Problématique des bons codes sur le corps à deux éléments

Daniel Augot

Problématique des bons codes sur le corps à deux éléments

Daniel Augot

1 Introduction

La théorie des codes correcteurs se préoccupe de transmission fiable à travers de canaux de communication soumis à des perturbations qui peuvent causer des erreurs. Le principe est de rajouter de la redondance au message utile, de sorte qu'on puisse reconstruire celui-ci en présence de modifications inconnues induites par le canal de transmission. La question pratique, mais aussi toute théorique, est de déterminer quelle quantité de redondance doit être introduite pour réduire efficacement le bruit du canal. On peut en effet répéter à loisir chaque symbole émis, et un nombre assez grand de répétitions permet de diminuer le bruit. Mais le problème avec cette approche est que le taux de transmission, qui est le rapport entre le nombre de symboles utiles et le nombre de symboles transmis, devient très faible quand on augmente le nombre de répétitions. En 1948, Shannon a montré que, pour tout canal de transmission sans mémoire, on peut communiquer avec un taux de transmission constant limite, dépendant du canal. La preuve de Shannon est non constructive, et le but principal de la théorie des codes a été de construire des codes, avec les algorithmes de correction d'erreur associés, permettant d'atteindre ce taux de transmission limite. Les *turbocodes* et les *codes à matrice de parité creuse — codes LDPC* ont menés à une solution effective au problème de Shannon. Ces résultats permettent de lutter contre des canaux stochastiques, avec une probabilité d'erreur résiduelle évanescente.

D'un autre coté, en 1950, Hamming a posé des définitions combinatoires pour les codes correcteurs, notamment celle de la distance minimale d'un code. La *théorie des codes algébriques* s'est alors développée, avec les codes de Hamming, les codes BCH, les codes de Reed-Solomon, etc, et enfin les codes géométriques de Goppa. La notion de distance minimale permet de lutter contre tous les bruits, même les plus difficiles, tant que ceux ci sont de faible importance, mais la contrepartie est que les taux de transmission obtenus dans ce contexte sont beaucoup plus faibles que dans la vision stochastique. De plus les codes stochastiques permettent de traiter le cas de canaux continus, qui sont mal pris en compte par les codes algébriques. Cela explique que les turbocodes et les codes LDPC sont maintenant préférés dans beaucoup d'applications, quand les codes algébriques restent prédominants dans d'autres.

Dans cet exposé, on traite la problématique introduite par Hamming, qui n'est pas encore complètement résolue, notamment dans le cas où l'espace des symboles à transmettre est l'alphabet binaire (les bits). Il est notable que ce cas est plus difficile que des alphabets plus gros (par exemple le corps fini à 256 éléments qui correspond aux octets), où la construction des codes géométriques

due à Goppa en 1981 donne d'excellent résultats.

Dans une dernière partie, je citerai les nouveaux résultats obtenus dans le domaine du codage algébrique, avec le *décodage en liste*. Cette relaxation du problème du décodage permet d'atteindre des taux de transmission proches de ceux prédits par Shannon, mais dans le cas du canal dit *adversaire*, c'est-à-dire dans *le pire cas* pour l'algorithme de décodage. Ces travaux ont été obtenus par les informaticiens théoriques (Sudan, Guruswami, etc) [4], qui ont rafraîchi et renouvelé la théorie algébrique des codes en utilisant des notions modernes d'algorithmique et de complexité introduites bien après 1948. Il est toutefois frappant de voir que certaines techniques classiques sont quasiment réutilisées à l'identique dans le contexte du décodage en liste (arguments sur les codes aléatoires, codes concaténés, etc).

Cet exposé se voulant introductif et destiné à un public non spécialiste, j'ai donné peu de références aux articles des revues de recherche, donnant si possible la préférence à des ouvrages où figure une introduction pédagogique aux sujets présentés : Berrou [2] pour une approche stochastique du décodage, Roth [8] pour les codes algébriques classiques, dont notamment les codes concaténés, Vladut, Nogin et Tsfasman [9] pour les bornes combinatoires usuelles et les codes géométriques, et Guruswami [4] pour le décodage en liste.

2 Shannon et Hamming

2.1 Théorème de Shannon

Définition 1. *Un* canal de transmission sans mémoire *est la donnée de deux alphabets $\mathcal{A}$ et $\mathcal{B}$, et de* probabilités de transitions

$$\mathrm{P}(y \mid x)$$

qui sont les probabilités de recevoir $y \in B$ sachant que $x \in A$ a été émis.

Définition 2. *Un* code correcteur C *est sous ensemble de $\mathcal{A}^n$. Si $\mathcal{A}$ est de cardinal q, on note $k = \log_q |C|$. Le* taux de transmission *de C est le rapport $k/n \in [0,1]$. On dit que n est la* longueur *de C.*

Un élément dans le code est appelé *mot de code* (il est de tradition de dire « mot »plutôt que vecteur, et d'utiliser une notation horizontale). Le principe du code est d'introduire de la redondance, c'est-à-dire que les mots de C contiennent typiquement des symboles $x \in A$ redondants. On utilisera ensuite cette redondance pour retrouver le mot de code correct, après une transmission pendant laquelle certaines positions ont été faussées.

