نظریه اطلاعات کلاسیک

۱ مقدمه

در درسهای گذشته با مبانی نظری آلگوریتم های کوانتومی آشنا شدیم. اما آلگوریتم های کوانتومی[1] و محاسبه کوانتومی [2] تمامی آنچیزی نیستند که در این حوزه جدید از دانش وجود دارد. اگر محاسبه کوانتومی و آلگوریتم کوانتومی را پردازش داده‌ها و اطلاعات کوانتومی به طرز موثر[3] تعریف کنیم آنگاه سوال اساسی که پیش روی ماست آن است که اساساً اطلاعات کوانتومی چیست؟ چگونه‌می توان آن را اندازه گرفت؟ اطلاعات موجود در یک حالتِ کوانتومی چقدراست؟ تا چه حد می توان اطلاعات کوانتومی موجود در یک منبع را فشرده کرد بدون اینکه به محتوای آن صدمه وارد کرد؟ چگونه نوفه‌ی یک کانال کوانتومی موجب از بین رفتن اطلاعات کوانتومی فرستاده شده می شود؟ چقدر این اطلاعات از بین می رود؟ چگونه می توان فهمید که اطلاعات کوانتومی دریافتی آیا نزدیک به اطلاعات ارسال شده هست یا نه؟ چگونه می توان به خطاهای ایجاد شده روی اطلاعات کوانتومی پی برد و آنها را آشکار و تصحیح کرد؟ چگونه می توان ظرفیت یک کانال کوانتومی را تعریف کرد؟ این سوال ها و سوال های بسیارِ دیگر و تلاش برای پاسخ گویی به آنها موضوع اطلاعات کوانتومی را تشکیل می دهد. بقیه این درس نامه به این موضوع خواهد پرداخت. اما هرنوع مطالعه‌ای در باره اطلاعات کوانتومی نیازمند آموختن تعاریف و اصول اساسی در باره اطلاعات کلاسیک است. دراین درس و درس بعدی با مفاهیم بنیادی اطلاعات کلاسیک آشنا می شویم. اگرچه مشابهت های جدی بین نظریه اطلاعات کلاسیک و کوانتومی وجود دارد، ولی بدیهی است که تفاوت های مهم و اساسی بین این دو نظریه نیز وجود دارد. بعد از آشنایی مقدماتی با نظریه اطلاعات کلاسیک به مطالعه نظریه اطلاعات کوانتومی خواهیم پرداخت.

---

[1] Quantum Algorithm
[2] Quantum Computation
[3] Efficient

۲ تعاریف اساسی

فرض کنیدکه $X = \{x_1, x_2, \cdots x_n\}$ یک متغیرتصادفی بااحتمالات $\{p_1, p_2, \cdots p_n\}$ باشد. به این متغیر تصادفی می توان تابعی به شکل زیرنسبت داد.

$$H(X) := \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}. \tag{1}$$

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.

۳ مفهوم واندازه اطلاعات

۱.۳ 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 وکسی نتیجه این واقعه را به مابگوید، مثلاً بگوید که پیشامد x_i رخ داده است. دراین صورت می توان پرسید که آن شخص چه مقدار به مااطلاع داده است وچه مقدارازبی اطلاعی ما کاسته شده است. 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 درنتیجه 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 $\{(x_i, y_j), \quad i = 1 \cdots m, \ j = 1 \cdots n\}$ 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

$$h(p_i q_j) = h(p_i) + h(q_j). \tag{2}$$

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

$$h(p_i) = \log_\alpha \frac{1}{p_i}, \tag{3}$$

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:

$$H(X) = \frac{1}{N}\sum_{i=1}^{n} Np_i \log_2 \frac{1}{p_i} = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}. \tag{4}$$

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.

۲.۳ 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:

$$H(X,Y) := -\sum_{i,j} p(x_i,y_j)\log_2 p(x_i,y_j) \tag{5}$$

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

این تعریف به همین شکل به بیش از دومتغیرتصادفی تعمیم می یابد.

