# Reinforcement learning in finance: online portfolio allocation and policy gradient approaches

Gabriel Turinici, P. Brugiere

CEREMADE, Université Paris Dauphine - PSL

NAMMAT Conference
Cluj, Nov $9^{th}$, 2023

## Executive summary

Reinforcement learning (RL) algorithms have been used very successfully to find good strategies based on available information.

However few works investigated applications in finance, especially in online portfolio allocation.

Exploiting "flow gradient"-type techniques we discuss a more formal setting for the implicit policy gradient schemes.

The procedure is further adapted to take into account transaction fees.

## Disclamer

What follows is a scientific presentation and not an invitation to use one approach or another in a professional or personal framework, the reader is encouraged to use her/his common sense and critical views.
In particular past performances does not guarantee future performance.
Moreover, the result can depend on hyper-parameters and their robustness should be investigated in practice.

# Outline

1. Reinforcement learning
   - Basic examples
   - Policy gradient approaches : the multi-armed bandit
2. Reinforcement learning in finance
3. Online flow portfolio algorithm

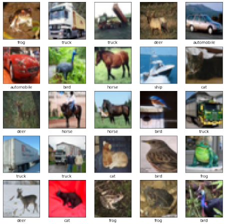
# Reminders : types of "learning"

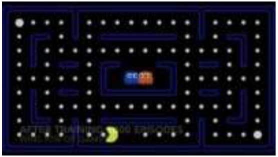
• Supervised learning : e.g. classification: the labels are given i.e. we know the value function;

• Unsupervised learning : e.g. generative : no labels, only an objective e.g. clustering or generate objects similar to a given set

• Reinforcement learning : e.g. game play : based on the interaction with the environment; any action executed within an environment; a signal is received that indicates whether the action has been positive or negative. The good actions are **reinforced** encouraged and bad actions are "punished"; note that in the beginning good/bad is not always defined (e.g. 0.5 is good ?)

# Reminders : types of "learning"




**Left** : supervised learning e.g. classification, e.g. CIFAR10/100 labels. (source: Tensorflow); **Middle :** generative learning from Midjourney (source wikipedia, sept 2023 https://en.wikipedia.org/wiki/Generative_artificial_intelligence ) ; **Right :** reinforcement learning, credits : https://www.youtube.com/watch?v=QilHGSYbjDQ and https://www.youtube.com/watch?v=VMp6pq6_QjI.

- We will focus on reinforcement learning.

# Multi-armed bandit

- $k$-armed bandit : has $k$ options to choose from
- the problem is to allocate limited resources (time, money, turns etc.) among terms of a given list. Goal is to maximize expected rewards. Other situations: choice among medical treatments, for a series of patients
- rewards information: each action '$a$' has a random reward $q(a)$ with a fixed but unknown mean $q_\star(a)$; the means = "values" of the arms.
- Notations $t$ : turn or time; $R_t$ : reward at step $t$ (random variable), $A_t$ : action at step $t$, $\mathcal{A}$ : set of possible actions
- Name: from slot-machines (one-armed bandit); example of goal maximize return over $n = 1000$ steps.

References : [1, 2] etc.

# Multi-armed bandit : (policy) gradient algorithms

Choice of arm: probability law $\pi_t$; auxiliary variables $H_t$,
$\pi_t = \mathrm{s}oftmax(H_t) : P(A_t = a) = \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} =: \pi_t(a)$

• Perspective: stochastic optimization approach (e.g. like Stochastic Gradient Descent [3]) to maximize the expected reward $\mathcal{R} = \mathbb{E}[R_t] = \sum_b q_*(b)\pi_t(b)$ w/r to $H_t$ which define $\pi_t$.

• softmax derivation rule : $\nabla_{H_t(a)}\pi_t(b) = \pi_t(b)(\mathbb{1}_{b=a} - \pi_t(a))$

• Recall: SGD uses a non-biased version of the gradient, possibly involving some random variable here $A_t$

• Final update formula $H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t)(\mathbb{1}_{a=A_t} - \pi_t(a))$ as expected.

