Κωδικοποίηση

1. Εισαγωγή και γενικότητες

Ορισμός 1.1

(i) **Αλφάβητο** ονομάζεται το πεπερασμένο σύνολο των συμβόλων (πολλές φορές θα τα ονομάζουμε **γράμματα**) που χρησιμοποιούμε για να καταγράψουμε-διατυπώσουμε ένα μήνυμα. Το αλφάβητό μας θα είναι το πεπερασμένο σώμα $\mathbf{F}_q$ όπου το q είναι δύναμη πρώτου αριθμού.

(ii) Ένα **k-μήνυμα** αποτελείται από μια ακολουθία γραμμάτων των αλφαβήτου μας μήκους k. Είναι δηλαδή της μορφής: $a_1, a_2 \ldots, a_k$ με $a_i \in \mathbf{F}_q$.

(iii) Η αντίστοιχη **κωδική λέξη** x ενός k-μηνύματος είναι μια ακολουθία μήκους n. Είναι δηλαδή της μορφής $x = x_1, x_2 \ldots, x_n$ με $x_i \in \mathbf{F}_q$ και $n \geq k$. Όπου (σχεδόν) πάντα θα ισχύει ότι $x_1 = a_1, x_2 = a_2, \ldots, x_k = a_k$ ενώ τα υπόλοιπα n – k σύμβολα $(x_{k+1}, x_{k+2}, \ldots, x_n)$ θα τα λέμε **σύμβολα ελέγχου** (**check symbols** ή **control symbols**).

Συμβολισμός 1.2 Οι κωδικές λέξεις θα γράφονται x ή $x_1, x_2, \ldots, x_n$ ή $(x_1, x_2, \ldots, x_n)$ ή $x_1x_2\ldots x_n$.

Ορισμός 1.3 Θα ονομάζουμε **διάνυσμα λήψης** (ή **μήνυμα λήψης**) το διάνυσμα $y = y_1, y_2, \ldots, y_n$ που λαμβάνουμε. Το y εν γένει είναι διαφορετικό από το μήνυμα x που μας στέλνουνε. Το $e := y - x = e_1e_2\ldots e_n$ θα λέγεται **διάνυσμα λάθους** (ή απλά **λάθος**).

Ορισμός 1.4 Ένας **n-κώδικας** C είναι ένα υποσύνολο του $\mathbf{F}_q^n$. Ακριβέστερα ο κώδικας θα λέγεται **(n, k)-κώδικας**, όπου k το μήκος του μηνύματος που κωδικοποιούμε. Αν ο κώδικας C είναι $\mathbf{F}_q$-διανυσματικός υπόχωρος του $\mathbf{F}_q^n$ τότε θα λέγεται **(n, k)-γραμμικός κώδικας**. Τα στοιχεία του C θα είναι οι κωδικές λέξεις.

Όταν πάρουμε το y θα πρέπει να αποφασίσουμε ποια κωδική λέξη μας έχουν στείλει. Θα διαλέγουμε από το σύνολο C μια κωδική λέξη που διαφέρει λιγότερο από το y. Αυτό το αξίωμα ονομάζεται **αποκωδικοποίηση μέγιστης πιθανότητας**.

Ορισμός 1.5 **Απόσταση Hamming** d(x, y) δύο διανυσμάτων x, y στο $\mathbf{F}_q^n$, με

$$x = x_1, x_2 \ldots, x_n \text{ και } y = y_1, y_2 \ldots, y_n,$$

είναι το πλήθος των συντεταγμένων στις οποίες τα x και y διαφέρουν. Δηλαδή

$$d(x, y) = \#\left\{i \in \mathbf{N},\ 1 \leq i \leq n \middle|\ x_i \neq y_i\right\}$$

Ορισμός 1.6 **Βάρος (weight) Hamming** w(x) ενός διανύσματος $x = x_1, x_2 \ldots, x_n$ στο $\mathbf{F}_q^n$ είναι το πλήθος των μη-μηδενικών συντεταγμένων του x. Δηλαδή,

$$w(x) = \#\{i \in \mathbf{N},\ 1 \le i \le n \mid x_i \neq 0\}$$

Προφανώς, w(x) = d(x, 0).
Παράδειγμα 1.7 Έστω $C \subseteq \mathbf{F}_3^4$.
Το βάρος Hamming του 1201 είναι w(1201) = 3.
Η απόσταση Hamming των 1201 και 2211 είναι d(1201, 2211) = 2.

Παρατήρηση 1.8 Η απόσταση Hamming d(C) είναι μια μετρική στον $\mathbf{F}_q^n$ και το βάρος Hamming w είναι μια νόρμα στον $\mathbf{F}_2^n$.

