

نظریه اطلاعات کلاسیک - بخش اول

وحیدکریمی پور- دانشکده فیزیک - دانشگاه صنعتی شریف

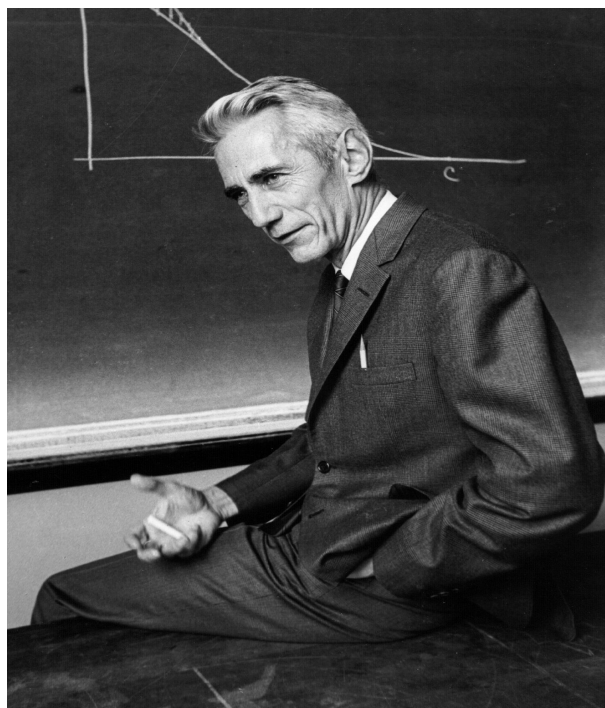
۱۸ آذر ۱۳۹۸

۱ مقدمه

فرض کنید که $X = \{x_1, x_2, \dots, x_n\}$ یک متغیر تصادفی با احتمالات $\{p(x_1), p(x_2), \dots, p(x_n)\}$ باشد. به این متغیر تصادفی می توان تابعی به شکل زیر نسبت داد.

$$H(X) := - \sum_{i=1}^n p(x_i) \log_2 p(x_i). \quad (1)$$

بدون اغراق می توان گفت که تمام نظریه اطلاعات کلاسیک بر روی این تابع که آن را تابع آنتروپی می خوانند و خواص و تعبیرهای آن بنا شده است. هدف ما در این درس آن است که اولاً خواص ریاضی این تابع و توابع وابسته به آن را استخراج کنیم، ثانیاً تعبیر و تفسیرهای این توابع را بفهمیم. نخستین کاری که باید بکنیم آن است که نشان دهیم این تابع سنج مناسبی برای اطلاعات است. این کاری است که در نخستین بخش این درس انجام می دهیم. در بخش های بعدی این درس مفاهیمی مثل اطلاعات شرطی و اطلاعات متقابل را معرفی می کنیم. پس از بررسی خواص ریاضی توابعی که برای اندازه گیری اطلاعات معرفی کرده ایم به فشرده سازی اطلاعات و حدی که برای این فشرده سازی وجود دارد می پردازیم.



شکل ۱: کلاود شانون، بنیانگذار نظریه اطلاعات

۲ مفهوم و اندازه اطلاعات

۱.۲ اطلاعات یک متغیر تصادفی

فرض کنید که آزمایش یا واقعه ای مثل X که نتایج پایشامدهای ممکن آن را با مجموعه $\{x_1, x_2, \dots, x_n\}$ نشان می دهیم اتفاق بیفتد و یک نتیجه معین مثل x_i حاصل شود. در این صورت می توانیم بررسییم که ما به عنوان ناظر یا مشاهده گر چه مقدار اطلاع حاصل کرده ایم، یا چه مقدار از عدم یقین ما نسبت به نتیجه های ممکن کاسته شده است. فرض ما این است که احتمالات وقوع یعنی $p(x_i)$ ها معلوم هستند. طبیعی است که با دانستن این احتمالات نمی توان یقیناً گفت که چه پیشامدی رخ خواهد داد. میزان عدم یقینی که نسبت به نتیجه داریم و در نتیجه مقدار اطلاعی که از مشاهده خود دریافت می کنیم، طبیعتاً تابعی از این احتمالات است. به عنوان مثال اگر داشته باشیم

$$P(x_1) = 1, \quad P(x_i) = 0, \quad i = 2, 3, \dots, N, \quad (۲)$$

آنگاه نتیجه هر آزمایشی از قبل معلوم است و ما از مشاهده آزمایش هیچ اطلاعی حاصل نمی کنیم، زیرا از قبل و با محاسبه تحلیلی می توانیم

بگوییم که همواره نتیجه x_1 حاصل خواهد شد. اما اگر داشته باشیم

$$P(x_i) = \frac{1}{N}, \quad (۳)$$

آنگاه هربار که آزمایش را انجام می دهیم یک نتیجه بدست می آید که به دانش ما اضافه می کند، دانشی که از قبل نداشتیم و نمی توانستیم با محاسبه ریاضی به آن برسیم. از نظر شهودی هرچقدر که پیشامدی که بوقوع پیوسته است محتمل تر بوده باشد اطلاعی که ماکسب کرده ایم کمتر و هرچقدر که آن پیشامد دور از انتظار بوده باشد تعجب ما از وقوع آن بیشتر و اطلاعی که ماکسب کرده ایم بیشتر خواهد بود. بنابراین اگر میزان اطلاع خود از وقوع پیشامد x_i را با h_i نشان دهیم می توانیم بگوییم که h_i می بایست نسبت معکوس با احتمال وقوع آن پیشامد یعنی p_i داشته باشد.

حال فرض کنید که یک آزمایش مرکب از دو واقعه مستقل (X, Y) شود که نتایج ممکن آن را با زوج های $i = 1 \dots m, j = 1 \dots n$ نشان می دهیم. هرگاه احتمال وقوع x_i را با p_i و احتمال وقوع y_j را با q_j نشان دهیم احتمال هرپیشامد (x_i, y_j) برابر خواهد بود با $p_i q_j$ و میزان اطلاعی که از وقوع این پیشامد کسب می کنیم برابر خواهد بود با $h(p_i q_j)$. انتظار داریم که میزان اطلاع ما در این مورد که دوپیشامد مستقل x_i و y_j رخ داده اند برابر با مجموع اطلاعاتی باشد که از وقوع پیشامد x_i به تنهایی و y_j به تنهایی کسب می کنیم بنابراین انتظار داریم که

$$h(p_i q_j) = h(p_i) + h(q_j). \quad (۴)$$

تنها تابعی که شرط فوق را برآورده کند و ضمناً نزولی باشد، تابع لگاریتم است بنابراین خواهیم داشت:

$$h(p_i) = \log_{\alpha} \frac{1}{p_i}, \quad (۵)$$

که در آن α ثابت است. ثابت α را می توان با شرط بهنجارش تعیین کرد. قرار می نهمیم که میزان اطلاع کسب شده ما از وقوع یک پدیده دو حالتی متساوی الاحتمال برابر با یک باشد، یعنی $h(1/2) = 1$. در نتیجه میزان ثابت α برابر می شود با ۲.

اگر یک آزمایش X را N بار انجام دهیم به طور متوسط $N p_i$ بار نتیجه x_i رخ خواهد داد و میزان اطلاعی که در هربار کسب می کنیم برابر خواهد بود با $\log_2(\frac{1}{p_i})$. میزان اطلاعی که ما به طور متوسط از وقوع نتایج آزمایش تصادفی X کسب می کنیم برابر خواهد بود با:

$$H(X) = -\frac{1}{N} \sum_x N p(x) \log_2 p(x) = -\sum_x p(x) \log_2 p(x). \quad (۶)$$

این تابع، تابع آنتروپی یا تابع شانون نیز خوانده می شود. دقت کنید که تابع $p \log \frac{1}{p}$ در فاصله $p \in [0, 1]$ یک تابع مثبت است بنابراین $H(X)$

یک تابع مثبت است.

■ تمرین: با مراجعه به گوگل، فرکانس حروف انگلیسی را پیدا کرده و سپس تابع آنتروپی را برای آن پیدا کنید.

۲.۲ اطلاعات دو متغیر تصادفی

هرگاه دو متغیر تصادفی (X, Y) داشته باشیم که لزوماً از هم مستقل نباشند تابع آنتروپی یا اطلاعات به طور طبیعی به شکل زیر تعریف می شود:

$$H(X, Y) := - \sum_{x,y} p(x, y) \log_2 p(x, y) \quad (۷)$$

درحالتی که دو متغیر تصادفی مستقل باشند یعنی $p(x, y) = p(x)q(y)$ ، رابطه بالا بدست می دهد که $H(X, Y) = H(X) + H(Y)$.

این تعریف به همین شکل به بیش از دو متغیر تصادفی تعمیم می یابد به این معنا که تعریف می کنیم:

$$H(X, Y, Z) = - \sum_{x,y,z} p(x, y, z) \log_2 p(x, y, z). \quad (۸)$$

۳.۲ اطلاعات شرطی

دو متغیر تصادفی X, Y که با توزیع آنها تابع $P(x, y)$ مشخص می شود در نظر می گیریم. فرض کنید که مقدار یکی از متغیرهای تصادفی مثل Y را می دانیم و این مقدار برابر است با y . در این صورت توزیع متغیر تصادفی X عوض خواهد شد و تبدیل خواهد شد به توزیع $P(X|y)$ که در آن y یک پارامتر است و X مقادیر متغیر را بخود می گیرد. می دانیم که:

$$P(x|y) := \frac{P(x, y)}{p(y)}, \quad \sum_x p(x|y) = 1. \quad (۹)$$

در نتیجه اطلاعات باقیمانده در متغیر تصادفی X برابر خواهد بود با:

$$H(X|y) := - \sum_x P(x|y) \log_2 P(x|y) \quad (۱۰)$$

اگر بخواهیم بدانیم که به طور متوسط دانستن یک مقدار از Y چه مقدار اطلاعات در X باقی می‌گذارد باید روی $H(X|y_j)$ متوسط بگیریم. بنابراین خواهیم داشت:

$$\begin{aligned} H(X|Y) &= \sum_y p(y)H(X|y) = - \sum_{x,y} P(y)P(x|y) \log_2 P(x|y) \\ &= - \sum_{x,y} P(x,y) \log_2 P(x|y) = - \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(y)} \\ &= H(X,Y) - H(Y). \end{aligned} \quad (11)$$

دقت کنید که به همان دلیلی که تابع $H(X)$ مثبت است تابع $H(X|y)$ و در نتیجه تابع $H(X|Y)$ نیز مثبت خواهند بود. $H(X|Y)$ را اطلاعات X مشروط به Y می‌خوانیم و این کمیت بیان‌کننده میزان اطلاعات باقیمانده در X است هرگاه ما مقادیر Y را دانسته باشیم. باید توجه داشت که این تابع متقارن نیست یعنی $H(X|Y) \neq H(Y|X)$.

از رابطه (11) به نتیجه زیر می‌رسیم:

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X). \quad (12)$$

اگر دو متغیر تصادفی X, Y مستقل باشند آنگاه دانستن Y هیچ تاثیری در اطلاعات باقیمانده در X نخواهد داشت و در نتیجه $H(X|Y) = H(X)$ و بنابر (12)، $H(X,Y) = H(X) + H(Y)$. بالعکس هرگاه X و Y کاملاً به هم وابسته باشند انتظار داریم که دانستن Y برای دانستن X نیز کفایت کند یعنی هیچ اطلاعی در X باقی نگذارد یعنی $H(X|Y) = 0$ که با توجه به رابطه (12) به این معناست که $H(X,Y) = H(Y)$. این رابطه نیز معنای شهودی روشنی دارد.

۴.۲ اطلاعات متقابل

اطلاعات متقابل در دو متغیر تصادفی X و Y به شکل زیر تعریف می‌شود:

$$I(X : Y) := H(X) + H(Y) - H(X,Y). \quad (13)$$

این کمیت نسبت به دو متغیر تصادفی X و Y متقارن است. با توجه به رابطه (12) می‌توان آن را به شکل زیر بازنویسی کرد:

$$I(X : Y) := H(X) - H(X|Y). \quad (14)$$

این رابطه معرف چه چیزی است؟ قبل از آنکه مقدار Y را بدانیم، اطلاعات موجود در X با $H(X)$ سنجیده می‌شد. با دانستن Y این اطلاعات به $H(X|Y)$ تقلیل پیدا می‌کند. بنابراین تفاوت این دو میزان اطلاعی است که Y درباره X حمل می‌کند. بعداً نشان خواهیم داد که $I(X : Y)$ یک کمیت نامنفی است.

■ مثال: منبع

$$X = \{000(1/2), 111(1/2)\} \quad (15)$$

را که در آن اعداد داخل پرانتز احتمالات رشته‌ها را نشان می‌دهند در نظر بگیرید. برای این منبع داریم $H(X) = 1$ می‌توانیم رشته‌ها را خود را به ترتیب زیر تنظیم کنیم:

۱ - آیا همه اعداد صفر هستند؟

در هر دو صورت جواب آری یا خیر ما به رشته مورد نظری که سوال کننده در نظر گرفته است پی می‌بریم. یعنی یک سوال برای رسیدن به رشته مورد نظر کفایت می‌کند.

حال منبع زیر را در نظر بگیرید:

$$X = \{000(1/4), 111(1/4), 001(1/4), 110(1/4)\} \quad (16)$$

حال می‌توانیم سوالات خود را به شکل زیر تنظیم کنیم:

۱ - آیا اکثریت بیت‌ها صفر هستند؟

۲ - آیا همه بیت ها مثل هم هستند؟

در این صورت با دوسوال به رشته مورد نظری رسمیم و $H(X)$ نیزبرابریا ۲ است.

