

# Programmation dynamique

Fabian Bastin

IFT-6521 – Hiver 2011

# Contenu de l'introduction

- 1 Modalités pratiques.
- 2 Qu'est-ce que la programmation dynamique (PD)?
- 3 Exemples simples.
- 4 Modèle de base: versions déterministe et stochastique.
- 5 Principe d'optimalité et algorithme de la PD.
- 6 Contrôle en boucle ouverte vs boucle fermée, et valeur de l'information.
- 7 Reformulations pour se ramener au modèle de base.
- 8 Fonction d'utilité et mesure de risque.

Disponible sur rendez-vous à mon bureau (3367).

courriel: `bastin@iro.umontreal.ca`

Les transparents servant de support au cours sont disponibles à l'adresse

`http://www.iro.umontreal.ca/~bastin/Cours/IFT6521` et sont susceptibles d'être mis à jour.

Les notes sont inspirés du cours donné préalablement par Pierre L'Ecuyer.

Les figures ont été fournies par D. P. Bertsekas.

Les modalités d'évaluation précises seront discutées en classe, mais elle se répartira comme suit:

- 4 devoirs, chacun comptant pour 12.5% de la note finale;
- un examen intra, comptant pour 20%;
- un projet final, comptant pour 30%.

# Qu'est-ce que la programmation dynamique?

Processus de décision séquentiel (PDS):

- Suite de **décisions** à prendre, avec un **objectif à optimiser**.
- **Temps discret**: À chaque étape, on prend une décision, selon l'information disponible.
- Liens (interactions) entre les décisions.
- En général, on peut recevoir de l'**information** additionnelle à chaque étape, et il y a des aléas dans l'évolution entre les prises de décision.
- **Temps continu**: la suite de décisions est remplacée par un contrôle, qui est une fonction du temps.

**Exemples**: gestion d'un inventaire; conduite automobile; pilotage d'un avion (pilote automatique), d'un robot, etc.; entretien préventif; gestion d'un fond d'investissement; option financière de type américaine; partie de tennis ou de football; jeu d'échecs; etc.  
La vie en général!

## Programmation dynamique:

Ensemble d'outils mathématiques et algorithmiques pour étudier les processus de décision séquentiels et calculer éventuellement des stratégies optimales (exactes ou approximatives).

Une **politique** (ou **stratégie**) est une règle de prise de décisions qui, pour chaque situation possible (état du système), nous dit quelle décision (ou action) prendre dans le but d'optimiser une fonction objectif globale.

Souvent, la fonction objectif est une espérance mathématique. Parfois, on pourra caractériser la **politique optimale** par des théorèmes (théorie); souvent, on pourra la calculer, ou en calculer une approximation; dans certains cas la résolution sera trop difficile.

## Exemple: modèle d'inventaire simple pour un seul produit

- $x_k$  = stock en inventaire au début de la période  $k$ .
- $u_k$  = quantité commandée au début de la période  $k$ , et obtenue immédiatement. Contrainte:  $u_k \geq 0$ .
- $w_k$  = demande durant la période  $k$ .  
Supposons que les  $w_k$  sont des v.a.'s indépendantes.
- Évolution:  $x_{k+1} = x_k + u_k - w_k = f(x_k, u_k, w_k)$ .  
Note:  $x_k$  peut être négatif (la demande est en attente).
- $r(x_k)$  = coût associé à l'inventaire  $x_k$  au début de la période  $k$ .  
Coût de stockage si  $x_k > 0$ ; coût de pénurie si  $x_k < 0$ .
- $cu_k$  = coût d'approvisionnement pour la période  $k$ .
- $R(x_N)$  = coût pour l'inventaire terminal.

Coût espéré total, à minimiser:

$$\mathbb{E} \left[ R(x_N) + \sum_{k=0}^{N-1} (r(x_k) + cu_k) \right].$$

$$\mathbb{E} \left[ R(x_N) + \sum_{k=0}^{N-1} (r(x_k) + cu_k) \right].$$

Contrôle en **boucle ouverte**: on choisit  $u_0, \dots, u_{N-1}$  à l'avance.

Contrôle en **boucle fermée**: on choisit  $u_k$  après avoir observé  $x_k$ .

Dans le second cas, on cherche une **politique**  $\pi = (\mu_0, \mu_1, \dots, \mu_{N-1})$  qui nous indique comment prendre les décisions:  $u_k = \mu_k(x_k)$ .

Le coût espéré associé à la politique  $\pi$  est

$$J_\pi(x_0) = \mathbb{E} \left[ R(x_N) + \sum_{k=0}^{N-1} (r(x_k) + c\mu_k(x_k)) \right]$$

Comment trouver  $\pi = \pi^*$  qui minimise  $J_\pi(x_0)$ ? Sous certaines hypothèses raisonnables, on peut montrer que  $\pi^*$  a la forme

$$\mu_k(x_k) = \max(0, S_k - x_k).$$

Dans ce cas, il suffit d'optimiser les  $S_k$  (plus simple).

$$\begin{array}{ll} x_0 & \rightarrow u_0 = \mu_0(x_0), & w_0 & \rightarrow x_1 = x_0 + u_0 - w_0 \\ x_1 & \rightarrow u_1 = \mu_1(x_1), & w_1 & \rightarrow x_2 = x_1 + u_1 - w_1 \\ x_2 & \rightarrow u_2 = \mu_2(x_2), & w_2 & \rightarrow x_3 = x_2 + u_2 - w_2 \\ x_3 & \dots & & \end{array}$$

## Principaux ingrédients d'un PDS (usuel):

Temps discret; aléas indépendants; contraintes sur les décisions et politiques; coût additif; on cherche à optimiser une politique.

# Un modèle de PDS déterministe

À l'étape  $k$ , le système est dans un état  $x_k \in X_k$ .

Un décideur observe  $x_k$  et choisit une décision (action)  $u_k \in U_k(x_k)$ .

Il paye un coût  $g_k(x_k, u_k)$  pour cette étape, puis le système transite dans un nouvel état  $x_{k+1} = f_k(x_k, u_k)$  à l'étape  $k + 1$ .

À l'étape  $k$ , on a donc:

- $X_k$  = espace d'états;
- $U_k(x)$  = ensemble des décisions admissibles dans l'état  $x$ ;
- $g_k$  = fonction de coût;
- $f_k$  = fonction de transition;
- $x_k$  = état du système à l'étape  $k$ ;
- $u_k$  = décision prise à l'étape  $k$ .

On veut minimiser la somme des coûts de l'étape 0 à l'étape  $N$ :

$$\min \quad g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k)$$

$$\text{s.l.c.} \quad u_k \in U_k(x_k) \text{ et } x_{k+1} = f_k(x_k, u_k), \quad k = 0, \dots, N - 1; x_0 \text{ fixé.}$$

$$\begin{aligned} \min \quad & g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k) \\ \text{s.l.c.} \quad & u_k \in U_k(x_k) \text{ et } x_{k+1} = f_k(x_k, u_k), \quad k = 0, \dots, N-1 \\ & x_0 \text{ fixé.} \end{aligned}$$

On peut illustrer cela par un arbre de décision.

