Apprentissage par renforcement en environnement complexe et incertain

Rémi Munos

Centre de Mathématiques Appliquées,

Ecole Polytechnique

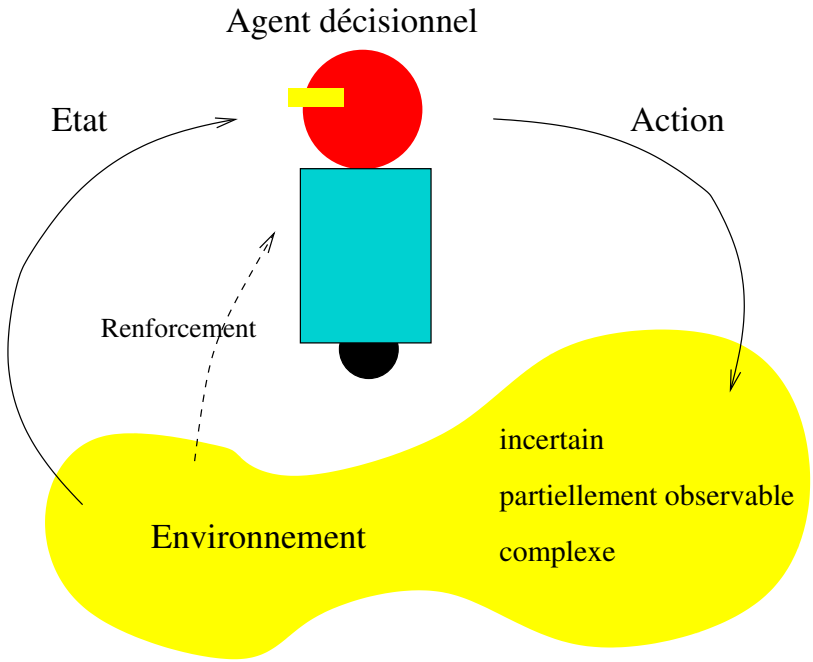
Orsay, le 27 avril 2006

Plan de la présentation:

- Introduction à l'apprentissage par renforcement
- Méthodes de résolution exacte
  - Exemple du temps continu
  - Méthodes de discrétisation adaptative
- Méthodes de résolution approchée
  - Analyse L_p en programmation dynamique
  - Sensibilité par rapport à des paramètres de contrôle
- Projets en cours

Introduction à l'apprentissage par renforcement

Comment une machine peut-elle apprendre à prendre de bonnes décisions, **par l'expérience** ?



Apprentissage par renforcement [Sutton et Barto, 1998]

- Prise de décision en milieu complexe et incertain
- Apprentissage dans toute sa complexité : ce n'est pas un apprentissage supervisé, dynamiques et récompenses *a priori* inconnues, partiellement observable, ...

Applications réalisées et potentielles

- Robotique autonome

- Apprentissage de jeux

- Domotique intelligente

- Contrôle de procédés chimiques,
- Recherche opérationnelle,
- Finance mathématique, ...

Toute tâche de prise de décisions séquentielles que l'on désire améliorer par l'expérience.

Deux approches de résolution

- **Méthodes de résolution exacte**
  Garantie de convergence vers la performance optimale lorsque les ressources computationnelles (mémoire, temps CPU) tendent vers l'infini.

- **Méthodes de résolution approchée**
  Recherche approximative des fonctions d'intérêt dans des espaces fonctionnels donnés et estimation des performances en fonction de la capacités d'approximation des espaces considérés.

Exemple : le temps continu

Dynamique stochastique contrôlée

$$dx_t = f(x_t, a_t)dt + \sigma(x_t, a_t)dW_t.$$

Déterminer un **contrôle** $(a_t)_{t\geq 0}$ qui maximise le **gain**, par exemple

$$J(x, a.) = \mathbb{E}\big[\int_0^T \gamma^t r(x_t, a_t)dt + \gamma^T R(x_T) \,|\, x_0 = x\big],$$

où r et R sont les **fonctions récompense** courante et finale, $0 \leq \gamma \leq 1$ est un coefficient d'actualisation et T un temps de sortie du domaine.

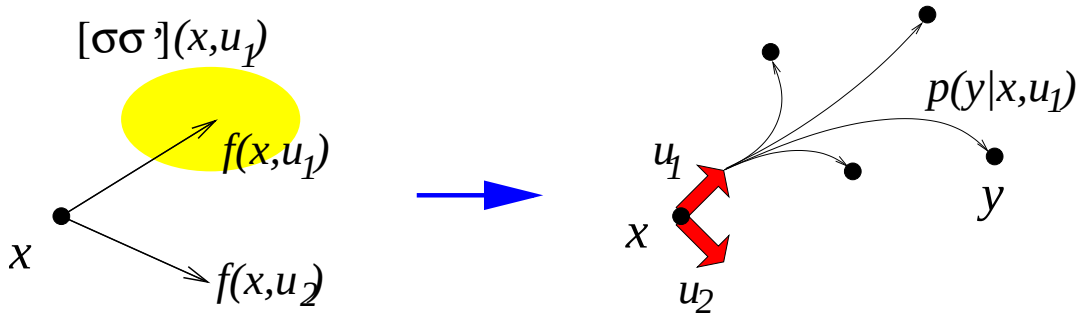
La **fonction valeur**

$$V(x) = \sup_{u.} J(x, a.)$$

est solution d'une équation de *Hamilton-Jacobi-Bellman* (HJB) (au sens des solutions de viscosité).

