Reinforcement learning in finance: implicit policy gradient for portfolio allocation

Gabriel Turinici
with partial contributions
L. Laguzet, G. Legendre, S. Anita, P. Brugiere

CEREMADE, Université Paris Dauphine

ACDSDE Conference
Academy of Sciences of Iasi, sept 30^{th}, 2023

Executive summary

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

However few works investigated implicit type algorithms.

Exploiting previous JKO-style "flow gradient" techniques, we discuss a more formal setting for the implicit policy gradient schemes.

The procedure is further adapted to two situations: portfolio optimization in finance and the classical multi-armed bandit RL problem. Preliminary numerical results are encouraging.

Outline

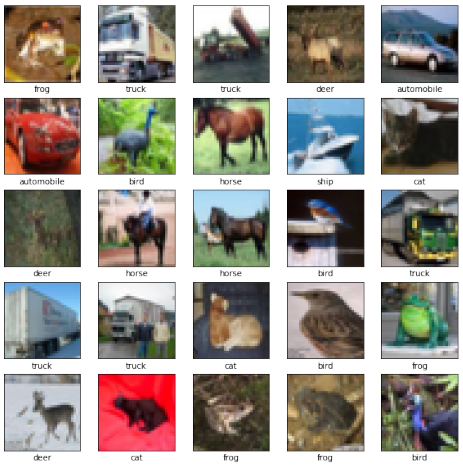
Reminders : types of "learning"

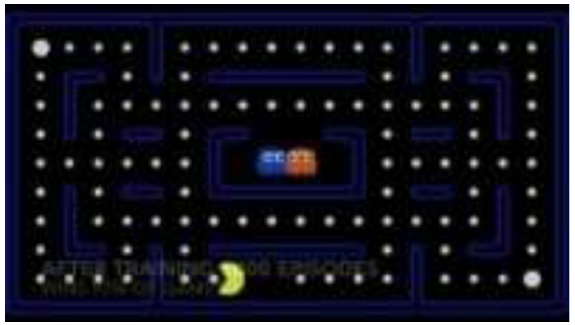
• Supervised learning : e.g. classification: the labels are given i.e. we know the value function;

• 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

- the problem is to allocate limited resources (time, money, turns etc.) among terms of a given list. Goal is to maximize expected rewards.
- Name: from slot-machines (one-armed bandit); example of goal maximize return over $n = 1000$ steps.

References : [1, 2] etc.

Multi-armed bandit

- k-armed bandit : has k options to choose from
- Other situations: choice among medical treatments, for a series of patients
- rewards information: each action has a random reward with a fixed but unknown mean;
- the means will be called "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
- value function $q_* : \mathcal{A} \to \mathbb{R}$ is unknown; in particular $q_*(a) := \mathbb{E}[R_t | A_t = a]$. (note : here "*" stands for the "true" or "optimal" or "most precise")

Multi-armed bandit : (policy) gradient algorithms

Choice of arm: probability law π_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 π_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
- $\nabla_{H_t(a)}\mathcal{R} = \nabla_{H_t}\left(\sum_b q_*(b)\pi_t(b)\right) = \sum_b q_*(b)\pi_t(b)(\mathbb{1}_{b=a} - \pi_t(a))$
$= \mathbb{E}_{A_t}[q_*(A_t)(\mathbb{1}_{A_t=a} - \pi_t(a))]$
- $R_t(\mathbb{1}_{a=A_t} - \pi_t(a)) =$ unbiased estimator for $\nabla_{H_t(a)}\mathcal{R}$ because $q_*(A_t) = \mathbb{E}[R_t|A_t]$; we use it in the SGD update of H_{t+1}.

Multi-armed bandit : theoretical insights into gradient algorithms

- Next idea: minimize $\mathcal{R}$ or $\mathcal{R} - c$ is the same (cst. c independent of A_t)
- $\mathcal{R} - c = \sum_b (q_*(b) - c)\pi_t(b)$
- $\nabla_{H_t(a)}(\mathcal{R} - c) = ... = \mathbb{E}_{A_t}[(q_*(A_t) - c)(\mathbb{1}_{A_t=a} - \pi_t(a))]$
- Choice for c ? Idea: consistency "if we are already in the solution move the least possible" : $c = \bar{R}_t$ (baseline) (e.g. take a situation with 2 or 3 actions having same q_*); can be seen as variance reduction technique [4, 5].
- 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

- we will take the example of portfolio optimization : choose among K assets
- notation : $\pi_t(k)$ will be the proportion of wealth allocated to asset k;
- total portfolio return $\sum_k \pi(k) R_t(k)$
- at first sight it is similar to reward maximization in MAB with the distinction all K rewards are available simultaneously at each step
- we will ask the question of "implicit" stochastic gradient schemes (see next section why) that are less sensitive to the choice of the learning rate α
- then we will move to more particular considerations regarding risk metrics (such as volatility etc) on one side and on higher order policy gradient algorithms on the other side.

Reinforcement learning : implicit gradient schemes

Stochastic implicit schemes may behave bad ...

SDE : $dX = a(X)dt + b(X)dW_t$

- Euler-Maruyama (explicit) $X_{n+1} = X_n + a(X_n)\Delta t + b(X_n)\Delta W_n$, ok
- "implicit" Euler-Maruyama : $X_{n+1} = X_n + a(X_{n+1})\Delta t + b(X_{n+1})\Delta W_n$
- For $a = 0, b(X) = X$: $X_{n+1} = \frac{X_n}{1-\Delta W_n}$; $\mathbb{E}|X_{n+1}| = \infty$ because $\mathbb{E}\left|\frac{1}{\mathcal{N}(\mu,\sigma^2)}\right| = \infty$, $\mathcal{N}(\mu, \sigma^2)$ = normal variable.

