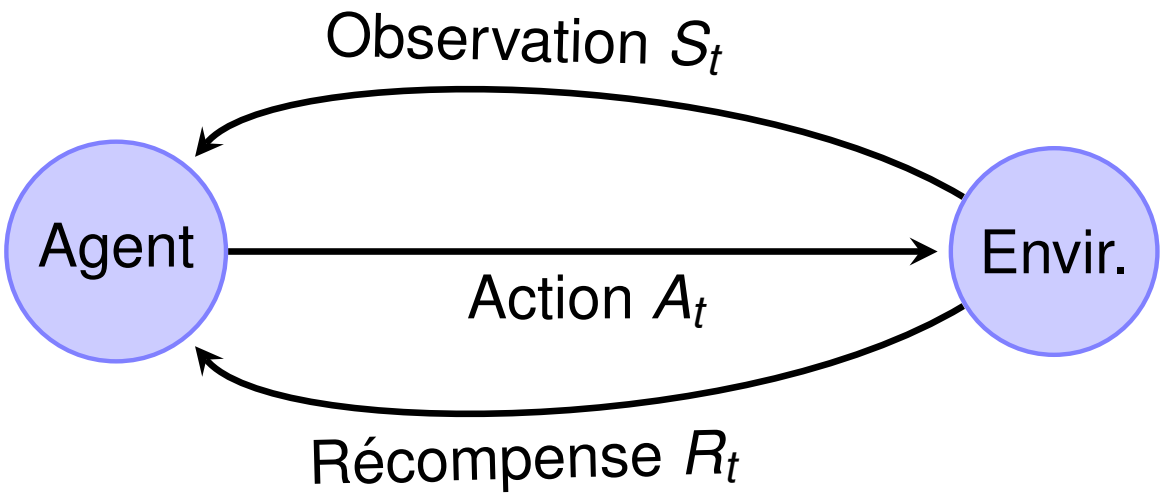
Optimism in Reinforcement Learning and Kullback-Leibler Divergence

Sarah Filippi, Olivier Cappé, Aurélien Garivier

TELECOM ParisTech

Plan

Apprentissage par Renforcement

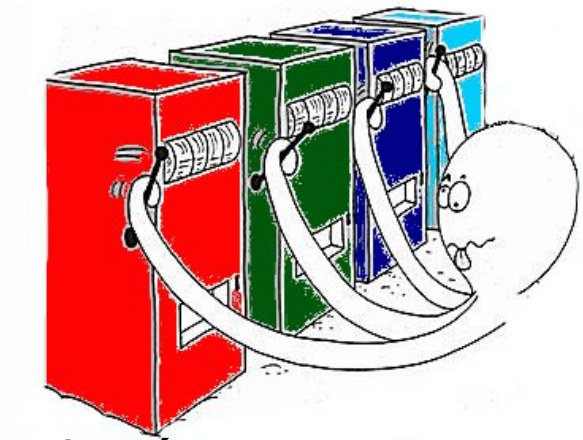


dilemme
exploration
|
exploitation

- L'agent est acteur et pas spectateur [Sutton '92; Bertsekas '95]
- A chaque instant t, il choisit une action $A_t \in A$ en fonction des observations et récompenses passés $(S_s, R_s)_{s<t}$ pour maximiser la récompense cumulée $\sum_{t=1}^{n} R_t$
- Exemples: essais médicaux, robotique, proposition de contenu, finance, publicité, internet mobile, ...

Problèmes de bandits

- Environnement constant
- Conditionnellement aux actions $(A_t)_{1\leq t\leq n}$, les récompenses $(R_t)_{1\leq t\leq n}$ sont i.i.d. de moyenne μ_{A_t}



- But : jouer l'action a^* qui a la plus grande récompense moyenne :

$$\mu_{a^*} = \max_{a\in A} \mu_a$$

- Mesure de performance : *regret cumulé*

$$\text{Regret}(n) = \sum_{t=1}^{n} \mu_{a^*} - \mu_{A_t}$$

Upper Confidence Bound (UCB)

- Algorithmes optimistes : [Lai&Robins '85; Agrawal '95]

  Fais comme si tu te trouvais dans l'environnement qui t'est le plus favorable parmi tous ceux qui rendent les observations assez vraisemblables

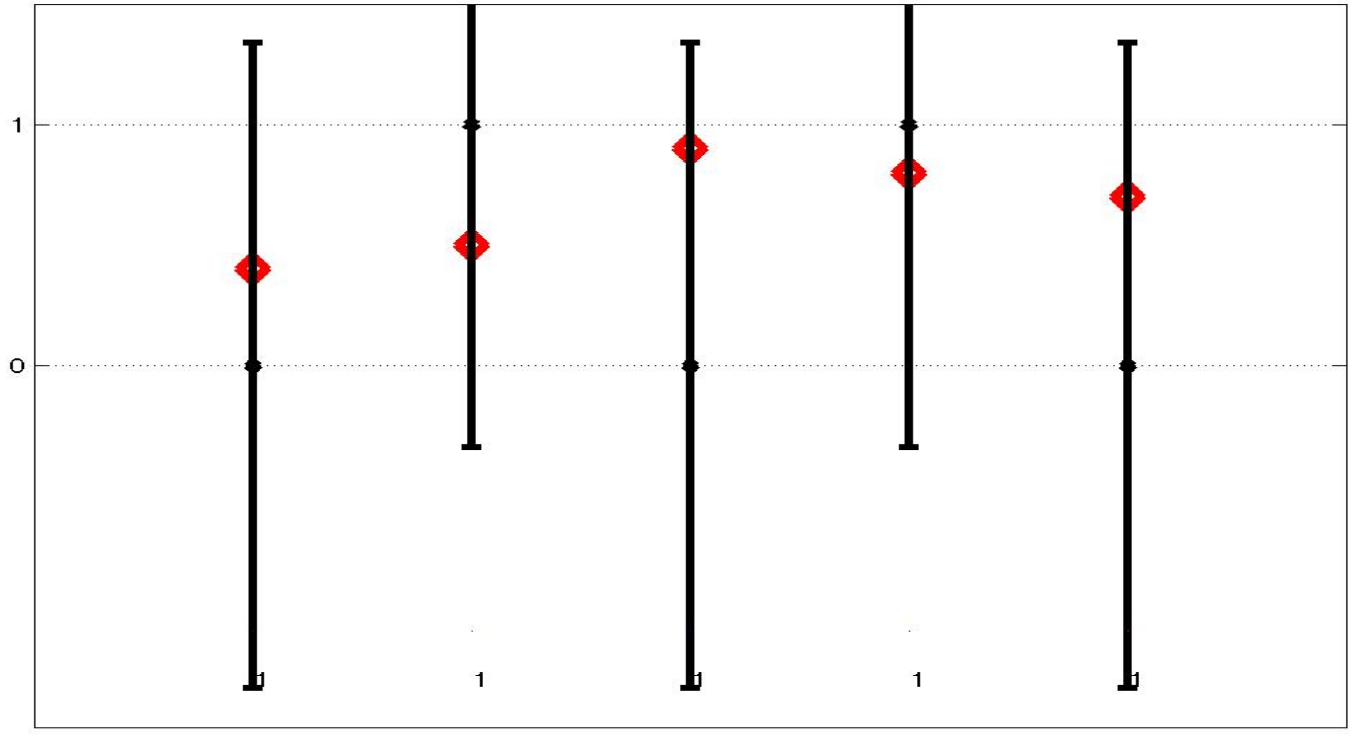
- Ici : UCB (Upper Confidence Bound) = établir une borne supérieure de l'intérêt de chaque action, et choisir celle qui est la plus prometteuse [Auer&al '02; Audibert&al '07]
- Avantage : comportement facilement interprétable et "acceptable"