۳.۳ 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_j. دراین صورت توزیع متغیرتصادفی X عوض خواهد شد

وتبدیل خواهد شد به توزیع $P(X|y_j)$ که درآن y_j یک پارامتراست و X مقادیرمتغیر را بخود می گیرد. می دانیم که:

$$P(x_i|y_j) := \frac{P(x_i, y_j)}{p(y_j)}, \qquad \sum_i p(x_i|y_j) = 1. \tag{6}$$

درنتیجه اطلاعات باقیمانده در متغیرتصادفی X برابرخواهد بود با:

$$H(X|y_j) := -\sum_{x_i} P(x_i|y_j) \log_2 P(x_i|y_j) \tag{7}$$

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:

$$\begin{aligned} H(X|Y) &= \sum_{y_j} p(y_j) H(X|y_j) = -\sum_{x_i, y_j} P(y_j) P(x_i|y_j) \log_2 P(x_i|y_j) \\ &= -\sum_{x_i, y_j} P(x_i, y_j) \log_2 P(x_i|y_j) = -\sum_{x_i, y_j} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(y_j)} \\ &= H(X, Y) - H(Y). \end{aligned} \tag{8}$$

دقت کنید که به همان دلیلی که تابع $H(X)$ مثبت است تابع $H(X|y_j)$ 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 (8) به نتیجه زیر می رسیم:

$$H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X). \tag{9}$$

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 (9)، $H(X, Y) = H(X) + H(Y)$.
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 (9) به این معناست که $H(X, Y) = H(Y)$. این رابطه نیز معنای شهودی روشنی دارد.

۴.۳ اطلاعات متقابل

اطلاعات متقابل دردومتغیرتصادفی X و Y به شکل زیرتعریف می شود:

$$I(X : Y) := H(X) + H(Y) - H(X, Y). \tag{10}$$

این کمیت نسبت به دومتغیرتصادفی X و Y متقارن است. باتوجه به رابطه (9) می توان آن را به شکل زیربازنویسی کرد:

$$I(X:Y) := H(X) - H(X|Y). \tag{11}$$

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.

۴ یک تذکر مهم درباره اطلاعات منبع

تاکنون اطلاعات منبع X را ازروی فرکانس حروف الفبایی که درآن بکاررفته بود حساب کردیم. بنابراین اگر منبع X را یک متن ادبیات انگلیسی درنظر بگیریم که الفبای آن را بااحتساب نقطه، علامت سوال، فاصله و نظایرآن متشکل از ۳۰ حرف حساب کنیم ازروی فرکانس این سی حرف تابع $H(X)$ را می توانیم محاسبه کنیم. درزبان انگلیسی حرف e بالاترین فرکانس را دارد، پس ازآن حرف t ودیگرحروف قرارگرفته اند. اما بدلیل ساختارمتن ها دوحرف e پشت سرهم یادوحرف t پشت سرهم بندرت قرارمی گیرند. بنابراین اگراطلاعات رابراساس رشته های یک حرفی ودوحرفی هردودرنظربگیریم نتیجه متفاوتی بدست می آوریم. درنتیجه راه بهترومطمئن تربرای محاسبه اطلاعات منبع آن است که اطلاعات را نه براساس تک تک حروف الفبابلکه براساس هجاها، کلمات وحتی جملات حساب کنیم. درنتیجه ازاین به بعد وقتی از منبع X والفبای A صحبت می کنیم منظورما یک الفبای تعمیم یافته است که عناصرآن نه تک تک حروف بلکه هجاهاوکلمات نیزهستند. این پیچیدگی ها درعمل همواره مورد ملاحظه قرارمی گیرند.

۵ 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: فرض کنید که $\{p_i\}$ و $\{q_i\}$ دو توزیع احتمال باشند. دراین صورت

$$\sum_i p_i \log \frac{1}{p_i} \leq \sum_i p_i \log \frac{1}{q_i}, \tag{12}$$

که درآن تساوی فقط وقتی برقرارمی شود که دو توزیع احتمال یکی باشند. نامساوی بالا را می توان به شکل گویاترزیرنوشت:

$$\langle \log \frac{1}{p} \rangle_p \leq \langle \log \frac{1}{q} \rangle_p. \tag{13}$$

زیرنویس p بیان می کند که هردومتوسط با توزیع احتمال p محاسبه شده اند.

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

$$\log x \leq x - 1, \tag{14}$$

که درآن تساوی فقط برای $x = 1$ برقرارمی شود.
حال قرارمی دهیم $x = \frac{q_i}{p_i}$ و درنتیجه

$$\log \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1, \ \ \forall \ \ i, \tag{15}$$

که درآن تساوی فقط وقتی برقرارمی شود که $q_i = p_i$. درنتیجه

$$\sum_i p_i \log \frac{q_i}{p_i} \leq \sum_i q_i - \sum_i p_i = 0, \tag{16}$$

که همان نامساوی ای است که می خواستیم ثابت کنیم. حال فرض کنید که

$$\sum_i p_i \log \frac{q_i}{p_i} = 0. \tag{17}$$

این تساوی را به شکل زیربازنویسی می کنیم

$$\sum_i p_i (\log \frac{q_i}{p_i} - (\frac{q_i}{p_i} - 1)) = 0. \tag{18}$$

حال دقت می کنیم که بنابرنامساوی (15) 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 (14) به معنای آن است که برای همه i ها $q_i = p_i$. یعنی دوتابع توزیع احتمال یکی هستند.

