Master MVA: Apprentissage par renforcement **Lecture: 3**

Introduction aux algorithmes de bandit

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

Références bibliographiques:

- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. Machine Learning, 47(2/3):235-256, 2002.
- Auer, Cesa-Bianchi, Freund, Schapire. The nonstochastic multi-armed bandit problem. 2003.
- T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6:4-22, 1985.

1 Le bandit stochastique à K bras

Considérons K bras (actions, choix) définis par des distributions $(\nu_k)_{1\leq k\leq K}$ à valeurs dans $[0,1]$, de loi inconnues. A chaque instant, l'agent choisit un bras $I_t \in \{1,\dots,K\}$ et observe une récompense $x_t \sim \nu_{I_t}$, réalisation indépendante (des récompenses passées) générée selon la loi du bras I_t. Son objectif est de maximiser la somme des récompenses qu'il reçoit, en espérance.

Notons $\mu_k = \mathbb{E}_{X\sim\nu_k}[X]$ l'espérance de récompense de chaque bras, et $\mu^* = \max_k \mu_k$ la valeur moyenne du meilleur bras. Si l'agent connaissait les lois, il choisirait alors le meilleur bras à chaque instant et obtiendrait une récompense moyenne μ^*. Comme il ne connait pas les lois initialement, il doit explorer les différents bras pour acquérir de l'information (exploration) qui lui servira ensuite pour agir optimalement (exploitation). Cela illustre le compromis exploration-exploitation.

Pour évaluer la performance d'une stratégie donnée, on va définir à quelle vitesse cette stratégie permet d'atteindre un taux de récompense moyen optimal. Pour cela on définit le **regret cumulé** à l'instant n:

$$R_n = n\mu^* - \sum_{t=1}^{n} x_t,$$

qui représente la différence en récompenses cumulées entre ce qu'il a obtenu et ce qu'il aurait pu obtenir en moyenne s'il avait joué optimalement dès le début. On s'intéresse à définir des stratégies qui ont un petit regret cumulé moyen $\mathbb{E}R_n$.

Remarquons que

$$\mathbb{E}R_n = n\mu^* - \mathbb{E}\sum_{t=1}^{n}\mu_{I_t} = \mathbb{E}\sum_{k=1}^{K} T_k(n)(\mu^* - \mu_k) = \mathbb{E}\sum_{k=1}^{K} T_k(n)\Delta_k,$$

où $\Delta_k = \mu^* - \mu_k$ est le "gap" entre le bras optimal et le bras k, et où $T_k(n) = \sum_{t=1}^{n} \mathbf{1}\{I_t = k\}$ est le nombre de fois que le bras k a été tiré jusqu'à l'instant n.

Ainsi un bon algorithme de bandit devra tirer peu souvent les bras sous-optimaux.

1.1 Stratégie UCB

La stratégie UCB (pour Upper Confidence Bound) [Auer et. al, 2002] consiste à choisir le bras:

$$I_t = \arg\max_k B_{t,T_k(t-1)}(k), \text{ avec } B_{t,s}(k) = \hat{\mu}_{k,s} + \sqrt{\frac{2\log t}{s}},$$

où $\hat{\mu}_{k,s} = \frac{1}{s}\sum_{i=1}^{s} x_{k,i}$ est la moyenne empirique des récompenses reçues en ayant tiré le bras k (i.e., $x_{k,i}$ est la i-ème récompense reçue en ayant tiré le bras k).

Il s'agit d'une stratégie dite "optimiste dans l'incertain". La valeur $B_{t,T_k(t-1)}(k)$ représente une borne supérieure de μ_k en forte probabilité. Ainsi on choisit le bras qui serait le meilleur si les valeurs des bras étaient les meilleures possibles (en forte proba), sachant ce qui a été obvervé.

