Master MVA: Apprentissage par renforcement **Lecture: 1**

Introduction à l'apprentissage par renforcement

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

Page web La page web du cours est `http://researchers.lille.inria.fr/∼munos/master-mva/`
Voir la page web pour les horaires, détails du cours, annonces, références, description des mini-projets.

Le cours portera sur des aspects théoriques de l'apprentissage par renforcement (A/R), les algorithmes de bandit pour le compromis exploration-exploitation, et la programmation dynamique avec approximation (PDA), dans le cadre des processus de décision markoviens (PDM).

Plan

1. Introduction générale à l'A/R
2. Processus de décision markoviens et programmation dynamique
   (a) Itération sur les valeurs
   (b) Itération sur les politiques
   (c) Programmation linéaire
3. Algorithmes d'apprentissage par renforcement
   (a) Introduction à l'approximation stochastique
   (b) Algorithmes TD(λ) et Q-learning
4. Introduction aux algorithmes de bandit
   (a) Bandits stochastiques: UCB
   (b) Bandits contre un adversaire: Exp3
5. Programmation dynamique avec approximation
   (a) Analyse en norme L_∞
      - Itération sur les valeurs avec approximation
      - Itération sur les politiques avec approximation
      - Minimisation du résidu de Bellman
   (b) Analyse de quelques algorithmes: LSTD, Bellman residual, LSPI, Fitted Q-iterations
   (c) Extension à une analyse en norme L_p
6. Analyse en temps fini de l'A/R et la PDA
   (a) Outils statistiques
   (b) Analyse de LSTD, Bellman residual minimization

Références bibliographiques:

- *Neuro Dynamic Programming*, Bertsekas et Tsitsiklis, 1996.
- *Introduction to Reinforcement Learning*, Sutton and Barto, 1998.
- *Markov Decision Problems*, Puterman, 1994.
- *Processus Décisionnels de Markov en Intelligence Artificielle*, ouvrage collectif édité par Sigaud et Buffet, 2004. Voir `http://researchers.lille.inria.fr/~munos/papers/files/bouquinPDMIA.pdf`
- *Algorithms for Reinforcement Learning*, Szepesvári, 2009.
  Voir `http://researchers.lille.inria.fr/~munos/master-mva/docs/Csaba_book.pdf`

1 Introduction générale à l'A/R

Objectifs de l'A/R

- Acquisition automatisée de compétences pour la prise de décisions (**actions** ou **contrôle**) en milieu complexe et incertain.
- Apprendre par l'"expérience" une stratégie comportementale (appelée **politique**) en fonction des échecs ou succès constatés (les **renforcements** ou **récompenses**).
- **Exemples**: jeu du chaud-froid, apprentissage sensori-moteur, jeux (backgammon, échecs, poker, go), robotique mobile autonome, gestion de portefeuille, recherche opérationnelle, ...

Naissance du domaine : Rencontre fin années 1970 entre

- **Neurosciences computationnelles**. Renforcement des poids synaptiques des transmissions neuronales (règle de Hebbs, modèles de Rescorla et Wagner dans les années 60, 70).
  Renforcement = corrélations activités neuronales.
- **Psychologie expérimentale**. Modèles de conditionnement animal: renforcement de comportement menant à une satisfaction (recherches initiées vers 1900 par Pavlov, Skinner et le courant béhavioriste).
  Renforcement = satisfaction, plaisir ou inconfort, douleur.

Cadre mathématique adéquat: **Programmation dynamique** de Bellman (années 50, 60), en théorie du contrôle optimal. Renforcement = critère à maximiser.

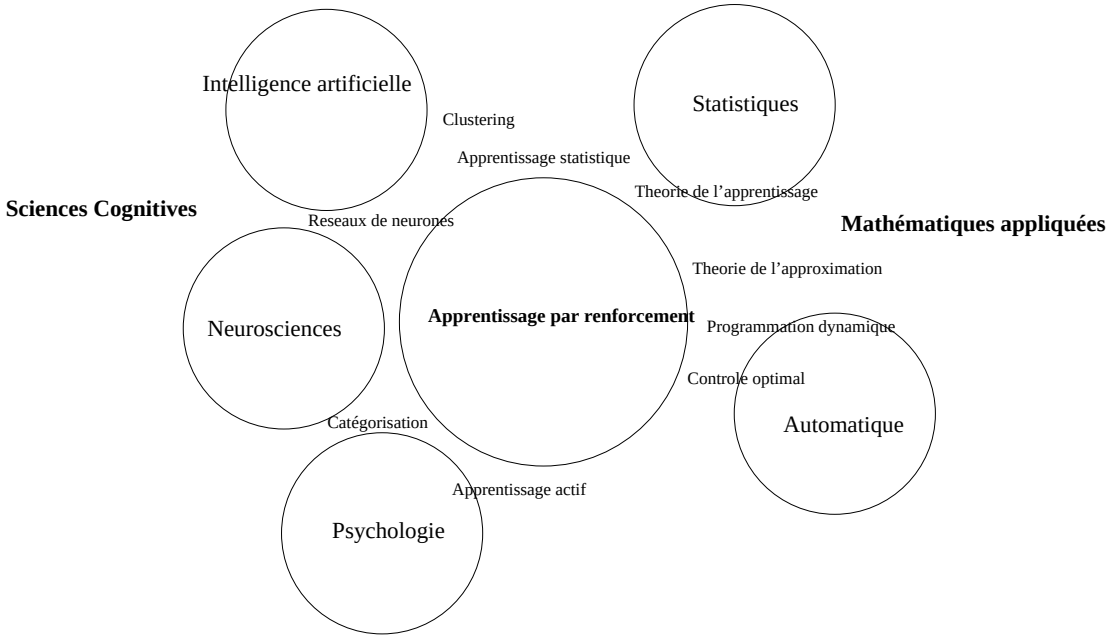
Liens avec la psychologie expérimentale Thorndike (1911) *Loi des effets*:

> "Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond"

