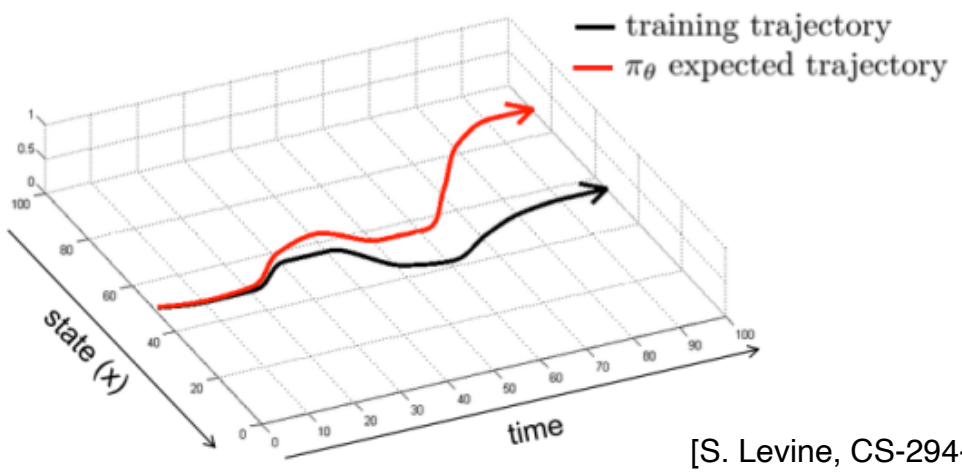
$$\arg\min_{\theta} E_{x\sim\mathcal{D}}\ \text{loss}(f_\theta(x), y(x))$$

Reinforcement learning objective

$$\arg\min_{\theta} E_{s\sim\mathcal{D}^{\pi_\theta}}\ \text{cost}(s, \pi_\theta(s))$$

e.g. cost = -reward

Compounding error:



[S. Levine, CS-294-112-2]

Demo

from http://videolectures.net/aistats2011_ross_reduction/

Behavior cloning limitation - formally

Consider *T*-step reinforcement learning with bounded reward [0,1]

$$J(\theta) = E_{s,a,r\sim\pi_\theta}[\sum_{t=1}^{T} r_t]$$

We have data from the optimal policy

$$s_0 \rightarrow a_1^* \rightarrow s_1 \rightarrow a_2^* \rightarrow s_2 \rightarrow \cdots \rightarrow a_T^* \rightarrow s_T$$

We apply BC(SL) to imitate the policy with a small classification error

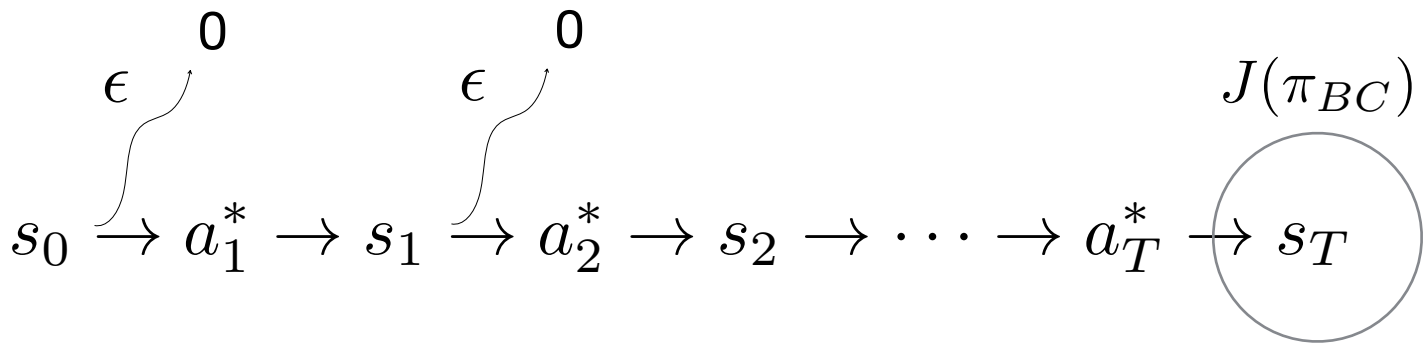
$$E_{s,a^*}[\pi(s) \neq a^*] \leq \epsilon$$

Then the BC policy has a return as

$$J(\pi_{BC}) \geq J(\pi^*) - \frac{T^2+T}{2}\epsilon$$

[John Langford and Bianca Zadrozny Reducing T-step Reinforcement Learning to Classification.]

