Projet Mathématique et informatiques : Modèle d'Eden Richardson

ALLAIN Maëliss et LI Afonso
Projet dirigé par LUCAS Cyrille

3 Juin 2021

Table des matières

1 PRÉSENTATION DU MODELE D'EDEN-RICHARDSON

Nous allons étudier le modèle d'Eden-Richardson. Ce modèle a été présenté en 1961 par Eden pour expliquer la prolifération d'une tumeur, mais il s'applique à la croissance de clusters en tout genre comme les épidémies ou encore les feux de forêts.
Notre objectif dans ce projet est d'étudier le comportement de ce modèle par rapport aux différents paramètres qui sont donnés.
Dans un premier temps nous étudierons ce modèle unidimensionnellement, c'est à dire sur $\mathbb{Z}$ puis nous nous placerons dans $\mathbb{Z}^2$
Pour commencer, expliquons comment fonctionne ce modèle.
On pose E_n une suite de sous-ensembles de $\mathbb{Z}^d$ avec $E_0=\{0\}$. À chaque itération n, on regarde les voisins les plus proches de chaque élément de la suite E_{n-1} et on les ajoute à E_n, indépendamment des uns et des autres, avec une probabilité **p**.
On remarque le caractère markovien du premier ordre de E_n, puisque pour déterminer E_{n+1}, nous aurons besoin que de E_n, i.e. que de l'ensemble précédent.

2 ÉTUDE DU MODÈLE DANS $\mathbb{Z}$

Nous pouvons désormais étudier notre modèle dans $\mathbb{Z}$. Nous avons commencé par simuler sur ordinateur l'évolution de E_n jusqu'à l'étape 100000 en faisant varier la probabilité p.

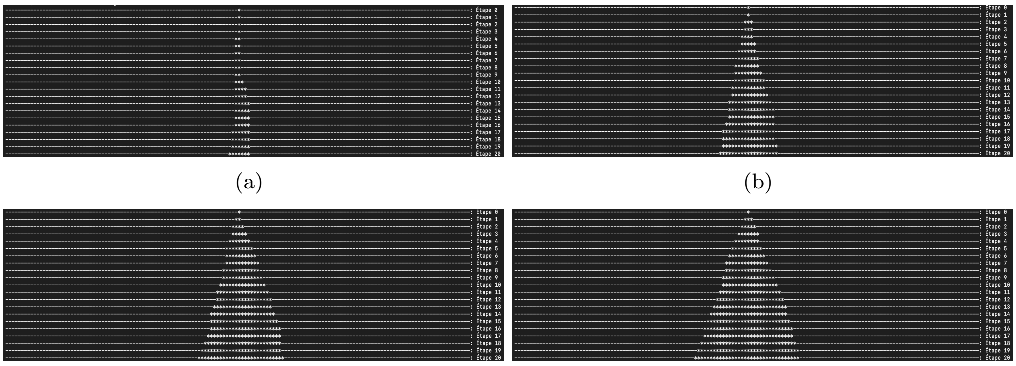
```python
def evolution(proba,n) :
    point_gauche = 0
    point_droite = 0
    for _ in range(n) :
        proba_gauche = random()
        proba_droite = random()
        if(proba_gauche <= proba) :
            point_gauche = point_gauche-1
        if(proba_droite <= proba) :
            point_droite = point_droite+1
    res = "["+str(point_gauche)+";"+str(point_droite)+"]"
    print(res)
```

FIGURE 1 – Code en Python qui affiche l'intervalle occupé par E_n

Cela semble semble logique de penser que l'ensemble des sites occupés est un intervalle d'entiers. Mais nous allons le montrer de façon plus précise.

Définition : Soit I une partie de $\mathbb{Z}$. On dit que I est un intervalle si pour tout x, y appartenant à I tels que $x<y$ et pour tout z tel que $x<z<y$ alors z est un élément de I.
Montrons que E_n est un intervalle comme défini ci-dessus.
Nous allons procéder par récurrence.
Initialisation : Soit n=0, nous avons que $E_0 = \{0\}$, donc l'ensemble E_0 est bien un intervalle d'entiers.
Hérédité : Supposons que la proposition est vraie au rang n cela signifie que E_n est un intervalle d'entiers. Montrons que E_{n+1} est un intervalle d'entiers.
À l'instant n+1, chaque point ayant un plus proche voisin dans E_n est ajouté dans E_{n+1} avec une probabilité p. On se situe dans Z, alors a est un voisin proche de b si a+1=b ou a-1=b. Si un singleton est ajouté dans notre E_n, qui est de la forme $\{\ldots,a\text{-}2,a\text{-}1,a,a+1,a+2,\ldots\}$, on est bien en train d'ajouter un plus proche voisin.
Alors l'ensemble des sites occupés est bien un intervalle d'entiers.
Conclusion : Par héridité, nous avons montré que E_n est un intervalle.

