RECONSTITUTION D'UNE CHAÎNE À PARTIR D'UNE TRACE

Emmanuel Kammerer Victor Dubach
sous la direction de Thomas Budzinski

2018/2019

Table des matières

1 Introduction

On s'intéresse à la reconstruction d'un mot à partir de plusieurs échantillons qui ont été altérés. Plus précisément : considérons un mot binaire $\mathbf{t} = \mathrm{t}_1\mathrm{t}_2 \dots \mathrm{t}_n \in \{0,1\}^n$ de longueur n. Supprimons aléatoirement des bits de $\mathbf{t}$: chaque bit est supprimée avec probabilité q de manière indépendante. On obtient alors un nouveau mot $\widetilde{\mathbf{t}}$, qui est la concaténation des bits n'ayant pas été supprimés. On parlera d'algorithme de délétion pour désigner ce procédé, et les mots ainsi obtenus seront appelés des échantillons ou des mots reçus. On dira également que $\mathbf{t}$ est le mot transmis. Par exemple, l'algorithme de délétion appliqué au mot $\mathbf{t} = 0111001$ peut donner l'échantillon 01100. Notre but va alors être d'étudier la reconstruction d'un mot initial $\mathbf{t}$ à partir d'échantillons $\widetilde{\mathbf{t}}^1, \dots, \widetilde{\mathbf{t}}^m$ obtenus en appliquant m fois de manière indépendante l'algorithme de délétion à $\mathbf{t}$. Étant donnée la probabilité q de délétion (dépendant éventuellement de la longueur n du mot), on cherche un algorithme et une suite (m_n) tels que la probabilité que l'algorithme retrouve $\mathbf{t} \in \{0,1\}^n$ à partir de m_n mots reçus indépendants tende vers 1 quand n tend vers l'infini.

Cette question est issue du problème de la reconstitution de l'ADN d'un ancêtre commun, en ayant accès à l'ADN de ses descendants. En effet, l'ADN peut être vu comme un mot subissant des délétions, des insertions et des substitutions de lettres. Si l'on se donne des ADN de plusieurs êtres vivants, on peut se demander quel serait l'ADN de leur plus proche ancêtre commun en supposant que les ADN observés ont subi les délétions, des insertions et des substitutions selon une certaine loi. Ici, on se restreint à des mots binaires et les seules modifications que l'on autorise sont des délétions. Pour une étude plus exhaustive, le lecteur est invité à lire l'article [5] qui s'intéresse aussi au cas des substitutions et insertions.

Dans la première partie, qui correspond à l'article [1], on suppose que la probabilité de délétion q tend vers 0 lorsque la longueur n des mots transmis $\mathbf{t}$ tend vers $+\infty$.

Le premier résultat important qu'on établira est le théorème 2.9. Il montre qu'en supposant la décroissance de q assez rapide, en $1/\log n$, on peut reconstruire la plupart des mots transmis avec grande probabilité à l'aide d'un nombre d'échantillons de l'ordre de $\log n$ en utilisant un algorithme assez simple. Cet algorithme plutôt naïf, appelé algorithme de majorité bit par bit, parcourt simultanément tous les échantillons en les faisant successivement voter pour chaque bit, en avançant dans les mots reçus qui votent pour le bit majoritaire. L'avancement dans chacun des échantillons peut être représenté par un compteur, qui est incrémenté à chaque fois que le mot reçu correspondant vote pour le bit majoritaire. Des explications plus précises ainsi qu'un exemple sont donnés en section 2.1. Reformulons l'énoncé du théorème en précisant ce que l'on entend par "la plupart" et par "avec grande probabilité". On considère pour chaque entier n un sous-ensemble de $\{0,1\}^n$ de mots "gentils", dont la proportion tend vers 1 quand n tend vers l'infini. Il existe alors des constantes c, d telles que, si $m_n \geq c \log n$ et la probabilité de délétion q_n pour des mots de longueur n vérifie $q_n \leq \frac{1}{d \log n}$, alors la probabilité pour que l'algorithme de majorité bit par bit reconstitue un mot transmis "gentil" $\mathbf{t} \in \{0,1\}^n$ à partir de m_n échantillons tend vers 1 quand n tend vers l'infini. Pour montrer ce théorème, on s'appuiera sur les bornes de Chernoff que l'on énonce au début du mémoire ainsi que des considérations sur les marches aléatoires sur $\mathbb{Z}$, notamment un théorème du scrutin géométrique et visuel qui servira à contrôler l'avancement des compteurs.

Puis on évoquera un autre résultat de [1], portant sur la reconstruction de mots $\mathbf{t}$ arbitraires. Dans ce cas là, on suppose que q décroit encore plus rapidement, de l'ordre de $\frac{1}{n^{1/2+\varepsilon}}$. Alors, à l'aide d'une légère modification de l'algorithme de majorité bit par bit, on obtient que $\frac{c}{\varepsilon} nq \log n$ échantillons suffisent pour reconstruire tout mot de longueur n avec probabilité tendant vers 1 lorsque n tend vers l'infini, avec c une constante indépendante de ε. Voici une idée de la modification de l'algorithme : les mots transmis $\mathbf{t}$ qui ne sont pas "gentils" sont les mots comportant des longues suites de 0 consécutifs ou de 1 consécutifs car il y aura très probablement des délétions dans ces suites. L'algorithme de majorité bit par bit nous permet de retrouver $\mathbf{t}$ sauf les longueurs de ces longues suites de 1 ou de 0. Le problème consistera donc à retrouver la longueur de ces longues suites.

Le dernier point abordé dans l'article [1] est une borne inférieure de $\Omega(nq(1-q))$ pour le nombre d'échantillons nécessaire à la reconstruction de mots $\mathbf{t}$ arbitraires de longueur n. L'idée pour trouver une borne inférieure est d'exhiber pour tout $n \in \mathbb{N}$ deux mots de $\{0,1\}^n$ qui soient difficiles à distinguer.

On montrera ensuite deux résultats dus à l'article [5]. Cette fois-ci, la probabilité de délétion q est fixée et ne dépend plus de n. Tout d'abord, le théorème 4.2 dit que tout mot de longueur n peut être reconstitué à partir de $\exp\left(O(n^{1/3} \log n)\right)$ échantillons avec grande probabilité. L'algorithme utilisé est un algorithme de statistiques

bit par bit au sens où il ne considère que les lois des bits des mots reçus $\widetilde{\mathrm{t}}_i^j$. Ensuite, le théorème 4.3 assure que $\Omega\left(\exp(c\log n)\right)$ échantillons sont nécessaires (où c est une constante ne dépendant que de q) si l'on veut reconstituer le mot transmis $\mathbf{t}$ à partir des lois des $\widetilde{\mathrm{t}}_i^j$ uniquement, et non de leurs lois conjointes.

Ce mémoire comporte quatre grandes sections. Dans la première, on rappelle des résultats classiques de probabilité dont on se servira à plusieurs reprises comme les bornes de Chernoff ou la notion de distance en variation totale. On y fixe également des notations et définitions utiles tout au long de notre étude.

La deuxième section s'inspire grandement de [1]. On y introduit d'abord l'algorithme de majorité bit par bit, puis des études de marches aléatoires nous permettent de prouver le théorème 2.9. Ensuite, on s'intéresse à une modification de l'algorithme pour reconstituer des mots qui ne sont pas forcément "gentils". Enfin, on étudie une borne inférieure du nombre d'échantillons nécessaires.

Pour démontrer les théorèmes 4.2 et 4.3, l'article [5] propose de passer par les séries génératrices des mots et échantillons. C'est pourquoi dans la troisième section, on établit de nombreux résultats d'analyse complexe. Le but général y est de maîtriser des polynômes sur des parties simples du plan complexe telles des droites, des cercles ou des ellipses.

Enfin, on utilise ces résultats dans la quatrième section pour prouver les théorèmes 4.2 et 4.3. Pour cela, on commence par prouver le lemme 4.1 qui permettra de faire le lien entre les lois des bits reçus et le mot transmis en terme de séries génératrices. C'est donc ce résultat qui transforme le problème de probabilité en un problème d'analyse complexe.

Les résultats présentés ne sont pas les plus récents. Voici les dernières bornes obtenues sur ce problème par des articles qui ne seront pas abordés dans ce mémoire. Ainsi, avec q constant, l'article [6] montre qu'il est possible de reconstruire la plupart des mots transmis $\mathbf{t}$ avec grande probabilité à l'aide de $e^{O(\sqrt{\log n})}$ échantillons. L'article [7] montre qu'on peut même le faire avec $e^{O(\log^{1/3}(n))}$ échantillons. L'article [8] donne une meilleure borne inférieure pour le nombre de mots reçus nécessaires à la reconstruction d'un mot $\mathbf{t}$ arbitraire, de l'ordre de $\Omega(n^{5/4})$, en trouvant un exemple de deux mots particulièrement difficiles à distinguer. Le même exemple donne même une borne inférieure plus élevée de l'ordre de $\Omega(n^{3/2})$, comme le montre l'article [9].

1.1 Quelques outils techniques

Ce premier résultat classique de probabilités donne des bornes exponentielles pour maîtriser les écarts à leurs moyennes de certaines variables aléatoires.

Théorème 1.1. ***Bornes de Chernoff***
Soient n un entier, $X_1, X_2, \ldots, X_n$ des variables de Bernoulli indépendantes de paramètres respectifs $p_1, p_2, \ldots, p_n$, et $X = \sum_{i=1}^n X_i$. On pose également $\mu = \mathbb{E}[X]$. Alors :

1) $\forall \delta > 0,\ \ \mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(\frac{-\delta^2}{2+\delta}\mu\right)$.

2) $\forall \delta \in [0,1],\ \ \mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(\frac{-\delta^2}{2}\mu\right)$.

En particulier : $\forall \delta \in [0,1],\ \ \mathbb{P}(|X-\mu| \geq \delta\mu) \leq 2\exp\left(\frac{-\delta^2}{3}\mu\right)$.

Démonstration. Soit $i \in [\![1, n]\!]$. Par l'inégalité de Markov, on a pour $a \in \mathbb{R}$ et $s > 0$

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{sX} \geq e^{sa}) \leq e^{-sa}\mathbb{E}[e^{sX}].$$

Or, par convexité : $\mathbb{E}[e^{sX_i}] = (1-p_i) + p_i e^s = 1 + p_i(e^s - 1) \leq \exp(p_i(e^s-1))$.
Par indépendance des X_i, il en découle

$$\mathbb{E}[e^{sX}] \leq \prod_{i=1}^n \exp(p_i(e^s-1)) = \exp(\mu(e^s-1)).$$

Pour obtenir la première inégalité, on applique ces résultats à $a = (1+\delta)\mu$ et $s = \ln(1+\delta)$:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-sa}\exp(\mu(e^s-1)) = \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^\mu.$$

Or pour $x > 0$, on a : $\ln(1+x) \geq \frac{x}{1+x/2}$. On en déduit : $(\delta - (1+\delta)\ln(1+\delta))\mu \leq \frac{-\delta^2}{2+\delta}\mu$.
En passant à l'exponentielle, on obtient finalement : $\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(\frac{-\delta^2}{2+\delta}\mu\right)$.
La deuxième inégalité se prouve de manière similaire. On utilise que, si $s > 0$,

$$\mathbb{P}(X \leq a) = \mathbb{P}(e^{-sX} \geq e^{-sa}) \leq e^{sa}\mathbb{E}[e^{-sX}].$$

L'inégalité de convexité reste vraie si l'on remplace s par $-s$, et on peut appliquer cela à $a = (1+\delta)\mu$ et $s = -\ln(1-\delta)$:

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp(-\mu(\delta + (1-\delta)\ln(1-\delta)))$$

où, pour $\delta \in [0,1]$: $(1-\delta)\ln(1-\delta) \geq -\delta + \delta^2/2$. D'où la deuxième inégalité. La troisième inégalité voulue est une simple conséquence des deux précédentes, en majorant $2 \leq 3$ et $2 + \delta \leq 3$.

□

On va maintenant introduire la notion de distance en variation totale, et établir des résultats classiques. Cette notion sera importante dans les sections 2.5 et 4.2, qui s'intéressent à des bornes inférieures sur le nombre d'échantillons nécessaires pour reconstituer le mot transmis. En effet, la définition qui suit apparaît naturellement lorsque l'on veut formaliser l'expression "ces deux lois sont indistinguables" ; distinguer la loi de X de la loi de Y signifie alors trouver un événement A tel que $\mathbb{P}(X \in A)$ soit éloigné de $\mathbb{P}(Y \in A)$.

Définition. Si μ et ν sont des mesures de probabilité sur un espace mesuré $(E, \mathcal{E})$, on définit leur distance en variation totale par

$$\mathrm{d_{vt}}(\mu, \nu) = \sup_{A \in \mathcal{E}} |\mu(A) - \nu(A)|.$$

Si X et Y sont des variables aléatoires à valeurs dans $(E, \mathcal{E})$, on définit leur distance en variation totale $\mathrm{d_{vt}}(X, Y)$ comme la distance en variation totale entre leurs lois.

Lemme 1.2. *Si X et Y sont des variables aléatoires à valeurs dans $(E, \mathcal{E})$, alors*

$$\mathrm{d_{vt}}(X, Y) \leq \mathbb{P}(X \neq Y).$$

Démonstration. Soit $A \in \mathcal{E}$. Si $X(\omega) \neq Y(\omega)$, alors $\mathbf{1}_{X \in A}(\omega) - \mathbf{1}_{Y \in A}(\omega) \leq 1 = \mathbf{1}_{X \neq Y}(\omega)$. Si $X(\omega) = Y(\omega)$, alors $\mathbf{1}_{X \in A}(\omega) - \mathbf{1}_{Y \in A}(\omega) = 0 = \mathbf{1}_{X \neq Y}(\omega)$. Le résultat en découle en passant à l'espérance. □

Définition. On dit qu'un couple de variables aléatoires (X, Y) sur un même espace de proba à valeurs dans $(E, \mathcal{E})$ est un couplage de μ et ν si X suit la loi μ et Y suit la loi ν.

