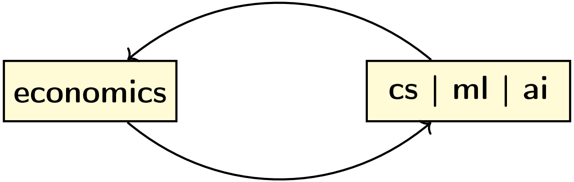
# Reinforcement Learning in Economics and Finance

(a modest state-of-the-art)

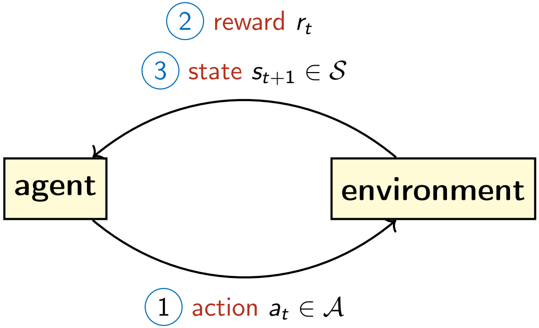
Arthur Charpentier, Romuald Elie & Carl Remlinger

# The authors

- Arthur Charpentier, professor, maths dpt UQAM (Montréal, Canada), previously econ. dpt at Université de Rennes (France). Works on actuarial modeling, insurance & data science (see freakonometrics)
- Romuald Elie, professor, maths dpt Université Gustave Eiffel (Paris, France), visiting UC Berkeley (U.S.).
- Carl Remlinger, PhD student, maths dpt Université Gustave Eiffel (Paris, France).

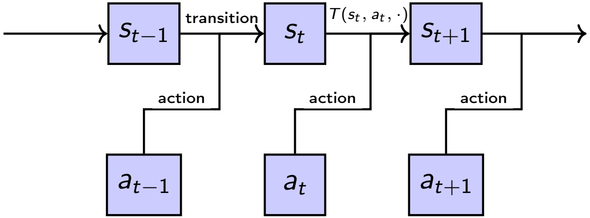


# Reinforcement Learning




- the learner takes an action $a_t \in \mathcal{A}$ (while at state $s_t$)
- the learner obtains a (short-term) reward $r_t \in \mathcal{R}$
- then the state of the world becomes $s_{t+1} \in \mathcal{S}$

# Reinforcement (Sequential) Learning




Let $T$ be a transition function $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ (Markov dynamics) where:

$$\mathbb{P}\left[s_{t+1} = s' \middle| s_t = s, a_t = a, a_{t-1}, a_{t-2}, \dots\right] = T(s, a, s') \quad \text{(see 3)}.$$

A policy is an action, decided at some state of the world.

- either $\pi : \mathcal{S} \to \mathcal{A}$, $\pi(s) \in \mathcal{A}$
- or $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$, i.e. probability to choose action $a \in \mathcal{A}$

Given a policy $\pi$, its expected reward, is

$$V^\pi(s_t) = \mathbb{E}_{\mathbb{P}}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k} \middle| s_t, \pi\right) \text{ where } a \sim \pi(s_t, \cdot)$$

# Machine Learning

Data chunk $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ on $\mathcal{X} \times \mathcal{Y}$.
E.g. classification, $\mathcal{Y} = \{0, 1\}$. With a logistic regression

$$f(x) = \left(1 + e^{-\boldsymbol{x}^\top \boldsymbol{\beta}}\right)^{-1} \in [0, 1] =: \mathcal{A}$$

Given a loss $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}_+$, define regret as

$$R_n = \frac{1}{n} \sum_{i=1}^{n} \ell(\widehat{f}(\boldsymbol{x}_i), y_i) - \underbrace{\inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i) \right\}}_{\text{optimal oracle risk}}$$

As proved in Robbins (1952), minimizing regret $\longleftrightarrow$ maximizing a reward

# Online Learning

Consider a dynamic setting, with sequential data $(\boldsymbol{x}_t, y_t)$
Define regret for some forecasting rule $\widehat{f}_t$

$$R_T = \frac{1}{T}\sum_{t=1}^{T}\ell(\widehat{f}_t(\boldsymbol{x}_t), y_{t+1}) - \inf_{f\in\mathcal{F}}\left\{\frac{1}{T}\sum_{t=1}^{T}\ell(f(\boldsymbol{x}_t), y_{t+1})\right\}$$

