Apprentissage par renforcement: Processus de Décision Markovien

INFMDI 341

Aurélien Garivier, Eric Moulines

May 16, 2014

Plan

Le modèle 1/2

Le système est dans un état S_t qui évolue de façon markovienne :

$$S_{t+1} \sim P(\cdot; S_t, A_t) \text{ et } (R_{t+1}, S_{t+1}) \sim P_0(S_t, A_t; \cdot)$$

Le modèle 2/2

[Bellman 1957, Howard 1960, Dubins et Savage 1965, Fleming et Rishel 1975, Bertsekas 1987, Puterman 1994]

Défini par $(\mathcal{S}, \mathcal{A}, p, r)$, où:

- $\mathcal{S}$ espace d'états (supposé fini ici mais peut être dénombrable, continu)
- $\mathcal{A}$ espace d'actions (supposé fini aussi)
- $p(y|x,a)$: probabilités de transition d'un état $x \in \mathcal{S}$ à $y \in \mathcal{S}$ lorsque l'action a est choisie:

$$p(y|x,a) = P(X_{t+1} = y | X_t = x, A_t = a)$$

- $r(x,a,y)$: récompense obtenue lors de la transition de l'état x à y en ayant choisi l'action a.

Principes

- Un processus de décision Markovien permet de modéliser un problème de prise de décision séquentielle, où l'agent interagit avec le système de façon séquentielle
- A l'instant t, le système est dans l'état X_t et l'agent choisi l'action A_t. L'agent observe alors le nouvel etat X_{t+1} et le nouveau reward, qui sont deux variables aléatoires.
- L'objectif de l'agent est de maximiser la moyenne de ses récompenses futures.
- L'agent va définir à cette fin une politique.

Politique

Politique π = une règle de décision, une stratégie comportementale, qui détermine, à un instant donné quelle action doit être choisie.

On distingue deux types de politiques :

- *déterministe* : $\pi : \mathcal{S} \to A$
  $\pi(x)$ = action choisie en x.
- *stochastique* : $\pi : \mathcal{S} \to \mathfrak{M}_1(A)$
  $\pi(a|x)$ = probabilité de choisir a en x.

Valeur d'une politique

Horizon temporel fini :

$$V^\pi(x,t) = E_\pi\left[\sum_{s=t}^{T} r(X_s,\pi(X_s))|X_t = x;\right]$$

Horizon temporel infini avec critère actualisé

$$V^\pi(x) = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(X_s,\pi(X_s))|X_0 = x;\right]$$

où $\gamma \in (0,1)$ est un coefficient d'actualisation.

- $\gamma \ll 1$: approche myope
- $\gamma \approx 1$: approche long-terme

Horizon temporel infini avec critère moyen

$$V^\pi(x) = \lim \frac{1}{T} E_\pi\left[\sum^{T-1} r(X_s,\pi(X_s))|X_0 = x;\right]$$

Politique optimale: horizon fini

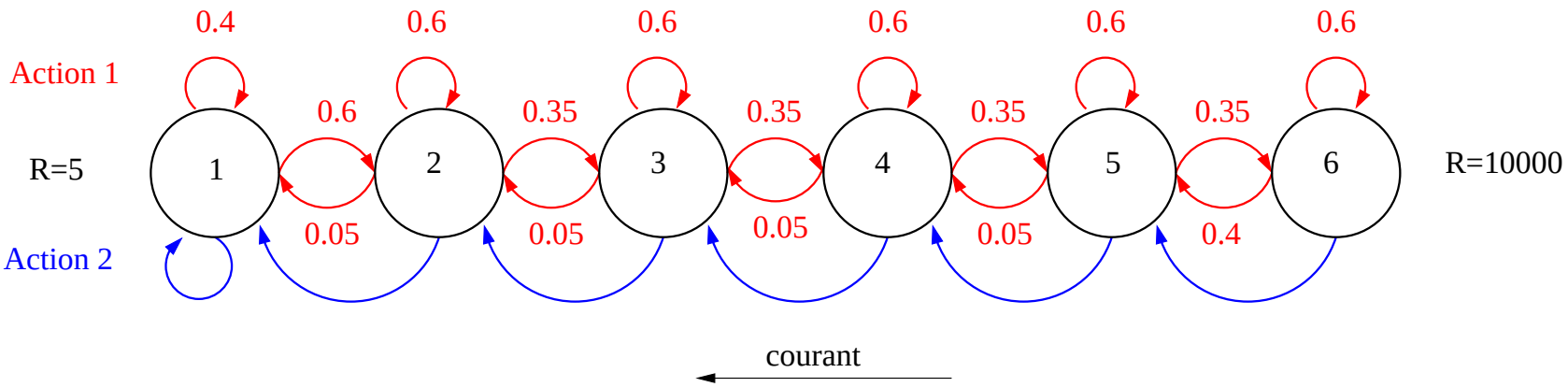
- But : trouver la politique $\pi : \mathcal{S} \to$ A qui a la plus grande *récompense moyenne* :

$$\rho^{\pi} = \lim_{n \to \infty} \frac{1}{n} \mathbb{E}^{\pi} \left[\sum_{t=0}^{n} R_t \right]$$

- Même en connaissant les paramètres, trouver la politique optimale n'est pas évident : c'est le problème dit de *planification*
  $\Rightarrow$ Programmation Dynamique

Plan

Le problème



- deux actions: Un nageur peut nager avec le courant, ou à contre courant
- quand il nage avec le courant, il passe avec une probabilité 1 de passer à l'état immédiatement à gauche
- quand il nage contre le courant, il peut aller à l'état de droite avec une certaine probabilité, rester dans le même état, ou passer dans l'état de gauche

Plan

Le problème

Le responsable d'un entrepot dispose d'un stock X_t d'une marchandise. Il doit satisfaire la demande (aléatoire) D_t des clients. Pour cela, il peut, tous les mois, décider de commander une quantité A_t supplémentaire à son fournisseur.

Il paye un coût de maintenance du stock $h(x)$, un coût de commande du produit $C(a)$

Il reçoit un revenu $f(q)$, où q est la quantité vendue

Le stock restant à la fin procure un revenu $g(x)$

Contrainte: l'entrepot à une capacité limitée M

Objectif: maximiser le profit sur une durée donnée T

Modélisation simplifiée

Modélisation de la demande D_t par une variable aléatoire i.i.d.

Etat: $X_t \in \mathcal{S} = 0, 1, \ldots, M$: quantité (discrète) de produit en stock

Décisions: $a \in \mathcal{A} = \{0, 1, ...M\}$: commande supplémentaire du produit (Rq: ici l'ensemble des actions disponibles à chaque instant dépend de l'état)

Dynamique: $X_{t+1} = [(X_t + A_t) \wedge M - D_t]_+$ (ce qui définit les probabilités de transition $p(y|x, a)$).

Récompense:
$R_{t+1} = -C(A_t) - h(X_t + A_t) + f\left([(X_t + A_t) \wedge M - X_{t+1}]_+\right)$

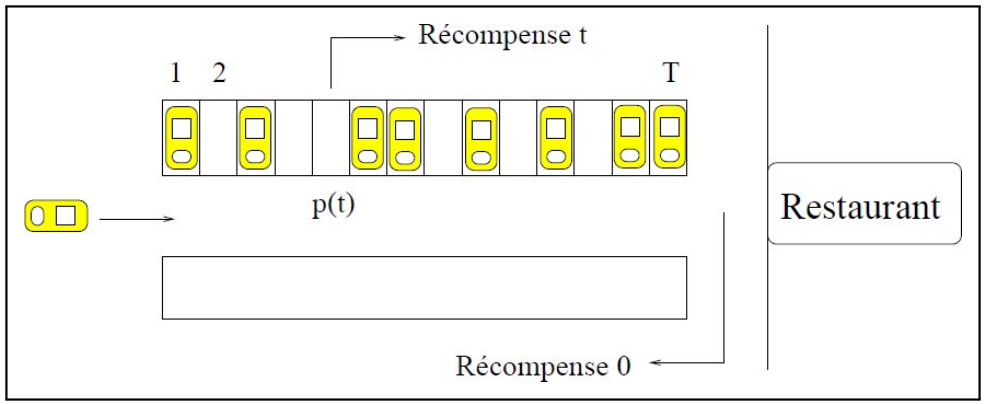
Critère à maximiser:

$$\mathbb{E}\left[\sum_{t=1}^{T-1} R_t + g(X_T)\right]$$

Problème du parking

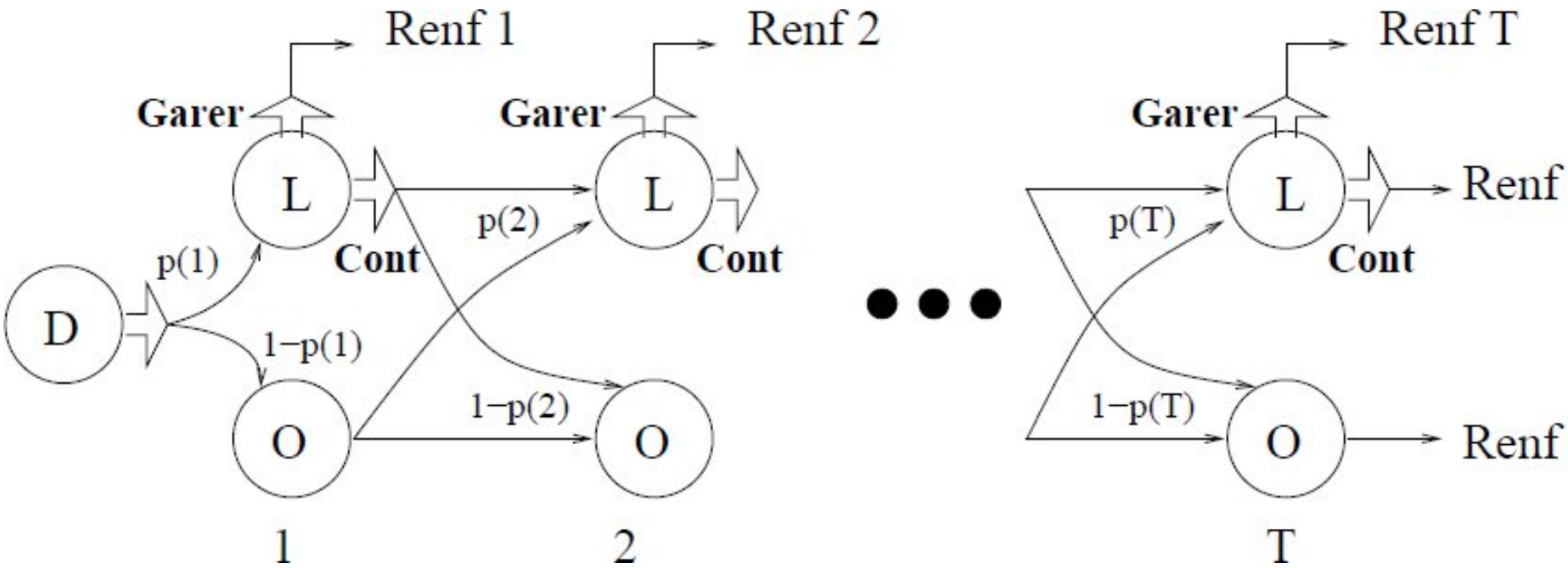
Un conducteur souhaite se garer le plus près possible du restaurant. A chaque instant, l'agent possède 2 actions: continuer ou arrêter.

- Chaque place i est libre avec une probabilité $p(i)$.
- Le conducteur ne peut voir si la place est libre que lorsqu'il est devant. Il décide alors de se garer ou de continuer.
- La place t procure une récompense t. Si pas garé, récompense nulle.



Quelle stratégie maximise le gain espéré ?

MDP pour le problème du parking



Jeux épisodiques

- Dans un jeu épisodique, certains états sont absorbants, *i.e.* le jeu s'arrête lorsque le système atteint un état (ou un ensemble d'états)
- Un épisode du jeu est le temps (aléatoire) mis par le système pour atteindre cet état.

Problème du joueur

- A chaque étape, un joueur parie une fraction $A_t \in [0,1]$ de sa fortune $X_t \geq 0$.
- Il double sa mise avec une probabité p; autrement il perd sa mise:

$$X_{t+1} = (1 + S_{t+1} A_t) X_t$$

où $S_{t+1} \in \{-1, +1\}$ est une suite de variables aléatoires telle que $P(S_{t+1} = 1) = 1$.

- L'objectif du joueur est de maximiser la probabilité que sa fortune atteigne un niveau $w^* > 0$ (on suppose que $X_0 \in [0, w^*]$).

Problème du joueur

- Espace d'état $\mathcal{S} = [0, w^*]$ et l'espace d'actions (continue) est $\mathcal{A} = [0, 1]$
- On définit

$$X_{t+1} = (1 + S_{t+1} A_t)_+ X_t \wedge w^* \,.$$

où w^* est l'état terminal où le jeu s'arrête.

- La récompense immédiate est nulle si $X_{t+1} < w^*$ et est égale à 1 si $X_{t+1} = w^*$ pour la première fois.
- L'espérance de la récompense cumulée est égale à la probabilité d'atteindre w^*.

Plan

Politique stationnaire

- Politique déterministe stationnaire fonction associant à chaque état $x \in \mathcal{S}$ une action $a \in \mathcal{A}$; la politique du joueur est stationnaire

$$A_t = \pi(X_t) \, , t \geq 0 \, .$$

- Politique randomisée stationnaire noyau de transition associant à chaque état $x \in \mathcal{S}$ une distribution de probabilités sur l'espace des actions $\pi(\cdot|x)$; à chaque instant $t \geq 0$, l'agent choisit une action suivant cette distribution

$$A_t \sim \pi(\cdot|X_t) \, , t \geq 0 \, .$$

Processus de récompense markovien

- Une politique stationnaire (aléatoire ou randomisée) appliquée à processus de décision Markovien définissent un processus de récompense markovien $\{(X_t, R_t)\}_{t\geq 0}$.
- La transition de ce processus de récompense markovien est donnée par

$$P_0^\pi(\cdot|x) = \sum_{a\in\mathcal{A}} \pi(a|x) P_0(\cdot|x,a)$$

Fonction valeur

A toute politique stationnaire on peut associer une fonction valeur définie par

$$V^\pi(x) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \middle| X_0 = x \right] , x \in \mathcal{S}$$

où $(R_t, t \geq 1)$ est la suite des récompenses associées à la suite des états et des actions associées à la politique stationnaire π.
La fonction action-valeur $Q^\pi(x,a)$ est définie de façon similaire par

$$Q^\pi(x,a) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \middle| X_0 = x, A_0 = a \right] , x \in \mathcal{S}, a \in \mathcal{A} .$$

Equations de Bellman pour une politique stationnaire

Pour la politique π et une fonction valeur V définie sur $\mathcal{S}$, l'opérateur de Bellman est défini par

$$T^\pi V(x) = r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{S}} P(x, \pi(x); y) V(y) \ , x \in \mathcal{S}$$

Theorem

La fonction valeur V^π

$$V^\pi(x) = r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{S}} P(x, \pi(x); y) V^\pi(y) \ , \quad x \in \mathcal{S}$$

est un point-fixe de l'opérateur de Bellman.

Contraction

Theorem

Pour $0 < \gamma < 1$, l'opérateur de Bellman est une contraction pour la norme $\|z\|_\infty = \max(|z_i|)$

Contraction

Theorem

Pour $0 < \gamma < 1$, l'opérateur de Bellman est une contraction pour la norme $\|z\|_\infty = \max(|z_i|)$

Proof.

$$\begin{aligned}
\|T^\pi U - T^\pi V\| &\leq \gamma \sup_{x\in\mathcal{S}} \left| \sum_{y\in\mathcal{S}} P(x,\pi(x);y)\{U(y)-V(y)\} \right| \\
&\leq \gamma \sup_{x\in\mathcal{S}} \sum_{y\in\mathcal{S}} P(x,\pi(x);y)\,|U(y)-V(y)| \\
&\leq \gamma \sup_{x\in\mathcal{S}} \sum_{y\in\mathcal{S}} P(x,\pi(x);y)\|U-V\|_\infty \\
&= \gamma \|U-V\|_\infty .
\end{aligned}$$

Point fixe de Banach

Theorem

Soit $(E, \|\cdot\|)$ un espace de Banach et $T : E \mapsto E$ une contraction, $\|T(x-y)\| \leq \gamma \|x-y\|$. Alors l'opérateur T a un unique point fixe v, $Tv = v$. De plus, pour $v_0 \in E$, la suite définie récursivement par $v_{n+1} = Tv_n$ converge vers v, $\lim_{n\to\infty} \|v_n - v\| = 0$, et la convergence de cette suite est géométrique

$$\|v_n - v\| \leq \gamma^n \|v_0 - v\| \ .$$

Calcul de la fonction valeur d'une politique stationnaire

- Comme l'opérateur de Bellman T^π est une contraction stricte, l'opérateur T^π possède un unique point fixe V^π, la fonction valeur associée à la politique π
- La suite de fonctions valeurs définie récursivement par $V_{k+1} = T^\pi V_k$, où V_0 est choisie de façon arbitraire converge exponentiellement vite vers la fonction valeur V^π.

Calcul de la fonction valeur: cas fini

- Dans le cas où l'espace des états et actions est fini, l'équation de Bellman est une équation linéaire

$$V^\pi = R^\pi + \gamma P^\pi V^\pi$$
$$(I - \gamma P^\pi) V^\pi = R^\pi$$
$$V^\pi = (I - \gamma P^\pi)^{-1} V^\pi \ .$$

Calcul de la fonction valeur: cas fini

- Dans le cas où l'espace des états et actions est fini, l'équation de Bellman est une équation linéaire

$$\begin{aligned} V^\pi &= R^\pi + \gamma P^\pi V^\pi \\ (I - \gamma P^\pi) V^\pi &= R^\pi \\ V^\pi = (I - \gamma P^\pi)^{-1} V^\pi &\,. \end{aligned}$$

- Complexité dans le cas à n états: $O(n^3)$.

