

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



پروژه کارشناسی

تعمیم اندازه اطلاع کای-دو گسسته و کاربرد آن در پردازش تصویر

نگارش

آذرمیدخت کیانی

استاد راهنما

دکتر زهرا صابری

تابستان ۱۴۰۵

چکیده: در این پروژه، با در نظر گرفتن اندازه‌های اطلاع کای-دو تعمیم‌یافته و کای-دو تعمیم‌یافته نسبی، نسخه‌های گسسته‌ای از این معیارهای اطلاع معرفی می‌شود. سپس تعمیم‌هایی از این کمیت‌های اطلاع براساس خاصیت تحدب آن‌ها ارائه می‌شود. برخی از ویژگی‌های مهم این اندازه‌های جدید و روابط بین آن‌ها مورد مطالعه قرار می‌گیرد. به علاوه، کارکرد این معیارهای جدید اطلاع برای برخی از مدل‌های شناخته شده و پرکاربرد در شاخه‌های نظریه کدگذاری و ترمودینامیک از قبیل توزیع های اسکورت و اسکورت تعمیم یافته مطالعه می شود. در پایان هم دو کاربرد برای اندازه اطلاع معرفی شده کای-دو تعمیم یافته گسسته در زمینه ارزیابی کیفیت تصاویر بررسی می‌شود. علاوه بر این، نتایج به دست آمده از کارکرد این معیارها با عملکرد معیار مهم نسبت سیگنال به نویز پیک که یک کمیت شناخته شده در ارزیابی کیفیت تصاویر است، مقایسه می‌شوند. نشان داده می‌شود که اندازه واگرایی کای-دو تعمیم‌یافته عملکردی مشابه با معیار نسبت سیگنال به نویز پیک دارد و می‌تواند به عنوان یک معیار جایگزین به کار رود.

واژه‌های کلیدی: اندازه واگرایی کای-دو، تابع محدب، نامساوی جنسن، توزیع اسکورت.

فهرست مطالب

۱ فصل اول: مقدماتی از نظریه اطلاع	۳
۱-۱ مقدمه	۳
۱-۲ نظریه اطلاع و ارتباط با سایر علوم	۴
۱-۳ آنتروپی	۱۶
۱-۴ آنتروپی توام و شرطی	۲۰
۱-۵ آنتروپی نسبی و اطلاع متقابل	۲۲
۱-۶ رابطه بین آنتروپی و اطلاع متقابل	۲۳
۱-۷ قواعد زنجیره‌ای برای آنتروپی، آنتروپی نسبی و اطلاع متقابل	۲۵
۱-۸ نامساوی جنسن و نتایج آن	۲۸
۱-۹ نامساوی مجموع لگاریتمی و کاربردهای آن	۳۳
۱-۱۰ نامساوی پردازش داده‌ها	۳۷
۱-۱۱ آماره بسنده	۳۸
۱-۱۲ نامساوی فانو	۴۱
۲ فصل دوم: تعمیم اندازه اطلاع کای-دو گسسته و کاربرد در پردازش تصویر	۴۵
۲-۱ مقدمه	۴۵
۲-۲ اندازه اطلاع χ^2_α گسسته و تعمیم آن	۵۰
۲-۳ اندازه اطلاع جنسن- χ^2_α گسسته	۵۴
۲-۴ اندازه اطلاع جنسن- $\chi^2_\alpha(w, \lambda)$	۵۶
۲-۵ کاربرد در پردازش تصویر	۵۹
۲-۵-۱ تصویر دریاچه	۵۹

۶۳.....	۲-۵-۲ تصویر جاده‌ای در نروژ.....
۶۴.....	۲-۶ بحث و نتیجه‌گیری.....
۶۵.....	فهرست منابع.....

فصل اول: مقدماتی از نظریه اطلاع

۱-۱ مقدمه

در این فصل، تعاریف پایه‌ای مورد نیاز برای فصل دوم را ارائه می‌کنیم. پس از بخش مقدمه، ارتباط نظریه اطلاع با سایر شاخه‌های علوم به طور مختصر مورد بحث قرار می‌گیرد. در بخش‌های بعدی به تعریف آنتروپی و اطلاع متقابل، قواعد زنجیره‌ای، نامنفی بودن اطلاع متقابل، و نامساوی پردازش داده پرداخته شده و با بررسی آماره‌های بسنده و نامساوی فانو، این تعاریف را تشریح می‌نماییم.

مفهوم «اطلاع»، چنان گسترده است که نمی‌توان آن را به‌طور کامل در یک تعریف واحد گنجاند. با این حال، برای هر توزیع احتمال، کمیتی به نام «آنتروپی» تعریف می‌کنیم که دارای ویژگی‌های بسیاری است و با مفهوم شهودی آنچه که یک معیار اطلاع باید باشد، مطابقت دارد. این مفهوم به تعریف «اطلاع متقابل» تعمیم می‌یابد، معیاری برای سنجش مقدار اطلاعی که یک متغیر تصادفی درباره یک متغیر تصادفی دیگر در خود دارد. در این چارچوب، آنتروپی به «خود-اطلاع»^۱ یک متغیر تصادفی تبدیل می‌شود.

اطلاع متقابل، حالت خاصی از کمیتی عمومی‌تر به نام «آنتروپی نسبی» است که معیاری برای اندازه‌گیری فاصله بین دو توزیع احتمال می‌باشد. تمام این کمیت‌ها پیوند نزدیکی با یکدیگر دارند

^۱ Self-information

و از تعدادی ویژگی‌های ساده مشترک برخوردارند که برخی از آن‌ها را در این فصل استخراج خواهیم کرد. در بخش بعدی به صورت مختصر نشان خواهیم داد که چگونه این کمیت‌ها به عنوان پاسخ‌های طبیعی به پرسش‌های متعددی در زمینه‌های ارتباطات، آمار، پیچیدگی و شرط‌بندی پدیدار می‌شوند.

۲-۱ نظریه اطلاع و ارتباط با سایر علوم

نظریه اطلاع به دو پرسش بنیادی در نظریه ارتباطات پاسخ می‌دهد: غایت فشرده‌سازی داده چیست (پاسخ: آنتروپی^۱)، و غایت نرخ مخابراتی ارتباط چیست (پاسخ: ظرفیت کانال^۲). به همین دلیل، برخی نظریه اطلاع را زیرمجموعه‌ای از نظریه ارتباطات می‌دانند، اما استدلال می‌کنیم که نظریه اطلاع بسیار فراتر از مطالب مربوط به نظریه ارتباطات است (کاور و توماس، ۱۹۹۱). در واقع، این نظریه سهم بنیادی در مکانیک آماری (ترمودینامیک)، علوم کامپیوتر (پیچیدگی کولموگوروف یا پیچیدگی الگوریتمی)، فلسفه علم («ساده‌ترین تبیین، بهترین است»)، و احتمالات و آمار (توان‌های خطا برای آزمون فرض بهینه و برآورد) دارد. شکل ۱-۱ رابطه نظریه اطلاع با سایر حوزه‌ها را ترسیم می‌کند. همان‌طور که این شکل نشان می‌دهد، نظریه اطلاع با فیزیک (مکانیک آماری)، ریاضیات (نظریه احتمالات)، مهندسی برق (نظریه ارتباطات) و علوم کامپیوتر (پیچیدگی الگوریتمی) اشتراک دارد. در ادامه حوزه‌های اشتراک را با جزئیات بیشتری توصیف می‌کنیم.

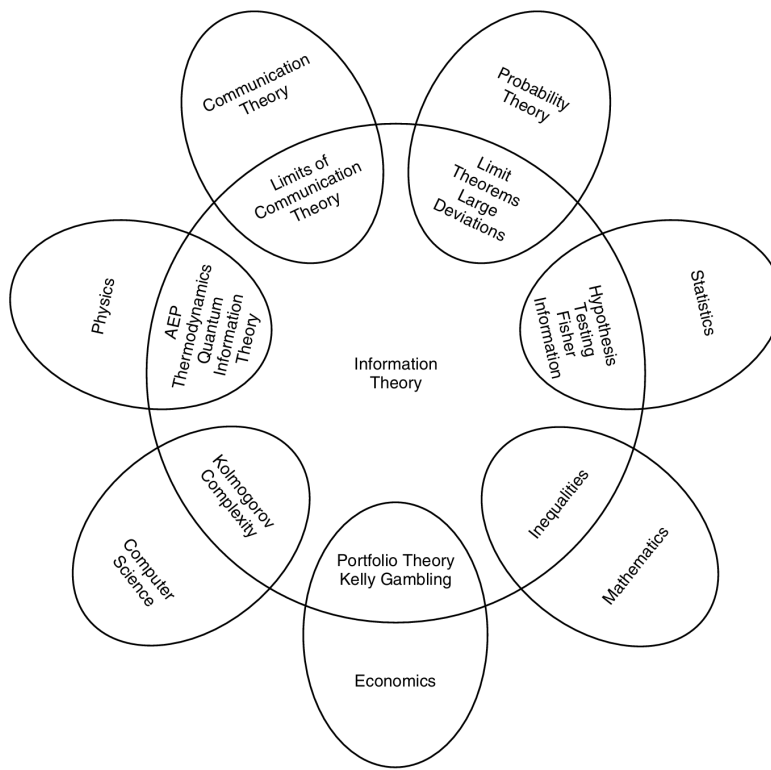
مهندسی برق (نظریه ارتباطات^۳). در اوایل دهه ۱۹۴۰، تصور می‌شد که ارسال اطلاع با نرخ مثبت و احتمال خطای ناچیز، ناممکن است. شانون با اثبات اینکه احتمال خطا را می‌توان برای تمامی نرخ‌های ارتباطی زیر ظرفیت کانال، تقریباً صفر کرد، جامعه نظریه ارتباطات را شگفت‌زده کرد. ظرفیت کانال را می‌توان به سادگی از روی ویژگی‌های نویز کانال محاسبه کرد. شانون در ادامه استدلال کرد که فرآیندهای تصادفی نظیر موسیقی و گفتار، دارای پیچیدگی تقلیل‌ناپذیری هستند که سیگنال نمی‌تواند فراتر از آن فشرده شود. او با احترام به کاربرد موازی این واژه در ترمودینامیک، آن را آنتروپی نامید و استدلال کرد که اگر آنتروپی منبع، کمتر از ظرفیت کانال

¹ Entropy

² Channel Capacity

³ Communication Theory

باشد، می‌توان به یک ارتباط مجانبی بدون خطا دست یافت.



شکل ۱-۱: ارتباط نظریه اطلاع با سایر علوم.



شکل ۱-۲: نظریه اطلاع به مثابه نقاط حدی در نظریه ارتباطات.

همان‌طور که در شکل ۱-۲ نشان داده شده است، نظریه اطلاع امروزه نمایانگر نقاط حدی مجموعه تمامی طرح‌های ممکن برای ارتباطات است. حداقل فشردسازی داده یعنی $I(X; \hat{X})$ یک نقطه حدی مجموعه ایده‌های ارتباطی است. تمامی طرح‌های فشردسازی داده به نرخ‌های توصیفی نیاز دارند که حداقل برابر با این مقدار کمینه باشد. یک نقطه حدی دیگر، حداکثر انتقال داده یعنی $I(X; Y)$ است که به عنوان ظرفیت کانال شناخته می‌شود. بنابراین، تمامی طرح‌های مدولاسیون و فشردسازی داده در میان این محدوده‌ها قرار می‌گیرند.

نظریه اطلاع همچنین روش‌هایی برای دستیابی به این حدهای نهایی در نظریه ارتباطات پیشنهاد می‌کند. با این حال، این طرح‌های ارتباطی که از نظر تئوری بهینه هستند، هرچقدر هم زیبا باشند، ممکن است از نظر محاسباتی دستیابی به آنها غیرعملی باشد. تنها به دلیل امکان‌پذیری محاسباتی طرح‌های ساده مدولاسیون و دمودولاسیون است که به جای کدگذاری تصادفی و قاعده رمزگشایی نزدیک‌ترین همسایه که توسط اثبات شانون برای قضیه ظرفیت کانال پیشنهاد شده، از آنها استفاده می‌کنیم. پیشرفت در مدارهای مجتمع و طراحی کد، ما را قادر ساخته تا از برخی دستاوردهای پیشنهادی نظریه شانون بهره‌برداری کنیم. عملی بودن محاسباتی سرانجام با ظهور توربو کدها حاصل شد. یک نمونه خوب از کاربرد ایده‌های نظریه اطلاع، استفاده از کدهای تصحیح‌کننده خطا در لوح‌های فشرده (CD) و DVD ها است.

کارهای اخیر در جنبه‌های ارتباطی نظریه اطلاع، روی نظریه اطلاع شبکه متمرکز شده‌اند: یعنی نظریه نرخ‌های هم‌زمان ارتباط از تعداد زیادی فرستنده به تعداد زیادی گیرنده در حضور تداخل و نویز. برخی از داد و ستدهای نرخ بین فرستنده‌ها و گیرنده‌ها غیرمنتظره هستند و همگی دارای نوعی سادگی ریاضی می‌باشند. با این حال، هنوز یک نظریه واحد برای آن یافت نشده است.

علوم کامپیوتر¹ (پیچیدگی کولموگوروف). کولموگوروف، چایتین و سالومونوف این ایده را مطرح کردند که پیچیدگی یک رشته داده را می‌توان با طول کوتاه‌ترین برنامه رایانه‌ای دودویی برای محاسبه آن رشته تعریف کرد. بنابراین، پیچیدگی برابر با حداقل طول توصیف است. این تعریف پیچیدگی در نهایت جهانی از آب درمی‌آید، یعنی مستقل از رایانه است، و از اهمیتی بنیادی برخوردار است. از این‌رو، پیچیدگی کولموگوروف شالوده نظریه پیچیدگی توصیفی است. خوشبختانه، اگر دنباله‌ای به‌طور تصادفی از توزیعی با آنتروپی شانون H انتخاب شده باشد، پیچیدگی کولموگوروف K تقریباً برابر با آنتروپی شانون H است. بنابراین، پیوند میان نظریه اطلاع و پیچیدگی کولموگوروف کامل است. در واقع، پیچیدگی کولموگوروف را بنیادی‌تر از آنتروپی شانون می‌دانیم. این همان فشرده‌سازی نهایی داده است و به روشی از استنباط می‌انجامد که از نظر منطقی سازگار است.

¹ Computer Science

میان پیچیدگی الگوریتمی و پیچیدگی محاسباتی رابطه‌ای مکمل و دلیلی وجود دارد. می‌توان پیچیدگی محاسباتی (پیچیدگی زمانی) و پیچیدگی کولموگوروف (طول برنامه یا پیچیدگی توصیفی) را به‌مثابه دو محور در نظر گرفت که به‌ترتیب متناظر با زمان اجرای برنامه و طول برنامه‌اند. پیچیدگی کولموگوروف بر کمینه‌سازی در امتداد محور دوم تمرکز دارد و پیچیدگی محاسباتی بر کمینه‌سازی در امتداد محور اول. کار اندکی بر کمینه‌سازی هم‌زمان این دو انجام شده است.

فیزیک^۱ (ترمودینامیک). مکانیک آماری زادگاه آنتروپی و قانون دوم ترمودینامیک است. آنتروپی همیشه افزایش می‌یابد. قانون دوم ترمودینامیک به ما اجازه می‌دهد ادعاهای مربوط به ماشین‌های دایما در حال حرکت را کنار بگذاریم.

ریاضیات^۲ (نظریه احتمالات و آمار). کمیت‌های بنیادی نظریه اطلاع (آنتروپی، آنتروپی نسبی، و اطلاع متقابل) به‌صورت تابعی‌هایی از توزیع‌های احتمالی تعریف می‌شوند. به نوبه خود، این کمیت‌ها رفتار بلندمدت دنباله‌های متغیرهای تصادفی را مشخص می‌کنند و به ما اجازه می‌دهند احتمال رویدادهای نادر (نظریه انحرافات بزرگ) را برآورد کنیم و بهترین توان آزمون و خطا را در آزمون‌های فرض بیابیم.

فلسفه علم^۳ (تیغ اوکام). ویلیام اوکام گفت: «علت‌ها نباید فراتر از ضرورت تکثیر شوند»، یا به بیان دیگر، «ساده‌ترین تبیین بهترین تبیین است». سالومونوف و چایتین به‌طور قانع‌کننده‌ای استدلال کردند که اگر ترکیب وزنی همه برنامه‌هایی را که داده‌ها را توضیح می‌دهند در نظر بگیریم و خروجی بعدی آن‌ها را مشاهده کنیم، یک روش پیش‌بینی خوب به دست می‌آید. افزون بر این، این استنباط در بسیاری از مسائلی که آمار از عهده آن‌ها بر نمی‌آید نیز کار خواهد کرد. برای مثال، این روش سرانجام ارقام بعدی π را پیش‌بینی خواهد کرد. وقتی این روش بر پرتاب سکه‌هایی که با احتمال $0/7$ شیر می‌آیند اعمال شود، همین نیز استنباط خواهد شد. وقتی بر بازار سهام اعمال شود، این روش باید اساساً همه «قوانین» بازار سهام را بیابد و آن‌ها را به‌طور بهینه تعمیم دهد. از نظر اصولی، چنین روشی می‌توانست قوانین نیوتن در فیزیک را بیابد. البته، چنین استنباطی

¹ Physics

² Mathematics

³ Philosophy of Science

بسیار غیرعملی است، زیرا غربال کردن همه برنامه‌های رایانه‌ای که در تولید داده‌های موجود شکست می‌خورند، بینهایت زمان خواهد برد و آنچه را فردا رخ می‌دهد، صد سال دیگر پیش‌بینی می‌کردیم.

اقتصاد^۱ (سرمایه‌گذاری). سرمایه‌گذاری مکرر در یک بازار سهام ایستا به رشد نمایی ثروت می‌انجامد. نرخ رشد ثروت، دوگانه‌ای از نرخ آنتروپی بازار سهام است. شباهت‌های میان نظریه سرمایه‌گذاری بهینه در بازار سهام و نظریه اطلاع چشمگیر هستند.

محاسبه در برابر ارتباط^۲. هنگامی که رایانه‌های بزرگ‌تر را از اجزای کوچک‌تر می‌سازیم، هم با یک حد محاسباتی و هم با یک حد ارتباطی روبه‌رو می‌شویم. محاسبه مقید به ارتباط است و ارتباط مقید به محاسبه. این دو در هم تنیده می‌شوند، و از این‌رو همه پیشرفت‌ها در نظریه ارتباطات از طریق نظریه اطلاع، باید اثری مستقیم بر نظریه محاسبه داشته باشند.

پرسش‌های اولیه‌ای که نظریه اطلاع به آن‌ها می‌پردازد، در حوزه‌های فشرده‌سازی و انتقال داده قرار دارند. پاسخ‌ها کمیت‌هایی نظیر آنتروپی و اطلاع متقابل هستند که توابعی از توزیع‌های احتمالی زیربنای فرایند ارتباط‌اند. چند تعریف به بحث اولیه کمک خواهند کرد.

آنتروپی (یا آنتروپی شانون) یک متغیر تصادفی X با تابع جرم احتمال $p(x)$ به صورت زیر تعریف می‌شود:

$$H(X) = - \sum_x p(x) \log p(x),$$

در کل این پروژه از لگاریتم در مبنای ۲ استفاده می‌کنیم. در این صورت، آنتروپی با واحد بیت اندازه‌گیری خواهد شد. آنتروپی معیاری برای میانگین عدم قطعیت در متغیر تصادفی است. این مقدار، تعداد بیت‌هایی است که به‌طور میانگین برای توصیف متغیر تصادفی مورد نظر لازم است.

به عنوان مثال، یک متغیر تصادفی را در نظر بگیرید که دارای توزیع یکنواخت روی ۳۲ برآمد ممکن است. برای شناسایی یک برآمد، به برجسبی نیاز داریم که ۳۲ مقدار مختلف بگیرد. بنابراین،

¹ Economics

² Computation vs. Communication

رشته‌های ۵ بیتی به عنوان برچسب کافی هستند. آنتروپی این متغیر تصادفی برابر است با

$$H(X) = - \sum_{i=1}^{32} p(i) \log p(i) = - \sum_{i=1}^{32} \frac{1}{32} \log \frac{1}{32} = \log 32 = 5 \text{ bits.}$$

که با تعداد بیت‌های لازم برای توصیف X یکسان است. در این حالت، تمام برآمدها نمایش‌هایی با طول یکسان دارند.

به عنوان مثالی دیگر، فرض کنید یک مسابقه اسب‌دوانی با شرکت هشت اسب داریم. فرض کنید احتمالات برنده شدن این هشت اسب به صورت $P = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$ باشد. می‌توان آنتروپی این مسابقه اسب‌دوانی را به صورت زیر محاسبه کرد:

$$H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{16} \log \frac{1}{16} - 4 \frac{1}{64} \log \frac{1}{64} = 2 \text{ bits.}$$

فرض کنید بخواهیم پیامی بفرستیم که مشخص کند کدام اسب برنده مسابقه شده است. یک راه این است که اندیس اسب برنده را ارسال کنیم. این توصیف برای هر یک از اسب‌ها به ۳ بیت نیاز دارد. اما احتمال‌های برد یکسان نیستند. بنابراین منطقی است که برای اسب‌های با احتمال برنده شدن بالاتر، توصیف‌های کوتاه‌تر و برای اسب‌ها با احتمال برنده شدن کمتر، توصیف‌های بلندتر به کار برده شود، به طوری که طول متوسط توصیف کاهش یابد. برای مثال، می‌توان مجموعه زیر از رشته‌های بیتی را برای نمایش هشت اسب به کار برد: ۰، ۱۰، ۱۱۰، ۱۱۱۰، ۱۱۱۱۰۰، ۱۱۱۱۰۱، ۱۱۱۱۱۰، ۱۱۱۱۱۱. طول متوسط توصیف در این حالت ۲ بیت است، در مقابل ۳ بیت برای کد یکنواخت. توجه کنید که طول متوسط توصیف در این حالت برابر با آنتروپی است. می‌توان نشان داد (کاور و توماس، ۱۹۹۱) که آنتروپی یک متغیر تصادفی، کران پایین تعداد متوسط بیت‌های لازم برای نمایش آن متغیر تصادفی است.

