Tentamen Informatietheorie en Codes (kans A)

Opgave 1. *(7 punten)*
Gegeven is de kansverdeling $P = (0.5,\ 0.25,\ 0.1,\ 0.05.\ 0.05,\ 0.05)$.

(i) Bepaal de entropie $H(P)$ van P.

(ii) Laat zien dat er voor P twee echt verschillende binaire Huffman-coderingen bestaan (d.w.z. met verschillende codewoordlengtes) en geef de gemiddelde codewoordlengte L_H van deze coderingen aan.

(iii) Een *quaternaire* codering is een codering over een alfabet $\mathcal{A}$ met 4 elementen, bijvoorbeeld $\mathcal{A} = \{a, b, c, d\}$.

Bepaal een quaternaire Huffman-codering voor P en geef de gemiddelde codewoordlengte hiervan aan.

(iv) Voor een willekeurige stochast X laat zich een quaternaire codering $\mathcal{C}_4$ met gemiddelde codewoordlengte $L_{\mathcal{C}_4}(X)$ omzetten in een binaire codering $\mathcal{C}_2$ met gemiddelde codewoordlengte $L_{\mathcal{C}_2}(X) = 2L_{\mathcal{C}_4}(X)$ door de vier symbolen uit $\mathcal{A}$ te vervangen door de vier verschillende blokken van 2 bits, bijvoorbeeld $a \to 00$, $b \to 01$, $c \to 10$, $d \to 11$.

Bewijs dat er een quaternaire codering $\mathcal{C}_4$ bestaat zo dat voor de gemiddelde codewoordlengte $L_{\mathcal{C}_2}(X)$ van de hieruit verkregen binaire codering geldt dat

$$L_{\mathcal{C}_2}(X) \leq L_H(X) + 1$$

voor $L_H(X)$ de gemiddelde codewoordlengte van de binaire Huffman-codering van X.

Hint: Verleng binaire codewoorden van oneven lengte tot codewoorden van even lengte en vat deze op als codewoorden van een quaternaire codering.

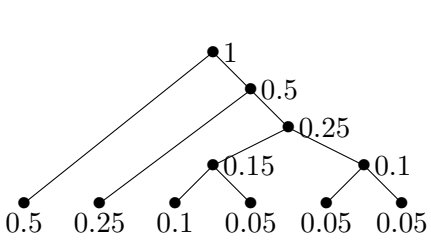
(v) Bewijs dat voor de gemiddelde codewoordlengtes L_{H_2} van de binaire en L_{H_4} van de quaternaire Huffman-codering geldt dat

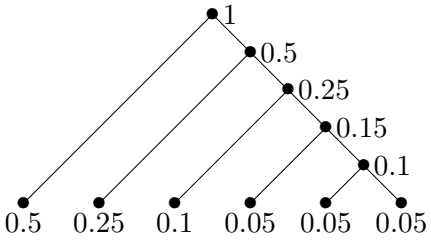
$$L_{H_2}(X) \leq 2L_{H_4}(X) \leq L_{H_2}(X) + 1.$$

Oplossing:

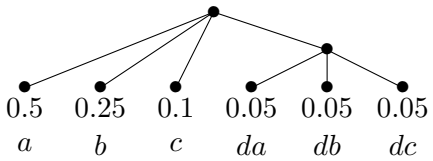
(i) De entropie is $0.5 + 0.25 \cdot 2 + 0.1 \cdot \log 10 + 3 \cdot 0.05 \cdot \log 20 \approx 1.98048$.

(ii) In de eerste stap van het Huffman algoritme worden noodzakelijk twee van de kansen 0.05 samengevoegd, maar in de tweede stap is er een keuze: de derde kans 0.05 wordt samengevoegd met de losse kans 0.1 of met de in de eerste stap samengevoegde kansen. Vervolgens zijn weer alle stappen eenduidig. De eerste optie leidt tot de linkerboom met bladhoogtes $1, 2, 4, 4, 4, 4$ en gemiddelde codewoordlengte $0.5 \cdot 1 + 0.25 \cdot 2 + (0.1 + 3 \cdot 0.05) \cdot 4 = 2$, de tweede tot de rechterboom met bladhoogtes $1, 2, 3, 4, 5, 5$ en gemiddelde codewoordlengte $0.5 \cdot 1 + 0.25 \cdot 2 + 0.1 \cdot 3 + 0.05 \cdot 4 + 2 \cdot 0.05 \cdot 5 = 2$, dus is de gemiddelde codewoordlengte van de Huffman-codering $L_H = 2$ (minder dan 0.2 boven de entropie).