Stochastic optimization : $L(x) = \mathbb{E}_\omega[x^2 Z(\omega)/2]$, $Z(\omega) \sim \mathcal{N}(1, 1)$

- $L(x) = x^2/2$, minimum at $x = 0$.
- For ρ small enough the (explicit) stochastic gradient decent (SGD) converges : $x_{n+1} = x_n - \rho x_n Z_n = x_n(1 - \rho Z_n)$.
- ISGD (implicit SGD) : $x_{n+1} = x_n - \rho x_{n+1} Z_n$ thus $x_{n+1} = \frac{x_n}{1+\rho Z_n}$. Cannot make any step because $\mathbb{E}_{Z_n}|x_{n+1}| = \infty$!

Reinforcement learning : implicit gradient schemes

General conclusion : implicit schemes may be unstable if stochastic character is present.

They are however special cases [6] where ISGD is more stable and asymptotically of same order as SGD : "the implicit method is unconditionally stable under any specification of the learning rate, whereas standard SGD can deviate arbitrarily when the learning rate is misspecified" (from [6]).
Procedural problem : in [6] the stochastic variable is draws from the **future** distribution, can pose problems in practice.

Question : do implicit algorithm work for our application ?

Which one, how to design them ?

Outline

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 defintion, huge litterature (cf. books by Ambrosio et al. , Villani, Santambroggio) [7, 8]...

- Euclidian space (under some regularity assumptions):

$$\frac{d}{dt} F(x_t) = \langle \nabla F(x_t), x_t' \rangle \geq - \left| x_t' \right| \cdot \left| \nabla F \right| (x_t) \geq -\frac{1}{2} \left| x_t' \right|^2 - \frac{1}{2} \left| \nabla F \right|^2 (x_t),$$

$$\text{or equivalently } \frac{d}{dt} F(x_t) + \frac{1}{2} \left| x_t' \right|^2 + \frac{1}{2} \left| \nabla F \right|^2 (x_t) \geq 0,$$

with equality only if $x_t' = -\nabla F(x_t)$.

- Conclusion: $\frac{d}{dt} F(x_t) + \frac{1}{2} \left| x_t' \right|^2 + \frac{1}{2} \left| \nabla F \right|^2 (x_t) \leq 0$ a.e. is equivalent with $x_t' = -\nabla F(x_t)$.

Gradient flows: theory

- Euclidian space formulation: $\frac{d}{dt}F(x_t) + \frac{1}{2}|x_t'|^2 + \frac{1}{2}|\nabla F|^2(x_t) \leq 0$ a.e.
- the (local metric) slope of F at x:

$$|\nabla F|(x) = \limsup_{z\to x} \frac{[F(x)-F(z)]_+}{d(x,z)} = \max\left\{\limsup_{z\to x} \frac{F(x)-F(z)}{d(x,z)}, 0\right\}.$$

- the metric derivative of x at t: $|x_t'| = \lim_{h\to 0} \frac{d(x_{t+h},x_t)}{|h|}$, exists a.e. as soon as $t \mapsto x_t$ is absolutely continuous. Moreover $|x'| \in L^1(0,1)$.
- EDI ∇-flow (pointwise): $\frac{d}{dt}F(x_t) + \frac{1}{2}|x_t'|^2 + \frac{1}{2}|\nabla F|^2(x_t) \leq 0$ a.e.
- EDI ∇-flow from $\bar{x}$: an absolutely continuous curve such that:

$$\forall s \geq 0,\ F(x_s) + \frac{1}{2}\int_0^s |x_r'|\,\mathrm{d}r + \frac{1}{2}\int_0^s |\nabla F|^2(x_r)\,\mathrm{d}r \leq F(\bar{x}),$$

$$\text{a.e. } t > 0,\ \forall s \geq t,\ F(x_s) + \frac{1}{2}\int_t^s |x_r'|\,\mathrm{d}r + \frac{1}{2}\int_t^s |\nabla F|^2(x_r)\mathrm{d}r \leq F(x_t).$$

- EVI form for λ-convex (i.e., when smooth $F'' \geq \lambda Id$ ...) functionals:
$F(x_t) + \frac{1}{2}\frac{d}{dt}d^2(x_t,y) + \frac{\lambda}{2}d^2(x_t,y) \leq F(y), \forall y,\ a.e.\ t \geq 0.$

Gradient flows examples: heat flow (Fokker-Planck)

$\mathcal{X} = \mathcal{P}_2(\mathbb{R})$ (the set of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with finite second-order moment, endowed with the Wasserstein distance $\mathcal{W}_2$)
Consider for $\sigma \in \mathbb{R}$ $F : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R} \cup \{+\infty\}$:
$F(\nu) = \int_{\mathbb{R}} V(x)\rho(x) + \frac{\sigma^2}{2} \int_{\mathbb{R}} \rho(x) \log(\rho(x)) dx$, if $\nu \ll dx$, $\nu = \rho(x)dx$
$F(\nu) = +\infty$, if $\nu \not\ll dx$.
For smooth V, the gradient flow $t \mapsto \nu(t) \in \mathcal{P}_2(\mathbb{R})$ of F satisfies $\nu(t) = \rho(t, \cdot)dx$ and:

$$\frac{\partial \rho}{\partial t}(t, x) = \frac{\partial}{\partial x}[V'(x)\rho(t, x)] + \frac{\sigma^2}{2} \frac{\partial^2 \rho}{\partial x^2}(t, x), \tag{1}$$

i.e., Fokker-Planck of the SDE: $dX(t) = -V'(X(t))dt + \sigma dW(t)$.

Remark: also a L^2 flow (term $\int |\nabla \rho|^2$)...

