Master MVA: Apprentissage par renforcement **Lecture: 4**

Programmation dynamique avec approximation

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

Références bibliographiques:

- Références générales:
  - Livre PDMIA, chapitre 11: "Programmation dynamique avec approximation".
  - Bertsekas et Tsitsiklis: *Neuro Dynamic Programming*, 1996.
- Minimisation du résidu de Bellman:
  - Baird: *Residual Algorithms: Reinforcement Learning with Function Approximation,* 1995.
  - Williams et Baird: *Tight performance bounds on greedy policies based on imperfect value functions*, 1993.
- Itérations sur les valeurs:
  - Munos and Szepesvári: *Finite time bounds for sampling based fitted value iteration*, 2008.
  - Munos: *Performance bounds in L_p norm for approximate value iteration*, 2007.
- Itérations sur les politiques:
  - Tsitsiklis et Van Roy: *An analysis of Temporal Difference learning with function approximation*, 1996.
  - Bradtke et Barto: *Linear Least-Squares algorithms for Temporal Difference learning*, 1996.
  - Lagoudakis et Parr: *Least squares policy Iteration*, 2003.
  - Munos: *Error Bounds for Approximate Policy Iteration*, 2003.

Plan:

1. Approximation de la fonction valeur
2. Minimisation du résidu de Bellman
3. Itération sur les valeurs avec approximation
4. Itération sur les politiques avec approximation

1 Approximation de la fonction valeur

Dans ce chapitre on considérera le cas d'un critère à horizon temporel infini avec récompenses actualisées ($\gamma < 1$).

En général la fonction valeur optimale V^* n'est pas calculable exactement (l'espace d'état $\mathcal{X}$ peut être grand, voire infini) donc il est nécessaire de considérer des représentations approchées, et de savoir évaluer l'impact de ces approximations sur la performance des politiques qui s'en déduisent.

Hypothèse implicite de la programmation dynamique avec approximation: une bonne approximation de la fonction valeur optimale induit une politique dont la performance est proche de l'optimum. Cette hypothèse est validée par la proposition suivante.

Considérons une approximation V de la fonction valeur optimale V^*. Par simplicité de notations, on considère que $\mathcal{X}$ est un espace fini avec N états et que A est aussi fini. Donc V peut être considérée comme un vecteur de $\mathbb{R}^N$. Notons π une politique déduite de V (i.e., gloutonne par rapport à V), c'est à dire telle que:

$$\pi(x) \in \arg\max_{a \in A} \sum_y p(y|x,a)\big[r(x,a,y) + \gamma V(y)\big].$$

le résultat suivant donne une majoration sur la perte en performance $\|V^* - V^\pi\|_\infty$ résultante de l'utilisation de la politique π plutôt que de la politique optimale, en fonction de l'erreur d'approximation $\|V^* - V\|_\infty$.

Proposition 1. Considérons un problème à horizon temporel infini avec actualisation. Soit $V \in \mathbb{R}^N$ une fonction et notons π une politique déduite de V. Alors

$$\|V^* - V^\pi\|_\infty \leq \frac{2\gamma}{1-\gamma}\|V - V^*\|_\infty.$$

De plus, il existe $\epsilon > 0$ tel que si $\|V - V^*\|_\infty \leq \epsilon$, alors π est optimale.

Preuve. D'après les définitions des opérateurs $\mathcal{T}$ et $\mathcal{T}^\pi$ et leur propriété de contraction (et le fait que $\mathcal{T}V = \mathcal{T}^\pi V$ car π est déduite de V), on a

$$\begin{aligned}
\|V^* - V^\pi\|_\infty &\leq \|\mathcal{T}V^* - \mathcal{T}^\pi V\|_\infty + \|\mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi\|_\infty \\
&\leq \|\mathcal{T}V^* - \mathcal{T}V\|_\infty + \gamma\|V - V^\pi\|_\infty \\
&\leq \gamma\|V^* - V\|_\infty + \gamma(\|V - V^*\|_\infty + \|V^* - V^\pi\|_\infty) \\
&\leq \frac{2\gamma}{1-\gamma}\|V^* - V\|_\infty.
\end{aligned}$$

Soit $\delta = \min_\pi \|V^\pi - V^*\|_\infty$ où le min est pris sur toute politique non-optimale. Puisqu'il y a un nombre fini d'états et d'actions, il y a aussi un nombre fini de politiques, donc $\delta > 0$. Pour ϵ assez petit, i.e.

$$\frac{2\gamma}{1-\gamma}\epsilon < \delta,$$

si $\|V - V^*\|_\infty \leq \epsilon$, on a $\|V^* - V^\pi\|_\infty < \delta$ donc π est optimale.

□

2 Minimisation du résidu de Bellman

Nous avons vu que la fonction valeur optimale V^* est l'unique solution de l'équation de PD: $\mathcal{T}V = V$. Cherchons une approximation de V^* dans un espace de fonctions donné $\mathcal{F}$.

Soit $\mathcal{F}$ un espace de fonctions et $\|\cdot\|$ une norme, on peut s'intéresser à déterminer la fonction $V \in \mathcal{F}$ qui minimise la norme du résidu de Bellman:

$$\inf_{V\in\mathcal{F}} \|\mathcal{T}V - V\|.$$

2.1 Erreur d'approximation et borne sur la performance

Considérons la norme L_∞. On peut majorer l'erreur d'approximation de la fonction valeur optimale $\|V^* - V\|_\infty$ et la perte en performance $\|V^* - V^\pi\|_\infty$ (où π est déduite de V) en fonction du résidu de Bellman $\|\mathcal{T}V - V\|_\infty$. De plus le résidu de Bellman minimum $\inf_{V\in\mathcal{F}} \|\mathcal{T}V - V\|_\infty$ est majoré par la distance de V^* à $\mathcal{F}$.

Proposition 2. *[Williams et Baird, 1993]* Soit V une fonction.

1. On a
$$\|V^* - V\|_\infty \leq \frac{1}{1-\gamma}\|\mathcal{T}V - V\|_\infty. \tag{1}$$

2. Soit π la politique gloutonne par rapport à V. Alors
$$\|V^* - V^\pi\|_\infty \leq \frac{2}{1-\gamma}\|\mathcal{T}V - V\|_\infty.$$

3. Supposons qu'il existe un minimiseur du résidu de Bellman: $V_{BR} = \arg\min_{V\in\mathcal{F}} \|\mathcal{T}V - V\|_\infty$. Alors
$$\|\mathcal{T}V_{BR} - V_{BR}\|_\infty \leq (1+\gamma) \inf_{V\in\mathcal{F}} \|V^* - V\|_\infty. \tag{2}$$

Ainsi, en combinant 2 et 3, et en notant π_{BR} une politique déduite de V_{BR}, il vient:

$$\|V^* - V^{\pi_{BR}}\|_\infty \leq \frac{2(1+\gamma)}{1-\gamma} \inf_{V\in\mathcal{F}} \|V^* - V\|_\infty.$$

Preuve.

