IFT-3655, Modèles Stochastiques

Processus de décision Markoviens

Prof. Pierre L'Ecuyer
DIRO, Université de Montréal

Référence pour ce chapitre:

D. P. Bersekas, Dynamic Programming and Optimal Control, volume 1;
athenasc.com/dpbook.html

Idée générale d'un processus de décision Markovien

Un processus de décision Markovien (PDM) est une structure contenant une "chaîne de Markov" pour laquelle à chaque étape, on observe l'état de la chaîne et on choisit une action ou décision qui influence les probabilités pour la prochaine transition.

À chaque étape, on a aussi un coût qui dépend de l'état actuel et de la décision prise. En réalité, pour un état et une décision donnés, ce coût peut être aléatoire (dépendre par exemple du prochain état ou d'autre information inconnue au moment de prendre la décision courante), mais on le remplace alors par son espérance conditionnelle à l'état et la décision de l'étape courante.

Idée générale d'un processus de décision Markovien

Un processus de décision Markovien (PDM) est une structure contenant une "chaîne de Markov" pour laquelle à chaque étape, on observe l'état de la chaîne et on choisit une action ou décision qui influence les probabilités pour la prochaine transition.

À chaque étape, on a aussi un coût qui dépend de l'état actuel et de la décision prise. En réalité, pour un état et une décision donnés, ce coût peut être aléatoire (dépendre par exemple du prochain état ou d'autre information inconnue au moment de prendre la décision courante), mais on le remplace alors par son espérance conditionnelle à l'état et la décision de l'étape courante.

Exemples:
Un avion, une auto, un robot, ..., qui se conduisent tout seuls;
gestion d'un portefeuille d'investissement en finance;
gestion d'un système de production, d'un système d'inventaire, etc.
un match de tennis ou de football;
Etc.

L'objectif est d'optimiser la prise des décisions (ou la commande du système), disons pour minimiser l'espérance du coût total. Ce coût total peut être pour un nombre fini d'étapes fixé à l'avance, ou encore un nombre aléatoire d'étapes (par exemple jusqu'à ce que l'état de la chaîne atteigne un certain sous-ensemble de l'espace d'états), ou pour un nombre infini d'étapes mais avec une actualisation des coûts (un coût payé au temps t est multiplié par $e^{-\rho t}$ pour un certain $\rho > 0$), ou encore ce peut-être le coût moyen par unité de temps sur un horizon infini.

Une politique de prise de décisions est une fonction (une règle) qui à chaque état associe une décision à prendre. Quand l'horizon est fini et fixé, la règle peut dépendre du numéro d'étape. On cherche une politique optimale, qui minimise le coût total espéré.

Dans certains cas, on voudra considérer une politique randomisée, qui à chaque état associe une loi de probabilité sur l'espace des décisions. Parfois nécessaire s'il y a des contraintes, par exemple. Ou dans le cas où un adversaire prend aussi des décisions.

PDM en temps discret sur horizon fini

On a un modèle qui ressemble à celui de CMTD vu précédemment, sauf que l'on doit maintenant prendre une décision à chaque étape et que les probabilités de transition de la chaîne dépendent aussi de la décision prise, à chaque étape.

$\mathcal{X} \subseteq \{0, 1, 2, \dots\}$: espace d'états fini;
$\mathcal{A}$: espace des décisions (ou actions) fini;
$\mathcal{X}_n \subseteq \mathcal{X}$: espace d'états à l'étape n;
X_n: état à l'étape n;
$A_n(X_n)$: ensemble des décisions admissibles dans l'état X_n à l'étape n;
a_n: action (décision) prise à l'étape n;
$c_n(X_n, a_n)$: coût (espéré) à l'étape n si on est dans l'état X_n et on prend la décision a_n;
$P_{i,j}(n, a) = \mathbb{P}[X_{n+1} = j \mid X_n = i, a_n = a]$.

Le processus est supposé Markovien: si on est dans l'état $X_n \in \mathcal{X}_n$ et que l'on prend une décision admissible a_n à l'étape n, la loi de probabilité de l'évolution future conditionnelle à $(n, X_0, a_0, X_1, a_1, \dots, X_n, a_n)$ est la même que celle conditionnelle à (n, X_n, a_n).

À l'étape n, on observe l'état X_n et on prend une décision $a_n \in A_n(X_n)$, puis on paye un coût[5] (espéré) $c_n(X_n, a_n)$. La loi de probabilité du prochain état X_{n+1} dépend de (X_n, a_n):

$$\mathbb{P}[X_{n+1} = j \mid X_n = i, a_n = a] = P_{i,j}(n, a).$$

À l'étape n, on observe l'état X_n et on prend une décision $a_n \in A_n(X_n)$, puis on paye un coût[5] (espéré) $c_n(X_n, a_n)$. La loi de probabilité du prochain état X_{n+1} dépend de (X_n, a_n):

$$\mathbb{P}[X_{n+1} = j \mid X_n = i, a_n = a] = P_{i,j}(n, a).$$

Coût total (aléatoire) additif pour un horizon de N étapes:

$$\sum_{n=0}^{N} c_n(X_n, a_n).$$

À l'étape N, on paye un coût mais on ne prend habituellement pas de décision, car c'est terminé, mais pour éviter d'introduire une notation additionnelle, on peut supposer simplement dans ce cas qu'il y a une seule décision a_N admissible: ne rien faire.