Ορισμός 1.9 Αν $C \subseteq \mathbf{F}_q^n$ ένας (n, k)-κώδικας, **η ελάχιστη απόσταση $d_{min}(C)$ του κώδικα** είναι

$$d_{min}(C) = \min_{\substack{u,v \in C \\ u \neq v}} d(u, v)$$

Άρα, όταν παίρνουμε το y πρέπει να ελέγχουμε τις q^k κωδικές λέξεις για να βρούμε ποια έχει την μικρότερη απόσταση Hamming από το y. Προφανώς, αυτή η διαδικασία είναι αδύνατη για μεγάλα k και ένας από τους στόχους της θεωρίας κωδίκων είναι να βρει κώδικες με γρηγορότερους αλγόριθμους αποκωδικοποίησης.

Ορισμός 1.10 Το σύνολο $S_r(x) := \{y \in \mathbf{F}_q^n \mid d(x, y) \le r\}$ θα λέγεται η **σφαίρα ακτίνας r ως προς το x** $\in \mathbf{F}_q^n$.

Παράδειγμα 1.11 Έστω $C = \mathbf{F}_2^3$ τότε ο κύκλος με ακτίνα 1 ως προς το 100 είναι

$$S_1(100) = \{100, 000, 110, 101\}.$$

Στόχος Παίρνοντας σφαίρες κατάλληλης ακτίνας r κέντρου κωδικής λέξης θα πρέπει κατά το δυνατό να καλύπτει η ένωσή τους όλο το χώρο $\mathbf{F}_q^n$ ώστε να μπορούμε να αποκωδικοποιούμε όλα τα κωδικοποιημένα μηνύματα που λαμβάνουμε ενώ συγχρόνως η ακτίνα r θα πρέπει να είναι αρκετά μικρή ώστε οι σφαίρες να μην τέμνονται (ή εφάπτονται) και να μπορούμε να αποκωδικοποιούμε μονοσήμαντα.

Πρέπει πάντως να ισχύει ότι $r < \frac{1}{2} d_{min}(C)$.

Η σημασία της ιδέας της ελάχιστης απόστασης δίνεται από την

Πρόταση 1.12 Υποθέτουμε ότι ο C είναι γραμμικός κώδικας με ελάχιστη απόσταση $d_{min}(C) = d$. Ο C **ανιχνεύει** την ύπαρξη d–1 ή λιγότερων λαθών και **διορθώνει** το πολύ e λάθη όπου e τέτοιο ώστε $2e + 1 \le d$.

Απόδειξη Έστω ότι λάβαμε το μήνυμα y με απόσταση f από την κωδική λέξη x, όπου $f \le d–1$. Φανταζόμαστε ότι η x είναι η μεταδιδόμενη (αρχική) λέξη και y η λέξη που πήραμε τελικά. Δηλαδή έχουμε f λάθη κατά την μεταφορά. Επειδή d είναι η ελάχιστη

απόσταση του C η λέξη y καταλαβαίνουμε αμέσως ότι δεν μπορεί να είναι κωδική λέξη. Δηλαδή, ο κώδικας C ανακαλύπτει d – 1 ή λιγότερα λάθη.
Αν τώρα το μήνυμα y έχει απόσταση e από την κωδική λέξη x και $2e + 1 \leq d$ τότε δεν υπάρχει άλλη κωδική λέξη πιο κοντά στη y, διότι αν $d(y, x_1) \leq e$ για κάποια x_1 τότε θα ίσχυε

$$d(x, x_1) \leq d(x, y) + d(y, x_1) \leq e + e < d$$

άτοπο, διότι η ελάχιστη απόσταση του κώδικα C είναι d. Επομένως, υπάρχει μοναδική κοντινότερη λέξη του y και συνεπώς ο C διορθώνει e λάθη σ' αυτήν την περίπτωση.

Ένα από τα βασικά προβλήματα στη θεωρία κωδίκων είναι να ελαχιστοποιηθούν τα λάθη αλλά χωρίς να μειωθεί υποχρεωτικά **η αναλογία της πληροφορίας** $\frac{k}{n}$.
Κεντρικό πρόβλημα της Θεωρίας Κωδίκων είναι το εξής:

Δίνονται d, n, q *φυσικοί αριθμοί όπου q δύναμη πρώτου αριθμού. Να υπολογιστεί ο μέγιστος αριθμός διανυσμάτων, έστω* $A_q(n, d)$*, του διανυσματικού χώρου* $\mathbf{F}_q^n$ *τα οποία ανά δυο να έχουν απόσταση μεγαλύτερη ή ίση με* d. *Φυσικά, αν είναι δυνατόν να βρεθούν τα διανύσματα.*

Το πρόβλημα αυτό χαρακτηρίζεται και σαν discrete sphere packing problem (Conway & Sloan, 1988)

Ο επόμενος πίνακας μας δίνει κάποιες τιμές του $A_2(n, d)$ για d = 3

n	3	4	5	6	7	8	9	10
$A_2(n, 3)$	2	2	4	8	16	20	40	άγνωστος, μεταξύ 72 και 79

Ορισμός 1.13 Ένας κώδικας C ο οποίος **διορθώνει** την ύπαρξη t λαθών θα λέγεται **t-κώδικας διόρθωσης λαθών** (**t-error-correcting code**), ενώ ένας κώδικας C που **ανιχνεύει** e λάθη θα λέγεται **e-κώδικας ανίχνευσης λαθών** (**e-error-detecting code**).

