

Génération aléatoire d'automates non déterministes par flots de bits

J.-M. CHAMPARNAUD, G. HANSEL, T. PARANTHOËN et D. ZIADI

LIFAR, Université de Rouen, 76821 Mont-Saint-Aignan, France

[champarnaud,hansel,paranth,ziadi]@dir.univ-rouen.fr

Résumé

L'objectif de cette étude est la génération aléatoire des automates non-déterministes. Nous nous intéressons plus particulièrement à la méthode de génération par flots de bits dont nous présentons une analyse probabiliste. Nous montrons que les déterministes des automates de taille n générés à partir de flots de bits équiprobables ont une probabilité d'avoir une taille égale à 2 qui tend vers 1 quand n tend vers l'infini. Cette propriété fournit une justification asymptotique des résultats expérimentaux obtenus par L. van Zijl concernant la concision de diverses représentations des langages rationnels. Nous calculons d'autre part la probabilité de transition d'un état équiprobablement choisi parmi les 2^n états possibles du déterminisé vers un état donné. Nous en déduisons que les cibles des transitions déterministes sont équiprobablement distribuées dans le cas de flots de bits générés avec la probabilité $2 - 2^{\frac{n-1}{n}}$. Ce résultat conforte la conjecture de T. Leslie et D. Wood selon laquelle le nombre d'états du déterminisé est maximum quand la densité déterministe est 2. Enfin nous étendons cette étude probabiliste au cas des \star -*nfas* définis par L. van Zijl.

1 Introduction

La génération aléatoire de structures permet de réaliser des tests de performance d'algorithmes, ou de donner une illustration de résultats théoriques. Une bonne connaissance de l'espace des structures que l'on désire générer est indispensable pour concevoir un algorithme de génération. C. Nicaud a étudié la génération aléatoire équiprobable des automates déterministes ayant n états [7].

Notre objectif est de réaliser une étude similaire en ce qui concerne la génération des automates non-déterministes. Nous considérons ici la méthode de génération des NFAs par flots de bits aléatoires. L. van Zijl a utilisé cette méthode, avec des flots de bits équiprobables, pour comparer la concision de diverses représentations des langages rationnels [8]. Nous développons ici une analyse probabiliste des tables de transition non déterministes produites par

cette méthode et nous en déduisons certaines propriétés des automates non-déterministes associés. Nous nous intéressons en particulier au nombre d'états du déterminisé de tels automates.

Dans une première partie, l'approche est inspirée de la construction par accessibilité du déterminisé. Nous calculons la probabilité que l'image d'une partie de taille i donnée par un symbole donné soit de taille k . Nous en déduisons la probabilité que l'image de l'ensemble des initiaux par un mot de longueur t donné soit de taille k . Ces calculs fournissent une justification asymptotique des résultats expérimentaux obtenus par L. van Zijl [8]. Nous montrons en particulier que les NFAs produits à partir de flots de bits équiprobables sont asymptotiquement accessibles et coaccessibles et que les déterminisés de ces NFAs sont asymptotiquement de taille $m + 2$, où m est la taille de l'alphabet.

Dans une seconde partie, l'approche est inspirée de la construction exhaustive du déterminisé. Nous calculons la probabilité que l'image d'une partie équiprobablement choisie parmi les 2^n parties possibles contienne un état donné. Cela nous permet de montrer que les parties produites par détermination sont équiprobablement distribuées dans le cas de flots de bits générés avec la probabilité $2 - 2^{\frac{n-1}{n}}$. Ce résultat conforte la conjecture de T. Leslie et D. Wood [4] selon laquelle le nombre d'états du déterminisé est maximum quand la densité déterministe est 2.

Enfin nous étendons cette étude probabiliste au cas des \star -*nfas* définis par L. van Zijl. Dans le cas des flots de bits équiprobables, nos calculs justifient les résultats expérimentaux rapportés en [8, 10, 9]. Les \cap -*nfas* ont un comportement analogue à celui des \cup -*nfas*. Pour les \oplus -*nfas*, les cibles des transitions déterministes sont équiprobablement distribuées dans le cas de flots de bits équiprobables.

La section suivante rassemble quelques définitions et conventions. La section 3 fournit une description de la méthode de génération par flots de bits et passe en revue les travaux de L. van Zijl et ceux de T. Leslie. Les sections 4 et 5 présentent l'analyse probabiliste de cette méthode et les principaux résultats qui en découlent. La section 4 est consacrée aux flots équiprobables et la section 5 aux flots non équiprobables.

2 Définitions et notations

Un automate est un quintuplet $\mathcal{A} = \langle Q, \Sigma, \delta, I, F \rangle$ où Q est l'ensemble fini des états de l'automate, Σ est l'alphabet sur lequel est défini l'automate, δ est la *fonction de transition* ($\delta : Q \times \Sigma \rightarrow 2^Q$) qui associe à chaque élément de $Q \times \Sigma$ une partie de Q , I est un sous-ensemble non-vide de Q dont les éléments sont les états *initiaux*, et F est un sous-ensemble de Q dont les éléments sont les états *finaux*.

Un automate \mathcal{A} est *unaire* si son alphabet est restreint à un seul symbole.

Un automate est dit *accessible* si et seulement si pour tout état $q \in Q$, il existe un chemin menant d'un des états initiaux vers cet état. Un automate est dit *coaccessible* si et seulement si pour tout état $q \in Q$, il existe un chemin menant

de cet état vers un des états finaux. Un automate accessible et co-accessible est dit *émondé*.