Convergence des schémas numériques

Discrétisation: Soit X_h une grille de résolution h.



Processus continu	$->$	Processus de Décision Markovien
$dx = f(x,u)dt + \sigma(x,a)dW_t$	$->$	$p(y\|x,a)$
V solution de HJB	$->$	V^h solution d'une équation de PD

La consistance du schéma implique la convergence $V^h \overset{h\to 0}{\longrightarrow} V$, i.e.

$$\begin{cases} \sum_y p(y|x,a)(y-x) &= f(x,a)h + o(h) \\ \sum_y p(y|x,a)(y-x)(y-x)' &= [\sigma\sigma'](x,a)h + o(h) \end{cases}$$

Convergence: Itération sur les valeurs: $V_n^h \overset{n\to\infty}{\longrightarrow} V^h \overset{h\to 0}{\longrightarrow} V$.

Convergence d'algorithmes d'A/R

Si les dynamiques d'état sont inconnues de l'agent –> estimation des probabilités $p(y|x,a)$ par "observation".

Travail de thèse :

- Lien avec les solutions de viscosité
- Utilisation de schémas aux DF et aux EF
- Algorithmes simples, dont la convergence est garantie: $V_n^h \longrightarrow V$ lorsque $n \to \infty$ et $h \to 0$ [Munos, 2000].

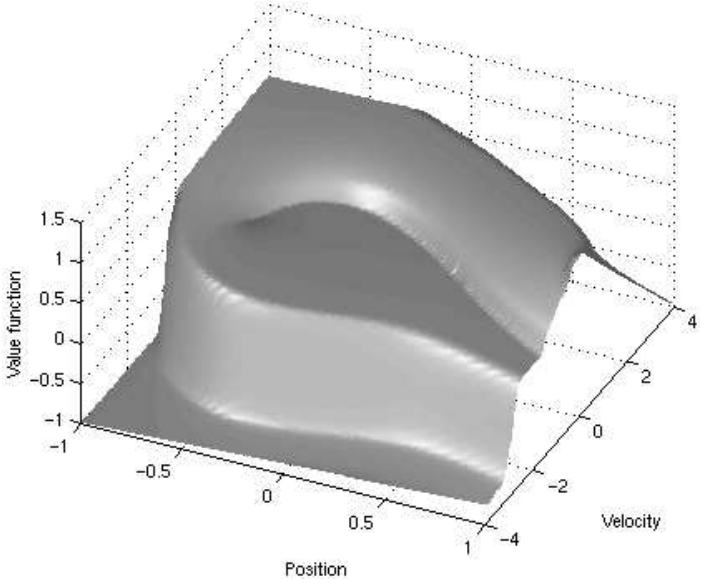
Problème : malédiction de la dimension ! [Bellman, 1957]

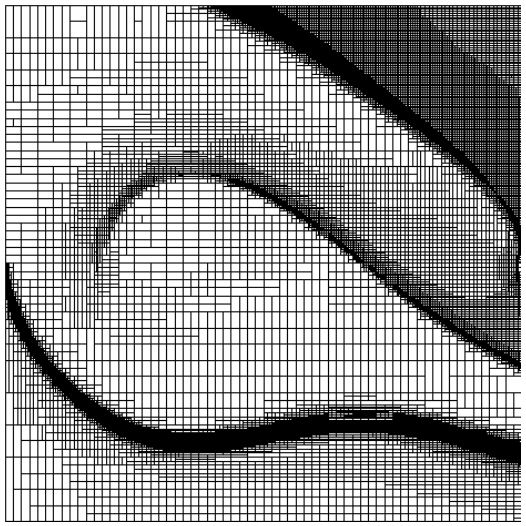
(Complexité en terme de ressources computationnelles $O(1/h^d)$)

Méthodes adaptatives d'allocation de ressources

Travail réalisé avec Andrew Moore
(Robotics Institute, CMU).

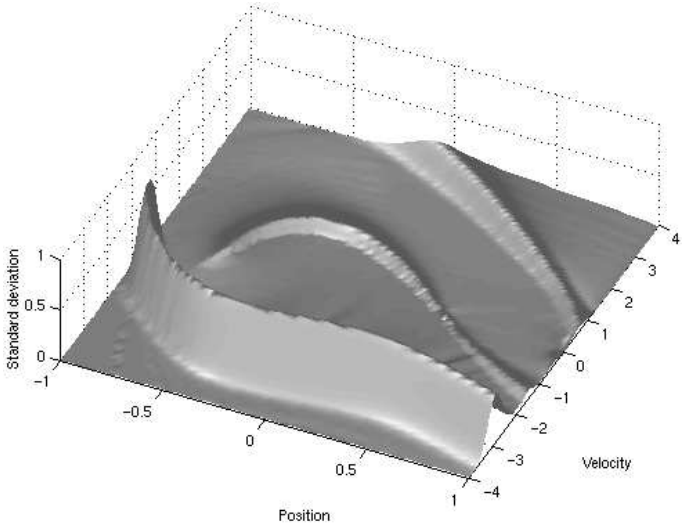
Critère local de raffinement de maillage :

