Stratégie optimale pour le Mastermind

On note m le nombre de trous et n le nombre couleurs[1]. Le nombre total de combinaison est $N = n^m$.

1 Modélisation par un Processus de Décision Markovien

L'idée est de modéliser le problème par le MDP (Markov Decision Process = Processus de Décision Markovien) suivant :

- Soit $\mathcal{C} \simeq [\![1, N]\!]$ l'ensemble des combinaisons. L'espace d'état $\mathcal{S}$ est l'ensemble des distributions catégorielles sur $\mathcal{C}$. L'état courant nous fournit ainsi la probabilité que chaque combinaison soit la vraie combinaison c_* sachant les données observées jusqu'ici. Les états finaux $\mathcal{S}_F$ sont les distributions chargeant une seule combinaison.
- Pour tout $\mathbb{P} \in \mathcal{S}$, l'ensemble des actions est $\mathcal{C}$. Cela correspond à choisir une combinaison à jouer parmi toutes les distributions possibles.
- Pour chaque couple d'état action $(\mathbb{P}, c) \in \mathcal{S} \times \mathcal{C}$, la récompense est toujours égale à -1 puisque l'objectif est d'atteindre un état final avec le moins d'étapes possibles.
- Les transitions d'un état à un autre étant totalement déterministes, on considère la fonction de transition f suivante : $\mathbb{P}_{t+1} = f(\mathbb{P}_t, c_t, g(c_*, c_t))$ où c_t est l'action choisie. Les fonctions f et g sont explicitées dans la section 2.

Soit $\pi : \mathcal{S} \times \mathcal{C} \to [0, 1]$ une stratégie (i.e., une distribution de probabilités sur les couples état-action). Soit T_F la variable aléatoire correspondant au nombre d'étapes d'un épisode obtenue en utilisant π, et R_t la variable aléatoire correspondant à la récompense reçue à l'étape t (on a $R_t = -1$ presque sûrement). Le gain G est

$$G = \sum_{t=1}^{T_F} R_t = -T_F.$$

On suppose qu'on dispose d'une distribution initiale (voir section 3) $\mathbb{P}_0$. On note

$$v_\pi(\mathbb{P}_0) = \mathbb{E}_\pi[G|\mathbb{P}_0],$$

l'espérance du gain lorsque les actions sont choisies avec la stratégie π et lorsque l'état initial est $\mathbb{P}_0$, appelé fonction valeur pour la stratégie π (appliquée en $\mathbb{P}_0$). On cherche la stratégie optimale π_* définie par:

$$\pi_* = \arg\max_\pi v_\pi(\mathbb{P}_0).$$

Je ne vois pas comment optimiser $v_\pi(\mathbb{P}_0)$ de façon exacte. Peut-être est-il possible d'utiliser la programmation dynamique ? Cependant, l'espace d'état est infini donc il n'est pas possible d'utiliser des méthodes tabulaires. La section 2 montre que, pour une distribution initiale $\mathbb{P}_0$ fixée, l'espace d'état se réduit à 2^N, ce qui est énorme[2]. Encore une fois, les méthodes tabulaires sont inutilisables. Il est bien sûr possible d'utiliser des méthodes approchées, mais je souhaite trouver la stratégie optimale, pas une approximation.

[1] L'absence de couleur (le "trou") est aussi une couleur !

[2] Dans la configuration classique, on a $m = 5$ et $n = 9$ donc $N = 59049$.

Supposons qu'on soit à l'étape t et qu'on connaisse la distribution $\mathbb{P}_t$. On définit la distribution :

$$\mu_{t+1}(c) = \sum_{c' \in \mathcal{C}} \pi(\mathbb{P}_t, c') f(\mathbb{P}_t, c', g(c_*, c')).$$

On considère l'entropie de la distribution μ_{t+1} définie par :

$$H(\mu_{t+1}) = -\sum_{c \in \mathcal{C}} \mu_{t+1}(c) \ln(\mu_{t+1}(c)).$$

Conjecture 1. La stratégie optimale est (au moins approximativement) égale à

$$\pi_* = \arg\min_{\pi} H(\mu_{t+1}),$$

et elle ne dépend pas de t.