Point 1: On a

$$\begin{aligned}
\|V^* - V\|_\infty &\leq \|V^* - \mathcal{T}V\|_\infty + \|\mathcal{T}V - V\|_\infty \\
&\leq \gamma\|V^* - V\|_\infty + \|\mathcal{T}V - V\|_\infty \\
&\leq \frac{1}{1-\gamma}\|\mathcal{T}V - V\|_\infty
\end{aligned}$$

Point 2: On a $\|V^* - V^\pi\|_\infty \leq \|V^* - V\|_\infty + \|V - V^\pi\|_\infty$. Puisque $\mathcal{T}V = \mathcal{T}^\pi V$, on a

$$\begin{aligned}
\|V - V^\pi\|_\infty &\leq \|V - \mathcal{T}V\|_\infty + \|\mathcal{T}V - V^\pi\|_\infty \\
&\leq \|\mathcal{T}V - V\|_\infty + \gamma\|V - V^\pi\|_\infty \\
&\leq \frac{1}{1-\gamma}\|\mathcal{T}V - V\|_\infty,
\end{aligned}$$

donc, en utilisant le point 1, il vient

$$\|V^* - V^\pi\|_\infty \leq \frac{2}{1-\gamma}\|\mathcal{T}V - V\|_\infty.$$

Point 3: On a

$$\begin{aligned}\|\mathcal{T}V - V\|_\infty &\leq \|\mathcal{T}V - V^*\|_\infty + \|V^* - V\|_\infty \\ &\leq (1+\gamma)\|V^* - V\|_\infty.\end{aligned}$$

Ainsi, le minimiseur du résidu de Bellman satisfait:

$$\begin{aligned}\|T^\pi V_{BR} - V_{BR}\|_\infty &= \inf_{V\in\mathcal{F}} \|T^\pi V - V\|_\infty \\ &\leq (1+\gamma) \inf_{V\in\mathcal{F}} \|V^* - V\|_\infty\end{aligned}$$

□

2.2 Implémentation numérique

Soit $\mathcal{F}$ un ensemble de fonctions f_α paramétrées par un paramètre α. Problèmes:

- Il est difficile de minimiser la norme L_∞ du résidu (car celà signifie minimiser le résidu en tous états)
- En choisissant une norme L_2 avec poids μ (distribution sur $\mathcal{X}$), la fonction $\alpha \mapsto g(\alpha) = \|\mathcal{T}V_\alpha - V_\alpha\|_\mu^2$ n'est pas une fonction convexe.

On peut cependant utiliser une méthode de descente de gradient, qui converge vers un minimum local:

$$\alpha \leftarrow \alpha - \eta \nabla g(\alpha)$$

On peut calculer un estimateur du gradient $\nabla g(\alpha)$ de la manière suivante:

1. On tire n états $X_i \sim \mu$,
2. On définit le résidu empirique

   $$\hat{g}(\alpha) = \frac{1}{n}\sum_{i=1}^n \left[\mathcal{T}V_\alpha(X_i) - V_\alpha(X_i)\right]^2$$

   et on effectue un pas de descente selon le (sous) gradient:

   $$\nabla_\alpha \hat{g}(\alpha) = \frac{2}{n}\sum_{i=1}^n [\mathcal{T}V_\alpha - V_\alpha](X_i)(\gamma P^{\pi_\alpha} - I)\nabla V_\alpha(X_i),$$

   où π_α est une politique déduite de V_α.

Problème: dans une approche apprentissage par renforcement où la fonction récompense et les probabilités de transition ne sont pas connues, il n'est pas facile de calculer $\mathcal{T}V_\alpha(X_i)$ ou $P^{\pi_\alpha}V_\alpha(X_i)$.

3 Itération sur les valeurs avec approximation

La fonction valeur optimale V^* est l'unique solution de l'équation de programmation dynamique $V = \mathcal{T}V$, et elle peut être calculée par itération sur les valeurs:

$$V_{k+1} = \mathcal{T}V_k.$$

avec V_0 quelconque. Alors $V_k \to V^*$ (car $\|V^* - V_{k+1}\|_\infty \leq \gamma \|V^* - V_k\|_\infty$).

Lorsque le nombre d'états N est grand ou quand l'espace est continu, on peut définir l'algorithme d'**itération sur les valeurs avec approximation** (IVA):

$$V_{k+1} = \mathcal{A}\mathcal{T}V_k,$$

où $\mathcal{T}$ est l'opérateur de Bellman et $\mathcal{A}$ un **opérateur d'approximation**. Par exemple si $\mathcal{F}$ est un espace de fonctions représentables, et soit $\|\cdot\|$ une norme, alors $\mathcal{A}$ peut être la projection sur $\mathcal{F}$ de $\mathcal{T}V_k$, i.e.,

$$V_{k+1} = \inf_{V \in \mathcal{F}} \|\mathcal{T}V_k - V\|. \tag{3}$$

Remarque: En notant $\Pi_\infty g = \arg\min_{f \in \mathcal{F}} \|f - g\|_\infty$ la projection sur $\mathcal{F}$ selon la norme L_∞, alors (3) s'écrit $V_{k+1} = \Pi_\infty \mathcal{T}V_k$. L'opérateur $\mathcal{T}$ est une contraction en norme L_∞ et l'opérateur Π_∞ est une non-expansion en L_∞. Donc l'opérateur composé $\Pi_\infty \mathcal{T}$ est une contraction, et il existe un unique point fixe $\tilde{V} \in \mathcal{F}$ de $\Pi_\infty \mathcal{T}$. Ainsi on a $\|V_{k+1} - \tilde{V}\|_\infty \leq \gamma \|V_k - \tilde{V}\|_\infty$ donc $V_k \to \tilde{V}$.

Maintenant, si on considère une autre norme, par exemple une norme $L_2(\mu)$, alors $\Pi_{2,\mu}\mathcal{T}$ n'est pas une contraction et l'algorithme IVA ne converge pas nécessairement.

3.1 Résultat de majoration d'erreur pour IVA

Proposition 3. *[Bertsekas & Tsitsiklis, 1996]* La perte en performance $\|V^* - V^{\pi_k}\|_\infty$ résultante de l'utilisation de la politique π_K (déduite de l'approximation V_K) au lieu de la politique optimale est majorée selon:

$$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < K} \|\mathcal{T}V_k - \mathcal{A}\mathcal{T}V_k\|_\infty + \frac{2\gamma^{K+1}}{1-\gamma} \|V^* - V_0\|_\infty.$$

Preuve. Notons $\varepsilon = \max_{0 \leq k < K} \|\mathcal{T}V_k - \mathcal{A}\mathcal{T}V_k\|_\infty$. Pour tout $0 \leq k < K$, nous avons

$$\begin{aligned} \|V^* - V_{k+1}\|_\infty &\leq \|\mathcal{T}V^* - \mathcal{T}V_k\|_\infty + \|\mathcal{T}V_k - V_{k+1}\|_\infty \\ &\leq \gamma \|V^* - V_k\|_\infty + \varepsilon, \end{aligned}$$

donc

$$\begin{aligned} \|V^* - V_K\|_\infty &\leq (1 + \gamma + \cdots + \gamma^{K-1})\varepsilon + \gamma^K \|V^* - V_0\|_\infty \\ &\leq \frac{1}{1-\gamma}\varepsilon + \gamma^K \|V^* - V_0\|_\infty \end{aligned}$$

Mais d'après la proposition 1, i.e. $\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{1-\gamma} \|V^* - V_K\|_\infty$, on a

$$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\varepsilon + \frac{2\gamma^{K+1}}{1-\gamma} \|V^* - V_0\|_\infty.$$

□

3.2 Implémentation avec des fonctions Q-valeur

Modèle génératif: L'étape 2 de l'algo précédent ($Z_i = \mathcal{T}V_k(X_i)$) nécessite le calcul d'une espérance, ainsi que la connaissance des fonctions récompense et probabilités de transition (non supposées connues dans un cadre apprentissage par renforcement).