Définition 3. *Un* algorithme de décodage *du code C associe à tout mot de $\mathcal{B}^n$ un élément de $C \cup \{\perp\}$, où $\perp$ correspond à un* échec de l'algorithme.

Il faut distinguer un échec de l'algorithme, d'une *erreur de décodage*, qui correspond au cas où le mot décodé ne correspond pas au mot reçu. Il peut être préférable d'avoir un échec contrôlé à une erreur non contrôlée. Le théorème de Shannon s'exprime comme suit.

Théorème 1. *Pour tout canal de transmission sans mémoire, il existe une quantité $C \in [0,1]$ appelée* capacité du canal*, tel qu'il existe un famille infinie*

de codes C_n, de longueur croissante, de taux de transmission $R < C$, telle que la probabilité maximale d'erreur de décodage tende vers zéro.

La capacité d'un canal de transmission s'exprime en termes issus de la théorie de l'information (information mutuelle). Pour certains canaux simples, comme le *canal q-aire symétrique* de probabilité de transition $p \in [0, 1 - 1/q]$ (c'est la probalibité comme symbole émis x soit changé en un symbole $y \neq x$), la capacité est $1 - H_q(p)$, où H_q est la *fonction d'entropie q-aire* :

$$H_q(p) = p \log_q(q-1) - p \log_q p - (1-p) \log_q(1-p).$$

Le théorème de Shannon est non constructif, et la famille des codes n'est pas spécifiée, même pour les canaux q-aires symétriques.

2.2 Approche combinatoire de Hamming

On considère un alphabet fini $\mathcal{A}$ que nous verrons comme un corps fini $\mathbb{F}_q$. En pratique, le corps le plus utilisé est $\mathbb{F}_2$, et on parle alors de codes binaires. Dans ce contexte, un code est *linéaire* s'il est un $\mathbb{F}_q$-sous espace vectoriel de $\mathbb{F}_q^n$.

Définition 4. *La* distance de Hamming *entre deux mots x, y de $\mathbb{F}_q^n$ est*

$$d(x, y) = |\{i;\ x_i \neq y_i\}|.$$

Les isométries pour la distance de Hamming sont les permutations des indices, composées avec des permutations de $\mathbb{F}_q$ agissant séparèment sur les composantes. On a peu de codes isométriques à un code donné, et cela justifie la terminologie « code » plutôt que sous-espace vectoriel.

Définition 5. *La* distance minimale *d'un code C est*

$$d = d(C) = \min_{x \neq y \in C} d(x, y).$$

Sa capacité de correction *est $t = \lfloor \frac{d-1}{2} \rfloor$.*

Lors de la transmission d'un mot de code c, s'il y a eu moins de t symboles de c perturbés par le canal de transmission, alors un algorithme qui donne le mot de code de plus proche d'un mot $y \in \mathbb{F}_q^n$ retrouvera de manière correcte le mot émis. Il n'y aura pas d'erreur de décodage, mais l'algorithme peut déclarer un échec si le nombre d'erreurs est supérieur à t. L'objectif combinatoire de la théorie des codes au sens de Shannon est de construire des codes tels que leur dimension k et leur distance minimale d soient toutes les deux grandes. Ces deux objectifs sont antagonistes.

Pour un code linéaire de longueur n, dimension k, et distance minimale d sur $\mathbb{F}_q$, on parle de code de paramètres $[n, k, d]_q$. Pour un code non linéaire de longueur n, cardinal M, distance minimale d sur un alphabet de taille q, on parle de code de paramètres $(n, M, d)_q$.

2.3 Interprétation

On dit souvent que le modèle étudié par Shannon correspond à un modèle de canal aléatoire, par opposition à la définition de Hamming, que l'on associe

maintenant au canal dit *adversaire*. En effet, dans le canal aléatoire, le modèle d'erreur est stochastique et ne tient pas compte du code utilisé. Dans le cas de Hamming, on requiert que le code soit capable de décoder *toute configuration de t erreurs*, même les plus défavorables. Une autre manière de considérer les choses est de se placer dans un contexte algorithmique : on dira qu'un bon code au sens de Shannon traitera le cas aléatoire avec une probabilité résiduelle très faible ; alors qu'un bon code au sens de Hamming traitera tous les cas, même le cas le pire, tant que le nombre d'erreur est borné par t. En revanche, ce dernier échouera systématiquement quand le nombre d'erreurs est trop élevé.

Cela limite sévèrement les performances des codes issus des considérations de Hamming. En revanche, dans le modèle aléatoire, les progrès dus à Berrou et Glavieux (les turbocodes) et à MacKay et Spielman (redécouverte des codes à matrice de parité creuses de Gallager — codes « LDPC ») ont permis d'atteindre pour certains canaux la capacité donnée par la théorie de Shannon, avec des codes qui n'ont pas forcément une bonne distance minimale [2].