(iii) Omdat P uit 6 kansen bestaat, maar $6 \equiv 0 \bmod (4-1)$ en niet $\equiv 1 \bmod (4-1)$, moeten we in de eerste stap virtueel een extra punt met kans 0 toevoegen. In de eerste stap worden daarom slechts drie kansen samengevoegd. Het Huffman algoritme geeft dan de volgende boom.



De gemiddelde codewoordlengte is $(0.5 + 0.25 + 0.1) \cdot 1 + 3 \cdot 0.05 \cdot 2 = 1.15$.

Merk op: Als in de eerste stap niet 3 maar 4 kansen worden samengevoegd, komt de kans $p = 0.1$ op een blad van hoogte 2 terecht, hierdoor is de boom niet optimaal en heeft een gemiddelde codewoordlengte van 1.25.

(iv) Een codewoord van oneven lengte laat zich verlengen tot een codewoord van even lengte door (bijvoorbeeld) een 0 aan te hangen. Merk op dat zich de zo veranderde codering nog steeds door een binaire boom laat representeren, bij de bladeren op oneven hoogte wordt gewoon een enkele tak ingevoegd. Een codewoord van even lengte laat zich direct omzetten naar een quarternair codewoord, omdat ieder blok van twee bits eenduidig met een letter uit $\mathcal{A}$ correspondeert.

De gemiddelde codewoordlengte verandert bij het verlengen van $L_H(X) = \sum_i p_i l_i$ naar $\sum_{i,l_i\text{even}} p_i l_i + \sum_{i,l_i\text{oneven}} p_i(l_i+1) = L_H(X) + \sum_{i,l_i\text{oneven}} p_i \leq L_H(X) + 1$, waarbij de ongelijkheid volgt omdat de p_i een kansverdeling met som 1 vormen.

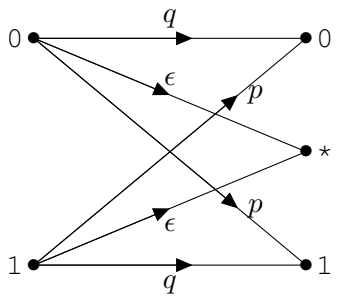
(v) De eerste ongelijkheid volgt uit de optimaliteit van de binaire Huffman-codering, want $L_{H_2}(X) \leq L_{\mathcal{C}_2}(X) = 2L_{H_4}(X)$ voor $\mathcal{C}_2$ de binaire codering verkregen uit de quaternaire Huffman-codering.

Volgens deel (iv) is er een quaternaire codering $\mathcal{C}_4$ met $2L_{\mathcal{C}_4}(X) \leq L_{H_2}(X) + 1$ en wegens de optimaliteit van de quaternaire Huffman-codering is $L_{H_4}(X) \leq L_{\mathcal{C}_4}(X)$, dus geldt $2L_{H_4}(X) \leq 2L_{\mathcal{C}_4}(X) \leq L_{H_2}(X) + 1$.

Opgave 2. *(6 punten)*
Een zeker kanaal heeft `0` en `1` als inputs en outputs, maar met een zekere kans ϵ is de output onleesbaar, dat noteren we met $\star$. De overgangsmatrix voor inputs in de volgorde `0`, `1` en outputs in de volgorde `0`, `1`, $\star$ is

$$\begin{pmatrix} 1-p-\epsilon & p & \epsilon \\ p & 1-p-\epsilon & \epsilon \end{pmatrix},$$

een schema voor dit kanaal is hieronder weergegeven, waarin $q = 1 - p - \varepsilon$.



(i) Bereken voor een input verdeling $P(X) = \begin{pmatrix}\pi & \pi'\end{pmatrix}$ de output verdeling $P(Y)$ en leg uit waarom de entropie $H(Y)$ van de output verdeling maximaal is voor een uniforme verdeling op de inputs.

(ii) Bewijs dat dit kanaal capaciteit $C = (1-\epsilon)(1-\log(1-\epsilon)) + q\log q + p\log p$ heeft.

Hint: Dit is een *uniform dispersief* kanaal (d.w.z. iedere rij van de overgangsmatrix is een permutatie van de eerste rij) en we weten dat hiervoor $H(Y|X) = H(a_1, \ldots, a_n)$ voor $a_1, \ldots, a_n$ de eerste rij van de overgangsmatrix.