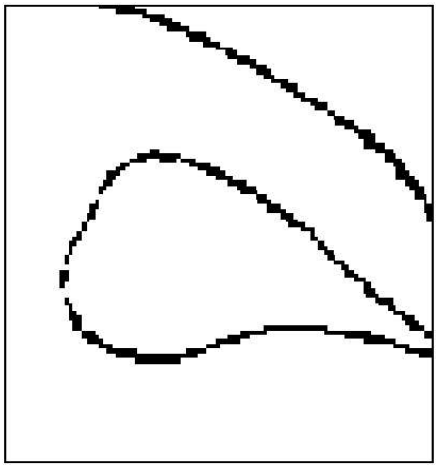




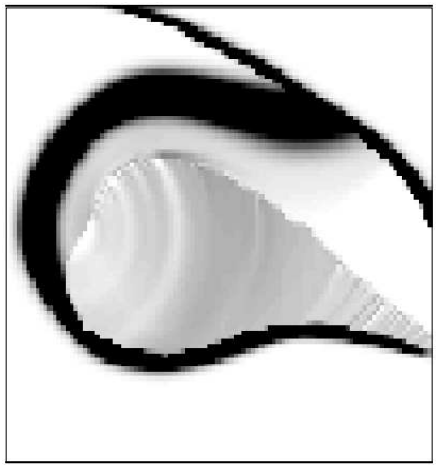
Critère globale de raffinement

Affiner les régions où l'erreur due à l'interpolation a l'influence la plus préjudiciable sur la qualité d'approximation de la fonction valeur aux frontières de transitions de la commande optimale [Munos & Moore, 2002].

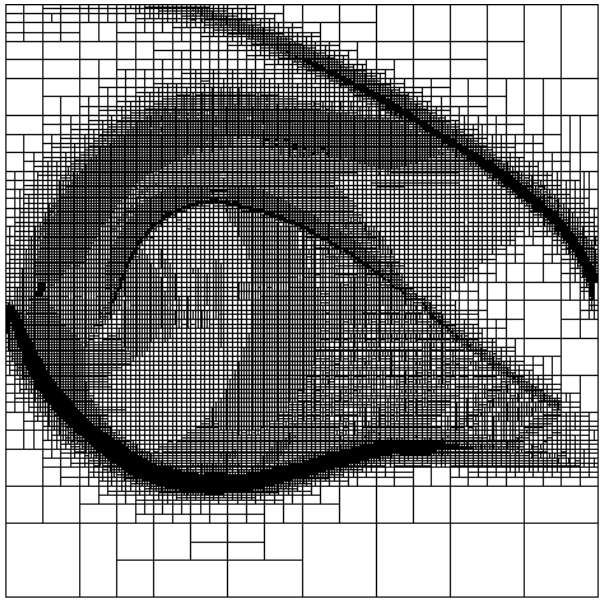




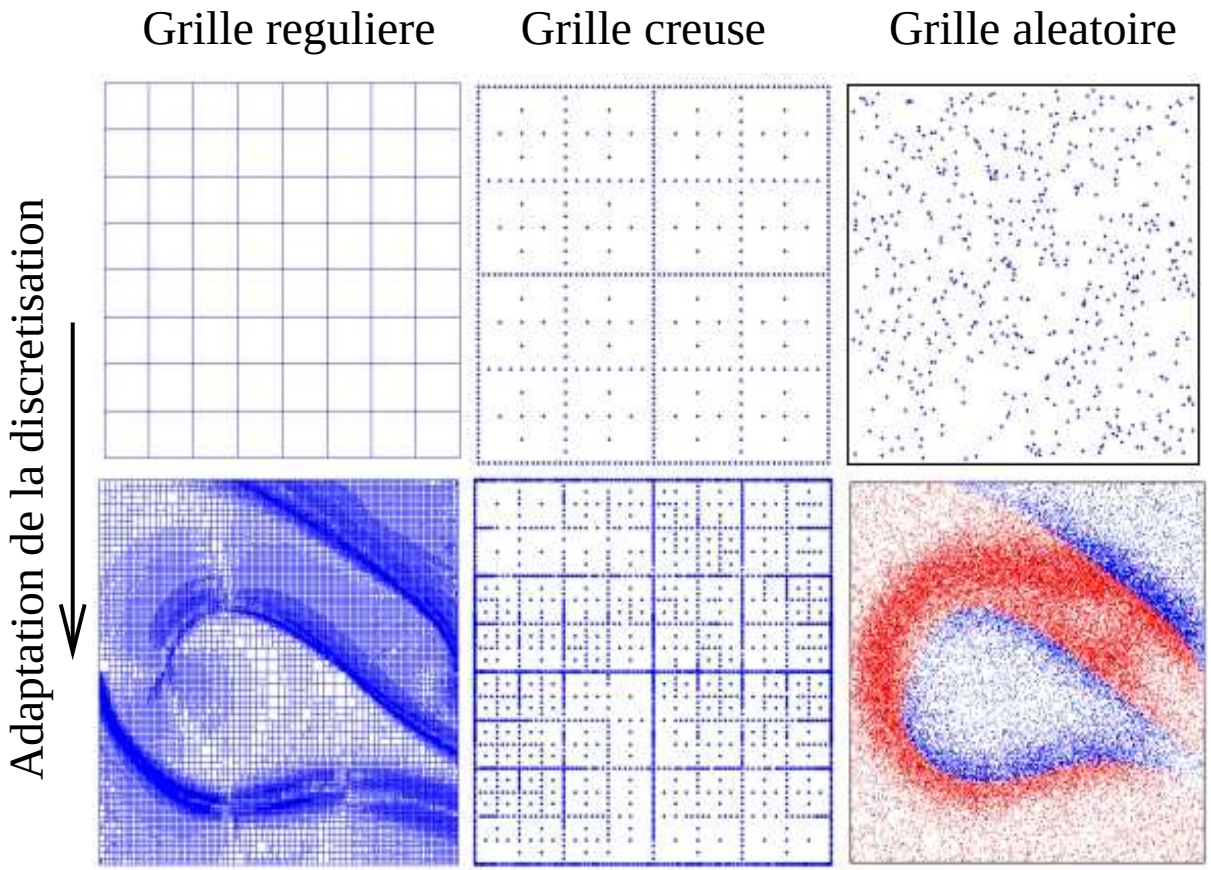
(a) Frontière de transition de la commande