Un automate \mathcal{A} est dit *déterministe* s'il possède un unique état initial et s'il existe au plus, pour chaque état et pour chaque symbole de l'alphabet, une transition sortant de cet état. Un automate déterministe (resp. non déterministe) est un DFA (resp. un NFA). Pour tout NFA \mathcal{A} reconnaissant un langage L , on sait construire un DFA équivalent (i.e. reconnaissant le même langage) appelé *déterminisé* de \mathcal{A} . L'algorithme utilisé pour convertir un NFA en un DFA équivalent est l'algorithme de *construction ensembliste* [1, 11]. Comme cet algorithme peut, dans le pire des cas, engendrer à partir d'un NFA à n états un DFA à 2^n états [5], on a l'habitude de dire que les NFAs sont plus petits que les DFAs.

Suivant L. van Zijl, on appelle \star -*nfa* un NFA dans lequel l'opération d'union (utilisée pour calculer la fonction de transition du déterminisé) est remplacée par une opération *binnaire commutative et associative*. Ainsi un \cup -*nfa* n'est rien d'autre qu'un NFA classique. Dans le cas de l'*intersection* on obtient un \cap -*nfa*. La fonction de transition a pour expression : $\delta(P, a) = \delta(p_1, a) \cap \delta(p_2, a) \cap \dots \cap \delta(p_k, a)$, où $P = \{p_1, p_2, \dots, p_k\}$ est une partie de Q . Un mot u est reconnu par un \cap -*nfa* si et seulement si $F \subseteq \delta(I, u)$. Dans le cas de la *différence symétrique* on obtient un \oplus -*nfa*. La fonction de transition a pour expression : $\delta(P, a) = \delta(p_1, a) \oplus \delta(p_2, a) \oplus \dots \oplus \delta(p_k, a)$. Un mot u est reconnu par un \oplus -*nfa* si et seulement si $|\delta(I, u) \cap F|$ est impair.

La définition de la taille d'un NFA est différente selon les auteurs. Le choix du nombre de transitions [2] est justifié par l'espace mémoire nécessaire pour implémenter un automate. Certaines propriétés de complexité dépendent de la somme du nombre d'états et du nombre de transitions [3]. Cependant la plupart des études utilisent le nombre d'états, et nous adopterons cette convention.

3 Génération des NFAs par flots de bits

La méthode utilisée par L. van Zijl [10] pour construire aléatoirement un automate déterministe est la suivante :

- on considère un alphabet $\Sigma = \{1, \dots, m\}$ de taille $m = 1$ ou 2 et un ensemble d'états $Q = \{1, \dots, n\}$,
- on génère un flot de bits de taille mn^2 décrivant la fonction de transition δ , chaque bit étant tiré de manière équiprobable. L'apparition d'un bit non nul à la position $(l-1)n^2 + (i-1)n + j$ indique l'existence d'une transition de l'état i vers l'état j portant l'étiquette de la $l^{\text{ème}}$ lettre de l'alphabet Σ ,
- l'automate possède un état initial unique, l'état 1 ,
- l'ensemble des état finaux est tiré au hasard, chaque état possède une chance égale d'être final ou non.

Cette méthode fournit un NFA $\langle Q, \Sigma, \delta, I, F \rangle$ qui n'est pas nécessairement émondé. L. van Zijl utilise cette méthode pour mesurer la concision des NFAs de taille n fixée. A cet effet elle construit une séquence de NFAs et pour chaque

NFA \mathcal{A} elle exécute les opérations suivantes :

- vérifier si \mathcal{A} est émondé ou non (on rejette les automates qui ne sont pas émondés),
- calculer l'automate \mathcal{M} déterministe minimal équivalent à \mathcal{A} ,
- vérifier si \mathcal{M} est isomorphe ou non aux automates minimaux précédemment calculés.

Ces calculs permettent, pour une taille de NFA fixée, d'évaluer la répartition en taille des automates minimaux équivalents aux NFAs produits aléatoirement, comme l'indique la Figure 1.

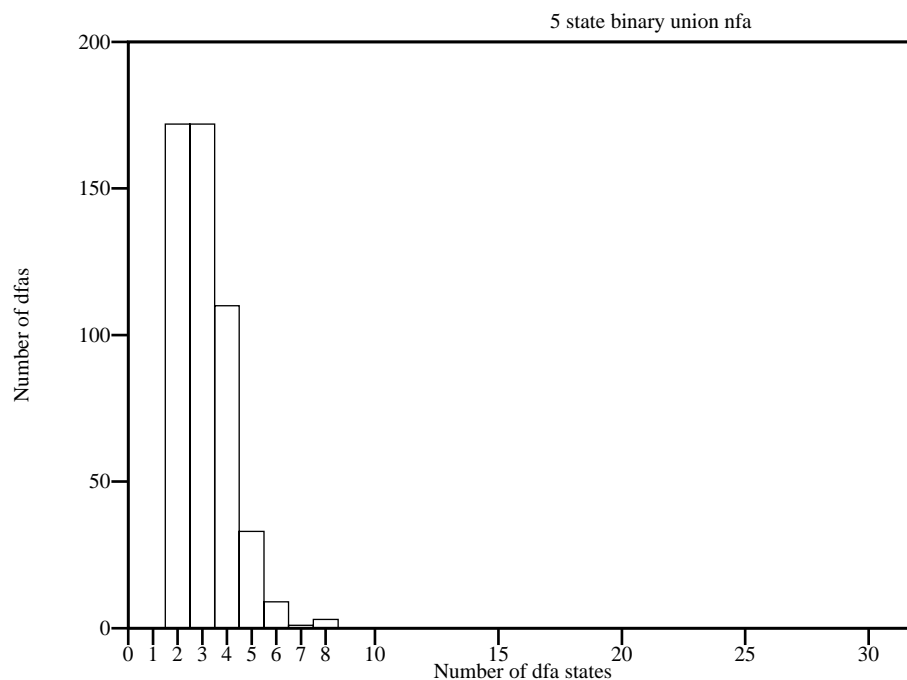


FIG. 1 – Résultats expérimentaux [8] représentant le nombre de langages reconnus par un automate minimal de taille n' construit à partir d'un NFA de taille 5 sur un alphabet binaire.