À l'étape n, on observe l'état X_n et on prend une décision $a_n \in A_n(X_n)$, puis on paye un coût[5] (espéré) $c_n(X_n, a_n)$. La loi de probabilité du prochain état X_{n+1} dépend de (X_n, a_n):

$$\mathbb{P}[X_{n+1} = j \mid X_n = i, a_n = a] = P_{i,j}(n, a).$$

Coût total (aléatoire) additif pour un horizon de N étapes:

$$\sum_{n=0}^{N} c_n(X_n, a_n).$$

À l'étape N, on paye un coût mais on ne prend habituellement pas de décision, car c'est terminé, mais pour éviter d'introduire une notation additionnelle, on peut supposer simplement dans ce cas qu'il y a une seule décision a_N admissible: ne rien faire.

Une politique admissible est une suite de N fonctions $\pi = (\mu_0, \ldots, \mu_N)$ telle que $\mu_n : \mathcal{X} \to \mathcal{A}$ et $\mu_n(x) \in A_n(x)$ pour tout $x \in \mathcal{X}_n$, $0 \leq n \leq N$. Une politique est dite optimale si elle minimise l'espérance mathématique du coût total:

$$\min_{\pi} \mathbb{E}_{\pi} \left[\sum_{n=0}^{N} c_n(X_n, a_n) \right].$$

Pour $0 \leq n \leq N$ et $x \in \mathcal{X}_n$, posons

$$\begin{aligned} V_{\pi,n}(x) &= \text{coût espéré total de l'étape } n \text{ à la fin si on est dans l'état } x \text{ à l'étape } n \\ &\quad \text{et si on utilise la politique } \pi \\ &= \mathbb{E}_{\pi,x}\left[\sum_{k=n}^{N} c_k(X_k, a_k)\right] = \mathbb{E}_{\pi}\left[\sum_{k=n}^{N} c_k(X_k, a_k) \mid X_n = x\right] \end{aligned}$$

où $\mathbb{E}_{\pi,x}$ indique l'espérance lorsqu'on est dans l'état x et on suit la politique π jusqu'à la fin: $X_n = x$ et $a_k = \mu_k(X_k)$ pour $k = n, \ldots, N$.

Pour une politique π donnée, on a l'équation de récurrence

$$\begin{aligned} V_{\pi,N}(x) &= c_N(x, \mu_N(x)) \quad \text{pour tout } x \in \mathcal{X}_N, \\ V_{\pi,n}(x) &= \mathbb{E}_{\pi,x}\left[c_n(x, \mu_n(x)) + V_{\pi,n+1}(X_{n+1})\right] \qquad \text{pour } 0 \leq n < N,\ x \in \mathcal{X}_n. \end{aligned}$$

En effet:

$$\begin{aligned} V_{\pi,n}(x) &= \mathbb{E}_{\pi,x}\left[\sum_{k=n}^{N} c_k(X_k, a_k)\right] \\ &= \mathbb{E}_{\pi,x}\left[c_n(x, \mu_n(x)) + \mathbb{E}_{\pi,x}\left[\sum_{k=n+1}^{N} c_k(X_k, u_k) \mid X_{n+1}\right]\right] \\ &= \mathbb{E}_{\pi,x}\left[c_n(x, \mu_n(x)) + V_{\pi,n+1}(X_{n+1})\right]. \end{aligned}$$

En effet:

$$\begin{aligned} V_{\pi,n}(x) &= \mathbb{E}_{\pi,x}\left[\sum_{k=n}^{N} c_k(X_k, a_k)\right] \\ &= \mathbb{E}_{\pi,x}\left[c_n(x, \mu_n(x)) + \mathbb{E}_{\pi,x}\left[\sum_{k=n+1}^{N} c_k(X_k, u_k) \mid X_{n+1}\right]\right] \\ &= \mathbb{E}_{\pi,x}\left[c_n(x, \mu_n(x)) + V_{\pi,n+1}(X_{n+1})\right]. \end{aligned}$$

On cherche une politique π qui minimise $V_{\pi,0}(x_0)$, l'espérance mathématique de la somme des coûts de l'étape 0 à l'étape N, si $X_0 = x_0$.
Notons $\pi^* = (\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*)$ une telle politique optimale. Posons

$$\begin{aligned} V_n^*(x) &= \text{coût espéré total optimal de l'étape } n \text{ à la fin,} \\ &\quad \text{si on est dans l'état } x \text{ à l'étape } n \\ &= \min_{\pi} V_{\pi,n}(x) \\ &= \min_{\mu_n,\ldots,\mu_N} V_{\mu_n,\ldots,\mu_N,n}(x). \end{aligned}$$

Proposition.
(A) On a $V_n^* \equiv V_n$, où les fonctions V_n sont définies par les équations de récurrence (les équations de la programmation dynamique):

$$\begin{aligned} V_{N+1}(x) &= 0 \quad \forall x \in \mathcal{X}, \\ V_n(x) &= \min_{a \in A_n(x)} \mathbb{E}\left[c_n(x,a) + V_{n+1}(X_{n+1})\right] \qquad \text{pour } 0 \le n \le N,\ x \in \mathcal{X}_n, \end{aligned}$$

où l'espérance $\mathbb{E}$ est par rapport aux probabilités $P_{i,j}(n,a)$.