Proof idea



reward loss at step 1 $< T\epsilon$

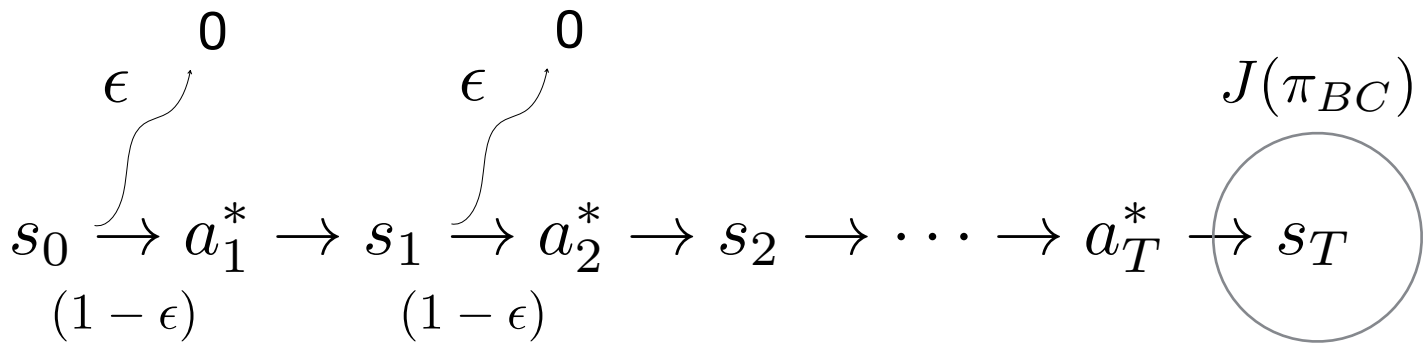
reward loss at step 2 $< (T-1)\epsilon$

reward loss at step *t* $< (T-t)\epsilon$

total loss $< \dfrac{T^2+T}{2}\epsilon$ (union bound)

[John Langford and Bianca Zadrozny Reducing T-step Reinforcement Learning to Classification.]

Consider cumulative error



reward at step 1 $$r(s_0, a_{BC}) \geq r(s_0, a^*) - \epsilon \cdot 1$$

$$r_0^{BC} \geq r_0^* - \epsilon$$

cumulative reward at step 2

$$r_0^{BC}(1-\epsilon) + r_1^{BC} \geq (1-\epsilon)(r_0^* - \epsilon) + (r_1^* - \epsilon)$$

cumulative reward at step T

$$J(\pi_{BC}) \geq \sum_{i=1}^{T} (1-\epsilon)^{i-1}(r_{T-i}^* - \epsilon)$$

[Stéphane Ross, Drew Bagnell: Efficient Reductions for Imitation Learning. In: AISTATS 2010, 661-668.]

Consider cumulative reward (return)

$$J(\pi_{BC}) \geq \sum_{i=1}^{T}(1-\epsilon)^{i-1}(r^*_{T-i}-\epsilon) = \sum_{i=1}^{T}(1-\epsilon)^{i-1}r^*_{T-i} - \sum_{i=1}^{T}(1-\epsilon)^{i-1}\epsilon$$

$$J(\pi^*) = \sum_{i=1}^{T} r^*_{T-i} = \sum_{i=1}^{T}(1-\epsilon)^{i-1}r^*_{T-i} + \sum_{i=1}^{T}\left(1-(1-\epsilon)^{i-1}\right)r^*_{T-i}$$

$$J(\pi_{BC}) \geq J(\pi^*) - \sum_{i=1}^{T}\left(1-(1-\epsilon)^{i-1}\right)r^*_{T-i} - \sum_{i=1}^{T}(1-\epsilon)^{i-1}\epsilon$$