En suivant des probabilités variées, l'ensemble évolue de façon différente, il s'agit aussi de notre but d'étude. Visualisons ceci avec des affichages provanant d'un programme en Java.



(a) (b) (c) (d)

FIGURE 2 – (a) p=0.1 (b)p=0.3 (c) p=0.7 (d) p=0.9

Peut-on mieux comprendre cet ensemble E_n ? La réponse est évidemment que oui. En effet on peut réécrire la suite E_n de la façon suivante : $E_n = \{G_n,...,D_n\}$, cela signifie que E_n est un rassemblement des points qui sont à droite et à gauche de l'origine. Ce qui semble logique. Mais nous pouvons remarquer que les D_n suivent une loi binomiale de paramètres n et p. Si nous arrivons à comprendre son comportement quand n tend vers l'infini, c'est à dire laisser l'ensemble E_n évoluer infinimment, alors nous pouvons apporter des informations sur E_n.

Ainsi nous allons montrer que

$$\lim_{n\to\infty} \frac{D_n}{n} = p$$

Nous pouvons d'abord décomposer D_n :

$$D_n = \sum_{k=1}^{n} X_k$$

$\forall k$, X_k suit une loi de Bernoulli de paramètre p.
Revenons sur ce qui nous intéresse

$$\lim_{n\to\infty} \frac{D_n}{n} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} X_k}{n}$$

D'après la loi forte des grands nombres.

$$\lim_{n\to\infty} \frac{D_n}{n} = \mathbb{E}(X_k)$$

$$\lim_{n\to\infty} \frac{D_n}{n} = p$$

Conclusion : La probabilité p est obtenue avec la limite que nous venons de calculer. Et ceci va prédire la taille de notre intervalle.

Pour conclure notre étude sur $\mathbb{Z}$, nous avons cherché à faire un histogramme de $\frac{D_t}{t}$ pour un t fixé, que nous avons choisi égal à 10^4.
Nous avons alors effectué 10^3 réalisations de copies indépendantes de la variable aléatoire. Chaque bâton représente le nombre de réalisations dont le résultat est compris dans l'intervalle sur lequel le bâton est situé. Nous nous sommes placés dans un premier temps sur l'intervalle [0,1] puis nous nous sommes concentrés sur un intervalle plus petit centré en $\frac{1}{2}$ car toutes nos réalisations se situaient dans cet intervalle.

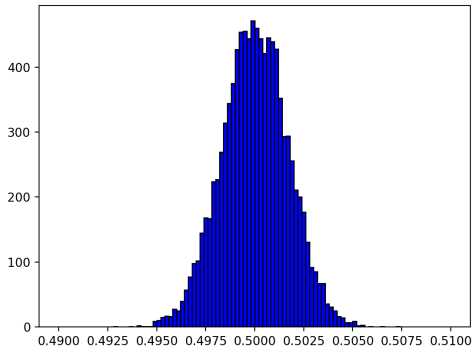


FIGURE 3 – Histogramme de l'expérience proposée

Hmmm étrange... On dirait la densité d'une loi normale. Nous allons alors centrer notre histogramme sur l'origine et le comparer à une courbe de Gauss.

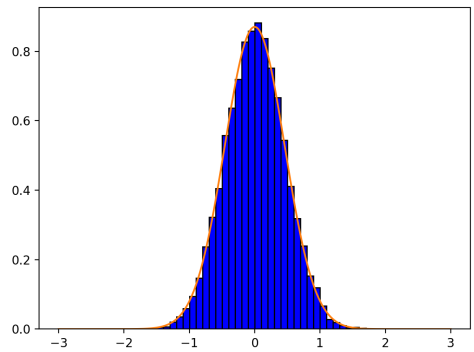


FIGURE 4 – Histogramme centralisé à l'origine et comparaison avec une courbe gaussienne

La courbe représente la loi normale qui dépend de deux paramètres : son espérance, un nombre réel noté μ, et son écart type, un nombre réel positif noté σ, dans notre cas, la fonction de densité de la loi est :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$$

où μ=0 et σ=0.5(1-0.5)=0.25 dans notre cas.