En effet, rappelons l'inégalité de Chernoff-Hoeffding: Soient $X_i \in [0,1]$ variables aléatoires indépendantes de moyenne $\mu = \mathbb{E}X_i$. Alors

$$\mathbb{P}\big(\frac{1}{s}\sum_{i=1}^{s} X_i - \mu \geq \epsilon\big) \leq e^{-2s\epsilon^2}, \quad \text{et} \quad \mathbb{P}\big(\frac{1}{s}\sum_{i=1}^{s} X_i - \mu \leq -\epsilon\big) \leq e^{-2s\epsilon^2}. \tag{1}$$

Donc pour $1 \leq s \leq t$ fixés,

$$\mathbb{P}(\hat{\mu}_{k,s} + \sqrt{\frac{2\log t}{s}} \leq \mu_k) \leq e^{-4\log(t)} = t^{-4}. \tag{2}$$

Et aussi:

$$\mathbb{P}(\hat{\mu}_{k,s} - \sqrt{\frac{2\log t}{s}} \geq \mu_k) \leq e^{-4\log(t)} = t^{-4}. \tag{3}$$

Proposition 1. Chaque bras sous-optimal k est tiré en moyenne au plus

$$\mathbb{E}T_k(n) \leq 8\frac{\log n}{\Delta_k^2} + \frac{\pi^2}{3}$$

fois. Donc le regret cumulé moyen d'UCB est borné selon

$$\mathbb{E}R_n = \sum_k \Delta_k \mathbb{E}T_k(n) \leq 8 \sum_{k:\Delta_k>0} \frac{\log n}{\Delta_k} + K\frac{\pi^2}{3}.$$

Ce résultat établit que le regret cumulé moyen est logarithmique en n.

Proof. Intuition de la preuve: supposons qu'à l'instant t les moyennes empiriques des bras sont dans leur intervalle de confiance respectifs, c'est à dire

$$\mu_k - \sqrt{\frac{2\log t}{s}} \overset{(a)}{\leq} \hat{\mu}_{k,s} \overset{(b)}{\leq} \mu_k + \sqrt{\frac{2\log t}{s}}. \tag{4}$$

pour $s = T_k(t-1)$. Alors soit k un bras sous-optimal et k^* un bras optimal. Si le bras k est tiré à l'instant t, cela signifie que $B_{t,T_k(t-1)}(k) \geq B_{t,T_{k^*}(t-1)}(k^*)$, soit

$$\hat{\mu}_{k,s} + \sqrt{\frac{2\log t}{s}} \geq \hat{\mu}_{k^*,s^*} + \sqrt{\frac{2\log t}{s^*}}, \tag{5}$$

pour $s = T_k(t-1)$ et $s^* = T_{k^*}(t-1)$ donc d'après (4), $\mu_k + 2\sqrt{\frac{2\log t}{s}} \geq \mu^*$, c'est à dire $s \leq \frac{8\log t}{\Delta_k^2}$.

Maintenant, pour tout entier u, nous avons:

$$\begin{aligned} T_k(n) &\leq u + \sum_{t=u+1}^{n} \mathbf{1}\{I_t = k; T_k(t) > u\} \\ &\leq u + \sum_{t=u+1}^{n} \mathbf{1}\{\exists s : u < s \leq t, \exists s^* : 1 \leq s^* \leq t, B_{t,s}(k) \geq B_{t,s^*}(k^*)\} \end{aligned} \quad (6)$$

Maintenant, d'après le raisonnement précédent, l'évènement $\{B_{t,s}(k) \geq B_{t,s^*}(k^*)\}$ (c'est à dire (5)) implique que $s \leq \frac{8\log t}{\Delta_k^2}$ ou bien qu'une des deux inégalités (a) ou (b) dans (4) n'est pas satisfaite. Donc en choisissant $u = \frac{8\log(n)}{\Delta_k^2}$, on en déduit que (a) ou (b) n'est pas satisfaite. Mais d'après (2), l'inégalité (a) n'est pas satisfaite avec une probabilité $\leq t^{-4}$, et d'après (3) l'inégalité (b) n'est pas vraie avec une probabilité $\leq t^{-4}$.

