**Master MVA: Apprentissage par renforcement** **Lecture: 2**

# Algorithmes d'apprentissage par renforcement

*Professeur: Rémi Munos* *http://researchers.lille.inria.fr/∼munos/master-mva/*

**Références bibliographiques:**


- Sutton et Barto, *Reinforcement Learning, an introduction*, 1998.
- Bertsekas et Tsitsiklis, *Neuro-Dynamic Programming*, 1996.
- Livre PDMIA, chapitre 2.
- Sutton, *Learning to predict by the method of temporal differences*, 1988.
- Watkins et Dayan, *Q-learning*, 1992.
- Jaakola, Jorda et Singh, *On the convergence of Stochastic Iterative Dynamic Programming Algorithms*, 1994.


**Problématique de l'apprentissage par renforcement**: Les probabilités de transition et les récompenses sont initialement inconnues de l'agent. Différentes hypothèses sont possibles pour explorer l'environnement tout en permettant la convergence vers une politique optimale:

- **Apprentissage en-ligne:** l'agent se trouve en un état $x_t$, choisit une action $a_t$, observe une récompense $r_t = r(x_t, a_t)$ et se retrouve en $x_{t+1} \sim p(\cdot|x_t, a_t)$, et recommence. Pour espérer apprendre une stratégie optimale il faut s'assurer que le PDM permet de revenir à un état où on peut obtenir le taux optimal de récompenses même si on fait des actions (exploration) sous-optimales au début. Voir algorithme UCRL [Auer et al. 2006-2009]
- **Apprentissage épisodique:** On génére une séquence d'épisodes où à chaque étape on on suit une politique pendant un certain temps (éventuellement aléatoire), ce qui nous permet d'explorer, puis on réinitialise l'état et on passe à l'épisode suivant.
- **Apprentissage avec modèle génératif:** On dispose d'un simulateur de l'environnement (type boîte noire) qui permet, en entrant $x$ et $a$, de générer une récompense $r = r(x, a)$ et un état suivant $y \sim p(\cdot|x, a)$.

# 1 Préliminaires: outils statistiques

## 1.1 Rappels sur les modes de convergence de variables aléatoires

Soient $X, X_1, X_2, \ldots$ des v.a. Alors

**(a)** $X_n$ converge presque sûrement vers $X$, $X_n \xrightarrow{p.s.} X$, si

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1,$$

**(b)** $X_n$ converge vers $X$ en probabilité, $X_n \xrightarrow{P} X$, si pour tout $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0,$$

**(c)** $X_n$ converge vers $X$ en loi (ou en distribution), $X_n \xrightarrow{D} X$, si pour toute fonction $f$ continue, bornée,

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

On a les propriétés: $X_n \xrightarrow{p.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$.

## 1.2 Inégalité de Chernoff-Hoeffding

**Inégalité de Markov:**

*Proposition* 1. Soit $Y$ une v.a. positive. Alors pour tout $a > 0$,

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}Y}{a}.$$

*Proof.* On a $\mathbb{P}(Y \geq a) = \mathbb{E}[\mathbf{1}\{Y \geq a\}] = \mathbb{E}[\mathbf{1}\{Y/a \geq 1\}] \leq \mathbb{E}[Y/a]$. □

**Inégalité de Hoeffding:**

*Proposition* 2. Soit $Y$ une v.a. centrée à valeurs dans $[a, b]$. Alors pour tout $s \in \mathbb{R}$,

$$\mathbb{E}[e^{sY}] \leq e^{s^2(b-a)^2/8}.$$

*Proof.* Par convexité de l'exponentielle, on a pour tout $a \leq x \leq b$,

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}.$$

En notant $p = -a/(b-a)$ on a

$$\begin{aligned} \mathbb{E}e^{sx} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= (1-p+pe^{s(b-a)})e^{-ps(b-a)} = e^{\phi(u)} \end{aligned}$$

avec $u = s(b-a)$ et $\phi(u) = -pu + \log(1-p+pe^u)$. La dérivée de $\phi$ est $\phi'(u) = -p + \frac{p}{p+(1-p)e^{-u}}$. De plus $\phi(0) = \phi'(0) = 0$. De plus, $\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq 1/4$.

Donc d'après le théorème de Taylor, il existe $\theta \in [0, u]$ tel que

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

□

**Inégalité de Chernoff-Hoeffding**

*Proposition* 3. Soient $X_i \in [a_i, b_i]$ variables aléatoires indépendantes. Soit $\mu_i = \mathbb{E}X_i$. Alors

$$\mathbb{P}\big(\big|\sum_{i=1}^{n} X_i - \mu_i\big| \geq \epsilon\big) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}. \tag{1}$$

*Proof.* On a:

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} X_i - \mu_i \geq \epsilon) &= \mathbb{P}(e^{s\sum_{i=1}^{n} X_i-\mu_i} \geq e^{s\epsilon}) \\
&\leq e^{-s\epsilon}\mathbb{E}[e^{s\sum_{i=1}^{n} X_i-\mu_i}], \text{par Markov} \\
&= e^{-s\epsilon}\prod_{i=1}^{n}\mathbb{E}[e^{s(X_i-\mu_i)}], \text{ par indépendence des v.a.} \\
&\leq e^{-s\epsilon}\prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8}, \text{ par Hoeffding} \\
&= e^{-s\epsilon+s^2\sum_{i=1}^{n}(b_i-a_i)^2/8}
\end{aligned}$$