مفهوم آنتروپی در نظریه اطلاع با مفهوم آنتروپی در مکانیک آماری مرتبط است. اگر دنباله‌ای از n متغیر تصادفی مستقل و هم‌توزیع (i.i.d) را در نظر بگیریم، نشان خواهیم داد که احتمال رخ دادن یک دنباله «معمولی» تقریباً $2^{-nH(X)}$ است و حدود $2^{nH(X)}$ دنباله‌ی معمولی از این نوع وجود دارد. این خاصیت که با نام خاصیت تجزیه‌ی مقارن مجانبی (AEP) شناخته می‌شود، مبنای

بسیاری از اثبات‌ها در نظریه اطلاع است. بعداً مسائل دیگری را نیز مطرح خواهیم کرد که در آن‌ها آنتروپی به‌طور طبیعی به‌عنوان پاسخ ظاهر می‌شود (برای مثال، تعداد پرتاب‌های منصفانه سکه که برای تولید یک متغیر تصادفی لازم است).

مفهوم پیچیدگی توصیفی یک متغیر تصادفی را می‌توان گسترش داد تا پیچیدگی توصیفی یک رشته‌ی منفرد را تعریف کند. پیچیدگی کولموگوروف یک رشته‌ی دودویی، طول کوتاه‌ترین برنامه رایانه‌ای است که آن رشته را چاپ می‌کند. اگر رشته واقعا تصادفی باشد، پیچیدگی کولموگوروف به آنتروپی نزدیک است. پیچیدگی کولموگوروف چارچوبی طبیعی برای بررسی مسائل استنباط و مدل‌سازی آماری است و به درک روشن‌تری از تیغ اوکام می‌انجامد: «ساده‌ترین توضیح بهترین توضیح است.» برخی ویژگی‌های ساده پیچیدگی کولموگوروف را در ادامه این فصل بررسی می‌کنیم.

آنتروپی، عدم قطعیت یک متغیر تصادفی منفرد است. می‌توانیم آنتروپی شرطی $H(X|Y)$ را تعریف کنیم که آنتروپی یک متغیر تصادفی به شرط آگاهی از مقدار یک متغیر تصادفی دیگر است. کاهش عدم قطعیت ناشی از یک متغیر تصادفی دیگر را اطلاع متقابل می‌نامند. برای دو متغیر تصادفی X و Y ، این کاهش عدم قطعیت همان اطلاع متقابل است.

$$I(X; Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

اطلاع متقابل $I(X; Y)$ معیاری برای وابستگی میان دو متغیر تصادفی است. این کمیت نسبت به X و Y متقارن است، همواره نامنفی است و فقط و فقط در صورتی برابر با صفر است که X و Y مستقل باشند.

کانال ارتباطی سامانه‌ای است که در آن خروجی به‌صورت احتمالاتی به ورودی وابسته است. این کانال با یک ماتریس گذار احتمالاتی $p(y|x)$ مشخص می‌شود که توزیع شرطی خروجی را در حالت معلوم بودن ورودی تعیین می‌کند. برای یک کانال ارتباطی با ورودی X و خروجی Y ، می‌توان ظرفیت C را به صورت زیر تعریف کرد:

$$C = \max_{p(x)} I(X; Y).$$

ظرفیت بیشترین نرخ است که با آن می‌توان اطلاع را از طریق کانال ارسال کرده و اطلاع را در

خروجی با احتمال خطایی که به طور نامتناهی کوچک می شود بازیابی کرد. این موضوع را با چند مثال نشان می دهیم.

کانال دودویی بدون نویز: در این کانال، ورودی دودویی دقیقاً در خروجی بازتولید می شود. این کانال در شکل ۱-۳ نشان داده شده است. در اینجا، هر بیت ارسال شده بدون خطا دریافت می شود. بنابراین، در هر ارسال، می توان ۱ بیت را به طور قابل اعتماد برای گیرنده فرستاد، و ظرفیت کانال برابر با ۱ بیت است. همچنین می توان ظرفیت اطلاعی را محاسبه کرد:

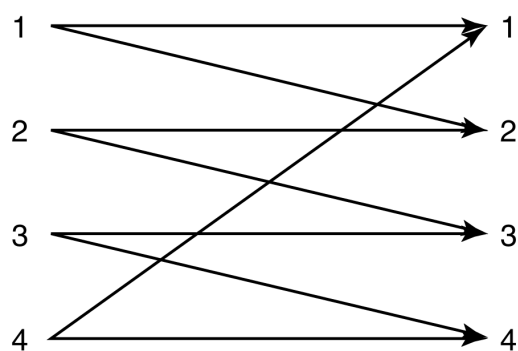
$$C = \max I(X; Y) = 1 \text{ bit.}$$

کانال چهارنمادی نویزی: کانالی را در نظر بگیرید که در شکل ۱-۴ نشان داده شده است. در این کانال، هر حرف ورودی یا با احتمال $\frac{1}{2}$ با همان حرف دریافت می شود، یا با احتمال $\frac{1}{2}$ به حرف بعدی. اگر از هر چهار نماد ورودی استفاده شود، بررسی خروجی با قطعیت نشان نخواهد داد که کدام نماد ورودی ارسال شده است. اما اگر فقط از دو ورودی استفاده شود (مثلاً ۱ و ۳)، می توان بلافاصله از روی خروجی تشخیص داد که کدام نماد ورودی ارسال شده است. در این صورت، این کانال مانند کانال بدون نویز عمل می کند و می توان در هر ارسال ۱ بیت را بدون خطا از طریق این کانال منتقل کرد. می توان ظرفیت کانال را در این حالت به صورت $C = \max I(X; Y)$ محاسبه کرد، و این مقدار برابر ۱ بیت در هر ارسال است که با تحلیل بالا سازگار است.

$$0 \longrightarrow 0$$

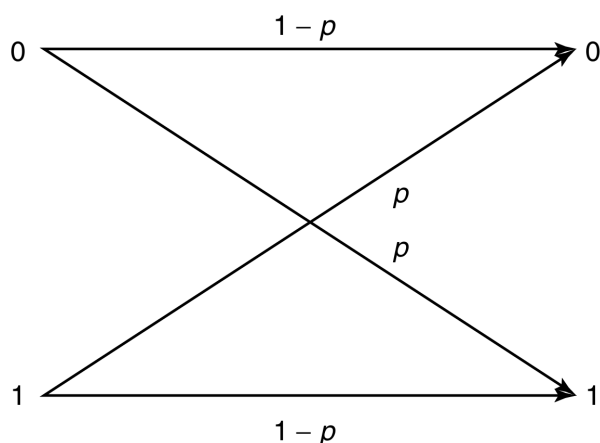
$$1 \longrightarrow 1$$

شکل ۱-۳: کانال دودویی بی نویز $C = 1 \text{ bit}$.



شکل ۴-۱: کانال نویزی

به طور کلی، کانال‌های ارتباطی معمولاً ساختار ساده این مثال را ندارند، بنابراین همیشه نمی‌توان زیرمجموعه‌ای از ورودی‌ها را برای ارسال اطلاع بدون خطا شناسایی کرد. اما اگر دنباله‌ای از ارسال‌ها را در نظر بگیریم، تمام کانال‌ها شبیه این مثال به نظر می‌رسند و در این صورت می‌توان زیرمجموعه‌ای از دنباله‌های ورودی (کدواژه‌ها) را شناسایی کرد که می‌توان از آن‌ها برای ارسال اطلاع از طریق کانال استفاده کرد، به طوری که مجموعه‌های دنباله‌های خروجی ممکن مرتبط با هر یک از کدواژه‌ها، تقریباً مجزا باشند. سپس می‌توان به دنباله خروجی نگاه کرده و دنباله ورودی را با احتمال خطای ناچیز (میل‌کننده به صفر) شناسایی کرد.



شکل ۵-۱: کانال دودویی متقارن.

کانال متقارن دودویی: این مثال یک مثال پایه از یک سیستم ارتباطی نویزی است. این کانال در شکل ۵-۱ نشان داده شده است. کانال دارای یک ورودی دودویی است و خروجی آن با احتمال $1-p$ برابر با ورودی است. از سوی دیگر، با احتمال p ، یک 0 به صورت 1 دریافت

می‌شود و بالعکس یک 1 به صورت 0 دریافت می‌شود. در این مورد، می‌توان ظرفیت کانال را به صورت $C = 1 + p \log p + (1 - p) \log(1 - p)$ بیت در هر ارسال محاسبه کرد. با این حال، دیگر بدیهی نیست که چگونه می‌توان به این ظرفیت دست یافت. اما اگر از کانال دفعات زیادی استفاده شود، کانال شروع به شبیه‌شدن به کانال چهارنمادی نویزی می‌کند و می‌توان اطلاعات را با نرخ C بیت در هر ارسال و با احتمال خطای به دلخواه کوچک، منتقل کرد.

حد نهایی نرخ انتقال اطلاع در یک کانال، توسط ظرفیت کانال تعیین می‌شود. قضیه کدگذاری کانال (کاور و توماس، ۱۹۹۱) نشان می‌دهد که این حد با استفاده از کدهایی با طول بلوک زیاد قابل دستیابی است. در سیستم‌های ارتباطی عملی، محدودیت‌هایی بر پیچیدگی کدهایی که می‌توان استفاده کرد وجود دارد و به همین دلیل، ممکن است قادر به دستیابی به ظرفیت نباشیم.

اطلاع متقابل، در واقع حالت خاصی از کمیتی عمومی‌تر به نام «آنتروپی نسبی» یا $D(p \parallel q)$ است که معیاری برای «فاصله» میان دو تابع جرم احتمال p و q به شمار می‌رود. این کمیت به صورت زیر تعریف می‌شود:

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

اگرچه آنتروپی نسبی یک متریک حقیقی نیست، اما برخی از خواص یک متریک را دارد. به‌ویژه، همواره نامنفی است و فقط و فقط زمانی صفر می‌شود که $p = q$. آنتروپی نسبی به‌عنوان نمای احتمال خطا در یک آزمون فرض میان توزیع‌های p و q ظاهر می‌شود. همچنین می‌توان از آنتروپی نسبی برای تعریف یک هندسه برای توزیع‌های احتمال استفاده کرد، که به ما اجازه می‌دهد بسیاری از نتایج نظریه انحرافات بزرگ را تفسیر کنیم.

میان نظریه اطلاع و نظریه سرمایه‌گذاری در بازار سهام، شباهت‌های متعددی وجود دارد. بازار سهام با یک بردار تصادفی X تعریف می‌شود که مولفه‌های آن اعدادی نامنفی هستند که برابر با نسبت قیمت یک سهم در پایان روز به قیمت آن در آغاز روزند. برای یک بازار سهام با توزیع $F(x)$ ، می‌توان نرخ دوبرابر شدن W را به صورت زیر تعریف کرد:

$$W = \max_{b: b_i \geq 0, \sum b_i = 1} \int \log b^T x dF(x).$$

نرخ دوبرابرشدن، بیشینه نمای مجانبی در رشد ثروت است. این نرخ، دارای ویژگی‌های متعددی است که با ویژگی‌های آنتروپی هم‌ارز (مشابه) هستند.

کمیت‌های H (آنتروپی)، I (اطلاع متقابل)، C (ظرفیت)، D (آنتروپی نسبی)، K (پیچیدگی کولموگوروف) و W (نرخ دوبرابرشدن) به‌طور طبیعی در حوزه‌های زیر پدیدار می‌شوند:

فشرده‌سازی داده‌ها^۱: آنتروپی H برای یک متغیر تصادفی، کران پایینی بر طول متوسط کوتاه‌ترین توصیف این متغیر تصادفی است. می‌توان توصیفاتی با طول متوسط در حد یک بیت آنتروپی ساخت. اگر قید بازیابی کامل منبع را سست کنیم، می‌توان پرسید که برای توصیف منبع با اعوجاج D ^۲ به چه نرخ‌های ارتباطی نیاز است؟ و چه ظرفیت‌های کانالی برای انتقال این منبع از طریق کانال و بازسازی آن با اعوجاج کمتر یا مساوی D کافی هستند؟ این موضوع، حوزه نظریه نرخ-اعوجاج^۳ است. وقتی می‌کوشیم مفهوم کوتاه‌ترین توصیف را برای اشیاء غیرتصادفی صورت‌بندی کنیم، به تعریف پیچیدگی کولموگوروف (K) می‌رسیم.

انتقال داده‌ها^۴: مسئله انتقال اطلاعات به نحوی که گیرنده بتواند پیام را با احتمال خطای اندک رمزگشایی کند، مطرح است. در اصل، به دنبال «کدواژه‌ها» (دنباله‌هایی از نمادهای ورودی کانال) هستیم که به لحاظ متمایز بودن نسخه‌های نویزی آن‌ها (در خروجی کانال)، از یکدیگر «دور» باشند؛ این امر با مسئله‌ی بسته‌بندی کره^۵ در یک فضا با ابعاد بالا معادل است. شانون با استفاده از کدهایی که به‌صورت تصادفی تولید شده بودند نشان داد که می‌توان اطلاع را با هر نرخ کمتر از ظرفیت کانال (C)، با احتمال خطای به دلخواه کوچک ارسال کرد. یکی از ایده‌های کلیدی در این اثبات، مفهوم دنباله‌های نمونه^۶ است. ظرفیت C ، لگاریتم تعداد سیگنال‌های ورودی قابل تمایز است.

نظریه اطلاع شبکه^۷: مسائلی چون فشرده‌سازی توام منابع (قضیه اسلپین-ولف)، کانال‌های

^۱ Data Compression

^۲ Distortion

^۳ Rate Distortion Theory

^۴ Data Transmission

^۵ Sphere Packing

^۶ Typical Sequences

^۷ Network Information Theory

دسترسی چندگانه (تحقیقات لیانو و آلسود)، کانال‌های پخش و به‌طور کلی محیط‌های دارای تداخل و نویز، همگی در حوزه کلی نظریه اطلاع چندکاربره قرار می‌گیرند. اگرچه ممکن است امید به یک نظریه جامع برای شبکه‌ها فراتر از تکنیک‌های پژوهشی فعلی باشد، اما همچنان این امید وجود دارد که تمام پاسخ‌ها تنها شامل اشکال پیچیده‌ای از اطلاع متقابل و آنتروپی نسبی باشند.

نظریه ارگودیک^۱: قضیه حد مرکزی مجانبی (AEP) بیان می‌کند که اکثر n -دنباله‌های نمونه از یک فرآیند ارگودیک، احتمالی در حدود 2^{-nH} دارند و در حدود 2^{nH} دنباله این چنینی وجود دارد.

آزمون فرض^۲: آنتروپی نسبی D به‌مثابه یک نما در احتمال خطا در آزمون فرض میان دو توزیع ظاهر می‌شود و معیاری طبیعی برای «فاصله» میان توزیع‌هاست.

مکانیک آماری^۳: آنتروپی H در مکانیک آماری به‌مثابه معیاری از عدم قطعیت یا بی‌نظمی در یک سیستم فیزیکی مطرح است. قانون دوم ترمودینامیک بیان می‌دارد که آنتروپی یک سیستم بسته نمی‌تواند کاهش یابد.

مکانیک کوانتومی^۴: در اینجا، آنتروپی فون نویمان ($S = \text{tr}(\rho \ln \rho)$) نقش آنتروپی کلاسیک شانون-بولتزمن را ایفا می‌کند.

استنباط^۵: می‌توان از پیچیدگی کولموگوروف برای یافتن کوتاه‌ترین توصیف داده‌ها و استفاده از آن به‌عنوان مدلی برای پیش‌بینی رخداد‌های بعدی استفاده کرد. مدلی که عدم قطعیت یا آنتروپی را بیشینه می‌کند، منجر به رویکرد بیشینه آنتروپی در استنباط می‌شود.

شرط‌بندی و سرمایه‌گذاری^۶: نمای بهینه در نرخ رشد ثروت توسط نرخ دوبرابر شدن W داده می‌شود. در یک مسابقه اسب‌دوانی با احتمالات یکنواخت، مجموع نرخ دوبرابر شدن W و آنتروپی H ثابت است. نتایج مشابهی برای سرمایه‌گذاری در بازار سهام برقرار است.

¹ Ergodic Theory

² Hypothesis Testing

³ Statistical Mechanics

⁴ Quantum Mechanics

⁵ Inference

⁶ Gambling and Investment

نظریه احتمالات^۱: در نظریه انحرافات بزرگ، احتمال یک مجموعه تقریباً 2^{-nD} است که در آن D فاصله آنتروپی نسبی بین نزدیک‌ترین عنصر در مجموعه و توزیع واقعی است.

نظریه پیچیدگی^۲: پیچیدگی کولموگوروف (K) معیاری از پیچیدگی توصیفی یک شیء است که با پیچیدگی محاسباتی (زمان یا فضای مورد نیاز برای محاسبه) مرتبط اما متفاوت است.

۳-۱ آنتروپی

ابتدا مفهوم آنتروپی را معرفی می‌کنیم، که معیاری برای عدم قطعیت یک متغیر تصادفی است. فرض کنید X یک متغیر تصادفی گسسته با تکیه‌گاه $\mathcal{X}$ و تابع جرم احتمال $p(x) = \Pr\{X = x\}$ ، برای $x \in \mathcal{X}$ ، باشد. برای سهولت، تابع جرم احتمال را به صورت $p(x)$ می‌نویسیم و نه $p_X(x)$. بنابراین، $p(x)$ و $p(y)$ به دو متغیر تصادفی متفاوت اشاره دارند و در واقع تابع‌های جرم احتمال متفاوتی، یعنی به ترتیب $p_X(x)$ و $p_Y(y)$ ، هستند.

تعریف ۱-۱: آنتروپی شانون $H(X)$ برای یک متغیر تصادفی گسسته X به صورت زیر تعریف می‌شود:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \blacksquare$$

همچنین این کمیت را به صورت $H(p)$ نیز می‌نویسیم. لگاریتم در پایه ۲ است و بنابراین آنتروپی برحسب بیت اندازه‌گیری می‌شود. برای مثال، آنتروپی پرتاب یک سکه سالم برابر با ۱ بیت است. در این پروژه قرارداد می‌کنیم که $0 \log 0 = 0$ ، که به سادگی با استفاده از مفهوم پیوستگی توجیه می‌شود، زیرا وقتی $x \rightarrow 0$ داریم $x \log x \rightarrow 0$. افزودن جمله‌هایی با احتمال صفر، آنتروپی را تغییر نمی‌دهد.

اگر پایه لگاریتم b باشد، آنتروپی را با $H_b(X)$ نمایش می‌دهیم. اگر پایه لگاریتم e باشد، آنتروپی برحسب نات اندازه‌گیری می‌شود. در این پروژه همه لگاریتم‌ها را در پایه ۲ در نظر می‌گیریم، مگر آن‌که خلاف آن تصریح شود. در نتیجه همه آنتروپی‌ها برحسب بیت خواهند بود.

¹ Probability Theory

² Complexity Theory

توجه کنید که آنتروپی تابعی از توزیع X است. این کمیت به مقادیر واقعی که متغیر تصادفی X می‌تواند اختیار کند وابسته نیست، بلکه فقط به احتمال‌ها وابسته است. امید ریاضی را با E نمایش می‌دهیم. بنابراین، اگر $X \sim p(x)$ آنگاه امید ریاضی یا به طور معادل، مقدار مورد انتظار متغیر تصادفی $g(X)$ به صورت زیر تعریف می‌شود:

$$E_p[g(X)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

یا به شکل ساده‌تر، زمانی که تابع جرم احتمال از سیاق متن قابل درک باشد، آن را به صورت $E[g(X)]$ می‌نویسیم. در اینجا توجه ویژه‌ای به امید $g(X)$ تحت توزیع $p(x)$ خواهیم داشت، هنگامی که $g(X) = \log \frac{1}{p(X)}$.

توجه کنید که آنتروپی X را همچنین می‌توان به عنوان امید ریاضی متغیر تصادفی $\log \frac{1}{p(X)}$ تفسیر کرد، جایی که X طبق تابع جرم احتمال $p(x)$ توزیع شده است. بنابراین:

$$H(X) = E_p \left[\log \frac{1}{p(X)} \right]$$

این تعریف آنتروپی با تعریف آنتروپی در ترمودینامیک مرتبط است. می‌توان تعریف آنتروپی را به صورت اصل موضوعی نیز به دست آورد، یعنی با تعریف مجموعه‌ای از خواصی که آنتروپی یک متغیر تصادفی باید آن‌ها را ارضا کند. با این حال، در این پروژه برای توجیه تعریف آنتروپی از روش اصل موضوعی استفاده نمی‌کنیم، بلکه نشان می‌دهیم که آنتروپی به عنوان پاسخ طبیعی پرسش‌های متعددی پدیدار می‌شود، مانند: «میانگین طول کوتاه‌ترین توصیف یک متغیر تصادفی چقدر است؟»

ابتدا چند پیامد فوری این تعریف را استخراج می‌کنیم.

$$\text{لم ۱-۱: } H(X) \geq 0$$

$$\text{اثبات: چون } 0 \leq p(x) \leq 1 \text{ داریم } \log \frac{1}{p(x)} \geq 0 \quad \blacksquare$$

$$\text{لم ۱-۲: } H_b(X) = (\log_b a) H_a(X)$$

$$\log_b p = (\log_b a) \log_a p \quad \blacksquare$$

اثبات:

خاصیت دوم آنتروپی به ما امکان می‌دهد که پایه لگاریتم را در تعریف تغییر دهیم. بنابراین، آنتروپی را می‌توان از یک پایه به پایه دیگر با ضرب در ضریب مناسب تبدیل کرد.