Les fonctions  $g_k$  et  $f_k$  peuvent être non linéaires et compliquées. Si  $g_k(x_k, u_k)$  est un **revenu** au lieu d'un coût, on remplace "min" par "max" (ou "inf" par "sup").

Une **politique (ou stratégie) admissible** est une suite de fonctions  $\pi = (\mu_0, \dots, \mu_{N-1})$  tel que  $\mu_k : X_k \rightarrow U_k$  et  $\mu_k(x) \in U_k(x)$  pour tout  $x \in X_k$ ,  $0 \leq k \leq N-1$ . La décision à l'étape  $k$  est  $\mu_k(x_k)$ .

Pour une stratégie  $\pi$  fixée, posons

$J_{\pi,k}(x)$  = coût total pour les étapes  $k$  à  $N$  si on est dans l'état  $x$  à l'étape  $k$  et si on utilise la politique  $\pi$

$$= g_N(x_N) + \sum_{n=k}^{N-1} g_n(x_n, u_n)$$

s.l.c.  $u_n = \mu_n(x_n)$  et  $x_{n+1} = f_n(x_n, u_n)$ ,  $n = k, \dots, N-1$   
 $x_k = x$  (fixé).

Ces valeurs satisfont les **équations de récurrence** (ou **équations fonctionnelles**) suivantes:

$$J_{\pi,N}(x) = g_N(x) \quad \forall x \in X_N$$

$$J_{\pi,k}(x) = g_k(x, \mu_k(x)) + J_{\pi,k+1}(f_k(x, \mu_k(x))) \quad \text{pour tout } x \in X_k, \\ \text{pour } k = N-1, N-2, \dots, 1, 0.$$

Sans fixer la politique, pour  $0 \leq k \leq N$  et  $x \in X_k$ , soient

$J_k(x)$  = coût optimal pour les étapes  $k$  à  $N$  si on est dans l'état  $x$  à l'étape  $k$

$$= \min_{u_k, \dots, u_{N-1}} \left( g_N(x_N) + \sum_{n=k}^{N-1} g_n(x_n, u_n) \right)$$

s.l.c.  $u_n \in U_n(x_n)$  et  $x_{n+1} = f_n(x_n, u_n)$ ,  $n = k, \dots, N-1$   
 $x_k = x$  (fixé).

Une **politique admissible**  $\pi^* = (\mu_0^*, \dots, \mu_{N-1}^*)$  telle que

$$J_{\pi^*,0}(x) = J_0(x) \stackrel{\text{def}}{=} J(x) \quad \text{pour tout } x$$

s'appelle une **politique optimale**.

On a les **équations de récurrence** (équations de Bellman):

$$\begin{aligned} J_N(x) &= g_N(x) \quad \text{pour tout } x \in X_N \\ J_k(x) &= \min_{u \in U_k(x)} \{g_k(x, u) + J_{k+1}(f_k(x, u))\} \quad \text{pour tout } x \in X_k, \\ &\quad \text{pour } k = N - 1, N - 2, \dots, 1, 0. \end{aligned}$$

On cherche  $J_0(x_0)$  pour  $x_0$  fixé (problème de la valeur initiale).

On peut résoudre par fixation itérative, **chaînage arrière**:

Calculer  $J_{N-1}(x)$  pour tout  $x \in X_{N-1}$ ,

puis  $J_{N-2}(x)$  pour tout  $x \in X_{N-2}$ , etc.

Comment retrouver la solution optimale ?

Durant les calculs, on mémorise, pour chaque  $k$  et  $x \in X_k$ ,

$$\mu_k^*(x) = \arg \min_{u \in U_k(x)} \{g_k(x, u) + J_{k+1}(f_k(x, u))\},$$

qui est la **décision optimale** à prendre dans l'état  $x$  à l'étape  $k$ .

Note: La notation pourra varier un peu selon les problèmes.

# Procédure Chaînage Arrière

pour tout  $x \in X_N$ ,  $J_N(x) \leftarrow g_N(x)$ ;  
pour  $k = N - 1, \dots, 0$  faire  
  pour tout  $x \in X_k$  faire

$$J_k(x) \leftarrow \min_{u \in U_k(x)} \{g_k(x, u) + J_{k+1}(f_k(x, u))\};$$
$$\mu_k^*(x) \leftarrow \arg \min_{u \in U_k(x)} \{g_k(x, u) + J_{k+1}(f_k(x, u))\};$$

# Principe d'optimalité de Bellman

Si  $\pi^* = (\mu_0^*, \dots, \mu_{N-1}^*)$  est une politique optimale pour le problème initial et si  $0 \leq i \leq j \leq N - 1$ , alors la politique tronquée  $\pi_{i,j}^* = (\mu_i^*, \dots, \mu_j^*)$  est une politique optimale pour le sous-problème qui consiste à minimiser

$$\sum_{k=i}^j g_k(x_k, u_k)$$

pour  $x_i$  et  $x_j$  fixés (en posant  $g_N(x_N, u_N) \equiv g_N(x_N)$ ).

**Hypothèses importantes:** Temps discret et coûts additifs.

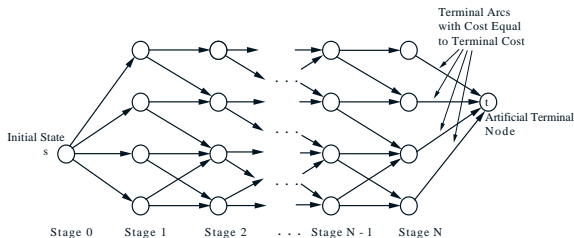
Si le coût n'est pas additif, le principe d'optimalité ne tient pas nécessairement. Exemple: Si on remplace la somme par le maximum, i.e., on veut minimiser

$$\max [g_k(x_k, u_k), \dots, g_{N-1}(x_{N-1}, u_{N-1}), g_N(x_N)].$$

# Plus court chemin dans un réseau.

Si  $X_k$  et chaque  $U_k(x)$  sont des ensembles **finis**, le problème se ramène à un problème de plus court chemin dans un réseau.

Les **noeuds** du réseau sont tous les couples  $(k, x_k)$  possibles, pour  $0 \leq k \leq N$  et  $x_k \in X_k$ , et les **arcs** partant d'un noeud correspondent aux décisions  $u_k$  que l'on peut prendre à partir de ce noeud.



Ce réseau est sans cycle et ordonné topologiquement.

Tout algorithme pour trouver le plus court chemin dans un tel réseau s'applique pour calculer une politique optimale pour ce PDS.

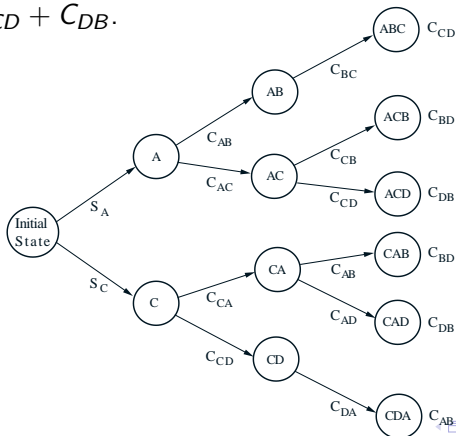
**Important:** Il ne faut pas que le nombre d'états devienne trop grand!

