Evalutation Apprentissage par renforcement

Pierre Cavalier

January 22, 2024

1 Etude théorique d'un nouvel algorithme de points fixes

Soit $d \geq 1$, $F \colon \mathbb{R}^d \to \mathbb{R}^d$ une application admettant un point fixe $x_* \in \mathbb{R}^d$, $\eta > 0$, et $x_0 \in \mathbb{R}^d$. On définit l'algorithme Ada-FP comme suit :

$$x_{k+1} = x_k + \eta \frac{Fx_k - x_k}{\sqrt{\sum_{\ell=0}^{k} \|Fx_\ell - x_\ell\|_2^2}}, \quad k \geq 0, \quad \text{(Ada-FP)}$$

avec la convention $0/0 = 0$. On note pour tout $k \geq 0$,

$$u_k = Fx_k - x_k,$$

$$\eta_k = \eta \frac{1}{\sqrt{\sum_{\ell=0}^{k} \|Fx_\ell - x_\ell\|_2^2}},$$

$$D_k = \max_{0 \leq \ell \leq k} \frac{1}{2} \|x_\ell - x_*\|_2^2.$$

1.1

Montrer que pour tout $k \geq 0$,

$$\|x_{k+1} - x_*\|_2^2 \leq \|x_k - x_*\|_2^2 + 2\eta_k u_k^T (x_k - x_*) + \eta_k^2 \|u_k\|_2^2.$$

En utilisant le fait que $x_{k+1} = x_k + \eta_k u_k$:

$$\begin{aligned}
\|x_{k+1} - x_*\|_2^2 &= \|x_k + \eta_k u_k - x_*\|_2^2 \\
&= \langle x_k - x_* + \eta_k u_k, x_k - x_* + \eta_k u_k \rangle \\
&= \langle x_k - x_*, x_k - x_* \rangle + 2\langle x_k - x_*, \eta_k u_k \rangle + \langle \eta_k u_k, \eta_k u_k \rangle \\
&= \|x_k - x_*\|_2^2 + 2\eta_k u_k^T (x_k - x_*) + \eta_k^2 \|u_k\|_2^2
\end{aligned}$$

Et ce par définition de la norme $\|.\|_2$ dans $\mathbb{R}^d$ issue du produit scalaire euclidien.
Ce qui permet, en particulier, de trouver le résultat voulu:

$$\|x_{k+1} - x_*\|_2^2 \leq \|x_k - x_*\|_2^2 + 2\eta_k u_k^T (x_k - x_*) + \eta_k^2 \|u_k\|_2^2.$$

1.2

En déduire que pour tout $k \geq 0$,

$$\sum_{\ell=0}^{k} u_\ell^T (x_* - x_\ell) \leq \frac{D_k}{\eta_k} + \sum_{\ell=0}^{k} \frac{\eta_\ell \|u_\ell\|_2^2}{2}.$$

Pour un ℓ fixé:

$$\|x_{\ell+1} - x_*\|_2^2 \leq \|x_\ell - x_*\|_2^2 + 2\eta_\ell u_l^T(x_\ell - x_*) + \eta_\ell^2\|u_\ell\|_2^2$$

$$\implies \frac{1}{\eta_\ell}(\|x_{\ell+1} - x_*\|_2^2 - \|x_\ell - x_*\|_2^2) - \eta_\ell\|u_\ell\|_2^2 \leq 2u_l^T(x_\ell - x_*) \qquad (\text{car } \eta_\ell > 0)$$

$$\implies u_l^T(x_* - x_\ell) \geq \frac{1}{2\eta_\ell}(\|x_\ell - x_*\|_2^2 - \|x_{\ell+1} - x_*\|_2^2) + \frac{\eta_\ell\|u_\ell\|_2^2}{2}$$

En sommant les inégalités pour ℓ allant de 0 à k on obtient:

$$\sum_{\ell=0}^{k} u_\ell^T(x_* - x_\ell) \leq \sum_{\ell=0}^{k} \frac{1}{2\eta_\ell}(\|x_\ell - x_*\|_2^2 - \|x_{\ell+1} - x_*\|_2^2) + \sum_{\ell=0}^{k} \frac{\eta_\ell\|u_\ell\|_2^2}{2}.$$

