VI. Un lemme d'échantillonnage.

Dans cette partie, on se fixe un langage S de $\mathbf{P/poly}$.

Notons $S^{=n}$ l'ensemble des mots de longueur n qui sont dans S.

Disons qu'un sous-ensemble T d'un ensemble E est *grand* si et seulement si $|T| > |E|/2$. Le lemme d'échantillonnage va énoncer qu'il existe une machine de Turing $\mathcal{M}$, probabiliste, en temps polynomial en n et avec un oracle décidant SAT, qui retourne un ensemble S' de $S^{=n}$ tel que pour tout grand sous-ensemble T de $S^{=n}$, S' intersecte T avec forte probabilité. C'est ce que nous allons démontrer dans cette partie; la question 8 en donne l'énoncé précis.

Notons que $\mathcal{M}$ doit calculer en temps polynomial en n: on lui fournira donc en entrée le nombre n lui-même, écrit *en unaire*. La machine $\mathcal{M}$ prendra aussi en argument un circuit $\mathcal{C}_n$ de taille polynomiale en n décidant de l'appartenance à $S^{=n}$. (L'existence de $\mathcal{C}_n$ est garantie par le fait que $S \in \mathbf{P/poly}$.)

1. Montrer que la famille des fonctions de hachage linéaires h de $\{0,1\}^m$ vers $\{0,1\}^{m'}$ est *2-universelle*, au sens où, pour tous $\alpha, \beta \in \{0,1\}^{m'}$, pour tous $x, y \in \{0,1\}^m$ avec $x \neq y$, $Pr[h(x) = \alpha \wedge h(y) = \beta] = 1/2^{2m'}$. (Les probabilités sont prises sur l'espace des fonctions h, tirées uniformément.)

2. Fixons $n \in \mathbb{N}$. Si $|S^{=n}| \leq n^2$, montrer comment trouver un ensemble S' qui intersecte tout grand sous-ensemble T de $S^{=n}$. Indication: prendre $S' = S^{=n}$ lui-même, et montrer que l'on peut énumérer les éléments de S' en temps $O(n^{2+k})$ pour une certaine constant $k \geq 0$, en appelant l'oracle SAT, sur une formule construite à l'aide du circuit $\mathcal{C}_n$. Pensez à l'auto-réductibilité de SAT.

3. Raffiner la question précédente, et montrer que l'on peut décider si $|S^{=n}| \leq n^2$, et si c'est le cas, énumérer tous ses éléments, en temps polynomial en faisant appel à un oracle SAT et au circuit $\mathcal{C}_n$.

4. On suppose maintenant $|S^{=n}| > n^2$, n étant fixé. Comme $|S^{=n}| > n^2$, il existe un unique entier k_e tel que $2^{k_e-1} < |S^{=n}| \leq 2^{k_e}$, avec $2\log_2 n < k_e \leq n$. En utilisant le lemme de codage de Sipser, montrer que l'on peut estimer k_e avec forte probabilité. Plus précisément, montrer que l'on peut calculer en temps polynomial (en n, pour un polynôme fixé), avec un oracle SAT et en utilisant le circuit $\mathcal{C}_n$, un entier k_0 avec $2\log_2 n < k_0 \leq n$ et tel que $k_0 - 1 \leq k_e \leq k_0 + \log_2 n + 2$ avec probabilité au moins $1 - 1/2^{3n}$.

5. Soit $U = 4n2^{k_0}$, où k_0 est l'entier trouvé à la question précédente. Observer qu'avec forte probabilité, $\frac{U}{16n} < |S^{=n}| \leq U$. Pour toute fonction de hachage linéaire h de $\{0,1\}^n$ vers $\{0,1\}^{k_0+2\log_2 n+1}$, soit $C(h)$ le nombre de ses paires en collision dans $S^{=n}$, c'est-à-dire le nombre de paires $\{x, y\}$ d'éléments distincts ($x \neq y$) de $S^{=n}$ telles que $h(x) = h(y)$. En utilisant le fait que la famille des fonctions de hachage linéaire est 2-universelle, montrer que $Pr[C(h) \geq |S^{=n}|/8] \leq 8/n$. (Penser à l'inégalité de Markov.)