Théorème 1.3. *On suppose ici que E est fini ou dénombrable et on le munit de la tribu discrète $\mathcal{P}(E)$. Si l'on note $a \wedge b$ le minimum entre a et b, on a alors :*

$$\begin{aligned}
\mathrm{d_{vt}}(\mu, \nu) &= \max\{|\mu(A) - \nu(A)| \mid A \subset E\} \\
&= \frac{1}{2}\sum_{x \in E} |\mu(x) - \nu(x)| \\
&= 1 - \sum_{x \in E} \mu(x) \wedge \nu(x) \\
&= \min\{\mathbb{P}(X \neq Y) \mid (X, Y) \textit{ couplage de } \mu \textit{ et } \nu\}
\end{aligned}$$

Démonstration. Pour la première égalité, on parachute le bon événement :

$$B = \{\mu > \nu\}$$

et on note $m := \mu(B^c) - \nu(B^c)$ et $M := \mu(B) - \nu(B) = -m$. On a pour tout $A \subset E$:

$$m \leq \mu(A) - \nu(A) \leq M.$$

En effet,

$$\mu(A)-\nu(A)=\sum_{x\in A}\mu(x)-\nu(x)=\sum_{x\in A\cap B}\mu(x)-\nu(x)+\sum_{x\in A\cap B^c}\mu(x)-\nu(x)\leq\sum_{x\in A\cap B}\mu(x)-\nu(x)\leq M$$

et de même pour l'autre inégalité. Il vient alors que $\mathrm{d_{vt}}(\mu,\nu)\leq M$, et donc l'événement B atteint bien le supremum.
La deuxième égalité résulte de

$$\mathrm{d_{vt}}(\mu,\nu)=M=\frac{1}{2}M+\frac{1}{2}M=\frac{1}{2}(M-m)=\frac{1}{2}\sum_{x\in B}(\mu(x)-\nu(x))+\frac{1}{2}\sum_{x\in B^c}(\nu(x)-\mu(x))=\frac{1}{2}\sum_{x\in E}|\mu(x)-\nu(x)|.$$

La troisième en découle :

$$\frac{1}{2}\sum_{x\in E}|\mu(x)-\nu(x)|=\frac{1}{2}\sum_{x\in E}(\mu(x)+\nu(x)-2(\mu(x)\wedge\nu(x)))=1-\sum_{x\in E}\mu(x)\wedge\nu(x).$$

Montrons la quatrième égalité. On sait déjà par le lemme 1.2 que

$$\mathrm{d_{vt}}(\mu,\nu)\leq\inf\{\mathbb{P}(X\neq Y)\,|\,(X,Y)\text{ couplage de }\mu\text{ et }\nu\}.$$

Si $\mathrm{d_{vt}}(\mu,\nu)=0$ alors $\mu=\nu$. On prend alors une variable X de loi μ et le couplage (X,X) convient. Si $\mathrm{d_{vt}}(\mu,\nu)=1$, il existe d'après la première égalité du théorème 1.3 un événement $B\subset E$ tel que $\mu(B)=1$ et $\nu(B)=0$ (quitte à prendre le complémentaire). Dans ce cas tout couplage convient. Supposons désormais $\mathrm{d_{vt}}(\mu,\nu)\in]0,1[$. On pose

$$\alpha=1-\mathrm{d_{vt}}(\mu,\nu)=\sum_{x\in E}\mu(x)\wedge\nu(x)$$

et on prend U,V,W trois variables aléatoires indépendantes dont les lois sont données par :

$$\begin{aligned}\mathbb{P}(U=x)&=\frac{\mu(x)\wedge\nu(x)}{\alpha}\\ \mathbb{P}(V=x)&=\frac{\mu(x)-\mu(x)\wedge\nu(x)}{1-\alpha}\\ \mathbb{P}(W=x)&=\frac{\nu(x)-\mu(x)\wedge\nu(x)}{1-\alpha}.\end{aligned}$$

On se donne également B indépendante de U,V,W et suivant une loi de Bernoulli de paramètre α. On définit enfin X et Y par : si $B=1$ alors $X=Y=U$ et si $B=0$ alors $X=V$ et $Y=W$. Alors

$$\mathbb{P}(X=x)=\mathbb{P}(B=0)\mathbb{P}(U=x)+\mathbb{P}(B=1)\mathbb{P}(V=x)=\mu(x)$$

et de même Y suit la loi de ν. De plus :

$$\mathbb{P}(X\neq Y)\leq\mathbb{P}(B=0)=\mathrm{d_{vt}}(\mu,\nu)$$

donc c'est en fait une égalité par le lemme 1.2, et le théorème est démontré. □

1.2 Définitions

Définition. Soit $\mathbf{t}$ un mot de longueur n. On appelle run de 0 (resp. de 1) toute sous-chaîne maximale de 0 (resp. de 1) consécutifs dans $\mathbf{t}$. On note alors L_i le i-ième run de $\mathbf{t}$ et ℓ_i sa longueur, de sorte que $\mathbf{t}=L_1L_2\dots L_k$ et $\ell_1+\ell_2+\dots+\ell_k=n$.

On remarquera :

- Si L_i est un run de $\varepsilon\in\{0,1\}$, alors L_{i+1} est un run de $(1-\varepsilon)$.
- Un échantillon $\widetilde{\mathbf{t}}$ comporte au plus autant de runs que $\mathbf{t}$.

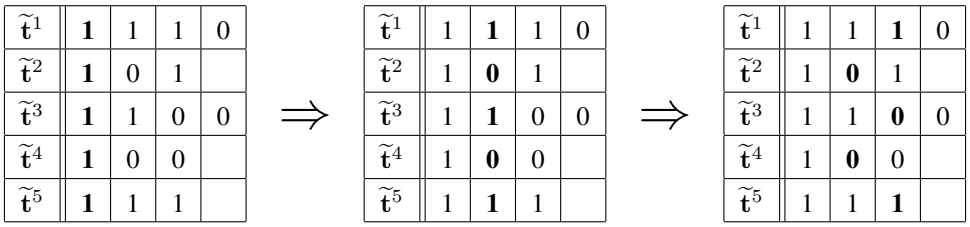

FIGURE 1 – Un exemple des premières étapes de l'algorithme de majorité bit par bit avec cinq échantillons. Les bits sur lesquels pointent les compteurs sont en gras. Le mot finalement reconstitué est 11010. Les cases blanches ne sont pas comptées dans le vote.

Définition. Si $\mathbf{t} = L_1 L_2 \dots L_k$ est le mot transmis et $\widetilde{\mathbf{t}}$ un échantillon, on dit que les runs L_i et L_{i+2} ont fusionné dans $\widetilde{\mathbf{t}}$ si les bits de L_{i+1} ont tous été supprimés au cours de l'algorithme de délétion.

On introduit enfin deux notations :
Si $\varepsilon \in \{0,1\}$ et $\ell \in \mathbb{N}$, on note ε^ℓ le mot constitué de ℓ fois le même bit ε.
Si $\mathbf{t} \in \{0,1\}^n$ et $k \leq n$, on note $\mathbf{t}|_k$ le mot constitué des k premiers bits de $\mathbf{t}$.

2 Étude du problème lorsque la probabilité de délétion tend vers 0

L'article [1] aborde le problème de cette façon : si l'on fait tendre n vers l'infini et q_n vers 0, de combien d'échantillons a-t-on besoin pour retrouver $\mathbf{t}$ avec une probabilité tendant vers 1 ? Mais précisons d'abord ce que l'on entend par "tant d'échantillons suffisent pour retrouver $\mathbf{t}$".

Définition. On dit que (m_n) échantillons suffisent si l'on dispose d'un algorithme et d'une suite $(p_n) \in [0,1]^{\mathbb{N}}$ tendant vers 1 tels que pour tout entier n et pour tout mot $\mathbf{t}$ de longueur n, la probabilité que l'algorithme retrouve $\mathbf{t}$ à partir de m_n échantillons est d'au moins p_n.

Définition. On dit que (m_n) échantillons suffisent dans la plupart des cas si l'on dispose d'un algorithme, d'une suite $(p_n) \in [0,1]^{\mathbb{N}}$ tendant vers 1 et d'une suite $(s_n) \in [0,1]^{\mathbb{N}}$ tendant vers 1 tels que pour tout entier n, il existe un ensemble de mots de longueur n noté J_n tel que $\frac{|J_n|}{2^n} \geq s_n$ et pour tout mot $\mathbf{t} \in J_n$, la probabilité que l'algorithme retrouve $\mathbf{t}$ à partir de m_n échantillons est d'au moins p_n.

Par souci de légèreté, on ne notera désormais plus m_n ou q_n mais simplement m et q, même lorsque ces grandeurs dépendent de n.

2.1 L'algorithme de majorité bit par bit

Pour toute constante d assez grande, un algorithme simple fonctionne pour la plupart des mots si $q \leq \frac{1}{d \log n}$. Voici comment il procède. À chaque échantillon $\widetilde{\mathbf{t}}^j$ est associé un compteur c_j^τ initialisé à $c_j^0 = 0$. À chaque itération τ, soit w_τ le bit donné par la majorité des $\widetilde{\mathrm{t}}^j_{c_j^\tau}$. Pour chaque j ayant voté pour w_τ, on incrémente $c_j^{\tau+1} = c_j^\tau + 1$ et les autres restent inchangés : $c_j^{\tau+1} = c_j^\tau$. Un exemple est donné en figure 1.

Pour $q = O(\frac{1}{\log n})$, plus précisément avec $q \leq \frac{1}{d \log n}$ où d est une constante assez grande, l'article [1] montre que $m = O(\log n)$ échantillons suffisent dans la plupart des cas en appliquant cet algorithme simple.

2.2 Un ensemble de mots qui conviennent

On décrit dans cette partie deux conditions qui sont satisfaites par une proportion tendant vers 1 de mots binaires de longueur n. On montrera dans la suite que l'algorithme de majorité bit par bit reconstitue les mots $\mathbf{t}$

vérifiant ces conditions avec probabilité tendant vers 1. Si $\mathbf{t} = L_1 \dots L_r$ est tel que

$$\forall i \in [\![1, r - 2k \log n + 1]\!], \forall h \in [\![1, k' \log n]\!], \exists i' \in [\![i, i - 1 + 2k \log n]\!], l_{i'} < l_{i'+2h} \tag{1}$$

pour certaines constantes k et k', on dit qu'il satisfait la condition (1) pour k, k'. Cela revient à dire que, dans tout segment de $2k \log n$ runs consécutifs de $\mathbf{t}$, pour tout pas $h \leq k' \log n$, il existe deux runs dans ce segment distants de $2h$ (donc ayant tous les deux des 0 ou tous les deux des 1) tels que le second soit strictement plus long que le premier. L'idée est que ce run plus long servira à "ralentir" les compteurs ayant pris de l'avance lors de l'algorithme.

On pose $J_n(k, k')$ l'ensemble des mots de longueur n dont les runs sont petits, *i.e.* de longueur au plus $2 \log n$, et satisfaisant la condition (1) pour k, k'. Montrons que pour tout k' et pour tout $k \geq k'$ assez grands, les ensembles $J_n(k, k')$ conviennent.

Lemme 2.1. *On a* $\frac{|J_n(k,k')|}{2^n} \geq 1 - \frac{2}{n}$. *Autrement dit, si* $\mathbf{t}$ *est un mot binaire aléatoire uniforme de longueur* n, *alors* $\mathbb{P}(\mathbf{t} \in J_n(k, k')) \geq 1 - \frac{2}{n}$.

Démonstration. La première condition est vérifiée par un mot binaire aléatoire uniforme $\mathbf{t} = X_1 \dots X_n$ de longueur n avec probabilité au moins $1 - \frac{1}{n}$ car

$$\begin{aligned}
\mathbb{P}\left(\bigcup_i \{X_i = \dots = X_{i+2\log n}\}\right) &\leq n\mathbb{P}(X_1 = \dots = X_{1+2\log n}) \\
&\leq 2n\mathbb{P}(X_1 = \dots = X_{1+2\log n} = 1) \\
&= \frac{2n}{2^{1+2\log n}} \\
&= \frac{1}{n}.
\end{aligned}$$

Étant donnés k et $k' \leq k$, calculons maintenant la probabilité pour que (1) soit mise en défaut. On a envie de dire que les ℓ_j suivent des lois géométriques de paramètre $\frac{1}{2}$ indépendantes car cela permettrait de calculer cette probabilité facilement. Bien malheureusement ce n'est pas exactement le cas (par exemple car la longueur de $\mathbf{t}$ est fixée). On va donc se donner une variable Y suivant une loi de Bernoulli de paramètre $\frac{1}{2}$ et des variables ℓ'_j pour $j \in [\![1, n]\!]$ suivant des lois géométriques de paramètre $\frac{1}{2}$, toutes indépendantes. Puis on considère $\mathbf{t}' = Y^{\ell'_1}(1-Y)^{\ell'_2} Y^{\ell'_3} \dots$, qui est mot binaire aléatoire de longueur au moins n. On peut alors vérifier que $\mathbf{t}'|_n$ suit la même loi que $\mathbf{t}$. Majorons maintenant la probabilité que (1) soit fausse pour $\mathbf{t}'$. On a

$$\forall j_1, j_2\,, \quad \mathbb{P}(\ell'_{j_1} \geq \ell'_{j_2}) = \sum_{p \geq 1} \mathbb{P}(\ell'_{j_2} = p)\mathbb{P}(\ell'_{j_1} \geq p) = \sum_{p \geq 1} \frac{1}{2^{p+p-1}} = \frac{2}{3}.$$

Soit $A_{i,h} = \{\forall j \in [\![0, 2k \log n - 2h - 1]\!], \ell_{i+j} \geq \ell_{i+j+2h}\}$. Pour tout $i \in [\![1, n - 2k \log n + 1]\!]$, pour tout $h \in [\![1, k' \log n]\!]$, on a $\mathbb{P}(A_{i,h}) \leq \mathbb{P}(A_{i,k' \log n}) \leq \left(\frac{2}{3}\right)^{\frac{2k \log n - 2k' \log n - 1}{2}}$ en ne considérant que la moitié des paires $\{i + j, i + j + 2h\}$ pour qu'elles soient deux à deux disjointes et que les évènements $\{\ell_{i+j} \geq \ell_{i+j+2h}\}$ soient indépendants. On a ainsi

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{1 \leq i \leq n - 2k \log n + 1} \bigcup_{1 \leq h \leq k' \log n} A_{i,j}\right) &\leq nk' \log n \left(\frac{2}{3}\right)^{\frac{2k \log n - 2k' \log n - 1}{2}} \\
&= \sqrt{\frac{3}{2}} nk' \log(n) n^{(1 - \log 3)(k - k')}
\end{aligned}$$

qui est bien inférieure à $1/n$ pour k assez grand. Reste à montrer que, quitte à prendre k légèrement plus grand, on a le même résultat pour $\mathbf{t}$. En appliquant le résultat sur $\mathbf{t}'$ et en remplaçant par exemple k par $k + 1$, on a le résultat voulu. En effet, si on se donne un segment de $2(k + 1) \log n$ runs consécutifs de $\mathbf{t}'|_n$, alors les $2k \log n$ premiers

sont aussi des runs de $\mathbf{t}'$ car seul le dernier peut être tronqué. Donc la probabilité que (1) pour $k+1, k'$ ne soit pas vérifiée par $\mathbf{t}'|_n$ est inférieure à celle que (1) pour k, k' ne soit pas vérifiée par $\mathbf{t}'$, et on a majoré cette dernière par $\frac{1}{n}$. Ainsi, $\mathbf{t}$ vérifie lui aussi la condition (1) avec probabilité au moins $1 - \frac{1}{n}$.