L'entropie est maximale pour une distribution uniforme et minimale (i.e., nulle) pour une distribution ne chargeant qu'une combinaison. L'entropie permet de faire un compromis entre réduire le nombre de combinaisons possibles (quitte à avoir une distribution quasi-uniforme sur les combinaisons restantes), et garder plus de combinaisons tout en ayant certaines combinaisons très probables.

Conjecture 2. Si π_* est déterminée conformément à la conjecture 1, alors π_* charge uniformément les combinaisons ayant la probabilité maximale selon $\mathbb{P}_t$ et ne charge aucune autre combinaison.

2 Détermination de la fonction de transition

On se place dans le cadre bayésien. La vraie distribution c_* est vue comme une variable aléatoire dont la distribution à l'étape t est l'état $\mathbb{P}_t$. On utilise la stratégie π pour choisir une combinaison à jouer $c_t \in \mathcal{C}$. Le correcteur va nous fournir le nombre de couleurs correctes bien placées (nombre de "noirs") et le nombre de couleurs correctes mal placées (nombre de "blancs"). On modélise le correcteur par une fonction de comparaison $g(c_*, c_t)$ fournissant le nombre de noirs et de blancs modélisant la différence entre c_* et c_t. Ainsi, à l'étape t, on obtient les données D suivantes :

$$D = \{c_t, g(c_*, c_t)\},$$

correspondant à la combinaison jouée et au résultat de la comparaison de cette combinaison avec la vraie combinaison.

On utilise la formule de Bayes pour mettre à jour la distribution de c_*. Pour tout $c \in \mathcal{C}$, on a :

$$\begin{aligned}
\mathbb{P}_{t+1}(c) &= \mathbb{P}(c|D) \\
&= \frac{\mathbb{P}(D|c)\mathbb{P}(c)}{\sum_{c' \in \mathcal{C}} \mathbb{P}(D|c')\mathbb{P}(c')} \\
&= \frac{\mathbb{P}(D|c)\mathbb{P}_t(c)}{\sum_{c' \in \mathcal{C}} \mathbb{P}(D|c')\mathbb{P}_t(c')}
\end{aligned}$$

Il reste à déterminer la vraisemblance $\mathbb{P}(D|c)$. Cela consiste à se demander quelle serait la probabilité d'avoir obtenu les données D si la vraie combinaison était c, i.e., si $c_* = c$. Pour répondre à cette question, il suffit de comparer c avec la combinaison jouée c_t :

- Si $g(c, c_t) = g(c_*, c_t)$ alors on aurait certainement obtenu les mêmes données si $c_* = c$, donc $\mathbb{P}(D|c) = 1$.
- Si $g(c, c_t) \neq g(c_*, c_t)$ alors il est impossible d'obtenir les mêmes données si $c_* = c$, donc $\mathbb{P}(D|c) = 0$.

On introduit la notation $[a:b]=1$ si $a=b$ et 0 si $a\neq b$. On a donc pour tout $c\in\mathcal{C}$:

$$\mathbb{P}_{t+1}(c)=\frac{[g(c_*,c_t):g(c,c_t)]\mathbb{P}_t(c)}{\sum_{c'\in\mathcal{C}}[g(c_*,c_t):g(c',c_t)]\mathbb{P}_t(c')}.$$

Ainsi, le calcul de $\mathbb{P}_{t+1}$ consiste à supprimer les combinaisons impossibles d'après les données et à renormaliser les probabilités des combinaisons restantes sans changer les rapports entre elles.