6. Disons qu'un point $\alpha \in \{0,1\}^{k_0+2\log_2 n+1}$ est une *image unique* par la fonction h si et seulement s'il existe un unique $x \in S^{=n}$ tel que $h(x) = \alpha$. (Attention: ne pas oublie que x

doit être dans $S^{=n}$.) Montrer que, si $C(h) \leq |S^{=n}|/8$, alors pour tout grand sous-ensemble T de $S^{=n}$ ($|T| > |S^{=n}|/2$), au moins $|S^{=n}|/4$ des éléments x de T sont tels que $h(x)$ est une image unique par h.

7. On définit la procédure d'échantillonnage suivante. On tire au hasard $5n$ fonctions de hachage linéaires $h_1, \ldots, h_{5n}$. Pour i variant de 1 à $5n$, on calcule un point $x_i \in \{0,1\}^n \cup \{\perp\}$ comme suit.

   (a) On tire au hasard un point $\alpha \in \{0,1\}^{k_0+2\log_2 n+1}$;

   (b) on teste si α est une image unique par h_i, et si oui, on pose x_i l'unique x dans $S^{=n}$ tel que $h_i(x) = \alpha$;

   (c) sinon, on reboucle à l'étape 1.

   Si aucun des α tirés ci-dessus en $12n^3$ tours n'est une image unique, poser $x_i = \perp$ (informellement, le calcul de x_i échoue). On pose S' l'ensemble des x_i, $1 \leq i \leq 5n$, qui sont différents de $\perp$.

   Indiquer comment réaliser l'étape 2 dans la boucle ci-dessus à l'aide de l'oracle SAT et du circuit $\mathcal{C}_n$.

8. Soit T un grand sous-ensemble quelconque de $S^{=n}$. Montrer qu'avec probabilité très forte, l'ensemble S' calculé à la question précédente intersecte T. On observera que $e^{-1.5n} \ll \frac{1}{2^{2n}}$. (Le rapport des deux vaut $e^{2n\log 2 - 1.5n} = e^{-0.113\ldots n}$. On notera aussi que $\log 2 = 0.694\ldots$, et $\log(3/4) = -0.288\ldots$) Plus précisément, montrer que, pour n assez grand, avec probabilité au moins $1 - 1/2^{2n}$, les $5n$ éléments tirés au hasard sont tous différents de $\perp$, et qu'alors l'ensemble S' de ces (au plus) $5n$ éléments intersecte T avec probabilité (conditionnelle) au moins $1 - 1/2^{2n}$.

   Nous appellera ceci le *lemme d'échantillonnage*. Il est valable dès que $S^{=n}$ a des circuits $\mathcal{C}_n$ de taille polynomiale en n pour tout n, lorsque n est assez grand et $|S^{=n}| > n^2$. Alors S' est calculé en temps polynomial (randomisé) sur une machine de Turing avec oracle SAT sur l'entrée formée de n écrit en unaire et du circuit $\mathcal{C}_n$.

VII. Le théorème de Cai: $\mathsf{S}_2^p \subseteq \mathbf{ZPP^{NP}}$

Informellement, la classe d'*alternance symétrique* S_2^p est la classe des langages L que l'on peut décider en mettant en compétition deux prouveurs Y et Z, qui sur l'entrée x de taille n fournissent une "preuve" chacun, y pour Y et z pour Z, toutes les deux de taille polynomiale $p(n)$. Y cherche à montrer que $x \in L$, et Z cherche à prouver que $x \notin L$. L'idée est que L est dans S_2^p si et seulement si:

- si $x \in L$, alors il existe une preuve y de Y (de $x \in L$) qui est plus convaincante que n'importe quelle "preuve" z de Z (de $x \notin L$);
- si $x \notin L$, alors il existe une preuve z de Z (de $x \notin L$) qui est plus convaincante que n'importe quelle "preuve" y de Y (de $x \in L$).