Supposons qu'on dispose d'un modèle génératif:

Rappel Q-valeurs: la fonction $Q-$valeur optimale définie selon

$$Q^*(x,a) = \sum_y p(y|x,a)\big[r(x,a,y) + \gamma V^*(y)\big].$$

est le point fixe de l'opérateur de Bellman $\mathcal{T}$ pour des fonctions Q définies sur $X \times A$:

$$\mathcal{T}Q(x,a) = \sum_y p(y|x,a)[r(x,a,y) + \gamma \max_b Q(y,b)].$$

Algorithme d'itérations sur les Q-valeurs avec approximation (fitted Q-iteration):

$$Q_{k+1} = \mathcal{A}\mathcal{T}Q_k.$$

On implémente cette itération à l'aide d'un algorithme de régression statistique dont les données sont obtenues par appel au modèle génératif. Ainsi un algorithme d'IVA est implémenté par une séquence de problèmes de régression. Deux exemples sont décrit ci-dessous.

Approximation linéaire: Soit $\mathcal{F}$ l'espace vectoriel de fonctions définies sur $\mathcal{X} \times A$ engendré par les fonctions de base $\phi_1, \ldots, \phi_d : \mathcal{X} \times A \to \mathbb{R}$:

$$\mathcal{F} \stackrel{\text{def}}{=} \{Q_\alpha(x,a) = \sum_{j=1}^d \alpha_j \phi_j(x,a), \alpha \in \mathbb{R}^d\}.$$

Soit μ une distribution sur $\mathcal{X}$. On considère une projection en norme $L_2(\mu)$. L'algorithme suivant implémente une approximation empirique de l'itération suivante:

$$Q_{k+1} = \arg\min_{Q \in \mathcal{F}} \|Q - \mathcal{T}Q_k\|_\mu^2.$$

On construit une séquence de fonctions $Q_k : \mathcal{X} \times A \to \mathbb{R}$. A l'étape k:

1. On sample n états-actions (X_i, A_i) où $X_i \sim \mu$ et A_i choisi uniformément aléatoirement, et on appelle le modèle génératif pour générer des transitions $(X_i, A_i) \longrightarrow (R_i, Y_i)$, avec $Y_i \sim p(\cdot|X_i, A_i)$ et $R_i = r(X_i, A_i, Y_i)$,
2. On calcule les valeurs $Z_i \stackrel{\text{def}}{=} R_i + \gamma \max_{a \in A} Q_k(Y_i, a)$ (qui vérifient $\mathbb{E}[Z_i|X_i, A_i] = \mathcal{T}Q_k(X_i, A_i)$).
3. On calcule Q_{k+1}, solution de

$$Q_{k+1} = \arg\min_{Q_\alpha \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \big[Q_\alpha(X_i, A_i) - Z_i\big]^2. \tag{4}$$

(il s'agit d'un problème de minimisation d'une fonction quadratique dont la solution est obtenue en résolvant un système linéaire de taille d.

Ainsi l'étape (4) est une régression de type moindres-carrés à partir des données constituées des entrées (X_i, A_i) et des sorties Z_i.

Il s'agit d'une version empirique du problème de minimisation $Q_{k+1} = \arg\min_{Q \in \mathcal{F}} \|Q - \mathcal{T}Q_k\|_{\mu \times u}^2$, où $\mu \times u$ est une distribution produit sur $\mathcal{X} \times A$, où u est une distribution unforme sur A.

k-plus proches voisins: Avant de commencer l'algorithme, on génère les échantillons:

1. On sample n états $X_i \sim \mu$,
2. Pour chaque action a, on fait appel au modèle génératif pour générer un état suivant $Y_{i,a}$ et une récompense $R_{i,a}$ à partir de (X_i, a)

La donnée de $n \times |A|$ valeurs $Q(X_i, a)$ aux points X_i pour toute action a permet de définir une fonction en tout (x, a) à partir des k-plus proches voisins de x:

$$Q(x,a) = \frac{1}{k} \sum_{i=1}^{k} Q(X_{i(x)}, a), \tag{5}$$

où $i(x)$ est l'indice du $i$ème état de la grille $\{X_i, 1 \leq i \leq n\}$ le plus proche de x. On en déduit un algorithme qui itère des Q-fonctions de la manière suivante. A partir de la fonction Q-valeur Q_k on définit $Q_{k+1}(X_i, a)$ aux points X_i pour toute action a, selon

$$Q_{k+1}(X_i, a) = R_{i,a} + \gamma \max_{b \in A} Q_k(Y_{i,a}, b),$$

(ce qui définit une Q-fonction sur $\mathcal{X} \times A$ grâce à (5)).

Remarque: le nombre de voisins k doit être bien choisi... (k trop petit -> overfitting, k trop grand -> biais!).

Remarque: On peut utiliser un noyau $k(\cdot, \cdot)$ pour régulariser l'approximation des fonctions f, selon

$$Q(x,a) = \sum_{i=1}^{n} \frac{k(x, x_i)}{\sum_{j=1}^{n} k(x, x_j)} Q(X_i, a).$$

Autres méthodes de régression: approximation linéaire avec pénalisation (ℓ_2, ℓ_1), régression linéaire locale, approximation non-linéaire (ex: à partir d'ondelettes), réseaux de neurones, régression avec Support Vector Machines, méthodes à noyaux (RKHS), etc.

Remarque: L'analyse de ces algorithmes sera faite plus tard dans le chapitre "Sample complexity en apprentissage par renforcement".

3.3 Illustration: problème de remplacement optimal

Etat : usure d'un bien courant (ex. automobile). **Décisions**:

- **Remplacer**: coût de remplacement: C. Nouvel état $y \sim \exp(\beta)$ (densité $d(y) = \beta e^{-\beta y} \mathbf{1}_{y \geq 0}$),
- **Conserver**: coût d'entretien (régulier + ponctuel): $c(x)$. Nouvel état $y \sim x + \exp(\beta)$ (densité $d(y-x)$).

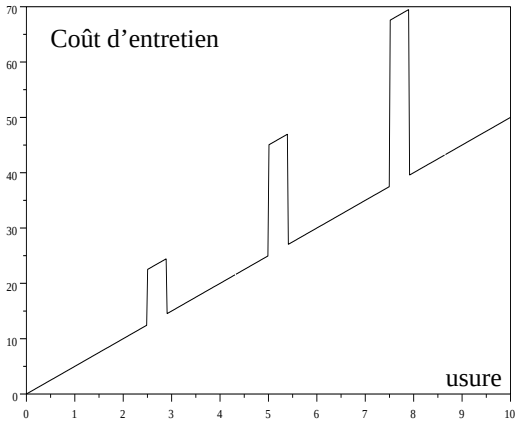
Problème: Minimiser l'espérance de coût (actualisé) sur le long-terme.

