5. Reinforcement learning

Apprentissage :

données

labels

- supervisé : (X_i, Y_i)

  on cherche à prédire le label (la classe...) d'une nouvelle donnée

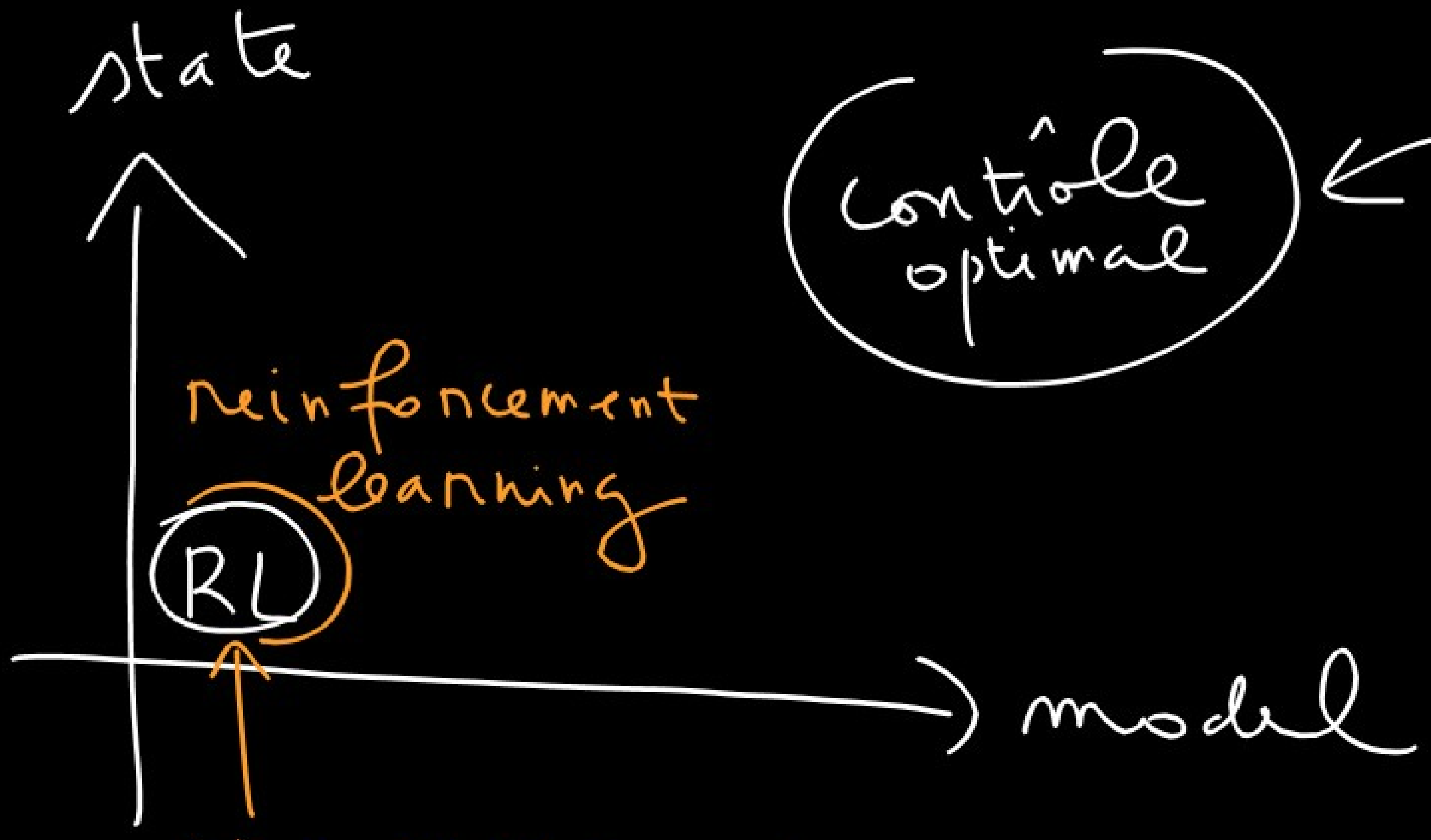
  (SVM...) → classification supervisée

- non-supervisé :

  → clustering : "faire des paquets" (en postulant un nombre de classes) (K-means, spectral clustering...)

- par renforcement

Approximate dynamic learning (ADP) :



contrôle optimal

$$\underbrace{x(t) = f(x(t), u(t))}_{\text{model}}$$

état et modèle connus

ex. : maps avec
- carte partiellement connue (modèle imparfait)
- position (état) imparfaitement mesuré

→ Contrôle optimal stochastique discret :

$$\mathbb{E}\left(\sum_{t=0}^{N-1} R(x_t, u_t)\right) \longrightarrow \max$$

reward

stochastique (e_t : v.a.)

$$x_{t+1} = f(x_t, u_t, e_t), \quad t = 0, 1 \dots \quad \text{(temps discret)}$$

x_0 connu, $u_t \in U$

R, f : modèle inconnu

$$x_t \xrightarrow[u_t]{e_t} x_{t+1}$$

contrôle ("action")

