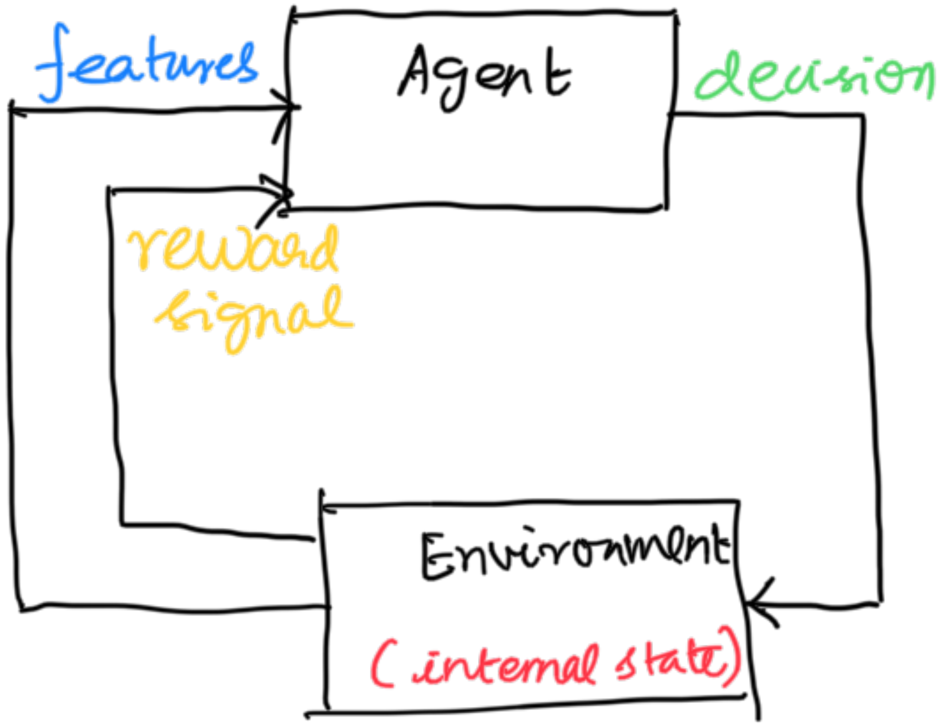
Statistical Decision Theory (Birds eye view)



max (over decision) total reward (Decision Theoretic Goal)

Statistical → probabilistic / frequency of occurrence

Statistical → probabilities / frequency of occurances
and co-occurances

Probability Model (Motivating example)

Say I would like to model an environment / world which has 360 days and 20% days it rains.

$\boxed{\text{Environment (internal state)}}$

1 = rain

0 = no rain

y_t : internal state (rain / no-rain)

Deterministic 1) model

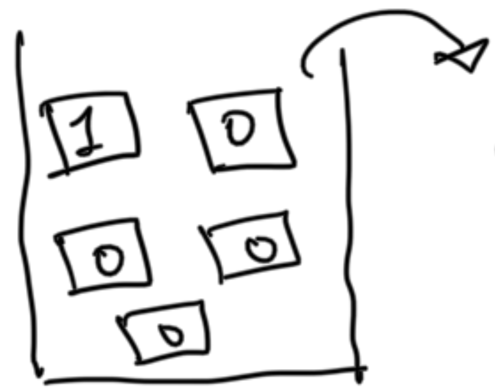
$$y_t = \underbrace{1, 1, \ldots, 1}_{72 \text{ days}}, 0, 0, \ldots, 0$$

↑ day 360

Deterministic Model 2)