Préhistoire de l'A/R computationnel

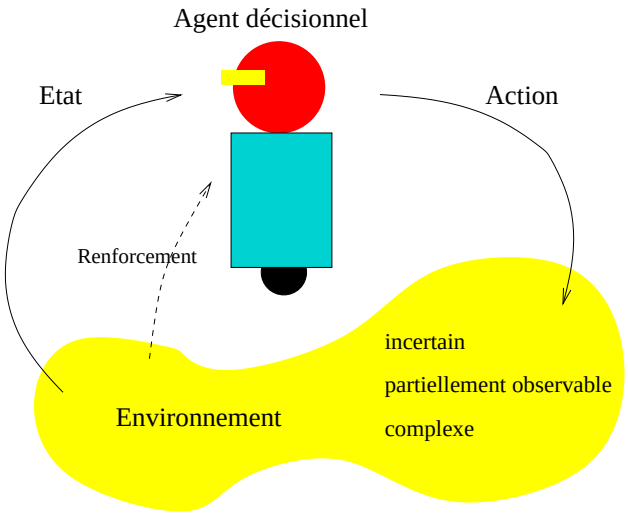
- Shannon 1950: Programming a computer for playing chess.
- Minsky 1954: Theory of Neural-Analog Reinforcement Systems.
- Samuel 1959: Studies in machine learning using the game of checkers.
- Michie 1961: Trial and error. -> joueur de tic-tac-toe.
- Michie et Chambers 1968: Adaptive control -> pendule inversé.
- Widrow, Gupta, Maitra 1973: Punish/reward: learning with a critic in adaptive threshold systems -> règles neuronales.
- Sutton 1978: Théories d'apprentissage animal : règles dirigées par des modifications dans prédictions temporelles successives.
- Barto, Sutton, Anderson 1983: règles neuronales Actor-Critic pour le pendule inversé.
- Sutton 1984: Temporal Credit Assignment in Reinforcement Learning.
- Klopf 1988: A neuronal model of classical conditioning.
- Watkins 1989: Q-learning.
- Tesauro 1992: TD-Gammon

Domaine pluridisciplinaire



Differents types d'apprentissage

- **Apprentissage supervisé:** à partir de l'observation de données $(X_i, Y_i)_i$ où $Y_i = f(X_i) + \epsilon_i$, où f est la fonction cible (inconnue), estimer f afin de faire des prédictions de $f(x)$
- **Apprentissage non-supervisé:** à partir de données $(X_i)_i$, trouver des structures dans ces données (ex. des classes), estimer des densités, ...
- **Apprentissage par renforcement**



Information disponible pour l'apprentissage: le renforcement. Les dynamiques sont aléatoires et partiellement inconnues. *Objectif:* maximiser la somme des récompenses reçues.

L'environnement

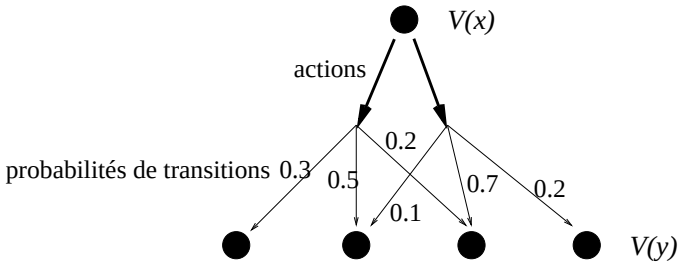
- Déterministe ou stochastique (ex: backgammon)
- Hostile (ex: jeu d'échecs) ou non (ex: jeu Tétris)
- Partiellement observable (ex: robotique mobile)
- Connu ou inconnu (ex: vélo) de l'agent décisionnel

Le renforcement:

- Peut récompenser une séquence d'actions –> problème du "credit-assignment" : quelles actions doivent être accréditées pour un renforcement obtenu au terme d'une séquence de décisions?
- Comment sacrifier petit gain à court terme pour privilégier meilleur gain à long terme?

Fonction valeur:

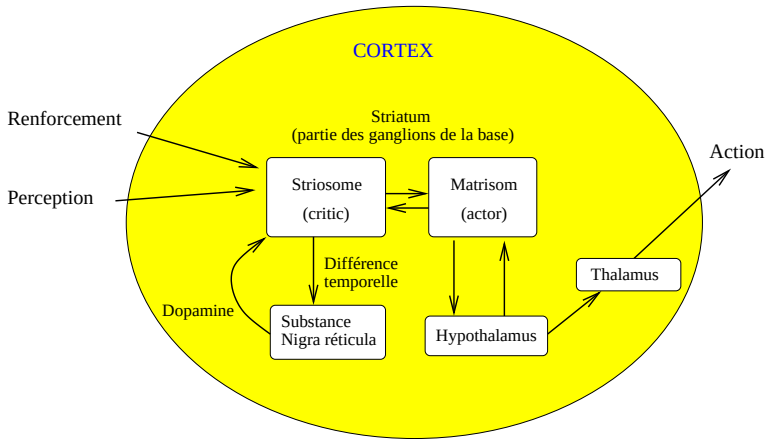
- Attribue une valeur à chaque état = ce que l'agent peut espérer de mieux en moyenne s'il est dans cet état.
- La valeur $V(x)$ en un état dépend de la récompense immédiate et de la valeur des états résultants $V(y)$



- V doit être telle que sa valeur en un état doit être la récompense immédiate plus la valeur moyenne de l'état suivant si je choisis l'action optimale: $V(x) = \max_a[r(x,a) + \mathbb{E}[V(Y)|x,a]$ (équation de Bellman).
- Si je connais V, "en moyenne, pas de surprise" !
- Comment apprendre la fonction valeur? Grâce à la surprise (différence temporelle).
- Si V est connue, cela permet de choisir, à chaque instant, la meilleure action $\arg\max_a[r(x,a) + \mathbb{E}[V(Y)|x,a]$. Donc maximiser localement la fonction valeur revient à maximiser le renforcement à long terme.

Lien avec les neurosciences:

- *Théorie des émotions.* Lien entre juste appréciation des émotions en fonction de la situation vécue et capacités de prises de décisions adéquates [Damasio, L'erreur de Descartes, la raison des émotions, 2001].
- Neurotransmetteurs du renforcement: dopamine –> surprise.
- Modèle des ganglions de la base (inspiré de [Doya, 1999])



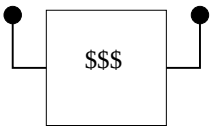
On en dira pas plus dans ce cours...

Quelques problématiques de l'A/R

- A/R = résoudre de manière adaptative un problème de contrôle optimal lorsque les dynamique d'état ou les récompenses sont partiellement inconnues. Deux approches possibles:
  - **A/R indirect**: apprentissage préalable d'un modèle des dynamiques (forme d'apprentissage supervisé), puis utilisation du modèle pour faire de la planification
  - **A/R direct**: 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.

Dilemme Exploration / Exploitation:

- **Exploiter** (agir en maximisant) la connaissance actuelle, ou **explorer** (améliorer notre connaissance).
- Exemple simple: **Le bandit à 2 bras**



  - A chaque instant t, le joueur choisit un bras k, reçoit récompense $r_t \sim \nu_k$, où les lois ν_k (une loi pour chaque bras) sont inconnues.
  - Objectif: maximiser $\sum_t r_t$.
  - Ex: récompenses déjà reçues: $6\$, 7\$, 5\$, 4\$$ pour le bras gauche, $5\$, 0\$$ pour le bras droit. Quel bras choisir?
  - Propriété : Il ne faut jamais s'arrêter d'explorer, mais il faut explorer de moins en moins fréquemment ($\log n/n$).
  - Différentes stratégies: ϵ-greedy, Upper-Confidence-Bounds, règles bayésiennes, échantillonnage de Gibbs, indices de Gittings, ...
- A/R = bandit avec dynamique sur l'état.