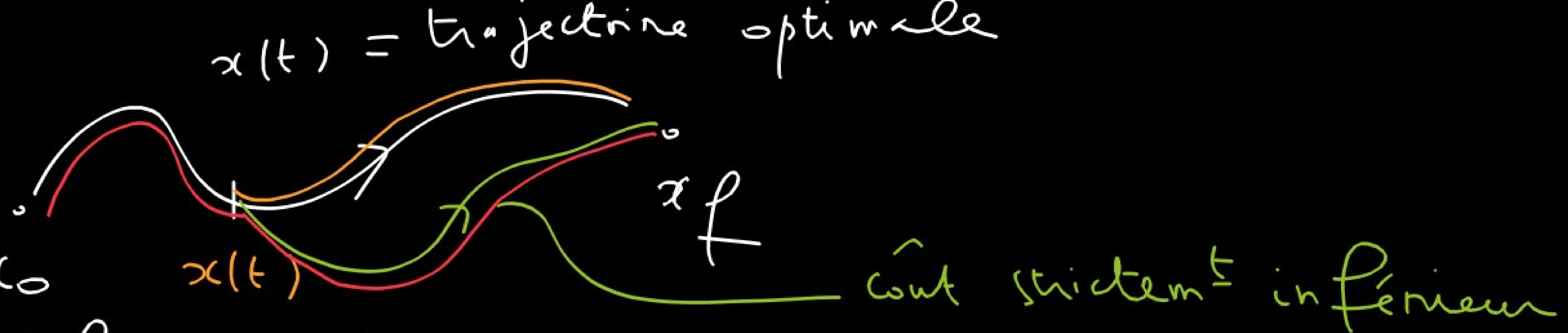
(MDP = Markov Decision Process)

— sur la base de plusieurs épisodes ("réalisations"), on fait des moindres carrés pour apprendre R et f :

$$\left.\begin{array}{l} (\tilde{x}_0, \tilde{x}_1, \dots, \tilde{x}_L) \\ (\tilde{u}_0, \dots, \tilde{u}_{L-1}) \\ (\tilde{r}_0, \dots, \tilde{r}_{L-1}) \end{array}\right\} \text{1 épisode}$$

— autre approche : Q-learning $\Longrightarrow$ on ne cherche pas à apprendre R ou f (= le modèle) mais on utilise le fait que le modèle sous-jacent est du contrôle (optimal stochastique discret)

Th. (Bellman / programmation dynamique) :



$x : [0, t_f] \longrightarrow \mathbb{R}^n$ optimale

$\Rightarrow (\forall t \in [0, t_f]) : x_{|[t, t_f]} : [t, t_f] \longrightarrow \mathbb{R}^n$ est encore optimale

pour relier $x(t)$ à x_f.

<u>dém</u> : cf. ci-dessous. $\square$

Déf. : Q-function : on considère le pb précédent mais à horizon infini :

$$\mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t R(x_t, u_t)\right) \longrightarrow \max$$

$\gamma \in \,]0,1[$ (discount factor)

$x_{t+1} = f(x_t, u_t, e_t)$

x_0 connu, $u_t \in U \subset \mathbb{R}^m$, $t = 0, 1, 2, \ldots$

$$Q(x, u) := \max_{\underbrace{u_1, u_2, \ldots}_{\text{suite}}} \left\{ \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t R(x_t, u_t)\right) \,\middle|\, \begin{array}{l} x_{t+1} = f(x_t, u_t),\ x_0 = x \\ u_0 = u \end{array} \right\}$$

Rq : i) $v(x) = \max_{u \in U} Q(x,u)$: fonction valeur

ii) si on connaît Q, on a accès à l'action (la "politique") optimale à jouer dans l'état x :

$$u \in \arg\max_{v \in U} Q(x,v)$$

si on apprend la fonction Q directement (sans apprendre le modèle R, f...), on sait quelle action optimale prendre dans l'état x.

On :

Bellman

$$Q(x,u) = R(x,u) + E\left(\max_{u', u_2, \dots} \left\{ E\left(\sum_{t=1}^{\infty} \gamma^t R(x_t, u_t)\right) \middle| \begin{array}{l} x_{t+1} = f(x_t, u_t, e_t),\ t = 1,2\dots \\ x_1 = f(x,u,e_0),\ u_1 = u' \end{array} \right\}\right)$$

$$\Longrightarrow Q(x,u) = R(x,u) + \underset{=}{E}\left(\max_{u' \in U} \gamma \cdot Q(f(x,u,\underset{=}{e_0}), u')\right)$$: la fonction est solution de cette équation de pt fixe.

$\longrightarrow$ <u>Apprentissage</u> : $\left\{\begin{array}{l} \tilde{x}_0, \dots, \tilde{x}_L \\ \tilde{u}_0, \dots, \tilde{u}_{L-1} \\ \tilde{r}_0, \dots, \tilde{r}_{L-1} \end{array}\right)$ 1 épisodes $\longrightarrow$ K épisodes

Si $\hat{Q}(x,u,y_k)$ est un estimateur (*) (paramétré par y_k), on l'améliore à chaque nouvel épisode en cherchant y_{k+1} tq :

(*) en pratique $\hat{Q}(x,u,y) := NN_y(x,u)$ (y = poids synaptiques)

$$\hat{Q}(\tilde{x}_t, \tilde{u}_t, y_{k+1}) = (1-\eta_k)\cdot\left(\tilde{r}_t + \max_{u' \in U} \gamma \cdot \hat{Q}(\tilde{x}_{t+1}, u', y_{k+1})\right)$$

$$+ \eta_k \cdot \hat{Q}(\tilde{x}_t, \tilde{u}_t, y_k), \quad t = 0, \dots, L-1$$

Si $(y_k)_k \longrightarrow \bar{y}$, on a CV vers $\hat{Q}(\cdot, \cdot, \bar{y})$ qui vérifie le pt fixe souhaité sur les réalisations.

$$\longrightarrow \quad u \in \arg\max_{v \in U} \hat{Q}(x, v, \bar{y})$$