Gradient flows examples: heat flow (Fokker-Planck)

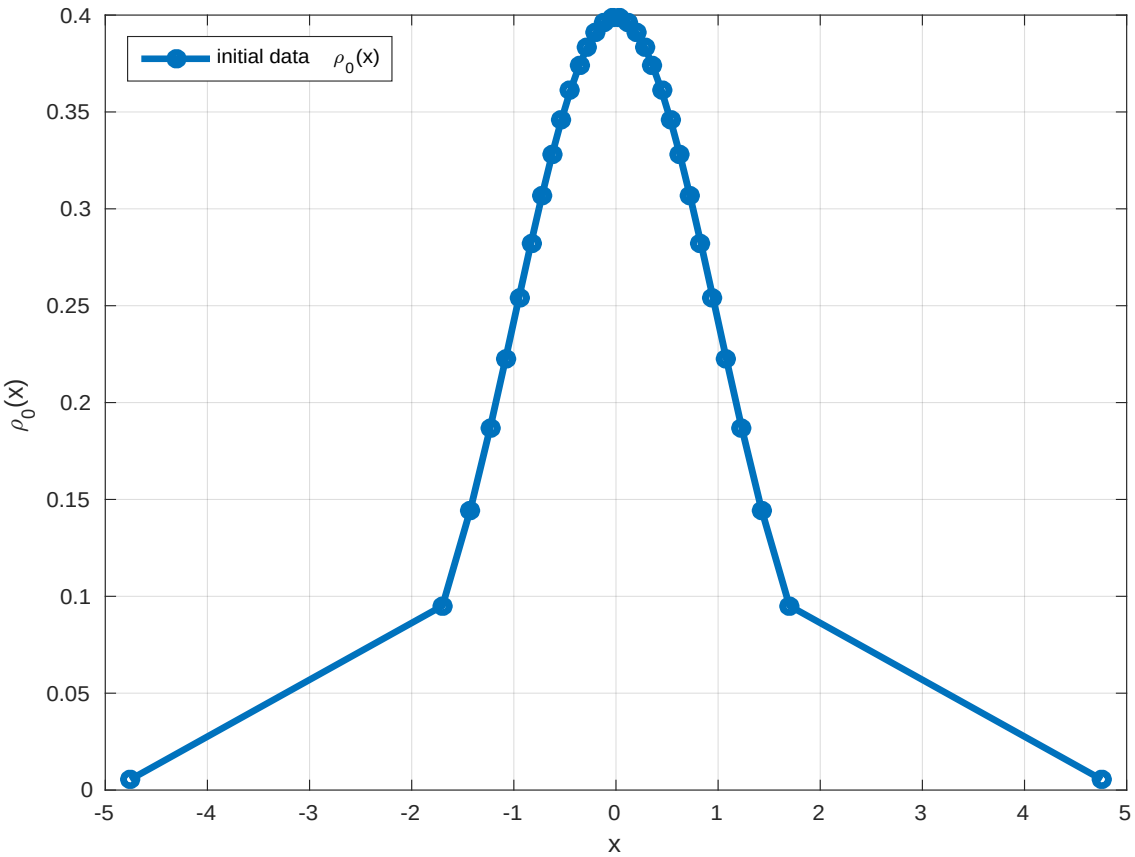


Figure: Initial data for the heat flow (FP) model and its evolution (VIDEO).

Gradient flows examples : a 1D Patlak-Keller-Segel model

• the (modified) Patlak–Keller–Segel system (Perthame-Calvez-Sharifi Tabar 2007, Blanchet-Calvez-Carrillo 2008), is a PDE model for diffusion-aggregation competition in biological applications (chemotaxis).

• Free energy functional:

$$\mathcal{G}[\rho] = \int \rho(t,x)\log(\rho(t,x))\,\mathrm{d}x + \frac{\chi}{\pi}\int\int \rho(t,x)\rho(t,y)\log|x-y|\,\mathrm{d}x\mathrm{d}y$$

• the resulting Patlak-Keller-Segel equation:

$$\frac{\partial \rho}{\partial t} = \Delta\rho - \nabla(\chi\,\rho\nabla c),\ t > 0,\ x \in \Omega \subset \mathbb{R}^d \tag{2}$$

$$c = -\frac{1}{d\pi}\log|z| \star \rho \tag{3}$$

$\rho =$ cell density, $c =$ concentration of chemo-attractant, $\chi =$ sensitivity of the cells to the chemo-attractant.

Gradient flows examples: 1D Patlak-Keller-Segel model

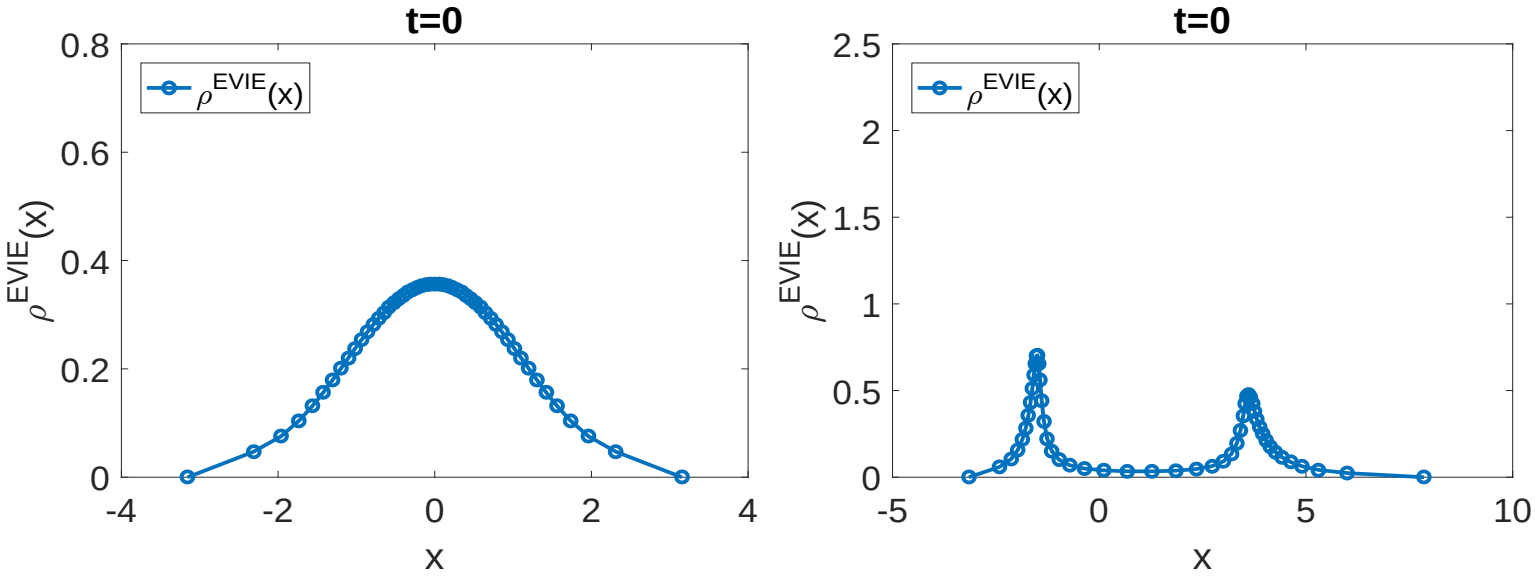


Figure: Initial data for the PKS model: $\chi = \pi$ (left), $\chi = 1.9\pi$ (right) and its evolution (VIDEO $T = 2$). Implementation : G. Legendre; ∇-flow JKO PKS code : courtesy A. Blanchet.