$$y_t = 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, \ldots$$

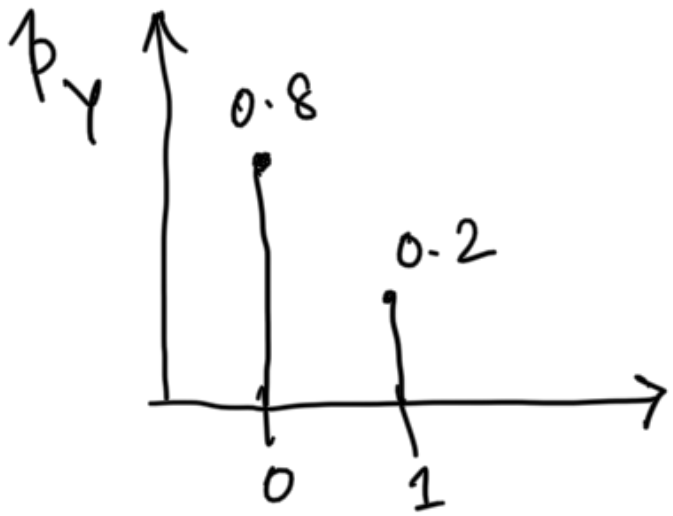
Stochastic (Probabilistic) Model



at time t

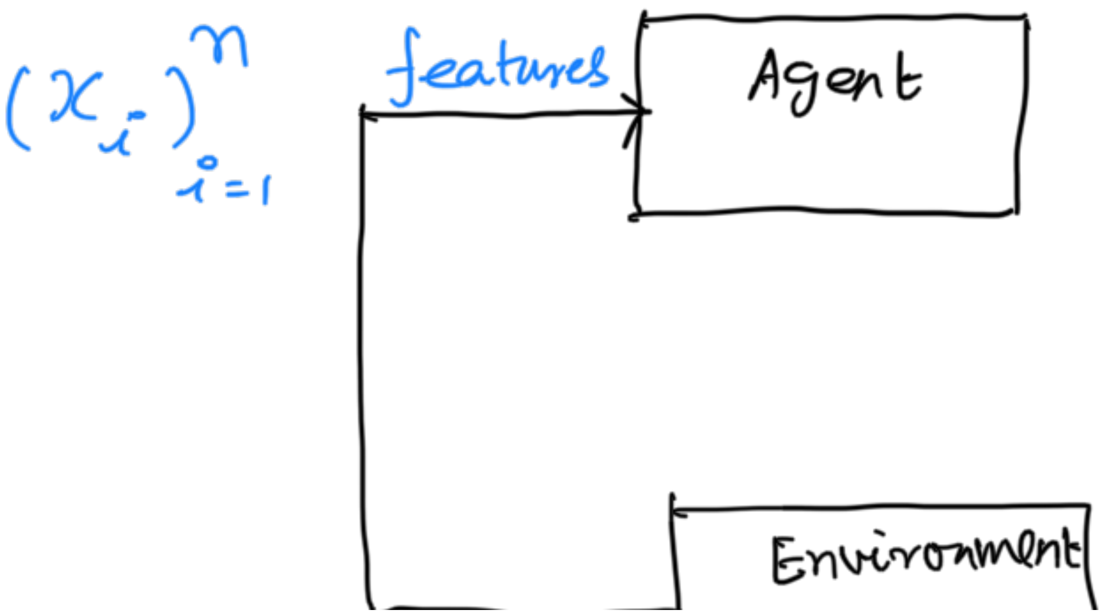
* shuffle well
* pick a chit and read out

$$y_t \overset{i.i.d}{\sim} p_Y(\cdot), \quad \text{where}$$



Descriptive Task

(Data Representation)



(internal state)

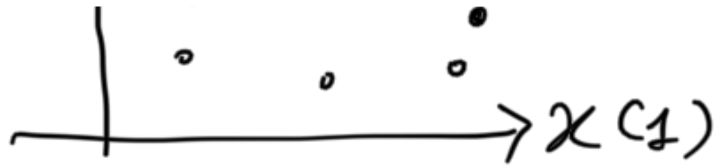
Agent does not make any decision

- Features : $(x_i)_{i=1}^{n} \in \mathbb{R}^d$

Eg 1) clustering : Goal is to say group articles in a newspaper by topic

n: articles, $x_i \in \mathbb{R}^d$ (d = size of vocabulary)

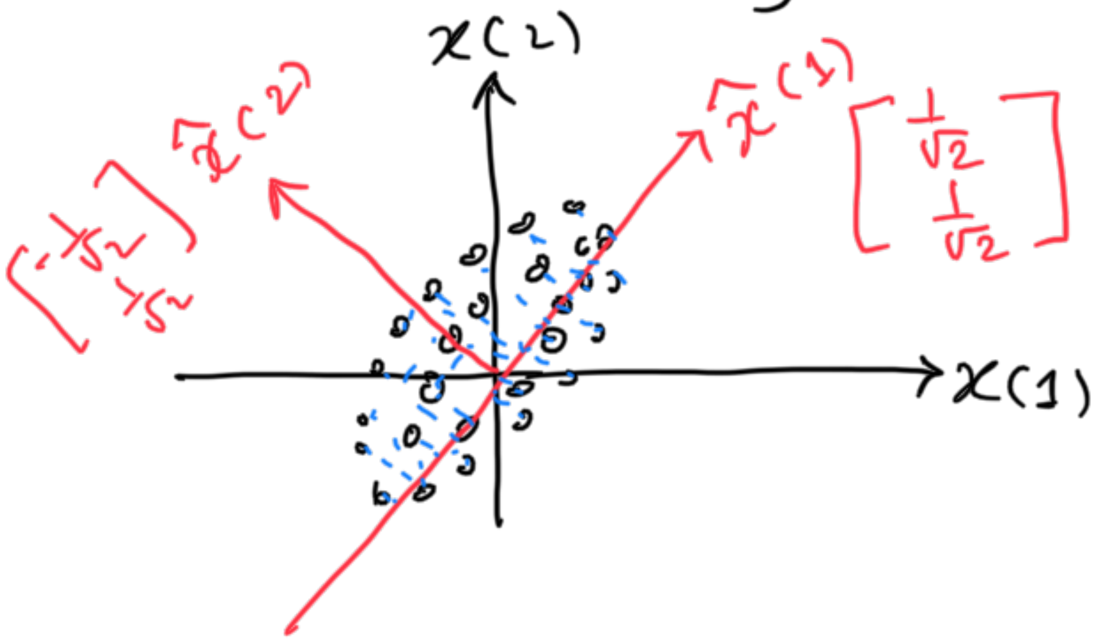
$x_i(j)$ = # word j occured in article i.