Classical model averaging (see online aggregator)
$k$ models providing forecasts ${}_t\widehat{\boldsymbol{y}}_{t+1} = {}_t\widehat{y}^1_{t+1}, \cdots, {}_t\widehat{y}^k_{t+1}$, define ${}_t\widehat{y}^{\boldsymbol{\omega}}_{t+1} = \boldsymbol{\omega}^\top {}_t\widehat{\boldsymbol{y}}_{t+1}$

$$R_T = \frac{1}{T}\sum_{t=1}^{T}\ell(\widehat{y}^*_t, y_t) - \inf_{\omega\in\Omega}\left\{\frac{1}{T}\sum_{t=1}^{T}\ell({}_t\widehat{y}^{\boldsymbol{\omega}}_{t+1}, y_{t+1})\right\}$$

# (multi-armed) Bandits

Pulling arm $k$ yields (random) reward $R_k$, with mean $Q(k)=\mathbb{E}(R_k)$. Optimal policy is

$$a^\star = \underset{a\in\{1,\ldots,K\}}{\text{argmax}}\left\{Q(a)\right\},\text{ with return }Q^\star = Q(a^\star)$$

Consider a sequential game, with $a_{t+1}=f_t(a_t,r_t,\cdots,a_1,r_1)$
The regret of a bandit algorithm is thus:

$$R_T(f) = TQ^\star - \mathbb{E}\left[\sum_{t=1}^T r_t\right] = \underbrace{TQ^\star}_{\text{oracle}} - \mathbb{E}\left[\sum_{t=1}^T Q(a_t)\right]$$

Classical exploration-exploitation tradeoff.
See Rothschild (1974) or Weitzman (1979) for economic applications.

# Reinforcement Learning Framework (1)

Consider the (infinite time horizon) discounted return

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} = r_{t+1} + \gamma G_{t+1}$$

where $0 \leq \gamma \leq 1$ is the discount factor

To quantify the performance of an action, define the $Q$-value on $\mathcal{S} \times \mathcal{A}$:

$$Q^\pi(s_t, a_t) = \mathbb{E}_{\mathbb{P}}\left[G_t \middle| s_t, a_t, \pi\right] \tag{1}$$

In order to maximize the reward, as in bandits, the optimal strategy is characterized by the optimal policies

$$\pi^\star(s_t) = \underset{a \in \mathcal{A}}{\text{argmax}}\left\{Q^\star(s_t, a)\right\}, \quad \text{where } Q^\star(s_t, a_t) = \max_{\pi \in \Pi}\left\{Q^\pi(s_t, a_t)\right\}$$

while $V^\star(s_t) = \max_{a \in \mathcal{A}}\left\{Q^\star(s_t, a)\right\}$ (see Sutton and Barto (1998)).

# Reinforcement Learning Framework (2)

Bellman's equation is here

$$V^\pi(s_t) = \sum_{a\in\mathcal{A}} \pi(a|s)\left(r(s_t,a) + \gamma \sum_{s'\in\mathcal{S}} T(s_t,a,s')V^\pi(s')\right),$$

or, with $n$ states of the world, if $T^\pi_{ij} = \sum_{a\in\mathcal{A}} \pi(a|i)T(i,a,j)$ and $r^\pi_i = \sum_{a\in\mathcal{A}} \pi(a|i)r(i,a)$

$$\begin{bmatrix} V^\pi(1) \\ \vdots \\ V^\pi(n) \end{bmatrix} = \begin{bmatrix} r^\pi_1 \\ \vdots \\ r^\pi_n \end{bmatrix} + \gamma \begin{bmatrix} T^\pi_{11} & \cdots & T^\pi_{1n} \\ \vdots & \ddots & \vdots \\ T^\pi_{n1} & \cdots & T^\pi_{nn} \end{bmatrix} \begin{bmatrix} V^\pi(1) \\ \vdots \\ V^\pi(n) \end{bmatrix}$$

i.e. $\boldsymbol{V}^\pi = \boldsymbol{r}^\pi + \gamma \boldsymbol{T}^\pi \boldsymbol{V}^\pi = \mathcal{T}_\pi(\boldsymbol{V}^\pi)$, using Bellman's operator $\mathcal{T}_\pi$
(then use the contraction mapping theorem, see Denardo (1967)).