Formellement, L est dans S_2^p si et seulement s'il existe un langage $D \in \mathbf{P}$ tel que, pour toute entrée x de taille n,

- si $x \in L$ alors il existe y de taille $p(n)$ tel que pour tout z de taille $p(n)$, $(x, y, z) \in D$;
- si $x \notin L$ alors il existe z de taille $p(n)$ tel que pour tout y de taille $p(n)$, $(x, y, z) \notin D$;

où $p(n)$ est un polynôme fixé.

Noter que ceci n'est pas une classe randomisée.

Le théorème de Cai, $\mathsf{S}_2^p \subseteq \mathbf{ZPP}^{\mathbf{NP}}$, fait l'objet de ce devoir.

1. On se donne un langage L de S_2^p, défini comme ci-dessus, et l'on fixe une entrée $x \in \{0,1\}^n$. Pour tout ensemble Y_k d'au plus k "preuves" y de Y, notons $Z(Y_k)$ l'ensemble des "preuves" z de Z qui *battent* tous les y de Y_k:

$$Z(Y_k) \quad = \quad \{z \in \{0,1\}^{p(n)} | \forall y \in Y_k \cdot (x, y, z) \notin D\}$$

   On notera que, dans la suite, k sera, avec très forte probabilité, borné par un polynôme fixé en n: on peut penser à l'expression $\forall y \in Y_j$ comme à une conjonction d'un nombre polynomial de tests d'appartenance à D.

   Pour tout $y' \in \{0,1\}^{p(n)}$, disons que y' est *mauvais* (pour $Z(Y_k)$) s'il *bat* moins de la moitié de $Z(Y_k)$, c'est-à-dire si le nombre des $z \in \{0,1\}^{p(n)}$ tels que $(x, y', z) \in D$ est strictement inférieur à $|Z(Y_k)|/2$. (Noter que les notions d'un y qui bat un ensemble de z et d'un z qui bat un ensemble de y sont proches, mais distinctes.)

   Pour tout mauvais y', considérons l'ensemble

$$T_{y'} \quad = \quad Z(Y_k \cup \{y'\})$$

   En utilisant le lemme d'échantillonnage, montrer que si $|Z(Y_k)| > p(n)^2$, on peut calculer, en temps polynomial sur une machine randomisée à oracle SAT, un ensemble Z' d'au plus $5p(n)$ éléments de $Z(Y_k)$ qui intersecte $T_{y'}$ avec forte probabilité pour tout mauvais y' (et ce pour n assez grand).

2. Ayant calculé un ensemble Z' à au plus $5p(n)$ éléments de $\{0,1\}^{p(n)}$, montrer comment calculer en temps polynomial avec oracle SAT, un élément y^* de $\{0,1\}^{p(n)}$ qui bat tous les éléments de Z'. L'algorithme peut échouer à trouver un tel élément, mais uniquement si $x \notin L$.

3. Supposons que $Z(Y_k) > p(n)^2$, et que l'on ait calculé un ensemble Z' comme spécifié à la question 1, et trouvé un y^* qui bat tous les éléments de Z' comme demandé à la question 2. Montrer que $|Z(Y_{k+1})| \leq |Z(Y_k)|/2$, où $Y_{k+1} = Y_k \cup \{y^*\}$.

4. En déduire un algorithme qui décide si $x \in L$ de façon *exacte* en temps *moyen* polynomial probabiliste avec oracle SAT. Autrement dit, $L \in \mathbf{ZPP}^{\mathbf{NP}}$.

   Indication: poser $Y_0 = \emptyset$, et itérer la construction de la suite Y_k comme suggéré plus haut.