Université de Montréal

Réconciliation et complexité de la communication de données corrélées

par

Hugues Mercier

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en informatique

Novembre 2002

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Réconciliation et complexité de la communication de données corrélées

présenté par :

Hugues Mercier

a été évalué par un jury composé des personnes suivantes :

Président-rapporteur
Monsieur Alain Tapp

Membre du jury
Monsieur Geňa Hahn

Directeur de recherche
Monsieur Pierre McKenzie

Codirecteur
Monsieur Stefan Wolf

Mémoire accepté le 23 janvier 2003

Sommaire

Ce mémoire considère le problème où deux interlocuteurs distants possèdent chacun une chaîne de bits, le but étant qu'un des interlocuteurs apprenne la chaîne de son vis-à-vis en minimisant la communication. Contrairement au modèle original de la complexité de la communication, la difficulté est due au fait que les chaînes sont corrélées. Ce modèle est inhérent à plusieurs applications pratiques incluant la synchronisation de données mobiles, la réconciliation de séquences de symboles comme les séquences de nucléotides dans les molécules d'ADN, le calcul distribué, la sauvegarde de fichiers et la distribution quantique de clés secrètes.

Nous analysons les modèles déterministe, déterministe amorti, probabiliste avec générateurs aléatoires privés et probabiliste avec générateur aléatoire public. Nous montrons entre autres que pour les applications mentionnées précédemment, tous ces modèles sont équivalents. Nous considérons également le nombre de messages que les interlocuteurs doivent échanger pour réconcilier leurs chaînes, car la non-interactivité est nécessaire pour certaines applications.

Mots-clés : réconciliation, complexité de la communication, données corrélées, théorie de l'information, codes correcteurs d'erreurs.

Abstract

This thesis considers the problem where two distant parties each possess a chain of bits, the goal being for one party to learn his encounter's input while minimizing the communication. Unlike the original communication complexity model, the difficulty arises because the inputs are correlated. This model is inherent to several practical applications including synchronisation of mobile data, reconciliation of sequences of symbols such as nucleotides sequences in DNA molecules, distributed computations, remote file storage and quantum key distribution.

We analyse the deterministic, amortized deterministic, private coin randomized and public coin randomized models. Among other things, we show that all these models are equivalent for the applications previously mentioned. We also consider the number of messages the parties need to exchange to reconcile their inputs, since non-interactivity is required for some applications.

Keywords : reconciliation, communication complexity, correlated data, information theory, error-correcting codes.

Table des matières

Liste des tableaux

Mesures de complexité

Notation	Remarques	Définition
$D(f)$	Complexité de la communication déterministe de la fonction f	Définition 1.3
$D(S)$	Complexité de la communication déterministe du problème avec données corrélées S	Définition 1.10
$D^k(S)$	Complexité de la communication déterministe à k rondes	Définition 1.10
$D_{x\|y}(S)$	Complexité de la communication déterministe lorsqu'Alice connaît la chaîne de Bob	Définition 1.11
$D(S,T)$	Complexité de la communication déterministe simultanée de S et T	Définition 2.1
$D(S^l)$	Complexité de la communication déterministe simultanée de l exemplaires de S	Définition 2.2
$\overline{D}(S)$	Complexité de la communication déterministe amortie	Définition 2.3
$R_\epsilon(S)$	Complexité de la communication probabiliste avec erreur ϵ	Définition 3.5
$R_0(S)$	Complexité de la communication probabiliste sans erreur	Définition 3.5
$R_{\epsilon,pub}(S)$	Complexité de la communication probabiliste publique avec erreur ϵ	Définition 3.13
$D_{\epsilon,\mu}(S)$	Complexité distributionnelle avec erreur ϵ	Définition 3.24
$D_{0,\mu}(S)$	Complexité distributionnelle sans erreur	Définition 3.25

Notation

$S_{X,Y}$	Support de (X, Y)	Définition 1.7
$a_A(x)$	Ambiguité de x	Définition 1.8
$\widehat{a_A}$	Ambiguité maximale de X	Définition 1.9
G_S	Hypergraphe caractéristique de S	Définition 1.12
$\sigma_{\mathcal{P}}(x, y)$	Concaténation des messages transmis lors de l'exécution de $\mathcal{P}$ sur (x, y)	Définition 1.20

Remerciements

Mes premiers remerciements s'adressent à mes directeurs de recherche, Pierre McKenzie et Stefan Wolf, pour leur professionnalisme, leur dynamisme et leurs compétences académiques. Je remercie Pierre de m'avoir fait confiance dès les premiers moments de ce qui était pour moi un retour aux études, et Stefan pour ses idées originales et sa porte toujours ouverte.

Ce mémoire a été entrepris suite à une série de discussions avec Hervé Caussinus. Je le remercie d'avoir suggéré un sujet de recherche si intéressant.

Je remercie Alain Tapp pour les discussions que nous avons partagées, ainsi que Martin Sauerhoff et Pascal Tesson de m'avoir finalement arrêté de travailler sur des problèmes résolus il y a une décennie.

Je remercie le Fonds québécois de la recherche sur la nature et les technologies (NATEQ) de m'avoir accordé une bourse, et le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) du soutien financier complémentaire par l'entremise des subventions de recherche de Pierre McKenzie.

Je remercie mes parents de m'avoir transmis leur soif de savoir et d'avoir patiemment répondu aux innombrables questions que je leur ai posées durant ma jeunesse. Finalement, je remercie Isabelle de sa spontanéité, sa joie de vivre, son authenticité, en somme de faire partie de ma vie ...

Avant-propos

Lorsque des humains, des ordinateurs ou des parties d'un système veulent résoudre un problème conjointement, ils doivent communiquer. Cela peut être nécessaire lorsque la tâche à effectuer est trop lourde pour être réalisée par une seule partie, ou encore parce que les données du problème sont décentralisées. La communication peut être explicite comme dans le cas de deux internautes s'échangeant des fichiers sur Internet, ou implicite dans le cas d'un processeur qui accède à des données sur un disque dur.

La complexité de la communication est la théorie mathématique des processus requérant de la communication. Elle mesure la quantité d'information qui doit être échangée pour calculer une fonction ou résoudre un problème. La complexité d'un problème est inhérente à celui-ci et ne dépend donc pas d'un protocole particulier le résolvant. Contrairement à la complexité du calcul, la complexité de la communication ne tient pas compte de la puissance de calcul des participants. Il existe d'ailleurs des problèmes distribués pouvant être résolus à l'aide de protocoles requérant peu de communication, mais nécessitant un temps de calcul exponentiel. Évidemment, un compromis entre la communication et le temps de calcul est nécessaire pour quiconque désire obtenir des protocoles pouvant être utilisés en pratique.

La théorie de l'information définit un modèle dans lequel un émetteur envoie un message à un récepteur, celui-ci recevant une version bruitée du message original. L'entropie quantifie l'incertitude du receveur par rapport au message envoyé par l'émetteur, et l'information est la réduction de l'entropie. La théorie de l'information s'intéresse au taux de transmission de données qui peut être atteint et à la façon d'encoder les données efficacement afin que le récepteur puisse décoder le message original malgré la présence de bruit.

Ce mémoire traite de la complexité de la communication de données corrélées, située à mi-chemin entre la complexité de la communication et la théorie de l'information. Dans notre modèle, deux interlocuteurs distants possèdent chacun une chaîne de bits, les deux chaînes étant corrélées, et le but est qu'un des interlocuteurs apprenne la chaîne de son vis-à-vis en minimisant la communication. Nous supposons que la communication est effectuée sur un canal de transmission sans erreur et qu'elle alterne d'un interlocuteur à l'autre selon un protocole sur lequel ils se sont entendus initialement.

Le premier chapitre présente le modèle déterministe de communication de données corrélées. Le deuxième chapitre analyse la complexité déterministe amortie, c'est-à-dire la résolution simultanée de plusieurs exemplaires d'un problème ou de plusieurs problèmes différents. Le troisième chapitre étudie la complexité de la communication probabiliste, ce qui à notre connaissance n'a jamais été fait auparavant. Finalement, le quatrième chapitre traite de la complexité de la communication de problèmes équilibrés, modèle qui englobe toutes les applications pratiques mentionnées au sommaire. Mentionnons que nous avons obtenu plusieurs nouveaux résultats, parmi lesquels la classification presque complète du modèle probabiliste ainsi que la démonstration que les modèles déterministe et

probabiliste sont équivalents pour les problèmes équilibrés sont certainement les plus intéressants.

Bien que le modèle de communication de données corrélées ait été originalement étudié dans le but de résoudre plusieurs problèmes pratiques de communication, il n'en demeure pas moins qu'il peut être abstrait mathématiquement de façon élégante. Toutefois, comme les préalables mathématiques nécessaires à la compréhension de ce mémoire sont assez variés, nous avons choisi de les regrouper en annexe afin de ne pas alourdir inutilement le texte. Mentionnons finalement que tous les logarithmes sont en base 2.

Chapitre 1

Complexité déterministe

Dans ce chapitre, nous introduisons un modèle déterministe pour les problèmes de communication dont les données sont corrélées, ainsi que les mesures de complexité que nous utilisons pour en faire l'analyse. À l'exception du théorème 1.25, du lemme 1.28 et de la preuve du lemme 1.22, ce chapitre ne contient pas de nouveau matériel et permet surtout d'introduire les notions dont nous aurons besoin pour les chapitres subséquents.

1.1 Modèle original de la complexité de la communication

Le modèle original de la complexité de la communication classique à deux participants a été introduit par Yao [54] en 1979 et est traité en détails dans l'excellent ouvrage de Kushilevitz et Nisan [27]. Il a été initialement étudié afin d'obtenir des bornes inférieures pour la communication dans les puces VLSI [50] ainsi que pour la communication lors de calculs distribués [1].

Soient X, Y, Z des ensembles finis et $f : X \times Y \to Z$ une fonction. Deux interlocuteurs distants, Alice et Bob[1], possèdent respectivement des éléments $x \in X$ et $y \in Y$. Alice n'a aucune information sur y, Bob n'a aucune information sur x, et tous deux veulent calculer $f(x, y)$ en minimisant la communication.

Pour calculer f, Alice et Bob exécutent un protocole $\mathcal{P}$ sur lequel ils se sont entendus initialement. $\mathcal{P}$ est en fait un arbre binaire où chaque noeud interne i est étiqueté par une fonction $a_i : X \to \{0, 1\}$ ou $b_i : Y \to \{0, 1\}$, et où chaque feuille est étiquetée par un élément $z \in Z$. Lors de l'exécution de $\mathcal{P}$ sur l'entrée (x, y), les interlocuteurs parcourent l'arbre de la racine vers les feuilles en fonction des valeurs de a_i et de b_i. Lorsqu'un noeud est étiqueté par une fonction a_i, Alice calcule $a_i(x)$; si $a_i(x) = 0$, le noeud suivant du protocole est l'enfant gauche du noeud i, tandis que si $a_i(x) = 1$, le protocole se poursuit avec l'enfant droit. Comme Bob ne connaît pas x, Alice doit transmettre un bit à Bob afin qu'il sache quel est le noeud suivant. Lorsqu'un noeud est étiqueté par une fonction b_i, c'est Bob qui calcule $b_i(y)$ et qui envoie le résultat à Alice. La valeur de $f(x, y)$ est la valeur de la feuille atteinte par le protocole à partir de la racine sur l'entrée (x, y), et le nombre de bits de communication en pire cas entre Alice et Bob correspond exactement à la hauteur de l'arbre.

Définition 1.1. Le *coût* d'un protocole $\mathcal{P}$ est la hauteur de l'arbre défini par ce protocole.

Définition 1.2. Soit $f : X \times Y \to Z$ une fonction. La complexité de la communication déterministe de f, notée $D(f)$, est le coût minimal de $\mathcal{P}$, parmi tous les protocoles $\mathcal{P}$ calculant f.

[1]Les prénoms Alice et Bob sont utilisés partout dans la littérature ; il en est de même dans ce mémoire.

La définition 1.3, équivalente à la précédente, exprime mieux la communication intrinsèque au modèle.

Définition 1.3. La *complexité de la communication déterministe* de f, $D(f)$, est le nombre minimal de bits[2] qu'Alice et Bob doivent échanger pour calculer f à coup sûr pour toute paire (x, y).

Le lemme suivant illustre une façon triviale qui permet aux interlocuteurs de calculer f. Un des objectifs de l'étude de la complexité de la communication est évidemment de trouver des protocoles plus performants ou de démontrer qu'il n'est pas possible de faire mieux.

Lemme 1.4. *Pour toute fonction* $f : X \times Y \to Z$,

$$D(f) \leq \lceil \log |X| \rceil + \lceil \log |Z| \rceil.$$

Démonstration. Alice envoie x à Bob, ce qui nécessite $\lceil \log |X| \rceil$ bits de communication. De son côté, Bob calcule $z = f(x, y)$ et envoie la réponse à Alice, ce qui peut être fait avec $\lceil \log |Z| \rceil$ bits. □

Le lemme précédent implique que la valeur de f après l'exécution du protocole $\mathcal{P}$ est connue d'Alice et de Bob. Cette contrainte peut augmenter $D(f)$ d'au plus $\lceil \log |Z| \rceil$ bits, mais nous ne l'imposons pas pour ce mémoire pour une raison évidente que sera expliquée à la section 1.2.

Une autre variante du modèle original de Yao est la notion de rondes. Il peut être intéressant de considérer non pas le nombre de bits de communication entre les interlocuteurs, mais plutôt l'interaction véritable entre ceux-ci. En effet, pour

[2] Nous supposons que toutes les communications entre les interlocuteurs sont binaires.

certaines fonctions, il existe des protocoles efficaces pour lesquels Alice envoie une seule série de bits à Bob, tandis que pour d'autres fonctions, de nombreux échanges sont nécessaires afin de réduire la communication. En pratique, il est avantageux d'avoir des protocoles ayant le moins de rondes possibles, et il existe même plusieurs problèmes qui requièrent des protocoles non interactifs.

Définition 1.5. Un protocole à k rondes est un protocole tel que pour toute entrée, il y a au plus $k-1$ alternances entre les bits envoyés par Alice et les bits envoyés par Bob. La *complexité de la communication déterministe d'un protocole à k rondes*, notée $D^k(f)$, est le coût du meilleur protocole à k rondes pour f.

Définition 1.6. Un protocole à plus d'une ronde est dit *interactif*, tandis qu'un protocole à une ronde est dit *non interactif*.

1.2 Complexité de la communication de données corrélées

Le sujet de ce mémoire est la complexité de la communication de données corrélées. Ce modèle a été introduit par Orlitsky [35, 36, 37, 38], même si des articles antérieurs, entre autres par Witsenhausen [53], se sont attaqués à certains problèmes sans considérer le modèle de façon globale. Les appellations «communication interactive» et «communication avec information partielle» sont utilisées dans la littérature mais, selon nous, l'expression «communication de données corrélées» est plus appropriée.

Le modèle original de Yao requiert qu'Alice et Bob puissent calculer $f(x, y)$ pour toutes les paires (x, y) possibles. Dans le modèle avec données corrélées, chaque interlocuteur possède de l'information sur la chaîne de son vis-à-vis en fonction de sa propre chaîne, et cela permet d'éliminer certaines paires.

Définition 1.7. Le *support* de (X, Y), noté $S_{X,Y} \subseteq X \times Y$, est l'ensemble des paires possibles entre Alice et Bob. La notation S est utilisée lorsqu'il n'y a pas de confusion possible.

Définition 1.8. L'*ambiguité* de x, notée $a_A(x)$, est l'ensemble des $y \in Y$ possibles pour un $x \in X$. Formellement, $a_A(x) \stackrel{\text{déf}}{=} \{y \in Y \mid (x, y) \in S\}$. L'ambiguité de y, notée $a_B(y)$, est définie de manière analogue.

Autrement dit, l'ambiguité d'Alice, $a_A(x)$, est l'ensemble de toutes les chaînes possibles chez Bob lorsqu'elle possède la chaîne x.

Définition 1.9. L'*ambiguité maximale* de X, $\widehat{a_A}(S) \stackrel{\text{déf}}{=} \max_{x \in X}\{|a_A(x)|\}$, est le nombre maximal de valeurs possibles de Y pour toute valeur de X. L'ambiguité maximale de Y, notée $\widehat{a_B}(S)$, est définie de manière analogue. Nous écrivons $\widehat{a_A}$ et $\widehat{a_B}$ lorsqu'il n'y a pas de confusion possible.

Les paramètres du modèle sont les suivants : Alice possède une chaîne $x \in X$ et Bob possède une chaîne $y \in Y$ avec la restriction que $(x, y) \in S$, et le but est que Bob apprenne la valeur de x. Il n'est pas nécessaire qu'Alice apprenne la valeur de y. Le modèle analyse la communication entre les interlocuteurs et suppose que ceux-ci ont une puissance de calcul illimitée.

Définition 1.10. La *complexité de la communication déterministe* d'un problème ayant un ensemble de support $S \subseteq X \times Y$, notée $D(S)$, est le nombre minimal

de bits qu'Alice et Bob doivent échanger afin que Bob apprenne la chaîne d'Alice à coup sûr pour toute paire $(x, y) \in S$. Nous écrivons $D^k(S)$ lorsque le nombre de rondes est borné par k. Comme le modèle de communication est asymétrique, nous supposons que la dernière ronde est effectuée d'Alice vers Bob (le but étant que Bob apprenne x, il est en effet inutile qu'il envoie le dernier message).

Définition 1.11. Notons $D_{x|y}$ la *complexité de la communication déterministe de S lorsqu'Alice connaît la chaîne de Bob.*

1.3 Le problème de la ligue sportive

Le problème de la ligue sportive a été introduit pas Orlitsky [35] et vaut la peine d'être présenté en guise d'introduction. Une ligue sportive a 2^n équipes, chacune ayant comme nom une chaîne de n bits. Bob est un maniaque de sport et connaît les deux équipes qui participent à la grande finale de la ligue, mais une panne d'électricité l'a empêché de regarder le match et d'apprendre l'identité de l'équipe gagnante. Alice, de son côté, a entendu à la radio le nom de l'équipe gagnante, mais n'a aucune idée de l'équipe qu'elle a vaincue en finale. Bob veut apprendre d'Alice le nom de l'équipe gagnante en minimisant la communication avec elle. Formellement, $S = \{(e_1, \{e_1, e_2\}) \mid e_1 \neq e_2\}$, où $e_1, e_2 \in \{0,1\}^n$.