(B) Une valeur de a qui fait atteindre le minimum ci-haut est une décision optimale à prendre lorsqu'on est dans l'état x à l'étape n. On peut définir une politique optimale (si elle existe) par

$$\mu_n^*(x) = \arg \min_{a \in A_n(x)} \mathbb{E}\left[c_n(x,a) + V_{n+1}(X_{n+1})\right] \qquad \text{pour tout } x \in \mathcal{X}_n.$$

On a alors $V_n \equiv V_{\pi^*,n}$ pour tout n.

Preuve informelle de (A) et (B).

Pour $\pi = (\mu_1, \ldots, \mu_N)$, on note $\pi^n = (\mu_n, \ldots, \mu_N)$. On a

$$V_n^*(x) \quad = \quad \min_{\pi^n} \mathbb{E}_{\pi^n, x} \left[\sum_{k=n}^{N} c_k(X_k, \mu_k(X_k)) \right] \qquad \text{pour } 0 \leq n \leq N,\ x \in \mathcal{X}_n.$$

Preuve informelle de (A) et (B).

Pour $\pi = (\mu_1, \ldots, \mu_N)$, on note $\pi^n = (\mu_n, \ldots, \mu_N)$. On a

$$V_n^*(x) \quad = \quad \min_{\pi^n} \mathbb{E}_{\pi^n, x} \left[\sum_{k=n}^{N} c_k(X_k, \mu_k(X_k)) \right] \qquad \text{pour } 0 \leq n \leq N,\ x \in \mathcal{X}_n.$$

On vérifie facilement que $V_N^* = V_N$.

On montre ensuite par induction sur n (pour $n = N-1, \ldots, 0$) que $V_n^* = V_n$.

Supposons que $V_{n+1}^* = V_{n+1}$. On écrit $\pi^n = (\mu_n, \pi^{n+1})$.

$$V_n^*(X_n) = \min_{(\mu_n, \pi^{n+1})} \mathbb{E}_{\pi^n, X_n} \left[c_n(X_n, \mu_n(X_n)) + \sum_{k=n+1}^{N} c_k(X_k, \mu_k(X_k)) \right]$$

$$
\begin{aligned}
V_n^*(X_n) &= \min_{(\mu_n, \pi^{n+1})} \mathbb{E}_{\pi^n, X_n} \left[c_n(X_n, \mu_n(X_n)) + \sum_{k=n+1}^{N} c_k(X_k, \mu_k(X_k)) \right] \\
&= \min_{\mu_n} \mathbb{E}_{\pi^n, X_n} \left(c_n(X_n, \mu_n(X_n)) + \min_{\pi^{n+1}} \mathbb{E}_{\pi^{n+1}, X_{n+1}} \left[\sum_{k=n+1}^{N} c_k(X_k, \mu_k(X_k)) \mid X_{n+1} \right] \right)
\end{aligned}
$$

$$
\begin{aligned}
V_n^*(X_n) &= \min_{(\mu_n, \pi^{n+1})} \mathbb{E}_{\pi^n, X_n} \left[c_n(X_n, \mu_n(X_n)) + \sum_{k=n+1}^{N} c_k(X_k, \mu_k(X_k)) \right] \\
&= \min_{\mu_n} \mathbb{E}_{\pi^n, X_n} \left(c_n(X_n, \mu_n(X_n)) + \min_{\pi^{n+1}} \mathbb{E}_{\pi^{n+1}, X_{n+1}} \left[\sum_{k=n+1}^{N} c_k(X_k, \mu_k(X_k)) \mid X_{n+1} \right] \right) \\
&= \min_{\mu_n} \mathbb{E}_{\pi^n, X_n} \left[c_n(X_n, \mu_n(X_n)) + V_{n+1}^*(X_{n+1}) \right]
\end{aligned}
$$

$$
\begin{aligned}
V_n^*(X_n) &= \min_{(\mu_n, \pi^{n+1})} \mathbb{E}_{\pi^n, X_n} \left[c_n(X_n, \mu_n(X_n)) + \sum_{k=n+1}^{N} c_k(X_k, \mu_k(X_k)) \right] \\
&= \min_{\mu_n} \mathbb{E}_{\pi^n, X_n} \left(c_n(X_n, \mu_n(X_n)) + \min_{\pi^{n+1}} \mathbb{E}_{\pi^{n+1}, X_{n+1}} \left[\sum_{k=n+1}^{N} c_k(X_k, \mu_k(X_k)) \mid X_{n+1} \right] \right) \\
&= \min_{\mu_n} \mathbb{E}_{\pi^n, X_n} \left[c_n(X_n, \mu_n(X_n)) + V_{n+1}^*(X_{n+1}) \right] \\
&= \min_{a_n \in A_n(X_n)} \mathbb{E}_{\pi^n, X_n} \left[c_n(X_n, a_n) + V_{n+1}^*(X_{n+1}) \right] \\
&= \min_{a_n \in A_n(X_n)} \mathbb{E}_{\pi^n, X_n} \left[c_n(X_n, a_n) + V_{n+1}(X_{n+1}) \right] \;=\; V_n(X_n).
\end{aligned}
$$