Il suffit maintenant de prouver que $\sum_{\ell=0}^{k} \frac{1}{2\eta_\ell}(\|x_\ell - x_*\|_2^2 - \|x_{\ell+1} - x_*\|_2^2) \leq \frac{D_k}{\eta_k}$ et l'inégalité souhaitée sera obtenue.

$$\begin{aligned}
\sum_{\ell=0}^{k} \frac{1}{2\eta_\ell}(\|x_\ell - x_*\|_2^2 - \|x_{\ell+1} - x_*\|_2^2) &= \sum_{\ell=0}^{k} \frac{1}{2\eta_\ell}(\|x_\ell - x_*\|_2^2) + \sum_{\ell=1}^{k+1} \frac{1}{2\eta_{\ell-1}}(\|x_\ell - x_*\|_2^2) \\
&= \frac{1}{2\eta_0}(\|x_0 - x_*\|_2^2) + \sum_{\ell=1}^{k}(\frac{1}{2\eta_\ell} - \frac{1}{2\eta_{\ell-1}})\|x_\ell - x_*\|_2^2 - \underbrace{\frac{1}{2\eta_k}(\|x_{k+1} - x_*\|_2^2)}_{>0} \\
&\leq \frac{1}{\eta_0}D_k + D_k\sum_{\ell=1}^{k}(\frac{1}{\eta_\ell} - \frac{1}{\eta_{\ell-1}}) \\
&\leq D_k(\frac{1}{\eta_0} + \frac{1}{\eta_k} + \frac{1}{\eta_0}) \\
&\leq \frac{D_k}{\eta_k}
\end{aligned}$$

Ce qui achève la démonstration.

1.3

1.3.1

Soit $(a_k)_{k\geq 0}$ une suite positive. Montrer que pour tout $k \geq 0$,

$$\sum_{\ell=0}^{k} \frac{a_\ell}{\sqrt{\sum_{m=0}^{\ell} a_m}} \leq 2\sqrt{\sum_{\ell=0}^{k} a_\ell},$$

Procédons par récurrence, pour $k = 0$ on obtient:

$$\sum_{\ell=0}^{0} \frac{a_\ell}{\sqrt{\sum_{m=0}^{\ell} a_m}} = \frac{a_0}{\sqrt{a_0}} = \sqrt{a_0} \leq 2\sqrt{\sum_{\ell=0}^{0} a_\ell},$$

On suppose l'hypothèse vraie pour k fixé. Montrons qu'elle est vraie pour $k+1$:

$$\begin{aligned}
\sum_{\ell=0}^{k+1} \frac{a_\ell}{\sqrt{\sum_{m=0}^{\ell} a_m}} &= \sum_{\ell=0}^{k} \frac{a_\ell}{\sqrt{\sum_{m=0}^{\ell} a_m}} + \frac{a_{k+1}}{\sqrt{\sum_{m=0}^{k+1} a_m}} \\
&\leq 2\sqrt{\sum_{\ell=0}^{k} a_\ell} + \frac{a_{k+1}}{\sqrt{\sum_{m=0}^{k+1} a_m}} \qquad (\text{Par hypothèse de récurrence})
\end{aligned}$$

La fonction $x \mapsto \sqrt{x}$ est concave et sa dérivée est la fonction $x \mapsto \frac{1}{2\sqrt{x}}$ défini sur $\mathbb{R}_*^+$. En un point a fixé de $\mathbb{R}_*^+$, on peut écrire l'équation de la tangente de paramètre b qui, par propriété des fonctions concaves, est au dessus de la courbe de la fonction:

$$\sqrt{a} \leq \sqrt{b} + \frac{1}{2\sqrt{b}}(a-b) \iff 2\sqrt{b} \geq 2\sqrt{a} + \frac{1}{\sqrt{b}}(b-a)$$