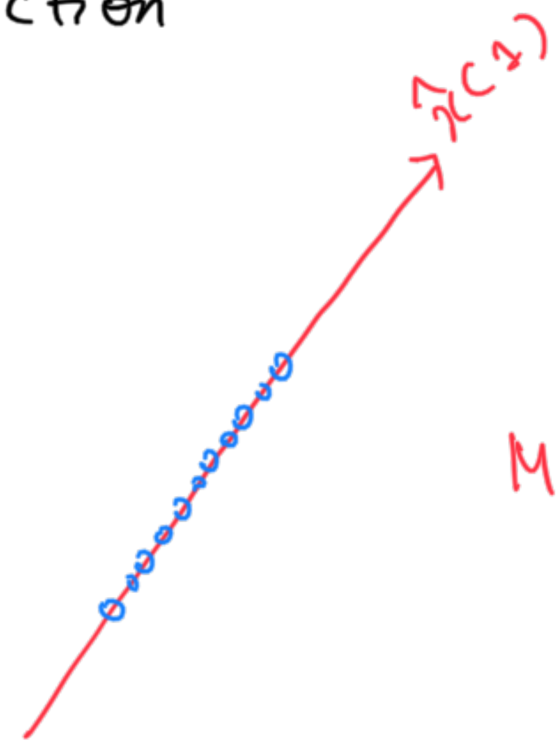


clustering $\quad \mathcal{C} : \{1, \dots, n\} \longrightarrow \{1, \dots, \underset{\substack{\uparrow \\ \text{total clusters}}}{k}\}$

$$x_i \in \text{ cluster } \mathcal{C}(i)$$

Eg (2) Dimensionality Reduction



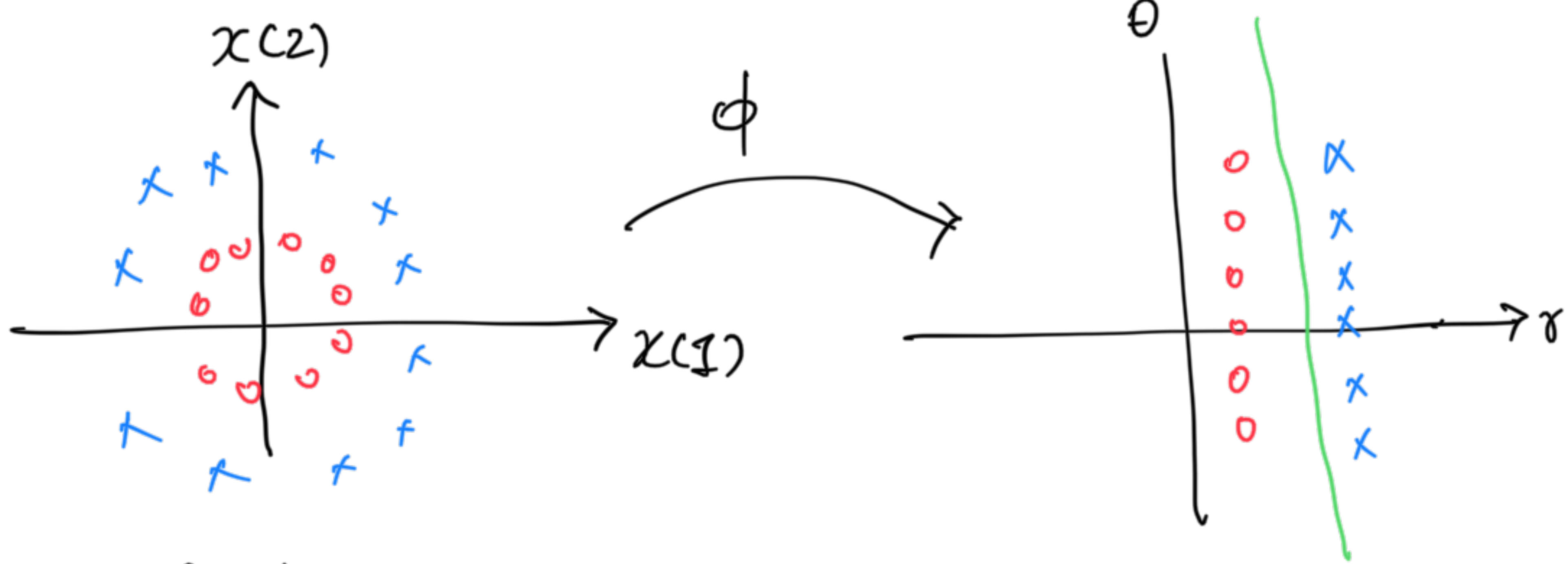


$$M = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \checkmark$$

$$(x_i)_{i=1}^{n} \in \mathbb{R}^d \xrightarrow{M} (\hat{x}_i)_{i=1}^{n} \in \mathbb{R}^{d'} \quad (d' << d)$$

$$\hat{x}_i = M x_i \qquad \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Eg (3) : Representation Learning



