نظریه اطلاعات کلاسیک ـ بخش اول

وحیدکریمی پور ـ دانشکده فیزیک ـ دانشگاه صنعتی شریف

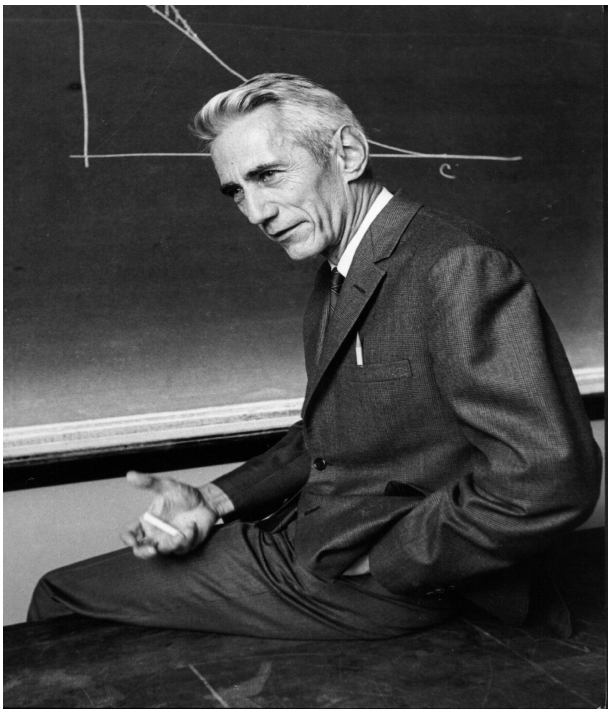
۱۶ اردیبهشت ۱۴۰۰

۱ مقدمه

فرض کنیدکه $X = \{x_1, x_2, \cdots x_n\}$ یک متغیرتصادفی بااحتمالات $\{)p(x_1), p(x_2), \cdots p(x_n\}$ باشد. به این متغیر تصادفی می توان تابعی به شکل زیرنسبت داد.

$$H(X) := -\sum_{i=1}^{n} p(x) \log_2 p(x). \qquad (۱)$$

بدون اغراق می توان گفت که تمام نظریه اطلاعات کلاسیک برروی این تابع که آن را تابع آنتروپی می خوانند و خواص وتعبیرهای آن بناشده است. هدف ما دراین درس آن است که اولاً خواص ریاضی این تابع وتوابع وابسته به آن رااستخراج کنیم، ثانیاً تعبیروتفسیرهای این توابع را بفهمیم. نخستین کاری که باید بکنیم آن است که نشان دهیم این تابع سنجه مناسبی برای اطلاعات است. این کاری است که درنخستین بخش این درس انجام می دهیم. دربخش های بعدی این درس مفاهیمی مثل اطلاعات شرطی و اطلاعات متقابل را معرفی می کنیم. پس ازبررسی خواص ریاضی توابعی که برای اندازه گیری اطلاعات معرفی کرده ایم به فشرده سازی اطلاعات و حدی که برای این فشرده سازی وجود داردمی پردازیم.



شکل ۱: کلاود شانون، بنیانگذار نظریه اطلاعات

۲ مفهوم واندازه اطلاعات

۱.۲ اطلاعات یک متغیرتصادفی

فرض کنید که آزمایش یا واقعه ای مثل X که نتایج یاپیشامدهای ممکن آن را بامجموعه $\{x_1, x_2, \cdots x_n\}$ نشان می دهیم اتفاق بیفتد و یک نتیجه معین مثل x_i حاصل شود. دراین صورت می توانیم بپرسیم 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 یقینا 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، طبیعتا تابعی از این احتمالات است. به عنوان مثال اگر داشته باشیم

$$P(x_1) = 1, \qquad P(x_i) = 0, \quad i = 2,\ 3,\ \cdots N, \tag{۲}$$

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

$$P(x_i) = \frac{1}{N}, \tag{۳}$$

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. ازنظرشهودی هرچقدرکه پیشامدی که بوقوع پیوسته است محتمل تر بوده باشد اطلاعی که ماکسب کرده ایم کمتروهرچقدرکه آن پیشامد دورازانتظاربوده باشد تعجب ماازوقوع آن بیشتر و اطلاعی که ماکسب کرده ایم بیشتر خواهد بود. بنابراین اگر میزان اطلاع خودازوقوع 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 دوواقعه مستقل (X, Y) شود که نتایج ممکن آن را با زوج های $\{(x_i, y_j),\ \ i = 1 \cdots m,\ j = 1 \cdots n\}$ نشان می دهیم. هرگاه احتمال وقوع x_i را با p_i واحتمال وقوع y_j را با q_j نشان دهیم احتمال هرپیشامد (x_i, y_j) برابرخواهد بود با $p_i q_j$ ومیزان اطلاعی که ازوقوع این پیشامد کسب می کنیم برابرخواهد بود با $h(p_i q_j)$. انتظارداریم که میزان اطلاع ما دراین مورد که دوپیشامد مستقل x_i و y_j رخ داده اند برابربا مجموع اطلاعاتی باشد که از وقوع پیشامد x_i به تنهایی و y_j به تنهایی کسب می کنیم بنابراین انتظارداریم که

$$h(p_i q_j) = h(p_i) + h(q_j). \tag{۴}$$

تنهاتابعی که شرط فوق رابرآورده کند و ضمناً نزولی باشد، تابع لگاریتم است بنابراین خواهیم داشت:

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

که درآن α ثابت است. ثابت α رامی توان با شرط بهنجارش تعیین کرد. قرارمی نهیم که میزان اطلاع کسب شده ما ازوقوع یک پدیدهٔ دوحالته متساوی الاحتمال برابربایک باشد، یعنی $h(1/2) = 1$. درنتیجه میزان ثابت α برابرمی شود با ۲.

اگر یک آزمایش X را N بارانجام دهیم به طور متوسط Np_i بارنتیجه x_i رخ خواهد داد و میزان اطلاع عی که درهربارکسب می کنیم برابرخواهد بود با $\log_2(\frac{1}{p_i})$. میزان اطلاعی که ما به طورمتوسط از وقوع نتایج آزمایش تصادفی X کسب می کنیم برابرخواهد بود با:

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

این تابع ، تابع آنتروپی یا تابع شانون نیزخوانده می شود. دقت کنید که تابع $p \log \frac{1}{p}$ درفاصله $p \in [0, 1]$ 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 متغیرتصادفی

هرگاه دومتغیرتصادفی (X, Y) داشته باشیم که لزوماً ازهم مستقل نباشند تابع آنتروپی یا اطلاعات به طورطبیعی به شکل زیرتعریف می شود:

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

درحالتی که دومتغیرتصادفی مستقل باشند یعنی $p(x,y) = p(x)q(y)$، رابطه بالابدست می دهد که $H(X,Y) = H(X) + H(Y)$.

این تعریف به همین شکل به بیش از دومتغیرتصادفی تعمیم می یابد به این معنا که تعریف می کنیم:

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

۳.۲ اطلاعات شرطی

دومتغیرتصادفی X, Y که باتوزیع آنهابا تابع $P(x,y)$ مشخص می شود درنظر می گیریم. فرض کنید که مقدار یکی از متغیرهای تصادفی مثل Y را می دانیم و این مقدار برابراست با y. دراین صورت توزیع متغیرتصادفی X عوض خواهد شد وتبدیل خواهد شد به توزیع $P(X|y)$ که درآن y یک پارامتراست و X مقادیرمتغیر را بخود می گیرد. می دانیم که:

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

درنتیجه اطلاعات باقیمانده در متغیرتصادفی X برابرخواهد بود با:

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

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

$$\begin{aligned} H(X|Y) &= \sum_y p(y)H(X|y) = -\sum_{x,y} P(y)P(x|y)\log_2 P(x|y) \\ &= -\sum_{x,y} P(x,y)\log_2 P(x|y) = -\sum_{x,y} P(x,y)\log_2 \frac{P(x,y)}{P(y)} \\ &= H(X,Y) - H(Y). \end{aligned} \quad (۱۱)$$

دقت کنید که به همان دلیلی که تابع $H(X)$ مثبت است تابع $H(X|y)$ ودرنتیجه تابع $H(X|Y)$ نیزمثبت خواهند بود. $H(X|Y)$ را اطلاعات X مشروط به Y می خوانیم و این کمیت بیان کننده میزان اطلاعات باقیمانده درX است هرگاه ما مقادیر Y را دانسته باشیم. باید توجه داشت که این تابع متقارن نیست یعنی $H(X|Y) \neq H(Y|X)$. ازرابطه (11) به نتیجه زیر می رسیم:

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X). \quad (۱۲)$$

اگر دومتغیرتصادفی X, Y مستقل باشند آنگاه دانستن Y هیچ تاثیری دراطلاعات باقیمانده در X نخواهد داشت و درنتیجه $H(X|Y) = H(X)$ و بنابر (12)، $H(X,Y) = H(X) + H(Y)$.

بالعکس هرگاه X و Y کاملاً به هم وابسته باشند انتظارداریم که دانستن Y برای دانستن X نیزکفایت کند یعنی هیچ اطلاعی درX باقی نگذارد یعنی $H(X|Y) = 0$ که باتوجه 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 دردومتغیرتصادفی X و Y به شکل زیرتعریف می شود:

$$I(X:Y) := H(X) + H(Y) - H(X,Y). \quad (۱۳)$$

این کمیت نسبت به دومتغیرتصادفی X و Y متقارن است. باتوجه به رابطه (12) می توان آن را به شکل زیربازنویسی کرد:

$$I(X:Y) := H(X) - H(X|Y). \quad (۱۴)$$

این رابطه معرف چه چیزی است ؟ قبل ازآنکه مقدار Y را بدانیم، اطلاعات موجود درX با $H(X)$ سنجیده می شد. بادانستن Y این اطلاعات به $H(X|Y)$ تقلیل پیدا می کند. بنابراین تفاوت این دو میزان اطلاعی است که Y درباره X حمل می کند. بعداً نشان خواهیم داد که $I(X:Y)$

یک کمیت نامنفی است.