مثال ۱-۱: فرض کنید

$$X = \begin{cases} 1 & \text{با احتمال } p \\ 0 & \text{با احتمال } 1-p \end{cases}$$

در این صورت

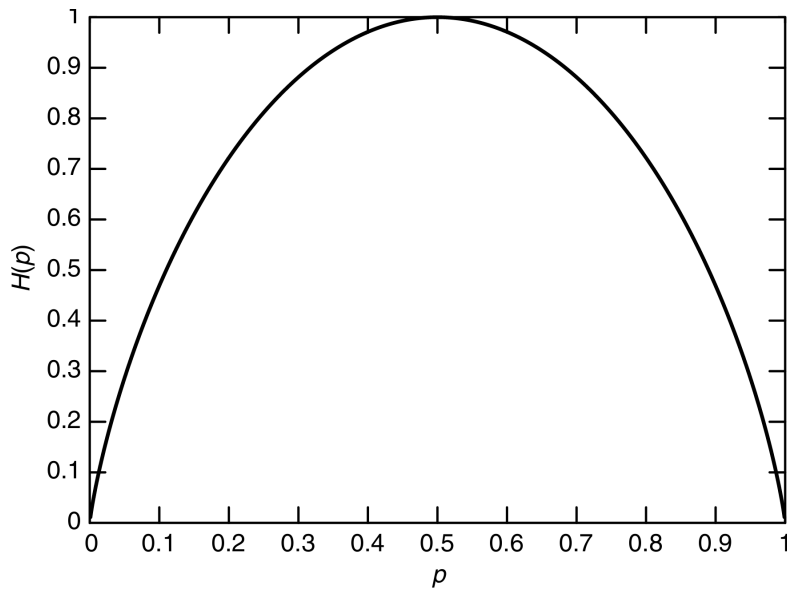
$$H(p) = -p \log p - (1-p) \log(1-p) = H(p)$$

به‌خصوص وقتی $p = \frac{1}{2}$ ، داریم $H(X) = 1 \text{ bit}$. نمودار تابع $H(p)$ در شکل ۶-۱ داده شده است. این شکل برخی از ویژگی‌های بنیادی آنتروپی را نمایش می‌دهد:

- آنتروپی، تابعی مقعر نسبت به توزیع است.

- وقتی $p = 0$ یا $p = 1$ ، مقدار آنتروپی صفر است.

این نتیجه کاملاً معقول است، زیرا وقتی $p = 0$ یا $p = 1$ ، متغیر تصادفی در واقع تصادفی نیست و هیچ عدم قطعیتی وجود ندارد. به‌طور مشابه، عدم قطعیت هنگامی بیشینه است که $p = \frac{1}{2}$ ، و این نیز با بیشینه بودن آنتروپی در همان نقطه سازگار است. ■



شکل ۶-۱: $H(p)$ در برابر p .

مثال ۲-۱: فرض کنید تابع احتمال متغیر تصادفی X به صورت زیر باشد:

x	a	b	c	d
$p(x)$	$1/2$	$1/4$	$1/8$	$1/8$

در این صورت آنتروپی X برابر است با

$$H(X) = - \left[\frac{1}{2} \log \frac{1}{2} + \frac{1}{4} \log \frac{1}{4} + \frac{1}{8} \log \frac{1}{8} + \frac{1}{8} \log \frac{1}{8} \right] = \frac{7}{4} \text{ bits}$$

فرض کنید می‌خواهیم مقدار X را با حداقل تعداد پرسش‌های دوتایی (بله/خیر) تعیین کنیم. پرسش اول و کارآمد این است: «آیا $X = a$ است؟» این پرسش، احتمال را به دو قسمت تقسیم می‌کند. اگر پاسخ به پرسش اول «نه» باشد، پرسش دوم می‌تواند این باشد: «آیا $X = b$ است؟» پرسش سوم می‌تواند «آیا $X = c$ است؟» باشد. تعداد مورد انتظار پرسش‌های دوتایی لازم در این روش، 1.75 است. این مقدار در واقع حداقل تعداد پرسش‌های دوتایی مورد انتظار برای تعیین مقدار X است. ثابت شده که حداقل تعداد مورد انتظار پرسش‌های دوتایی برای تعیین X ، عددی بین $H(X)$ و $H(X) + 1$ است. ■

۱-۴ آنتروپی توام و شرطی

در بخش قبل آنتروپی یک متغیر تصادفی منفرد را تعریف کردیم. اکنون این تعریف را به یک جفت متغیر تصادفی تعمیم می‌دهیم. در این تعریف، در واقع چیز تازه‌ای وجود ندارد، زیرا (X, Y) را می‌توان به صورت یک متغیر تصادفی برداری در نظر گرفت.

تعریف ۱-۲: آنتروپی توام $H(X, Y)$ برای یک جفت متغیر تصادفی گسسته (X, Y) با توزیع توام $p(x, y)$ به صورت زیر تعریف می‌شود:

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y),$$

که می‌توان آن را به صورت زیر نیز بیان کرد:

$$H(X, Y) = -E[\log p(X, Y)]$$

همچنین آنتروپی شرطی یک متغیر تصادفی، به شرط متغیر یک متغیر تصادفی دیگر را به عنوان امید ریاضی آنتروپی‌های توزیع‌های شرطی، که نسبت به متغیر تصادفی شرط‌کننده میانگین‌گیری شده‌اند، تعریف می‌کنیم. ■

تعریف ۱-۳: اگر $(X, Y) \sim p(x, y)$ ، آنتروپی شرطی $H(Y | X)$ به صورت زیر تعریف می‌شود:

$$\begin{aligned} H(Y | X) &= \sum_{x \in \mathcal{X}} p(x) H(Y | X = x) = - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y | x) \log p(y | x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) = -E[\log p(Y | X)]. \end{aligned}$$

طبیعی‌بودن تعریف آنتروپی توام و آنتروپی شرطی، با این واقعیت آشکار می‌شود که آنتروپی یک جفت متغیر تصادفی، برابر است با آنتروپی یکی از آن‌ها به علاوه آنتروپی شرطی دیگری. این موضوع در قضیه زیر اثبات شده است. ■

قضیه ۱-۱ (قاعده زنجیره‌ای): $H(X, Y) = H(X) + H(Y | X)$.

اثبات:

$$\begin{aligned}
H(X, Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log[p(x)p(y|x)] \\
&= - \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\
&= H(X) + H(Y | X)
\end{aligned}$$

هم‌ارز آن، می‌توان به صورت $\log p(X, Y) = \log p(X) + \log p(Y | X)$ نوشت. حال با گرفتن امید ریاضی طرفین معادله، قضیه اثبات می‌شود. ■

نتیجه ۱-۱: $H(X, Y|Z) = H(X|Z) + H(Y | X, Z)$.

اثبات: اثبات این نتیجه نیز مشابه اثبات قضیه ۱-۱ است. ■

مثال ۳-۱: فرض کنید (X, Y) دارای توزیع توأم زیر باشد:

$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

توزیع حاشیه‌ای برای X برابر با $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ و توزیع حاشیه‌ای برای Y برابر با $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ است، در نتیجه $H(X) = \frac{7}{4}$ بیت و $H(Y) = 2$ بیت. همچنین،

$$\begin{aligned}
H(X | Y) &= \sum_{i=1}^4 p(Y = i) H(X | Y = i) \\
&= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\
&= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 = \frac{11}{8} \text{ bits.} \\
&\quad \blacksquare \text{ به‌طور مشابه، } H(Y | X) = \frac{13}{8} \text{ بیت و } H(X, Y) = \frac{27}{8} \text{ بیت.}
\end{aligned}$$

توجه کنید که $H(Y | X) \neq H(X | Y)$. با این حال، $H(X) - H(X | Y) = H(Y) - H(Y | X)$.

یک ویژگی که در ادامه از آن بهره خواهیم برد.

۵-۱ آنتروپی نسبی و اطلاع متقابل

آنتروپی یک متغیر تصادفی، معیاری از عدم قطعیت آن متغیر تصادفی است، به بیان دیگر، معیاری از میزان اطلاعی است که به طور میانگین برای توصیف یک متغیر تصادفی مورد نیاز است. در این بخش، دو مفهوم مرتبط را معرفی می‌کنیم: آنتروپی نسبی و اطلاع متقابل.

آنتروپی نسبی معیاری برای سنجش فاصله میان دو توزیع است. در آمار، این مفهوم به صورت لگاریتم مورد انتظار نسبت درست‌نمایی ظاهر می‌شود. آنتروپی نسبی $D(p \parallel q)$ معیاری از ناسازگاری این فرضیه است که فرض کنیم توزیع q است، در حالی که توزیع واقعی p است. برای مثال، اگر توزیع واقعی p برای یک متغیر تصادفی را می‌دانستیم، می‌توانستیم کدی با میانگین طول توصیف $H(p)$ بسازیم. اما اگر به جای آن از کدی برای توزیع q استفاده می‌کردیم، به طور میانگین برای توصیف آن متغیر تصادفی به $H(p) + D(p \parallel q)$ بیت نیاز داشتیم.

تعریف ۴-۱: آنتروپی نسبی یا فاصله کولبک-لایبلر میان دو تابع جرم احتمال $p(x)$ و $q(x)$ به صورت زیر تعریف می‌شود:

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)} \blacksquare$$

در تعریف بالا، از قرارداد $0 \log \frac{0}{0} = 0$ استفاده شده است. همچنین (براساس استدلال‌های پیوستگی) قرارداد می‌کنیم که $0 \log \frac{0}{q} = 0$ و $p \log \frac{p}{0} = \infty$. بنابراین، اگر نقطه‌ای مانند $x \in X$ وجود داشته باشد که در آن $p(x) > 0$ و $q(x) = 0$ باشد، آنگاه $D(p \parallel q) = \infty$.

به زودی نشان خواهیم داد که آنتروپی نسبی همیشه نامنفی است و برابر با صفر است اگر و تنها اگر $p = q$. با این حال، این مفهوم یک «فاصله» واقعی بین توزیع‌ها نیست، زیرا مقارن نمی‌باشد و در نامساوی مثلث صدق نمی‌کند. با این وجود، اغلب مفید است که آنتروپی نسبی را به عنوان یک «فاصله» میان توزیع‌ها در نظر بگیریم.

اکنون اطلاع متقابل را معرفی می‌کنیم که معیاری از میزان اطلاعی است که یک متغیر تصادفی درباره متغیر تصادفی دیگر در بر دارد. این مفهوم، نشان‌دهنده کاهش عدم قطعیت در یک متغیر تصادفی، به دلیل داشتن دانش درباره متغیر تصادفی دیگر است.

تعریف ۱-۵: دو متغیر تصادفی X و Y را با تابع جرم احتمال توام $p(x, y)$ و توابع جرم احتمال حاشیه‌ای $p(x)$ و $p(y)$ در نظر بگیرید. اطلاع متقابل $I(X; Y)$ ، آنتروپی نسبی بین توزیع توام و حاصل ضرب $p(x)p(y)$ (یعنی توزیع توام در حالت استقلال دو متغیر تصادفی) است:

$$I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = D(p(x, y) \parallel p(x)p(y))$$

$$= E_{p(x, y)} \log \frac{p(X, Y)}{p(X)p(Y)} \blacksquare$$

مثال ۱-۴: فرض کنید $\mathcal{X} = \{0, 1\}$ و دو توزیع p و q را روی $\mathcal{X}$ در نظر بگیرید. فرض کنید $p(1) = r, p(0) = 1 - r$ همچنین $q(1) = s, q(0) = 1 - s$ در این صورت:

$$D(p \parallel q) = (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s}$$

$$D(q \parallel p) = (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r}$$

اگر $r = s$ در این صورت، $D(p \parallel q) = D(q \parallel p)$. همچنین اگر $r = \frac{1}{2}$ و $s = \frac{1}{4}$ ، داریم:

$$D(p \parallel q) = \frac{1}{2} \log \frac{1/2}{3/4} + \frac{1}{2} \log \frac{1/2}{1/4} = 1 - \frac{1}{2} \log 3 \approx 0.2075 \text{ bit}$$

در حالی که

$$D(q \parallel p) = \frac{3}{4} \log \frac{3/4}{1/2} + \frac{1}{4} \log \frac{1/4}{1/2} = \frac{3}{4} \log 3 - 1 \approx 0.1887 \text{ bit}$$

توجه داشته باشید که به طور کلی $D(p \parallel q) \neq D(q \parallel p)$. $\blacksquare$

۱-۶ رابطه بین آنتروپی و اطلاع متقابل

می‌توان تعریف اطلاع متقابل $I(X; Y)$ را به صورت زیر بازنویسی کرد:

$$\begin{aligned}
I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = \sum_{x, y} p(x, y) \log \frac{p(x | y)}{p(x)} \\
&= - \sum_{x, y} p(x, y) \log p(x) + \sum_{x, y} p(x, y) \log p(x | y) \\
&= - \sum_x p(x) \log p(x) - \left(- \sum_{x, y} p(x, y) \log p(x | y) \right) \\
&= H(X) - H(X | Y)
\end{aligned}$$

بنابراین، اطلاع متقابل یعنی $I(X; Y)$ ، کاهش عدم قطعیت در (X) است که ناشی از دانستن (Y) می باشد. به دلیل تقارن، نتیجه زیر نیز برقرار است:

$$I(X; Y) = H(Y) - H(Y | X)$$

بنابراین، X همان قدر درباره Y می گوید که Y درباره X می گوید. اما می دانیم که $H(X, Y) = H(X) + H(Y | X)$ ، بنابراین:

$$I(X; Y) = H(X) + H(Y) - H(Y | X)$$

در نهایت، خاطرنشان می کنیم که

$$I(X; X) = H(X) - H(X | X) = H(X)$$

بنابراین، اطلاع متقابل یک متغیر تصادفی با خودش برابر با آنتروپی همان متغیر تصادفی است. به همین دلیل است که آنتروپی گاهی «خود اطلاع» نیز نامیده می شود. با گردآوری این نتایج، قضیه زیر را داریم.

قضیه ۱-۲ (اطلاع متقابل و آنتروپی):

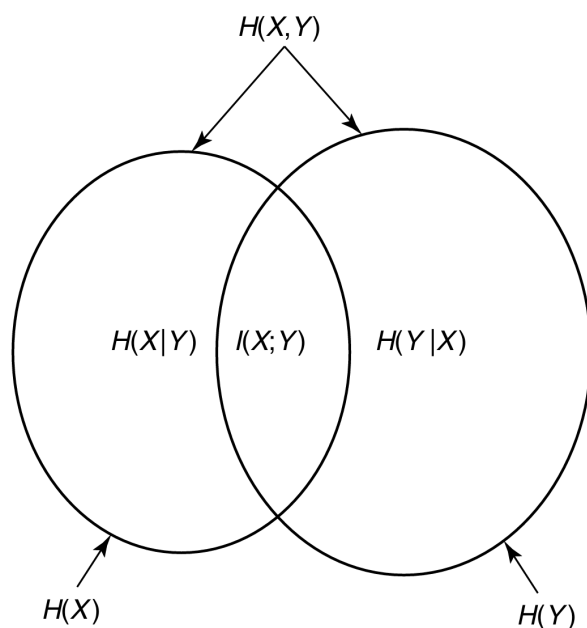
$$I(X; Y) = H(X) - H(X | Y),$$

$$I(X; Y) = H(Y) - H(Y | X),$$

$$I(X; Y) = H(X) + H(Y) - H(Y | X),$$

$$I(X; Y) = I(Y; X),$$

$$I(X; X) = H(X). \blacksquare$$



شکل ۱-۷: ارتباط بین آنترپی و اطلاع متقابل.

رابطه بین $H(X)$ ، $H(Y)$ ، $H(X, Y)$ ، $H(X|Y)$ ، $H(Y|X)$ و $I(X; Y)$ در شکل ۱-۷ نمایش داده شده است. توجه داشته باشید که اطلاع متقابل $I(X; Y)$ ، معادل با اشتراک اطلاع موجود در X با اطلاع موجود در Y است.

۱-۷ قواعد زنجیره‌ای برای آنترپی، آنترپی نسبی و اطلاع متقابل

قضیه ۱-۳ (قاعده زنجیره‌ای برای آنترپی): فرض کنید $X_1, X_2, \dots, X_n$ متغیرهای تصادفی با تابع احتمال توام $p(x_1, x_2, \dots, x_n)$ باشند. در این صورت داریم:

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1)$$

اثبات: با به‌کارگیری مکرر قاعده بسط آنترپی برای دو متغیر، داریم:

$$\begin{aligned} H(X_1, X_2) &= H(X_1) + H(X_2 | X_1) \\ H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3 | X_1) = H(X_1) + H(X_2 | X_1) + H(X_3 | X_2, X_1) \\ &\vdots \\ H(X_1, X_2, \dots, X_n) &= H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_{n-1}, \dots, X_1) \end{aligned}$$

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1)$$

اثبات جایگزین: داریم: $p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i | x_{i-1}, \dots, x_1)$ بنابراین:

$$\begin{aligned} H(X_1, X_2, \dots, X_n) &= - \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log p(x_1, x_2, \dots, x_n) \\ &= - \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log \prod_{i=1}^n p(x_i | x_{i-1}, \dots, x_1) \\ &= - \sum_{x_1, \dots, x_n} \sum_{i=1}^n p(x_1, x_2, \dots, x_n) \log p(x_i | x_{i-1}, \dots, x_1) \\ &= - \sum_{i=1}^n \sum_{x_1, \dots, x_n} p(x_1, x_2, \dots, x_n) \log p(x_i | x_{i-1}, \dots, x_1) \\ &= - \sum_{i=1}^n \sum_{x_1, \dots, x_i} p(x_1, x_2, \dots, x_i) \log p(x_i | x_{i-1}, \dots, x_1) \\ &= \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \blacksquare \end{aligned}$$

اکنون اطلاع متقابل شرطی را به صورت کاهش عدم قطعیت X به دلیل آگاهی از Y ، با فرض معلوم بودن Z ، تعریف می‌کنیم.

تعریف ۶-۱: اطلاع متقابل شرطی متغیرهای تصادفی X و Y با فرض معلوم بودن Z به صورت زیر تعریف می‌شود:

$$I(X; Y | Z) = H(X | Z) - H(X | Y, Z) = E_{p(x,y,z)} \log \frac{p(X, Y | Z)}{p(X | Z)p(Y | Z)} \blacksquare$$

اطلاع متقابل نیز از یک قاعده زنجیره‌ای پیروی می‌کند:

قضیه ۴-۱ (قاعده زنجیره‌ای برای اطلاع):

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, \dots, X_1)$$

اثبات:

$$\begin{aligned}
I(X_1, \dots, X_n; Y) &= H(X_1, \dots, X_n) - H(X_1, \dots, X_n | Y) \\
&= \sum_{i=1}^n [H(X_i | X_{i-1}, \dots, X_1) - H(X_i | X_{i-1}, \dots, X_1, Y)] \\
&= \sum_{i=1}^n I(X_i; Y | X_{i-1}, \dots, X_1) \blacksquare
\end{aligned}$$

در ادامه نسخه شرطی آنتروپی نسبی را تعریف می‌کنیم.

تعریف ۷-۱: برای توابع جرم احتمال توام $p(x, y)$ و $q(x, y)$ ، آنتروپی نسبی شرطی که آن را با $D(p(y | x) || q(y | x))$ نشان می‌دهیم، برابر با میانگین آنتروپی‌های نسبی بین توابع جرم احتمال شرطی $p(y | x)$ و $q(y | x)$ است که بر روی تابع جرم احتمال $p(x)$ میانگین‌گیری شده است. به بیان دقیق‌تر:

$$D(p(y | x) || q(y | x)) = \sum_x p(x) \sum_y p(y | x) \log \frac{p(y | x)}{q(y | x)} = E_{p(x, y)} \log \frac{p(Y | X)}{q(Y | X)} \blacksquare$$

نمادگذاری برای آنتروپی نسبی شرطی صریح نیست زیرا ذکر توزیع $p(x)$ متغیر تصادفی شرط‌ساز را حذف می‌کند. با این حال، معمولاً از متن درک می‌شود.

آنتروپی نسبی بین دو توزیع توام بر روی یک جفت متغیر تصادفی را می‌توان به صورت مجموع یک آنتروپی نسبی و یک آنتروپی نسبی شرطی بسط داد.

قضیه ۵-۱ (قاعده زنجیره‌ای برای آنتروپی نسبی):

$$D(p(x, y) || q(x, y)) = D(p(x) || q(x)) + D(p(y | x) || q(y | x))$$

اثبات:

$$\begin{aligned}
D(p(x, y) || q(x, y)) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} = \sum_x \sum_y p(x, y) \log \frac{p(x)p(y | x)}{q(x)q(y | x)} \\
&= \sum_x \sum_y p(x, y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x, y) \log \frac{p(y | x)}{q(y | x)} \\
&= D(p(x) || q(x)) + D(p(y | x) || q(y | x)). \blacksquare
\end{aligned}$$

۱-۸ نامساوی جنسن و نتایج آن

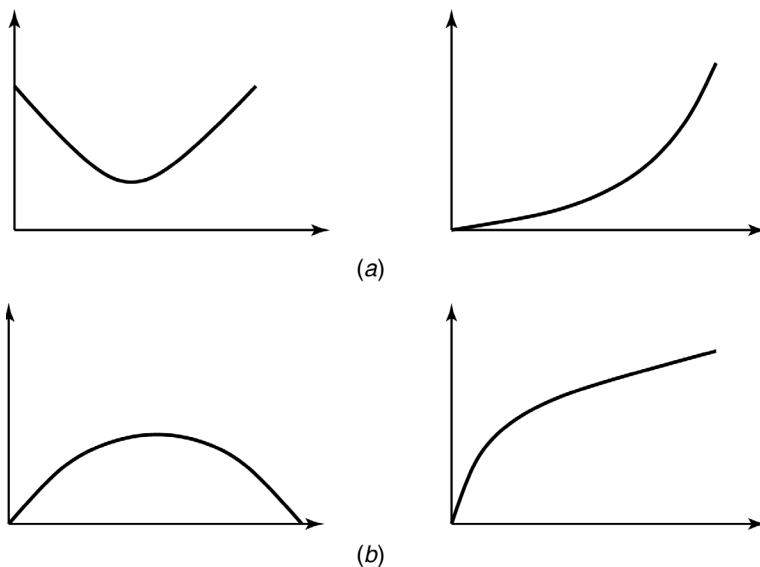
در این بخش برخی خواص ساده کمیت‌های تعریف شده بخش‌های قبل اثبات می‌شود. با خواص توابع محدب شروع می‌کنیم.