Quelques réalisations

- TD-Gammon. [Tesauro 1992-1995]: jeu de backgammon. Produit le meilleur joueur mondial!
- KnightCap [Baxter et al. 1998]: jeu d'échec ($\simeq$2500 ELO)
- Robotique: jongleurs, balanciers, acrobots, ... [Schaal et Atkeson, 1994]
- Robotique mobile, navigation: robot guide au musée Smithonian [Thrun et al., 1999], ...
- Commande d'une batterie d'ascenseurs [Crites et Barto, 1996],

- Routage de paquets [Boyan et Littman, 1993],
- Ordonnancement de tâches [Zhang et Dietterich, 1995],
- Maintenance de machines [Mahadevan et al., 1997],
- Computer poker (calcul d'un équilibre de Nash avec bandits adversarials), [Alberta, 2008]
- Computer go (algorithmes de bandits hiérarchiques), [Mogo, 2006]

2 Processus de décision markoviens et programmation dynamique

Référence bibliographique: Livre "Processus Décisionnels de Markov en Intelligence Artificielle", chapitre 1.

2.1 Définitions

Chaîne de Markov

- Système dynamique à temps discret $(x_t)_{t\in\mathbb{N}} \in X$, où X est l'espace d'états (supposé fini ici) tel que

$$\mathbb{P}(x_{t+1} = x \,|\, x_t, x_{t-1}, \ldots, x_0) = \mathbb{P}(x_{t+1} = x \,|\, x_t)$$

  toute l'information pertinente pour la prédiction du futur est contenue dans l'état présent (**propriété de Markov**)

- Ainsi une chaîne de Markov sur X est définie par un état initial x_0 et les *probabilités de transition*: $p(y|x) = \mathbb{P}(x_{t+1} = y|x_t = x)$.

Exemple: "état" x_t pour le vélo = toutes les variables pertinentes pour la prédiction de l'état suivant (position, vitesse, angles, vitesses angulaires...).

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

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

$$p(y|x, a) = \mathbb{P}(x_{t+1} = y|x_t = x, a_t = a),$$

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

Politique

- **Règle de décision**: π_t détermine, à l'instant t, une loi d'action en tout état:
  - déterministe: $\pi_t : X \to A$, $\pi_t(x)$= action choisie en x.
  - stochastique: $\pi_t : X \times A \to \mathbb{R}$, $\pi_t(a|x)$= probabilité de choisir a en x.
- **Politique** (ou **stratégie**, ou **plan**): séquence de règles de décision $\pi = (\pi_0, \pi_1, \pi_2, \dots)$. 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 $(x_t)_{t\geq 0}$ où l'action $a_t = \pi(x_t)$ est choisie selon la politique π est une chaîne de Markov (définie par les probabilités de transition $p(y|x) = p(y|x, \pi(x))$).

Critère à optimiser (ou **mesure de performance** ou **fonction valeur**)

- Horizon temporel fini

$$V^\pi(t,x) = \mathbb{E}\big[\sum_{s=t}^{T-1} r(x_s, \pi_s(x_s)) + R(x_T) | \, x_t = x; \pi\big],$$

  où R est une fonction récompense terminale.

- Horizon temporel infini avec critère actualisé

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

  où $0 \leq \gamma < 1$ est un coefficient d'actualisation.

- Horizon temporel inifi avec critère non-actualisé:

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

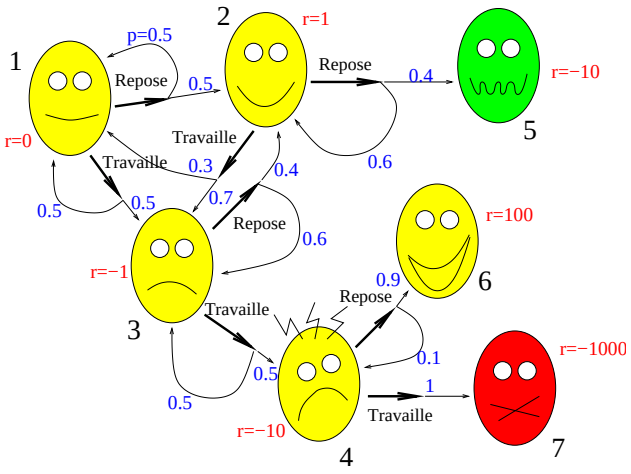
  où T est le premier instant (aléatoire) où l'on atteint un état absorbant.

- Horizon temporel infini avec critère moyen

$$V^\pi(x) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\big[\sum_{t=0}^{T-1} r(x_t, \pi(x_t)) \, | \, x_0 = x; \pi\big].$$

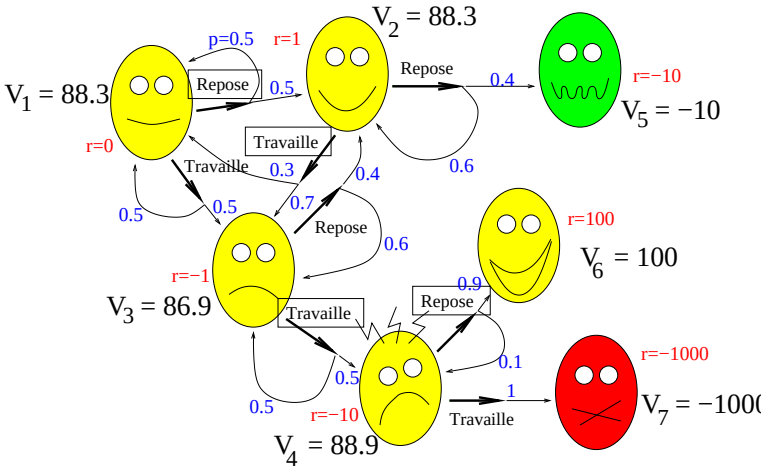
2.2 Exemples

Le dilemme de l'étudiant MVA



- Modélisation: les états 5, 6, et 7 sont des "états terminaux".
- Objectif: maximiser la somme des récompenses jusqu'à atteindre un état terminal.
- Supposons que l'étudiant connaisse les probabilités de transition et les fonctions récompenses. Comment résoudre ce problème?

Solution:



$V_5 = -10$, $V_6 = 100$, $V_7 = -1000$, $V_4 = -10 + 0.9V_6 + 0.1V_4 \simeq 88.9$. $V_3 = -1 + 0.5V_4 + 0.5V_3 \simeq 86.9$. $V_2 = 1 + 0.7V_3 + 0.3V_1$ et $V_1 = \max\{0.5V_2 + 0.5V_1, 0.5V_3 + 0.5V_1\}$, soit: $V_1 = V_2 = 88.3$.

Maintenant, supposons que l'étudiant ne connaisse pas les probabilités de transition et les fonctions récompenses. Peut-il apprendre à résoudre ce problème? C'est l'objectif de l'apprentissage par renforcement!