Si la communication entre les deux interlocuteurs est non interactive, Alice ne peut rien faire de mieux que de communiquer à Bob les n bits de l'équipe gagnante. En effet, si Alice communique moins de n bits à Bob, il existe deux équipes e_1 et e_2 pour lesquelles elle envoie le même message. Si e_1 et e_2 sont les deux équipes participant à la finale, alors Bob ne peut pas déduire l'équipe gagnante avec certitude à partir du message qu'il reçoit.

Par contre, une économie substantielle est possible lorsque deux rondes de communication sont permises. Bob envoie à Alice la position d'un des bits où les chaînes diffèrent, ce qui nécessite $\lceil \log n \rceil$ bits de communication. Alice n'a qu'à communiquer à Bob la valeur du bit de l'équipe gagnante pour la position demandée. Ce protocole requiert $\lceil \log n \rceil + 1$ bits de communication, un gain exponentiel par rapport au protocole non interactif.

1.4 Résultats préliminaires

Pour analyser mathématiquement les problèmes de communication avec données corrélées, il est fort utile d'utiliser leur représentation sous forme de graphes. Voici trois approches équivalentes.

- La première approche a été définie par Witsenhausen [53]. Soit $x \in X$ la chaîne d'Alice et $y \in Y$ la chaîne de Bob, toujours avec la condition que $(x, y) \in S$. Appelons G_{XY} le graphe bipartie formé des deux ensembles de sommets X et Y et dont les arêtes entre les sommets des ensembles X et Y correspondent aux paires $(x, y) \in S$. Définissons le graphe G_X pour lequel les sommets sont les éléments de l'ensemble X et où deux sommets x_1 et x_2 sont reliés par une arête si et seulement si il y a un sommet y de G_{XY} adjacent à la fois à x_1 et à x_2 .
- La deuxième approche est celle que nous utilisons dans ce mémoire et a été introduite par Orlitsky [35]. Étant donné le support $S \subseteq X \times Y$, introduisons l'hypergraphe G_S (voir l'annexe A.2). Les sommets de G_S sont les éléments de X, et pour tout $y \in Y$, il y a une hyperarête $E(y) = \{x \mid (x, y) \in S\}$.

- Une troisième approche, que nous ne définissons pas ici, a été présentée par Karpovsky, Levitin et Trachtenberg [26]. Même si elle peut être utilisée pour faire l'analyse de problèmes de communication avec données corrélées lorsque $X = Y$, cette approche est surtout utile pour des modèles de détection et de correction d'erreurs.

Définition 1.12. G_S est appelé l'*hypergraphe caractéristique* de S.

Il est important de remarquer que le graphe G_S est défini en sachant que Bob veut apprendre la chaîne d'Alice. Le graphe serait différent si Alice voulait apprendre la chaîne de Bob, à moins que le support S soit symétrique[3].

Remarque 1.13. L'ambiguité de Bob correspond au degré maximal des hyperarêtes de G_S (voir l'annexe A.2).

Définition 1.14. Un ensemble de support S est *trivial* si $D(S) = 0$, autrement dit si $\widehat{a_B} = 1$.

Nous présentons maintenant les premiers résultats sur la complexité de la communication de problèmes avec données corrélées. Ils ont tous été démontrés par Orlitsky [35, 38].

Lemme 1.15. *Pour tout problème de communication avec données corrélées,*

$$D^1(S) \geq D^2(S) \geq \cdots \geq D(S).$$

Démonstration. Trivial. □

[3] Un support symétrique est un support S tel que $(x, y) \in S \Leftrightarrow (y, x) \in S$.

Lemme 1.16. *Si $S_1 \subseteq S_2$, alors pour tout $k \in \mathbb{N}$,*

$$D^k(S_1) \leq D^k(S_2).$$

Démonstration. Il s'agit de remarquer que tout protocole pour S_2 est également un protocole pour S_1. □

Lemme 1.17. *Pour tout problème de communication avec données corrélées S,*

$$D(S) \geq \lceil \log \widehat{a_B} \rceil.$$

Démonstration. Supposons qu'il existe un protocole $\mathcal{P}$ requérant $\alpha < \lceil \log \widehat{a_B} \rceil$ bits de communication. Cela implique qu'il existe $y_i \in Y$ pour lequel $|\{x \mid (x, y) \in S\}| > 2^\alpha$. Si la chaîne de Bob est y_i, il existe donc au moins deux éléments distincts de X pour lesquels les bits transmis entre Alice et Bob sont les mêmes, ce qui est une contradiction. □

Lemme 1.18. *Pour tout problème de communication avec données corrélées S,*

$$D^1_{x|y}(S) = D_{x|y}(S) = \lceil \log \widehat{a_B} \rceil.$$

Démonstration. Par le lemme 1.17, $D(S) \geq \lceil \log \widehat{a_B} \rceil$. Montrons que $D(S) \leq \lceil \log \widehat{a_B} \rceil$. Soit G_S l'hypergraphe caractéristique de S. Initialement, Alice et Bob s'entendent sur un encodage des sommets des hyperarêtes de G_S, ce qui peut être fait avec $\lceil \log \widehat{a_B} \rceil$ bits. Lors de l'exécution du protocole, Alice, qui connaît y et par le fait même l'hyperarête a_y correspondant à l'ambiguité de Bob, n'a qu'à envoyer l'encodage de x. Nous obtenons donc que $D(S) \leq D^1(S) \leq \lceil \log \widehat{a_B} \rceil$. □

Lemme 1.19. *Pour tout problème de communication S,*

$$D^1(S) = \lceil \log \chi(G_S) \rceil,$$

où $\chi(G_S)$ est le nombre chromatique de G_S.

Démonstration.

- $D^1(S) \leq \lceil \log \chi(G_S) \rceil$

  Alice et Bob s'entendent sur un coloriage de G_S utilisant $\chi(G_S)$ couleurs. Alice envoie la couleur du sommet x à Bob. Par construction de G_S, tous les sommets x tels que $(x, y) \in S$ sont de couleur différente, ce qui permet à Bob de déduire x à partir de sa couleur et de y.

- $D^1(S) \geq \lceil \log \chi(G_S) \rceil$

  Supposons qu'il existe un protocole $\mathcal{P}$ requérant $\alpha < \lceil \log \chi(G_S) \rceil$ bits de communication. Cela veut dire qu'il existe un y pour lequel Alice envoie le même message pour au moins deux éléments de X, ce qui ne permet pas à Bob de les distinguer avec certitude.

□

Pour la majorité des problèmes, le lemme 1.19 est inutilisable en pratique, car le calcul exact du nombre chromatique d'un graphe est un problème $\mathcal{NP}$-difficile (consulter [42] pour plus de détails sur les classes de complexité du calcul). En fait, Feige et Kilian [17] ont montré que si $\mathcal{NP} \not\subseteq \mathcal{ZPP}$, il est impossible d'approximer en temps polynomial le nombre chromatique d'un graphe de n sommets à un facteur $n^{1-\epsilon}$ près, pour toute constante $\epsilon > 0$. Malgré ce résultat peu encourageant, il existe des méthodes heuristiques, par exemple des algorithmes évolutifs hybrides [20], qui permettent d'obtenir des bornes supérieures

intéressantes pour le nombre chromatique de graphes de taille raisonnable (moins de 1000 sommets). Par le fait même, de telles méthodes permettent d'obtenir des bornes supérieures intéressantes pour $D^1(S)$.

Le lemme 1.22 illustre qu'un protocole interactif peut permettre jusqu'à un gain exponentiel sur le nombre de bits communiqués par rapport à un protocole non interactif. Pour le démontrer, nous aurons besoin du fait que tout problème de communication avec données corrélées respecte la propriété d'absence de préfixe[4].

Définition 1.20. Notons $\sigma_{\mathcal{P}}(x, y)$ la concaténation de tous les messages transmis lors de l'exécution d'un protocole $\mathcal{P}$ sur l'entrée (x, y).

Lemme 1.21 (Propriété d'absence de préfixe). *Soit $S = X \times Y$ un problème de communication avec données corrélées, et soient $(x', y), (x, y), (x, y') \in S$ avec $x \neq x'$. Alors $\sigma_{\mathcal{P}}(x', y)$ n'est pas un préfixe de $\sigma_{\mathcal{P}}(x, y')$, et $\sigma_{\mathcal{P}}(x, y')$ n'est pas un préfixe de $\sigma_{\mathcal{P}}(x', y)$ (en particulier ils ne peuvent pas être égaux).*

Démonstration. Voir [35]. □

La propriété d'absence de préfixe est importante, car les modèles qui ne la respectent pas peuvent comprimer les messages et ainsi réduire la communication, tel qu'illustré par Papadimitriou et Sipser [43]. De plus, elle élimine la nécessité d'avoir un symbole spécial pour indiquer la fin de l'exécution du protocole.

Lemme 1.22 (Mercier 2002, démonstration seulement). *Pour tout problème de communication avec données corrélées S,*

$$D(S) \geq \lceil \log D^1(S) \rceil.$$

[4] «Prefix-freeness property» en anglais.

Démonstration. Soit $\mathcal{P}$ un protocole avec support S dont la complexité de la communication est $D(S)$. Construisons un protocole non interactif $\mathcal{P}$' de complexité $2^{D(S)}$. Alice considère toutes les $2^{D(S)}$ séquences possibles de $D(S)$ bits. Pour chacune de ces séquences α, elle transmet $f(\alpha)$ à Bob, où

$$f(\alpha) = \begin{cases} 1 \text{ si } \exists y' \in a_A(x) \mid \sigma_{\mathcal{P}}(x, y') \text{ est un préfixe de } \alpha \\ 0 \text{ sinon} \end{cases}$$

Bob trouve l'unique $x' \in a_B(y)$ tel que $\sigma_{\mathcal{P}}(x', y)$ est un préfixe d'un α pour lequel $f(\alpha) = 1$, et il conclut que c'est la chaîne d'Alice.

Pour justifier que l'algorithme fonctionne, il faut montrer qu'il existe un et un seul $x' \in a_B(y)$ tel que $\sigma_{\mathcal{P}}(x', y)$ est un préfixe d'un α pour lequel $f(\alpha) = 1$. Il est clair qu'il en existe au moins un, car $f(\sigma_{\mathcal{P}}(x, y)) = 1$. Supposons qu'il existe un $x' \in X, x' \neq x$, tel que $\sigma_{\mathcal{P}}(x', y)$ est un préfixe de α' pour lequel $f(\alpha') = 1$. Le fait que $f(\alpha') = 1$ implique qu'il existe un $y' \in a_A(x)$ tel que $\sigma_{\mathcal{P}}(x, y')$ est un préfixe de α'. Comme $\sigma_{\mathcal{P}}(x, y')$ et $\sigma_{\mathcal{P}}(x', y)$ sont des préfixes de α', il suit que $\sigma_{\mathcal{P}}(x, y')$ est un préfixe de $\sigma_{\mathcal{P}}(x', y)$ ou que $\sigma_{\mathcal{P}}(x', y)$ est un préfixe de $\sigma_{\mathcal{P}}(x, y')$. Ceci est une contradiction, car S respecte la propriété d'absence de préfixe. □

Le lemme 1.23 améliore d'un bit la borne du lemme précédent, mais la preuve de ce résultat un peu plus fort, bien que conceptuellement simple, prend plusieurs pages.

Lemme 1.23. *Pour tout problème de communication avec données corrélées S,*

$$D(S) \geq \lceil \log D^1(S) \rceil + 1.$$

Démonstration. Voir [35]. □

Terminons cette section par un retour sur le problème de la ligue sportive de la section 1.3. Nous avons montré que $D^1(S) = n$ et que $D^2(S) \leq \lceil \log n \rceil + 1$. Le lemme 1.23 nous permet de conclure que le protocole à deux rondes est optimal, c'est-à-dire que $D^2(S) = D^3(S) = \cdots = D(S) = \lceil \log n \rceil + 1$. Dans un autre ordre d'idées, si Alice connaît y, alors un seul bit de communication suffit d'après le lemme 1.18.

1.5 Deux rondes sont presque optimales

Le dernier exemple de la section précédente semble assez étonnant : pour le problème de la ligue sportive, il est inutile d'utiliser un protocole à plus de deux rondes, car cela ne permet pas de diminuer le nombre de bits communiqués. Dans cette section, nous présentons un des résultats les plus intéressants de la complexité de la communication de données corrélées, à savoir que pour tout problème, deux messages sont presque optimaux. Ce résultat contraste avec le modèle original de Yao : Duris, Galil et Schnitger [15] ont en effet démontré que pour tout $k \in \mathbb{N}$, il existe des suites de fonctions $(f_i)_{i \in \mathbb{N}}$ avec $f_i : \{0,1\}^i \times \{0,1\}^i \to \{0,1\}$ pour lesquelles $D^k(f_n)$ est exponentiellement plus grand que $D^{k+1}(f_n)$. Avant de démontrer le résultat principal de cette section, nous avons besoin de quelques notions préliminaires.

Lemme 1.24. $1 \cdot \frac{p-1}{p} \cdot \frac{p-2}{p} \cdots \frac{p-t+1}{p} \geq \left(1 - \frac{t}{p}\right)^t$, *où* $p \in \mathbb{N}$.

Démonstration.

$$\begin{aligned} 1 \cdot \frac{p-1}{p} \cdot \frac{p-2}{p} \cdots \frac{p-t+1}{p} &\geq 1 \cdot \left(1-\frac{1}{p}\right) \cdot \left(1-\frac{2}{p}\right) \cdots \left(1-\frac{t-1}{p}\right) \\ &\geq \left(1-\frac{t}{p}\right) \cdot \left(1-\frac{t}{p}\right) \cdots \left(1-\frac{t}{p}\right) \geq \left(1-\frac{t}{p}\right)^t. \end{aligned}$$

□

Le prochain théorème garantit l'existence d'une famille de fonctions de «hachage» qui permettront par la suite de démontrer plusieurs résultats intéressants. Une famille $H_{m,t}$ avec des paramètres légèrement différents que ceux présentés a été explicitement construite par Fredman, Komlòs et Szemerèdi [19].

Théorème 1.25 (Mercier 2002). *Soient m et t deux entiers supérieurs à 1. Il existe une famille de $k = 4t\lceil \log m \rceil$ fonctions, $H_{m,t}$, telles que :*

1. *Toute fonction $h \in H_{m,t}$ va de $\{1, \ldots, m\}$ à $\{1, \ldots, p\}$, où $p = 4t^2$;*
2. *Pour tout sous-ensemble $A \subseteq \{1, \ldots, m\}$ de taille au plus t, au moins la moitié des fonctions dans $H_{m,t}$ sont injectives sur A.*

Démonstration. Choisissons au hasard une fonction $h : \{1, \ldots, m\} \to \{1, \ldots, p\}$. La probabilité que h soit injective sur un ensemble $A \subseteq \{1, \ldots, m\}$ de taille au plus t est bornée par $1 \cdot \frac{p-1}{p} \ldots \frac{p-t+1}{p} \geq \left(1-\frac{t}{p}\right)^t = \left(1-\frac{1}{4t}\right)^t \geq \frac{3}{4}$ (voir le lemme 1.24). Choisissons maintenant au hasard k fonctions $h_1, h_2 \ldots, h_k$: $\{1, \ldots, m\} \to \{1, \ldots, p\}$ et définissons les variables aléatoires Z_i prenant la valeur 0 si h_i est injective sur A et 1 sinon. Il est clair que $E[Z_i] \leq 3/4$. La probabilité qu'au moins la moitié des fonctions h_i ne soient pas injectives sur A est :

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{k} Z_i \geq \frac{k}{2}\right) &= \Pr\left(\frac{\sum_{i=1}^{k} Z_i}{k} \geq \frac{1}{2}\right) \\
&= \Pr\left(\frac{\sum_{i=1}^{k} Z_i}{k} - \frac{1}{4} \geq \frac{1}{4}\right) \\
&\leq \Pr\left(\left|\frac{\sum_{i=1}^{k} Z_i}{k} - \frac{1}{4}\right| \geq \frac{1}{4}\right) \\
&\leq \Pr\left(\left|\frac{\sum_{i=1}^{k} Z_i}{k} - \frac{1}{4}\right| \geq \frac{3}{16}\right) \\
&\leq 2e^{-\frac{k(3/16)^2}{2\cdot 3/16}} \text{ par l'inégalité de Chernoff (voir l'annexe A.4)} \\
&= 2e^{-\frac{3k}{32}} \\
&< 1 \text{ pour } k > 7,39.
\end{aligned}$$

La dernière inégalité est toujours vraie, car $k = 4t\lceil \log m \rceil$ et $m, t \geq 2$. Il existe donc une famille de k fonctions pour lesquelles au moins la moitié sont injectives sur A. □

Utilisons maintenant une famille de fonctions de hachage pour démontrer que pour tout problème de communication avec données corrélées, il existe un protocole à deux rondes qui nécessite au plus quatre fois le nombre de bits de communication requis par le protocole optimal.

Lemme 1.26. *Pour tout problème de communication avec données corrélées S,*

$$D^2(S) \leq \lceil \log\lceil \log \chi(G_S)\rceil\rceil + 3\lceil \log \widehat{a_B}\rceil + 4.$$

Démonstration. Soit S un problème de communication non trivial (il est clair que le lemme est vrai si S est trivial) et G_S son hypergraphe caractéristique. Alice et Bob s'entendent sur un coloriage ψ de G_S avec $\chi(G_S)$ couleurs ainsi que sur une famille de fonctions $H = H_{\chi(G_S),\widehat{a_B}}$ possédant les propriétés énoncées au théorème 1.25.