اطلاعات متقابل را برای احتمالات شرطی هم می توان تعریف کرد. به عنوان مثال:

$$I(X:Y|z) := H(X|z) + H(Y|z) - H(X,Y|z) \qquad (۱۵)$$

که در آن آنتروپی های شرطی مطابق با رابطه (۱۰) تعریف شده اند. آنتروپی شانون برای یک توزیع احتمال را می توان به صورت زیر نیز نوشت:

$$H(X) = -\sum_x p_x \log p_x = -\langle \log p_x \rangle, \qquad (۱۶)$$

که در آن معنای $\langle\ \rangle$ متوسط نسبت به همان تابع توزیع احتمال است. به این ترتیب آنتروپی شانون برای هر توزیع احتمالی، متوسط $\log \frac{1}{p_x}$ روی همان تابع توزیع احتمال است. با استفاده از این تعبیر می توان آنتروپی شرطی را نیز به صورت زیر بازنویسی کرد:

$$\begin{aligned} H(X|Y) &= H(X,Y) - H(Y) = -\sum_{x,y} p(x,y)\log p(x,y) + \sum_y p(y)\log p(y) \\ &= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_{x,y} p(x,y)\log p(y) \\ &= -\left\langle \log \frac{p(x,y)}{p(y)} \right\rangle. \end{aligned} \qquad (۱۷)$$

توجه کنید که در این رابطه ها $\langle\ \rangle$ به معنای متوسط روی تابع توزیع $p(x,y)$ است. به طور کلی در همه روابط مشابه متوسط روی بزرگ ترین تابع توزیع است. هم چنین اطلاعات متقابل دو تابع توزیع کلاسیک به صورت ساده زیر نوشته می شود:

$$I(X:Y) = H(X) + H(Y) - H(X,Y) = -\left\langle \log \frac{p(x)p(y)}{p(x,y)} \right\rangle. \qquad (۱۸)$$

■ مثال: منبع

$$X = \{000(1/2), 111(1/2)\} \qquad (۱۹)$$

را که درآن اعداد داخل پرانتز احتمالات رشته ها رانشان می دهند در نظر بگیرید. برای این منبع داریم $H(X) = 1$ می توانیم رشته سوالات خودرا به ترتیب زیرتنظیم کنیم:

۱ ـ آیا همه اعداد صفر هستند؟

درهردوصورت جواب آری یا خیر ما به رشته موردنظری که سوال کننده درنظرگرفته است پی می بریم. یعنی یک سوال برای رسیدن به رشته

مورد نظر کفایت می کند.

حال منبع زیررادرنظربگیرید:

$$X = \{000(1/4), 111(1/4), 001(1/4), 110(1/4)\} \qquad (۲۰)$$

حال می توانیم سوالات خود را به شکل زیرتنظیم کنیم:

۱ ـ آیا اکثریت بیت ها صفرهستند؟

۲ ـ آیا همه بیت ها مثل هم هستند؟

دراین صورت با دوسوال به رشته مورد نظرمی رسیم و $H(X)$ نیزبرابربا 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

$$H(X \mid 1) = 1, \qquad H(X \mid 0) = 1, \longrightarrow H(X|Y) = 1. \qquad (۲۱)$$

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

$$I(X : Y) = H(X) - H(X \mid Y) = 2 - 1 = 1. \qquad (۲۲)$$

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:

$$H(X|00) = 0, \quad H(X|01) = 0, \quad H(X|10) = 0, \quad H(X|11) = 0 \quad \longrightarrow H(X|Y) = 0 \qquad (۲۳)$$

در نتیجه خواهیم داشت:

$$I(X : Y) = H(X) - H(X \mid Y) = 2 - 0 = 2. \qquad (۲۴)$$

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

دراین بخش خواص ریاضی توابع اطلاعات را بررسی می کنیم. تقریباً همه این خواص از یک قضیه ساده ولی مهم بدست می آیند.

■ **قضیه :** تابع انتروپی شانون در رابطه زیر صدق می کند که در آن q هر تابع توزیع احتمال دلخواهی ا ست:

$$H(X) \leq -\sum_{x} p(x) \log_2 q(x). \qquad (۲۵)$$

که درآن تساوی فقط وقتی برقرارمی شود که دو توزیع احتمال یکی باشند.

■ **اثبات:** بارسم کردن تابع لگاریتم و تابع $x-1$، می توان نشان داد که تابع لگاریتم درخاصیت زیرصدق می کند:

$$\ln x \leq (x) - 1, \qquad (۲۶)$$

که درآن تساوی فقط برای $x = 1$ برقرارمی شود. حال قرارمی دهیم $x = \frac{q(x)}{p(x)}$ و درنتیجه

$$\ln \frac{q(x)}{p(x)} \leq \frac{q(x)}{p(x)} - 1, \quad \forall \ x, \qquad (۲۷)$$

که درآن تساوی فقط وقتی برقرارمی شود که $q(x) = p(x)$. درنتیجه

$$\sum_{x} p(x) \ln \frac{q(x)}{p(x)} \leq \sum_{x} q(x) - \sum_{x} p(x) = 0, \qquad (۲۸)$$

که همان نامساوی ای است که می خواستیم ثابت کنیم. دقت کنید که نامساوی (۲۶) فقط برای $\ln(x)$ درست است و نه برای لگاریتم در پایه ۲. ولی پس از بدست آوردن رابطه (۲۸) ما می توانیم 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

$$\sum_x p(x) \log \frac{q(x)}{p(x)} = 0. \tag{۲۹}$$

این تساوی را به شکل زیربازنویسی می کنیم

$$\sum_x p(x)(\log \frac{q(x)}{p(x)} - (\frac{q(x)}{p(x)} - 1)) = 0. \tag{۳۰}$$

حال دقت می کنیم که بنابرنامساوی (27) جملات داخل پرانتز همگی کوچک تراز یا مساوی باصفر هستند. صفرشدن این جمع به این معناست که همه این جملات برابرباصفرهستند که باتوجه به نامساوی (26) به معنای آن است که برای همه i ها $q(x) = p(x)$. یعنی دوتابع توزیع احتمال یکی هستند.

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

■ **اثبات:** درقضیه قبلی قرارمی دهیم $q(x) = \frac{1}{M}$. درنتیجه خواهیم داشت :

$$\sum_{x=1}^{M} p(x) \log \frac{\frac{1}{M}}{p_x} = H - \log M \leq 0, \tag{۳۱}$$

که درآن تساوی فقط وقتی برقرارمی شود که $p(x) = \frac{1}{M}$.

■ **نتیجه ۲ : خاصیت زیر جمع پذیری** [۱] برای دو متغیرتصادفی X, Y نامساوی زیربرقراراست

$$H(X,Y) \leq H(X) + H(Y), \tag{۳۲}$$

که درآن تساوی فقط وقتی برقرارمی شود که X, Y متغیرهای مستقل باشند.

[۱] **Property Subadditivity**

■ **اثبات :** تابع توزیع دومتغیررا با $p(x,y)$ نشان می دهیم. درنتیجه خواهیم داشت:

$$p_1(x) := \sum_y p(x,y), \qquad p_2(y) := \sum_x p(x,y). \tag{۳۳}$$

حال تابع توزیع $q(x,y) := p_1(x)p_2(y)$ رادرنظرمی گیریم واز قضیه ای که ثابت کردیم استفاده می کنیم :

$$\sum_{x,y} p(x,y) \log \frac{q(x,y)}{p(x,y)} \leq 0 \tag{۳۴}$$

که درآن تساوی فقط وقتی برقرارمی شود که $p(x,y) = q(x,y) = p_1(x)p_2(y)$. اما نامساوی بالا راوقتی بازنویسی کنیم چیزی نیست جز

$$H(X,Y) \leq H(X) + H(Y), \tag{۳۵}$$

که می توان آن را به شکل زیر نیز نوشت:

$$H(X|Y) \leq H(X). \tag{۳۶}$$

این نامساوی در واقع بیان می کند که دانستن یک متغیر تصادفی دیگر مثل Y همواره از انتروپی موجود در متغیر X کم می کند (چیزی در باره آن به ما می گوید و اطلاعات ما را آفزایش می دهد). اگر بخواهیم از زبان زندگی روزمره کمک بگیریم می توانیم بگوییم که معنای نامساوی (۳۵) به عنوان مثال این است که : اطلاعات موجود درجمله « فردا هواابری است وباران می بارد» کمترازمجموع اطلاعاتی است که دردوجمله « فرداهوا ابری است» و « فردا هوابارانی 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 دوبرابرکسی که فقط یکی ازجملات را به ما می گوید به ما اطلاع نمی دهد. این مثال طبیعتا یک مثال کلامی است و کمی نیست.

■ **نتیجه۳:** اطلاعات متقابل یک کمیت نامنفی است. این نتیجه ازتعریف اطلاعات متقابل و نتیجه ۲ بدست می آید.

■ **تمرین:** احتمالات نسبی دو متغیرتصادفی 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: $P(x \mid y) \leq P(x)$ و برای بعضی دیگر داشته باشیم $P(x \mid y) \geq P(x)$.

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

$$H_0(X) \geq \lambda H_1(X) + (1-\lambda)H_2(X). \qquad (۳۷)$$

به اصطلاح می گوییم که اطلاعات یک تابع محدب روبه پایین است که به یادماندن شکل آن را نیزدرذهن 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 ازنامساوی اساسی ای که ثابت کردیم استفاده می کنیم. باکمی خلاصه نویسی درنمادها خواهیم داشت:

$$H_0 - \lambda H_1 - (1-\lambda)H_2$$

$$= \sum p_0 \log \frac{1}{p_0} - \lambda \sum p_1 \log \frac{1}{p_1} - (1-\lambda) \sum p_2 \log \frac{1}{p_2}$$

$$= \sum (\lambda p_1 + (1-\lambda)p_2) \log \frac{1}{\lambda p_1 + (1-\lambda)p_2} - \lambda \sum p_1 \log \frac{1}{p_1} - (1-\lambda) \sum p_2 \log \frac{1}{p_2}$$

$$= \lambda \sum p_1 \log \frac{p_1}{\lambda p_1 + (1-\lambda)p_2} + (1-\lambda) \sum p_2 \log \frac{p_2}{\lambda p_1 + (1-\lambda)p_2} \geq 0, \qquad (۳۸)$$

که درخط آخر ازنامساوی 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:

$$\{P(0) = 1/2,\ P(1) = 1/2\}, \qquad ,\{Q(0) = 1/3,\ Q(1) = 2/3\}. \qquad (۳۹)$$

حال درستی رابطه تحدب را برای آنتروپی شانون تحقیق کنید.