En choisissant $s = 4\epsilon/\sum_{i=1}^{n}(b_i-a_i)^2$ on déduit $\mathbb{P}\big(\sum_{i=1}^{n} X_i - \mu_i \geq \epsilon\big) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$. En refaisant le même calcul pour $\mathbb{P}\big(\sum_{i=1}^{n} X_i - \mu_i \leq -\epsilon\big)$ on déduit (1). □

# 2 Approximation stochastique pour l'estimation d'une moyenne

## 2.1 Monte-Carlo

Soit $X$ une variable aléatoire de moyenne $\mu = \mathbb{E}[X]$ et de variance $\sigma^2 = \mathrm{Var}[X]$. Soient $x_n \sim X$ des réalisations i.i.d. Définissons la moyenne empirique

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Alors $\mathbb{E}[\mu_n] = \mu$, $\mathrm{Var}[\mu_n] = \frac{\mathrm{Var}[X]}{n}$, et on a

- **Loi faible des grands nombres:** $\mu_n \xrightarrow{P} \mu$.
- **Loi forte des grands nombres**: $\mu_n \xrightarrow{p.s.} \mu$.
- **Théorème Central Limite**: $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} \mathcal{N}(0, \mathrm{Var}[X])$.

## 2.2 Estimation d'une moyenne inconnue

Soit $X$ une variable aléatoire à valeurs dans $[0, 1]$, de moyenne $\mu = \mathbb{E}[X]$ inconnue que l'on souhaite estimer. Soit $x_n \sim X$ des réalisations i.i.d. Soit l'estimateur $\mu_n$ construit selon: $\mu_1 = x_1$, et pour $n > 1$,

$$\mu_n = (1 - \eta_n)\mu_{n-1} + \eta_n x_n \tag{2}$$

où $(\eta_n)$ sont des *pas d'apprentissage*.

*Proposition* 4. Sous l'hypothèse que les pas sont positifs et vérifient

$$\sum_{n\geq 0} \eta_n = \infty, \tag{3}$$

$$\sum_{n\geq 0} \eta_n^2 < \infty, \tag{4}$$

alors on a $\mu_n \xrightarrow{p.s.} \mu$. (on dit que l'estimateur $\mu_n$ est **consistant**).

**Remarque:** Les pas $\eta_n = \frac{1}{n}$ satisfont les conditions précédentes, et on a alors la moyenne empirique $\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. La **loi forte des grands nombres** est donc une conséquence de cette proposition.

**Lemme de Borel-Cantelli:** Si $(E_n)_{n\geq 1}$ est une suite d'événements tels que $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, alors la probabilité qu'une infinité d'entre eux se réalisent simultanément est nulle.

*Proof.* Si la probabilité qu'une infinité $\mathcal{E}$ d'entre eux se réalisent était $\geq a > 0$, alors la probabilité que chaque $E_i \in \mathcal{E}$ se réalise serait $\geq a$ et donc $\sum_{n\geq 1} \mathbb{P}(E_n) \geq \sum_{E_n \in \mathcal{E}} a = \infty$. □

*Preuve de la proposition 4.* Preuve partielle. Nous allons considérer uniquement des pas $\eta_n = n^{-\alpha}$ avec $\alpha \leq 1$ (afin de satisfaire (3)) et $\alpha > 1/2$ (afin de satisfaire (4)). La preuve générale peut être trouvée dans *On Stochastic Approximation*, Dvoretzky, 1956.

**Cas** $\alpha = 1$ (i.e. $\mu_n$ est la moyenne empirique des $x_i$). D'après l'inégalité de Chernoff-Hoeffding, on a

$$\mathbb{P}\big(\big|\mu_n - \mu\big| \geq \epsilon\big) \leq 2e^{-2n\epsilon^2}. \tag{5}$$

Nous utilisons le lemme de Borel-Cantelli avec les évènements $E_n = \{\big|\mu_n - \mu\big| \geq \epsilon\}$. On déduit de (5) que $\sum_{n\geq 1} \mathbb{P}(E_n)$ est finie et donc qu'avec probabilité 1, il n'existe qu'un nombre fini d'instants $n$ tels que $\big|\mu_n - \mu\big| \geq \epsilon$. On choisit une suite d'$\epsilon_k \to O$. Donc pour tout $\epsilon_k$, il existe $n_k$ tel que pour $\mathbb{P}(\forall n \geq n_k, \big|\mu_n - \mu\big| \leq \epsilon_k) = 1$. L'union dénombrable d'évènements de mesure nulle est toujours de mesure nulle, donc

$$\mathbb{P}(\forall k, \exists n_k, \forall n \geq n_k, \big|\mu_n - \mu\big| \leq \epsilon_k) = 1,$$

ce qui est la définition que $\lim_{n\to\infty} \mu_n = \mu$ presque sûrement.

**Cas** $1/2 < \alpha < 1$**.** Nous écrivons

$$\mu_n = \sum_{i=1}^{n} \lambda_i x_i, \text{ avec } \sum_{i=1}^{n} \lambda_i = 1, \tag{6}$$

avec $\lambda_i = \eta_i \prod_{j=i+1}^{n} (1 - \eta_j)$. On a

$$\log \lambda_i = \log \eta_i + \sum_{j=i+1}^{n} \log(1 - \eta_j) \leq \log \eta_i - \sum_{j=i+1}^{n} \eta_j$$

donc $\lambda_i \leq \eta_i e^{-\sum_{j=i+1}^{n} \eta_j}$.