تعریف ۱-۸: تابع $f(x)$ روی بازه (a, b) محدب گفته می‌شود اگر برای هر $x_1, x_2 \in (a, b)$ و $0 \leq \lambda \leq 1$ داشته باشیم:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2), \quad (1-1)$$

تابع f را اکیدا محدب گوئیم هرگاه تساوی تنها زمانی برقرار باشد که $\lambda = 0$ یا $\lambda = 1$. ■

تعریف ۱-۹: تابع f مقعر است اگر $-f$ محدب باشد. تابع محدب تابعی است که همواره زیر هر وتر خود قرار می‌گیرد اما تابع مقعر تابعی است که همواره بالای هر وتر خود قرار می‌گیرد. ■



شکل ۱-۸: نمونه‌هایی از (a) توابع محدب و (b) توابع مقعر.

x^2 ، $|x|$ ، e^x ، $x \log x$ (برای $x \geq 0$) و غیره مثال‌هایی از توابع محدب هستند. همچنین $\log x$ و $\sqrt{x}$ برای $x \geq 0$ مثال‌هایی از توابع مقعر هستند. شکل ۱-۸ تعدادی از توابع محدب و مقعر را نشان می‌دهد. توجه داشته باشید که تابع خطی $ax + b$ هم محدب و هم مقعر است. تحدب

زیربنای بسیاری از خواص اساسی کمیت‌های نظریه اطلاع مانند آنتروپی و اطلاع متقابل است. پیش از اثبات برخی از این خواص، نتایج ساده‌ای برای توابع محدب استخراج می‌شود.

قضیه ۶-۱: اگر تابع f دارای مشتق دوم باشد که در یک بازه نامنفی (مثبت) است، آنگاه تابع در آن بازه محدب (اکیدا محدب) است.

اثبات: از بسط سری تیلور تابع حول x_0 استفاده می‌کنیم:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2$$

که در آن x^* بین x_0 و x قرار دارد. طبق فرض، $f''(x^*) \geq 0$ و در نتیجه جمله آخر برای تمام x ها نامنفی است. در رابطه بالا فرض کنید $x_0 = \lambda x_1 + (1 - \lambda)x_2$ و $x = x_1$ در این صورت:

$$f(x_1) \geq f(x_0) + f'(x_0)[(1 - \lambda)(x_1 - x_2)], \quad (1-2)$$

به طور مشابه، با در نظر گرفتن $x = x_2$ ، داریم:

$$f(x_2) \geq f(x_0) + f'(x_0)[\lambda(x_2 - x_1)], \quad (1-3)$$

با ضرب طرفین رابطه (۱-۲) در λ و (۱-۳) در $1 - \lambda$ و جمع کردن آن‌ها، رابطه (۱-۱) به دست می‌آید و اثبات کامل می‌شود. اثبات برای توابع اکیدا محدب نیز به همین روال پیش می‌رود. ■

قضیه بالا بلافاصله به این امکان را می‌دهد تا تحدب اکید x^2 ، e^x و $x \log x$ برای $x \geq 0$ ، و تقعر اکید $\log x$ و $\sqrt{x}$ برای $x \geq 0$ را تأیید کنیم.

فرض کنید E نشان‌دهنده عملگر امید ریاضی باشد. در حالت گسسته $E(X) = \sum_{x \in X} xp(x)$ و در حالت پیوسته $E(X) = \int_{-\infty}^{\infty} xf(x)dx$. نامساوی بعدی یکی از پرکاربردترین نامساوی‌ها در ریاضیات است و اساس بسیاری از نتایج در نظریه اطلاع را تشکیل می‌دهد.

قضیه ۷-۱ (نامساوی جنسن): اگر f تابعی محدب و X متغیری تصادفی باشد، آنگاه:

$$E(f(X)) \geq f(E(X)).$$

علاوه بر این، اگر f اکیدا محدب باشد، تساوی در رابطه بالا مستلزم آن است که با احتمال ۱،

$X = E(X)$ (یعنی X یک مقدار ثابت باشد).

اثبات: این قضیه را برای توزیع‌های گسسته و با استقرای ریاضی روی تعداد نقاط جرم اثبات می‌کنیم. اثبات شرایط برابری در حالتی که f اکیدا محدب است نیز مشابه می‌باشد. برای توزیعی با دو نقطه جرم، نامساوی به صورت زیر تبدیل می‌شود:

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2),$$

که مستقیماً از تعریف توابع محدب به دست می‌آید. فرض کنید قضیه برای توزیع‌هایی با $k-1$ نقطه جرم درست باشد. آنگاه با نوشتن $p_i = \frac{p'_i}{1-p_k}$ برای $i = 1, 2, \dots, k-1$ داریم:

$$\begin{aligned} \sum_{i=1}^k p_i f(x_i) &= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \\ &\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) \\ &\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p'_i x_i\right) \\ &= f\left(\sum_{i=1}^k p_i x_i\right) \end{aligned}$$

که نامساوی اول از فرض استقرا و نامساوی دوم از تعریف تحدب به دست می‌آید. ■

این اثبات را می‌توان با استفاده از استدلال‌های پیوستگی به توزیع‌های پیوسته نیز تعمیم داد. اکنون از این نتایج برای اثبات برخی از خواص آنتروپی و آنتروپی نسبی استفاده می‌کنیم. قضیه زیر از اهمیت اساسی برخوردار است.

قضیه ۸-۱ (نامساوی اطلاع): فرض کنید $p(x)$ و $q(x)$ دو تابع جرم احتمال روی $x \in \mathcal{X}$ باشند. در این صورت داریم:

$$D(p||q) \geq 0$$

و حالت تساوی رخ می‌دهد اگر و تنها اگر برای هر x ، $p(x) = q(x)$.

اثبات: فرض کنید $A = \{x: p(x) > 0\}$ تکیه‌گاه $p(x)$ باشد. آنگاه:

$$\begin{aligned} -D(p \parallel q) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \\ &= \log \sum_{x \in A} q(x) \leq \log \sum_{x \in X} q(x) = \log 1 = 0 \end{aligned}$$

که نامساوی در خط دوم از نامساوی جنسن حاصل می‌شود. از آنجا که $\log t$ یک تابع اکیدا مقعر بر اساس t است، این نامساوی به تساوی تبدیل می‌شود اگر و تنها اگر $q(x)/p(x)$ در همه جا ثابت باشد (یعنی $q(x) = cp(x)$ برای هر x). بنابراین $\sum_{x \in A} p(x) = c \sum_{x \in A} q(x)$. اما نامساوی استفاده شده در خط سوم نیز تنها زمانی به تساوی تبدیل می‌شود که $\sum_{x \in A} q(x) = \sum_{x \in X} q(x) = 1$ ، که مستلزم آن است که $c = 1$. از این رو، اگر $D(p \parallel q) = 0$ و تنها اگر برای هر x ، $p(x) = q(x)$. ■

نتیجه ۱-۲ (نامنفی بودن اطلاع متقابل): برای هر دو متغیر تصادفی X, Y ,

$$I(X; Y) \geq 0,$$

و تساوی رخ می‌دهد اگر و تنها اگر X و Y مستقل باشند.

اثبات: $I(X; Y) = D(p(x, y) \parallel p(x)p(y)) \geq 0$ ، و تساوی برقرار است اگر و تنها اگر $p(x, y) = p(x)p(y)$ (یعنی X و Y مستقل باشند). ■

نتیجه ۱-۳: داریم: $D(p(y|x) \parallel q(y|x)) \geq 0$ و حالت تساوی رخ می‌دهد اگر و تنها اگر برای هر y و همه x هایی که $p(x) > 0$ ، داشته باشیم $p(y|x) = q(y|x)$. ■

نتیجه ۱-۴: $I(X; Y|Z) \geq 0$ و تساوی رخ می‌دهد اگر و تنها اگر X و Y به شرط Z مستقل باشند. ■

اکنون نشان می‌دهیم که توزیع یکنواخت روی دامنه X توزیع حداکثر آنتروپی بر روی این دامنه است. در نتیجه هر متغیر تصادفی با این دامنه، دارای آنتروپی نابیشتر از $\log |X|$ است.

قضیه ۹-۱: $H(X) \leq \log |\mathcal{X}|$ ، که در آن $|\mathcal{X}|$ تعداد عناصر مجموعه $\mathcal{X}$ را نشان می‌دهد، و تساوی رخ می‌دهد اگر و تنها اگر X دارای توزیع یکنواخت روی $\mathcal{X}$ باشد.

اثبات: فرض کنید $u(x) = \frac{1}{|\mathcal{X}|}$ تابع جرم احتمال یکنواخت روی $\mathcal{X}$ و $p(x)$ تابع جرم احتمال X باشد. در این صورت:

$$D(p \parallel u) = \sum_x p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X)$$

حال با توجه به نامنفی بودن آنروپی نسبی (واگرایی کولبک-لایبلر) داریم:

$$0 \leq D(p \parallel u) = \log |\mathcal{X}| - H(X) \blacksquare$$

قضیه ۱۰-۱ (شرطی‌سازی آنروپی را کاهش می‌دهد به طور معادل اطلاع ضرری نمی‌رساند):

$$H(X|Y) \leq H(X)$$

و تساوی رخ می‌دهد اگر و تنها اگر X و Y مستقل باشند.

اثبات: $0 \leq I(X; Y) = H(X) - H(X|Y)$ ■

به طور شهودی، قضیه بیان می‌کند که دانستن مقدار یک متغیر تصادفی مانند Y تنها می‌تواند عدم قطعیت در متغیر تصادفی X را کاهش دهد. توجه داشته باشید که این گزاره فقط به طور متوسط درست است. به خصوص، $H(X|Y=y)$ ممکن است بزرگتر، کوچکتر یا مساوی $H(X)$ باشد، اما به طور متوسط $H(X|Y) = \sum_y p(y)H(X|Y=y) \leq H(X)$. برای مثال، در یک پرونده قضایی، شواهد جدید ممکن است عدم قطعیت را افزایش دهند، اما انتظار می‌رود که شواهد جدید به طور متوسط باعث کاهش عدم قطعیت شوند.

مثال ۵-۱: فرض کنید (X, Y) دارای توزیع توأم زیر باشند:

$Y \backslash X$	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

پس $H(X) = H(1/8, 7/8) = 0.544$ بیت، $H(X | Y = 1) = 0$ بیت، و $H(X | Y = 2) = 1$ بیت. بنابراین

$$H(X | Y) = \frac{3}{4}H(X | Y = 1) + \frac{1}{4}H(X | Y = 2) = 0.25 \text{ bit.}$$

در نتیجه اگر $Y = 2$ مشاهده شود، عدم قطعیت در X افزایش و اگر $Y = 1$ مشاهده شود، عدم قطعیت در X کاهش می‌یابد، اما به‌طور متوسط، شاهد کاهش عدم قطعیت در X هستیم. ■

قضیه ۱۱-۱ (کران استقلال بر آنتروپی): فرض کنید $X_1, X_2, \dots, X_n$ دارای توزیع توام $p(x_1, x_2, \dots, x_n)$ باشند. در این صورت:

$$H(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$$

و تساوی رخ می‌دهد اگر و تنها اگر X_i ها مستقل باشند.

اثبات: با استفاده از قاعده زنجیره‌ای برای آنتروپی داریم:

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \leq \sum_{i=1}^n H(X_i)$$

و تساوی رخ می‌دهد اگر و تنها اگر X_i از $X_{i-1}, \dots, X_1$ (برای هر i) مستقل باشد (یعنی، اگر و تنها اگر X_i ها مستقل باشند). ■

۹-۱ نامساوی مجموع لگاریتمی و کاربردهای آن

اکنون یک نتیجه ساده از تقعر لگاریتم را اثبات می‌کنیم که برای اثبات برخی نتایج تقعر برای آنتروپی استفاده خواهد شد.

قضیه ۱۲-۱ (نامساوی مجموع لگاریتمی): برای اعداد نامنفی $a_1, a_2, \dots, a_n$ و $b_1, b_2, \dots, b_n$ داریم:

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

و تساوی رخ می‌دهد اگر و تنها اگر برای هر i ، a_i/b_i مقداری ثابت باشد.

دوباره از قرارداد $0 \log 0 = 0$ ، $a \log a / 0 = \infty$ اگر $a > 0$ و $0 \log 0 / 0 = 0$ استفاده می‌کنیم. این موارد به راحتی از پیوستگی به دست می‌آیند.

اثبات: بدون از دست دادن کلیت، فرض کنیم $a_i > 0$ و $b_i > 0$. تابع $f(t) = t \log t$ اکیدا محدب است، زیرا برای همه t های مثبت، $f''(t) = 1/t \log e > 0$. بنابراین، با استفاده از نامساوی جنسن، برای $\alpha_i \geq 0$ به طوری که $\sum_{i=1}^n \alpha_i = 1$ داریم:

$$\sum_{i=1}^n \alpha_i f(t_i) \geq f\left(\sum_{i=1}^n \alpha_i t_i\right)$$

با قرار دادن $t_i = a_i/b_i$ و $\alpha_i = b_i / \sum_{j=1}^n b_j$ به دست می‌آوریم:

$$\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_j} \log \sum \frac{a_i}{\sum b_j}$$

که همان نامساوی مجموع لگاریتمی است.

اکنون از نامساوی مجموع لگاریتمی برای اثبات نتایج مختلف تقعر استفاده می‌کنیم. ابتدا قضیه ۸-۱ را مجدداً اثبات می‌کنیم که بیان می‌کند: $D(p \parallel q) \geq 0$ و تساوی رخ می‌دهد اگر و تنها اگر $p(x) = q(x)$. با استفاده از نامساوی مجموع لگاریتمی داریم،

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq \left(\sum_x p(x) \right) \log \frac{\sum_x p(x)}{\sum_x q(x)} \geq 1 \cdot \log \left(\frac{1}{1} \right) = 1 \cdot 0 = 0$$

و تساوی رخ می‌دهد اگر و تنها اگر $p(x)/q(x) = c$ از آنجا که p و q توابع جرم احتمال هستند،

نتیجه می‌گیریم که $c = 1$ ، پس $D(p \parallel q) = 0$ اگر و تنها اگر $p(x) = q(x)$ برای هر x . ■

قضیه ۱۳-۱ (محدب بودن آنترپی نسبی): $D(p \parallel q)$ در زوج (p, q) محدب است. یعنی،

اگر (p_1, q_1) و (p_2, q_2) دو زوج تابع جرم احتمال باشند، آنگاه برای هر $0 \leq \lambda \leq 1$:

$$D(\lambda p_1 + (1 - \lambda)p_2 \parallel \lambda q_1 + (1 - \lambda)q_2) \leq \lambda D(p_1 \parallel q_1) + (1 - \lambda)D(p_2 \parallel q_2)$$

اثبات: نامساوی مجموع لگاریتمی را بر یکی از جملات سمت چپ رابطه فوق اعمال می‌کنیم:

$$\begin{aligned}
& (\lambda p_1(x) + (1 - \lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1 - \lambda)p_2(x)}{\lambda q_1(x) + (1 - \lambda)q_2(x)} \\
& \leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)}
\end{aligned}$$

با جمع کردن این عبارت روی تمام x ها، رابطه مورد نظر به دست می آید. ■

قضیه ۱۴-۱ (مقعر بودن آنتروپی): $H(p)$ یک تابع مقعر نسبت به p است.

اثبات: $H(p) = \log |X| - D(p || u)$ که در آن u دارای توزیع یکنواخت روی مجموعه $|X|$ است. بنابراین مقعر بودن H مستقیماً از محدب بودن D نتیجه می شود.

اثبات جایگزین: فرض کنید X_1 یک متغیر تصادفی با توزیع p_1 باشد که مقادیرش را در مجموعه A می گیرد. فرض کنید X_2 نیز متغیر تصادفی دیگر با توزیع p_2 روی مجموعه A باشد. بالاخره فرض کنید $Z = X^\theta$ و $P(\theta = 1) = \lambda, P(\theta = 2) = 1 - \lambda$. در این صورت توزیع Z برابر با $\lambda p_1 + (1 - \lambda)p_2$ خواهد بود. اکنون با توجه به اینکه شرطی سازی باعث کاهش آنتروپی می شود داریم:

$$H(Z) \geq H(Z|\theta),$$

یا به طور معادل:

$$H(\lambda p_1 + (1 - \lambda)p_2) \geq \lambda H(p_1) + (1 - \lambda)H(p_2),$$

که مقعر بودن آنتروپی را به عنوان تابعی از توزیع اثبات می کند. ■

یکی از پیامدهای مقعر بودن آنتروپی این است که مخلوط کردن دو گاز با آنتروپی برابر، گازی با آنتروپی بالاتر ایجاد می کند.

قضیه ۱۵-۱: فرض کنید $(X, Y) \sim p(x, y) = p(x)p(y|x)$. برای $p(y|x)$ ثابت، اطلاع متقابل یعنی $I(X; Y)$ تابعی مقعر از $p(x)$ است. همچنین برای $p(x)$ ثابت، $I(X; Y)$ تابعی محدب از $p(y|x)$ است.

اثبات: برای اثبات بخش اول، اطلاع متقابل را بسط می دهیم:

$$I(X; Y) = H(Y) - H(Y | X) = H(Y) - \sum_x p(x) H(Y | X = x).$$

حال اگر $p(y | x)$ ثابت باشد، آنگاه $p(y)$ تابعی خطی از $p(x)$ است. بنابراین $H(Y)$ که تابعی مقعر از $p(y)$ است، تابعی مقعر از $p(x)$ نیز خواهد بود. جمله دوم رابطه بالا تابعی خطی از $p(x)$ است. در نتیجه تفاضل نیز تابعی مقعر از $p(x)$ است. برای اثبات بخش دوم، $p(x)$ را ثابت نگه داشته و دو توزیع شرطی متفاوت $p_1(y | x)$ و $p_2(y | x)$ را در نظر می‌گیریم. توزیع‌های توام متناظر عبارتند از $p_1(x, y) = p(x)p_1(y | x)$ و $p_2(x, y) = p(x)p_2(y | x)$ و توزیع‌های حاشیه‌ای متناظر آن‌ها $p_1(y)$ ، $p(x)$ و $p_2(y)$ هستند. یک توزیع شرطی به صورت زیر را در نظر بگیرید:

$$p_\lambda(y|x) \geq \lambda p_1(y|x) + (1 - \lambda)p_2(y|x),$$

که یک ترکیب خطی از $p_1(y | x)$ و $p_2(y | x)$ است و $0 \leq \lambda \leq 1$. توزیع توام متناظر نیز یک ترکیب خطی از توزیع‌های توام متناظر خواهد بود:

$$p_\lambda(x, y) \geq \lambda p_1(x, y) + (1 - \lambda)p_2(x, y),$$

و توزیع Y نیز یک ترکیب خطی به صورت زیر خواهد بود:

$$p_\lambda(y) \geq \lambda p_1(y) + (1 - \lambda)p_2(y).$$

بنابراین اگر $q_\lambda(x, y) = p(x)p_\lambda(y)$ را حاصلضرب توزیع‌های حاشیه‌ای در نظر بگیریم، داریم:

$$q_\lambda(x, y) \geq \lambda q_1(x, y) + (1 - \lambda)q_2(x, y).$$

از آنجا که اطلاع متقابل، در واقع آنتروپی نسبی بین توزیع توام و حاصلضرب توزیع‌های حاشیه‌ای است، خواهیم داشت:

$$I(X; Y) = D(p_\lambda(x, y) || q_\lambda(x, y)),$$

و با توجه به اینکه آنتروپی نسبی $D(p || q)$ تابعی محدب از (p, q) است، نتیجه می‌گیریم که اطلاع متقابل تابعی محدب از توزیع شرطی است. ■

۱-۱۰ نامساوی پردازش داده‌ها

نامساوی پردازش داده‌ها را می‌توان برای نشان دادن اینکه هیچ دستکاری هوشمندانه‌ای در داده‌ها نمی‌تواند استنادات قابل انجام از داده‌ها را بهبود بخشد، به کار برد.

تعریف ۱-۱۰: گوییم متغیرهای تصادفی X ، Y و Z (با ترتیبی که ذکر شده‌اند) تشکیل یک زنجیر مارکوف می‌دهند (و با $X \rightarrow Y \rightarrow Z$ نشان داده می‌شود) اگر توزیع شرطی Z تنها به Y وابسته باشد و از نظر شرطی مستقل از X باشد. به طور خاص، X ، Y و Z یک زنجیر مارکوف تشکیل می‌دهند اگر تابع جرم احتمال توام را بتوان به صورت زیر نوشت:

$$p(x, y, z) = p(x)p(y|x)p(z|y). \blacksquare$$

برخی نتایج ساده به شرح زیر هستند:

۱. $X \rightarrow Y \rightarrow Z$ اگر و تنها اگر X و Z با معلوم بودن مقدار Y مستقل شرطی باشند. مارکوفی بودن استقلال شرطی را نتیجه می‌دهد زیرا

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|z)p(z|y).$$

این مشخصه زنجیرهای مارکوف است که می‌توان آن را برای تعریف میدان‌های مارکوف گسترش داد. میدان‌های مارکوف، فرآیندهای تصادفی n -بعدی هستند که در آن‌ها قسمت داخلی و خارجی با داشتن مقادیر روی مرز، مستقل هستند.

۲. $X \rightarrow Y \rightarrow Z$ دلالت بر این دارد که $Z \rightarrow Y \rightarrow X$. بنابراین، این شرط گاهی اوقات به صورت $X \leftrightarrow Y \leftrightarrow Z$ نیز نوشته می‌شود.