Procédure ChaînageArrière

pour tout $x \in \mathcal{X}$, faire $V_{N+1}(x) \leftarrow 0$;
pour $n = N, \ldots, 0$ faire
    pour tout $x \in \mathcal{X}_n$ faire

$$V_n(x) \leftarrow \min_{a \in A_n(x)} \mathbb{E}\left[c_n(x,a) + V_{n+1}(X_{n+1})\right];$$
$$\mu_n^*(x) \leftarrow \arg\min_{a \in A_n(x)} \mathbb{E}\left[c_n(x,a) + V_{n+1}(X_{n+1})\right];$$

Principe d'optimalité de Bellman (cas probabiliste):

Si $\pi^* = (\mu_0^*, \ldots, \mu_N^*)$ est une politique optimale pour le problème initial et si $0 < n \leq N$, alors la politique tronquée $\pi_n^* = (\mu_n^*, \ldots, \mu_N^*)$ est une politique optimale pour le sous-problème "des décisions futures", qui consiste à minimiser

$$\mathbb{E}_{\mu_n,\ldots,\mu_N}\left[\sum_{k=n}^{N} c_k(X_k, a_k) \mid X_n\right].$$

par rapport à $\mu_n, \ldots, \mu_N$.

Hypothèses: Temps discret, modèle markovien, coûts additifs.

Si le coût n'est pas additif, le principe d'optimalité ne tient pas nécessairement.

Exemple: il ne tient pas si on veut minimiser

$$\mathbb{E}_{\mu_n,\ldots,\mu_N}\left[\max\left(c_n(X_n,a_n),\ \ldots,\ c_{N-1}(x_{N-1},a_{N-1}),\ c_N(x_N,a_N)\right) \mid X_n\right].$$

Si le coût n'est pas additif, le principe d'optimalité ne tient pas nécessairement.

Exemple: il ne tient pas si on veut minimiser

$$\mathbb{E}_{\mu_n,\ldots,\mu_N}\left[\max\left(c_n(X_n,a_n),\ \ldots,\ c_{N-1}(x_{N-1},a_{N-1}),\ c_N(x_N,a_N)\right) \mid X_n\right].$$

Le principe ne tient pas non plus (dans le sens qu'une politique optimale pour le problème entier n'est peut-être pas optimale pour le sous-problème) pour un sous-problème de la forme:

$$\text{minimiser} \quad \mathbb{E}_{\mu_n,\ldots,\mu_j}\left[\sum_{k=n}^{j} c_k(X_k,a_k) \mid X_n\right]$$

si $j < N$ et l'état X_j n'est pas déterminé, car il peut arriver que la politique optimale π^* amène des coûts un peu plus élevés pour les étapes n à j que la politique optimale pour le sous-problème, afin d'éviter un gros coût à l'étape N, par exemple.

Commande en boucle fermée: on prend chaque décision le plus tard possible, lorsqu'on a le maximum d'information.
Commande en boucle ouverte: on prend toutes les décisions $a_0, \ldots, a_N$ dès le départ.

La différence de coût espéré entre les deux est la valeur de l'information additionnelle. Cette différence peut être grande.

Dans le cas où tout est déterministe: pas de différence, car aucune information additionnelle à chaque étape.

Commande en boucle fermée: on prend chaque décision le plus tard possible, lorsqu'on a le maximum d'information.
Commande en boucle ouverte: on prend toutes les décisions $a_0, \ldots, a_N$ dès le départ.

La différence de coût espéré entre les deux est la valeur de l'information additionnelle. Cette différence peut être grande.

Dans le cas où tout est déterministe: pas de différence, car aucune information additionnelle à chaque étape.

Ce modèle de PDM possède de nombreuses généralisations:
— Introduction d'un facteur d'actualisation;
— Horizon infini;
— Revenu moyen par unité de temps sur horizon infini;
— Espaces d'états et de décisions infinis et non dénombrables;
— Évolution en temps continu;
— État partiellement observé (POMDP); Etc.

Exemple: Gestion d'un inventaire.