Toutefois, la recherche de bons codes au sens de Hamming s'est avéré être un problème combinatoire riche et difficile, en relation avec d'autres domaines des mathématiques discrètes (graphes, plans d'expériences — designs, empilements de sphères, etc). Le volant algorithmique s'est avéré lui aussi riche de problèmes informatiques, tant on ne sait pas bien construire des algorithmes de décodage efficaces pour des codes linéaires génériques. Cette difficulté du problème du décodage a fourni à la cryptographie le chiffrement de McEliece, contemporain du système de chiffrement RSA, et qui résiste toujours à la cryptanalyse.

3 Bornes combinatoires

3.1 La fonction α_q

Pour une taille d'alphabet donnée q, en considérant les codes de longueur n, et de distance minimale d, on se pose la question de connaître le cardinal maximal d'un tel code. On a aussi une version asymptotique de ce problème.

Définition 6. *La quantité $A_q(n,d)$ est le cardinal maximal d'un code de longueur n et de distance minimale d sur l'alphabet q-aire. Pour $\delta \in [0,1]$, on définit $\alpha_q(\delta)$:*

$$\alpha_q(\delta) = \limsup \frac{\log_q A_q(n, \lfloor \delta n \rfloor)}{n}.$$

L'intuition que la taille d'un code doit décroître avec sa distance minimale est vraie.

Théorème 2. *La fonction $\delta \mapsto \alpha_q(\delta)$ est continue et décroissante sur $[0,1]$.*

On a la version linéaire $\alpha_q^{\text{lin}}(\delta)$ qui correspond aux mêmes questions posées dans le cas des codes linéaires. On sait peut de choses sur les fonctions $\alpha_q(\delta)$ et $\alpha_q^{\text{lin}}(\delta)$: sont-elles dérivables, sont-elles concaves ? A-t-on $\alpha_q(\delta) = \alpha_q^{\text{lin}}(\delta)$? Toutefois de nombreuses bornes existent, inférieures ou supérieures sur ces fonctions.

3.2 Bornes supérieures

Proposition 1 (Borne de Singleton)**.** *Tout code C (linéaire ou non) q-aire de longueur n et de distance minimale d vérifie $|C| \leqslant q^{n-d+1}$. On a donc $\alpha_q(\delta) \leqslant 1-\delta$.*

Proposition 2 (Borne de Plotkin)**.** *Tout code C (linéaire ou non) q-aire de longueur n et de distance minimale d vérifie*

$$d \leqslant n \cdot \frac{|C|}{|C|-1} \cdot \frac{q-1}{q}$$

On a donc $\alpha_q(\delta) = 0$, pour $\delta \in [(q-1)/q, 1]$.

Proposition 3 (Borne de Hamming)**.** *Tout code C (linéaire ou non), de distance minimale d et de capacité de correction $t = \lfloor \frac{d-1}{2} \rfloor$, vérifie*

$$|C| \cdot V_q(n,t) \leqslant q^n,$$

où $V_q(n,t) = \sum_{i=0}^{t} (q-1)^i \binom{n}{i}$ est le volume de la sphère de Hamming de rayon t en longueur n sur l'alphabet q-aire.

Pour obtenir la version asymptotique, il y a l'équivalent suivant, pour $\delta \in [0,1]$:

$$\frac{\log_q V_q(n, \lfloor \delta n \rfloor)}{n} \sim H_q(\delta).$$

Proposition 4. $\alpha_q(\delta) \leqslant 1 - H_q(\delta/2)$.

Ces bornes sont défavorables, notamment dans le cas binaire. Le point de vue de Shannon montre qu'on peut avoir des taux de transmission proche $1-H_2(p)$, pour un taux d'erreur de $p \in [0, 1/2]$, alors que le point de Hamming limite sévèrement la taux de transmission. Par exemple, pour un taux d'erreur de $1/4$, le point de vue de Shannon dit qu'on peut communiquer à taux $R = 1 - H_2(1/4) \approx 0,188 > 0$, alors que, du point de Hamming, on doit avoir une distance minimale de $1/2$. Mais la borne de Plotkin entraîne que $\alpha_2(1/2) = 0$. On ne peut donc pas corriger, au sens combinatoire, un taux d'erreur de $1/4$ avec un taux de transmission positif.

3.3 Bornes inférieures

Si $|C| \cdot V_q(n, d-1) < q^n$, il existe un code de paramètres $(n, M, d)_q$. On construit ce code comme avec l'algorithme (non efficace) suivant :

1. (Initialisation) On tire un premier mot de longueur n, qui est le premier élément du code ;
2. (Itération) On considère tous les mots à distance inférieure à $d-1$ des mots déjà construits. Si cet ensemble ne recouvre pas tout Q^n, alors on peut ajouter un mot au code construit.

On a de même un borne très proche quand on se restreint à la famille des codes linéaires.