Nous bornons maintenant la somme des amplitudes au carré: pour tout $1 \leq m \leq n$,

$$\begin{aligned}\sum_{i=1}^{n} \lambda_i^2 &\leq \sum_{i=1}^{n} \eta_i^2 e^{-2\sum_{j=i+1}^{n}\eta_j} \\ &\leq \sum_{i=1}^{m} e^{-2\sum_{j=i+1}^{n}\eta_j} + \sum_{i=m+1}^{n} \eta_i^2 \\ &\leq me^{-2(n-m)\eta_n} + (n-m)\eta_m^2 \\ &= me^{-2(n-m)n^{-\alpha}} + (n-m)m^{-2\alpha}\end{aligned}$$

On choisit $m = n^\beta$ avec $\beta = (1+\alpha/2)/2$ (donc $1 - 2\alpha\beta = 1/2 - \alpha$):

$$\sum_{i=1}^{n} \lambda_i^2 \leq ne^{-2(1-n^{-1/4})n^{1-\alpha}} + n^{1/2-\alpha} \leq 2n^{1/2-\alpha}$$

pour $n$ assez grand.

A partir de l'écriture de $\mu_n$ selon (6), on utilise l'inégalité de Chernoff-Hoeffding pour déduire une majoration de $\mu_n - \mu$ en forte probabilité:

$$\mathbb{P}\big(|\mu_n - \mu| \geq \epsilon\big) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}\lambda_i^2}} \leq e^{-\frac{\epsilon^2}{n^{1/2-\alpha}}}.$$

pour $n$ assez grand.

On utilise ensuite le même argument (lemme de Borel-Cantelli) que dans le cas $\alpha = 1$ (car $\sum_{n\geq 1}\mathbb{P}(|\mu_n - \mu| \geq \epsilon) < \infty$) pour en déduire la convergence presque sûre de $\mu_n$ vers $\mu$. □

## 2.3 Autres algorithmes d'approximation stochastique

**AS de Robbins-Monro (1951):** Résoudre le système $f(x) = 0$ lorsque l'évaluation de la fonction $f$ est bruitée, i.e. en $x_n$ on observe $y_n = f(x_n) + b_n$ où $b_n$ est un bruit indépendant centré. Soit:

$$x_{n+1} = x_n - \eta_n y_n.$$

Supposons $f$ croissante et notons $x^*$ la solution. Sous les mêmes hypothèses $\sum \eta_n = \infty$, $\sum \eta_n^2 < \infty$, on a la convergence $x_n \xrightarrow{p.s.} x^*$.

**AS de Kiefer-Wolfowitz (1952):** On veut trouver le minimum local d'une fonction $f$ lorsque l'évaluation de son gradient est bruité, i.e. en $x_n$, on dispose de $g_n = \nabla f(x_n) + b_n$. Alors:

$$x_{n+1} = x_n - \eta_n g_n.$$

alors sous les mêmes hypothèses sur $(\eta_n)$ (et une condition de positivité de la Hessienne $\nabla^2 f$), on a $x_n \xrightarrow{p.s.} x^*$ avec $x^*$minimum de $f$. Il s'agit d'un algorithme de **gradient stochastique.**

## 2.4 Approximation stochastique de point fixe

Lorsque $\mathcal{T}$ est un opérateur contractant, nous avons vu que l'algorithme d'itérations sur les valeurs: $V_{n+1} = \mathcal{T}V_n$ génère une séquence de fonctions qui converge vers $V^*$.

Maintenant supposons qu'on ne puisse pas calculer exactement $\mathcal{T}V$ (rappelons que dans le cas de la programmation dynamique, cet opérateur fait intervenir le calcul d'espérances) mais que l'on sache calculer $\hat{\mathcal{T}}V = \mathcal{T}V + b$, où $b$ est un "bruit" centré: $\mathbb{E}b = 0$. Alors on voudrait utiliser un algorithme stochastique pour calculer le point fixe de $\mathcal{T}$.

Soit $\mathcal{T} : \mathbb{R}^N \to \mathbb{R}^N$ contraction en norme $L_\infty$ pondérée ($\exists \beta < 1$, $\exists \mu \in \mathbb{R}^N_{+,*}$, $\forall V_1, V_2 \in \mathbb{R}^N$, $||\mathcal{T}V_1 - \mathcal{T}V_2||_\mu \leq \beta ||V_1 - V_2||_\mu$.) Considérons l'AS défini, pour $x \in \{1, \ldots, N\}$, par:

$$V_{n+1}(x) = (1 - \eta_n(x))V_n(x) + \eta_n(x)(\mathcal{T}V_n(x) + b_n(x)),$$

Notons $\mathcal{F}_n = \{V_0, \ldots, V_n, b_0, \ldots, b_{n-1}, \eta_0, \ldots, \eta_n\}$ l'histoire de l'algo jusqu'à l'instant $n$.

*Proposition* 5. On suppose que le bruit est centré ($\mathbb{E}[b_n(x)|\mathcal{F}_n] = 0$) et de variance majorée selon $\mathbb{E}[b_n^2(x)|\mathcal{F}_n] \leq c(1 + ||V_n||^2)$ (pour une constante $c$) et que les pas sont positifs et satisfont en tout $x$: $\sum_{n\geq 0} \eta_n(x) = \infty$ et $\sum_{n\geq 0} \eta_n^2(x) < \infty$. Alors, $\forall x$,

$$V_n(x) \xrightarrow{p.s.} V^*(x).$$

*Proof.* La preuve n'est pas reproduite ici (un peu longue). Mentionnons qu'il y a plusieurs approches possibles pour prouver ce résultat. Une preuve utilisant la notion de fonction de Lyapunov peut être trouvée dans le livre *Neuro Dynamic Programming de Bertsekas et Tsitsiklis*, 1996. La méthode la plus générale est basée sur une approximation en temps continue: ODE (Ordinary Differential Equation), voir aussi le livre de Kushner et Yin *Stochastic Approximation and Recursive Algorithms and Applications*, 2003. Une première preuve a été rédigée simultanément dans les articles de Jaakola, Jordan et Singh, *On the convergence of Stochastic Iterative Dynamic Programming Algorithms*, 1994, et de Tsitsiklis, *Asynchronous Stochastic Approximation and Q-Learning*, 1994.