نتیجه ۱: مقدار بیشینه تابع اطلاعات $H = \sum_{i=1}^{M} p_i \log \frac{1}{p_i}$ برابراست با $\log M$ واین مقداربیشینه را فقط برای توزیع یکنواخت $\{p_i = \frac{1}{M}\}$ اختیارمی کند.

اثبات: درقضیه قبلی قرارمی دهیم $q_i = \frac{1}{M}$. درنتیجه خواهیم داشت :

$$\sum_{i=1}^{M} p_i \log \frac{\frac{1}{M}}{p_i} = H - \log M \leq 0, \tag{19}$$

که درآن تساوی فقط وقتی برقرارمی شود که $p_i = \frac{1}{M}$.

نتیجه ۲: برای دو متغیرتصادفی X, Y نامساوی زیربرقراراست

$$H(X,Y) \leq H(X) + H(Y), \tag{20}$$

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

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

$$p_1(x) := \sum_y p(x,y), \qquad p_2(y) := \sum_x p(x,y). \tag{21}$$

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

$$\sum_{x,y} p(x,y) \log \frac{q(x,y)}{p(x,y)} \leq 0 \tag{22}$$

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

$$H(X,Y) \leq H(X) + H(Y). \tag{23}$$

این عبارت بیان می کند که به عنوان مثال 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 هردوجمله را به ما می گوید دوبرابرکسی که فقط یکی ازجملات را به ما می گوید به ما اطلاع نمی دهد.

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

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

$$H_0(X) \geq \lambda H_1(X) + (1-\lambda)H_2(X). \tag{24}$$

به اصطلاح می گوییم که اطلاعات یک تابع محدب روبه پایین است که به یادماندن شکل آن را نیزدرذهن آسان می کند.

اثبات: بازهم ازنامساوی اساسی ای که ثابت کردیم استفاده می کنیم. باکمی خلاصه نویسی درنمادها خواهیم داشت:

$$\begin{aligned}
& H_0 - \lambda H_1 - (1-\lambda)H_2 \\
= \quad & \sum p_0 \log \frac{1}{p_0} - \lambda \sum p_1 \log \frac{1}{p_1} - (1-\lambda) \sum p_2 \log \frac{1}{p_2} \\
= \quad & \sum (\lambda p_1 + (1-\lambda)p_2) \log \frac{1}{\lambda p_1 + (1-\lambda)p_2} - \lambda \sum p_1 \log \frac{1}{p_1} - (1-\lambda) \sum p_2 \log \frac{1}{p_2} \\
= \quad & \lambda \sum p_1 \log \frac{p_1}{\lambda p_1 + (1-\lambda)p_2} + (1-\lambda) \sum p_2 \log \frac{p_2}{\lambda p_1 + (1-\lambda)p_2} \geq 0,
\end{aligned} \tag{25}$$

که درخط آخر ازنامساوی اساسی استفاده کرده ایم.

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:

$$\begin{aligned}
P(x_i|y_j) &= \frac{P(x_i, y_j)}{P(y_j)} = \frac{P(y_j, x_i)}{\sum_{x_i} P(y_j, x_i)} \\
&= \frac{P(y_j|x_i)P(x_i)}{\sum_{x_i} P(y_j|x_i)P(x_i))}
\end{aligned} \tag{26}$$

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

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

اثبات: 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:

$$\begin{aligned} P(y) &= \sum_x P(y|x)P(x), \\ P(x,y) &= P(y|x)P(x), \end{aligned} \tag{27}$$

خواهیم داشت

$$\begin{aligned} P_0(x,y) &= \lambda P_1(x,y) + (1-\lambda)P_2(x,y) \\ P_0(y) &= \lambda P_1(y) + (1-\lambda)P_2(y). \end{aligned} \tag{28}$$

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

مثال: جفت متغیرتصادفی (X,Y) را مطابق جدول زیر درنظربگیرید: Y ناشی ازانداختن یک طاس است که مقادیر ۱ تا ٦ 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

(29)

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

٦ 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 .

$$\begin{aligned} A &\longrightarrow 00 \\ B &\longrightarrow 01 \\ C &\longrightarrow 10 \\ D &\longrightarrow 11. \end{aligned} \tag{30}$$