3 ÉTUDE DU MODÈLE DANS $\mathbb{Z}^2$

Nous venons d'étudier notre modèle dans $\mathbb{Z}$, nous pouvons alors nous placer dans $\mathbb{Z}^2$. Commençons par un schéma explicatif pour notre modèle :

Nous avons commencé par simuler sur ordinateur notre modèle pour différentes valeurs de p afin d'avoir un aperçu de son évolution dans $\mathbb{Z}^2$.

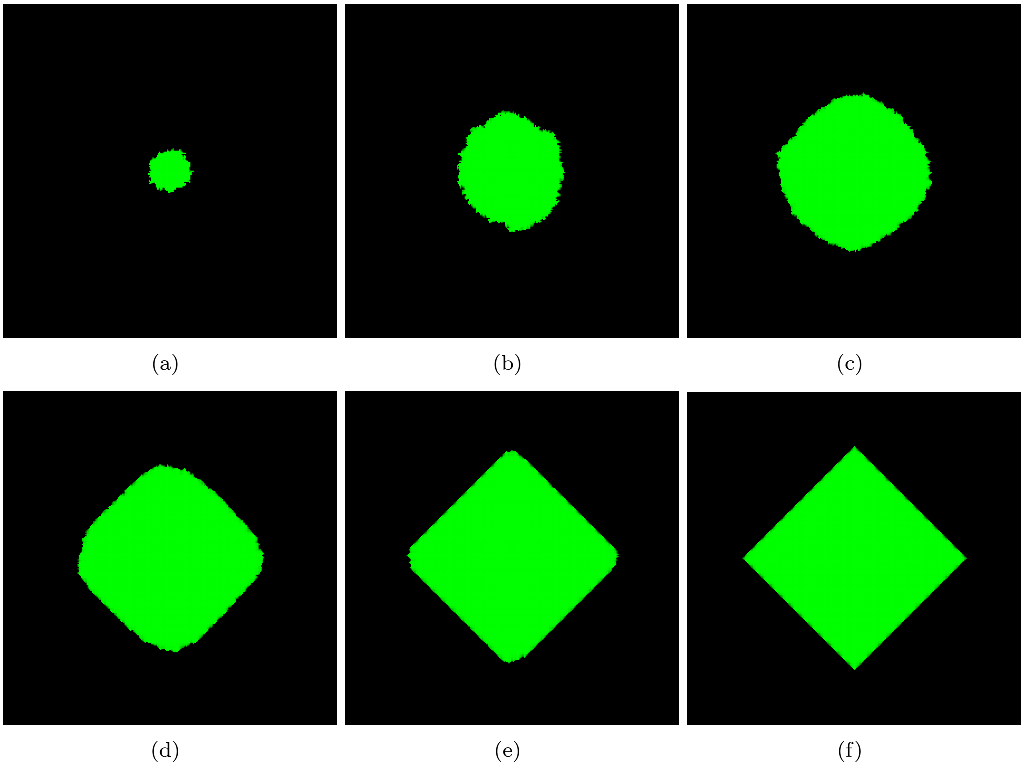


FIGURE 5 – Évolution de l'ensemble E_n en dimension 2 suivant des probabilités différentes (a) p=0.1 (b)p=0.3 (c) p=0.5 (d) p=0.7 (e) p=0.9 (f) p=1

3.1 Une évolution proportionnelle à n^2

Pour chaque expérience montrée ci-dessus, nous avons calculé le rapport $\frac{\#E_{200}}{\#E_{100}}$, le résultat donne toujours 4 en arrondi. Nous expliquons cela car nous nous plaçons en dimension 2. Si on multiple la largeur et la longueur par un entier k, l'aire de l'objet après transformation sera k^2 plus grand que l'originale.

3.2 Convergence presque sûrement vers une forme limite

Lorsque nous laissons le programme tourner, nous remarquons que sa forme se converge très vite vers une forme limite. Voici 6 figures du même ensemble E_n avec paramètre 0.7 comme probabilité pour contaminer ses voisins, mais aux temps différents.