Cartesian

Polar

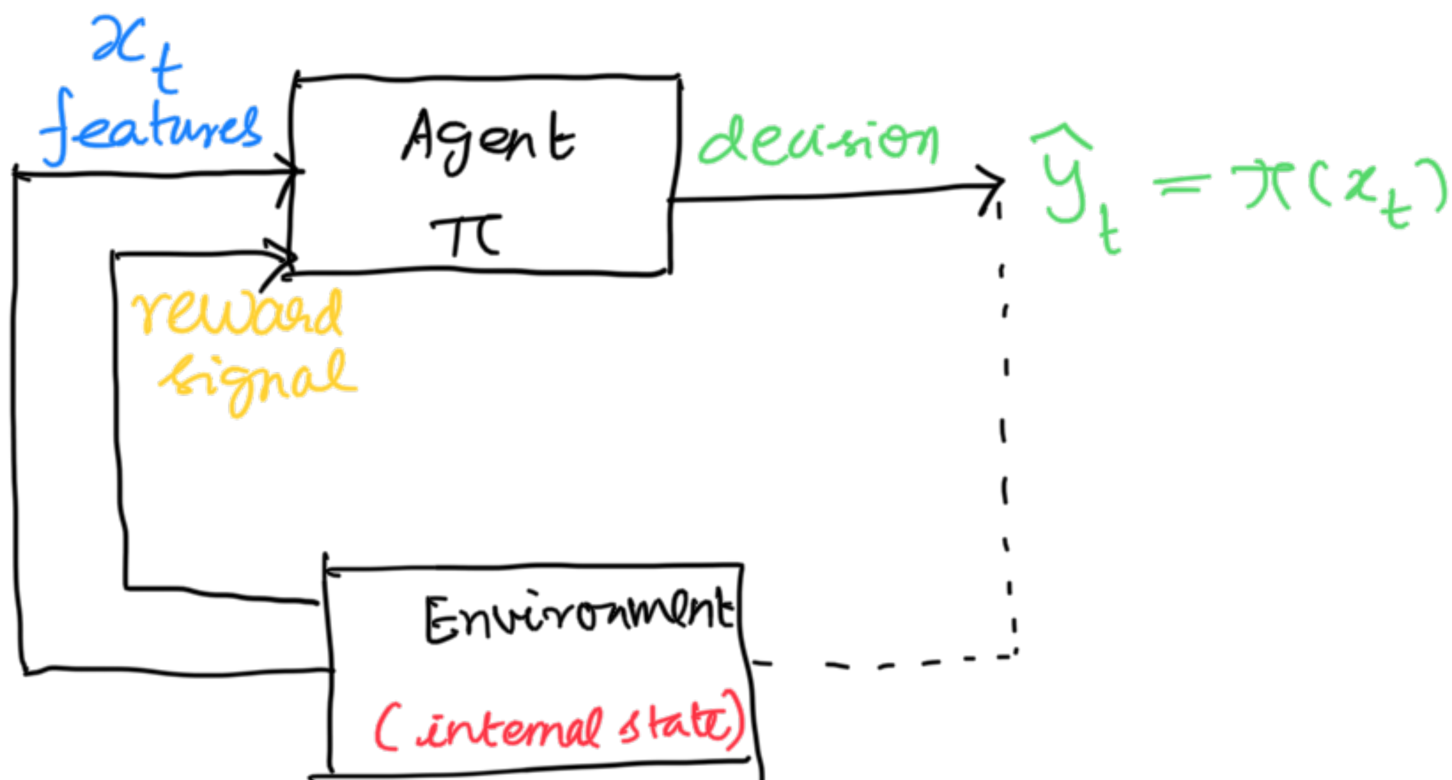
$$\phi : \mathbb{R}^d \to \mathbb{R}^m$$

$$\phi : \mathbb{R} \to \mathbb{R}$$

$$\hat{x}_i = \phi(x_i)$$

Predictive Task

Static Prediction



$y_t \sim p_y(\cdot)$ (distribution of categories)

$x_t \sim p_{X|Y}(\cdot|y_t)$ (distribution of image given category)

(Static Prediction)

Multi-class classification (object classification)