FIG. 2 – Résultats expérimentaux [8] représentant le nombre de langages reconnus par un automate minimal de taille n' construit à partir d'un NFA de taille 5 sur un alphabet binaire.

T. Leslie et D. Wood proposent une autre approche [4]. Ils utilisent la notion de *densité* d'un NFA pour construire la structure de transition. La densité d'un NFA possédant e transitions est égale à $da = \frac{e}{|\Sigma|n^2}$, sa *densité déterministe* est égale à $dd = \frac{e}{|\Sigma|n}$. L'étude expérimentale de T. Leslie et D. Wood porte sur la

taille du déterminisé d'un NFA en fonction de sa densité. Une fois la taille du NFA fixée, le protocole expérimental utilisé est le suivant :

- on fixe initialement une densité da de l'automate. Le nombre de transitions de l'automate est donc $e = da * |\Sigma|n^2$,
- on construit une structure de transition connexe aléatoire,
- on génère aléatoirement les $e - (n - 1)$ transitions restantes de l'automate. Dans le cas où une transition apparaît une seconde fois on effectue un nouveau tirage (utilisation d'un algorithme avec rejet),

Les résultats expérimentaux de T. Leslie font apparaître une densité déterministe optimale qui est égale à 2. Ils proposent la conjecture suivante :

Conjecture 1 *Pour un NFA donné on peut calculer une bornes du nombre d'états et du nombre de transitions du DFA équivalent produit par construction ensembliste. De plus, la relation entre la taille du DFA et la densité déterministe de l'automate donne lieu à une courbe de Poisson dont le pic est approximativement à une densité déterministe égale à 2.*

4 Analyse probabiliste : cas équiprobable

Nous considérons ici un \star -nfa \mathcal{A} de taille n sur un alphabet Σ de taille m associé à un flot de mn^2 bits équiprobable. Nous supposons dans notre étude que l'état initial est unique.

Soit \mathcal{D} l'automate obtenu par construction ensembliste de \mathcal{A} . Nous montrons que la probabilité que la taille de \mathcal{D} soit égale à $m + 2$ tend vers 1 quand n tend vers l'infini. Nous nous inspirons ici de la construction par accessibilité du déterminisé. L'idée est que la taille moyenne de l'image de l'état initial par une lettre est égale à $n/2$ et que la taille des images par un mot de longueur t croît très rapidement vers n [6]. Ainsi pour une partie X de taille i et un symbole a donnés nous calculons la probabilité que l'image $\delta(X, a)$ soit de taille k . Nous en déduisons la probabilité que l'image de l'ensemble des initiaux par un mot de longueur t soit de taille k .

4.1 Etude des NFAs classiques

Dans cette partie \mathcal{A} est un \cup -nfa.

Proposition 4.1 *Soit $\mathcal{A} = \langle Q, \Sigma, \delta, I, F \rangle$ un NFA de taille n associé à un flot de bits équiprobable. Soit X une partie de taille i de Q et a un symbole de Σ . La probabilité $P(i \rightarrow k)$ que l'image $\delta(X, a)$ de X par a soit de taille k est donnée par :*

$$P(0 \rightarrow 0) = 1 \text{ et } P(0 \rightarrow k) = 0, \forall k \neq 0$$

$$P(i \rightarrow k) = \frac{C_n^k (2^i - 1)^k}{2^{in}}, \forall i \neq 0$$

Preuve

Soit Z la variable aléatoire dans $\{0, \dots, n\}$ qui a pour valeur le cardinal de $\delta(X, a)$. Soit Z_j , $1 \leq j \leq n$, la variable aléatoire dans $\{0, 1\}$ qui a pour valeur 1 si l'état j appartient à $\delta(X, a)$ et 0 sinon. Comme la taille de X est égale à i , on a : $P(Z_j = 0) = 1/2^i$ et $P(Z_j = 1) = 1 - 1/2^i$. D'autre part les variables Z_j sont indépendantes. Comme $Z = \sum_{j=1}^n Z_j$, Z est une binômiale de paramètres n et $1 - 1/2^i$. On a donc : $P(Z = k) = C_n^k (1 - 1/2^i)^k (1/2^i)^{n-k}$. D'où le résultat. \square

Dans le cas d'un alphabet unaire, la probabilité $P(k, t)$ d'atteindre une partie Y de taille k par t transitions successives à partir de la partie initiale I peut être calculée à l'aide de la matrice de probabilité suivante :

$$M = \begin{pmatrix} P(0 \rightarrow 0) & \dots & \dots & \dots & P(0 \rightarrow n) \\ \vdots & \ddots & & & \vdots \\ \vdots & & P(i \rightarrow k) & & \vdots \\ \vdots & & & \ddots & \vdots \\ P(n \rightarrow 0) & \dots & \dots & \dots & P(n \rightarrow n) \end{pmatrix}$$

où le coefficient $M_{i,k}$ représente la probabilité que l'image d'une partie de taille i par un mot de taille 1 soit de taille k .