En prenant $a = \sum_{\ell=0}^{k} a_\ell$ et $b = \sum_{\ell=0}^{k+1} a_\ell$, on obtient (en remarquant que $b - a = a_{k+1}$):

$$2\sqrt{\sum_{\ell=0}^{k+1} a_\ell} \geq 2\sqrt{\sum_{\ell=0}^{k} a_\ell} + \frac{a_{k+1}}{\sqrt{\sum_{m=0}^{k+1} a_m}} \geq \sum_{\ell=0}^{k+1} \frac{a_\ell}{\sqrt{\sum_{m=0}^{\ell} a_m}}$$

L'initialisation et l'hérédité étant vérifiée, l'hypothèse de récurrence est donc vraie pour tout $k \geq 0$ ce qui conclut la question.

1.3.2

En déduire que pour $k \geq 0$,

$$\sum_{\ell=0}^{k} u_\ell^T (x_* - x_\ell) \leq \left(\eta + \frac{D_k}{\eta}\right) \sqrt{\sum_{\ell=0}^{k} \|u_\ell\|_2^2}.$$

En partant de la question 2 et en utilisant la 3.a sur la suite $a_\ell = \|u_\ell\|_2^2$ qui est bien une suite positive et que $\eta_\ell = \frac{\eta}{\sqrt{\sum_{\ell=0}^{k} \|u_\ell\|_2^2}}$, on a:

$$\begin{aligned}
\sum_{\ell=0}^{k} u_\ell^T (x_* - x_\ell) &\leq \frac{D_k}{\eta_k} + \sum_{\ell=0}^{k} \frac{\eta \|u_\ell\|_2^2}{2\sqrt{\sum_{\ell=0}^{k} \|u_\ell\|_2^2}} \\
&\leq \frac{D_k \sqrt{\sum_{\ell=0}^{k} \|u_\ell\|_2^2}}{\eta} + \eta \sqrt{\sum_{\ell=0}^{k} \|u_\ell\|_2^2} \\
&\leq \left(\eta + \frac{D_k}{\eta}\right) \sqrt{\sum_{\ell=0}^{k} \|u_\ell\|_2^2}
\end{aligned}$$

1.4

On suppose que F est γ_F-Lipschitzienne pour un certain $0 \leq \gamma_F < 1$ i.e:

$$\forall x, y \in \mathbb{R}^d, \quad \|F(x) - F(y)\|_2 \leq \gamma_F \|x - y\|_2$$

1.4.1

Pour tout $k \geq 0$, montrons que

$$\|Fx_k - x_k\|_2^2 \leq 2(Fx_k - x_k)^T (x^* - x_k).$$

En partant du fait $x_* = Fx_*$ on obtient:

$$
\begin{aligned}
\|Fx_k - x_k\|_2^2 &= \|Fx_k + Fx_* - x_* - x_k\|_2^2 \\
&= \|Fx_k - Fx_*\|_2^2 + \|x_* - x_k\|_2^2 + 2\langle x_k - x_*, Fx_k - Fx_*\rangle \\
&\leq \underbrace{\gamma_F}_{\leq 1} \|x_* - x_k\|_2^2 + \|x_* - x_k\|_2^2 + 2\langle x_k - x_*, Fx_k - x_k + x_k - x_*\rangle \\
&\leq 2\|x_* - x_k\|_2^2 + 2\langle x_k - x_*, Fx_k - x_k\rangle + 2\langle x_k - x_*, x_k - x_*\rangle \\
&\leq 2\|x_* - x_k\|_2^2 + 2\langle x_k - x_*, Fx_k - x_k\rangle - 2\underbrace{\langle x_k - x_*, x_* - x_k\rangle}_{\|x_* - x_k\|_2^2} \\
&\leq 2\langle x_k - x_*, Fx_k - x_k\rangle \\
&\leq 2(Fx_k - x_k)^T(x^* - x_k)
\end{aligned}
$$

1.4.2

En déduire que pour tout $k \geq 0$, on a

$$\min_{0\leq \ell \leq k} \|Fx_\ell - x_\ell\|_2 \leq \frac{2}{\sqrt{k}}(\eta + \frac{D_k}{\eta}).$$