(b) Influence sur ces états



Combinaison avec grilles parsimonieuses



- Résolution fine de problèmes en 4d et 5d. Et 6d avec grilles aléatoires.
- Application DASSAULT-AVIATION.

Méthodes de résolution approchée

Les méthodes de discrétisation sont vouées à l'échec en grande dimension !

–> Nécessaire utilisation de représentations approchées :

- Approches de type **programmation dynamique** de Bellman : utiliser des approximateurs de fonctions pour représenter la fonction valeur,
- Approches de type **principe du maximum** de Pontryagin : recherche directe de politique par optimisation paramétrique.

Programme de recherche :

Méthodes d'apprentissage pour la prise de décision dans l'incertain en grande dimension

Activités de recherche en cours

Approches de type **programmation dynamique**

- Résolution numérique approchée des équations HJB avec réseaux de neurones (Y. Tassa). Délicat mais possible !
- Combinaison de Monte-Carlo et approximation de fonction. Réduction géométrique de variance (avec E. Gobet, S. Maire)
- Application au go (Y. Wang, O. Teytaud, R. Coulom, P.A. Coquelin)
- Méthodes de dérandomisation (avec O. Teytaud).
- **Analyse en norme L_p de la programmation dynamique**

Approches de type **principe du maximum**

- Analyse de sensibilité par outils de calcul stochastique (avec E. Gobet)
- **Extension lorsque les dynamiques sont *a priori* inconnues**

Analyse L_p de la programmation dynamique

Extension de l'analyse usuelle en norme L_∞.

Intérêts :

- Majoration des performances d'algorithmes de programmation dynamique avec approximation en fonction de la capacité et puissance de représentation des espaces fonctionnels considérés
- Lien avec l'Apprentissage Statistique (avec Cs. Szepesvári et P.A. Coquelin)
  - Analyse de complexité, majoration d'erreur de type PAC
  - Représentation des fonctions à l'aide de données

Apprentissage statistique

Analyse L_p en PD

Analyse A/R avec approximation de fonction

Exemple : algorithme d'itérations sur les valeurs

Processus de Décision Markovien : espace d'état X, espace d'action A, noyau de transition $p(dy|x,a)$ et la fonction récompense $r(x,a)$.

Pour une **politique** $\pi : X \to A$, soit un critère de performance (fonction valeur) actualisé ($0 \leq \gamma < 1$) avec horizon temporel infini

$$V^\pi(x) = \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(x_t, a_t) \,|\, x_0 = x,\, a_t = \pi(x_t)\big]$$

La fonction valeur optimale $V^* = \max_\pi V^\pi$ est le point fixe $V^* = \mathcal{T}V^*$ de l'**opérateur de Bellman**:

$$\mathcal{T}f(x) = \max_{a\in A} \big[r(x,a) + \gamma \int p(dy|x,a) f(y)\big].$$

L'opérateur $\mathcal{T}$ est une contraction en L_∞, donc V^* être calculé par itérations sur les valeurs $V_{n+1} = \mathcal{T}V_n$.

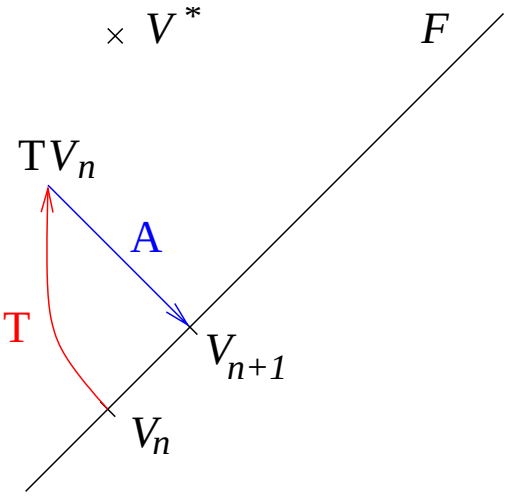
Itération sur les valeurs avec approximation

La fonction valeur optimale V^* est approchée par **itérations sur les valeurs avec approximation**:

$$V_{n+1} = \mathcal{A}\mathcal{T}V_n,$$

où $\mathcal{T}$ est l'*opérateur de Bellman* et $\mathcal{A}$ un *opérateur d'approximation.*

Exemple : $\mathcal{F}$ est un sous-espace vectoriel de dimension finie, et $\mathcal{A}$ une projection orthogonale (norme L_2) sur $\mathcal{F}$.



Propriétés :

- $\mathcal{T}$ est une contraction en norme L_∞,
- $\mathcal{A}$ est non-expansif en norme L_2

Problème : on ne peut rien dire de l'opérateur combiné $\mathcal{A}\mathcal{T}$!