دراین صورت به ازای هرحرف دوبیت مخابره کرده ایم. حال سوال این است که آیا می توانیم یک روش کد کردن به کارببریم 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:

$$P(A) = \frac{1}{8} \qquad P(B) = \frac{1}{8} \qquad P(C) = \frac{1}{4} \qquad P(D) = \frac{1}{2}. \tag{31}$$

حال روش کدگذاری زیررا به کارمی بریم:

$$\begin{aligned} D &\longrightarrow 0 \\ C &\longrightarrow 10 \\ B &\longrightarrow 110 \\ A &\longrightarrow 111. \end{aligned} \tag{32}$$

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:

$$\langle l \rangle = \sum_{i=1}^{4} l_i \times p_i = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = \frac{7}{4}. \tag{33}$$

بنابراین با یک کدگذاری مناسب توانسته ایم طول متوسط رشته بیت هایی را که برای مخابره پیام بکاربرده ایم از ۲ به ۷/۴ 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

$$0\,1\,0\,0\,0\,1\,0\,0\,0\,1\,1\,0\,1\,1\,1. \tag{34}$$

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

$$D\ C\ D\ D\ C\ D\ D\ B\ A. \tag{35}$$

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

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

$$= \frac{1}{2}\times 1+\frac{1}{4}\times 2+\frac{1}{8}\times 3+\frac{1}{8}\times 3=\frac{7}{4}. \tag{36}$$

بنابراین دراین مثال خاص طول متوسط کدگزاری ای که به کاربردیم با میزان اطلاعات موجود درمتن برابراست. آیا این یک خصلت عمومی است؟

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 زیربکارببریم.

به جای اینکه تک تک حروف الفبا را کدگذاری کنیم، رشته های طولانی به طول m حرف رادرنظرمی گیریم وبرای هررشته یک کد انتخاب می کنیم. تعداد کل رشته های m حرفی ممکن که ازالفبای A تشکیل شده اند برابراست با N^m. ولی نکته دراین است 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. درهررشته 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 وجود دارد. تعداداین رشته های متعارف تقریباً برابراست با تعداد طرقی که می توان از یک رشته m بیتی mp_1 تا را به صورت x_1 انتخاب کرد و mp_2 تا را به صورت x_2 والی آخر. این تعداد برابراست با:

$$K=\frac{m!}{(mp_1)!(mp_2)!\cdots(mp_N)!} \tag{37}$$

اما بااستفاده از تقریب استرلینگ می توانیم بنویسیم:

$$\log_2 K=\log_2\left(\frac{m!}{(mp_1)!(mp_2)!\cdots(mp_n)!}\right)\approx m\left(\sum_{i=1}^{N}p_i\log_2\frac{1}{p_i}\right)\equiv mH(X) \tag{38}$$

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

$$H(X):=\sum_{i=1}^{N}p_i\log_2\left(\frac{1}{p_i}\right) \tag{39}$$

بنابراین تعداد جملات متعارف باتقریب بسیارخوب برابرخواهد بود با

$$K \approx 2^{mH(X)} \tag{40}$$

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.

۷ تعریف دقیق از رشته های متعارف

تعریف: 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 کوچک باشد. بنابراین به ازای هر ϵ عدد k را چنان انتخاب می کنیم که $\frac{1}{k^2} < \frac{\epsilon}{N}$. دراین صورت رشته α رشته متعارف خوانده می شود اگر شرط زیر برقرارباشد:

$$\left| \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}} \right| < k \quad \forall \ i = 1, 2, \cdots N. \tag{41}$$

تعداد رشته های متعارف برای m های بزرگ درحدود 2^{mH} است که هرکدام با احتمال 2^{-mH} دراین مجموعه پدیدارمی شوند. به عبارت دقیق تر قضیه زیربرقراراست.

قضیه:

الف: احتمال کل رشته های غیرمتعارف از ϵ کمتراست.

ب: یک عدد A وجود دارد به قسمی که به ازای هر رشته متعارف α

$$2^{-mH-A\sqrt{m}} < P(\alpha) < 2^{-mH+A\sqrt{m}}. \tag{42}$$

پ: تعداد رشته های متعارف به طول m برابر است با $2^{m(H+\delta_m)}$ که درآن $\lim_{m\longrightarrow\infty}\delta_m = 0$.

اثبات: نخست به یک لم احتیاج داریم.