Proposition 5 (Borne de Gilbert-Varshamov)**.** *Si $V_q(n-1, d-2) < q^{n-k}$, alors il existe un code linéaire de paramètres $[n,k,d]_q$. Asymptotiquement, $\alpha_q(\delta) \geq 1 - H_q(\delta)$.*

Cette borne peut être démontrée avec des arguments probabilistes, qui permettent de plus d'obtenir que la probabilité qu'un code tiré uniformément au hasard soit en dessous de la borne de Varshamov-Gilbert tend exponentiellement vers zéro avec la longueur n. Il a été longtemps cru que la version asymptotique de cette borne était en fait une borne *supérieure* : on pensait qu'on ne pouvait pas obtenir de meilleurs codes que ceux construits au hasard. La construction de codes géométriques de Goppa a montré que non, au moins pour les alphabets pas trop petits. C'est l'objet de la partie suivante.

Un objectif plus modeste est de construire une famille de *codes asymptotiquement bonne*.

Définition 7. *À alphabet fixé $\mathbb{F}_q$, une famille de codes C_i, de paramètres $[n_i, k_i, d_i]_q$ est asymptotiquement bonne si*

$$\limsup \frac{k_i}{n_i} > 0, \text{ et } \limsup \frac{d_i}{n_i} > 0.$$

Cette définition est motivée par la constatation que les codes binaires algébriques classiques (BCH, Reed-Muller etc), qui peuvent être optimaux pour des paramètres finis, voient leur distance minimale relative ou leur taux de transmission tendre vers zéro. De plus les codes asymptotiquement bons voient les algorithmes de décodage triviaux (énumération de toutes les erreurs ou énumération de tous les mots de codes) être de complexité pûrement exponentielle.

4 Codes algébriques

4.1 Codes de Reed-Solomon

Définition 8. *Soient $x_1, \ldots, x_n \in \mathbb{F}_q$, deux à deux distincts, avec $n < q$. La* fonction d'évaluation *associée à $x_1, \ldots, x_n$ est*

$$\begin{array}{rccc} \text{ev} : & \mathbb{F}_q[X] & \to & \mathbb{F}_q^n \\ & f & \mapsto & (f(x_1), \ldots, f(x_n)) \end{array}$$

Définition 9 (Codes de Reed-Solomon)**.** *Soient $x_1, \ldots, x_n \in \mathbb{F}_q$, deux à deux distincts, avec $n < q$, et* ev *la fonction d'évaluation associée. Soit*

$$L_k = \{f \in \mathbb{F}_q[X];\ \deg f < k\},$$

l'ensemble des polynômes de degré inférieur à k avec $0 \leqslant k \leqslant n$. Le code de Reed-Solomon *de dimension k est $C = \text{ev}(L_k)$.*

On voit que la distance minimale de ces codes vérifie $d = n - k + 1$. Ils atteignent donc la borne de Singleton, et de tels codes sont dits *MDS* (Maximum Distance Separable). Ils sont en un sens optimaux, mais ils ne répondent pas complètement à la problématique posée, car leur longueur est petite : $n < q$. On ne peut pas les considérer du point asymptotique, en gardant une taille d'alphabet fixée. En particulier pour le corps à deux éléments, il n'existe pas de codes de Reed-Solomon ou MDS pertinents. En revanche, pour l'alphabet $\mathbb{F}_{256} = \mathbb{F}_{2^8}$, qui est commode en informatique (les octets), les codes de Reed-Solomon sont de longueur correcte, et ont de nombreuses applications (CD, DVD, Blu-Ray, TNT, etc).

Le problème de déterminer à alphabet fixé, la plus grande longueur possible des codes MDS, est résumé par la difficile conjecture MDS, qui a des connexions les géométries finies [6].

Conjecture 1 (Conjecture MDS)**.** *Tous les codes MDS de paramètres* $[n, k, n-k+1]_q$ *vérifient* $n \leqslant q+1$*, sauf si* $k = 3$ *ou* $k = q-1$*, avec* q *pair, auquel cas on a* $n \leqslant q+2$*.*

4.2 Codes de Reed-Muller

Une manière naturelle d'augmenter la longueur des codes est de considérer un contexte multivarié. On note $n = q^m$, et $\{P_1, \ldots, P_n\}$ une énumération des points de $\mathbb{F}_q^m$.

Définition 10. *La fonction d'évaluation associée aux points* $\{P_1, \ldots, P_n\}$ *est*

$$\begin{array}{rccc} \mathrm{ev}: & \mathbb{F}_q[X_1, \ldots, X_m] & \to & \mathbb{F}_q^n \\ & f & \mapsto & (f(P_1), \ldots, f(P_n)) \end{array}$$

Définition 11 (Codes de Reed-Muller généralisés)**.** *Soit* $n = q^m$*,* $\{P_1, \ldots, P_n\}$ *les points de* $\mathbb{F}_q^m$ *et* ev *la fonction d'évaluation associée. On considère l'espace*

$$L_r = \{f \in \mathbb{F}_q[X_1, \ldots, X_m];\ \deg f \leqslant r\}.$$

Le code de Reed-Muller *d'ordre* r *est* $C = \mathrm{ev}(L_r)$*, pour* $0 \leqslant r \leqslant m(q-1)$*.*

En comptant le nombre de monômes de degré inférieur à r, et en utilisant le lemme de Schwartz-Zippel sur le nombre de zéros d'un polynôme multivarié sur un corps fini, on peut déterminer la dimension et la distance minimale des codes de Reed-Muller. Dans le cas $r < q$, les formules sont simples.