- $y_t \in Y = \{1, \dots, c\}$ (label space) categories

For instance, object classification

$Y = \{$house, person, dog, cat, $\dots$, car, plane $\dots\}$

- $x_t \in X$ (feature space) $\subseteq \mathbb{R}^d$

In object classification x_t is an image

$x_t \in \mathbb{R}^{\#\text{ pixels}}$

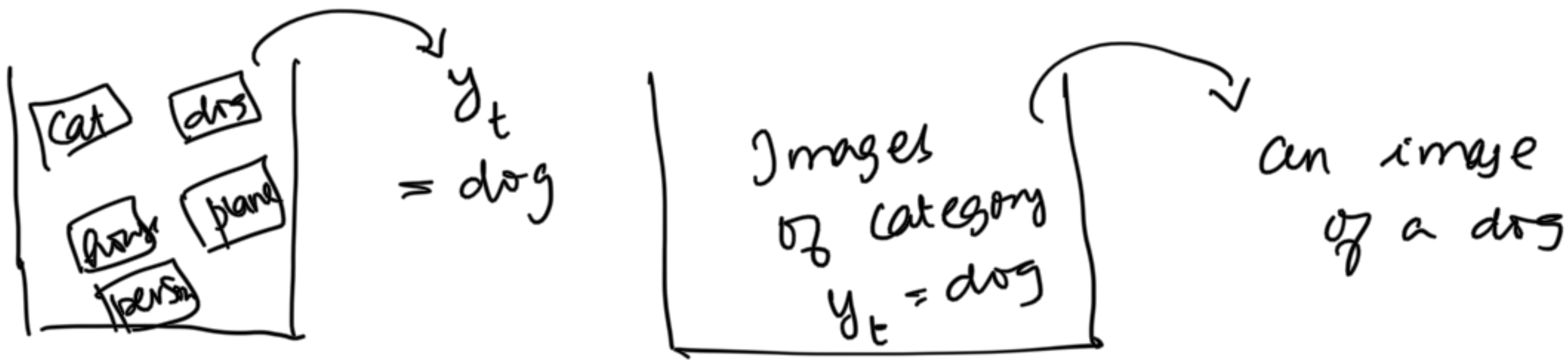
- $\pi : X \rightarrow Y$

- $\ell_t = L(\text{state, decision})$

  $= L(\text{label, predicted label})$

- $L(y_t, \hat{y}_t) = 0, \quad \hat{y}_t = y_t$

  $= 1, \quad \hat{y}_t \neq y_t$



$p_Y(\cdot) \qquad\qquad p_{X|Y}(\cdot \mid y_t = \text{dog})$

Goal: $\min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right]$

Model is known, i.e., p_Y, $p_{X|Y}$

$$\hat{y}_t = \pi_*(x_t) = \underset{y \in Y}{\arg\max}\; p_Y(y) \cdot p_{X|Y}(x_t | y)$$

(Bayesian Decision Theory)

Regression (Static Prediction)

- $X \subseteq \mathbb{R}^d$ (feature)
- $Y \subseteq \mathbb{R}^m$
- $(x_t, y_t) \overset{iid}{\sim} p_{XY}(\cdot, \cdot)$
- $\ell_t = L(y_t, \hat{y}_t)$