Opérateur de Bellman optimal

$$T^*V(x) = \sup_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \sum_{y \in \mathcal{S}} P(x,a;y) V(y) \right\} , \quad x \in \mathcal{S} .$$

Pour toute fonction état-action π et toute fonction V, $T^\pi V \leq T^* V$.

Theorem

L'opérateur de Bellman optimal est une contraction stricte par rapport à la norme $\| \cdot \|_\infty$.

Preuve

On note que

$$\left| \sup_{a \in \mathcal{A}} f(a) - \sup_{a \in \mathcal{A}} g(a) \right| \leq \sup_{a \in \mathcal{A}} |f(a) - g(a)|$$

On en déduit

$$\begin{aligned} \|T^* u - T^* V\|_\infty &\leq \gamma \sup_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{y \in \mathcal{S}} P(x, a : y) \{U(y) - V(y)\} \\ &\leq \gamma \sup_{x \in \mathcal{S}} \sum_{y \in \mathcal{S}} P(x, a; y) \|U - V\|_\infty \\ &\leq \gamma \|U - V\|_\infty \ . \end{aligned}$$

Politique optimale

Comme T^* est une contraction, T^* admet un unique point fixe V^*.

Theorem

Pour toute politique stationnaire π, $V^\pi \preceq V^*$, i.e., pour tout $x \in \mathcal{S}$, $V^\pi(x) \leq V^*(x)$, $x \in \mathcal{S}$. S'il existe une fonction état-action π^* telle que $T^{\pi^*} V^* = T^* V^*$, alors $V^\pi \leq V^{\pi^*}$ pour toute politique stationnaire: la politique π^* est optimale.

Politique optimale (preuve)

Proof.

- Pour toute fonction état-action π, $V^\pi = T^\pi V^\pi \preceq T^* V^\pi$.

Politique optimale (preuve)

Proof.

- Pour toute fonction état-action π, $V^\pi = T^\pi V^\pi \preceq T^* V^\pi$.
- Comme T^* est monotone (si $U \preceq V$, $T^* U \preceq T^* V$), $V^\pi \preceq (T^*)^k V^\pi$.

Politique optimale (preuve)

Proof.

- Pour toute fonction état-action π, $V^\pi = T^\pi V^\pi \preceq T^* V^\pi$.
- Comme T^* est monotone (si $U \preceq V$, $T^*U \preceq T^*V$), $V^\pi \preceq (T^*)^k V^\pi$.
- Comme $\lim_{k\to\infty} (T^*)^k V^\pi \to V^*$, on a $V^\pi \preceq V^*$.

Politique optimale (preuve)

Proof.

- Pour toute fonction état-action π, $V^\pi = T^\pi V^\pi \preceq T^* V^\pi$.
- Comme T^* est monotone (si $U \preceq V$, $T^* U \preceq T^* V$), $V^\pi \preceq (T^*)^k V^\pi$.
- Comme $\lim_{k\to\infty} (T^*)^k V^\pi \to V^*$, on a $V^\pi \preceq V^*$.
- ${T^\pi}^* V^* = T^* V^* = V^*$ et comme le point fixe de ${T^\pi}^*$ est unique, on a $V^* = V^{\pi^*}$.

□

Une caractérisation des politiques optimales

Theorem

Si $T^*V^{\pi_0} = V^{\pi_0}$ alors π_0 est une politique optimale

Proof.

Une caractérisation des politiques optimales

Theorem

Si $T^*V^{\pi_0} = V^{\pi_0}$ alors π_0 est une politique optimale

Proof.

- Comme $T^*V^{\pi_0} = V^{\pi_0}$, alors V^{π_0} est un point fixe de T^*.

Une caractérisation des politiques optimales

Theorem

Si $T^*V^{\pi_0} = V^{\pi_0}$ alors π_0 est une politique optimale

Proof.

- Comme $T^*V^{\pi_0} = V^{\pi_0}$, alors V^{π_0} est un point fixe de T^*.
- Comme le point fixe est unique, $V^{\pi_0} = V^*$

Une caractérisation des politiques optimales

Theorem

Si $T^*V^{\pi_0} = V^{\pi_0}$ alors π_0 est une politique optimale

Proof.

- Comme $T^*V^{\pi_0} = V^{\pi_0}$, alors V^{π_0} est un point fixe de T^*.
- Comme le point fixe est unique, $V^{\pi_0} = V^*$
- Pour tout π, $V^{\pi} \leq V^* = V^{\pi_0}$ et donc la politique est optimale.

□

Itération sur les politiques

Idée : on part d'une politique quelconque, on l'évalue, et on l'améliore !

- Politique initiale π_0 quelconque
- Evaluation de la politique π_k : on calcule $V_k^\pi(x)$ pour tout $x \in \mathcal{S}$.
- Amélioration de la politique : on choisit la politique *gloutonne*

$$\pi_{k+1}(x) = \operatorname*{argmax}_a r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y)$$

- Critère d'arrêt : $V^{\pi_{k+1}} = V^{\pi_k}$ (ou $\pi_{k+1} = \pi_k$)

On parle d'algorithme *acteur-critique*

Algorithme Acteur-Critique: amélioration d'une fonction valeur

Theorem

Soit π_0 et π deux politiques telles que $T^{\pi} V^{\pi_0} = T^* V^{\pi_0}$. Alors $V^{\pi_0} \leq V^{\pi}$

Algorithme Acteur-Critique: amélioration d'une fonction valeur

Theorem

Soit π_0 et π deux politiques telles que $T^\pi V^{\pi_0} = T^* V^{\pi_0}$. Alors $V^{\pi_0} \leq V^\pi$

Proof.

- $T^\pi V^{\pi_0} = T^* V^{\pi_0} \geq T^{\pi_0} V^{\pi_0} = V^{\pi_0}$

Algorithme Acteur-Critique: amélioration d'une fonction valeur

Theorem

Soit π_0 et π deux politiques telles que $T^\pi V^{\pi_0} = T^* V^{\pi_0}$. Alors $V^{\pi_0} \leq V^\pi$

Proof.

- $T^\pi V^{\pi_0} = T^* V^{\pi_0} \geq T^{\pi_0} V^{\pi_0} = V^{\pi_0}$
- $(T^\pi)^k V^{\pi_0} \geq V^{\pi_0}$ et donc $V^\pi = \lim_{k\to\infty} (T^\pi)^k V^{\pi_0} \geq V^{\pi_0}$

□

Algorithme acteur-critique

Theorem

S'il existe $x \in \mathcal{S}$ *tel que* $T^* V^{\pi_0}(x) > V^{\pi_0}(x)$ *alors* $V^{\pi}(x) > V^{\pi_0}(x)$

Algorithme acteur-critique

Theorem

S'il existe $x \in \mathcal{S}$ *tel que* $T^* V^{\pi_0}(x) > V^{\pi_0}(x)$ *alors* $V^{\pi}(x) > V^{\pi_0}(x)$

Proof.

- $T^{\pi} V^{\pi_0}(x) = T^* V^{\pi_0}(x) > V^{\pi_0}(x)$ et donc $(T^{\pi})^n V^{\pi_0}(x) > V^{\pi_0}(x)$

Algorithme acteur-critique

Theorem

S'il existe $x \in \mathcal{S}$ *tel que* $T^* V^{\pi_0}(x) > V^{\pi_0}(x)$ *alors* $V^{\pi}(x) > V^{\pi_0}(x)$

Proof.

- $T^{\pi} V^{\pi_0}(x) = T^* V^{\pi_0}(x) > V^{\pi_0}(x)$ et donc $(T^{\pi})^n V^{\pi_0}(x) > V^{\pi_0}(x)$
- en passant à la limite, $V^{\pi}(x) > V^{\pi_0}(x)$.

□

Du local au global

Equation de Bellman

$$V(x) = \max_a r(x,a) + \gamma \mathbb{E}[V(Y)|x,a]$$

Si je connais V , "en moyenne, pas de surprise" !

Comment apprendre la fonction valeur? Grâce à la surprise (TD)

Si V est connue, cela permet de choisir, à chaque instant, la meilleure action

$$\operatorname*{argmax}_a r(x,a) + E[V(Y)|x,a]$$

$\Longrightarrow$ maximiser localement la fonction valeur revient à maximiser le renforcement à long terme.