■ **خاصیت زیرجمع پذیری قوی:**

به ازای هر سه متغیر تصادفی X, Y و Z نامساوی زیر برقرار است:

$$H(X|Y,Z) \leq H(X|Y). \qquad (۴۰)$$

این رابطه در واقع بیان می کند که ایجاد شرط اضافه باعث کاهش انتروپی شانون می شود. به عبارت دیگر اگر علاوه بر مقدار Y مقدار Z را نیز تعیین کنیم، اطلاعات کمتری در X باقی می ماند. این نامساوی را به صورت زیر نیز می توان نوشت:

$$H(X,Y,Z) - H(Y,Z) \leq H(X,Y) - H(Y) \qquad (۴۱)$$

■ **اثبات:** آنتروپی شانون برای یک توزیع احتمال را می توان به صورت زیر نیز نوشت:

$$H(X) = -\sum_x p_x \log p_x = -\langle \log p_x \rangle, \qquad (۴۲)$$

که در آن معنای $\langle\ \rangle$ متوسط نسبت به همان تابع توزیع احتمال است. به این ترتیب آنتروپی شانون برای هر توزیع احتمالی، متوسط $\log \frac{1}{p_x}$ روی همان تابع توزیع احتمال است. با استفاده از این تعبیر می توان آنتروپی شرطی را نیز به صورت زیر بازنویسی کرد:

$$\begin{aligned} H(X|Y) &= H(X,Y) - H(Y) = -\sum_{x,y} p(x,y)\log p(x,y) + \sum_{y} p(y)\log p(y) \\ &= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_{x,y} p(x,y)\log p(y) \\ &= -\left\langle \log \frac{p(x,y)}{p(y)} \right\rangle. \end{aligned} \tag{۴۳}$$

هم چنین اطلاعات متقابل دو تابع توزیع کلاسیک به صورت ساده زیر نوشته می شود:

$$I(X:Y) = H(X) + H(Y) - H(X,Y) = -\left\langle \log \frac{p(x)p(y)}{p(x,y)} \right\rangle. \tag{۴۴}$$

این بازنویسی به ما امکان می دهد که خاصیت زیرجمع پذیری قوی را برای آنتروپی شانون براحتی ثابت کنیم. می نویسیم:

$$I(X,Z:Y) = H(X,Z) + H(Y) - H(X,Z,Y) = -\left\langle \log \frac{p(x,z)p(y)}{p(x,y,z)} \right\rangle. \tag{۴۵}$$

برای این منظور توجه می کنیم که خاصیت زیرجمع پذیری قوی چیزی نیست جز رابطه ی زیر:

$$I(X:YZ) \geq I(X:Y) \tag{۴۶}$$

که از نظر شهودی معنای روشنی نیز دارد: دانستن Y,Z مسلما بیش از دانستن Y به تنهایی به ما کمک می کند که در باره X اطلاعات بدست آوریم. برای اثبات این رابطه با کمی محاسبه بدست می آوریم

$$\begin{aligned} I(X:YZ) - I(X:Y) &= -\left\langle \log \frac{p(x)p(y,z)}{p(x,y,z)} - \log \frac{p(x)p(y)}{p(x,y)} \right\rangle \\ &= -\left\langle \log \frac{p(y,z)p(x,y)}{p(x,y,z)p(y)} \right\rangle \\ &= -\left\langle \log \frac{p(x,y)}{p(y)} \frac{p(y,z)}{p(y)} \frac{p(y)}{p(x,y,z)} \right\rangle \\ &= -\left\langle \log \frac{p(x|y)p(z|y)}{p(x,z|y)} \right\rangle \end{aligned} \tag{۴۷}$$

حال باید ثابت کنیم که عبارت آخر مثبت است. اما اگر به رابطه ی (۴۴) نگاه کنیم، جمله آخر چیزی نیست جز:

$$\sum_y p(y) I(X : Z|y) \geq 0, \tag{۴۸}$$

که در نامساوی آخری از این استفاده کرده ایم که اطلاعات متقابل همواره مثبت است.

---

۴ تعریف کانال کلاسیک

منظوراز یک کانال کلاسیک عملگری است که یک آنزامبل تصادفی X را به آنزامبل تصادفی Y تبدیل می کند. بهترین مثال آن هرنوع کانال مخابراتی کلاسیک است. X را ورودی کانال و Y راخروجی آن می نامیم. یک کانال بدون نوفه کانالی است که خروجی آن دقیقاً باورودی آن برابراست. بجزاین کانال ایده آل هرکانال دیگری علائم ورودی $x_i \in X$ را بااحتمالات معینِ $P(y_j|x_i)$ به علائم خروجی $y_j \in Y$ تبدیل می کند. هرگاه‌درخروجی کانال علامت y_j رادریافت کنیم می توانیم احتمال شرطی این که چه علامت x_i ای منجربه این علامت درخروجی شده است را حساب کنیم. درواقع داریم:

$$\begin{aligned} P(x_i|y_j) &= \frac{P(x_i, y_j)}{P(y_j)} = \frac{P(y_j, x_i)}{\sum_{x_i} P(y_j, x_i)} \\ &= \frac{P(y_j|x_i)P(x_i)}{\sum_{x_i} P(y_j|x_i)P(x_i))} \end{aligned} \tag{۴۹}$$

درآخرین عبارت $P(x_i)$ مشخصه منبع X و $P(y_j|x_i)$ مشخصه کانال است وهردومعلوم هستند.

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

$$P(0|0) = 1 - p, \qquad P(1|1) = 1 - q. \tag{۵۰}$$

هرگاه آنزامبل ورودی به صورت

$$X = \{P(0) = a,\ P(1) = 1 - a\}, \tag{۵۱}$$

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

$$H(X), \qquad H(Y), \qquad H(X|Y), \qquad H(Y|X), \qquad I(Y\ :\ X), \qquad I(X\ :\ Y). \qquad (۵۲)$$

■ **قضیه:** اطلاعات پردازش شده دریک کانال $I(X;Y)$ تابع محدبی از احتمالات ورودیِ X است.

دریک کانال آنزامبل ورودی را با X وآنزامبل خروجی را با Y نشان می دهیم. احتمالات شرطی $P(y|x)$ درواقع مشخصه کانال هستند و احتمال تبدیل پیام x به y را درطول کانال نشان می دهند وربطی به احتمال پیام های ورودی ندارند. حال هرگاه برای آنزامبل 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 درنظربگیریم آنگاه باتوجه به تعاریف زیر:

$$\begin{aligned} P(y) &= \sum_x P(y|x)P(x), \\ P(x,y) &= P(y|x)P(x), \end{aligned} \qquad (۵۳)$$

خواهیم داشت

$$\begin{aligned} P_0(x,y) &= \lambda P_1(x,y) + (1-\lambda)P_2(x,y) \\ P_0(y) &= \lambda P_1(y) + (1-\lambda)P_2(y). \end{aligned} \qquad (۵۴)$$

باترکیب این روابط با تعریف اطلاعات متقابل و هم چنین محدب 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:

$$P(0|0) = 1-p, \qquad P(1|1) = 1-q. \qquad (۵۵)$$

آنزامبل های ورودی را به صورت زیر در نظر بگیرید:

$$X_0 := \{P(0) = 1/2,\ P(1) = 1/2\}, \qquad , X_1 := \{Q(0) = 1/3,\ Q(1) = 2/3\}. \qquad (۵۶)$$

الف: درستی رابطه تحدب 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 $x_1x_2x_3$ ارسال شده باشند چقدر است؟

■ **تمرین:** جفت متغیرتصادفی (X,Y) را مطابق جدول زیر درنظربگیرید: Y ناشی از انداختن یک طاس است که مقادیر ۱ تا ۶ را به خود می گیرد و X نیز دومقدارمتفاوت یک سکه است که مقادیر a یا b را اختیارمی کند.

