Projet de Thèse
Complexité en moyenne, entropies, et fonctions à sens unique

Directeurs de thèse :
LIPN : Stefano Guerrini (Professeur), Thomas Seiller (CR CNRS)
IRIF : Geoffroy Couteau (CR CNRS, `https://geoffroycouteau.github.io`)

Candidat :
Ulysse Léchine

Éléments de contexte scientifique local :
Cette thèse encadrée par Stefano Guerrini, Thomas Seiller, et Geoffroy Couteau porte sur un thème de la théorie de complexité "standard" (complexité "Track A") : la complexité algorithmique *en moyenne*. Les thèmes historiques et l'expertise des membres de l'équipe portent sur la théorie de la complexité implicite, et plus généralement les méthodes logiques et sémantiques en complexité (complexité "Track B"). Le sujet de thèse que nous détaillons ci-dessous consiste à adapter certaines méthodes provenant de la complexité implicite pour les utiliser dans le cadre de la complexité algorithmique en moyenne. Ce travail permettra donc d'accentuer un mouvement thématique visant à se rapprocher de la complexité "Track A" souhaité par l'équipe, initié depuis quelques années, et soutenu par certains résultats récents. L'encadrement par Geoffroy Couteau, qui apportera son expertise sur les fonctions à sens unique et leur lien avec les classes de complexité permettra d'assurer la bonne réussite du projet sur ces aspects les plus éloignés des thèmes historiques de l'équipe. Ces thèmes nouveaux ouvriront par ailleurs la possibilité d'interactions nouvelles entre le LIPN et le LAGA sur les thèmes de la cryptographie (en particulier, Sihem Mesnager et Claude Carlet qui, sans être co-auteurs de Geoffroy Couteau, ont des domaines d'intérêts proches des travaux de celui-ci).

Éléments de contexte sur le candidat :
Ulysse Léchine est un ancien étudiant de l'École Normale Supérieure de Lyon. Il a notamment suivi en master de nombreux cours sur des sujets particulièrement approprié pour travailler sur le sujet proposé, notamment les cours suivants pour lesquels il a été classé premier : Cryptographie, Topologie combinatoire, Théorie de l'information, et Complexité algorithmique. Ulysse a effectué de nombreux stages validant et renforçant ces compétences : en complexité algorithmique avec Akitoshi Kawamura en 2018 (stage de 3 mois), puis avec Shuichi Hirahara en 2020 (stage de 9 mois), ainsi que sur des aspects de dynamique symbolique (automates cellulaires et pavages auto-assemblants) avec Pablo Arrighi en 2017 (stage de 2 mois) et Damien Woods en 2020 (stage de 5 mois).

Contexte général. La théorie de la complexité a pour objet d'étude les quantités de resources (temps, espace, etc.) utilisées par un programme, et par extension quantifier les ressources nécessaires pour résoudre un problème donné. Traditionnellement, la notion de complexité considérée par défaut est la complexité "dans le pire cas", où la complexité d'un programme correspond aux ressources utilisées pour résoudre l'instance la plus difficile du problème. Cette notion n'est cependant pas toujours significative, notamment en pratique : si la plupart des instances du problème considéré nécessitent moins de ressources, par exemple un temps n^2 plutôt que n^5, alors la veritable complexité du programme *à l'usage* sera plus proche de n^2 que de n^5.

La théorie de la complexité "en moyenne" vise donc à étudier la moyenne des ressources nécessaires, pondérée par une distribution de probabilité sur les instances du problème. Au delà de l'intérêt pratique mis

en avant ci-dessus qui consisterait à trouver des problèmes pour lesquels la complexité en moyenne est plus faible qu'espéré (pour avoir des algorithmes plus rapides à l'usage), la complexité en moyenne a également des applications en cryptographie.

Complexité en moyenne et cryptographie. La sureté d'un protocole cryptographique n'est pas assurée par une complexité dans le pire cas, mais bien par une complexité en moyenne qui permet d'assurer qu' il est couteux à l'usage de casser le protocole – et non dans le cas peut-être improbable où le problème est très difficile à résoudre.

En effet, s'il est établi que l'existence de protocoles cryptographiques robustes impliquerait que PTIME $\neq$ NPTIME, cette condition nécessaire n'est pas suffisante : la séparation PTIME $\neq$ NPTIME implique seulement que le schéma d'encryption est difficile à casser dans le pire des cas mais il n'élimine pas la possibilité que ce dernier soit facile à casser dans presque tous les cas. Une condition nécessaire pour l'existence de schéma d'encryption sûrs est dont l'existence de langages dans NPTIME qui soient difficile en moyenne. Par ailleurs, les connaissances actuelles ne permettent pas d'établir que PTIME $\neq$ NPTIME impliquerait l'existence de langages dans NPTIME qui soient difficiles en moyenne.