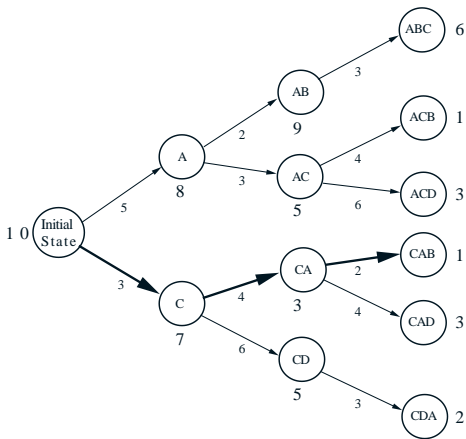
# Exemple: ordonnancement de tâches

(DPOC, pages 7 et 19) On a 4 opérations,  $\{A, B, C, D\}$ , à effectuer sur une machine. On doit faire A avant B, et C avant D.

Il y a un coût  $S_m$  si on débute avec l'opération  $m$ , puis un coût  $C_{mn}$  pour passer de l'opération  $m$  à l'opération  $n$ .

Par exemple, pour la séquence ACDB, le coût total est  $S_A + C_{AC} + C_{CD} + C_{DB}$ .





Une seule tâche à accomplir: aucune décision à prendre.

Par le principe d'optimalité, s'il ne reste que deux tâches, l'ordonnancement de ces deux tâches doit minimiser la coût associé.

Même chose s'il ne reste que trois tâches.

## Exemple: allocation d'équipes médicales

On dispose de 5 équipes médicales à allouer à 3 pays en développement. Plus on alloue d'équipes médicales à un pays, plus on augmente son nombre d'années-personnes de vie espérée. Cette augmentation n'est pas linéaire.

Milliers d'années-personnes de vie espérée additionnelle			
Nombre d'équipes allouées, $u_k$	Pays 0	Pays 1	Pays 2
0	0	0	0
1	45	20	50
2	70	45	70
3	90	75	80
4	105	110	100
5	120	150	130

Soit  $u_k$  (un entier) le nombre d'équipes allouées au pays  $k$ .  
On veut maximiser le nombre total d'années-personnes de vie espérée additionnelle pour les 3 pays.

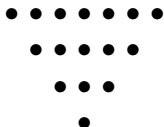
PDS:  $N = 3$ . À l'étape  $k \leq 2$ , il reste  $u_k, \dots, u_2$  à fixer,  $x_k$  équipes sont encore disponibles, et  $J_k(x_k)$  est le revenu optimal pour les pays  $k, \dots, 2$ .

Les  $g_k(x_k, u_k) = g_k(u_k)$  sont les valeurs données dans le tableau.  
(Ici, elles ne dépendent que de  $u_k$  et pas de  $x_k$ .)

Le noeud  $(k, x)$  correspond à l'état  $x_k$ : il reste  $x$  équipes pour les pays  $k, \dots, 2$ . Le chaînage arrière revient à rechercher le plus long chemin de  $(0, 5)$  à  $(3, 0)$ .

# Exemple: jeu de Nim

Règles du jeu: on place des pièces dans 4 rangées comme ci-dessous.



Deux joueurs jouent à tour de rôle.

Lorsque c'est son tour, un joueur enlève un nombre arbitraire de jetons dans un même rangée. Il doit en enlever au moins 1. Celui qui enlève le dernier jeton perd.

Quelle est la stratégie optimale et comment la calculer?

# Jeu de Nim (suite)

**Etat** du jeu  $x = (x_1, x_2, x_3, x_4)$ , où  $x_i$  est le nombre de jetons dans la rangée  $i$ .

Le numéro d'étape n'a pas d'importance ici.

Posons

$$g(x) = \begin{cases} 1 & \text{si } x_1 = \dots = x_4 = 0 \text{ (situation gagnante),} \\ 0 & \text{sinon (situation perdante ou intermédiaire).} \end{cases}$$

Chaque joueur applique la même stratégie.

**Idée:**  $J(x) = 1$  [=0] si on est dans une position gagnante [perdante].

On essaie de mettre l'adversaire dans une position perdante.

# Jeu de Nim (suite)

Posons

$$\begin{aligned} J(x) &= J(x_1, \dots, x_4) \\ &= \begin{cases} g(x_1, \dots, x_4) = 1 & \text{si } x_1 = x_2 = x_3 = x_4 = 0, \\ 1 - \min_{u \in U(x)} J(x_1 - u_1, \dots, x_4 - u_4) & \text{sinon,} \end{cases} \end{aligned}$$

où  $U(x) = \{u = (u_1, \dots, u_4) : 0 \leq u_i \leq x_i \text{ pour tout } i \text{ et exactement un seul des } u_i \text{ est positif}\}$ .

Ainsi,  $J(x)$  vaut 1 en cas de situation gagnante, sinon, on joue le coup qui minimise  $J$ , vu qu'après le coup, c'est à l'adversaire de jouer. Remarque:

$$\arg \min_{u \in U(x)} 1 - J(x_1 - u_1, \dots, x_4 - u_4) \neq \arg \max_{u \in U(x)} J(x_1 - u_1, \dots, x_4 - u_4).$$

On peut calculer un tableau qui pour chaque  $x$  nous donne  $J(x)$  et un coup optimal à jouer  $\mu^*(x)$ . On peut simplifier les calculs en **agrégant les états**; e.g., ne considérer que les états  $x$  où  $x_1 \geq x_2 \geq x_3 \geq x_4$ . On a

$$J(x_1, \dots, x_4) = \begin{cases} 1 & \text{si } x_1 = x_2 = x_3 = x_4 = 0, \\ 0 & \text{si } x_1 = 1 \text{ et } x_2 = x_3 = x_4 = 0, \\ 1 & \text{si } x_1 = 2 \text{ et } x_2 = x_3 = x_4 = 0, \\ \text{etc.} \end{cases}$$

état $x$	décision	prochain état	valeur $J(x)$	condition
(0 0 0 0)	—	—	1	
(1 0 0 0)	(1 0 0 0)	(0 0 0 0)	0	
( $x_1$ 0 0 0)	( $x_1 - 1$ 0 0 0)	(1 0 0 0)	1	$x_1 \geq 2$
( $x_1$ 1 0 0)	( $x_1$ 0 0 0)	(1 0 0 0)	1	$x_1 \geq 1$
(2 2 0 0)	(0 2 0 0)	(2 0 0 0)	0	
( $x_1$ 2 0 0)	( $x_1 - 2$ 0 0 0)	(2 2 0 0)	1	$x_1 \geq 3$
⋮	⋮	⋮	⋮	

## Exemple: une fonction objectif sous forme produit.

$N$  équipes de chercheurs travaillent (indépendamment) sur le même problème d'ingénierie.