Exemple 1 Pour un automate de taille 3, on obtient la matrice suivante :

$$M_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \\ \frac{1}{64} & \frac{9}{64} & \frac{27}{64} & \frac{27}{64} \\ \frac{1}{512} & \frac{21}{512} & \frac{147}{512} & \frac{343}{512} \end{pmatrix}$$

Une fois cette matrice construite, la probabilité $P(k, t)$ peut être calculée selon la formule :

$$P(k, t) = \sum_{i=0}^n P(i, t-1) \times P(i \rightarrow k)$$

D'où la proposition :

Proposition 4.2 On suppose que l'alphabet est unaire. Soit V_t le vecteur de taille n dont le $k^{\text{ème}}$ élément correspond à la probabilité $P(k, t)$ que l'image $\delta(I, w)$ soit de taille k , avec $|w| = t$. En prenant $V_0 = (0, 1, 0, \dots, 0)$, ce qui correspond au choix d'un état initial unique, on a :

$$V_t = V_0 M^t$$

Le tableau suivant, issu de calculs réalisés sous Maple, montre la croissance de la probabilité $P(n, 2)$ en fonction de la taille n du NFA. Il indique par exemple que $P(n, 2)$ est supérieur à 0.9 dès que n est supérieur à 17.

$n >$	17	28	37	46	54	100
$P(n, 2) >$	$1 - 10^{-1}$	$1 - 10^{-2}$	$1 - 10^{-3}$	$1 - 10^{-4}$	$1 - 10^{-5}$	$1 - 10^{-10}$

Il apparaît clairement que $P(n, 2)$ tend vers 1 quand n tend vers l'infini. On en déduit le résultat suivant.

Proposition 4.3 *Soit \mathcal{A} un NFA unaire de taille n associé à un flot de bits équiprobable. La probabilité que le déterminisé de \mathcal{A} soit de taille 3 tend vers 1 quand n tend vers l'infini.*

Remarquons que la taille du déterminisé est 2 si l'on prend un ensemble d'initiaux produit de façon aléatoire et équiprobable. D'autre part, ce résultat se généralise facilement au cas d'un alphabet de taille arbitraire de la manière suivante.

Proposition 4.4 *Soit \mathcal{A} un NFA de taille n associé à un flot de bits équiprobable possédant un unique état initial. Alors la probabilité que le déterminisé de \mathcal{A} soit de taille $m + 2$ tend vers 1 quand n tend vers l'infini.*

Preuve

Les images $\delta(I, a), \forall a \in \Sigma$ sont en moyenne de taille $n/2$ et on peut les considérer comme étant toutes différentes si n est grand devant m . D'autre part la probabilité que l'image $\delta(I, w)$ soit de taille n , $\forall w \in \Sigma^*, |w| \geq 2$, tend vers 1 quand n tend vers l'infini.

□

Comme dans le cas unaire, la taille du déterminisé est 2 si l'on prend un ensemble d'initiaux produit de façon aléatoire et équiprobable. D'autre part, un corollaire de cette proposition est que la probabilité que \mathcal{A} soit accessible tend vers 1 quand n tend vers l'infini, puisque l'on atteint la partie contenant tout les états du DFA de départ. Deplus, si les états finaux de \mathcal{A} sont choisis équiprobablement, alors \mathcal{A} est co-accessible. En effet si \mathcal{A} possède des états finaux aléatoires alors son miroir possède des initiaux aléatoires et satisfait donc toutes les propriétés que nous venons d'énoncer.

4.2 Etude des \star -nfas

Une analyse probabiliste peut également être menée en ce qui concerne les \cap -nfas et les \oplus -nfas associés à des flots de bits équiprobables.

Proposition 4.5 *Soit \mathcal{A} un \cap -nfa de taille n associé à un flot de bits équiprobable. Si l'état initial est unique, la probabilité que le déterminisé de \mathcal{A} soit de taille $m + 2$ tend vers 1 quand n tend vers l'infini. Sinon la probabilité que le déterminisé soit de taille 2 tend vers 1 quand n tend vers l'infini.*

Preuve

On utilise le même raisonnement que pour les \cup -*nfas*, le rôle des valeurs 1 et 0 étant inversé.

□

Proposition 4.6 *Soit \mathcal{A} un \oplus -*nfa* de taille n associé à un flot de bits équiprobable. La probabilité $P(i \rightarrow k)$ que l'image $Y = \delta(X, a)$ d'une partie X de taille i soit de taille k est indépendante de i . On a :*

$$P(i \rightarrow k) = \frac{C_n^k}{2^n}$$

Preuve

Soit O (resp. E) l'ensemble des entiers impairs (resp. pairs) de $\{1, \dots, n\}$. Un état y appartient (resp. n'appartient pas) à Y s'il existe un nombre impair (resp. pair) d'états de X menant à y . On a par conséquent :

$$P(i \rightarrow k) = \frac{C_n^k \times A^k \times B^{n-k}}{2^{in}}$$

où A et B sont définis de la manière suivante :

$$A = \sum_{k \in E} C_i^k = 2^{i-1} = \sum_{k \in O} C_i^k = B$$

D'où le résultat.

□

En conclusion de cette analyse du cas où le flot de bits est équiprobable, remarquons que nos résultats sont conformes aux observations expérimentales de L. van Zijl. Pour les \oplus -*nfas* chaque partie conserve une chance égale d'apparaître au cours des transitions successives. Cette propriété implique que les \oplus -*nfas* sont plus concis que les \cup -*nfas*. D'autre part, les \cap -*nfas* ont un comportement analogue à celui des \cup -*nfas*.

5 Analyse probabiliste : cas non équiprobable

Nous supposons ici que la probabilité pour qu'un bit soit égal à 1, dans le flot utilisé pour construire la fonction de transition, est égale à $1/x$. L'approche est inspirée de la construction exhaustive du déterminisé. Nous calculons la probabilité que l'image d'une partie équiprobablement choisie parmi les 2^n parties possibles contienne un état donné. Cela nous permet de montrer que les parties produites par déterminisation sont équiprobablement distribuées dans le cas où le flot de bits est généré avec la probabilité $\frac{1}{x(n)} = 2 - 2^{\frac{n-1}{n}}$.