On rappelle que $u_\ell = Fx_\ell - x_\ell$:

$$
\begin{aligned}
\sum_{\ell=0}^{k} \|u_\ell\|_2^2 = \sum_{\ell=0}^{k} \|Fx_\ell - x_\ell\|_2^2 &\leq \sum_{\ell=0}^{k} 2u_\ell^T(x^* - x_\ell) \qquad \text{(D'après 4.a)} \\
&\leq 2\left(\eta + \frac{D_k}{\eta}\right)\sqrt{\sum_{\ell=0}^{k} \|u_\ell\|_2^2} \qquad \text{(D'après 3.b)}
\end{aligned}
$$

Si tout les $\|u_\ell\|_2^2$ sont nul alors l'inégalité est vérifiée, sinon on divise les deux parties de l'inégalité par $\sqrt{\|u_\ell\|_2^2}$ qui est strictement positif:

$$
\begin{aligned}
&0 \leq \sqrt{\sum_{\ell=0}^{k} \|u_\ell\|_2^2} \leq 2\left(\eta + \frac{D_k}{\eta}\right) \\
&\implies \sum_{\ell=0}^{k} \|u_\ell\|_2^2 \leq \left(2\left(\eta + \frac{D_k}{\eta}\right)\right)^2 \qquad \text{La fonction carré étant croissante sur } \mathbb{R}^+
\end{aligned}
$$

Le terme de gauche est une somme de $k+1$ termes ce qui implique que le plus petit terme de cette somme doit être à minima plus petit que la moyenne du terme de droite à savoir $\left(2\left(\eta + \frac{D_k}{\eta}\right)\right)^2/(k+1)$. En particulier $k+1$ étant plus quand grand que k on peut le remplacer dans l'expression:

$$
\begin{aligned}
\min_{0\leq \ell \leq k} \|u_\ell\|_2^2 &\leq \frac{(2\left(\eta + \frac{D_k}{\eta}\right)^2)}{k} \\
\min_{0\leq \ell \leq k} \|u_\ell\|_2 &\leq \frac{2}{\sqrt{k}}\left(\eta + \frac{D_k}{\eta}\right)
\end{aligned}
$$

Ce qui montre le résultat voulu car $u_\ell = Fx_\ell - x_\ell$.

1.4.3

Qu'en déduire sur $\min_{0\leq\ell\leq k}\|x_\ell - x_*\|_2$?

On rappelle que dans un contexte de programmation dynamique (i.e. la dynamique de transition p du MDP est disponible sous forme explicite) les itérations valeur (synchrones) pour l'évaluation d'une politique $\pi \in \Pi_0$ sont données par

$$v_{k+1} = B_\pi v_k, \quad k \geq 1, \quad (\mathrm{VI}_\pi^V)) \tag{$\mathrm{VI}_\pi^{(V)}$}$$

et pour le contrôle, les itérations sont données par

$$v_{k+1} = B_* v_k, \quad k \geq 1, \quad (\mathrm{VI}_*^V)) \tag{$\mathrm{VI}_\pi^{(V)}$}$$

respectivement.

1.5

En utilisant (Ada-FP), définir des algorithmes analogues aux intérations valeur synchrones classiques ($\mathrm{VI}_\pi^{(V)}$) et ($\mathrm{VI}_*^{(V)}$). On appellera (Ada-$\mathrm{VI}_\pi^{(V)}$) et (Ada-$\mathrm{VI}_\pi^{(V)}$) les algorithmes ainsi obtenus.