۳. اگر $Z = f(Y)$ ، آنگاه $X \rightarrow Y \rightarrow Z$.

اکنون می‌توان یک قضیه مهم و مفید را اثبات کرد که نشان می‌دهد هیچ پردازشی از Y ، چه قطعی و چه تصادفی، نمی‌تواند اطلاعاتی را که Y درباره X دارد، افزایش دهد.

قضیه ۱-۱۶ (نامساوی پردازش داده‌ها): اگر $X \rightarrow Y \rightarrow Z$ ، آنگاه $I(X; Y) \geq I(X; Z)$.

اثبات: با استفاده از قاعده زنجیره‌ای، می‌توان اطلاع متقابل را به دو روش مختلف بسط داد:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z) = I(X; Y) + I(X; Z|Y).$$

از آنجایی که X و Z با داشتن Y مستقل شرطی هستند، داریم $I(X; Z|Y) = 0$. از آنجایی که $I(X; Y|Z) \geq 0$ ، داریم $I(X; Y) \geq I(X; Z)$ که همان نتیجه مورد نظر است. در این نامساوی حالت تساوی رخ می‌دهد اگر و تنها اگر $I(X; Y|Z) = 0$ (یعنی، $X \rightarrow Z \rightarrow Y$ یک زنجیر مارکوف تشکیل دهد). به طور مشابه، می‌توان ثابت کرد که $I(Y; Z) \geq I(X; Z)$. ■

نتیجه ۵-۱: به طور خاص، اگر $Z = g(Y)$ ، داریم $I(X; Y) \geq I(X; g(Y))$.

اثبات: $X \rightarrow Y \rightarrow g(Y)$ یک زنجیر مارکوف تشکیل می‌دهد. بنابراین، توابع داده Y نمی‌توانند اطلاع مربوط به X را افزایش دهند. ■

نتیجه ۶-۱: اگر $X \rightarrow Y \rightarrow Z$ ، آنگاه $I(X; Y|Z) \leq I(X; Y)$.

اثبات: به دلیل خاصیت مارکوفی بودن، و اینکه $I(X; Z) \geq 0$ داریم: $I(X; Z|Y) = 0$. بنابراین،

$$I(X; Y|Z) \leq I(X; Y). \quad \blacksquare$$

در نتیجه وابستگی X و Y با مشاهده یک متغیر تصادفی «پایین‌دستی» (Z) کاهش می‌یابد (یا بدون تغییر باقی می‌ماند). توجه داشته باشید که زمانی که X ، Y و Z یک زنجیر مارکوف تشکیل ندهند نیز این امکان وجود دارد که $I(X; Y|Z) > I(X; Y)$. به عنوان مثال، فرض کنید X و Y متغیرهای تصادفی دودویی با توزیع یکنواخت و مستقل باشند و $Z = X + Y$. در این صورت $I(X; Y) = 0$ ، اما

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) = H(X|Z) = P(Z=1)H(X|Z=1) \\ &= \frac{1}{2} \text{ bit.} \end{aligned}$$

۱۱-۱ آماره بسنده

این بخش یک توضیح جانبی است که قدرت نامساوی پردازش داده‌ها را در روشن کردن یک ایده مهم در آمار نشان می‌دهد. فرض کنید خانواده‌ای از توابع جرم احتمال $\{f_\theta(x)\}$ داشته باشیم که با

θ اندیس‌گذاری شده‌اند. فرض کنید X نمونه‌ای از توزیعی در این خانواده باشد. فرض کنید $T(X)$ یک آماره (تابعی از نمونه) مانند میانگین نمونه یا واریانس نمونه باشد. آنگاه $\theta \rightarrow X \rightarrow T(X)$ ، و بنابراین طبق نامساوی پردازش داده‌ها، برای هر توزیعی روی θ داریم:

$$I(\theta; T(X)) \leq I(\theta; X).$$

با این حال، اگر تساوی برقرار باشد، هیچ اطلاعی از دست نمی‌رود.

آماره $T(X)$ برای θ بسنده نامیده می‌شود اگر تمام اطلاعات درباره θ که در نمونه (X) موجود است را در خود داشته باشد.

تعریف ۱-۱۱: گوییم تابع $T(X)$ یک آماره بسنده نسبت به خانواده $\{f_\theta(x)\}$ است اگر توزیع X به شرط $T(X)$ مستقل از θ باشد (یعنی، $\theta \rightarrow T(X) \rightarrow X$ یک زنجیر مارکوف تشکیل دهد). ■

این همان شرط تساوی در نامساوی پردازش داده‌ها است.

$$I(\theta; T(X)) = I(\theta; X)$$

بنابراین، آماره‌های بسنده، اطلاع متقابل را حفظ می‌کنند و برعکس.

در ادامه چند مثال از آماره‌های بسنده آمده است:

۱. فرض کنید $X_i \in \{0,1\}$ ، $X_1, X_2, \dots, X_n$ یک دنباله مستقل و هم‌توزیع (i.i.d) از پرتاب‌های یک سکه با پارامتر نامعلوم $\theta = \Pr(X_i = 1)$ باشند. با دانستن مقدار n ، تعداد ۱ها یک آماره بسنده برای θ است. در اینجا $T(X_1, X_2, \dots, X_n) = \sum_{i=1}^n X_i$. در واقع، می‌توان نشان داد که با دانستن مقدار T ، همه دنباله‌هایی که همان تعداد ۱ را دارند، به یک اندازه محتمل‌اند و از پارامتر θ مستقل هستند. به‌طور مشخص،

$$\Pr\left\{(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n) \mid \sum_{i=1}^n X_i = k\right\} = \begin{cases} \frac{1}{\binom{n}{k}} & \sum x_i = k, \\ 0 & \text{o.w} \end{cases}$$

بنابراین، $\theta \rightarrow \sum_{i=1}^n X_i \rightarrow (X_1, X_2, \dots, X_n)$ یک زنجیر مارکوف تشکیل می‌دهد و T یک آماره

بسندۀ برای θ است.

دو مثال بعدی به جای توابع جرم احتمال، شامل چگالی‌های احتمال هستند، اما نظریه همچنان صادق است.

۲. فرض کنید متغیر تصادفی X دارای توزیع نرمال با میانگین θ و واریانس ۱ باشد، یعنی

$$f_{\theta}(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} = \mathcal{N}(\theta, 1),$$

و $X_1, X_2, \dots, X_n$ به‌طور مستقل از این توزیع انتخاب شده باشند، یک آماره بسندۀ برای θ میانگین نمونه $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ است. می‌توان تایید کرد که توزیع شرطی $X_1, X_2, \dots, X_n$ به شرط $\bar{X}_n$ و n به θ بستگی ندارد.

۳. فرض کنید توزیع X یکنواخت باشد یعنی $f_{\theta} = \text{Unif}(\theta, \theta + 1)$. یک آماره بسندۀ برای θ عبارت است از $T(X_1, X_2, \dots, X_n) = (\max\{X_1, X_2, \dots, X_n\}, \min\{X_1, X_2, \dots, X_n\})$. اثبات بسندۀ بودن این آماره با استفاده از تعریف، کمی پیچیده‌تر است، اما باز هم می‌توان نشان داد که توزیع داده‌ها به شرط آماره T ، مستقل از پارامتر است.

آماره بسندۀ مینیمال، آماره بسندۀ ای است که تابعی از تمام آماره‌های بسندۀ دیگر باشد.

تعریف ۱۲-۱: آماره $T(X)$ یک آماره بسندۀ مینیمال نسبت به $\{f_{\theta}(x)\}$ است اگر تابعی از هر آماره بسندۀ دیگر مانند U باشد. ■

با تفسیر این موضوع براساس نامساوی پردازش داده، این بدان معناست که

$$\theta \rightarrow T(X) \rightarrow U(X) \rightarrow X.$$

بنابراین، یک آماره بسندۀ مینیمال، اطلاع مربوط به θ را در نمونه به‌صورت حداکثری فشرده می‌کند. سایر آماره‌های بسندۀ ممکن است شامل اطلاع نامرتب اضافی باشند. برای مثال، برای یک توزیع نرمال با میانگین θ ، زوج تابع‌هایی که میانگین همه نمونه‌های فرد و میانگین همه نمونه‌های زوج را می‌دهند، یک آماره بسندۀ است، اما یک آماره بسندۀ مینیمال نیست. در مثال‌های پیشین، آماره‌های بسندۀ، مینیمال نیز هستند.

۱۲-۱ نامساوی فانو

فرض کنید مقدار متغیر تصادفی Y را می‌دانیم و می‌خواهیم مقدار یک متغیر تصادفی همبسته مانند X را حدس بزنیم. نامساوی فانو، احتمال خطا در حدس زدن متغیر تصادفی X را به آنتروپی شرطی آن $H(X|Y)$ مرتبط می‌کند. می‌دانیم که آنتروپی شرطی یک متغیر تصادفی مانند X به شرط یک متغیر تصادفی دیگر مانند Y صفر است اگر و تنها اگر X تابعی از Y باشد. بنابراین، می‌توان X را از روی Y با احتمال خطای صفر برآورد کرد اگر و تنها اگر $H(X|Y) = 0$.

با گسترش این استدلال، انتظار داریم تنها زمانی قادر به برآورد مقدار X با احتمال خطای پایین باشیم که آنتروپی شرطی $H(X|Y)$ کوچک باشد. نامساوی فانو این ایده را کمی‌سازی می‌کند. فرض کنید می‌خواهیم متغیر تصادفی X با توزیع $p(x)$ را برآورد کنیم. متغیر تصادفی Y را مشاهده می‌کنیم که از طریق توزیع شرطی $p(y|x)$ با X مرتبط است. از روی Y ، تابعی به صورت $g(Y)$ محاسبه می‌کنیم، که در آن $\hat{X}$ یک برآورد از X است و مقادیری را در مجموعه $\hat{\mathcal{X}}$ می‌پذیرد. در اینجا مجموعه $\hat{\mathcal{X}}$ را به برابر بودن با $\mathcal{X}$ محدود نمی‌کنیم و همچنین اجازه می‌دهیم که تابع $g(Y)$ تصادفی باشد. می‌خواهیم احتمالی که در آن $\hat{X} \neq X$ است را کران‌دار کنیم. مشاهده می‌کنیم که $X \rightarrow Y \rightarrow \hat{X}$ یک زنجیر مارکوف تشکیل می‌دهد. احتمال خطا را به صورت زیر تعریف می‌کنیم:

$$P_e = \Pr\{\hat{X} \neq X\}.$$

قضیه ۱۷-۱ (نامساوی فانو): برای هر برآوردگر $\hat{X}$ که رابطه $X \rightarrow Y \rightarrow \hat{X}$ برقرار باشد، داریم:

$$H(P_e) + P_e \log|\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y).$$

که در آن $P_e = \Pr(X \neq \hat{X})$. این نامساوی را می‌توان به گونه‌ای تضعیف کرد که

$$1 + P_e \log|\mathcal{X}| \geq H(X|Y)$$

یا

$$P_e \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|}.$$

از نامساوی فانو در می‌یابیم که $P_e = 0$ نتیجه می‌دهد $H(X|Y) = 0$ ، همانطور که انتظار می‌رود.

اثبات: ابتدا نقش Y را نادیده می‌گیریم و نامساوی اول را اثبات می‌کنیم. سپس از نامساوی پردازش داده برای اثبات شکل سنتی‌تر نامساوی فانو (که توسط نامساوی دوم داده شده است) استفاده خواهیم کرد. یک متغیر تصادفی خطا را به صورت زیر تعریف می‌کنیم:

$$E = \begin{cases} 1 & X \neq \hat{X} \\ 0 & X = \hat{X} \end{cases}$$

حال به منظور بسط $H(E, X | \hat{X})$ به دو روش مختلف، با استفاده از قاعده زنجیره‌ای برای آنروپی‌ها داریم:

$$H(E, X | \hat{X}) = H(X | \hat{X}) + H(E | X, \hat{X}) = H(X | \hat{X}) + 0$$

$$H(E, X | \hat{X}) = H(E | \hat{X}) + H(X | E, \hat{X}) \leq H(P_e) + P_e \log |\mathcal{X}|$$

چون شرطی‌سازی آنروپی را کاهش می‌دهد، پس $H(E | \hat{X}) \leq H(E) = H(P_e)$. اکنون از آنجایی که E تابعی از X و $\hat{X}$ است، آنروپی شرطی $H(E | X, \hat{X})$ برابر با ۰ است. همچنین، از آنجایی که E یک متغیر تصادفی با مقادیر دودویی است، $H(E) = H(P_e)$. برای عبارت باقی‌مانده، یعنی $H(X | E, \hat{X})$ به صورت زیر می‌توان یک کران بالا پیدا کرد:

$$\begin{aligned} H(X | E, \hat{X}) &= \Pr(E = 0)H(X | \hat{X}, E = 0) + \Pr(E = 1)H(X | \hat{X}, E = 1) \\ &\leq (1 - P_e) \cdot 0 + P_e \log |\mathcal{X}|, \end{aligned}$$

زیرا اگر $E = 0$ ، می‌توان نتیجه گرفت که $X = \hat{X}$ ، و اگر $E = 1$ ، می‌توان آنروپی شرطی را با لگاریتم تعداد حالت‌های ممکن محدود کرد. با ترکیب این نتایج، به دست می‌آوریم:

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X | \hat{X}).$$

به کمک نامساوی پردازش داده داریم $I(X; \hat{X}) \leq I(X; Y)$ زیرا $X \rightarrow Y \rightarrow \hat{X}$ یک زنجیر مارکوف است. در نتیجه $H(X | \hat{X}) \geq H(X | Y)$. بنابراین، داریم:

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X | \hat{X}) \geq H(X | Y) \blacksquare$$

نتیجه ۷-۱: برای دو متغیر تصادفی X و Y ، فرض کنید $p = \Pr(X = Y)$. در این صورت:

$$H(p) + p \log |\mathcal{X}| \geq H(X | Y).$$

اثبات: در نامساوی فانو فرض کنید $\hat{X} = Y$. برای دو متغیر تصادفی X و Y ، اگر برآوردگر $g(Y)$ مقادیری در مجموعه $\mathcal{X}$ بگیرد، می‌توان با جایگزینی $\log |\mathcal{X}|$ با $\log(|\mathcal{X}| - 1)$ نامساوی را اندکی تقویت کرد. ■

نتیجه ۸-۱: فرض کنید $P_e = \Pr(X \neq \hat{X})$ و $\hat{X}: \mathcal{Y} \rightarrow \mathcal{X}$ آنگاه:

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X | Y).$$

اثبات: قضیه بدون تغییر ادامه می‌یابد، جز اینکه

$$\begin{aligned} H(X | E, \hat{X}) &= \Pr(E = 0) H(X | \hat{X}, E = 0) + \Pr(E = 1) H(X | \hat{X}, E = 1) \\ &\leq (1 - P_e) \cdot 0 + P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

از آنجایی که با فرض $E = 0$ ، داریم $X = \hat{X}$ ، و با فرض $E = 1$ ، تعداد مقادیر ممکن برای X برابر $|\mathcal{X}| - 1$ است، می‌توان آنتروپی شرطی را با $\log(|\mathcal{X}| - 1)$ که لگاریتم تعداد حالات ممکن است، کران دار کرد. جایگزینی این مقدار، نامساوی قوی‌تری را نتیجه می‌دهد. ■

فرض کنید هیچ اطلاعی از Y وجود نداشته باشد. بنابراین، X باید بدون هیچ اطلاعی حدس زده شود. فرض کنید $X \in \{1, 2, \dots, m\}$ و $p_1 \geq p_2 \geq \dots \geq p_m$. در این حالت، بهترین حدس $\hat{X} = 1$ و احتمال خطای حاصل $P_e = 1 - p_1$ است و نامساوی فانو به صورت زیر تبدیل می‌شود:

$$H(P_e) + P_e \log(m - 1) \geq H(X).$$

در نامساوی بالا تساوی رخ می‌دهد اگر تابع جرم احتمال را به صورت زیر در نظر بگیریم:

$$(p_1, p_2, \dots, p_m) = \left(1 - P_e, \frac{P_e}{m - 1}, \dots, \frac{P_e}{m - 1}\right)$$

بنابراین، نامساوی فانو دقیق است.

اکنون نامساوی جدیدی را که احتمال خطا و آنتروپی را به هم مرتبط می‌کند، معرفی می‌کنیم. فرض کنید X و X' دو متغیر تصادفی مستقل با توزیع یکسان و با آنتروپی $H(X)$ باشند. احتمال $X = X'$ به صورت زیر داده می‌شود:

$$\Pr(X = X') = \sum_x p^2(x).$$

بنابراین نامساوی زیر را داریم:

لم ۳-۱: اگر X و X' مستقل و هم‌توزیع (*i.i.d*) با آنتروپی $H(X)$ باشند داریم:

$$\Pr(X \neq X') \geq 2^{-H(X)},$$

و تساوی رخ می دهد اگر و تنها اگر X دارای توزیع یکنواخت باشد.

اثبات: فرض کنید $X \sim p(x)$. با استفاده از نامساوی جنسن، داریم:

$$2^{E(\log p(X))} \leq E(2^{\log p(X)}),$$

که دلالت بر این دارد که

$$2^{-H(X)} = 2^{\sum p(x) \log p(x)} \leq \sum p(x) 2^{\log p(x)} = \sum p^2(x). \blacksquare$$

نتیجه ۹-۱: فرض کنید X و X' دو متغیر تصادفی مستقل باشند، به طوری که برای $x \in \mathcal{X}$ ، داشته باشیم $X \sim p(x)$ و $X' \sim r(x)$. آنگاه:

$$\Pr(X = X') \geq 2^{-H(p) - D(p||r)},$$

$$\Pr(X = X') \geq 2^{-H(r) - D(r||p)}.$$

اثبات: داریم:

$$\begin{aligned} 2^{-H(p) - D(p||r)} &= 2^{\sum p(x) \log p(x) + \sum p(x) \log \frac{r(x)}{p(x)}} = 2^{\sum p(x) \log r(x)} \leq \sum p(x) 2^{\log r(x)} \\ &= \sum p(x) r(x) = \Pr(X = X') \end{aligned}$$

که نامساوی موجود در خط اول از نامساوی جنسن و تحدب تابع $f(y) = 2^y$ حاصل می شود. ■

فصل دوم: تعمیم اندازه اطلاع کای-دو گسسته و کاربرد در پردازش تصویر

۲-۱ مقدمه

در بسیاری از کاربردهای دنیای واقعی، منابع اطلاعاتی دارای طبیعت گسسته هستند. به عنوان مثال، داده‌های متنی که هر کلمه یک مقدار گسسته از یک واژگان محدود را شامل می‌شود. در مقابل، اندازه‌های اطلاع پیوسته به طور معمول در مواردی استفاده می‌شوند که منبع اطلاع پیوسته است. به عنوان مثال، می‌توان به منبع اطلاعاتی در پردازش سیگنال اشاره کرد (برونی و همکاران، ۲۰۱۳). یکی از مزایای مرتبط با معیارهای اطلاع گسسته این است که می‌توانند، بدون نیاز به دانستن توزیع احتمالی زیربنایی، به سادگی از طریق فراوانی‌های مشاهده شده نمادها یا داده‌های مختلف در یک منبع اطلاعاتی، محاسبه شوند. در حقیقت با استفاده از این ویژگی‌ها، امکان به کارگیری این معیارها در کاربردهای متنوعی مانند فشرده‌سازی داده، تصحیح خطا و یادگیری ماشین فراهم می‌شود (گری، ۲۰۱۱؛ آنجالی و گوپتا، ۲۰۲۵).

در هر دو نوع اندازه اطلاع پیوسته و گسسته، مزایا و محدودیت‌های خاصی وجود دارد که انتخاب بین آن‌ها را وابسته به کاربرد خاص و ویژگی‌های منبع اطلاع موردنظر می‌سازد. در حالت پیوسته، معمولاً به دنبال برآورد مدل احتمال برای متغیرهای تصادفی هستیم، زیرا داده‌ها به صورت یک توزیع پیوسته بیان می‌شوند (صانعی طبس و محتشمی برزادران، ۲۰۱۷؛ احراری و همکاران،

۲۰۱۹). این برآورد ممکن است از طریق برازش یک تابع چگالی احتمال^۱ یا استفاده از روش‌های هسته^۲ انجام شود. در این فرآیند، گسسته‌سازی داده‌ها نیز ممکن است به عنوان یک گام واسطه‌ای به کار رود تا اندازه‌های اطلاع بتوانند روی مجموعه داده‌های پیوسته اعمال شوند. با این حال، این مراحل می‌توانند منجر به خطای برآورد شوند، زیرا هرگونه خطا در برآورد توزیع احتمال اولیه مستقیماً بر محاسبات تأثیر خواهد گذاشت (وانگ و همکاران، ۲۰۲۵). از سوی دیگر، در کاربردهایی که با داده‌های گسسته سر و کار داریم، نیاز به برآورد توزیع احتمال کمتر احساس می‌شود، زیرا می‌توان مستقیماً از فراوانی‌های مشاهده شده در داده‌های گسسته برای محاسبه اندازه‌های اطلاع استفاده کرد (وانگ و فان، ۲۰۲۴). به عنوان مثال، در مسائل مرتبط با پردازش زبان طبیعی، دسته‌بندی متون و داده‌های ژنتیکی، داده‌ها به طور ذاتی ماهیت گسسته دارند و معیارهای اطلاع گسسته می‌توانند بدون نیاز به برآورد توزیع احتمالاتی پیوسته، مستقیماً از توزیع تجربی استفاده شوند. از لحاظ نظری، کار کردن با مدل‌های پیوسته به ویژه زمانی که معیارهای اطلاع به صورت انتگرالی تعریف شده‌اند و تابع درون انتگرال را می‌توان با استفاده از توابع ریاضی پیوسته یا کمیت‌هایی مرتبط با مشتقات آن‌ها توصیف کرد گاهی ساده‌تر است. برای مثال، در اطلاع فشر پیوسته، از مشتق تابع چگالی احتمال نسبت به پارامترهای آن استفاده می‌شود، درحالی که در حالت گسسته، به جای مشتق معمولی، از مشتق تفاضلی در فضای گسسته بهره می‌گیریم که می‌تواند پیچیدگی بیشتری به همراه داشته باشد. بنابراین، بسته به نوع داده‌ها و اهداف مورد نظر، انتخاب میان معیارهای اطلاع پیوسته و گسسته تأثیر قابل توجهی بر کارایی و دقت تحلیل‌های اطلاعی خواهد داشت (شانون، ۱۹۴۸).