On voudrait qu'au moins une équipe résolve le problème d'ici 2 mois. On dispose de  $b$  nouveaux brillants chercheurs à leur affecter. Soit  $p_k(u_k)$  la probabilité que l'équipe  $k$  échoue si on lui alloue  $u_k$  chercheurs additionnels.

La probabilité que toutes les équipes échouent est  $p_1(u_1) \times \cdots \times p_N(u_N)$ . On veut minimiser cette probabilité.

Exemple de Hillier et Lieberman:  $N = 3$ ,  $b = 2$ .

Probabilité d'échouer $p_k(u)$			
Nb. $u$ de chercheurs additionnels	Équipe 1	Équipe 2	Équipe 3
0	0.40	0.60	0.80
1	0.20	0.40	0.50
2	0.15	0.20	0.30

$J_k(x)$  = Prob. que toutes les équipes  $k, \dots, N$  échouent, si on leur alloue  $x$  chercheurs additionnels de façon optimale.

On peut formuler le problème sous forme déterministe classique:

$$\min_x \prod_{i=1}^3 p_i(x_i) = p_1(x_1)p_2(x_2)p_3(x_3),$$

$$\text{t.q. } \sum_{i=1}^3 x_i = 2, \text{ et } x_i \geq 0, \quad i = 1, 2, 3.$$

Notons qu'on peut reconstruire une forme linéaire du problème, en utilisant l'opérateur logarithmique:

$$\min \sum_{i=1}^3 \ln p_i(x_i),$$
$$\text{t.q. } \sum_{i=1}^3 x_i = 2, \text{ et } x_i \geq 0, \quad i = 1, 2, 3.$$

Réécrivons le problème en considérant une approche par programmation dynamique séquentielle.

On a:

$$J_N(x) = p_N(x) \text{ pour tout } x;$$
$$J_k(x) = \min_{u \in \{0, \dots, x\}} p_k(u) \times J_{k+1}(x - u),$$
$$x = 0, \dots, b; \quad k = N - 1, \dots, 1.$$

# Un modèle de PDS probabiliste

## Processus de décision markovien sur horizon fini

À l'étape  $k$ , on observe l'état  $x_k$  et prend une décision  $u_k \in U_k(x_k)$ . Puis une variable aléatoire  $\omega_k$  est générée selon une loi de probabilité  $\mathbb{P}_k(\cdot \mid x_k, u_k)$  qui peut dépendre de  $(k, x_k, u_k)$ . On suppose que les valeurs précédentes  $\{(x_n, u_n, \omega_n), n < k\}$  ne nous donnent pas d'information additionnelle sur cette loi  $\mathbb{P}_k$  lorsqu'on connaît  $(k, x_k, u_k)$ .

On observe  $\omega_k$ , on paye un coût  $g_k(x_k, u_k, \omega_k)$ , et l'état à la prochaine étape est  $x_{k+1} = f_k(x_k, u_k, \omega_k)$ . Coût total (aléatoire):

$$g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k, \omega_k).$$

Une politique admissible est une suite de fonctions

$\pi = (\mu_0, \dots, \mu_{N-1})$  telle que  $\mu_k : X_k \rightarrow U_k$  et  $\mu_k(x) \in U_k(x)$  pour tout  $x \in X_k$ ,  $0 \leq k \leq N-1$  (+ détails techniques:  $\mu_k$  doit être une fonction mesurable, etc.).

À l'étape  $k$ , on a :

$X_k$  = espace d'états;

$U_k(x)$  = ensemble des décisions admissibles dans l'état  $x$ ;

$D_k$  = l'espace des perturbations  $\omega_k$ ;

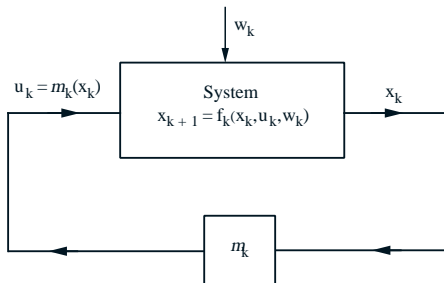
$g_k$  = fonction de coût;

$f_k$  = fonction de transition;

$x_k$  = état du système à l'étape  $k$ ;

$u_k$  = décision prise à l'étape  $k$ .

$\omega_k$  = perturbation (variable aléatoire) produite à l'étape  $k$ .



Pour  $0 \leq k \leq N + 1$  et  $x \in X_k$ , posons

$$\begin{aligned} J_{\pi,k}(x) &= \text{coût espéré total de l'étape } k \text{ à la fin,} \\ &\quad \text{si on est dans l'état } x \text{ à l'étape } k \\ &\quad \text{et si on utilise la politique } \pi \\ &= \mathbb{E}_{\pi,x} \left[ g_N(x_N) + \sum_{n=k}^{N-1} g_n(x_n, u_n, \omega_n) \right] \end{aligned}$$

où  $\mathbb{E}_{\pi,x}$  indique l'espérance lorsque  $x_k = x$ ,  $u_n = \mu_n(x_n)$  et  $x_{n+1} = f_n(x_n, u_n, \omega_n)$  pour  $n = k, \dots, N - 1$ .

Pour une politique  $\pi$  donnée, on a l'équation de **récurrence**

$$\begin{aligned} J_{\pi,N}(x) &= g_N(x) \quad \text{pour tout } x \in X_N \\ J_{\pi,k}(x) &= \mathbb{E}_{\pi,x} [g_k(x, \mu_k(x), \omega_k) + J_{\pi,k+1}(f_k(x, \mu_k(x), \omega_k))] \\ &\quad \text{pour } 0 \leq k \leq N, x \in X_k. \end{aligned}$$

où l'espérance est par rapport à  $\omega_k$  qui suit la loi  $\mathbb{P}_k(\cdot \mid x, \mu_k(x))$ .

En effet:

$$\begin{aligned} & J_{\pi,k}(x) \\ = & \mathbb{E}_{\pi,x} \left[ g_N(x_N) + \sum_{n=k}^{N-1} g_n(x_n, u_n, \omega_n) \right] \\ = & \mathbb{E}_{\pi,x} \left[ \mathbb{E}_{\pi,x} \left[ g_N(x_N) + \sum_{n=k}^{N-1} g_n(x_n, u_n, \omega_n) \mid \omega_k \right] \right] \\ = & \mathbb{E}_{\pi,x} \left[ g_k(x, \mu_k(x), \omega_k) + \mathbb{E}_{\pi,x} \left[ g_N(x_N) + \sum_{n=k+1}^{N-1} g_n(x_n, u_n, \omega_n) \mid \omega_k \right] \right] \\ = & \mathbb{E}_{\pi,x} \left[ g_k(x, \mu_k(x), \omega_k) + J_{\pi,k+1}(f_k(x, \mu_k(x), \omega_k)) \right]. \end{aligned}$$

On cherche une politique  $\pi$  qui **minimise**  $J_{\pi,0}(x_0)$ , l'**espérance mathématique** de la somme des coûts de l'étape 0 à l'étape  $N$ .