Il reste encore à expliciter la fonction de comparaison g. On note σ_i la permutation circulaire de $[\![0,m-1]\!]$ consistant à un "décalage" de $i-1$ vers la droite, i.e., $\sigma_i(j)=j+i-1 \quad [m]$. On sait que le nombre de blancs est compris entre 0 et m. On peut alors encoder le nombre de noirs et le nombre de blancs dans un entier naturel k de telle sorte que le nombre de noirs soit le quotient de k par $m+1$ et le nombre de blancs soit le reste de k dans la division euclidienne de k par $m+1$. Cela définit une injection de $[\![0,m]\!]\times[\![0,m]\!]$ dans $\mathbb{N}$. Le lecteur pourra se convaincre que g est de la forme :

$$g(c_*,c_t)=\sum_{i=1}^{m}\mu_i\sum_{j=1}^{m}\{[c_*^j:c_t^{\sigma_i(j)}]\prod_{l=1}^{i-1}(1-[c_*^j:c_t^{\sigma_l(j)}])(1-[c_*^{\sigma_l(j)}:c_t^{\sigma_i(j)}])\},$$

avec $\mu_1=m+1$ et $\mu_i=1$ pour $i>1$ et où on a noté c^j la j-ième couleur de la combinaison c. Le raisonnement derrière cette formule consiste à remarquer que lorsqu'une correspondance a été trouvé entre deux couleurs (une de c_* et l'autre de c_t) alors aucune de ces deux couleurs ne peut être réutilisée dans une autre correspondance. De plus, une correspondance de type "noir" a priorité sur une correspondance de type "blanc". Dans la formule, la première permutation circulaire σ_1 (qui n'est autre que l'identité) correspond à rechercher les correspondances "noir". Toutes les autres permutations permettent de rechercher les correspondances "blanc". Le produit permet de s'assurer que, si une correspondance est trouvée, elle ne sera pas prise en compte si les couleurs ont déjà participé à une autre correspondance dans les permutations précédentes.

3 Choix d'une distribution a priori

Comme dans toute méthode bayésienne, le choix de la distribution a priori $\mathbb{P}_0$ est subjectif[3]. Il dépend de l'idée qu'on se fait de la distribution utilisée par le correcteur pour choisir c_*. Si le correcteur est un logiciel, il est raisonnable de penser que $\mathbb{P}_0$ est la distribution uniforme sur $\mathcal{C}$. Si le correcteur est un être humain, je pense qu'il aura tendance à privilégier les combinaisons les plus "compliquées". Par exemple, il est peut probable qu'il utilise une combinaison avec m couleurs identiques. Il est plus probable qu'il choisisse une combinaison contenant cinq couleurs différentes.

Ce comportement sera rationnel si les combinaisons c_* "compliquées" conduisent à des indices éliminant en moyenne (sur tous les c_t) moins de combinaisons que les c_* "simples". Cela suggère une nouvelle stratégie. A chaque étape t, on considère d'une part l'ensemble des combinaisons jouables (i.e., $\mathcal{C}$), et d'autre part l'ensemble $\mathcal{C}_t$ des combinaisons encore possibles, i.e., les combinaisons ayant une probabilité non nulle selon $\mathbb{P}_t$. Pour tout $(c_1,c_2)\in\mathcal{C}_t\times\mathcal{C}$, on considère la fonction $\lambda_t(c_1,c_2)$ donnant le nombre de combinaisons qui seraient supprimées de $\mathcal{C}_t$ si c_1 était la vraie combinaison et qu'on jouait c_2, i.e., si $c_*=c_1$ et $c_t=c_2$. On calcule ensuite pour chaque $c\in\mathcal{C}$, la moyenne du nombre de combinaisons supprimées :

$$\overline{\lambda_t}(c)=\sum_{c'\in\mathcal{C}_t}\mathbb{P}_t(c')\lambda_t(c',c)$$

Conjecture 3. La stratégie optimale du Mastermind est

$$c_t=\arg\max_c\overline{\lambda_t}(c).$$

Cette conjecture semble entrer en contradiction avec les conjectures 1 et 2.

Une autre remarque. Si le joueur peut jouer plusieurs parties, il pourra découvrir quelle est la distribution utilisée par le correcteur, et donc privilégier les combinaisons les plus probables selon cette distribution. Le correcteur pourra en retour s'adapter en changeant sa distribution. Il s'agit d'un problème de théorie des jeux.