Proposition 6. *Pour* $r < q$*, le code de Reed-Muller d'ordre* r *est de dimension* $\binom{r+m}{m}$ *et de distance minimale* $(q-r)q^{m-1}$*.*

Pour r plus élevé, la taille du corps intervient et les formules pour la dimension et distance minimale sont connues [1]. Dans tous les cas, les paramètres du code de Reed-Muller généralisé vérifient : $k \leqslant (r+1)^m$, et $d = q^m(1 - \frac{r}{q})$. Du point de vue asymptotique, à r fixé, on obtient

$$R = \frac{k}{n} \leqslant \left(\frac{r+1}{q}\right)^m,$$

et une distance minimale relative constante $\delta = \frac{d}{n} = 1 - \frac{r}{q}$. Le taux de transmission R se dégrade exponentiellement avec le nombre de variables, et les codes de Reed-Muller ne sont donc pas asymptotiquement bons.

4.3 Codes géométriques

Une manière d'améliorer la situation précédente est de considérer l'évaluation des polynômes multivariés sur des points bien choisis $P_1, \ldots, P_n$ de $\mathbb{F}_q^m$, avec $n \ll q^m$. Cela revient à considérer des codes plus courts : les paramètres n, k et

d vont décroître, mais on espère que le taux $R = k/n$ et la distance minimale relative $\delta = d/n$ vont devenir bons. On a alors la fonction d'évaluation :

$$\begin{array}{rccc} \text{ev}: & \mathbb{F}_q[X_1, \ldots, X_m] & \to & \mathbb{F}_q^n \\ & f & \mapsto & (f(P_1), \ldots, f(P_n)) \end{array},$$

et le code associé est

$$C = \text{ev}(L_r),$$

où $L_r = \{f \in \mathbb{F}_q[x];\ \deg f < r\}$. Pour avoir suffisamment de structure, les points sont pris sur une courbe algébrique $\mathcal{C}$ définie sur $\mathbb{F}_q$, de genre g. C'est la construction de Goppa. Soit $\mathbb{F}_q(\mathcal{C})$ le corps des fonctions de $\mathcal{C}$. Pour simplifier je présente les codes dits « à un point ». Soit Q un point de la courbe différent de $P_1, \ldots, P_n$, et r un entier, l'*espace associé* à rQ, noté $L(rQ)$ est l'ensemble des fonctions $f \in \mathbb{F}_q(\mathcal{C})$ n'admettant qu'un unique pôle en Q, d'ordre au plus r. Une conséquence du théorème de Riemann-Roch [9] donne une borne sur la dimension des codes.

Théorème 3. *Pour $r > 2g - 2$, on a $\dim L(rQ) = r - g + 1$.*

Comme une fonction ne peut avoir plus de zéros que de pôles, on a :

Proposition 7. *Pour $r > 2g - 2$, le code de Goppa associé à l'espace $L(rQ)$ et à la fonction* ev *ci dessus est un code $[n, k = r - g + 1, d \geqslant n - r]_q$.*

Les codes géométriques vérifient donc $k \geqslant n - d + 1 - g$, et leur défaut par rapport à la borne de Singleton est le genre g de $\mathcal{C}$. Dans le cas du genre zéro, on retrouve les codes de Reed-Solomon. Le point de vue asymptotique donne, pour une famille de courbes $\mathcal{C}_i$, de genre g_i, dont le nombre de points n_i tend vers l'infini, l'inégalité

$$R \geqslant 1 - \delta - \limsup \frac{g_i}{n_i}.$$

On cherche donc une famille de courbes telle que le nombre de points croisse correctement par rapport à g. Une borne inférieure sur le rapport g/n est donnée par la borne de Hasse-Weil.

Théorème 4. *Soit $\mathcal{C}$ une courbe défini sur $\mathbb{F}_q$, de genre g. Le nombre de points de $\mathcal{C}$ à coefficients dans $\mathbb{F}_q$ vérifie*

$$|N_q(\mathcal{C}) - (q + 1)| \leq 2g\sqrt{q}.$$

Le résultat qui a montré l'importance et la pertinence des codes géométriques est dû à Tsfasman, Vladut et Zink.

Théorème 5. *Si q est un carré, alors* $\limsup \frac{n}{g} = \sqrt{q} - 1$.

Cela entraîne que, si q est un carré, il existe une famille de codes $[n_i, k_i, d_i]_q$ telle que $\limsup \frac{k_i}{n_i} = R$ et $\limsup \frac{d_i}{n_i} = \delta$, avec $R = 1 - \delta - \frac{1}{\sqrt{q}-1}$. Dans le cas où $q \geqslant 49$, cette nouvelle borne dépasse la borne inférieure de Varshamov-Gilbert : on a ainsi une famille de codes non seulement asymptotiquement bonne, mais de plus meilleure que les codes aléatoires. Le cas binaire reste en suspens.