Monsieur D. Taillant vend des Zyx à Loinville. Les acheteurs arrivent au hasard. Au début de chaque mois, l'avion peut apporter une commande de Zyx. Soient:

x_k	=	Niveau des stocks au début du mois k, avant de commander (c'est l'état à l'étape k, en minuscule ici);
a_k	=	Nombre de Zyx commandés (et reçus) au début du mois k;
ω_k	=	Nombre de Zyx demandés par les clients durant le mois k. On suppose que les ω_k sont des variables aléatoires discrètes indépendantes;
$C + ca$	=	Coût d'une commande de a Zyx;
v	=	Prix de vente d'un Zyx (encaissé à la fin du mois);
B	=	Borne supérieure sur le niveau des stocks.
$r_k(x_k)$	=	Coût d'inventaire pour x_k Zyx au début du mois k;
$-g_N(x_N)$	=	Valeur de revente de x_N Zyx au début du mois N;

Posons:

$$V_k(x) \quad = \quad \text{coût espéré total pour les mois } k \text{ à } N, \text{ si } x_k = x \text{ et que l'on suit une politique optimale;}$$

Si on permet les inventaires négatifs ("backlogs"), on a

$$x_{k+1} = x_k + a_k - \omega_k.$$

Récurrence:

$$V_N(x) = g_N(x), \quad \text{pour } x \le B;$$

$$V_k(x) = \min_{0 \le a \le B-x} \left(r_k(x) + \mathbb{I}(a > 0)C + ca - v\mathbb{E}[\omega_k] + \sum_{i \ge 0} \mathbb{P}[\omega_k = i] V_{k+1}(x + a - i) \right), \; x \le B; \; k = N-1, \ldots, 0;$$

$$\mu_k^*(x) = \arg \min_{0 \le a \le B-x} \left(\mathbb{I}(a > 0)C + ca + \sum_{i \ge 0} \mathbb{P}[\omega_k = i] V_{k+1}(x + a - i) \right).$$

(On peut enlever $r_k(x) - v\mathbb{E}[\omega_k]$, car indépendant de a.)
(On peut aussi le sortir du min plus haut.)

$$
\begin{aligned}
V_N(x) &= g_N(x), \quad \text{pour } x \le B; \\
V_k(x) &= \min_{0\le a\le B-x} \Bigg(r_k(x) + \mathbb{I}(a>0)C + ca - v\mathbb{E}[\omega_k] \\
&\qquad + \sum_{i\ge 0} \mathbb{P}[\omega_k = i] V_{k+1}(x+a-i) \Bigg), \ x \le B; \ k = N-1, \ldots, 0; \\
W_k(x) &= V_k(x) - r_k(x) \quad \text{(éviter de recalculer la somme pour chaque } a\text{)} \\
&= \min \Bigg(-v\mathbb{E}[\omega_k] + \sum_{i\ge 0} \mathbb{P}[\omega_k = i] V_{k+1}(x-i), \quad \text{(cas } a = 0\text{)} \\
&\qquad C + c + W_k(x+1), \ \ldots, \ C + (B-x)c + W_k(B) \Bigg) \ \text{si } x \le B. \\
\mu_k^*(x) &= \arg \min_{0\le a\le B-x} \Bigg[\Bigg(-v\mathbb{E}[\omega_k] + \sum_{i\ge 0} \mathbb{P}[\omega_k = i] V_{k+1}(x-i) \Bigg) \mathbb{I}[a=0], \\
&\qquad (C + ca + W_k(x+a))\, \mathbb{I}[a>0] \Bigg].
\end{aligned}
$$

Dans le cas où $C = 0$, on peut simplifier les calculs davantage:

$$\begin{aligned} W_k(x) &\stackrel{\text{def}}{=} V_k(x) - r_k(x) \\ &= \min\left(-v\mathbb{E}[\omega_k] + \sum_{i\geq 0} \mathbb{P}[\omega_k = i] V_{k+1}(x-i),\ c + W_k(x+1)\right). \end{aligned}$$

Coûts de calcul: supposons que la somme sur i (valeurs possibles de ω_k) a T termes non négligeables. Les coûts de calcul sont
$O(NB^2T)$ pour la récurrence sur V_k;
$O(NB(B+T))$ pour la récurrence sur W_k;
$O(NBT)$ pour la cas simplifié où $C = 0$.

Si les inventaires négatifs ne sont pas permis, on a

$$x_{k+1} = \max(0,\, x_k + a_k - \omega_k)$$

et les équations de récurrence deviennent:

$$\begin{aligned}
V_N(x) &= g_N(x) \quad \text{pour } 0 \le x \le B; \\
V_k(x) &= \min_{0 \le a \le B-x} \Bigg(r_k(x) + \mathbb{I}(a > 0) C + ca \\
&\qquad + \sum_{i \ge 0} P[\omega_k = i][-v \min(i, x+a) + V_{k+1}(\max(0,\, x + a - i))] \Bigg) \\
&\qquad \text{pour } 0 \le x \le B;\ \ k = N-1, \ldots, 0; \\
W_k(x) &= \min \Bigg(\sum_{i \ge 0} P[\omega_k = i][-v \min(i, x) + V_{k+1}(\max(0,\, x - i))], \\
&\qquad C + c + W_k(x+1),\ \ldots,\ C + (B-x)c + W_k(B) \Bigg) \qquad \text{si } x < B.
\end{aligned}$$

Dans le cas où $C = 0$:

$$W_k(x) = \min \left(\sum_{i \geq 0} \mathbb{P}[\omega_k = i][-v \min(i, x) + V_{k+1}(\max(0, x - i))], \right.$$
$$\left. c + W_k(x + 1) \right).$$

Exemple: taille d'un lot de pièces à fabriquer.