  $\Rightarrow$ le regret grandit comme $C\log(n)$, où C dépend de

$$\Delta = \min_{\mu_a < \mu_{a^*}} \mu_{a^*} - \mu_a$$

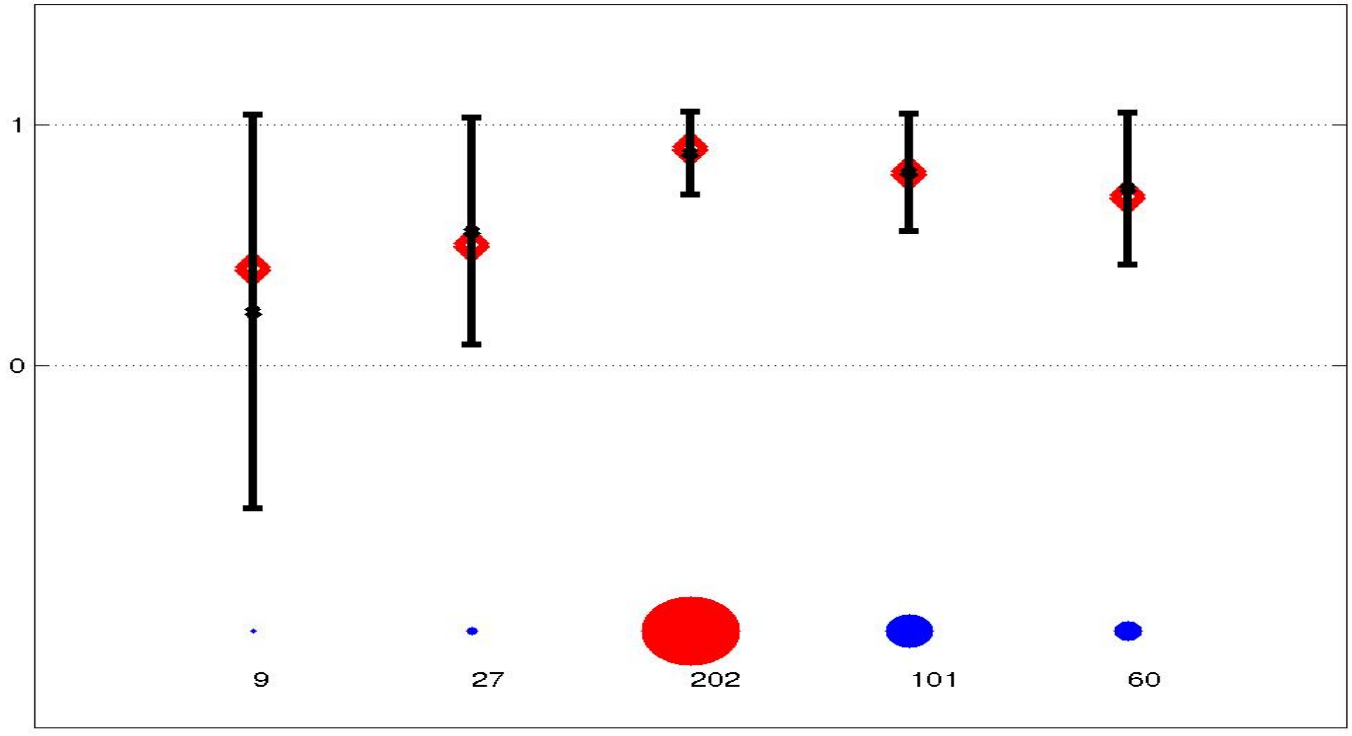
et c'est (presque) optimal

UCB en action



Début

UCB en action



Début

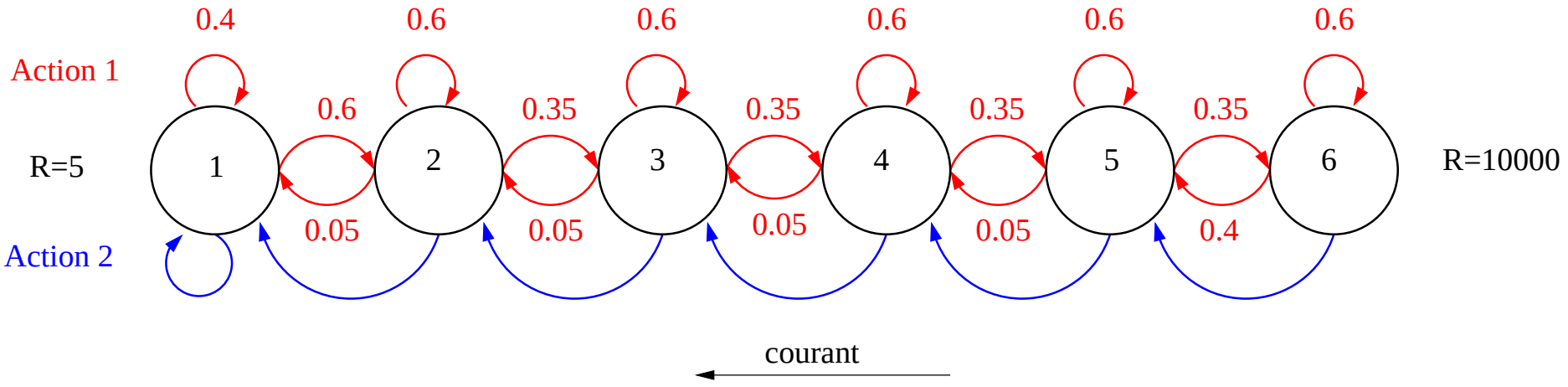
Plan

Processus de Décision Markoviens

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

$$S_{t+1} \sim P(\cdot; S_t, A_t) \text{ et } R_t = r(S_t, A_t) + \varepsilon_t$$