Maintenance d'un stock: Le responsable d'un entrepot dispose d'un stock x_t d'une marchandise. Il doit satisfaire la demande D_t des clients. Pour cela, il peut, tous les mois, décider de commander une quantité a_t supplémentaire à son fournisseur.

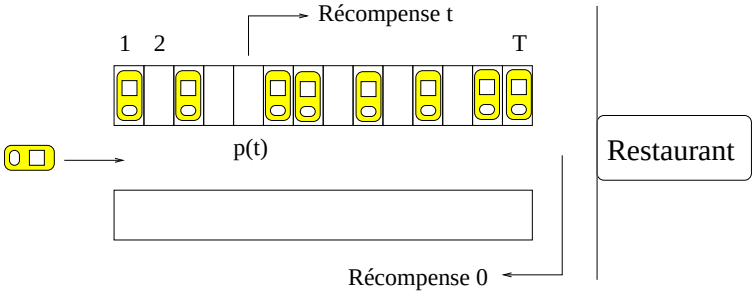
- Il paye un coût de maintenance du stock $h(x)$, un coût de commande du produit $C(a)$,

- Il reçoit un revenu $f(q)$, où q est la quantité vendue.
- Si la demande est supérieure au stock actuel, le client va s'approvisionner ailleurs.
- 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.

Modèle simplifié:

- Modélisation de la demande D_t par une variable aléatoire i.i.d.
- Etat: $x_t \in X = \{0, 1, \ldots, M\}$: quantité (discrète) de produit en stock,
- Décisions: $a_t \in A_{x_t} = \{0, 1, \ldots M - x_t\}$: commande supplémentaire du produit (Rq: ici l'ensemble des actions disponibles à chaque instant dépend de l'état),
- Dynamique: $x_{t+1} = [x_t + a_t - D_t]^+$ (ce qui définit les probabilités de transition $p(x_{t+1}|x_t, a_t)$).
- Récompense: $r_t = -C(a_t) - h(x_t + a_t) + f([x_t + a_t - x_{t+1}]^+)$.
- Critère à maximiser: $\mathbb{E}\big[\sum_{t=1}^{T-1} r_t + g(x_T)\big]$

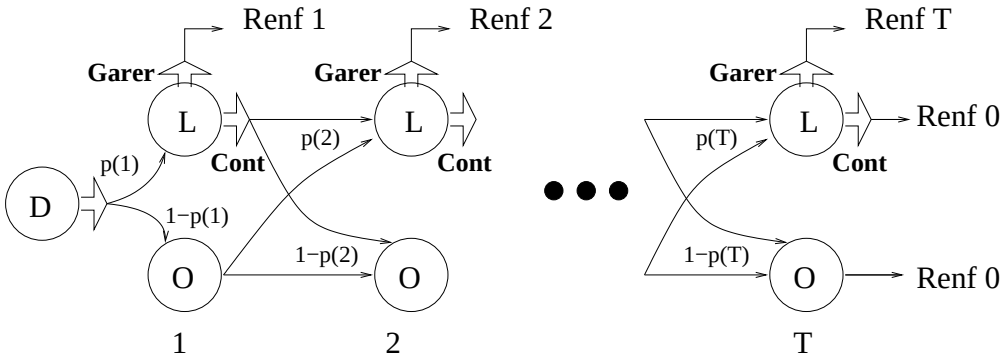
Problème du parking Un conducteur souhaite se garer le plus près possible du restaurant.



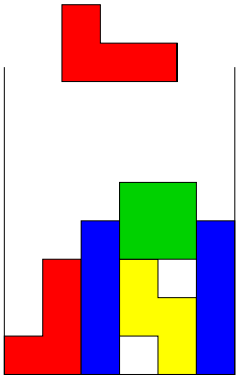
A chaque instant, l'agent possède 2 actions: *continuer* ou *arrêter*.

- 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 ?

Modélisation du parking par un PDM:



Tétris



- Etat: configuration du mur + nouvelle pièce
- Action: positions possibles de la nouvelle pièce sur le mur,
- Récompense: nombre de lignes supprimées
- Etat suivant: nouvelle configuration du mur + aléa sur la nouvelle pièce.

Il est prouvé que pour toute stratégie, le jeu se finit avec probabilité 1. Donc l'espérance de la somme des récompenses à venir est finie.

Difficulté de ce problème: espace d'états très grand (ex: 10^{61} pour hauteur 20, largeur 10, et 7 pièces différentes).

2.3 Problèmes à horizon temporel fini

Considérons un horizon temporel T. Pour une politique $\pi = (\pi_0, \dots, \pi_{T-1})$ donnée, le gain en partant de x à l'instant $t \in \{0, \dots, T\}$, est:

$$V^\pi(t,x) = \mathbb{E}\big[\sum_{s=t}^{T-1} r(x_s, \pi_s(x_s)) + R(x_T) \,|\, x_t = x; \pi\big].$$

Définitions:

- La **fonction valeur optimale** est

$$V^*(t,x) = \max_\pi V^\pi(t,x).$$

- Une politique π^* est dite **optimale** si $V^{\pi^*}(t,x) = V^*(t,x)$.

Proposition 1. La fonction valeur optimale $V^*(t,x)$ est la solution de l'**équation de Bellman**:

$$\begin{aligned} V^*(t,x) &= \max_{a\in A}\big[r(x,a)+\sum_{y\in X} p(y|x,a)V^*(t+1,y)\big],\ \text{pour } 0\le t<T \\ V^*(T,x) &= R(x) \end{aligned} \tag{1}$$

De plus, la politique définie par

$$\pi_t^*(x)\in \arg\max_{a\in A}\big[r(x,a)+\sum_{y\in X} p(y|x,a)V^*(t+1,y)\big],\ \text{pour } 0\le t<T.$$

est une politique optimale.

Proof. On a $V^*(T,x)=R(x)$. Puis résolution rétrograde de V^* pour $t<T$. Toute politique $\pi=(\pi_t,\pi_{t+1},\dots,\pi_{T-1})$ utilisée à partir de l'état initial x à l'instant t est de la forme $\pi=(a,\pi')$ avec $a\in A$ et $\pi'=(\pi_{t+1},\dots,\pi_{T-1})$. Donc

$$\begin{aligned} V^*(t,x) &= \max_{\pi}\mathbb{E}\big[\sum_{s=t}^{T-1} r(x_s,\pi_s(x_s))+R(x_T)|x_t=x;\pi\big] \\ &= \max_{(a,\pi')}\Big[r(x,a)+\sum_{y\in X} p(y|x,a)V^{\pi'}(t+1,y)\Big] \\ &= \max_{a\in A}\Big[r(x,a)+\sum_{y\in X} p(y|x,a)\max_{\pi'}V^{\pi'}(t+1,y)\Big] \\ &= \max_{a\in A}\big[r(x,a)+\sum_{y\in X} p(y|x,a)V^*(t+1,y)\big]. \end{aligned} \tag{2}$$

où (2) se justifie par:

- L'inégalité triviale $\max_{\pi'}\sum_y p(y|x,a)V^{\pi'}(y)\le\sum_y p(y|x,a)\max_{\pi'}V^{\pi'}(y)$
- Soit $\bar\pi=(\bar\pi_{t+1},\dots)$ une politique telle que $\bar\pi_{t+1}(y)=\arg\max_{b\in A}\max_{(\pi_{t+2},\dots)}V^{(b,\pi_{t+2},\dots)}(t+1,y)$. Donc

$$\sum_y p(y|x,a)\max_{\pi'}V^{\pi'}(t+1,y)=\sum_y p(y|x,a)V^{\bar\pi}(t+1,y)\le\max_{\pi'}\sum_y p(y|x,a)V^{\pi'}(y).$$

De plus la politique π_t^* réalise le max à chaque itération donc de manière rétrograde dans le temps on a $V^*=V^{\pi^*}$. □

Exemple du parking: Soit $V^*(t,L)$ (resp. $V^*(t,O)$) la récompense maximale moyenne à la position t lorsque la place est Libre (resp. Occupée). Alors

$$\begin{aligned} V^*(T,L) &= \max\{T,0\}=T,\quad V^*(T,O)=0, \\ V^*(T-1,L) &= \max\{T-1,p(T)V^*(T,L)+(1-p(T))V^*(T,O))\} \\ V^*(t,L) &= \max\{t,p(t+1)V^*(t+1,L)+(1-p(t+1))V^*(t+1,O)\} \end{aligned}$$

et une politique optimale est donnée par l'argument du max.

2.4 Problèmes à horizon temporel infini et critère actualisé

Soit $\pi = (\pi_0, \pi_1, \dots)$ une politique. Considérons la fonction valeur pour la politique π

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

où $0 \leq \gamma < 1$ est un coefficient d'actualisation (ce qui garantit la convergence de la série).

Définissons la fonction valeur optimale

$$V^* = \sup_{\pi=(\pi_0,\pi_1,\dots)} V^\pi$$

Proposition 2. On a:

- Pour une politique π stationnaire (i.e. $\pi = (\pi, \pi, \dots)$) donnée, la fonction valeur V^π satisfait l'**équation de Bellman**:
$$V^\pi(x) = r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) V^\pi(y).$$
- La fonction valeur optimale V^* satisfait l'**équation de programmation dynamique**:
$$V^*(x) = \max_{a \in A} \big[r(x, a) + \gamma \sum_y p(y|x, a) V^*(y)\big].$$

Proof. On a

$$\begin{aligned}
V^\pi(x) &= \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\big] \\
&= r(x, \pi(x)) + \mathbb{E}\big[\sum_{t\geq 1} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\big] \\
&= r(x, \pi(x)) + \gamma \sum_y \mathbb{P}(x_1 = y \,|\, x_0 = x; \pi(x_0)) \mathbb{E}\big[\sum_{t\geq 1} \gamma^{t-1} r(x_t, \pi(x_t)) \,|\, x_1 = y; \pi\big] \\
&= r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) V^\pi(y).
\end{aligned}$$

Et aussi, pour toute politique $\pi = (a, \pi')$ (non-nécessairement stationnaire),

$$\begin{aligned}
V^*(x) &= \max_\pi \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\big] \\
&= \max_{(a,\pi')} \Big[r(x, a) + \gamma \sum_y p(y|x, a) V^{\pi'}(y)\Big] \\
&= \max_a \Big[r(x, a) + \gamma \sum_y p(y|x, a) \max_{\pi'} V^{\pi'}(y)\Big] \qquad (3) \\
&= \max_a \Big[r(x, a) + \gamma \sum_y p(y|x, a) V^*(y)\Big].
\end{aligned}$$

où (3) se justifie par:

- L'inégalité triviale $\max_{\pi'} \sum_y p(y|x, a) V^{\pi'}(y) \leq \sum_y p(y|x, a) \max_{\pi'} V^{\pi'}(y)$

- Soit $\bar{\pi}$ la politique définie par $\bar{\pi}(y) = \arg\max_{\pi'} V^{\pi'}(y)$. Donc

$$\sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(y) = \sum_y p(y|x,a) V^{\bar{\pi}}(y) \leq \max_{\pi'} \sum_y p(y|x,a) V^{\pi'}(y).$$

□

Opérateurs $\mathcal{T}^\pi$ et $\mathcal{T}$ Définissons l'**opérateur de Bellman** $\mathcal{T}^\pi : \mathbb{R}^N \to \mathbb{R}^N$: pour tout $W \in \mathbb{R}^N$,

$$\mathcal{T}^\pi W(x) = r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) W(y)$$

et l'**opérateur de programmation dynamique** $\mathcal{T} : \mathbb{R}^N \to \mathbb{R}^N$:

$$\mathcal{T} W(x) = \max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) W(y) \right].$$

Notations: considérons V^π comme un vecteur de taille N. Notons r^π le vecteur de composantes $r^\pi(x) = r(x, \pi(x))$ et P^π la matrice (stochastique) $N \times N$ d'éléments $P^\pi(x,y) = p(y|x, \pi(x))$.

Proposition 3. On a:

1. Pour une politique π, la fonction valeur s'écrit $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$,
2. V^π est l'unique point-fixe de $\mathcal{T}^\pi$.
3. V^* est l'unique point-fixe de $\mathcal{T}$.
4. Toute politique $\pi^*(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^*(y) \right]$ est optimale et stationnaire. (ce qui nous permet de nous intéresser uniquement aux politiques stationnaires).
5. Pour tout vecteur $W \in \mathbb{R}^N$, pour toute politique stationnaire π,

$$\begin{aligned} \lim_{k \to \infty} (\mathcal{T}^\pi)^k W &= V^\pi, \\ \lim_{k \to \infty} (\mathcal{T})^k W &= V^*. \end{aligned}$$

Propriétés des opérateurs $\mathcal{T}^\pi$ et $\mathcal{T}$:

- **Monotonicité**: Si $W_1 \leq W_2$ (composante par composante), alors

$$\begin{aligned} \mathcal{T}^\pi W_1 &\leq \mathcal{T}^\pi W_2, \\ \mathcal{T} W_1 &\leq \mathcal{T} W_2. \end{aligned}$$

- **Contraction en norme sup**: Pour tous vecteurs W_1 et W_2,

$$\begin{aligned} ||\mathcal{T}^\pi W_1 - \mathcal{T}^\pi W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty, \\ ||\mathcal{T} W_1 - \mathcal{T} W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty. \end{aligned}$$

En effet, pour tout $x \in X$,

$$\begin{aligned} |\mathcal{T}W_1(x) - \mathcal{T}W_2(x)| &= \left| \max_a \left[r(x,a) + \gamma \sum_y p(y|x,a) W_1(y) \right] - \max_a \left[r(x,a) + \gamma \sum_y p(y|x,a) W_2(y) \right] \right| \\ &\leq \gamma \max_a \sum_y p(y|x,a) |W_1(y) - W_2(y)| \leq \gamma ||W_1 - W_2||_\infty \end{aligned}$$

Proof. (Proposition 3)

1. D'après la proposition 2, on a $V^\pi = r^\pi + \gamma P^\pi V^\pi$. Donc $(I - \gamma P^\pi)V^\pi = r^\pi$. La matrice P^π est une matrice stochastique donc ses valeurs propres sont de module ≤ 1. Donc les v.p. de $(I - \gamma P^\pi)$ sont de module $\geq 1 - \gamma$ et cette matrice est donc inversible.