[3]Il est néanmoins impératif que toute combinaison de $\mathcal{C}$ ait une probabilité non nulle selon $\mathbb{P}_0$.

Conjecture 4. La distribution du joueur et celle du correcteur convergent. Autrement dit, il existe (au moins) un équilibre de Nash.

Pour le moment, je suppose que le joueur ne va jouer qu'une seule partie contre un correcteur humain. Idéalement, la distribution a priori devrait être de la forme :

$$\mathbb{P}_0^* = \arg\max_{\mathbb{P}_0} \min_{\pi} v_\pi(\mathbb{P}_0),$$

mais c'est encore une fois impossible à calculer. La conjecture 3 suggère que la distribution a priori soit de la forme :

$$\mathbb{P}_0^* = \arg\min_{\mathbb{P}_0} \max_{c} \overline{\lambda_0}(c).$$

Cette idée motive le développement suivant. Soit $\mathcal{L}$ un langage de programmation Turing-complet. On définit une "complexité de Kolmogorov" $\eta : \mathcal{C} \to \mathbb{R}^+$ comme étant la longueur dans $\mathcal{L}$ du plus petit algorithme générant une combinaison $c \in \mathcal{C}$. Je connais mal cette théorie, donc je ne sais pas comment calculer explicitement η, ni à quel point η va dépendre du choix de $\mathcal{L}$. Il me semble que le choix de $\mathcal{L}$ n'impacte pas η si le nombre de trous m (i.e., la longueur de la séquence) tend vers l'infini. Comme le nombre de trous est faible dans le cadre d'un Mastermind classique, il est possible que η dépende fortement du choix de $\mathcal{L}$. La distribution a priori $\mathbb{P}_0$ est définie pour tout $c \in \mathcal{C}$ par :

$$\mathbb{P}_0(c) = \frac{\exp(\eta(c))}{\sum_{c' \in \mathcal{C}} \exp(\eta(c'))}, \tag{1}$$

de sorte que plus $\eta(c)$ est grand (i.e., plus c est "compliquée"), plus c est probable. A la place de exp, on peut prendre n'importe quelle fonction croissante de $\mathbb{R}^+$ dans $\mathbb{R}^+$ mais il me semble que exp a de bonnes propriétés (mais je ne me souviens plus lesquelles).

Conjecture 5. Si $\mathbb{P}_0^*$ est définie selon l'équation 1 alors :

$$\mathbb{P}_0^* \approx \arg\min_{\mathbb{P}_0} \max_{c} \overline{\lambda_0}(c).$$

Comme je ne sais pas calculer η, je propose un proxy pour η. On considère l'action du groupe symétrique $\mathfrak{S}_m$ sur $\mathcal{C}$.

Conjecture 6. Pour tout $c \in \mathcal{C}$, $\eta(c) \approx h(|\mathrm{Fix}_c|)$ où Fix_c est l'ensemble des éléments de $\mathfrak{S}_m$ laissant c invariant et h est une fonction décroissante de $\mathbb{R}^+$ dans $\mathbb{R}$.

Par exemple, une combinaison c_1 formée de m couleurs identiques sera invariante par l'action de tout élément de $\mathfrak{S}_m$ donc $|\mathrm{Fix}_{c_1}| = m!$, alors qu'une combinaison c_2 formée de m couleurs différentes ne sera invariante que par l'identité donc $|\mathrm{Fix}_{c_2}| = 1$. Il est aussi possible d'utiliser un sous-groupe de $\mathfrak{S}_m$, par exemple le groupe des rotations (permutations circulaires) ou le groupe diédral. Finalement, une proposition pour la distribution a priori $\mathbb{P}_0$ est, pour tout $c \in \mathcal{C}$:

$$\mathbb{P}_0(c) = \frac{\exp(-|\mathrm{Fix}_c|)}{\sum_{c' \in \mathcal{C}} \exp(-|\mathrm{Fix}_{c'}|)},$$

en utilisant $h : x \mapsto -x$.