Notons  $\pi^* = (\mu_0^*, \mu_1^*, \dots, \mu_{N-1}^*)$  une telle **politique optimale**. Posons

$$\begin{aligned} J_k^*(x) &= \text{coût espéré total optimal de l'étape } k \text{ à la fin,} \\ &\quad \text{si on est dans l'état } x \text{ à l'étape } k \\ &= \min_{\pi} J_{\pi,k}(x) \\ &= \min_{\mu_k, \dots, \mu_{N-1}} J_{\mu_k, \dots, \mu_{N-1}, k}(x). \end{aligned}$$

## Proposition.

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

$$\begin{aligned} J_N(x) &= g_N(x) \quad \forall x \in X_N \\ J_k(x) &= \min_{u \in U_k(x)} \mathbb{E} [g_k(x, u, \omega_k) + J_{k+1}(f_k(x, u, \omega_k))] \\ &\text{pour } 0 \leq k \leq N - 1, x \in X_k, \end{aligned}$$

où l'espérance  $\mathbb{E}$  est par rapport à  $\omega_k$  qui suit la loi  $\mathbb{P}_k(\cdot \mid x, u)$ .

(B) Une valeur de  $u$  qui fait atteindre l'infimum est une décision optimale à prendre lorsqu'on est dans l'état  $x$  à l'étape  $k$ . On peut définir une politique optimale (si elle existe) par

$$\mu_k^*(x) = \arg \min_{u \in U_k(x)} \mathbb{E} [g_k(x, u, \omega_k) + J_{k+1}(f_k(x, u, \omega_k))].$$

On a alors  $J_k \equiv J_{\pi^*, k}$  pour tout  $k$ .

**Preuve informelle de (A) et (B):** DPOC pages 23 et 44–46.

Pour  $\pi = (\mu_1, \dots, \mu_{N-1})$ , on note  $\pi^k = (\mu_k, \dots, \mu_{N-1})$ . On a

$$J_k^*(x) = \min_{\pi^k} \mathbb{E}_{\pi^k, x} \left[ g_N(x_N) + \sum_{i=k}^{N-1} g_i(x_i, \mu_i(x_i), \omega_i) \right]$$

pour  $0 \leq k \leq N$ ,  $x \in X_k$ .

Pour  $k = N$ , on pose  $J_N^*(x_N) = g_N(x_N)$ .

On montre par induction sur  $k$  (pour  $k = N-1, \dots, 0$ ) que  $J_k^* = J_k$ .  
Supposons que c'est vrai pour  $k+1$ . On écrit  $\pi^k = (\mu_k, \pi^{k+1})$ .

Preuve informelle (on suppose que tout est fini et que le min est toujours atteint):

$$\begin{aligned}
& J_k^*(x_k) \\
= & \min_{(\mu_k, \pi^{k+1})} \mathbb{E}_{\pi^k, x} \left[ g_k(x_k, \mu_k(x_k), \omega_k) \right. \\
& \left. + g_N(x_N) + \sum_{i=k+1}^{N-1} g_i(x_i, \mu_i(x_i), \omega_i) \right] \\
= & \min_{\mu_k} \mathbb{E}_{\pi^k, x_k} \left( g_k(x_k, \mu_k(x_k), \omega_k) \right. \\
& \left. + \min_{\pi^{k+1}} \left[ \mathbb{E}_{\pi^{k+1}, x_{k+1}} \left[ g_N(x_N) + \sum_{i=k+1}^{N-1} g_i(x_i, \mu_i(x_i), \omega_i) \mid \omega_k \right] \right] \right) \\
= & \min_{\mu_k} \mathbb{E}_{\pi^k, x_k} (g_k(x_k, \mu_k(x_k), \omega_k) + J_{k+1}^*(f_k(x_k, \mu_k(x_k), \omega_k))) \\
= & \min_{u_k \in U_k(x_k)} \mathbb{E}_{\pi^k, x_k} (g_k(x_k, u_k, \omega_k) + J_{k+1}^*(f_k(x_k, u_k, \omega_k))) \\
= & J_k(x_k).
\end{aligned}$$

# Procédure Chaînage Arrière

pour tout  $x \in X_N$ ,  $J_N(x) \leftarrow g_N(x)$ ;  
pour  $k = N - 1, \dots, 0$  faire  
  pour tout  $x \in X_k$  faire

$$J_k(x) \leftarrow \min_{u \in U_k(x)} \mathbb{E} [g_k(x, u, \omega_k) + J_{k+1}(f_k(x, u, \omega_k))];$$
$$\mu_k^*(x) \leftarrow \arg \min_{u \in U_k(x)} \mathbb{E} [g_k(x, u, \omega_k) + J_{k+1}(f_k(x, u, \omega_k))];$$

# Principe d'optimalité de Bellman (cas probabiliste)

Si  $\pi^* = (\mu_0^*, \dots, \mu_{N-1}^*)$  est une politique optimale pour le problème initial et si  $0 < k < N$ , alors la politique tronquée

$\pi_k^* = (\mu_k^*, \dots, \mu_{N-1}^*)$  est une politique optimale pour le sous-problème qui consiste à minimiser

$$\mathbb{E}_{\mu_k, \dots, \mu_{N-1}} \left[ g_N(x_N) + \sum_{n=k}^{N-1} g_n(x_n, u_n, \omega_n) \mid x_k \right].$$

par rapport à  $\mu_k, \dots, \mu_{N-1}$ .

**Hypothèses:** Temps discret, modèle markovien, coûts additifs.

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

**Exemple:** si on veut minimiser

$$\mathbb{E}_{\mu_k, \dots, \mu_{N-1}} [\max(g_k(x_k, u_k, \omega_k), g_{N-1}(x_{N-1}, u_{N-1}, \omega_{N-1}), g_N(x_N)) \mid x_k].$$

Le principe ne tient pas non plus pour le sous-problème: minimiser

$$\mathbb{E}_{\mu_k, \dots, \mu_j} \left[ \sum_{n=k}^j g_n(x_n, u_n, \omega_n) \mid x_k \right]$$

si  $j < N$  et l'état  $x_j$  n'est pas déterminé, car il peut arriver que la politique optimale  $\pi^*$  amène des coûts un peu plus élevés pour les étapes  $k$  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.

Par opposition, **commande en boucle ouverte:** on prend toutes les décisions  $u_0, \dots, u_{N-1}$  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 déterministe: pas de différence.

Ce modèle de PDS 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;
- Évolution en temps continu;
- Etc.

# Exemple (modifié) de gestion d'un inventaire.

Monsieur D. Taillant vend des Zyx à Loinville.

Les clients arrivent au hasard pour acheter des Zyx.