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

$$x \mapsto P_F^{JKO}(x; x_n^\tau, \tau) := \frac{1}{2\tau} d^2(x_n^\tau, x) + F(x). \tag{4}$$

• 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

• JKO scheme only (!!) was then used numerically to compute the gradient flow (J.K.O '99, Blanchet et al. 2009, Benamou et al 2016, ...). What about other numerical schemes ?

• JKO = first-order. Dynamics is regular with respect to time ! What about higher (second) order ?

High order schemes

- Idea 1: by Runge-Kutta : ODE $x'(t) = f(x(t))$ $(f = -\nabla F)$

Crank-Nicholson $x^\tau_{k+1} = x^\tau_k + \tau \frac{f_{k+1}+f_k}{2}$, 2nd order.

For ∇-flows in a metric space: no gradient 'f', no vector calculus.

- Idea 2 (from symplectic integrators): increase the order by composition
  - ⋆ here: take F quadratic, equation $x'(t) = Ax(t)$ (linear);
  - ⋆ Implicit Euler: $\frac{x^\tau_{k+1}-x^\tau_k}{\tau} = Ax^\tau_{k+1}$ thus $x^\tau_{k+1} = (I - \tau A)^{-1} x^\tau_k$;
  - ⋆ composition of IE steps $\alpha_1 h$, ... $\alpha_n h$: multiplication by

$(I - \alpha_n \tau A)^{-1} ... (I - \alpha_1 \tau A)^{-1}$;

  - ⋆ condition to be the same as $exp(A\tau)$:

first order $\sum_{\ell=1}^n \alpha_\ell = 1$;

second order $\sum_{\ell=1}^n \alpha_\ell^2 + \sum_{1 \le \ell < m \le n} \alpha_\ell \alpha_m = 1/2$.

Second condition implies $\left(\sum_{\ell=1}^n \alpha_\ell\right)^2 - \sum_{1 \le \ell < m \le n} \alpha_\ell \alpha_m = 1/2$ thus $\sum_{1 \le \ell < m \le n} \alpha_\ell \alpha_m = 1/2$, $\sum_{\ell=1}^n \alpha_\ell^2 = 0$.

CANNOT obtain second order from composition of I.E. schemes.

Second order schemes for gradient flows: the VIM scheme

- Recall for ODE $x'(t) = f(x(t))$: consistency error ($t_k = k\tau$).

consistency error for Implicit Euler $\frac{x(t_{k+1})-x(t_k)}{\tau} - f(x(t_{k+1})) =$
$x'(t_{k+1/2}) + O(\tau^2) - f(x(t_{k+1})) = \underbrace{f(x(t_{k+1/2})) - f(x(t_{k+1}))}_{O(\tau)} + O(\tau^2)$.

- (modified) Midpoint method: $x^\tau_{k+1} = x^\tau_{k-1} + 2\tau f_k$, 2nd order.

- idea: use the variational formulation : Variational Implicit Midpoint (VIM) scheme (G. Legendre, G.T., '16 [9]):

$x^\tau_{k+1} \in argmin_{x\in\mathcal{A}} \frac{d(x^\tau_k,y)^2}{2\tau} + 2F(\frac{x^\tau_k+y}{2})$

"$\frac{x+y}{2}$" = the midpoint of the geodesic from x to y.

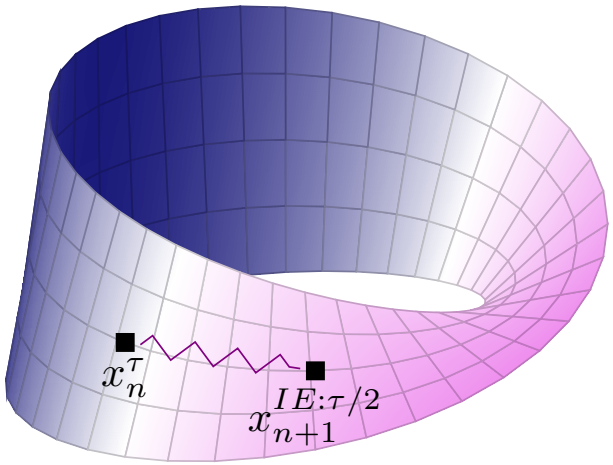
- Hilbert space critical point equation: $\frac{x^\tau_{n+1}-x^\tau_n}{\tau} + \nabla F(\frac{x^\tau_{n+1}+x^\tau_n}{2}) = 0$.

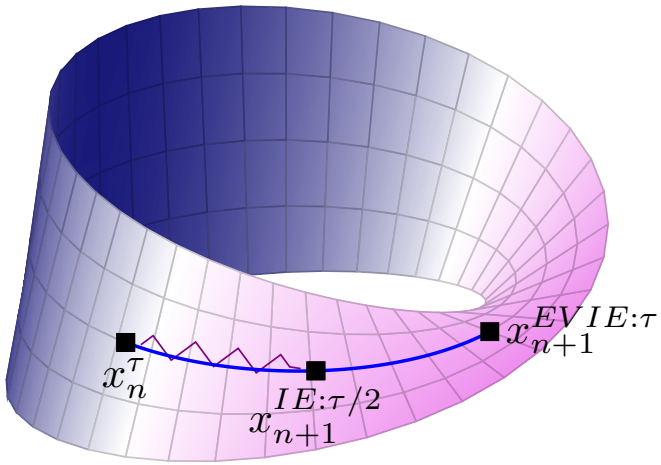
Consistency error $= O(\tau^2)$.