□

2.3 Correction de l'algorithme dans ce cadre

On ne considère désormais que le cas où le mot transmis est dans $J_n(k, k')$. Le désavantage de l'algorithme de majorité bit par bit est que lorsqu'un vote donne le mauvais résultat, cela décale mal les compteurs et le vote suivant a encore moins de chance d'être le bon. Donc le résultat d'un vote dépend fortement des précédents. On introduit donc un autre algorithme de reconstruction qui "triche" dans le sens où, à chaque étape τ, au lieu de procéder à un vote, il impose comme résultat du vote le τ-ième bit de $\mathbf{t}$ et incrémente les curseurs c_j^τ selon ce bit. Remarquons que dans cet algorithme, chaque compteur c_j ne dépend que de $\widetilde{\mathbf{t}}^j$. Dans presque toute cette partie, on considèrera l'algorithme "qui triche", et non l'algorithme de majorité bit par bit. On verra qu'avec grande probabilité, la majorité des compteurs de l'algorithme qui triche sont bien positionnés à chaque étape. Puis, on en déduira par récurrence qu'avec grande probabilité, l'algorithme de majorité bit par bit fonctionne exactement de la même manière que l'algorithme qui triche, au sens où il fait les bons votes.

Définition. On dira qu'un compteur c_j est en avance (resp. retard) au temps τ lorsque le bit $\widetilde{\mathrm{t}}^j_{c_j^\tau}$ provient du bit t_i pour un certain $i > \tau$ (resp. $i < \tau$).

Lors de l'exécution de l'algorithme "qui triche", les compteurs risquent toujours d'être décalés. On introduit un processus aléatoire qui va contrôler ce décalage : étant donné un mot $\mathbf{t}$ fixé de longueur n et $\widetilde{\mathbf{t}}^1, \dots, \widetilde{\mathbf{t}}^m$ des échantillons, on lance l'algorithme "qui triche" sur l'ensemble de ces $\widetilde{\mathbf{t}}^j$. Pour chaque $j \in [\![1, m]\!]$, on définit un processus aléatoire $\left(R_u'^{(j)}\right)_u$. Fixons $j \in [\![1, m]\!]$. Pour tout i, on note $\varphi(i)$ la position dans $\mathbf{t}$ du i-ème bit de $\widetilde{\mathbf{t}}^j$ lui correspondant. On pose

$$\forall j \in [\![1, m]\!], \forall u \in \mathbb{N}, \quad R_u'^{(j)} := \text{ nombre de runs de } \mathbf{t} \text{ entre les positions } \varphi(c_j^{\tau(u)}) \text{ et } \tau(u)$$

avec $\tau(u) = \sum_{i=1}^{u\lfloor 2k \log n \rfloor} \ell_i$. Autrement dit, une étape du processus correspond à la lecture de $\lfloor 2k \log n \rfloor$ runs de $\mathbf{t}$ et $R_u'^{(j)}$ est l'avance en terme de runs de la lecture de $\widetilde{\mathbf{t}}^j$ par rapport à la lecture de $\mathbf{t}$ à l'étape de l'algorithme correspondante. Les compteurs ne sont jamais en retard au sens où $\forall \tau,\ \varphi(c_j^\tau) \geq \tau$ (ils n'ont pas de retard au temps $\tau = 1$ et si c_j^τ est en avance, alors $c_j^{\tau+1}$ n'est pas en retard, et si c_j^τ pointe vers le bon bit, alors $c_j^{\tau+1} = c_j^\tau + 1$) mais peuvent prendre de l'avance. Les compteurs n'étant jamais en retard, on a $R_u'^{(j)} \geq 0$ pour tout j et à toute étape u. On va maintenant montrer que les $R'^{(j)}$ passent la plupart du temps en 0. Ainsi, on aura contrôlé le décalage aux temps $\tau(u)$. Puis, on le contrôlera entre les $\tau(u)$.

On a ainsi défini, à $\mathbf{t}$ fixé, m marches aléatoires indépendantes qui dépendent chacune d'un des $\widetilde{\mathbf{t}}^j$ (car on a défini ces marches à partir de l'algorithme qui triche). Il sera plus aisé de travailler avec la marche aléatoire R de matrice de transition P telle que :

$$P_{i,j} = \begin{cases} \alpha^{j-i} & \text{si } i < j \text{ et } i < k' \log n \\ \beta & \text{si } j = i-1 \text{ et } 0 < i < k' \log n \\ \dfrac{1}{1-\beta}\alpha^{j-i} & \text{si } i < j \text{ et } i \geq k' \log n \\ 1 - \displaystyle\sum_{l \neq i} P_{i,l} & \text{si } i = j \end{cases}$$

où $\alpha = \frac{2k}{d}$ et $\beta = e^{-2\alpha}$. Précisons tout de même que les indices i et j ici présents n'ont aucun lien avec les précédents. Pour d assez grand, on aura $\alpha \approx 0$ et $\beta \approx 1$ et on montrera que la marche a tendance à repasser très

0	1	0	**0**	1	1	1
0	1	0	0	**1**	1	1
0	1	0	0	1	**1**	1
0	1	0	0	1	**1**	1
0	1	0	0	1	**1**	1
0	1	0	0	1	**1**	1
0	1	0	0	1	1	**1**

FIGURE 2 – Un exemple d'application de l'algorithme qui triche. On considère ici un segment 001000111 d'un mot $\mathbf{t}$ qui a donné 0100111 dans l'échantillon. La i-ième colonne correspond au i-ième bit du segment et la τ-ième ligne correspond à la τ-ième étape de l'algorithme dans ce segment. Le compteur, qui pointe le bit en gras, finit donc par perdre un run d'avance.

souvent en 0. Montrons déjà que l'on peut se ramener à cette marche R.

Définition. Soient deux variables aléatoires X et Y à valeurs dans un espace ordonné. On dit que X domine Y s'il existe X' et Y' respectivement de mêmes lois que X et Y telles que $X' \geq Y'$ presque sûrement.

Lemme 2.2. *Soit R' l'un des $R'^{(j)}$. Alors pour n assez grand, R domine R'. En particulier, pour toute étape u, on a $\mathbb{P}(R_u = 0) \leq \mathbb{P}(R'_u = 0)$.*

Démonstration. Si $R'_u = 1$, alors, en supposant qu'aucun run du segment ne soit totalement supprimé, $R'_{u+1} = 0$. En effet, le compteur pointe par exemple sur un run de 0 alors que dans $\mathbf{t}$, l'algorithme lit un run de 1. Il finit donc de lire le run de 1 avant de commencer à avancer le compteur. Supposons maintenant que $R'_u = h \in [\![2, k' \log n]\!]$. On sait qu'il existe des runs L_i et $L_{i+2\lfloor h/2 \rfloor}$ dans le segment de $2k \log n$ runs considéré tels que $\ell_i < \ell_{i+2\lfloor h/2 \rfloor}$. Alors, si aucun run n'est totalement supprimé parmi les $2k \log n$ runs après le run sur lequel le compteur pointe et si $L_{i+2\lfloor h/2 \rfloor}$ est intact, on aura $R'_{u+1} \leq h - 1$. En effet, l'algorithme n'aura pas fini de lire $L_{i+2\lfloor h/2 \rfloor}$ à la place de L_i dans $\widetilde{\mathbf{t}}^j$ et donc perdra un run d'avance. Un exemple est donné en figure 2.
Or la probabilité que ces conditions soient réunies est au moins

$$(1-q)^{2\log n}(1-q)^{2k\log n} \geq \left(1 - \frac{1}{d\log n}\right)^{(2k+2)\log n} \geq \left(1 - \frac{1}{d\log n}\right)^{3k\log n} \geq e^{\frac{-4k}{d}} = e^{-2\alpha} = \beta$$

pour n assez grand.

Par ailleurs, si $R'_u = h$, alors on ne peut avoir $R'_{u+1} = h' > h$ que si exactement $h' - h$ runs sont supprimés dans le segment considéré des runs $L_{v+1}, \ldots, L_{v+2k\log n}$ de $\widetilde{\mathbf{t}}^j$. Cela n'arrive qu'avec une proba égale à :

$$\begin{aligned}
\sum_{\substack{A \subset [\![v+1, v+2k\log n]\!] \\ \text{tels que } |A| = h'-h}} q^{\sum_{i\in A}\ell_i}(1-q)^{\sum_{i\in[\![v+1,v+2k\log n]\!]\setminus A}\ell_i} &\leq \binom{2k\log n}{h'-h} q^{h'-h}(1-q)^{2k\log n-(h'-h)} \\
&\leq (2k\log n)^{h'-h}\left(\frac{1}{d\log n}\right)^{h'-h} \\
&= \left(\frac{2k}{d}\right)^{h'-h} = \alpha^{h'-h}
\end{aligned}$$

(la majoration ci-dessus est aussi valable pour $h \geq k' \log n$).

□

Montrons maintenant que la marche R reste souvent en 0. Le lemme 2.2 permettra d'en déduire que les marches $R'^{(j)}$ restent toutes souvent en 0.

Lemme 2.3. *Notons N le nombre d'étapes de la marche R. On peut choisir k, k' et d de sorte que*

$$\forall u \in [\![1, N]\!], \quad \mathbb{P}(R_u = 0) \geq \frac{19}{20}.$$

Nous présentons ici une démonstration différente de [1] s'appuyant sur le théorème du scrutin (en anglais *ballot theorem*) suivant que l'on peut par exemple trouver dans [2]. C'est un "théorème du scrutin" dans le sens où, si l'on a deux candidats tels que, lors du décompte, la marche monte lorsque le premier reçoit une voix et la marche descend lorsque le second en a une, alors le théorème donne la probabilité que le premier candidat soit en tête tout au long du scrutin.

Théorème 2.4. *Soit (X_n) une suite de v.a.i.i.d. à valeurs entières, d'espérance μ telles que $X_n \leq 1$ et soit $S_n = X_1 + \ldots + X_n$. Alors*

$$\mathbb{P}(\forall n \in \mathbb{N}^*, S_n > 0) = \begin{cases} \mu & \text{si } \mu > 0 \\ 0 & \text{sinon.} \end{cases}$$

Remarquons que d'après la loi forte des grands nombres, si $\mu > 0$, alors $\mathbb{P}(S_n > 0 \text{ a.p.c.r}) = 1 > 0$. Ainsi il existe $N \in \mathbb{N}$ tel que $\mathbb{P}(\forall n \geq N,\ S_n > 0) > 0$ et donc $\mathbb{P}(\forall n \in \mathbb{N}^*, S_n > 0) > 0$. Inversement si $\mu < 0$, la loi forte des grands nombres assure que $\mathbb{P}(\forall n \in \mathbb{N}^*, S_n > 0) = 0$. Une preuve du théorème 2.4 est donnée à la fin de cette section. On peut maintenant démontrer le lemme 2.3.

Démonstration du lemme 2.3 à l'aide du théorème 2.4. Soient X_i des v.a.i.i.d telles que

$$\begin{aligned} &\mathbb{P}(X_i = -1) = \beta \\ &\mathbb{P}(X_i = j) = \alpha^j \text{ si } j \geq 1 \\ &\mathbb{P}(X_i = 0) = 1 - \beta - \sum_{j \geq 1} \alpha^j \end{aligned}$$

Soit $S_i = X_1 + \ldots + X_i$. Cette marche se comporte à peu près de la même manière que R à deux différences près : elle peut passer en dessous de 0 et elle se comporte toujours de la même façon lorsqu'elle dépasse $k' \log n$. Précisons cette remarque.

Soit $W_u = S_u - \min_{i \in [\![0,u]\!]} S_i$. Notons A_R l'évènement $\{\forall i \in [\![0, N]\!], R_i < k' \log n\}$. On définit de même A_W. Les lois conditionnelles de R sachant A_R et de W sachant A_W sont les mêmes car R et W sont des chaînes de Markov qui ont les mêmes probabilités de transitions tant qu'elles restent en dessous de $k' \log n$.