لم: نامساوی 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،

$$P(X \geq \alpha) \leq \frac{\overline{X}}{\alpha} \tag{43}$$

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

اثبات:

$$P(X \geq \alpha) = \sum_{x=\alpha}^{\infty} P(x) \leq \sum_{x=\alpha}^{\infty} \frac{x}{\alpha} P(x) \leq \frac{\overline{X}}{\alpha}. \tag{44}$$

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

$$P((X - \overline{X})^2 \geq k^2\sigma_x^2) \leq \frac{1}{k^2}. \tag{45}$$

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:

$$P(T \geq \alpha) \leq \frac{\overline{T}}{\alpha}. \tag{46}$$

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

$$P((X - \overline{X})^2 \geq k^2\sigma_x^2) \leq \frac{\sigma_x^2}{k^2\sigma_x^2} = \frac{1}{k^2}. \tag{47}$$

این نامساوی را به شکل زیر نیز می توان نوشت:

$$P(|X - \overline{X}| \geq k\sigma_x) \leq \frac{1}{k^2}, \tag{48}$$

پس از این لم به اثبات قضیه اولیه می پردازیم:

اثبات الف: احتمال اینکه یک رشته α متعارف نباشد را با P_0 نشان می دهیم. بنابرتعریف داریم:

$$P_0 = Prob\{| \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}} |\geq k, \ \ i \text{ برای حداقل یک } \} = \sum_{i=1}^{N} P(| \frac{f_i(\alpha) - mp_i}{\sqrt{mp_i(1-p_i)}} |\geq k). \tag{49}$$

بااستفاده از نامساوی چبیشف نتیجه می گیریم که

$$P_0 \leq \sum_{i=1}^{N} \frac{1}{k^2} = \frac{N}{k^2} \leq \epsilon. \tag{50}$$

بنابراین قسمت الف قضیه ثابت می شود.

اثبات ب: فرض کنید که $\alpha = (\alpha_1, \alpha_2, \cdots \alpha_m)$ یک رشته متعارف باشد. دراین صورت با استفاده ازتعریف داریم

$$mp_i - k\sqrt{mp_i(1-p_i)} \leq f_i(\alpha) \leq mp_i + k\sqrt{mp_i(1-p_i)}, \quad \forall\, i. \tag{51}$$

حال باتوجه به عدم همبستگی بین حروف، احتمال وقوع رشته α برابراست با:

$$P(\alpha) = {p_1}^{f_1(\alpha)} {p_2}^{f_2(\alpha)} \cdots {p_N}^{f_N(\alpha)}. \tag{52}$$

ازاینجا نتیجه می گیریم

$$\log P(\alpha) = \sum_{i=1}^{N} f_i(\alpha) \log p_i \tag{53}$$

ودرنتیجه ترکیب با (51) بدست می آوریم:

$$\sum_{i=1}^{N} (mp_i - k\sqrt{mp_i(1-p_i)}) \log p_i \leq \log P(\alpha) \leq \sum_{i=1}^{N} (mp_i + k\sqrt{mp_i(1-p_i)}) \log p_i. \tag{54}$$

حال قرارمی دهیم

$$A := -k \sum_{i=1}^{N} \sqrt{p_i(1-p_i)} \log p_i. \tag{55}$$

درنتیجه نامساوی قبلی به شکل زیردرمی آید:

$$-mH + A\sqrt{m} \leq \log P(\alpha) \leq -mH - A\sqrt{m}, \tag{56}$$

که ازآن نتیجه می گیریم

$$2^{-mH-A\sqrt{m}} \leq P(\alpha) \leq 2^{-mH+A\sqrt{m}}. \tag{57}$$

اثبات پ: قبلاً نشان دادیم که احتمال اینکه یک رشته متعلق به زیرمجموعه رشته های متعارف باشد از $1-\epsilon$ بیشتراست. ازطرفی نشان دادیم که احتمال اینکه یک رشته خاص α، رشته متعارف باشد از $2^{-mH-A\sqrt{m}}$ بیشتراست. پس اگر مجموعه رشته های متعارف را به صورت زیردرنظربگیریم:

$$S := \{\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, \cdots \alpha^{(T)}\}, \tag{58}$$

که درآن T تعداد رشته های متعارف است، آنگاه داریم:

$$1-\epsilon \leq P(\alpha^{(1)}) + P(\alpha^{(2)}) + \cdots P(\alpha^{(T)}) \leq T(2^{-mH+A\sqrt{m}}), \tag{59}$$