5.1 Etude des NFAs classiques

Proposition 5.1 Soit $\mathcal{A} = \langle Q, \Sigma, \delta, I, F \rangle$ un NFA de taille n associé à un flot de bits de probabilité $1/x$. Soit X une partie de Q équiprobablement choisie. Soit $q \in Q$ et $a \in \Sigma$. La probabilité $P_{\cup}(x, n)$ que q appartienne à $\delta(X, a)$ est égale à :

$$P_{\cup}(x, n) = \frac{(2x)^n - (2x - 1)^n}{(2x)^n}$$

Preuve

Soit V le vecteur de $\{0, 1\}^n$ associé à X . La partie X étant choisie de manière équiprobable parmi les 2^n parties de Q , la probabilité que $V[i]$ soit égal à 1 est égale à $1/2$, pour tout i dans $\{1, \dots, n\}$. Soit V' le vecteur de $\{0, 1\}^n$ associé à $\delta^{-1}(q, a)$. Par hypothèse, la probabilité que $V'[i]$ soit égal à 1 est égale à $1/x$, pour tout i dans $\{1, \dots, n\}$. Par conséquent, la probabilité que $(V[i], V'[i])$ soit différent de $(1, 1)$ est égale à $1 - 1/2x$, pour tout i dans $\{1, \dots, n\}$.

Comme \mathcal{A} est un \cup -nfa, on a : $q \in \delta(X, a) \Leftrightarrow \exists y \in X \mid q \in \delta(y, a)$. Par conséquent, $q \in \delta(X, a) \Leftrightarrow \exists i \in \{1, \dots, n\} \mid (V[i], V'[i]) = (1, 1)$. La probabilité P' que q n'appartienne pas à $\delta(X, a)$ est donc égale à : $P' = (1 - 1/2x)^n$. D'où le résultat. □

Lorsque le flot de bits est équiprobable, on obtient : $P_{\cup}(2, n) = 1 - (3/4)^n$. Il est clair que la probabilité que $V'[i]$ soit égal à 1, pour i dans l'ensemble $\{1, \dots, n\}$, est alors différente de $1/2$. On vérifie ainsi que l'utilisation d'un flot de bits équiprobable ne permet pas de produire par déterminisation les parties $\delta(X, a)$ de manière équiprobable (Proposition 4.1).

Cependant, la résolution de l'équation :

$$\frac{(2x)^n - (2x - 1)^n}{(2x)^n} = \frac{1}{2}$$

donne comme solution la fonction $x(n)$ suivante :

$$x(n) = \frac{1}{2 - 2^{\frac{n-1}{n}}}$$

Par conséquent, l'utilisation de flots de bits de probabilité $\frac{1}{x(n)}$ permet d'assurer qu'à chaque étape de la déterminisation n'importe quelle partie de Q possède une chance égale d'apparaître. Intuitivement, cela indique que, pour n fixé, la valeur $\frac{1}{x(n)}$ permet de maximiser le nombre moyen d'états dans le déterminisé. Les \cup -nfas issus de flots de bits de probabilité égale à $\frac{1}{x(n)}$ sont donc plus concis que les \cup -nfas issus de flots équiprobables (de même que dans le cas équiprobable les \oplus -nfas sont plus concis que les \cup -nfas et les \cap -nfas). Les résultats expérimentaux que nous avons obtenus confirment ces deux propriétés.

5.2 Résultats expérimentaux

Nous avons réalisé une première série de tests pour étudier la répartition en taille des déterminisés des NFAs produits aléatoirement. Cette étude est de nature sensiblement différente de celle de L. van Zijl qui s'intéresse à la répartition en taille des DFAs minimaux équivalents aux NFAs produits aléatoirement. Par contre, elle est proche de celle de T. Leslie et D. Wood qui analysent la distribution de la taille du déterminisé pour des NFAs ayant une densité de transitions donnée [4].

Les Figures 3, 4 et 5 illustrent les résultats obtenus. Sur les deux premières figures l'abscisse des graphes représente le nombre d'états des déterminisés de NFA générés aléatoirement ; l'ordonnée représente le pourcentage de déterminisés construits, possédant une taille fixée.

Sur la Figure 5, l'abscisse représente la taille des sous-ensembles produits durant la construction ensembliste à partir de NFAs de taille 50 ; l'ordonnée représente le pourcentage de ces ensembles d'une taille fixée.

Les Figures 3.c et 3.d permettent de comparer les deux choix possibles pour l'ensemble des états initiaux (dans le cas d'un flot de bits équiprobable). Lorsque l'état initial est unique (Figure 3.d), la répartition des tailles a un spectre légèrement plus grand que lorsque l'ensemble des états initiaux est tiré équiprobablement (Figure 3.c).

Les Figures 3.e et 3.f mettent en évidence l'influence de la taille de l'alphabet, qui est égale à 2 pour la Figure 3.e et à 4 pour la Figure 3.f. Le flot de bits a une probabilité $\frac{1}{x(n)}$ dans les deux cas. Le spectre de la répartition des tailles s'agrandit quand le nombre de symboles augmente.