Equation de Bellman pour le critère moyen

- Pour tout MDP $\mathcal{M} = (\mathcal{S}, \mathsf{A}, P, r)$ *faiblement communicant*, la *récompense moyenne* $\rho^*(\mathcal{M})$ est indépendante de l'état initial.
- Il existe un vecteur de biais h^* tel que, pour tout $s \in \mathcal{S}$,

$$h^*(s) + \rho^*(\mathcal{M}) = \max_{a \in \mathsf{A}} \left(\mathbb{E}[R(s,a)] + \sum_{s' \in \mathcal{S}} P(s'; s, a) h^*(s') \right)$$

Algorithme d'itération sur les valeurs

Pour trouver une politique proche de l'optimale, il suffit de résoudre l'équation de Bellman :

- Soit $k = 0$. Fixons $V_k \in \mathbb{R}^{|\mathcal{S}|}$ et $\varepsilon > 0$
- Tant que $\max_s(V_{k+1}(s) - V_k(s)) - \min_s(V_{k+1}(s) - V_k(s)) > \varepsilon$,

$$\forall s \ , \ V_{k+1}(s) = \max_{a \in \mathsf{A}} \left(\mathbb{E}[R(s,a)](s,a) + \sum_{s' \in \mathcal{S}} P(s';s,a) V_k(s') \right)$$

- Une politique ε-optimale est donnée par

$$\forall s \ , \ \pi^*(s) \in \operatorname*{argmax}_{a \in \mathsf{A}} \left(\mathbb{E}[R(s,a)](s,a) + \sum_{s' \in \mathcal{S}} P(s';s,a) V_k(s') \right)$$

Convergence et complexité

Proposition : L'algorithme d'IP génère une séquence de politiques de performances croissantes qui se termine en un nombre fini d'étapes avec une politique optimale π^*
Complexité :

- résolution directe du système

$$V^\pi = r + \gamma P V^\pi$$

  par la méthode de Gauss : en $O(|S|^3)$ (ou un peu moins)
- itération sur les valeurs : $V^\pi_{k+1} = r + \gamma P V^\pi_k$ en $O(|S|^2 \frac{\log \varepsilon}{\log \gamma})$ pour une valeur ε-approchée
- Monte-Carlo : on tire n trajectoires au hasard, erreur en $1/\sqrt{n}$

Plan

Stratégie

Stratégie : séquence de règles de décision $\Pi = (\pi_0, \pi_1, \pi_2, \dots)$ telle que la politique π_t est choisie uniquement à partir des observations $y_1, \dots, y_{t-1}$.

Si Π est constante dans le temps, on parle de politique stationnaire ou Markovienne: $\Pi = (\pi, \pi, \pi, \dots)$

Pour une politique markovienne fixée Π, le processus $(X_t)_t$ est une chaîne de Markov (définie par les probabilités de transition $p(y|x) = p(y|x, \pi(x))$).

Attention : stratégie $\neq$ politique !

Deux grandes familles de méthodes

Méthodes indirectes apprentissage préalable d'un modèle des dynamiques (forme d'apprentissage supervisé), puis utilisation du modèle pour faire de la planification

Méthodes directes apprentissage direct d'une stratégie d'action sans étape préliminaire de modélisation (peut être intéressant quand les dynamiques d'état sont complexes alors que le contrôleur est simple).

Même si les dynamiques sont connues, le problème de planification peut être très complexe! On cherche alors une solution approchée (programmation dynamique avec approximation), ex: le programme TD-gammon.

Méthode directe : SARSA

SARSA = State-Action-Reward-State-Action
Idée : On construit une table action-valeur qu'on met à jour au fur et à mesure des observations suivant la règle :

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right)$$

$\Longrightarrow$ on ajuste l'estimation de $Q(s_t, a_t)$ suivant la "surprise" qu'on reçoit.

Variante : Q-learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(R_{t+1} + \gamma \max_a \{Q(s_{t+1}, a)\} - Q(s_t, a_t) \right)$$

Plan

L'algorithme UCRL-2 [Auer et al, '09]

- Stratégie optimiste : à l'instant t
  1. considère l'ensemble de tous les MDP (transitions + lois des récompenses) qui rendent les observations assez vraisemblables
  2. trouve le MDP (dit *optimiste*) dont la valeur est la plus grande
  3. joue *pendant un certain temps* la politique optimale de ce MDP
- Le MDP *optimiste* maximise les équations d'optimalité :

$$\forall s,\ h^*(s) + \rho^* = \max_{P,r} \max_{a \in \mathsf{A}} \left(r(s,a) + \sum_{s' \in \mathcal{S}} P(s';s,a) h^*(s') \right)$$

$$\text{tel que } \forall s,\ \forall a,\ \left\| \hat{P}_t(.;s,a) - P(.;s,a) \right\|_1 \leq \delta_P$$

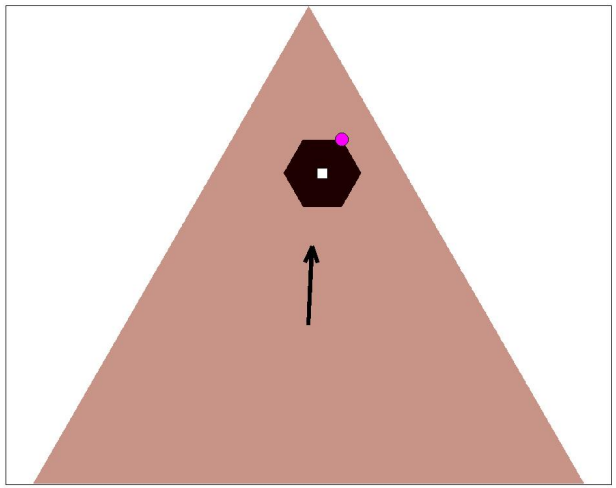
$$\forall s,\ \forall a,\ |\hat{r}_t(s,a) - r(s,a)| \leq \delta_R$$

$\Rightarrow$ *Extended* Value Iteration

Propriétés de l'algorithme UCRL-2

- On doit résoudre à chaque étape des problèmes du type : pour une loi empirique p et pour un vecteur de biais V, trouver

$$q^* = \underset{\|p-q\|_1 \leq \delta}{\operatorname{argmax}}\ q'V$$



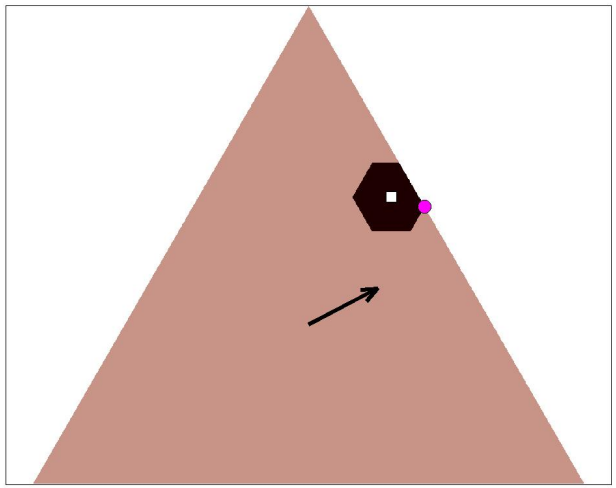
Solution :

- on gonfle la probabilité de transition vers l'état le plus "prometteur"

Propriétés de l'algorithme UCRL-2

- On doit résoudre à chaque étape des problèmes du type : pour une loi empirique p et pour un vecteur de biais V, trouver

$$q^* = \underset{\|p-q\|_1 \leq \delta}{\operatorname{argmax}}\ q'V$$



Solution :

- on gonfle la probabilité de transition vers l'état le plus "prometteur"

Propriétés de l'algorithme UCRL-2

Mesure de performance : *regret cumulé*

$$\text{Regret}(n) = \sum_{t=1}^{n} \rho^* - R_t$$

De plus, on peut montrer les *bornes de regret* suivantes:

$$\mathbb{E}(\text{Regret}(n)) \leq C|\mathcal{S}|^2|\text{A}|\log(n)\ ,$$

C étant une constante dépendant de

$$\Delta(\mathcal{M}) = \min_{\rho^\pi < \rho^*} \rho^* - \rho^\pi$$

Propriétés du modèle optimiste

Mais le modèle optimiste a quelques propriétés indésirables

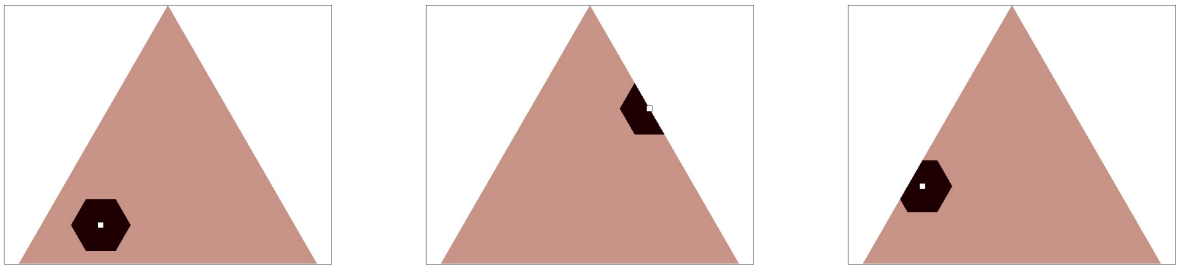
- il ne dépend pas continument des observations
- peut mettre à 0 des transitions observées
- ne peut pas mettre à 0 des transitions vers le "paradis"
- les voisinages L^1 n'ont pas beaucoup de sens pour des lois de probabilités

$$\Rightarrow \text{comportement difficilement explicable}$$

Les voisinages de UCRL-2

Théorème (Weissman, Ordentlich, Seroussi, Verdu, Weinberger '03)
si $X_1, \ldots, X_n$ sont des v.a. iid à valeur dans $\mathcal{S}$ et de loi $p = (p(1), \ldots, p(|\mathcal{S}|))$, l'estimateur $\hat{p}_n = (\hat{p}_n(1), \ldots, \hat{p}_n(|\mathcal{S}|))$ défini par $n\hat{p}_n(i) = \sum_{j=1}^{n} \mathbb{1}_{\{i\}}(X_j)$ vérifie

$$P\left(\|\hat{p}_n - p\|_1 > \delta\right) \leq (2^{|\mathcal{S}|} - 2) \exp\left(-\frac{n\delta^2}{2}\right)$$



Voisinages invariants par translation dans le simplexe.