که ازآن نتیجه می گیریم

$$2^{mH-A\sqrt{m}+\log(1-\epsilon)} \leq T, \tag{60}$$

ویا

$$2^{m(H+\delta_m)} \leq T, \tag{61}$$

که درآن $\lim_{m\longrightarrow\infty} \delta_m = 0$.
ازطرفی می دانیم که

$$T2^{-mH-A\sqrt{m}} \leq P(\alpha^{(1)}) + P(\alpha^{(2)}) + \cdots P(\alpha^{(T)}) \leq 1, \tag{62}$$

که ازآن نتیجه می گیریم

$$T \leq 2^{m(H+\frac{A}{\sqrt{m}})}. \tag{63}$$

ازترکیب (63 ،61) نتیجه می گیریم که

$$T = 2^{m(H+\delta_m)} \tag{64}$$

که درآن $\lim_{m\longrightarrow\infty} \delta_m = 0$. این همان نتیجه ای است که می خواستیم یعنی تعداد رشته های متعارف برابراست با 2^{mH}.

۸ 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 وازمابخواهد 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. به عنوان مثال منبع

$$X = \{000(1/2), 111(1/2)\} \tag{65}$$

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

۱ — 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:

$$X = \{000(1/4), 111(1/4), 001(1/4), 110(1/4)\} \tag{66}$$

حال می توانیم سوالات خود را به شکل زیرتنظیم کنیم:

۱ — آیا اکثریت بیت ها صفرهستند؟

۲ — 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 فرستنده ای پیام های منبع X را ازدرون یک کانال ارسال می کند که دراثرنوفه درون کانال این پیام هابه صورت Y توسط گیرنده دریافت می شوند. قبل از دریافت هرنوع پیامی اطلاعات منبع توسط آنتروپی $H(X)$ مشخص می شده است که نمایانگرحداقل سوالهای آری یانه ای بوده است که باپرسیدن آنهامی توانستیم به یک پیام خاص پی ببریم. وقتی که پیام های Y دریافت می شود آنتروپی منبع توسط تابع $H(X|Y)$ مشخص می شود که نشان دهنده تعداد جدید حداقل سوال

هابرای پی بردن به پیام اصلی است. هرگاه $H(X|Y) = 0$ باشد به معنای آن است به دقت می توانیم پیام منبع را تشخیص دهیم ودرنتیجه اطلاعات متقابل Y و X زیاداست وهرگاه $H(X|Y) = H(X)$ باشد یعنی دانستن Y موجب هیچ کاهشی درتعداد سوالات لازم نشده است. بنابراین تفاضل $H(X) - H(X|Y)$ یعنی تفاضل تعداد سوالات لازم برای شناسایی پیام های X ملاک خیلی خوبی از میزان اطلاعات متقابل X و Y است.

۹ ضمیمه: نحوه‌ی عملی ساختن کد کلمه ها برای یک منبع

دراین ضمیمه که خواندن آن برای فهم درس های آینده ضروری نیست نحوه عملی ساختن کدهای یکتا گشا و فشرده را برای یک منبع توضیح می دهیم.

منبعی را درنظرمی گیریم که پیام های یک آنزامبل که توسط متغیر تصادفی X مشخص می شود را مخابره می کند. داریم

$$X = \{x_1, x_2, \cdots x_n\}. \tag{67}$$

دراین آنزامبل هر متغیر x_k با احتمال $P(x_k)$ اختیار می شود. ازاین به بعد x_k هاراکلمات می نامیم اگر چه ممکن است ازتک حرف ها هجاها ویاجملات نیزتشکیل شده باشند.
هرکلمه x_k را با یک کد متشکل از علائم 0 و 1 کدگذاری می کنیم. کدی که برای یک کلمه x به کارمی بریم را با $w(x)$ نشان می دهیم و آن را یک کد کلمه یا $(Code\ word)$ می خوانیم.
رشته ای متشکل از کلمات را یک پیام می خوانیم و رشته ای متشکل از کدکلمه هارایک کدپیام یا $(Codemessage)$ می خوانیم.
طول یک کدکلمه $w(x)$ را با $l(x)$ نشان می دهیم. هدف ما کمینه کردن طول متوسط کدکلمه ها یعنی عبارت زیراست:

$$\bar{l} := \sum_{i=1}^{n} P(x_i) l(x_i). \tag{68}$$

۱۰ یکتایی کد گشایی

اولین مسئله ای که باآن مواجه هستیم یکتایی کد گشایی است. برای مثال به جدول زیر 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.