Bob considère l'hyperarête a_y qui détermine $a_B(y)$, les valeurs de x possibles chez Alice. Il choisit une fonction $h \in H$ qui est injective sur les couleurs de a_y et envoie sa description à Alice, ce qui requiert $\lceil \log(4\lceil \log \chi(G_S)\rceil \cdot \widehat{a_B})\rceil$ bits de communication. Une telle fonction existe grâce aux propriétés de H et parce que le nombre de sommets de l'hyperarête a_y est au plus $\widehat{a_B}$. Alice envoie ensuite $h(\psi(x))$ à Bob, ce qui nécessite $\lceil \log(4(\widehat{a_B})^2)\rceil$ bits de communication. Bob utilise alors $h(\psi(x))$ pour calculer $\psi(x)$, et comme tous les noeuds de a_y sont de couleur différente, il peut obtenir la valeur de x. La communication totale est donc $\lceil \log(4\lceil \log \chi(G_S)\rceil \cdot \widehat{a_B})\rceil + \lceil \log(4(\widehat{a_B})^2)\rceil \leq \lceil \log\lceil \log \chi(G_S)\rceil\rceil + 3\lceil \log \widehat{a_B}\rceil + 4$. □

Corollaire 1.27. *Pour tout problème avec données corrélées S,*

$$D^2(S) \leq 4D(S) + 3.$$

Démonstration.

$$\begin{aligned} D^2(S) &\leq \lceil \log \lceil \log \chi(G_S) \rceil \rceil + 3 \lceil \log \widehat{a_B} \rceil + 4 && \text{(lemme 1.26)} \\ &= \lceil \log D^1(S) \rceil + 3 \lceil \log \widehat{a_B} \rceil + 4 && \text{(lemme 1.19)} \\ &\leq D(S) - 1 + 3 \lceil \log \widehat{a_B} \rceil + 4 && \text{(lemme 1.23)} \\ &\leq D(S) - 1 + 3D(S) + 4 && \text{(lemme 1.17)} \\ &= 4D(S) + 3. \end{aligned}$$

□

Le corollaire 1.27 a été démontré par Orlitsky [35], qui par la suite a réussi à obtenir $D^2(S) \leq 4D(S) + 2$ [38]. Le lemme 1.28 nous permet d'améliorer ce résultat pour les problèmes de communication dont la complexité est assez grande.

Lemme 1.28 (Mercier 2002). *Soit S un problème avec données corrélées dont la complexité de la communication dépend de n, la taille de la chaîne d'Alice. Pour tout $c \in \mathbb{N}$, il existe un $n_0 \in \mathbb{N}$ tel que pour tout $n \geq n_0$,*

$$D^2(S) \leq 4D(S) - c.$$

Démonstration. Soit G_S l'hypergraphe caractéristique de S. Pour démontrer le lemme 1.26, nous avons utilisé une famille de $k = 4\widehat{a_B} \lceil \log \chi(G_S) \rceil$ fonctions de hachage. Il est possible de diminuer le nombre de fonctions, et par le fait même le nombre de bits de communication, en augmentant la base du logarithme précédent. Posons $k' = \lceil 4\widehat{a_B} \log_b \chi(G_S) \rceil$. Si nous appliquons un protocole similaire à la preuve du lemme 1.26 avec une famille de k' fonctions de hachage, cela entraîne que $D^2(S) \leq \lceil \log \lceil \log \chi(G_S) \rceil \rceil + 3 \lceil \log \widehat{a_B} \rceil + 6 - \log \log b$, et en appliquant la démarche utilisée pour démontrer le corollaire 1.27, il suit

que $D^2(S) \leq 4D(S) + 5 - \log\log b$. En posant $b = 2^{\left(2^{5+c}\right)}$, nous obtenons $D^2(S) \leq 4D(S) - c$.

Le prix à payer est que pour montrer l'existence d'une famille de k' fonctions de hachage possédant les propriétés énoncées au théorème 1.25, il faut que $k' = \lceil 4\widehat{a_B} \log_b \chi(G_S) \rceil > 7,39$, ce qui entraîne qu'il est nécessaire que $\widehat{a_B} \cdot \log \chi(G_S) > 1,585 \cdot 2^{5+c}$. Si S est un problème dont la complexité de la communication dépend de n, alors il existe un n_0 assez grand pour lequel $\widehat{a_B} \cdot \chi(G_S) > 1,585 \cdot 2^{5+c}$ pour tout $n \geq n_0$. □

1.6 En augmentant le nombre de rondes

Nous venons de démontrer que pour tout problème de communication avec données corrélées, il existe un protocole à deux rondes qui est presque optimal. Cette section résume les principaux résultats connus lorsque le nombre de rondes augmente, ainsi que les principaux problèmes ouverts reliés à cette question. Nous n'avons pas jugé bon d'inclure les preuves, car elles sont assez longues, plutôt techniques et ne sont pas utiles pour les chapitres subséquents de ce mémoire.

Une variante du problème de la ligue sportive a été formulée par Orlitsky [36]. Une ligue sportive possède $d \cdot e$ équipes réparties également en d divisions. Les deux meilleures équipes de chaque division participent aux séries éliminatoires, et toutes ces équipes sont connues de Bob. Alice, de son côté, connaît l'identité de l'équipe championne de la ligue. Le but est que Bob apprenne l'identité de l'équipe gagnante. Soit S l'ensemble de support de ce problème avec données corrélées. Considérons le protocole à trois rondes suivant : Alice envoie la division dans laquelle évolue l'équipe gagnante, ce qui nécessite $\lceil \log d \rceil$ bits de communication.

Bob considère ensuite les deux équipes de la division reçue par Alice participant aux séries éliminatoires. Il envoie à Alice une position pour laquelle les deux chaînes diffèrent, ce qui requiert $\lceil \log e \rceil$ bits de communication. Finalement, Alice envoie la valeur du bit de l'équipe gagnante à la position demandée. La complexité du protocole est donc $D^3(S) \leq \lceil \log d \rceil + \lceil \log e \rceil + 1$.

Orlitsky [36] a démontré que lorsque e et d sont choisis judicieusement, tout protocole à deux rondes nécessite plus de communication que le protocole à trois rondes[5]. En fait, il a démontré le résultat suivant :

Théorème 1.29. *Pour tout $\epsilon > 0$ et pour tout $c \geq 0$, il existe un problème avec données corrélées S tel que*

$$D^2(S) \geq (2 - \epsilon) D^3(S) \geq c.$$

Ce résultat et le corollaire 1.27 permettent de poser la question suivante :

Problème ouvert 1.30. Quel est le rapport maximal entre $D^2(S)$ et $D(S)$?

Zhang et Xia [56] et Ahlswede, Cai et Zhang [3] ont ensuite démontré que trois messages n'étaient pas optimaux :

Théorème 1.31. *Pour tout $\epsilon > 0$ et pour tout $c \geq 0$, il existe un problème de communication avec données corrélées S tel que*

$$D^3(S) \geq (2 - \epsilon) D^4(S) \geq c.$$

Zhang et Xia ont également émis la conjecture que $D^4(S) \leq D(S) + o(D(S))$, mais sans parvenir à la démontrer.

[5] Le problème considéré par Orlitsky est légèrement différent de celui présenté ici.

Problème ouvert 1.32. Existe-t-il un k tel que $D^k(S) \leq D(S) + o(D(S))$?

Nous avons essayé de résoudre cette conjecture en généralisant le problème de la ligue sportive, mais nous n'avons pas réussi à obtenir des résultats qui valent la peine d'être mentionnés ici.

Chapitre 2

Complexité déterministe amortie

La résolution simultanée de plusieurs exemplaires[1] d'un problème est parfois plus efficace du point de vue de la communication que la résolution séquentielle des exemplaires de façon optimale. En fait, cela peut même être vrai pour des problèmes différents. Ce phénomène contraire à l'intuition s'applique autant à la communication interactive que non interactive et est appelé le problème de la somme directe[2].

Le problème de la somme directe en complexité de la communication a été introduit par Karchmer, Raz et Wigderson [25] comme une approche prometteuse pour séparer $\mathcal{NC}^1$ de $\mathcal{NC}^2$. Il a été abondamment étudié (consulter entre autres [16, 24, 23]) et pourrait certainement faire l'objet d'un mémoire à lui seul. Dans ce chapitre, nous nous contentons de présenter les principales notions reliées au problème de la somme directe pour les protocoles de communication avec données corrélées (consulter également [4, 5]). Les nouveaux résultats que nous avons obtenus sont les théorèmes 2.8, 2.11, 2.13 et 2.18, les corollaires 2.9, 2.14 et 2.15,

[1] Nous supposons toujours que tous les exemplaires sont indépendants.
[2] «Direct-sum problem» en anglais

la conjecture 2.24 ainsi que toute la section 2.4.

2.1 Définitions

Définition 2.1. Notons $D(S,T)$ la *complexité de la communication déterministe simultanée* des problèmes avec données corrélées S et T, et $D(f,g)$ la complexité de la communication déterministe simultanée des fonctions f et g.

Définition 2.2. Notons $D(S^l)$ la complexité de la communication déterministe simultanée de l exemplaires d'un problème avec données corrélées S, et $D(f^l)$ la complexité de la communication déterministe simultanée de l exemplaires d'une fonction f.

Lorsque plusieurs exemplaires d'un même problème sont résolus simultanément, il peut être utile d'utiliser une mesure de complexité qui représente la complexité de la communication moyenne par exemplaire.

Définition 2.3. La *complexité de la communication déterministe amortie* de S, notée $\overline{D}(S)$, est donnée par l'expression

$$\overline{D}(S) \stackrel{\text{déf}}{=} \lim_{l\to\infty} \frac{1}{l} D(S^l).$$

Il n'est pas difficile de voir que la limite existe par sous-additivité, et que $\overline{D}(S) \leq D(S)$, $D(S,T) \leq D(S) + D(T)$ et $D(S^l) \leq l \cdot D(S)$.

Définition 2.4. Soient $S \subseteq X_S \times Y_S$ et $T \subseteq X_T \times Y_T$ des problèmes de communication avec données corrélées dont les hypergraphes caractéristiques sont $G_S = (V_S, A_S)$ et $G_T = (V_T, A_T)$. Notons $G_S \times G_T = (V_S \times V_T, A)$ le produit

des hypergraphes G_S et G_T. Les hyperarêtes de $G_S \times G_T$ sont définies de la façon suivante : pour tout $(y_s, y_t) \in Y_S \times Y_T$, il y a une hyperarête $A(y_s, y_t) = \{(x_s, x_t) \mid (x_s, y_s) \in S \wedge (x_t, y_t) \in T\}$.

Définition 2.5. Appelons G_S^l le produit, au sens de la définition 2.4, de l copies de l'hypergraphe G_S.

2.2 Communication non interactive

Lorsque plusieurs problèmes doivent être résolus de façon non interactive, il peut être avantageux de les résoudre simultanément plutôt que de les résoudre de façon séquentielle.

Lemme 2.6. *Soient $S_1, S_2, \ldots, S_l$ des problèmes de communication avec données corrélées. Alors*

$$D^1(S_1, S_2, \ldots, S_l) = \lceil \log \chi(G_{S_1} \times G_{S_2} \times \cdots \times G_{S_l}) \rceil.$$

Démonstration. Il s'agit d'appliquer la preuve du lemme 1.19 à l'entrée $((x_{s_1}, x_{s_2}, \ldots, x_{s_l}), (y_{s_1}, y_{s_2}, \ldots, y_{s_l}))$. □

Corollaire 2.7. *Pour tout problème de comunication avec données corrélées S,*

$$D^1(S^l) = \lceil \log \chi(G_S^l) \rceil.$$

Démonstration. Découle directement du lemme précédent. □

L'intérêt du corollaire 2.7 est dû au fait qu'il existe des graphes G tels que $\chi(G^l) < (\chi(G))^l$. En fait, ce corollaire a été démontré par Witsenhausen [53]

dans le but d'analyser un graphe célèbre présenté par Shannon [47] en relation avec la capacité sans erreur de canaux de communication. Nous présentons ce graphe comme exemple d'introduction. Alice et Bob possèdent respectivement des chaînes $x \in \mathbb{Z}_5$ et $y \in \mathbb{Z}_5$ avec $y \equiv x \pmod 5$ ou $y \equiv x+1 \pmod 5$, et Bob veut apprendre la valeur de x. L'ensemble de support de ce problème avec données corrélées est donc $S = \{(x,y) \mid y \equiv x+1 \pmod 5 \vee y \equiv x \pmod 5\}$. Il n'est pas difficile de voir que G_S est un pentagone et que $\chi(G_S) = 3$. Or, avec un peu de travail et beaucoup de patience, on peut montrer que $\chi(G_S^2) = 5 < 9$. Par conséquent, nous pouvons appliquer le corollaire 2.7 et obtenir que $D^1(S) = 2$ et que $D^1(S^2) = 3$. La résolution simultanée des deux exemplaires permet de sauver un bit de communication.

Il est important de remarquer que les économies possibles dépendent de la structure des graphes. Par exemple, si G_S est un graphe complet de k sommets, alors G_S^l est un graphe complet de k^l sommets. Autrement dit, lorsque $S = X \times Y$, il n'y a rien à gagner à résoudre simultanément les exemplaires d'un problème.

Nous pouvons également utiliser le lemme 2.6 pour résoudre simultanément des problèmes différents. Soit $T = \{(x,y) \mid y \equiv x+1 \pmod 3 \vee y \equiv x \pmod 3\}$, où $x, y \in \mathbb{Z}_3$. G_T est un triangle, et donc $\chi(G_T) = 3$ et $\chi(G_S) \cdot \chi(G_T) = 9$. Or, nous pouvons montrer que $\chi(G_S \times G_T) = 8$, ce qui entraîne que $D^1(G_S \times G_T) = 3 < D^1(G_S) + D^1(G_T) = 4$. La résolution simultanée des deux problèmes différents permet de sauver un bit de communication.

Le prochain théorème limite la communication qui peut être sauvée en résolvant simultanément deux problèmes avec données corrélées. Nous l'avons démontré indépendamment à partir des résultats obtenus par Witsenhausen [53] et Linial et Vazirani [28], mais le résultat a déjà été publié par Feder, Kushilevitz,

Naor et Nisan [16].

Théorème 2.8 (Feder, Kushilevitz, Naor, Nisan 1995 [16] ; Mercier 2002). *Soient S et T deux problèmes avec données corrélées dont les hypergraphes caractéristiques sont respectivement G_S et G_T, $|G_S| \leq |G_T|$. Alors*

$$D^1(S,T) \geq D^1(S) + D^1(T) - \log\log|G_S| - 4.$$

Démonstration. Linial et Vazirani [28] ont montré que pour deux graphes[3] quelconques G et H avec $|G| \leq |H|$, $\chi(G \times H) \geq \frac{(\chi(G)-1)\chi(H)}{\ln|G|}$. En appliquant ce résultat à G_S et G_T et en utilisant le lemme 2.6, nous obtenons :

$$\begin{aligned} D^1(S,T) &\geq \left\lceil \log\left(\frac{(\chi(G_S)-1)\chi(G_T)}{\ln|G_S|}\right)\right\rceil \\ &\geq \log(\chi(G_S)-1) + \log\chi(G_T) - \log\ln|G_S| \\ &\geq \lceil\log(\chi(G_S))\rceil + \lceil\log\chi(G_T)\rceil - \log\ln|G_S| - 3 \\ &\geq D^1(S) + D^1(T) - \log\log|G_S| - 4. \end{aligned}$$

□

Corollaire 2.9 (Feder, Kushilevitz, Naor, Nisan 1995 [16] ; Mercier 2002). *Pour tout problème de communication avec données corrélées S dont l'hypergraphe caractéristique est G_S,*

$$D^1(S^2) \geq 2D^1(S) - \log\log|G_S| - 4.$$

[3] Le résultat est également valide pour les hypergraphes.

Pour tout problème avec données corrélées dont les entrées sont de longueur n, le gain maximal possible pour la résolution de deux exemplaires simultanément est donc un terme additif de $\log n$ bits par rapport à la résolution séquentielle des exemplaires. Linial et Vazirani [28] ont montré qu'il existe des graphes de n sommets pour lesquels cette borne pouvait être atteinte.

Corollaire 2.10. *Pour tout problème de communication avec données corrélées S dont l'hypergraphe caractéristique est G_S,*

$$D^1(S) \geq \overline{D^1}(S) \geq D^1(S) - \log\log|G_S| - 4.$$

Démonstration. Le première inégalité est triviale. Pour la deuxième inégalité, un résultat similaire a été démontré par récurrence sur l par Feder, Kushilevitz, Naor et Nisan [16]. □

Ce dernier résultat signifie que pour un problème de communication avec données corrélées S dont les entrées sont de longueur n, le gain maximal possible pour la résolution simultanée de plusieurs exemplaires est un terme additif de $\log n + 4$ bits par exemplaire.

2.3 Communication interactive

Dans cette section, nous analysons le problème de la somme directe pour les protocoles interactifs. Encore une fois, il est parfois possible de réduire la communication en résolvant simultanément plusieurs exemplaires d'un même problème ou même plusieurs problèmes différents. La communication maximale qui peut être économisée n'est toutefois pas connue ; c'est d'ailleurs une des principales

questions ouvertes en complexité de la communication.

Théorème 2.11 (Mercier 2002). *Soient $S_1, S_2, \ldots, S_l$ des problèmes de communication avec données corrélées. Alors*

$$D^2(S_1, S_2, \ldots, S_l) \in O\left(l \cdot \log \max_{1\leq i\leq l}(\widehat{a_B}(S_i)) + \log l \cdot \log\log \max_{1\leq i\leq l}(\chi(G_{S_i}))\right).$$

Démonstration. Alice possède des chaînes $x_1 \in X_1, \ldots, x_l \in X_l$ et Bob des chaînes $y_1 \in Y_1, \ldots, y_l \in Y_l$ avec la restriction que $(x_i, y_i) \in S_i$ pour $1 \leq i \leq l$. Soient $G_{S_1}, G_{S_2}, \ldots, G_{S_l}$ les hypergraphes associés respectivement aux problèmes avec données corrélées $S_1, S_2, \ldots, S_l$. Pour chaque G_{S_i}, Alice et Bob s'entendent sur un coloriage ψ_i avec $\chi(G_{S_i})$ couleurs. Ils s'entendent également sur une famille de fonctions $H = H_{\max_{1\leq i\leq l}(\chi(G_{S_i})), \max_{1\leq i\leq l}(\widehat{a_B}(S_i))}$ possédant les propriétés énoncées au théorème 1.25.