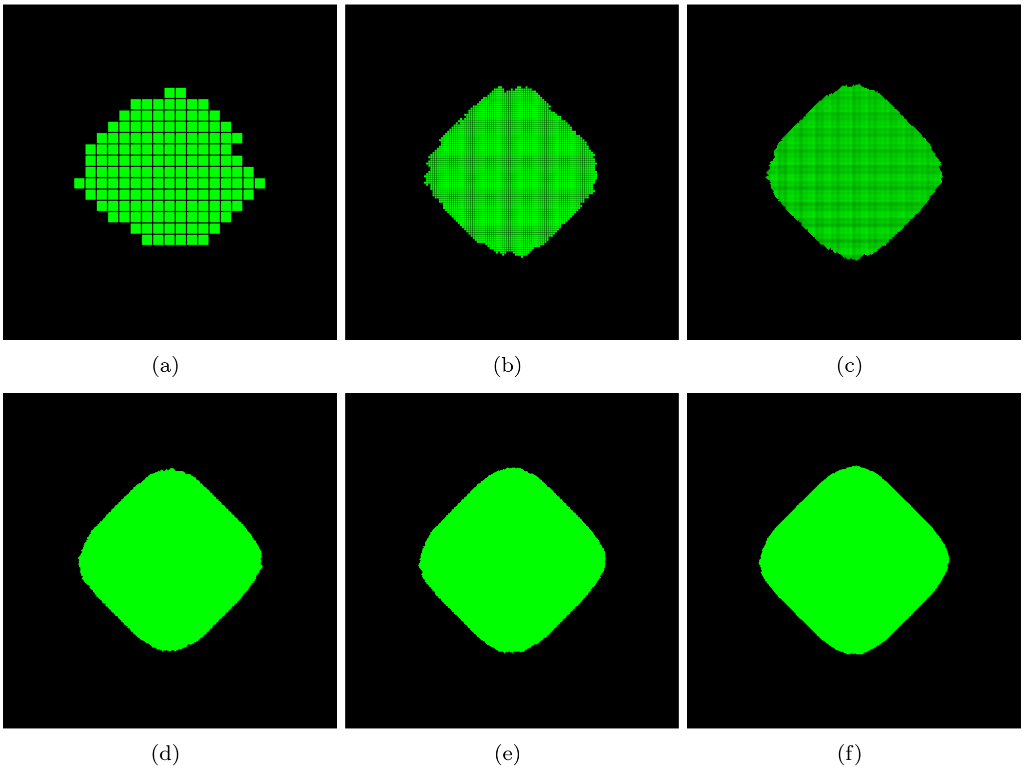


FIGURE 6 – Évolution de l'ensemble E_n en dimension 2 aux temps différés.
(a) 10 itérations (b) 50 itérations (c) 100itérations (d) 200 itérations (e) 300 itérations (f) 500 itérations

4 COMPARAISON AVEC UN ARBRE GÉNÉALOGIQUE

4.1 Brève explication et but

Soit ε_n un ensemble défini comme ceci $\varepsilon_n := \{(x,y) \in E_n \mid x+y=n\}$.
Nous allons maintenant construire un arbre généalogique τ, de la façon suivante : à la première génération, l'arbre ne contient qu'un seul individu. Puis à chaque génération, chaque individu peut avoir exactement un descendant avec une probabilité de **2p(1-p)**, deux descendants avec une probablité de $\mathbf{p}^2$, et l'individu peut avoir zéro enfant avec $\mathbf{(1\text{-}p)}^2$. Nous allons noter T_n le nombre d'enfants à la génération n.

Notre but final est de comparer cette variable aléatoire avec l'ensemble ε_n que nous allons expliquer son intérêt plus tard. Mais commençons par s'intéresser au comportement de T_n. Que vaut par exemple son espérance ?

4.2 Calculs

$$\mathbf{E}(T_n) = (2p)^n$$

Comment le trouver ? Nous allons raisonner par récurrence.

Initialisation : n=0,

$$\mathbf{E}(T_0) = 1$$

Supposons l'hypothèse vraie, montrons au rang n+1
Nous aurons besoin de

$$\mathbf{E}(T_{n+1}|T_n = k) = 2pk$$

Comment peut-on trouver ce résultat ? Déjà, nous allons interpréter l'expression $\mathbf{E}(T_{n+1}|T_n = k)$, En sachant qu'à la génération n, nous avons k enfants, quelle sera la moyenne de nombre d'enfants que nous pouvons obtenir à partir de ces k enfants qui sont devenu des parents à la génération n+1 ? On rappelle que chaque individu peut avoir exactement un enfant avec une probabilité **2p(1-p)**, deux enfants avec probabilité $\mathbf{p}^2$ et zéro enfant avec une probabilité $\mathbf{(1\text{-}p)}^2$. Ainsi nous pouvons décomposer T_{n+1} en somme finie de variables aléatoires Y_i^n où chaque Y_i^n va suivre la loi suivante, indépendantes les unes des autres :

$$\begin{aligned} P(Y_i^n = 0) &= (1-p)^2 \\ P(Y_i^n = 1) &= 2p(1-p) \\ P(Y_i^n = 2) &= p^2 \\ \text{Ainsi, nous avons : } T_{n+1} &= \sum_{i=1}^{T_n} Y_i^n \end{aligned}$$