Fonctions à sens unique. L'existence de problèmes (dans NPTIME) difficiles en moyenne ne suffit pas pour avoir un protocole cryptographique sûr. Afin de pouvoir utiliser un tel problème, il faut avoir un moyen de générer de telles instances difficile avec une information auxilliaire qui permet de résoudre ces instances facilement. L'existence de protocole sûrs nécessite donc d'avoir un moyen efficace – un algorithme probabiliste en temps polynomial – de générer de telles instances avec une information auxilliaire tels que :

1. il est facile de résoudre ces instances avec l'information auxilliaire ;
2. il est difficile en moyenne de résoudre ces instance sans l'information auxilliaire

Les fonctions à sens unique proposent une manière de produire de tels protocoles : il s'agit de fonctions "faciles" à calculer mais "difficiles" à inverser. Plus précisément, étant donné un algorithme probabiliste en temps polynomial A, celui-ci ne pourra inverser la fonction à sens unique f sur un élément de son image qu'avec une probabilité négligeable, où la probabilité est calculée sur l'ensemble des éléments dans le domaine de la fonction f et les choix probabilistes de l'algorithme A.

Objectifs de la thèse. L'objectif principale de cette thèse est d'apporter de nouveaux éléments pour éclairer la relation entre la complexité en moyenne et les fonctions à sens unique. L'apport technique principal sera d'introduire dans l'étude de ces deux questions une même notion, celle d'entropie (mesurable). Plus précisément, nous explorerons les liens entre l'entropie et :

1. l'existence de problèmes NPTIME-difficiles en moyenne. En effet, un résultat récent de Seiller en collaboration avec Pellissier établit une méthode de preuve pour obtenir des bornes inférieures de complexité *dans le pire cas* faisant intervenir la notion d'entropie topologique. Nous adapterons cette méthode pour l'obtention de bornes inférieures en moyenne en utilisant cette fois la notion d'entropie mesurable. Plus de détails expliquant comment l'entropie mesurable permettra d'exprimer la notion de complexité en moyenne sont fournis ci-dessous.
2. l'existence de fonctions à sens unique. Ce travail consistera à comprendre et adapter un certain nombre de résultats très récents reliant l'existence de fonctions à sens unique avec des questions relatives à la complexité de Kolmogorov. Nous exploiterons donc la relation entre complexité de Kolmogorov et entropie pour adapter les méthodes et établir des résultats similaires faisant la connexion entre l'existence de fonctions à sens unique et des questions relatives à l'entropie. Comme expliqué plus en détails ci-dessous, couplé aux résultats de la première direction de recherche concernant l'entropie et la complexité en moyenne ce résultat permettra d'apporter de nouveaux éléments sur une question fondamentale encore ouverte : est-il possible que certains problèmes NPTIME-difficiles en moyenne existent mais qu'il n'existe aucuns protocole cryptographiques sûrs ? Cette partie de la thèse s'appuiera sur l'expertise de Geoffroy Couteau sur la relation entre fonctions à sens unique et complexité.

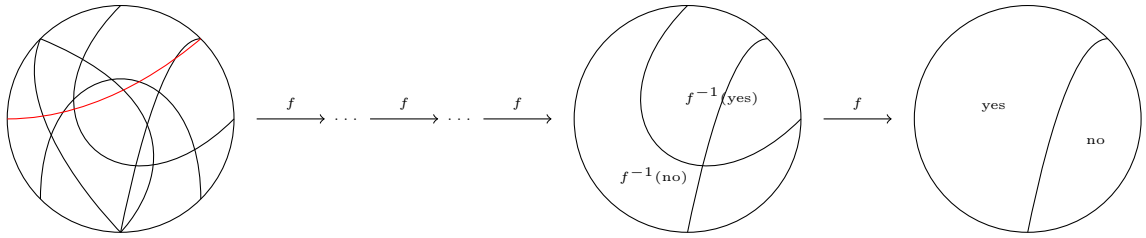

FIGURE 1 – Représentation géométrique de la décomposition en cellules et du problème W.
La décomposition est obtenue comme l'intersection des pré-images par f^i $(i = 0, 1, \ldots, k)$ des états acceptant et rejetant de la machine (représentée par la fonction f), et le problème W (un sous-ensemble de l'espace des entrées) est représenté par la "frontière" en rouge qui sépare les entrées appartenant à W et celles qui n'y appartiennent pas.