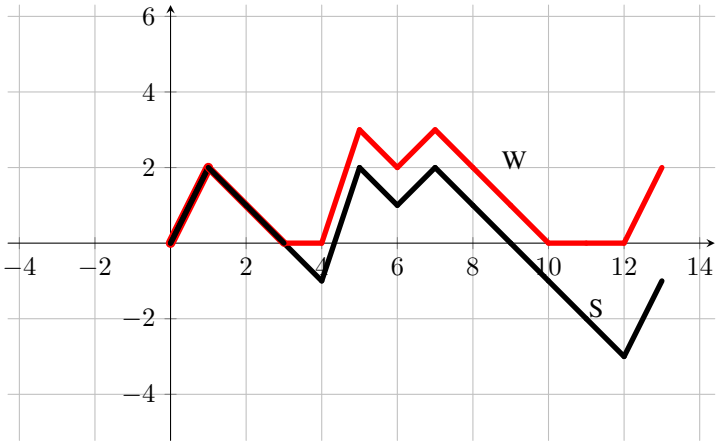

FIGURE 3 – Les marches S (en noir) et W (en rouge)

On a donc

$$\begin{aligned}
\mathbb{P}(R_u = 0) &\geq \mathbb{P}(R_u = 0 \text{ et } A_R) \\
&= \mathbb{P}(W_u = 0 \text{ et } A_W) \\
&= \mathbb{P}(S_u = \min_{i \in [\![0,u]\!]} S_i \text{ et } \forall i \in [\![0,N]\!], S_i < k' \log n + \min_{i \in [\![0,u]\!]} S_i) \\
&= \mathbb{P}\left(\forall i \in [\![0,u]\!], S_u \leq S_i \text{ et } \forall i < j \in [\![1,N]\!], \sum_{i \leq \ell \leq j} X_\ell < k' \log n \right) \\
&\geq \mathbb{P}(\forall i \in [\![0,u]\!], S_u \leq S_i) - \mathbb{P}\left(\exists i < j \in [\![1,N]\!], \sum_{i \leq \ell \leq j} X_\ell \geq k' \log n \right).
\end{aligned}$$

Majorons le terme de droite. On a pour tous $i < j$

$$\mathbb{P}\left(\sum_{i \leq \ell \leq j} X_\ell \geq k' \log n \right) = \mathbb{P}\left(e^{\sum_{i \leq \ell \leq j} X_\ell} \geq e^{k' \log n} \right) \leq \frac{\mathbb{E}\left[e^{\sum_{i \leq \ell \leq j} X_\ell} \right]}{e^{k' \log n}} = \frac{\mathbb{E}\left[e^{X_1} \right]^{j-i+1}}{e^{k' \log n}}.$$

Or

$$\mathbb{E}[e^{X_1}] = \frac{\beta}{e} + 1 - \beta - \sum_{j \geq 1} \alpha^j + \sum_{j \geq 1} \alpha^j e^j = 1 - \beta + \frac{\beta}{e} - \frac{\alpha}{1-\alpha} + \frac{\alpha e}{1 - \alpha e}$$

donc est proche de $\frac{1}{e}$ pour d assez grand, donc plus petit que 1 pour d assez grand.

Ainsi $\mathbb{P}\left(\sum_{i \leq \ell \leq j} X_\ell \geq k' \log n \right) \leq \frac{1}{e^{k' \log n}}$ et, pour k' assez grand,

$$\mathbb{P}\left(\exists i < j \in [\![1,N]\!], \sum_{i \leq \ell \leq j} X_\ell \geq k' \log n \right) \leq N^2 e^{-k' \log n} \leq n^2 n^{-k' \log e} \leq \frac{1}{n}.$$

Il ne reste plus qu'à minorer le terme de gauche. Pour cela, appliquons le théorème 2.4 aux $-X_i$. On calcule

$$\mathbb{E}[-X_i] = \beta - \frac{\alpha}{(1-\alpha)^2}.$$

Par dualité, $(S_i - S_u)_{0 \leq i \leq u}$ a même loi que $(-S_{u-i})_{0 \leq i \leq u}$. Alors

$$\mathbb{P}(\forall i \in [\![0,u]\!], S_i - S_u \geq 0) = \mathbb{P}(\forall i \in [\![0,u]\!], -S_i \geq 0) \geq \mathbb{P}(\forall i \geq 1, -S_i > 0) = \beta - \frac{\alpha}{(1-\alpha)^2}.$$

Ainsi, étant données les expressions de α et β, on peut prendre d grand pour que la probabilité soit arbitrairement proche de 1.

□

Maintenant que l'on a montré que R reste souvent en 0, on va montrer qu'avec grande probabilité, à chaque étape, la majorité des marches $R'^{(j)}$ est à l'état 0 et vote donc pour le bon bit.

Lemme 2.5. *Il existe une constante c telle que, si on considère $m \geq c \log n$ marches aléatoires i.i.d. de même loi que R, alors :*

$$\mathbb{P}\Big(\textit{pour chacune des } N \textit{ étapes de la marche } R, \textit{ au moins les } 3/4 \textit{ des marches sont à l'état } 0 \Big) \underset{n \to +\infty}{\longrightarrow} 1$$

Le lemme est aussi valable pour les $R'^{(j)}$ puisqu'elles sont indépendantes et chacune est dominée par la marche R. On peut donc toutes les dominer par m copies indépendantes de R.

Démonstration. Considérons $X_1, \dots, X_m$ des variables de Bernoulli i.i.d. de paramètre $p = 1/20$. Posons $X = \sum_{i=1}^{m} X_i$. Alors, par Chernoff :

$$\begin{aligned}
\mathbb{P}\Big(\exists u\,,\ \text{au moins } m/4 \text{ marches ne sont pas en 0 à l'instant } u\Big) &\leq N\mathbb{P}(X \geq m/4) \\
&\leq n\mathbb{P}(X \geq m/4) \text{ car } N \leq n \\
&= n\mathbb{P}(X \geq mp(1+\delta)) \\
&\leq n \exp\left(\frac{-\delta^2}{2+\delta} mp\right) \\
&\leq n^{1-\frac{\delta^2}{2+\delta} cp}
\end{aligned}$$

où $\delta = \frac{1}{4p} - 1$ vérifie $\delta > 0$. Pour c assez grand, cela tend vers 0.

□

On a montré qu'avec grande probabilité, au début de chaque segment de $2k \log n$ runs, les trois quarts des compteurs pointent vers le bon run. On veut montrer que, à tout instant τ, avec grande probabilité, la majorité des compteurs pointent vers la bonne lettre.

Lemme 2.6. *Si au moins $m' \geq c' \log n$ compteurs parmi les m sont bien positionnés au début de chaque segment (les segments de $2k \log n$ runs) avec c' assez grand, alors avec grande probabilité, au moins $\frac{8m'}{9}$ compteurs sont bien positionnés au début de chaque run.*

Démonstration. En fixant j et un segment,

$$\begin{aligned}
&\mathbb{P}(c_j \text{ ne pointe pas vers le bon run au début d'un run du segment}) \\
&\leq \mathbb{P}(\text{au moins un run du segment est supprimé}) \\
&\leq 2qk \log n.
\end{aligned}$$

Maintenant, j n'est plus fixé mais le segment l'est toujours. Notons

$$X'_j = \mathbf{1}_{\{c_j \text{ ne pointe pas vers le bon run au début d'un run du segment}\}}$$

et $X' = \sum_{j=1}^{m'} X'_j$, où on a supposé, quitte à réordonner les $\widetilde{\mathbf{t}}^j$, que les m' premiers $\widetilde{\mathbf{t}}^j$ ont leur compteur bien positionné au début du segment. Soient X_j des Bernoulli indépendantes de paramètre $2kq \log n$ et $X = \sum_{j=1}^{m'} X_j$. Alors, par Chernoff,

$$\begin{aligned}
\mathbb{P}(X' \geq \frac{m'}{9}) &\leq \mathbb{P}(X \geq \frac{\mu}{18qk \log n}) \text{ où } \mu = \mathbb{E}[X] \\
&= \mathbb{P}(X \geq (1+\delta)\mu) \text{ où } \delta = \frac{1}{18qk \log n} - 1 \\
&\leq e^{-\frac{\delta^2 \mu}{2+\delta}} \\
&\leq e^{-\frac{1}{18} c' \log n} \text{ pour } d \text{ assez grand} \\
&\leq \frac{1}{n^2} \text{ pour } c' \text{ assez grand.}
\end{aligned}$$

Donc en ne fixant plus le segment, on a

$$\begin{aligned}
&\mathbb{P}\left(\text{il existe un segment et un run dans ce segment tel qu'au moins } \frac{m'}{9} \text{ compteurs soient mal positionnés}\right)\\
&\leq N\frac{1}{n^2}\\
&\leq \frac{1}{n} \underset{n\to+\infty}{\longrightarrow} 0.
\end{aligned}$$

□

On en déduit le résultat suivant :

Lemme 2.7. *Pour c assez grand, avec grande proba, au début de tout run au moins deux tiers des compteurs (toujours pour l'algorithme "qui triche") sont bien positionnés. On notera E_1^n cet évènement.*

Démonstration. Soit c grande pour que c satisfasse le lemme 2.5 et que $c' = \frac{3}{4}c'$ satisfasse le lemme 2.6. D'après le lemme 2.5 la probabilité qu'au début de chaque segment de $2k\log n$ runs, au moins $3/4$ des compteurs soient bien positionnés tend vers 1 quand n tend vers $+\infty$. Par ailleurs, sachant cet évènement réalisé, en appliquant le lemme 2.6 avec $c' = \frac{3}{4}c$, on sait qu'au début de chaque run, $\frac{8}{9} \times \frac{3}{4} = \frac{2}{3}$ des compteurs sont bien positionnés avec une probabilité tendant vers 1 quand n tend vers $+\infty$. □

Reste à regarder l'intérieur des runs.

Lemme 2.8. *Pour c, d assez grands, la probabilité que pour tout run L_i, L_i soit intact dans au moins $\frac{11m}{12}$ échantillons tend vers 1 quand n tend vers l'infini. On note E_2^n cet évènement.*

Démonstration. C'est encore une conséquence des bornes de Chernoff car la probabilité que le run L_i ne soit pas intact est inférieure à $2q\log n$ (puisque sa longueur est au plus $2\log n$). □

Théorème 2.9. *Pour c assez grand, $m \geq c\log n$ échantillons conviennent dans la plupart des cas.*

Démonstration. Fixons n. Soit $\mathtt{t} \in J_n(k,k')$. Supposons E_1^n et E_2^n réalisés. Montrons par récurrence sur les runs que l'algorithme de majorité bit par bit reconstitue bien le $\mathtt{t}$, c'est-à-dire qu'il fait les bons votes à chaque étape. Supposons que les runs $L_1, \ldots, L_{i-1}$ soient bien reconstruits par l'algorithme de majorité. On sait que deux tiers des compteurs de l'algorithme qui triche sont bien positionnés au début du run L_i si E_1^n se produit. Or, jusqu'à présent, l'algorithme de majorité a voté pour les bons bits et a donc procédé comme l'algorithme qui triche. Ses compteurs pointent donc au mêmes endroits. Donc, sachant l'évènement E_1^n, deux tiers des compteurs de l'algorithme de majorité sont bien positionnés au début du run L_i. Par l'autre évènement E_2^n, $\frac{2}{3} - \frac{1}{12} = \frac{7}{12}$ des compteurs sont bien positionnés et pointent vers des runs intacts. Donc l'algorithme de majorité reconstruit bien le run L_i. □

Donnons enfin une démonstration du théorème 2.4.

Démonstration du théorème 2.4. Soit $n \in \mathbb{N}^*$ et $k \in \mathbb{Z}$. Montrons pour cela un résultat préliminaire, dit lemme cyclique :

$$\mathbb{P}\Big(\forall i \in [\![1,n]\!], S_i > 0 \;\Big|\; S_n = k\Big) = \begin{cases} \frac{k}{n} & \text{si } k \in [\![1,n]\!] \\ 0 & \text{sinon.} \end{cases} \tag{2}$$

Le théorème en résulte directement. En effet, comme $S_n \leq n$, on a :

$$\begin{aligned}
\mathbb{P}\left(\forall i \in [\![1,n]\!], S_i > 0\right) &= \sum_{k=1}^{n} \mathbb{P}\Big(\forall i \in [\![1,n]\!], S_i > 0 \;\Big|\; S_n = k\Big)\, \mathbb{P}(S_n = k)\\
&= \sum_{k=1}^{n} \frac{k}{n}\mathbb{P}(S_n = k)\\
&= \mathbb{E}\left[\frac{S_n}{n}\mathbf{1}_{S_n>0}\right].
\end{aligned}$$

Et comme $\frac{S_n}{n} \underset{n\to+\infty}{\longrightarrow} \mu$ en loi et $x \mapsto \min(x\mathbf{1}_{x>0}, 1)$ est continue bornée, on obtient (par définition de la convergence en loi) :

$$\mathbb{P}(\forall i \in \mathbb{N}^*, S_i > 0) = \lim_{n\to+\infty} \mathbb{P}(\forall i \in [\![1,n]\!], S_i > 0) = \lim_{n\to+\infty} \mathbb{E}\left[\frac{S_n}{n}\mathbf{1}_{S_n>0}\right] = \mu\mathbf{1}_{\mu\geq 0}.$$

Montrons maintenant (2). Pour ce faire, on pose $\forall i \in [\![1,n]\!], X'_i = X_i$ et $\forall i \in \mathbb{N}^*, X'_{n+i} = X'_i$ et $S'_i = X'_1 + \ldots + X'_i$. On pose aussi $\forall k \in \mathbb{N}, Y_k = \mathbf{1}_{\{\forall i>k, S'_i > S'_k\}}$. Les couples (Y_k, S_n) pour k dans $[\![1,n]\!]$ ont aussi même loi. En effet, $(Y_k, S_n) = (Y_k, S'_n) = (Y_k, S'_{n+k} - S'_k)$ et $(Y_k, S'_{n+k} - S'_k)$ suit la même loi que (Y_0, S'_n) car $(X'_{k+i})_{i\geq 1}$ suit la même loi que $(X'_i)_{i\geq 1}$.
De plus, $Y_0 = \mathbf{1}_{\{\forall i\in[\![1,n]\!], S'_i>0\}}$. En effet, si $S'_n \leq 0$, $Y_0 = 0$ et si $S'_n = k \in [\![1,n]\!]$, alors $\forall i > n, S'_i = S'_n + S'_{i-n} = k + S'_{i-n} > 0$ par récurrence. On a alors :

$$\mathbb{P}\Big(\forall i \in [\![1,n]\!] S'_i > 0 \,\Big|\, S'_n = k\Big) = \mathbb{E}\Big[Y_0 \,\Big|\, S'_n = k\Big] = \frac{1}{n}\mathbb{E}\left[\sum_{i=0}^{n-1} Y_i \,\middle|\, S'_n = k\right].$$

Montrons que $\sum_{i=0}^{n-1} Y_i = S'_n\mathbf{1}_{S'_n>0}$. Si $S'_n \leq 0$, alors $\forall i \in [\![1,n]\!], Y_i = 0$ et donc $\sum_{i=0}^{n-1} Y_i = 0$. Si $S'_n = k \in [\![1,n]\!]$, alors soit $m = \min_{0\leq i\leq n-1} S'_i$. Soient i_1 le dernier temps de passage de m à $m+1$,..., i_k le dernier temps de passage de $m+k-1$ à $m+k$. En d'autres termes, $\forall j \in [\![1,k]\!], i_j = \max\{i \in \mathbb{N}, X'_{i+1} = 1 \text{ et } S'_i = m+j-1\}$.