حال فرض کنید که از قبل کسی به ما گفته است که رقم سمت راست این رشته برابر با 1 است. در این صورت می دانیم که رشته مورد نظر یکی از رشته های $\{001, 111\}$ است. اکنون با دانستن رقم سمت راست که آن را یک متغیر تصادفی مثل Y در نظر می گیریم، کافی است که با پرسیدن تنها یک سوال به رشته مورد نظر دست پیدا کنیم. در واقع داریم

$$H(X | 1) = 1, \quad H(X | 0) = 1, \quad \rightarrow H(X|Y) = 1. \quad (17)$$

یعنی وقتی رقم سمت راست تعیین می شود، اطلاعات لازم (تعداد سوال های لازم) برای رسیدن به رشته مورد نظر از ۲ به ۱ تقلیل پیدا می کند. به این دلیل می گوییم که اطلاعات متقابل X ، Y برابر است با

$$I(X : Y) = H(X) - H(X | Y) = 2 - 1 = 1. \quad (18)$$

به این ترتیب دانستن یک رقم سمت راست به اندازه یک بیت در مورد کل رشته به ما اطلاع داده است. حال فرض کنید که کسی به ما دو رقم سمت راست را بگوید. در این صورت می بینیم که کل رشته به طور کامل تعیین می شود و سوالی برای پرسیدن باقی نمی ماند. در این جا داریم:

(19)

$$H(X|00) = 0, \quad H(X|01) = 0, \quad H(X|10) = 0, \quad H(X|11) = 0 \quad \rightarrow H(X|Y) = 0$$

در نتیجه خواهیم داشت:

$$I(X : Y) = H(X) - H(X | Y) = 2 - 0 = 2. \quad (20)$$

در این جا اطلاعات متقابل بین دو رقم سمت راست و کل رشته زیاد و برابر با ۲ بیت است.

۳ خواص ریاضی توابع اطلاعات

در این بخش خواص ریاضی توابع اطلاعات را بررسی می‌کنیم. تقریباً همه این خواص از یک قضیه ساده ولی مهم بدست می‌آیند.

■ قضیه: تابع انتروپی شانون در رابطه زیر صدق می‌کند که در آن q هر تابع توزیع احتمال دلخواهی است:

$$H(X) \leq - \sum_x p(x) \log_2 q(x). \quad (21)$$

که در آن تساوی فقط وقتی برقرار می‌شود که دو توزیع احتمال یکی باشند.

■ اثبات: بارسم کردن تابع لگاریتم و تابع $x - 1$ ، می‌توان نشان داد که تابع لگاریتم در خاصیت زیر صدق می‌کند:

$$\ln x \leq (x) - 1, \quad (22)$$

که در آن تساوی فقط برای $x = 1$ برقرار می‌شود. حال قرار می‌دهیم $x = \frac{q(x)}{p(x)}$ و در نتیجه

$$\ln \frac{q(x)}{p(x)} \leq \frac{q(x)}{p(x)} - 1, \quad \forall x, \quad (23)$$

که در آن تساوی فقط وقتی برقرار می‌شود که $q(x) = p(x)$. در نتیجه

$$\sum_x p(x) \ln \frac{q(x)}{p(x)} \leq \sum_x q(x) - \sum_x p(x) = 0, \quad (24)$$

که همان نامساوی ای است که می‌خواستیم ثابت کنیم. دقت کنید که نامساوی (۲۲) فقط برای $\ln(x)$ درست است و نه برای لگاریتم

در پایه ۲. ولی پس از بدست آوردن رابطه (۲۴) ما می‌توانیم طرفین آن را در هر عددی ضرب کنیم و رابطه ای بدست آوریم که در هر پایه

ای برای لگاریتم صحیح است. حال فرض کنید که

$$\sum_x p(x) \log \frac{q(x)}{p(x)} = 0. \quad (25)$$

این تساوی را به شکل زیر بازنویسی می‌کنیم

$$\sum_x p(x) \left(\log \frac{q(x)}{p(x)} - \left(\frac{q(x)}{p(x)} - 1 \right) \right) = 0. \quad (26)$$

حال دقت می کنیم که بنا بر نامساوی (23) جملات داخل پرانتز همگی کوچک تراز یا مساوی با صفر هستند. صفر شدن این جمع به این معناست که همه این جملات برابر با صفر هستند که با توجه به نامساوی (22) به معنای آن است که برای همه i ها $q(x) = p(x)$. یعنی دو تابع توزیع احتمال یکی هستند.

■ نتیجه ۱: مقدار بیشینه تابع اطلاعات $H = \sum_{x=1}^M p(x) \log \frac{1}{p(x)}$ برابر است با $\log M$ و این مقدار بیشینه فقط برای توزیع یکنواخت $\{p(x) = \frac{1}{M}\}$ برقرار می شود

■ اثبات: در قضیه قبلی قرار می دهیم $q(x) = \frac{1}{M}$. در نتیجه خواهیم داشت:

$$\sum_{x=1}^M p(x) \log \frac{1}{p_x} = H - \log M \leq 0, \quad (27)$$

که در آن تساوی فقط وقتی برقرار می شود که $p(x) = \frac{1}{M}$.

■ نتیجه ۲: برای دو متغیر تصادفی X, Y نامساوی زیر برقرار است

$$H(X, Y) \leq H(X) + H(Y), \quad (28)$$

که در آن تساوی فقط وقتی برقرار می شود که X, Y متغیرهای مستقل باشند.

■ اثبات: تابع توزیع دو متغیر را با $p(x, y)$ نشان می دهیم. در نتیجه خواهیم داشت:

$$p_1(x) := \sum_y p(x, y), \quad p_2(y) := \sum_x p(x, y). \quad (29)$$

حال تابع توزیع $q(x, y) := p_1(x)p_2(y)$ را در نظریه گیریم و از قضیه ای که ثابت کردیم استفاده می کنیم:

$$\sum_{x,y} p(x, y) \log \frac{q(x, y)}{p(x, y)} \leq 0 \quad (30)$$

که در آن تساوی فقط وقتی برقرار می شود که $p(x, y) = q(x, y) = p_1(x)p_2(y)$. اما نامساوی بالا را وقتی بازنویسی کنیم چیزی نیست جز

$$H(X, Y) \leq H(X) + H(Y), \quad (31)$$

که می توان آن را به شکل زیر نیز نوشت:

$$H(X|Y) \leq H(X). \quad (32)$$

این نامساوی در واقع بیان می کند که دانستن یک متغیر تصادفی دیگر مثل Y همواره از انترپی موجود در متغیر X کم می کند (چیزی در باره آن به ما می گوید و اطلاعات ما را افزایش می دهد). اگر بخواهیم از زبان زندگی روزمره کمک بگیریم می توانیم بگوییم که معنای نامساوی (31) به عنوان مثال این است که: اطلاعات موجود در جمله « فردا هوا ابری است و باران می بارد » کمتر از مجموع اطلاعاتی است که در دو جمله « فردا هوا ابری است » و « فردا هوا بارانی است » می باشد. دلیل این امر آن است که معمولاً بین ابری بودن هوا و بارانی بودن آن یک همبستگی وجود دارد که به ما اجازه می دهد از اولی بتوانیم وجود دومی را حدس بزنیم. بنابراین کسی که هر دو جمله را به ما می گوید دوبرابر کسی که فقط یکی از جملات را به ما می گوید به ما اطلاع نمی دهد. این مثال طبیعتاً یک مثال کلامی است و کمی نیست.

■ نتیجه ۳: اطلاعات متقابل یک کمیت نامنفی است. این نتیجه از تعریف اطلاعات متقابل و نتیجه ۲ بدست می آید.

■ تمرین: احتمالات نسبی دو متغیر تصادفی مطابق با جدول زیر داده شده است: منظور از احتمال نسبی این است که برای بدست آوردن احتمال می بایست اعداد درون جدول را بهنجار کنید طوری که مجموع تمام احتمالات برابر با یک شود.

y_6	y_5	y_4	y_3	y_2	y_1	$p(x, y)$
2	4	2	5	0	2	x_1
5	1	0	6	3	0	x_2
0	3	0	0	4	9	x_3
1	7	4	3	1	3	x_4
3	1	2	0	2	0	x_5
0	5	3	2	7	0	x_6

الف: تابع آنتروپی شانون $H(X, Y)$ را حساب کنید.

ب: تابع های آنتروپی $H(X)$ ، $H(Y)$ ، $H(X|Y)$ و $H(Y|X)$ را محاسبه کنید.

■ تمرین: مثالی از یک تابع توزیع احتمال $P(x, y)$ ارائه دهید که برای بعضی از مقادیر متغیرها داشته باشیم: $P(x | y) \leq P(x)$ و برای بعضی دیگر داشته باشیم $P(x | y) \geq P(x)$.

■ قضیه: اطلاعات تابع محدبی از توزیع احتمال است. به عبارت دیگر اگر P_1 و P_2 دو تابع توزیع احتمال و $P_0(x) = \lambda P_1(x) + (1 - \lambda) P_2(x)$ ترکیب خطی محدب آنها باشد آنگاه

$$H_0(X) \geq \lambda H_1(X) + (1 - \lambda) H_2(X). \quad (۳۳)$$

به اصطلاح می گوئیم که اطلاعات یک تابع محدب روبه پایین است که به یادماندن شکل آن را نیز در ذهن آسان می کند.

معنای فیزیکی این نامساوی این است که مخلوط کردن دو تابع توزیع احتمال همواره باعث افزایش آنتروپی می شود. به عنوان مثال می توانید فرض کنید که P_1 نشان دهنده وضعیتی است که در آن مولکول های یک گاز همه در طرف چپ یک ظرف جمع شده اند و P_2 نشان دهنده وضعیتی است که در آن همه مولکول ها در طرف راست ظرف جمع شده اند. در این صورت واضح است که وقتی دیواره جدا کننده را از وسط ظرف بر

می داریم وضعیتی بوجود می آید که مولکول ها تمام ظرف را اشغال می کنند و و این وضعیت جدید دارای انتروپی بیشتری نسبت به وضعیت های قبلی است.

■ اثبات : بازهم از نامساوی اساسی ای که ثابت کردیم استفاده می کنیم. با کمی خلاصه نویسی در نمادها خواهیم داشت:

$$\begin{aligned}
 & H_0 - \lambda H_1 - (1 - \lambda) H_2 \\
 = & \sum p_0 \log \frac{1}{p_0} - \lambda \sum p_1 \log \frac{1}{p_1} - (1 - \lambda) \sum p_2 \log \frac{1}{p_2} \\
 = & \sum (\lambda p_1 + (1 - \lambda) p_2) \log \frac{1}{\lambda p_1 + (1 - \lambda) p_2} - \lambda \sum p_1 \log \frac{1}{p_1} - (1 - \lambda) \sum p_2 \log \frac{1}{p_2} \\
 = & \lambda \sum p_1 \log \frac{p_1}{\lambda p_1 + (1 - \lambda) p_2} + (1 - \lambda) \sum p_2 \log \frac{p_2}{\lambda p_1 + (1 - \lambda) p_2} \geq 0, \quad (34)
 \end{aligned}$$

که در خط آخر از نامساوی اساسی استفاده کرده ایم.

■ تمرین: برای یک سکه که دو روی آن با اعداد 0 و 1 نشان داده می شوند، دو تابع توزیع احتمال به شکل زیر در نظر بگیرید:

$$\{P(0) = 1/2, P(1) = 1/2\}, \quad \{Q(0) = 1/3, Q(1) = 2/3\}. \quad (35)$$

حال درستی رابطه تحدب را برای آنتروپی شانون تحقیق کنید.

۱.۳ تعریف کانال کلاسیک

منظور از یک کانال کلاسیک عملگری است که یک آنزامل تصادفی X را به آنزامل تصادفی Y تبدیل می کند. بهترین مثال آن هر نوع کانال مخابراتی کلاسیک است. X را ورودی کانال و Y را خروجی آن می نامیم. یک کانال بدون نوفه کانالی است که خروجی آن دقیقاً با ورودی آن برابر است. بجزاین کانال ایده آل هر کانال دیگری علائم ورودی $x_i \in X$ را با احتمالات معین $P(y_j|x_i)$ به علائم خروجی $y_j \in Y$ تبدیل می کند. هرگاه در خروجی کانال علامت y_j را دریافت کنیم می توانیم احتمال شرطی این که چه علامت x_i ای منجر به این علامت در خروجی شده است

را حساب کنیم. در واقع داریم:

$$\begin{aligned} P(x_i|y_j) &= \frac{P(x_i, y_j)}{P(y_j)} = \frac{P(y_j, x_i)}{\sum_{x_i} P(y_j, x_i)} \\ &= \frac{P(y_j|x_i)P(x_i)}{\sum_{x_i} P(y_j|x_i)P(x_i)} \end{aligned} \quad (۳۶)$$

در آخرین عبارت $P(x_i)$ مشخصه منبع X و $P(y_j|x_i)$ مشخصه کانال است و هر دو معلوم هستند.

■ تمرین: یک کانال کلاسیک با احتمالات شرطی زیر توصیف می شود:

$$P(0|0) = 1 - p, \quad P(1|1) = 1 - q. \quad (۳۷)$$

هرگاه آنزامل ورودی به صورت

$$X = \{P(0) = a, P(1) = 1 - a\}, \quad (۳۸)$$

آنزامل خروجی را پیدا کنید. سپس کمیت های زیر را حساب کنید:

$$H(X), \quad H(Y), \quad H(X|Y), \quad H(Y|X), \quad I(Y : X), \quad I(X : Y). \quad (۳۹)$$

■ قضیه: اطلاعات پردازش شده در یک کانال $I(X; Y)$ تابع محدبی از احتمالات ورودی X است.