Plan

Plan

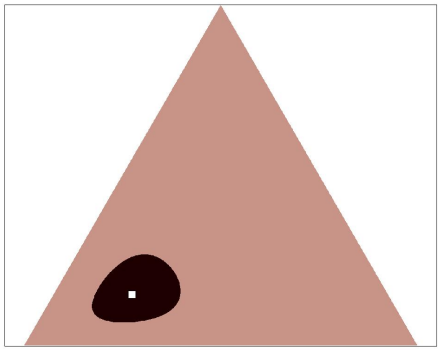
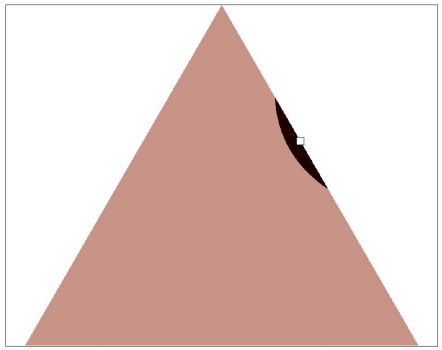
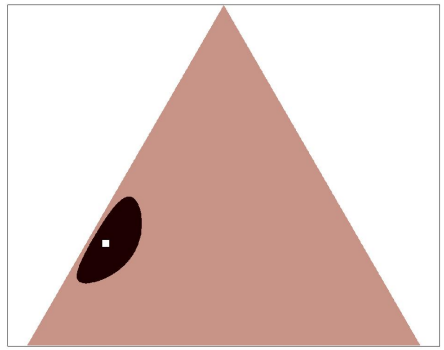
Inégalité de concentration

Théorème: si $X_1, \ldots, X_n$ sont des v.a. iid à valeur dans $\mathcal{S}$ et de loi $p = (p(1), \ldots, p(|\mathcal{S}|))$, l'estimateur $\hat{p}_n = (\hat{p}_n(1), \ldots, \hat{p}_n(|\mathcal{S}|))$ défini par $n\hat{p}_n(i) = \sum_{j=1}^n \mathbb{1}_{\{i\}}(X_j)$ vérifie

$$P\left(\forall t \leq n,\ KL(\hat{p}_t, p) > \frac{\delta}{t}\right) \leq 2e(\delta \log(n) + |\mathcal{S}|)e^{-\delta/|\mathcal{S}|}$$

Contrôle pour tout $1 \leq t \leq n$.
Preuve : à base d'incrément de martingales, cf. Hoeffding-Azuma mais sans majoration uniforme (en particulier, on garde la variance).

Géométrie des voisinages

Le voisinage KL est adapté à la géométrie et aux propriétés probabilistes du simplexe.

Plan

"Küllback-Leibler UCRL"

- Stratégie optimiste similaire à l'algorithme UCRL-2
- Voisinages du maximum de vraisemblance : utilisation de l'*information de Küllback-Leibler*.
  Le modèle optimiste maximise les équations d'optimalité :

$$\forall s,\ h^*(s) + \rho^* = \max_{P,r} \max_{a \in \mathsf{A}} \left(r(s,a) + \sum_{s' \in \mathcal{S}} P(s';s,a) h^*(s') \right)$$

$$\text{tel que } \forall s,\ \forall a,\ KL(\hat{P}_t(.;s,a); P(.;s,a)) \leq \delta_P$$

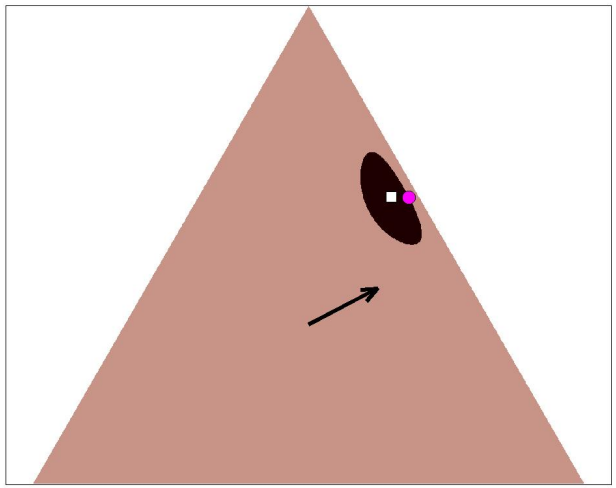
$$\forall s,\ \forall a,\ |\hat{r}_t(s,a) - r(s,a)| \leq \delta_R$$

$\Rightarrow$ *Extended* Value Iteration

La maximisation

- On doit résoudre à chaque étape des problèmes du type : pour une loi empirique p et pour un vecteur de biais V, trouver

$$q^* = \underset{KL(p;q)\leq\delta}{\operatorname{argmax}} \; q'V$$



- Solution explicite de cette maximisation : maximisation d'une fonction linéaire sur un espace convexe.

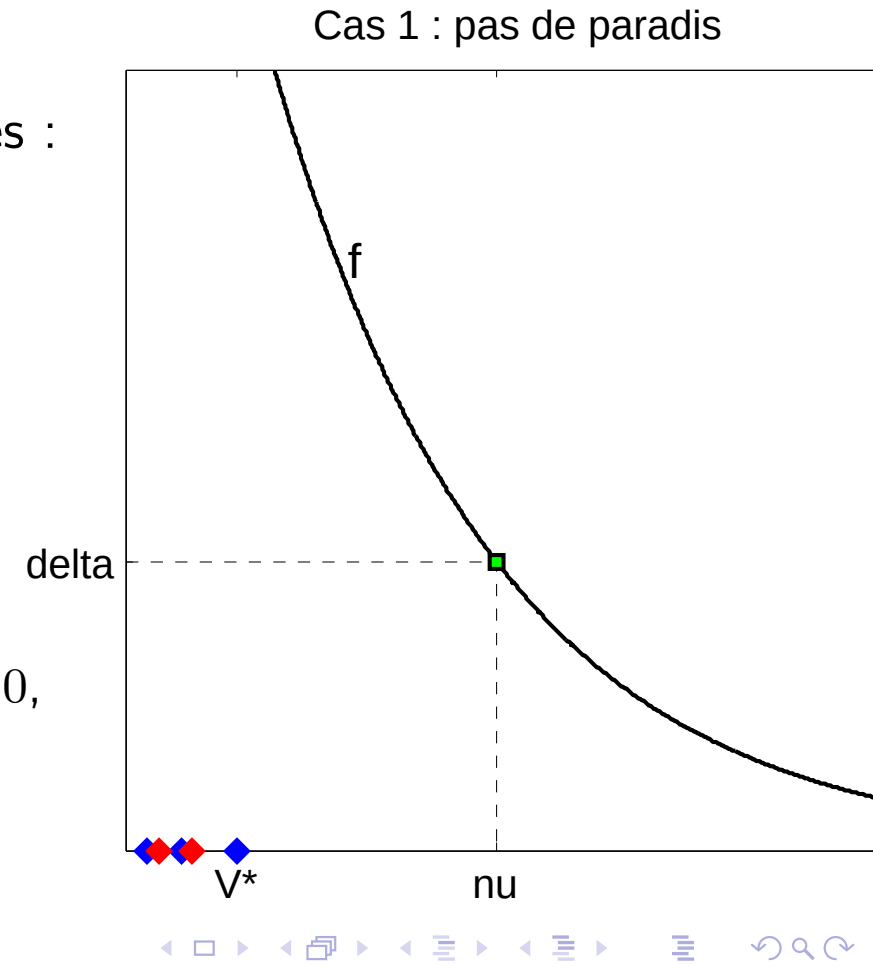
Trouver le maximum

- Soit $i^* = \operatorname{argmax} V_i$. Deux possibilités :
- Cas 1: si $p_{i^*} > 0$ alors $f(\nu) = \delta$ et

$$q_i \propto \frac{p_i}{\nu - V_i}$$

- Cas 2: Si $p_{i^*} = 0$, 2 cas :
  - Cas 2.A: si $f(V_{i^*}) \geq \delta$, alors cf. Cas 1
  - Cas 2.B: si $f(V_{i^*}) < \delta$, alors $q_{i^*} > 0$, $\nu = V_{i^*}$ et