• $\alpha =$ "learning rate" to be set, may be difficult to fit

# Outline

# Reinforcement learning in finance

• portfolio optimization : choose / combine $K$ assets

• when statistics of assets performance (mean, covariance) is known, classical Markovitz portfolio theory gives complete answers in the case of quadratic utility functions

• in general such statistics are unknown and misspecification has huge impact on the result; statistics has to be learned on-the-fly = online portfolio selection, cf. review [4].

• other relevant literature : Cover "universal portfolio" (UP) [5], OLMAR algorithm (under mean reverting hypothesis) [6], multiplicative updates style of Helmbold et al. [7], ...

# Reinforcement learning in finance

• notation : $\pi_t(k)$ = proportion of wealth allocated to asset $k$ at time $t$; $\sum_k \pi_t(k) = 1$, $\pi_t(k) \geq 0\ \forall k$.

• notation : price relatives factors: price of asset $k$ multiplies by $f_t(k)$ when advancing from time $t$ to time $t+1$); $f_t(k) - 1$ is also known as the 'return' over the interval $[t, t+1]$.

• total portfolio value change from $t$ to $t+1$ : $w \to w \sum_k \pi_t(k) f_t(k)$

• total wealth $w_{t+1}$ at time $t$ (assuming $w_0 = 1$): $w_t = \prod_{k=0}^{t-1} \langle \pi_k, f_k \rangle$ ... to be maximized

# Gradient flows: theory

• $F : \mathbb{R}^d \to \mathbb{R}$ = a smooth convex function, $\bar{x} \in \mathbb{R}^d$; gradient flow from $\bar{x}$ = a curve $(x_t)_{t\geq 0}$: $x_t' = -\nabla F(x_t)$ for $t > 0$, $x_0 = \bar{x}$.
• Polish metric space $(\mathcal{X}, d)$, functional $F : (\mathcal{X}, d) \to \mathbb{R} \cup \{+\infty\}$: non-trivial definition, huge literature (cf. books by Ambrosio et al. , Villani, Santambroggio) [8, 9]...
$\mathcal{X} = \mathcal{P}_2(\mathbb{R})$ (the set of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with finite second-order moment, endowed with the Wasserstein distance $\mathcal{W}_2$)

**Gradient flows: the JKO scheme**

• Jordan, Kinderlehrer and Otto '98, (JKO) numerical scheme: time step $= \tau > 0$, $x_0^\tau = \bar{x} \in \mathcal{X}$, by recurrence $x_{n+1}^\tau$ = a minimizer of the functional

$$x \mapsto P_F^{JKO}(x; x_n^\tau, \tau) := \frac{1}{2\tau} d^2(x_n^\tau, x) + F(x). \tag{1}$$

• If $\mathcal{X}$= Hilbert, $F$= smooth, JKO = implicit Euler (IE) scheme, i.e., $\frac{x_{n+1}^\tau - x_n^\tau}{\tau} = -\nabla F(x_{n+1}^\tau)$.
• JKO scheme was initially used theoretically to prove the existence of a gradient flow (see [10, 11] for higher order schemes).

# Outline

# OFPA : online flow portfolio algorithm (+ P. Brugiere)

- Financial application : portfolio optimization : maximize final wealth $w_T$

- $\pi$ is a distribution ( $\in$ probability simplex), "softmax" representation
$\pi(a) = softmax(H) = \frac{e^{H(a)}}{\sum_b e^{H(b)}} =: \mathcal{S}(H)$

IMPLICIT NOTATION $\pi = \mathcal{S}(H)$, $\pi_t = \mathcal{S}(H_t)$, etc.

- to maximize $\log(w_t) = \sum_{k=0}^{t-1} \log(\langle \pi_k, f_k \rangle)$, at each time step maximize $F_k(H) = \log(\langle \mathcal{S}(H), f_k \rangle)$.

- in finance one also includes some risk measures, and the reward will rather be $g(\log(\langle \mathcal{S}(H), f_k \rangle))$, e.g., $g(y) = y - \lambda y^2$ (risk $\sim$ volatility), $g(y) = y - \lambda (y_-)^2$ (risk $\sim$ drawdown),...