□

# 3 Evaluation d'une politique $\pi$ par Monte-Carlo

Considérons un problème à critère non-actualisé (notons 0 l'état absorbant). Soit $\pi$ une politique propre. On désire évaluer

$$V^\pi(x) = \mathbb{E}\big[\sum_{k=0}^{K-1} r^\pi(x_k) \,|\, x_0 = x; \pi\big],$$

où $K$ est le temps d'atteinte de l'état absorbant.

On génère $n$ trajectoires $(x_0^i = x, x_1^i, \ldots, x_{K_i}^i = 0)_{i\leq n}$ à partir d'un état initial $x$ et on forme la moyenne empirique de la somme des récompenses reçues le long des trajectoires:

$$V_n(x) = \frac{1}{n}\sum_{i=1}^{n}\big[r^\pi(x_0^i) + r^\pi(x_1^i) + \cdots + r^\pi(x_{K_i-1}^i)\big]$$

Alors $V_n(x)$ est un estimateur non biaisé de $V^\pi(x)$ et d'après la loi forte des grands nombres,

$$V_n(x) \xrightarrow{p.s.} V^\pi(x).$$

**Remarque:** La trajectoire $(x_0, x_1, x_2, \ldots, x_K)$ contient toute sous-trajectoire $(x_k, x_{k+1}, \ldots, x_K)$, donc la somme $r^\pi(x_k) + \cdots + r^\pi(x_{K-1})$ fournit un estimateur de $V^\pi(x_k)$. Donc une seule trajectoire peut fournir un estimateur pour tous les $\{V^\pi(x_k)\}$. Mais que se passe t-il si un état $x$ est visité plusieurs fois au cours d'une même trajectoire?

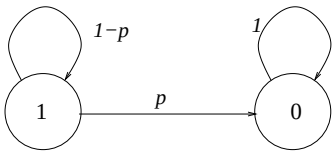
**Monte-Carlo à chaque visite:** Les sous trajectoires issues de $x$ sont dépendantes car portions d'une même trajectoire. De plus, le nombre d'échantillons de récompense accumulée est une variable aléatoire –> leur moyenne peut être biaisée.

**Monte-Carlo à première visite:** On ne considère que la première sous-trajectoire visitant $x$. Cet estimateur est non-biaisé, mais on ne dispose que d'une réalisation.

Mais quelle méthode donne la meilleure erreur quadratique moyenne?

$$\mathbb{E}\big[[\widehat{V} - V]^2\big] = \underbrace{\big(\mathbb{E}[\widehat{V}] - V\big)^2}_{\text{Biais}^2} + \underbrace{\mathbb{E}\big[\big(\widehat{V} - \mathbb{E}[\widehat{V}]\big)^2\big]}_{\text{Variance}}$$

**Exemple à 2 états:** Considérons l'exemple suivant:




On obtient une récompense 1 lorsqu'on est dans l'état 1. Les trajectoires sont du type $(x_0, x_1, \ldots, x_K)$ avec $x_k = 1$ pour $0 \leq k < K$ et $x_K = 0$.
$K$ est une v.a. géométrique de moyenne $\mathbb{E}[K] = \frac{1}{p}$.
On a $V^\pi(1) = 1 + (1-p)V^\pi(1) = \frac{1}{p}$.

**Monte-Carlo à première visite:** Le gain obtenu le long de la trajectoire est $K$, donc

$$\mathbb{E}[K] = \frac{1}{p} = V^\pi(1) : \textbf{L'estimateur est non-biaisé.}$$

L'erreur quadratique moyenne (= variance de $K$):

$$\mathbb{E}\big[\big(K - \frac{1}{p}\big)^2\big] = \frac{1}{p^2} - \frac{1}{p},$$

**Monte-Carlo à chaque visite:** Pour une trajectoire, le gain accumulé moyen sur toutes les sous-trajectoires est $\frac{1}{K}\sum_{k=0}^{K-1}(K-k) = \frac{K+1}{2}$. En moyenne:

$$\mathbb{E}\big[\frac{K+1}{2}\big] = \frac{1+p}{2p} \neq V^\pi(1) : \textbf{ l'estimateur est biaisé}.$$

Maintenant, si on génère plusieurs trajectoires et que $K_i$ est le temps d'arrêt de la $i^e$ trajectoire, alors le gain empirique moyen est

$$\frac{\sum_{i=1}^n \sum_{k=0}^{K_i-1} K_i - k}{\sum_{i=1}^n K_i} = \frac{\sum_{i=1}^n K_i(K_i+1)}{2\sum_{i=1}^n K_i} \xrightarrow{p.s.} \frac{\mathbb{E}[K^2] + \mathbb{E}[K]}{2\mathbb{E}[K]} = \frac{1}{p} = V^\pi(1)$$

**L'estimateur est consistant**. Et l'erreur quadratique moyenne:

$$\mathbb{E}\big[\big(\frac{K+1}{2} - \frac{1}{p}\big)^2\big] = \frac{1}{2p^2} - \frac{3}{4p} + \frac{1}{4},$$

est inférieure à celle de MC à première visite.