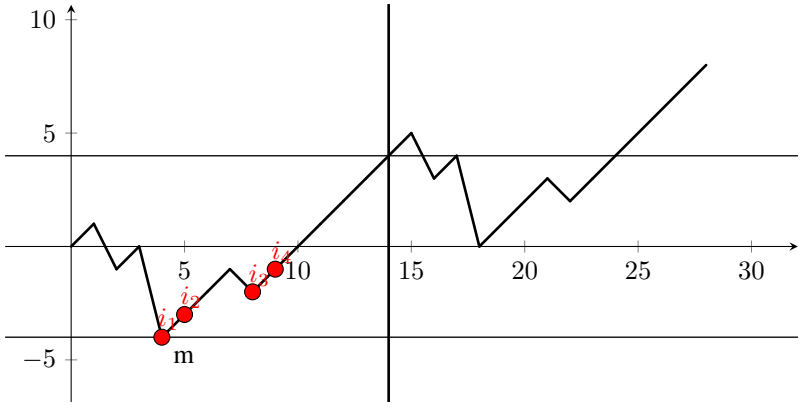

FIGURE 4 – Exemple de marche dans le cas où $n = 14$ et $k = 4$

On constate que $\forall j \in [\![1,k]\!], Y_{i_j} = 1$ et $\forall i \in [\![0,n-1]\!] \setminus \{i_1, \ldots i_k\}, Y_i = 0$, comme on le voit sur la figure 4. Ainsi, $\sum_{i=0}^{n-1} Y_i = S'_n\mathbf{1}_{S'_n>0}$ et donc

$$\mathbb{P}\Big(\forall i \in [\![1,n]\!], S'_i > 0 \,\Big|\, S'_n = k\Big) = \begin{cases} \frac{k}{n} & \text{si } k \in [\![1,n]\!] \\ 0 & \text{sinon.} \end{cases}$$

On en déduit le résultat voulu, puisque $\forall i \in [\![1,n]\!], S_i = S'_i$.

□

2.4 Une modification de l'algorithme pour reconstruire des mots arbitraires

On cherche dorénavant une suite (m_n) telle que m échantillons suffisent (voir définition). On suppose que $q \leq \frac{1}{n^{1/2+\varepsilon}}$, ce qui est une hypothèse bien plus forte sur la décroissance de q que précédemment. On va montrer, sans faire la preuve complète, que $m \geq m'nq\log n$ (où $m' \geq \frac{c}{\varepsilon}$) échantillons suffisent, où c est une constante assez grande indépendante de n et de ε. Le plus gros obstacle à la reconstitution de $\mathbf{t}$ est de retrouver la longueur

des runs longs (que l'on définit comme les runs de longueur supérieure à $\sqrt{n}$). Déjà, l'article [1] montre que l'on peut, avec grande probabilité, distinguer les runs longs des runs courts. Il montre aussi, de manière élémentaire mais laborieuse, que si l'on connaît la longueur des runs longs, alors l'algorithme de majorité bit par bit permet de reconstruire $\mathbf{t}$ à l'aide de m' échantillons avec proba supérieure à p'_n où (p'_n) est une suite tendant vers 1. Le tout est donc de récupérer la longueur des runs longs. L'idée est de répéter suffisamment de fois l'algorithme avec à chaque fois et de dire que la longueur du run reconstruit sera la moyenne des longueurs observées divisée par $(1-q)$. En effet, si on note N la longueur du run, alors en moyenne la longueur observée sera $(1-q)N$.

Notons L_i le run long que l'on cherche à reconstruire. Notons X_j pour $j \in [\![1, m]\!]$ les longueurs de L_i observées lors de toutes les répétitions de l'algorithme. Si on suppose que le premier bit de L_i n'est pas supprimé et que les runs L_{i-1} et L_{i+1} ne sont pas totalement supprimés dans le mot reçu r_j, alors la variable aléatoire $X_j - 1$ suit une loi binomiale de paramètres $(\ell_i - 1, 1-q)$. Dans le cas contraire, il y a fusion de runs et la longueur observée est donc bruitée. Mais puisque q est petit, la probabilité que l'un des deux runs L_{i-1} et L_{i+1} soit supprimé est très petite, donc on peut négliger le bruit.

Lemme 2.10. *Supposons qu'aucun des deux runs L_{i-1} et L_{i+1} ne soit supprimé dans les mots reçus. Alors, avec grande probabilité, la moyenne des $X_j - 1$ divisée par $(1-q)$ est dans $]\ell_i - 1 - \frac{1}{3}, \ell_i - 1 + \frac{1}{3}[$.*

Démonstration. Nous donnons une preuve dans le cas où $m \geq m' n \log n$. Le résultat plus fort peut être montré à partir d'une variante des bornes de Chernoff. Avec les bornes de Chernoff classiques, on obtient

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^{m}(X_j - 1) - (1-q)m(\ell_i - 1)\right| \geq \frac{(1-q)m}{3}\right) \\
&= \mathbb{P}(|X - \mu| \geq \delta\mu) \text{ avec } X = \sum_{j=1}^{m}(X_j - 1), \mu = \mathbb{E}[X] \text{ et } \delta = \frac{1}{3(\ell_i - 1)} \\
&\leq 2e^{-\frac{\delta^2\mu}{3}} \\
&\leq 2e^{-\frac{1}{3}\left(\frac{1}{3(\ell_i-1)}\right)^2(1-q)m' n \log n(\ell_i - 1)} \\
&\leq 2e^{-\frac{1}{54}m' \log n} \text{ pour } n \text{ assez grand .}
\end{aligned}$$

□

2.5 Une borne inférieure au nombre de représentants nécessaire

On note toujours q la proba de délétion et $p = 1 - q$. Jusqu'ici, on a donné une borne inférieure au nombre de représentants suffisant dans la plupart des cas ainsi qu'une borne inférieure au nombre de représentants suffisant (dans tous les cas). Plaçons nous dans le deuxième cas et montrons que $\Omega(nqp)$ échantillons sont nécessaires, *i.e.* si (m_n) n'est pas dans $\Omega(nqp)$, alors (m_n) échantillons ne suffisent pas.

Considérons les mots $t_0 = 1^{n/2}0^{n/2}$ et $t_1 = 1^{(n/2)+1}0^{(n/2)-1}$. Les échantillons de t_0 seront de la forme $1^\ell 0^{\ell'}$ où ℓ et ℓ' suivent chacune une loi binomiale de paramètres $n/2, p$. Distinguer t_0 et t_1 revient à distinguer $1^{n/2}$ et $1^{(n/2)+1}$ et donc à distinguer les binomiales $\mathcal{B}(n/2, p)$ de $\mathcal{B}((n/2)+1, p)$ à partir d'échantillons indépendants, c'est-à-dire distinguer m binomiales $\mathcal{B}(n/2, p)$ indépendantes et m binomiales $\mathcal{B}(n/2+1, p)$ indépendantes. Notons $d = \mathrm{d_{vt}}(\mathcal{B}(n/2, p), \mathcal{B}(n/2+1, p))$. D'après le lemme 1.3, on peut trouver X et Y des variables suivant les lois $\mathcal{B}(n/2, p)$ et $\mathcal{B}(n/2+1, p)$ telles que $\mathbb{P}(X \neq Y) = d$. On se donne m couples (X_j, Y_j) pour $j \in [\![1, m]\!]$ de v.a.i.i.d. de même loi que (X, Y). Alors

$$\mathrm{d_{vt}}\left(\mathcal{B}(n/2, p)^{\otimes m}, \mathcal{B}(n/2+1, p)^{\otimes m}\right) \leq \mathbb{P}\left((X_j)_{1\leq j\leq m} \neq (Y_j)_{1\leq j\leq m}\right) \leq m\mathbb{P}(X \neq Y) = md.$$

Ainsi, si $m = o\left(\frac{1}{d}\right)$ quand n tend vers l'infini, alors la distance $\mathrm{d_{vt}}\left(\mathcal{B}(n/2, p)^{\otimes m}, \mathcal{B}(n/2+1, p)^{\otimes m}\right)$ tend vers 0. Donc, si m n'est pas dans $\Omega\left(\frac{1}{d}\right)$, alors à extraction près, la distance tend vers 0. Donc (m_n) mots ne suffisent pas.

Déjà, on peut calculer la distance en variation totale entre les deux lois $\mathcal{B}(n,p)$ et $\mathcal{B}(n+1,p)$. D'après le lemme 1.3, elle est égale à

$$\begin{aligned}\mathrm{d_{vt}}\left(\mathcal{B}(n,p),\mathcal{B}(n+1,p)\right) &= \frac{1}{2}\sum_{k=0}^{n+1}\left|\binom{n}{k}p^k(1-p)^{n-k} - \binom{n+1}{k}p^k(1-p)^{n-k+1}\right| \\ &= \frac{1}{2}\sum_{k=0}^{n+1}\left|\binom{n}{k}p^{k+1}(1-p)^{n-k} - \binom{n}{k-1}p^k(1-p)^{n-k+1}\right|.\end{aligned}$$

Comparons les deux termes dans la valeur absolue :

$$\frac{\binom{n}{k}p^{k+1}(1-p)^{n-k}}{\binom{n}{k-1}p^k(1-p)^{n-k+1}} \geq 1 \iff \frac{(n-k+1)p}{k(1-p)} \geq 1 \iff k \leq (n+1)p$$

On peut séparer la somme en deux et on obtient

$$\begin{aligned}&\frac{1}{2}\sum_{k=0}^{\lfloor (n+1)p\rfloor}\left(\binom{n}{k}p^{k+1}(1-p)^{n-k} - \binom{n}{k-1}p^k(1-p)^{n-k+1}\right) \\ &+ \frac{1}{2}\sum_{k=\lfloor (n+1)p\rfloor+1}^{n+1}\left(\binom{n}{k-1}p^k(1-p)^{n-k+1} - \binom{n}{k}p^{k+1}(1-p)^{n-k}\right) \\ &= \frac{1}{2}\binom{n}{\lfloor (n+1)p\rfloor}p^{\lfloor (n+1)p\rfloor+1}(1-p)^{n-\lfloor (n+1)p\rfloor} + \frac{1}{2}\binom{n}{\lfloor (n+1)p\rfloor}p^{\lfloor (n+1)p\rfloor+1}(1-p)^{n-\lfloor (n+1)p\rfloor} \\ &= \binom{n}{\lfloor (n+1)p\rfloor}p^{\lfloor (n+1)p\rfloor+1}(1-p)^{n-\lfloor (n+1)p\rfloor}.\end{aligned}$$

En remplaçant n par $n/2$, on obtient la distance d. Pour simplifier les calculs, on se place dans le cas où $p = 1/2$. Remarquons que ce n'est même pas un cas particulier des cas traités auparavant, puisque jusqu'à présent, q tendait vers 0. Alors il existe une constante $c > 0$ telle que $d \sim \frac{c}{\sqrt{n}}$. Donc $\Omega(\sqrt{n})$ échantillons sont nécessaires. Avec ce raisonnement, on n'obtient malheureusement pas le $\Omega(nqp)$.

Le $\Omega(nqp)$ semble malgré tout naturel puisque "distinguer (X_j) de (Y_j) c'est un peu comme distinguer $X' = \sum X_j$ de $Y' = \sum Y_j$". Mais pour les distinguer, on voudrait que les masses de leurs lois soient disjointes, donc que $|\mathbb{E}[X'] - \mathbb{E}[Y']| \approx mp$ soit grand devant $\mathrm{Var}[X'] \approx \mathrm{Var}[Y'] \approx nmqp$, c'est-à-dire que m soit grand devant $\frac{nq}{p}$, qui ressemble à nqp.

3 Résultats d'analyse complexe

Dans la section 4, on va comparer les mots en étudiant leurs séries génératrices. Il est alors naturel d'introduire l'ensemble suivant : si $n \in \mathbb{N}$, on note

$$\mathcal{A}_n := \left\{P : z \mapsto \sum_{k=0}^{n} a_k z^k \,|\, \forall k \in [\![0,n]\!],\ a_k \in \{0,1\}\right\}.$$

$\mathcal{A}_n$ n'étant cependant pas stable par différence, on introduit également

$$\mathcal{F}_n := \left\{P : z \mapsto \sum_{k=0}^{n} a_k z^k \,|\, \forall k \in [\![0,n]\!],\ a_k \in \{-1,0,1\}\right\}.$$

Si $L > 0$, on pose aussi

$$\gamma_L := \{e^{i\theta} \,|\, |\theta| \leq \pi/L\}.$$

On établit, dans le reste de cette section, des résultats d'analyse complexe permettant de maîtriser ces polynômes. Plus précisément, notre but sera d'établir les lemmes 3.1, 3.3 et 3.4 ainsi que le corollaire 3.9. Tout d'abord,

le lemme 3.1 sera utile dans le théorème 4.2 pour justifier que des mots différents sont distinguables. En effet, il montre que tout polynôme non nul de $\mathcal{F}_n$ admet un point de γ_L en lequel il n'est "pas trop petit". Le lemme technique 3.4 servira ensuite à contrôler les séries génératrices. Pour le théorème 4.3, on cherchera des polynômes qui sont "petits" sur certaines ellipses. Le lemme 3.3 servira pour cela à exhiber un polynôme "petit" sur $[0, 1]$, et le corollaire 3.9 étendra cette majoration à une ellipse.