Exemple / Benchmark : RiverSwim [StrehI&Littman'08]



Politique optimale

- But : trouver la politique π : S $\rightarrow$ 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

Equation de Bellman

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

$$h^*(s) + \rho^*(\mathbf{M}) = \max_{a \in \mathrm{A}} \left(r(s,a) + \sum_{s' \in \mathrm{S}} P(s'; s, a) h^*(s') \right) .$$

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

$$\forall s\ ,\ V_{k+1}(s) = \max_{a \in A} \left(r(s,a) + \sum_{s' \in S} P(s';s,a) V_k(s') \right) .$$

- Une politique ε-optimale est donnée par

$$\forall s\ ,\ \pi^*(s) \in \operatorname{argmax}_{a \in A} \left(r(s,a) + \sum_{s' \in S} P(s';s,a) V_k(s') \right) .$$

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 :

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

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

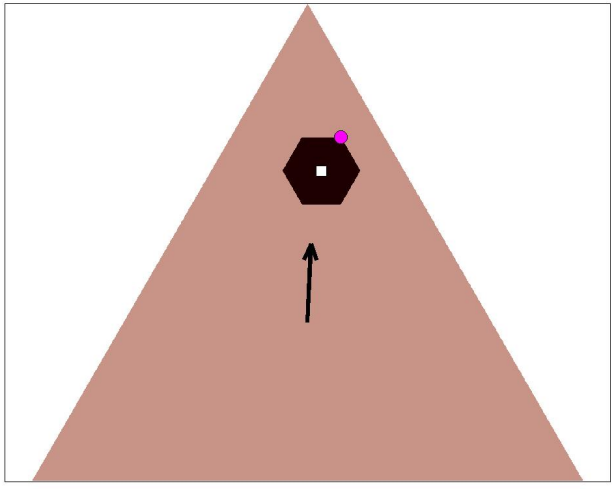
$$\forall s,\ \forall a,\ |\hat{r}_t(s,a) - r(s,a)| \leq \delta_R$$

$\Rightarrow$ *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

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



Solution :

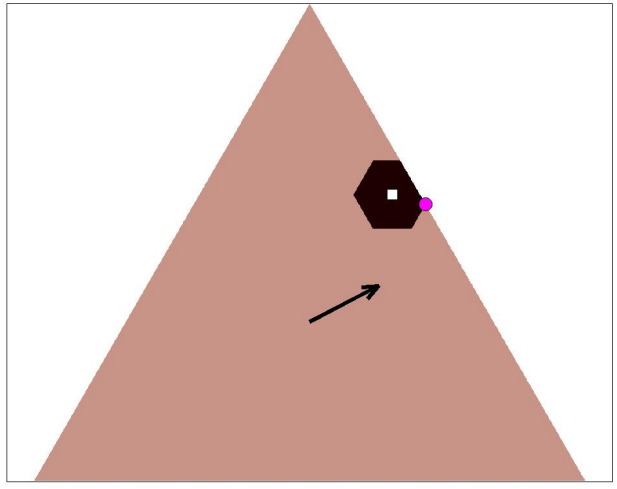
- on gonfle la probabilité de transition vers l'état le plus "prometteur"
- on diminue celle du plus faible, et au besoin du second, etc...

$\Rightarrow$ algorithmiquement trivial, facilement interprétable

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

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



Solution :

- on gonfle la probabilité de transition vers l'état le plus "prometteur"
- on diminue celle du plus faible, et au besoin du second, etc...

$\Rightarrow$ algorithmiquement trivial, facilement interprétable

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:

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

C étant une constante dépendant de

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

Propriétés du modèle optimiste

Mais le modèle optimiste a quelques propriétés indésirables

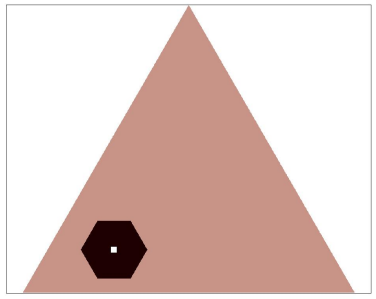
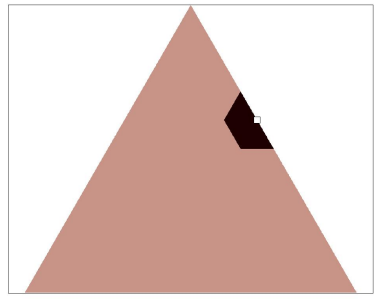
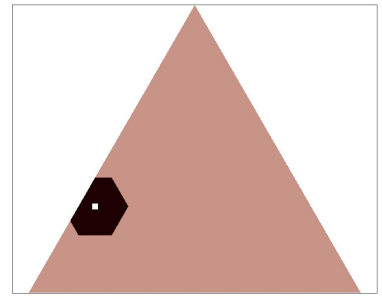
- il ne dépend pas continument des observations
- peut mettre à 0 des transitions observées
- ne peut pas mettre à 0 des transitions vers le ”paradis”
- les voisinages L^1 n’ont pas beaucoup de sens pour des lois de probabilités

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

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 S et de loi $p = (p(1), \ldots, p(|S|))$, l'estimateur $\hat{p}_n = (\hat{p}_n(1), \ldots, \hat{p}_n(|S|))$ défini par $n\hat{p}_n(i) = \sum_{j=1}^{n} \mathbb{1}_{\{i\}}(X_j)$ vérifie

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

Voisinages invariants par translation dans le simplexe.