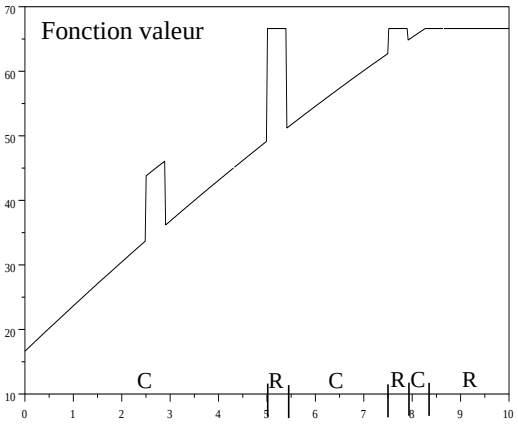
La fonction valeur optimale satisfait l'équation de PD:

$$V^*(x) = \min \left\{ c(x) + \gamma \int_0^\infty d(y-x) V^*(y) dy,\ C + \gamma \int_0^\infty d(y) V^*(y) dy \right\}$$

et la commande optimale est l'argument du min.

Coût d'entretien et fonction valeur:

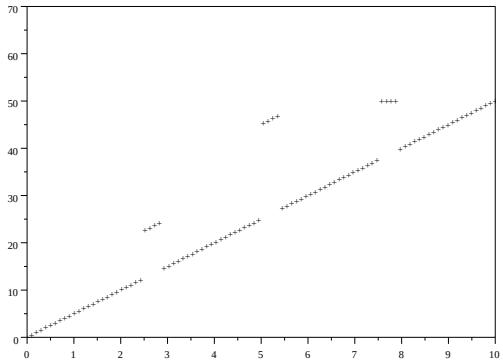




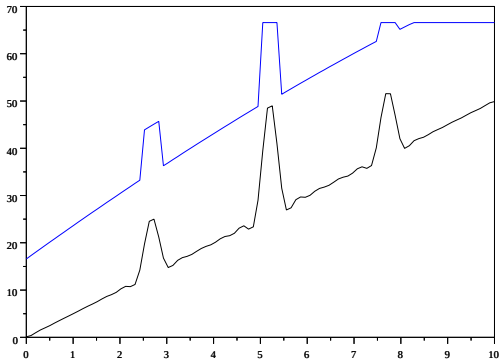
Ici, $\gamma = 0.6$, $\beta = 0.6$, $C = 50$. Coût d'entretien = fonction régulièrement croissante + coût ponctuels réguliers.

Approximation linéaire: Espace de fonctions $\mathcal{F} := \left\{V_n(x) = \sum_{k=1}^{20} \alpha_k \cos(k\pi \frac{x}{x_{\max}})\right\}$. Discrétisation sur une grille uniforme de N points.

Première itération: $V_0 = 0$,

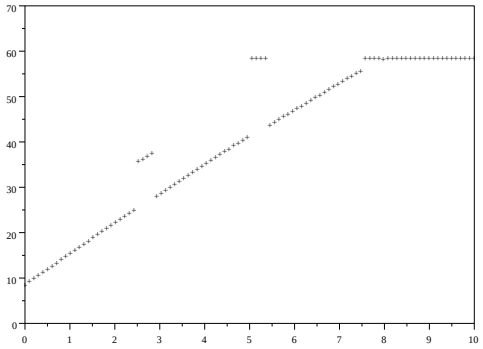


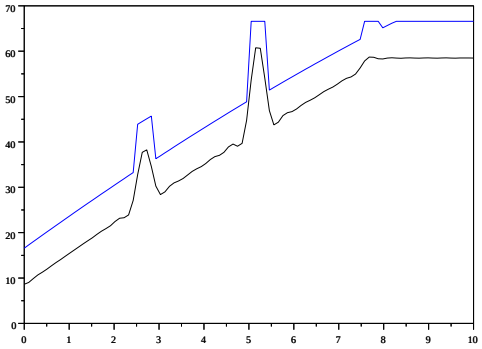
Valeurs itérées $\{\mathcal{T}V_0(x_n)\}_{1\leq n\leq N}$

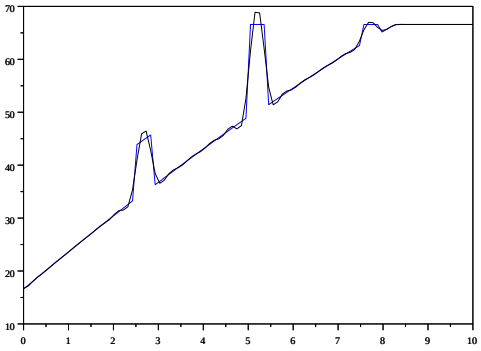


Approximation $V_1 \in \mathcal{F}$ de $\mathcal{T}V_0$

Itérations suivantes





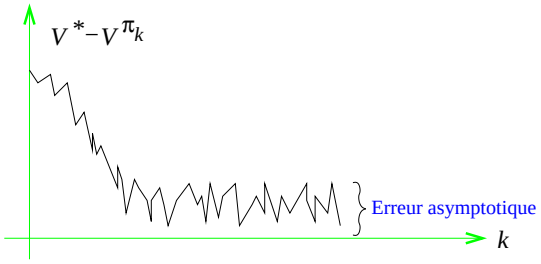


4 Itération sur les politiques avec approximation

On choisit une politique initiale π_0, et on itère les deux étapes:

1. **Evaluation de la politique π_k:** On calcule une approximation V_k de la fonction valeur V^{π_k}.
2. **Amélioration de la politique**: π_{k+1} est déduite de V_k selon:

$$\pi_{k+1}(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_{y \in \mathcal{X}} p(y|x,a) V_k(y) \right].$$



Cet algorithme ne converge pas nécessairement, mais on peut s'intéresser à la performance asymptotique des politiques déduites.

4.1 Majoration d'erreur pour l'algorithme IPA

Nous considérons le cas d'un critère à horizon temporel infini, avec actualisation.

Si les erreurs d'approximation $\|V_k - V^{\pi_k}\|$ sont petites, alors la performance asymptotique des politiques déduites est proche de l'optimum:

Proposition 4. La performance asymptotique des politiques π_k déduite de l'algorithme IPA est majorée en fonction de la qualité d'approximation $\|V_k - V^{\pi_k}\|$ des fonctions valeur:

$$\limsup_{k\to\infty} \|V^* - V^{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \|V_k - V^{\pi_k}\|_\infty$$

Preuve. Notons $e_k = V_k - V^{\pi_k}$ l'*erreur d'approximation*, $g_k = V^{\pi_{k+1}} - V^{\pi_k}$ le *gain en performance* entre les itérations k et $k+1$, et $l_k = V^* - V^{\pi_k}$ *la perte* en performance dûe à l'utilisation de la politique π_k au lieu de π^*.