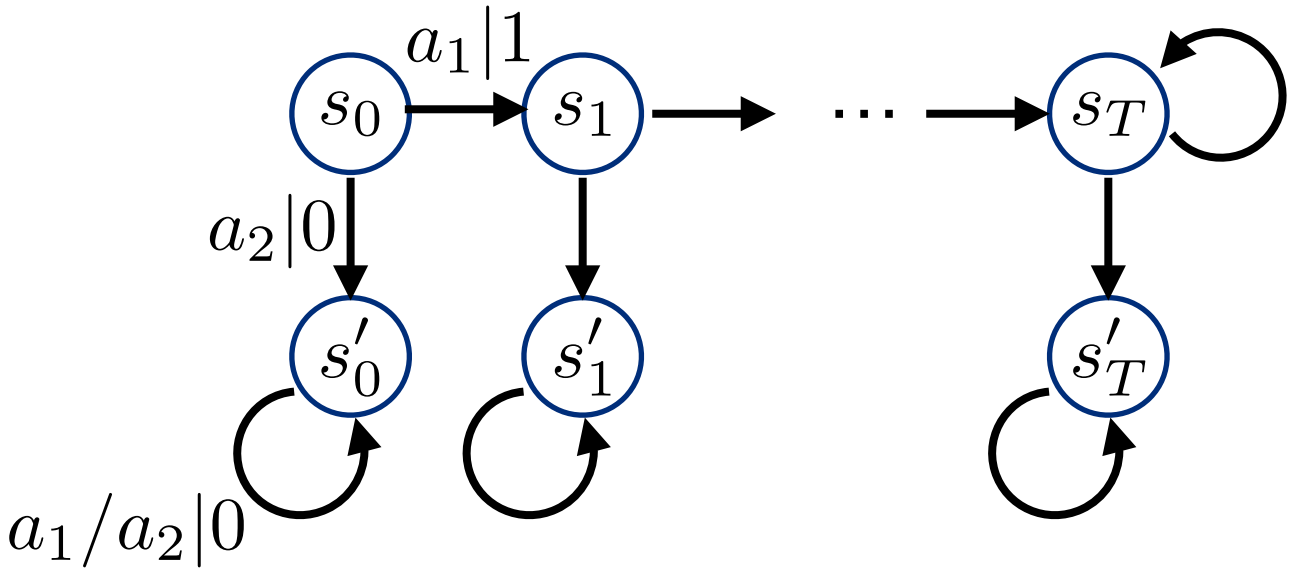
$$1-(1-\epsilon)^{T-1} \leq (T-1)\epsilon$$

$$\geq J(\pi^*) - \sum_{i=1}^{T}(T-1)\epsilon - \sum_{i=1}^{T}\epsilon = J(\pi^*) - T^2\epsilon$$

(compounding error)

[Stéphane Ross, Drew Bagnell: Efficient Reductions for Imitation Learning. In: AISTATS 2010, 661-668.]

Worst case



$$r_1 = 1 - \epsilon$$

$$r_1 + r_2 = 2 - \epsilon - (1 - \epsilon)\epsilon$$

$$\cdots$$

$$\sum_{t=1}^{T} r_t = \sum_{t=1}^{T} 1 - (1 - \epsilon)^{t-1}\epsilon \leq \sum_{t=1}^{T} T\epsilon = T^2\epsilon$$

Can we reduce the compounding error?