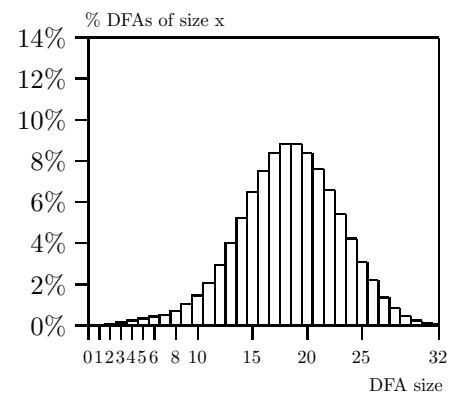
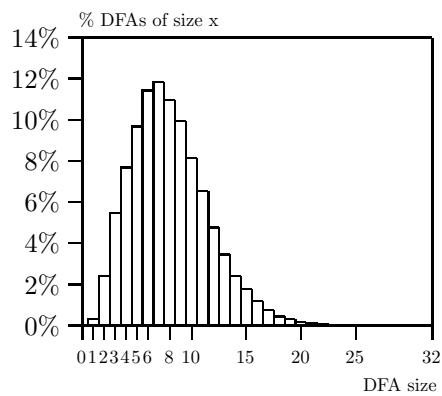
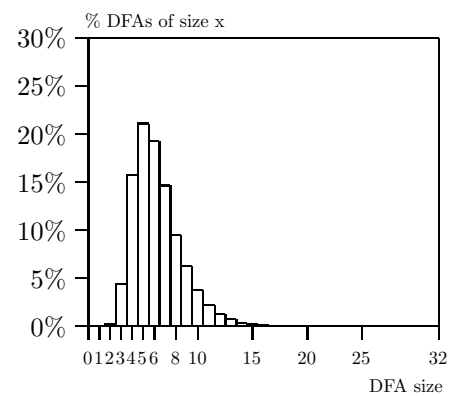
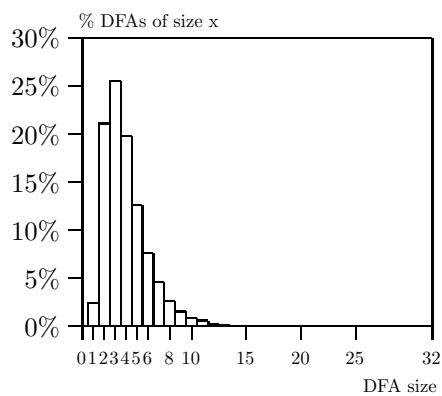
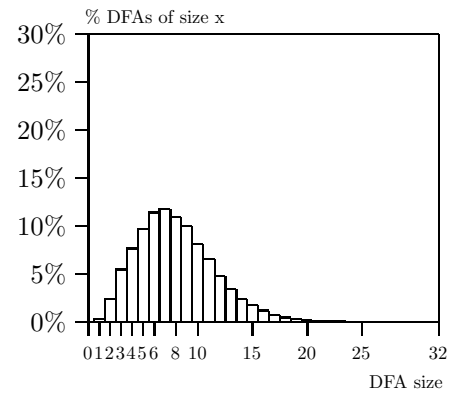
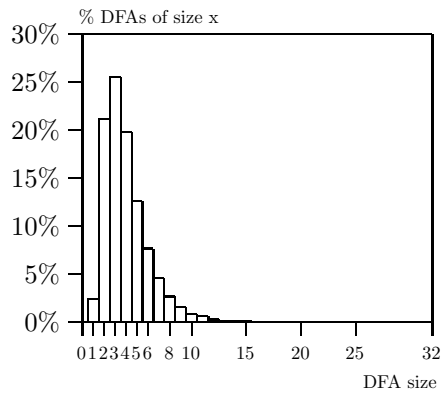


FIG. 3 – Tests sur les NFAs de taille 5.

Les Figures 4 et 5 illustrent la pertinence de la probabilité $\frac{1}{x(n)}$. La Figure 4 montre la distribution de la taille des DFAs obtenus par construction ensembliste à partir de NFAs de taille 20, 50 et 100 construits à partir de flux de bits de probabilité $1/x(n)$. Pour chacun des graphes un jeu de 100 000 NFAs a été utilisé, et les tailles des DFAs ont été regroupées par tranche de 10.

La Figure 5 met en évidence l'influence de petites variations de la probabilité $1/x(n)$ sur la taille des parties construites durant la déterminisation. Ainsi avec une probabilité plus grande ($1/x(n) + 0.01$, Figure 5.c), la taille des parties construites est plus grande, et, avec une probabilité plus petite ($1/x(n) - 0.01$, Figure 5.a), la taille des parties construites est plus petite. Pour chacun de ces graphes, un jeu de 1 000 NFAs accessibles a été utilisé.

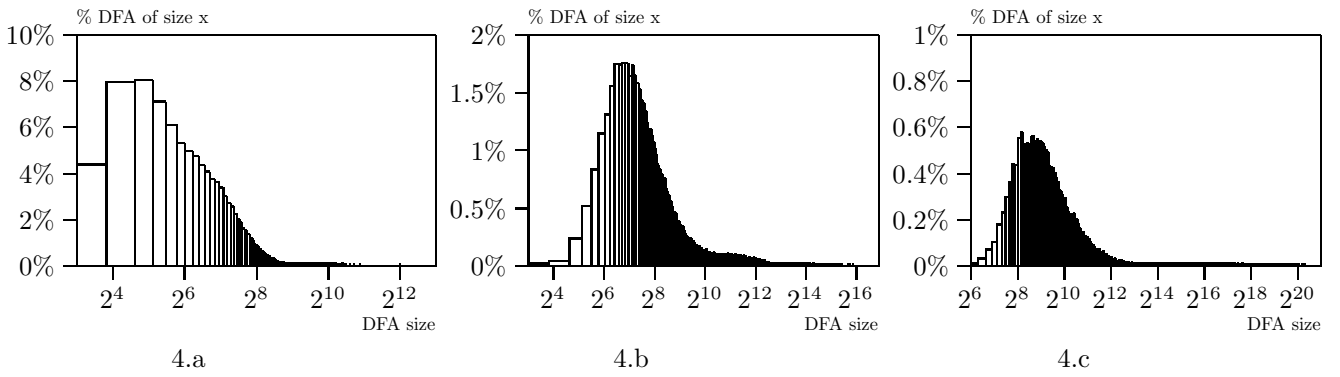


FIG. 4 – Répartition en taille des déterminisés de NFAs de taille 20 (a), 50 (b) et 100 (c), construits avec la probabilité $\frac{1}{x(n)}$.