Bob considère les couleurs des sommets des hyperarêtes $a_{y_1}, a_{y_2}, \ldots, a_{y_l}$. Pour chaque a_{y_i}, le nombre de sommets est au plus $\max_{1\leq i\leq l}(\widehat{a_B}(S_i))$. Par le théorème 1.25, il suit que pour chaque a_{y_i}, au moins la moitié des fonctions de H sont injectives sur les couleurs des sommets. Par le lemme A.15, il existe donc une fonction $h_1 \in H$ qui est injective pour au moins la moitié des hyperarêtes $a_{y_1}, a_{y_2}, \ldots, a_{y_l}$. Bob considère ensuite la moitié restante des hyperarêtes pour lesquelles h_1 n'est pas injective. Il trouve une fonction $h_2 \in H$ qui est injective pour au moins la moitié de ces hyperarêtes, et ainsi de suite. De cette façon, Bob trouve $\lceil \log(l+1) \rceil$ fonctions telles que pour toute hyperarête a_{y_i}, au moins une des fonctions est injective sur les couleurs de ses sommets. Bob envoie le nom de ces fonctions à Alice, ce qui nécessite $\lceil \log(l+1) \rceil \lceil \log(4\lceil \log \max_{1\leq i\leq l}(\chi(G_{S_i})) \rceil \cdot \max_{1\leq i\leq l}(\widehat{a_B}(S_i))) \rceil$ bits de communication.

Bob doit également communiquer à Alice quelle fonction h_j doit être utilisée avec chaque a_{y_i}. Comme chaque h_j est injective pour $\frac{1}{2^j}$ des hyperarêtes, il est avantageux de coder les fonctions en unaire, c'est-à-dire d'utiliser la chaîne 1^j pour h_j. Cette étape requiert donc $\sum_{i=1}^{\lceil \log (l+1) \rceil} \frac{l}{2^i} \cdot i \leq 2l-1$ bits de communication pour les fonctions et l zéros servant de séparateurs.

Finalement, lors de la deuxième ronde, Alice envoie $h_j(\psi_i(x_i))$ pour chacun des x_i, ce qui nécessite $l\lceil \log(4(\max_{1\leq i\leq l}(\widehat{a_B}(S_i)))^2) \rceil$ bits de communication. Comme les fonctions h_j sont injectives, Bob peut calculer les valeurs $\psi_i(x_i)$, et puisque tous les noeuds des hyperarêtes a_{y_i} sont de couleurs différentes, il peut en déduire la valeur des x_i. La communication totale est donc :

$$\begin{aligned}
D^2(S_1 + \cdots + S_l) &\leq \lceil \log (l+1) \rceil \cdot \left\lceil \log \left(4 \left\lceil \log \max_{1\leq i\leq l}(\chi(G_{S_i})) \right\rceil \cdot \max_{1\leq i\leq l}(\widehat{a_B}(S_i)) \right) \right\rceil \\
&\quad + 2l - 1 + l + l \left\lceil \log \left(4 \left(\max_{1\leq i\leq l}(\widehat{a_B}(S_i)) \right)^2 \right) \right\rceil \\
&= \lceil \log (l+1) \rceil \left\lceil \log \left\lceil \log \max_{1\leq i\leq l}(\chi(G_{S_i})) \right\rceil \right\rceil \\
&\quad + \lceil \log (l+1) \rceil \left\lceil \log \max_{1\leq i\leq l}(\widehat{a_B}(S_i)) \right\rceil + 2\lceil \log (l+1) \rceil \\
&\quad + 2l \left\lceil \log \max_{1\leq i\leq l}(\widehat{a_B}(S_i)) \right\rceil + 5l - 1 \\
&\in O\left(l \cdot \log \max_{1\leq i\leq l}(\widehat{a_B}(S_i)) + \log l \cdot \log \log \max_{1\leq i\leq l}(\chi(G_{S_i})) \right)
\end{aligned}$$

□

Le corollaire suivant a été initialement démontré par Feder, Kushilevitz, Naor et Nisan [16] et découle directement du théorème 2.11.

Corollaire 2.12. *Pour tout problème de communication avec données corrélées S,*

$$\begin{aligned} D^2(S^l) &\leq \lceil \log(l+1) \rceil \lceil \log \lceil \log \chi(G_S) \rceil \rceil + \lceil \log(l+1) \rceil \lceil \log \widehat{a_B} \rceil \\ &\quad + 2\lceil \log(l+1) \rceil + 2l \lceil \log \widehat{a_B} \rceil + 5l - 1 \\ &\in O(l \cdot \log \widehat{a_B} + \log l \cdot \log \log \chi(G_S)). \end{aligned}$$

Démonstration. Il s'agit de remarquer que $\max_{1 \leq i \leq l}(\widehat{a_B}(S_i)) = \widehat{a_B}$ et que $\max_{1 \leq i \leq l}(\chi(G_{S_i})) = \chi(G_S)$. □

Pour certains cas que nous analysons sous peu, le théorème 2.13 permet de sauver plus de bits de communication que le théorème 2.11.

Théorème 2.13 (Mercier 2002). *Soient $S_1, S_2, \ldots, S_l$ des problèmes de communication avec données corrélées. Alors*

$$D^2(S_1, S_2, \ldots, S_l) \in O\left(l \log \sum_{i=1}^{l} \widehat{a_B}(S_i) + \log \log \max_{1 \leq i \leq l}(\chi(G_{S_i})) \right).$$

Démonstration. Alice possède des chaînes $x_1 \in X_1, \ldots, x_l \in X_l$ et Bob des chaînes $y_1 \in Y_1, \ldots, y_l \in Y_l$ avec la restriction que $(x_i, y_i) \in S_i$ pour $1 \leq i \leq l$. Soient $G_{S_1}, G_{S_2}, \ldots, G_{S_l}$ les hypergraphes associés aux problèmes avec données corrélées $S_1, S_2, \ldots, S_l$. Pour chaque G_{S_i}, Alice et Bob s'entendent sur un coloriage ψ_i avec $\chi(G_{S_i})$ couleurs. Alice et Bob s'entendent également sur une famille de fonctions $H = H_{\max_{1 \leq i \leq l}(\chi(G_{S_i})), \sum_{i=1}^{l} \widehat{a_B}(S_i)}$ possédant les propriétés énoncées au théorème 1.25.

Bob considère les couleurs des hyperarêtes $a_{y_1}, a_{y_2}, \ldots, a_{y_l}$. Il choisit une fonction $h \in H$ qui est injective sur ces couleurs et envoie sa description à Alice, ce qui requiert $\lceil \log(4 \lceil \log \max_{1 \leq i \leq l}(\chi(G_{S_i})) \rceil \cdot \sum_{i=1}^{l} \widehat{a_B}(S_i)) \rceil$ bits de communication.

Une telle fonction existe à cause des propriétés de H et parce que le nombre total de sommets des hyperarêtes $a_{y_1}, \ldots, a_{y_l}$ est au plus $\sum_{i=1}^{l} \widehat{a_B}(S_i)$.

Alice envoie ensuite les l valeurs $h(\psi_i(x_i))$ à Bob, ce qui nécessite $l \cdot \lceil \log(4(\sum_{i=1}^{l} \widehat{a_B}(S_i))^2) \rceil$ bits de communication. Comme la fonction h est injective, Bob peut calculer les valeurs $\psi_i(x_i)$, et puisque tous les noeuds des hyperarêtes a_{y_i} sont de couleurs différentes, il peut en déduire la valeur des x_i. La communication totale est donc :

$$\begin{aligned} D^2\left(S_1 + \cdots + S_l\right) &\leq \left\lceil \log\left(4\left\lceil \log \max_{1\leq i\leq l}\left(\chi(G_{S_i})\right)\right\rceil \cdot \sum_{i=1}^{l} \widehat{a_B}(S_i)\right)\right\rceil \\ &\quad + l \cdot \left\lceil \log\left(4\left(\sum_{i=1}^{l} \widehat{a_B}(S_i)\right)^2\right)\right\rceil \\ &= \left\lceil \log\left\lceil \log \max_{1\leq i\leq l}(\chi(G_{S_i}))\right\rceil\right\rceil + \left\lceil \log \sum_{i=1}^{l} \widehat{a_B}(S_i)\right\rceil \\ &\quad + 2l\left\lceil \log \sum_{i=1}^{l} \widehat{a_B}(S_i)\right\rceil + 2l + 2 \\ &\in O\left(\log\log \max_{1\leq i\leq l}(\chi(G_{S_i})) + l \log \sum_{i=1}^{l} \widehat{a_B}(S_i)\right) \end{aligned}$$

□

Corollaire 2.14 (Mercier 2002). *Pour tout problème de communication avec données corrélées S,*

$$\begin{aligned} D^2(S^l) &\leq \lceil \log\lceil \log \chi(G_S)\rceil\rceil + \lceil \log(l \cdot \widehat{a_B})\rceil + 2l\lceil \log(l \cdot \widehat{a_B})\rceil + 2l + 2 \\ &\in O(\log\log \chi(G_S) + l \log l + l \log \widehat{a_B}). \end{aligned}$$

Démonstration. Il s'agit de remarquer que $\sum_{i=1}^{l} \widehat{a_B}(S_i) = l \cdot \widehat{a_B}$ et que $\max_{1 \leq i \leq l}(\chi(G_{S_i})) = \chi(G_S)$. □

Voici un autre corollaire surprenant découlant du théorème 2.13.

Corollaire 2.15 (Mercier 2002). *Soit S un problème de communication avec données corrélées dont l'hypergraphe caractéristique est G_S et pour lequel $\widehat{a_B} \in O(1)$. Si $l \in O(1)$, alors*

$$D(S^l) \leq D(S) + O(1).$$

Démonstration. En utilisant dans l'ordre le corollaire 2.14 ainsi que les lemmes 1.19 et 1.23, nous obtenons :

$$\begin{aligned} D(S^l) &\leq \lceil \log \lceil \log \chi(G_S) \rceil \rceil + \lceil \log(l \cdot \widehat{a_B}) \rceil + 2l \lceil \log(l \cdot \widehat{a_B}) \rceil + 2l + 2 \\ &= \lceil \log D^1(S) \rceil + O(1) \\ &\leq D(S) + O(1) \end{aligned}$$

□

Le tableau 2.1 résume la communication requise pour résoudre l exemplaires d'un problème avec données corrélées S en utilisant les protocoles du lemme 1.26 et des corollaires 2.12 et 2.14. Si $l \in o(\log \log \chi(G_S))$, le protocole du corollaire 2.14 est le plus efficace, tandis que si $l \in \omega(\log \log \chi(G_S))$, le meilleur protocole est celui du corollaire 2.12. Notons que le protocole du corollaire 2.14 est le pire des trois si $l \in \omega(\log \chi(G_S))$, en fait pire encore que la résolution séquentielle des exemplaires. Il est donc essentiel de connaître le nombre d'exemplaires à résoudre avant de choisir un protocole.

TAB. 2.1 – Résolution de l exemplaires d'un problème avec données corrélées S

Protocole	**Communication requise**
Lemme 1.26	$O(l \cdot \log\log \chi(G_S) + l \cdot \log \widehat{a_B})$
Corollaire 2.12	$O(\log l \cdot \log\log \chi(G_S) + l \cdot \log \widehat{a_B})$
Corollaire 2.14	$O(\log\log \chi(G_S) + l \log \widehat{a_B} + l \log l)$

TAB. 2.2 – Résolution de plusieurs exemplaires du problème de la ligue sportive

Nombre d'exemplaires	**Résolution séquentielle**	**Corollaire 2.12**	**Corollaire 2.14**
16	$16\lceil \log n \rceil + 16$	$5\lceil \log n \rceil + 126$	$\lceil \log n \rceil + 199$
$\boldsymbol{\Theta(\log\log n)}$	$\Theta(\log n \log\log n)$	$O(\log n \log\log\log n)$	$O(\log n)$
$\boldsymbol{\Theta(\log n)}$	$\Theta(\log^2 n)$	$O(\log n \log\log n)$	$O(\log n \log\log n)$
$\boldsymbol{\Theta(n)}$	$\Theta(n \log n)$	$O(n)$	$O(n \log n)$

Quant au tableau 2.2, il analyse la performance des trois protocoles pour résoudre plusieurs exemplaires du problème de la ligue sportive. Rappelons que pour ce problème, $\widehat{a_B} = 2$ et $\chi(G_S) = 2^n$. Deux résultats méritent d'être soulignés. Premièrement, le corollaire 2.15 implique que $D(S^l) \leq D(S) + O(1)$ lorsque $l \in O(1)$. Deuxièmement, le corollaire 2.12 nous permet de déduire que la complexité de la communication amortie est constante. En fait, nous verrons à la section 2.5 que $\overline{D}(S) = 1$ pour le problème de la ligue sportive.

2.4 Le problème de la ligue sportive, prise 2

Les théorèmes 2.11 et 2.13 peuvent être utilisés pour résoudre simultanément plusieurs problèmes avec données corrélées différents plus efficacement que s'ils étaient résolus séparément de façon optimale. Dans cette section, nous construisons une famille de problèmes différents pour lesquels la communication peut

être réduite lorsqu'ils sont résolus simultanément. À notre connaissance, une telle famille n'avait jamais été contruite.

Le problème de la ligue sportive défini à la section 1.3 est modifié de la façon suivante : Bob connaît les k équipes participant aux séries éliminatoires de la ligue et veut apprendre d'Alice le nom de l'équipe gagnante. Appelons L_k le problème de la ligue sportive ayant $k+1$ équipes participant aux séries éliminatoires, L_1 étant le problème initial.

Lemme 2.16. *$D(L_k) \in \Theta(\log n)$ lorsque $k \in O(\log n)$.*

Démonstration. Soit G_{L_k} l'hypergraphe caractéristique de L_k. Il est clair que $\chi(G_{L_k}) = 2^n$ et que $\widehat{a_B}(L_k) = k+1$. Le lemme 1.19 implique que $D^1(L_k) = n$, et il suit par le lemme 1.23 que $D(L_k) \geq \lceil \log n \rceil + 1$. Comme $D(L_k) \leq D^2(L_k) \leq \lceil \log n \rceil + 3\lceil \log(k+1) \rceil + 4$ par le lemme 1.26, il suit que $D(L_k) \in \Theta(\log n)$ lorsque $k \in O(\log n)$. □

Les deux exemples suivants montrent qu'il est possible d'avoir $D(L_1, L_2, \ldots, L_l) < D(L_1) + D(L_2) + \ldots D(L_l)$.

- Si $l \in O(1)$, alors $D(L_1) + D(L_2) + \cdots + D(L_k) \geq l\lceil \log n \rceil$ par les lemmes 1.19 et 1.23, alors que par le théorème 2.13, nous obtenons que $D(L_1, L_2, \ldots, L_k) \leq \lceil \log n \rceil + O(1)$.
- Si $l \in \Theta(\log n)$, alors $D(L_1) + D(L_2) + \cdots + D(L_k) \in \Theta(\log^2 n)$ bits de communication par le lemme 2.16, alors que les théorèmes 2.11 et 2.13 impliquent que $D(L_1, L_2, \ldots, L_k) \in O(\log n \log\log n)$.

2.5 Quatre rondes sont optimales

Dans cette section, nous montrons que quatre rondes de communication sont optimales pour la complexité de la communication déterministe amortie.

Le théorème 2.17 a été démontré par Naor, Orlitsky et Shor [33]. La démonstration, qui n'est pas présentée ici, n'est pas très complexe et utilise encore une fois une famille de fonctions de hachage.

Théorème 2.17. *Soit S un problème de communication dont l'hypergraphe caractéristique G_S possède $a(G_S)$ hyperarêtes. Alors*

$$D^4(S) \leq \log\log a(G_S) + \log \widehat{a_B} + 3\log\log \widehat{a_B} + 7.$$

Le théorème 2.17 peut être utilisé pour calculer plusieurs exemplaires d'un même problème, ou encore plusieurs problèmes différents, comme en font foi le théorème et le corollaire suivants :

Théorème 2.18 (Mercier 2002). *Soient $S_1, S_2, \ldots, S_l$ des problèmes de communication avec données corrélées. Alors*

$$D^4(S_1, S_2, \ldots, S_L) \leq \log\log \prod_{i=1}^{l} a(G_{S_i}) + \log \prod_{i=1}^{l} \widehat{a_B}(S_i) + 3\log\log \prod_{i=1}^{l} \widehat{a_B}(S_i) + 7.$$

Démonstration. Soit $G_S = G_{S_1} \times G_{S_2} \times \cdots \times G_{S_l}$ l'hypergraphe caractéristique des problèmes $(S_1, S_2, \ldots, S_l)$ tel qu'introduit à la définition 2.4. Il n'est pas difficile de voir que $a(G_S) = a(G_{S_1}) \times a(G_{S_2}) \times \cdots \times a(G_{S_l})$ et que $\widehat{a_B}(S) = \widehat{a_B}(S_1) \times \widehat{a_B}(S_2) \times \cdots \times \widehat{a_B}(S_l)$. En appliquant le théorème précédent à l'hypergraphe G_S, nous obtenons le résultat souhaité. □

Corollaire 2.19. *Pour tout problème de communication non trivial avec données corrélées S,*

$$D^4(S^l) \leq l \log \widehat{a_B} + 4 \log l + \log \log a(G_S) + 3 \log \log \widehat{a_B} + 7.$$

Démonstration. Il s'agit d'appliquer directement le théorème 2.18 en remarquant que $\prod_{i=1}^{l} \widehat{a_B}(S_i) = (\widehat{a_B})^l$ et que $\prod_{i=1}^{l} a(G_{S_i}) = (a(G_S))^l$. □

Ce corollaire nous permet de démontrer que quatre rondes de communication sont optimales pour la complexité de la communication amortie déterministe, résultat qui a été démontré par Naor, Orlitsky et Shor [33].