En prenant l'espérance des deux cotés de (6),

$$\begin{aligned} \mathbb{E}[T_k(n)] &\leq \frac{8\log(n)}{\Delta_k^2} + \sum_{t=u+1}^{n} \Big[\sum_{s=u+1}^{t} t^{-4} + \sum_{s=1}^{t} t^{-4} \Big] \\ &\leq \frac{8\log(n)}{\Delta_k^2} + \frac{\pi^2}{3} \end{aligned}$$

□

Proposition 2. Nous avons la borne uniforme sur le regret

$$\mathbb{E}R_n \leq \sqrt{8Kn(\log n + \frac{\pi^2}{3})}$$

Proof. Par Cauchy-Schwarz,

$$\begin{aligned} \mathbb{E}R_n &= \sum_k \Delta_k \sqrt{\mathbb{E}T_k(n)} \sqrt{\mathbb{E}T_k(n)} \\ &\leq \sqrt{\sum_k \Delta_k^2 \mathbb{E}T_k(n) \sum_k \mathbb{E}T_k(n)} \\ &\leq \sqrt{8Kn(\log n + \frac{\pi^2}{3})}. \end{aligned}$$

□

1.2 Bornes inférieures

Nous avons la borne inférieure asymptotique (pour une classe de distributions assez riche) [Lai et Robbins, 1985] fonction des distributions:

$$\limsup_n \frac{\mathbb{E}T_k(n)}{\log n} \geq \frac{1}{KL(\nu_k||\nu^*)},$$

avec la distance de Kullback Leibler $KL(\nu||\nu') = \int d\nu \log(d\nu/d\nu')$. Donc $\mathbb{E}R_n = \Omega(\log n)$.

Nous avons aussi une borne inférieure minimax non-asymptotique (voir par exemple [Cesa-Bianchi et Lugosi, Prediction, Learning, and Games, 2006]):

$$\inf_{\text{Algo}} \sup_{\text{Problème}} R_n = \Omega(\sqrt{nK}).$$

1.3 Améliorations

- Utilisation de la variance empirique pour affiner les intervalles de confiance [Audibert, Munos, Szepesvari, Use of variance estimation in the multi-armed bandit problem, 2008].
- Bornes minimax améliorées [Audibert, Bubeck, Minimax Policies for Adversarial and Stochastic Bandits, 2009]. On obtient la borne inf $\sqrt{Kn}$.
- Application au "online learning" (classification ou régression en-ligne) avec information partielle.

1.4 Extensions

Il existe de nombreuses extensions au problème du bandit stochastique à K bras:

- **Bandit dans un MDP** [Jaksch, Ortner, Auer. Near-optimal regret bounds for reinforcement learning, 2010]. Stratégie d'exploration "optimiste dans l'incertain" (basée sur UCB) pour explorer un Processus de décision markovien. $->$ **possibilité de mini-projet**
- **Bandits avec information contextuelle**. A chaque instant t, on observe une information $x_t \in X$ et on prend une décision $a_t \in A$. La récompense est une foncion de a_t et de x_t. On se compare à une classe de stratégies $\pi : X \to A$.
- **Bandit avec un nombre de bras dénombrable**. [Wang, Audibert, Munos, Algorithms for inifinitely many-armed bandits, 2008]. Chaque nouveau bras a une probabilité ϵ^β d'être ϵ-optimal. Compromis entre exploration - exploitation - découverte. $->$ **possibilité de mini-projet**
- **Bandits linéaires** [Dani, Hayes, Kakade, Stochastic Linear Optimization under Bandit Feedback, 2008] On choisit un bras $x_t \in X \subset \mathbb{R}^d$. La récompense moyenne est une fonction linéaire $r_t = x_t \cdot \alpha$, où $\alpha \in \mathbb{R}^d$ est un paramètre inconnu. On se compare à $\max_{x \in X} x \cdot \alpha$. $->$ **possibilité de mini-projet**
- **Bandits en espace métrique** [Kleinberg, Slivkins, Upfal, Multi-armed bandits in metric spaces, 2008], [Bubeck, Munos, Stoltz, Szepesvari, Online optimization in X-armed bandits, 2008]. On choisit un bras $x_t \in X$ dans un espace métrique. La récompense moyenne $f(x_t)$ est supposée Lipschitz. On se compare à $\sup_{x \in X} f(x)$. Il s'agit d'un problème d'optimisation online.
- **Bandits hiérarchiques** Algorithmes UCT [Kocsis et Szepesvari. Bandit based monte-carlo planning., 2006], BAST [Coquelin et Munos, Bandit algorithms for tree search, 2007], HOO [Bubeck, Munos, Stoltz, Szepesvari, Online optimization in X-armed bandits, 2008]. Application au jeu de go (programme MoGo) [Gelly, Wang, Munos, Teytaud. Modification of UCT with patterns in monte-carlo go, 2006]. Utilisation d'algorithmes de bandits de manière hiérarchique pour effectuer une recherche dans des grands arbres. $->$ **possibilité de mini-projet**