Second order schemes for gradient flows: the EVIE scheme

- Re-writing of the VIM scheme: $x^\tau_{k+1} \in argmin_{x\in\mathcal{A}} \frac{d(x^\tau_k,y)^2}{2\tau} + 2F(\frac{x^\tau_k+y}{2})$
- Notation $z = \frac{x^\tau_k+y}{2}$, then y is the 2-geodesic-extrapolate of x^τ_k with respect to z, "$y = 2z - x^\tau_k$"; $d(x^\tau_k, y) = 2d(x^\tau_k, z)$.
- $\min_{z\in\mathcal{A}}... \frac{d(x^\tau_k,z)^2}{2(\tau/2)} + F(z)$: BUT this is I.E. of step $\tau/2$!
- Extrapolated Variational Implicit Euler (EVIE) scheme : do a $\tau/2$ IE (= JKO) step and then extrapolate on the geodesic.

EASY to implement in an existing JKO / IE code ! OK in Hilbert spaces ...

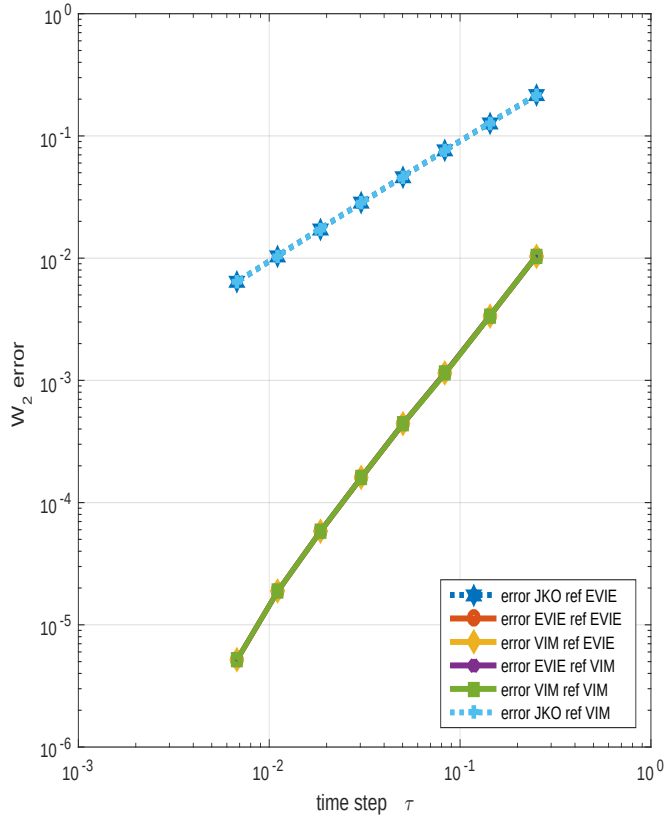




Numerical results for VIM and EVIE schemes: heat flow

Numerical results for $F(\nu) = \int_{\mathbb{R}} V(x)\rho(x) + \frac{\sigma^2}{2}\int_{\mathbb{R}} \rho(x)\log(\rho(x))dx$,

$V(x) = \theta\frac{(x-\mu)^2}{2}$, $T = 1$, $\sigma = 1$, $\theta = \frac{1}{2}$, $\mu = 5$, $M = 32$ spatial discretization points.



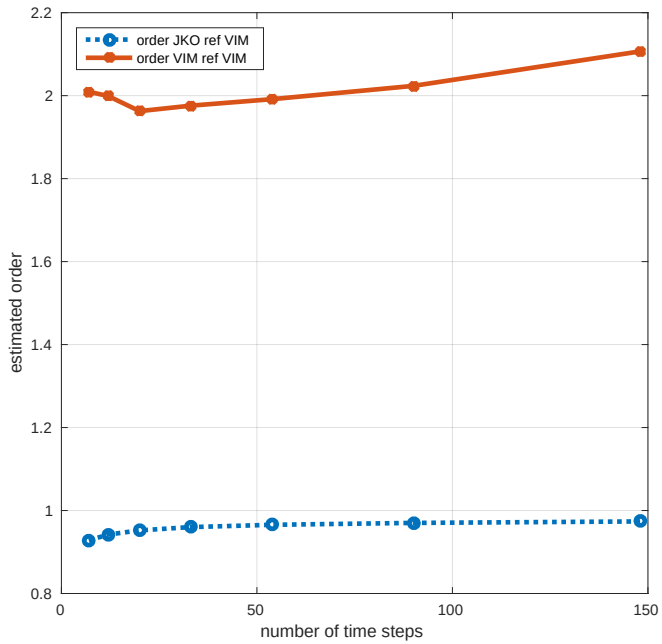


Figure: # of time steps: 4, 7, 12, 20, 33, 54, 90 and 148 (reference 244). **Left:** error for JKO (dotted line) and VIM / EVIE (solid lines) schemes. **Right:** order of convergence: JKO (dotted line), VIM / EVIE (solid lines). 4 steps VIM/EVIE = 90 steps JKO/IE.