Ainsi nous pouvons traduire $\mathbf{E}(T_{n+1}|T_n = k)$ de la façon suivante :

$$\begin{aligned} E(T_{n+1}|T_n = k) &= E(\sum_{i=1}^{T_n} Y_i^n | T_n = k) \\ &= E(\sum_{i=1}^{k} Y_i^n) \\ &= \sum_{i=1}^{k} E(Y_i^n) \\ &= \sum_{i=1}^{k} (0 * (1-p)^2 + 1 * 2p(1-p) + 2 * p^2) \\ &= \sum_{i=1}^{k} 2p \\ &= 2pk \end{aligned}$$

Nous sommes maintenant capable de trouver par récurrence le résultat de $\mathbf{E}(T_{n+1})$

$$
\begin{aligned}
\textbf{Prérequis : } E(T_{n+1}|T_n = k) &= \frac{E(T_{n+1} * 1_{T_n=k})}{P(T_n = k)} \\
\textbf{Sachant que : } E(T_{n+1}) &= E(T_{n+1} \sum_{n=0}^{2^n} 1_{T_n=k}) \\
&= \sum_{n=0}^{2^n} E(T_{n+1} 1_{T_n=k}) \\
&= \sum_{n=0}^{2^n} E(T_{n+1}|T_n = k) P(T_n = k) \\
&= \sum_{n=0}^{2^n} 2pkP(T_n = k) \\
&= 2p \sum_{n=0}^{2^n} kP(T_n = k) \\
&= 2pE(T_n) = (2p)^{n+1}
\end{aligned}
$$

Ainsi par récurrence, nous avons montré que $\mathbb{E}(\mathrm{T}_n)=(2\mathrm{p})^n$.

4.3 Comportement de l'arbre suivant certaines probabilités

Montrons que

$$\mathbf{P}(T_n > 0) \leq (2p)^n$$

Qui va nous aider à déterminer les comportements essentiels de T_n.
On emploie l'inégalité de Markov, soit α>0

$$\mathbf{P}(T_n \geq \alpha) \leq \frac{E(T_{n+1})}{\alpha}$$

Soit α=1,

$$\mathbf{P}(T_n \geq 1) \leq E(T_n)$$

$$\mathbf{P}(T_n > 0) \leq E(T_n)$$

$$\mathbf{P}(T_n > 0) \leq (2p)^n$$

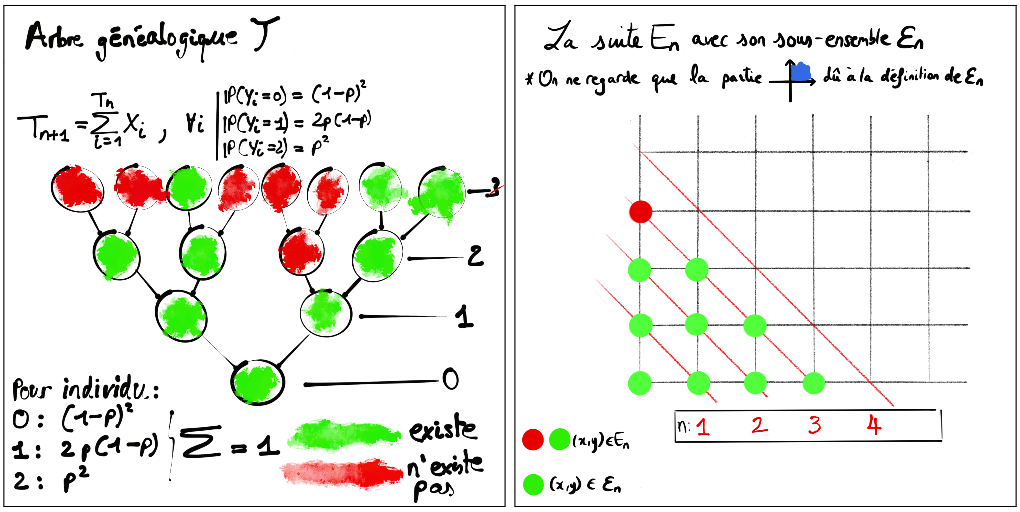
Montrons que si la probabilité p est strictement inférieure à $\frac{1}{2}$, la probabilité que l'arbre s'éteigne est 1.
On pose T un ensemble qui est défini de la façon suivante :
T := $\{\omega \in \Omega \ \omega$ appartient à une infinité de (T_n>0) $\}$.
Par le lemme de Borel-Cantelli en supposant que p<$\frac{1}{2}$, nous avons que $\sum_n$ P(T_n>0)< ∞ car d'après ce que nous avons calculé avant, on peut majorer la somme par une suite géométrique avec sa raison qui est strictement inférieure à 1 donc c'est sommable ce qui implique que

$\sum_n$ P(T$_n$>0) est fini.