2. D'après la proposition 2, V^π est un point fixe de $\mathcal{T}^\pi$. L'unicité découle de la contraction de $\mathcal{T}^\pi$.

3. D'après la proposition 2, V^* est un point fixe de $\mathcal{T}$. L'unicité découle de la contraction de $\mathcal{T}$.

4. D'après la définition de π^*, on a $\mathcal{T}^{\pi^*} V^* = \mathcal{T} V^* = V^*$. Donc V^* est le point fixe de $\mathcal{T}^{\pi^*}$. Mais comme par définition V^{π^*} est le point fixe de $\mathcal{T}^{\pi^*}$ et qu'il y a unicité de point fixe, donc $V^{\pi^*} = V^*$ et la politique π^* est optimale.

5. Considérons la suite définie par récurrence $W_{k+1} = \mathcal{T}W_k$ avec $W_0 = W$ quelconque. Alors les W_k sont bornés: $||W_{k+1}||_\infty \leq r_{\max} + \gamma ||W_k||_\infty$, soit $||W_k||_\infty \leq \frac{r_{\max}}{1-\gamma}$. De plus, pour $k \geq p$,

   $$||W_k - W_p||_\infty \leq \gamma ||W_{k-1} - W_{p-1}||_\infty \leq \cdots \leq \gamma^p ||W_{k-p} - W_0||_\infty \overset{k,p\to\infty}{\longrightarrow} 0$$

   donc (W_k) est de Cauchy. L'espace des fonctions sur $\mathbb{R}^N$ muni de la norme L_∞ est complet (espace de Banach), donc la suite (W_k) converge vers $\tilde{W}$. Par passage à la limite dans la définition de W_k, il vient $\tilde{W} = \mathcal{T}\tilde{W}$. D'après l'unicité de solution de point-fixe de $\mathcal{T}$, on a $\lim_{k\to\infty} W_k = \lim_{k\to\infty} (\mathcal{T})^k W = V^*$.

   Le même raisonnement tient pour montrer $\lim_{k\to\infty} (\mathcal{T}^\pi)^k W = V^\pi$.

   (Remarque: on a redémontré le théorème de point-fixe de Banach pour un opérateur contractant).

□

2.5 Algorithmes de programmation dynamique

2.5.1 Itérations sur les valeurs (IV)

Construisons une séquence de fonctions (V_k), avec V_0 quelconque, et V_k calculée selon:

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

Alors $\lim_{k\to\infty} V_k = V^*$.

En effet, $||V_{k+1} - V^*|| = ||\mathcal{T}V_k - \mathcal{T}V^*|| \leq \gamma ||V_k - V^*|| \leq \gamma^{k+1} ||V_0 - V^*|| \to 0$.

Nombreuses variantes existent selon l'ordre selon lequel on met à jour les valeurs du tableau V_k. Ex: itération asynchrone, à chaque itération k, on choisit un état x_k que l'on itère: $V_{k+1}(x_k) = \mathcal{T}V_k(x_k)$, les autres restant inchangés. En supposant que tous les états sont sélectionnés infiniment souvent, alors $V_k \to V^*$.

2.5.2 Itérations sur les politiques (IP)

On construit une séquence de politiques. Politique initiale π_0 quelconque. A chaque étape k,

1. **Evaluation de la politique** π_k : on calcule V^{π_k}.
2. **Amélioration de la politique**: on calcule π_{k+1} déduite de V^{π_k}:

$$\pi_{k+1}(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y) \right],$$

(on dit que π_{k+1} est gloutonne par rapport à V^{π_k}, c'est à dire $\mathcal{T}^{\pi_{k+1}} V^{\pi_k} = \mathcal{T} V^{\pi_k}$).

On s'arrête quand $V^{\pi_k} = V^{\pi_{k+1}}$.

Proposition 4. L'algorithme d'IP génère une séquence de politiques de performances croissantes ($V^{\pi_{k+1}} \geq V^{\pi_k}$) qui se termine en un nombre fini d'étapes avec une politique optimale π^*.

Proof. D'après la définition des opérateurs $\mathcal{T}$, $\mathcal{T}^{\pi_k}$, $\mathcal{T}^{\pi_{k+1}}$ et celle de π_{k+1},

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{4}$$

et par la monotonicité de $\mathcal{T}^{\pi_{k+1}}$, il vient

$$V^{\pi_k} \leq \lim_{n \to \infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}.$$

Donc $(V^{\pi_k})_k$ est une suite croissante. Comme il y a un nombre fini de politiques possibles, le critère d'arrêt est nécessairement satisfait pour un certain k; on a alors égalité dans (4), donc

$$V^{\pi_k} = \mathcal{T} V^{\pi_k}$$

et donc $V^{\pi_k} = V^*$ et π_k est une politique optimale. □

L'algorithme d'itération sur les politique peut être vu comme un algorithme de type *Actor-Critic*.

Evaluation de la politique L'algorithme d'IP nécessite de calculer la fonction valeur V^π de la politique courante π, soit le point fixe de $\mathcal{T}^\pi$.

Méthodes de résolution possibles:

- **Résolution directe** du système linéaire $(I - \gamma P^\pi) V^\pi = r^\pi$. Méthode d'élimination de Gauss –> complexité en $O(N^3)$ (ou $O(N^{2.807})$ pour l'algorithme de Strassen).
- **Itération sur les valeurs pour une politique fixe**: On itère l'opérateur $\mathcal{T}^\pi$. Soit V_0 quelconque, $V_{n+1} = \mathcal{T}^\pi V_n$. Alors convergence de V_n vers V^π. Problème: convergence asymptotique. Avantage: complexité $O(N^2 \frac{\log 1/\epsilon}{\log 1/\gamma})$ pour une ϵ-approximation (intéressant lorsque γ n'est pas trop proche de 1).
- **Monte-Carlo**: on simule n trajectoires $((x_t^i)_{t \geq 0,})_{1 \leq i \leq n}$, partant de x et suivant la politique π: $x_{t+1}^i \sim p(\cdot|x_t^i, \pi(x_t^i))$, alors

$$V^\pi(x) \simeq \frac{1}{n} \sum_{i=1}^{n} \sum_{t \geq 0} \gamma^t r(x_t^i, \pi(x_t^i)).$$

Intéressant lorsqu'on s'intéresse à l'évaluation d'un seul état. Erreur d'approximation en $O(1/\sqrt{n})$.