La compagnie Essai-Erreur doit fabriquer M exemplaires d'une pièce pour remplir une commande. Les critères de qualité sont très élevés. La compagnie estime que chaque pièce produite sera acceptable avec probabilité p. Les pièces sont fabriquées par lots ("batches"). Pour fabriquer un lot de a pièces, il en coûte $C + ca$. Il nous faut M pièces acceptables. Dans un lot de taille a, le nombre Y de pièces acceptables est une variable aléatoire binomiale:

$$\mathbb{P}[Y = y] = \binom{a}{y} p^y (1-p)^{a-y}, \qquad y = 0, \ldots, a.$$

En pratique, on pourra fabriquer un lot de taille $> M$, car il y aura probablement des pièces défectueuses (des rejets).

Si le nombre de pièces acceptables est quand même inférieur à M, on devra produire un second lot, peut-être même un troisième, etc.

Supposons qu'on a assez de temps pour produire N lots.

Si on n'a pas toutes les pièces requises après N lots, on doit payer une énorme pénalité K.

x_k = Nombre de pièces encore requises avant de produire le lot $k+1$;
a_k = Taille du lot $k+1$;
y_k = Nb de pièces acceptables dans le lot $k+1$;
$V_k(x)$ = Coût espéré minimal à partir de maintenant, si on a k lots de produits et qu'il manque encore x pièces.

x_k = Nombre de pièces encore requises avant de produire le lot $k+1$;
a_k = Taille du lot $k+1$;
y_k = Nb de pièces acceptables dans le lot $k+1$;
$V_k(x)$ = Coût espéré minimal à partir de maintenant, si on a k lots de produits et qu'il manque encore x pièces.

On cherche le coût total espéré $V_0(M)$ et une politique optimale.
Pour tout k et $x \leq 0$, on a $V_k(x) = 0$. Pour $x > 0$:

$$
\begin{aligned}
V_N(x) &= K; \\
V_k(x) &= \min_{a \geq x} \left(C + ca + \sum_{y=0}^{a} \binom{a}{y} p^y (1-p)^{a-y} V_{k+1}(x-y) \right) \\
\mu_k^*(x) &= \arg\min_{a \geq x} \left(\quad '' \quad \right).
\end{aligned}
$$

Exemple numérique.

$M = 1$, $N = 4$, $p = 1/2$, $C = 3$, $c = 1$, $K = 16$.
On obtient alors:

$$\begin{aligned} V_4(1) &= 16; \\ V_n(1) &= \min_{a \geq 1} \left(3 + a + \binom{a}{0} p^0 (1-p)^{a-0} V_{n+1}(1) \right) \\ &= \min_{a \geq 1} \left(3 + a + (1/2)^a V_{n+1}(1) \right) \end{aligned}$$

Exemple numérique.

$M = 1$, $N = 4$, $p = 1/2$, $C = 3$, $c = 1$, $K = 16$.
On obtient alors:

$$\begin{aligned} V_4(1) &= 16; \\ V_n(1) &= \min_{a \geq 1} \left(3 + a + \binom{a}{0} p^0 (1-p)^{a-0} V_{n+1}(1) \right) \\ &= \min_{a \geq 1} \left(3 + a + (1/2)^a V_{n+1}(1) \right) \end{aligned}$$

$$\begin{aligned} V_3(0) &= 0 \\ V_3(1) &= \min_{a \geq 1} \left(3 + a + 16/2^a \right) = \min(4+8,\ 5+4,\ 6+2,\ 7+1,\ \ldots) = 8 \text{ (avec } a = 3 \text{ ou } 4) \\ V_2(1) &= \cdots \\ V_1(1) &= \cdots \\ V_0(1) &= \cdots \end{aligned}$$

Exemple: Commande d'une file d'attente finie.

On a une file d'attente avec un seul serveur, avec de la place pour n clients au maximum dans le système, qui évolue en temps discret.

Exemple: Commande d'une file d'attente finie.

On a une file d'attente avec un seul serveur, avec de la place pour n clients au maximum dans le système, qui évolue en temps discret.

Le serveur a 2 vitesses: rapide et lent.
On peut choisir la vitesse au début de chaque période.
Pour une période en mode rapide [lent], le coût du serveur est c_f $[c_s]$,
et si le système n'est pas vide,
on sert 1 client avec probabilité q_f $[q_s]$ et 0 clients avec probabilité $1 - q_f$ $[1 - q_s]$.
On doit payer $r(i)$ à chaque période où il y a i clients dans le système au début de la période.

Durant chaque période, $\mathbb{P}[m \text{ clients arrivent}] = p_m, \quad m \geq 0$.
Ces m clients sont dans la file au début de la période suivante (mais pas plus de n au total dans le système).