در یک کانال آنزامل ورودی را با X و آنزامل خروجی را با Y نشان می دهیم. احتمالات شرطی $P(y|x)$ در واقع مشخصه کانال هستند و احتمال تبدیل پیام x به y را در طول کانال نشان می دهند و ربطی به احتمال پیام های ورودی ندارند. حال هرگاه برای آنزامل ورودی دو تابع توزیع احتمال $P_1(x)$ و $P_2(x)$ و جمع محدب آنها یعنی $P_0(x) = \lambda P_1(x) + (1 - \lambda)P_2(x)$ را در نظر بگیریم آنگاه با توجه به تعاریف زیر:

$$\begin{aligned} P(y) &= \sum_x P(y|x)P(x), \\ P(x, y) &= P(y|x)P(x), \end{aligned} \quad (۴۰)$$

خواهیم داشت

$$\begin{aligned} P_0(x, y) &= \lambda P_1(x, y) + (1 - \lambda) P_2(x, y) \\ P_0(y) &= \lambda P_1(y) + (1 - \lambda) P_2(y). \end{aligned} \quad (41)$$

باترکیب این روابط با تعریف اطلاعات متقابل و هم چنین محذب بودن تابع اطلاعات اثبات قضیه کامل می شود.

■ تمرین: این اثبات را به صورت کامل بنویسید.

■ تمرین: یک کانال در نظر بگیرید که به صورت زیر عمل می کند:

$$P(0|0) = 1 - p, \quad P(1|1) = 1 - q. \quad (42)$$

آزمایش های ورودی را به صورت زیر در نظر بگیرید:

$$X_0 := \{P(0) = 1/2, P(1) = 1/2\}, \quad X_1 := \{Q(0) = 1/3, Q(1) = 2/3\}. \quad (43)$$

الف: درستی رابطه تحذب را برای اطلاعات متقابل تحقیق کنید.

ب: هرگاه که در مقصد، گیرنده رشته خروجی 000 را دریافت کند حساب کنید که احتمال اینکه در مبداء هرکدام از رشته های $x_1 x_2 x_3$ ارسال شده باشند چقدر است؟

■ تمرین: جفت متغیر تصادفی (X, Y) را مطابق جدول زیر در نظر بگیرید: Y ناشی از انداختن یک طاس است که مقادیر ۱ تا ۶ را به خود می گیرد و X نیز دو مقدار متفاوت یک سکه است که مقادیر a یا b را اختیاری می کند.

6	5	4	3	2	1	(X, Y)
0.05	0.05	0.04	0.08	0.1	0.2	a
0.05	0.1	0.06	0.15	0.02	0.1	b

(44)

کمیت های زیر را حساب کنید: الف: $H(X)$ ، $H(Y)$ ، $H(X, Y)$ ، $H(X|Y)$ ، $H(Y|X)$ و $I(X; Y)$.

۴ فشرده سازی اطلاعات درغیاب نوفه

بهترین کاربرد برای فهم فشرده سازی اطلاعات مطالعه یک مثال ساده است. فرض کنید که هدف ما ارسال متن هایی است که تنها از چهار حرف الفبا به نام های A ، B ، C و D تشکیل شده است. یک روش برای ارسال این متن ها آن است که حرف های چهارگانه فوق را با بیت های 0 و 1 که در مخابرات دیجیتال معمول است، به ترتیب زیر کد کنیم.

$$A \rightarrow 00$$

$$B \rightarrow 01$$

$$C \rightarrow 10$$

$$D \rightarrow 11. \quad (45)$$

در این صورت به ازای هر حرف دوبیت مخابره کرده ایم. حال سوال این است که آیا می توانیم یک روش کد کردن به کار ببریم که در آن طول

به ازای هر حرف تعداد بیت هایی که به طور متوسط مخابره می کنیم کمتر از 2 باشد؟

فرض کنید که این حروف در متن های یاد شده با احتمالات زیر ظاهر می شوند:

$$P(A) = \frac{1}{8} \quad P(B) = \frac{1}{8} \quad P(C) = \frac{1}{4} \quad P(D) = \frac{1}{2}. \quad (46)$$

حال روش کدگذاری زیر را به کار می بریم:

$$D \rightarrow 0$$

$$C \rightarrow 10$$

$$B \rightarrow 110$$

$$A \rightarrow 111. \quad (47)$$

در این روش کدگذاری برای بعضی از حروف بیش از دو بیت به کار برده ایم ولی اگر طول متوسط کدهایی را که برای حروف به کار برده ایم محاسبه کنیم نتیجه جالب توجه خواهد بود. این طول متوسط برابر است با:

$$\langle l \rangle = \sum_{i=1}^4 l_i \times p_i = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = \frac{7}{4}. \quad (48)$$

بنابراین با یک کدگذاری مناسب توانسته ایم طول متوسط رشته بیت هایی را که برای مخابره پیام بکار برده ایم از ۲ به ۴/۷ تقلیل دهیم. ضمناً باید دقت کنیم که این نحوه کدگذاری هیچ نوع ابهامی درباره متنی که مخابره شده است در بر ندارد و هر رشته ای از بیت ها به طور یکتا به متن اولیه بازگشایی می شود. به عنوان مثال رشته زیر

$$010001000110111. \quad (49)$$

بدون ابهام به متن زیرگشوده می شود و متن دیگری برای بازگشایی آن قابل تصور نیست

$$DCDDCDDBA. \quad (50)$$

این که چه نوع کد هایی یکتاگشاهستند موضوعی است که مادردهای آینده به آن خواهیم پرداخت و فعلاً موضوع بحث مانیت. ولی یک نکته مهم را باید ذکر کنیم: هرگاه آنتروپی متغیر تصادفی $X = \{A, B, C, D\}$ را با احتمالات ذکر شده حساب کنیم حاصل آن برابر خواهد بود با:

$$\begin{aligned} H(X) &= \sum_{i=1}^4 p_i \log_2 \left(\frac{1}{p_i} \right) = \frac{1}{2} \times \log_2(2) + \frac{1}{4} \times \log_2(4) + \frac{1}{8} \times \log_2(8) + \frac{1}{8} \times \log_2(8) \\ &= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4}. \end{aligned} \quad (51)$$

بنابراین در این مثال خاص طول متوسط کدگذاری ای که به کار بردیم با میزان اطلاعات موجود در متن برابر است. آیا این یک خصلت عمومی است؟ ادامه این درس و ضمیمه آن پاسخی به این سوال را در بر دارد.

■ تمرین: فرض کنید که متنی که می خواهیم ذخیره کنیم از همان الفبای ساده چهارحرفی با همان احتمالات تشکیل شده است اما این بار می خواهیم حروف را به صورت دوتایی کد کنیم. ضمناً می دانیم که بین حروف یک همبستگی وجود دارد: این همبستگی به احتمالات زیر مشخص شده است:

$$P(x|x) = 5/8, \quad P(y \neq x|x) = 1/8. \quad (52)$$

الف: احتمالات مربوط به تمام حروف دوتایی را محاسبه کنید.

ب: حال حروف دوتایی را طوری کد کنید که بیشترین فشردگی حاصل شود. تعداد بیت های لازم برای ذخیره هر حرف به طور متوسط چقدر است؟ اگر این همبستگی وجود نداشت تعداد بیت های لازم برای ذخیره هر حرف چقدر می شد؟

ج: احتمالات مربوط به تمام حروف سه تایی را بدست بیاورید. فرض کنید که همبستگی ها فقط دوتایی است.

■ تمرین: فرض کنید که یک متن از حروف زیر با احتمالات نوشته شده تشکیل شده است. یک کد بهینه برای این حروف بنویسید به نحوی که هم طول متوسط پایین باشد و هم رشته ای صفر و یک ها به صورت یکتا به حروف نگاشته شود.

$$P(A) = \frac{1}{32} \quad P(B) = \frac{1}{32} \quad P(C) = \frac{1}{16} \quad P(D) = \frac{1}{8} \quad P(E) = \frac{1}{4} \quad P(F) = \frac{1}{2} \quad (53)$$

■ تمرین: فرض کنید که یک متن از حروف زیر با احتمالات نوشته شده تشکیل شده است. یک کد بهینه برای این حروف بنویسید به نحوی که هم طول متوسط پایین باشد و هم رشته ای صفر و یک ها به صورت یکتا به حروف نگاشته شود.

$$\begin{aligned} P(A) &= \frac{1}{128} & P(B) &= \frac{1}{128} & P(C) &= \frac{1}{64} & P(D) &= \frac{1}{32} \\ P(E) &= \frac{1}{16} & P(F) &= \frac{1}{8} & P(G) &= \frac{1}{4} & P(H) &= \frac{1}{2} \end{aligned} \quad (54)$$

■ تمرین: فرض کنید که یک متن از حروف زیر با احتمالات نوشته شده تشکیل شده است. یک کد بهینه برای این حروف بنویسید به نحوی که هم طول متوسط پایین باشد و هم رشته ای صفر و یک ها به صورت یکتا به حروف نگاشته شود.

$$\begin{aligned} P(A) &= \frac{1}{256} & P(B) &= \frac{1}{256} & P(C) &= \frac{1}{128} & P(D) &= \frac{1}{128} \\ P(E) &= \frac{1}{128} & P(F) &= \frac{1}{32} & P(G) &= \frac{1}{16} & P(H) &= \frac{1}{8} \\ P(K) &= \frac{1}{4} & P(L) &= \frac{1}{2}. \end{aligned} \quad (55)$$

بعد از ذکر این مثال می خواهیم بفهمیم که در حالت کلی چگونه می توان اطلاعات موجود در یک منبع X را فشرده کرد. فرض کنید که منبع متن هایی تولید می کند که این متن ها از الفبای $A = \{x_1, x_2, \dots, x_N\}$ تشکیل شده اند و احتمال ظاهر شدن هر حرف مثل x_i در این متن ها با p_i داده

می شود. بنابراین یک منبع رامی توان به عنوان یک متغیر تصادفی با اطلاعات معین $H(X)$ در نظر گرفت. برای سادگی فرض کنید که N توانی از 2 است یعنی $N = 2^n$. حال اگر بدون توجه به احتمالات ظاهر شدن حروف مختلف بخواهیم متن هارامخبره کنیم می توانیم هر حرف الفبای A را بایک رشته n تایی از بیت های 0 و 1 کدگذاری کنیم. در این صورت برای هر متن که شامل M حرف است تعداد Mn بیت مصرف می کنیم یابه عبارت دیگر به ازای هر حرف الفبا n بیت مصرف می کنیم. ولی می توانیم روش کدگذاری بهتری به ترتیب زیر بکار ببریم.

به جای اینکه تک تک حروف الفبا را کدگذاری کنیم، سعی می کنیم که رشته M تایی را به K رشته کوچکتر یعنی رشته هایی به طول m تقسیم کنیم. بنابراین داریم

$$M = Km.$$

دقت کنید که m نیز به اندازه کافی بزرگ است. تعداد کل رشته های m حرفی برابر است با N^m . ولی نکته این است که ما تنها می بایست رشته های نمونه را کد کنیم. بعنوان مثال درست است که هر حرف از حروف الفبای انگلیسی با یک فرکانس مشخص در نوشتارهای انگلیسی ظاهر می شود اما رشته هایی m حرفی مثل

AAAAAAAAAAAAAAAAAAAAA

یا

AAABBBBAAABBBAAABBB

رشته هایی هستند که به ندرت ظاهر می شوند. دقت کنید که فعلا کاری به معنای جملات نداریم بلکه تنها به فرکانس ظاهر شدن حروف توجه داریم. در دو مثال بالا منظور ما این نیست که چنین رشته هایی از نظر معنایی نادر هستند بلکه منظور ما این است که این رشته ها از نظر فراوانی حروف ظاهر شده نایاب هستند. در عوض رشته ای مثل

ABQUQIPQUTNVIABURQOR (۵۶)

از نظر فراوانی حروف یک رشته نمونه است. یعنی اینکه اگر یک رشته بلند از یک متن انگلیسی را انتخاب کنیم و فراوانی حروف آن را با رشته بالا مقایسه کنیم اختلاف چندانی مشاهده نمی کنیم.

اگر تعداد حروف انگلیسی را همراه با حروف اضافه و فاصله ها تعداد 32 تا بگیریم آنگاه هر کدام از حروف را با 5 بیت می توانیم کد کنیم. به این ترتیب یک رشته با طول N می بایست با $5N$ بیت کد کنیم.



شکل ۲: یک رشته بلند را به رشته های با طول m تقسیم و سپس هرکدام را به عنوان یک رشته نمونه کد می کنیم.

حال دقت می کنیم که احتمال ظاهر شدن بسیاری از رشته ها آنقدر ناچیز است که نیازی به کد کردن آنها نیست و با کد کردن تنها رشته های متعارف (رشته هایی که زیاد ظاهر می شوند) چیزی ازدست نمی دهیم. به این ترتیب یعنی با کد کردن تنها رشته های متعارف ما قادر می شویم که بیت های کمتری برای مخابره متن های منبع X مصرف کنیم. اما رشته های متعارف کدام ها هستند؟ و کد کردن آنها چقدر باعث فشرده شدن پیام هاست. در هر رشته m حرفی به شرطی که m به اندازه کافی بزرگ باشد به تقریب تعداد mp_1 حرف آن x_1 ، mp_2 حرف آن x_2 و mp_N تا حرف آن x_N خواهد بود. هر قدر که طول رشته یعنی m بیشتر باشد، افت و خیز تعداد واقعی حرف ها حول این مقادیر متوسط کمتر خواهد بود. حال سوال می کنیم که چه تعداد رشته متعارف با طول m وجود دارد. اگر این تعداد را با Q_m نشان دهیم خواهیم داشت

$$Q_m = \frac{m!}{(mp_1)!(mp_2)! \cdots (mp_N)!} \quad (57)$$

اما با استفاده از تقریب استرلینگ می توانیم بنویسیم:

$$\log_2 Q_m = \log_2 \left(\frac{m!}{(mp_1)!(mp_2)! \cdots (mp_n)!} \right) \approx m \left(\sum_{i=1}^N p_i \log_2 \frac{1}{p_i} \right) \equiv mH(X) \quad (58)$$

که در آن تابع $H(X)$ به صورت زیر تعریف شده است:

$$H(X) := \sum_{i=1}^N p_i \log_2 \left(\frac{1}{p_i} \right) \quad (59)$$

بنابراین تعداد جملات متعارف با طول m با تقریب بسیار خوب برابر خواهد بود با

$$Q_m \approx 2^{mH(X)} \quad (60)$$

حال اگر تعداد جملات متعارف برابر باشد با مقدار فوق، می توانیم هر کدام از این جملات را با یک رشته بیت های 0 و 1 کدگذاری کنیم و مسلم است که تعداد بیت هایی که برای این کار احتیاج داریم برابر است با $mH(X)$. از آنجا که هر رشته دارای m حرف بوده است مثل این است که در عمل برای مخابره هر حرف $H(X) := k$ بیت بکار برده ایم. از آنجا که $H(X) \leq \log_2 N = n$ نتیجه می گیریم که در ارسال بیت ها برای مخابره پیام صرفه جویی مهمی انجام داده ایم زیرا با این روش کد کردن که آن را *Block coding* می گوئیم برای هر حرف به جای n بیت $H(X)$ بیت مصرف کرده ایم که از n کمتر است.