- **Différences temporelles TD(λ)** [Sutton, 1988]: méthode intelligente pour utiliser des trajectoires pour évaluer tous les états traversés par ces trajectoires, en évaluant la valeur d'un état x_s par la somme des différences temporelles $r_t + \gamma V(x_{t+1}) - V(x_t)$ observées aux instants t futurs pondérées par une "trace" λ^{t-s}.
  - lorsque $\lambda = 1$ on retrouve Monte-Carlo
  - lorsque $\lambda = 0$ on retrouve Itérations sur les valeurs (asynchrone,incrémentale)

  On détaillera TD(λ) plus loin dans la section sur les algorithmes d'apprentissage par renforcement.

Comparaison algos IV / IP

- **Itération sur les valeurs:** Chaque itération est rapide ($O(N^2A)$ opérations), mais nécessite $O(\frac{\log 1/\epsilon}{\log 1/\gamma})$ itérations pour obtenir une approximation à ϵ près de V^*.
- **Itération sur les politiques:** converge en un nombre fini d'étapes (habituellement faible mais théoriquement peut être très grand...), mais chaque étape nécessite une évaluation de la politique.
- **Itération sur les politiques modifié**: Il n'est souvent pas besoin de connaître exactement V^{π_k} pour améliorer la nouvelle politique π_{k+1} -> étape d'évaluation grossière (à l'aide de quelques itérations sur les valeurs) de la politique. Voir [Puterman, 1994]

2.5.3 Représentation alternative: les fonctions Q-valeurs.

Définissons pour toute politique π la fonction Q-valeur $Q^\pi : X \times A \mapsto \mathbb{R}$:

$$Q^\pi(x,a) = \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(x_t,a_t) | x_0 = x, a_0 = a, a_t = \pi(x_t) \text{ pour } t \geq 1\big]$$

et la fonction Q-valeur optimale:

$$Q^*(x,a) = \max_\pi Q^\pi(x,a).$$

Remarquons les liens entre les fonctions valeur et Q-valeur:

$$\begin{aligned} Q^\pi(x,a) &= r(x,a) + \gamma \sum_{y\in X} p(y|x,a) V^\pi(y) \\ V^\pi(x) &= Q^\pi(x,\pi(x)) \\ Q^*(x,a) &= r(x,a) + \gamma \sum_{y\in X} p(y|x,a) V^*(y) \\ V^* &= Q^*(x,\pi^*(x)) = \max_{a\in A} Q^*(x,a) \end{aligned}$$

Et l'on a $\pi^*(x) \in \arg\max_{a\in A} Q^*(x,a)$.

On déduit les équations de Bellman et de PD pour les fonctions Q-valeurs:

$$\begin{aligned} Q^\pi(x,a) &= r(x,a) + \gamma \sum_{y\in X} p(y|x,a) Q^\pi(y,\pi(y)) \\ Q^*(x,a) &= r(x,a) + \gamma \sum_{y\in X} p(y|x,a) \max_{b\in A} Q^*(y,b) \end{aligned}$$

Donc en définissant les opérateurs de Bellman $\mathcal{T}^\pi$ et de programmation dynamique $\mathcal{T}$ (pour simplifier nous utilisons les mêmes notations que pour les fonctions valeurs) pour les fonctions Q-valeurs, c'est à dire portant sur des fonctions définies sur l'espace produit $X \times A$:

$$\begin{aligned} \mathcal{T}^\pi W(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) W(y, \pi(y)) \\ \mathcal{T} W(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) \max_{b \in A} W(y,b) \end{aligned}$$

alors nous avons l'analogue de la Proposition 3 (preuve identique): pour toute politique π, Q^π est l'unique point-fixe de $\mathcal{T}^\pi$, et Q^* est l'unique point-fixe de $\mathcal{T}$. De plus nous déduisons les algorithmes suivants:

- **Itération sur les Q-valeurs**: Q_0 quelconque, et $Q_{k+1} = \mathcal{T} Q_k$. Converge vers Q^*.
- **Itérations sur les politiques**: π_0 quelconque, puis itérations à étape k:
  - **Evaluation de la politique**: calcul de la Q-fonction valeur Q^{π_k}
  - **Amélioration de la politique**: calcul de la politique améliorée:

$$\pi_{k+1}(x) = \arg\max_{a \in A} Q^{\pi_k}(x,a).$$

  Intérêt : le calcul du max dans l'étape d'amélioration de la politique ne nécessite pas le calcul d'une espérance (ce qui est intéressant dans un cadre apprentissage par renforcement où les probabilités de transition ne sont pas connues).

2.5.4 Programmation linéaire

Commençons par donner une interprétation de l'algorithme d'itérations sur les politiques comme une méthode de type Newton pour trouver un zéro du résidu de Bellman.

Interprétation géométrique de l'algo d'IP Trouver un point fixe de $\mathcal{T}$ est équivalent à trouver un zéro de l'opérateur résidu de Bellman $B = \mathcal{T} - I$.

Proposition 5. Les π_k générées par l'algo IP vérifient

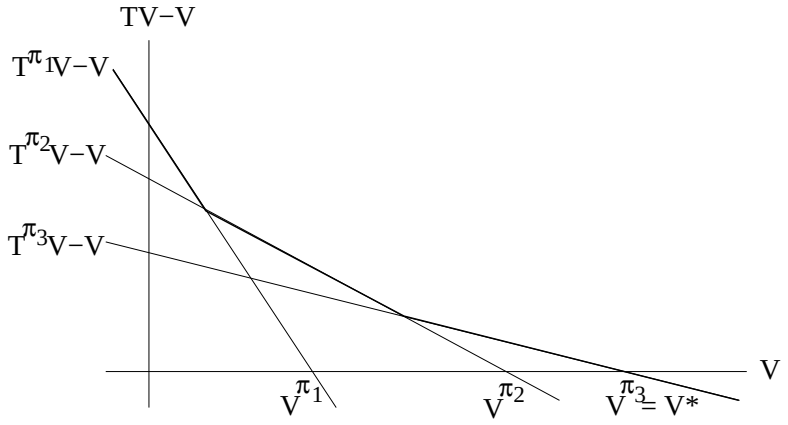
$$\begin{aligned} V^{\pi_{k+1}} &= V^{\pi_k} - (\gamma P^{\pi_{k+1}} - I)^{-1} [\mathcal{T}^{\pi_{k+1}} V^{\pi_k} - V^{\pi_k}] \\ &= V^{\pi_k} - [B']^{-1} B V^{\pi_k} \end{aligned}$$

Proof. En effet,

$$\begin{aligned} V^{\pi_{k+1}} &= (I - \gamma P^{\pi_{k+1}})^{-1} r^{\pi_{k+1}} - V^{\pi_k} + V^{\pi_k} \\ &= V^{\pi_k} + (I - \gamma P^{\pi_{k+1}})^{-1} [r^{\pi_{k+1}} + (\gamma P^{\pi_{k+1}} - I) V^{\pi_k}] \end{aligned}$$

□

Il s'agit d'une méthode de type Newton pour trouver un zéro de B.



