TP 3 : Résolution de MDP à espace d'état infini

Le Gypaète Barbu (Gypaetus barbatus Fig. 1) est un oiseau classé dans la liste rouge au bord de l'extinction en Suisse. Les autorités vous demandent de les aider à concevoir un plan

Figure 1 – Gypaetus Barbatus

optimisé de protection de cette espèce. Vous leur proposez une modélisation par un processus markovien décisionnel de caractéristiques suivantes :

- l'espace d'états est $\mathbb{X} = \mathbb{N}$, un état correspondant à un nombre de gypaètes,
- l'espace d'actions est $\mathbb{A} = \{0, 1, 2\}$:
  - l'action 0 correspond à ne rien faire,
  - l'action 1 correspond à mettre en place des mesures actives de protection de l'espèce qui vont favoriser les naissances d'oiseaux,
  - l'action 2 correspond à mettre en place des mesures actives de protection de l'espèce qui vont favoriser la survie des oiseaux existants,
- l'espace de contraintes est $\mathbb{K} = \mathbb{X} \times \mathbb{A}$, toutes les actions sont possibles dans tous les états,
- La matrice de transitions Q sur $\mathbb{X}$ sachant $\mathbb{K}$ est définie de la façon suivante : un pas de temps correspond à une année, et pour $x = n$ et $a \in \mathbb{A}$, la loi $Q(\cdot|x, a)$ est la loi de
$$n + B^a - D^a,$$
où B^a et D^a sont des variables aléatoires indépendantes de loi binomiale de paramètres respectifs (n, p_b^a) et (n, p_d^a). En effet, dans une année chaque oiseau peut donner naissance à un nouvel individu avec probabilité p_b^a et mourir avec probabilité p_d^a, indépendamment les uns des autres. En considérant que les mesures de protections sont efficaces, on a $p_b^0 = p_b^2 < p_b^1$ et $p_d^0 = p_d^1 > p_d^2$,
- la fonction de récompense instantanée est $c : \mathbb{K} \to \mathbb{R}$ définie pour tout $x \in \mathbb{X}$ par
  - $c(x, 0) = 0$: ne rien faire ne coûte rien et ne rapporte rien,
  - $c(x, 1) = -\alpha_1$, $c(x, 2) = -\alpha_2$: protéger génère un coût annuel fixe dépendant de l'action choisie.
- la fonction de récompense terminale $C : \mathbb{X} \to \mathbb{R}$ est définie pour tout $x \in \mathbb{X}$ par $C(x) = x$ si $x > 0$ et $C(0) = -\alpha_0$ car plus le nombre d'oiseaux est élevé en fin de programme, mieux le plan de protection a marché et on ne veut éviter (si possible) l'extinction de l'espèce.
- l'horizon d'optimisation est $N = 20$ ans.

Après discussion avec les expertes et experts, et les autorités vous choisissez les valeurs des paramètres de la Table 1. Au début de la période d'optimisation, il y a $X_0 = 2$ gypaètes en Suisse.

Simulations

1. Construire un simulateur du MDP pour la politique π_0 qui consiste à ne jamais rien faire.

Table 1 – Paramètres du MDP

$p_b^0 = p_b^2 = 0.45$	$p_d^0 = p_d^1 = 0.45$	$p_b^1 = 0.5$	$p_d^2 = 0.35$	$\alpha_0 = 5$	$\alpha_1 = 0.1$	$\alpha_2 = 0.2$

2. Tracer quelques trajectoires du MDP contrôlé par la politique π_0. Commenter.
3. A l'aide de votre simulateur, estimer le coût de la politique π_0 par la méthode de Monte Carlo.
4. Construire un simulateur du MDP pour la politique π_1 qui consiste à choisir l'action 1 à chaque pas de temps.
5. Tracer quelques trajectoires du MDP contrôlé par la politique π_1. Commenter.
6. A l'aide de votre simulateur, estimer le coût de la politique π_1 par la méthode de Monte Carlo.
7. Construire un simulateur du MDP pour la politique π_2 qui consiste à choisir l'action 2 à chaque pas de temps.
8. Tracer quelques trajectoires du MDP contrôlé par la politique π_2. Commenter.
9. A l'aide de votre simulateur, estimer le coût de la politique π_2 par la méthode de Monte Carlo.

Programmation dynamique

On cherche maintenant à mettre en oeuvre l'algorithme de programmation dynamique pour calculer la fonction valeur et une politique optimale.

10. Si $X_0 = 2$, quel est le nombre $N_{\max}$ maximum d'oiseaux qu'on pourra obtenir en 20 années quelle que soit la stratégie choisie ?
11. Est-il envisageable de faire tourner l'algorithme de programmation dynamique sur un espace d'état de cardinal $N_{\max}$?

On modifie un peu la dynamique du processus pour réduire la taille de l'espace d'états. On pose maintenant $\mathbb{X} = \{0, 1, \dots, M\}$ où le dernier état correspond maintenant à M *individus ou plus.* Pour $x = n \in \mathbb{X}$ et $a \in \mathbb{A}$, $Q(\cdot|x, a)$ est donc la loi de

$$\min\{M, n + B^a - D^a\}.$$

Tous les autres paramètres du MDP sont inchangés. On cherche maintenant à calibrer M.

12. Utilisez vos simulateurs pour calibrer une valeur de M satisfaisante.
13. Pour la valeur de M que vous avez choisie, construire numériquement les matrices Q.
14. Implémenter l'algorithme de programmation dynamique.
15. Quelles est la performance optimale pour $X_0 = 2$?
16. Commenter la forme de la politique optimale que vous avez obtenue.