D'après la proposition 2.2.4.(iv), les opérateurs de Bellman B_π et B_* sont γ-lipschitz avec $\gamma \in]0,1[$ ce qui remplit les conditions requises pour obtenir les résultats de la partie plus haute. On obtient ainsi les algorithmes suivants:

$$v_{k+1} = v_k + \eta \frac{B_\pi v_k - v_k}{\sqrt{\sum_{\ell=0}^k \|B_\pi v_\ell - v_\ell\|_2^2}}, \quad k \geq 0, \tag{$\mathrm{Ada-VI}_\pi^{(V)}$}$$

$$v_{k+1} = v_k + \eta \frac{B_* v_k - v_k}{\sqrt{\sum_{\ell=0}^k \|B_* v_\ell - v_\ell\|_2^2}}, \quad k \geq 0, \tag{$\mathrm{Ada-VI}_*^{(V)}$}$$

2 Comparaison en pratique avec les algorithmes classiques

Choisir un MDP de taille raisonnable, c'est-à-dire dont on puisse calculer les fonctions valeurs $v_\pi, v_*(\pi \in \Pi_0)$ en un temps raisonnable. On pourra soit reprendre un MDP vu en TP, soit en trouver un dans un livre, sur internet, dans une librairie (e.g. Gymnasium) ou encore en créer un soi-même, mais pour cette partie, il est nécessaire de connaître les transitions de façon explicite.

2.1

Se donner une politique stationnaire $\pi \in \Pi_0$ quelconque, ainsi qu'une fonction valeur initiale $v_0 \in \mathbb{R}^{\mathcal{S}}$ tirée aléatoirement une fois pour toutes. Comparer en pratique la vitesse de convergence de $\text{Ada}-\text{VI}_\pi^{(V)}$ avec celle de $\text{VI}_\pi^{(V)}$. On pourra tracer, avec une échelle logarithmique en ordonnée, les quantités

$$\|v_k - v^\pi\|_\infty \text{ et } \|v_k - B_\pi v_k\|_\infty$$

en fonction de k. Essayer différentes valeurs pour $\eta > 0$.

Pour cette partie, nous étudirons le problème du labyrinthe vu en TP1 dont nous rappellerons brièvement l'essence. On part en haut à gauche d'un labyrinthe de 30×30 dont l'objectif est d'arriver en haut à droite. Deux murs occupant deux tiers de la hauteur (respectivement partant du haut et du bas) sont placés (à respectivement un tier et deux tiers de la largeur). Les autres cases ont une probabilité 0.15 d'être remplacées par des murs.

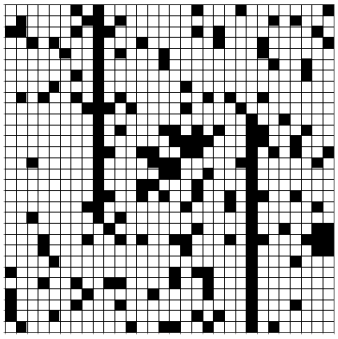
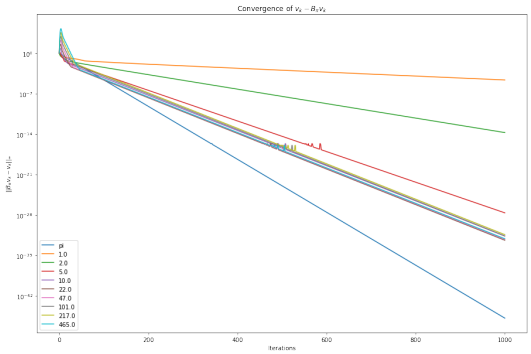
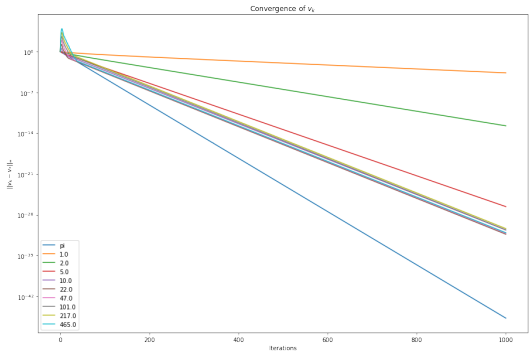


Figure 1: Exemple de labyrinthe

On estime les quantités voulu à partir de différente valeurs de η choisies entre 1 et 500 ainsi que pour l'algorithme "classique" du cours se basant sur l'itération $v_{k+1} = B_\pi v_k$. Pour chaque valeur de η, on fait tourner l'algorithme sur 1000 itérations avec une initialisation aléatoire (mais toujours la même). Pour estimer v_*, on fait tourner l'algorithme sur 10000 itérations en partant du principe qu'il aura convergé. Pour ce qui est de la politique initiale, celle-ci est décidée aléatoirement dès le départ.