État x_k: nombre i de clients dans le système à l'étape k.
L'espace des décisions est $\mathcal{A} = \{\text{rapide, lent}\}$.
Soit $\xi_k \in \{0, 1\}$ le nombre de clients servis à la période k.

$$
\begin{aligned}
V_N(i) &= r(i), \quad \text{pour } 0 \le i \le n; \\
V_k(0) &= r(0) + c_s + W_k(0); \qquad (\text{ici } \xi_k = 0) \\
V_k(i) &= r(i) + \min[c_f + q_f W_k(i-1) + (1-q_f) W_k(i), \\
&\qquad\qquad c_s + q_s W_k(i-1) + (1-q_s) W_k(i)] \\
&\qquad \text{pour } 0 \le k \le N-1,\ 1 \le i \le n,
\end{aligned}
$$

où

$$
\begin{aligned}
W_k(i) &= \mathbb{E}\left[V_{k+1}(x_{k+1}) \mid x_k - \xi_k = i\right] \\
&= \sum_{m=0}^{n-i-1} p_m V_{k+1}(i+m) + V_{k+1}(n) \sum_{m=n-i}^{\infty} p_m.
\end{aligned}
$$

Exemple: choix du niveau de risque à chaque étape.

Un match est constitué d'une suite d'étapes.
Décisions: à chaque étape, le joueur 1 peut adopter une stratégie prudente (conservatrice) ou agressive (risquée).
Stratégie prudente [agressive]: on marque i points de plus que l'adversaire avec probabilité p_i [q_i], disons pour $-b \leq i \leq b$.
La variance de la loi des q_i est plus grande que celle des p_i.

On suppose que le joueur 2 joue toujours de la même façon.
Note: si le joueur 2 optimisait aussi sa stratégie: théorie des jeux. Plus compliqué.

Exemple: choix du niveau de risque à chaque étape.

Un match est constitué d'une suite d'étapes.
Décisions: à chaque étape, le joueur 1 peut adopter une stratégie prudente (conservatrice) ou agressive (risquée).
Stratégie prudente [agressive]: on marque i points de plus que l'adversaire avec probabilité p_i [q_i], disons pour $-b \leq i \leq b$.
La variance de la loi des q_i est plus grande que celle des p_i.

On suppose que le joueur 2 joue toujours de la même façon.
Note: si le joueur 2 optimisait aussi sa stratégie: théorie des jeux. Plus compliqué.

Jeu de type A: Celui ou celle ayant le plus de points après N étapes gagne; en cas d'égalité on ajoute des étapes jusqu'à ce que l'un des joueurs devance l'autre.

Jeu de type B: Le premier joueur qui devance l'autre par au moins K points gagne le match.

État x: nombre de points d'avance du joueur 1 sur le joueur 2.
$V_k(x)$ = probabilité que le joueur 1 gagne s'il a x points d'avance sur le joueur 2 après k étapes de jeu et s'il prend ses décisions de façon optimale, i.e., pour maximiser sa probabilité de gain.

Pour un jeu de type B, $V_k \equiv V$ ne dépend pas de k et on a:

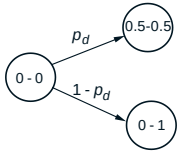
$$V(x) = \begin{cases} 1 & \text{pour } x \geq K; \\ 0 & \text{pour } x \leq -K; \\ \max\left(\sum_{i=-b}^{b} p_i V(x+i), \ \sum_{i=-b}^{b} q_i V(x+i) \right) & \\ \qquad\qquad \text{pour } -K < x < K. & \end{cases}$$

Applications possibles:
— Une série de la coupe Stanley ($N = 7$).
— Un match de hockey divisé en blocs (étapes) de 5 secondes.
— Une course cycliste par étapes.
— Une stratégie d'investissement en finance: fonction objectif différente.
— Etc.

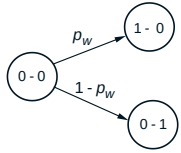
Applications possibles:
— Une série de la coupe Stanley ($N = 7$).
— Un match de hockey divisé en blocs (étapes) de 5 secondes.
— Une course cycliste par étapes.
— Une stratégie d'investissement en finance: fonction objectif différente.
— Etc.

Match d'échecs de N parties.
À chaque partie, le joueur 1 peut gagner ($i = 1$), perdre ($i = -1$), ou annuler ($i = 0$). Après N parties, si un joueur devance l'autre, il gagne le match, tandis que si le score est égal, on continue et le premier joueur qui gagne une partie gagne le match.

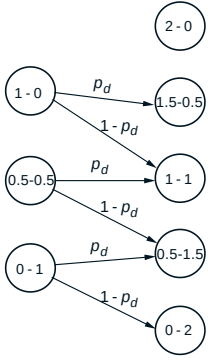
Supposons que $p_1 = 0$ et $p_{-1} = 1 - p_0$ (en mode prudent, on peut seulement annuler ou perdre) et que $q_0 = 0$ et $q_{-1} = 1 - q_1$ (en mode agressif, on peut gagner ou perdre).