Plan

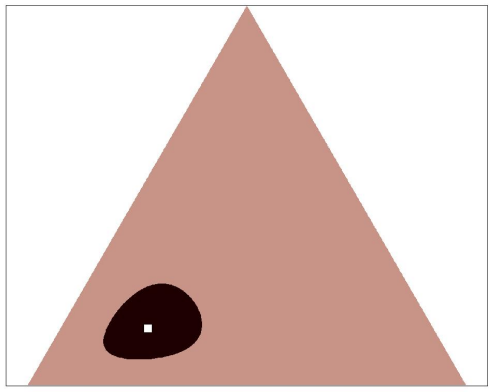
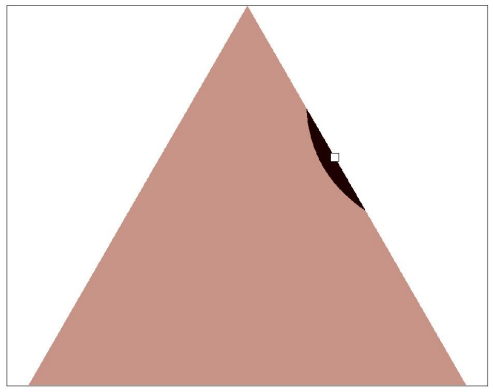
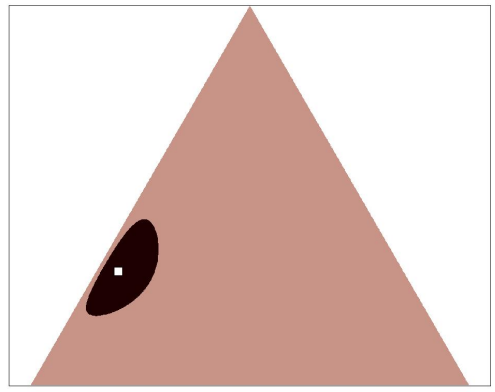
Inégalité de concentration

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

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

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 :

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

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

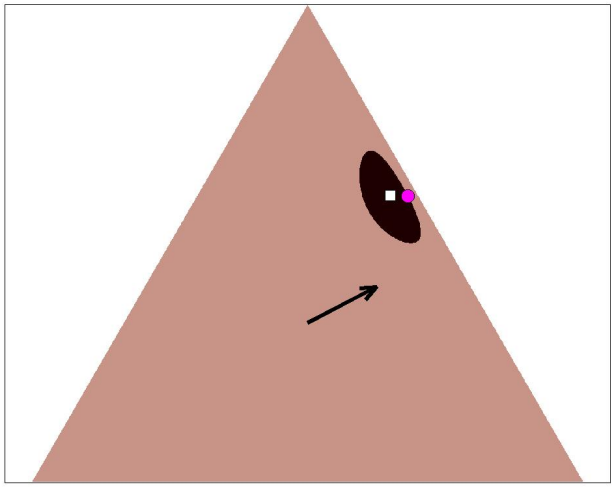
$$\forall s,\ \forall a,\ |\hat{r}_t(s,a) - r(s,a)| \leq \delta_R$$

$\Rightarrow$ *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

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



- Solution explicite de cette maximisation : maximisation d'une fonction linéaire sur un espace convexe.
- Pour $\nu > \max_{i:p_i>0} V_i$ on définit :

$$f(\nu) = \sum_i p_i \log(\nu - V_i) + \log\left(\sum_i \frac{p_i}{\nu - V_i}\right)$$

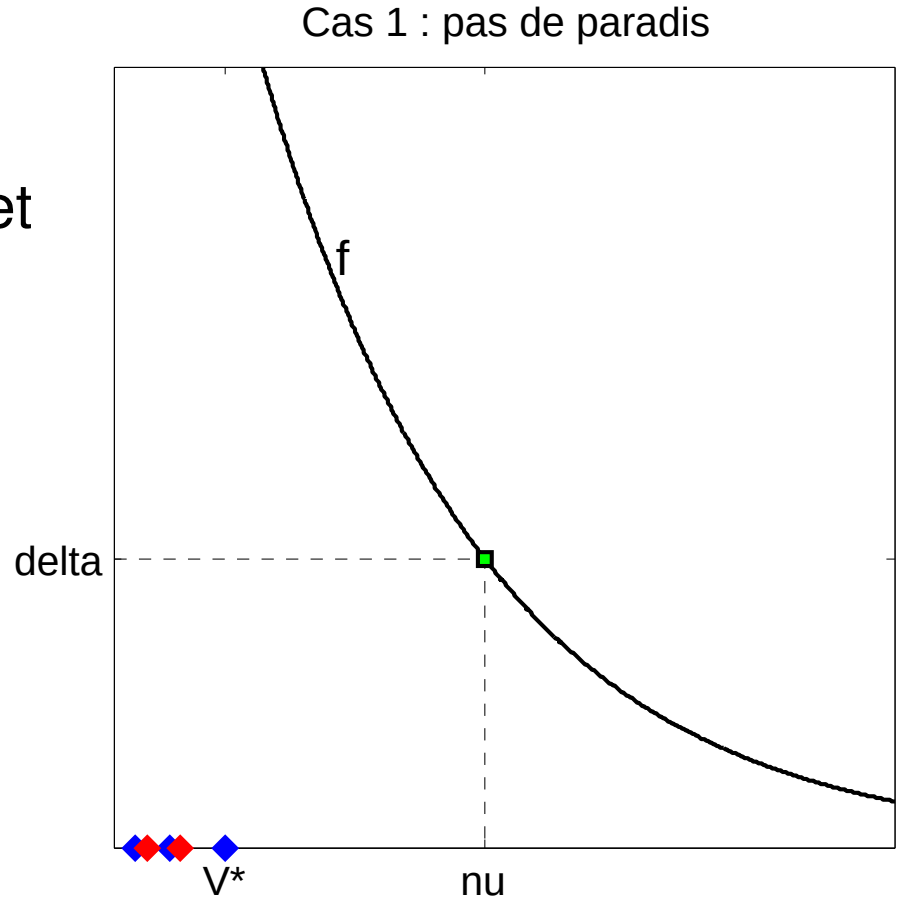
Trouver le maximum

- Soit $i^* = \text{argmax } V_i$. 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