Entropie et complexité en moyenne Un travail récent de Seiller et Pellissier [18, 17] a établi un lien fort entre la notion d'entropie topologique et la complexité algorithmique "dans le pire cas". En effet, ce travail décrit une méthode abstraite permettant d'obtenir des bornes inférieures de complexité pour des machines algébriques, recouvrant plusieurs résultats de la littérature : les bornes inférieures sur les arbres algébriques de décision de Steele et Yao [21], celles sur les arbres de calcul algébriques de Ben-Or [3], le résultat de séparation de Cucker entre $\mathrm{NC}_{\mathbf{R}}$ et $\mathrm{PTIME}_{\mathbf{R}}$ [6], et améliorant l'un des plus forts résultats de séparation connu du à Mulmuley [16]. Ce dernier résultat établissait que le problème `maxflow`, connu pour être PTIME-complet, ne pouvait être calculé en temps polylogarithmique par une certaine notion de machine parallèle nommée "PRAMs sans opérations de bits". La technique abstraite introduite par Pellissier et Seiller permet d'étendre ce résultat pour montrer que `maxflow` n'est pas calculable en temps polylogarithmique par une PRAMs algébrique (calculant sur les entiers), effectuant un pas supplémentaire vers une éventuelle preuve que $\mathrm{NC} \neq \mathrm{PTIME}$.

L'essence de la technique consiste à représenter une machine M par un système dynamique, et le langage reconnu par cette machine en k étapes de calcul comme une partition de l'espace des entrées en *cellules* $(C_i)_{i\in I}$. Un langage W [1] est alors calculé par la machine en k étapes si et seulement si celui-ci est calculé comme l'union des C_i pour un sous-ensemble $K \subseteq I$ (cf. Figure 1), i.e. $W = \cup_{i\in K}$. Le nombre de cellules $\|I\|$ de ce partitionnement dépends alors de l'entropie topologique du système dynamique associé à la machine. On peut alors utiliser des invariants géométriques pour établir des bornes inférieures de complexité à partir de cette première borne sur le nombre de partitions. Par exemple dans le cas le plus simples des arbres de calcul algébriques, le nombre de composantes connexes (le nombre de Betti b_0) de chaque cellule peut alors être borné en utilisant le théorème de Milnor-Thom (qui borne la somme des nombres de Betti d'une variété (semi-)algébrique réelle), donnant une borne supérieure $B(k, M)$ sur le nombre de composantes connexes des langages reconnus en au plus k étapes de calcul. Ceci permet alors de montrer des bornes inférieures pour décider un langage L en fonction de son nombre de composantes connexes : si L possède un grand nombre de composantes connexes $b_0(L)$, alors le nombre d'étapes de calculs d'une machine décidant L devra au moins être assez grande pour que $B(k, M) \geqslant b_0(L)$. Dans le cas de l'extension de la borne inférieure de Mulmuley, la technique utilise le fait que le problème considéré a une trop grande "volatilité" en un sens précis (dépendant des dérivées successives de la frontière du problème).

Par ailleurs, comme décrit dans de nombreux travaux [4, 7, 20, 2], l'entropie topologique est fortement liée à la complexité de Kolmogorov (parfois nommée entropie algorithmique), un autre invariant utilisé pour obtenir des bornes inférieures, notamment via la technique d'incompressibilité [2] [11, Chapter 6]. La connexion entre entropie topologique et complexité vient donc s'ajouter aux nombreux résultats obtenus en utilisant la complexité de Kolmogorov, arguant de la grande pertinence de ces notions pour établir des bornes inférieures sur la complexité *dans le pire cas*.