از مهم‌ترین اندازه‌های اطلاع گسسته در نظریه اطلاع می‌توان به آنتروپی شانون^۳، اکستروپی^۴، شاخص جینی-سیمپسون^۵، اطلاع فشر گسسته^۶، و تابع مولد اطلاع گسسته^۷ اشاره کرد. فرض کنید X یک متغیر تصادفی گسسته با تابع جرم احتمال $P = (p_1, \dots, p_n)$ باشد. تعریف اندازه‌های

¹ Probability Density Function

² Kernel

³ Shannon entropy

⁴ Extropy

⁵ Gini-Simpson index

⁶ Discrete Fisher information

⁷ Discrete information generating function

اطلاع بالا به ترتیب عبارتند از:

آنتروپی شانون:

$$H(\mathbf{P}) = - \sum_{i=1}^n p_i \log p_i.$$

معیار اکستروپی:

$$\bar{H}(\mathbf{P}) = - \sum_{i=1}^n (1 - p_i) \log(1 - p_i).$$

شاخص جینی-سیمپسون:

$$GS(\mathbf{P}) = \sum_{i=1}^n p_i(1 - p_i) = 1 - \sum_{i=1}^n p_i^2.$$

اطلاع فیشر گسسته:

$$J(\mathbf{P}) = \sum_{i=1}^n \frac{(p_{i+1} - p_i)^2}{p_i}, \quad p_{n+1} = 0.$$

تابع مولد اطلاع گسسته:

$$G_\alpha(\mathbf{P}) = \sum_{i=1}^n p_i^\alpha, \quad \alpha > 0.$$

برای جزئیات بیشتر به شانون (۱۹۴۸)، خوارزمی و بالاکریشن (۲۰۲۱ و ۲۰۲۳a)، خوارزمی و همکاران (۲۰۲۳a)، خوارزمی و همکاران (۲۰۲۳c) و خوارزمی و بالاکریشن (۲۰۲۴) مراجعه کنید. این معیارها نه تنها در حوزه نظری، بلکه در بسیاری از کاربردهای عملی، از جمله یادگیری ماشین و پردازش تصویر، نقش مهمی ایفا می‌کنند. در یادگیری ماشین، این اندازه‌ها برای انتخاب ویژگی، بهینه‌سازی مدل‌های دسته‌بندی و سنجش شباهت داده‌ها مورد استفاده قرار می‌گیرند. به عنوان مثال، آنتروپی شانون در درخت‌های تصمیم برای اندازه‌گیری عدم قطعیت یک متغیر تصادفی به کار می‌رود و شاخص جینی در الگوریتم‌هایی مانند جنگل تصادفی برای تعیین خلوص

گره‌ها در فرایند دسته‌بندی استفاده می‌شود (ناوارو و همکاران، ۲۰۲۳). علاوه بر این، در پردازش تصویر، از این معیارها برای بهبود کیفیت تصویر، تشخیص الگو و فشرده‌سازی داده‌های تصویری استفاده می‌شود. به طور خاص، اطلاع فیشر گسسته در ارزیابی میزان تغییرات ساختاری تصاویر و بهینه‌سازی الگوریتم‌های تشخیص لبه مورد استفاده قرار می‌گیرد. همچنین، تابع مولد اطلاع گسسته می‌تواند در تحلیل و توصیف ویژگی‌های پیچیده تصاویر، به ویژه در ارزیابی شباهت و کیفیت تصویر، به کار برده شود. در مجموع، اندازه‌های اطلاع گسسته ابزارهای قدرتمندی برای تحلیل داده‌های گسسته محسوب می‌شوند و در کاربردهای مختلف مورد توجه قرار گرفته‌اند (سوراتی و همکاران، ۲۰۱۹).

فرض کنید X و Y دو متغیر تصادفی با توابع چگالی پیوسته به ترتیب f و g روی تکیه گاه مشترک $\mathcal{X}$ باشند. در این صورت اندازه‌های واگرایی کای-دو (نلسن و ناک، ۲۰۱۳) و کای-دو تعمیم یافته (باسو و همکاران، ۱۹۹۸) به ترتیب به صورت زیر تعریف می‌شوند:

$$\chi^2(f: g) = \int_{\mathcal{X}} \frac{[f(x) - g(x)]^2}{f(x)} dx, \quad (2-1)$$

$$\chi_{\alpha}^2(f: g) = \frac{1 + \alpha}{2} \int_{\mathcal{X}} \frac{[f(x) - g(x)]^2}{f^{1-\alpha}(x)} dx. \quad (2-2)$$

خوارزمی و علیزاده (۱۳۹۲) نتایجی براساس معیار کای-دو و ارتباط آن با اندازه واگرایی جنسن-فیشر ارائه دادند. اخیراً خوارزمی و بالاکریشن (۲۰۲۴) با در نظر گرفتن اندازه‌های اطلاع کای-دو در رابطه (۲-۱) و کای دو تعمیم یافته در رابطه (۲-۲) یک اندازه اطلاع جدید به صورت زیر معرفی کردند:

$$D_{\alpha}^{\psi}(f: g) = \frac{1 + \alpha}{2} \int_{\mathcal{X}} \frac{[f(x) - g(x)]^2}{\psi^{1-\alpha}(x)} dx, \quad (2-3)$$

که در آن f ، g و ψ توابع چگالی احتمال هستند. این کمیت اندازه اطلاع χ_{α}^2 -نسبی نامیده شد. آن‌ها نشان دادند که این معیار جدید اطلاع، کمیت‌های اطلاع از قبیل کای-دو (χ^2) و کای-دو تعمیم یافته (χ_{α}^2) بالاکریشن-سانگهاوی را به عنوان حالات خاص شامل می‌شود. علاوه بر این، نویسندگان اندازه واگرایی از نوع جنسن شامل جنسن- χ_{α}^2 را بر اساس معیار (۲-۱) به صورت

$$J_{\alpha}^{\psi}(f_1, f_2; \mathbf{p}) = p\chi_{\alpha}^2(\psi: f_1) + (1-p)\chi_{\alpha}^2(\psi: f_2) - \chi_{\alpha}^2(\psi: pf_1 + (1-p)f_2), \quad (2-4)$$

و اندازه واگرایی جنسن- $\chi_{\alpha}^2(p, w)$ را به عنوان یک تعمیم از آن معرفی کردند. در حقیقت، معیارهای واگرایی جنسن- χ_{α}^2 و جنسن- $\chi_{\alpha}^2(p, w)$ پیشنهاد شده تعمیم‌هایی از $D_{\alpha}^{\psi}(f: g)$ هستند که براساس خاصیت تحدب این معیارها ارائه شده است. خوارزمی و بالاکریشنان (۲۰۲۴) نشان دادند که واگرایی جنسن- χ_{α}^2 را می‌توان به صورت آمیخته‌ای از معیارهای واگرایی نسبی- χ_{α}^2 به صورت زیر نشان داد:

$$J_{\alpha}^{\psi}(f_1, f_2; \mathbf{p}) = pD_{\alpha}^{\psi}(f_1: f_T) + (1-p)D_{\alpha}^{\psi}(f_2: f_T),$$

که در آن $0 < p < 1$. نکته مهم و قابل توجه این است که تمام این مطالعات در حالت پیوسته و مرتبط با معیار کای-دو در رابطه (۲-۱) بوده‌اند، در حالی که مطالعه این معیار در حالت گسسته به طور مستقل بررسی نشده است. یکی از اهداف اصلی پروژه حاضر، پرداختن به این موضوع است. هدف اصلی در این پروژه، معرفی نسخه‌های گسسته از معیارهای اطلاع در رابطه‌های (۲-۲)، (۲-۳) و (۲-۴)، همچنین تعمیم آن‌ها براساس خاصیت تحدب و مطالعه روابط بین آن‌ها است. به ویژه اینکه علاقه‌مند هستیم تا این کمیت‌های اطلاع را برای برخی از مدل‌های معروف و پرکاربرد در شاخه‌های نظریه کدگذاری و ترمودینامیک از قبیل توزیع‌های اسکورت و اسکورت تعمیم یافته (برچر، ۲۰۰۹ و ۲۰۱۳) نیز مورد مطالعه و بررسی قرار دهیم. در پایان هم دو کاربرد در زمینه پردازش تصویر برای یکی از این معیارهای اطلاع معرفی شده ارائه می‌شود.

بر این اساس، ساختار این فصل به صورت زیر خواهد بود: ابتدا در بخش دوم با در نظر گرفتن اندازه واگرایی کای-دو در رابطه (۲-۲)، اندازه‌های واگرایی کای-دو تعمیم یافته گسسته و اندازه واگرایی نسبی- χ_{α}^2 بین دو تابع جرم احتمال معرفی می‌شوند. سپس این اندازه‌های جدید معرفی شده را برای برخی از توزیع‌های شناخته شده و پرکاربرد اسکورت و اسکورت تعمیم یافته بدست می‌آوریم. علاوه بر این یک مثال مرتبط با توزیع برنولی ارائه و نمودارهای سه بعدی نیز بررسی می‌شوند. در بخش سوم، با در نظر گرفتن اندازه اطلاع χ_{α}^2 گسسته، اندازه واگرایی جنسن- χ_{α}^2 را معرفی می‌کنیم. سپس نشان می‌دهیم که این اندازه اطلاع جدید را می‌توان به صورت یک ترکیب محدب از اندازه‌های واگرایی χ_{α}^2 گسسته نسبی بیان کرد. علاوه بر این نشان داده می‌شود که مشتق

مرتبه دوم کمیت اندازه واگرایی جنسن- χ^2_α دارای رابطه مستقیم با χ^2_α گسسته نسبی است. در پایان این بخش، یک تعمیم از تعریف اندازه واگرایی جنسن- χ^2_α برای حالت $n + 1$ متغیر تصادفی ارائه می‌شود. معرفی معیار واگرایی جنسن- $\chi^2_\alpha(w, \lambda)$ و مطالعه برخی از ویژگی‌های مهم آن در بخش چهارم ارائه می‌شود. در بخش پنجم، به منظور بررسی عملکرد معیار χ^2_α گسسته، از آن برای ارزیابی کیفیت تصویر استفاده می‌کنیم. علاوه بر این، کارایی این اندازه اطلاع را با کمیت نسبت سیگنال به نویز^۱ (PSNR) که یک معیار استاندارد برای ارزیابی کیفیت تصویر است، مقایسه و تحلیل می‌کنیم. بخش ششم نیز به نتیجه‌گیری اختصاص داده می‌شود. لازم به ذکر است که منبع اصلی این فصل خوارزمی و شیرازی (۲۰۲۶) است.

۲-۲ اندازه اطلاع χ^2_α گسسته و تعمیم آن

تعریف ۲-۱: فرض کنید $P = (p_1, \dots, p_n)$ و $Q = (q_1, \dots, q_n)$ دو تابع جرم احتمال باشند. در این صورت اندازه اطلاع χ^2_α گسسته بین توابع جرم احتمال P و Q به صورت زیر تعریف می‌شود:

$$\chi^2_\alpha(P, Q) = \frac{1 + \alpha}{2} \sum_{i=1}^n \frac{[p_i - q_i]^2}{p_i^{1-\alpha}}. \quad (۲-۵)$$

تعریف ۲-۲: فرض کنید $P = (p_1, \dots, p_n)$ ، $Q = (q_1, \dots, q_n)$ و $R = (r_1, \dots, r_n)$ تابع جرم احتمال باشند. در این صورت اندازه اطلاع χ^2_α نسبی گسسته بین توابع جرم احتمال P و Q نسبت به تابع جرم احتمال R به صورت زیر تعریف می‌شود:

$$D^R_\alpha(P, Q) = \frac{1 + \alpha}{2} \sum_{i=1}^n \frac{[p_i - q_i]^2}{r_i^{1-\alpha}}. \quad (۲-۶)$$

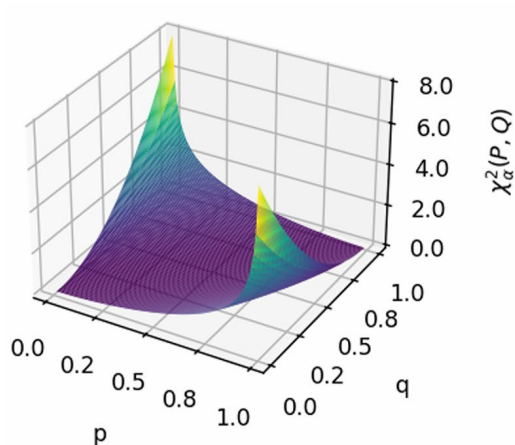
با توجه به تعریف اندازه‌های اطلاع در رابطه‌های (۲-۵) و (۲-۶)، دیده می‌شود که با قراردادن $R = P$ خواهیم داشت: $D^R_\alpha(P, Q) = \chi^2_\alpha(P, Q)$.

مثال ۲-۱: فرض کنید X و Y دو متغیر تصادفی برنولی به ترتیب با پارامترهای p و q باشند. بردارهای احتمال متناظر با این دو متغیر تصادفی به ترتیب به صورت $P = (p, 1 - p)$ و $Q = (q, 1 - q)$ هستند. اکنون با توجه به تعریف اندازه واگرایی $\chi^2_\alpha(P, Q)$ گسسته در رابطه (۲-۵)

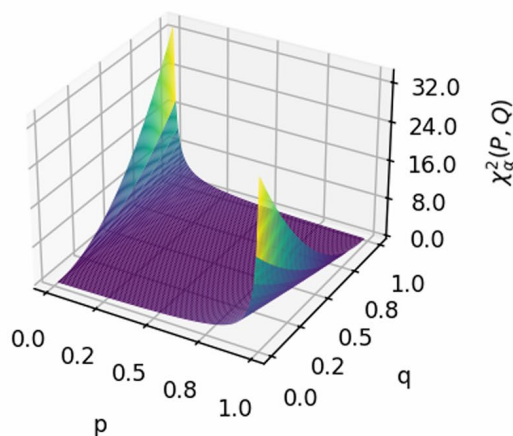
^۱ Peak Signal-to-Noise Ratio

خواهیم داشت:

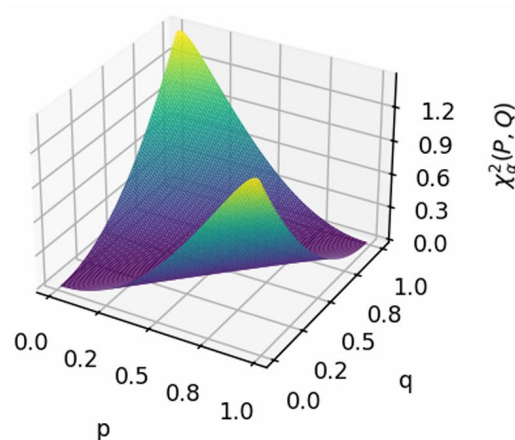
$$\chi_{\alpha}^2(P, Q) = \frac{1+\alpha}{2} \sum_{i=1}^n \frac{[p_i - q_i]^2}{r_i^{1-\alpha}} = \frac{1+\alpha}{2} (p - q)^2 \left[\frac{1}{p^{1-\alpha}} + \frac{1}{(1-p)^{1-\alpha}} \right].$$



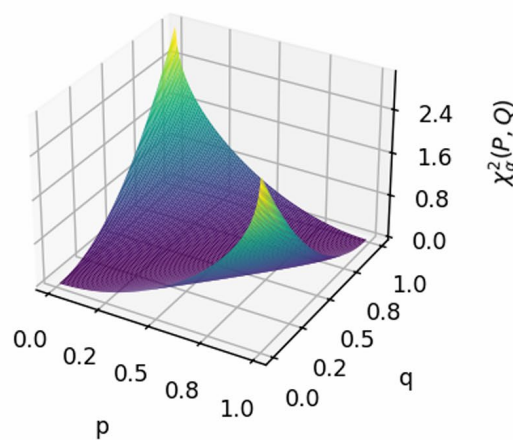
(ب)



(الف)



(د)



(ج)

شکل ۱-۲: نمودار سه بعدی برای الف. $\alpha = 0.1$ ، ب. $\alpha = 0.5$ ، ج. $\alpha = 0.8$ و د. $\alpha = 1.25$ اندازه واگرایی $\chi_{\alpha}^2(P, Q)$ برحسب احتمال‌های p و q .

یک نمودار سه بعدی از $\chi_{\alpha}^2(P, Q)$ برای مقادیر انتخاب شده α ، یعنی (0.1, 0.5, 0.8, 1.25) در شکل ۱-۲ رسم شده است. همانطور که ملاحظه می‌شود با افزایش α از 0.1 به 1.25، مقدار $\chi_{\alpha}^2(P, Q)$ کاهش یافته و نمودارها صاف‌تر می‌شوند، که بیانگر کاهش حساسیت معیار نسبت به

اختلاف بین p و q است. این واگرایی در حالت $p = q$ به حداقل می‌رسد و با افزایش فاصله بین آن‌ها افزایش می‌یابد، که خاصیت جداسازی آن را نشان می‌دهد. همچنین، برای مقادیر کوچک α ، قله‌های نمودار تیزتر و مقادیر واگرایی در نواحی دور از خط قطری بیشتر هستند. این رفتار نشان می‌دهد که انتخاب α نقش مهمی در میزان تمایز قائل شدن بین دو توزیع دارد و بسته به کاربرد، می‌توان α را برای کنترل حساسیت تنظیم کرد. فرض کنید که متغیر تصادفی Z نیز به صورت دارای توزیع برنولی با پارامتر r باشد. اکنون با در نظر گرفتن بردارهای احتمال متناظر با متغیرهای X, Y و Z به صورت $P = (p, 1 - p)$ ، $Q = (q, 1 - q)$ و $R = (r, 1 - r)$ و با توجه به تعریف اندازه واگرایی $\chi^2_\alpha(P, Q)$ گسسته نسبی در رابطه (۶-۲)، خواهیم داشت:

$$D_\alpha^R(P, Q) = \frac{1 + \alpha}{2} (p - q)^2 \left[\frac{1}{r^{1-\alpha}} + \frac{1}{(1-r)^{1-\alpha}} \right]. \blacksquare$$

فرض کنید $P = (p_1, \dots, p_n)$ و $Q = (q_1, \dots, q_n)$ دو تابع جرم احتمال باشند. در این صورت تابع جرم احتمال اسکورت مربوط به P و تابع جرم احتمال اسکورت تعمیم‌یافته P و Q برای $\eta > 0$ به ترتیب به صورت زیر تعریف می‌شوند (برچر، ۲۰۰۹؛ برچر، ۲۰۱۳؛ خوارزمی و بالاکریشنان، ۲۰۲۳a؛ خوارزمی و همکاران، ۲۰۲۳b):

$$P_\eta = \left(\frac{p_1^\eta}{\sum_{i=1}^n p_i^\eta}, \dots, \frac{p_n^\eta}{\sum_{i=1}^n p_i^\eta} \right), \quad (2-7)$$

$$S_\eta = \left(\frac{p_1^\eta q_1^{1-\eta}}{\sum_{i=1}^n p_i^\eta q_i^{1-\eta}}, \dots, \frac{p_n^\eta q_n^{1-\eta}}{\sum_{i=1}^n p_i^\eta q_i^{1-\eta}} \right). \quad (2-8)$$

قضیه ۲-۱: فرض کنید $P = (p, 1 - p)$ و $Q = (q, 1 - q)$ دو تابع جرم احتمال و P_η تابع جرم احتمال اسکورت متناظر با P باشد. در این صورت برای $0 < \alpha, \eta < 1$ و $R = P_\eta$ داریم:

$$D_\alpha^R(P; Q) = \frac{1 + \alpha}{1 + \beta} G_\eta^{1-\alpha}(P) \chi_\beta^2(P; Q),$$

که در آن $G_\eta(P) = \sum_{i=1}^n p_i^\eta$ و $\beta = 1 - \eta(1 - \alpha)$ است.

اثبات: با فرض اینکه $R = P_\eta$ داریم:

$$D_{\alpha}^R(\mathbf{P}; \mathbf{Q}) = \frac{1+\alpha}{2} \left(\sum_{i=1}^n p_i^{\eta} \right)^{1-\alpha} \sum_{i=1}^n \frac{[p_i - q_i]^2}{p_i^{\eta-\eta\alpha}} = \frac{1+\alpha}{2} G_{\eta}^{1-\alpha}(\mathbf{P}) \sum_{i=1}^n \frac{[p_i - q_i]^2}{p_i^{\eta-\eta\alpha}}$$

$$= \frac{1+\alpha}{1+\beta} G_{\eta}^{1-\alpha}(\mathbf{P}) \chi_{\beta}^2(\mathbf{P}; \mathbf{Q}),$$

که در آن $\beta = 1 - \eta(1 - \alpha)$. همچنین برای $R = S_{\eta}$ داریم:

$$2D_{\alpha=0}^R(\mathbf{P}; \mathbf{Q}) = \sum_{i=1}^n \frac{[p_i - q_i]^2}{S_{\eta}} = \left(\sum_{i=1}^n p_i^{\eta} q_i^{1-\eta} \right) \sum_{i=1}^n \frac{[p_i - q_i]^2}{p_i^{\eta} q_i^{1-\eta}}$$

$$= \Gamma_{\eta}(\mathbf{P}; \mathbf{Q}) \sum_{i=1}^n \frac{[p_i - q_i]^2}{p_i^{\eta} q_i^{1-\eta}}, \quad (2-9)$$

که در آن $\Gamma_{\eta}(\mathbf{P}; \mathbf{Q}) = \sum_{i=1}^n p_i^{\eta} q_i^{1-\eta}$ تابع مولد اطلاع نسبی بین توابع جرم احتمال $\mathbf{P}$ و $\mathbf{Q}$ است. ■

قضیه ۲-۲: یک کران پایین برای کمیت $D_{\alpha=0}^R(\mathbf{P}; \mathbf{Q})$ در رابطه (۲-۹) به صورت زیر است:

$$D_{\alpha=0}^R(\mathbf{P}; \mathbf{Q}) \geq \frac{\Gamma_{\eta}(\mathbf{P}; \mathbf{Q})}{2(1-\eta)^2} \chi^2(\mathbf{P}_T; \mathbf{P}),$$

که در آن $\mathbf{P}_T = \eta\mathbf{P} + (1-\eta)\mathbf{Q}$ جرم احتمال آمیخته و $\chi^2(\cdot; \cdot)$ اندازه واگرایی کای-دو است.