4.4 Codes concaténés

Une méthode permettant d'obtenir des codes sur des petits corps à partir de bons codes sur des grands alphabets est la technique des *codes concaténés*. La construction est la suivante (nous nous restreignons au cas binaire).

Définition 12. *Soit $\mathbb{F}_q = \mathbb{F}_{2^k}$, et C un code de paramètres $[N, K, D]_{2^k}$. Soit C_0 un code binaire de paramètres $[n, k, d]_2$, et $\phi_0 : \mathbb{F}_2^k \to C_0$ bijective. Le* code concaténé *de C par C_0 est le code dont les mots sont*

$$(\phi_0(c_1), \ldots, \phi_0(c_n)), \quad \text{pour } c = (c_1, \ldots, c_n) \in C,$$

où $\mathbb{F}_{2^k}$ est identifié arbitrairement avec $\mathbb{F}_2^k$.

D'un point de vue pratique, on encode d'abord un message $m \in \mathbb{F}_{2^k}^K$ en un mot de code $c \in \mathbb{F}_{2^k}^N$, dont on encode ensuite chaque composante séparément, en utilisant ϕ_0.

Proposition 8. *Soit $\mathbb{F}_q = \mathbb{F}_{2^k}$, et C un code de paramètres $[N, K, D]_{2^k}$, et C_0 un code binaire de paramètres $[n, k, d]_2$. Le code concaténé de C par C_0 a pour paramètres $[nN, kK, dD]_2$.*

Soit $r = \frac{k}{n}$, $\delta = \frac{d}{n}$. Supposons que C_0 ait des paramètres proches de la borne de Varshamov-Gilbert : $\delta \geqslant H_2^{-1}(1 - R) - \varepsilon$, et que C soit sur la borne de Singleton, en prenant par exemple un code de Reed-Solomon. Alors le code concaténé sera de taux rR, et distance minimale

$$(1 - R)\left(H_2^{-1}(1 - r) - \varepsilon\right).$$

Théorème 6. *Il existe une famille de codes binaires asymptotiquement bonne, celle des codes concaténés, constructible en temps polynomial en la longueur.*

Démonstration. La remarque cruciale est que la longueur du code C_0 est logarithmique en la longueur du code concaténé, quand le code C est un code de Reed-Solomon. On peut donc se permettre une recherche exhaustive sur tous les codes C_0, jusqu'à en trouver qui soit proche de la borne de Varshamov-Gilbert. Le code de Reed-Solomon C est complètement explicite et peut-être construit en temps polynomial. □

De plus les codes concaténés disposent d'un algorithme de décodage qui permet d'atteindre la capacité du canal binaire symétrique prédite par Shannon [8].

4.5 Résultats récents

Alors que l'on sait construire des codes meilleurs que les codes aléatoires sur des alphabets relativement gros (codes géométriques), et une famille de codes asymptotiquement bonne sur le corps à deux éléments, la question reste ouverte de savoir s'il existe des codes binaires meilleurs que les codes aléatoires, dont on sait qu'ils correspondent à la borne de Gilbert-Varshamov. Des résultats récents dus à Jiang et Vardy dans le cas non linéaire [7], et à Gaborit et Zémor dans le cas linéaire [3], montrent que la borne de Gilbert-Varshamov n'est pas la meilleure borne inférieure sur les paramètres des codes optimaux.

Proposition 9. *Il existe une constante $c > 0$, telle que, pour $\frac{d}{n} \leqslant 0,499$*

$$A_2(n,d) \geqslant cn \cdot \frac{2^n}{V(n,d-1)}.$$

Une telle borne est aussi vraie (avec une autre constante) pour les codes linéaires de paramètres $[n, n/2, d]$.

5 Décodage en liste

5.1 Définition et bornes

Le *décodage en liste* est une simple relaxation du problème de Hamming. On demande au décodeur de retourner tous les mots à distance $\lfloor \tau n \rfloor$ du mot reçu, avec $\tau \in [0,1]$. Quand $\lfloor \tau n \rfloor \leqslant t = \lfloor \frac{d-1}{2} \rfloor$, la capacité de correction du code, c'est le problème classique de Hamming. La question qui se pose pour le décodage en liste et de quantifier le gain que l'on peut espérer sur le rayon $\lfloor \tau n \rfloor$ en fonction de la taille de la liste renvoyée par le décodeur.

Définition 13. *Pour $\tau \in [0,1]$, et un entier $\ell \geqslant 1$, un code $C \in \mathbb{F}_q^n$ est dit (τ, ℓ)* décodable*, si toute boule de rayon $\lfloor \tau n \rfloor$ contient au plus ℓ mots de code.*

On a donc encore un problème combinatoire, et on retrouve le problème de Hamming quand $\ell = 1$. De même que pour la borne de Varshamov-Gilbert, on a une borne inférieure sur le taux de transmission.