2 Le bandit contre un adversaire

Ici les récompenses ne sont plus nécessairement i.i.d. mais choisies arbitrairement par un adversaire. Dans ce cas, on ne peut plus espérer faire aussi bien que si on avait connu à l'avance les récompenses (car l'adversaire n'a aucune envie de se laisser dévoiler). Mais on peut comparer les performances obtenues par notre algorithme aux performances obtenues par une classe de stratégies de comparaison, et tenter de faire presqu'aussi bien que la meilleure stratégie dans cette classe.

Nous pouvons distinguer 2 types de problèmes selon l'information reçue à chaque round: information parfaite ("full information") ou information partielle ("bandit information"). Voici un modèle de bandit contre un adversaire: On dispose de K bras. A chaque instant $t = 1 \ldots, n$,

- L'adversaire choisi des récompenses $x_t(1), \ldots, x_t(K)$ (supposées à valeurs dans $[0, 1]$, sans les dévoiler au joueur
- Le joueur choisit un bras I_t
- Dans le cas information parfaite, le joueur observe les récompenses de tous les bras $x_t(k)$, pour $k = 1 \ldots, K$.
- Dans le cas information partielle, le joueur n'observe que la récompense du bras choisi: $x_t(I_t)$.

La classe de stratégies de comparaison est l'ensemble des stratégies constantes. Ainsi le regret par rapport à la stratégie constante k est défini selon

$$R_n(k) = \sum_{t=1}^{n} x_t(k) - \sum_{t=1}^{n} x_t(I_t).$$

La stratégie du joueur pouvant être aléatoire, on considère le regret moyen pour la meilleure stratégie constante:

$$R_n = \max_{1 \le k \le K} \mathbb{E} R_n(k),$$

et l'on désire construire un algorithme qui soit bon pour toute séquence de récompenses fournies par l'adversaire, i.e. tel que

$$\sup_{x_1, \ldots, x_n} R_n$$

soit petit.

2.1 Information complete

Considérons l'algorithme EWF (Exponentially Weighted Forecaster):

- On initialise l'algorithme avec $w_1(k) = 1$ pour tout $k = 1, \ldots, n$.
- A chaque instant $t = 1, \ldots, n$, le joueur choisit le bras $I_t \sim p_t$, où $p_t(k) = \frac{w_t(k)}{\sum_{i=1}^{K} w_t(i)}$, avec

$$w_t(k) = e^{\eta \sum_{s=1}^{t-1} x_s(k)},$$

  où $\eta > 0$ est un paramètre de l'algorithme.

Proposition 3. Soit $\eta \le 1$. On a

$$R_n \le \frac{\log K}{\eta} + \frac{\eta n}{8}.$$

Ainsi, en choisissant $\eta = \sqrt{8 \frac{\log K}{n}}$, il vient

$$R_n \le \sqrt{\frac{n \log K}{8}}.$$

Preuve. Notons $W_t = \sum_{k=1}^K w_k(t)$. On a

$$\begin{aligned}\frac{W_{t+1}}{W_t} &= \sum_{k=1}^K \frac{w_k(t)e^{\eta x_t(k)}}{W_t} = \sum_{k=1}^K p_k(t)e^{\eta x_t(k)} = \mathbb{E}_{I\sim p_t}\left[e^{\eta x_t(I)}\right] \\ &= e^{\mathbb{E}[x_t(k)]}\mathbb{E}_{I\sim p_t}\left[e^{\eta(x_t(I)-\mathbb{E}_{J\sim p_t}[x_t(J)])}\right] \\ &\leq e^{\mathbb{E}[x_t(I)]}e^{\eta^2/8}, \text{ par le lemme de Hoeffding}\end{aligned}$$