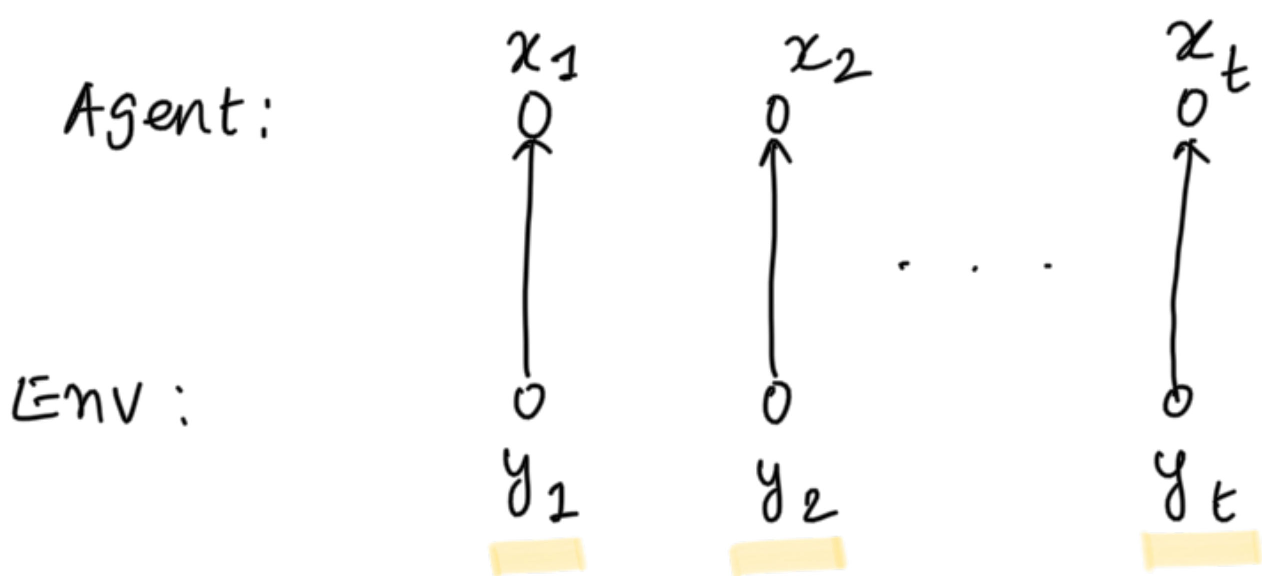
$$L(y, \hat{y}) = \| y - \hat{y} \|^2$$

$$L(y_t, \hat{y}_t) = \| y_t - \hat{y}_t \|_2$$

Goal : $\min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t \right]$

$$\hat{y}_t = \pi_*(x_t) = \mathbb{E}[Y | X = x_t]$$

(Bayesian Decision theory)



$y_t \sim p_Y(\cdot)$

$x_t \sim p_{X|Y}(\cdot | y_t)$

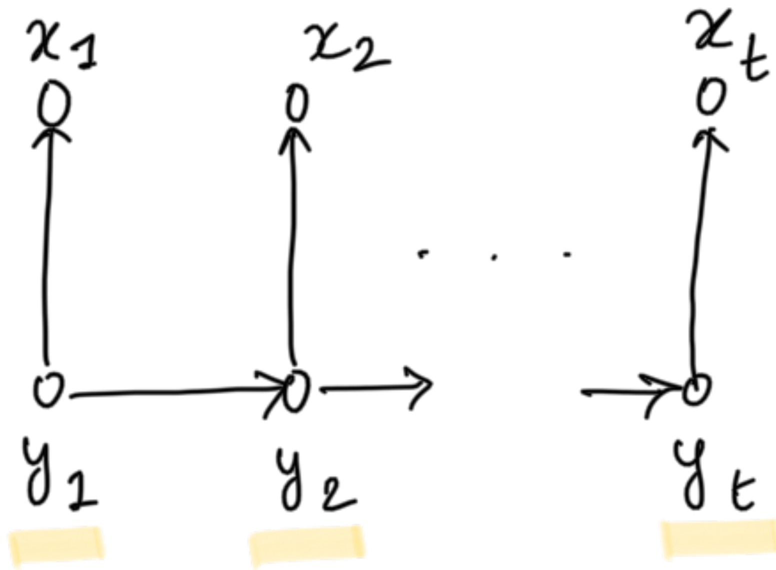
Static Prediction

Dynamic Prediction

Hidden Markov Models (HMM)

(features) Agent:

(internal state) Env:



$$y_t \sim p_Y(\cdot \mid y_{t-1})$$

$$x_t \sim p_{X|Y}(\cdot \mid y_t)$$

Dynamic Prediction

Example: Speech recognition

$$y_t \sim p_Y(\cdot \mid y_{t-1})$$

(Syntax & semantics grammatical rules)

is was of the when

$$p_Y(y_2 = \text{was} \mid y_1 = \text{is}) = 0$$

* Filtering problem:

Given $x_{1:t} = x_1, \dots, x_t$

Predict : $p(y_{t+k} = y)$, $k \geq 0$

(future)

Algo: Forward Algo

* Smoothing Problem

Given $x_{1:t}$

Predict $p(y_k = y)$, $1 \leq k < t$

(past)