Ainsi P(T)=0, ceci signifie que la probabilité que au plus un nombre fini de (T$_n$>0) se réalise est égal à 1, donc T$_n$ va se stationner à 0.
Donc avec une probabilité p $<\frac{1}{2}$, la probabilité d'extinction de l'arbre τ vaut 1.

4.4 Comparaison avec un sous-ensemble de ε_n

Nous rappellons que nous avons construit l'ensemble ε_n dans la section 4.1 sous cette forme : ε_n :={(x,y)∈E$_n$ | x+y=n}. Quel est le rapport de cet ensemble avec T$_n$?
Visualisons ceci tout de suite avec des schémas représentant ε_n et T$_n$.



(a) (b)

FIGURE 7 – Répresentation en schéma des ensembles E$_n$ et τ (a) arbre généalogique (b) E$_n$ avec ε_n

Nous pouvons établir un lien entre les deux ensembles. Nous savons déjà que dans l'arbre généalogique τ, chaque individu peut posséder au plus 2 enfants. Quant à l'ensemble ε_n, pour chaque instant n, les points qui se situent le plus loin de l'origine, peuvent contaminer au plus deux voisins. Vous allez vous demander : pourquoi pas trois ? Ceci est dû aux caractéristiques de l'ensemble ε_n, tout point (x,y) appartenant à ε_n doit vérifier x+y=n à l'instant n. Cette équation est une fonction affine passant par n quand x vaut 0 avec son coefficient qui vaut -1. Et il faut faire attention pour les contaminations des points appartenant à ε_n. Nous avons dit que tout point peut avoir au plus deux points pour contaminer, mais si un point a déjà été contaminé, on ne peut pas le re-contaminer.
Nous allons reprendre T$_{n+1}$ et majorer $\#\varepsilon_{n+1}$ de la façon suivante

$$\begin{aligned} T_{n+1} &= \sum_{i=1}^{T_n} Y_i^n \\ \#\varepsilon_{n+1} &\leq \sum_{i=1}^{\#\varepsilon_n} Y_i^n \end{aligned}$$

Nous voyons que T_{n+1} et $\#\varepsilon_n$ sont composés des mêmes éléments c'est à dire Y_i^n, nous pouvons déduire que toutes les deux sont construites sur un même espace probabilisé $(\Omega,\mathcal{A},\mathcal{P})$ où pour tout $\omega \in \Omega$, nous avons $\#\varepsilon_n(\omega) \leq \mathrm{T}_{n+1}(\omega)$.
Nous pouvons alors comparer T_{n+1} et $\#\varepsilon_{n+1}$ et déduire que $\#\varepsilon_n <_{st} \mathrm{T}_n$, cela siginifie que T_n domine stochastiquement $\#\varepsilon_n$.

Ainsi, nous pouvons prédire le comportement de ε_n grâce à la dominance stochastique par T_n. Nous avons dans la sous section 4.3 que si la probabilité p donnée à l'arbre est strictement inférieure à $\frac{1}{2}$ alors la probabilité d'extinction de l'arbre vaut 1 donc un événement certain.
Nous avons $\#\varepsilon_n <_{st} \mathrm{T}_n$ donc l'ensemble ε_n va lui être vide à partir d'un certain rang. Pour mieux comprendre, nous avons produit des schémas avec des algorithmes en Java.