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

(57)

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

۵ فشرده سازی اطلاعات درغیاب نوفه

بهترین کاربرای فهم فشرده سازی اطلاعات مطالعه یک مثال ساده است. فرض کنیدکه هدف ما ارسال متن هایی است که تنهاازچهارحرف 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 درمخابرات دیجیتال معمول است، به ترتیب زیر کدکنیم .

$$A \longrightarrow 00$$
$$B \longrightarrow 01$$
$$C \longrightarrow 10$$
$$D \longrightarrow 11. \qquad (۵۸)$$

دراین صورت به ازای هرحرف دوبیت مخابره کرده ایم. حال سوال این است که آیا می توانیم یک روش کد کردن به کاربریم که درآن طول به ازای هرحرف تعداد بیت هایی که به طور متوسط مخابره می کنیم کمتراز 2 باشد؟

فرض کنید که این حروف درمتن های یادشده بااحتمالات زیر ظاهرمی شوند:

$$P(A) = \frac{1}{8} \qquad P(B) = \frac{1}{8} \qquad P(C) = \frac{1}{4} \qquad P(D) = \frac{1}{2}. \tag{۵۹}$$

حال روش کدگذاری زیررا به کارمی بریم:

$$\begin{aligned} D &\longrightarrow 0 \\ C &\longrightarrow 10 \\ B &\longrightarrow 110 \\ A &\longrightarrow 111. \end{aligned} \tag{۶۰}$$

دراین روش کدگذاری برای بعضی ازحروف بیش از دوبیت به کاربرده ایم ولی اگر طول متوسط کدهایی را که برای حروف به کاربرده ایم محاسبه کنیم نتیجه جالب توجه خواهد بود. این طول متوسط برابراست با:

$$\langle l \rangle = \sum_{i=1}^{4} l_i \times p_i = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = \frac{7}{4}. \tag{۶۱}$$

بنابراین با یک کدگذاری مناسب توانسته ایم طول متوسط رشته بیت هایی را که برای مخابره پیام بکاربرده 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 دربرندارد وهررشته ای ازبیت ها به طوریکتا به متن اولیه بازگشایی می شود. به عنوان مثال رشته زیر

$$0\,1\,0\,0\,0\,1\,0\,0\,0\,1\,1\,0\,1\,1\,1. \tag{۶۲}$$

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

$$D\ C\ D\ D\ C\ D\ D\ B\ A. \tag{۶۳}$$

این که چه نوع کد هایی یکتاگشاهستند موضوعی است که مادردرسهای آینده به آن خواهیم پرداخت وفعلاً موضوع بحث مانیست. ولی یک نکته‌مهم رابابد ذکرکنیم: هرگاه آنتروپی متغیرتصادفی $X = \{A,\ B,\ C,\ D\}$ را بااحتمالات ذکرشده حساب کنیم حاصل آن برابرخواهد بودبا:

$$H(X) \quad = \quad \sum_{i=1}^{4} p_i \log_2(\frac{1}{p_i}) = \frac{1}{2} \times \log_2(2) + \frac{1}{4} \times \log_2(4) + \frac{1}{8} \times \log_2(8) + \frac{1}{8} \times \log_2(8)$$

$$= \frac{1}{2}\times 1+\frac{1}{4}\times 2+\frac{1}{8}\times 3+\frac{1}{8}\times 3=\frac{7}{4}. \qquad (۶۴)$$

بنابراین دراین مثال خاص طول متوسط کدگزاری ای که به کاربردیم با میزان اطلاعات موجود درمتن برابراست. 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. ضمنا می دانیم که بین حروف یک همبستگی وجود دارد : این همبستگی به احتمالات زیر مشخص شده است:

$$P(x|x)=5/8, \qquad P(y\neq x|x)=1/8. \qquad (۶۵)$$

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.

$$P(A)=\frac{1}{32} \qquad P(B)=\frac{1}{32} \qquad P(C)=\frac{1}{16} \qquad P(D)=\frac{1}{8} \qquad P(E)=\frac{1}{4} \qquad P(F)=\frac{1}{2} \qquad (۶۶)$$

■ تمرین: 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.

$$\begin{array}{llll} P(A)=\frac{1}{128} & P(B)=\frac{1}{128} & P(C)=\frac{1}{64} & P(D)=\frac{1}{32} \\ P(E)=\frac{1}{16} & P(F)=\frac{1}{8} & P(G)=\frac{1}{4} & P(H)=\frac{1}{2} \end{array} \qquad (۶۷)$$

■ تمرین: 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.

$$\begin{aligned} &P(A)=\frac{1}{256} \qquad P(B)=\frac{1}{256} \qquad P(C)=\frac{1}{128} \qquad P(D)=\frac{1}{128} \\ &P(E)=\frac{1}{128} \qquad P(F)=\frac{1}{32} \qquad P(G)=\frac{1}{16} \qquad P(H)=\frac{1}{8} \\ &P(K)=\frac{1}{4} \qquad P(L)=\frac{1}{2}. \end{aligned} \tag{۶۸}$$

بعدازذکراین مثال می خواهیم بفهمیم که درحالت کلی چگونه می توان اطلاعات موجود دریک منبع X را فشرده کرد. فرض کنید که منبع متن هایی تولید می کند که این متن هاازالفبای $A=\{x_1,x_2,\cdots x_N\}$ تشکیل شده‌اند و احتمال ظاهرشدن هرحرف مثل x_i دراین متن ها با p_i داده می شود. بنابراین یک منبع رامی توان به عنوان یک متغیرتصادفی با اطلاعات معین $H(X)$ درنظرگرفت. برای سادگی فرض کنید که N توانی از 2 است یعنی $N=2^n$. حال اگربدون توجه به احتمالات ظاهرشدن حروف مختلف بخواهیم متن هارامخابره کنیم می توانیم هرحرف الفبای A را بایک رشته n تای از بیت های 0 و1 کدگذاری کنیم. دراین صورت برای هر متن که شامل M حرف است تعداد Mn بیت مصرف می کنیم یابه عبارت دیگر به ازای هرحرف الفبا n بیت مصرف می کنیم. ولی می توانیم روش کدگذاری بهتری به ترتیب زیربکاربریم.

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 برابراست با N^m . ولی نکته این است که ما تنها می بایست رشته های نمونه را کد کنیم. بعنوان مثال درست است که هرحرف از حروف الفبای انگلیسی با یک فرکانس مشخص در نوشتارهای انگلیسی ظاهر می شود اما رشته هاییm حرفی مثل

$$AAAAAAAAAAAAAAAAAA$$

یا

$$AAABBBBAAABBBAAABBB$$

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

حروف ظاهرشده نایاب هستند. در عوض رشته ای مثل

$$ABQUQIPQUTNVIABURQOR \tag{۶۹}$$

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.

حال دقت می کنیم که احتمال ظاهرشدن بسیاری ازرشته هاآنقدر ناچیزاست که نیازی به کدکردن آنهانیست وباکدکردن تنها رشته های متعارف (رشته هایی که زیاد ظاهرمی شوند) چیزی ازدست نمی دهیم. به این ترتیب یعنی باکدکردن تنهارشته های متعارف ماقادرمی شویم که بیت های کمتری برای مخابره متن های منبع X مصرف کنیم. اما رشته های متعارف کدام هاهستند؟ و کدکردن آنها چقدرباعث فشرده شدن پیام هامی شود. در هررشته 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 خواهدبود. 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 $\mathcal{Q}_m$ نشان دهیم خواهیم داشت

$$\mathcal{Q}_m = \frac{m!}{(mp_1)!(mp_2)!\cdots(mp_N)!} \tag{۷۰}$$

اما بااستفاده از تقریب استرلینگ می توانیم بنویسیم:

$$\log_2 \mathcal{Q}_m = \log_2\left(\frac{m!}{(mp_1)!(mp_2)!\cdots(mp_n)!}\right) \approx m\left(\sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i}\right) \equiv mH(X) \tag{۷۱}$$

که درآن تابع $H(X)$ به صورت زیرتعریف شده است:

$$H(X) := \sum_{i=1}^{N} p_i \log_2\left(\frac{1}{p_i}\right) \tag{۷۲}$$

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.

بنابراین تعداد جملات متعارف با طول m باتقریب بسیارخوب برابرخواهد بود با

$$\mathcal{Q}_m \approx 2^{mH(X)} \tag{۷۳}$$

حال اگر تعداد جملات متعارف برابرباشدبا مقدارفوق، می توانیم هرکدام از این جملات را با یک رشته بیت های 0 و 1 کدگذاری کنیم و مسلم است که تعداد بیت هایی که برای این کاراحتیاج داریم برابراست با $mH(X)$. ازآنجا که هررشته دارای m حرف بوده است مثل این است که درعمل برای مخابره هرحرف $k := H(X)$ بیت بکاربرده ایم. ازآنجا 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 بااین روش کدکردن که آن را $Block\ coding$ می گوییم برای هرحرف به جای n بیت $H(X)$ بیت مصرف کرده ایم که ازn کمتراست.

آنچه که دربالاگفته شد محتوای کلی قضیه شانون درمورد 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 غیرمتعارف چه مقدارمرتکب خطامی شویم؟ آیا بیش از این هم می توان پیام های منبع X را فشرده کرد؟ برای پاسخ به این سوالات سعی می کنیم ابتداتعاریف 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	$P(x)$

(74)

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.

۶ نگاهی دوباره به اطلاعات شرطی واطلاعات متقابل

بعدازفهم رشته های متعارف وفشرده سازی می توانیم نگاهی دوباره به اطلاعات شرطی واطلاعات متقابل بیندازیم. ازاین زاویه جدید می توانیم تعریف متفاوتی برای تابع $H(X)$ پیداکنیم. یادگرفتیم که تعداد رشته های متعارف m حرفی برابراست با $2^{mH(X)}$. این حرف به این معناست که اگرکسی یک رشته معین را به عنوان سوال برای مادرنظرگرفته باشد و ازمابخواهد دریک مسابقه به اصطلاح « بیست سوالی» باپرسیدن سوال هایی که پاسخ آنهاتنها آری یا خیراست به آن رشته معین دست پیداکنیم دربهترین حالت می بایست تعداد $mH(X)$ بارسوال کنیم. زیرابهترین نحوه سوال کردن نحوه ای است که دران تعداد رشته های باقیمانده را به نصف مقدارقبلی کاهش می دهد و $2^{mH(X)}$ را به $2^{mH(X)-1}$، $2^{mH(X)-2}$ وسرانجام 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 $P(X|\mid y)$ 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

$$I(X:Y) = H(X) - H(X \mid Y) \tag{۷۵}$$

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. معمولا 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 (۷۵) می فهمیم که اگر کانال دارای هیچ نوع خطایی نباشد، آنگاه دانستن رشته خروجی دقیقا رشته ورودی را تعیین می کند و در نتیجه

$$H(X|Y) = 0 \qquad \longrightarrow I(X:Y) = H(X). \tag{۷۶}$$

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

$$H(X|Y) = H(X) \qquad \longrightarrow I(X:Y) = 0. \tag{۷۷}$$

---

۷ 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:

$$\mathcal{N}(m=18)=2^{18}=262144. \tag{۷۸}$$

حال از خود سوال می کنیم که معمولا 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 $AAAAA\cdots A$ را که کلا 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 $AAAABBBAAABB\cdots B$ که مخلوطی از دو حرف الفبا را دارند با احتمال کمتری تولید می کند ولی تعداد چنین رشته هایی آنقدر زیاد است که عملا 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. 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:

$$N_0=\binom{18}{12}=18564. \tag{۷۹}$$

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:

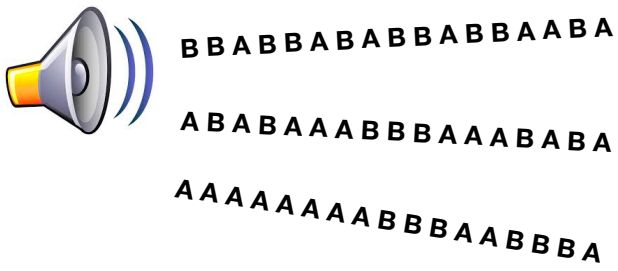
$$P_0=(\frac{2}{3})^{12}(\frac{1}{3})^6\binom{18}{12}=0.1962. \tag{۸۰}$$

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 دقیقا 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:

$$\sigma=\sqrt{mp(1-p)}=\sqrt{18\,\frac{2}{3}\,\frac{1}{3}}=2. \tag{۸۱}$$

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

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



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

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:

$$N_1 = \sum_{x=10}^{14} \binom{18}{x} = 105774. \qquad (۸۲)$$

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:

$$P_1 = \sum_{x=10}^{14} (\frac{2}{3})^x (\frac{1}{3})^{18-x} \binom{18}{x} = 0.7907. \qquad (۸۳)$$

بنابراین با در نظر گرفتن تنها یک انحراف معیار که مقدار کمِ 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

معیاریعنی به میزان ۴ را اختیار کنیم حاصل چه خواهد شد. پاسخ به سادگی بدست می آید.

$$N_2 = \sum_{x=8}^{16} \binom{18}{x} = 199121, \tag{۸۴}$$

و

$$P_2 = \sum_{x=8}^{16} (\frac{2}{3})^x (\frac{1}{3})^{18-x} \binom{18}{x} = 0.9788. \tag{۸۵}$$

یعنی تقریبا 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:

$$P(x) = (\frac{2}{3})^x (\frac{1}{3})^{18-x} \binom{18}{x} \qquad 10 \leq x \leq 12 \tag{۸۶}$$

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

$$P(10) = 0.1156 \qquad P(11) = 0.1682 \qquad P(12) = 0.1962 \qquad P(13) = 0.1811 \qquad P(14) = 0.1294. \tag{۸۷}$$

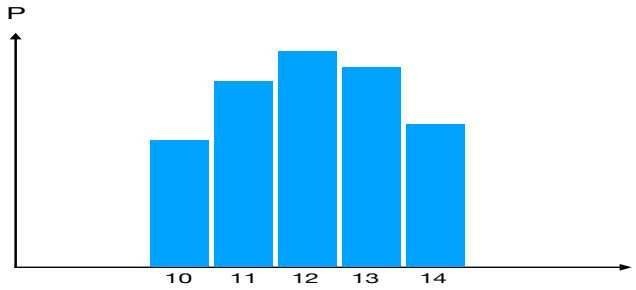
این احتمالات در شکل (۴) نشان داده شده اند. ب

به نظر می رسد که تابع توزیع احتمال تقریبا 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 ۴۴ تا ۵۲ است. برای این رشته ها خواهیم داشت:

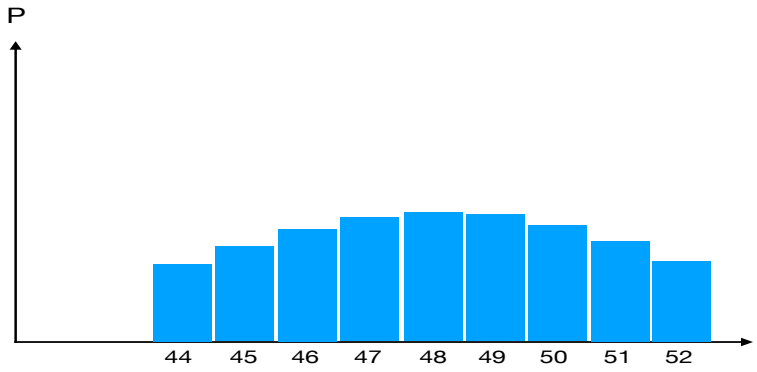
$$P(x) = (\frac{2}{3})^x (\frac{1}{3})^{72-x} \binom{72}{x} \qquad 44 \leq x \leq 52. \tag{۸۸}$$

و به طور صریح

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



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.



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.

$$P(49) = 0.0973 \qquad P(50) = 0.0895 \qquad P(51) = 0.0772 \qquad P(52) = 0.0624. \tag{۸۹}$$

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

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 ـ 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.
- دوـ برای طول های بسیار بزرگ همه این رشته های نمونه با احتمال تقریبا مساوی تولید می شوند.
- سه ـ 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:

$$\mathcal{N}_{typ} \approx 2^{NH(X)} \tag{۹۰}$$

و هر کدام از این رشته ها نیز با احتمال

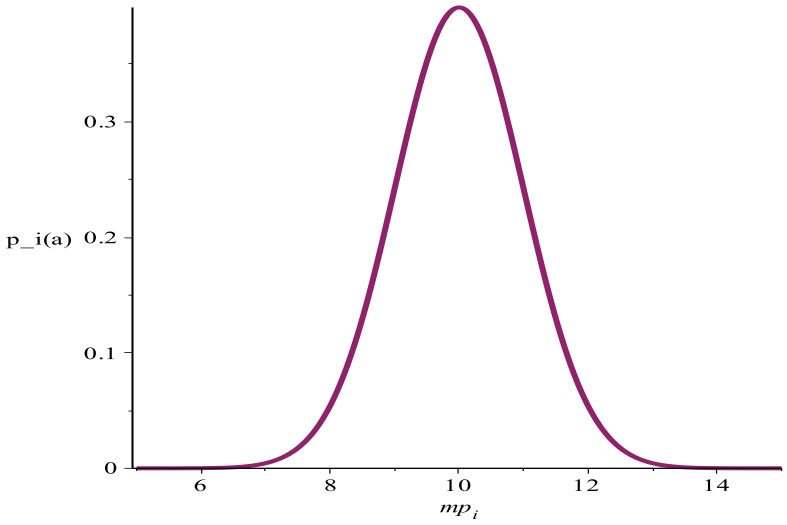
$$2^{-NH(X)} \tag{۹۱}$$

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 [۲] 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 دقیقا 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 دوجمله ای (و در حد m های بزرگ با یک تابع گاووسی) 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 دقیقا چه رشته ای است؟ در این تعریف حتما می بایست یک حد و اندازه وجود داشته باشید. بدون این حد و اندازه یا معیار نمی توان دقیقا گفت که یک رشته مثل آیا یک رشته نمونه است یا خیر؟

[۲] typical sequence



شکل ۶: تعداد حرف های x_i در یک رشته مثل a از یک تابع احتمال گاووسی 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.

رشته $\alpha = \alpha_1\alpha_2\alpha_3\cdots\alpha_m$ را درنظربگیرید. تعداد حروف x_j دراین رشته را با $f_j(\alpha)$ نشان دهید. تعداد حروف x_j در رشته های به طول m به طور متوسط برابراست با mp_j وواریانس توزیع احتمال حول این مقدارمتوسط برابراست با $\sigma_j := \sqrt{mp_j(1-p_j)}$. رشته متعارف رشته ای است که تفاوت تعداد واقعی هرکدام ازحروف مثل x_j ازتعداد متوسط آن یعنی mp_j درمقایسه باواریانس σ_j مقدار معینی باشد.

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

$$|\ \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}}\ | < k \quad \forall\ i = 1, 2, \cdots N. \qquad (۹۲)$$

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 $(Chebyshev\ inequlity)$:

الف : فرض کنید که متغیرتصادفی X مقادیر مثبت $\{x_1, x_2, \cdots x_N\}$ را بااحتمالات $\{p_1, p_2, \cdots p_N\}$ اختیارمی کند . دراین صورت به ازای هرعدد مثبت α،

$$P(X \geq \alpha) \leq \frac{\overline{X}}{\alpha} \tag{۹۳}$$

که درآن $\overline{X}$ متوسط متغیر تصادفی X است.

■ اثبات :

$$P(X \geq \alpha) = \sum_{x=\alpha}^{\infty} P(x) \leq \sum_{x=\alpha}^{\infty} \frac{x}{\alpha} P(x) \leq \frac{\overline{X}}{\alpha}. \tag{۹۴}$$

■ لم دوم : نامساوی دوم چبیشف $(Chebyshev\ inequlity)$:

حال فرض کنید که متغیرتصادفی X مقادیر دلخواه مثبت یامنفی اختیارمی کند. دراین صورت به ازای هرعدد k

$$P((X - \overline{X})^2 \geq k^2 \sigma_x^2) \leq \frac{1}{k^2}. \tag{۹۵}$$

اثبات : متغیر تصادفی $T = (X - \overline{X})^2$ را درنظرمی گیریم. این متغیرفقط مقادیر مثبت را اختیار می کند.

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

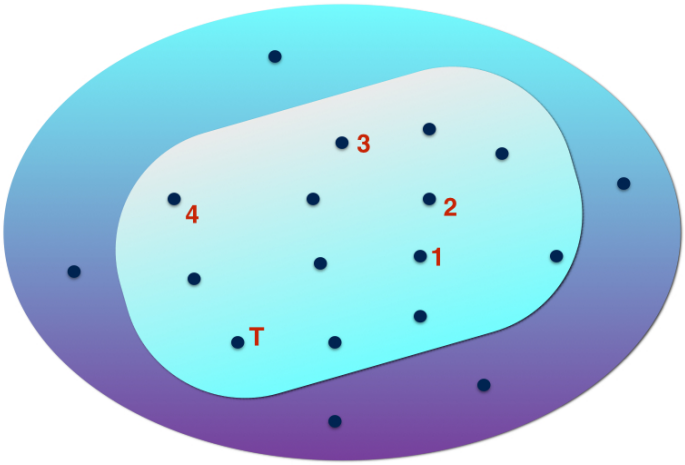
$$P(T \geq \alpha) \leq \frac{\overline{T}}{\alpha}. \tag{۹۶}$$

هرگاه به جای α درنامساوی اخیر قراردهیم $k^2\sigma_x^2$ بدست می آوریم:

$$P((X - \overline{X})^2 \geq k^2 \sigma_x^2) \leq \frac{\sigma_x^2}{k^2 \sigma_x^2} = \frac{1}{k^2}. \tag{۹۷}$$

این نامساوی را به شکل زیر نیز می توان نوشت:

$$P(|X - \overline{X}| \geq k\sigma_x) \leq \frac{1}{k^2}, \tag{۹۸}$$



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.