(iii) Geef de capaciteit van het kanaal aan voor de speciale gevallen

(a) $\epsilon = 0$;

(b) $p = 0$.

Kan je deze speciale gevallen met bekende kanalen identificeren?

Oplossing: We schrijven altijd $q = 1 - p - \epsilon$.

(i) Voor een input verdeling $P(X)$ met $p(X = 0) = \pi$ en $p(X = 1) = \pi'$ is de output verdeling $P(Y)$ gegeven door

$$\begin{pmatrix}\pi & \pi'\end{pmatrix}\begin{pmatrix}q & p & \epsilon\\ p & q & \epsilon\end{pmatrix} = \begin{pmatrix}\pi q + \pi' p & \pi p + \pi' q & \epsilon\end{pmatrix}$$

en omdat ϵ vast ligt is de entropie $H(Y)$ maximaal als de eerste twee componenten gelijk zijn. Dit is het geval voor $\pi = \pi' = \frac{1}{2}$ en hiervoor is $\frac{1}{2}q + \frac{1}{2}p = \frac{1}{2}(1-\epsilon)$ en de output verdeling is $P(Y) = (\frac{1}{2}(1-\epsilon), \frac{1}{2}(1-\epsilon), \epsilon)$.

(ii) Omdat dit een uniform dispersief kanaal is, is $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(q,p,\epsilon)$ en dit wordt gemaximaliseerd door $H(Y)$ te maximaliseren. De verdeling $P(Y)$ met maximale entropie voor de outputs hebben we in deel (i) al bepaald, deze is $P(Y) = (\frac{1}{2}(1-\epsilon), \frac{1}{2}(1-\epsilon), \epsilon)$ en de entropie hiervan is

$$H(Y) = 2\cdot\frac{1-\epsilon}{2}\log\frac{2}{1-\epsilon} + \epsilon\log\frac{1}{\epsilon} = (1-\epsilon)(1-\log(1-\epsilon)) + \epsilon\log\frac{1}{\epsilon}.$$

De entropie $H(q,p,\epsilon)$ is gewoon $q\log\frac{1}{q} + p\log\frac{1}{p} + \epsilon\log\frac{1}{\epsilon}$, dus is

$$\begin{aligned} C = I(X;Y) &= H(Y) - H(q,p,\epsilon) = (1-\epsilon)(1-\log(1-\epsilon)) - q\log\tfrac{1}{q} - p\log\tfrac{1}{p}\\ &= (1-\epsilon)(1-\log(1-\epsilon)) + q\log q + p\log p. \end{aligned}$$

(iii) (a) Voor $\epsilon = 0$ is dit een binair symmetrisch kanaal, er geldt $q = 1 - p$ en de capaciteit is $C = 1 + q \log q + p \log p = 1 - H(p, q)$.

(b) Voor $p = 0$ is dit een binair doorhalingskanaal, er geldt $q = 1 - \epsilon$ en dus $C = (1 - \epsilon)(1 - \log(1 - \epsilon) + \log(1 - \epsilon)) = 1 - \epsilon$.

Opgave 3. *(8 punten)*

Zij $\mathcal{C}$ de lineaire code over $\mathbb{F}_2 = \{0, 1\}$ met voortbrengermatrix $G = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$.

(i) Laat zien dat $\mathcal{C}$ minimum afstand $d = 3$ heeft.

(ii) Bepaal een check matrix H voor $\mathcal{C}$.

(iii) Voor de decodering gebruik je de methode van syndroom decodering (met restklassenrepresentanten van minimaal gewicht). Wegens $d = 3$ is het duidelijk dat je een enkele fout altijd kunt verbeteren.

(a) Hoeveel verschillende foutpatronen met 2 fouten kan je met deze code verbeteren?

(b) Je verstuurt de codewoorden van $\mathcal{C}$ via een binair symmetrisch kanaal (BSC) met omklapkans p.
Wat is de kans p_E op een decoderingsfout bij syndroom decodering?
Bepaal p_E concreet voor $p = 0.02 = 2\%$.