Corollaire 2.20. *Pour tout problème de communication non trivial S avec données corrélées,*

$$\overline{D^4}(S) = \overline{D^5}(S) = \cdots = \overline{D}(S) = \log \widehat{a_B}.$$

Démonstration. Pour la borne supérieure, le corollaire 2.19 entraîne que

$$\overline{D^4}(S) = \lim_{l \to \infty} \frac{D^4(S^l)}{l} \leq \log \widehat{a_B}.$$

Pour la borne inférieure, nous pouvons utiliser le lemme 1.17 et obtenir que $D(S) \geq \lceil \log \widehat{a_B}(S^l) \rceil \geq l \log \widehat{a_B}$. Par conséquent,

$$\overline{D^4}(S) \geq \overline{D}(S) = \lim_{l \to \infty} \frac{D(S^l}{l} \geq \log \widehat{a_B}.$$

□

En comparant le corollaire 2.20 et le lemme 1.18, nous obtenons directement le corollaire 2.21.

Corollaire 2.21. *La complexité de la communication amortie déterministe d'un problème avec données corrélées correspond exactement à la complexité de la communication déterministe lorsqu'Alice connaît la chaîne de Bob.*

Une remarque à propos du corollaire 2.21 est que c'est uniquement lorsqu'il y a une différence appréciable entre la complexité de la communication lorsqu'Alice ne connaît pas la chaîne de Bob et la complexité de la communication lorsqu'elle la connaît qu'il peut être avantageux de résoudre simultanément plusieurs exemplaires d'un problème avec données corrélées. Nous verrons au chapitre 4 que pour les applications comme la réconciliation de fichiers, la comparaison de séquences de nucléotides ou la distribution de clés secrètes, ce n'est malheureusement pas le cas.

2.6 Discussion

Plusieurs nouveaux résultats ont été présentés dans les sections précédentes, mais malheureusement, ils ne permettent pas d'améliorer beaucoup la compréhension des principales questions ouvertes reliées au problème de la somme directe en complexité de la communication. La présente section traite de deux de ces questions ouvertes en lien avec les résultats de ce chapitre. Nous énonçons également une nouvelle conjecture.

Problème ouvert 2.22. Existe-t-il des fonctions f et g telles que $D(f,g) < D(f) + D(g)$?

Il ne semble pas possible d'adapter les résultats obtenus pour les problèmes avec données corrélées afin de résoudre le problème de la somme directe pour le

modèle original de Yao. En effet, passer d'un protocole pour un problème avec données corrélées S à un protocole pour une fonction n'est pas aussi simple qu'il n'y paraît. Soit $\mathcal{P}$ un protocole pour S nécessitant $D(S)$ bits de communication. Si on simule $\mathcal{P}$ sur des données non corrélées jusqu'à ce que $D(S)$ bits aient été communiqués, la fonction calculée est la suivante :

$$f(x,y) = \begin{cases} x \text{ si } (x,y) \in S \\ \text{n'importe quoi sinon} \end{cases}$$

Si on n'exige pas qu'Alice apprenne la valeur de f, tous les résultats obtenus en résolvant simultanément plusieurs exemplaires de S (ou plusieurs problèmes différents) sont également valides pour f. Le problème est que f, malgré le fait que son domaine soit $X \times Y$, n'est pas une «fonction» mais plutôt une relation. Une autre particularité de f est que si $D(S) < n$ [4], alors Bob est condamné à ne pas savoir avec certitude si la sortie qu'il a obtenue est x. S'il pouvait différencier x de la sortie «n'importe quoi», il pourrait vérifier l'égalité de deux chaînes avec certitude avec moins de n bits de communication, ce qui est impossible (démontré par Yao [54]).

Lorsqu'un seul exemplaire d'un problème est résolu, il est possible de remplacer la partie «n'importe quoi sinon» par quelque chose de plus tangible et d'obtenir une vraie fonction. Par exemple, pour le problème original de la ligue sportive, «n'importe quoi sinon» pourrait être remplacé par «y tel que $y \in \{y_1, y_2\}$ et le i-ième bit de y est égal au i-ième bit de x, i étant la première position pour laquelle y_1 et y_2 diffèrent». Malheureusement, cette astuce ne fonctionne pas si nous voulons résoudre simultanément plusieurs exemplaires d'un problème

[4] Nous supposons que $X \subseteq \{0,1\}^n$.

à l'aide des fonctions de hachage utilisées précédemment.

Conjecture 2.23. $\overline{D}(S) \geq D(S) - O(\log n)$.

Il n'est pas difficile de voir que les lemmes 2.12 et 2.14 ne permettent pas de sauver plus de $\log n$ bits de communication par exemplaire, économie qui peut être atteinte pour le problème de la ligue sportive (voir la fin de la section 1.4). De plus, comme ces lemmes sont assez astucieux, ils laissent présager qu'il n'est pas possible de faire mieux. Il semble donc qu'il faille travailler sur une borne supérieure pour le nombre de bits qui peuvent être sauvés lorsque plusieurs exemplaires d'un problème avec données corrélées sont résolus simultanément. Dans le cas des fonctions, Feder, Kushilevitz, Naor et Nisan [16] ont montré que $D(f) \geq \overline{D}(f) \geq \sqrt{D(f)/2} - \log n - O(1)$. Malheureusement, leur preuve utilise la complexité de la communication non déterministe et ne peut pas être généralisée aux relations ou aux problèmes avec données corrélées.

Conjecture 2.24 (Mercier 2002). La différence entre la complexité de la communication lorsqu'Alice ne connaît pas la chaîne de Bob et la complexité de la communication lorsqu'elle la connaît est dans $O(\log n)$ bits.

Cette nouvelle conjecture est équivalente à la conjecture 2.23 par le corollaire 2.21, mais elle est toutefois plus naturelle et est selon nous une autre façon d'attaquer la problème de la somme directe qui mérite d'être étudiée.

Chapitre 3

Complexité probabiliste

Dans ce chapitre, nous analysons la complexité de la communication probabiliste de données corrélées. Dans ce modèle, Alice et Bob sont autorisés à «tirer à pile ou face» et à considérer le résultat des tirages pour décider des messages à envoyer. Cela implique que les bits de communication ainsi que la réponse de Bob ne sont plus fixés par l'entrée (x, y), mais deviennent plutôt des variables aléatoires.

Le complexité de la communication probabiliste de problèmes avec données corrélées n'a jamais été systématiquement étudiée auparavant, et c'est ce que nous faisons dans ce chapitre. Nous avons obtenu plusieurs nouveaux résultats, dont certains contrastent avec le modèle original de Yao. En plus de classifier presque entièrement le modèle probabiliste, notre contribution la plus intéressante est la démonstration que pour plusieurs classes de problèmes, les modèles probabiliste et déterministe sont équivalents du point de vue de la communication. Nous émettons la conjecture que cette propriété est également valide pour tous les problèmes avec données corrélées et montrons qu'elle permet de résoudre le

problème de la somme directe.

3.1 Générateurs aléatoires privés - définitions

Dans le modèle probabiliste de communication de données corrélées, Alice possède une chaîne $x \in X$ et Bob possède une chaîne $y \in Y$ avec la restriction que $(x, y) \in S$, et encore une fois le but est que Bob apprenne la valeur de x. La différence avec le modèle déterministe est qu'Alice et Bob possèdent respectivement des chaînes aléatoires indépendantes finies c_A et c_B de longueur arbitraire. Toute combinaison de $(x, y) \in R$, c_A et c_B détermine une feuille de l'arbre du protocole. Il est donc possible que pour une entrée (x, y), le protocole retourne des valeurs différentes pour différentes valeurs de c_A et c_B. La notation que nous utilisons est similaire à celle utilisée par Kushilevitz et Nisan [27].

Définition 3.1. Soient $f : X \times Y \to \{0, 1\}$ une fonction et $\mathcal{P}$ un protocole probabiliste pour lequel Alice possède une chaîne aléatoire c_A et Bob une chaîne aléatoire c_B.

- $\mathcal{P}$ calcule une fonction f *sans erreur* si, pour toute paire (x, y),

$$\Pr[\mathcal{P}(x, y) = f(x, y)] = 1.$$

- $\mathcal{P}$ calcule une fonction f avec *erreur* ϵ si, pour toute paire (x, y),

$$\Pr[\mathcal{P}(x, y) = f(x, y)] \geq 1 - \epsilon.$$

- $\mathcal{P}$ calcule une fonction f avec *erreur unilatérale*[1] ϵ si, pour toute paire

[1] «One-sided error» en anglais.

(x, y) telle que $f(x, y) = 0$,

$$\Pr[\mathcal{P}(x, y) = 0] = 1,$$

et pour toute paire (x, y) telle que $f(x, y) = 1$,

$$\Pr[\mathcal{P}(x, y) = 1] \geq 1 - \epsilon.$$

Il est important de remarquer que toutes les probabilités de la définition 3.1 sont distribuées sur les choix aléatoires de c_A et c_B et non sur les entrées x et y. Nous analysons cette dernière variante à la section 3.6.

Définition 3.2. La *communication en pire cas* d'un protocole probabiliste $\mathcal{P}$ sur l'entrée (x, y) est le nombre maximal de bits communiqués pour n'importe quel choix des chaînes aléatoires c_A et c_B. Le *coût en pire cas* de $\mathcal{P}$ est le maximum, pour toutes les entrées (x, y), de la communication en pire cas de $\mathcal{P}$ sur (x, y).

Définition 3.3. La *communication moyenne* d'un protocole probabiliste $\mathcal{P}$ sur l'entrée (x, y) est le nombre espéré de bits communiqués pour tous les choix des chaînes aléatoires c_A et c_B. Le *coût moyen* de $\mathcal{P}$ est le maximum, pour toutes les entrées (x, y), de la communication moyenne de $\mathcal{P}$ sur (x, y).

Définition 3.4. Soit $f : X \times Y \to \{0, 1\}$ une fonction.

- La *complexité de la communication probabiliste sans erreur* de f, notée $R_0(f)$, est le coût moyen minimal d'un protocole probabiliste qui calcule f sans erreur[2].
- La *complexité de la communication probabiliste avec erreur* ϵ de f, notée

[2]La lettre R est utilisée pour «randomized communication complexity».

$R_\epsilon(f)$, est le coût en pire cas minimal d'un protocole probabiliste qui calcule f avec erreur ϵ, pour $0 < \epsilon < 1/2$.

- La *complexité de la communication probabiliste avec erreur unilatérale* ϵ de f, notée $R_{\epsilon,uni}(f)$, est le coût en pire cas minimal d'un protocole probabiliste qui calcule f avec erreur unilatérale ϵ, pour $0 < \epsilon < 1$.
- Nous écrivons $R_0^k(f)$, $R_\epsilon^k(f)$ et $R_{\epsilon,uni}^k(f)$ lorsque le nombre de rondes est limité à k.

Définition 3.5. Soit S un problème avec données corrélées.

- La *complexité de la communication probabiliste sans erreur* de S, notée $R_0(S)$, est le coût moyen minimal d'un protocole probabiliste qui calcule S sans erreur.
- La *complexité de la communication probabiliste avec erreur* ϵ de S, notée $R_\epsilon(S)$, est le coût en pire cas minimal d'un protocole probabiliste qui calcule S avec erreur ϵ, pour $0 < \epsilon < 1/2$.
- Nous écrivons $R_0^k(S)$ et $R_\epsilon^k(S)$ lorsque le nombre de rondes est limité à k.

$R_\epsilon(S)$ est donc le nombre de bits transmis en pire cas par le meilleur protocole qui, pour toute paire $(x, y) \in R$, permet à Bob d'apprendre la chaîne d'Alice avec probabilité au moins $1 - \epsilon$.

Nous utilisons le coût en pire cas pour les protocoles avec erreur, car cette mesure est généralement plus intéressante à utiliser. De plus, pour ces protocoles, la complexité de la communication en pire cas est à un facteur multiplicatif près de la complexité de la communication moyenne. Par contre, pour les protocoles probabilistes sans erreur, la complexité de la communication en pire cas est égale à la complexité de la communication déterministe, car un protocole déterministe est simplement un protocole pour lequel les chaînes c_A et c_B ont été fixées initia-

lement. Pour les protocoles probabilistes sans erreur, le seul cas intéressant est donc la complexité de la communication moyenne.

3.2 Générateurs aléatoires privés - résultats

Dans cette section, nous présentons plusieurs résultats pour la complexité probabiliste de problèmes de communication avec données corrélées. Commençons par une borne supérieure triviale.

Lemme 3.6 (Mercier 2002).

$$R_\epsilon(S) \leq D(S).$$

Démonstration. Soit $\mathcal{P}$ un protocole déterministe pour S nécessitant $D(S)$ bits de communication. $\mathcal{P}$ peut être considéré comme un protocole probabiliste pour lequel Alice et Bob ne tiennent pas compte des chaînes aléatoires c_A et c_B. □

Essayons maintenant d'obtenir des bornes inférieures intéressantes pour la complexité de la communication probabiliste. Le lemme 3.7 démontre que pour tout problème avec données corrélées, la différence entre la complexité de la communication probabiliste et la complexité de la communication déterministe non interactive est au plus exponentielle.

Lemme 3.7 (Mercier 2002). *Pour tout problème avec données corrélées S,*

$$R_\epsilon(S) \in \Omega(\log D^1(S) - c(\epsilon)),$$

où $c(\epsilon)$ dépend uniquement de ϵ.

Démonstration. Le lemme 3.8 de Kushilevitz et Nisan [27] affirme que pour toute fonction booléenne $f : X \times Y \to \{0, 1\}$,

$$D(f) \leq 2^{R_\epsilon(f)} \left(\log \left(\frac{1}{2} - \epsilon \right)^{-1} + R_\epsilon(f) \right).$$

Sans démontrer ce résultat formellement, mentionnons que l'idée est, étant donné un protocole probabiliste nécessitant $R_\epsilon(f)$ bits de communication, de construire un protocole déterministe dont la complexité est $2^{R_\epsilon(f)} \left(\log \left(\frac{1}{2} - \epsilon \right)^{-1} + R_\epsilon(f) \right)$. Une analyse détaillée de la preuve nous permet de voir qu'elle s'applique aux problèmes avec données corrélées et que le protocole déterministe obtenu est non interactif. Nous obtenons donc :

$$\begin{aligned}
D^1(S) &\leq 2^{R_\epsilon(S)} \left(\log \left(\frac{1}{2} - \epsilon \right)^{-1} + R_\epsilon(S) \right) \\
R_\epsilon(S) &\geq \log D^1(S) - \log \left(\log \left(\frac{1}{2} - \epsilon \right)^{-1} + R_\epsilon(S) \right) \\
R_\epsilon(S) &\geq \log D^1(S) - \log R_\epsilon(S) - c_1(\epsilon) \\
2R_\epsilon(S) &\geq \log D^1(S) - c_1(\epsilon) \\
R_\epsilon(S) &\geq \frac{1}{2} \log D^1(S) - c(\epsilon).
\end{aligned}$$

□

Revenons au problème de la ligue sportive défini à la section 1.3, pour lequel $D(S) = \lceil \log n \rceil + 1$ et $D^1(S) = n$. D'après les lemmes 3.6 et 3.7, nous pouvons conclure que $R_\epsilon(S) \in \Theta(\log n)$ pour toute constante ϵ telle que $0 < \epsilon < \frac{1}{2}$, et donc que les modèles déterministe et probabiliste sont équivalents du point de vue de la communication.

Corollaire 3.8. *Lorsque la différence entre $D^1(S)$ et $D(S)$ est exponentielle, les modèles déterministe et probabiliste sont équivalents du point de vue de la communication, autrement dit $\Theta(D(S)) = \Theta(R_\epsilon(S))$.*

Il est intéressant de noter que les lemmes 3.6 et 3.7 entraînent que $\frac{1}{2} \log D^1(S) - c(\epsilon) \leq R_\epsilon(S) \leq D(S)$, ce qui démontre, comme nous l'avons déjà vu au lemme 1.23, que la différence entre $D^1(S)$ et $D(S)$ est au plus exponentielle.

Pour les problèmes dont la différence entre la complexité déterministe interactive et non interactive est petite, le lemme 3.7 n'exclut pas la possibilité qu'un algorithme permettant de réduire sensiblement la communication puisse exister. Afin d'obtenir une meilleure borne inférieure pour ces problèmes, nous utilisons le fait que Bob doit apprendre la chaîne d'Alice et non pas uniquement calculer une fonction.

Lemme 3.9 (Mercier 2002). *Pour tout problème de communication avec données corrélées S,*

$$R_\epsilon(S) \geq \lceil \log \widehat{a_B} \rceil.$$

Démonstration. Supposons que $R_\epsilon(S) \leq \lceil \log \widehat{a_B} \rceil - 1$. Par définition, il existe un protocole $\mathcal{P}$ pour S nécessitant au plus $\lceil \log \widehat{a_B} \rceil - 1$ bits de communication et tel que pour toute paire $(x, y) \in S$, la probabilité que Bob n'obtienne pas la valeur de x est au plus ϵ.