- تمرین: در الفبای انگلیسی حرفZ کمترین فرکانس را دارد و احتمال یافتن آن در متن های انگلیسی برابراست 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 $p << 1$ باشد استفاده کنید.

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

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 عددk را بزرگ تر انتخاب کنیم یعنی این که تعریف خود را از رشته نمونه فراخ تر کنیم که کاملا طبیعی به نظر می رسد. در نتیجه در یک رشته به طول m تعداد متوسط x_i دقیقا 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 (۹۲) پیروی نکند. با توجه به رابطه زیر در مورد احتمالات پیشامدهای گوناگون

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

داریم:

$$P_0 \leq Prob\{|\ \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}}\ | \geq k,\ \ i \text{ برای حداقل یک } \} = \sum_{i=1}^{N} P(|\ \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}}\ | \geq k). \qquad (۹۹)$$

پس اگر قرار دهیم

$$\epsilon = \frac{N}{k^2} \qquad (۱۰۰)$$

از نامساوی چبیشف نتیجه می گیریم که

$$P_0 \leq \sum_{i=1}^{N} \frac{1}{k^2} = \frac{N}{k^2} = \epsilon. \qquad (۱۰۱)$$

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:

$$1 - \epsilon \leq P_{typical} \leq 1. \qquad (۱۰۲)$$

هرگاه قرار دهیم $k = 10$ و حروف الفبای انگلیسی را ۳۲ تا بگیریم نتیجه می گیریم که احتمال اینکه یک رشته نمونه باشد از ۶۸.۰ بیشتر است. دقت کنید که این یک 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 ا نگلیسی 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 معمولا 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:

$$2^{-mH - A\sqrt{m}} \leq P(\alpha \in typical) \leq 2^{-mH + A\sqrt{m}}. \qquad (۱۰۳)$$

که در آن

$$A = -k \sum_{i=1}^{N} \sqrt{p_i(1 - p_i)} \log p_i. \qquad (۱۰۴)$$

در حد رشته های طولانی ، یعنی $(m >> 1)$ 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 برابراست با

$$P(\alpha \in typical) = 2^{-m[H(X) + \delta_m]} \qquad (۱۰۵)$$

که در آن

$$\lim_{m \in \infty} \delta_m = 0. \qquad (۱۰۶)$$

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 $\alpha = AABBCDBDADDAABCCCC$ چقدر است؟ هرگاه که حروف را مستقل از هم بگیریم پاسخ این سوال برابر است با

$$P(\alpha) = P_A P_A P_B P_B P_C P_C \cdots.$$

بنابراین در حالت کلی احتمال پیداکردن یک رشته مثل α برابراست با:

$$P(\alpha) = P_1^{f(\alpha_1)} P_2^{f(\alpha_2)} \cdots P_N^{f(\alpha_N)}. \qquad (۱۰۷)$$

یک رشته دلخواه از این نوع الزاما نمونه نیست ولی اگر پارامترهای $f(\alpha_i)$ آن در نامساوی (۹۲) 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:

$$\log P(\alpha) = \sum_{i=1}^{N} f_i(\alpha) \log p_i \qquad (۱۰۸)$$

ودرنتیجه ترکیب با (92) بدست می آوریم:

$$\sum_{i=1}^{N} (mp_i - k\sqrt{mp_i(1-p_i)}) \log p_i \le \log P(\alpha \in typical) \le \sum_{i=1}^{N} (mp_i + k\sqrt{mp_i(1-p_i)}) \log p_i. \qquad (۱۰۹)$$

حال قرارمی دهیم

$$A := -k \sum_{i=1}^{N} \sqrt{p_i(1-p_i)} \log p_i. \qquad (۱۱۰)$$

درنتیجه نامساوی قبلی به شکل زیردرمی آید:

$$-mH + A\sqrt{m} \le \log P(\alpha \in typical) \le -mH - A\sqrt{m}, \qquad (۱۱۱)$$

که ازآن نتیجه می گیریم

$$2^{-mH-A\sqrt{m}} \leq P(\alpha \in typical) \leq 2^{-mH+A\sqrt{m}}. \qquad (۱۱۲)$$

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

$$P_{min} \leq P(\alpha \in typical) \leq P_{max}, \qquad (۱۱۳)$$

که در آن

$$P_{min} = 2^{-mH-A\sqrt{m}}, \qquad P_{max} = 2^{-mH+A\sqrt{m}}. \qquad (۱۱۴)$$

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

$$(1-\epsilon)2^{mH-\sqrt{m}A} \leq T \leq 2^{mH+\sqrt{m}A}. \qquad (۱۱۵)$$

- اثبات:

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

$$P_{typical} = P(\alpha_1 \in typical) + P(\alpha_2 \in typical) + \cdots P(\alpha_T \in typical). \qquad (۱۱۶)$$

$P_{typical}$ 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

$$P_{min} = 2^{-mH-A\sqrt{m}} \leq P(\alpha \in typical) \leq 2^{-mH+A\sqrt{m}} = P_{max} \qquad (۱۱۷)$$

بنابراین با ترکیب این دو رابطه بدست می آوریم

$$TP_{min} \leq P_{typical} \leq TP_{max} \qquad (۱۱۸)$$

یا

$$(2^{-mH-A\sqrt{m}})T \leq P_{typical} \leq T(2^{-mH+A\sqrt{m}}). \quad (۱۱۹)$$

اما در قضیه اول ثابت کرده ایم که

$$1-\epsilon \leq P_{typical} \leq 1 \quad (۱۲۰)$$

با ترکیب مناسب نامساوی های اخیر بدست می آوریم:

$$(1-\epsilon)2^{mH-\sqrt{m}A} \leq T \leq 2^{mH+\sqrt{m}A}. \quad (۱۲۱)$$

- 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 (۵۹) 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 (۶۶) نیز انجام دهید.

- تمرین: تمرین قبلی را برای الفبای معرفی شده در رابطه (۶۸) نیز انجام دهید.

0	x_1
010	x_2
01	x_3
10	x_4

جدول ۱: مثالی از یک کد که درآن بعضی ازکد کلمه های پیشوند کد کلمه های دیگرند

۸ ضمیمه

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

که درآن ستون سمت چپ کلمه ها و ستون سمت راست کد کلمه هارا 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. درنتیجه این نوع کد گذاری دارای ابهام زیاد است و کد گذاری خوبی نیست. نخست باید یک صفت اساسی از هرنوع کدگذاری را مشخص کنیم.

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 کلمه‌ای پیشوند کد کلمهٔ 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 درآن C دلخواه است و لزومی ندارد که خود یک کد کلمه باشد. درجدول (8) x_1 پیشوند x_2 و x_3 است. x_3 نیز پیشوند x_2 است.

تعریف : یک کد که درآن 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 درمورد این نوع کد ها آن است که هرکد لحظه ای یکتا گشا ست. البته معکوس این قضیه درست نیست.

0	x_1
100	x_2
101	x_3
11	x_4

جدول ۲: مثالی ازیک کد لحظه ای

0	x_1
01	x_2

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

بازهم به کد نشان داده شده درجدول ?? دقت کنید. هرگاه کد پیامی مثل رشته

$$1\,0\,1\,1\,1\,0\,1\,0\,0\,1\,0\,1 \qquad (۱۲۲)$$

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

$$x_3\ x_4\ x_1\ x_2\ x_3. \qquad (۱۲۳)$$

حال کد زیر را درنظر بگیرید:

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

$$0\,0\,1\,0\,0\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1. \qquad (۱۲۴)$$

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

$$x_1\ x_2\ x_1\ x_1\ x_2\ x_2\ x_2\ x_1\ x_1\ x_1\ x_1\ x_2. \qquad (۱۲۵)$$

درزیر روشی را بیان می کنیم که به کمک آن می توانیم تشخیص بدهیم که آیا یک کد به صورت یکتا گشوده می شود یاخیر.

0	x_1
010	x_2
01	x_3
10	x_4

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

0	x_1
001	x_2

جدول ۵: مثالی از یک کد یکتاگشا

فرض کنیدکه S_0 مجموعه همه کد کلمه ها باشد. مجموعه تمام پسوندهایی را که درS_0 وجود دارد درمجموعه دیگری به نام S_1 قرارمی دهیم.

حال مجموعه $S_2, S_3, \cdots S_n$ 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 قرارمی دهیم.

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

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

مثال :کد زیر یکتا گشانیست.

زیرا:

$$S_0 = \{0, 010, 01, 10\} \qquad S_1 = \{10, 1, 0\} \tag{۱۲۶}$$

و $S_0 \cap S_1 \neq \phi$.

مثال : کد زیریکتا گشاست:

زیرا:

$$S_0 = \{0, 001\} \qquad S_1 = \{01\} \qquad S_2 = \{1\}, \tag{۱۲۷}$$

و $S_0 \cap [S_1 \cup S_2] = \phi$.

مثال : کد زیر را درنظرمی گیریم:

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

(۱۲۸)

برای این کد داریم:

$$\begin{aligned} S_0 &= \{a, c, ad, abb, bad, deb, bbcde\} \\ S_1 &= \{d, bb\} \\ S_2 &= \{eb, cde\} \\ S_3 &= \{de\} \\ S_4 &= \{b\} \\ S_5 &= \{ad, bcde\} \\ S_6 &= \{d\} \\ S_7 &= \{eb\} \end{aligned} \tag{۱۲۹}$$

با توجه به این روابط خواهیم دید که

$$S_0 \cap [S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6 \cup S_7] = \phi, \tag{۱۳۰}$$

و درنتیجه این کد یکتا گشا است.