Lemme 3.1. *Soit $n \in \mathbb{N}^*$, $L > 0$ et $A \in \mathcal{F}_{n-1} \setminus \{0\}$. Alors* $\max_{z\in\gamma_L} |A(z)| \geq n^{-L}$.

Démonstration. Quitte à diviser A par un monôme et changer le signe, ce qui ne change pas $|A(z)|$ sur le cercle unité, on peut supposer $a_0 = 1$. On considère alors la fonction entière :

$$F(z) = \prod_{j=0}^{\lfloor L \rfloor} A(ze^{2ij\pi/L})$$

qui vérifie $F(0) = 1$. Par le principe du maximum, $|F|$ atteint son $\max$ sur le disque unité en un point z du cercle unité, qui vérifie donc $|F(z)| \geq 1$. Or il existe $k \in [\![0, \lfloor L \rfloor]\!]$ tel que $ze^{2ik\pi/L} \in \gamma_L$; on peut alors majorer $|A(ze^{2ik\pi/L})| \leq \max_{w\in\gamma_L} |A(w)|$ et $|A(ze^{2ij\pi/L})| \leq n$ pour $j \neq k$. On a donc $n^{\lfloor L \rfloor} \max_{\gamma_L} |A| \geq 1$, d'où le résultat. □

On voudra également s'intéresser à une borne inférieure d'échantillons pour reconstituer le mot transmis. Le lemme 3.3, tiré du théorème 3.3 de [3], nous permettra de trouver des mots différents mais difficile à distinguer. On commence par prouver une formule utile, dite des différences divisées :

Lemme 3.2. *Soit $[a, b]$ un intervalle. f une fonction $n+1$ fois dérivable sur $[a, b]$. Soient $a \leq x_0 < \ldots < x_n \leq b$. Soit P le polynôme d'interpolation de Lagrange interpolant f en les x_i. Alors, pour tout $x \in [a, b]$, il existe $\xi \in [a, b]$ tel que*

$$f(x) - P(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i).$$

Ce lemme permet de contrôler l'écart entre f et son polynôme d'interpolation à l'aide de $f^{(n+1)}$.

Démonstration. Si $y_0, \ldots, y_k$ sont des éléments distincts de $[a, b]$, on note $P_{y_1,\ldots,y_k}$ le polynôme de degré k interpolant f en les y_i. On peut l'écrire

$$P_{y_0,\ldots,y_k} = \sum_{i=0}^{k} f_{y_0,\ldots,y_i} \prod_{j=0}^{i-1} (X - y_j).$$

On remarque que les coefficients $f_{y_0,\ldots,y_i}$ ne dépendent pas des y_j suivants ni de l'ordre des y_j.
Par ailleurs, par un calcul simple, on a, si $k \geq 1$,

$$P_{y_0,\ldots,y_k} = \frac{(X - y_0)P_{y_1,\ldots,y_k} - (X - y_k)P_{y_0,\ldots,y_{k-1}}}{y_k - y_0}.$$

Ainsi en regardant les coefficients dominants,

$$f_{y_0,\ldots,y_k} = \frac{f_{y_1,\ldots,y_k} - f_{y_0,\ldots,y_{k-1}}}{y_k - y_0}.$$

Par ailleurs, on a $f_{y_0} = P_{y_0} = f(y_0)$.

Remarquons que l'égalité voulue est évidemment vraie si x vaut un des x_i. Supposons maintenant x distinct des x_i. Soit $Q = P_{x_0,\ldots,x_n,x}$. Alors Q s'écrit

$$Q = P + f_{x_0,\ldots,x_n,x} \prod_{i=0}^{n} (X - x_i).$$

Soit $g = f - Q$. Alors g s'annule en $n+2$ points distincts de l'intervalle $[a, b]$. Donc, en appliquant $n+1$ fois le théorème de Rolle, il existe $\xi \in [a, b]$ tel que $g^{(n+1)}(\xi) = 0$. Ainsi $f^{(n+1)}(\xi) = Q^{(n+1)}(\xi) = (n+1)! f_{x_0,\dots,x_n,x}$. Or $Q(x) = P(x) + f_{x_0,\dots,x_n,x} \prod_{i=0}^{n}(x - x_i)$ et $Q(x) = f(x)$. On en déduit

$$f^{(n+1)}(\xi) = (n+1)! \frac{f(x) - P(x)}{\prod_{i=0}^{n}(x - x_i)}.$$

□

Lemme 3.3. *Il existe une constante $\tilde{c} > 0$ telle que pour tout entier $n \geq 2$, on a*

$$\inf_{P \in \mathcal{F}_n \setminus \{0\}} \max_{z \in [0,1]} |P(z)| \leq \exp(-\tilde{c}\sqrt{n}). \tag{3}$$

Remarque : un élément de $\mathcal{F}_1 \setminus \{0\}$ vaut toujours ± 1 en 0 ou en 1, d'où la condition $n \geq 2$.

Démonstration. On va en fait trouver une constante $\tilde{c}$ qui convient pour $n > n_0$ où n_0 est un certain entier suffisamment grand. Puis, comme pour tout entier n on a $z \mapsto z - z^2 \in \mathcal{F}_n \setminus \{0\}$,

$$\inf_{P \in \mathcal{F}_n \setminus \{0\}} \max_{z \in [0,1]} |P(z)| \ \leq \ \max_{z \in [0,1]} |z - z^2| \ \leq \ 1/2$$

et il existe une constante $\tilde{c}_1$ qui vérifie (3) pour tout $n \in [\![2, n_0]\!]$. En effet, il suffit de prendre $\tilde{c}_1$ telle que $e^{-\tilde{c}_1\sqrt{n_0}} \geq 1/2$. Enfin, $\tilde{c}_2 = \min\{\tilde{c}, \tilde{c}_1\}$ vérifie (3) pour tout entier $n \geq 2$.

On va maintenant prouver le résultat pour n suffisamment grand. Si $n \geq 2$, on pose $k = \lceil \frac{\sqrt{n}}{2} \rceil$ et $1 - \frac{k}{2n} =: y_0 < y_1 < \cdots < y_k := 1$ une subdivision régulière (de pas $\frac{1}{2n}$) de l'intervalle $[1 - \frac{k}{2n}, 1]$. On va d'abord montrer qu'il existe un entier n_1 tel que pour tout $n > n_1$, il existe $F \in \mathcal{F}_{n-1} \setminus \{0\}$ vérifiant

$$\forall j \in [\![0, k]\!], \ |F(y_j)| \leq 2^{1-\sqrt{n}}. \tag{4}$$

Pour cela on pose $m = \lceil 2^{\sqrt{n}-1} \rceil$ et

$$Q = \{(x_0, \dots, x_k) \,|\, \forall j \in [\![0, k]\!], \ x_j \in [0, n]\}.$$

On découpe ensuite le cube Q en $(nm)^{k+1}$ sous-cubes

$$Q_{i_0,\dots,i_k} = \{(x_0, \dots, x_k) \,|\, \forall j \in [\![0, k]\!], \ x_j \in [i_j/m, (i_j + 1)/m]\}$$

où $(i_0, \dots, i_k)$ parcourt $[\![0, nm-1]\!]^{k+1}$. Si $P \in \mathcal{A}_{n-1}$, on note $M(P) = (P(y_0), \dots, P(y_k))$. On a alors $M(P) \in Q$. Or $|\mathcal{A}_{n-1}| = 2^n$. Montrons qu'à partir d'un certain rang, $(nm)^{k+1} < 2^n$. Quand $n \to +\infty$, on a

$$(k+1)\log(nm) \sim \frac{\sqrt{n}}{2}\log(n2^{\sqrt{n}}) \sim \frac{\sqrt{n}}{2}\log(2^{\sqrt{n}}) \sim \frac{\log 2}{2}n$$

où $\frac{\log 2}{2} < \log 2$. Ainsi il existe un entier n_1 tel que, pour $n > n_1$, on a $(2nm)^{k+1} < 2^n$. Par principe des tiroirs, on peut alors trouver deux éléments $F_1, F_2 \in \mathcal{A}_{n-1}$ distincts et tels que $M(F_1), M(F_2)$ appartiennent au même sous-cube. Enfin, $F := F_1 - F_2 \in \mathcal{F}_{n-1} \setminus \{0\}$ vérifie (4).

Il faut maintenant maintenant maîtriser F sur $[y_0, 1]$ à l'aide des y_j. Or si $y \in [y_0, 1]$ n'est pas un y_j, par la formule des différences divisées, il existe $\xi \in [y_0, 1]$ tel que :

$$\frac{F(y)}{\prod_{j=0}^{k}(y - y_j)} + \sum_{i=0}^{k} \frac{F(y_i)}{(y_i - y)\prod_{j \neq i}(y_i - y_j)} = \frac{F^{(k+1)}(\xi)}{(k+1)!}.$$

De plus, comme $F \in \mathcal{F}_{n-1}$, on a $\max_{z\in[0,1]} |F^{(k+1)}(z)| \leq n^{k+2}$. Par ailleurs, notons i_0 l'indice tel que $y \in [y_{i_0}, y_{i_0+1}]$ Alors

$$\prod_{j=0}^{k} |y - y_j| \leq \prod_{j\leq i_0} |y_{i_0+1} - y_j| \prod_{j>i_0} |y_{i_0} - y_j| \leq \frac{(i_0+1)!}{(2n)^{i_0+1}} \frac{(k-i_0)!}{(2n)^{k-i_0}} \leq \frac{(k+1)!}{(2n)^{k+1}},$$

et par le même raisonnement, pour tout indice i :

$$\prod_{j\neq i} |y - y_j| \leq \frac{(k)!}{(2n)^k}.$$

Il en découle :

$$\begin{aligned}
|F(y)| &\leq \frac{|F^{(k+1)}(\xi)|}{(k+1)!} \prod_{j=0}^{k} |y - y_j| + \sum_{i=0}^{k} \frac{|F(y_i)| \prod_{j=0}^{k} |y - y_j|}{|y_i - y| \prod_{j\neq i} |y_i - y_j|} \\
&\leq \frac{n^{k+2}}{(k+1)!} \frac{(k+1)!}{(2n)^{k+1}} + 2^{1-\sqrt{n}} \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} \\
&= 2^{-(k+1)} n + 2^{1-\sqrt{n}} 2^k \\
&\leq 2^{-\sqrt{n}/2} n + 2^{1-\sqrt{n}} 2^{\sqrt{n}/2} \\
&\leq \exp(-\tilde{c}_3 \sqrt{n})
\end{aligned}$$

pour une certaine constante $\tilde{c}_3$. On a donc montré :

$$\max_{z\in[y_0,1]} |F(z)| \leq \exp(-\tilde{c}_3 \sqrt{n}).$$

On va ensuite maîtriser les valeurs sur $[0, y_0]$. Pour cela, on pose $G(x) = x^n F(x) \in \mathcal{F}_{2n} \setminus \{0\}$. Si $x \geq y_0$, $|G(x)| \leq |F(x)| \leq \exp(-\tilde{c}_3\sqrt{n})$. Supposons $x \leq y_0$. On a quand $n \to +\infty : 4\sqrt{n}(1 - y_0) \longrightarrow 1$. Soit alors n_2 tel que pour $n > n_2$, $y_0 \leq 1 - \frac{1}{5\sqrt{n}}$. On a alors :

$$|G(x)| \leq (1 - \frac{1}{5\sqrt{n}})^n n \leq \exp(-\tilde{c}_4 \sqrt{n})$$

pour n assez grand et une certaine constante $\tilde{c}_4 > 0$. Ainsi, on a trouvé un entier n_0, une constante $\tilde{c}_5 > 0$, et pour tout $n > n_0$ une fonction $G \in \mathcal{F}_{2n} \setminus \{0\}$ telle que

$$\max_{z\in[0,1]} |G(z)| \leq \exp(-\tilde{c}_5 \sqrt{n}).$$

Soit enfin $\tilde{c}_6$ tel que pour tout $n \geq 2$, $\sqrt{n} \leq \tilde{c}_6 \sqrt{n+1}$ et $\tilde{c} := \tilde{c}_5 \tilde{c}_6 / \sqrt{2}$. Alors pour $n > 2n_0$:

$$\begin{aligned}
\inf_{P\in\mathcal{F}_n\setminus\{0\}} \max_{z\in[0,1]} |P(z)| &\leq \inf_{P\in\mathcal{F}_{2\lfloor n/2\rfloor}\setminus\{0\}} \max_{z\in[0,1]} |P(z)| \\
&\leq \exp(-\tilde{c}_5 \sqrt{\lfloor n/2 \rfloor}) \leq \exp(-\tilde{c}_5 \sqrt{(n-1)/2}) \leq \exp(-\tilde{c}\sqrt{n})
\end{aligned}$$

ce qui conclut la preuve.