La performance de la politique suivante ne peut pas être bien pire que celle de la politique courante:

$$g_k \geq -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})\, e_k \tag{6}$$

En effet, puisque $T^{\pi_{k+1}} V_k \geq T^{\pi_k} V_k$ (car π_{k+1} est déduite de V_k), on a:

$$\begin{aligned} g_k &= T^{\pi_{k+1}} V^{\pi_{k+1}} - T^{\pi_{k+1}} V^{\pi_k} + T^{\pi_{k+1}} V^{\pi_k} - T^{\pi_{k+1}} V_k \\ &\quad + T^{\pi_{k+1}} V_k - T^{\pi_k} V_k + T^{\pi_k} V_k - T^{\pi_k} V^{\pi_k} \\ &\geq \gamma P^{\pi_{k+1}} g_k - \gamma(P^{\pi_{k+1}} - P^{\pi_k})\, e_k \\ &\geq -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})\, e_k \end{aligned}$$

La perte à l'itération suivante est majorée par celle courante selon:

$$l_{k+1} \leq \gamma P^{\pi^*} l_k + \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k}) - P^{\pi^*}]e_k$$

En effet, puisque $T^{\pi^*} V_k \leq T^{\pi_{k+1}} V_k$,

$$\begin{aligned} l_{k+1} &= T^{\pi^*} V^* - T^{\pi^*} V^{\pi_k} + T^{\pi^*} V^{\pi_k} - T^{\pi^*} V_k \\ &\quad + T^{\pi^*} V_k - T^{\pi_{k+1}} V_k + T^{\pi_{k+1}} V_k - T^{\pi_{k+1}} V^{\pi_k} \\ &\quad + T^{\pi_{k+1}} V^{\pi_k} - T^{\pi_{k+1}} V^{\pi_{k+1}} \\ &\leq \gamma[P^{\pi^*} l_k - P^{\pi_{k+1}} g_k + (P^{\pi_{k+1}} - P^{\pi^*})e_k] \end{aligned}$$

et en utilisant (6),

$$\begin{aligned} l_{k+1} &\leq \gamma P^{\pi^*} l_k + \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k}) + P^{\pi_{k+1}} - P^{\pi^*}]e_k \\ &\leq \gamma P^{\pi^*} l_k + \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k}) - P^{\pi^*}]e_k. \end{aligned}$$

En notant $f_k = \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k}) - P^{\pi^*}]e_k$, on a

$$l_{k+1} \leq \gamma P^{\pi^*} l_k + f_k.$$

Donc, en passant à la limite supérieure,

$$\begin{aligned} (I - \gamma P^{\pi^*}) \limsup_{k\to\infty} l_k &\leq \limsup_{k\to\infty} f_k \\ \limsup_{k\to\infty} l_k &\leq (I - \gamma P^{\pi^*})^{-1} \limsup_{k\to\infty} f_k, \end{aligned}$$

car $I - \gamma P^{\pi^*}$ est inversible. En norme L_∞, il vient

$$\begin{aligned}\limsup_{k\to\infty} \|l_k\| &\leq \frac{\gamma}{1-\gamma} \limsup_{k\to\infty} \|P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1}(I+\gamma P^{\pi_k}) + P^{\pi^*}\| \, \|e_k\| \\ &\leq \frac{\gamma}{1-\gamma}(\frac{1+\gamma}{1-\gamma}+1) \limsup_{k\to\infty} \|e_k\| = \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \|e_k\|.\end{aligned}$$

□

4.2 Evaluation de la politique avec approximation linéaire

Nous approfondissons maintenant la première étape de l'algorithme IPA, c'est-à-dire l'évaluation approchée de la politique dans le cas où l'on considère un espace d'approximation linéaire. Ainsi, pour une politique π donnée, nous souhaitons déterminer une bonne approximation de la fonction de valeur V^π par une fonction de $\mathcal{F}$, espace vectoriel engendré par un ensemble de fonctions de base (appelées *features*) $\phi_1, \ldots, \phi_d : \mathcal{X} \to \mathbb{R}$:

$$\mathcal{F} \stackrel{\text{def}}{=} \{V_\alpha(x) = \sum_{i=1}^{d} \alpha_i \phi_i(x), \alpha \in \mathbb{R}^d\}.$$

Notre objectif est ainsi de déterminer un paramètre $\alpha \in \mathbb{R}^d$ tel que V_α soit « proche » de V^π. Commençons par étudier l'extension directe de l'algorithme TD(λ).

4.2.1 TD(λ)

L'algorithme TD(λ) est défini de la même manière qu'au chapitre 2 du cours. On utilise un vecteur trace $z \in \mathbb{R}^d$ de même dimension que le paramètre α, initialisé à zéro. A partir d'un état initial x_0, on génère une trajectoire $(x_0, x_1, x_2, \ldots)$ en suivant la politique π. A chaque instant t, on calcule la différence temporelle pour l'approximation V_α courante :

$$d_t \stackrel{\text{def}}{=} r(x_t, \pi(x_t)) + \gamma V_\alpha(x_{t+1}) - V_\alpha(x_t)$$

et l'on met à jour, à la fois le paramètre α et la trace z, selon:

$$\begin{aligned}\alpha_{t+1} &= \alpha_t + \eta_t d_t z_t, \\ z_{t+1} &= \lambda\gamma z_t + \phi(x_{t+1}),\end{aligned}$$

où η_t est un pas d'apprentissage et $\phi : \mathcal{X} \to \mathbb{R}^d$ la fonction ayant pour composantes les ϕ_i.

Cet algorithme construit une séquence d'approximations V_{α_t} qui converge, sous une hypothèse d'ergodicité de la chaîne de Markov et une hypothèse sur la décroissance des pas d'apprentisage η_t, vers une fonction dont l'erreur d'approximation (par rapport à V^π) est majorée en fonction de la meilleure approximation possible dans $\mathcal{F}$.

Proposition 5 (Tsitsiklis et Van Roy, 1996). Supposons que les pas η vérifient $\sum_{t\geq 0} \eta_t = \infty$ et $\sum_{t\geq 0} \eta_t^2 < \infty$, qu'il existe une distribution μ sur $\mathcal{X}$ telle que $\forall x, x' \in \mathcal{X}$, $\lim_{t\to\infty} P(x_t = x' | x_0 = x) = \mu(x')$ et que les vecteurs $(\phi_i)_{1\leq k\leq K}$ soient linéairement indépendants. Alors α_t converge. Notons α^* sa limite. On a alors:

$$\|V_{\alpha^*} - V^\pi\|_\mu \leq \frac{1-\lambda\gamma}{1-\gamma} \inf_\alpha \|V_\alpha - V^\pi\|_\mu, \tag{7}$$

où $\|\cdot\|_\mu$ désigne la norme L^2 pondérée par la distribution μ, c'est-à-dire $\|f\|_\mu \stackrel{\text{def}}{=} \left[\sum_{x\in\mathcal{X}} f(x)^2 \mu(x)\right]^{1/2}$.

Lorsque $\lambda = 1$, on retrouve le résultat que l'estimateur Monte-Carlo donne la meilleure approximation de V^π dans $\mathcal{F}$, c'est-à-dire la projection de V^π sur $\mathcal{F}$. Maintenant si $\lambda < 1$, la qualité d'approximation se détériore (introduction d'un biais), mais la variance de l'estimateur est plus faible, donc sa détermination à une précision donnée peut être plus rapide.

De par sa nature d'algorithme d'approximation stochastique, TD(λ) est très coûteux en terme de données, au sens où il nécessite l'observation d'un grand nombre de transitions $x_t, a_t \to x_{t+1}$ pour que le paramètre α converge. Il faut plusieurs observations des mêmes transitions pour que la valeur du paramètre se stabilise. Ce problème a motivé l'introduction de méthodes de type moindres carrés décrites maintenant, qui sont beaucoup plus économes en terme de transitions observées.