(Q will be explained in later lectures, for now we can treat it as the policy)

given $\forall a, s, t : Q^{\pi^*}(s, a) - Q^{\pi^*}(s, a^*) \leq u$

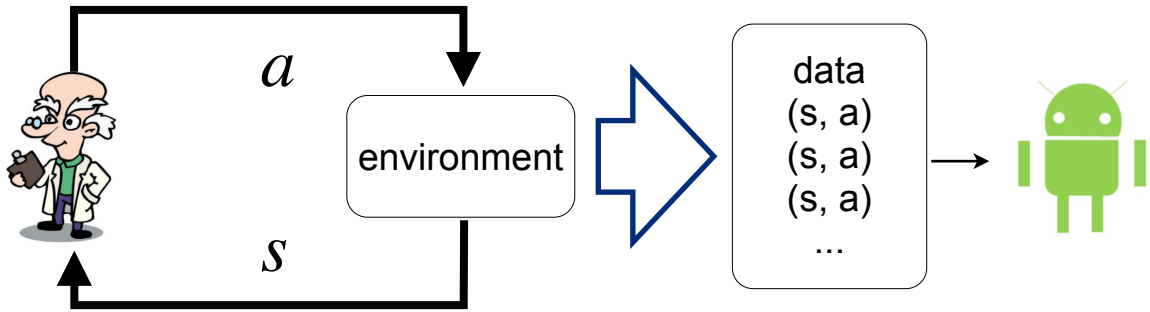
$$J(\pi_{BC}) \geq J(\pi^*) - uT\epsilon$$ (proof in the reference)

to reduce *u*, one way is to enlarge the training data

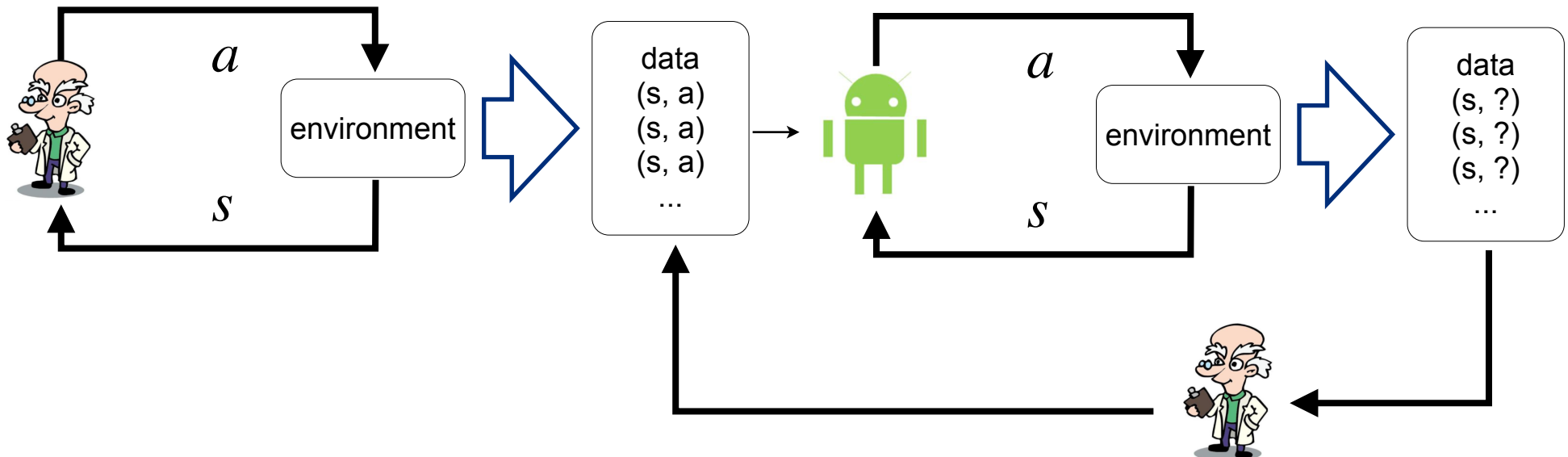
[Stéphane Ross, Geoffrey J. Gordon, Drew Bagnell: A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In: AISTATS 2011, 627-635]