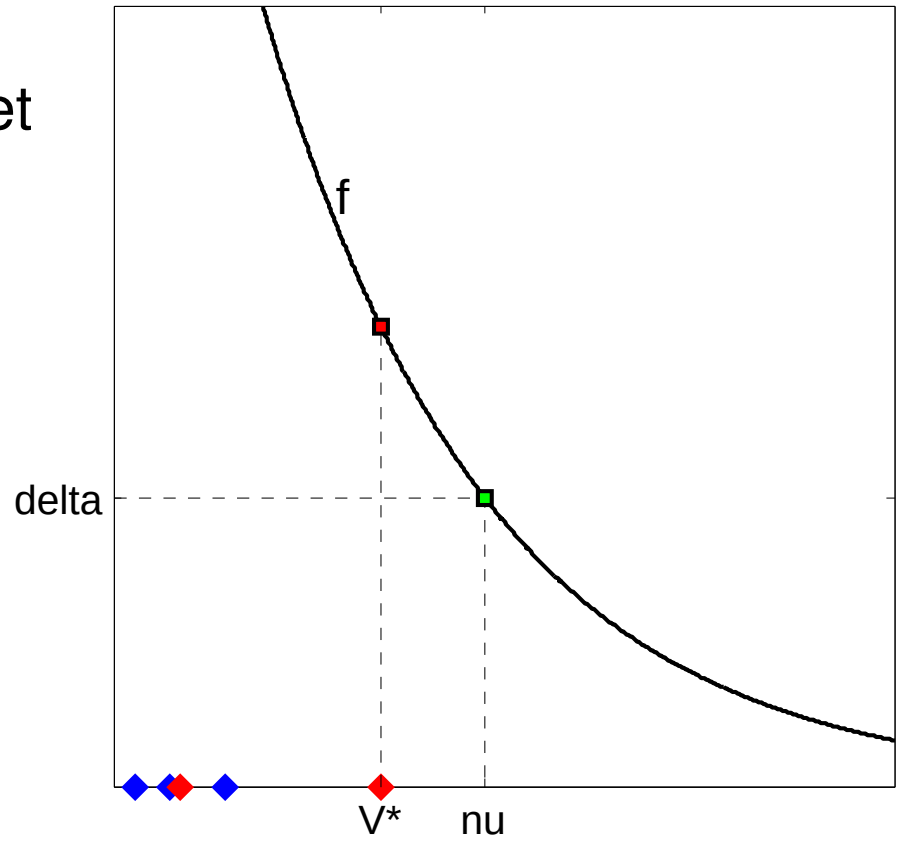
- Soit $i^* = \text{argmax}\, V_i$. Deux possibilités :
- **Cas 1:** si $p_{i^*} > 0$ alors $f(\nu) = \delta$ et

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

- **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



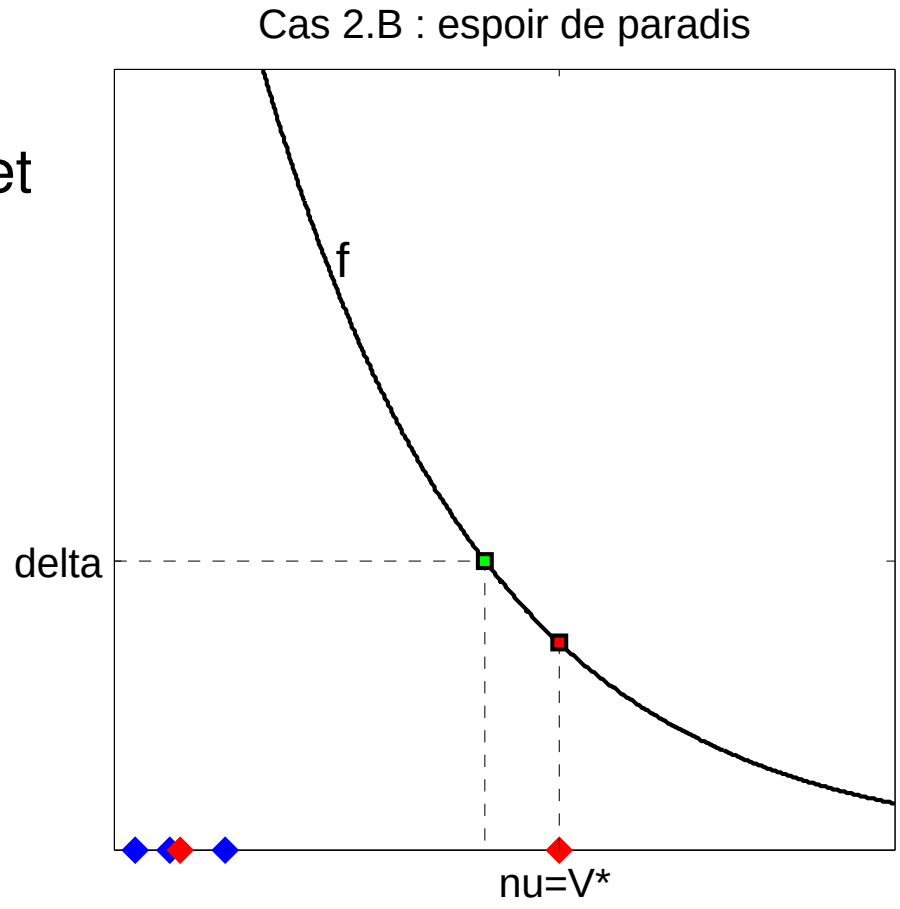
Trouver le maximum

- Soit $i^* = \text{argmax}\, V_i$. 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

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



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

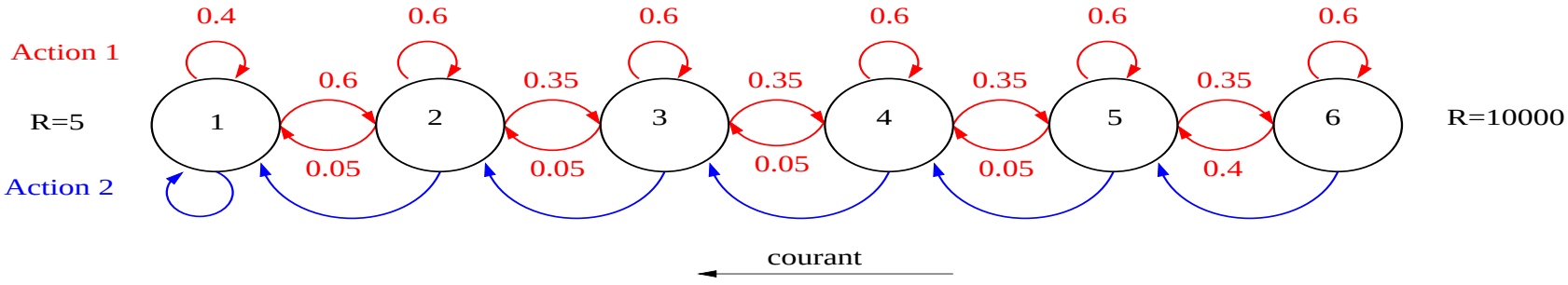
Théorème : Pour un horizon $n > 1$ assez grand, le regret moyen en utilisant l'algorithme KL-UCRL est borné par :