Algo: Forward - Backward

* Maximum likelihood sequence

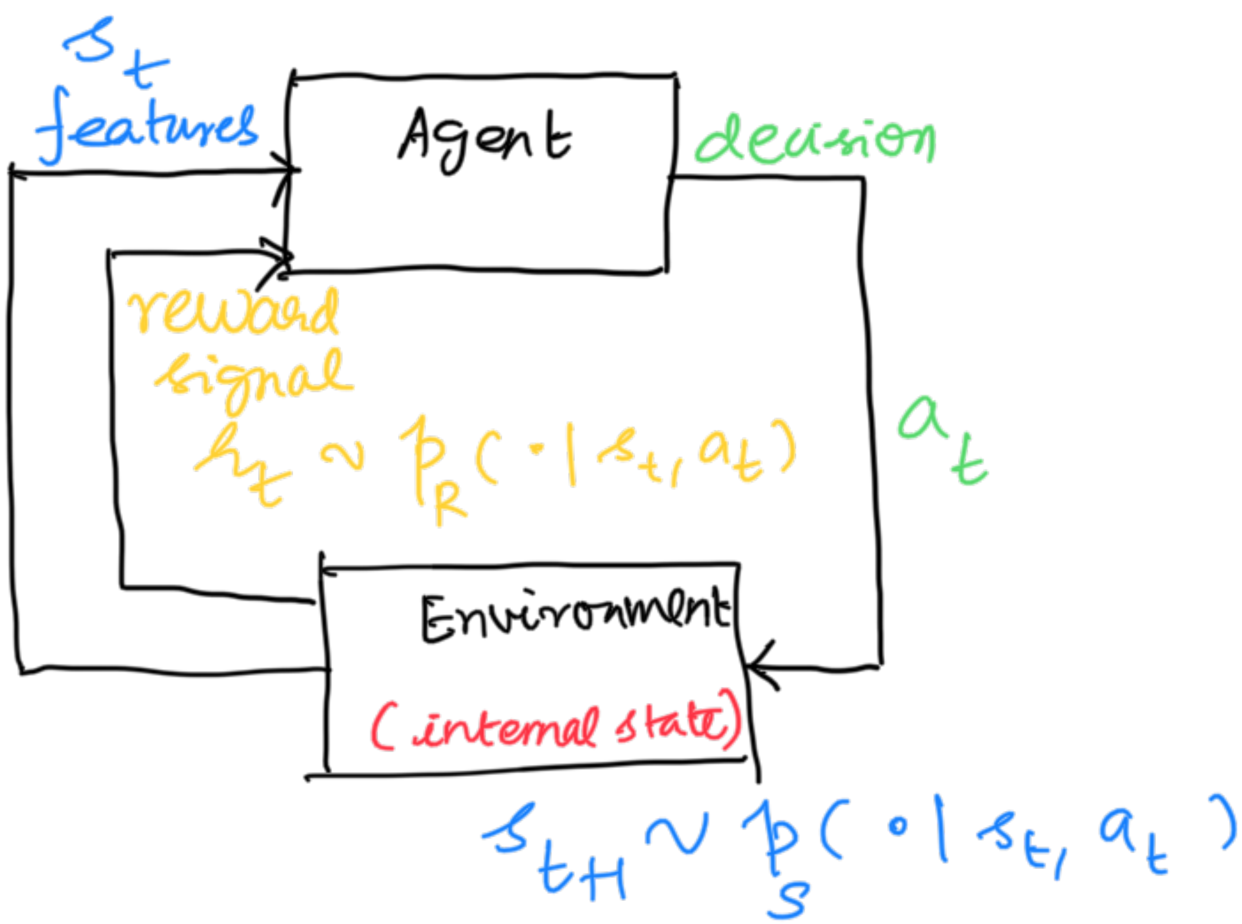
Given $x_{1:t}$

Predict $\operatorname{argmax} p(y_{1:t} \mid x_{1:t})$

(all of past) $y_1 \cdots y_t$

Algo: Viterbi

Dynamic Control



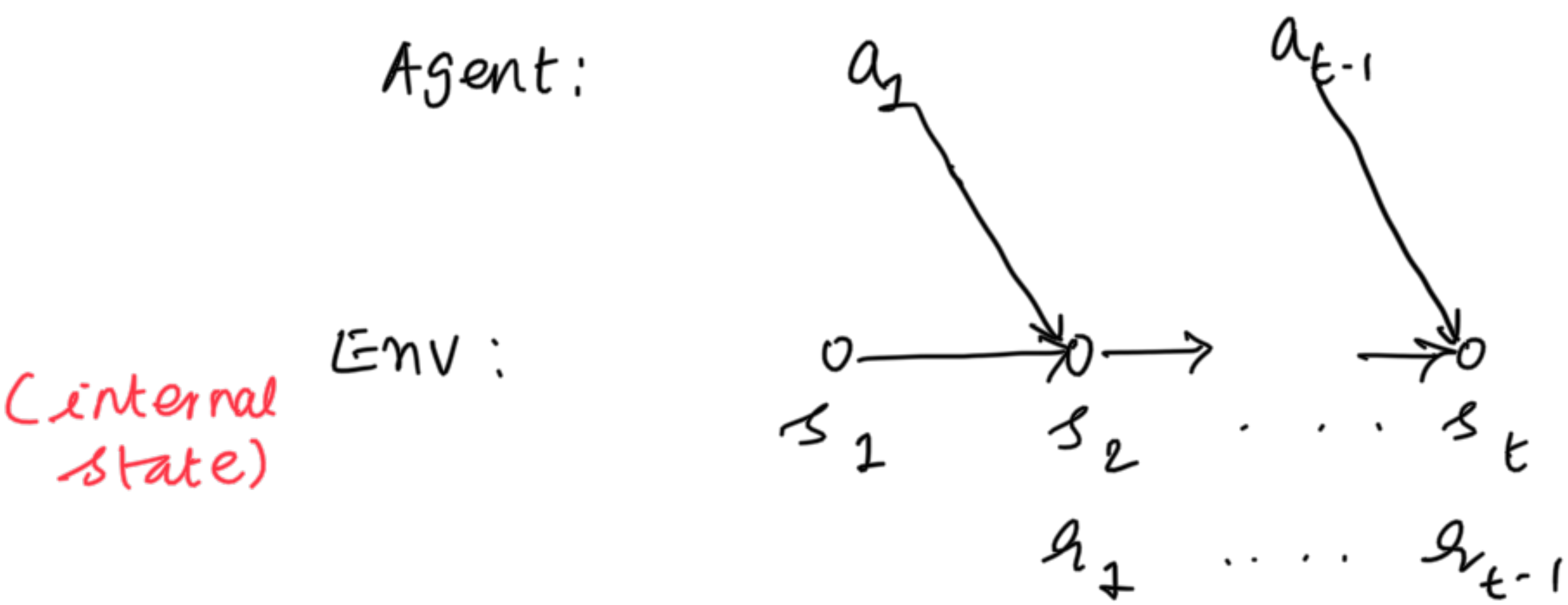
Eg: chess, automatic driving

- $s_t \in S$ (state space)
- $a_t \in A$ (action space)
- r_t is the reward

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$

- $\pi(s_t) = a_t$

Markov Decision Process (MDPs)

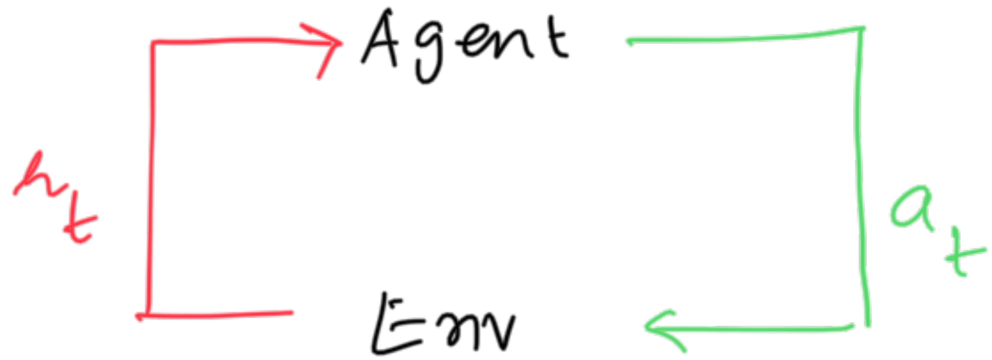


$$s_{t+1} \sim p_S(\cdot \mid s_t, a_t)$$

$$r_t \sim p_R(\cdot \mid s_t, a_t)$$

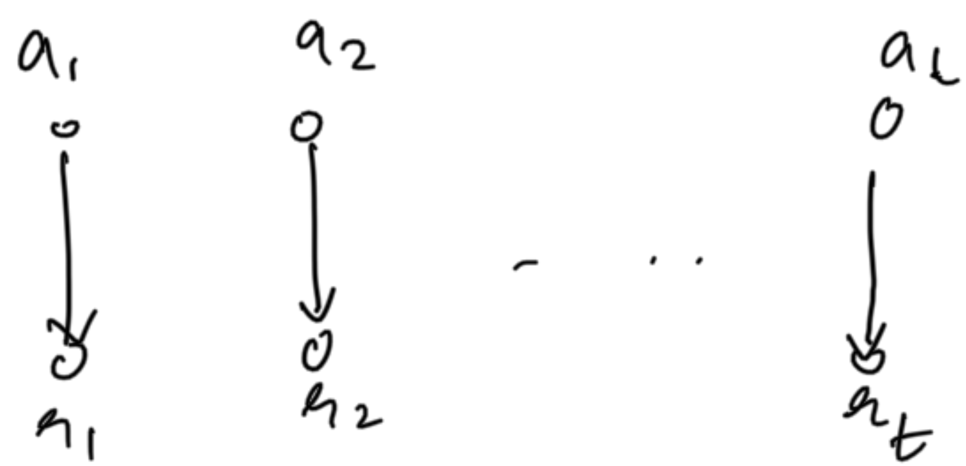
Algo: Value Iteration, Policy Iteration
Linear programming.

Static control



$$r_t \sim p_{a_t}(\cdot)$$

- $a_t \in A = \{1, \ldots, k\}$



$$\max_{a \in A} \mathbb{E}_{p_a}[r_t] \quad *$$

Model is known (*) is trivial

1) compute $\mu(i) = \mathbb{E}_{P_{a_i}}[r_t]$

2) $i_* = \arg\max_i \mu(i)$

Descriptive	Predictive	Control
(No output)	(output + no loop)	full loop

Motivation for Machine Learning

* we dont have model

* have loads of data

* can we do something / can we still solve tasks

Problem	Data	Machine Learning method
Descriptive	$(x_i)_{i=1}^{n}$	unsupervised learning
Predictive	$(x_i, y_i)_{i=1}^{n}$	Supervised learning
	feature label	
Dynamic control	$(s_t, a_t, r_t, s_{t+1})_{t=1}^{T}$	Reinforcement learning
static control	$(a_t, r_t)_{t=1}^{T}$	Multi-Armed Bandit Problem

Where does deep learning stand ?

* curse - of - dimensionality

board positions > # atoms in the universe
in Go