4.2.2 Least Squares Temporal Difference

Références: [BB96, Boy99, LP03, Sch03, Mun03]

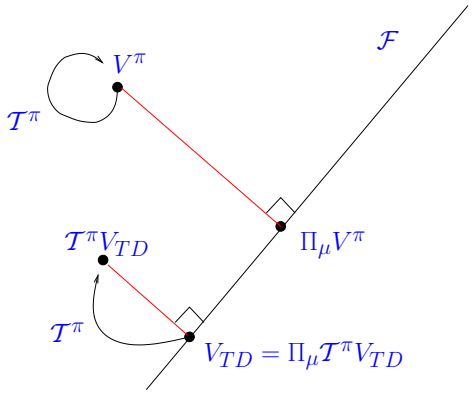
La fonction valeur V^π est le point fixe de $\mathcal{T}^\pi$. Comme V^π n'a pas de raison d'appartenir à l'espace d'approximation $\mathcal{F}$, on peut s'intéresser à chercher le point fixe de $\Pi\mathcal{T}^\pi$ où Π est un opérateur de projection sur $\mathcal{F}$.

Idéalement on souhaiterait considérer une projection Π_∞ en norme L_∞, afin que l'opérateur $\Pi_\infty\mathcal{T}^\pi$ soit une contraction en norme L_∞. Cependant la projection en norme L_∞ n'est pas numériquement envisageable lorsque le nombre d'états est grand. La plupart des outils d'approximation de fonction considèrent la minimisaton d'une norme L_2 (régression linéaire de type moindres-carrés, réseaux de neurones, etc.) ou L_1 (SVM, etc.).

Ici nous considérons une norme $L_2(\mu)$ où μ est une distribution sur $\mathcal{X}$, et notons Π_μ la projection correspondante:

$$\Pi_\mu g = \arg\min_{f\in\mathcal{F}} \|f - g\|_\mu.$$

Lorsque le point fixe de $\Pi_\mu T^\pi$ existe, on l'appelle la solution LSTD (Least Squares Temporal Difference), noté V_{TD}. Cela signifie que le résidu de Bellman $\mathcal{T}^\pi V_{TD} - V_{TD}$ est orthogonal à l'espace $\mathcal{F}$, i.e. $\langle \mathcal{T}^\pi V_{TD} - V_{TD}, \phi_i\rangle_\mu = 0$ pour tout $1 \le i \le d$, où le produit scalaire est $\langle f, g\rangle_\mu = \sum_{x\in\mathcal{X}} f(x)g(x)\mu(x)$.



Ainsi, lorsque V_{TD} existe, le paramètre α_{TD} correspondant satisfait pour tout $1 \leq i \leq d$:

$$\begin{aligned} \langle r^\pi + \gamma P^\pi V_{TD} - V_{TD}, \phi_i \rangle_\mu &= 0 \\ \langle r^\pi, \phi_i \rangle_\mu + \sum_{j=1}^{d} \langle \gamma P^\pi \phi_j - \phi_j, \phi_i \rangle_\mu \alpha_{TD,j} &= 0, \end{aligned}$$

on en déduit que α_{TD} est solution du système linéaire (de taille d):

$$A\alpha = b, \text{ avec} \left\{ \begin{array}{rcl} A_{i,j} &=& \langle \phi_i, \phi_j - \gamma P^\pi \phi_j \rangle_\mu \\ b_i &=& \langle \phi_i, r^\pi \rangle_\mu \end{array} \right. \tag{8}$$

Erreur d'approximation. En général on n'a pas de garantie qu'il existe un point fixe de $\Pi_\mu \mathcal{T}^\pi$, et même s'il existe une solution, on n'a pas de borne sur l'erreur d'approximation $\|V^\pi - V_{TD}\|$.

Cependant lorsque l'on considère la distribution stationnaire μ_π associée à la politique π (c'est-à-dire μ_π vérifie $\mu_\pi P^\pi = \mu_\pi$, i.e., $\mu_\pi(y) = \sum_x p(y|x, \pi(x))\mu_\pi(x)$ pour tout $y \in \mathcal{X}$), alors on a existence (et unicité) du point fixe de l'opérateur $\Pi_{\mu_\pi} \mathcal{T}^\pi$ et l'on déduit une majoration sur l'erreur d'approximation.

Proposition 6. Supposons que la politique π admette une distribution stationnaire μ_π. Alors l'operateur de Bellman est une contration en norme $L_2(\mu_\pi)$. Donc $\Pi_{\mu_\pi} \mathcal{T}^\pi$ est une contaction en $L_2(\mu_\pi)$ et il existe un unique point fixe V_{TD}. On a la borne suivante sur l'erreur d'approximation:

$$\|V^\pi - V_{TD}\|_{\mu_\pi} \leq \frac{1}{\sqrt{1-\gamma^2}} \inf_{V \in \mathcal{F}} \|V^\pi - V\|_{\mu_\pi}. \tag{9}$$

Preuve. Montrons d'abord que $\|P_\pi\|_{\mu_\pi} = 1$. En effet:

$$\begin{aligned} \|P^\pi V\|_{\mu_\pi}^2 &= \sum_x \mu_\pi(x) \big(\sum_y p(y|x, \pi(x)) V(y) \big)^2 \\ &\leq \sum_x \sum_y \mu_\pi(x) p(y|x, \pi(x)) V(y)^2 \\ &= \sum_y \mu_\pi(y) V(y)^2 = \|V\|_{\mu_\pi}^2. \end{aligned}$$

On en déduit que $\mathcal{T}^\pi$ est une contraction en norme $L_2(\mu_\pi)$:

$$\|\mathcal{T}^\pi V_1 - \mathcal{T}^\pi V_2\|_{\mu_\pi} = \gamma \|P^\pi (V_1 - V_2)\|_{\mu_\pi} \leq \gamma \|V_1 - V_2\|_{\mu_\pi},$$

et puisque Π_{μ_π} est une non-expansion en $L_2(\mu_\pi)$, l'opérateur $\Pi_{\mu_\pi} \mathcal{T}^\pi$ est une contraction en $L_2(\mu_\pi)$. Appelons V_{TD} son (unique) point fixe. On a

$$\|V^\pi - V_{TD}\|_{\mu_\pi}^2 = \|V^\pi - \Pi_{\mu_\pi} V^\pi\|_{\mu_\pi}^2 + \|\Pi_{\mu_\pi} V^\pi - V_{TD}\|_{\mu_\pi}^2,$$

mais

$$\|\Pi_{\mu_\pi} V^\pi - V_{TD}\|_{\mu_\pi}^2 = \|\Pi_{\mu_\pi} V^\pi - \Pi_{\mu_\pi} \mathcal{T}^\pi V_{TD}\|_{\mu_\pi}^2 \leq \|\mathcal{T}^\pi V^\pi - \mathcal{T} V_{TD}\|_{\mu_\pi}^2 \leq \gamma^2 \|V^\pi - V_{TD}\|_{\mu_\pi}^2.$$