Donc

$$\log\frac{W_{n+1}}{W_1} \leq \mathbb{E}\big[\sum_{t=1}^n x_t(I_t)\big] + n\frac{\eta^2}{8}.$$

Maintenant

$$\log\frac{W_{n+1}}{W_1} = \log\sum_{k=1}^K e^{\eta\sum_{t=1}^n x_t(k)} - \log K \geq \eta\sum_{t=1}^n x_t(k) - \log K,$$

pour n'importe quel $k = 1, \ldots, n$. Ainsi pour tout k, on a

$$\sum_{t=1}^n x_t(k) - \mathbb{E}\big[\sum_{t=1}^n x_t(I_t)\big] \leq \frac{\log K}{\eta} + n\frac{\eta}{8}.$$

□

Remarques:

- Bornes inf du même ordre: $R_n = \Omega(\sqrt{n\log K})$.
- Dépendance logarithmique en K
- Possibilité d'avoir un algorithme anytime (qui ne nécessite pas la connaissance de l'horizon temporel n) en choisissant un pas $\eta_t = O(\sqrt{\frac{\log K}{t}})$.
- Prédiction à base d'experts [Lugosi and Cesa-Bianchi, 2006].

2.2 Information partielle

Considérons l'algorithme EXP3 (Exploration-Exploitation using Exponential weights):

- $\eta > 0$ et $\beta > 0$ sont deux paramètres de l'algorithme.
- On initialise l'algorithme avec $w_1(k) = 1$ pour tout $k = 1, \ldots, n$.
- A chaque instant $t = 1, \ldots, n$, le joueur choisit le bras $I_t \sim p_t$, où

  $$p_t(k) = (1-\beta)\frac{w_t(k)}{\sum_{i=1}^K w_t(i)} + \frac{\beta}{K},$$

  avec

  $$w_t(k) = e^{\eta\sum_{s=1}^{t-1}\tilde{x}_s(k)},$$

  où $\tilde{x}_s(k) = \frac{x_s(k)}{p_s(k)}\mathbf{1}\{I_s = k\}$.

Proposition 4. Soit $\eta \leq 1$, et $\beta = \eta K$. On a

$$R_n \leq \frac{\log K}{\eta} + (e-1)\eta n K.$$

Ainsi, en choisissant $\eta = \sqrt{\frac{\log K}{(e-1)nK}}$, il vient

$$R_n \leq 2.63\sqrt{nK \log K}.$$

Preuve. Notons $W_t = \sum_{k=1}^K w_k(t)$. Remarquons que $\mathbb{E}_{I_s \sim p_s}[\tilde{x}_s(k)] = \sum_{i=1}^K p_s(i)\frac{x_s(k)}{p_s(k)}\mathbf{1}\{i=k\} = x_s(k)$, que $\mathbb{E}_{I_s \sim p_s}[\tilde{x}_s(I_s)] = \sum_{i=1}^K p_s(i)\frac{x_s(i)}{p_s(i)} \leq K$. On a

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{k=1}^K \frac{w_k(t)e^{\eta\tilde{x}_t(k)}}{W_t} = \sum_{k=1}^K \frac{p_k(t) - \beta/K}{1-\beta} e^{\eta\tilde{x}_t(k)} \\
&\leq \sum_{k=1}^K \frac{p_k(t) - \beta/K}{1-\beta}(1 + \eta\tilde{x}_t(k) + (e-2)\eta^2\tilde{x}_t(k)^2),
\end{aligned}$$

puisque $\eta\tilde{x}_t(k) \leq \eta K/\beta = 1$, et $e^x \leq 1 + x + (e-2)x^2$ pour $x \leq 1$.