آنچه که در بالا گفته شد محتوای کلی قضیه شانون در مورد کدگذاری بدون نوفه بود. ولی چگونه می توان این حرف را دقیق کرد؟ چگونه می توان تعریف دقیقی از رشته های متعارف بدست داد؟ با کد نکردن رشته های غیر متعارف چه مقدار مرتکب خطای شویم؟ آیا بیش از این هم می توان پیام های منبع X را فشرده کرد؟ برای پاسخ به این سوالات سعی می کنیم ابتدائاً تعریف دقیقی از مفاهیم گفته شده بدست دهیم.

■ تمرین: فرض کنید که الفبای مورد استفاده شما از حروف زیر با فرکانس های داده شده تشکیل شده است:

k	g	h	g	f	e	d	c	b	a	
1/32	1/32	1/32	1/32	1/16	1/16	1/8	1/8	1/4	1/4	$P(x)$

(61)

برای این حروف یک کد یکتا گشا و بهینه بنویسید.

تا کنون بحث ما در باره رشته های نمونه یا متعارف یک بحث تقریبی بود. حالا می خواهیم این تعریف و نتایج ناشی از آن را به طور دقیق تر

بررسی کنیم.

۵ نگاهی دوباره به اطلاعات شرطی و اطلاعات متقابل

بعد از فهم رشته های متعارف و فشرده سازی می توانیم نگاهی دوباره به اطلاعات شرطی و اطلاعات متقابل بیندازیم. از این زاویه جدید می توانیم تعریف متفاوتی برای تابع $H(X)$ پیدا کنیم. یاد گرفتیم که تعداد رشته های متعارف m حرفی برابر است با $2^{mH(X)}$. این حرف به این معناست که اگر کسی یک رشته معین را به عنوان سوال برای مادر نظر گرفته باشد و از ما بخواهد در یک مسابقه به اصطلاح « بیست سوالی » بپرسیدن سوال هایی که پاسخ آنها تنها آری یا خیر است به آن رشته معین دست پیدا کنیم در بهترین حالت می بایست تعداد $mH(X)$ بار سوال کنیم. زیرا بهترین نحوه سوال کردن نحوه ای است که در آن تعداد رشته های باقیمانده را به نصف مقدار قبلی کاهش می دهد و $2^{mH(X)}$ را به $2^{mH(X)-1}$ ، $2^{mH(X)-2}$ و سرانجام به ۱ تقلیل می دهد. مطالب بالا را می توانیم به شکل زیر تعمیم دهیم. فرض کنید که یک متغیر تصادفی با آنتروپی $H(X)$ داریم. رشته هایی طولانی با طول m در نظر می گیریم. مجموعه رشته های نمونه برابر است با $2^{mH(X)}$. ما با پرسیدن $mH(X)$ سوال می توانیم به یک رشته مورد نظر برسیم.

حال فرض کنید که کسی اطلاعی از یک متغیر تصادفی دیگر مثل Y به ما داده باشد. این متغیر تصادفی دیگر می تواند یک چیز با ربط مثل رقم های سمت راست این رشته یا تعداد صفرهای رشته و نظایر آن یا یک چیز بی ربط مثل وضع هوای امروز باشد. در هر صورت تابع توزیع رشته ها از $P(X)$ به $P(X|y)$ تغییر می کند. در این صورت تعداد رشته هایی که می بایست جستجو کنیم به $2^{mH(X|y)}$ تقلیل پیدا می کند. در نتیجه با پرسیدن $mH(X|y)$ سوال می توانیم به رشته مورد نظر برسیم. بنابراین دانستن مقدار y تعداد سوالات لازم برای رسیدن به رشته مورد نظر x را از $mH(X)$ به $mH(X|y)$ کاهش داده است. یعنی اینکه دانستن y به اندازه $mH(X) - mH(X|y)$ بیت به ما اطلاع داده است. اگر روی y متوسط بگیریم، و بر m تقسیم کنیم، چیزی که بدست می آوریم برابر است با

$$I(X : Y) = H(X) - H(X | Y) \quad (۶۲)$$

همان اطلاعات متقابل است.

در واقع مهمترین مثال مشخص از ای نوع وقتی است که متغیر تصادفی X رشته های ورودی یک کانال کلاسیک و متغیر تصادفی Y رشته های خروجی همان کانال را تعیین می کند. تابع توزیع $P(x, y)$ احتمال این است که رشته x فرستاده و رشته y دریافت شود. حال سوال می کنیم اگر رشته y دریافت شده باشد، به طور متوسط چه مقدار اطلاعات در مورد رشته ارسال شده داریم؟ یا به طور متوسط با چه تعداد سوال می توانیم رشته ورودی را بفهمیم. معمولاً یک کانال دارای خطاست به این معنی که وقتی رشته ای را دریافت می کنیم، احتمال دارد که رشته ای که فرستاده شده همین رشته نباشد بلکه در اثر خطای کانال، رشته x به این رشته تبدیل شده است. به طور متوسط تعداد این رشته ها برابر است با $2^{mH(X|Y)}$.

هدف ما یافتن رشته x از روی رشته دریافت شده است. تمام آنچه که در بالا گفتیم، در این جا معنا پیدا می کند به این معنا که از روی رابطه (۶۲) می فهمیم که اگر کانال دارای هیچ نوع خطایی نباشد، آنگاه دانستن رشته خروجی دقیقاً رشته ورودی را تعیین می کند و در نتیجه

$$H(X|Y) = 0 \quad \longrightarrow \quad I(X : Y) = H(X). \quad (۶۳)$$

هر چه که خطای کانال بیشتر باشد، استقلال رشته های ورودی و خروجی از هم بیشتر شده و در نهایت وقتی که خطای کانال به حدی می رسد که این دو رشته از هم مستقل می شوند خواهیم داشت

$$H(X|Y) = H(X) \quad \longrightarrow \quad I(X : Y) = 0. \quad (۶۴)$$

۶ تعریف دقیق از رشته های متعارف

بهترین کار برای فهم این بخش شروع از یک مثال است. این مثال را به دقت بررسی می کنیم و سپس مفاهیمی را که در این مثال طرح خواهیم کرد به حالت های کلی و دلخواه تعمیم می دهیم. فرض کنید منبع نشان داده شده در شکل (۳) رشته هایی به طول $m = 18$ تولید می کند. این رشته ها از دو حرف A و B تشکیل می شوند. بنابراین الفبایی که این پیام ها از آن ساخته می شود فقط این دو حرف را دارد. در هر رشته این دو حرف با احتمالات نشان داده شده در شکل تولید می شوند و احتمال تولید حرف ها نیز در هر رشته ای از هم مستقل است. می دانیم که تعداد کل رشته های به طول 18 برابر است با:

$$\mathcal{N}(m = 18) = 2^{18} = 262144. \quad (۶۵)$$

حال از خود سوال می کنیم که معمولاً این منبع چه نوع رشته هایی را تولید می کند؟ برای پاسخ به این سوال می بایست به احتمالات تولید رشته ها نگاه کنیم. این منبع رشته ای مثل $AAAAA \dots A$ را که کلاً از حرف A تشکیل شده است با احتمال خیلی بالا تولید می کند. اما چنین رشته ای بسیار نایاب است. در عوض هرکدام از رشته هایی مثل $AAAABBBBAAABB \dots B$ که مخلوطی از دو حرف الفبا را دارند با احتمال کمتری تولید می کند ولی تعداد چنین رشته هایی آنقدر زیاد است که عملاً مثل این است که این منبع فقط چنین رشته هایی را تولید می کند. در واقع تعداد متوسط حرف های A در این رشته ها برابر با $mp = 12$ و تعداد متوسط حرف B در این رشته ها برابر است با $m(1-p) = 6$ است.



ABABBABVBAABBAABA

BBABBABVBBABBAABA

ABABAABVBBAAABABA

AAAAAAABVBBAAABBA

$$P(A) = p = \frac{2}{3}$$

$$P(B) = q = \frac{1}{3}$$

شکل ۳: یک منبع که رشته های به طول ۱۸ را با احتمالات نشان داده شده صادر می کند.

بیشترین رشته ها نیز از نوعی هستند که این تعداد متوسط از حروف را در بر دارند. در نگاه اول این رشت ها را رشته های نمونه خوانده ایم. تعداد این رشته ها برابر است با:

$$N_0 = \binom{18}{12} = 18564. \quad (66)$$

اما خب این عدد نسبت به کل تعداد رشته های ممکن کمی کوچک است. حال از خود می پرسیم که احتمال کل تولید این رشته ها چقدر است. از آنجا که هر رشته از این نوع با احتمال $(\frac{2}{3})^{12}(\frac{1}{3})^6$ تولید می شود، احتمال کل تولید این رشته ها که آن را با P_0 نشان می دهیم برابر است با:

$$P_0 = \left(\frac{2}{3}\right)^{12} \left(\frac{1}{3}\right)^6 \binom{18}{12} = 0.1962. \quad (67)$$

یعنی ۲۰ درصد رشته ها از این نوع هستند که البته رقم قابل ملاحظه ای است. اما به هر حال ۸۰ درصد رشته ها از نوع دیگری هستند. دلیل اش هم خیلی روشن است چون فقط رشته هایی را در نظر گرفته ایم که دقیقاً تعداد ۱۲ تا حرف A و ۶ تا حرف B دارند. بهتر است رشته های نمونه را با سخت گیری کمتری تعریف کنیم. برای این کار احتیاج داریم که علاوه بر میانگین تعداد حرف های A میزان افت و خیز حول این تعداد را

نیز در نظر بگیریم. می دانیم که میزان افت و خیز برابر است با:

$$\sigma = \sqrt{mp(1-p)} = \sqrt{18 \frac{2}{3} \frac{1}{3}} = 2. \quad (68)$$

بنابراین اگر رشته های متعارف را طوری تعریف کنیم که علاوه بر مقدار دقیق میانگین یعنی ۱۲ تا حرف A به اندازه یک انحراف معیار از این تعداد را نیز در خود جای دهند، شاید تعریف بهتری از رشته های میانگین بدست بیاوریم. این تعداد را با N_1 نشان می دهیم. این تعداد برابر است با:

$$N_1 = \sum_{x=10}^{14} \binom{18}{x} = 105774. \quad (69)$$

از خود می پرسیم که احتمال کل تعداد این رشته ها چقدر است. در این جا باید دقت کنیم که این رشته ها همه هم احتمال نیستند. اگر چه احتمال آنها به هم نزدیک است. این احتمال کل برابر است با:

$$P_1 = \sum_{x=10}^{14} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{18-x} \binom{18}{x} = 0.7907. \quad (70)$$

بنابراین با در نظر گرفتن تنها یک انحراف معیار که مقدار کم ۲ نسبت به ۱۸ است توانسته ایم تعداد رشته های نمونه را از عدد ۱۸۵۶۴ به تعداد ۱۰۵۷۷۴ برسانیم و احتمال کل را از 0.1962 به 0.7907 برسانیم. از خود می پرسیم که اگر به جای یک انحراف معیار دو انحراف معیار یعنی به میزان ۴ را اختیار کنیم حاصل چه خواهد شد. پاسخ به سادگی بدست می آید.

$$N_2 = \sum_{x=8}^{16} \binom{18}{x} = 199121, \quad (71)$$

و

$$P_2 = \sum_{x=8}^{16} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{18-x} \binom{18}{x} = 0.9788. \quad (72)$$

یعنی تقریباً تمام رشته ها از این نوع هستند. به عبارت دیگر منبع مورد نظر با احتمال 97.88 درصد رشته هایی از این نوع تولید می کند و نکته مهم این است که تعداد کل این رشته ها یعنی N_2 هنوز کمتر از تعداد کل رشته های ممکن یعنی $2^{18} = 262144$ تا است.

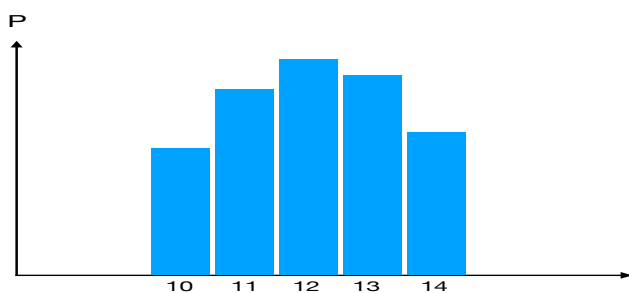
به یک خصلت مهم دیگر نیز باید توجه کنیم. بیایید احتمال رشته های مختلف را وقتی که طول رشته برابر با 18 است و رشته های نمونه به اندازه یک انحراف معیار انتخاب می شوند حساب کنیم. احتمال کل رشته هایی که در آن تعداد A ها برابر با x است برابر است با:

$$P(x) = \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{18-x} \binom{18}{x} \quad 10 \leq x \leq 12 \quad (73)$$

برای x های مختلف (با یک انحراف معیار) این احتمالات برابر است با:

$$P(10) = 0.1156 \quad P(11) = 0.1682 \quad P(12) = 0.1962 \quad P(13) = 0.1811 \quad P(14) = 0.1294. \quad (۷۴)$$

این احتمالات در شکل (۴) نشان داده شده اند. ب



شکل ۴: احتمال این که منبع رشته های نمونه مختلف را تولید کند. محور افقی تعداد A ها را در رشته ای به طول 18 نشان می دهد. دیده می شود که تابع توزیع به یک تابع یکنواخت نزدیک است.