**MC à chaque visite ou à première visite?** Considérons une chaîne de Markov arbitraire et un état particulier $x$. Observons cette chaîne seulement aux instants où elle vaut $x$ ou 0.

- Chaîne de Markov réduite d'espace $\{x, 0\}$.

- La probabilité $p(x|x)$ = probabilité dans la chaîne initiale de revenir en $x$ partant de $x$, et $p(0|x)$ = probabilité d'arriver en 0 sans repasser par $x$.
- La récompense $r(x)$= somme moyenne des récompenses obtenues lors d'une trajectoire partant de $x$.

–> L'exemple à 2 états capture l'essence des méthodes MC:

- Monte-Carlo avec visites multiples: estimateur biaisé, mais consistant.
- Monte-Carlo avec première visite: estimateur non-biaisé, mais erreur quadratique moyenne supérieure.

Quand l'espace est grand, peu de chances de repasser par les mêmes états –> méthodes comparables.

# 4 Algorithmes Stochastiques pour l'évaluation d'une politique

## 4.1 AS pour l'estimation de moyenne

La fonction valeur pour une politique $\pi$ (supposée propre) est:

$$V^\pi(x_0) = \mathbb{E}\Big[\sum_{i\geq 0} r^\pi(x_i)\big|\pi\Big].$$

**Algorithme TD(1):** Après avoir observé une trajectoire $(x_0, x_1, \ldots, x_K = 0)$, on itère les valeurs $V_n$ des états $(x_k)$ selon

$$V_{n+1}(x_k) = (1-\eta_n(x_k))V_n(x_k) + \eta_n(x_k)[r^\pi(x_k) + r^\pi(x_{k+1}) + \cdots + r^\pi(x_{K-1})]. \tag{7}$$

Puisque

$$\mathbb{E}\big[r^\pi(x_k) + r^\pi(x_{k+1}) + \cdots + r^\pi(x_{K-1})|x_k\big] = V^\pi(x_k),$$

alors on peut appliquer la proposition 4 (AS pour l'estimation d'une moyenne inconnue) et déduire que, si tous les états sont visités par une infinité de trajectoires et que les pas vérifient $\forall x$, $\sum_n \eta_n(x) = \infty$ et $\sum_n \eta_n(x)^2 < \infty$, alors $V_n(x) \overset{p.s.}{\to} V^\pi(x)$ pour tout $x \in X$.

Remarque: l'itération (7) peut se réécrire:

$$V_{n+1}(x_k) \quad = \quad V_n(x_k) + \eta_n(x_k)[d_k + d_{k+1} + \cdots + d_{K-1}]$$

où $d_k = r^\pi(x_k) + V_n(x_{k+1}) - V_n(x_k)$ est la **différence temporelle** d'évaluation par $V_n$ lors de la transition $x_k \to x_{k+1}$. On en déduit une implémentation incrémentale.

**Algorithme incrémental**: une fois la transition $x_l$ à $x_{l+1}$ observée, la différence temporelle $d_l$ est connue, et ainsi, pour tout $k \leq l$, on itère

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k)d_l.$$

La différence temporelle $d_k = r^\pi(x_k) + V_n(x_{k+1}) - V_n(x_k)$ fournit un indicateur de cohérence de l'estimation de $V_n$ lors de la transition $x_k \to x_{k+1}$.

Remarquons qu'asymptotiquement $V_n$ converge vers $V^\pi$ et que la différence temporelle $d_k = r^\pi(x_k) + V^\pi(x_{k+1}) - V^\pi(x_k)$ pour $V^\pi$ satisfait:

$$\mathbb{E}[r^\pi(x_k)+V^\pi(x_{k+1})-V^\pi(x_k)|x_k = x] = r(x,\pi(x))+\sum_y p(y|x,\pi(x))V^\pi(y)-V^\pi(x) = \mathcal{T}^\pi V^\pi(x)-V^\pi(x) = 0.$$

## 4.2 AS pour l'estimation du point fixe de $\mathcal{T}^\pi$

Puisque $V^\pi$ est le point fixe de $\mathcal{T}^\pi$ on peut utiliser un algorithme d'AS pour l'estimation du point fixe de $\mathcal{T}^\pi$, selon la Proposition 5.

**Idée**: un estimateur non-biaisé de $\mathcal{T}^\pi V(x)$ est $r^\pi(x_k) + V(x_{k+1})|x_k = x$. En effet, on a

$$\mathbb{E}[r^\pi(x_k) + V(x_{k+1})|x_k = x] = r(x, \pi(x)) + \sum_y p(y|x, \pi(x))V(y) = \mathcal{T}^\pi V(x).$$

**Algorithme TD(0):** Nous construisons une séquence de fonctions $V_n$ par l'algorithme d'AS suivant: après l'observation d'une trajectoire $(x_0, x_1, \ldots, x_K = 0)$, on met à jour $V_n$ aux états $(x_k)_{0 \le k < K}$ selon

$$\begin{aligned} V_{n+1}(x_k) &= (1 - \eta_n(x_k))V_n(x_k) + \eta_n(x_k)[r^\pi(x_k) + V_n(x_{k+1})] \\ &= V_n(x_k) + \eta_n(x_k) \underbrace{\left[r^\pi(x_k) + V_n(x_{k+1}) - V_n(x_k)\right]}_{d_k}. \end{aligned}$$

## 4.3 Temporal Differences TD($\lambda$)

Définissons l'opérateur de Bellman $\mathcal{T}_\lambda^\pi$ comme une combinaison convexe des opérateur de Bellman à $m$-pas $(\mathcal{T}^\pi)^m$ pondérés par des coefficients qui dépendent de $\lambda < 1$:

$$\mathcal{T}_\lambda^\pi = (1 - \lambda) \sum_{m \ge 0} \lambda^m (\mathcal{T}^\pi)^{m+1}. \tag{8}$$