اثبات: از تعریف $D_{\alpha}^R(\mathbf{P}; \mathbf{Q})$ و فرض اینکه

$$\mathbf{R} = \mathbf{S}_{\eta} = \left(\frac{p_1^{\eta} q_1^{1-\eta}}{\sum_{i=1}^n p_i^{\eta} q_i^{1-\eta}}, \dots, \frac{p_n^{\eta} q_n^{1-\eta}}{\sum_{i=1}^n p_i^{\eta} q_i^{1-\eta}} \right), \quad 0 \leq \eta \leq 1,$$

و با استفاده از نامساوی میانگین هندسی-میانگین حسابی بین توابع جرم احتمال $\mathbf{P}$ و $\mathbf{Q}$ که به صورت $\mathbf{P}^{\eta} \mathbf{Q}^{1-\eta} \leq \eta\mathbf{P} + (1-\eta)\mathbf{Q}$ است (استیل، ۲۰۰۴)، و توجه به اینکه $\mathbf{Q} - \mathbf{P} = \frac{\mathbf{P}_T - \mathbf{P}}{1-\eta}$ خواهیم داشت:

$$2D_{\alpha=0}^R(\mathbf{P}; \mathbf{Q}) = \left(\sum_{i=1}^n p_i^{\eta} q_i^{1-\eta} \right) \sum_{i=1}^n \frac{[p_i - q_i]^2}{p_i^{\eta} q_i^{1-\eta}} = \Gamma_{\eta}(\mathbf{P}; \mathbf{Q}) \sum_{i=1}^n \frac{[p_i - q_i]^2}{p_i^{\eta} q_i^{1-\eta}}$$

$$\geq \Gamma_{\eta}(\mathbf{P}; \mathbf{Q}) \sum_{i=1}^n \frac{[p_i - q_i]^2}{\eta p_i + (1-\eta) q_i} = \frac{\Gamma_{\eta}(\mathbf{P}; \mathbf{Q})}{(1-\eta)^2} \chi^2(\mathbf{P}_T; \mathbf{P}).$$

با توجه به اثبات قضیه ۲-۲ مشاهده می‌شود که تساوی در کران پایین ارائه شده زمانی برقرار است که $\eta = 0$ یا $\eta = 1$. ■

۲-۳ اندازه اطلاع جنسن- χ_α^2 گسسته

تعریف ۲-۳: فرض کنید $P = (p_1, \dots, p_n)$ ، $Q = (q_1, \dots, q_n)$ و $R = (r_1, \dots, r_n)$ تابع جرم احتمال باشند. در این صورت اندازه اطلاع χ_α^2 نسبی گسسته بین توابع جرم احتمال P و Q نسبت به تابع جرم احتمال R به صورت زیر تعریف می‌شود:

$$J_\alpha^R(P, Q, w) = w\chi_\alpha^2(R: P) + (1-w)\chi_\alpha^2(R: Q) - \chi_\alpha^2(R: wP + (1-w)Q) \quad (۲-۱۰) \quad \blacksquare$$

لم ۲-۱: اندازه جنسن- χ_α^2 نامنفی است.

اثبات: با توجه به این که تابع $\phi(x) = x^2$ محدب است، با استفاده از نامساوی جنسن خواهیم داشت:

$$\begin{aligned} & \frac{2}{1+\alpha} J_\alpha^R(P, Q, w) \\ &= w \sum_{i=1}^n \frac{(p_i - r_i)^2}{r_i^{1-\alpha}} + (1-w) \sum_{i=1}^n \frac{(q_i - r_i)^2}{r_i^{1-\alpha}} \\ & \quad - \sum_{i=1}^n \frac{(wp_i + (1-w)q_i - r_i)^2}{r_i^{1-\alpha}} \\ &= w \sum_{i=1}^n \frac{p_i^2}{r_i^{1-\alpha}} + (1-w) \sum_{i=1}^n \frac{q_i^2}{r_i^{1-\alpha}} - \sum_{i=1}^n \frac{(wp_i + (1-w)q_i)^2}{r_i^{1-\alpha}} \geq 0. \quad \blacksquare \end{aligned}$$

قضیه ۲-۳: فرض کنید متغیرهای تصادفی گسسته X_1 و X_2 دارای تابع جرم احتمال به ترتیب P و Q باشند. در این صورت اندازه اطلاع $J_\alpha^R(P, Q, w)$ را می‌توان به صورت آمیخته‌ای از اندازه اطلاع χ_α^2 گسسته نسبی، یعنی D_α^R به صورت زیر بیان کرد:

$$J_\alpha^R(P, Q, w) = wD_\alpha^R(P: P_T) + (1-w)D_\alpha^R(Q: P_T),$$

که در آن $D_\alpha^R(P: P_T)$ اندازه اطلاع داده شده در رابطه (۲-۶) و P_T تابع جرم احتمال آمیخته است.

اثبات: با توجه به تعریف $P_T = wP + (1-w)Q$ داریم:

$$\frac{2}{1+\alpha} J_{\alpha}^R(P, Q, w) = w \sum_{i=1}^n \frac{p_i^2}{r_i^{1-\alpha}} + (1-w) \sum_{i=1}^n \frac{q_i^2}{r_i^{1-\alpha}} - \sum_{i=1}^n \frac{(wp_i + (1-w)q_i)^2}{r_i^{1-\alpha}}.$$

از طرف دیگر، با قرار دادن $k = wD_{\alpha}^R(P: P_T) + (1-w)D_{\alpha}^R(Q: P_T)$ خواهیم داشت:

$$\begin{aligned} \frac{2}{1+\alpha} k &= w \sum_{i=1}^n \frac{(p_i - p_i^T)^2}{r_i^{1-\alpha}} + (1-w) \sum_{i=1}^n \frac{(q_i - p_i^T)^2}{r_i^{1-\alpha}} \\ &= w \sum_{i=1}^n \frac{p_i^2}{r_i^{1-\alpha}} - 2w \sum_{i=1}^n \frac{p_i p_i^T}{r_i^{1-\alpha}} + w \sum_{i=1}^n \frac{(p_i^T)^2}{r_i^{1-\alpha}} + (1-w) \sum_{i=1}^n \frac{q_i^2}{r_i^{1-\alpha}} \\ &\quad - 2(1-w) \sum_{i=1}^n \frac{q_i p_i^T}{r_i^{1-\alpha}} + (1-w) \sum_{i=1}^n \frac{(p_i^T)^2}{r_i^{1-\alpha}} \\ &= w \sum_{i=1}^n \frac{p_i^2}{r_i^{1-\alpha}} + (1-w) \sum_{i=1}^n \frac{q_i^2}{r_i^{1-\alpha}} - 2 \sum_{i=1}^n \frac{(p_i^T)^2}{r_i^{1-\alpha}} + \sum_{i=1}^n \frac{(p_i^T)^2}{r_i^{1-\alpha}}. \blacksquare \end{aligned}$$

قضیه ۴-۲: فرض کنید متغیرهای تصادفی گسسته X_1 و X_2 دارای تابع جرم احتمال به ترتیب P و Q باشند. در این صورت داریم:

$$-\frac{1}{2} \frac{\partial^2}{\partial w^2} J_{\alpha}^R(P, Q, w) = D_{\alpha}^R(P: Q).$$

اثبات: با استفاده از تعریف $J_{\alpha}^R(P, Q, w)$ در رابطه (۱۰-۲) داریم:

$$\begin{aligned} &-\frac{1}{2} \frac{\partial^2}{\partial w^2} J_{\alpha}^R(P, Q, w) \\ &= -\frac{1+\alpha}{4} \frac{\partial}{\partial w} \left(\sum_{i=1}^n \frac{(p_i - r_i)^2}{r_i^{1-\alpha}} - \sum_{i=1}^n \frac{(q_i - r_i)^2}{r_i^{1-\alpha}} \right) \\ &\quad + \frac{1+\alpha}{2} \frac{\partial}{\partial w} \left(\sum_{i=1}^n (p_i - q_i) \frac{wp_i + (1-w)q_i - r_i}{r_i^{1-\alpha}} \right) \\ &= \frac{1+\alpha}{2} \sum_{i=1}^n \frac{(p_i - q_i)^2}{r_i^{1-\alpha}}. \blacksquare \end{aligned}$$

اکنون تعریف اندازه اطلاع جنسن- χ_{α}^2 گسسته در رابطه (۱۰-۲) به حالت $n+1$ متغیر تصادفی تعمیم داده می‌شود.

تعریف ۲-۴: فرض کنید $Y, X_1, \dots, X_n$ متغیرهای تصادفی گسسته با توابع جرم احتمال به ترتیب $R, P_1, \dots, P_n$ باشند که $P_i = (p_{i1}, \dots, p_{in})$ و $R = (r_1, \dots, r_n)$ و $w_1, \dots, w_n$ اعداد حقیقی غیرمنفی باشند به طوری که $\sum_{i=1}^n w_i = 1$. در این صورت اندازه اطلاع جنسن- χ^2_α عبارت است از:

$$J_\alpha^R(P_1, \dots, P_n; w) = \sum_{i=1}^n w_i \chi_\alpha^2(R; P_i) - \chi_\alpha^2\left(R; \sum_{i=1}^n w_i P_i\right). \blacksquare$$

قضیه ۲-۵: اندازه اطلاع $J_\alpha^R(P_1, \dots, P_n; w)$ را می‌توان به صورت آمیخته‌ای از اندازه اطلاع نسبی D_α^R به صورت زیر بیان کرد:

$$J_\alpha^R(P_1, \dots, P_n; w) = \sum_{i=1}^n w_i D_\alpha^R\left(P_i; \sum_{i=1}^n w_i P_i\right), \quad (2-11)$$

اثبات: اثبات مشابه با قضیه ۲-۳ انجام می‌شود. $\blacksquare$

۲-۴ اندازه اطلاع جنسن- $\chi^2_\alpha(w, \lambda)$

فرض کنید f_1 و f_2 دو تابع چگالی باشند. اندازه واگرایی جنسن-شانون (w, λ) به صورت زیر تعریف می‌شود:

$$JS_{(w, \lambda)}(f_1, f_2) = H((1 - \bar{s})f_1 + \bar{s}f_2) - wH((1 - \lambda)f_1 + \lambda f_2) - (1 - w)H(\lambda f_1 + (1 - \lambda)f_2),$$

که در آن $\bar{s} = w\lambda + (1 - w)(1 - \lambda)$ و H آنترپی شانون تابع چگالی f است که به صورت زیر می‌باشد (نېسن و ناک، ۲۰۱۳؛ مبلورن و همکاران، ۲۰۲۲):

$$H(f) = - \int_{-\infty}^{\infty} f(x) \log f(x) dx.$$

تعریف ۲-۵: فرض کنید Y, X_1, X_2 متغیرهای تصادفی گسسته با توابع جرم احتمال به ترتیب R, Q, P باشند. در این صورت اندازه اطلاع جنسن- $\chi^2_\alpha(w, \lambda)$ به صورت زیر تعریف می‌شود:

$$J_\alpha^R(P, Q; w, \lambda) = \lambda \chi_\alpha^2(R; (1 - w)P + wQ) + (1 - \lambda) \chi_\alpha^2(R; wP + (1 - w)Q) - \chi_\alpha^2(R; (1 - \bar{w})P + \bar{w}Q),$$

که در آن $\bar{w} = w\lambda + (1-w)(1-\lambda)$. ■

قضیه ۶-۲: فرض کنید متغیرهای تصادفی گسسته X_1 و X_2 دارای تابع جرم احتمال به ترتیب P و Q باشند. در این صورت اندازه اطلاع $J_\alpha^R(P, Q; w, \lambda)$ را می‌توان به صورت آمیخته‌ای از اندازه اطلاع χ_α^2 گسسته نسبی در رابطه (۶-۲) به صورت

$$J_\alpha^R(P, Q; w, \lambda) = \lambda D_\alpha^R((1-w)P + wQ; P_{\bar{w}}^T) + (1-\lambda) D_\alpha^R(wP + (1-w)Q; P_{\bar{w}}^T),$$

نوشت که در آن $P_{\bar{w}}^T = (1-\bar{w})P + \bar{w}Q$ تابع جرم احتمال آمیخته است.

اثبات: با توجه به تعریف

$$J_\alpha^R(P, Q; w, \lambda) = \lambda \chi_\alpha^2(R; (1-w)P + wQ) + (1-\lambda) \chi_\alpha^2(R; wP + (1-w)Q) - \chi_\alpha^2(R; (1-\bar{w})P + \bar{w}Q),$$

با در نظر گرفتن $P_{\bar{w}}^T = (1-\bar{w})P + \bar{w}Q$ خواهیم داشت:

$$\begin{aligned} \frac{2}{1+\alpha} J_\alpha^R(P, Q; w, \lambda) &= \lambda \sum_{i=1}^n \frac{((1-w)p_i + wq_i)^2}{r_i^{1-\alpha}} + (1-\lambda) \sum_{i=1}^n \frac{(wp_i - (1-w)q_i)^2}{r_i^{1-\alpha}} \\ &\quad - \sum_{i=1}^n \frac{((1-\bar{w})p_i + \bar{w}q_i)^2}{r_i^{1-\alpha}}. \end{aligned}$$

از طرف دیگر با قرار دادن مقدار k در این رابطه

$$\frac{1+\alpha}{2} k = \lambda D_\alpha^R((1-w)P + wQ; P_{\bar{w}}^T) + (1-\lambda) D_\alpha^R(wP + (1-w)Q; P_{\bar{w}}^T),$$

و در نظر گرفتن مولفه‌های بردار احتمال به صورت $P_{\bar{w}}^T = (p_1^T, \dots, p_n^T)$ خواهیم داشت:

$$p_i^T = (1-\bar{w})p_i + \bar{w}q_i = \lambda((1-w)p_i + wq_i) + (1-\lambda)(wp_i + (1-w)q_i),$$

به طوری که

$$\begin{aligned}
k &= \lambda \sum_{i=1}^n \frac{((1-w)p_i + wq_i - p_i^T)^2}{r_i^{1-\alpha}} + (1-\lambda) \sum_{i=1}^n \frac{(wp_i - (1-w)q_i - p_i^T)^2}{r_i^{1-\alpha}} \\
&= \lambda \sum_{i=1}^n \frac{((1-w)p_i + wq_i)^2}{r_i^{1-\alpha}} - 2\lambda \sum_{i=1}^n \frac{((1-w)p_i + wq_i)p_i^T}{r_i^{1-\alpha}} \\
&\quad + \sum_{i=1}^n \frac{(p_i^T)^2}{r_i^{1-\alpha}} + (1-\lambda) \sum_{i=1}^n \frac{(wp_i - (1-w)q_i)^2}{r_i^{1-\alpha}} \\
&\quad - 2(1-\lambda) \sum_{i=1}^n \frac{(wp_i - (1-w)q_i)^2 p_i^T}{r_i^{1-\alpha}} \\
&= \lambda \sum_{i=1}^n \frac{((1-w)p_i + wq_i)^2}{r_i^{1-\alpha}} + (1-\lambda) \sum_{i=1}^n \frac{(wp_i - (1-w)q_i)^2}{r_i^{1-\alpha}} \\
&\quad + \sum_{i=1}^n \frac{(p_i^T)^2}{r_i^{1-\alpha}},
\end{aligned}$$

و اثبات قضیه کامل می شود. ■

نتیجه ۱-۲: از ارتباط بین $J_\alpha^R(P, Q; w, \lambda)$ و $J_\alpha^R(P, Q; \lambda)$ به رابطه زیر می رسم:

$$J_\alpha^R(P, Q; w, \lambda) = J_\alpha^R((1-w)P + wQ: wP + (1-w)Q; \lambda).$$

قضیه ۷-۲: فرض کنید متغیرهای تصادفی گسسته X_1 و X_2 دارای تابع جرم احتمال به ترتیب P و Q باشند. در این صورت:

الف.

$$J_\alpha^R(P, Q; w, \lambda) = J_\alpha^R((1-w)P + wQ: wP + (1-w)Q; \lambda).$$

ب.

$$-\frac{1}{2} \frac{\partial^2}{\partial \lambda^2} J_\alpha^R(P, Q; w, \lambda) = D_\alpha^R((1-w)P + wQ: wP + (1-w)Q; \lambda).$$

ج.

$$-\frac{1}{2} \frac{\partial^2}{\partial w^2} J_\alpha^R(P, Q; w, \lambda) = D_\alpha^R((1-\lambda)P + \lambda Q: \lambda P + (1-\lambda)Q) - D_\alpha^R(P: Q).$$

اثبات: اثبات رابطه (الف) با توجه به تعاریف کمیت‌های $J_{\alpha}^R(P, Q; \lambda)$ و $J_{\alpha}^R(P, Q; w, \lambda)$ به راحتی نتیجه می‌شود. برای اثبات قسمت (ب) با توجه به قضیه ۲-۴ و نتیجه ۲-۱، داریم:

$$-\frac{1}{2} \frac{\partial^2}{\partial \lambda^2} J_{\alpha}^R(P, Q; w, \lambda) = -\frac{1}{2} \frac{\partial^2}{\partial \lambda^2} J_{\alpha}^R((1-w)P + wQ; wP + (1-w)Q; \lambda) \\ = D_{\alpha}^R((1-w)P + wQ; wP + (1-w)Q),$$

که اثبات قسمت ب را کامل می‌کند. همچنین با توجه به نتیجه ۲-۱ داریم:

$$-\frac{1}{2} \frac{\partial^2}{\partial w^2} J_{\alpha}^R(P, Q; w, \lambda) = -\frac{1}{2} \frac{\partial^2}{\partial w^2} J_{\alpha}^R((1-w)P + wQ; wP + (1-w)Q; \lambda) \\ = D_{\alpha}^R((1-\lambda)P + \lambda Q; \lambda P + (1-\lambda)Q) - D_{\alpha}^R(P; Q),$$

که اثبات قسمت ج را کامل می‌کند. ■

۲-۵ کاربرد در پردازش تصویر

در این بخش، دو کاربرد برای اندازه واگرایی $\chi_{\alpha}^2(P; Q)$ در زمینه ارزیابی کیفیت تصویر ارائه می‌شود. علاوه بر این، نتایج را با عملکرد نسبت سیگنال به نویز پیک^۱ (PSNR) که یک معیار شناخته شده در ارزیابی کیفیت تصاویر است، مقایسه می‌شود (گنزالز، ۲۰۰۹)

۲-۵-۱ تصویر دریاچه

شکل ۲-۲ تصویر اصلی یک دریاچه را نشان می‌دهد که شامل یک شبکه 512×512 پیکسل است. هر پیکسل در این تصویر دارای سطح خاکستری در محدوده $[0, 1]$ است، به طوری که مقدار صفر نمایانگر رنگ سیاه و مقدار یک نمایانگر رنگ سفید است. این تصاویر به طور معمول برای تحلیل‌های پردازش تصویر و ارزیابی تکنیک‌های مختلف اصلاح تصویر مورد استفاده قرار می‌گیرند. تصویر اصلی با نام X همراه با سه نسخه تعدیل شده آن نمایش داده شده است. هر یک از این نسخه‌ها تغییراتی در ویژگی‌های تصویر اصلی اعمال می‌کنند تا تأثیرات مختلف فرآیندهای پردازش تصویر را نشان دهد و به صورت زیر تعریف می‌شوند:

$$1. \text{ نسخه } Y = 0.4X + 0.3$$

^۱ Peak Signal-to-Noise Ratio

۲. نسخه $Z = 0.5X$

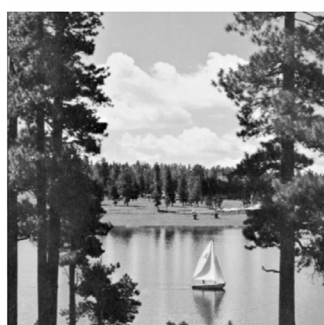
۱. نسخه $W = 5\sqrt{X}$

برای پیاده سازی این تغییرات، از بسته OpenCV در پایتون استفاده شده است که ابزارهای جامعی برای برای پردازش تصویر ارائه می دهد (کیلر و برادسکی، ۲۰۱۶؛ وود و گونزالس، ۲۰۰۱). هیستوگرام های استخراج شده برای تصاویر X, Y, Z, W در شکل ۳-۲ ترسیم شده اند. از شکل ۲-۲ و شکل ۳-۲ می توان مشاهده کرد که بالاترین درجه شباهت ابتدا مربوط به Y و سپس مربوط به W است. در حالی که Z بالاترین درجه واگرایی را از تصویر اصلی یعنی X دارد. در نهایت، با محاسبه اندازه واگرایی $\chi^2_\alpha(P:Q)$ و نسبت سیگنال به نویز پیک (PSNR) بین هیستوگرام های نرمال شده تصویر اصلی دریاچه و هر یک از تصاویر تعدیل شده، برای مقادیر انتخابی $\alpha = 0.5, 0.8, 1, 2, 3$ مقایسه می کنیم. مقدار PSNR یک معیار رایج برای ارزیابی کیفیت تصویر بازسازی شده یا فشرده شده در مقایسه با تصویر اصلی است. این معیار به واحد دسی بل (dB) بیان می شود و به صورت $PSNR = 20 \log_{10} \left(\frac{MAX}{\sqrt{MSE}} \right)$ محاسبه می شود که در آن MSE خطای میانگین مربعات است که تفاوت میانگین مربعات بین مقادیر پیکسل های تصویر اصلی I و تصویر تعدیل شده I^* را اندازه گیری می کند و به صورت زیر تعریف می شود:

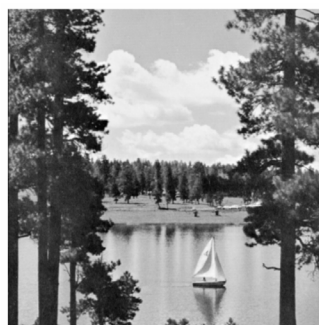
$$MSE = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n [I(i, j) - I^*(i, j)]^2$$

علاوه بر این MAX حداکثر مقدار پیکسل ممکن تصویر است (به عنوان مثال، ۲۵۵ برای تصاویر بیتی) و m و n ابعاد تصاویر هستند. PSNR بالاتر نشان دهنده شباهت بیشتر بین تصویر تعدیل شده و تصویر اصلی است. مقادیر بالای ۳۰ دسی بل معمولاً کیفیت خوبی برای فشرده سازی را نشان می دهند. برعکس، PSNR پایین تر نشان دهنده تفاوت های قابل توجه بین تصویر اصلی و تصویر تعدیل شده است، به طوری که مقادیر زیر ۲۰ دسی بل اغلب کیفیت پایین را نشان می دهند. مقادیر PSNR بین ۲۰ دسی بل و ۳۰ دسی بل بسته به زمینه و کاربرد خاص می توانند قابل قبول باشند (اسکیسیگولا و فیشر، ۱۹۹۵؛ گنزالز، ۲۰۰۹). در جدول ۱-۲، اندازه واگرایی کای-دو تعمیم یافته گسسته و کمیت PSNR بین تصویر اصلی و تصاویر نویزی Y, Z, W ارائه شده اند. اندازه

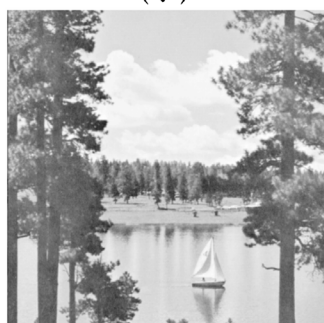
واگرایی کای-دو تعمیم یافته برای α های مختلف ($\alpha = 0.5, 0.8, 1, 2, 3$) محاسبه شده است. به نتایج این جدول دیده می شود که با افزایش α ، مقدار واگرایی به طور چشم گیری کاهش می یابد، و این کاهش برای تمامی تصاویر نویزی قابل مشاهده است (شکل ۴-۲). این امر نشان دهنده همگرایی بیشتر توزیع پیکسل ها با افزایش α است. از سوی دیگر، کمیت PSNR که معیاری برای اندازه گیری کیفیت بازسازی تصویر است، نشان می دهد که تصویر نویزی Y با PSNR برابر با $28.0274dB$ بالاترین کیفیت را در مقایسه با تصاویر Z, W دارد. تصاویر Z و W نیز به ترتیب PSNR برابر با $27.1208dB$ و $27.8731dB$ دارند. این نتایج نشان می دهند که تصویر Y کمترین میزان نویز را دارد و کیفیت بازسازی بالاتری نسبت به تصاویر Z, W دارد. بنابراین PSNR و واگرایی کای-دو تعمیم یافته هر دو به وضوح نشان می دهند که تصویر Y نزدیک ترین تصویر به تصویر اصلی است و تصویر Z دارای بیشترین انحراف است. اندازه واگرایی کای-دو تعمیم یافته عملکردی مشابه با PSNR دارد و می تواند به عنوان یک معیار جایگزین به کار برود.



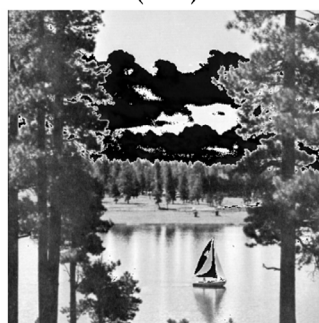
(ب)



(الف)

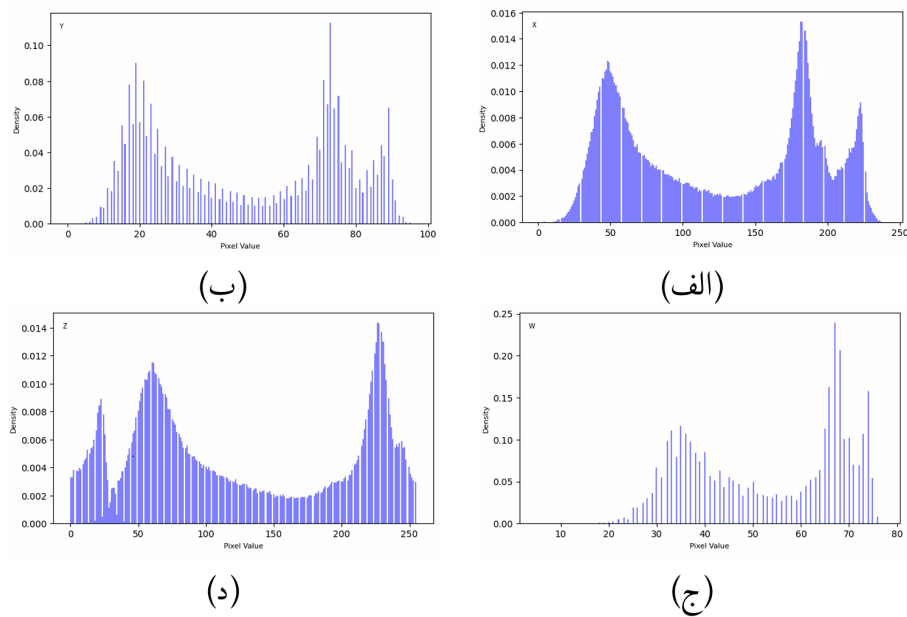


(د)



(ج)

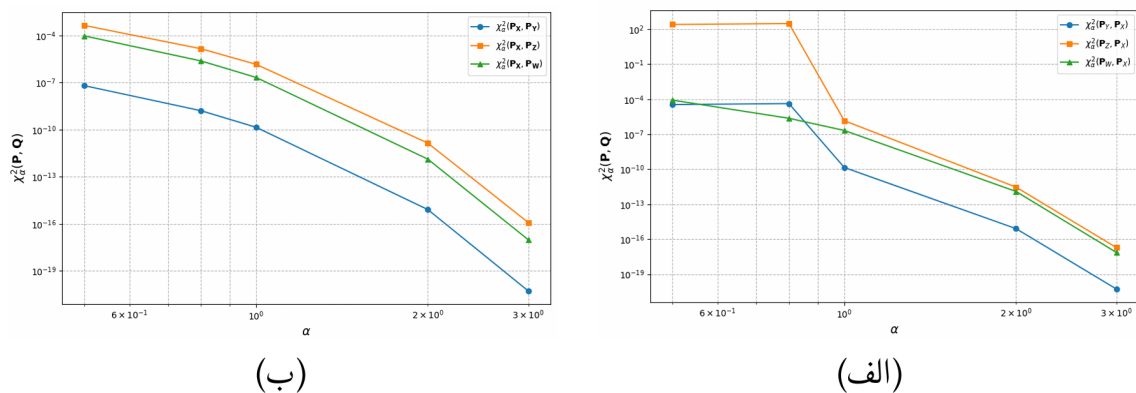
شکل ۲-۲: تصویر اصلی دریاچه یعنی X (الف) به همراه سه نسخه تعدیل شده آن: تصویر Y (ب)، تصویر Z (ج) و تصویر W (د).



شکل ۳-۲: هیستوگرام تصویر اصلی دریاچه یعنی X (الف) به همراه سه نسخه تعدیل شده آن: تصویر Y (ب)، تصویر Z (ج) و تصویر W (د).

جدول ۲-۱: اندازه واگرایی کای دو تعمیم یافته گسسته و کمیت PSNR بین تصویر اصلی و حالت‌های نویزی.

α					χ^2_α
۳	۲	۱	۰.۸	۰.۵	
5.33×10^{-21}	8.12×10^{-16}	1.42×10^{-10}	1.62×10^{-9}	6.34×10^{-8}	(P_X, P_Y)
1.21×10^{-16}	1.39×10^{-11}	1.45×10^{-6}	1.43×10^{-5}	4.33×10^{-4}	(P_X, P_Z)
9.70×10^{-18}	1.32×10^{-12}	2.12×10^{-7}	2.40×10^{-6}	9.34×10^{-5}	(P_X, P_W)
5.36×10^{-21}	8.12×10^{-16}	1.42×10^{-10}	4.23×10^{-5}	3.53×10^{-5}	(P_Y, P_X)
1.90×10^{-17}	2.95×10^{-12}	1.45×10^{-6}	3.08×10^{-2}	2.57×10^{-2}	(P_Z, P_X)
7.32×10^{-18}	1.25×10^{-12}	2.12×10^{-7}	2.33×10^{-6}	8.40×10^{-5}	(P_W, P_X)
کمیت PSNR بین تصویر اصلی و تصویر نویزی (dB)					
۲۸۰۲۷۴					$X \leftrightarrow Y$
۲۷۱۲۰۸					$X \leftrightarrow Z$
۲۷۸۷۳۱					$X \leftrightarrow W$



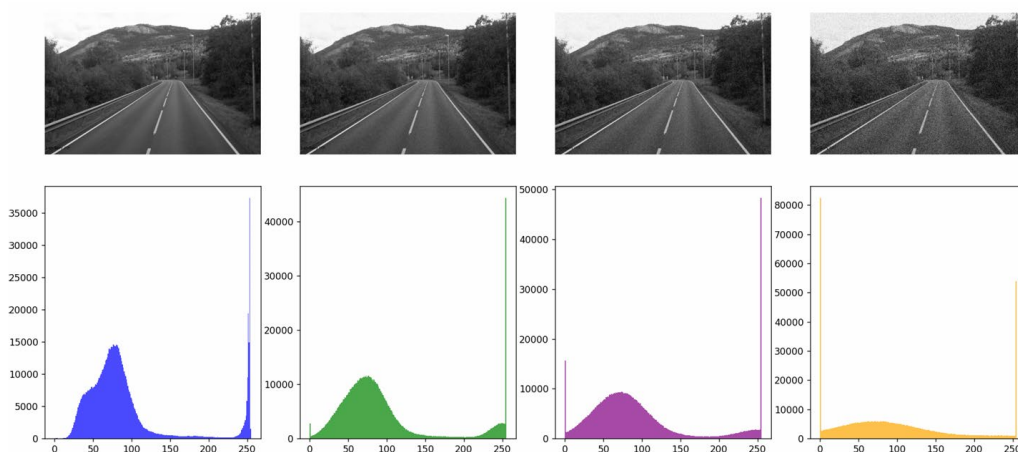
شکل ۴-۲: نمودار مقدار واگرایی کای-دو تعمیم یافته گسسته بین تصاویر اصلی (الف) و نویزی (ب).

۲-۵-۲ تصویر جاده‌ای در نروژ

در این زیربخش، یک کاربرد مرتبط با ارزیابی کیفیت تصویر با استفاده از معیار واگرایی اندازه χ^2_α برای تصویر «جاده ای در نروژ» ارائه شده است. در شکل ۲-۵ تصویر اصلی خاکستری با ابعاد 1346×882 و تصاویر نویزی ایجاد شده با نویز گاوسی در سطوح مختلف $\sigma = 15, 25, 50$ به همراه هیستوگرام‌های متناظر هر تصویر نمایش داده شده‌اند. برای یک تصویر خاکستری $I(x, y)$ با ابعاد $M \times N$ ، نسخه نویزی تصویر، که با $I_{noisy}(x, y)$ نمایش داده می‌شود، از طریق افزودن نویز گاوسی ϵ به هر پیکسل تصویر اصلی به صورت $I_{noisy}(x, y) = I_{original}(x, y) + \epsilon$ ایجاد می‌شود، که در آن $\epsilon \sim N(0, \sigma^2)$. مشابه زیر بخش قبل، اندازه واگرایی $\chi^2_\alpha(P: Q)$ و کمیت PSNR بین هیستوگرام‌های نرمال شده تصویر اصلی و هر یک از تصاویر تعدیل شده، برای مقادیر انتخابی $\alpha = 0.1, 1, 2, 4, 20$ محاسبه و مقایسه شده است. نتایج این تحلیل در جدول ۲-۲ ارائه شده است. مطابق نتایج این جدول، مشاهده می‌شود که با افزایش سطح نویز (σ)، مقادیر واگرایی χ^2_α در حالت‌های مختلف α به طور کلی افزایش یافته است، به ویژه برای مقادیر کوچک‌تر α مانند $\alpha = 0.1, 1$. این موضوع بیانگر آن است که اندازه واگرایی این موضوع بیانگر آن است که اندازه واگرایی χ^2_α بین تصویر اصلی و تصاویر نویزی با شدت نویز بیشتر افزایش یافته است. در مقابل، برای مقادیر بزرگ‌تر α مانند $\alpha = 20$ ، مشاهده می‌شود که مقدار واگرایی تقریباً ثابت و بسیار کوچک باقی مانده است که نشان می‌دهد واگرایی برای α های بزرگ حساسیت کمتری به تغییرات نویز دارد. از سوی دیگر، مشاهده می‌شود که مقادیر PSNR با افزایش شدت نویز کاهش یافته‌اند، که این رفتار نیز مطابق انتظار بوده و افت کیفیت تصویر نویزی نسبت به تصویر اصلی را نشان می‌دهد. بنابراین، روند تخریب کیفیت تصویر هم از دید واگرایی χ^2_α و هم از دید PSNR تایید شده است.

جدول ۲-۲: مقادیر واگرایی کای دو تعمیم یافته χ^2_α و کمیت PSNR بین تصویر اصلی و حالت‌های نویزی.

(dB) PSNR	α					σ
	۲۰	۴	۲	۱	۰/۱	
۲۹/۴۰۲۹	3.17×10^{-29}	2.19×10^{-7}	1.04×10^{-4}	4.32×10^{-3}	5.82×10^{-1}	۱۵
۲۸/۶۱۶۲	3.32×10^{-29}	2.34×10^{-7}	1.20×10^{-4}	5.74×10^{-3}	2.50×10^0	۲۵
۲۸/۱۳۳۶	3.47×10^{-29}	2.54×10^{-7}	1.57×10^{-4}	1.58×10^{-2}	4.76×10^1	۵۰



شکل ۵-۲: ردیف بالا: تصویر اصلی (سمت چپ) و تصاویر حاصل از افزودن نویز گاوسی با انحراف معیارهای ۱۵، ۲۵ و ۵۰، ردیف پایین: هیستوگرام های خاکستری متناظر با هر تصویر

۲-۶ بحث و نتیجه‌گیری

در این پروژه چندین اندازه جدید اطلاع از نوع کای-دو در زمینه توزیع‌های گسسته معرفی و مورد بررسی قرار گرفته‌اند. ابتدا، اندازه‌های واگرایی کای-دو تعمیم یافته گسسته و واگرایی نسبی χ^2_α معرفی و برای توزیع‌های شناخته شده و پرکاربرد در شاخه‌های نظریه کدگذاری و ترمودینامیک از قبیل توزیع‌های اسکورت و اسکورت تعمیم یافته مطالعه شده و یک مثال در این زمینه در مورد توزیع برنولی بررسی شده است. همچنین، اندازه اطلاع جنسن χ^2_α معرفی شده و نشان داده شده است که این اندازه را می‌توان به صورت ترکیب محدب از اندازه‌های واگرایی نسبی χ^2_α گسسته بیان کرد. رابطه‌ای مستقیم میان مشتق مرتبه دوم اندازه واگرایی جنسن χ^2_α و واگرایی نسبی χ^2_α گسسته ارائه شده است. علاوه بر این، تعمیمی از اندازه واگرایی جنسن χ^2_α برای حالت $n + 1$ متغیر تصادفی پیشنهاد شده است. معیار واگرایی جنسن $\chi^2_\alpha(w, \lambda)$ به عنوان تعمیمی از اندازه اطلاع جنسن χ^2_α معرفی و برخی ویژگی‌های مهم آن بررسی شده‌اند. در پایان، دو کاربرد از اندازه واگرایی $\chi^2_\alpha(P: Q)$ در ارزیابی کیفیت تصویر ارائه شده است. نتایج به دست آمده با عملکرد معیار نسبت سیگنال به نویز پیک (PSNR) مقایسه شده‌اند، که در آن اثربخشی اندازه واگرایی $\chi^2_\alpha(P: Q)$ نسبت به یک استاندارد معتبر بررسی شده است. نشان داده شده است که اندازه واگرایی کای-دو تعمیم یافته عملکردی مشابه PSNR دارد و می‌تواند به عنوان یک معیار جایگزین به کار رود.

فهرست منابع

- Ahrari, V. and Habibirad, A. (2019), Quantile Based Tsallis Residual Entropy and its Divergence Measure. *Journal of Statistical Sciences*, **12**(2), 295-321.
- Anjali, and Gupta, A. (2025), A Novel Shannon Entropy-Based Backward Cloud Model and Cloud K-Means Clustering, *The Journal of Supercomputing*, **81**(1), 65.
- Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998), Robust and Efficient Estimation by Minimizing a Density Power Divergence, *Biometrika*, **85**, 549-559.
- Bercher, J. F. (2009), Source Coding with Escort Distributions and Rényi Entropy Bounds, *Physics Letters A*, **36**, 3235-3238.
- Bercher, J. F. (2013), Some Properties of Generalized Fisher Information in the Context of Nonextensive Thermostatistics, *Journal of Physics A: Mathematical and Theoretical*, **392**, 3140-3154.
- Bruni, V., Rossi, E. and Vitulano, D. (2013), Jensen–Shannon Divergence for Visual Quality Assessment, *Signal, Image and Video Processing*, **7**, 411–421.
- Cover, T. M., Thomas, J. A., & Kieffer, J. (1991). *Elements of information theory* (Vol. 2, pp. 1-55). New York: wiley.
- Gray, R. M. (2011), *Entropy and Information Theory*, Springer Science & Business Media, New York.
- Gonzalez, R. C. (2009), *Digital Image Processing*, Prentice-Hall, New York.
- Kaehler, A. and Bradski, G. (2016), *Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library*, O'Reilly Media, Sebastopol.
- Kharazmi, O. and Balakrishnan, N. (2021a), Jensen-Information Generating Function and Its Connections to Some Well-Known Information Measures, *Statistics & Probability Letters*, **170**, 108995.
- Kharazmi, O., Balakrishnan, N. and Ozonur, D. (2023a), Jensen-Discrete Information Generating Function with an Application to Image Processing. *Soft Computing*, **27**, 4543-4552.
- Kharazmi, O., Contreras-Reyes, J. E. and Balakrishnan, N. (2023b), Jensen-Fisher Information and Jensen-Shannon Entropy Measures Based on Complementary Discrete Distributions with an Application to Conway's Game of Life, *Physica D*, **453**, 133822.

Kharazmi, O., Contreras-Reyes, J. E. and Balakrishnan, N. (2023c), Optimal Information, Jensen-RIG Function and α -Onicescu's Correlation Coefficient in Terms of Information Generating Functions, *Physica A: Statistical Mechanics and Its Applications*, **609**, 128362.

Kharazmi, O. and Balakrishnan, N. (2024a), On Jensen- χ^2 Divergence Measure, *Probability in the Engineering and Informational Sciences*, **38**, 403-427.

Kharazmi, O., & Shirazi-Niya, F. (2026). Generalization of Discrete Chi-Squared Information Measure and Its Application in Image Processing. *Journal of Statistical Sciences*, *19*(2), 351-373.

Melbourne, J., Talukdar, S., Bhaban, S., Madiman, M. and Salapaka, M. V. (2022), The Differential Entropy of Mixtures: New Bounds and Applications, *IEEE Transactions on Information Theory*, **68**, 2123-2146.

Navarro, J., Buono, F. and Arevalillo, J. M. (2023), A New Separation Index and Classification Techniques Based on Shannon Entropy, *Methodology and Computing in Applied Probability*, **25**(4), 78.

Nielsen, F. and Nock, R. (2013), On the Chi Square and Higher-Order Chi Distances for Approximating f -Divergences, *IEEE Signal Processing Letters*, **21**, 10-13.

Eskicioglu, A. M. and Fisher, P. S. (1995), Image Quality Measures and Their Performance, *IEEE Transactions on Communications*, **43**, 2959-2965.

Sanei Tabass, M., and Mohtashami Borzadaran, G. (2017), The Generalization of Maximum Entropy Principle for Generalized Information Measures, *Journal of Statistical Sciences*, **11**(1), 101-118.

Shannon, C. E. (1948), A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 379-423.

Steele, J. M. (2004), *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*, Cambridge University Press, Cambridge.

Sourati, J., Gholipour, A., Dy, J. G., Tomas-Fernandez, X., Kurugol, S. & Warfield, S. K. (2019), Intelligent Labeling Based on Fisher Information for Medical Image Segmentation Using Deep Learning, *IEEE Transactions on Medical Imaging*, **38**(11), 2642–2653.

Wang, S. and Fan, J. (2024), Image Thresholding Method Based on Tsallis Entropy Correlation, *Multimedia Tools and Applications*, 1–37.

Wang, N., Zhang, H., Kong, X. and Xu, D. (2025), Explicit Entropy Error Bound for Compressive DOA Estimation in Sensor Array, *Digital Signal Processing*, **156**, 104785.

Woods, R. E. and Gonzalez, R. C. (2001), *Digital Image Processing* (3rd ed.), Computer Science E-Books.