(a) $\|v_k - v_*\|_\infty$

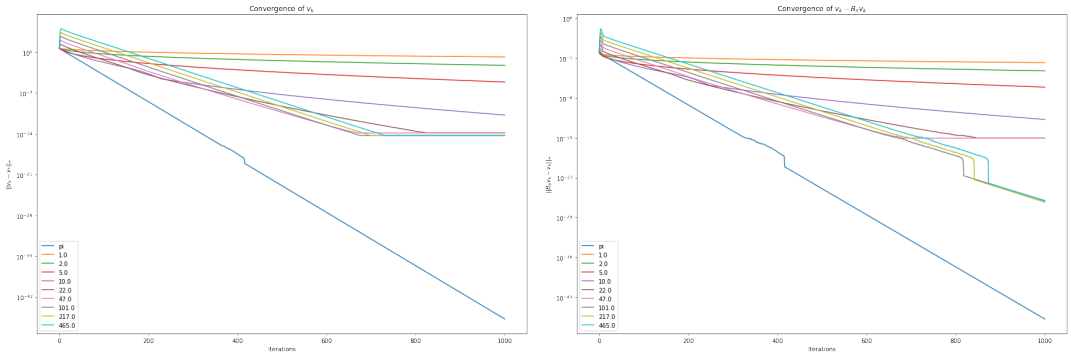


(b) $\|v_k - B_\pi v_k\|_\infty$

On remarque que l'algorithme initialement vu en cours est le plus performant. Tout les algorithmes semblent converger bien que cela se fasse à des vitesses différentes. On remarque notamment que que plus η est grand plus la convergence se fait rapidement pour un grand nombre d'itérations. Cependant, pour les premières itérations, un η plus petit permet une convergence plus rapide. Cela s'explique par le fait que le terme $\eta \frac{Fx_k - x_k}{\sqrt{\sum_{\ell=0}^{k} \|Fx_\ell - x_\ell\|_2^2}}$ est trop grand pour les premières itérations si η est grand.

2.2

Même question pour $(\text{Ada} - \text{VI}_*^{(V)})$ et $(\text{VI}_*^{(V)})$



(a) $\|v_k - v_*\|_\infty$ (b) $\|v_k - B_* v_k\|_\infty$

On obtient les mêmes résultats, la convergence est plus rapide pour les η dans un premier temps mais la convergence est plus lente pour un plus grand nombre d'itération. On remarque en outre que la converge des algorithmes Ada convergent vers des valeurs plus grandes que pour la question précédente même si, ces valeurs restent très faibles.

3 Extensions

Reprendre la démarche de la Section 2 en incorporant, par exemple, un ou plusieurs des aspects suivants :

- itérations de fonctions action-valeur,
- itérations asynchrones
- méthode d'apprentissage par renforcement utilisant des estimateurs stochastiques des opérateurs de Bellman
- approximation de la fonction valeur par une classe paramétrique
- variante de (Ada-FP) définie composante par composante

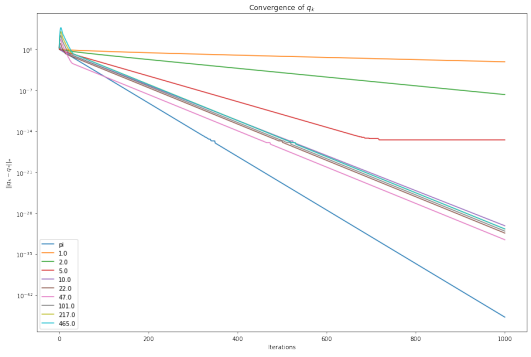
$$x_{k+1,j} = x_{k,j} + \eta \frac{F(x_k)_j - x_{k,j}}{\sqrt{\sum_{\ell=0}^{k} (F(x_\ell)_j - x_{\ell,j})^2}}, \quad 1 \leq j \leq d, \quad k \geq 0$$