به نظر می رسد که تابع توزیع احتمال تقریباً یک نواخت است. البته این یکنواختی خیلی تقریبی است. هر چه که طول رشته ها را بیشتر کنیم، این یکنواختی دقیق تر می شود. به عنوان مثال طول رشته ها را بجای 18، 72 انتخاب می کنیم که در این صورت یک انحراف معیار برابر خواهد بود با 4. در نتیجه رشته های نمونه ای که با یک انحراف معیار انتخاب می کنیم مقدار x آنها از 44 تا 52 است. برای این رشته ها خواهیم داشت:

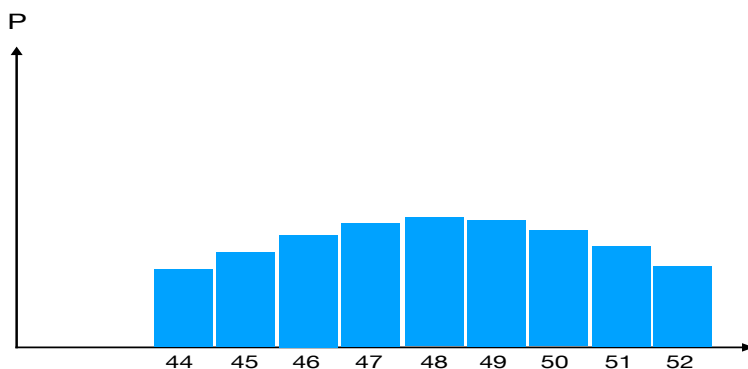
$$P(x) = \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{72-x} \binom{72}{x} \quad 44 \leq x \leq 52. \quad (۷۵)$$

و به طور صریح

$$P(44) = 0.0589 \quad P(45) = 0.0734 \quad P(46) = 0.0862 \quad P(47) = 0.0954 \quad P(48) = 0.0993$$

$$P(49) = 0.0973 \quad P(50) = 0.0895 \quad P(51) = 0.0772 \quad P(52) = 0.0624. \quad (76)$$

این احتمالات در شکل (۵) نشان داده شده اند.



شکل ۵: احتمال این که منبع رشته های نمونه مختلف را تولید کند. محور افقی تعداد A ها را در رشته ای به طول 72 نشان می دهد. دیده می شود که تابع توزیع به تابع یکنواخت خیلی نزدیک تر شده است.

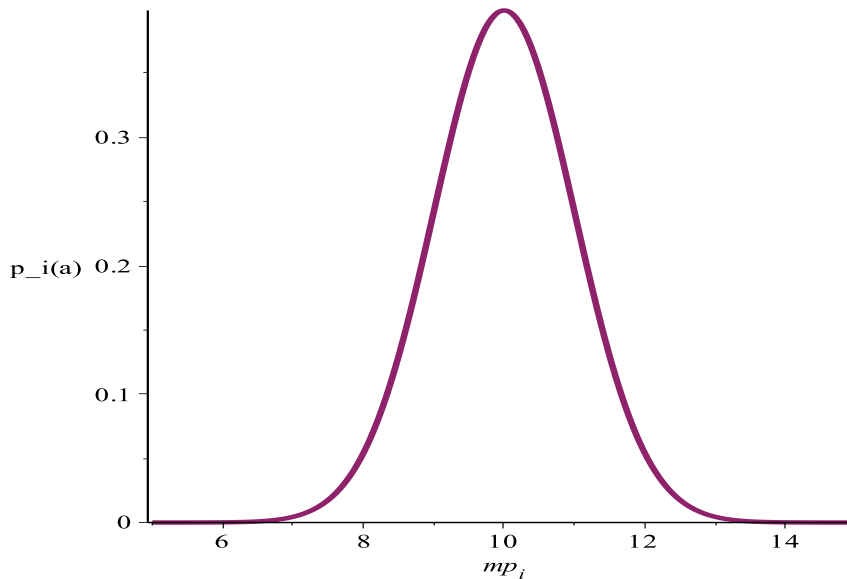
به وضوح معلوم است که تابع توزیع یکنواخت تر شده است. این روند با افزایش طول رشته ها ادامه می یابد و برای طول های بسیار بزرگ تابع توزیع واقعا یک نواخت می شود. از این مثال ها درس های زیر را می آموزیم:

■ یک - برای طول های بسیار بزرگ کافی است چند انحراف معیار را حول رشته میانگین در نظر بگیریم تا احتمال کل رشته های نمونه هرچقدر که می خواهیم به یک نزدیک شود.

■ دو- برای طول های بسیار بزرگ همه این رشته های نمونه با احتمال تقریبا مساوی تولید می شوند.

■ سه - برای طول های بسیار بزرگ و با در نظر گرفتن چند انحراف معیار منبع عملا فقط رشته های نمونه تولید می کند. تعداد این رشته های نمونه با تقریب بسیار خوبی برابر است با:

$$\mathcal{N}_{typ} \approx 2^{NH(X)} \quad (77)$$



شکل ۶: تعداد حرف های x_i در یک رشته مثل a از یک تابع احتمال گاوسی تبعیت می کند و بنابراین متوسط این تعداد برابر با mp_i است ولی این تابع توزیع یک پهنا به اندازه $\sigma_i = \sqrt{mp_i(1-p_i)}$ دارد که نشان دهنده این است در خیلی از رشته ها تعداد x_i با مقدار متوسط mp_i متفاوت است. در این نمودار $p_i(a)$ احتمال این است که یک رشته دارای یک تعداد معین x_i باشد.

و هر کدام از این رشته ها نیز با احتمال

$$2^{-NH(X)} \quad (78)$$

تولید می شود.

آنچه را که تا کنون در مورد این مثال گفتیم می توانیم به زبان دقیق تر برای حالت کلی بیان کنیم. تا کنون رشته های نمونه ^۱ را به طور شهودی تعریف کرده ایم و گفته ایم که این رشته ها رشته هایی هستند که تعداد حروف x_i در آنها برابر با mp_i باشد. اما می دانیم که تعداد x_i هیچگاه دقیقاً برابر با این مقدار نیست بلکه همواره یک افت و خیز حول این مقدار میانگین وجود دارد. اگر در هر مکان از یک رشته m تایی، وجود یک متغیر مثل x_i را با احتمال p_i و نبود آن را با احتمال $1-p_i$ در نظر بگیریم، آنگاه با یک تابع احتمال دو جمله ای (و در حد m های بزرگ با یک تابع گاوسی) روبرو هستیم که تعداد متوسط x_i را برابر با mp_i و واریانس آن را برابر با $\sigma_i = \sqrt{mp_i(1-p_i)}$ بدست می دهد، (شکل (۷)). از روی همین شکل واضح است که می بایست رشته های نمونه یا متعارف را به شکل بهتری تعریف کنیم. بنابراین از خود می پرسیم که یک رشته متعارف دقیقاً چه رشته ای است؟ در این تعریف حتماً می بایست یک حد و اندازه وجود داشته باشید. بدون این حد و اندازه یا معیار نمی توان دقیقاً گفت که یک رشته مثل آیا یک رشته نمونه است یا خیر؟

^۱ typical sequence

رشته $\alpha = \alpha_1 \alpha_2 \alpha_3 \dots \alpha_m$ را در نظر بگیرید. تعداد حروف x_j در این رشته را با $f_j(\alpha)$ نشان دهید. تعداد حروف x_j در رشته های به طول m به طور متوسط برابر است با mp_j و واریانس توزیع احتمال حول این مقدار متوسط برابر است با $\sigma_j := \sqrt{mp_j(1-p_j)}$. رشته متعارف رشته ای است که تفاوت تعداد واقعی هر کدام از حروف مثل x_j از تعداد متوسط آن یعنی mp_j در مقایسه با واریانس σ_j مقدار معینی باشد.

■ تعریف: رشته α رشته نمونه k یا k -typical خوانده می شود اگر شرط زیر برقرار باشد:

$$\left| \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}} \right| < k \quad \forall i = 1, 2, \dots, N. \quad (79)$$

برای ادامه بحث خود احتیاج به دو لم خیلی ساده در نظریه احتمال داریم. این لم ها دامنه کاربرد خیلی وسیعی دارند و یادگیری آنها اهمیت دارد.

■ لم اول: نامساوی اول چبیشف (*Chebyshev inequality*):

الف: فرض کنید که متغیر تصادفی X مقادیر مثبت $\{x_1, x_2, \dots, x_N\}$ را با احتمالات $\{p_1, p_2, \dots, p_N\}$ اختیار می کند. در این صورت به ازای هر عدد مثبت α ,

$$P(X \geq \alpha) \leq \frac{\bar{X}}{\alpha} \quad (80)$$

که در آن \bar{X} متوسط متغیر تصادفی X است.

■ اثبات:

$$P(X \geq \alpha) = \sum_{x=\alpha}^{\infty} P(x) \leq \sum_{x=\alpha}^{\infty} \frac{x}{\alpha} P(x) \leq \frac{\bar{X}}{\alpha}. \quad (81)$$

■ لم دوم: نامساوی دوم چبیشف (*Chebyshev inequality*):

حال فرض کنید که متغیر تصادفی X مقادیر دلخواه مثبت یا منفی اختیاری کند. در این صورت به ازای هر عدد k

$$P((X - \bar{X})^2 \geq k^2 \sigma_x^2) \leq \frac{1}{k^2}. \quad (۸۲)$$

اثبات: متغیر تصادفی $T = (X - \bar{X})^2$ را در نظریه گیریم. این متغیر فقط مقادیر مثبت را اختیار می کند.

ضمناً می دانیم که $\bar{T} = \sigma_x^2$. از قسمت الف داریم:

$$P(T \geq \alpha) \leq \frac{\bar{T}}{\alpha}. \quad (۸۳)$$

هرگاه به جای α در نامساوی اخیر قرار دهیم $k^2 \sigma_x^2$ بدست می آوریم:

$$P((X - \bar{X})^2 \geq k^2 \sigma_x^2) \leq \frac{\sigma_x^2}{k^2 \sigma_x^2} = \frac{1}{k^2}. \quad (۸۴)$$

این نامساوی را به شکل زیر نیز می توان نوشت:

$$P(|X - \bar{X}| \geq k \sigma_x) \leq \frac{1}{k^2}, \quad (۸۵)$$

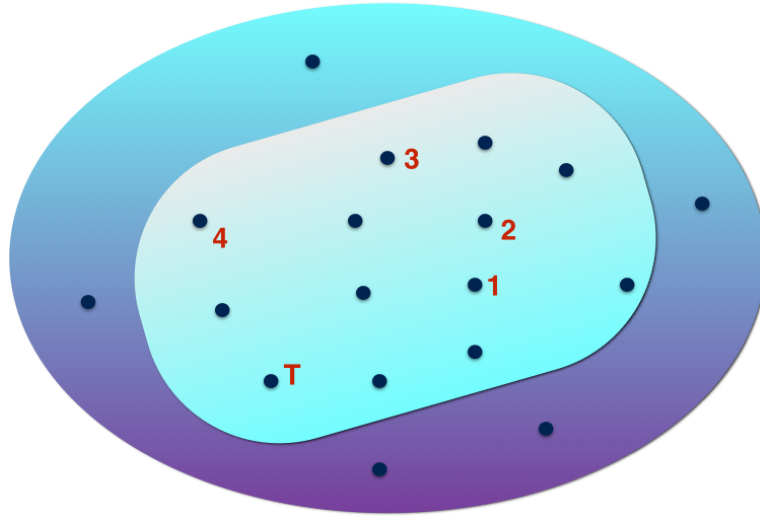
■ تمرین: در الفبای انگلیسی حرف Z کمترین فرکانس را دارد و احتمال یافتن آن در متن های انگلیسی برابر است با $P(z) = 0.074$.

الف: احتمال اینکه در یک متن که دارای N حرف است، تعداد k حرف z حضور داشته باشد چقدر است؟

ب: احتمال اینکه این تعداد از تعداد متوسط به اندازه کمتر از دو واریانس فاصله داشته باشد چقدر است؟

راهنمایی: می توانید از توزیع دوجمله ای یا توزیع پواسون که حد توزیع دوجمله ای برای وقتی است که $p \ll 1$ باشد استفاده کنید.

جواب قسمت ب را برای وقتی که $N = 100,000$ است بدست آورید.



شکل ۷: ناحیه وسط رشته های نمونه را نشان می دهد. این ناحیه شامل T تا رشته نمونه است. احتمال کل رشته های نمونه را حساب کرده ایم. هرگاه احتمال این را که هر رشته یک رشته نمونه باشد، می توانیم تعداد رشته های نمونه را حساب کنیم. برای توضیح دقیق تر به متن مراجعه کنید.

پس از این مقدمات می توانیم سه قضیه اساسی را در باره رشته های نمونه ثابت کنیم.

■ قضیه اول: به ازای هر ϵ می توان k را چنان انتخاب کرد که احتمال کل رشته های متعارف به طول m از $1 - \epsilon$ بیشتر باشد.

نکته مهم این است که عدد k را می توانیم مستقل از m انتخاب کنیم. در واقع اگر بخواهیم تعداد بیشتری رشته نمونه باشند ناچاریم که عدد k را بزرگ تر انتخاب کنیم یعنی این که تعریف خود را از رشته نمونه فراخ تر کنیم که کاملاً طبیعی به نظر می رسد. در نتیجه در یک رشته به طول m تعداد متوسط x_i دقیقاً برابر با mp_i نیست بلکه حول این مقدار افت و خیزی دارد که با کمتر شدن ϵ این افت و خیز بیشتر می شود. اما نکته مهم این است که نشان خواهیم داد در حد m های بزرگ، پهنای این افت و خیز در مقایسه با طول خود رشته به سمت صفر میل می کند. در واقع این گسترش متناسب با \sqrt{m} است که نسبت به طول خود رشته که متناسب با m افزایش می یابد بسیار کوچک است و در حد m های بزرگ این نسبت به سمت صفر میل می کند.