□

On utilisera aussi les résultats suivants :

Lemme 3.4. *Soit $q \in [0,1]$ et $p = 1-q$. Il existe alors une constante $C_1 > 0$ telle que, si $z \in \gamma_L$ et $w = (z-q)/p$, on a : $|w| \leq \exp(C_1/L^2)$.*

Démonstration. Écrivons $z = \cos\theta + i\sin\theta$. Grâce aux inégalités $\cos\theta \geq 1 - \frac{\theta^2}{2}$ et $e^x \geq 1 + x$, on obtient :

$$\begin{aligned} |w|^2 &= \frac{1 + q^2 - 2q\cos\theta}{p^2} = \frac{(1-q)^2 + 2q(1-\cos\theta)}{p^2} \\ &\leq 1 + \frac{q}{p^2}\theta^2 \leq \exp(q\theta^2/p^2) \end{aligned}$$

□

Lemme 3.5. ***Lemme des trois droites d'Hadamard***
Notons $S = \{z \in \mathbb{C} \,|\, \Re z \in]0,1[\}$. Soit f une fonction continue et bornée sur $\overline{S}$ et holomorphe sur S. On pose également, pour $x \in [0,1]$: $M_x = \sup_{\Re z = x} |f(z)|$. On a alors :

$$\forall x \in [0,1],\ M_x \leq M_0^{1-x} M_1^x.$$

Démonstration. On suppose d'abord $M_0 \neq 0$ et $M_1 \neq 0$. On peut alors poser, pour $\varepsilon > 0$, :

$$g_\varepsilon(z) = \frac{f(z)}{M_0^{1-z} M_1^z} e^{\varepsilon z^2}.$$

qui est holomorphe sur S et continue sur $\overline{S}$. Notons $K_R = \{z \in \overline{S} \,|\, |\Im z| \leq R\}$. Par le principe du maximum, on a :

$$\forall R > 0,\ \max_{z \in K} |g_\varepsilon(z)| = \max_{z \in \partial K} |g_\varepsilon(z)|.$$

Or si $z = x + iy$ avec $|x| \leq 1$, on a :

$$|g_\varepsilon(z)| \leq \frac{|f(z)|}{M_0^{1-x} M_1^x} e^{\varepsilon} e^{-\varepsilon y^2}.$$

Si $\Re z \in \{0,1\}$, on a donc :

$$|g_\varepsilon(z)| \leq e^{\varepsilon}$$

et si $|\Im z| = R$, alors :

$$|g_\varepsilon(z)| \leq e^{\varepsilon} \frac{||f||_\infty}{\min(M_0, M_1)} e^{-\varepsilon R^2}.$$

En prenant R assez grand, on en déduit donc que g_ε est bornée par e^ε sur $\overline{S}$. Ainsi :

$$\forall \varepsilon > 0, \forall z = x + iy \in \overline{S},\ |f(z)| \leq M_0^{1-x} M_1^x e^{\varepsilon} e^{\varepsilon(x^2 - y^2)}.$$

Le résultat en découle en faisant tendre ε vers 0.
Dans le cas où $M_0 = 0$ ou $M_1 = 0$, on applique le raisonnement précédent à $f + \delta$ puis on fait tendre δ vers 0. □

Corollaire 3.6. *Soient $x_1 < x_2 \in \mathbb{R}$ et notons $S = \{z \in \mathbb{C} \,|\, \Re z \in]x_1, x_2[\}$. Soit f une fonction continue et bornée sur $\overline{S}$ et holomorphe sur S. On pose également, pour $x \in [x_1, x_2]$: $M_x = \sup_{\Re z = x} |f(z)|$. On a alors :*

$$\forall x \in [x_1, x_2],\ M_x^{x_2 - x_1} \leq M_{x_1}^{x_2 - x} M_{x_2}^{x - x_1}.$$

Démonstration. Il suffit de se ramener au lemme précédent à l'aide d'une transformation affine. □

Corollaire 3.7. ***Théorème des trois cercles d'Hadamard***
Soient $0 \leq r_1 < r_2 \in \mathbb{R}$ et notons $A = \{z \in \mathbb{C} \,|\, |z| \in]r_1, r_2[\}$. Soit f une fonction continue sur $\overline{A}$ et holomorphe sur A. On pose également, pour $r \in [r_1, r_2]$: $M_r = \sup_{|z| = r} |f(z)|$. On a alors :

$$\forall r \in [r_1, r_2],\ M_r^{\log(r_2/r_1)} \leq M_{r_1}^{\log(r_2/r)} M_{r_2}^{\log(r/r_1)}.$$

Démonstration. On pose $g(z) = f(e^z)$ définie sur $\{z \in \mathbb{C} \,|\, \Re z \in [x_1, x_2]\}$ où $x_i = \log r_i$, et on applique le corollaire précédent. □

Ces derniers résultats s'intéressent au cas des ellipses, qui serviront pour le théorème 4.3. En effet, on voudra majorer un polynôme sur un arc de cercle en connaissant une majoration sur un segment. Un segment n'étant pas un cercle, on ne peut pas appliquer tel quel le théorème des trois cercles d'Hadamard : d'où l'intérêt de se ramener à des ellipses. Si $a \in]0, 1/8]$, on note

$$E_a = \{z \in \mathbb{C} \,|\, |z - (1-8a)| + |z-1| < 34a\}$$

l'ellipse ouverte de foyers $(1, 1-8a)$ et de grand axe $[(1-8a) - 17a, 1 + 17a]$, et

$$\widetilde{E_a} = \{z \in \mathbb{C} \,|\, |z - (1-8a)| + |z-1| < 20a\}$$

l'ellipse ouverte de foyers $(1, 1-8a)$ et de grand axe $[(1-8a) - 10a, 1 + 10a]$. On remarque que $\widetilde{E_a} \subset E_a$.

Corollaire 3.8. *Soient $a \in]0, 1/8]$ et f une fonction continue sur $\overline{E_a}$ et holomorphe sur E_a. Alors :*

$$\max_{z \in \partial \widetilde{E_a}} |f(z)| \leq \left(\max_{z \in \partial E_a} |f(z)| \right)^{1/2} \left(\max_{z \in [1-8a, 1]} |f(z)| \right)^{1/2}.$$

Démonstration. On applique le théorème 3.7 à $g(z) := f\left((1-4a) + 4a(\frac{z+z^{-1}}{2})\right)$, où $r_1 = 1$, $r = 2$ et $r_2 = 4$. Cette transformation envoie les cercles centrés en 0 de rayons r_1, r et r_2 respectivement sur $[1-8a, 1]$, $\partial\widetilde{E_a}$ et ∂E_a. En effet, par translation puis homothétie, cela revient à montrer que $z \mapsto z + \frac{1}{z}$ envoie les cercles de rayons 1, 2 et 4 de centre 0 sur le segment $[-2, 2]$, et les ellipses de centre 0, de foyers $(-2, 2)$ de grands axes 5 et $\frac{17}{2}$. C'est le cas puisque si $z = r\cos\theta + ir\sin\theta$, alors

$$z + \frac{1}{z} = \left(r + \frac{1}{r}\right)\cos\theta + i\left(r - \frac{1}{r}\right)\sin\theta.$$

□

Corollaire 3.9. *Soient $a \in]0, 1/8]$ et $P \in \mathcal{F}_n$. Alors :*

$$\max_{z \in \mathrm{adh}(\widetilde{E_a})} |P(z)| \leq \Big((n+1)\exp(13na)\Big)^{1/2} \left(\max_{z \in [1-8a, 1]} |P(z)| \right)^{1/2}.$$

Démonstration. Si $z \in \overline{E_a}$, on a

$$|2z - 2 + 8a| \leq |z - 1 + 8a| + |z - 1| \leq 34a,$$

d'où $|z| - (1-4a) \leq 17a$ et donc $|z| \leq 1 + 13a$. Ainsi,

$$|P(z)| \leq \sum_{k=0}^{n} |z|^k \leq (n+1)(1+13a)^n \leq (n+1)\exp(13na).$$

Le résultat suit par principe du maximum et application du corollaire précédent. □

4 Étude du problème lorsque la probabilité de délétion est constante

4.1 Une solution à coût exponentiel

Le lemme qui suit sera fondamental pour distinguer les mots à partir de leurs séries génératrices. Il montre que l'espérance de la série génératrice d'un mot reçu s'exprime simplement en fonction de la série génératrice du mot transmis et de la probabilité q de délétion.

Lemme 4.1. *Soit $w \in \mathbb{C}$ et $\mathbf{x} = (\mathrm{x}_0, \dots, \mathrm{x}_{n-1}) \in \mathbb{R}^n$ un mot réel. On note alors $\widetilde{\mathbf{x}}$ le vecteur aléatoire de $\mathbb{R}^n$ obtenu en appliquant l'algorithme de délétion à $\mathbf{x}$ puis en complétant avec des 0. Si q est la proba de délétion et $p = 1 - q$, on a alors :*

$$\mathbb{E}\left[\sum_{j=0}^{n-1} \widetilde{\mathrm{x}}_j w^j\right] = p \sum_{k=0}^{n-1} \mathrm{x}_k (pw + q)^k.$$

Démonstration. Soit $j \in [\![0, n-1]\!]$, on cherche $\mathbb{E}[\widetilde{\mathrm{x}}_j]$. Or $\widetilde{\mathrm{x}}_j$ vaut soit 0, soit l'un des x_k pour $k \geq j$. Le cas $\widetilde{\mathrm{x}}_j = \mathrm{x}_k$ arrive lorsqu'il y a eu exactement $k - j$ délétions parmi $\mathrm{x}_0, \dots, \mathrm{x}_{k-1}$ et que x_k n'est pas supprimé. On a donc :

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=0}^{n-1} \widetilde{\mathrm{x}}_j w^j\right] &= \sum_{j=0}^{n-1} \sum_{k=j}^{n-1} \mathrm{x}_k \mathbb{P}(\widetilde{\mathrm{x}}_j = \mathrm{x}_k) w^j \\
&= \sum_{j=0}^{n-1} \sum_{k=j}^{n-1} \mathrm{x}_k \binom{k}{k-j} p^j q^{k-j} p w^j \\
&= p \sum_{k=0}^{n-1} \mathrm{x}_k \sum_{j=0}^{k} \binom{k}{j} p^j q^{k-j} w^j \\
&= p \sum_{k=0}^{n-1} \mathrm{x}_k (pw + q)^k.
\end{aligned}$$

□

On montre maintenant que l'on peut reconstruire le mot transmis à l'aide de $\exp(Cn^{1/3})$ mots reçus. L'algorithme de statistiques bit par bit utilisé est décrit dans la preuve.

Théorème 4.2. *Soit $\delta > 0$ et $q \in [0, 1[$. Alors pour toute constante C assez grande, tout mot transmis $\mathbf{t} \in \{0,1\}^n$ peut être reconstitué correctement avec probabilité au moins $1 - \delta$ à partir de $m = \exp(Cn^{1/3} \log n)$ échantillons, en connaissant au préalable n.*

Démonstration. Soit $\mathbf{t} \in \{0,1\}^n$ un mot transmis, et notons $\widetilde{\mathbf{t}}^1, \dots, \widetilde{\mathbf{t}}^m$ les échantillons (ces derniers sont ici fixés et non des variables aléatoires, contrairement à $\widetilde{\mathbf{x}}$). L'idée est la suivante : si m est grand, alors les fonctions génératrices des échantillons vont en moyenne être proches de leur espérance. Étant donnés deux mots $\mathbf{x}$ et $\mathbf{y}$ de longueur n, on cherche à les distinguer pour déterminer le candidat le plus probable.

On pose alors $\mathbf{a} = \mathbf{x} - \mathbf{y}$ et $A(z) = \sum_{k=0}^{n-1} \mathrm{a}_k z^k$. Pour $L = \lfloor n^{1/3} \rfloor$ (choix permettant d'optimiser peu ou prou les inégalités qui suivent), on sait par le lemme 3.1 qu'il existe $z \in \gamma_L$ vérifiant $|A(z)| \geq n^{-L}$. Posons $w = (z - q)/p$. On a alors, par le lemme 4.1 appliqué aux mots $\mathbf{x}$ et $\mathbf{y}$:

$$\mathbb{E}\left[\sum_{k=0}^{n-1} (\widetilde{\mathrm{x}}_k - \widetilde{\mathrm{y}}_k) w^k\right] = pA(z).$$

Le lemme 3.4 nous donne que $|w|^k \leq \exp(kC_1/L^2) \leq \exp(nC_1/L^2)$. On a donc, pour une certaine constante $C_2 > 0$:

$$\mathbb{E}\left[\sum_{k=0}^{n-1} |\widetilde{\mathrm{x}}_k - \widetilde{\mathrm{y}}_k|\right] \geq \exp(-C_2 n^{1/3} \log n).$$

On en déduit que pour un certain $k(\mathbf{x}, \mathbf{y}) \in [\![0, n-1]\!]$, on a :

$$|\mathbb{E}[\widetilde{\mathrm{x}}_{k(\mathbf{x},\mathbf{y})} - \widetilde{\mathrm{y}}_{k(\mathbf{x},\mathbf{y})}]| \geq \frac{1}{n} \exp(-C_2 n^{1/3} \log n).$$

Étant donnés les échantillons, on va maintenant définir une relation d'ordre sur $\{0,1\}^n$. On dit que $\mathbf{x}$ est un meilleur candidat que $\mathbf{y}$ et on note $\mathbf{x} > \mathbf{y}$ si :

$$\left|\frac{1}{m}\sum_{j=1}^{m}\widetilde{\mathrm{t}}^j_{k(\mathbf{x},\mathbf{y})} - \mathbb{E}_{\mathbf{y}}[\widetilde{\mathrm{y}}_{k(\mathbf{x},\mathbf{y})}]\right| > \left|\frac{1}{m}\sum_{j=1}^{m}\widetilde{\mathrm{t}}^j_{k(\mathbf{x},\mathbf{y})} - \mathbb{E}_{\mathbf{x}}[\widetilde{\mathrm{x}}_{k(\mathbf{x},\mathbf{y})}]\right|.$$

Le plus grand élément, s'il existe, est unique. C'est alors notre meilleur pari : si l'on note $\widehat{\mathbf{x}}$ le candidat choisi *in fine*, on prend $\widehat{\mathbf{x}} = \mathbf{x}$ lorsqu'il existe un plus grand élément $\mathbf{x}$, et $\widehat{\mathbf{x}}$ arbitraire sinon. On a alors :

$$\mathbb{P}(\widehat{\mathbf{x}} \neq \mathbf{t}) \leq \mathbb{P}\big(\exists \mathbf{y} \in \{0,1\}^n \setminus \{\mathbf{t}\}\,,\ \mathbf{t} \not> \mathbf{y}\big) \leq \sum_{\mathbf{y}\neq\mathbf{t}} \mathbb{P}(\mathbf{t} \not> \mathbf{y}).$$

Or, si $\mathbf{y} \neq \mathbf{t}$:

$$\begin{aligned}\mathbb{P}(\mathbf{t} \not> \mathbf{y}) &= \mathbb{P}\left(\left|\frac{1}{m}\sum_{j=1}^{m}\widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} - \mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right| \geq \left|\frac{1}{m}\sum_{j=1}^{m}\widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} - \mathbb{E}[\widetilde{\mathrm{y}}_{k(\mathbf{t},\mathbf{y})}]\right|\right)\\ &\leq \mathbb{P}\left(\left|\sum_{j=1}^{m}\widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} - m\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right| \geq \frac{m}{2n}\exp(-C_2 n^{1/3}\log n)\right)\end{aligned}$$

car $\left|\frac{1}{m}\sum_{j=1}^{m}\widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} - \mathbb{E}[\widetilde{\mathrm{y}}_{k(\mathbf{t},\mathbf{y})}]\right| \geq \left|\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left[\widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} - \widetilde{\mathrm{y}}_{k(\mathbf{t},\mathbf{y})}\right]\right| - \left|\frac{1}{m}\sum_{j=1}^{m}\widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} - \mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right|$.