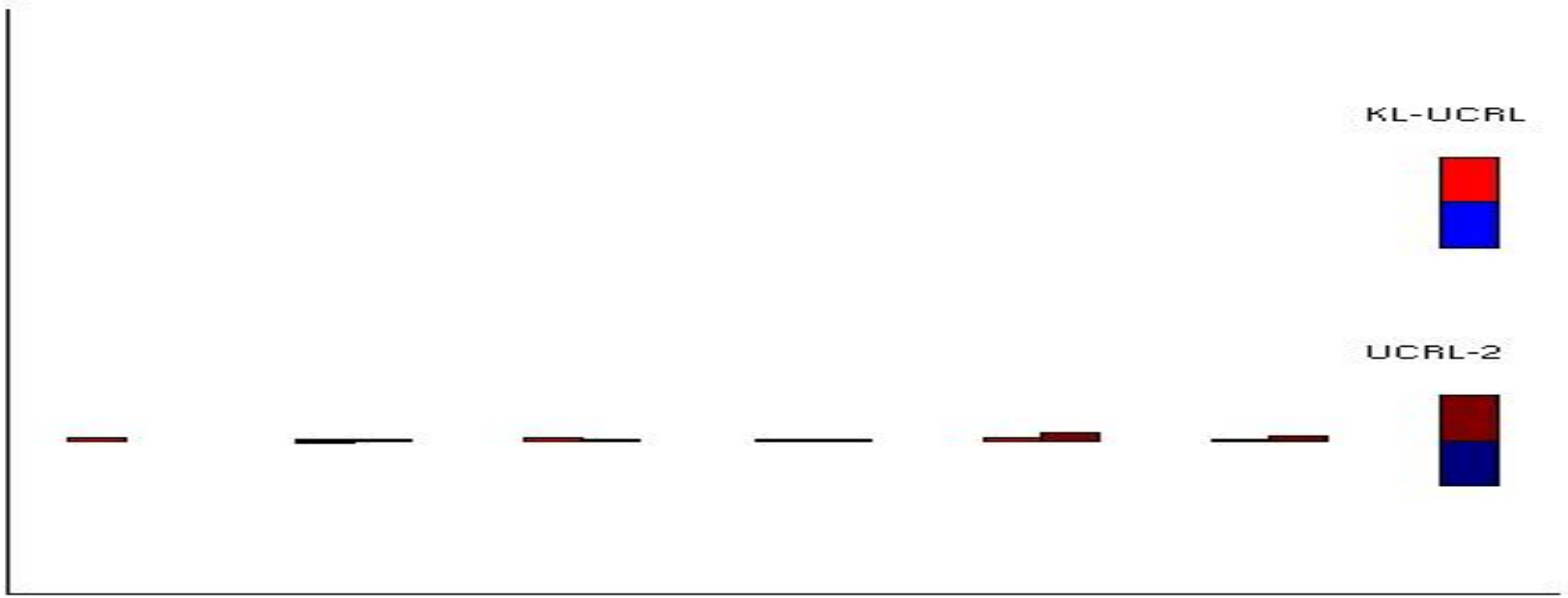
$$\mathbb{E}(\text{Regret}(n)) \leq C|\text{S}|^2|\text{A}|\log(n) \ ,$$

C étant une constante dépendant de

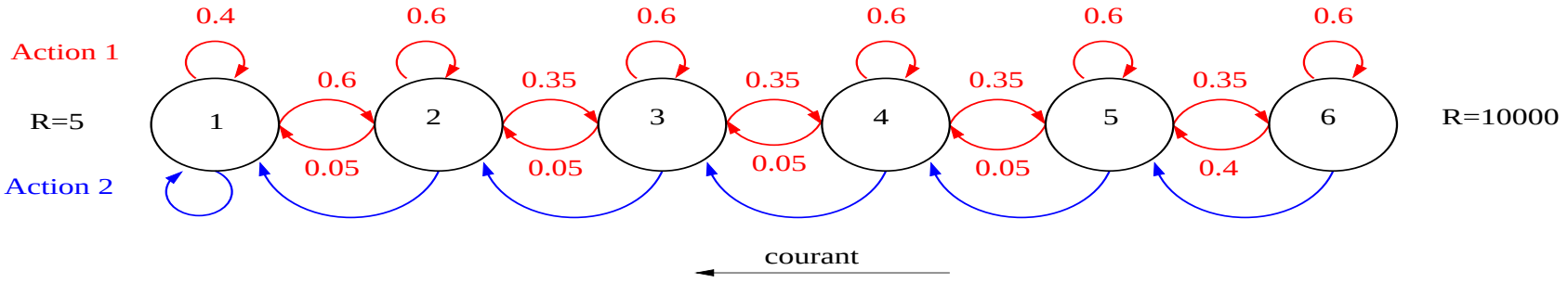
$$\Delta(\mathbf{M}) = \min_{\rho^\pi < \rho^*} \rho^* - \rho^\pi$$