Proposition 10. *Soit $\ell \geqslant 2$, et $\tau \in [0, 1 - 1/q]$. Il existe une famille de codes (τ, ℓ) décodables sur $\mathbb{F}_q$ avec $R \geq 1 - H_q(\tau) - 1/\ell$.*

Démonstration. Soit $n \in \mathbb{N}$ assez grand, et $R \in [0,1]$. Soit C un code de cardinal q^{Rn} tiré informément au hasard. Soit un ensemble fixé de $\ell + 1$ mots de C, la probabilité que ces mots soient dans une sphère de Hamming donnée de rayon e est

$$\left(\frac{V_q(n, \tau n)}{q^n}\right)^{\ell+1} \leqslant \left(q^{(H_q(\tau)-1)n}\right)^{\ell+1}.$$

La probabilité qu'il existe $\ell + 1$ mots de C dans une sphère de rayon $\lfloor \tau n \rfloor$ est au plus

$$|C|^{\ell+1} \cdot q^n \cdot \left(q^{(H_q(\tau)-1)n}\right)^{\ell+1}.$$

Pour tout R vérifiant $R < 1 - H_q(\tau) - 1/(\ell+1)$, cette quantité tend vers zéro avec n. Il existe donc des codes (τ, ℓ) décodables pour $R \geqslant 1 - H_q(\tau) - 1/\ell$. □

Pour les codes *linéaires*, il existe aussi une borne semblable : $R \geq 1 - H_q(\tau) - \varepsilon$, avec des listes de taille $1/\varepsilon$. Quand la taille de la liste est assez grande, la « capacité » du décodage en liste rejoint donc celle promise par Shannon pour le canal q-aire symétrique. Comme dans le problème originel de Shannon, il reste à établir le volet effectif de cette théorie, c'est-à-dire construire une famille explicite infinie de codes, constructibles en temps polymonial atteignant cette borne, et ensuite, trouver les algorithmes de décodage associés.

5.2 Décodage en liste des codes géométriques

La percée algorithmique de Guruswami et Sudan a été de construire une méthode capable de décoder tout code de Reed-Solomon de longueur n et de dimension k jusqu'à un rayon $\tau \leqslant n-\sqrt{kn}$. Ce rayon, obtenu avec un algorithme, est la « borne de Johnson ».

Proposition 11 (Borne de Johnson)**.** *Soit C un code quelconque de longueur n et de distance minimale d, et τ un entier tel que $\tau < n - \sqrt{n(n-d)}$. Le nombre de mots de C dans toute boule de rayon τ est borné par*

$$\frac{nd}{(n-\tau)^2 - n(n-d)}.$$

En particulier, un algorithme décodant en liste our un nombre d'erreurs inférieur au rayon de Johnson $J(n,d) = n - \sqrt{n(n-d)}$ retournera une liste de taille polynomiale. À l'opposé, il existe des codes C, tel qu'il existe des boules de rayon τ contenant un nombre exponentiel de mots de C dès que τ dépasse le rayon de Johnson. Il est à noter que c'est aussi le rayon obtenu par les mêmes auteurs pour les codes géométriques à un point, par une méthode généralisant le cas du genre zéro. C'est un résultat très spectaculaire car il permet notamment de recouvrer l'information même avec des taux d'erreurs approchant 100%.

Le rayon de Johnson $J(n,d) = n - \sqrt{n(n-\delta n)}$ ne prend pas en compte la taille de l'alphabet q. Il existe une version plus fine de la borne de Johnson, qui est plus pertinente pour les codes géométriques, puisque ceux-ci peuvent devenir longs sans croissance de la taille de l'alphabet.

5.3 Très grands alphabets

La capacité du décodage en liste, pour le canal q-aire symétrique étant $1 - H_q(p) - 1/\varepsilon$, pour des listes de tailles constantes $O(1/\varepsilon)$, on remarque que

$$1 - H_q(p) \geqslant 1 - p - \frac{1}{\log q}. \tag{1}$$

On peut donc obtenir un taux de transmission de $1 - p - \varepsilon$ pour corriger une fraction de p erreurs, pour une taille d'alphabet grande, exponentielle en $1/\varepsilon$.

Ce résultat a été approché algorithmiquement en utilisant des codes que Guruswami et Rudra ont appelés *codes de Reed-Solomon repliés*.

Définition 14. *Soit γ une racine n-ième de l'unité dans $\mathbb{F}_q$, n étant un entier divisant $q-1$, et $m \mid n$. Le* code de Reed-Solomon replié *m fois est l'ensemble des mots de longueur n/m à coefficients dans $\mathbb{F}_q^m$, telle que leur composantes est*

$$\left(f(\gamma^{jm}), \ldots, f(\gamma^{jm+(m-1)})\right), \quad i \in [0, \frac{n}{m} - 1]$$

pour $f \in \mathbb{F}_q[X]$ un polynôme de degré inférieur à k.

L'algorithme de Guruswami-Sudan se généralise alors.