Au début de chaque mois, l'avion vient à Loinville et peut apporter une commande de Zyx. Soient:

$x_k$  = Niveau des stocks au début du mois  $k$ ,  
avant de commander;

$u_k$  = Nombre de Zyx commandés (et reçus) au début du mois

$\omega_k$  = Nombre de Zyx demandés par les clients durant le  
mois  $k$ . On suppose que les  $\omega_k$  sont des variables  
aléatoires discrètes indépendantes;

$C + cu$  = Coût d'une commande de  $u$  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:

$J_k(x)$  = coût espéré total pour les mois  $k$  à  $N$ , si  $x_k = x$  et que l'on suit une politique optimale;

Si les inventaires négatifs ("backlogs") sont permis, on a

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

et on peut optimiser sans tenir compte des revenus de vente, car ceux-ci ne dépendent pas de la politique. Récurrence:

$$J_N(x) = g_N(x), \quad \text{pour } x \leq B;$$

$$J_k(x) = \min_{0 \leq u \leq B-x} \left( r_k(x) + \mathbb{I}(u > 0)C + cu - v\mathbb{E}[\omega_k] + \sum_{i \geq 0} \mathbb{P}[\omega_k = i] J_{k+1}(x + u - i) \right), \quad x \leq B; \quad k = N-1, \dots, 0;$$

$$\mu_k^*(x) = \arg \min_{0 \leq u \leq B-x} \left( \mathbb{I}(u > 0)C + cu + \sum_{i \geq 0} \mathbb{P}[\omega_k = i] J_{k+1}(x + u - i) \right).$$

$$J_N(x) = g_N(x), \quad \text{pour } x \leq B;$$

$$J_k(x) = \min_{0 \leq u \leq B-x} \left( r_k(x) + \mathbb{I}(u > 0)C + cu - v\mathbb{E}[\omega_k] + \sum_{i \geq 0} \mathbb{P}[\omega_k = i] J_{k+1}(x + u - i) \right), \quad x \leq B; \quad k = N-1, \dots, 0;$$

$$V_k(x) = J_k(x) - r_k(x) \quad (\text{éviter de recalculer la somme pour chaque } u)$$

$$= \min \left( -v\mathbb{E}[\omega_k] + \sum_{i \geq 0} \mathbb{P}[\omega_k = i] J_{k+1}(x - i), \right. \\ \left. C + c + V_k(x+1), \dots, C + (B-x)c + V_k(B) \right) \quad \text{si } x < B.$$

$$\mu_k^*(x) = \arg \min_{0 \leq u \leq B-x} (-\mathbb{I}[u = 0]v\mathbb{E}[\omega_k] + \mathbb{I}(u > 0)C + cu + V_k(x+u)).$$

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

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

**Coûts de calcul:** supposons que la somme sur  $i$  (valeurs possibles de  $\omega_k$ ) a  $T$  termes non négligeables. Les coûts de calcul sont

$O(NB^2T)$  pour la récurrence sur  $J_k$ ;

$O(NB(B + T))$  pour la récurrence sur  $V_k$ ;

$O(NBT)$  pour la cas simplifié où  $C = 0$ .

Supposons maintenant que les **inventaires négatifs ne sont pas permis**. On a

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

et les équations de récurrence deviennent:

$$J_N(x) = g_N(x) \quad \text{pour } 0 \leq x \leq B;$$

$$J_k(x) = \min_{0 \leq u \leq B-x} \left( r_k(x) + \mathbb{I}(u > 0)C + cu \right. \\ \left. + \sum_{i \geq 0} P[\omega_k = i] [-v \min(i, x + u) + J_{k+1}(\max(0, x + u - i))] \right) \\ \text{pour } 0 \leq x \leq B; \quad k = N - 1, \dots, 0;$$

$$V_k(x) = \min \left( \sum_{i \geq 0} P[\omega_k = i] [-v \min(i, x) + J_{k+1}(\max(0, x - i))], \right. \\ \left. C + c + V_k(x + 1), \dots, C + (B - x)c + V_k(B) \right) \\ \text{si } x < B.$$

Dans le cas où  $C = 0$ :

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

Exemple numérique: DPOC, pages 28–32.

## 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  $u$  pièces, il en coûte  $C + cu$ . Dans un lot de taille  $u$ , le nombre  $Y$  de pièces acceptables est une variable aléatoire binomiale:

$$\mathbb{P}[Y = y] = \binom{u}{y} p^y (1-p)^{u-y}, \quad y = 0, \dots, u.$$

En pratique, on va 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  $K$ .

# Taille d'un lot de pièces à fabriquer (suite)

- $x_k$  = Nb de pièces encore requises avant de produire le lot  $k + 1$ ;
- $u_k$  = Taille du lot  $k + 1$ ;
- $y_k$  = Nb de pièces acceptables dans le lot  $k + 1$ ;
- $J_k(x)$  = 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  $J_0(M)$  et une politique optimale.  
Pour tout  $k$  et  $x \leq 0$ , on a  $J_k(x) = 0$ . Pour  $x > 0$ :

$$J_N(x) = K;$$
$$J_k(x) = \min_{u \geq x} \left( C + cu + \sum_{y=0}^u \binom{u}{y} p^y (1-p)^{u-y} J_{k+1}(x-y) \right)$$
$$\mu_k^*(x) = \arg \min_{u \geq x} \left( \quad \right).$$

Peut-on simplifier ces équations pour réduire les coûts de calcul?

## Autre notation souvent utilisée: pas de $\omega_k$ dans la notation.

On définit un **noyau de transition** (famille de lois de probabilité) par

$$\begin{aligned} Q(A \mid x_k, u_k) &= \mathbb{P}[x_{k+1} \in A \mid x_k, u_k] \\ &= \mathbb{P}_k(\{\omega_k \in D_k : f_k(x_k, u_k, \omega_k) \in A\}). \end{aligned}$$

On peut remplacer le coût  $g_k(x_k, u_k, \omega_k)$  par

$$\tilde{g}_k(x_k, u_k) = \mathbb{E}[g_k(x_k, u_k, \omega_k) \mid x_k, u_k].$$

Si  $X_k \equiv X$  est fini ou **dénombrable**, les noyaux de transition deviennent des matrices de probabilité de transition: on a une **chaîne de Markov commandée** à espace d'états dénombrable.

On pourra dénoter, par exemple,

$$p_{ij}(u, k) = \mathbb{P}[x_{k+1} = j \mid x_k = i, u_k = u].$$

La matrice  $P(u, k)$  dont les éléments sont les  $p_{ij}(u, k)$  est la matrice des probabilités de transition à l'étape  $k$ , sous la décision  $u$ .

Si  $P(u, k)$ ,  $U_k$  et  $\tilde{g}_k$  ne dépendent pas de  $k$  (**modèle stationnaire**):

$$J_k(i) = \min_{u \in U(i)} \left( \tilde{g}(i, u) + \sum_j p_{ij}(u) J_{k+1}(j) \right).$$

# 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.

À chaque période,  $\mathbb{P}[m \text{ clients arrivent}] = p_m$ ,  $m \geq 0$ .

Le serveur a 2 vitesses: rapide et lent.

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  $1 - q_f$  [ $1 - q_s$ ].

Il y a aussi un coût de  $r(i)$  à chaque période où il y a  $i$  clients dans le système au début de la période.

État: nombre de clients dans le système.

L'espace des décisions est  $U = \{\text{rapide, lent}\}$ .