- قضیه(شرط لازم و کافی برای وجود کد های لحظه ای): مجموعه کلمه های $X = \{x_1, x_2, \cdots x_M\}$ و مجموعه حروف الفبای $A := \{a_1, a_2, \cdots a_D\}$ داده شده اند. مجموعه اعداد صحیح $\{n_1, n_2, \cdots n_M\}$ نیز مفروض اند. آیا یک کد لحظه ای می توان از الفبای A نوشت که طول های $\{n_1, n_2, \cdots n_M\}$ داشته باشند؟ پاسخ این سوال مثبت است اگر وفقط اگر شرط زیر برقرارباشد:

$$\sum_{i=1}^{M} \frac{1}{D^{n_i}} \leq 1. \qquad (۱۳۱)$$

این نامساوی به نامساوی $Kraft$ مشهوراست.

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

نتیجه: برای حروف الفبای انگلیسی بااحتساب نقطه، کاما ، ودیگرعلائم داریم: $M = 32$. هم چنین اگر بخواهیم از الفبای $\sum := \{0, 1\}$ استفاده کنیم داریم $D = 2$. بنابراین بایدداشته باشیم:

$$\sum_{i=1}^{32} \frac{1}{2^{n_i}} \leq 1 \longrightarrow n_{min} \geq 5. \qquad (۱۳۲)$$

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

حال به اثبات قضیه می پردازیم:

- اثبات: می توانیم ازنمودارهای درختی استفاده کنیم. یک درخت با مرتبه D و اندازه k درختی است که D ریشه دارد و ازهرریشه نیز D شاخه منشعب می شود واین کار ادامه می یابد تا $k-1$ مرحله . دراین صورت تعداد شاخه های آخرین مرحله عبارت است از D^k. D ریشه اول درخت متناسب با کد کلمه های تک حرفی $\{1, 2, 3 \cdots D\}$ هستند. شاخه های مرحله بعد متناسب با کد کلمات دو حرفی هستند مثل $\{11, 12, \cdots DD\}$ و همینطور تا آخر. به این ترتیب هر کدکلمه متناسب بایکی 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 هرگره که از این درخت انتخاب می کنیم می بایست تمام شاخه های منشعب از آن گره را کناربگذاریم 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 برابربا i باشد، تعداد شاخه هایی که ازآن منشعب می شود برابراست با D^{k-i}. بنابراین به ازای هر کدکلمه به طول i تعداد D^{k-i} تا از شاخه ها حذف می شوند. درنتیجه خواهیم داشت:

$$D^{k-i_1} + D^{k-i_2} + \cdots D^{k-i_M} \leq D^k \qquad (۱۳۳)$$

که باتقسیم طرفین بر D^k به رابطه (143) می رسیم. این رابطه را می توان به شکل زیر نیز نوشت:

$$\sum_i w_i D^{-i} \leq 1, \qquad (۱۳۴)$$

که درآن w_i تعداد کد کلمه های با طول i است.

حال معکوس قضیه را ثابت می کنیم: تاکنون ثابت کردیم که اگر کد لحظه ای باشد می بایست شرط (143) برقرارباشد. حال نشان می دهیم که به ازای هر $(n_1, n_2, \cdots n_M)$ که درشرط (143) صدق کند می توان یک کد لحظه ای ساخت. n_i را به شکل زیر مرتب می کنیم:

$$n_1 \leq n_2 \leq n_3 \leq \cdots \leq n_M. \qquad (۱۳۵)$$

حال یک نقطه به اندازه n_1 را روی درختِ با مرتبه D واندازه n_M انتخاب می کنیم. به این ترتیب تعداد $D^{n_M - n_1}$ نقطه حذف می شوند. تعداد نقاط باقیمانده برابراست با $D^{n_M} - D^{n_M - n_1}$. نقطه دوم را به طول n_2 انتخاب می کنیم. این نقطه تعداد $D^{n_M - n_2}$ نقطه دیگر را حذف می کند. تعدادنقاط باقیمانده برابراست با $D^{n_M} - D^{n_M - n_1} - D^{n_M - n_2}$. این کاررا ادامه می دهیم تا نقطه ماقبل آخر که طول آن n_{M-1} است. این نقطه نیز تعداد $D^{n_M - n_{M-1}}$ را حذف می کند. 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 برابراست با

$$\begin{aligned} Q &= D^{n_M} - D^{n_M - n_1} - D^{n_M - n_2} - \cdots D^{n_M - n_{M-1}} \\ &= D^{n_M}\left[1 - D^{-n_1} - D^{-n_2} - \cdots D^{-n_{M-1}}\right] \end{aligned} \qquad (۱۳۶)$$

اما چون شرط (143) برقراراست خواهیم داشت:

$$\sum_{i=1}^{M} D^{-n_i} \leq 1 \longrightarrow 1 - \sum_{i=1}^{M-1} D^{-n_i} \leq D^{-n_M} \qquad (۱۳۷)$$

و این رابطه به این معناست که حداقل یک انتخاب برای آخرین کد کلمه باقی می ماند. اثبات قضیه دراین جا کامل می شود.

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

اثبات: الف: اگر نامساوی کرافت برقرارباشد 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 برقراراست. بجای عبارت $\sum_i D^{-n_i}$ عبارت $\sum_{i=1}^{r} w_i D^{-i}$ را بکارمی بریم که درآن w_i تعدادکلمات با طول i است. حال عبارت اخیررامی توان به صورت یک تابع مولد تعبیرکرد. می توان دریافت که

$$\left(\sum_{i=1}^{r} w_i D^{-i}\right)^n = \sum_{k=r}^{nr} X_k D^{-k}, \qquad (۱۳۸)$$

که درآن X_k تعدادکل کد کلمه های با طول k درکدگذاری رشته های r تایی است. اما می دانیم که کد ازنوع یکتا گشودنی است. درضمن تعداد کل کد کلمه های باطول k برابراست با D^k. چون کد های یکتاگشا زیرمجموعه کلیه کد هاهستند نتیجه می گیریم که $X_k \leq D^k$. بنابراین خواهیم داشت:

$$\left(\sum_{i=1}^{r} w_i D^{-i}\right)^n \leq \sum_{k=r}^{nr} 1 = nr - r + 1. \qquad (۱۳۹)$$

و ازآنجا

$$\left(\sum_{i=1}^{r} w_i D^{-i}\right) \leq (1 + (n-1)r)^{\frac{1}{n}}. \qquad (۱۴۰)$$

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

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

$$\overline{n} := \sum_{i=1}^{M} p_i n_i. \tag{۱۴۱}$$

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

$$\overline{n} = \frac{H(X)}{\log D}. \tag{۱۴۲}$$

■ اثبات: نخست توجه می کنیم که کد موردنظرمایکتا گشاست اگر وفقط اگر شرط زیر برقراربباشد:

$$\sum_{i=1}^{M} \frac{1}{D^{n_i}} \leq 1. \tag{۱۴۳}$$

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

$$\overline{n} \geq \frac{H(X)}{logD} \tag{۱۴۴}$$

که درآن شرط تساوی برقرارمی شود اگر وفقط اگر $p_i = D^{-n_i}$.

درمرحله دوم تحقیق می کنیم که چقدر می توانیم به این حد پایین نزدیک شد. و بالاخره درمرحله سوم بهترین کد ممکن را می سازیم.

برای اثبات نامساوی (144) می بایست نامساوی زیررا ثابت کنیم:

$$\sum_{i=1}^{M} n_i p_i \geq -\sum_{i=1}^{M} p_i \frac{\log p_i}{\log D}, \tag{۱۴۵}$$

و یا

$$\sum_{i=1}^{M} (n_i \log D) p_i \geq -\sum_{i=1}^{M} p_i \log p_i. \tag{۱۴۶}$$

قبلا داشتیم که به ازای هردو توزیع احتمال $\{p_i\}$ و $\{q_i\}$، نامساوی زیربرقراراست:

$$\sum_i -p_i \log p_i \leq -\sum p_i \log q_i, \tag{۱۴۷}$$

وتساوی تنهاوقتی برقرارمی شود که $\{q_i\} = \{p_i\}$.

می توانیم یک توزیع احتمال مطابق بارابطه زیرتعریف کنیم:

$$q_i := \frac{D^{-n_i}}{\sum_{i=1}^{M} D^{-n_i}} \tag{۱۴۸}$$

و از رابطه (145) استفاده کنیم. یک محاسبه ساده منجربه رابطه زیرخواهد شد:

$$H(X) \leq \overline{n} \log D + \log(\sum_{i=1}^{M} D^{-n_i}), \tag{۱۴۹}$$

که تساوی وقتی برقرارمی شود که

$$p_i = \frac{D^{-n_i}}{\sum_{i=1}^{M} D^{-n_i}}. \tag{۱۵۰}$$

حال باتوجه به اینکه برای کد های یکتاگشانامساوی 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 ازآنجا بدست می آوریم که

$$H(X) \leq \overline{n} \log D. \tag{۱۵۱}$$

هرگاه بتوانیم یک کد را چنان انتخاب کنیم که طول کد کلمه های آن ازرابطه $n_i = \log_D \frac{1}{p_i}$ تبعیت کند، آنگاه خواهیم داشت : . $\overline{n} = \frac{H(X)}{\log D}$ معکوس این قضیه نیز صحیح است یعنی اینکه اگر رابطه $\overline{n} = \frac{H(X)}{\log D}$ برقرارباشد آنگاه $p_i = D^{-n_i}$. برای اثبات این نتیجه از رابطه 149 استفاده می کنیم و به این نتیجه می رسیم که

$$\overline{n} \log D \leq \overline{n} \log D + \log(\sum_{i=1}^{M} D^{-i}), \tag{۱۵۲}$$

وازآنجا باتوجه به اینکه $\sum_{i=1}^{M} D^{-n_i} \leq 1$، به این نتیجه می رسیم که $\sum_{i=1}^{M} D^{-j} = 1$. اماباتوجه به رابطه 150 این نتیجه به این معناست که $p_i = D^{-n_i}$.