# OFPA : online flow portfolio algorithm (+ P. Brugiere)

- at each time step maximize $F_t(H) = \log(\langle \mathcal{S}(H), f_t \rangle)$.

- JKO approach, i.e. minimization : $H_{t+1} = \arg\min \frac{d(H,H_t)^2}{2\tau} - F_t(H)$

magenta = specific to our algo
- asset price change $f_t$ induces a drift in $\pi_t$ ! New allocation that takes into account the prices at time $t+1$ : $\pi_{t+} = \frac{\pi_t \odot f_t}{\langle \pi_t \odot f_t, \mathbb{1} \rangle}$, $\odot$ = element-wise (Hadamard) product. Can obtain $H_{t+}$ from $\pi_{t+}$ (explicit formula).

- what about the distance $d(H, H_{t+})^2$ ? Use : makes $H_{t+1}$ close to $H_t$. In [7] they use relative entropy $KL(\pi_{t+1}||\pi_t)$ approximated to first order.
- $\xi$ = multiplicative transaction costs coefficient; distance related to the transaction costs : $\xi \sum_k |\pi(k) - \pi_{t+}(k)|$
- replace $|x-y|$ by $\sqrt{(x-y)^2 + a^2} - a$, $a > 0$ small ... cf. the Huber-energy distance [12, 13].

# OFPA : online flow portfolio algorithm (+ P. Brugiere)

$F_t(H) = \log(\langle \mathcal{S}(H), f_t \rangle)$. $G_t(H) := \xi \sum_k \sqrt{[\mathcal{S}(H)(k) - \pi_{t+}(k)]^2 + a^2} - a$

- minimize $G_t(H) - F_t(H) + dist(...)^2/2\tau$ thus $H_{t+1} \simeq$ solution of $\nabla_H(G_t - F_t + dist(...)^2/2\tau) = 0$:

- Explicit (first order) approximation ok for small $\tau$, unstable otherwise, but what is "small" $\tau$ ?

- the data is scarce $\Delta t = t + 1 - t = 1$ : use long 'time steps' $\tau$

- proposal (implicit scheme): use gradient flow: solve for $u \in [0, \tau]$ : $\mathcal{H}(u = 0) = H_{t+}$, $\frac{d}{du}\mathcal{H}(u) = \nabla_H(F_t(\mathcal{H}(u)) - G_t(\mathcal{H}(u)))$.

- Rq: can replace $f_t$ by a some mean-normalized version

# OFPA : online flow portfolio algorithm (+ P. Brugiere)

- ODE $\pi^u = \mathcal{S}(\mathcal{H}(u))$ :

$$\frac{d}{du}\mathcal{H}(u) = \frac{\pi^u \odot f_t}{\langle \pi^u \odot f_t, \mathbb{1}\rangle} - \pi^u - \xi \left( \sum_k \frac{\pi^u(k) - \pi_{t+}(k)}{\sqrt{(\pi^u(k) - \pi_{t+}(k))^2 + a^2}} \pi^u(k)(\mathbb{1}_{k=b} - \pi_b) \right)_b$$

- comparison with EC($\eta$) algorithm from [7] : in "$H$" formulation their update is of the form ($\xi = 0$) : $H_{t+1} = H_t + \tau \frac{f_t}{\langle \pi_t, f_t \rangle} + cst_t$

**Theoretical result**

For $\tau \to \infty$ and $\xi \to \infty$ we obtain a gradient flow (in some metric space of discrete probability related to the product Huber-energy metric). When $\tau \to 0$ and $\xi \to 0$ obtain a flow in a Hilbert space.