$$\text{pour } i \neq i^*, \quad q_i \propto \frac{p_i}{\nu - V_i}$$



Trouver le maximum

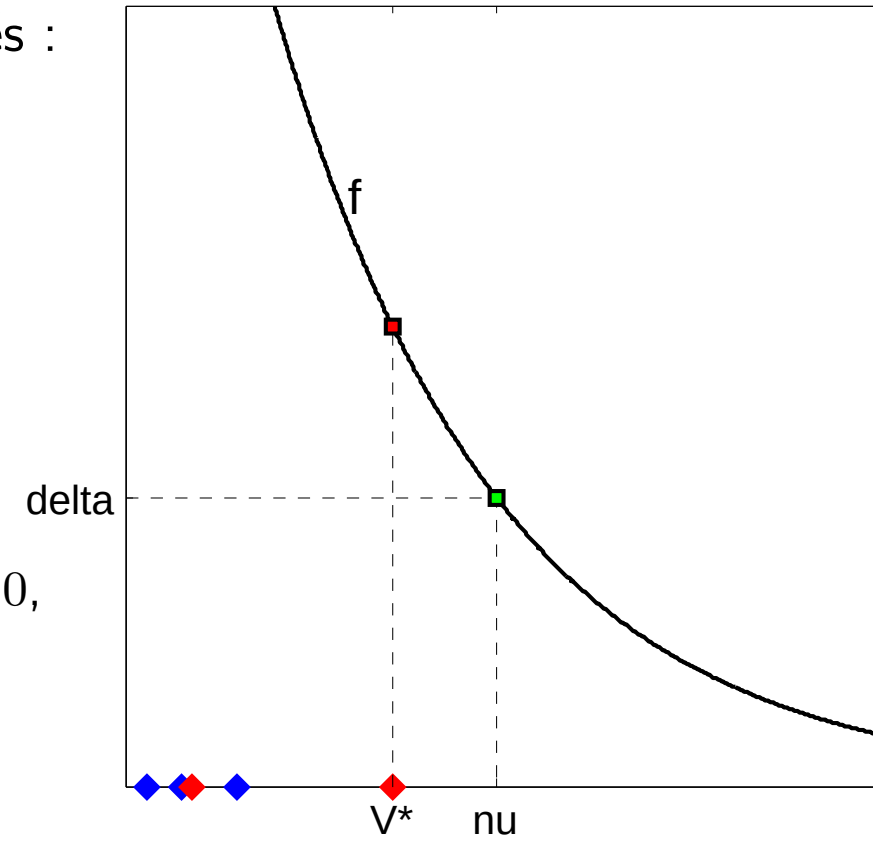
- Soit $i^* = \operatorname{argmax} V_i$. Deux possibilités :
- Cas 1: si $p_{i^*} > 0$ alors $f(\nu) = \delta$ et

$$q_i \propto \frac{p_i}{\nu - V_i}$$

- Cas 2: Si $p_{i^*} = 0$, 2 cas :
  - Cas 2.A: si $f(V_{i^*}) \geq \delta$, alors cf. Cas 1
  - Cas 2.B: si $f(V_{i^*}) < \delta$, alors $q_{i^*} > 0$, $\nu = V_{i^*}$ et

$$\text{pour } i \neq i^*, \quad q_i \propto \frac{p_i}{\nu - V_i}$$

Cas 2.A : renoncement au paradis



Trouver le maximum

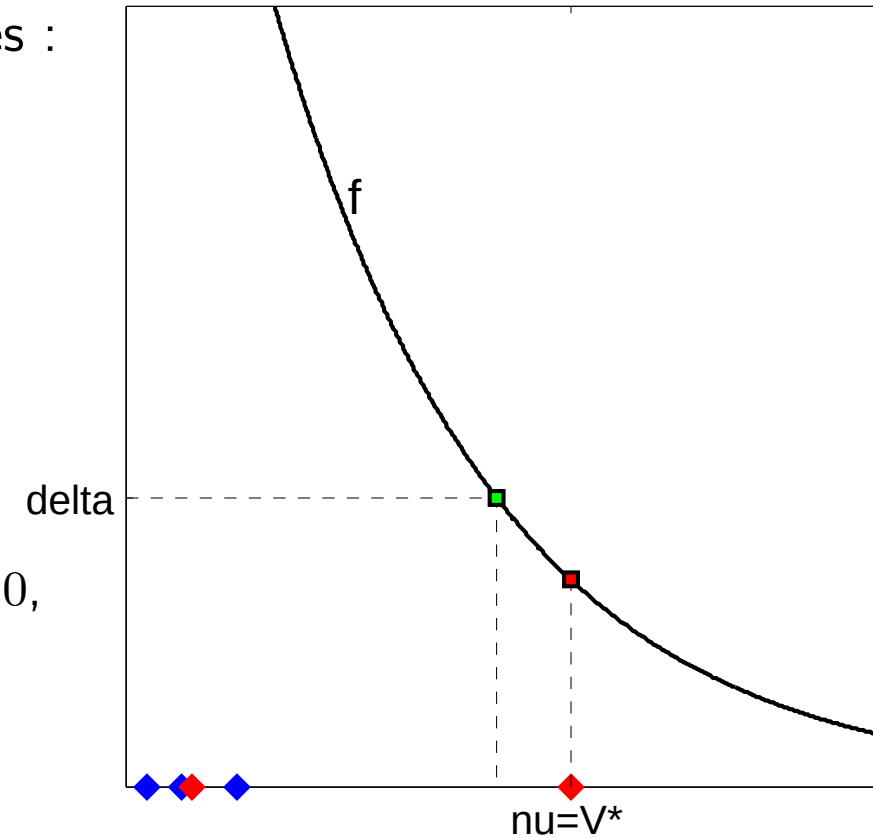
- Soit $i^* = \operatorname{argmax} V_i$. Deux possibilités :
- Cas 1: si $p_{i^*} > 0$ alors $f(\nu) = \delta$ et

$$q_i \propto \frac{p_i}{\nu - V_i}$$

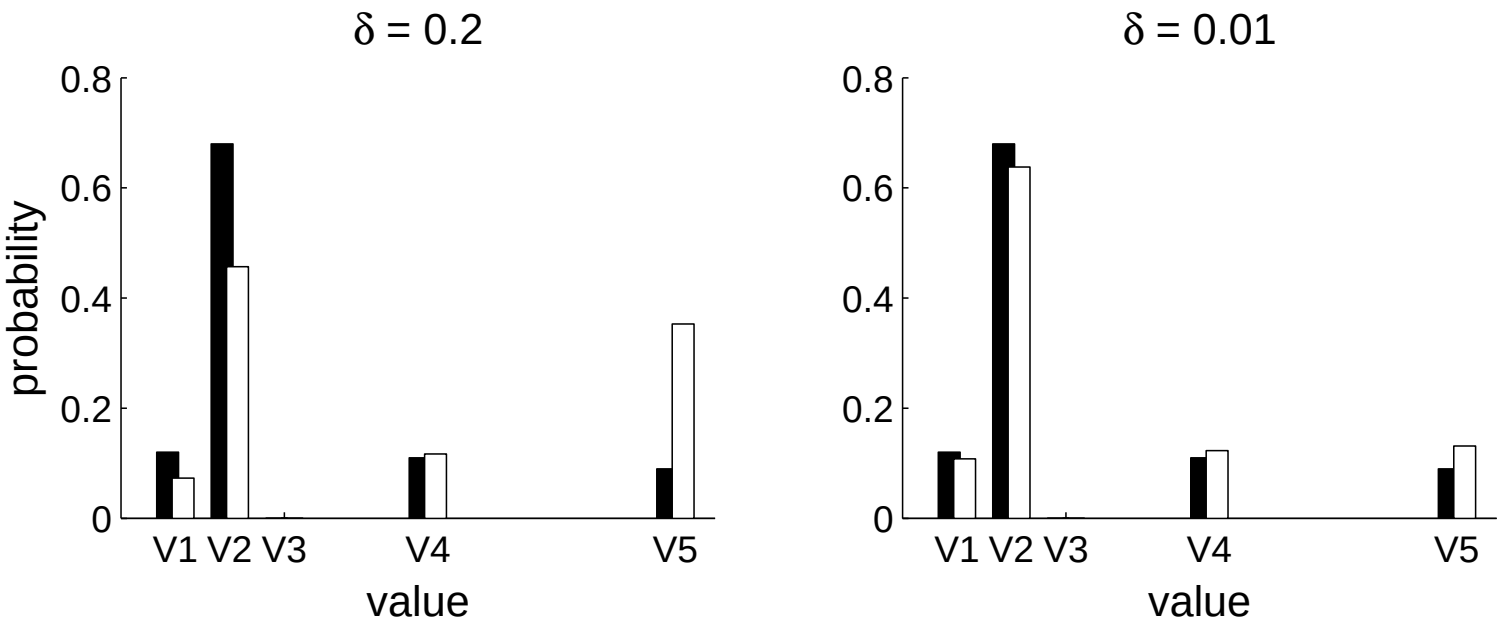
- Cas 2: Si $p_{i^*} = 0$, 2 cas :
  - Cas 2.A: si $f(V_{i^*}) \geq \delta$, alors cf. Cas 1
  - Cas 2.B: si $f(V_{i^*}) < \delta$, alors $q_{i^*} > 0$, $\nu = V_{i^*}$ et