x_1	0
x_2	010
x_3	01
x_4	10

مثالی از یک کد که درآن بعضی ازکد کلمه های پیشوند کد کلمه های دیگرند :1

x_1	0
x_2	100
x_3	101
x_4	11

مثالی ازیک کد لحظه ای :2

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 (۱۰) 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.
بازهم به کد نشان داده شده درجدول ?? دقت کنید. هرگاه کد پیامی مثل رشته

$$1\,0\,1\,1\,1\,0\,1\,0\,0\,1\,0\,1 \tag{69}$$

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

$$x_3\ x_4\ x_1\ x_2\ x_3. \tag{70}$$

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:

$$0\,0\,1\,0\,0\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1. \tag{71}$$

x_1	0
x_2	01

یک کد که به طور یکتا گشوده می شود ولی لحظه ای نیست. :3

x_1	0
x_2	010
x_3	01
x_4	10

مثالی از یک کد که به طوریکتا گشوده نمی شود. :4

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

$$x_1\ x_2\ x_1\ x_1\ x_2\ x_2\ x_2\ x_1\ x_1\ x_1\ x_1\ x_2. \tag{72}$$

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.

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:

$$S_0 = \{0, 010, 01, 10\} \qquad S_1 = \{10, 1, 0\} \tag{73}$$

و $S_0 \cap S_1 \neq \phi$.

مثال: کد زیریکتا گشاست:

x_1	0
x_2	001

مثالی از یک کد یکتاگشا :5

زیرا:

$$S_0 = \{0, 001\} \qquad S_1 = \{01\} \qquad S_2 = \{1\}, \tag{74}$$

و $S_0 \cap [S_1 \cup S_2] = \phi$.

مثال: کد زیر را درنظرمی گیریم:

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

(75)

برای این کد داریم:

$$\begin{aligned}
S_0 &= \{a, c, ad, abb, bad, deb, bbcde\} \\
S_1 &= \{d, bb\} \\
S_2 &= \{eb, cde\} \\
S_3 &= \{de\} \\
S_4 &= \{b\} \\
S_5 &= \{ad, bcde\} \\
S_6 &= \{d\} \\
S_7 &= \{eb\}
\end{aligned} \tag{76}$$

با توجه به این روابط خواهیم دید که

$$S_0 \cap [S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6 \cup S_7] = \phi, \tag{77}$$

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:

$$\sum_{i=1}^{M} \frac{1}{D^{n_i}} \leq 1. \tag{78}$$

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:

$$\sum_{i=1}^{32} \frac{1}{2^{n_i}} \leq 1 \longrightarrow n_{min} \geq 5. \tag{79}$$

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 کارادامه 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 همینطورتا 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:

$$D^{k-i_1} + D^{k-i_2} + \cdots D^{k-i_M} \leq D^k \tag{80}$$

که باتقسیم طرفین بر D^k به رابطه (90) می رسیم. این رابطه را می توان به شکل زیر نیز نوشت:

$$\sum_i w_i D^{-i} \leq 1, \tag{81}$$

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 (90) برقرارباشد. حال نشان می دهیم که به ازای هر $(n_1, n_2, \cdots n_M)$ که درشرط (90) صدق کند می توان یک کد لحظه ای ساخت. n_i را به شکل زیر مرتب می کنیم:

$$n_1 \leq n_2 \leq n_3 \leq \cdots \leq n_M. \tag{82}$$

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

$$\begin{aligned} Q &= D^{n_M} - D^{n_M-n_1} - D^{n_M-n_2} - \cdots D^{n_M-n_{M-1}} \\ &= D^{n_M}\left[1 - D^{-n_1} - D^{-n_2} - \cdots D^{-n_{M-1}}\right] \end{aligned} \tag{83}$$

اما چون شرط (90) برقراراست خواهیم داشت:

$$\sum_{i=1}^{M} D^{-n_i} \leq 1 \longrightarrow 1 - \sum_{i=1}^{M-1} D^{-n_i} \leq D^{-n_M} \tag{84}$$

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.

اثبات 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

$$\left(\sum_{i=1}^{r} w_i D^{-i}\right)^n = \sum_{k=r}^{nr} X_k D^{-k}, \tag{85}$$

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:

$$\left(\sum_{i=1}^{r} w_i D^{-i}\right)^n \leq \sum_{k=r}^{nr} 1 = nr - r + 1. \tag{86}$$

و ازآنجا

$$\left(\sum_{i=1}^{r} w_i D^{-i}\right) \leq (1 + (n-1)r)^{\frac{1}{n}}. \tag{87}$$