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

یک مثال از یک کد کاملا 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

(۱۵۳)

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

درحالت کلی معلوم نیست که بتوان کد راچنان طراحی کرد که حد $\overline{n} = \frac{H}{\log D}$ برقرارشود، زیرا اعداد $n_i = \log_D \frac{1}{p_i}$ معلوم نیست که صحیح باشند. بااین وجود می توان کاری کرد که شرط زیر برقرارشود:

$$\log_D \frac{1}{p_i} \leq n_i \leq \log_D \frac{1}{p_i} + 1. \quad (۱۵۴)$$

دراین صورت خواهیم داشت :

$$\frac{H(X)}{\log D} \leq \overline{n} \leq \frac{H(X)}{\log D} + 1. \quad (۱۵۵)$$

حال نکته این است که هرقدربخواهیم می توانیم به حدّ 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 راکد نگاری کنیم، یعنی رشته های $Y = (X_1, X_2, \cdots X_s)$ را. حال باید نشان دهیم که تحت این شرایط طول کد کلمه ها به ازای هر X پایین می آید. به رابطه (154) دقت می کنیم. ازآنجا که $Y = (X_1, X_2, \cdots X_s)$، کلمه ها به صورت s تایی های ازنوع $y_{ij} = (x_i, x_j, \cdots x_s)$ هستند. داریم

$$H(Y) = -\sum_{i,j,\cdots} p_{ij\cdots} \log p_{ij\cdots}. \qquad (۱۵۶)$$

چون کلمات پیام Y ازهم مستقل هستند خواهیم داشت: $H(Y) = sH(X)$. وبنابراین

$$\frac{H(Y)}{\log D} \leq \overline{n} \leq \frac{H(Y)}{\log D} + 1, \qquad (۱۵۷)$$

ویا

$$\frac{H(X)}{\log D} \leq \frac{1}{s}\overline{n} \leq \frac{H(X)}{\log D} + \frac{1}{s}. \qquad (۱۵۸)$$

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

۱.۸ ساختن کد های بهینه

حال باید آلگوریتمی را معرفی کنیم که کد های بهینه را به طور روشمند می سازد. نخست به یک لم احتیاج داریم:

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

اثبات: می دانیم که کد های لحظه ای زیرمجموعه کد های یکتا گشا ست. حال فرض کنید که یک کد یکتا گشای C' با طول کدکلمه های $n'_1, n'_2, \cdots n'_M$ وجود دارد که طول متوسط آن از طول متوسط C کمتراست. اولاً چون C' یکتا گشاست بنابرقضیه ای که قبلاً ثابت کردیم خواهیم داشت: $\sum_{i=1}^{M} D^{-n'_i} \leq 1$. اما دراین صورت بنابرقضیه قبل یک کد لحظه ای با طول کلمات $n'_1, n'_2, \cdots n'_M$ وجود خواهد داشت .

بدین ترتیب بهینه بودن کد C دردرون مجموعه کد های لحظه ای نیز نقض می شود.

از این به بعد توجه خود را به کد های 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 $n_1, n_2, \cdots n_M$ برای کد گذاری علامات $x_1, x_2, \cdots x_M$ باشد که این علامات نیز با احتمالات $p_1, p_2, \cdots p_M$ تکرارشوند. دراین صورت اگر کد C درون کد های لحظه ای بهینه باشد آنگاه خاصیت های زیربرقرارند:

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

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

ج: دربین کلماتی که بیشترین طول را دارند، حتماً باید دو کلمه وجود داشته باشند که فقط وفقط دریک رقم بایکدیگر تفاوت داشته باشند.

اثبات الف : فرض کنید که $p_1 \geq p_2$ که درآن p_2, p_1 به ترتیب احتمال ظهور علامات x_2, x_1 باشند. هم چنین فرض کنید که دراین کد لحظه ای C داشته باشیم $n_1 \geq n_2$. دراین صورت می توان یک کد بهتراز C ساخت. جای کد کلمه های مربوط به x_1 و x_2 را عوض می کنیم. کد هنوزلحظه ای است زیراشرط کرافت برقراراست. درکد جدید C' داریم :

$$\overline{n'} - \overline{n} = n_1 p_2 + n_2 p_1 - n_1 p_1 - n_2 p_2 = (n_1 - n_2)(p_2 - p_1) \leq 0. \qquad (۱۵۹)$$

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

$$\begin{aligned} S &= s_1 s_2 \cdots s_{n_{M-1}} \\ \tilde{S} &\equiv S'\tilde{S}' = s'_1 s'_2 \cdots s'_{n_{M-1}} (s'_{n_{M-1}+1} s_{n_{M-1}+2} \cdots s'_{n_M}) \end{aligned} \qquad (۱۶۰)$$

حال می توانیم قسمت اضافی را که درپرانتزقرارداده ایم حذف کنیم بدون اینکه به لحظه ای بودن کد خللی وارد شود. چون اگر کلمه ای پیشوند $S'\tilde{S}'$ نبوده است پیشوند S' نیز نخواهد بود. ضمناً S' نمی تواند پیشوند کدکلمه دیگری باشد،چون کلمات مربوط به x_{M-1} و x_M بزرگترین طول ها را دارند. تنهاامکانی که باقی می ماند آن است که کلمات با طول n_{M-1} بیش از دو تاباشند. دراین صورت تنها را ه برای پیشوند بودن S' آن است که S' دقیقاً بایکی از آن کلمات برابرباشد. ولی این بدان معناست که درکد اولیه که درآن حذفی صورت نگرفته بود، آن کلمه خاص پیشوند $\tilde{S}$ بوده است.

اثبات ج: حال فرض کنید که دوتاازبلندترین کلمات را درنظربگیریم. اگر تنها دررقم آخر اختلاف داشته باشند که این همان چیزی است که مطلوب ماست. اگر بیش از رقم آخر باهم 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.

۲.۸ روش هوفمان برای ساختن کد های لحظه ای بهینه

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

$$(X, P) = \{(x_1, p_1), (x_2, p_2), \cdots (x_M, p_M)\}. \quad (۱۶۱)$$

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

$$(\tilde{X}, \tilde{P}) = \{(x_1, p_1), (x_2, p_2), \cdots (x_{M-2}, p_{M-2}), (x_{M-1,M}, p_{M-1} + p_M)\}. \quad (۱۶۲)$$

سوال: منظور از $x_{M-1,M}$ چیست؟ منظوراین است که درذهن خود تفاوت بین x_{M-1} و x_M را ازبین ببریم. به عبارت دیگر می دانیم که تنها احتمالات مهم هستند و نه خود نمادها. بنابراین مجموعه $\{p_1, p_2, \cdots p_{M-1}, p_M\}$ را به مجموعه $\{p_1, p_2, \cdots p_{M-1} + p_M\}$ تقلیل داده ایم. حال فرض کنید که کد بهینه ای برای $(\tilde{X}, \tilde{P})$ دردست 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}
$n_{M-1,M}$	$w_{M-1,M}$	$p_{M-1}+p_M$	$x_{M-1,M}$

(۱۶۳)

حال کد C را برای (X,P) به شکل زیرمی 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

(۱۶۴)

حال ثابت می کنیم که اگر $\tilde{C}$ بهینه باشد آنگاه C نیز بهینه است. ازبرهان خلف استفاده می کنیم. فرض کنید که کدی مثل C' وجود داشته باشد که از کد C بهتر باشد. دراین صورت بااستفاده از کد C' می توان کدی مثل $\tilde{C}'$ ساخت که از $\tilde{C}$ بهترباشد. کد C' درجدول زیرنشان 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

(۱۶۵)

دراین کد $n'_m = n'_{M-1}$. w'_{M-1} و w'_M نیز تنها دررقم آخرباهم اختلاف دارند. حال کد $\tilde{C}'$ را مطابق جدول زیرمی سازیم:

$\tilde{N}'$	$\tilde{C}'$	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}	$\tilde{w'}_{M-1,M}$	$p_{M-1}+p_M$	$x_{M-1,M}$

(۱۶۶)

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

$$\overline{n}-\overline{\tilde{n}}=(p_{M_1}+p_M)(n_{M-1,M}+1-n_{M-1,M})=p_{M-1}+p_M, \tag{۱۶۷}$$

و

$$\overline{n'}-\overline{\tilde{n'}}=(p_{M_1}+p_M)(n'_M-n'_M-1)=p_{M-1}+p_M. \tag{۱۶۸}$$

درنتیجه

$$\overline{n}-\overline{\tilde{n}}=\overline{n'}-\overline{\tilde{n'}} \tag{۱۶۹}$$

که ازآن خواهیم داشت:

$$if\ \ \overline{n'}<\overline{n}\longrightarrow\overline{\tilde{n'}}<\overline{\tilde{n}}. \tag{۱۷۰}$$

بنابراین اگر کد C' از کد C بهتر باشد کد $\tilde{C}$ نیز از کد $\tilde{C}$ بهتراست 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

(۱۷۱)

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

(۱۷۲)

وازآنجا

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

(۱۷۳)

هرگاه تعدادکلمات بیشترباشد این کاررادرچندمرحله انجام می دهیم . درهرمرحله احتمالات راازبیشترین به کمترین مرتب می کنیم و آخرین دو کلمه را باهم مطابق باآنچه که دربالا گفته شدادغام می کنیم. این کارراآنقدرانجام می دهیم تا به یک مجموعه برسیم متشکل از دونماد و دواحتمال. به دو نماد آخر کلمه های 0 و 1 را نسبت می دهیم و سپس مراحل را درجهت عکس طی می کنیم تا به جدول اولیه برسیم وکد های تمام نماد ها را بدست آوریم.