Numerical results for the EVIE scheme: PKS

$$\mathcal{G}[\rho] = \int \rho(t,x)\log(\rho(t,x))\,\mathrm{d}x + \frac{\chi}{\pi}\int\int \rho(t,x)\rho(t,y)\log|x-y|\,\mathrm{d}x\mathrm{d}y$$

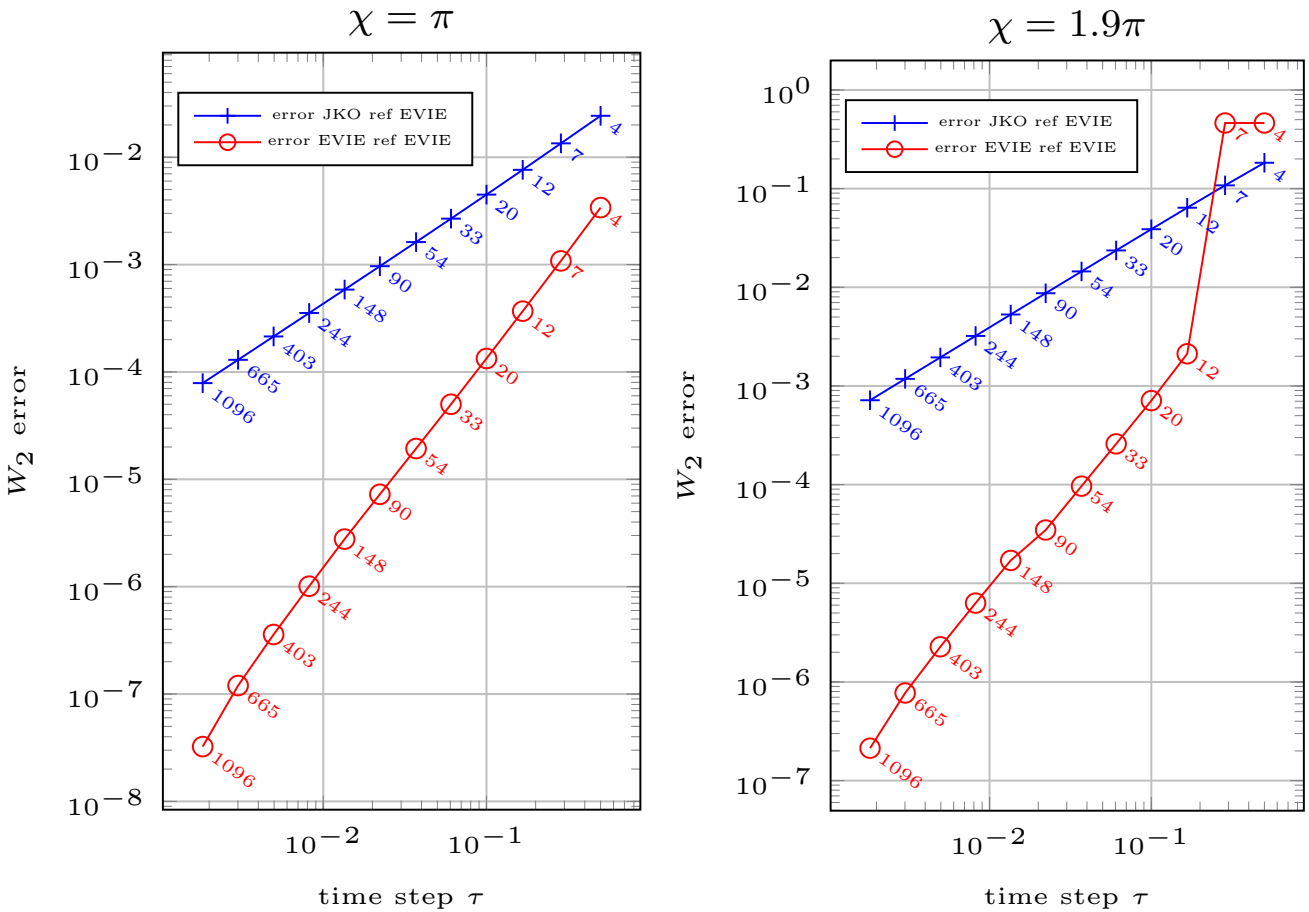


Figure: Error of the JKO and EVIE schemes for PKS model; $T = 2$, time steps: reference sol EVIE(1808). **Left:** $\chi = \pi$, order JKO $= 1.02$, order EVIE $= 2.01$; **Right:** $\chi = 1.9\pi$, order JKO $= 0.99$, order EVIE $= 2.00$ (corrected by excluding first two points).

Theoretical results for the VIM scheme

mid-slope : $\left|\nabla^M F\right|(x,y) = \limsup_{z\to y} \frac{\left(F\left(\frac{x+y}{2}\right)-F\left(\frac{x+z}{2}\right)\right)^+}{d\left(\frac{x+y}{2},\frac{x+z}{2}\right)}$.

Hypothesis (non standard):

- (geometric) $\forall x \in \mathcal{X}$ the set $\bigcup_{y\in\mathcal{X}} \frac{x+y}{2}$ is closed;
- (geometric) $\forall (x,y) \in \mathcal{X}^2$, the set $\frac{x+y}{2}$ is a singleton.
- (adaptation for $\left|\nabla^M F\right|$ instead of $|\nabla F|$) $\forall x \in D(F)$, $D(F) \supset (x_n)_{n\in\mathbb{N}} \to x$ and $D(F) \supset (y_n)_{n\in\mathbb{N}} \to x$ imply $\left|\nabla^M F\right|(x,x) \leq \liminf_{n\to\infty} \left|\nabla^M F\right|(x_n,y_n)$;
- (regularity for F) if any two of the elements x, y, $\frac{x+y}{2}$ belong to $D(F)$, then the third also does and:

$$\left|\frac{F(x)+F(y)-2\,F(\frac{x+y}{2})}{d^2(x,y)}\right| \leq H, \tag{5}$$

where H is a constant independent of x and y.

Sufficient condition: F and $-F$ are λ-convex.

Theoretical results for the VIM scheme

Hypothesis (standard):

- F is lower semicontinuous, bounded from below, and such that: $\forall r > 0, \forall c \in \mathbb{R}, \forall x \in \mathcal{X}$ the set $\{y \in \mathcal{X} \,|\, F(y) \leq c,\ d(x,y) \leq r\}$ is compact,
- F has the following continuity property
  if $x_n \to x$, and $\sup\{|\nabla F|(x_n), E(x_n)\} < \infty$ then $F(x_n) \to F(x)$;

Theorem (G. Legendre, G.T. 2016)

Let $T > 0$ be fixed and $(\mathcal{X}, d)$ be a Polish metric space. Under above hypotheses for some $\bar{\tau} > 0$, the set of curves $\{(x_t^\tau)_{t\in[0,T]}; 0 \leq \tau \leq \bar{\tau}\}$ is relatively compact (with respect to the local uniform convergence) and any limit curve is a gradient flow in the EDI formulation.