Donc

$$\|V^\pi - V_{TD}\|_{\mu_\pi}^2 \leq \|V^\pi - \Pi_{\mu_\pi} V^\pi\|_{\mu_\pi}^2 + \gamma^2 \|V^\pi - V_{TD}\|_{\mu_\pi}^2,$$

d'où l'on déduit (9). □

Implémentation numérique On observe une unique trajectoire $(X_0, X_1, \dots)$ en suivant la politique π (i.e., $X_{t+1} \sim p(\cdot|X_t, \pi(X_t))$). Notons $R_t = r(X_t, \pi(X_t))$. Alors on construit un estimateur de la matrice A et du vecteur b (définis en (8)) selon

$$\begin{aligned} \hat{A}_{ij} &= \frac{1}{n}\sum_{t=1}^{n} \phi_i(X_t)[\phi_j(X_t) - \gamma\phi_j(X_{t+1})], \\ \hat{b}_i &= \frac{1}{n}\sum_{t=1}^{n} \phi_i(X_t)R_t. \end{aligned}$$

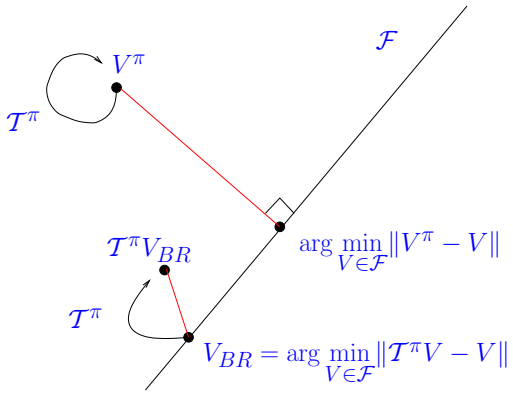
et on résout $\hat{A}\alpha = \hat{b}$. Lorsque la chaîne de Markov est ergodique alors la distribution empirique des états (X_t) tend vers la distribution stationnaire, donc $\hat{A} \to A$ et $\hat{b} \to b$ quand $n \to \infty$. Donc la solution du système empirique tend vers la solution α_{TD} quand le nombre de transitions observées $n \to \infty$.

4.2.3 Bellman Residual Minimization (BRM)

Une autre approche consiste à déterminer la fonction $V_{BR} \in \mathcal{F}$ qui minimise le résidu de Bellman pour la politique π:

$$V_{BR} = \arg\min_{V\in\mathcal{F}} \|T^\pi V - V\|, \tag{10}$$

pour une certaine norme $\|\cdot\|$.



On a le résultat suivant.

Proposition 7. On a

$$\|V^\pi - V_{BR}\| \le \|(I - \gamma P^\pi)^{-1}\|(1 + \gamma\|P^\pi\|) \inf_{V\in\mathcal{F}} \|V^\pi - V\|. \tag{11}$$

De plus si μ_π est la distribution stationnaire associée à la politique π, alors $\|P^\pi\|_{\mu_\pi} = 1$ et $\|(I-\gamma P^\pi)^{-1}\|_{\mu_\pi} = \frac{1}{1-\gamma}$, donc

$$\|V^\pi - V_{BR}\|_{\mu_\pi} \le \frac{1+\gamma}{1-\gamma} \inf_{V\in\mathcal{F}} \|V^\pi - V\|_{\mu_\pi}.$$

Preuve. Pour tout fonction V, on a

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

donc

$$\|V^\pi - V_{BR}\| \leq \|(I - \gamma P^\pi)^{-1}\| \|\mathcal{T}^\pi V_{BR} - V_{BR}\|$$

et

$$\|\mathcal{T}^\pi V_{BR} - V_{BR}\| = \inf_{V \in \mathcal{F}} \|\mathcal{T}^\pi V - V\| \leq (1 + \gamma \|P^\pi\|) \inf_{V \in \mathcal{F}} \|V^\pi - V\|,$$

et (11) s'en déduit.

Lorsque l'on considère la ditribution stationnaire μ_π, on a déjà vu que $\|P^\pi\|_{\mu_\pi} = 1$, et l'on déduit que $\|(I - \gamma P^\pi)^{-1}\|_{\mu_\pi} \leq \sum_{t \geq 0} \gamma^t \|P^\pi\|^t_{\mu_\pi} \leq \frac{1}{1-\gamma}$.

□

Soit μ une distribution et définissons V_{BR} le minimum du résidu de Bellman (10) en norme $L_2(\mu)$. L'application $\alpha \to \mathcal{T}^\pi V_\alpha - V_\alpha$ est affine, donc la fonction $\alpha \to \|\mathcal{T}^\pi V_\alpha - V_\alpha\|^2_\mu$ est quadratique. Son point de minimum (obtenu en écrivant que le gradient de cette fonction est nul) est donc la solution du système linéaire

$$\langle r^\pi + (\gamma P^\pi - I) \sum_{j=1}^{d} \phi_j \alpha_j, (\gamma P^\pi - I)\phi_i \rangle_\mu = 0, \quad \text{pour tout } 1 \leq i \leq d,$$

qui peut s'écrire:

$$A\alpha = b,$$

avec la matrice carrée A et le vecteur b (de taille d) définis par:

$$\left\{ \begin{array}{rcll} A_{i,j} & = & \langle \phi_i - \gamma P^\pi \phi_i, \phi_j - \gamma P^\pi \phi_j \rangle_\mu, & \text{pour } 1 \leq i,j \leq d \\ b_i & = & \langle \phi_i - \gamma P^\pi \phi_i, r^\pi \rangle_\mu, & \text{pour } 1 \leq i \leq d \end{array} \right. \tag{12}$$

Ce système possède toujours une solution lorsque la famille des ϕ_i est linéairement indépendante (sous la distribution μ). Remarquons que ce problème peut être considéré comme un problème de régression linéaire avec un ensemble de fonctions de base $\{\psi_i \stackrel{\text{def}}{=} \phi_i - \gamma P^\pi \phi_i\}_{i=1\ldots d}$ où il s'agit de déterminer α qui minimise $\|\alpha \cdot \psi - r^\pi\|_\mu$. Les méthodes usuelles en apprentissage supervisé peuvent alors être utilisées.

Implémentation numérique. Nous considérons une distribution μ quelconque (pas nécessairement la distribution stationnaire). Nous supposons ici que nous disposons d'un modèle génératif qui permet, pour tout (x,a) de déterminer la récompense $r(x,a)$ et de générer des états successeurs $y \sim p(\cdot|x,a)$. Nous cherchons à formuler un estimateur empirique du résidu de Bellman

$$\mathcal{B}(V) = \|\mathcal{T}^\pi V - V\|^2_\mu.$$

Pour cela, nous tirons n états $X_t \sim \mu$ et appelons le modèle génératif en (X_t, A_t) (où $A_t = \pi(X_t)$) pour obtenir la récompense $R_t = r(X_t, A_t)$ et un état successeur $Y_t \sim p(\cdot|X_t, A_t)$, et nous formulons l'estimateur empirique selon

$$\hat{\mathcal{B}}(V) = \frac{1}{n} \sum_{t=1}^{n} \big[V(X_t) - \underbrace{\big(R_t + \gamma V(Y_t)\big)}_{\text{noté } \hat{\mathcal{T}} V(X_t)}\big]^2.$$

Problème: cet estimateur est biaisé. En effet,

$$\begin{array}{rcl} \mathbb{E}[\hat{\mathcal{B}}(V)] & = & \mathbb{E}\Big[\big[V(X_t) - \mathcal{T}^\pi V(X_t) + \mathcal{T}^\pi V(X_t) - \hat{\mathcal{T}} V(X_t)\big]^2\Big] \\ & = & \|\mathcal{T}^\pi V - V\|^2_\mu + \mathbb{E}\Big[\big[\mathcal{T}^\pi V(X_t) - \hat{\mathcal{T}} V(X_t)\big]^2\Big] \end{array}$$

Ainsi, minimiser $\hat{\mathcal{B}}(V)$ n'est pas équivalent à minimiser $\mathcal{B}(V)$, même si $n \to \infty$.