# Inventory problem (Hellwig (1973))

Action $a_t \in \mathcal{A} = \{0, 1, 2, \ldots, m\}$ denote the number of ordered items arriving on the morning of day $t$, puchased at individual prices $\underline{p}$.

States $s_t = \mathcal{S} = \{0, 1, 2, \ldots, m\}$ are the number of items available at the end of the day (before ordering new items for the next day).

Then, state dynamics is

$$s_{t+1} = \big(\min\{(s_t + a_t), m\} - \varepsilon_t\big)_+$$

where $\varepsilon_t$ is the unpredictable demand, independent and identically distributed variables $(s_t)$ is a Markov chain,

$$T(s, a, s') = \mathbb{P}\big[s_{t+1} = s' \big| s_t = s, a_t = a\big] = \mathbb{P}\big[\varepsilon_t = \big(\min\{(s + a), m\} - s'\big)_+\big]$$

The reward function $R$ is such that, on day $t$, revenue made is

$$r_t = -\underline{p}a_t + \overline{p}\varepsilon_t = -\underline{p}a_t + \overline{p}\big(\min\{(s_t + a_t), m\} - s_{t+1}\big)_+ = R(s_t, a_t, s_{t+1})$$

where $\overline{p}$ is the price when items are sold to consumers (and $\underline{p}$ is the price when items are purchased).

# Econ: Consumption & Income Dynamics

Consider an infinitely living agent, with utility $u(c_t)$ when consuming $c_t \geq 0$ in period $t$. The agent receives random income $y_t$ at time $t$, and assume that $(y_t)$ is a Markov process, $T(s, s') = \mathbb{P}[y_{t+1} = s' | y_t = s]$.
Let $w_t$ denote the wealth of the agent, at time $t$, so that $w_{t+1} = w_t + y_t - c_t$. Assume that the wealth must be non-negative, so $c_t \leq w_t + y_t$. And for convenience, $w_0 = 0$, as in Lettau and Uhlig (1999). At time $t$, given state $s_t = (w_t, y_t)$, we seek $c_t^\star$ solution of

$$v(w_t, y_t) = \max_{c \in [0, w_t + y_t]} \left\{ u(c) + \gamma \sum_{y'} \left[ v(w_t + y_t - c, y') \right] T(y_t, y') \right\}$$

This is a standard recursive consumption model, discussed in Ljungqvist and Sargent (2018) or Hansen and Sargent (2013)

# Econ: Bounded Rationality & Experiments

With adaptative learning, Marcet and Sargent (1989a,b) proved that there was convergence to a rational expectations equilibrium.
Leimar and McNamara (2019) suggested that adaptive and reinforcement learning leads to bounded rationality, while Abel (2019) motivates reinforcement learning as a suitable formalism for studying bounded rational agents, since "*at a high level, Reinforcement Learning unifies learning and decision making into a single, general framework*".

Thompson (1933) introduced this idea of adaptive treatment assignment.
Weber (1992) proved that this problem can be expressed using multi-armed bandits, and the optimal solution to this bandit problem is to choose the arm with the to the highest Gittins index, that can be related to the so-called Thompson sampling strategy, intensively used for AB Testing and experimental economics, see Chattopadhyay and Duflo (2004).

# Operation Research & Stochastic Games

Maskin and Tirole (1988) introduced the concept of *Markov perfect equilibrium*,

*Econometrica*, Vol. 56, No. 3 (May, 1988), 571–599

**A THEORY OF DYNAMIC OLIGOPOLY, II: PRICE COMPETITION, KINKED DEMAND CURVES, AND EDGEWORTH CYCLES**

BY ERIC MASKIN AND JEAN TIROLE[1]

Brown (1951) suggested that firms could form beliefs about competitors' choice probabilities, using some fictitious plays, also called Cournot learning (studied more deeply in Hopkins (2002)).
Bernheim (1984) and Pearce (1984) added assumptions on firms beliefs, called rationalizability, under which we can end-up with Nash equilibria.