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &\leq 1 + \frac{1}{1-\beta}\sum_{k=1}^K p_k(t)(\eta\tilde{x}_t(k) + (e-2)\eta^2\tilde{x}_t(k)^2), \\
\log\frac{W_{t+1}}{W_t} &\leq \frac{1}{1-\beta}\sum_{k=1}^K p_k(t)(\eta\tilde{x}_t(k) + (e-2)\eta^2\tilde{x}_t(k)^2), \\
\log\frac{W_{n+1}}{W_1} &\leq \frac{1}{1-\beta}\sum_{t=1}^n\sum_{k=1}^K p_k(t)(\eta\tilde{x}_t(k) + (e-2)\eta^2\tilde{x}_t(k)^2).
\end{aligned}$$

Mais on a aussi

$$\log\frac{W_{n+1}}{W_1} = \log\sum_{k=1}^K e^{\eta\sum_{t=1}^n \tilde{x}_t(k)} - \log K \geq \eta\sum_{t=1}^n \tilde{x}_t(k) - \log K,$$

pour n'importe quel $k = 1, \ldots, n$. En prenant l'espérance (par rapport à la randomisation interne de l'algorithme), on a pour tout k,

$$\begin{aligned}
\mathbb{E}\Big[(1-\beta)\sum_{t=1}^n \tilde{x}_t(k) - \sum_{t=1}^n\sum_{i=1}^K p_i(t)\tilde{x}_t(i)\Big] &\leq (1-\beta)\frac{\log K}{\eta} + (e-2)\eta\mathbb{E}\Big[\sum_{t=1}^n\sum_{k=1}^K p_k(t)\tilde{x}_t(k)^2\Big] \\
\mathbb{E}\Big[\sum_{t=1}^n x_t(k) - \sum_{t=1}^n x_t(I_t)\Big] &\leq \beta n + \frac{\log K}{\eta} + (e-2)\eta n K \\
\mathbb{E}[R_n(k)] &\leq \frac{\log K}{\eta} + (e-1)\eta n K
\end{aligned}$$

□

Remarques:

- Bornes inf $R_n \geq \frac{1}{20}\sqrt{nK}$ (voir [Auer et al., 2002]).

- Algorithme INF [Audibert et Bubeck 2009] obtient un regret $O(\sqrt{nK})$.
- Exp3 donne un regret en espérance en $O(\sqrt{nK \log K})$ mais le regret peut être grand sur certaines réalisations de l'algorithme. En effet les récompenses repondérées $\tilde{x}_t(k)$ sont d'ordre $1/p_t(k)$ ce qui peut être (en négligeant la dépendance en K) aussi grand que $\sqrt{n}$ pour les valeurs de β considérées, ce qui entraîne un regret dont la variance est en $n^{3/2}$ et ainsi le regret peut être en $n^{3/4}$. On ne peut donc pas obtenir de regret en $\sqrt{n}$ avec forte probabilité. Cependant en modifiant légèrement l'algorithme afin d'augmenter l'exploration (en surestimant les récompenses reçues), on obtient un algorithme EXP3.P [Auer et al., 2002] qui dépend d'un paramètre $\delta > 0$ tel que

$$\max_{1 \leq k \leq K} \sum_{t=1}^{n} x_t(k) - \sum_{t=1}^{n} x_t(I_t) = O(\sqrt{nK \log(Kn/\delta)}). \tag{7}$$

2.3 Equilibre de Nash

Consider a 2-players zero-sum repeated game:

A and B play actions: 1 or 2 simultaneously, and receive the reward (for A):

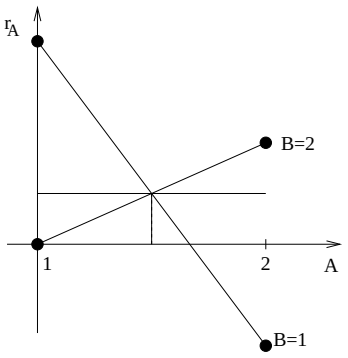
A \ B	1	2
1	2	0
2	-1	1

(A likes consensus, B likes conflicts)

Nash equilibrium: (mixed) strategy (p^*, q^*) for both players, such that no player has incentive for changing unilaterally his own strategy: for any p and q,

$$r_A(p, q^*) \leq r_A(p^*, q^*) \text{ and } r_B(p^*, q) \leq r_B(p^*, q^*).$$

Here the Nash is: A plays 1 with probability $p_A = 1/2$, B plays 1 with probability $p_B = 1/4$.