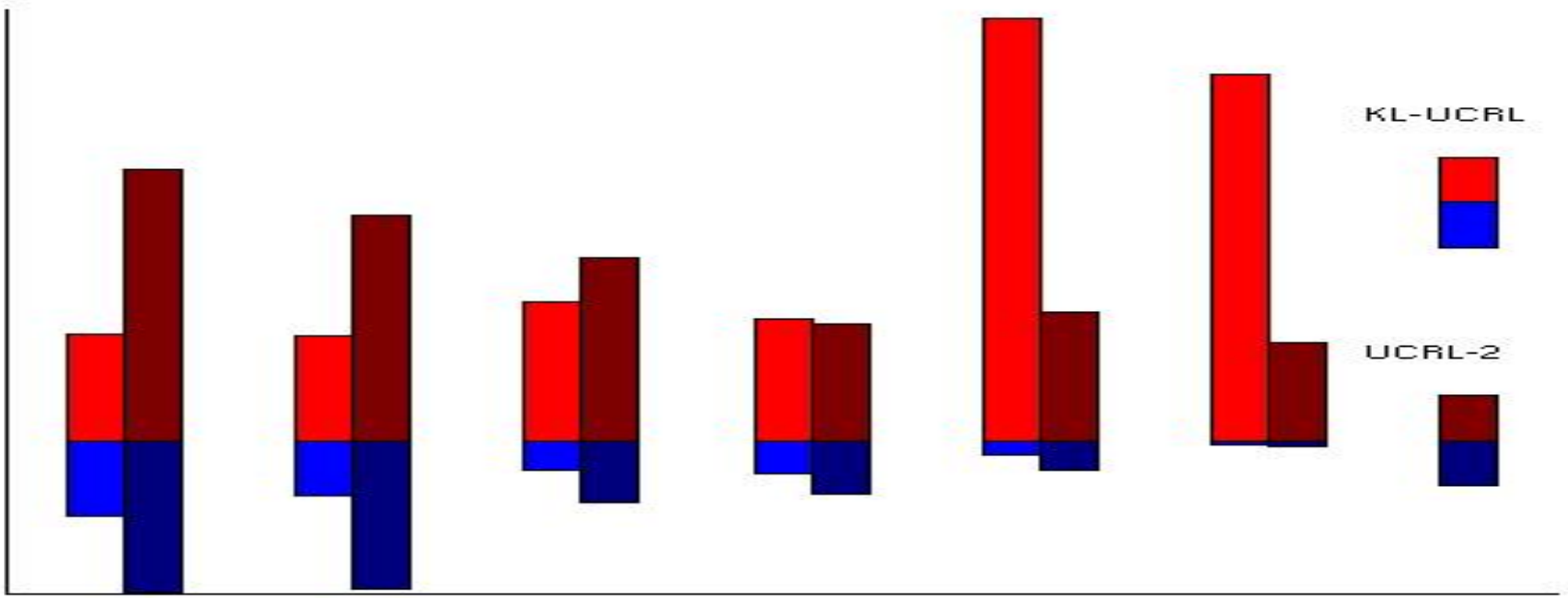
Simulations : RiverSwim





Simulations : RiverSwim





Simulations : RiverSwim

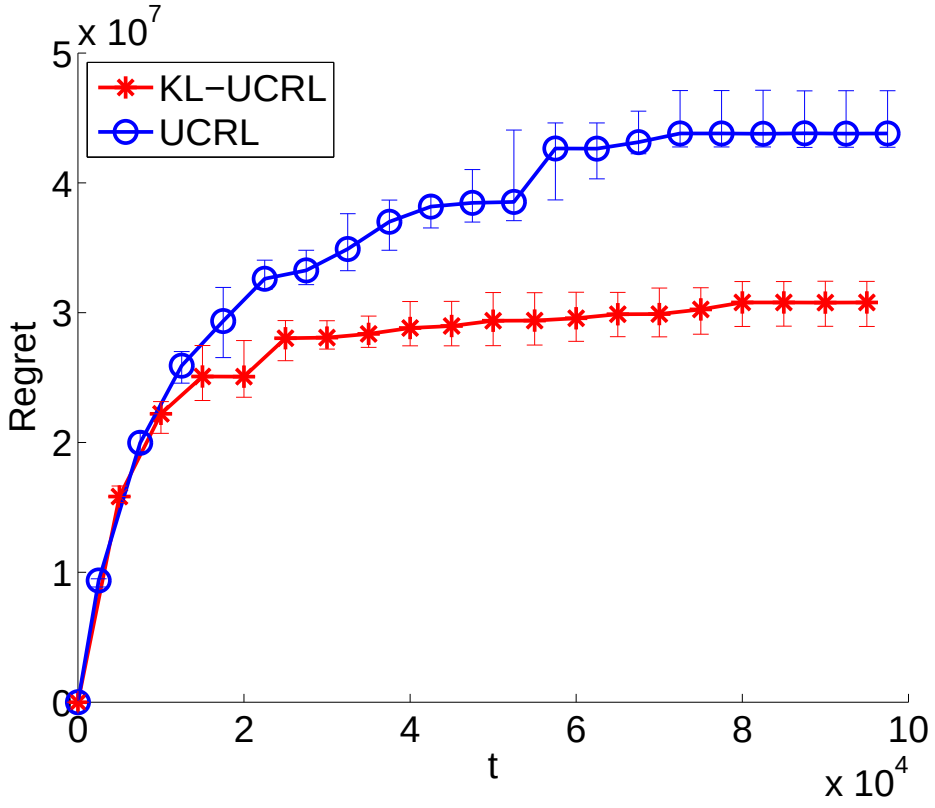


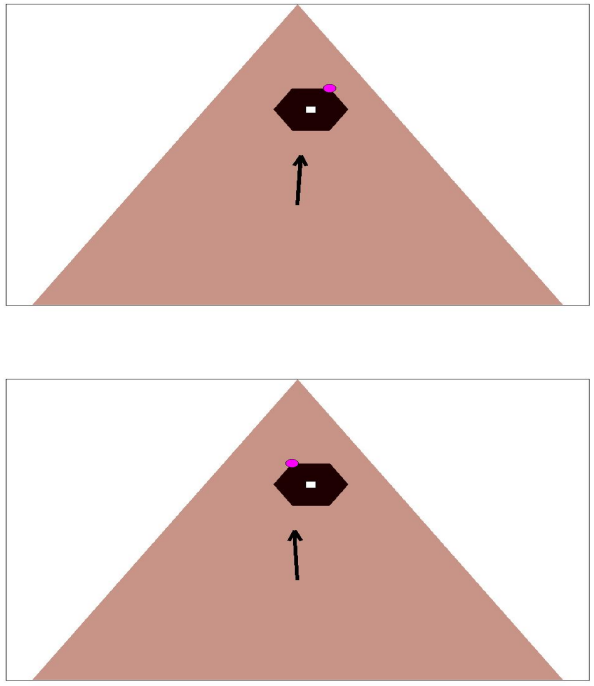
Figure: Comparaison des regrets des algorithmes UCRL-2 et KL-UCRL.