(iv) Omdat $\mathcal{C}$ drie informatiebits met drie checkbits aanvult, kan je middels deze code $2^3 = 8$ verschillende letters coderen. Je gebruikt de volgende idenficaties tussen blokken van drie informatiebits en letters:

$$spatie \mathrel{\widehat{=}} 000,\ \texttt{A} \mathrel{\widehat{=}} 100,\ \texttt{D} \mathrel{\widehat{=}} 010,\ \texttt{E} \mathrel{\widehat{=}} 001,\ \texttt{K} \mathrel{\widehat{=}} 110,\ \texttt{L} \mathrel{\widehat{=}} 101,\ \texttt{N} \mathrel{\widehat{=}} 011,\ \texttt{O} \mathrel{\widehat{=}} 111.$$

Bijvoorbeeld codeer je $\texttt{A}$ als 100110 en $\texttt{K}$ als 110011.

Je ontvangt de output

$$110001 \quad 111000 \quad 110101 \quad 100111 \quad 100101 \quad 110110.$$

Decodeer de oorspronkelijke boodschap (in letters) door eventuele transmissiefouten te verbeteren.

Oplossing:

(i) Ieder van de basisvectoren heeft gewicht 3 en een lineaire combinatie van twee van de basisvectoren heeft gewicht 4, de som van alle drie de basisvectoren heeft gewicht 3.

Alternatief kan men op basis van de check matrix H uit deel (ii) zien dat geen twee kolommen van H lineair afhankelijk zijn, daarom is de minimum afstand minstens 3.

Let op: Dat de kolommen van G verschillend en dus geen twee kolommen van G lineair afhankelijk zijn, is geen geldig argument, kijk bijvoorbeeld naar de parity-check code van lengte 3 met $G = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, deze heeft duidelijk $d = 2$.

(ii) De voortbrengermatrix is al in standaard vorm, daarom is $H = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$ een check matrix.

(iii) (a) Er zijn $2^3 = 8$ verschillende syndromen. De kolommen van H zijn alle verschillend en corresponderen met de foutpatronen met één fout. Naast de nulvector ontbreekt dan alleen het syndroom 111 en dit komt op verschillende manieren door een foutpatroon van gewicht 2 tot stand, namelijk door 100001, 010010 en 001100. Door één van deze drie vectoren als representant voor de restklasse met syndroom 111 te kiezen, laat zich dus één foutpatroon met 2 fouten corrigeren.

(b) Alle foutpatronen met 0 en 1 fouten en één foutpatroon met 2 fouten worden correct gedecodeerd, dus is $p_E = 1 - ((1-p)^6 + 6p(1-p)^5 + p^2(1-p)^4)$.
Voor $p = 0.02 = 2\%$ is $p_E \approx 0.00532 = 0.532\%$.

(iv) Natuurlijk kunnen we de fout middels syndroom decodering lokaliseren en vervolgens verbeteren. Maar als we ervan uitgaan dat per ontvangen woord hoogstens één fout optreedt, kunnen we het correcte codewoord makkelijk achterhalen. Hierbij helpt de observatie dat in ieder codewoord behalve 000000 en 111000 de laatste drie bits altijd twee 1en bevatten. De decodering van de zes outputs is dan snel gevonden:
$110001 \to 110011 =$ `K`
$111000 \qquad\quad =$ `O`
$110101 \to 010101 =$ `D`
$100111 \to 100110 =$ `A`
$100101 \to 101101 =$ `L`
$110110 \to 100110 =$ `A`.

De originele boodschap was dus `KODALA`.

De standaard manier middels syndroom decodering levert dezelfde decodering op:

$(110001)H^{tr} = (010)$, dit is de 5^{e} kolom van H, de gecorrigeerde output is 110011, de informatiebits zijn dus $110 \mathrel{\widehat{=}}$ `K`;

$(111000)H^{tr} = (000)$, er is geen fout opgetreden, de informatiebits zijn dus $111 \mathrel{\widehat{=}}$ `O`;

$(110101)H^{tr} = (110)$, dit is de 1^{e} kolom van H, de gecorrigeerde output is 010101, de informatiebits zijn dus $010 \mathrel{\widehat{=}}$ `D`;

$(100111)H^{tr} = (001)$, dit is de 6^{e} kolom van H, de gecorrigeerde output is 100110, de informatiebits zijn dus $100 \mathrel{\widehat{=}}$ `A`;

$(100101)H^{tr} = (011)$, dit is de 3^{e} kolom van H, de gecorrigeerde output is 101101, de informatiebits zijn dus $101 \mathrel{\widehat{=}}$ `L`;

$(110110)H^{tr} = (101)$, dit is de 2^{e} kolom van H, de gecorrigeerde output is 100110, de informatiebits zijn dus $100 \mathrel{\widehat{=}}$ `A`.