Now, let A and B be bandit algorithms, aiming at minimizing their regret, i.e. for player A:

$$R_n(A) \stackrel{\text{def}}{=} \max_{a \in \{1,2\}} \sum_{t=1}^{n} r_A(a, B_t) - \sum_{t=1}^{n} r_A(A_t, B_t).$$

and that of B accordingly.

In the case of zero-sum games (i.e. $r_A = -r_B$), the value of the game V is defined as

$$\begin{aligned} V &= \inf_{q \in [0,1]} \sup_{p \in [0,1]} r_A(p, q) \\ &= \sup_{p \in [0,1]} \inf_{q \in [0,1]} r_A(p, q) \end{aligned}$$

(minimax Theorem). Note that there may exists several Nash equilibria (p^*, q^*) but their value $r_A(p^*, q^*)$ is unique and equals V^*. Indeed:

$$V^* = \sup_{p\in[0,1]} \inf_{q\in[0,1]} r_A(p,q) \geq \inf_{q\in[0,1]} r_A(p^*,q) = r_A(p^*,q^*) = \sup_{p\in[0,1]} r_A(p,q^*) \geq \inf_{q\in[0,1]} r_A(p,q) = V.$$

Proposition 5. If both players perform a (Hannan) consistent regret-minimization strategy (i.e. $R_n(A)/n \to 0$ and $R_n(B)/n \to 0$), then the average value of the received rewards converges to the value of the game:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} r_A(A_t, B_t) = V.$$

In addition, the empirical frequencies of chosen actions of both players converge to a Nash equilibrium. (Note that EXP3.P is consistent!).

Preuve. Write $p_A^n \overset{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{A_t=1}$ and $p_B^n \overset{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{B_t=1}$ the empirical frequencies of played actions. Write $r_A(p,q) \overset{\text{def}}{=} \mathbb{E}_{A\sim p, B\sim Q} r_A(A,B)$.

Player A plays a Hannan-consistent regret-minimization strategy, thus

$$\begin{aligned}
\limsup_{n\to\infty} \max_{a\in\{1,2\}} \frac{1}{n}\sum_{t=1}^{n} r_A(a,B_t) - \frac{1}{n}\sum_{t=1}^{n} r_A(A_t,B_t) &\leq 0 \\
\limsup_{n\to\infty} \sup_{p\in[0,1]} \frac{1}{n}\sum_{t=1}^{n} r_A(p,B_t) - \frac{1}{n}\sum_{t=1}^{n} r_A(A_t,B_t) &\leq 0
\end{aligned}$$

since $p \to r_A(p,q)$ is a linear mapping. Thus

$$V = \inf_q \sup_p r_A(p,q) \leq \liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} r_A(A_t,B_t).$$

Now, since player B does the same, we also have that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} r_A(A_t,B_t) \leq \sup_p \inf_q r_A(p,q) = V.$$

Thus $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} r_A(A_t,B_t) = V$.

We also have that the empirical joint distribution $(p\times q)_n(a,b) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{A_t = a, B_t = b\}$ converges to the set of Nash equilibria $\{(p^*,q^*)\}$ since the set of joint distributions is compact (since we consider a finite number of actions) and any (p',q') such that $r_A(p',q') = V$ is a Nash equilibrium:

$$r_A(p',q') \geq \inf_{q\in[0,1]} r_A(p',q) \geq V^* \geq \sup_{p\in[0,1]} r_A(p,q') \geq r_A(p',q').$$

□

Example: Texas Hold'em Poker

- In the 2-players Poker game, the Nash equilibrium is interesting (zero-sum game)
- A policy:
  information set (my cards + board + pot) → probabilities over decisions (check, raise, fold)
- Space of policies is huge!

Idea: Approximate the Nash equilibrium by using bandit algorithms assigned to each information set.

This method provided the world best Texas Hold'em Poker program for 2-player with pot-limit [Zinkevitch et al., 2007].

Extension: In non-zero sum games or in games with more than 2 players, there are general convergence results towards correlated equilibrium [Foster and Vohra, 1997].