$$\text{pour } i \neq i^*, \quad q_i \propto \frac{p_i}{\nu - V_i}$$

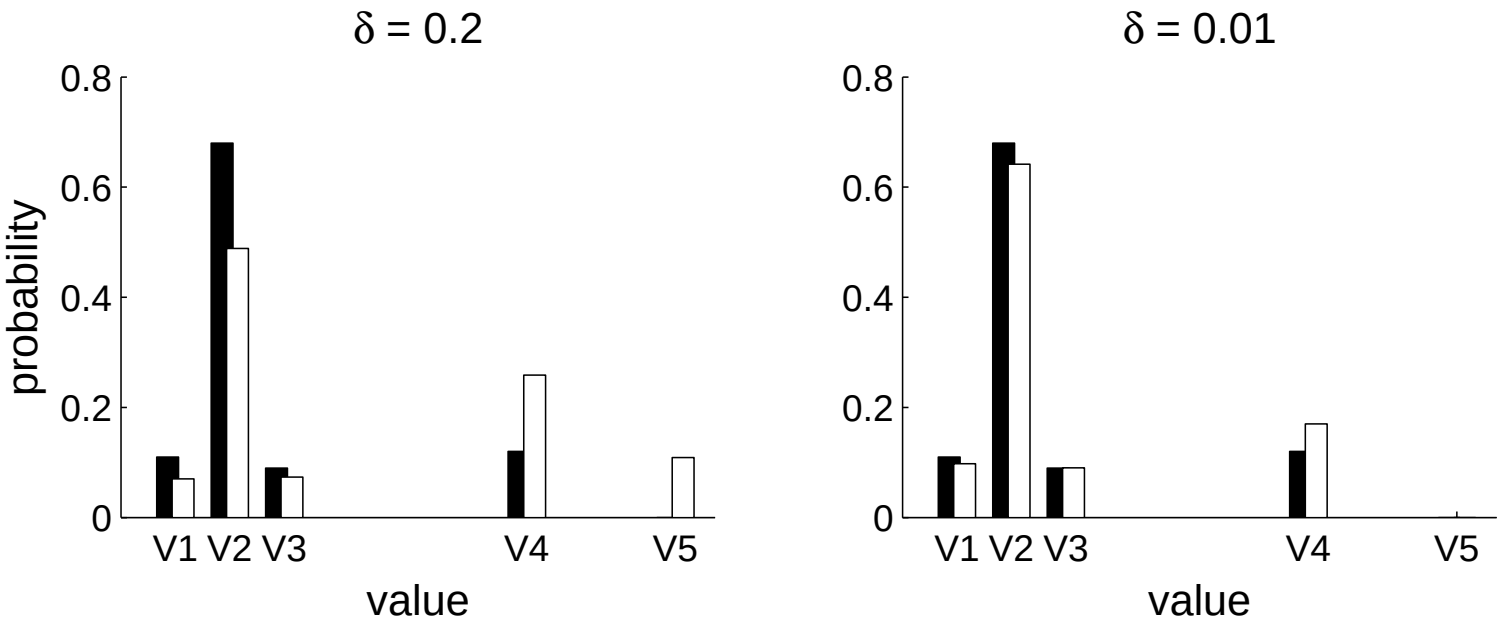
Cas 2.B : espoir de paradis



Règle I
"Les meilleurs états sont favorisés"



Règle II
"Pas de Paradis si δ est trop petit"



Commentaires

- La maximisation ne pose donc aucun problème algorithmique et peut être résolue très rapidement en quelques itérations de Newton.
- Si aucune transition vers un "paradis" n'a été observée, l'algorithme arbitre entre
  - ajouter de la probabilité à cette transition
  - reconnaître qu'elle est invraisemblable et ajouter de la probabilité à d'autres transitions

  en fonction
  - du *nombre de transitions* observées (dont dépend δ)
  - l'*intérêt relatif* de cet état (mesuré par son biais)

Plan

Majoration du regret

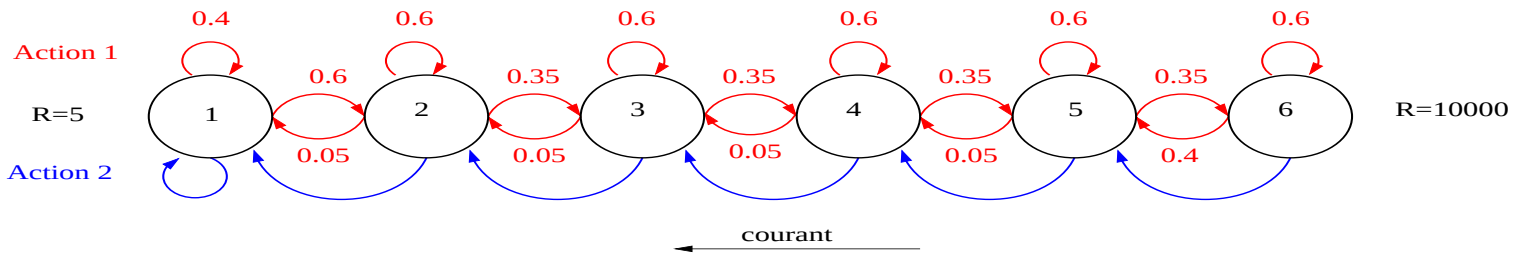
Théorème : Pour un horizon $n > 1$ assez grand, le regret moyen en utilisant l'algorithme KL-UCRL est borné par :

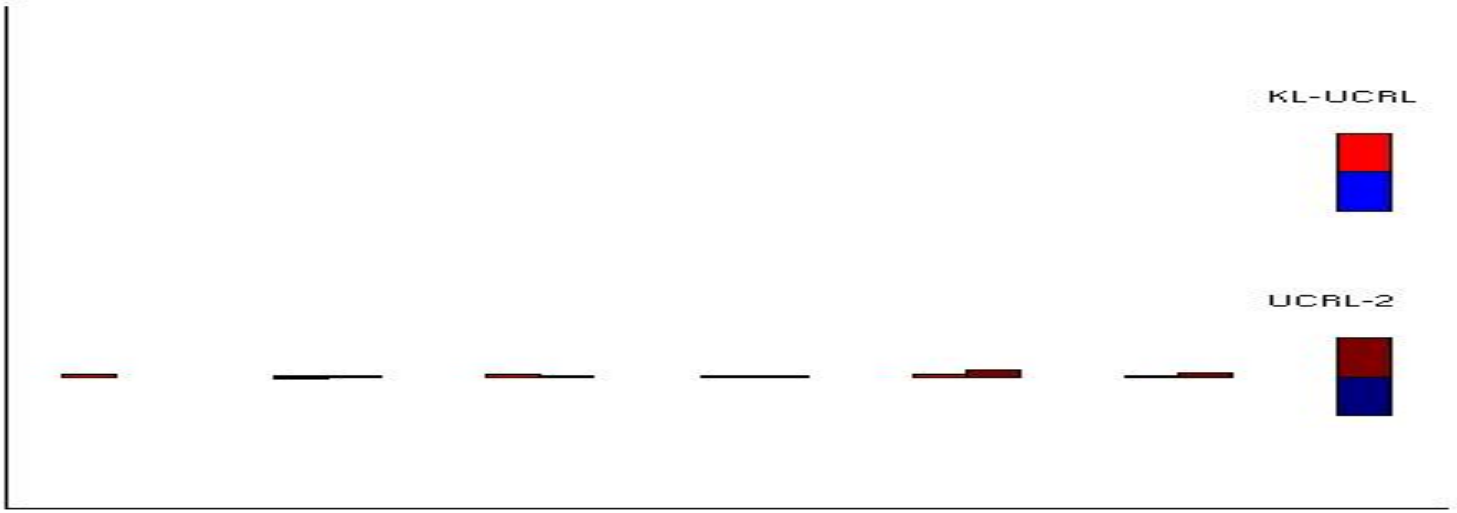
$$\mathbb{E}(\text{Regret}(n)) \leq C|\mathcal{S}|^2|\text{A}|\log(n)\ ,$$