Soit y une chaîne telle que $|a_B(y)| = \widehat{a_B}$. Comme $\mathcal{P}$ requiert au plus $\lceil \log \widehat{a_B} \rceil - 1$ bits de communication, cela veut dire qu'il existe deux chaînes $x_1, x_2 \in a_B(y)$, $x_1 \neq x_2$, pour lesquelles la communication entre Alice et Bob est la même. Il suit que peu importe la stratégie de Bob, il existe une chaîne $x_i \in \{x_1, x_2\}$ pour laquelle la probabilité d'erreur de $\mathcal{P}$ sur l'entrée (x_i, y) est au moins $\frac{1}{2}$, ce qui est une contradiction. □

Nous pouvons remarquer que la borne inférieure du lemme 3.9 correspond à la complexité déterministe amortie (voir le corollaire 2.20), ou encore à la complexité lorsqu'Alice connaît la chaîne de Bob (voir le lemme 1.18). Une autre constatation est que cette borne est très mauvaise pour les problèmes pour lesquels la différence entre la complexité de la communication déterministe interactive et non interactive est grande. Prenons par exemple le problème de la ligue sportive, pour lequel $\widehat{a_B} = 2$. La borne du lemme 3.9 ne donne pas mieux que $R_\epsilon(S) \geq 1$, ce qui est loin de la borne obtenue par le lemme 3.7. Par contre, pour une classe de problèmes avec une petite différence entre $D^1(S)$ et $D(S)$, nous verrons à la section 4.3 que la borne donnée par le lemme 3.9 peut être atteinte, et qu'encore une fois les modèles déterministe et probabiliste sont équivalents.

Le prochain théorème résume les résultats démontrés depuis le début de cette section.

Théorème 3.10 (Mercier 2002). *Pour tout problème de communication avec données corrélées S,*

$$\max\left(\lceil \log \widehat{a_B} \rceil, \frac{1}{2} \log D^1(S) - c(\epsilon)\right) \leq R_\epsilon(S) \leq D(S).$$

Démonstration. Découle directement des lemmes 3.6, 3.7 et 3.9. □

Bien que le modèle probabiliste ne semble pas permettre de diminuer le nombre de bits de communication (du moins c'est ce que nous pensons), il permet par contre d'obtenir des protocoles non interactifs efficaces. Le théorème 3.11 permet d'obtenir une borne supérieure pour la complexité de la communication probabiliste non interactive.

Théorème 3.11. *Pour tout problème de communication avec données corrélées S,*

$$R^1_\epsilon(S) \leq 4D(S) + \left\lceil 2 \log \frac{1}{\epsilon} \right\rceil .$$

Démonstration. Nous pensons avoir une preuve très élégante du résultat, mais comme elle contient une petite faille que nous n'avons pas encore réussi à corriger, nous ne pouvons malheureusement pas l'inclure ici. Voir [35] pour la démonstration originale.

□

3.3 Générateur aléatoire public - définitions

Pour le modèle probabiliste que nous avons considéré jusqu'à maintenant, Alice et Bob ont chacun leur pièce de monnaie. Bob ne peut pas voir les résultats des tirages d'Alice, et vice-versa. Dans cette section, nous supposons qu'Alice et Bob ont accès à une pièce de monnaie «publique». Ce modèle est appelé modèle probabiliste avec générateur aléatoire public, ou plus simplement *modèle probabiliste public*. Formellement, Alice et Bob possèdent une chaîne aléatoire commune c obéissant à une distribution de probabilité Π. Les bits de communication envoyés par Alice dépendent de x et de c, tandis que ceux de Bob dépendent de y et de c.

Un protocole probabiliste avec générateur aléatoire public peut également être vu comme une distribution $\{\mathcal{P}_c\}_{c\in\Pi}$ de protocoles déterministes. Alice et Bob choisissent conjointement une chaîne c indépendamment de l'entrée (x, y) et exécutent ensuite le protocole déterministe $\mathcal{P}_c$. La probabilité de succès d'un tel protocole sur l'entrée $(x, y) \in S$ est la probabilité de choisir un protocole déterministe, selon la distribution de probabilité Π, qui calcule S correctement.

Définition 3.12. Soit $f : X \times Y \to \{0, 1\}$ une fonction. La *complexité de la communication probabiliste publique avec erreur* ϵ de f, notée $R_{\epsilon,pub}(f)$, est le coût minimal d'un protocole avec générateur aléatoire public qui calcule f avec une probabilité d'erreur bornée par ϵ pour toute paire (x, y). Nous écrivons $R^k_{\epsilon,pub}(f)$ lorsque le nombre de rondes est borné par k.

Définition 3.13. Soit S un problème avec données corrélées. La *complexité de la communication probabiliste publique avec erreur* ϵ de S, notée $R_{\epsilon,pub}(S)$, est le coût minimal d'un protocole avec générateur aléatoire public qui permet à Bob d'apprendre la chaîne d'Alice avec une probabilité d'erreur bornée par ϵ pour toute paire $(x, y) \in S$. Nous écrivons $R^k_{\epsilon,pub}(S)$ lorsque le nombre de rondes est borné par k.

3.4 Générateur aléatoire public - résultats

Regardons tout d'abord comment un générateur aléatoire public peut nous aider à résoudre le problème de la ligue sportive. Alice choisit un sous-ensemble aléatoire des bits de l'équipe gagnante et envoie le $\oplus$ de ces bits à Bob, ce qui nécessite un bit de communication. Bob calcule ensuite le $\oplus$ équivalent pour chacune des deux équipes finalistes. Il n'est pas difficile de voir que la probabilité de

distinguer deux chaînes différentes à l'aide de cette méthode est $\frac{1}{2}$. Donc, si Alice et Bob exécutent ce processus k fois, la probabilité de ne pas pouvoir éliminer l'équipe perdante devient $\frac{1}{2^k}$. Appelons Z la variable aléatoire représentant le nombre d'équipes qui ne sont pas éliminées après k itérations. Comme l'équipe gagnante n'est jamais éliminée, nous obtenons que $E[Z] = 1 + \frac{1}{2^k}$. En supposant que Bob choisisse au hasard si l'équipe perdante n'est pas éliminée après k itérations, la probabilité qu'il apprenne correctement l'identité de l'équipe gagnante est

$$\begin{aligned}
\Pr[\text{succès}] &= E\left[\frac{1}{\text{nombre d'équipes qui ne sont pas éliminées}}\right] \\
&= E\left[\frac{1}{Z}\right] \\
&\geq \frac{1}{E[Z]} \text{ par l'inégalité de Jensen (voir l'annexe A.4)} \\
&\geq \frac{1}{1+\frac{1}{2^k}}.
\end{aligned}$$

En posant $k = \lceil \log(\frac{1}{\epsilon} - 1) \rceil$, nous obtenons

$$\begin{aligned}
\Pr[\text{succès}] &\geq \frac{1}{1+\frac{1}{2^{\left\lceil \log(\frac{1-\epsilon}{\epsilon}) \right\rceil}}} \\
&\geq \frac{1}{1+\frac{1}{\frac{1-\epsilon}{\epsilon}}} \\
&= \frac{1}{1+\frac{\epsilon}{1-\epsilon}} \\
&= 1-\epsilon,
\end{aligned}$$

ce qui nous permet de conclure que

$$\begin{aligned} R^1_{\epsilon,pub}(S) &\leq k \\ &= \left\lceil \log\left(\frac{1}{\epsilon} - 1\right)\right\rceil. \end{aligned}$$

Nous avons vu à la section précédente que le problème de la ligue sportive requiert $\Theta(\log n)$ bits de communication dans le modèle avec générateurs privés pour toute constante ϵ telle que $0 < \epsilon < \frac{1}{2}$. Comme $R^1_{\epsilon,pub}(S) \in \Theta(1)$, nous pouvons affirmer que le modèle probabiliste public peut être meilleur que le modèle probabiliste privé.

Remarque 3.14. Il est possible d'obtenir $R^1_{\epsilon,pub}(S) \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil - 1$ pour le problème de la ligue sportive en calculant la probabilité de succès de façon exacte plutôt que de la minorer à l'aide de l'inégalité de Jensen. Malheureusement, cela ne permet pas d'économiser plus d'un bit de communication.

Le théorème 3.15 montre que tout protocole probabiliste public peut être transformé en protocole probabiliste privé dont la probabilité d'erreur est un peu plus grande et qui communique un peu plus de bits. Notre résultat s'inspire d'un théorème similaire pour les fonctions booléennes qui a été démontré par Newman [34].

Théorème 3.15 (Mercier 2002). *Soit $S \subseteq X \times Y$ un problème avec données corrélées pour lequel $X = \{0,1\}^n$. Pour tout $\delta > 0$ et pour tout $\epsilon > 0$,*

$$R_{\epsilon+\delta}(S) \leq R_{\epsilon,pub}(S) + O\left(\log \frac{1}{\delta} + \log(n + \log \widehat{a_A})\right).$$

Démonstration. Soit $\mathcal{P}$ un protocole probabiliste public pour S dont l'erreur est bornée par ϵ et nécessitant $R_\epsilon^{pub}(S)$ bits de communication. Nous supposons que le générateur aléatoire obéit à une distribution de probabilité μ. Considérons $Z(x, y, c)$, une variable aléatoire égale à 1 si la réponse donnée par Bob suite à l'exécution de $\mathcal{P}$ sur l'entrée (x, y) avec la chaîne aléatoire c est mauvaise (différente de x) et égale à 0 sinon. Comme $\mathcal{P}$ résout S avec erreur au plus ϵ, il suit que $\underset{c\in\mu}{E}[Z(x, y, c)] \leq \epsilon$ pour toute paire $(x, y) \in S$.

Construisons un nouveau protocole qui utilise moins de bits aléatoires. Soient t un paramètre à être fixé plus tard et $c_1, c_2, \ldots, c_t$ des chaînes binaires. Définissons le protocole $\mathcal{P}_{c_1,c_2,\ldots,c_t}$ suivant : Alice et Bob choisissent un i au hasard entre 1 et t et exécutent le protocole $\mathcal{P}$ avec la chaîne c_i comme chaîne aléatoire commune. Montrons qu'il existe des chaînes $c_1, c_2, \ldots, c_t$ telles que $\underset{i}{E}[Z(x, y, c_i)] \leq \epsilon + \delta$ pour toute paire $(x, y) \in S$. Choisissons les t chaînes $c_1, c_2, \ldots, c_t$ au hasard selon la distribution de probabilité μ. Considérons une paire $(x, y) \in S$ arbitraire, et calculons la probabilité que $\underset{i}{E}[Z(x, y, c_i)] > \epsilon + \delta$ (où i est uniformément distribué). Ceci est exactement la probabilité que $\frac{1}{t}\sum_{i=1}^{t} Z(x, y, c_i) > \epsilon + \delta$. Comme $\underset{c\in\mu}{E}[Z(x, y, c)] \leq \epsilon$, nous obtenons par l'inégalité de Chernoff (voir l'annexe A.4) que

$$\Pr_{c_1,\ldots,c_t}\left[\frac{1}{t}\sum_{i=1}^{t} Z(x, y, c_i) - \epsilon > \delta\right] \leq 2e^{-2\delta^2 t}.$$

En choisissant $t = \frac{1}{\delta^2}\left((n + 1)\ln 2 + \ln \widehat{a_A}\right)$, nous obtenons :

$$\begin{aligned} 2e^{-2\delta^2 t} &\leq 2e^{-2(n+1)\ln 2 - 2\ln \widehat{a_A}} \\ &\leq 2 \cdot 2^{-2(n+1)} \cdot \widehat{a_A}^{-2} \\ &< \frac{2^{-n}}{\widehat{a_A}}. \end{aligned}$$

Donc, pour un choix aléatoire de $c_1, \ldots, c_t$, la probabilité qu'il existe au moins une paire $(x, y) \in S$ (il y a au plus $2^n \cdot \widehat{a_A}$ telles paires) telle que $\underset{i}{E}[Z(x, y, c_i)] > \epsilon + \delta$ est plus petite que $2^n \cdot \widehat{a_A} \cdot \frac{2^{-n}}{\widehat{a_A}} = 1$. Par conséquent, il existe un choix de $c_1, \ldots, c_t$ tel que pour toute paire $(x, y) \in S$, l'erreur du protocole $\mathcal{P}_{c_1,c_2,\ldots,c_t}$ est au plus $\epsilon + \delta$.

Le nombre de bits aléatoires utilisés par $\mathcal{P}_{c_1,c_2,\ldots,c_t}$ est $\lceil \log t \rceil$. Autrement dit, pour transformer le protocole public $\mathcal{P}_{c_1,c_2,\ldots,c_t}$ en protocole privé, Alice n'a qu'à choisir un i au hasard entre 1 et t et à l'envoyer à Bob, ce qui nécessite $\lceil \log t \rceil$ bits de communication. Nous obtenons donc :

$$\begin{aligned} R_{\epsilon+\delta}(S) &\leq R_{\epsilon,pub}(S) + \lceil \log t \rceil \\ &= R_{\epsilon,pub}(S) + \left\lceil \log \left(\frac{1}{\delta^2} \left((n+1) \ln 2 + \ln \widehat{a_A} \right) \right) \right\rceil \\ &\leq R_{pub,\epsilon}(S) + O\left(\log \frac{1}{\delta} + \log(n + \log \widehat{a_A}) \right). \end{aligned}$$

□

Corollaire 3.16 (Mercier 2002). *Soit $S \subseteq \{0,1\}^n \times \{0,1\}^n$ un problème de communication avec données corrélées. Pour tout $\delta > 0$ et pour tout $\epsilon > 0$,*

$$R_{\epsilon+\delta}(S) \leq R_{\epsilon,pub}(S) + O\left(\log \frac{1}{\delta} + \log n \right).$$

Démonstration. Comme $Y = \{0,1\}^n$, il suit $\widehat{a_A} \leq 2^n$. En appliquant le théorème 3.15, nous obtenons

$$\begin{aligned} R_{\epsilon+\delta}(S) &\leq R_{\epsilon,pub}(S) + O\left(\log\frac{1}{\delta} + \log(n + \log 2^n)\right) \\ &\leq R_{\epsilon,pub}(S) + O\left(\log\frac{1}{\delta} + \log n\right). \end{aligned}$$

□

Le corollaire 3.16 signifie que lorsque $S \subseteq \{0,1\}^n \times \{0,1\}^n$ (ou lorsque $\widehat{a_A} \in O(2^n)$), la différence entre la complexité probabiliste privée et la complexité probabiliste publique est au plus un terme additif dans $O(\log n)$ bits.

Essayons maintenant de borner la complexité probabiliste publique comme nous l'avons fait à la section précédente pour la complexité probabiliste privée. Le lemme 3.17 nous donne une borne supérieure triviale que nous améliorons ensuite avec le lemme 3.18.

Lemme 3.17 (Mercier 2002).

$$R_{\epsilon,pub}(S) \leq R_\epsilon(S).$$

Démonstration. Un protocole avec générateurs aléatoires privés peut être simulé par un protocole public dont la chaîne aléatoire commune c est la concaténation de c_A et c_B. □

Lemme 3.18 (Mercier 2002). *Pour tout problème de communication non trivial avec données corrélées S,*

$$R^1_{\epsilon,pub}(S) \leq \lceil \log(\widehat{a_B} - 1) \rceil + \left\lceil \log\frac{1-\epsilon}{\epsilon} \right\rceil.$$

Démonstration. Nous utilisons un argument similaire à celui utilisé pour résoudre le problème de la ligue sportive. Rappelons qu'Alice possède une chaîne x et que Bob possède une chaîne y définissant $a_B(y) = \{x_1, \ldots, x_l\} = \{x \mid (x, y) \in S\}$. Alice choisit k sous-ensembles aléatoires de bits de sa chaîne x et, pour chacun de ces sous-ensembles, envoie le $\oplus$ des bits à Bob. Celui-ci calcule les k $\oplus$ équivalents pour chacune des chaînes x_i de $a_B(y)$. Chaque fois que le $\oplus$ d'un sous-ensemble des bits d'un x_i diffère du résultat correspondant envoyé par Alice, Bob déduit que $x_i \neq x$ et élimine ce sommet. Lorsque Bob a terminé les comparaisons, il tire au hasard une chaîne parmi celles qui n'ont pas été éliminées et conclut que c'est la chaîne d'Alice.

La probabilité de ne pas éliminer un sommet x_i est 1 si $x_1 = x$ et $\frac{1}{2^k}$ sinon. Appelons Z la variable aléatoire représentant le nombre de chaînes qui ne sont pas éliminées après k itérations. Comme il y a $|a_B(y)| - 1$ chaînes à éliminer, il suit que $E(Z) = 1 + \frac{1}{2^k}(|a_B(y)| - 1) \leq 1 + \frac{1}{2^k}(\widehat{a_B} - 1)$. La probabilité que Bob apprenne correctement la chaîne d'Alice est donc

$$\begin{aligned}
\Pr[\text{succès}] &= E\left[\frac{1}{\text{nombre de chaînes qui ne sont pas éliminées}}\right] \\
&= E\left[\frac{1}{Z}\right] \\
&\geq \frac{1}{E[Z]} \text{ par l'inégalité de Jensen (voir l'annexe A.4)} \\
&\geq \frac{1}{1 + \frac{1}{2^k}(\widehat{a_B} - 1)}.
\end{aligned}$$

En posant $k = \lceil \log(\widehat{a_B} - 1) \rceil + \left\lceil \log \frac{1-\epsilon}{\epsilon} \right\rceil$, nous obtenons

$$\begin{aligned}
\Pr[\text{succès}] &\geq \frac{1}{1 + \frac{1}{2^{\lceil \log(\widehat{a_B}-1)\rceil + \left\lceil \log \frac{1-\epsilon}{\epsilon} \right\rceil}} \cdot (\widehat{a_B} - 1)} \\
&\geq \frac{1}{1 + \frac{1}{\frac{1-\epsilon}{\epsilon}}} \\
&= \frac{1}{1 + \frac{\epsilon}{1-\epsilon}} \\
&= 1 - \epsilon,
\end{aligned}$$

ce qui nous permet de conclure que

$$\begin{aligned}
R^1_{\epsilon,pub}(S) &\leq k \\
&= \lceil \log(\widehat{a_B} - 1) \rceil + \left\lceil \log \frac{1-\epsilon}{\epsilon} \right\rceil .
\end{aligned}$$

□

Le prochain théorème résume les résultats démontrés depuis le début de cette section.