Analyse L_∞ de l'algorithme IVA

[Bertsekas & Tsitsiklis, 1996] Majoration sur la *perte* $V^* - V^{\pi_n}$ en fonction des *erreurs d'approximation* $e_n = \mathcal{T}V_n - \mathcal{A}\mathcal{T}V_n$:

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{n\to\infty} ||e_n||_\infty.$$

Problème : la norme L_∞ est très peu appropriée à l'approximation de fonction, surtout à partir de données :

- L'erreur uniforme $||e_n||_\infty = \sup_{x\in X} |e_n(x)|$ est difficilement évaluable en fonction de l'erreur empirique $|e_n(x_k)|$ aux données $\{x_k\}_{1\leq k\leq K}$,
- En général, un opérateur d'approximation réalise un problème de minimisation empirique en norme L_p (ex. régression linéaire, réseaux de neurones, Support Vectors, méthodes à noyaux, ...).

Analyse L_p de l'algorithme IVA

Soit μ une distribution sur X. Rappel : $||f||_{p,\mu} = \left(\int \mu(dx)|f(x)|^p \right)^{1/p}$.

Définissons $C(\mu)$ telle que $p(\cdot|x,a) \leq C(\mu)\mu(\cdot)$, pour tous $x \in X$, $a \in A$. Alors [Munos, 2006] :

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} C(\mu)^{1/p} \limsup_{n\to\infty} ||e_n||_{p,\mu}.$$

Majoration en fonction de l'erreur d'approximation en norme L_p.

Apprentissage statistique : majoration de $||e_n||_{p,\mu}$, *risque structurel* (ou erreur en généralisation), en fonction du *risque empirique* (ou erreur en apprentissage) $\left[\frac{1}{K}\sum_{k=1}^K |e_n(x_k)|^p\right]^{1/p}$ réellement minimisé,

$$||e_n||_{p,\mu} \leq \left[\frac{1}{K}\sum_{k=1}^K |e_n(x_k)|^p\right]^{1/p} + E\big(K, \mathrm{VC}(\mathcal{F}), \dots\big)$$

et de la capacité de $\mathcal{F}$ (dimension de VC, nombre de couverture, ...).

Exemple d'algorithme d'A/R avec approximation

Hypothèse : on dispose d'un modèle génératif, i.e. tirer $y \sim p(\cdot|x,a)$.

Exemple : Construit N approximations $\{V_n\}_{1\leq n\leq N}$ successivement, selon

1. Selection de K états $(x_k)_{k=1\ldots K}$ tirés selon μ,
2. En chaque état x_k, pour chaque action a, tirage de M états suivants $\{y_{k,a}^m\}_{1\leq m\leq M} \sim p(\cdot|x_k,a)$ utilisant un modèle génératif, et calcul de l'opérateur de Bellman échantillonné

$$v_k = \max_a \left[r(x_k,a) + \gamma \frac{1}{M} \sum_{m=1}^{M} V_n(y_{k,a}^m) \right]$$

3. Détermine $V_{n+1} \in \mathcal{F}$, par un problème de minimisation:

$$V_{n+1} = \arg\min_{f\in\mathcal{F}} \frac{1}{K} \sum_{k=1}^{K} l\big(f(x_k) - v_k\big)$$

où l est une fonction risque (quadratique, absolue, $\epsilon-$insensitive, ...).

Résultats de majoration PAC

[Munos & Szepesvári, 2006] Soient $\epsilon > 0$ et $\delta > 0$. Supposons que $\mathcal{F}$ est tel que la distance entre $\mathcal{T}\mathcal{F}$ et $\mathcal{F}$ est $\leq \epsilon$, i.e.

$$\sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} ||\mathcal{T} g - f||_{p,\mu} \leq \epsilon.$$

Alors il existe des entiers N (nombre d'itérations), K (nombre de données) et M (nombre d'états suivants) avec N linéaire en $\log \frac{1}{\epsilon}$, K et M polynomiaux en $\frac{1}{\epsilon}$, $\log \frac{1}{\delta}$, $\log \mathcal{N}(\epsilon/C(\mu), \mathcal{F})$, tels que

$$\mathcal{P}\Big(||V^* - V^{\pi_N}||_\infty < \epsilon\Big) \geq 1 - \delta.$$

Dépendance par rapport à la dimension : on en déduit une complexité d'échantillons en fonction de la dimension $O\big((\log 1/\epsilon)^d\big)$ au lieu du résultat de [Chow et Tsitsiklis, 1989] en $O\big(1/\epsilon^d\big)$.

Autres algorithmes de programmation dynamique

Toute l'analyse L_∞ en programmation dynamique avec approximation se généralise à la norme L_p.

- **Itérations sur les politiques** [Munos, 2003]. Performance asymptotique

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} C(\mu)^{1/p} \varepsilon_{\mathcal{F}}.$$

  majorée en fonction de la capacité représentationnelle de l'architecture d'approximation $\varepsilon_{\mathcal{F}} = \limsup_n \inf_{f\in\mathcal{F}} ||V^{\pi_n} - f||_{p,\mu}$.

- **Minimisation du résidu de Bellman** [Munos, 2006]

$$||V^* - V^{\pi}||_\infty \leq \frac{2}{1-\gamma} C(\mu)^{1/p} ||\mathcal{T}V - V||_{p,\mu}$$

- et toutes leurs variantes apprentissage par renforcement.