■ اثبات: احتمال اینکه یک رشته α متعارف نباشد را با P_0 نشان می دهیم. این احتمال برابر با این است که تعداد رخ دادن حداقل یکی از

حروف الفبا در رشته مزبور از تعریف (۷۹) پیروی نکند. با توجه به رابطه زیر در مورد احتمالات پیشامدهای گوناگون

$$P(A \cup B) \leq P(A) + P(B)$$

داریم:

$$P_0 \leq \text{Prob}\left\{ \left| \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}} \right| \geq k, i \text{ برای حداقل یک } i \right\} = \sum_{i=1}^N P\left(\left| \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}} \right| \geq k \right). \quad (۸۶)$$

پس اگر قرار دهیم

$$\epsilon = \frac{N}{k^2} \quad (۸۷)$$

از نامساوی چیشف نتیجه می گیریم که

$$P_0 \leq \sum_{i=1}^N \frac{1}{k^2} = \frac{N}{k^2} = \epsilon. \quad (۸۸)$$

این رابطه می گوید که احتمال کل رشته های نمونه از $1 - \epsilon$ بیشتر است. بنابراین اگر یک رشته m تایی به طور تصادفی از یک متن برداریم احتمال اینکه این رشته نمونه باشد از این مقدار بیشتر است. اگر این احتمال را با P_{typical} نشان دهیم داریم:

$$1 - \epsilon \leq P_{\text{typical}} \leq 1. \quad (۸۹)$$

هرگاه قرار دهیم $k = 10$ و حروف الفبای انگلیسی را ۳۲ تا بگیریم نتیجه می گیریم که احتمال اینکه یک رشته نمونه باشد از ۶۸.۰۰ بیشتر است. دقت کنید که این حد پایین است. ممکن است که احتمال واقعی بیشتر از این مقدار باشد.

■ نکته: اگر به معنای جملات توجه نکنیم، می بینیم که یک عبارت سه حرفی مثل *eee* در زبان انگلیسی فراوانی بیشتری از عبارت *the* دارد زیرا حرف *e* فراوانی بیشتری از دو حرف دیگر دارد. به همین دلیل است که در عمل و برای فشرده سازی بهتر معمولاً هجاهای چند حرفی را به عنوان یک الفبای بزرگ تر در نظر می گیرند و فشرده سازی را بر اساس فراوانی آنها انجام می دهند.

■ قضیه دوم: احتمال این که یک رشته مشخص مثل α یک رشته نمونه باشد در نامساوی زیر صدق می کند:

$$2^{-mH-A\sqrt{m}} \leq P(\alpha \in \text{typical}) \leq 2^{-mH+A\sqrt{m}}. \quad (90)$$

که در آن

$$A = - \sum_{i=1}^N \sqrt{p_i(1-p_i)} \log p_i. \quad (91)$$

در حد رشته های طولانی، یعنی ($m \gg 1$) جمله دوم در مقایسه با رشته اول به سمت صفر میل می کند و در نتیجه این رابطه می گوید که احتمال این که یک رشته خاص مثل α یک رشته نمونه باشد بستگی به نوع آن رشته ندارد و این احتمال برابریست با

$$P(\alpha \in \text{typical}) = 2^{-mH(X)+\delta_m} \quad (92)$$

که در آن

$$\lim_{m \in \infty} \delta_m = 0. \quad (93)$$

معنای این رابطه این است که تابع توزیع احتمال روی رشته های نمونه یکنواخت است، زیرا احتمال اینکه رشته α نمونه باشد بستگی به α ندارد و تنها به انتروپی منبع X بستگی دارد.

■ اثبات: به شکل (۷) نگاه می کنیم. احتمال این که یک رشته نمونه باشد، یعنی، احتمال این که یک رشته درون ناحیه رنگی باشد را یک بار با استفاده از قضیه چبیشف حساب کرده ایم. حال یک بار دیگر هم این احتمال را به شیوه متفاوتی حساب می کنیم و از آن برای بدست آوردن تعداد رشته های نمونه استفاده می کنیم. فرض کنید یک منبع به صورت تصادفی رشته ها را تولید می کند. می توانیم به صورت مجازی تصور کنیم که رشته ها به صورت تیرهای یک بازی دارت هستند که به صورت تصادفی شلیک می شوند و ممکن است به درون ناحیه رنگی اصابت کنند یا نکنند. در مثال ساده ای که از ابتدای این درس به آن اشاره کرده ایم، می توانیم از خود بپرسیم که احتمال تولید یک رشته m تایی معین مثل $\alpha = AABBCDBDADDAABC\text{CCC}$ چقدر است؟ هرگاه که حروف را مستقل از هم بگیریم پاسخ این سوال برابر است با

$$P(\alpha) = P_A P_A P_B P_B P_C P_C \dots$$

بنابراین در حالت کلی احتمال پیدا کردن یک رشته مثل α برابر است با:

$$P(\alpha) = P_1^{f(\alpha_1)} P_2^{f(\alpha_2)} \dots P_N^{f(\alpha_N)}. \quad (94)$$

یک رشته دلخواه از این نوع الزاما نمونه نیست ولی اگر پارامترهای $f(\alpha_i)$ آن در نامساوی (79) صدق کنند آنوقت حتما نمونه است. پس احتمال اینکه این رشته نمونه باشد (یعنی در درون ناحیه رنگی قرار بگیرد) را می توانیم به طریق زیر بدست بیاوریم:

$$\log P(\alpha) = \sum_{i=1}^N f_i(\alpha) \log p_i \quad (95)$$

و در نتیجه ترکیب با (79) بدست می آوریم:

$$\sum_{i=1}^N (mp_i - k\sqrt{mp_i(1-p_i)}) \log p_i \leq \log P(\alpha \in \text{typical}) \leq \sum_{i=1}^N (mp_i + k\sqrt{mp_i(1-p_i)}) \log p_i. \quad (96)$$

حال قرار می دهیم

$$A := -k \sum_{i=1}^N \sqrt{p_i(1-p_i)} \log p_i. \quad (97)$$

در نتیجه نامساوی قبلی به شکل زیر درمی آید:

$$-mH + A\sqrt{m} \leq \log P(\alpha \in \text{typical}) \leq -mH - A\sqrt{m}, \quad (98)$$

که از آن نتیجه می گیریم

$$2^{-mH - A\sqrt{m}} \leq P(\alpha \in \text{typical}) \leq 2^{-mH + A\sqrt{m}}. \quad (99)$$

بنابراین یک حد بالا و پایین برای احتمال این که یک رشته نمونه باشد را بدست آوردیم یعنی

$$P_{min} \leq P(\alpha \in \text{typical}) \leq P_{max}, \quad (100)$$

که در آن

$$P_{min} = 2^{-mH - A\sqrt{m}}, \quad P_{max} = 2^{-mH + A\sqrt{m}}. \quad (101)$$

■ قضیه سوم: هرگاه تعداد کل رشته های نمونه را با T نشان دهیم آنگاه

$$(1 - \epsilon)2^{mH - \sqrt{m}A} \leq T \leq 2^{mH + \sqrt{m}A}. \quad (1.02)$$

■ اثبات:

اگر تعداد T تا رشته داشته باشیم و آنها را از 1 تا T شماره گذاری کرده باشیم داریم:

$$P_{\text{typical}} = P(\alpha_1 \in \text{typical}) + P(\alpha_2 \in \text{typical}) + \dots + P(\alpha_T \in \text{typical}). \quad (1.03)$$

P_{typical} در واقع احتمال کلی این است که یک رشته نمونه باشد که ما در قضیه اول حدودی برای آن تعیین کرده ایم.

اما در قضیه دوم ثابت کرده ایم که

$$P_{\min} = 2^{-mH - A\sqrt{m}} \leq P(\alpha \in \text{typical}) \leq 2^{-mH + A\sqrt{m}} = P_{\max} \quad (1.04)$$

بنابراین با ترکیب این دو رابطه بدست می آوریم

$$TP_{\min} \leq P_{\text{typical}} \leq TP_{\max} \quad (1.05)$$

یا

$$(2^{-mH - A\sqrt{m}})T \leq P_{\text{typical}} \leq T(2^{-mH + A\sqrt{m}}). \quad (1.06)$$

اما در قضیه اول ثابت کرده ایم که

$$1 - \epsilon \leq P_{\text{typical}} \leq 1 \quad (1.07)$$

با ترکیب مناسب نامساوی های اخیر بدست می آوریم:

$$(1 - \epsilon)2^{mH - \sqrt{m}A} \leq T \leq 2^{mH + \sqrt{m}A}. \quad (1.08)$$

■ تمرین: با استفاده از ویکیپدیا یا هر منبع دیگری که می دانید فرکانس حروف مختلف انگلیسی را بدست آورید. تابع آنتروپی شانون را برای این متغیر تصادفی حساب کنید. سپس تابع $A(X) = k \sum_i \sqrt{p_i(1-p_i)} \log_2 p_i$ را برای آن حساب کنید. سپس کمیت های $2^{-mH+\sqrt{m}A}$ و $2^{-mH-\sqrt{m}A}$ را برای مقادیر مختلف k و m حساب کنید. این کمیت ها را به عنوان توابعی از دو مقدار m و k رسم کنید.

■ تمرین: الفبایی که در رابطه (۴۶) داده شده است را در نظر بگیرید.

الف: می خواهیم که احتمال کل رشته های نمونه از 0.95 بیشتر باشد. حساب کنید که تا چند تا واریانس نسبت به متوسط تعداد حروف را می بایست جز رشته های نمونه در نظر بگیریم؟

ب: اگر طول رشته ها برابر با 100 باشد، حدود بالا و پایین را برای تعداد کل رشته های نمونه حساب کنید. حساب کنید که این رشته ها را با چند تا بیت می توانیم کد کنیم یعنی چقدر می توانیم آنها را فشرده کنیم.

ج: مقدار فشرده سازی را برای وقتی که طول رشته ها برابر با 500 است نیز حساب کنید.

■ تمرین: تمرین قبلی را برای الفبای معرفی شده در رابطه (۵۳) نیز انجام دهید.

■ تمرین: تمرین قبلی را برای الفبای معرفی شده در رابطه (۵۵) نیز انجام دهید.

۷ ضمیمه

اولین مسئله ای که با آن مواجه هستیم یکتایی کد گشایی است. برای مثال به جدول شماره یک توجه کنید:

که در آن ستون سمت چپ کلمه ها و ستون سمت راست کد کلمه ها را نشان می دهد. حال هرگاه کد پیام 010 را دریافت کنیم می توانیم آن را به کدی برای هرکدام از پیام های x_2, x_3x_1, x_1x_4 تعبیر کنیم. در نتیجه این نوع کد گذاری دارای ابهام زیاد است و کد گذاری خوبی نیست. نخست باید یک صفت اساسی از هر نوع کدگذاری را مشخص کنیم.

0	x_1
010	x_2
01	x_3
10	x_4

جدول ۱: مثالی از یک کد که در آن بعضی از کد کلمه های پیشوند کد کلمه های دیگرند

0	x_1
100	x_2
101	x_3
11	x_4

جدول ۲: مثالی از یک کد لحظه ای

تعریف: یک کد یکتا گشاست اگر هر کد پیام حداکثرمتناظر با یک پیام باشد.

یک راه برای نوشتن کد های یکتا گشا آن است که تقاضا کنیم هیچ کد کلمه ای پیشوند کد کلمه دیگری نباشد.

تعریف: یک کد کلمه A پیشوند یک کد کلمه B خوانده می شود اگر B را بتوان به صورت $B = AC$ نوشت که در آن C دلخواه است و لزومی ندارد که خود یک کد کلمه باشد. در جدول (7) x_1 پیشوند x_2 و x_3 است. x_3 نیز پیشوند x_2 است.

تعریف: یک کد که در آن هیچ کد کلمه ای پیشوند کد کلمه دیگری نباشد یک کد لحظه ای خوانده می شود.

مثال: کد زیر یک کد لحظه ای است.

نکته مهم در مورد این نوع کد ها آن است که هر کد لحظه ای یکتا گشا است. البته معکوس این قضیه درست نیست.

0	x_1
01	x_2

جدول ۳: یک کد که به طور یکتا گشوده می شود ولی لحظه ای نیست.

باز هم به کد نشان داده شده در جدول ?? دقت کنید. هرگاه کد پیامی مثل رشته

$$101110100101 \quad (109)$$

را دریافت کنیم تنها می توانیم آن را به صورت پیام زیر بازگشایی کنیم:

$$x_3 x_4 x_1 x_2 x_3. \quad (110)$$

حال کد زیر را در نظر بگیرید:

این کد لحظه ای نیست زیرا x_1 پیشوند x_2 است. باین وجود این کد به طور یکتا گشوده می شود. زیرا هر رشته ای را که دریافت می کنیم رشته ای از 0 هاست که در بعضی جاهای آن 1 های منفرد قرار گرفته اند، مثل رشته زیر:

$$001000101010000001. \quad (111)$$

چنین رشته ای به آسانی قابل گشایش است و کدی برای پیام زیر است:

$$x_1 x_2 x_1 x_1 x_2 x_2 x_2 x_1 x_1 x_1 x_1 x_2. \quad (112)$$

در زیر روشی را بیان می کنیم که به کمک آن می توانیم تشخیص بدهیم که آیا یک کد به صورت یکتا گشوده می شود یا خیر.

فرض کنید که S_0 مجموعه همه کد کلمه ها باشد. مجموعه تمام پسوند هایی را که در S_0 وجود دارد در مجموعه دیگری به نام S_1 قرار می دهیم.