Enlarge the data by experts

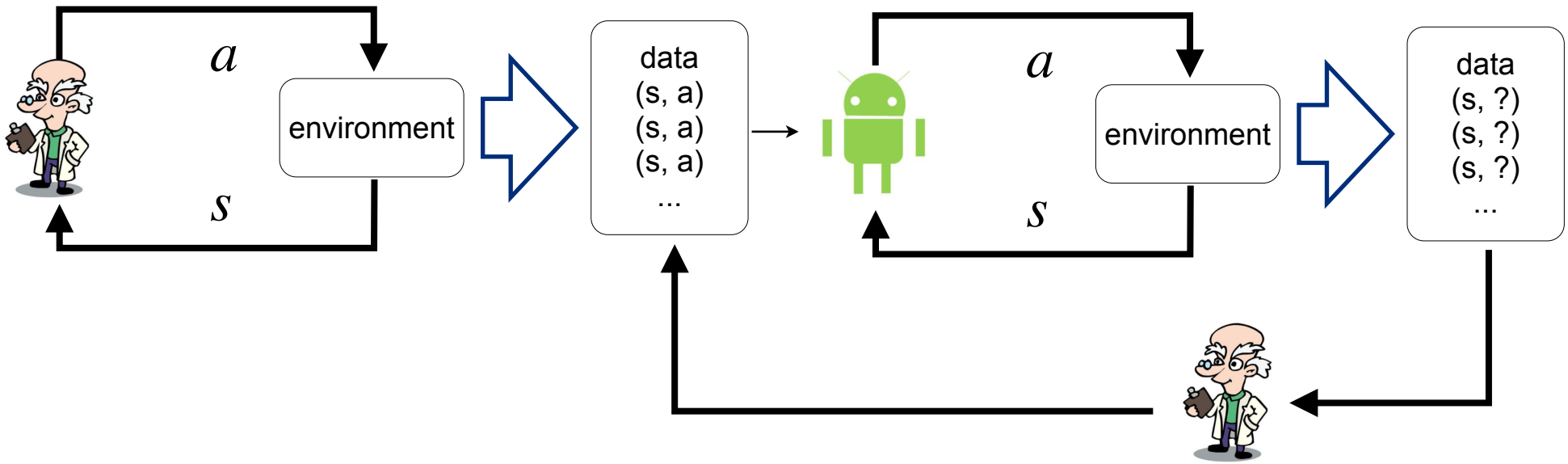
Simple imitation learning



Iteratively collecting more data



Dagger



π^* is the expert

β_i can be 0

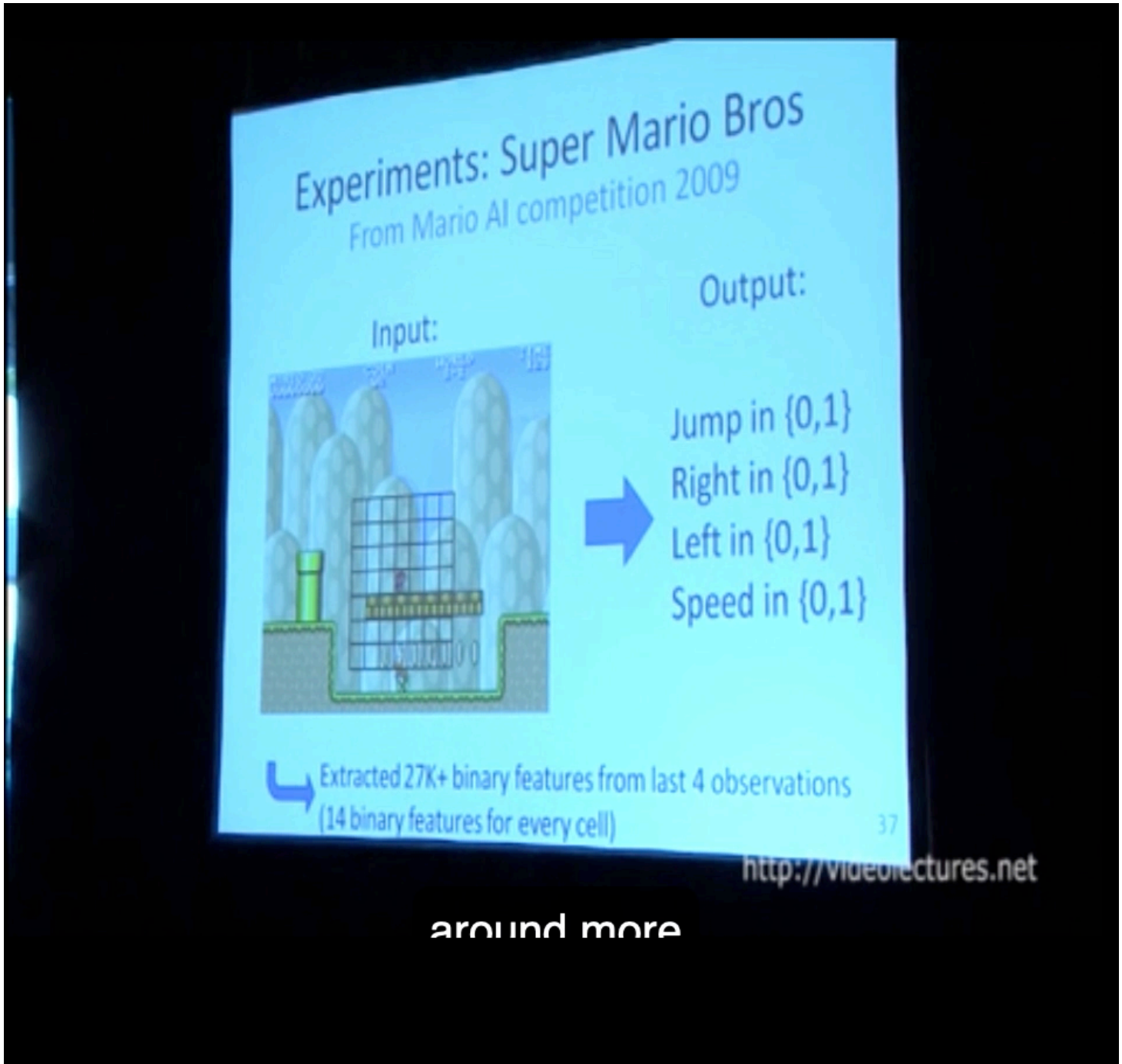
$$\frac{1}{N}\sum_{i=1}^{N}\beta_i \to 0 \text{ as } N \to \infty$$

Initialize $\mathcal{D} \leftarrow \emptyset$.
Initialize $\hat{\pi}_1$ to any policy in Π.
for $i = 1$ **to** N **do**
  Let $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$.
  Sample T-step trajectories using π_i.
  Get dataset $\mathcal{D}_i = \{(s, \pi^*(s))\}$ of visited states by π_i and actions given by expert.
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.
  Train classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$.
end for
Return best $\hat{\pi}_i$ on validation.