Έστω τώρα C κώδικας ως προς το $\mathbf{F}_q$ μήκους n με πλήθος κωδικών λέξεων M. Υποθέτουμε ότι ο κώδικας είναι ένας t-κώδικας διόρθωσης λαθών. Υπάρχουν $\binom{n}{m}$ διανύσματα του $\mathbf{F}_q^n$ τα οποία να έχουν βάρος m στο $\mathbf{F}_q$. Αν $c \in C$ τότε μέσα στην σφαίρα $S_t(c)$ υπάρχουν $1 + (q-1)2c\binom{n}{1} + \ldots + (q-1)^t\binom{n}{t}$ διανύσματα του $\mathbf{F}_q^n$.

Θεώρημα 1.14 (Φράγμα του Hamming) Οι παράμετροι q, n, t, M ενός t-κώδικα διόρθωσης λαθών C ορισμένου στο σώμα $\mathbf{F}_q$ μήκους n με M κωδικές λέξεις ικανοποιούν την ανισότητα

$$M\left(1+(q-1)\binom{n}{1}+\ldots+(q-1)^t\binom{n}{t}\right)\le q^n.$$

Αν όλα τα διανύσματα του $\mathbf{F}_q^n$ είναι μέσα σε σφαίρες ακτίνας t κέντρου κωδικών λέξεων ενός (n, k)-γραμμικού κώδικα τότε παίρνουμε μια ειδική κατηγορία κωδίκων:

Στην περίπτωση όπου ο κώδικας είναι $\mathbf{F}_2$ – διανυσματικός υπόχωρος η ανισότητα γράφεται:

$$M\left(1+\binom{n}{1}+\ldots+\binom{n}{t}\right)\le 2^n.$$

Ορισμός 1.15 Ένας t-κώδικας διόρθωσης λαθών ορισμένος στο σώμα $\mathbf{F}_q$ θα ονομάζεται **τέλειος** αν στο θεώρημα 1.16 ισχύει η ισότητα.

Αν ο C είναι κώδικας όπως αυτός του θεωρήματος 1.14 με $d_{min}(C) = d = 2t + 1$, τότε αν διαγράψουμε τα τελευταία $d - 1$ σύμβολα πάλι έχουμε έναν κώδικα με όλες τις κωδικές λέξεις διαφορετικές. Ο κώδικας που προκύπτει έχει μήκος $n - d + 1$, και παίρνουμε το

Θεώρημα 1.16 (Φράγμα του Singleton) Αν ένας κώδικας $C \subseteq \mathbf{F}_q^n$ έχει ελάχιστη απόσταση d, τότε $|C| \le q^{n-d+1}$ ή αλλιώς $k \le n - d + 1$.

Ορισμός 1.17 Ένας κώδικας C θα λέγεται **διαχωρίσιμος μέγιστης απόστασης** (**maximum distance separable**) ή πιο απλά **κώδικας MDS** αν στο θεώρημα 1.16 ισχύει η ισότητα.

Παράδειγμα 1.18 Έστω ο κώδικας

$$C=\{000000,\ 001011,\ 010101,\ 011110,\ 100110,\ 101101,\ 110011,\ 111000\}\subseteq \mathbf{F}_2^6$$

όπου $d_{min}(C) = 3$. Έχουμε επομένως, $M = 8$, $q = 2$, $n = 6$, $d = 3$, $t = 1$. Το φράγμα του Hamming δίνει την ανισότητα $8\left(1+\binom{6}{1}\right)\le 2^6$, δηλαδή $56 < 64$. Αυτό σημαίνει ότι μόνο $64 – 56 = 8$ λέξεις μήκους 6 στο $\mathbf{F}_2^6$ βρίσκονται έξω από κάποια σφαίρα και δεν μπορούν να διορθωθούν σωστά (αυτό είναι προφανές και από το γεγονός ότι έχουμε 8 μη τεμνόμενες σφαίρες με 7 στοιχεία η κάθε μια). Ένα παράδειγμα μιας από τις 8 λέξεις που δεν μπορούν να διορθωθούν σωστά είναι η 100100 η οποία έχει απόσταση μεγαλύτερη ή ίση του 2 από όλες τις κωδικές λέξεις. Το φράγμα του Singleton μας δίνει $8 \le 2^4 = 16$, άρα ο C δεν είναι MDS.

Τα φράγματα μας δείχνουν τα όρια μας. Μεγάλη έκπληξη αποτέλεσε το **θεώρημα του Shannon** (1948) ότι υπάρχουν όσο καλοί κώδικες θέλουμε.