Perspectives

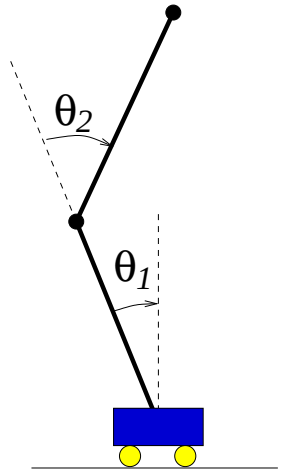
- Analyse en normes L_p identiques à celle utilisée pour le problème de minimisation –> finesse et utilité applicative.
- **Bénéficier des avancées récentes en Apprentisage Statistique**
  - Analyse A/R avec représentation de fonctions à partir de données (SVM, méthodes à noyaux dans RKHS, ...)
  - Méthodologie méthodes à noyaux (avec P.A. Coquelin)
  - Intégration dans plateforme *OpenDP* (O. Teytaud et S. Gelly), et combinaison avec méthodes dérandomisation (ANR ARPRODY soumise, application Cemagref).
- Projet INRIA Futurs SequeL (Sequential Learning), (avec Ph. Preux, M. Davy, R. Coulom, E. Duflos, Ph. Vanheeghe),
- Workshop ICML 2006, *Kernel Methods and Reinforcement Learning* (avec Ph. Preux, M. Davy, Cs. Szepesvári)

Analyse de sensibilité du critère par rapport à des paramètres de contrôle

Recherche directe d'une politique dans une classe de fonctions approchées

- Problème d'optimisation paramétrique
- Processus partiellement observables
- Domaine très actif actuellement

[Ng, 2005] 



Robotique mobile

- **Résultats** : convergence de l'estimateur du gradient lorsque les dynamiques d'état sont inconnues [Munos, 2006]

Problème d'optimisation pramétrique

Dynamique déterministe temps continu :

$$\frac{dx_t}{dt} = f(x_t, a_t), \quad x_0 = x.$$

On cherche une politique $a_t = \pi_\alpha(t, x_t)$ paramétrée qui maximise le gain.
Problème d'optimisation: déterminer α qui maximise

$$\alpha \to V^{\pi_\alpha} = \int \gamma^t r(x_t) dt.$$

Si l'on utilise un algorithme de type gradient : $\alpha \leftarrow \alpha + \eta \nabla_\alpha V^{\pi_\alpha}$, on a besoin de déterminer le **gradient** $\nabla_\alpha V^{\pi_\alpha}$. Dérivée trajectorielle :

$$\nabla_\alpha V^{\pi_\alpha} = \int \gamma^t \nabla r(x_t) \nabla_\alpha x_t dt.$$

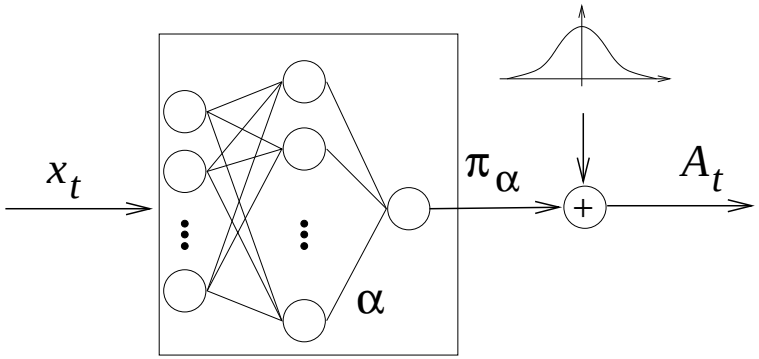
En notant $z_t = \nabla_\alpha x_t$, on a $z_0 = 0$ et avec $f_\alpha(x) = f(x, \pi_\alpha(x))$,

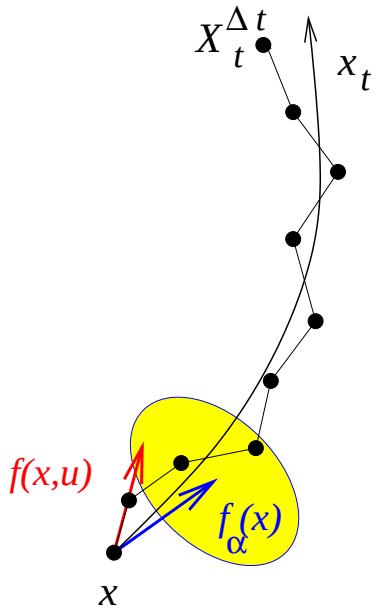
$$\frac{dz_t}{dt} = \nabla_\alpha f_\alpha(x_t) + \nabla_x f_\alpha(x_t) z_t.$$

Algorithmes d'apprentissage par renforcement ?

Comment estimer $z_t = \nabla_\alpha x_t$ lorsque la dynamique d'état f_α est inconnue ?

Astuce : perturber aléatoirement la politique (i.e. utiliser une *politique stochastique* en gardant un contrôle constant sur des intervalles Δt).





Processus stochastique $(X_t^{\Delta t})$:

- A_t est choisie selon la politique perturbée
- Cette action est maintenue constante pendant Δt.

Propriété de consistance locale $\mathbb{E}[X_{t+\Delta t}^{\Delta t} - X_t^{\Delta t}] = f_\alpha(X_t^{\Delta t})\Delta t + o(\Delta t)$,

Alors : [Munos, 2006] $(X_t^{\Delta t}) \overset{\Delta t \to 0}{\longrightarrow} x_t$ presque sûrement.