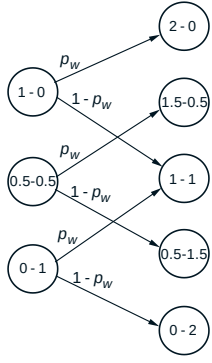


1st Game / Timid Play



1st Game / Bold Play





On a ici

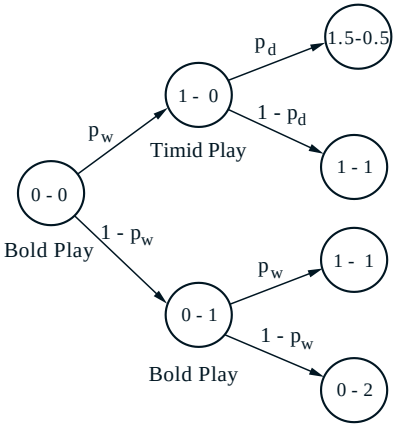
$$
\begin{aligned}
V_k(x) &= V_N(x) \quad \text{pour } k > N; \\
V_N(x) &= \begin{cases} 1 & \text{si } x > 0; \\ q_1 & \text{si } x = 0; \\ 0 & \text{si } x < 0; \end{cases} \\
V_{N-1}(x) &= \begin{cases} 1 & \text{si } x > 1; & \\ p_0 + (1-p_0)q_1 & \text{si } x = 1; & \text{(jeu prudent);} \\ q_1 & \text{si } x = 0; & \text{(jeu agressif);} \\ q_1^2 & \text{si } x = -1; & \text{(jeu agressif);} \\ 0 & \text{si } x < -1; & \end{cases} \\
V_k(x) &= \max[p_0 V_{k+1}(x) + (1-p_0) V_{k+1}(x-1), \\
&\qquad q_1 V_{k+1}(x+1) + (1-q_1) V_{k+1}(x-1)] \\
&\qquad \text{pour } 0 \le k < N \text{ et } -k \le x \le k.
\end{aligned}
$$

Si $N = 2$, au début du match on a

$$\begin{aligned} V_0(0) &= \max\left[p_0 V_1(0) + (1-p_0)V_1(-1),\ q_1 V_1(1) + (1-q_1)V_1(-1)\right] \\ &= \max\left[p_0 q_1 + (1-p_0)q_1^2,\ p_0 q_1 + (1-p_0)q_1^2 + (1-q_1)q_1^2\right] \\ &= p_0 q_1 + (1-p_0)q_1^2 + (1-q_1)q_1^2 \qquad \text{(jeu agressif)}. \end{aligned}$$

La politique optimale si $N = 2$ est donc:

jouer prudent si on est en avance, jouer agressif sinon.


0 - 0
Bold Play
p_w
Timid Play
1 - 0
p_d
1.5-0.5
$1 - p_d$
1 - 1
$1 - p_w$
0 - 1
p_w
1 - 1
Bold Play
$1 - p_w$
0 - 2

Intéressant: On pourrait croire que $q_1 < 1/2$ implique que $V_0(0) < 1/2$, mais non. Notre probabilité de gagner le match peut dépasser $1/2$ même si notre probabilité de gagner une partie est toujours $< 1/2$.
Par exemple, si $q_1 = 0.45$ et $p_0 = 0.90$, alors $V_0(0) \approx 0.537$.

Intéressant: On pourrait croire que $q_1 < 1/2$ implique que $V_0(0) < 1/2$, mais non. Notre probabilité de gagner le match peut dépasser $1/2$ même si notre probabilité de gagner une partie est toujours $< 1/2$.
Par exemple, si $q_1 = 0.45$ et $p_0 = 0.90$, alors $V_0(0) \approx 0.537$.

Explication: Le joueur 1 choisit son style de jeu à chaque étape et peut adapter sa stratégie au pointage, ce qui lui donne un avantage sur le joueur 2, qui n'a aucun choix.
Le joueur 1 utilise une politique en boucle fermée. S'il était forcé de choisir toutes ses décisions à l'avance (politique en boucle ouverte), on aurait:

décisions	prob. de gagner
prudent, prudent	$p_0^2 q_1$
prudent, agressif	$p_0 q_1 + (1-p_0) q_1^2$
agressif, prudent	$p_0 q_1 + (1-p_0) q_1^2$
agressif, agressif	$q_1^2 + 2(1-q_1) q_1^2$

En supposant que $p_0 \geq 2q_1$, la meilleure politique en boucle ouverte est de jouer prudent pour une étape et agressif pour l'autre.
La prob. de gagner est alors

$$\tilde{V}_0(0) = V_0(0) - (1 - q_1)q_1^2.$$

Cette différence de $(1 - q_1)q_1^2$ est la valeur de l'information.

En supposant que $p_0 \geq 2q_1$, la meilleure politique en boucle ouverte est de jouer prudent pour une étape et agressif pour l'autre.
La prob. de gagner est alors

$$\tilde{V}_0(0) = V_0(0) - (1 - q_1)q_1^2.$$

Cette différence de $(1 - q_1)q_1^2$ est la valeur de l'information.

Par exemple, si $q_1 = 0.45$ et $p_0 = 0.90$, alors $(1 - q_1)q_1^2 \approx 0.1114$ et la probabilité de gain avec la meilleure politique en boucle ouverte est ≈ 0.425.

Conclusion: **fixer toutes nos décisions à l'avance est une bien mauvaise idée!**

Pour la suite: IFT6521, Programmation Dynamique