حال مجموعه S_2, S_3, \dots, S_n را به طریق زیر تشکیل می دهیم:

الف: اگر یک کد کلمه $A \in S_0$ پیشوند کد کلمه ای مثل $w = AB \in S_{n-1}$ باشد، B را در S_n قرار می دهیم.

ب: اگر یک کد کلمه $A \in S_{n-1}$ پیشوند کد کلمه ای مثل $w = AB \in S_0$ باشد، B را در S_n قرار می دهیم.

0	x_1
010	x_2
01	x_3
10	x_4

جدول ۴: مثالی از یک کد که به طور یکتا گشوده نمی شود.

0	x_1
001	x_2

جدول ۵: مثالی از یک کد یکتا گشا

■ قضیه: یک کد به صورت یکتا گشوده می شود اگر فقط اگر $\phi = S_0 \cap [S_1 \cup S_2 \cup S_3 \dots]$.

مثال: کد زیر یکتا گشان نیست.

زیرا:

$$S_0 = \{0, 010, 01, 10\} \quad S_1 = \{10, 1, 0\} \quad (113)$$

$$S_0 \cap S_1 \neq \phi$$

مثال: کد زیر یکتا گشاست:

زیرا:

$$S_0 = \{0, 001\} \quad S_1 = \{01\} \quad S_2 = \{1\}, \quad (114)$$

$$S_0 \cap [S_1 \cup S_2] = \phi$$

مثال: کد زیر را در نظر می گیریم:

a	x_1	
c	x_2	
ad	x_3	
abb	x_4	(۱۱۵)
bad	x_5	
deb	x_6	
$bbcde$	x_7	

برای این کد داریم:

$$\begin{aligned}
S_0 &= \{a, c, ad, abb, bad, deb, bbcde\} \\
S_1 &= \{d, bb\} \\
S_2 &= \{eb, cde\} \\
S_3 &= \{de\} \\
S_4 &= \{b\} \\
S_5 &= \{ad, bbcde\} \\
S_6 &= \{d\} \\
S_7 &= \{eb\}
\end{aligned}
\tag{۱۱۶}$$

با توجه به این روابط خواهیم دید که

$$S_0 \cap [S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6 \cup S_7] = \phi, \tag{۱۱۷}$$

و در نتیجه این کد یکتا گشا است.

■ قضیه (شرط لازم و کافی برای وجود کد های لحظه ای): مجموعه کلمه های $X = \{x_1, x_2, \dots, x_M\}$ و مجموعه حروف الفبای $A := \{a_1, a_2, \dots, a_D\}$ داده شده اند. مجموعه اعداد صحیح $\{n_1, n_2, \dots, n_M\}$ نیز مفروض اند. آیا یک کد لحظه ای می توان از

الفبای A نوشت که طول های $\{n_1, n_2, \dots, n_M\}$ داشته باشند؟ پاسخ این سوال مثبت است اگر فقط اگر شرط زیر برقرار باشد:

$$\sum_{i=1}^M \frac{1}{D^{n_i}} \leq 1. \quad (118)$$

این نامساوی به نامساوی $Kraft$ مشهور است.

قبل از اثبات این قضیه به یک نتیجه ساده آن توجه می کنیم:

نتیجه: برای حروف الفبای انگلیسی با احتساب نقطه، کاما، و دیگر علائم داریم: $M = 32$. هم چنین اگر بخواهیم از الفبای $\{0, 1\}$ استفاده کنیم داریم $D = 2$. بنابراین باید داشته باشیم:

$$\sum_{i=1}^{32} \frac{1}{2^{n_i}} \leq 1 \rightarrow n_{min} \geq 5. \quad (119)$$

بنابراین نمی توان هیچ حرفی را با کمتر از 5 بیت کد کرد.

حال به اثبات قضیه می پردازیم:

■ اثبات: می توانیم از نمودارهای درختی استفاده کنیم. یک درخت با مرتبه D و اندازه k درختی است که D ریشه دارد و از هر ریشه نیز D شاخه منشعب می شود و این کار ادامه می یابد تا $k-1$ مرحله. در این صورت تعداد شاخه های آخرین مرحله عبارت است از D^k . D ریشه اول درخت متناسب با کد کلمه های تک حرفی $\{1, 2, 3, \dots, D\}$ هستند. شاخه های مرحله بعد متناسب با کد کلمات دو حرفی هستند مثل $\{11, 12, \dots, DD\}$ و همینطور تا آخر. به این ترتیب هر کد کلمه متناسب با یکی از گره های این درخت می شود. حال اگر بخواهیم یک کد لحظه ای بسازیم می بایست کد کلمه های خود را از شاخه های این درخت به نحو خاصی انتخاب کنیم. هر کد کلمه یا هر گره که از این درخت انتخاب می کنیم می بایست تمام شاخه های منشعب از آن گره را کنار بگذاریم زیرا همه کد کلمه های مربوط به آن شاخه ها کلمه مربوط به این گره را به عنوان پیشوند خود دارند. اگر طول یک کد کلمه که انتخاب می کنیم برابر با i باشد، تعداد شاخه هایی که از آن منشعب می شود برابر است با D^{k-i} . بنابراین به ازای هر کد کلمه به طول i تعداد D^{k-i} تا از شاخه ها حذف می شوند. در نتیجه خواهیم داشت:

$$D^{k-i_1} + D^{k-i_2} + \dots + D^{k-i_M} \leq D^k \quad (120)$$

که با تقسیم طرفین بر D^k به رابطه (130) می‌رسیم. این رابطه را می‌توان به شکل زیر نیز نوشت:

$$\sum_i w_i D^{-i} \leq 1, \quad (121)$$

که در آن w_i تعداد کد کلمه‌های با طول i است.

حال معکوس قضیه را ثابت می‌کنیم: تاکنون ثابت کردیم که اگر کد لحظه‌ای باشد می‌بایست شرط (130) برقرار باشد. حال نشان می‌دهیم که به ازای هر (n_1, n_2, \dots, n_M) که در شرط (130) صدق کند می‌توان یک کد لحظه‌ای ساخت. n_i را به شکل زیر مرتب می‌کنیم:

$$n_1 \leq n_2 \leq n_3 \leq \dots \leq n_M. \quad (122)$$

حال یک نقطه به اندازه n_1 را روی درخت با مرتبه D و اندازه n_M انتخاب می‌کنیم. به این ترتیب تعداد $D^{n_M - n_1}$ نقطه حذف می‌شوند. تعداد نقاط باقیمانده برابر است با $D^{n_M} - D^{n_M - n_1}$. نقطه دوم را به طول n_2 انتخاب می‌کنیم. این نقطه تعداد $D^{n_M - n_2}$ نقطه دیگر را حذف می‌کند. تعداد نقاط باقیمانده برابر است با $D^{n_M} - D^{n_M - n_1} - D^{n_M - n_2}$. این کار را ادامه می‌دهیم تا نقطه ماقبل آخر که طول آن n_{M-1} است. این نقطه نیز تعداد $D^{n_M - n_{M-1}}$ را حذف می‌کند. آیا درخت موردنظر این همه جا دارد؟ برای پاسخ به این سوا ل کافی است که تعداد نقاط باقیمانده را بعد از مرحله ماقبل آخر بشماریم: این تعداد برابر است با

$$\begin{aligned} Q &= D^{n_M} - D^{n_M - n_1} - D^{n_M - n_2} - \dots - D^{n_M - n_{M-1}} \\ &= D^{n_M} [1 - D^{-n_1} - D^{-n_2} - \dots - D^{-n_{M-1}}] \end{aligned} \quad (123)$$

اما چون شرط (130) برقرار است خواهیم داشت:

$$\sum_{i=1}^M D^{-n_i} \leq 1 \longrightarrow 1 - \sum_{i=1}^{M-1} D^{-n_i} \leq D^{-n_M} \quad (124)$$

و این رابطه به این معناست که حداقل یک انتخاب برای آخرین کد کلمه باقی می‌ماند. اثبات قضیه در این جا کامل می‌شود.

■ قضیه: نامساوی کرافت شرط لازم و کافی برای ساختن کد‌های یکتاگشاست.

اثبات: الف: اگر نامساوی کرافت برقرار باشد می توانیم یک کد لحظه ای مطابق با قضیه قبل بسازیم و می دانیم که کد های لحظه ای یکتا گشاهستند.

ب: حال فرض کنید که یک کد یکتا گشاداریم. می خواهیم نشان دهیم که حتماً نامساوی کرافت برقرار است. بجای عبارت $\sum_i D^{-n_i}$ عبارت $\sum_{i=1}^r w_i D^{-i}$ را بکار می بریم که در آن w_i تعداد کلمات با طول i است. حال عبارت اخیر را می توان به صورت یک تابع مولد تعبیر کرد. می توان دریافت که

$$\left(\sum_{i=1}^r w_i D^{-i} \right)^n = \sum_{k=r}^{nr} X_k D^{-k}, \quad (125)$$

که در آن X_k تعداد کل کد کلمه های با طول k در کدگذاری رشته های r تایی است. اما می دانیم که کد از نوع یکتا گشودنی است. در ضمن تعداد کل کد کلمه های با طول k برابر است با D^k . چون کد های یکتاگشا زیرمجموعه کلیه کد ها هستند نتیجه می گیریم که $X_k \leq D^k$. بنابراین خواهیم داشت:

$$\left(\sum_{i=1}^r w_i D^{-i} \right)^n \leq \sum_{k=r}^{nr} 1 = nr - r + 1. \quad (126)$$

و از آنجا

$$\left(\sum_{i=1}^r w_i D^{-i} \right) \leq (1 + (n-1)r)^{\frac{1}{n}}. \quad (127)$$

در حد n های بزرگ این رابطه تبدیل می شود به نامساوی کرافت.

■ قضیه کدگذاری بدون نوفه: مجموعه کلمات $X = \{x_1, x_2, \dots, x_M\}$ که در آن نماد x_i با احتمال $P(x_i) := P_i$ ظاهر می شود و مجموعه حروف الفبای $A := \{a_1, a_2, \dots, a_D\}$ داده شده اند. این نماد ها با کد کلمه های $\{w_1, w_2, \dots, w_M\}$ کد شده اند و طول هر کد کلمه w_i برابر است با $l(w_i) := n_i$. هدف ما آن است که طول متوسط کد کلمه ها را کمینه کنیم یعنی کمیت زیر را:

$$\bar{n} := \sum_{i=1}^M p_i n_i. \quad (128)$$

مجموعه اعداد صحیح $\{n_1, n_2, \dots, n_M\}$ نیز مفروض اند. بهترین کد یکتا گشایی که می توان برای کد کردن این الفباساخت، یعنی کد یکتا گشایی که کمترین طول متوسط را داشته باشد کدی است با طول متوسط

$$\bar{n} = \frac{H(X)}{\log D}. \quad (129)$$

■ اثبات: نخست توجه می کنیم که کد مورد نظر ما یکتا گشاست اگر و فقط اگر شرط زیر برقرار باشد:

$$\sum_{i=1}^M \frac{1}{D^{n_i}} \leq 1. \quad (130)$$

بقیه اثبات را در سه مرحله انجام می دهیم. از این به بعد نیز ما فقط درباره کد های یکتا گشا حرف می زنیم. در مرحله اول یک حد پایین برای \bar{n} پیدا می کنیم و نشان می دهیم که

$$\bar{n} \geq \frac{H(X)}{\log D} \quad (131)$$

که در آن شرط تساوی برقراری شود اگر و فقط اگر $p_i = D^{-n_i}$.

در مرحله دوم تحقیق می کنیم که چقدر می توانیم به این حد پایین نزدیک شد. و بالاخره در مرحله سوم بهترین کد ممکن را می سازیم. برای اثبات نامساوی (131) می بایست نامساوی زیر را ثابت کنیم:

$$\sum_{i=1}^M n_i p_i \geq - \sum_{i=1}^M p_i \frac{\log p_i}{\log D}, \quad (132)$$

و یا

$$\sum_{i=1}^M (n_i \log D) p_i \geq - \sum_{i=1}^M p_i \log p_i. \quad (133)$$

قبلا داشتیم که به ازای هر دو توزیع احتمال $\{p_i\}$ و $\{q_i\}$ ، نامساوی زیر برقرار است:

$$\sum_i -p_i \log p_i \leq - \sum_i p_i \log q_i, \quad (134)$$

و تساوی تنها وقتی برقراری شود که $\{q_i\} = \{p_i\}$.

می توانیم یک توزیع احتمال مطابق بارابطه زیرتعریف کنیم:

$$q_i := \frac{D^{-n_i}}{\sum_{i=1}^M D^{-n_i}} \quad (135)$$

و از رابطه (132) استفاده کنیم. یک محاسبه ساده منجر به رابطه زیر خواهد شد:

$$H(X) \leq \bar{n} \log D + \log\left(\sum_{i=1}^M D^{-n_i}\right), \quad (136)$$

که تساوی وقتی برقرار می شود که

$$p_i = \frac{D^{-n_i}}{\sum_{i=1}^M D^{-n_i}}. \quad (137)$$

حال باتوجه به اینکه برای کدهای یکتاگشانا مساوی کرافت برقرار است یعنی $\sum_{i=1}^M D^{-n_i} \leq 1$ نتیجه می گیریم که $\log \sum_{i=1}^M D^{-n_i} \leq 0$ و از آنجا بدست می آوریم که

$$H(X) \leq \bar{n} \log D. \quad (138)$$

هرگاه بتوانیم یک کد را چنان انتخاب کنیم که طول کد کلمه های آن از رابطه $\frac{1}{p_i} = \log_D n_i$ تبعیت کند، آنگاه خواهیم داشت :

$\bar{n} = \frac{H(X)}{\log D}$ معکوس این قضیه نیز صحیح است یعنی اینکه اگر رابطه $\bar{n} = \frac{H(X)}{\log D}$ برقرار باشد آنگاه $p_i = D^{-n_i}$ برای اثبات این نتیجه

از رابطه 136 استفاده می کنیم و به این نتیجه می رسیم که

$$\bar{n} \log D \leq \bar{n} \log D + \log\left(\sum_{i=1}^M D^{-n_i}\right), \quad (139)$$

و از آنجا باتوجه به اینکه $\sum_{i=1}^M D^{-n_i} \leq 1$ ، به این نتیجه می رسیم که $\sum_{i=1}^M D^{-n_i} = 1$. اما باتوجه به رابطه 137 این نتیجه به این

معناست که $p_i = D^{-n_i}$.