Convergence algorithme d'A/R

Algorithme : au cours de la trajectoire $(X_t^{\Delta t})$, on calcule $(Z_t^{\Delta t})$, discrétisation du gradient $z_t = \nabla_\alpha x_t$, selon

$$\begin{aligned} Z_{t+\Delta t}^{\Delta t} &= Z_t^{\Delta t} + \nabla_\alpha \log \pi_\alpha(A_t|X_t^{\Delta t})\Delta X_t^{\Delta t} \\ &\quad + \nabla_x \log \pi_\alpha(A_t|X_t^{\Delta t})\Delta X_t^{\Delta t} Z_t^{\Delta t} + \widehat{\nabla_x f}(X_t^{\Delta t}, A_t) Z_t^{\Delta t} \Delta t, \end{aligned}$$

où $\widehat{\nabla_x f}$ est un estimateur moindres-carrés :

$$\widehat{\nabla_x f}(X_t^{\Delta t}, A_t) = (\Delta t)^{-1} \left(\overline{\Delta X\, X'} - \overline{\Delta X}\,\overline{X'}\right)\left(\overline{X\, X'} - \overline{X}\,\overline{X'}\right)^{-1}.$$

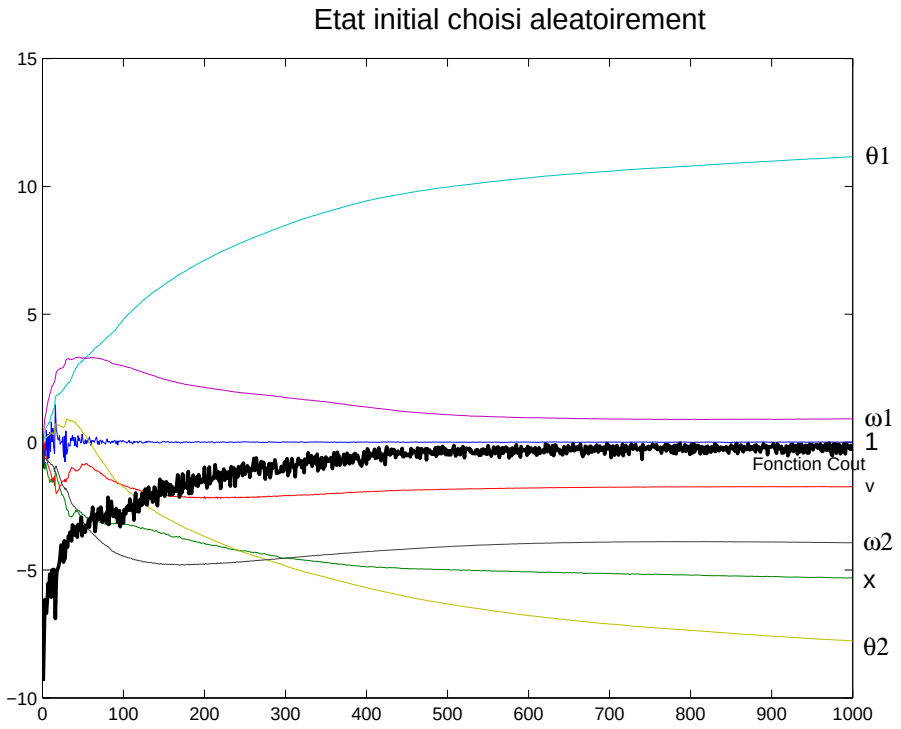
Convergence : [Munos, 2006] On a $(X_t^{\Delta t}, Z_t^{\Delta t}) \overset{\Delta t \to 0}{\longrightarrow} (x_t, z_t)$ presque sûrement, et l'estimateur trajectoriel converge :

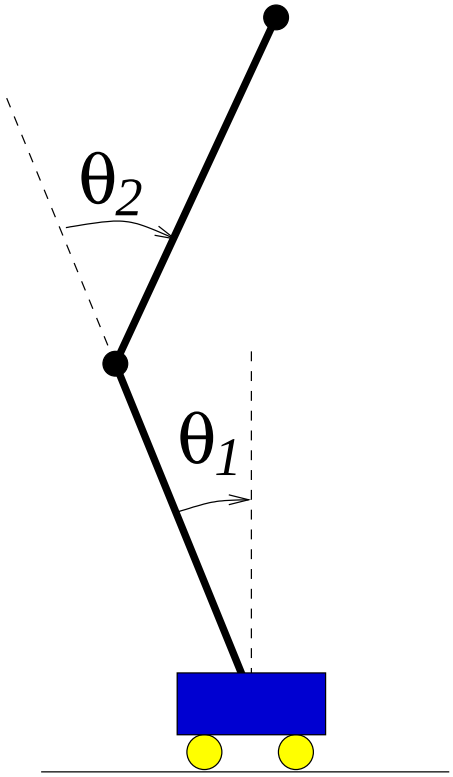
$$\int \gamma^t \nabla r(X_t^{\Delta t}) Z_t^{\Delta t} dt \overset{\Delta t \to 0}{\longrightarrow} \nabla_\alpha V^{\pi_\alpha}.$$

Le calcul du gradient ne nécessite que la connaissance de la politique π_α et l'observation de la trajectoire $X_t^{\Delta t}$.

Exemple : double pendule inversé

Etat 6 dimensions: x, v, θ_1, ω_1, θ_2, ω_2. Contrôle : force appliquée au chariot.