# Finance

- Market Micro-Structure - see order book dynamics as aggregation of other traders actions ( buy or sell orders) - see Vyetrenko and Xu (2019)
- Portfolio Allocation - see Li and Hoi (2014)
- Risk Management, with realistic market frictions or imperfections.

But finance is related to risk measures...

Discounted return $G_t$ is a random variable, function of $s_t$, $a_t$ and $\pi$, that can be denoted $\Phi^\pi(s_t, a_t)$. We defined

$$Q^\pi(s_t, a_t) = \mathbb{E}_{\mathbb{P}}\left[\Phi^\pi(s_t, a_t)\middle| s_t, a_t, \pi\right]$$

but we can consider another functional of the distribution $\Phi^\pi(s_t, a_t)$, see Distributional Reinforcement Learning, Bellemare et al. (2017).

# Wrap-Up

- Intensive recent literature related to Reinforcement Learning (CS, AI, ML)
- Focus on algorithms
  ($Q$-learning – Watkins and Dayan (1992) – $TD(\lambda)$, deep RL, etc)
- Connexions with many applications in economics and finance
  dynamic programming, operation research, stochastic games, risk measures, etc.
- Several recent extensions
  inverse reinforcement learning (Miller (1984), Pakes (1986))
  distributional reinforcement learning
- see arXiv:2003.10014 for more details, and references

# References I


Abel, D. (2019). *Concepts in Bounded Rationality: Perspectives from Reinforcement Learning*. PhD thesis, Brown University.

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. *arXiv:1707.06887*.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, 52(4):1007–1028.

Brown, G. W. (1951). Iterative solutions of games by fictitious play. In Koopmans, T., editor, *Activity Analysis of Production and Allocation*, pages 374–376. John Wiley & Sons, Inc.

Chattopadhyay, R. and Duflo, E. (2004). Women as policy makers: Evidence from a randomized policy experiment in india. *Econometrica*, 72(5):1409–1443.

Denardo, E. V. (1967). Contraction mappings in the theory underlying dynamic programming. *SIAM Review*, 9(2):165–177.

Hansen, L. P. and Sargent, T. J. (2013). *Recursive Models of Dynamic Linear Economies*. The Gorman Lectures in Economics. Princeton University Press.

Hellwig, M. F. (1973). *Sequential models in economic dynamics*. PhD thesis, Massachusetts Institute of Technology, Department of Economics.

# References II


Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica*, 70(6):2141–2166.

Leimar, O. and McNamara, J. (2019). Learning leads to bounded rationality and the evolution of cognitive bias in public goods games. *Nature Scientific Reports*, 9:16319.

Lettau, M. and Uhlig, H. (1999). Rules of thumb versus dynamic programming. *American Economic Review*, 89(1):148–174.

Li, B. and Hoi, S. C. (2014). Online portfolio selection: A survey. *ACM Computing Surveys (CSUR)*, 46(3):1–36.

Ljungqvist, L. and Sargent, T. J. (2018). *Recursive Macroeconomic Theory*. MIT Press, 4 edition.

Marcet, A. and Sargent, T. J. (1989a). Convergence of least-squares learning in environments with hidden state variables and private information. *Journal of Political Economy*, 97(6):1306–1322.

Marcet, A. and Sargent, T. J. (1989b). Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory*, 48(2):337 – 368.

Maskin, E. and Tirole, J. (1988). A theory of dynamic oligopoly, I: Overview and quantity competition with large fixed costs. *Econometrica*, 56:549–569.

Miller, R. A. (1984). Job matching and occupational choice. *Journal of Political Economy*, 92(6):1086–1120.

# References III


Pakes, A. (1986). Patents as options: Some estimates of the value of holding european patent stocks. *Econometrica*, 54(4):755–784.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):1029–1050.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Rothschild, M. (1974). A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185 – 202.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIP Press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Vyetrenko, S. and Xu, S. (2019). Risk-sensitive compact decision trees for autonomous execution in presence of simulated market response. *arXiv preprint arXiv:1906.02312*.

Watkins, C. J. C. H. and Dayan, P. (1992). $q$-learning. *Machine Learning*, 8(3):279–292.

Weber, R. (1992). On the gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033.

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica*, 47(3):641–654.