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:

$$\overline{n} := \sum_{i=1}^{M} p_i n_i. \tag{88}$$

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

$$\overline{n} = \frac{H(X)}{\log D}. \tag{89}$$

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

$$\sum_{i=1}^{M} \frac{1}{D^{n_i}} \leq 1. \tag{90}$$

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

$$\overline{n} \geq \frac{H(X)}{logD} \tag{91}$$

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 (91) می بایست نامساوی زیررا ثابت کنیم:

$$\sum_{i=1}^{M} n_i p_i \geq -\sum_{i=1}^{M} p_i \frac{\log p_i}{\log D}, \tag{92}$$

و یا

$$\sum_{i=1}^{M} (n_i \log D) p_i \geq -\sum_{i=1}^{M} p_i \log p_i. \tag{93}$$

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

$$\sum_i -p_i \log p_i \leq -\sum p_i \log q_i, \tag{94}$$

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

می توانیم یک توزیع احتمال مطابق بارابطه زیرتعریف کنیم:

$$q_i := \frac{D^{-n_i}}{\sum_{i=1}^{M} D^{-n_i}} \tag{95}$$

و از رابطه (92) استفاده کنیم. یک محاسبه ساده منجربه رابطه زیرخواهد شد:

$$H(X) \leq \overline{n} \log D + \log(\sum_{i=1}^{M} D^{-n_i}), \tag{96}$$

که تساوی وقتی برقرارمی شود که

$$p_i = \frac{D^{-n_i}}{\sum_{i=1}^{M} D^{-n_i}}. \tag{97}$$

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

$$H(X) \leq \overline{n} \log D. \tag{98}$$

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 96 استفاده می کنیم و به این نتیجه می رسیم که

$$\overline{n} \log D \leq \overline{n} \log D + \log(\sum_{i=1}^{M} D^{-i}), \tag{99}$$

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 97 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:

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

(100)

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:

$$\log_D \frac{1}{p_i} \leq n_i \leq \log_D \frac{1}{p_i} + 1. \tag{101}$$

دراین صورت خواهیم داشت :

$$\frac{H(X)}{\log D} \leq \overline{n} \leq \frac{H(X)}{\log D} + 1. \tag{102}$$

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 (101) 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

$$H(Y) = -\sum_{i,j,\cdots} p_{ij\cdots} \log p_{ij\cdots}. \tag{103}$$

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

$$\frac{H(Y)}{\log D} \leq \overline{n} \leq \frac{H(Y)}{\log D} + 1, \tag{104}$$

ویا

$$\frac{H(X)}{\log D} \leq \frac{1}{s}\overline{n} \leq \frac{H(X)}{\log D} + \frac{1}{s}. \tag{105}$$

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.

۱۱ 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.

پ: 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 :

$$\overline{n'} - \overline{n} = n_1 p_2 + n_2 p_1 - n_1 p_1 - n_2 p_2 = (n_1 - n_2)(p_2 - p_1) \leq 0. \tag{106}$$

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

$$\begin{aligned} S &= s_1 s_2 \cdots s_{n_{M-1}} \\ \tilde{S} &\equiv S'\tilde{S}' = s'_1 s'_2 \cdots s'_{n_{M-1}} (s'_{n_{M-1}+1} s_{n_{M-1}+2} \cdots s'_{n_M}) \end{aligned} \tag{107}$$

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 پ: 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.

۱۲ روش هوفمان برای ساختن کد های لحظه ای بهینه

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

$$(X,P) = \{(x_1,p_1),(x_2,p_2),\cdots(x_M,p_M)\}. \tag{108}$$

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

$$(\tilde{X},\tilde{P}) = \{(x_1,p_1),(x_2,p_2),\cdots(x_{M-2},p_{M-2}),(x_{M-1,M},p_{M-1}+p_M)\}. \tag{109}$$

سوال: منظوراز $x_{M-1,M}$ 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:

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

(110)

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

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

(111)

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:

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

(112)

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:

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

(113)

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:

$$\overline{n}-\overline{\tilde{n}}=(p_{M_1}+p_M)(n_{M-1,M}+1-n_{M-1,M})=p_{M-1}+p_M, \tag{114}$$

و

$$\overline{n'}-\overline{\tilde{n'}}=(p_{M_1}+p_M)(n'_M-n'_M-1)=p_{M-1}+p_M. \tag{115}$$

درنتیجه

$$\overline{n}-\overline{\tilde{n}}=\overline{n'}-\overline{\tilde{n'}} \tag{116}$$

که ازآن خواهیم داشت:

$$if\ \ \overline{n'}<\overline{n}\longrightarrow\overline{\tilde{n'}}<\overline{\tilde{n}}. \tag{117}$$

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:

X	P
x_1	0.5
x_2	0.35
x_3	0.15

(118)

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

(119)

وازآنجا

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

(120)

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.