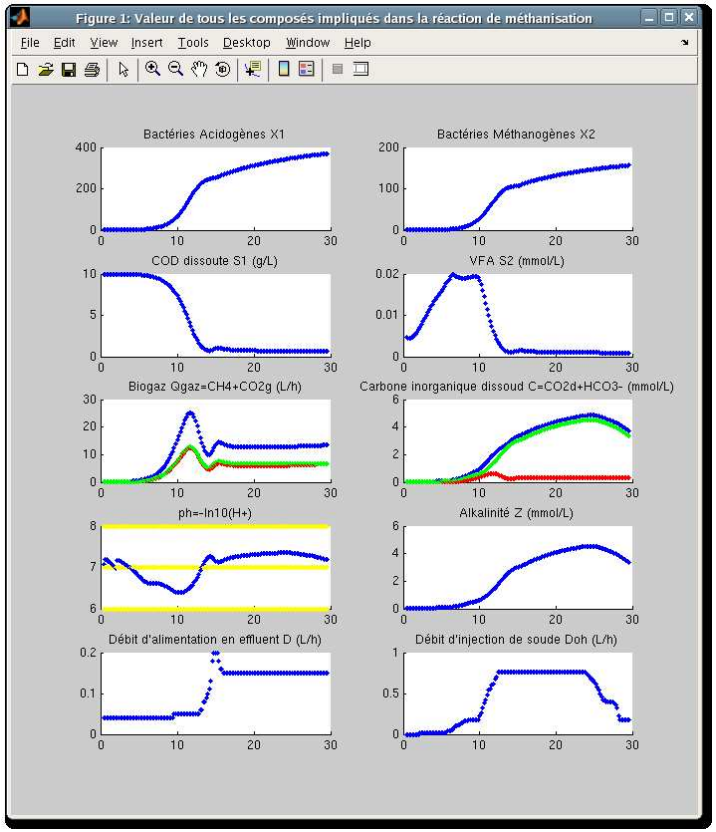
Contrôle obtenu:

$$\pi_\alpha = -0.0023 - 5.31y - 1.74v + 11.16\theta_1 + 0.92\omega_1 - 7.77\theta_2 - 3.94\omega_2.$$

Application en chimie industrielle

Avec P.A. Coquelin, F. Champlain (CMAP)

- **Naskeo Environnement** : convertir pollution industrielle organique en énergie renouvelable par un procédé de digestion anaérobie.
- **Brevet 2006** : A/R pour la commande adaptative du débit de l'effluent (et alcalinisation)
- **Méthode** : recherche directe de politique paramétrée par dérivée trajectorielle avec estimation de l'état par filtrage particulaire.



Autres projets en cours

Actuels :

- Workshop **Kernel Machines and Reinforcement Learning**, ICML 2006 (avec M. Davy, Ph. Preux, Cs. Szepesvári)
- IEEE International Symposium on **Approximate Dynamic Programming and Reinforcement Learning**, 2007, (avec D. Fogel, D. Liu, J. Si, D.C. Wunsch) célébrant le 50e anniversaire de [Bellman, 1957]. **"Towards breaking the curse!"**
- ANR **ADAP'MC** (Méthodes de Monte-Carlo adaptatives). Responsable : E. Moulines, 2006-2008.
- Groupe **PDMIA**, depuis 2001.

Avec le LRI :

- ANR **ARPRODY**, avec O. Teytaud, S. Gelly, F. Garcia.
- OMT dans un cadre ANR proposé par **Digitéo Labs**. Projet de type "synergie" entre CMAP et LRI (avec O. Teytaud).
- Co-encadrement Y. Wang : programmation dynamique et MC pour le go (avec O. Teytaud, S. Gelly, P.A. Coquelin, R. Coulom).

Au sein du LRI

Equipe Inférence et Apprentissage (Michèle Sebag)

- Création d'un groupe **prise de décision dans l'incertain, pour la grande dimension**
  - Renforcer compétences programmation dynamique (O. Teytaud et collègues *OpenDP*)
  - Méthodes d'apprentissage pour la PD (SVAPI, Adaptive MC, KMP & RL)
  - Méthodes pour l'estimation de performance
  - Traiter applications complexes (Naskeo, go, Cemagref, Robotique)
- Interactions
  - Optimisation (M. Schoenauer)
  - Fouille de données (M. Sebag)
  - Apprentissage statistique (P. Massart, Probas et Stats)

Enseignements proposés

- Fondements mathématiques pour l'informatique
- Programmation mathématique
- Optimisation dynamique
- Recherche opérationnelle
- Fouille de données, Apprentissage automatique
- Intelligence artificielle et sciences cognitives
- Traitement du signal, vision, parole, théorie de l'information

Collaborateurs proches

En France :

- CMAP (E. Gobet, R. Douc, P.A. Coquelin, Y. Wang, F. Champlain)
- CREA (P. Bourgine, H. Frankowska)
- LRI (O. Teytaud, S. Gelly)
- INRA (F. Garcia)
- LMPT (G. Barles)
- GRAPPA (Ph. Preux, R. Coulom, M. Davy)
- LORIA (B. Scherrer)
- LIP6 (O. Sigaud)
- CEMAGREF (G. Deffuant, S. Martin)
- ENSTA (H. Zidani)
- LJLL (O. Bokanowski)

A l'étranger :

- SZTAKI (Cs. Szepesvári, A. Antos)
- HUJ (Y. Tassa)
- NASA (N. Meuleau)
- CMU (A. Moore, L. Baird, J. Bagnell, J. Schneider, G. Gordon)