■ تعریف: یک کد کاملاً بهینه کدی است که برای آن $\bar{n} = \frac{H(X)}{\log D}$.

یک مثال از یک کد کاملاً بهینه درجدول زیر داده شده است:

Cw	P	X
0	$\frac{1}{2}$	x_1
10	$\frac{1}{4}$	x_2
110	$\frac{1}{8}$	x_3
111	$\frac{1}{8}$	x_4

(۱۴۰)

این کد دارای این خاصیت است که $n_i = \log \frac{1}{p_i}$.

در حالت کلی معلوم نیست که بتوان کد را چنان طراحی کرد که حد $\bar{n} = \frac{H}{\log D}$ برقرار شود، زیرا اعداد $n_i = \log_D \frac{1}{p_i}$ معلوم نیست که صحیح باشند. با این وجود می توان کاری کرد که شرط زیر برقرار شود:

$$\log_D \frac{1}{p_i} \leq n_i \leq \log_D \frac{1}{p_i} + 1. \quad (141)$$

در این صورت خواهیم داشت :

$$\frac{H(X)}{\log D} \leq \bar{n} \leq \frac{H(X)}{\log D} + 1. \quad (142)$$

حال نکته این است که هر قدر بخواهیم می توانیم به حد پایین نامساوی بالا نزدیک شویم. برای این کار می بایست از کدهای چندتایی یا کدهای بلوکی استفاده کنیم. فرض کنید به جای کد نگاری X رشته های s تایی از X ها را کد نگاری کنیم، یعنی رشته های $Y = (X_1, X_2, \dots, X_s)$ را. حال باید نشان دهیم که تحت این شرایط طول کد کلمه ها به ازای هر X پایین می آید. به رابطه (141) دقت می کنیم. از آنجا که

کلمه ها به صورت s تایی های از نوع $y_{ij} = (x_i, x_j, \dots, x_s)$ هستند. داریم

$$H(Y) = - \sum_{i,j,\dots} p_{ij\dots} \log p_{ij\dots} \quad (143)$$

چون کلمات پیام Y از هم مستقل هستند خواهیم داشت: $H(Y) = sH(X)$. و بنابراین

$$\frac{H(Y)}{\log D} \leq \bar{n} \leq \frac{H(Y)}{\log D} + 1, \quad (144)$$

و یا

$$\frac{H(X)}{\log D} \leq \frac{1}{s} \bar{n} \leq \frac{H(X)}{\log D} + \frac{1}{s}. \quad (145)$$

در این رابطه $\frac{1}{s} \bar{n}$ طول متوسط هر کد کلمه به ازای هر کلمه در X است و در حد s های بزرگ دیده می شود که ما به حد بهینه نزدیک می شویم.

۱.۷ ساختن کدهای بهینه

حال باید الگوریتمی را معرفی کنیم که کدهای بهینه را به طور روشمند می سازد. نخست به یک لم احتیاج داریم:

لم: فرض کنید که برای احتمالات P_1, P_2, \dots, P_M یک کد C در درون مجموعه کدهای لحظه ای بهینه باشد. یعنی هیچ کد لحظه ای دیگری با طول متوسط کمتر از طول متوسط مربوط به C وجود نداشته باشد. در این صورت این کد در درون مجموعه کدهای یکتا گشا نیز بهینه است.

اثبات: می دانیم که کدهای لحظه ای زیرمجموعه کدهای یکتا گشا است. حال فرض کنید که یک کد یکتا گشای C' با طول کد کلمه های n'_1, n'_2, \dots, n'_M وجود دارد که طول متوسط آن از طول متوسط C کمتر است. اولاً چون C' یکتا گشاست بنابراین قضیه ای که قبلاً ثابت کردیم خواهیم داشت: $\sum_{i=1}^M D^{-n'_i} \leq 1$. اما در این صورت بنا بر قضیه قبل یک کد لحظه ای با طول کلمات n'_1, n'_2, \dots, n'_M وجود خواهد داشت. بدین ترتیب بهینه بودن کد C در درون مجموعه کدهای لحظه ای نیز نقض می شود.

از این به بعد توجه خود را به کدهای لحظه ای و دوتایی *binary* معطوف می کنیم. نخست به یک لم احتیاج داریم:

لم: فرض کنید که C یک کد لحظه ای با طول کد کلمه های n_1, n_2, \dots, n_M برای کد گذاری علامات x_1, x_2, \dots, x_M باشد که این علامات

نیز با احتمالات p_1, p_2, \dots, p_M تکرارشوند. در این صورت اگر کد C درون کد های لحظه ای بهینه باشد آنگاه خاصیت های زیر برقرارند:

الف: علامت های با احتمال بیشتر طول کمتر دارند. یعنی اگر $p_i \geq p_j$ آنگاه $n_i \leq n_j$.

ب: دوتا از کد کلمه هایی که کمترین احتمال ها را دارند حتماً دارای طول مساوی هستند.

ج: در بین کلماتی که بیشترین طول را دارند، حتماً باید دو کلمه وجود داشته باشد که فقط و فقط در یک رقم بایکدیگر تفاوت داشته باشند.

اثبات الف: فرض کنید که $p_1 \geq p_2$ که در آن p_1, p_2 به ترتیب احتمال ظهور علامت x_1, x_2 باشند. هم چنین فرض کنید که در این کد لحظه ای C داشته باشیم $n_1 \geq n_2$. در این صورت می توان یک کد بهتر از C ساخت. جای کد کلمه های مربوط به x_1 و x_2 را عوض می کنیم. کد هنوز لحظه ای است زیرا شرط کرافت برقرار است. در کد جدید C' داریم:

$$\bar{n}' - \bar{n} = n_1 p_2 + n_2 p_1 - n_1 p_1 - n_2 p_2 = (n_1 - n_2)(p_2 - p_1) \leq 0. \quad (146)$$

اثبات ب: فرض کنید که کمترین احتمالات عبارت باشند از $P_M, P_{M-1}, P_{M-1} \geq P_M$. حال می خواهیم حالت $n_{M-1} < n_M$ را حذف کنیم. کد کلمه های مربوط به علامت های x_M و x_{M-1} را به ترتیب با S و \bar{S} نشان می دهیم. فرض کنید که

$$\begin{aligned} S &= s_1 s_2 \cdots s_{n_{M-1}} \\ \bar{S} &\equiv S' \bar{S}' = s'_1 s'_2 \cdots s'_{n_{M-1}} (s'_{n_{M-1}+1} s_{n_{M-1}+2} \cdots s'_{n_M}) \end{aligned} \quad (147)$$

حال می توانیم قسمت اضافی را که در پراپرتی قراردادده ایم حذف کنیم بدون اینکه به لحظه ای بودن کد خلی وارد شود. چون اگر کلمه ای پیشوند $S' \bar{S}'$ نبوده است پیشوند S' نیز نخواهد بود. ضمناً S' نمی تواند پیشوند کد کلمه دیگری باشد، چون کلمات مربوط به x_M و x_{M-1} بزرگترین طول ها را دارند. تنها مکانی که باقی می ماند آن است که کلمات با طول n_{M-1} بیش از دو تا باشند. در این صورت تنها راه برای پیشوند بودن S' آن است که S' دقیقاً بایکی از آن کلمات برابر باشد. ولی این بدان معناست که در کد اولیه که در آن حذفی صورت نگرفته بود، آن کلمه خاص پیشوند \bar{S} بوده است.

اثبات ج: حال فرض کنید که دوتا از بلندترین کلمات را در نظر بگیریم. اگر تنها در رقم آخر اختلاف داشته باشند که این همان چیزی است که مطلوب ماست. اگر بیش از رقم آخر با هم اختلاف داشته باشند ما می توانیم رقم آخر را حذف کنیم و یک کد لحظه ای بهتر بدست بیاوریم. استدلال این که لحظه ای بودن کد به هم نمی خورد مثل قسمت ب است.

۲.۷ روش هوفمان برای ساختن کد های لحظه ای بهینه

ازاین به بعد نمادها واحتمالات را با (X, P) نمایش می دهیم:

$$(X, P) = \{(x_1, p_1), (x_2, p_2), \dots, (x_M, p_M)\}. \quad (148)$$

مرحله اول: از (X, P) یک (\tilde{X}, \tilde{P}) به ترتیب زیری سازیم:

$$(\tilde{X}, \tilde{P}) = \{(x_1, p_1), (x_2, p_2), \dots, (x_{M-2}, p_{M-2}), (x_{M-1, M}, p_{M-1} + p_M)\}. \quad (149)$$

سوال: منظور از $x_{M-1, M}$ چیست؟ منظور این است که در ذهن خود تفاوت بین x_{M-1} و x_M را ازین ببریم. به عبارت دیگر می دانیم که

تنها احتمالات مهم هستند و نه خود نمادها. بنابراین مجموعه $\{p_1, p_2, \dots, p_{M-1}, p_M\}$ را به مجموعه $\{p_1, p_2, \dots, p_{M-1} + p_M\}$ تقلیل داده

ایم. حال فرض کنید که کد بهینه ای برای (\tilde{X}, \tilde{P}) در دست باشد با مشخصات زیر:

\tilde{N}	\tilde{C}	\tilde{P}	\tilde{X}
n_1	w_1	p_1	x_1
n_2	w_2	p_2	x_2
.	.	.	.
n_{M-2}	w_{M-2}	p_{M-2}	x_{M-2}
$n_{M-1, M}$	$w_{M-1, M}$	$p_{M-1} + p_M$	$x_{M-1, M}$

(150)

حال کد C را برای (X, P) به شکل زیر می سازیم.

\tilde{N}	\tilde{C}	\tilde{P}	\tilde{X}
n_1	w_1	p_1	x_1
n_2	w_2	p_2	x_2
.	.	.	.
n_{M-2}	w_{M-2}	p_{M-2}	x_{M-2}
$n_{M-1, M} + 1$	$w_{M-1, M} 0$	p_{M-1}	x_{M-1}
$n_{M-1, M} + 1$	$w_{M-1, M} 1$	p_M	x_M

(۱۵۱)

حال ثابت می کنیم که اگر \tilde{C} بهینه باشد آنگاه C نیز بهینه است. از برهان خلف استفاده می کنیم. فرض کنید که کدی مثل C' وجود داشته باشد که از کد C بهتر باشد. در این صورت با استفاده از کد C' می توان کدی مثل \tilde{C}' ساخت که از \tilde{C} بهتر باشد. کد C' در جدول زیر نشان داده شده است:

N'	C'	P	X
n'_1	w'_1	p_1	x_1
n'_2	w'_2	p_2	x_2
.	.	.	.
n'_{M-2}	w'_{M-2}	p_{M-2}	x_{M-2}
n'_{M-1}	w'_{M-1}	p_{M-1}	x_{M-1}
n'_M	w'_M	p_M	x_M

(۱۵۲)

در این کد $n'_m = n'_{M-1}$ و w'_M نیز تنها در رقم آخری ما اختلاف دارند. حال کد \tilde{C}' را مطابق جدول زیر می‌سازیم:

\tilde{N}'	\tilde{C}'	P	X
n'_1	w'_1	p_1	x_1
n'_2	w'_2	p_2	x_2
.	.	.	.
n'_{M-2}	w'_{M-2}	p_{M-2}	x_{M-2}
n'_{M-1}	$\tilde{w}'_{M-1,M}$	$p_{M-1} + p_M$	$x_{M-1,M}$

(۱۵۳)

که در آن $\tilde{w}'_{M-1,M}$ همان w'_M یا w'_{M-1} است که رقم آخر آن برداشته شده است. حال بدست می آوریم:

$$\bar{n} - \bar{\tilde{n}} = (p_{M_1} + p_M)(n_{M-1,M} + 1 - n_{M-1,M}) = p_{M-1} + p_M, \quad (154)$$

و

$$\bar{n}' - \bar{\tilde{n}}' = (p_{M_1} + p_M)(n'_{M-1} - n'_{M-1} - 1) = p_{M-1} + p_M. \quad (155)$$

در نتیجه

$$\bar{n} - \bar{\tilde{n}} = \bar{n}' - \bar{\tilde{n}}' \quad (156)$$

که از آن خواهیم داشت:

$$if \bar{n}' < \bar{n} \rightarrow \bar{\tilde{n}}' < \bar{\tilde{n}}. \quad (157)$$

بنابراین اگر کد C' از کد C بهتر باشد کد \tilde{C} نیز از کد \tilde{C} بهتر است و این خلاف بهینه بودن کد \tilde{C} است.

این قضایا به ما می آموزند که چگونه کد های بهینه بسازیم.

مثال یک: روش ساخت درجدول های زیر نشان داده شده است:

P	X
0.5	x_1
0.35	x_2
0.15	x_3

(۱۵۸)

\tilde{C}	\tilde{P}	\tilde{X}
0	0.5	x_1
1	0.5	$x_{2,3}$

(۱۵۹)

وازانجا

C	P	X
0	0.5	x_1
10	0.35	x_2
11	0.15	x_3

(۱۶۰)

هرگاه تعداد کلمات بیشتر باشد این کار را در چند مرحله انجام می دهیم . در هر مرحله احتمالات را از بیشترین به کمترین مرتب می کنیم و آخرین دو کلمه را با هم مطابق با آنچه که در بالا گفته شد ادغام می کنیم. این کار را آنقدر انجام می دهیم تا به یک مجموعه برسیم متشکل از دو نماد و دو احتمال. به دو نماد آخر کلمه های 0 و 1 را نسبت می دهیم و سپس مراحل را در جهت عکس طی می کنیم تا به جدول اولیه برسیم و کد های تمام نماد ها را بدست آوریم.