# OFPA : numerical results

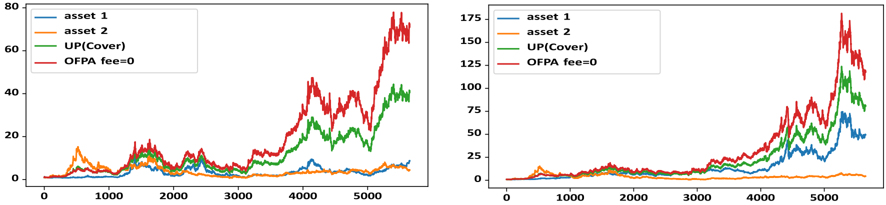
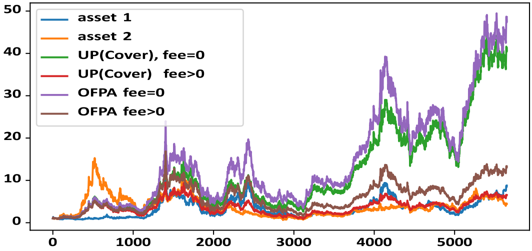
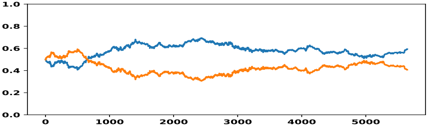



Figure: Preliminary numerical results comparing the performance of individual assets, UP of Cover and OFPA approach: **left** Iroquois vs. Kin Ark (cf. Cover paper), **right:** Commercial Metals vs. Kin Ark. Consistent with the the literature [7] we set $\tau = 0.05$, but this may not be transferable to other data.

CAUTION: these results are comparable with those from the literature [7] and depend on the data used. The performance vary greatly and the applicability domain is still to be investigated !

# OFPA : further numerical results



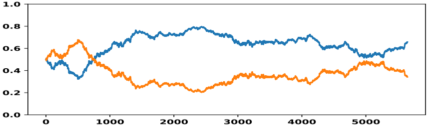


Figure: Preliminary numerical results comparing the performance with and without fees for Iroquois vs. Kin Ark (cf. Cover paper), $\tau = 0.5$. **Top:** comparison of the impact of fees between the UP and OFPA at $\xi = 2\%$. **Bottom:** evolution of OFPA portfolio allocation for $\xi = 0$ **(left)** and $\xi = 2\%$ **(right)**. The OFPA portfolio performs better when fees are taken into account while the UP does not improve over individual asset performance; the fee level is seen to influence its composition over time.

# References I

[1] Richard S. Sutton and Andrew G. Barto.
*Reinforcement learning. An introduction*.
Adapt. Comput. Mach. Learn. Cambridge, MA: MIT Press, 2nd expanded and updated edition edition, 2018.
http://incompleteideas.net/book/the-book-2nd.html.

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer.
Finite-time Analysis of the Multiarmed Bandit Problem.
*Machine Learning*, 47(2):235–256, May 2002.

[3] Gabriel Turinici.
The convergence of the Stochastic Gradient Descent (SGD) : a self-contained proof.
*arXiv:2103.14350 [cs, math, stat]*, March 2021.

# References II

[4] Bin Li and Steven C. H. Hoi.
Online portfolio selection: A survey.
*ACM Comput. Surv.*, 46(3), jan 2014.

[5] Thomas M Cover.
Universal portfolios.
*Mathematical finance*, 1(1):1–29, 1991.

[6] Bin Li and Steven CH Hoi.
On-line portfolio selection with moving average reversion.
*arXiv preprint arXiv:1206.4626*, 2012.

[7] David P Helmbold, Robert E Schapire, Yoram Singer, and Manfred K Warmuth.
On-line portfolio selection using multiplicative updates.
*Mathematical Finance*, 8(4):325–347, 1998.

# References III


[8] Cédric Villani.
*Optimal transport, Old and new*, volume 338.
Springer-Verlag, Berlin, 2009.

[9] Luigi Ambrosio and Nicola Gigli.
*Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009, Editors: Benedetto Piccoli, Michel Rascle*, chapter A User's Guide to Optimal Transport, pages 1–155.
Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[10] Guillaume Legendre and Gabriel Turinici.
Second-order in time schemes for gradient flows in wasserstein and geodesic metric spaces.
*Comptes Rendus Mathematique*, 355(3):345–353, 2017.

# References IV


[11] Laetitia Laguzet.
High order variational numerical schemes with application to Nash–MFG vaccination games.
*Ricerche di Matematica*, 67:247–269, 2018.

[12] Gabriel Turinici.
Radon–Sobolev Variational Auto-Encoders.
*Neural Networks*, 141:294–305, 2021.

[13] Gabriel Turinici.
Huber-energy measure quantization, 2022.