En notant $P^\pi$ la matrice de probabilités de la chaîne de Markov induite par $\pi$, on a

$$\begin{aligned} T_\lambda^\pi V &= (1 - \lambda)\Big[\sum_{m \ge 0} \lambda^m \sum_{i=0}^{m} (P^\pi)^i\Big] r^\pi + (1 - \lambda) \sum_{m \ge 0} \lambda^m (P^\pi)^{m+1} V \\ &= \Big[\sum_{m \ge 0} \lambda^m (P^\pi)^m\Big] r^\pi + (1 - \lambda) \sum_{m \ge 0} \lambda^m (P^\pi)^{m+1} V \\ &= (I - \lambda P^\pi)^{-1} r^\pi + (1 - \lambda) \sum_{m \ge 0} \lambda^m (P^\pi)^{m+1} V. \end{aligned}$$

Donc si la politique $\pi$ est propre, il existe $\mu \in \mathbb{R}^N$ positif et $\beta < 1$ tel que $||P^\pi V||_\mu \le \beta ||V||_\mu$ (donc $||(P^\pi)^m V||_\mu \le \beta^m ||V||_\mu$) et

$$||(1 - \lambda) \sum_{m \ge 0} \lambda^m (P^\pi)^{m+1} V||_\mu \le (1 - \lambda) \sum_{m \ge 0} \lambda^m ||(P^\pi)^{m+1} V||_\mu \le \frac{(1 - \lambda)\beta}{1 - \beta\lambda} ||V||_\mu,$$

et $\mathcal{T}$ est une contraction par rapport à la norme $L_{\infty,\mu}$ (de coefficient de contraction $\frac{(1-\lambda)\beta}{1-\beta\lambda} \in [0, \beta]$), et son point fixe est $V^\pi$. Un algorithme d'AS pour trouver le point fixe de $\mathcal{T}_\lambda^\pi$ s'écrit:

**Algorithme TD($\lambda$)** [Sutton, 1988]: Après l'observation d'une trajectoire $(x_0, x_1, \ldots, x_K = 0)$, on met à jour $V_n$ aux états $(x_k)_{0 \le k < K}$ selon:

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) \sum_{l=k}^{K-1} \lambda^{l-k} d_l, \text{ où } d_l = r^\pi(x_l) + V_n(x_{l+1}) - V_n(x_l).$$

En effet, $\sum_{l=k}^{K-1} \lambda^{l-k} d_l | x_k = x$ est un estimateur non-biaisé de $\mathcal{T}_\lambda^\pi V_n(x) - V_n(x)$, car pour $l \geq k$,

$$\begin{aligned}
\mathbb{E}[d_l | x_k = x] &= \mathbb{E}\Big[r^\pi(x_l) + V_n(x_{l+1}) - V_n(x_l)\Big| x_k = x\Big] \\
&= \mathbb{E}\Big[\sum_{i=k}^{l} r^\pi(x_i) + V_n(x_{l+1})\Big| x_k = x\Big] - \mathbb{E}\Big[\sum_{i=k}^{l-1} r^\pi(x_i) + V_n(x_l)\Big| x_k = x\Big] \\
&= (\mathcal{T}^\pi)^{l-k+1} V_n(x) - (\mathcal{T}^\pi)^{l-k} V_n(x)
\end{aligned}$$

et donc

$$\begin{aligned}
\mathbb{E}[\sum_{l=k}^{K-1} \lambda^{l-k} d_l | x_k = x] &= \sum_{l=k}^{K-1} \lambda^{l-k} \Big[(\mathcal{T}^\pi)^{l-k+1} V_n(x) - (\mathcal{T}^\pi)^{l-k} V_n(x)\Big] \\
&= \sum_{m \geq 0} \lambda^m \Big[(\mathcal{T}^\pi)^{m+1} V_n(x) - (\mathcal{T}^\pi)^m V_n(x)\Big] \\
&= (1-\lambda) \sum_{m \geq 0} \lambda^m (\mathcal{T}^\pi)^{m+1} V_n(x) - V_n(x) = \mathcal{T}_\lambda^\pi V_n(x) - V_n(x)
\end{aligned}$$

**TD($\lambda$)**: l'impact des différences temporelles des transitions futures sur l'estimation de la valeur de l'état courant est actualisé par $\lambda$. TD($\lambda$) réalise un compromis entre

- **TD(0)** (AS pour estimer le point fixe de $T^\pi$): $V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) d_k$.
- **TD(1)** (AS pour estimer la moyenne): $V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) \sum_{l \geq k} d_l$.

**Convergence de TD($\lambda$)** Considérons l'algorithme TD($\lambda$) qui, après l'observation de trajectoires $(x_0, x_1, \ldots, x_K = 0)$ obtenues en suivant la politique $\pi$ (supposée propre), itère $V_n$ aux états $(x_k)_{0 \leq k < K}$ selon

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) \sum_{l \geq k} \lambda^{l-k} d_l$$

avec $d_l = r(x_l) + V_n(x_{l+1}) - V_n(x_l)$.

*Proposition* 6. Supposons que tous les états sont visités par une infinité de trajectoires et que les pas $\eta$ sont positifs et satisfont, pour tout $x$, $\sum_{n \geq 0} \eta_n(x) = \infty$, $\sum_{n \geq 0} \eta_n^2(x) < \infty$, alors $V_n \xrightarrow{p.s.} V^\pi$.