Soit  $\xi_k$  le nombre de clients servis à la période  $k$ .

$$J_N(i) = r(i), \quad \text{pour } 0 \leq i \leq n;$$

$$J_k(0) = r(0) + c_s + V_k(0);$$

$$J_k(i) = r(i) + \min[c_f + q_f V_k(i-1) + (1 - q_f) V_k(i), \\ c_s + q_s V_k(i-1) + (1 - q_s) V_k(i)] \\ \text{pour } 0 \leq k \leq N-1, 1 \leq i \leq n,$$

où

$$V_k(i) = \mathbb{E}[J_{k+1}(x_{k+1}) \mid x_k - \xi_k = i] \\ = \sum_{m=0}^{n-i-1} p_m J_{k+1}(i+m) + J_{k+1}(n) \sum_{m=n-i}^{\infty} p_m.$$

# 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. On y reviendra.

**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.

$J_k(x)$  = 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,  $J_k \equiv J$  ne dépend pas de  $k$  et on a :

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

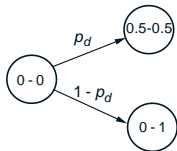
## Applications possibles:

- 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.

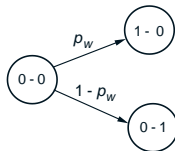
DPOC, Exemples 1.1.5, 1.3.3: [match d'échecs de  \$N\$  parties](#).

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.

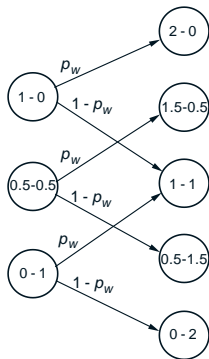
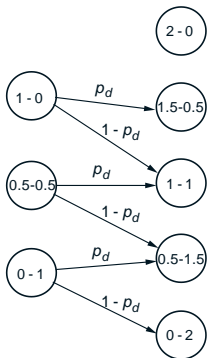
On suppose que  $p_1 = 0$  et  $p_{-1} = 1 - p_0$  (en mode prudent, on peut seulement annuler ou perdre) et que  $q_0 = 0$  et  $q_{-1} = 1 - q_1$  (en mode agressif, on peut gagner ou perdre).



1st Game / Timid Play



1st Game / Bold Play



On a ici

$$J_k(x) = J_N(x) \quad \text{pour } k > N;$$

$$J_N(x) = \begin{cases} 1 & \text{si } x > 0; \\ q_1 & \text{si } x = 0; \\ 0 & \text{si } x < 0; \end{cases}$$

$$J_{N-1}(x) = \begin{cases} 1 & \text{si } x > 1; \\ p_0 + (1 - p_0)q_1 & \text{si } x = 1; & \text{(jeu prudent);} \\ q_1 & \text{si } x = 0; & \text{(jeu agressif);} \\ q_1^2 & \text{si } x = -1; & \text{(jeu agressif);} \\ 0 & \text{si } x < -1; \end{cases}$$

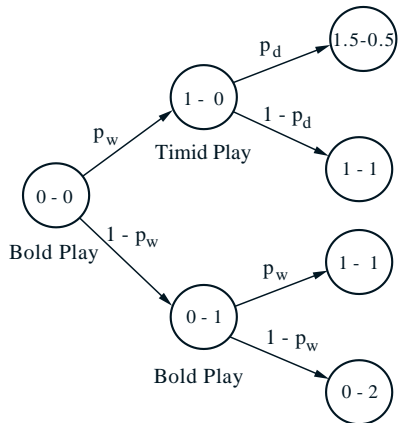
$$J_k(x) = \max[p_0 J_{k+1}(x) + (1 - p_0) J_{k+1}(x - 1), \\ q_1 J_{k+1}(x + 1) + (1 - q_1) J_{k+1}(x - 1)] \\ \text{pour } 0 \leq k < N \text{ et } -k \leq x \leq k.$$

Si  $N = 2$ , au début du match on a

$$\begin{aligned} J_0(0) &= \max [p_0 J_1(0) + (1 - p_0) J_1(-1), q_1 J_1(1) + (1 - q_1) J_1(-1)] \\ &= \max [p_0 q_1 + (1 - p_0) q_1^2, q_1 p_0 + (1 - p_0) q_1^2 + (1 - q_1) q_1^2] \\ &= q_1 p_0 + (1 - p_0) q_1^2 + (1 - q_1) q_1^2 \quad (\text{jeu agressif}). \end{aligned}$$

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

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



**Intéressant:** On pourrait croire que  $q_1 < 1/2$  implique que  $J_0(0) < 1/2$ , 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  $J_0(0) \approx 0.530$ .

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

$$\tilde{J}_0(0) = J_0(0) - (1 - q_1)q_1^2.$$

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  $(1 - q_1)q_1^2 \approx 0.105$  et la probabilité de gain avec la meilleure politique en boucle ouverte est  $\approx 0.425$ .

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

**Exemple: DPOC, Exercice 1.26.**

# Reformulation pour “markovianiser”

Que faire si les hypothèses ne sont pas vérifiées, e.g., si  $f_k$ ,  $g_k$  et la loi de  $\omega_k$  dépendent des états et décisions précédant  $x_k$  et  $u_k$ ?

En général, on peut toujours se ramener au modèle de base (que nous avons décrit) en redéfinissant judicieusement les espaces d'états, de décisions, et de perturbations.

Il suffit d'incorporer suffisamment d'information dans l'état  $x_k$  (“state augmentation”).

Par exemple, si  $f_k$  [ou  $g_k$ ] dépend de  $(x_{k-1}, u_{k-1}, x_k, u_k, \omega_k)$ , on peut redéfinir l'état  $x_k$  par  $\tilde{x}_k = (x_{k-1}, u_{k-1}, x_k)$  et la fonction de transition devient

$$\tilde{x}_{k+1} = \tilde{f}_k(\tilde{x}_k, u_k, \omega_k) = (x_k, u_k, f_k(x_{k-1}, u_{k-1}, x_k, u_k, \omega_k)).$$

À la limite, on peut définir l'état comme étant toute l'histoire du processus observée jusqu'à date.

On peut ainsi traiter (en principe) des modèles non additifs.

Mais si l'espace d'états est trop grand, on ne pourra pas résoudre les équations de récurrence! La [malédiction des grandes dimensions](#) ("the curse of dimensionality")

Il faut être parcimonieux dans la définition de l'état.

Exemple: DPOC, p.38.

Supposons qu'à l'étape  $k$ , juste avant de prendre la décision  $u_k$ , on obtient une **prévision**  $y_k$  nous donnant une information plus précise sur la loi de  $\omega_k$ .

Par exemple, supposons que pour  $i = 1, \dots, m$ ,  $\mathbb{P}[y_k = i] = p_i$ , et  $\omega_k$  suit la loi  $Q_i$  lorsque  $y_k = i$ .

On a donc  $\mathbb{P}[\omega_k \in \cdot] = \sum_{i=1}^m p_i Q_i[\cdot]$ .