Le premier objectif de cette thèse sera donc d'explorer l'extension des techniques basées sur l'entropie, en

1. Ici nous sommes dans le cadre algébrique, et un langage est donc un sous-ensemble de $\mathbf{R}^n$.
2. Notons par ailleurs que la technique d'incompressibilité a également été utilisée pour une preuve constructive alternative du Lemme Local de Lovász [15].

particulier l'entropie topologique, pour obtenir des résultats de bornes inférieures sur la complexité algorithmique **en moyenne**. La complexité en moyenne ne s'intéressant pas à l'instance du problème à résoudre qui demande le plus de ressources, mais plutôt à la quantité de resources nécessaire "en moyenne", un problème est considéré avec une distribution de probabilité associée (souvent supposée calculable en temps raisonnable). L'utilisation de la technique de Pellissier et Seiller dans ce cadre passera par la notion d'entropie mesurable, définie pour les systèmes dynamiques (mesurables) sur un espace mesuré $(X, \mathcal{B}, \mu)$, dépend de la mesure μ considérée. Le lien entre l'entropie mesurable $h_\mu(f)$ d'un système dynamique $f : X \to X$ et son entropie topologique est établi par le "principe variationnel" : $h_{\text{top}}(f) = sup_\mu h_\mu(f)$ [8]. Par ailleurs, nous chercherons à établir des résultats de bornes inférieures en moyenne grâce à une variante quantitative de la méthode décrite ci-dessus permettant de considérer les machines décidant d'un langage L à une fraction ϵ près en k étapes [3]. En d'autres termes, on pourra quantifier un degré d'erreur du calcul de la machine en k étapes en mesurant les entrées "échappant" au calcul de la machine, permettant de parler de machines calculant, e.g. en temps polynomial, sur une certaine fraction de ses entrées.

Entropie et fonctions à sens unique Faisant un point sur l'état des connaissances sur la complexité en moyenne, et quelles seraient les implications des diverses réponses aux questions alors ouvertes, Impagliazzo a défini en 1998 [10] cinq mondes possibles, mutuellement exclusifs, dans lesquels nous pourrions nous trouver : Algorithmica, Heuristica, Pessiland, Minicrypt, Cryptomania. Aujourd'hui encore, 20 ans plus tard, les questions sont toujours ouvertes et nous n'avons toujours pas exclu un seul des ces mondes possibles. Le monde nommé Pessiland est celui où certains problèmes de NPTIME seraient durs en moyenne mais où les fonctions à sens unique (OWF) n'existeraient pas, rendant la cryptographie sévèrement limitée.

Un certain nombre de résultats récents [1, 13, 14, 12, 19, 9] ont commencé à avancer dans la direction d'une élimination de la possibilité d'existence de Pessiland. Plus précisément, ces résultats établissent des équivalences entre la complexité moyenne de problèmes NP-complets et l'existence de fonction à sens unique, via la complexité de Kolmogorov temporelle (complexité KT). Ces résultats montrent donc l'équivalence entre l'existence des fonctions à sens unique et la complexité en moyenne d'un problème NPTIME-complet donné, et présente un premier pas pour éliminer Pessiland. Cependant, prouver l'impossibilité de Pessiland nécessite de montrer que l'existence d'un problème NPTIME-difficile en moyenne (quel qu'il soit) implique l'existence de fonctions à sens unique, et la NPTIME-complétude dans le pire cas n'implique pas une complétude vis à vis de la complexité en moyenne. Cette approche récente ne parvient donc pas encore à adresser pleinement le problème d'existence de Pessiland car la supposition qu'un problème particulier, même NPTIME-complet (dans le pire cas), soit NPTIME-difficile en moyenne est trop forte. Il n'est pas clair à ce point si cette limitation est dûe à des raisons purement techniques ou sont conséquences d'un manquement crucial de l'approche.

Le deuxième objectif de cette thèse sera de chercher à donner un éclairage nouveau à ces résultat au travers des liens mutuels entre complexité de Kolmogorov et entropie. Plus précisément, nous chercherons à établir un lien direct entre l'entropie et les fonctions à sens unique, en établissant une connection entre l'existence de fonctions à sens unique et des résultats de calcul et/ou de minimisation/maximisation d'entropie. Nous nous appuierons à la fois sur les récents résultats et sur l'expertise de Geoffroy Couteau dans l'étude de fonctions à sens uniques pour les petites classes de complexité [5]. Introduire l'entropie permettra notamment l'utilisation de techniques jusqu'alors inconsidérées. Notamment, couplant cet objectif avec les résultats détaillés précédemment établissant un lien entre la complexité en moyenne et entropie, celles-ci pourront être utilisées pour tenter d'établir une double équivalence :

$$\begin{array}{c}\text{Existence de problèmes} \\ \text{NPTIME-complet en moyenne}\end{array} \Leftrightarrow \begin{array}{c}\text{Minimisation / Maximisation} \\ \text{d'entropie}\end{array} \Leftrightarrow \begin{array}{c}\text{Existence de fonctions} \\ \text{à sens unique}\end{array}$$

Ceci permettra soit de raffiner les méthodes employées par les travaux récents basés sur la complexité de Kolmogorov, afin de se rapprocher d'une preuve d'impossibilité de Pessiland, soit d'établir une faille majeure (un résultat d'impossibilité) dans les approches récentes.

