TD12 : Algorithmes de flux

Exercice 1. *Calcul par échantillonage*

On note $f_1, \dots, f_N$ la fréquence des entiers $1, \dots N$ dans un flux de longueur $x_1, \dots, x_m$. On souhaite calculer le $k^{\text{ème}}$ moment $F_k = \sum_i f_i^k$ pour un entier k.

1. Quelle est la complexité en mémoire de l'algorithme naïf ?

Pour améliorer la complexité, on tire j aléatoirement entre 1 et m (la longueur du flux) et on calcule $r = \#\{t \geq j : x_t = x_j\}$. On renvoie $X = m(r^k - (r-1)^k)$.

2. Quelle est la complexité en mémoire de l'algorithme proposé ?
3. Montrer que $\mathbb{E}[X|x_j = i] = \sum_{r=1}^{f_i} \frac{1}{f_i} m(r^k - (r-1)^k)$.
4. En déduire que $\mathbb{E}[X] = \sum_i f_i^k$.
5. Pour quelle(s) valeur(s) de k a-t-on vu un autre algorithme en cours ? Comparer avec celui-ci.
6. Quelle stratégie peut-on mettre en œuvre pour réduire la variance de l'estimation ? Analyser aussi précisément que possible cette stratégie.
7. *(plus difficile)* En déduire qu'on peut calculer une $(1 \pm \epsilon)$-approximation de F_k avec probabilité au moins $1 - \delta$, en espace $\tilde{O}(\frac{1}{\epsilon^2} n^{1-1/k})$.