On peut se ramener à notre modèle de base en redéfinissant l'état et la perturbation par  $\tilde{x}_k = (y_k, x_k)$  et  $\tilde{\omega}_k = (\omega_k, y_{k+1})$ . On obtient

$$\tilde{J}_k(y_k, x) = \min_{u \in U_k(x)} \mathbb{E} \left[ g_k(x, u, \omega_k) + \sum_{i=1}^m p_i \tilde{J}_{k+1}(i, f_k(x, u, \omega_k)) \mid y_k \right]$$

pour  $0 \leq k \leq N - 1$ ,  $x \in X_k$ ,  $y_k \in \{1, \dots, m\}$ .

Mais dans ce cas-ci, il n'est pas nécessaire de mettre  $y_k$  dans l'état. Au lieu d'écrire la récurrence en termes des fonctions  $\tilde{J}_k(y_k, x_k)$ , on peut l'écrire en termes des fonctions

$$\hat{J}_k(x_k) = \sum_{i=1}^m p_i \tilde{J}_k(i, x).$$

On obtient

$$\hat{J}_k(x) = \sum_{i=1}^m p_i \min_{u \in U_k(x)} \mathbb{E} \left[ g_k(x, u, \omega_k) + \hat{J}_{k+1}(f_k(x, u, \omega_k)) \mid y_k = i \right].$$

Cela équivaut à observer l'état  $x_k$  **avant** d'obtenir la prévision, puis de prendre la décision  $u_k$  **après** avoir observé la prévision.

La décision est prise avec **davantage** d'information que l'état  $x_k$ , mais on peut quand même écrire la récurrence en termes de fonctions de  $x_k$  seulement. **C'est plus économique.**

# Subtilités mathématiques.

Pour que les espérances  $\mathbb{E}_\pi$  et  $\mathbb{E}_{\mu_k, \dots, \mu_{N-1}}$  soient bien définies, on doit faire des **hypothèses de mesurabilité** sur les fonctions  $f_k$ ,  $g_k$ , et  $\mu_k$ , et les espaces  $S_k$ ,  $U_k$ , et  $D_k$  doivent avoir une structure additionnelle (espaces métriques complets mesurables, etc.). Et on doit s'assurer qu'il existe une politique optimale qui satisfait les conditions de mesurabilité.

Dans le cas où les  $D_k$  sont **dénombrables**: pas de problème.

Cas général: beaucoup plus complexe. Nous n'allons pas aborder ces questions ici.

Questions intéressantes du point de vue mathématique, mais pas beaucoup d'impact du point de vue pratique.

Références: Bertsekas et Shreve (1978) et Hernández-Lerma et Lasserre (1995).

# Fonction d'utilité et mesure de risque.

L'espérance mathématique du coût (ou du revenu) n'est pas toujours la mesure appropriée à optimiser.

## Exemple: le paradoxe de St-Petersbourg:

Vous payez  $x$  dollars pour jouer au jeu suivant. On tire à pile ou face, on compte le nombre  $Y$  de faces avant l'obtention du premier pile, et vous recevez  $2^Y$  dollars. L'espérance de gain net est

$$\sum_{y=0}^{\infty} 2^y \frac{1}{2^{y+1}} - x = \infty$$

quelque soit  $x$ . Mais est-on vraiment prêt à payer un montant arbitrairement grand pour jouer à ce jeu? **Non.**

On a une très grande probabilité de recevoir un montant modeste, et une probabilité minuscule de recevoir un montant gigantesque.

Mais l'utilité d'un gain gigantesque n'est pas proportionnelle au gain.

## Critère min-max

Approche pessimiste: dans le cas où il y a de l'incertitude, on considère toujours le pire cas (au lieu de l'espérance). C'est le cas extrême d'aversion au risque.

Au lieu de minimiser (par rapport à  $\pi$ , pour  $x_0 = x$ )

$$\mathbb{E}_{\pi, x} \left[ g_N(x_N) + \sum_{n=0}^{N-1} g_n(x_n, \mu_n(x_n), \omega_n) \right]$$

on voudra minimiser

$$\max_{\omega_1, \dots, \omega_N} \left[ g_N(x_N) + \sum_{n=0}^{N-1} g_n(x_n, \mu_n(x_n), \omega_n) \right].$$

Rarement la bonne solution. Sauf si  $\omega_n$  est la décision d'un adversaire. Avec un tel critère, un investisseur ne va jamais investir!

Solution beaucoup plus intéressante: **fonction d'utilité**.

Le preneur de décision veut maximiser  $\mathbb{E}[U(X)]$  où  $X$  est le gain net et  $U : \mathbb{R} \rightarrow \mathbb{R}$  est sa **fonction d'utilité**.

On se ramène au cas précédent, en remplaçant  $X$  par  $U(X)$ .

Habituellement, la fonction  $U$  est **croissante et concave**.

Dans l'appendice G.2 de DPOC, Proposition G.1, on donne des conditions suffisantes pour l'existence d'une telle fonction  $U$ , dans le cas où  $X$  prend les valeurs  $o_1, \dots, o_n$  avec probabilités  $p_1, \dots, p_n$ .

$$\mathbb{E}[U(X)] = \sum_{i=1}^n p_i U(o_i).$$

Le rôle de la fonction d'utilité consiste essentiellement à **modifier la valeur** d'un gain, selon son utilité, de manière à pouvoir exprimer l'objectif comme une espérance mathématique.

Si  $o_i \geq 0$  représente le gain, et si  $\sum_i p_i U(o_i)/o_i = 1$  (il suffit de multiplier  $U$  par une constante pour obtenir cela), alors on peut obtenir un résultat équivalent en modifiant les probabilités à la place de prendre une fonction d'utilité: on remplace  $p_i$  par  $q_i = p_i U(o_i)/o_i$ :

$$\mathbb{E}[U(X)] = \sum_{i=1}^n p_i U(o_i) = \sum_{i=1}^n q_i o_i.$$

C'est ce que l'on fait pour l'évaluation d'options financières.

L'importance sampling (en simulation) équivaut à changer les deux (les  $p_i$  et l'utilité) sans changer l'espérance.

## Exemple.

On a deux possibilités d'investissement pour notre capital:

(A) placement **sûr** qui rapportera **1.5** dollar par dollar investi;

(B) placement **risqué** qui rapporte **1** dollar par dollar investi avec probabilité  $p$  et **3** dollars par dollar investi avec probabilité  $1 - p$ .

Supposons que l'on place une fraction  $d$  du capital dans l'option A, et  $1 - d$  dans l'option B. L'**utilité espérée** sera

$$\mathbb{E}[U(X)] = p U(1.5d + (1 - d)) + (1 - p) U(1.5d + 3(1 - d)).$$

Quelqu'un qui n'aime pas le risque choisira un  $d$  plus grand, et vice-versa.

En général, plus on est riche ou jeune, plus on a raison de préférer le risque (on aura une fonction  $U(x)$  dont la dérivée seconde sera plus proche de 0 pour les grandes valeurs de  $x$ ). Pour les valeurs négatives, l'utilité  $U(-x)$  peut devenir une constante lorsque  $x \rightarrow \infty$ .