Plan

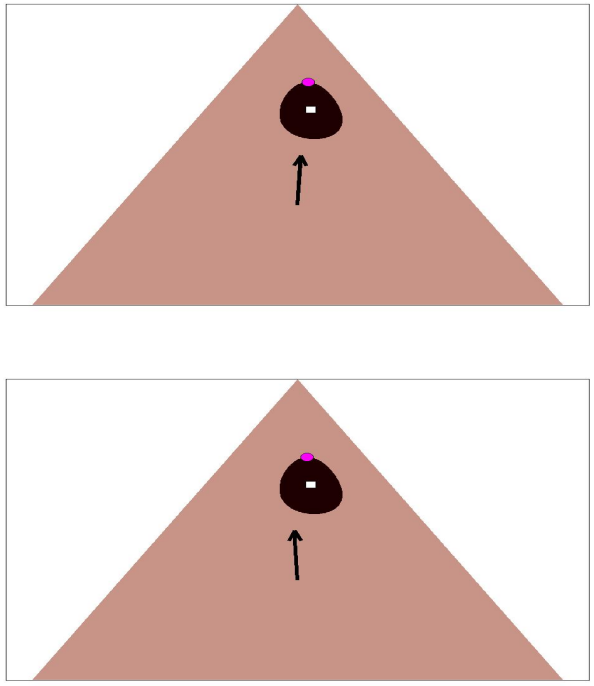
Continuité du modèle optimiste

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

Voisinage L^1



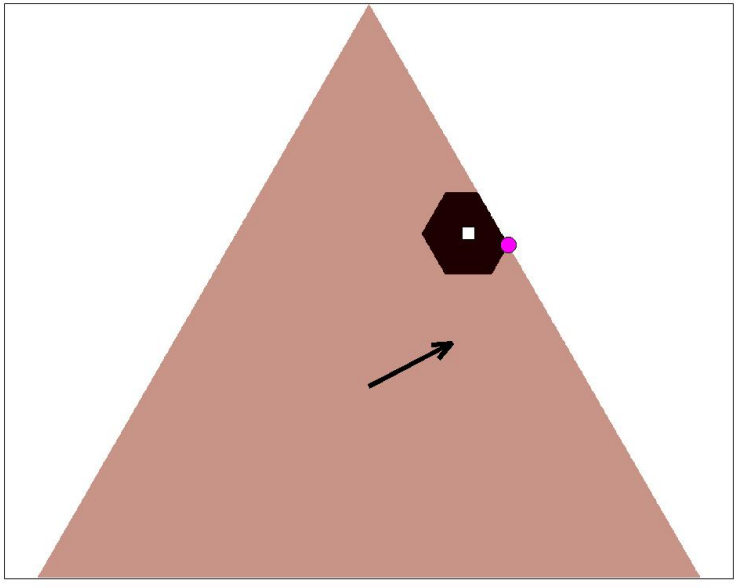
Voisinage KL



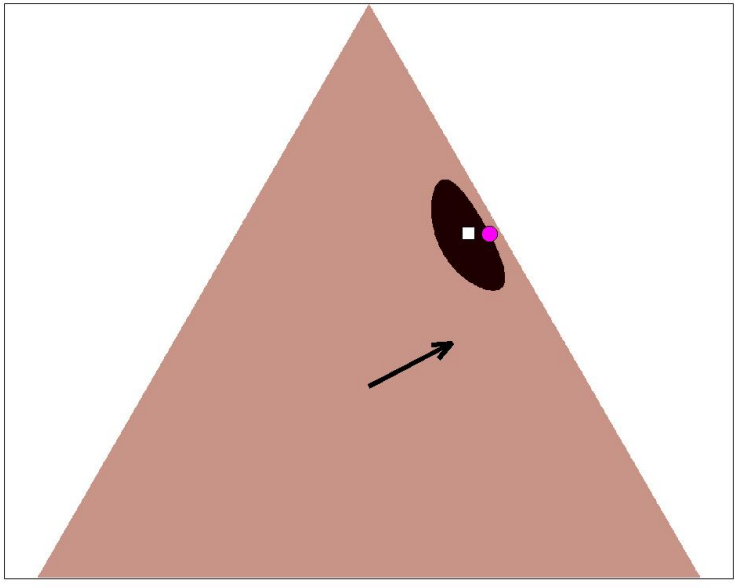
Compatibilité avec les observations

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

Voisinage L^1



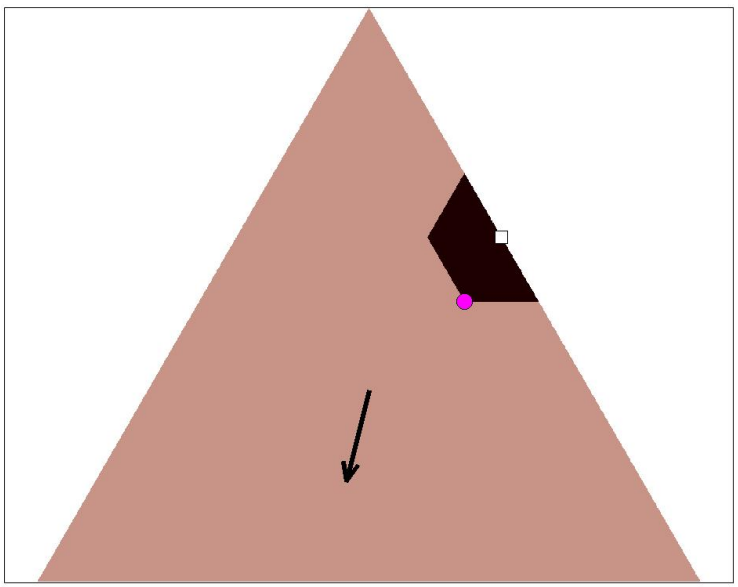
Voisinage KL



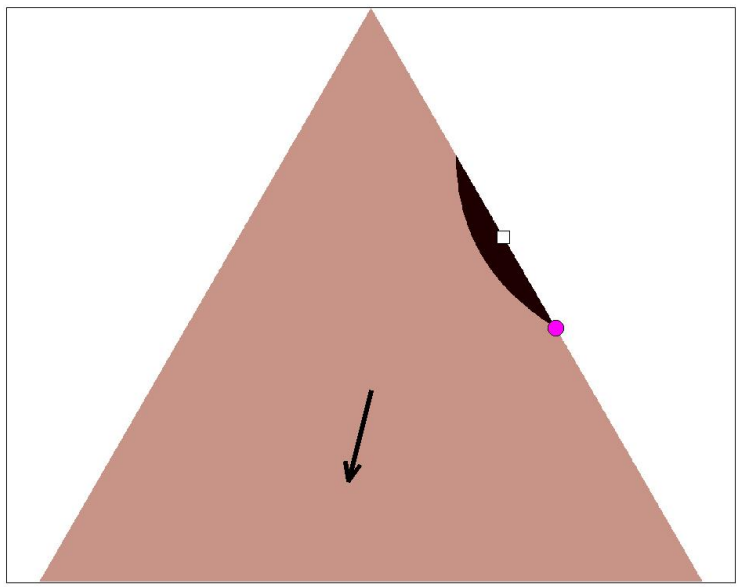
Propriétés des modèles optimistes

Quand une transition de x vers y n'a pas été observée, l'algorithme arbitre entre l'attractivité relative de y et les preuves statistiques accumulées contre l'existence d'une telle transition.

Voisinage L^1



Voisinage KL



Conclusion

- Le calcul du modèle optimiste peut se faire très efficacement avec quelques itérations de Newton
- L'analyse de l'algorithme peut facilement être adaptée aux voisinages KL grâce à l'inégalité de Pinsker
- Il ne nécessite aucune connaissance a priori de la structure du MDP
- Les simulations montrent un comportement significativement meilleur en pratique