C étant une constante dépendant de

$$\Delta(\mathcal{M}) = \min_{\rho^\pi < \rho^*} \rho^* - \rho^\pi$$

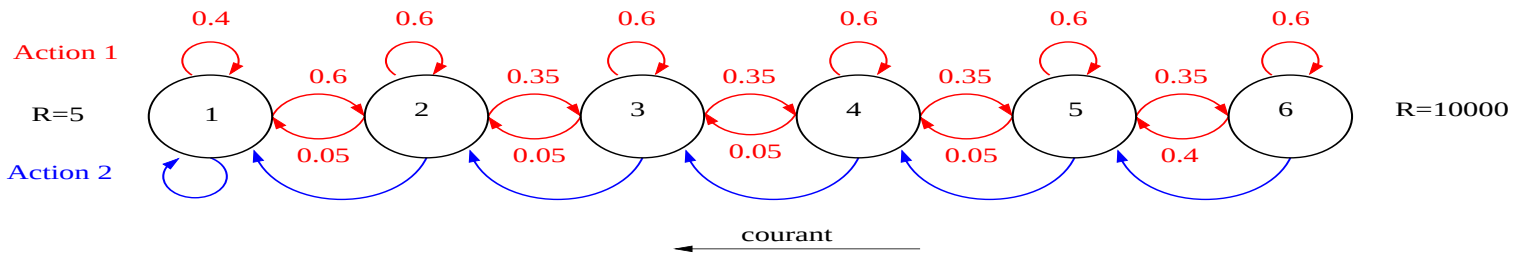
Simulations : RiverSwim

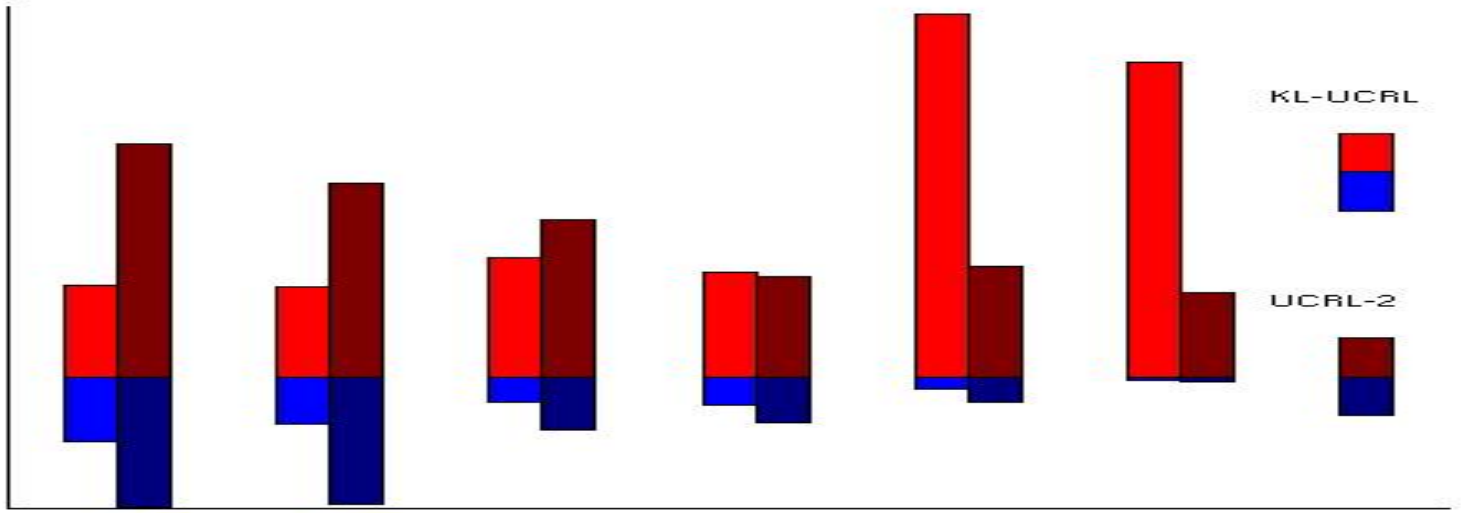




Début

Simulations : RiverSwim





Début

Simulations : RiverSwim

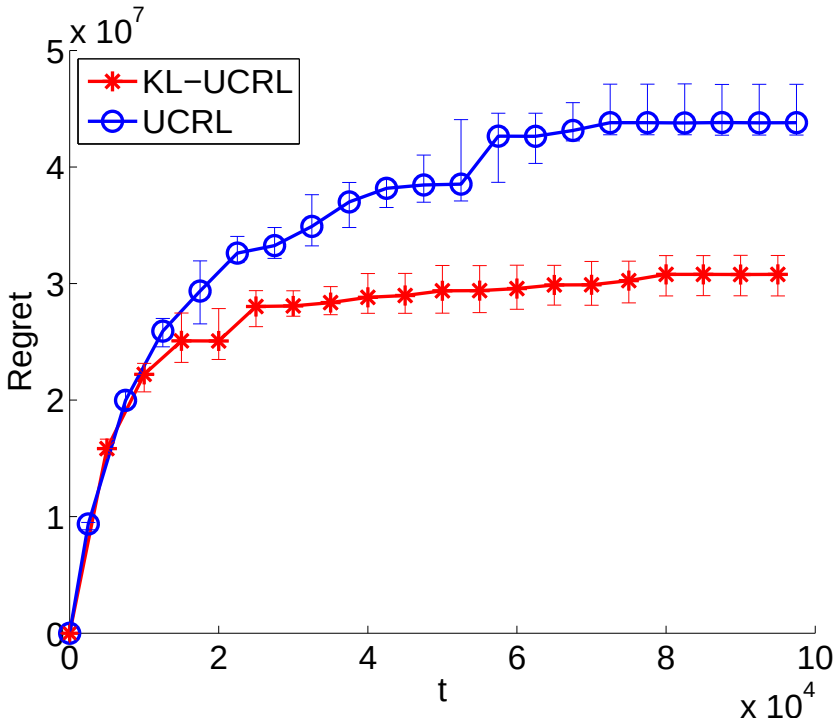


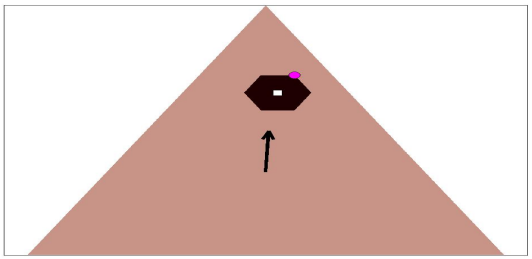
Figure : Comparaison des regrets des algorithmes UCRL-2 et KL-UCRL.

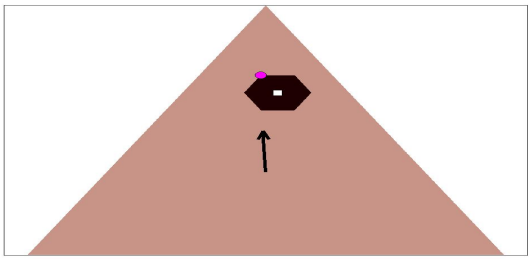
Plan

Continuité du modèle optimiste

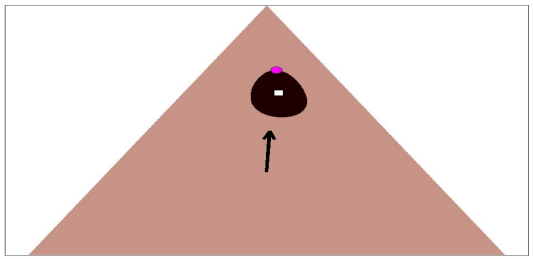
De plus, le voisinage KL dépend plus continument des observations.

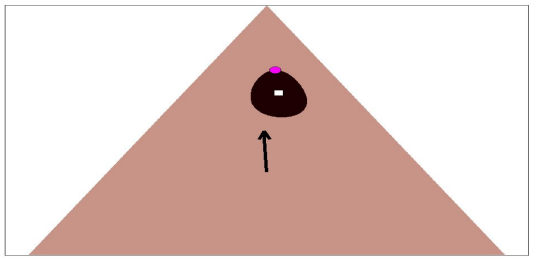
Voisinage L^1





Voisinage KL

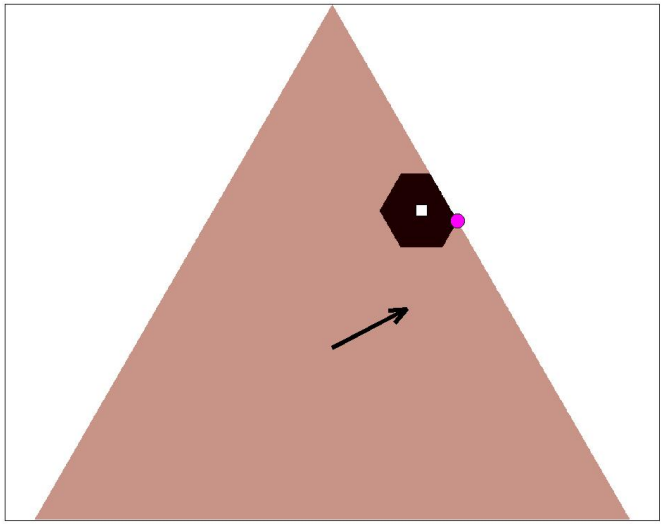




Compatibilité avec les observations

Le modèle optimiste donne toujours une probabilité non-nulle aux évènements observés

Voisinage L^1



Voisinage KL