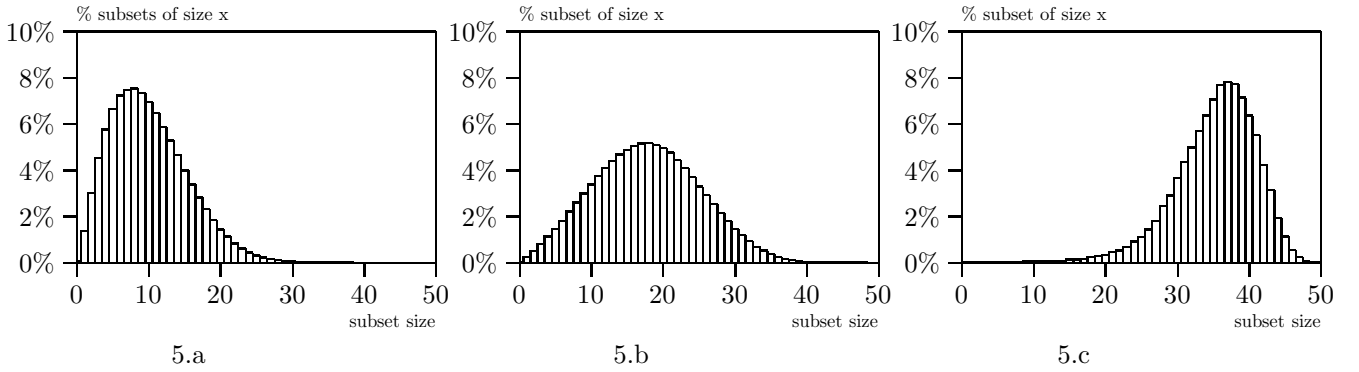


FIG. 5 – Répartition en taille des sous-ensembles construits durant la construction ensembliste à partir de NFAs de taille 50. (a) : $\frac{1}{x(50)} - 0.01$. (b) : $\frac{1}{x(50)}$ (c) : $\frac{1}{x(50)} + 0.01$.

5.3 Etude des \star -nfas

La probabilité que l'image d'une partie équiprobablement choisie contienne un état donné peut être également calculée dans le cas des \cap -nfas et des \oplus -nfas.

Proposition 5.2 Soit $\mathcal{A} = \langle Q, \Sigma, \delta, I, F \rangle$ un \cap -nfa de taille n associé à un flot de bits de probabilité $1/x$. Soit X une partie de Q équiprobablement choisie. Soit $q \in Q$ et $a \in \Sigma$. La probabilité $P_{\cap}(x, n)$ que q appartienne à $\delta(X, a)$ est égale à :

$$P_{\cap}(x, n) = \left(\frac{2x-1}{2x} \right)^n - \frac{1}{x^n}$$

Preuve

Soit V (resp. V') le vecteur de $\{0, 1\}^n$ associé à X (resp. à $\delta^{-1}(q, a)$). Pour tout i dans $\{1, \dots, n\}$, la probabilité que $V[i]$ (resp. $V'[i]$) soit égal à 1 est égale à $1/2$ (resp. $1/x$). Par conséquent, la probabilité que $(V[i], V'[i])$ soit différent de $(1, 0)$ est égale à $\frac{2x-1}{2x}$, pour tout i dans $\{1, \dots, n\}$.

Comme \mathcal{A} est un \cap -nfa, on a : $q \in \delta(X, a) \Leftrightarrow \forall y \in X, q \in \delta(y, a)$. Par conséquent,

$$q \in \delta(X, a) \Leftrightarrow \forall i \in \{1, \dots, n\}, (V[i], V'[i]) \neq (1, 0) \text{ et } \exists i \in \{1, \dots, n\} \mid V[i] = 1$$

La probabilité que q appartienne à $\delta(X, a)$ est donc égale à :

$$P_{\cap}(x, n) = \left(\frac{2x-1}{2x} \right)^n - \frac{1}{x^n}$$

□

Lorsque le flot de bits est équiprobable, on obtient $P_{\cap}(2, n) = (3/4)^n - 1/2^n$.

Proposition 5.3 Soit $\mathcal{A} = \langle Q, \Sigma, \delta, I, F \rangle$ un \oplus -nfa de taille n associé à un flot de bits de probabilité $1/x$. Soit X une partie de Q équiprobablement choisie. Soit $q \in Q$ et $a \in \Sigma$. La probabilité $P_{\oplus}(x, n)$ que q appartienne à $\delta(X, a)$ est égale à :

$$P_{\oplus}(x, n) = \frac{1}{2(x-1)} - \frac{1}{2x^{n-1}(x-1)}$$

Preuve

Soit V (resp. V') le vecteur de $\{0, 1\}^n$ associé à X (resp. à $\delta^{-1}(q, a)$). Pour tout i dans $\{1, \dots, n\}$, la probabilité que $V[i]$ (resp. $V'[i]$) soit égal à 1 est égale à $1/2$ (resp. $1/x$). Par conséquent, la probabilité que $(V[i], V'[i])$ soit différent de $(1, 1)$ est égale à $\frac{2x-1}{2x}$, et la probabilité que $(V[i], V'[i])$ soit égal à $(1, 1)$ est égale à $\frac{1}{2x}$.