Théorème 7. *Pour tout $\varepsilon > 0$, et $R \in]0,1[$, il existe une famille de codes de Reed-Solomon repliés de taux de transmission R, pouvant corriger une fraction $1 - R - \varepsilon$ d'erreurs. La taille de leur alphabet croît en*

$$\left(\frac{N}{\varepsilon^2}\right)^{O(\frac{1}{\varepsilon^2})}.$$

Par rapport à la promesse donnée par l'équation 1, on voit que la taille de l'alphabet croît avec la longueur. Dans un objectif de garder une taille d'alphabet fixe, il est naturel de considérer les codes géométriques. On peut ainsi « replier » les codes géométriques, et notamment ceux obtenus par Tsfasman Vladut et Zink avec les courbes asymptotiquement optimales. On obtient une meilleure dépendance en ε de la taille de l'alphabet :

$$\left(\frac{1}{\varepsilon}\right)^{O(\log(\frac{1}{\varepsilon}))}$$

qui notamment ne croît pas en fonction de la longueur du code [5].

5.4 Cas binaire

Bien que le résultat existentiel dise qu'il existe des codes binaires (τ, ℓ) décodable de taux de transmission supérieur à $1 - H_2(\tau) - \frac{1}{\ell}$, on ne sait en produire explicitement. Toutefois, en utilisant la méthode des codes concaténés, on a le résultat suivant.

Théorème 8. *Soit $\rho_1 \in [0, 1]$ et $\rho_2 \in [0, 1/2]$, il existe une famille de codes binaires linéaires, constructibles en temps polynomial en leur longueur, de taux*

$$(1 - \rho_1)\,(1 - H_2(\rho_2)) - \varepsilon$$

corrigeant en liste une fraction $\rho_1\rho_2$ d'erreur. L'algorithme de décodage est de complexité polynomiale en la longueur des codes.

L'idée est de concaténer un code de Reed-Solomon replié, défini sur un gros alphabet, avec un code binaire, décodable en liste optimal, dont on sait qu'il existe par un argument probabiliste. On remarque comme précédemment que la longueur du code binaire est logarithmique en la longueur désirée, et qu'une recherche exhaustive exponentielle en $\log n$, donc polynomiale en n, permettra de trouver ce bon code binaire décodable en liste. On combine donc un code explicite à forte structure sur un grand alphabet, avec un code non structuré mais « bon », obtenu par force brute.

6 Conclusion

Il est frappant que dans le contexte de la théorie algébrique des codes, le cas binaire ($q = 2$) est le cas le plus difficile, pour lequel la fonction $\alpha_q(\delta)$ est la moins connue, alors que pour des alphabets plus gros, les codes géométriques donnent d'excellent résultats, en pouvant de plus être considérés de manière asymptotique, ce qui n'est pas le cas des codes de Reed-Solomon. Cette difficulté de l'alphabet binaire existe tout aussi bien dans le cas classique (Hamming) que dans le cas du décodage en liste (Guruswami).

Un des intérêts du décodage en liste est qu'il permet de lutter contre le bruit adverse, avec le même taux que le codage statistique. Enfin, à toutes fins pratiques, dans le cas de codes de Reed-Solomon, il a été montré que la liste est de taille 1 avec un probabilité très proche de un. Ainsi, en relaxant le problème pour traiter les cas pathologiques, on tolère des taux d'erreur élevés,

avec quasiment unicité de la solution du mot de code retourné dans les cas non pathologiques.

Toutefois, les algorithmes de décodage en liste, bien que de complexité polynomiale, sont encore en l'état trop lourds pour être utilisés dans des schémas classiques de communication.

Références

[1] Ed F. Assmus and Jennifer D. Key. Polynomial codes and finite geometries. In Vera S. Pless, W.C. Huffman, and R.A. Brualdi, editors, *Handbook of coding theory*, volume II. North-Holland, 1998.

[2] Claude Berrou. *Codes et turbocodes*. Collection IRIS. Springer, March 2007.

[3] Philippe Gaborit and Gilles Zémor. Asymptotic improvement of the Gilbert-Varshamov bound for linear codes. *IEEE Transactions on Information Theory*, 54(9) :3865–3872, September 2008.

[4] Venkatesan Guruswami. *Algorithmic results in list decoding*. Now Publishers Inc, January 2007.

[5] Venkatesan Guruswami and Anindya C. Patthak. Correlated algebraic-geometric codes : Improved list decoding over bounded alphabets. *Mathematics of computation*, September 2007.

[6] James Hirschfeld. *Projective Geometries over Finite Fields*. Oxford University Press, USA, March 1998.

[7] Tao Jiang and Alexander Vardy. Asymptotic improvement of the Gilbert-Varshamov bound on the size of binary codes. *IEEE Transactions on Information Theory*, 50(8) :1655–1664, July 2004.

[8] Ron Roth. *Introduction to Coding Theory*. Cambridge University Press, 2006.

[9] Serge Vladut, Dmitry Nogin, and Michael Tsfasman. *Algebraic Geometric Codes : Basic Notions*. American Mathematical Society, September 2007.