3.1 Itérations de fonctions action-valeur

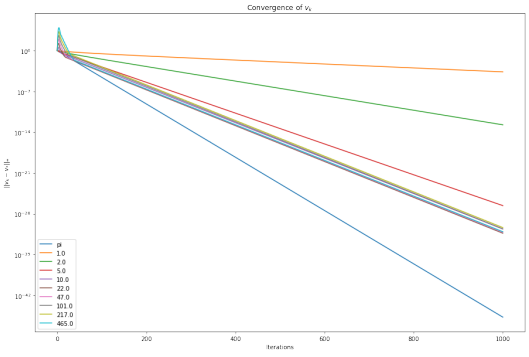
On reprend reprend l'algorithme Ada et on l'adapte pour les fonctions action-valeur de la façon suivante:

$$q_{k+1} = q_k + \eta \frac{B_\pi q_k - q_k}{\sqrt{\sum_{\ell=0}^{k} \|B_\pi q_\ell - q_\ell\|_2^2}}, \quad k \geq 0, \qquad (\text{Ada} - \text{VI}_*^{(V)})$$

De la même manière que précédemment on calcule les différences de normes en remplaçant les v_k par des q_k:



(a) $\|q_k - q_\pi\|_\infty$



(b) $\|q_k - B_\pi q_k\|_\infty$

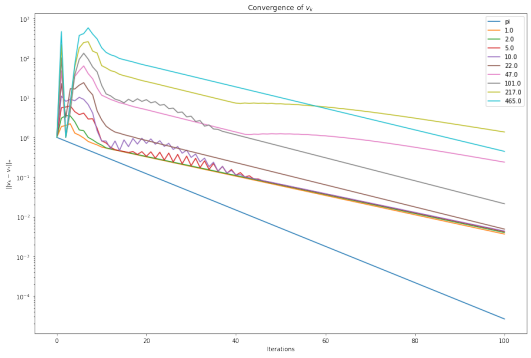
On observe les même résultats, les taux d'apprentissage plus élévé favorisent une convergence plus précises au fur et à mesure des itérations.

3.1.1 Variante de (Ada-FP) définie composante par composante

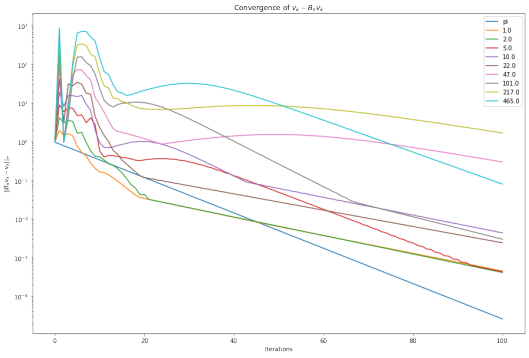
On définit la variante de (Ada-FP) définie composante par composante de la façon suivante:

$$v_{k+1,j} = v_{k,j} + \eta \frac{B_\pi(v_k)_j - v_{k,j}}{\sqrt{\sum_{\ell=0}^{k}(B_\pi(v_\ell)_j - v_{\ell,j})^2}}, \quad 1 \leq j \leq d, \quad k \geq 0$$

Le processus étant plus long à tourner, on se limitera à cent itération, ce qui est suffisant pour voir les tendances se former bien que rendant la comparaison avec les autres méthodes plus laborieuses.



(a) $\|v_k - v_\pi\|_\infty$



(b) $\|v_k - B_\pi v_k\|_\infty$

On remarque comme précédemment qu'un η trop grand entraîne des variations trop importantes dans les premières itérations. On remarque de plus qu'un learning rate trop élevé n'entraine pas des convergence spécialement plus rapide.

Au final, les différents algorithmes n'entrainent pas spécialement des résultats plus probant, cela peut être du au fait que le MDP est peut être trop simple et la convergence est trop rapide.