Comme \mathcal{A} est un \oplus -nfa, on a :

$$q \in \delta(X, a) \Leftrightarrow |\{x \in X \text{ tel que } q \in \delta(x, a)\}| \equiv 1 \pmod{2}$$

Supposons $n = 2p$ et notons P la probabilité cherchée. Soit E (resp. O) l'ensemble des entiers pairs (resp. impairs) de $\{1, \dots, n\}$. Soit E' l'ensemble des entiers impairs compris entre 1 et $n - 2$. On a :

$$q \in \delta(X, a) \Leftrightarrow \begin{cases} |\{i \in \{1, \dots, n\} \text{ tel que } (V[i], V'[i]) = (1, 1)\}| \equiv 1 \pmod{2} \\ |\{i \in \{1, \dots, n\} \text{ tel que } (V[i], V'[i]) \neq (1, 1)\}| \equiv 1 \pmod{2} \end{cases}$$

D'où :

$$\begin{aligned} P &= \frac{\sum_{k \in O} C_n^k (2x - 1)^k}{(2x)^n} \\ P &= \frac{\sum_{k \in O} C_{n-1}^k (2x - 1)^k + \sum_{k \in O} C_{n-1}^{k-1} (2x - 1)^k}{(2x)^n} \\ P &= \frac{\sum_{k=0}^n C_{n-1}^k (2x - 1)^k + 2 \sum_{k \in E'} C_{n-1}^k (2x - 1)^k}{(2x)^n} \\ P &= \frac{1}{2x} + \frac{2 \sum_{k \in E'} C_{n-1}^k (2x - 1)^k}{(2x)^n} \\ P &= \frac{1}{2x} + \frac{1}{2x^2} + \dots + \frac{1}{2x^{n-1}} \end{aligned}$$

D'où finalement :

$$P_{\oplus}(x, n) = \frac{1}{2(x-1)} - \frac{1}{2x^{n-1}(x-1)}$$

La démonstration est similaire pour $n = 2p + 1$, avec les hypothèses :

$$q \in \delta(X, a) \Leftrightarrow \begin{cases} |\{i \in \{1, \dots, n\} \text{ tel que } (V[i], V'[i]) = (1, 1)\}| \equiv 1 \pmod{2} \\ |\{i \in \{1, \dots, n\} \text{ tel que } (V[i], V'[i]) \neq (1, 1)\}| \equiv 0 \pmod{2} \end{cases}$$

□

Lorsque le flot de bits est équiprobable, on obtient $P_{\oplus}(2, n) = \frac{1}{2} - \frac{1}{2^n}$. Le terme $-\frac{1}{2^n}$ correspond au cas où X est vide. Partant d'une partie non vide, on atteint donc un état q avec une probabilité égale à $1/2$. Un \oplus - nfa généré à partir d'un flux de bits équiprobable engendre donc, lors de sa détermination, toutes les parties de façon équiprobables. Ce qui est cohérent avec la probabilité $P = \frac{C_n^k}{2^n}$ d'obtenir une partie de taille k à partir d'une partie de taille i (Proposition 4.6).

6 Conclusion

L'analyse probabiliste des tables de transition non déterministes associées à des flots de bits aléatoires et équiprobables donne une justification asymptotique des résultats expérimentaux obtenus par L. van Zijl. D'une part un algorithme de rejet des NFAs non émondés est possible, avec une complexité polynomiale.

D'autre part les déterminisés des NFAs produits par ce type de génération sont asymptotiquement de taille $m + 2$. Remarquons que le comportement asymptotique est en fait obtenu très rapidement, dès que le NFA possède quelques dizaines d'états. Les résultats expérimentaux obtenus par L. van Zijl portent sur des NFAs ayant moins de 10 états, ce qui explique en partie que la concision observée est relativement plus forte. D'autre part, l'analyse de la répartition en taille des déterminisés semble prometteuse. La génération des NFAs à partir de flots de bits de probabilité $\frac{1}{x(n)} = 2 - 2^{\frac{n-1}{n}}$ a la propriété de répartir la taille des déterminisés de la façon la plus large possible. Ce résultat conforte la conjecture de T. Leslie et D. Wood selon laquelle le nombre d'états du déterminisé est maximum quand la densité du NFA est égale à $2/n$.

Références

- [1] J.E. Hopcroft and J.D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, 1979.
- [2] J. Hromkovič, S. Seibert, and T. Wilke. Translating regular expressions into small ϵ -free nondeterministic finite automata. *Journal of Computer and System Sciences*, 62(4) :565–588, June 2001.
- [3] L. Ilie and S. Yu. Constructing NFAs by optimal use of positions in regular expressions. In *CPM'2002, Lecture Notes in Computer Science*, pages 279–288, 2002.
- [4] T. Leslie. Efficient approaches to subset construction. Master's thesis, University of Waterloo, Ontario, Canada, 1995. Supervised by D. Raymond and D. Wood.
- [5] A.R. Meyer and M.J. Fisher. Economy of description by automata grammars and formal systems. *FOCS*, 12 :188–191, 1971.
- [6] T. Paranthoën. Génération aléatoire d'automates non-déterministes, et application aux systèmes multi-agents. Master's thesis, Université de Rouen, France, 2001.
- [7] C. Nicaud. *Etude du comportement en moyenne des automates finis et des langages rationnels*. PhD thesis, Université Paris 7, 2000.
- [8] L. van Zijl. *Generalized Nondeterminism and the Succinct Representation of Regular languages*. PhD thesis, University of Stellenbosch, 1997.
- [9] L. van Zijl. The quantification of succinctly representable regular languages. Technical report, Stellenbosch University, 1999.
- [10] L. van Zijl *et al.* www.cs.sun.ac.za/~lynette/MERLin/MerlinManRev.ps, Merlin 1.1 help. Technical report.
- [11] S. Yu. Regular languages. in : G. Rozenberg, A. Salomaa (Eds.), *Handbook of Formal Languages, Word, Grammar*, I :41–110, 1997.