Théorème 3.19 (Mercier 2002). *Pour tout problème de communication non trivial avec données corrélées S,*

$$\lceil \log \widehat{a_B} \rceil \leq R^1_{\epsilon,pub}(S) \leq \lceil \log(\widehat{a_B} - 1) \rceil + \left\lceil \log \frac{1-\epsilon}{\epsilon} \right\rceil .$$

Démonstration. La borne inférieure provient du fait que le lemme 3.9 s'applique également au modèle probabiliste public, et la borne supérieure a été démontrée au lemme 3.18. □

TAB. 3.1 – Modèles équivalents pour les problèmes avec données corrélées

Complexité de la communication déterministe lorsqu'Alice connaît la chaîne de Bob
Complexité de la communication déterministe amortie
Complexité de la communication probabiliste avec générateur aléatoire public

À la lumière des résultats précédents, le théorème 3.20 et le tableau 3.1 résument les modèles équivalents pour les problèmes de communication avec données corrélées.

Théorème 3.20. *Pour tout problème de communication avec données corrélées S,*

$$\Theta(\overline{D}(S)) = \Theta(D_{x|y}(S)) = \Theta(R_{\epsilon,pub}(S)).$$

Plus précisément,

$$\begin{aligned} D_{x|y}(S) &\leq R_{\epsilon,pub}(S) \leq D_{x|y}(S) + \left\lceil \log \frac{1-\epsilon}{\epsilon} \right\rceil ; \\ \overline{D}(S) &\leq R_{\epsilon,pub}(S) \leq \overline{D}(S) + \left\lceil \log \frac{1-\epsilon}{\epsilon} \right\rceil + 1. \end{aligned}$$

Démonstration. Il s'agit de combiner le théorème 3.19, le lemme 1.18 et le corollaire 2.20. □

La différence entre les trois modèles est que les protocoles permettant de minimiser la communication amortie nécessitent au moins deux rondes de communication et un nombre d'exemplaires à résoudre qui tend vers l'infini, tandis que pour la communication probabiliste publique et la communication lorsqu'Alice connaît la chaîne de Bob, il existe des protocoles optimaux qui peuvent être

exécutés sur un seul exemplaire et ne nécessitent qu'une ronde de communication.

3.5 L'équivalence des modèles déterministe et probabiliste permet de résoudre le problème de la somme directe

Nous n'avons pas réussi à trouver de problème avec données corrélées pour lequel il existe un algorithme probabiliste avec générateurs aléatoires privés qui est plus efficace que le meilleur des protocoles déterministes. Récapitulons : d'une part, à la section 3.2, nous avons démontré que les modèles probabiliste et déterministe sont équivalents pour les problèmes dont la différence entre la complexité déterministe interactive et non interactive est maximale. D'autre part, à la section 4.3, nous montrons que les modèles probabiliste et déterministe sont équivalents pour une classe de problèmes dont la différence entre la complexité déterministe interactive et non interactive est petite. Ces résultats nous incitent à formuler la conjecture suivante :

Conjecture 3.21 (Mercier 2002). Pour tout problème avec données corrélées, le modèle probabiliste avec générateurs aléatoires privés et le modèle déterministe sont équivalents du point de vue de la communication.

Il y a de bonnes raisons de croire que cette conjecture est difficile à démontrer. En particulier, elle permet de résoudre le problème de la somme directe pour les problèmes de communication avec données corrélées tels que $\widehat{a_A} \in O(2^n)$, ce qui inclut les problèmes ayant un support de la forme $S \subseteq \{0,1\}^n \times \{0,1\}^n$.

Théorème 3.22 (Mercier 2002). *Si les modèles déterministe et probabiliste sont équivalents du point de vue de la communication, alors la communication qui peut être économisée en résolvant simultanément plusieurs exemplaires d'un problème avec données corrélées est un terme additif dans $O(\log n)$ bits par exemplaire.*

Démonstration. Démontrons plutôt la contraposée du théorème. Supposons que $D(S) - \overline{D}(S) \in \omega(\log n)$ (sans perte de généralité, nous ne considérons pas les fonctions dans $\overline{\omega(\log n) \cup O(\log n)}$).

Par le corollaire 2.20, il suit que $D(S) - \log \widehat{a_B} \in \omega(\log n)$, ce qui est équivalent à $D(S) \in \omega(\widehat{a_B} + \log n)$. Comme $\log \widehat{a_B} \in \Theta(R_{\epsilon,pub}(S))$ par le théorème 3.19, nous obtenons que

$$D(S) \in \omega(R_{\epsilon,pub}(S) + \log n). \tag{3.1}$$

Le corollaire 3.16 entraîne que

$$R_\epsilon(S) \in O(R_{\epsilon,pub}(S) + \log n), \tag{3.2}$$

et en combinant (3.1) et (3.2), nous obtenons que $D(S) \in \omega(R_\epsilon(S))$. □

Notons que la réciproque du théorème n'est pas nécessairement vraie : il est possible que les modèles probabiliste et déterministe ne soient pas équivalents et que la communication qui puisse être économisée en résolvant simultanément plusieurs exemplaires d'un problème avec données corrélées soit quand même un terme additif dans $O(\log n)$ bits par exemplaire.

3.6 Complexité distributionnelle

Le modèle de la complexité de la communication distributionnelle, introduit par Yao [55] et traité en détails par Kushilevitz et Nisan [27], considère la distribution de probabilité de $S \subseteq X \times Y$. Cela contraste avec le modèle probabiliste, pour lequel nous avons considéré l'espace des probabilités sur les choix aléatoires effectués par Alice et Bob et considéré les entrées en pire cas. Cette section présente brièvement les résultats qu'Orlitsky [37] a obtenus sur la complexité de la communication distributionnelle de problèmes avec données corrélées.

Définition 3.23. Le support de S, noté $S_{X,Y}$, est défini de la façon suivante :

$$S_{X,Y} \stackrel{\text{déf}}{=} \{(x, y) \mid p(x, y) > 0\}.$$

La notation S est utilisée lorsqu'il n'y a pas de confusion possible.

La définition 3.23 ressemble à la définition 1.7 ; en fait, nous aurions pu définir S de cette façon dès le départ. Si nous ne l'avons pas fait, c'est que comme nous considérons la communication en pire cas pour la complexité déterministe et la complexité probabiliste, la distribution de probabilité de S n'a pas d'importance.

Définition 3.24. Soit $S \subseteq X \times Y$ un ensemble de support qui obéit à une distribution de probabilité μ. $D^k_{\epsilon,\mu}(S)$ est le coût du meilleur protocole déterministe à k rondes qui calcule S pour au moins une fraction $1 - \epsilon$ de toutes les entrées de S, pondérées par μ.

Définition 3.25. $D^k_{0,\mu}(S)$ est la communication espérée (pondérée par μ) du meilleur protocole déterministe à k rondes pour S.

Lemme 3.26.

$$H(X|Y) \leq D_{0,\mu}(S) \leq \cdots \leq D^2_{0,\mu}(S) \leq D^1_{0,\mu}(S) \leq H(X) + 1.$$

Même si $H(X|Y)$ peut être beaucoup plus petit que $H(X)$, ces bornes sont les meilleures qui puissent être exprimées en terme d'entropie de Shannon (voir l'annexe A.5). La borne supérieure est atteinte si S est un produit cartésien, et la borne inférieure est atteinte si S est uniformément distribué.

Lemme 3.27. *Si S est uniformément distribué, alors quatre rondes sont asymptotiquement optimales :*

$$D^4_{0,\mu}(S) \leq D_{0,\mu}(S) + o(D_{0,\mu}(S)).$$

Chapitre 4

Problèmes équilibrés

Depuis le début de ce mémoire, nous ne nous sommes pas préoccupés des liens entre l'ambiguité maximale d'Alice $\widehat{a_A}$ et l'ambiguité maximale de Bob $\widehat{a_B}$. Pour certains problèmes, comme par exemple le problème de la ligue sportive, la différence entre $\widehat{a_A}$ et $\widehat{a_B}$ est grande. Dans ce chapitre, nous analysons la complexité de la communication des problèmes avec données corrélées pour lesquels l'ambiguité maximale d'Alice est égale à l'ambiguité maximale de Bob. Le principal résultat est que les modèles déterministe, déterministe amorti, probabiliste avec générateurs aléatoires privés et probabiliste avec générateur aléatoire public sont équivalents.

4.1 Définitions

Définition 4.1. Un ensemble de support S est *équilibré* si et seulement si $\widehat{a_B} = \widehat{a_A}$.

Définition 4.2. Un ensemble de support *symétrique* S est un support tel que $(x, y) \in S$ si et seulement si $(y, x) \in S$.

Tous les résultats de ce chapitre s'appliquent aux ensembles de support équilibrés, et donc aux supports symétriques (car tout support symétrique est équilibré). Si nous mentionnons les problèmes de communication symétriques avec données corrélées, c'est qu'ils apparaissent de façon naturelle dans toutes les applications mentionnées au sommaire, c'est-à-dire les problèmes pour lesquels les données d'Alice et de Bob sont séparées par une certaine «distance».

- x et y sont deux chaînes binaires dont la distance de Hamming est bornée (par exemple si x a été transmis sur un canal imparfait) ;
- x et y sont des mesures de la même quantité, des entiers dont la valeur absolue de la différence est bornée ;
- x et y sont deux versions d'un fichier original à partir duquel certaines modifications ont été effectuées ;
- etc.

4.2 Résultats

Dans cette section, nous présentons les bornes pour la complexité de la communication déterministe de problèmes équilibrés avec données corrélées. Tous ces résultats ont été démontrés dans un excellent article d'Orlitsky [38]. Nous n'avons pas jugé bon d'inclure toutes les preuves, car certaines sont assez longues et apportent peu à la compréhension de ce mémoire.

Lemme 4.3. *Pour tout problème de communication avec données corrélées S,*

$$D^1(S) \leq \log \widehat{a_A} + \log \widehat{a_B} + 1.$$

Démonstration. Chaque sommet de G_S appartient à au plus $\widehat{a_A}$ hyperarêtes, et chaque hyperarête de G_S contient au plus $\widehat{a_B}$ sommets. Par conséquent, $\chi(G_S) \leq \widehat{a_A} \cdot (\widehat{a_B} - 1) + 1 \leq \widehat{a_A} \cdot \widehat{a_B}$, et

$$\begin{aligned} D^1(S) &= \lceil \log \chi(G_S) \rceil \text{ (lemme 1.19)} \\ &\leq \lceil \log \widehat{a_A} \cdot \widehat{a_B} \rceil \\ &\leq \log \widehat{a_A} + \log \widehat{a_B} + 1. \end{aligned}$$

□

Le lemme précédent nous permet de démontrer que lorsqu'un problème avec données corrélées est équilibré, la complexité de la communication déterministe non interactive est au plus deux fois plus grande que la complexité de la communication déterministe interactive. Ce résultat est encore plus impressionnant que le corollaire 1.27 et n'a évidemment aucun équivalent dans le modèle de Yao.

Corollaire 4.4. *Pour tout problème équilibré de communication avec données corrélées S,*

$$D^1(S) \leq 2D(S) + 1.$$

Démonstration.

$$\begin{aligned} D^1(S) &\leq \log \widehat{a_A} + \log \widehat{a_B} + 1 \\ &= 2 \log \widehat{a_B} + 1 \\ &\leq 2D(S) + 1. \text{ (lemme 1.17)} \end{aligned}$$

□

Orlitsky a démontré que la borne donnée par le corollaire 4.4 était pratiquement optimale.

Lemme 4.5. *Pour tout $\alpha \geq 0$, il existe un problème équilibré avec données corrélées S tel que*

$$D^1(S) \geq 2D(S) - 6 \geq \alpha.$$

Démonstration. Consulter [38]. □

Il a également démontré, en utilisant une famille de fonctions de hachage similaire à celles utilisées dans les chapitres précédents, que trois rondes de communication étaient optimales pour tout problème équilibré.

Théorème 4.6. *Pour tout problème équilibré de communication avec données corélées S,*

$$D(S) \leq D^3(S) \leq \log \widehat{a_B} + 3 \log \log \widehat{a_B} + 11.$$

Démonstration. Consulter [38]. □

Corollaire 4.7. *Pour tout ensemble de support équilibré S,*

$$D^3(S) \leq D(S) + o(D(S)).$$

Démonstration. Il s'agit d'utiliser le fait que $D(S) \geq \log \widehat{a_B}$ (lemme 1.17). □

4.3 Les modèles de communication sont équivalents

Les résultats de la section 4.2 et des trois premiers chapitres nous permettent de démontrer que pour les problèmes équilibrés avec données corrélées, tous les modèles que nous avons considérés sont équivalents du point de vue de la communication. Le tableau 4.1 et le théorème 4.8 résument ce résultat.

Théorème 4.8 (Mercier 2002). *Soit une constante $0 < \epsilon < \frac{1}{2}$. Pour tout problème équilibré de communication avec données corrélées S,*

$$\Theta(D(S)) = \Theta(\overline{D}(S)) = \Theta(D_{x|y}(S)) = \Theta(R_\epsilon(S)) = \Theta(R_{\epsilon,pub}(S)).$$

Démonstration. Nous savons que $\Theta(\overline{D}(S)) = \Theta(D_{x|y}(S)) = \Theta(R_{\epsilon,pub}(S))$ par le théorème 3.20. De plus, le lemme 4.6 et le corollaire 2.20 entraînent que $\Theta(D(S)) = \Theta(\overline{D}(S))$. Finalement, remarquons que $R_{\epsilon,pub}(S) \leq R_\epsilon(S) \leq D(S)$ par les lemmes 3.6 et 3.17. □

Le théorème 4.8 ne tient pas compte des constantes multiplicatives, qui sont en fait toutes égales à 1. Nous présentons donc un théorème plus précis.

TAB. 4.1 – Modèles équivalents pour les problèmes équilibrés avec données corrélées

Complexité de la communication déterministe
Complexité de la communication déterministe lorsqu'Alice connaît la chaîne de Bob
Complexité de la communication déterministe amortie
Complexité de la communication probabiliste avec générateurs aléatoires privés
Complexité de la communication probabiliste avec générateur aléatoire public

Théorème 4.9 (Mercier 2002). *Pour tout problème équilibré de communication avec données corrélées S,*

$$\begin{aligned} D_{x|y}(S) \leq \quad & D(S) \quad && \leq D_{x|y}(S) + o(D_{x|y}(S)), \\ D_{x|y}(S) \leq \quad & R_{\epsilon}(S) \quad && \leq D_{x|y}(S) + o(D_{x|y}(S)), \\ D_{x|y}(S) \leq \quad & R_{\epsilon,pub}(S) \quad && \leq D_{x|y}(S) + o(D_{x|y}(S)). \\ D_{x|y}(S) - 1 \leq \quad & \overline{D}(S) \quad && \leq D_{x|y}(S) + 1. \end{aligned}$$

Démonstration. La première équation peut être déduite à partir du lemme 1.17, du lemme 1.18 et du théorème 4.6 ; la deuxième équation à partir des lemmes 1.18, 3.6 et 3.9 ; la troisième équation à partir du théoreme 3.20 et du lemme 3.17 ; la quatrième équation à partir du corollaire 2.20 et du lemme 1.18. □

Remarquons que contrairement au modèle probabiliste public et au modèle lorsqu'Alice connaît la chaîne de Bob, le lemme 4.5 implique qu'il faut parfois plus d'une ronde de communication pour obtenir un protocole déterministe optimal. Cela dit, le corollaire 4.4 nous assure qu'il existe un protocole déterministe non

interactif requérant au plus deux fois le nombre de bits de communication du protocole optimal.

Pour terminer ce mémoire, il est important de mentionner que les théorèmes 4.8 et 4.9 n'affirment absolument rien sur le temps de calcul des protocoles. Il est possible qu'un problème n'admette pas d'algorithme déterministe efficace du point de vue de la communication et fonctionnant en temps polynomial, mais qu'il en possède un probabiliste.

Bibliographie

[1] Harold Abelson. Lower bounds on information transfer in distributed computations. *Journal of the ACM*, 27(2), 1980.

[2] Sachin Agarwal, David Starobinski, and Ari Trachtenberg. On the scalability of data synchronization protocols for PDAs and mobile devices. *IEEE Network Magazine : Scalability in Communication Networks*, July/August 2002.

[3] Rudolf Ahlswede, Ning Cai, and Zhen Zhang. On interactive communication. *IEEE Transactions on Information Theory*, 43(1) :22–37, 1997.

[4] Noga Alon and Alon Orlitsky. Repeated communication and Ramsey graphs. *IEEE Transactions on Information Theory*, 41(5) :1276–1289, 1995.

[5] Noga Alon and Alon Orlitsky. Source coding and graph entropies. *IEEE Transactions on Information Theory*, 42(5) :1329–1339, 1996.

[6] Charles H. Bennett, Gilles Brassard, Claude Crépeau, and Marie-Hélène Skubiszewska. Practical quantum oblivious transfer. In *Advances in Cryptology : Proceedings of Eurocrypt '91*, volume 576 of *Lecture Notes in Computer Science*, pages 351–366. Springer-Verlag, 1992.

[7] Béla Bollobás. *Extremal Graph Theory*. Academic Press, 1978.

[8] Gilles Brassard. Crusade for a better notation. *ACM Sigact News*, 17(1) :60–64, 1985.

[9] Gilles Brassard and Paul Bratley. *Fundamentals of Algorithmics*. Prentice Hall, 1996.

[10] Gilles Brassard and Louis Salvail. Secret-key reconciliation by public discussion. In *Advances in Cryptology : Proceedings of Eurocrypt '93*, volume 765 of *Lectures Notes in Computer Science*, pages 410–423. Springer-Verlag, 1994.

[11] Graham Cormode, Mike Paterson, Süleyman Sahinalp, and Uzi Vishkin. Communication complexity of document exchange. In *Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*, pages 197–206, 2000.

[12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[13] Claude Crépeau. Efficient cryptographic protocols based on noisy channels. In *Advances in Cryptology : Proceedings of Eurocrypt '97*, volume 1233 of *Lecture Notes in Computer Science*, pages 306–317. Springer-Verlag, 1997.

[14] Jean-Marie De Koninck et Armel Mercier. *Introduction à la théorie des nombres*. Modulo Éditeur, 1994.