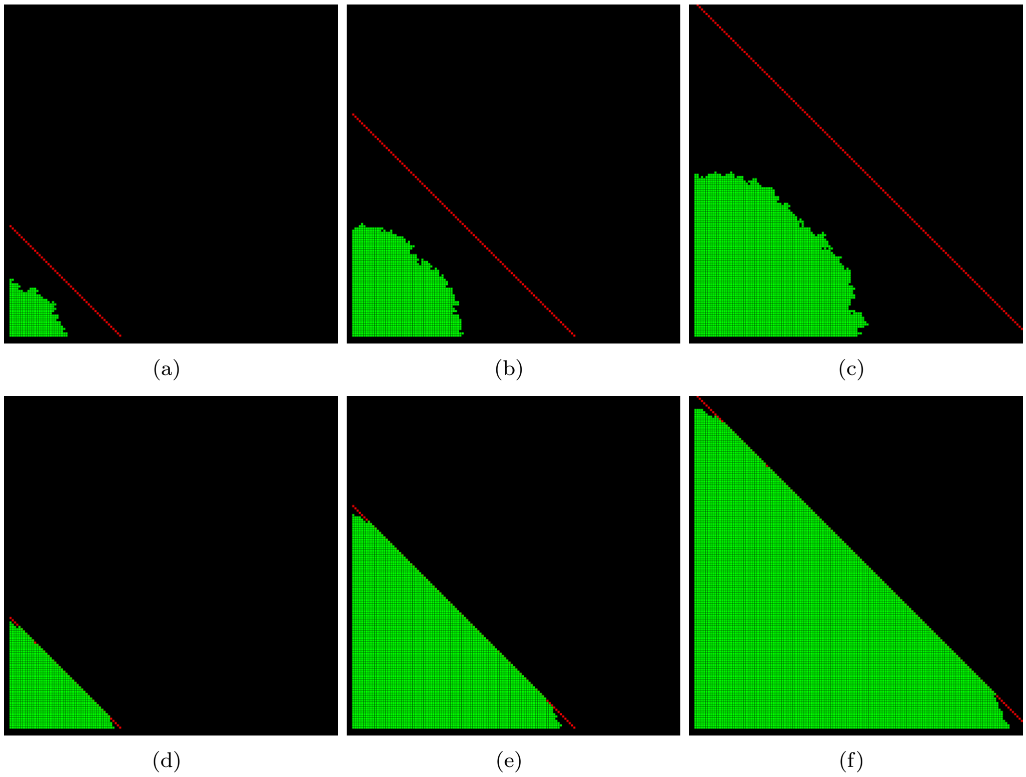


FIGURE 8 – Évolution de l'ensemble E_n en suivant des probabilités différentes avec la droite y=-x+n, (1ère ligne) p=0.3 (2ème ligne) p=0.9

Observation : Quand nous mettons 0.3 en tant que paramètre de l'algorithme, l'ensemble E_n représenté en vert a du mal à suivre la droite d'équation y=-x+n, et lorsque la probabilité p est 0.9, E_n reste collé à la droite rouge.

Nous pouvons alors nous demander une question, pour quelle valeur p(probabilité) que l'on peut fournir à E_n pour qu'il reste toujours collé à la droite rouge ?
En faisant environ 20 simulations pour chaque probabilité donnée entre 0.6 et 0.7, nous avons obtenu des données intéressantes que l'on met sous forme de tableau.

Probabilité donnée en paramètre	Probabilité de survie au temps 200	Probabilité de survie au temps 500
0.6	10%	5%
0.62	15%	10%
0.64	55%	40%
0.66	75%	50%
0.68	80%	70%
0.7	95%	95%

Observation : Nous remarquons que si la probabilité donné en paramètre est strictement inférieure à 0.7, la probabilité de survie pour l'ensemble ε_n va diminuer par rapport au temps, et finalement, on aura une extinction, i.e. ε_n sera vide à partir d'un certain temps.

Conjecture : Nous pouvons alors conjecturer que si la probilité p est strictement supérieure à 0.7, l'ensemble ε_n a une grande chance de survivre plus le temps avance.

5 COMPLEXITÉ

Compléxité des algorithmes utilisés

Nom du fichier ou de la methode	utilité	complexité en temps
EdenRicharson.Java	Évolution en dimension 1	$O(n^2)$
evolution(python)	Évolution en dimension 1 sous forme d'intervalle	O(n)
liste(python)	Tracer un histogramme de l'ensemble $\frac{D_n}{n}$	$O(n^2)$
RandomWalk.Java	Évolution en dimension 2	$O(n^3)$
VRandomWalk.Java	Évolution en dimension 2 de E_n en intégrant la droite y=-x+n	$O(n^3)$
classify.py	Classifier les ensembles suivant leurs paramètre p (p pour la probabilité)	$O(n^2)$