*Proof.* La preuve est une conséquence immédiate de l'application de la Proposition 5 pour la convergence de l'AS pour estimer le point fixe de l'opérateur contractant $T_\lambda^\pi$. □

**Implémentation de TD($\lambda$)** Définissons la *trace d'éligibilité* $z_n \in \mathbb{R}^N$.

Une fois la transition $x_k$ à $x_{k+1}$ observée, on calcule la différence temporelle $d_k = r^\pi(x_k) + V_n(x_{k+1}) - V_n(x_k)$ et on met à jour la trace d'éligibilité:

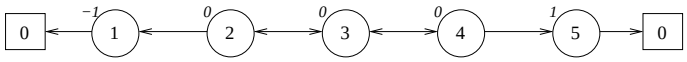
$$z_n(x) = \begin{cases} \lambda z_{n-1}(x) & \text{si } x \neq x_k \\ 1 + \lambda z_{n-1}(x) & \text{si } x = x_k \\ 0 & \text{si } x_k = 0 \text{ (remise à 0 des traces)} \end{cases}$$

et on itère: pour tout $x$,

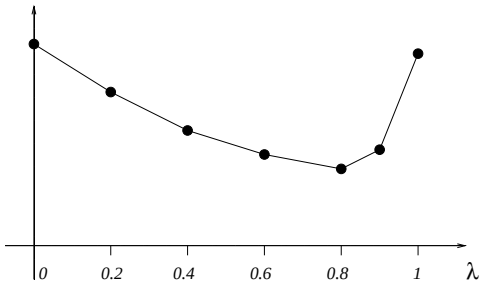
$$V_{n+1}(x) = V_n(x) + \eta_n(x) z_n(x) d_n.$$

D'autres variantes possibles (traces avec remplacement)

**Choix de $\lambda$:** Exemple de la chaîne linéaire:




Erreur quadratique moyenne (pour 100 trajectoires):




- $\lambda < 1$ permet de réduire la variance des estimateurs par rapport à $\lambda = 1$.
- $\lambda > 0$ permet de propager plus rapidement les récompenses par rapport à $\lambda = 0$.

**TD($\lambda$) dans le cas actualisé** Opérateur de Bellman ($\lambda$):

$$\begin{aligned}
\mathcal{T}V_n(x_0) &= (1-\lambda)\mathbb{E}\big[\sum_{k\geq 0}\lambda^k\big(\sum_{i=0}^{k}\gamma^i r^\pi(x_i) + \gamma^{k+1}V_n(x_{k+1})\big)\big] \\
&= \mathbb{E}\big[(1-\lambda)\sum_{i\geq 0}\gamma^i r^\pi(x_i)\sum_{k\geq i}\lambda^k + \sum_{k\geq 0}\gamma^{k+1}V_n(x_{k+1})(\lambda^k - \lambda^{k+1})\big] \\
&= \mathbb{E}\big[\sum_{i\geq 0}\lambda^i\big(\gamma^i r^\pi(x_i) + \gamma^{i+1}V_n(x_{i+1}) - \gamma^i V_n(x_i)\big)\big] + V_n(x_0) \\
&= \mathbb{E}\big[\sum_{i\geq 0}(\gamma\lambda)^i d_i\big] + V_n(x_0),
\end{aligned}$$

avec la différence temporelle $d_i = r^\pi(x_i) + \gamma V_n(x_{i+1}) - V_n(x_i)$.

**Algorithme TD($\lambda$)**: $V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k)\sum_{l\geq k}(\gamma\lambda)^{l-k}d_l$.

# 5 Q-learning

Dans un cadre d'un algorithme d'itérations sur les politiques, à chque étape $k$, une fois la fonction valeur $V^{\pi_k}$ approchée par notre estimation $V_n$ (à l'aide de TD($\lambda$)) par exemple, on souhaite définir la nouvelle politique $\pi_{k+1}$ à partir de $V_n$ selon:

$$\pi_{k+1}(x) \in \arg\max_a \big[r(x,a) + \sum_y p(y|x,a)V_n(y)\big]$$

Cependant, dans un cadre apprentissage par renforcement, les probabilités de transition ne sont pas connues, et le calcul de cette politique gloutonne introduit une nouvelle approximation.

Nous introduisons maintenant un algorithme qui construit une approximation de la fonction Q-valeur plutôt que de la fonction valeur.

**Algorithme du Q-learning:** [Watkins, 1989] Construisons une séquence de Q-valeurs $Q_n$ (fonctions définies sur l'espace produit $X \times A$) de la manière suivante: lors d'une transition d'un état $x$ vers $y \sim p(\cdot|x,a)$ en ayant choisi l'action $a$, et observé la récompense $r$, on itère la fonction Q-valeur de l'état $(x,a)$ selon:

$$Q_{n+1}(x,a) = (1-\eta_n(x,a))Q_n(x,a) + \eta_n(x,a)\big[r + \max_{b\in A} Q_n(y,b)\big].$$

*Proposition* 7. *[Watkins et Dayan, 1992]* Supposons que toutes les politiques sont propres. De plus nous supposons que tous les états-actions $(x,a)$ sont itérés une infinité de fois et que les pas satisfont $\forall x,a$, $\sum_{n\geq 0}\eta_n(x,a) = \infty$, $\sum_{n\geq 0}\eta_n^2(x,a) < \infty$. Alors pour tout $x \in X$, $a \in A$, $Q_n(x,a) \xrightarrow{p.s.} Q^*(x,a)$.