V^{π_k} est zéro de l'opérateur linéaire $\mathcal{T}^{\pi_k} - I$. L'application $V \to \mathcal{T}V - V = \max_\pi T^\pi V - V$ est convexe –> convergence de l'algorithme de Newton pour tout V_0 tel que $\mathcal{T}V_0 - V_0 \geq 0$.

Programmation linéaire En reprenant l'interprétation géométrique, V^* est le plus petit V tel que $V \geq \mathcal{T}V$. En effet,

$$V \geq \mathcal{T}V \quad \Longrightarrow \quad V \geq \lim_{k\to\infty} (\mathcal{T})^k V = V^*.$$

Donc V^* est solution du **programme linéaire**:

- Min $\sum_x V(x)$,
- Sous les contraintes (système fini d'inégalités linéaires qui définit un polyèdre dans $\mathbb{R}^N$):

$$V(x) \geq r(x,a) + \gamma \sum_y p(y|x,a)V(y), \quad \forall x \in X, \forall a \in A$$

(qui comprend N variables et $N \times |A|$ contraintes).

2.6 Problèmes à horizon temporel infini non actualisés

Pour une politique stationnaire π:

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

On suppose qu'il existe un *état terminal absorbant*, noté 0. Une fois cet état atteint, le système y reste indéfiniment, avec récompense nulle.

Exemple: En remplaçant max par min, il s'agit d'un problème de plus court chemin stochastique dans un graphe dirigé où il s'agit de trouver le chemin le + court en moyenne (où le coût = longueur du chemin) qui parte d'un nœud donné et arrive à destination (état absorbant).

Pour s'assurer de la convergence de la série on fait une hypothèse sur la probabilité d'atteindre l'état terminal:

Politiques propres: une politique stationnaire π est dite *propre*, s'il existe un entier n tel qu'avec une probabilité strictement positive l'état terminal est atteint en au plus n étapes, quelque soit l'état initial:

$$\rho_\pi = \max_x \mathbb{P}(x_n \neq 0 \,|\, x_0 = x, \pi) < 1.$$

Propriété: sous une politique propre, $\mathbb{P}(x_{2n} \neq 0 \,|\, x_0 = x, \pi) = \mathbb{P}(x_{2n} \neq 0 \,|\, x_n \neq 0, \pi) \times \mathbb{P}(x_n \neq 0 \,|\, x_0 = x, \pi) \leq \rho_\pi^2$. Ainsi, $\mathbb{P}(x_t \neq 0 \,|\, x_0 = x, \pi) \leq \rho_\pi^{\lfloor t/n \rfloor}$, donc l'état terminal est atteint avec probabilité 1, et la fonction valeur

$$||V^\pi|| \leq \sum_{t \geq 0} \rho_\pi^{\lfloor t/n \rfloor} r_{\max}.$$

Equation de programmation dynamique On fait l'hypothèse qu'il existe au moins une politique propre, et que pour toute politique non-propre π, la fonction valeur V^π est égale à moins l'infini en au moins un état (existence d'un cycle avec somme des récompenses négatives).

Exemple: problème de plus court chemin: de tout état initial, il existe un chemin qui mène à destination, et tous les cycles du graphe ont un coût positif (coût = - récompense).

Proposition 6 (Bertsekas et Tsitsiklis, 1996). Sous cette hypothèse, la fonction valeur optimale V^* a des composantes finies et est l'unique point fixe de l'opérateur de Bellman $\mathcal{T}$, défini pour tout $W \in \mathbb{R}^N$, par

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

De plus, elle est la limite $V^* = \lim_{k \to \infty} (\mathcal{T})^k W$, pour tout $W \in \mathbb{R}^N$.

Méthodes de programmation dynamique: Les méthodes de résolution dans le cas actualisé s'appliquent ici: la fonction valeur optimale peut être calculée par itération sur les valeurs, itération sur les politiques, et leurs versions asynchrones ou modifiées.

Les preuves de convergence sont plus fines (voir les références [Bertsekas et Tsitsiklis, 1996] et [Puterman, 1994]) car ici $\mathcal{T}$ n'est pas une contraction en L_∞ mais en norme L_∞ pondérée. Définition norme pondérée avec poids μ (vecteur strictement positif): $||W||_\mu = \max_{x \in X} \frac{|W(x)|}{\mu(x)}$.

Contraction en norme pondérée

Proposition 7. Supposons que toutes les politiques sont propres. Il existe un vecteur μ, de composantes > 0, et un réel $\beta < 1$ tels que, $\forall x, y \in X_N$, $\forall a \in A$,

$$\sum_y p(y|x,a) \mu(y) \leq \beta \mu(x).$$

On en déduit que les opérateurs $\mathcal{T}$ et $\mathcal{T}^\pi$ sont des contractions en norme L_∞ pondérée avec le poids μ:

$$||\mathcal{T}W_1 - \mathcal{T}W_2||_\mu \leq \beta ||W_1 - W_2||_\mu$$

Proof. Soit μ défini par le max (sur toutes les politiques) du temps moyen d'atteinte de l'état terminal. Il s'agit d'un problème de décision markovien avec récompense immédiate 1 quelle que soit l'action choisie et sous l'hypothèse que toutes les politiques sont propres on a que μ est fini et est solution de l'équation de PD:

$$\mu(x) = 1 + \max_a \sum_y p(y|x,a) \mu(y).$$

Donc $\mu(x) \geq 1$ et on a, pour tout $a \in A$, $\mu(x) \geq 1 + \sum_y p(y|x,a)\mu(y)$. Ainsi,

$$\sum_y p(y|x,a)\mu(y) \leq \mu(x) - 1 \leq \beta\mu(x),$$

pour

$$\beta = \max_x \frac{\mu(x) - 1}{\mu(x)} < 1.$$

D'où l'on déduit la propriété de contraction de $\mathcal{T}$ en norme $L_{\infty,\mu}$:

$$\begin{aligned} ||\mathcal{T}W_1 - \mathcal{T}W_2||_\mu &= \max_x \frac{|\mathcal{T}W_1(x) - \mathcal{T}W_2(x)|}{\mu(x)} \\ &\leq \max_{x,a} \frac{\sum_y p(y|x,a)}{\mu(x)} |W_1(y) - W_2(y)| \\ &\leq \max_{x,a} \frac{\sum_y p(y|x,a)\mu(y)}{\mu(x)} \|W_1 - W_2\|_\mu \\ &\leq \beta \|W_1 - W_2\|_\mu \end{aligned}$$

□