[15] Pavol Duris, Zvi Galil, and Georg Schnitger. Lower bounds on communication complexity. In *Proceedings of the 16th annual ACM Symposium on Theory of Computing*, pages 81–91, 1984.

[16] Tomàs Feder, Eyal Kushilevitz, Moni Naor, and Noam Nisan. Amortized communication complexity. *SIAM Journal on Computing*, 24(4) :736–750,

1995.

[17] Uriel Feige and Joe Kilian. Zero knowledge and the chromatic number. *Journal of Computer and System Sciences*, 57(2) :187–199, 1998.

[18] G. David Forney. *Concatenated Codes*. The M.I.T. Press, 1966.

[19] M. Fredman, J. Komlòs, and E. Szemerèdi. Storing a sparse table with O(1) access time. *Journal of the ACM*, 31 :538–544, 1984.

[20] Philippe Galinier and Jin-Kao Hao. Hybrid evolutionary algorithms for graph coloring. *Journal of Combinatorial Optimization*, 3(4) :379–397, 1999.

[21] Abbas El Gamal and Alon Orlitsky. Interactive data compression. In *Proceedings of the 25th IEEE Symposium on Foundations of Computer Science*, pages 100–108, 1984.

[22] Johan Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. In *Proceedings of the 37th IEEE Symposium on Foundations of Computer Science*, pages 627–636, 1996.

[23] Russell Impagliazzo, Ran Raz, and Avi Wigderson. A direct product theorem. In *Proceedings of the 9th Annual Structure in Complexity Theory Conference*, pages 88–96, 1994.

[24] Mauricio Karchmer, Eyal Kushilevitz, and Noam Nisan. Fractional covers and communication complexity. In *Proceedings of the 7th Annual Structure in Complexity Theory Conference*, pages 262–274, 1992.

[25] Mauricio Karchmer, Ran Raz, and Avi Wigderson. Super-logarithmic depth lower bounds via the direct sum in communication complexity. *Computational Complexity*, 5(3/4) :191–204, 1995.

[26] Mark G. Karpowsky, Lev B. Levitin, and Ari Trachtenberg. Data verification and reconciliation with generalized error-control codes. *39th Annual Allerton Conference on Communication, Control, and Computing*, 2001.

[27] Eyan Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[28] Nati Linial and Umesh Vazirani. Graph products and chromatic numbers. In *Proceedings of the 30th IEEE Symposium on Foundations of Computer Science*, pages 124–128, 1989.

[29] John J. Metzner. Efficient replicated remote file comparison. *IEEE Transactions on Computers*, 40(5) :651–660, 1991.

[30] Yaron Minsky and Ari Trachtenberg. Practical set reconciliation. Technical report, Boston University, 2002.

[31] Yaron Minsky, Ari Trachtenberg, and Richard Zippel. Set reconciliation with nearly optimal communication complexity. In *2001 IEEE International Symposium on Information Theory*, page 232, 2001.

[32] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge University Press, 1995.

[33] Moni Naor, Alon Orlitsky, and Peter Shor. Three results on interactive communication. *IEEE Transactions on Information Theory*, 39(5) :1608–1615, 1993.

[34] Ilan Newman. Private vs. common random bits in communication complexity. *Information Processing Letters*, pages 67–71, 1991.

[35] Alon Orlitsky. Worst-case interactive communication I : Two messages are almost optimal. *IEEE Transactions on Information Theory*, 36 :1111–1126, 1990.

[36] Alon Orlitsky. Worst-case interactive communication II : Two messages are not optimal. *IEEE Transactions on Information Theory*, 37 :995–1005, 1991.

[37] Alon Orlitsky. Average case interactive communication. *IEEE Transactions on Information Theory*, 38 :1534–1547, 1992.

[38] Alon Orlitsky. Interactive communication of balanced distribution and of correlated files. *SIAM Journal on Discrete Mathematics*, 6(4) :548–564, 1993.

[39] Alon Orlitsky and Abbas El Gamal. Communication with secrecy constraints. In *Proceedings of the 16th annual ACM Symposium on Theory of Computing*, pages 217–224, 1984.

[40] Alon Orlitsky and Krishnamurthy Viswanathan. Practical protocols for interactive communication. In *Proceedings of the IEEE International Symposium on Information Theory*, page 115, 2001.

[41] King F. Pang and Abbas El Gamal. Communication complexity of computing the Hamming distance. *SIAM Journal on Computing*, 15(4) :932–947, 1986.

[42] Christos H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[43] Christos H. Papadimitriou and Michael Sipser. Communication complexity. In *Proceedings of the 14th annual ACM Symposium on Theory of Computing*, pages 196–200, 1982.

[44] Vera Pless. *Introduction to the Theory of Error-Correcting Codes*. Wiley-Interscience, third edition, 1998.

[45] Sheldon M. Ross. *Initiation aux probabilités*. Presses polytechniques romandes, 1987.

[46] Louis Salvail. Le problème de réconciliation en cryptographie. Mémoire de maîtrise, Département d'informatique et de recherche opérationnelle, Université de Montréal, 1992.

[47] Claude E. Shannon. The zero-error capacity of a noisy channel. *IRE Transactions on Information Theory*, IT-2(3) :8–19, 1956.

[48] Neil J. A. Sloane. The on-line encyclopedia of integer sequences. http ://www.research.att.com/˜njas/sequences/.

[49] Neil J. A. Sloane. On single-deletion-correcting codes (à paraître). Ray-Chaudhuri Festschrift, 2002.

[50] Clark D. Thompson. Area-time complexity for VLSI. In *Proceedings of the 11th ACM Symposium on Theory of Computing*, pages 81–88, 1979.

[51] Ari Trachtenberg and David Starobinski. Towards global synchronisation. Workshop on New Visions for Large-Scale Networks : Research and Applications, 2001.

[52] J. H. van Lint. *Introduction to Coding Theory*. Springer-Verlag, 1982.

[53] Hans S. Witsenhausen. The zero-error side information problem and chromatic numbers. *IEEE Transactions on Information Theory*, 22(5) :592–593, 1976.

[54] Andrew Chi-Chih Yao. Some complexity questions related to distributed computing. In *Proceedings of the 11th ACM Symposium on Theory of Computing*, pages 209–213, 1979.

[55] Andrew Chi-Chih Yao. Lower bounds by probabilistic arguments. In *Proceedings of the 24th IEEE Symposium on Foundations of Computer Science*, pages 420–428, 1983.

[56] Zhen Zhang and Xiang-Gen Xia. Three messages are not optimal in worst case interactive communication. *IEEE Transactions on Information Theory*, 40(1) :3–10, 1994.

Annexe A

Préalables mathématiques

A.1 Notation asymptotique

Lorsqu'il est nécessaire d'analyser l'efficacité d'un algorithme, il est souhaitable de déterminer mathématiquement la quantité de ressources nécessaires en fonction de la taille des exemplaires considérés. La notation asymptotique, introduite à cette fin, évalue le comportement des fonctions à la limite, c'est-à-dire pour des exemplaires assez grands. Cela permet entre autres de comparer entre elles plusieurs fonctions difficilement comparables autrement (par exemple lorsque $f(n_1) < g(n_1)$ et $f(n_2) > g(n_2)$). Bien sûr, il peut arriver que l'analyse asymptotique d'une fonction ne soit d'aucune utilité pratique sur des exemplaires de la vie de tous les jours, néanmoins dans la grande majorité des cas, un algorithme asymptotiquement supérieur à un autre le sera également en pratique. Un autre avantage indéniable de l'analyse asymptotique est qu'elle permet de simplifier grandement l'écriture de la complexité des algorithmes.

Voici les principales définitions utilisées dans ce mémoire pour faire l'analyse asymptotique d'algorithmes. Pour un traitement détaillé du sujet, consulter un ouvrage d'algorithmique, par exemple le livre de Brassard et Bratley [9].

Définition A.1. Soient $f, g : \mathbb{N} \to \mathbb{R}^+$ deux fonctions.

$$
\begin{aligned}
f(n) \in O(g(n)) &\Leftrightarrow (\exists c > 0)(\exists n_0 > 0) \mid (\forall n \geq n_0)[f(n) \leq c \cdot g(n)] \\
f(n) \in \Omega(g(n)) &\Leftrightarrow (\exists c > 0)(\exists n_0 > 0) \mid (\forall n \geq n_0)[f(n) \geq c \cdot g(n)] \\
f(n) \in \Theta(g(n)) &\Leftrightarrow f(n) \in O(g(n)) \wedge f(n) \in \Omega(g(n)) \\
f(n) \in o(g(n)) &\Leftrightarrow (\forall c > 0)(\exists n_0 > 0) \mid (\forall n \geq n_0)[f(n) \leq c \cdot g(n)] \\
f(n) \in \omega(g(n)) &\Leftrightarrow (\forall c > 0)(\exists n_0 > 0) \mid (\forall n \geq n_0)[f(n) \geq c \cdot g(n)]
\end{aligned}
$$

Remarque A.2. Tel que suggéré par Brassard [8], la notation asymptotique est définie en termes d'ensembles dans ce mémoire, ce qui est selon nous plus facile à manipuler, mais surtout beaucoup moins choquant que des égalités dont les deux membres ne peuvent pas être permutés.

A.2 Graphes et hypergraphes

Voici quelques définitions élémentaires reliées aux graphes et aux hypergraphes. Pour plus de détails, consulter un ouvrage sur la théorie des graphes, par exemple «Extremal Graph Theory» de Bollobás [7].

Définition A.3. Un *graphe* $G = (S, A)$ est une paire ordonnée formée d'un ensemble de sommets $S \neq \emptyset$ et d'un ensemble d'arêtes A. Les arêtes sont de la forme $\{s_1, s_2\}$, où $s_1 \neq s_2$ et $s_1, s_2 \in S$.

Définition A.4. Deux sommets s_1 et s_2 sont *adjacents* s'il existe une arête $a = \{s_1, s_2\} \in A$. Deux arêtes sont adjacentes si elles ont un sommet commun. Deux graphes sont *isomorphes* s'il existe une bijection entre leurs ensembles de sommets qui préserve l'adjacence.

Définition A.5. L'*ordre* d'un graphe G, noté $|G|$, est le nombre de sommets de G. La *taille* d'un graphe G, notée $a(G)$, est le nombre d'arêtes de G.

Définition A.6. $\Gamma(s)$ est l'ensemble des sommets adjacents à un sommet s et $d(s) = |\Gamma(s)|$ est le *degré* de s.

Définition A.7. Le *degré minimal* des sommets de G est noté $\delta(G)$, alors que le *degré maximal* des sommets de G est noté $\Delta(G)$. Si $\delta(G) = \Delta(G) = k$, alors G est dit *k-régulier*.

Définition A.8. Un *k-coloriage* de $G = (S, A)$ est une fonction $f : S \to \{1, 2, \ldots, k\}$ telle que $f(s_1) \neq f(s_2)$ pour toute arête $\{s_1, s_2\}$. Le *nombre chromatique* de G est $\chi(G) \stackrel{\text{déf}}{=} \min\{k \mid G \text{ est } k\text{-coloriable}\}$.

Définition A.9. Un *coloriage de deuxième ordre* de G est un coloriage propre de G avec la propriété additionnelle que les voisins de tout sommet du graphe ont des couleurs différentes. Le *nombre chromatique de deuxième ordre* de G est $\chi_2(G) \stackrel{\text{déf}}{=} \min\{k \mid G^2 \text{ est } k\text{-coloriable}\}$.

Définition A.10. Un *ensemble indépendant* d'un graphe $G = (S, A)$ est un ensemble de sommets $S' \subseteq S$ tel qu'aucune paire de sommets de S' n'est reliée par une arête de A. Notons $\alpha(G)$ le nombre maximal de sommets d'un ensemble indépendant de G.

Définition A.11. Un *hypergraphe* H est un ensemble S avec une famille Σ de sous-ensembles non vides de S. Évidemment, $s \in S$ est un sommet de H et $A \in \Sigma$ est une hyperarête de H. Autrement dit, une hyperarête de l'hypergraphe peut contenir plus de deux sommets distincts.

Définition A.12. L'*ordre* d'un hypergraphe, noté $|H|$, est le nombre de sommets de H. La *taille* d'un hypergraphe, notée $a(H)$ est le nombre d'hyperarêtes de H.

Définition A.13. Le *degré minimal* des sommets de l'hypergraphe H est noté $\delta_S(H)$, alors que le *degré maximal* des sommets de H est noté $\Delta_S(H)$. De manière analogue, $\delta_A(H)$ et $\Delta_A(H)$ sont respectivement utilisés pour les degrés minimal et maximal des hyperarêtes de H, où le degré d'une hyperarête est le nombre de sommets qu'elle contient.

Définition A.14. Un *k-coloriage* d'un hypergraphe H ayant un ensemble de sommets S est une fonction $f : S \to \{1, 2, \ldots, k\}$ telle que pour toute hyperarête $\{a_1, \ldots, a_l\}$ de l'hypergraphe, $f(a_i) \neq f(a_j)$ pour tout $1 \leq i < j \leq l$. Autrement dit pour chaque hyperarête, tous les sommets qu'elle contient sont de couleur différente. Le *nombre chromatique* de H est $\chi(H) \stackrel{\text{déf}}{=} \min\{k \mid H \text{ est } k\text{-coloriable}\}$.

A.3 Principe de Dirichlet

Voici une version améliorée du principe de Dirichlet, souvent appelé principe du pigeonnier.

Lemme A.15. *Soient des pigeons de k couleurs ainsi que s trous. Nous supposons qu'un pigeon est d'une seule couleur et que chaque trou ne peut pas contenir plus d'un pigeon de la même couleur. Si chaque trou contient au moins* $\lceil \frac{k}{2} \rceil$ *pigeons,*

alors il existe une couleur telle qu'au moins $\lceil \frac{s}{2} \rceil$ *trous contiennent des pigeons de cette couleur.*

Démonstration. Remarquons d'abord qu'il y a au moins $s\lceil \frac{k}{2} \rceil$ pigeons dans les trous. Supposons que chaque couleur est présente dans au plus $\lfloor \frac{s-1}{2} \rfloor$ trous. Cela entraîne que le nombre de pigeons dans les trous est au plus $k\lfloor \frac{s-1}{2} \rfloor < (\frac{s}{2})k \leq s\lceil \frac{k}{2} \rceil$. Contradiction. □

A.4 Probabilités

Cette section contient quelques inégalités provenant de la théorie des probabilités qui sont utiles pour ce mémoire. Consulter [32, 45] pour les preuves ainsi que pour des informations additionnelles sur le sujet.

Théorème A.16 (Inégalité de Markov). *Pour toute variable aléatoire* X *supérieure ou égale à 0 et pour tout* $\alpha > 0$,

$$\Pr[X \geq \alpha] \leq \frac{E[X]}{\alpha}.$$

L'inégalité de Markov peut également être exprimée comme

$$\Pr[X \geq \alpha \cdot E[X]] \leq \frac{1}{\alpha}.$$

Lorsque des algorithmes probabilistes sont analysés, il est essentiel de pouvoir borner la probabilité qu'une variable aléatoire X s'éloigne de son espérance $E(X)$. En fait, lorsqu'une variable aléatoire est générée plusieurs fois, cela revient à borner le vitesse de convergence de la valeur moyenne des tirages obtenus.

Théorème A.17 (Inégalité de Chernoff). *Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires booléennes indépendantes pour lesquelles $\Pr[X_i = 1] = p \leq 1/2$. Pour tout δ tel que $0 < \delta \leq p(1-p)$,*

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - p\right| \geq \delta\right] \leq 2e^{-\frac{\delta^2 n}{2p(1-p)}}.$$

L'inégalité de Chernoff peut être généralisée pour des variables aléatoires continues.

Théorème A.18 (Inégalité de Hoeffding). *Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes ayant la même distribution de probabilité sur l'intervalle réel $[a, b]$. Si $E[X] = p$, alors*

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - p\right| \geq \delta\right] \leq 2e^{-\frac{2\delta^2 n}{b-a}}.$$

La dernière inégalité dont nous aurons besoin porte sur des espérances plutôt que des probabilités.

Théorème A.19 (Inégalité de Jensen). *Soit f une fonction convexe. Alors*

$$E[f(X)] \geq f(E[X])$$

pour autant que ces espérances existent et soient finies.

A.5 Entropie

L'entropie est une mesure de l'incertitude d'une variable aléatoire qui possède plusieurs propriétés cohérentes avec la notion intuitive d'information. Consulter

l'ouvrage de Cover et Thomas [12] pour une introduction à la théorie de l'information.

Soit X une variable aléatoire discrète avec alphabet Σ telle que $p(x) = \Pr(X = x)$, $x \in \Sigma$.

Définition A.20. L'*entropie* d'une variable aléatoire discrète X, notée $H(X)$, est définie par :

$$H(X) \stackrel{\text{déf}}{=} -\sum_{x\in\Sigma} p(x) \log p(x).$$

L'entropie est exprimée en bits, et par convention $0 \log 0 = 0$.

Soit

$$X = \begin{cases} 0 \text{ avec probabilité } p \\ 1 \text{ avec probabilité } 1-p \end{cases}$$

Alors $H(X) = -p \log p - (1-p) \log(1-p) \stackrel{\text{déf}}{=} h(p)$.

Définition A.21. Soit (X, Y) une paire de variables aléatoires avec une distribution de probabilité $p(x, y)$. L'*entropie conjointe* de (X, Y), notée $H(X, Y)$, est définie par :

$$H(X,Y) \stackrel{\text{déf}}{=} -\sum_{x\in X}\sum_{y\in Y} p(x,y) \log p(x,y).$$

Définition A.22. L'*entropie conditionnelle* de Y sachant X, notée $H(Y|X)$, est définie par :

$$\begin{aligned} H(Y|X) &\stackrel{\text{déf}}{=} \sum_{x\in X} p(x) H(Y|X=x) \\ &= -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x) \log p(y|x). \end{aligned}$$

Le théorème A.23 permet de relier les trois définitions précédentes.

Théorème A.23.

$$H(X,Y) = H(X) + H(Y|X).$$