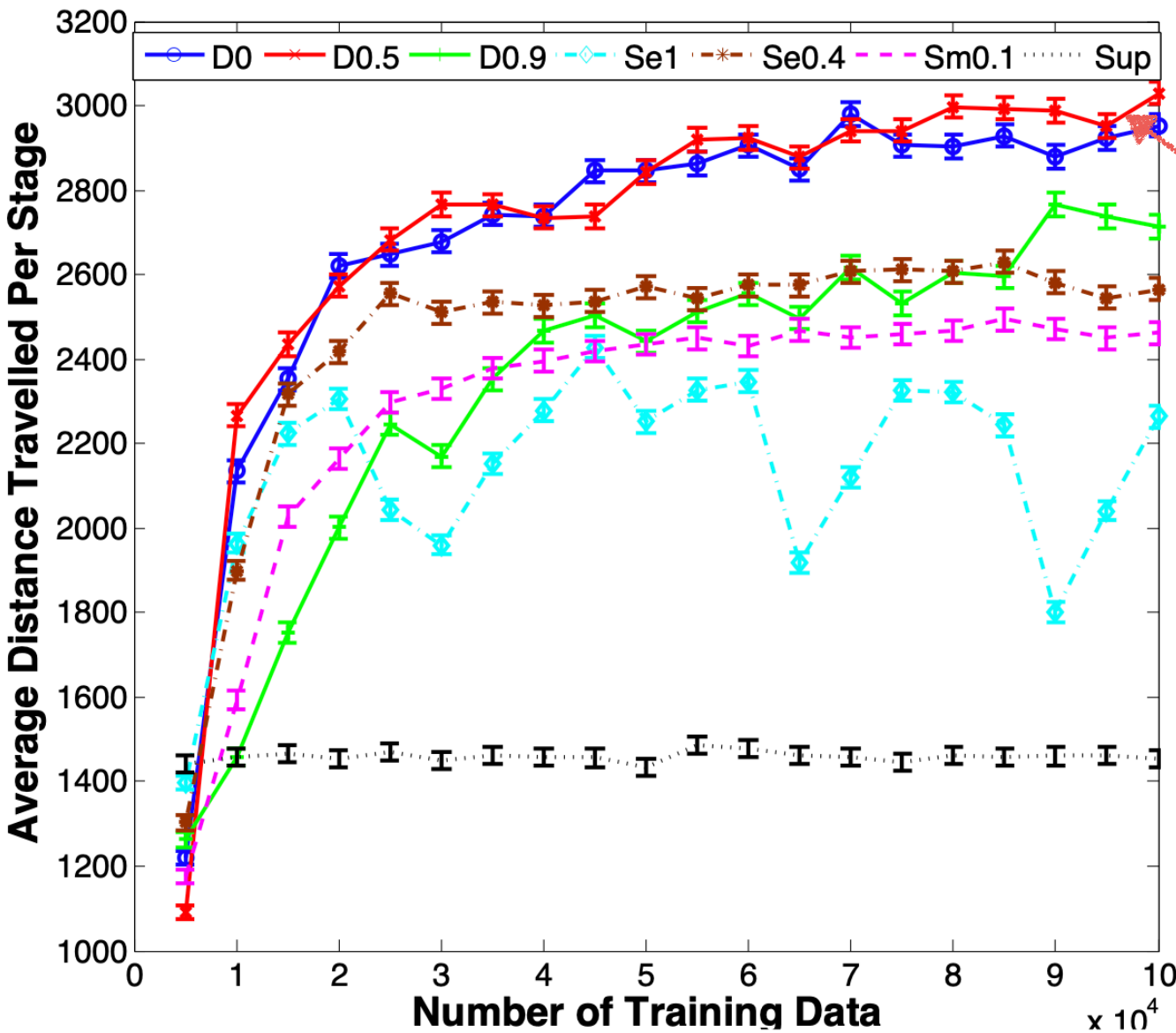
[Stéphane Ross, Geoffrey J. Gordon, Drew Bagnell: A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In: AISTATS 2011, 627-635]

Demo

from http://videolectures.net/aistats2011_ross_reduction/



Dagger



$\beta_0 = 1,\ \beta_{>0} = 0$

$\beta_i = \text{best constant}$

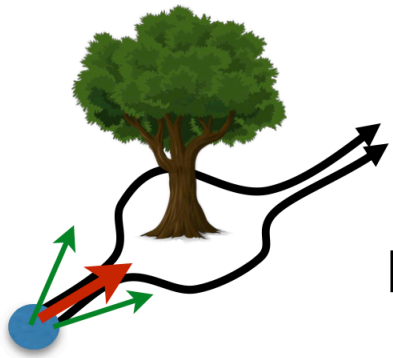
$\beta_i = 0.9$

[Stéphane Ross, Geoffrey J. Gordon, Drew Bagnell: A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In: AISTATS 2011, 627-635]

More issues to think

What if the partial observation?

What if the expert is stochastic?



[from CMU 10-403]

What if the expert is not optimal?