- this is consistency
- what about the (second) order of convergence ?

Equilibrium metric flows: theoretical results (GT '17)

• Hilbert space: $\partial_\tau \xi(\tau,t) + \nabla_1 \mathcal{C}(\xi(\tau,t),\xi(\tau,t)) = 0$; metric space equivalent ?

• literature: ∇-flows for $E(t,x)$: Ferreira-Valencia-Guevara '15, Rossi-Mielke-Savaré '08, C. Jun '12, Kopfer-Sturm '16

• EDI (pointwise) formulation (G.T. '17) [10]
$\frac{d}{d\tau}\mathcal{C}(\xi_\tau,\nu)\Big|_{\nu=\xi_\tau} + \frac{1}{2}|\xi'_\tau|^2 + \frac{1}{2}|\nabla_1\mathcal{C}|^2(\xi_\tau,\xi_\tau) \leq 0$ a.e.
does not use convexity but uses regularity hypothesis for $\mathcal{C}$.

• EVI formulation (G.T. '17)
$\mathcal{C}(\xi_\tau,\xi_\tau) + \frac{1}{2}\frac{d}{d\tau}d^2(\xi_\tau,y) + \frac{\lambda}{2}d^2(\xi_\tau,y) \leq \mathcal{C}(y,\xi_\tau), \forall y,\ a.e.\ \tau \geq 0.$
does not use much regularity uses λ-convexity.

• both are the limit of numerical schemes (under hyp.)

High order schemes : results by L. Laguzet

What about 2nd order schemes for metric gradient flows ? Cf. work by L. Laguzet, 3 high order schemes inspired from Heun, RK3, RK4: [11]
The (standard, Hilbert space) Heun:

$$p_1 = x_k + \tau f(t_k, x_k),\ x_{k+1} = x_k + \frac{\tau}{2}\Big[f(t_k, x_k) + f(t_{k+1}, p_1)\Big].$$

The variational (metric space) Heun scheme

$$\widetilde{\xi}_{k+1} \in \operatorname*{argmin}_{\eta \in \mathcal{A}} \left\{ \frac{d(\eta, \xi_k)^2}{2\tau} + \mathcal{C}(\eta, \xi_k) \right\}, \tag{6}$$

$$\xi_{k+1} \in \operatorname*{argmin}_{\eta \in \mathcal{A}} \left\{ \frac{d(\eta, \xi_k)^2}{2\tau} + \frac{1}{2}\mathcal{C}(\eta, \xi_k) + \frac{1}{2}\mathcal{C}(\eta, \widetilde{\xi}_{k+1}) \right\}. \tag{7}$$

Two minimizations are required in order to obtain ξ_{k+1}.

Outline

Implicit stochastic schemes in finance (partially with P. Brugiere)

- Financial application : portfolio optimization, maximize return $\sum_k \pi(k) R_t(k)$ (r.v.); r.v. R_t of mean $\bar{R}$ and covariance Σ
- π is a distribution ($\in$ probability simplex), "softmax" representation $\pi(a) = softmax(H) = \frac{e^{H(a)}}{\sum_b e^{H(b)}}$
- we use implicit schemes to go from step t to step $t+1$.
- to maximize $F_{R_t}(H) = \sum_k \pi_H(k) R_t(k)$, $F(H) = \mathbb{E}_{R_t}[\sum_k \pi_H(k) R_t(k)]$
- once $R_t(k)$ is sampled for each k, $F_{R_t}(\cdot)$ is a bounded function of H; moreover its gradient $\nabla_H F_{R_t}(H)$ is also bounded :
$\nabla_H F_{R_t}(H) = (\nabla_H F_{R_t}(H))_{a=1}^K = (\nabla_{H_a} \sum_k R_t(k) \pi_H(k))_{a=1}^K$
$= (\sum_k R_t(k) \pi_H(k) (\mathbb{1}_{k=a} - \pi_H(a)))_{a=1}^K = R_t . * \pi_H - \langle R_t, \pi_H \rangle \pi_H$ ($.*$ is the elementwise product = Hadamard)
- Thus F_{R_t} is bounded, its gradient bounded (i.e. conditional on R_t)
- in finance one also includes some risk measures, and the reward will rather be $g(\sum_k \pi(k) R_t(k))$, e.g., $g(y) = y - \lambda y^2$ (risk $\sim$ volatility), $g(y) = y - \lambda (y_-)^2$ (risk $\sim$ drawdown),...

Implicit stochastic schemes in finance (partially with P. Brugiere)

- notation $\pi_t = softmax(H_t)$, $\tau =$ "time" step
- explicit scheme

$H_{t+1} = H_t + \tau \nabla_H F_{R_t}(H_t) = H_t + \tau(R_t. * \pi_t - \langle R_t, \pi_t \rangle \pi_t)$

- ok for small τ, unstable otherwise, but what is "small" τ ?
- How to write an implicit scheme ?
- implicit scheme candidate $H_{t+1} = H_t + \tau(R_t. * \pi_{t+1} - \langle R_t, \pi_{t+1} \rangle \pi_{t+1})$
- Does a solution exist ?
- Picard iterations for small τ ; using boundedness of the gradient Brower for all τ ...
- asymptotic for $\tau \to \infty$? explicit $O(\tau)$, implicit ?
- JKO approach, i.e. minimization : $H_{t+1} = \arg\min \frac{d(H,H_t)^2}{2\tau} - F_{R_t}(H)$
- Consequences : existence for any τ, bound $d(H, H_t)^2 \leq C\tau$, asymptotic behavior at most $O(\sqrt{\tau})$

Implicit stochastic schemes in finance (partially with P. Brugiere)

- how to find the solution ?
- Choice 1 : standard numerical algorithm
- Choice 2 : for small τ : Picard iterations on the critical point equations
- Choice 3: for all τ : use JKO **again** to minimize $\underbrace{\frac{d(H,H_t)^2}{2\tau} - F_{R_t}(H)}_{\text{the new function to minimize}}$

$$Y_{l+1} = \arg\min \frac{d(Y,Y_\ell)^2}{2\rho} + \left(\frac{d(Y,H_t)^2}{2\tau} - F_{R_t}(Y)\right),\ Y_0 = H_t$$

Advantages: we choose the ρ; when ρ is small enough the minimum is unique, the critical point equation $=$ fixed point of a contraction $\Longrightarrow$ Picard; can limit the effort in the number of iterations in ℓ.