On pose alors $\eta > 0$ tel que $\frac{1}{2n}\exp(-C_2 n^{1/3}\log n) = \eta\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]$ (le cas où l'espérance est nulle se traite à part) et on distingue deux cas pour appliquer les bornes de Chernoff :

- Si $\eta \leq 1$. Alors on a :

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{j=1}^{m}\widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} - m\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right| \geq \frac{m}{2n}\exp(-C_2 n^{1/3}\log n)\right)\\ &\leq 2\exp\left(\frac{-\eta^2}{3}m\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right)\\ &= 2\exp\left(\frac{-m\exp(-2C_2 n^{1/3}\log n)}{12n^2\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]}\right)\\ &\leq 2\exp\left(\frac{-m\exp(-2C_2 n^{1/3}\log n)}{12n^2}\right).\end{aligned}$$

Puis : $\mathbb{P}\big(\widehat{\mathbf{x}} \neq \mathbf{t}\big) \leq 2^{n+1}\exp\left(\frac{-m}{12n^2}\exp(-2C_2 n^{1/3}\log n)\right)$.

- Si $\eta > 1$. Comme $\sum_{j=1}^{m} \widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} > 0$, on a en fait :

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^{m} \widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} - m\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right| \geq \frac{m}{2n}\exp(-C_2 n^{1/3}\log n)\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{m} \widetilde{\mathrm{t}}^j_{k(\mathbf{t},\mathbf{y})} \geq m\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}] + \frac{m}{2n}\exp(-C_2 n^{1/3}\log n)\right) \\
&\leq \exp\left(\frac{-\eta^2}{2+\eta} m\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right) \\
&= \exp\left(\frac{-\eta}{1+2/\eta} m\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right) \\
&\leq \exp\left(\frac{-\eta}{3} m\mathbb{E}[\widetilde{\mathrm{t}}_{k(\mathbf{t},\mathbf{y})}]\right) \\
&= \exp\left(\frac{-m}{6n}\exp(-C_2 n^{1/3}\log n)\right).
\end{aligned}$$

  Puis : $\mathbb{P}(\widehat{\mathbf{x}} \neq \mathbf{t}) \leq 2^n \exp\left(\frac{-m}{6n}\exp(-C_2 n^{1/3}\log n)\right)$.

Dans tous les cas, on a $\mathbb{P}(\widehat{\mathbf{x}} \neq \mathbf{t}) \underset{n\to+\infty}{\longrightarrow} 0$ en prenant $C_3 > 2C_2$ et $m = \exp(C_3 n^{1/3}\log n)$.

□

4.2 Un résultat d'optimalité

L'article [5] montre en fait une version plus forte de notre théorème 4.2 : il suffit de $m = \exp(Cn^{1/3})$ échantillons pour reconstituer le mot transmis avec grande probabilité. La preuve est exactement la même, en remplaçant le lemme 3.1 par un raffinement prouvé dans [4]. Le théorème suivant assure alors que ce résultat est optimal parmi les algorithmes de statistique bit par bit, dans le sens où l'on ne regarde que les lois marginales des $\widetilde{\mathrm{t}}^j_k$ et non leurs lois conjointes.

Théorème 4.3. *Fixons une probabilité de délétion $q \in [0,1]$. Pour tout n assez grand, il existe deux mots distincts $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ tels que, pour tout k, la distance en variation totale entre les lois de m copies indépendantes de $\widetilde{\mathrm{x}}_k$ et de m copies indépendantes de $\widetilde{\mathrm{y}}_k$ soit au plus $me^{-cn^{1/3}}$, où c ne dépend que de q.*

Démonstration. On note $p := 1 - q$. On pose $L = \lfloor n^{1/3} \rfloor$. D'après le lemme 3.3, il existe alors une constante $c_3 > 0$ (indépendante de n) et $Q \in \mathcal{F}_{L^2} \setminus \{0\}$ tels que

$$\max_{z\in[0,1]} |Q(z)| \leq \exp(-c_3 L).$$

On écrit alors $Q = \varphi - \psi$ où $\varphi, \psi \in \mathcal{A}_{L^2}$. On pose ensuite $\ell := n - L^2$ et on définit $\mathbf{x} \in \{0,1\}^n$ (resp. $\mathbf{y}$) le mot dont les ℓ premiers bits sont des zéros, suivis des coefficients de φ (resp. ψ). Ces deux mots sont distincts car $Q \neq 0$, et on va montrer qu'il est difficile de les distinguer à partir des échantillons. Pour cela on pose $b_k := \mathbb{E}[\widetilde{\mathrm{x}}_k - \widetilde{\mathrm{y}}_k]$ et $B(w) := \sum_{k=0}^{n-1} b_k w^k$. Par la formule de Cauchy, on a

$$b_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ik\theta} B(e^{i\theta}) d\theta$$

et on cherche donc à maîtriser B sur le cercle unité. Or d'après le lemme 4.1 : $B(w) = pA(pw+q)$, où on a posé $A(z) = \sum_{k=0}^{n-1}(\mathrm{x}_k - \mathrm{y}_k)z^k$. Si l'on note Γ le cercle de centre q et de rayon p, on cherche alors à maîtriser A sur Γ. Observons que $A(z) = z^\ell Q(z)$, et que

$$|b_k| \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} |B(e^{i\theta})| d\theta \leq p \max_{z\in\Gamma} |A(z)|.$$

Soit $c_2 > 0$, dont on fixera la valeur plus tard. On considère l'ellipse pleine fermée $\widehat{E_L} := \mathrm{adh}(\widetilde{E}_{c_2/L})$. Le corollaire 3.9 donne alors, avec c_2 assez petite, l'existence d'une constante c_4 (indépendante de n) telle que :

$$\max_{z \in \widehat{E_L}} |Q(z)| \leq \exp(-c_4 L).$$

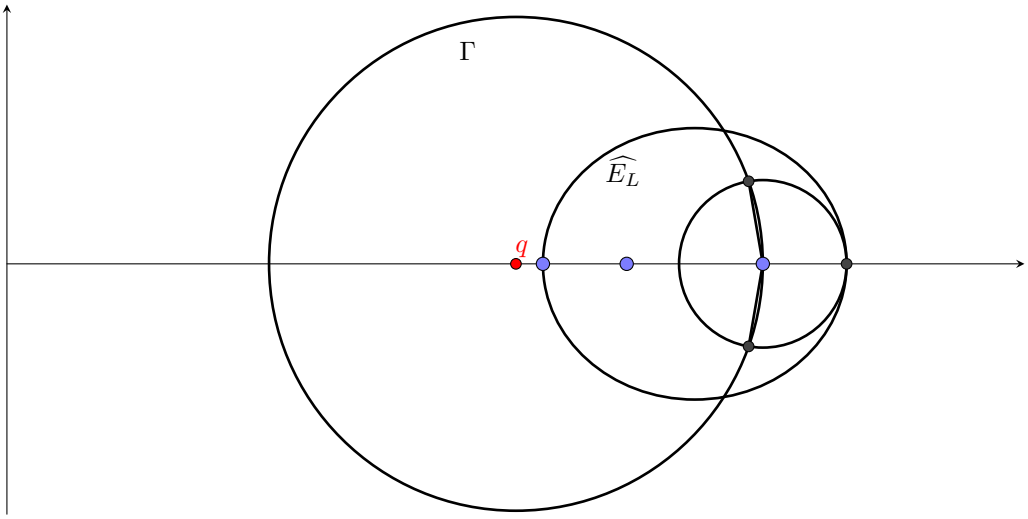

FIGURE 5 – L'ellipse $\widehat{E_L}$, le cercle Γ et le cercle de centre 1 et de rayon $\frac{6}{L}$

$\widehat{E_L}$ contient le disque de centre 1 et de rayon $\frac{6c_2}{L}$, qui intersecte Γ en un arc de longueur au moins $\frac{12c_2}{L}$ et centré en 0. Posons alors

$$\Gamma_L := \{pe^{i\theta} + q \,|\, |\theta| \leq \frac{6c_2}{pL}\}$$

qui est à la fois dans Γ et dans $\widehat{E_L}$. Ainsi, pour $z \in \Gamma_L$: $|A(z)| \leq \exp(-c_4 L)$. Reste donc à étudier le cas $z \in \Gamma \setminus \Gamma_L$. Or si $z = pe^{i\theta} + q$ avec $\theta \in [-\pi, \pi] \setminus [-\frac{6c_2}{pL}, \frac{6c_2}{pL}]$:

$$\begin{aligned}
|z|^2 &= (p\cos\theta + q)^2 + (p\sin\theta)^2 = p^2\cos^2\theta + 2pq\cos\theta + q^2 + p^2\sin^2\theta \\
&= (p+q)^2 + 2pq(\cos\theta - 1) \leq 1 + 2pq(\cos\frac{6c_2}{pL} - 1) \\
&\leq 1 + 2pq(\frac{-1}{2}\frac{(6c_2)^2}{p^2L^2} + \frac{1}{24}\frac{(6c_2)^4}{p^4L^4}) \leq 1 - \frac{c_5}{L^2}
\end{aligned}$$

pour une constante $c_5 > 0$ et si n est assez grand. Puis, il existe $c_6 > 0$ telle que, pour n assez grand :

$$|A(z)| \leq \frac{|z|^\ell}{1-|z|} \leq (1 - \frac{c_5}{L^2})^{\ell/2} c_6 L^2 \leq e^{c_6 L}.$$

Il en découle enfin qu'il existe une certaine constante $c_7 > 0$ ne dépendant que de p telle que pour n assez grand :

$$\forall k \in [\![0, n-1]\!], \quad |b_k| \leq e^{-c_7 L}.$$

Soient maintenant $\widetilde{\mathrm{x}}_k^1, \ldots, \widetilde{\mathrm{x}}_k^m$ des variables aléatoires suivant la même loi que $\widetilde{\mathrm{x}}_k$. On se donne de même $\widetilde{\mathrm{y}}_k^1, \ldots, \widetilde{\mathrm{y}}_k^m$. On veut majorer la distance de variation totale entre $\left(\widetilde{\mathrm{x}}_k^j\right)_{1 \leq j \leq m}$ et $\left(\widetilde{\mathrm{y}}_k^j\right)_{1 \leq j \leq m}$. Pour cela, on va les coupler. Soit $(\xi_j)_{1 \leq j \leq m}$ une famille de m v.a.i.i.d. uniformes sur $[0,1]$. Alors $\left(\mathbf{1}_{\xi_j \leq \mathbb{E}[\widetilde{X}_k]}\right)_j$ a la même loi que $\left(\widetilde{\mathrm{x}}_k^j\right)_j$ et

$\left(\mathbf{1}_{\xi_j \leq \mathbb{E}[\widetilde{\mathrm{y}}_k]}\right)_j$ a la même loi que $\left(\widetilde{\mathrm{y}}_k^j\right)_j$. De plus

$$\begin{aligned}
&\mathbb{P}\left(\left(\mathbf{1}_{\xi_j \leq \mathbb{E}[\widetilde{\mathrm{x}}_k]}\right)_j \neq \left(\mathbf{1}_{\xi_j \leq \mathbb{E}[\widetilde{\mathrm{y}}_k]}\right)_j\right) \\
&= \mathbb{P}\left(\exists j \in [\![1, m]\!],\ \mathbf{1}_{\xi_j \leq \mathbb{E}[\widetilde{\mathrm{x}}_k]} \neq \mathbf{1}_{\xi_j \leq \mathbb{E}[\widetilde{\mathrm{y}}_k]}\right) \\
&\leq m\mathbb{P}\left(\mathbf{1}_{\xi_j \leq \mathbb{E}[\widetilde{\mathrm{x}}_k]} \neq \mathbf{1}_{\xi_j \leq \mathbb{E}[\widetilde{\mathrm{y}}_k]}\right) \\
&= m|b_k|.
\end{aligned}$$

Ainsi, la distance de variation totale entre $\left(\widetilde{\mathrm{x}}_k^j\right)_{1 \leq j \leq m}$ et $\left(\widetilde{\mathrm{y}}_k^j\right)_{1 \leq j \leq m}$ est inférieure à $me^{-c_7 n^{1/3}}$, ce qu'on voulait. □

Références

[1] Tugkan Batu, Sampath Kannan, Sanjeev Khanna, and Andrew McGregor. Reconstructing strings from random traces. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA), pages 910–918. Society for Industrial and Applied Mathematics, 2004.

[2] Lajos Takács. Combinatorial Methods in the Theory of Stochastic Processes. John Wiley & Sons, Inc, New York, NY, first edition, 1967.

[3] Peter Borwein, Tamás Erdélyi, and Géza Kós. Littlewood-type problems on $[0, 1]$. Proc. London Math. Soc. (3), 79(1) :22–46, 1999.

[4] Borwein, P., and T. Erdélyi. "Littlewood-Type Problems on Subarcs of the Unit Circle." *Indiana University Mathematics Journal*, vol. 46, no. 4, 1997, pp. 1323–1346.

[5] Fedor Nazarov, Yuval Peres. (2017). Trace reconstruction with $\exp(O(n^{1/3}))$ samples. 1042-1046. 10.1145/3055399.3055494.

[6] Y. Peres and A. Zhai. Average-case reconstruction for the deletion channel : subpolynomially many traces suffice, 2017, FOCS.

[7] N. Holden, R. Pemantle, and Y. Peres. Subpolynomial trace reconstruction for random strings and arbitrary deletion probability. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1799–1840. PMLR, 06–09 Jul 2018.

[8] N. Holden and R. Lyons. Lower bounds for trace reconstruction, https ://arxiv.org/abs/1808.02336, 2018.

[9] Zachary Chase. New Lower Bounds for Trace Reconstruction, https ://arxiv.org/abs/1905.03031, 2019