*Proof.* $Q^*$ est point fixe de l'opérateur $\mathcal{T}$, défini sur $X \times A$, par:

$$\mathcal{T}W(x,a) = r(x,a) + \sum_y p(y|x,a)\max_{b\in A} W(y,b).$$

Puisque toutes les politiques sont propres, il existe $\mu \in \mathbb{R}^N$ positif et $\beta < 1$ tels que $\sum_y p(y|x,a)\mu(y) \leq \beta\mu(x)$. Donc $|\mathcal{T}Q_1(x,a) - \mathcal{T}Q_2(x,a)| \leq \sum_y p(y|x,a)\max_b |Q_1(y,b) - Q_2(y,b)| \leq \beta||Q_1 - Q_2||_\mu \mu(x)$, et $\mathcal{T}$ est une contraction en norme $L_\infty$ pondérée.

L'algorithme du Q-learning s'exprime ainsi

$$Q_{n+1}(x,a) = (1-\eta_n(x,a))Q_n(x,a) + \eta_n[\mathcal{T}Q_n(x,a) + b_n(x,a)],$$

où $b_n(x,a)$ est une v.a. centrée et telle que $\mathbb{E}[b_n^2(x,a)] \leq c(1 + \max_{y,b} Q_n^2(y,b))$ (où $c$ est une constante). La proposition 5 prouve la convergence de l'algorithme. □

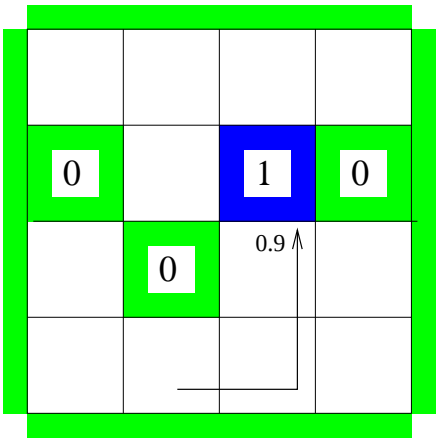
**Q-learning dans le cas actualisé** Remarquons que les Q-valeurs optimales vérifient $Q^*(x,a) = r(x,a) + \gamma\sum_y p(y|x,a)\max_{b\in A} Q^*(y,b)$.

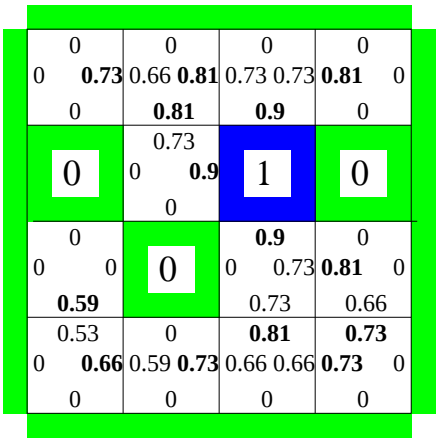
**Algorithme**: lors d'une transitions $x, a \xrightarrow{r} y$, on itère la Q-valeur:

$$Q_{n+1}(x,a) = Q_n(x,a) + \eta_n(x,a)\big[r + \gamma\max_{b\in A} Q_n(y,b) - Q_n(x,a)\big].$$

Convergence sous les mêmes conditions que précédemment.

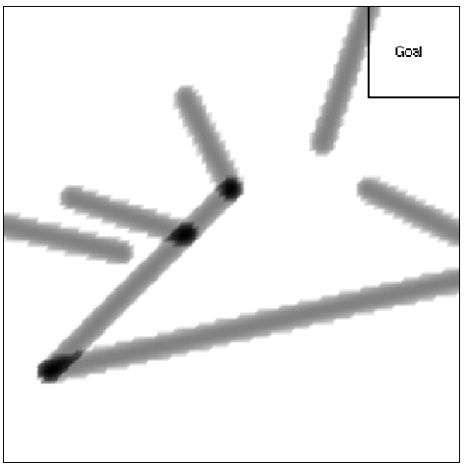
**Exemple: le labyrinthe** Cas déterministe, actualisé $\gamma = 0.9$. Prenons des pas $\eta = 1$. Après la transition $x, a \xrightarrow{r} y$, on itère: $Q_{n+1}(x,a) = r + \gamma\max_{b\in A} Q_n(y,b)$.



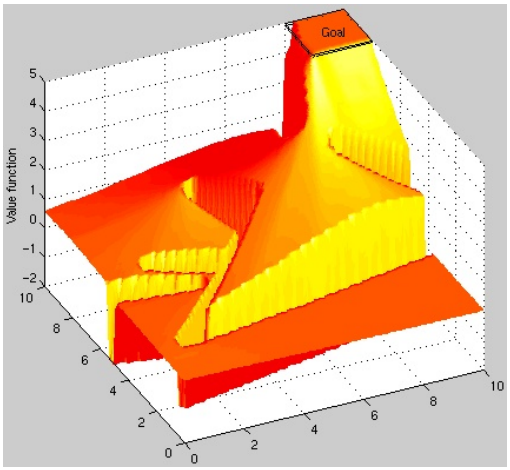



Les Q-valeurs optimales vérifient $Q^*(x,a) = \gamma \max_{b \in A} Q^*(\text{état-suivant}(x,a),b)$.

**Espace de grande taille voire continu** Que faire lorsque l'espace est de grande taille, voire continu? Comment représenter la fonction valeur?




Fonction coût




Fonction valeur

On a besoin de considérer des représentations approchées de nos fonctions (fonction valeur ou Q-valeur) et étudier l'impact de ces approximations sur la qualité des politiques résultantes –> **Programmation dynamique avec approximation**.