Numerical results for MAB (with Stefana Anita)

MAB : explicit update formula
$H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t)(\mathbb{1}_{a=A_t} - \pi_t(a))$

naive implicit update formula
$H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t)(\mathbb{1}_{a=A_t} - \pi_{t+1}(a))$, can be solved by Picard for small α

For large α does this corresponds to a JKO-style minimization ?

Remark: the unbiasedness of the gradient is not discussed here yet, hence the denomination "semi-implicit"

JKO-style minimization
$H_{t+1} = \arg\min \frac{d(H,H_t)^2}{2\alpha} - (R_t - \bar{R}_t) \log\left[softmax(H)(A_t)\right]$
$H_{t+1} = \arg\min \frac{d(H,H_t)^2}{2\alpha} - (R_t - \bar{R}_t) \left[H(A_t) - \log(\sum_a e^{H(a)})\right]$

Numerical results for MAB (with Stefana Anita)

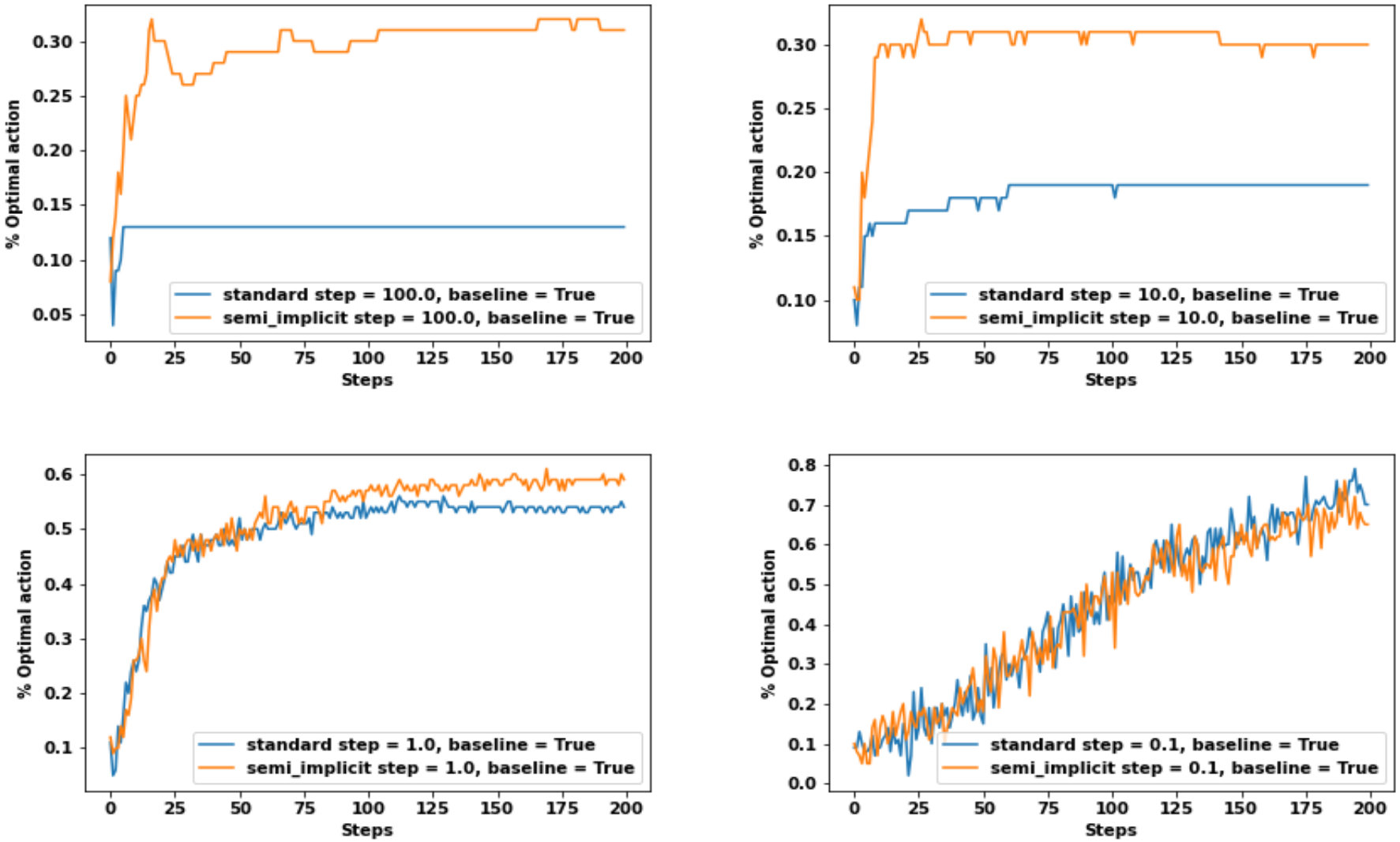


Figure: Numerical results comparing the explicit and semi-implicit scheme. The semi-implicit scheme appears more stable for large values of the step size and comparable for smaller values.

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] Lex Weaver and Nigel Tao.
The optimal reward baseline for gradient-based reinforcement learning, 2013.
arxiv:1301.2315.

[5] Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter.
Variance reduction techniques for gradient estimates in reinforcement learning.
J. Mach. Learn. Res., 5:1471–1530, dec 2004.

[6] Panagiotis Toulis, Edoardo Airoldi, and Jason Rennie.
Statistical analysis of stochastic gradient methods for generalized linear models.
In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 667–675, Bejing, China, June 2014. PMLR.

References III

Issue: 2.

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

[8] 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.

[9] 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

[10] Gabriel Turinici.
Metric gradient flows with state dependent functionals: The Nash-MFG equilibrium flows and their numerical schemes.
Nonlinear Analysis, 165:163–181, 2017.

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