3. Formellement, en utilisant les notations ci-dessus, le langage W est reconnu à ϵ près en k étapes s'il existe $K \subseteq I$ tels que $\mu(W \setminus \cup_{i \in K} C_i) + \mu(\bar{W} \setminus \cup_{i \notin K} C_i) < \epsilon$, avec $\bar{K}$ le complémentaire de K.

Références

[1] E. Allender, J. Gouwar, S. Hirahara, and C. Robelle. Cryptographic hardness under projections for time-bounded kolmogorov complexity. *Electron. Colloquium Comput. Complex.*, 28 :10, 2021.

[2] A. V. Alpeev. An announce of results linking kolmogorov complexity to entropy for amenable group actions. *Journal of Mathematical Sciences*, 240(5), 2019.

[3] M. Ben-Or. Lower bounds for algebraic computation trees. In *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*, STOC '83, pages 80–86, New York, NY, USA, 1983. ACM.

[4] A. A. Brudno. Topological entropy, and complexity in the sense of a. n. kolmogorov. *Uspekhi Mat. Nauk*, 29 :157–158, 1974.

[5] C. Brzuska and G. Couteau. Towards fine-grained one-way functions from strong average-case hardness. Cryptology ePrint Archive, Report 2020/1326, 2020. `https://eprint.iacr.org/2020/1326`.

[6] F. Cucker. $\mathrm{P}_{\mathbf{R}} \neq \mathrm{NC}_{\mathbf{R}}$. *J. Complexity*, 8(3) :230–238, 1992.

[7] S. Galatolo, M. Hoyrup, and C. Rojas. Effective symbolic dynamics, random points, statistical behavior, complexity and entropy. *Information and Computation*, 208(1) :23–41, 2010.

[8] T. N. T. Goodman. Relating topological entropy and measure entropy. *Bulletin of the London Mathematical Society*, 3(2) :176–180, 1971.

[9] S. Hirahara. Average-case hardness of NP from exponential worst-case hardness assumptions. *Electron. Colloquium Comput. Complex.*, 28 :58, 2021.

[10] R. Impagliazzo. A personal view of average-case complexity. In *Proceedings of the 10th Annual Structure in Complexity Theory Conference (SCT'95)*. IEEE Computer Society, 1995.

[11] M. Li and P. M. B. Vitnyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Publishing Company, Incorporated, 3 edition, 2008.

[12] Y. Liu and R. Pass. Cryptography from sublinear-time average-case hardness of time-bounded kolmogorov complexity. *Electron. Colloquium Comput. Complex.*, 28 :55, 2021.

[13] Y. Liu and R. Pass. On one-way functions from np-complete problems. Cryptology ePrint Archive, Report 2021/513, 2021. `https://eprint.iacr.org/2021/513`.

[14] Y. Liu and R. Pass. On the possibility of basing cryptography on $\exp \neq \bpp$. *Electron. Colloquium Comput. Complex.*, 28 :56, 2021.

[15] J. Messner and T. Thierauf. A kolmogorov complexity proof of the lovász local lemma for satisfiability. In B. Fu and D.-Z. Du, editors, *Computing and Combinatorics*, pages 168–179, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[16] K. Mulmuley. Lower bounds in a parallel model without bit operations. *SIAM Journal of Compution*, 28(4) :1460–1509, 1999.

[17] L. Pellissier and T. Seiller. Lower bounds for algebraic machines, semantically. `https://hal.archives-ouvertes.fr/hal-02487667`, 2021.

[18] L. Pellissier and T. Seiller. Prams over integers do not compute maxflow efficiently. `https://hal.archives-ouvertes.fr/hal-01921942`, 2021.

[19] R. Santhanam. Pseudorandomness and the minimum circuit size problem. In T. Vidick, editor, *11th Innovations in Theoretical Computer Science Conference, ITCS 2020, January 12-14, 2020, Seattle, Washington, USA*, volume 151 of *LIPIcs*, pages 68 :1–68 :26. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[20] S. G. Simpson. Symbolic Dynamics : Entropy = Dimension = Complexity. *Theory of Computing Systems*, 56(3) :527–543, 2015.

[21] J. M. Steele and A. Yao. Lower bounds for algebraic decision trees. *Journal of Algorithms*, 3 :1–8, 1982.