Pour remédier à ce problème, en chaque état X_t, on peut considérer deux échantillons indépendants Y_t et $Y'_t \sim p(\cdot|X_t, A_t)$, et formuler l'estimateur empirique

$$\hat{\mathcal{B}}(V) = \frac{1}{n}\sum_{t=1}^{n} \big[V(X_t) - \big(R_t + \gamma V(Y_t)\big)\big]\big[V(X_t) - \big(R_t + \gamma V(Y'_t)\big)\big].$$

Cet estimateur nécessite $2n$ appels au modèle génératif, mais fournit un estimateur non biaisé du résidu de Bellman: $\mathbb{E}\hat{\mathcal{B}}(V) = \mathcal{B}(V)$.

Ainsi, lorsque $n \to \infty$, le minimum de $\hat{\mathcal{B}}$ est aussi minimum de $\mathcal{B}$.

La fonction $\alpha \to \hat{\mathcal{B}}(V_\alpha)$ est aussi quadratique, donc le paramètre $\hat{\alpha}_{BR}$ du minimum de $\hat{\mathcal{B}}(V_\alpha)$ est solution du système linaire $A\alpha = b$ avec

$$\begin{aligned}
\widehat{A}_{i,j} &= \frac{1}{n}\sum_{t=1}^{n} \big[\phi_i(X_t) - \gamma\phi_i(Y_t)\big]\big[\phi_j(X_t) - \gamma\phi_j(Y'_t)\big], \\
\widehat{b}_i &= \frac{1}{n}\sum_{t=1}^{n} \big[\phi_i(X_t) - \gamma\frac{\phi_i(Y_t) + \phi_i(Y'_t)}{2}\big]\, R_t.
\end{aligned}$$

Les algorithmes présentés dans cette section sont efficaces en termes d'utilisation des données expérimentales, puisque les transitions $x, a \to y, r$ sont mémorisées et permettent de déterminer directement le paramètre α par la résolution d'un système linéaire.

4.2.4 Avantages et inconvénients de LSTD et BRM

- **Les hypothèses sont différentes:** BRM nécessite un modèle génératif alors que LSTD ne nécessite que l'observation d'une seule trajectoire.
- **Les performances sont évaluées différemment:** La performance de l'approximation obtenue $\|V^\pi - \hat{V}\|_\mu$ est évaluée selon la distribution d'échantillonnage considérée. BRM permet de choisir la distribution d'échantillonnage μ alors que LSTD ne considère que la distribution stationnaire μ_π. Donc V_{TD} peut être une mauvaise approximation de V^π aux endroits non visités par la politique π. Ce qui peut être dangereux lors de l'étape d'amélioration de la politique.

4.3 Etape d'amélioration de la politique

A la section précédente nous avons étudié l'étape d'évaluation d'une politique π_k avec approximation : notons V_k l'approximation de V^{π_k}. L'étape d'amélioration de la politique est définie selon

$$\pi_{k+1}(x) \in \arg\max_{a \in A}[r(x,a) + \gamma \sum_y p(y|x,a)V_k(y)].$$

Dans un cadre apprentissage par renforcement, pour éviter de devoir calculer les espérances dans l'équation précédente, nous pouvons à nouveau considérer des approximations des fonctions Q-valeurs et déduire des algorithmes très similaires.

Ainsi, soit $\mathcal{F}$ l'espace vectoriel de fonctions définies sur $\mathcal{X} \times A$ engendré par les fonctions de base $\phi_1, \ldots, \phi_d$: $\mathcal{X} \times A \to \mathbb{R}$:

$$\mathcal{F} \stackrel{\text{def}}{=} \{Q_\alpha(x,a) = \sum_{j=1}^{d} \alpha_j \phi_j(x,a), \alpha \in \mathbb{R}^d\}.$$

Algorithme LSTD: On observe une trajectoire $(X_0, X_1, \dots)$ en suivant la politique π_k (i.e., $X_{t+1} \sim p(\cdot|X_t, \pi_k(X_t))$). Notons $R_t = r(X_t, \pi_k(X_t)$. Nous construisons la matrice $\hat{A}$ et le vecteur $\hat{b}$ selon

$$\begin{aligned}\hat{A}_{ij} &= \frac{1}{n}\sum_{t=1}^{n}\phi_i(X_t, A_t)[\phi_j(X_t, A_t) - \gamma\phi_j(X_{t+1}, A_{t+1})],\\ \hat{b}_i &= \frac{1}{n}\sum_{t=1}^{n}\phi_i(X_t, A_t)R_t.\end{aligned}$$

et on résout $\hat{A}\alpha = \hat{b}$ pour en déduire le paramètre $\hat{\alpha}_{TD}$. L'étape d'amélioration de la politique s'écrit alors

$$\pi_{k+1}(x) \in \arg\max_{a\in A} Q_{\hat{\alpha}_{TD}}(x, a).$$

Algorithme BRM: on génère n états $X_t \sim \mu$, $A_t = \pi_k(X_t)$, $R_t = r(X_t, A_t)$ et on fait un double appel au modèle génératif pour générer deux échantillons indépendants Y_t et $Y'_t \sim p(\cdot|X_t, A_t)$. Notons $B_t = \pi_k(Y_t)$ et $B'_t = \pi_k(Y'_t)$. On construit la matrice $\hat{A}$ et le vecteur $\hat{b}$ selon

$$\begin{aligned}\widehat{A}_{i,j} &= \frac{1}{n}\sum_{t=1}^{n}\big[\phi_i(X_t, A_t) - \gamma\phi_i(Y_t, B_t)\big]\big[\phi_j(X_t, A_t) - \gamma\phi_j(Y'_t, B'_t)\big],\\ \widehat{b}_i &= \frac{1}{n}\sum_{t=1}^{n}\big[\phi_i(X_t, A_t) - \gamma\frac{\phi_i(Y_t, B_t) + \phi_i(Y'_t, B'_t)}{2}\big]R_t.\end{aligned}$$

On résout $\hat{A}\alpha = \hat{b}$ pour en déduire le paramètre $\hat{\alpha}_{BR}$ et l'étape d'amélioration de la politique s'écrit selon

$$\pi_{k+1}(x) \in \arg\max_{a\in A} Q_{\hat{\alpha}_{BR}}(x, a).$$

References

[BB96] S. Bradtke and A. Barto. Linear least-squares algorithms for temporal difference learning. *Journal of Machine Learning Research*, 22:33–57, 1996.

[Boy99] Justin Boyan. Least-squares temporal difference learning. In *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*, pages 49–56, 1999.

[LP03] Michail Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

[Mun03] Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 19th International Conference on Machine Learning (ICML'03)*, 2003.

[Sch03] R. Schoknecht. Optimality of reinforcement learning algorithms with linear function approximation. In *Advances in Neural Information Processing Systems 15 (NIPS'02)*, 2003.