6 CONCLUSION ET OUVERTURE

6.1 Notre récolte

Lors du semestre 5, nous avons pu travailler sur le système de Lindenmayer pour modéliser la croissance d'une plante. Ceci nous a motivé à choisir le modèle d'Eden-Richardson qui décrit la croissance de types spécifiques de clusters bactériens. En intégrant le modèle par des graphismes qui proviennent des codes en Java et Python, nous avons pu avoir une meilleure vision d'un ensemble aléatoire mathématique. En laissant les programmes tourner pendant un certain temps, le modèle va converger vers une forme limite qui dépend du paramètre p qui représente la probabilité qu'un individu puisse infecter ses voisins. La construction d'un type spécifique d'arbre génealogique, et de sa comparaison avec le modèle d'Eden-Richardson nous permet de déduire l'aspect d'extinction de notre modèle à un certain temps.
De nombreuses notions mathématiques sont employées dans ce projet, et surtout des notions en probabilités telles que, la probibilité conditionnelle, l'espérance conditionnelle, les variables aléatoires, les chaînes de Markov, la loi gaussienne etc...

6.2 Extra fun

Nous avons construit une sorte de machine qui arrive à déterminer la probabilité que vous avez donné au modèle d'Eden-Richardson en utilisant un toolkit qui s'appelle TensorFlow qui nous a permis de créer un modèle qui prend en entrée des images des ensembles E_n avec probabilités différentes. Le machine va ainsi s'entraîner sur ces données là que l'on appelle "training-data". Puis nous pouvons tester notre "machine" avec des données qui sont différentes de notre "training-data" que l'on nomme "testing-data". Enfin le modèle que nous avons construit va classifier les données à tester. Comme vous le savez, nous adorons les schémas. Voici des schémas pour comprendre notre programme.

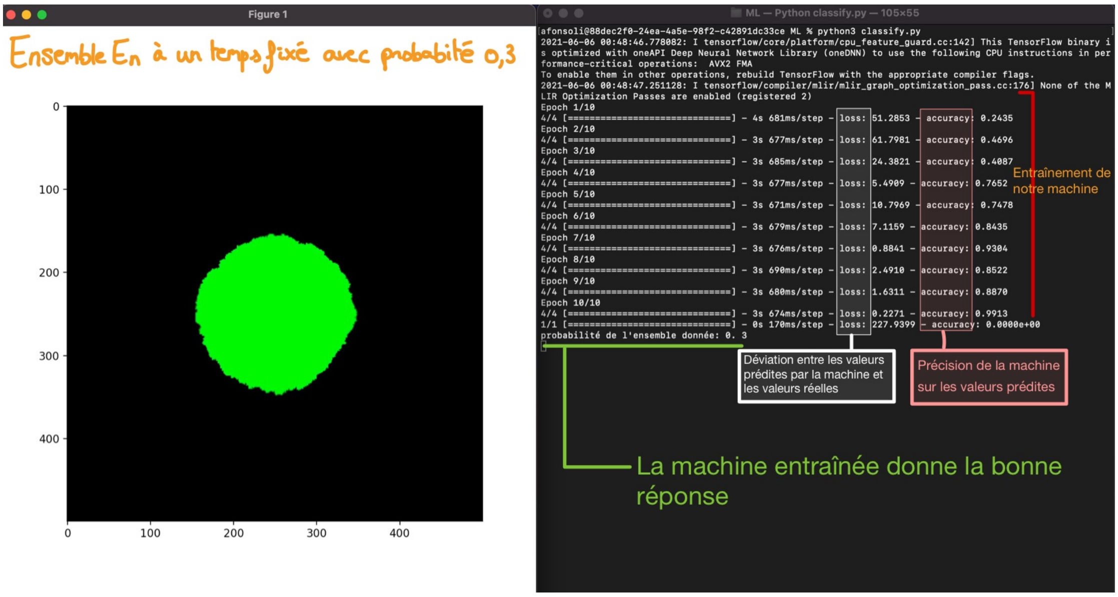


FIGURE 9 – (Gauche) Image de E_n (Droite) Classification de l'image en donnant le paramètre p(probabilité) initial

Une idée sympa serai de donner à la machine, un E_n à n'importe quel moment de son évolution, et de nous donner le paramètre p(probabilité donnée à E_n). Il est curieux de savoir comment la machine remarque les légères différences chez les ensembles E_n. Ceci est sûrement dû à la convergence des E_n vers une forme limite.

Finalement, nous aimerons remercier M. Lucas pour nous avoir guider tout au long du projet, c'est un plaisir de travailler avec un professeur avec qui nous pouvons partager nos idées, et surtout discuter pour trouver les solutions adaptées pour réaliser ces idées. Le défi pour nous n'était pas de comprendre le but du projet mais de donner des réponses plus précises à chaque aspect du projet. Et nous nous sommes beaucoup amusés lors de ce processus.