بسندگی احتمالاتی و بسندگی الگوریتمی از دیدگاه نظریهٔ اطلاع

مهدی شمس[۱]

تاریخ دریافت: ۱۳۹۵/۱۰/۷
تاریخ پذیرش: ۱۳۹۶/۱/۳۰

چکیده:
با توجه به اهمیت زنجیر مارکوف در نظریهٔ اطلاع، تعریف احتمال شرطی این فرایند تصادفی می‌تواند بر حسب اطلاع متقابل نیز تعریف شود. در این مقاله ارتباط بین مفهوم بسندگی و زنجیر مارکوف از دیدگاه اصول نظریهٔ اطلاع، و همچنین ارتباط بین بسندگی احتمالاتی و بسندگی الگوریتمی مشخص می‌شود.

واژه‌های کلیدی: تابع بسنده، زنجیر مارکوف، اطلاع متقابل، بسندگی احتمالاتی، بسندگی الگوریتمی.

۱ مقدمه

امروزه نظریهٔ فرایندهای مارکوف یک بخش اصلی از نظریهٔ احتمال است و کاربردهای گسترده‌ای در بسیاری از علوم دیگر دارد [۲۳]. مارکوف [۱۴] آغازکنندهٔ این نظریهٔ در سال ۱۹۰۷ بوده است [۱]. یکی از اوّلین کاربردهای این تئوری، تحقیق روی حروف صدادار و بی‌صدا در زبان روسی بوده است [۱۶]؛ اگرچه زبان‌شناسان مدرن نشان داده‌اند که در برخی زبان‌ها خصوصیت زنجیر مارکوف برقرار نیست. شاخهٔ نظریهٔ اطلاع در سال ۱۹۴۸ [۱۸] توسط شانون معرفی شد و در این شاخه از زنجیرهای مارکوف به‌عنوان یک ابزار استفاده می‌شود. یکی از کاربردهای نظریهٔ اطلاع زمانی است که احتمال وقوع یک نماد در یک پیام به تعداد متناهی از نمادهای قبلی بستگی داشته باشد. در این حالت می‌توان از منبع اطلاع با حافظه[۲] نام برد و دنبالهٔ تولید شده توسط چنین منبعی را یک زنجیر مارکوف در نظر گرفت. در [۷] ارتباط بین اطلاع شرطی و بسندگی زیرمیدان‌ها بررسی شده و ثابت می‌شود اطلاع شرطی برابر با اختلاف بین اطلاعات حاوی میدان و زیرمیدان است. در [۱۰] ارتباط بین بسندگی و اطلاع از این جهت که اطلاعات نمونه برای آمارهٔ بسنده نسبت به نمونهٔ اصلی نه کم می‌شود و نه زیاد، مورد بررسی قرار گرفته است. در [۲۴] راجع به ارتباط بین مفاهیم نسبت درست‌نمایی مجانبی و آنتروپی و بسندگی، مطالبی گفته شده است. در [۱۰، ۲۲] آمارهٔ بسندهٔ مینیمال الگوریتمی معرفی شده است.

در این مقاله با تعریف تابع بسنده و بررسی مفهوم آن از طریق احتمال شرطی، ارتباط بین مفهوم بسندگی و زنجیر مارکوف بیان می‌شود. همچنین ارتباط بین بسندگی احتمالاتی و بسندگی الگوریتمی تعیین می‌گردد و ثابت می‌شود که اگر یک آمارهٔ بسندهٔ الگوریتمی باشد، این آماره، بسندهٔ احتمالاتی نیز هست. بسندگی احتمالاتی در شاخه‌هایی نظیر قابلیت اطمینان [۱۷] و آمار فازی [۱۳] کاربرد دارد. بسندگی الگوریتمی هم در شاخه‌هایی مثل خوشه‌بندی [۴] و یادگیری ماشین [۲۲] مورد استفاده قرار می‌گیرد. در بخش ۳ خواننده باید با مفاهیمی از نظریهٔ کدگذاری آشنا باشد، که برای درک بهتر این بخش، مطالعهٔ [۱۲] و فصل ۱۴ از [۳] توصیه می‌شود.

فضای احتمال $(\Omega, \mathcal{F}, P)$ را در نظر بگیرید، که فضای نمونهٔ Ω متناهی است. متغیر تصادفی گسستهٔ X با مقادیر x_1، ...، x_N را در نظر بگیرید که $P(X = x_k) = p_k$، $k = 1, \ldots, N$. آنتروپی X به‌صورت $H(X) = \sum_{k=1}^{N} p_k log_2 \frac{1}{p_k}$ تعریف می‌شود [۳]. از این رو $H(X) \geq 0$ و شرط لازم وکافی برای $H(X) = 0$ آن است که $N = 1$، یعنی X یک متغیر تصادفی تباهیده باشد. آنتروپی، مقدار اطلاعات دریافت‌شده از X را بعد از مشاهدهٔ مقدار واقعی X نشان می‌دهد. حال اگر f تابعی با دامنهٔ $\{x_1, \ldots, x_N\}$ باشد به‌طوری که برای $i \neq j$ داشته باشیم $f(x_i) \neq f(x_j)$، خواهیم داشت $P(f(X) = f(x_k)) = P(X = x_k) = p_k$ و از این رو

[۱] عضو هیئت علمی گروه آمار، دانشگاه کاشان، ایران

[۲] information source with memory

$H(f(X)) = H(X)$. در حالت کلّی‌تر برای هر تابع f داریم $H(f(X)) \leq H(X)$. از این‌که $H(X)$ فقط به توزیع $\mathcal{P} = \{p_1, \ldots, p_N\}$ مربوط به X بستگی دارد، برخی مواقع آنتروپی به‌صورت $H(\mathcal{P}) = \sum_{k=1}^{N} p_k log_2 \frac{1}{p_k}$ نیز نوشته می‌شود. حال اگر $\mathcal{Q} = \{q_1, \ldots, q_N\}$ احتمال‌های مربوط به متغیر تصادفی دیگری مانند Y باشد و $N \geq 2$، واگرایی اطلاع توزیع $\mathcal{P}$ از $\mathcal{Q}$ را به‌صورت $D(\mathcal{P} \parallel \mathcal{Q}) = \sum_{k=1}^{N} p_k log_2 \frac{p_k}{q_k}$ تعریف می‌کنند که آن را فاصلهٔ کولبک-لایبلر[۳] [۱۰] یا آنتروپی نسبی یا آنتروپی متقابل[۴] نیز می‌گویند [۳].

گزارهٔ ۱.۱. [۳] برای دو توزیع دلخواه $\mathcal{P}$ و $\mathcal{Q}$ داریم $D(\mathcal{P} \parallel \mathcal{Q}) \geq 0$ و تساوی برقرار است اگر و تنها اگر برای هر $k = 1, \ldots, N$، داشته باشیم $p_k = q_k$.

برای متغیر تصادفی گسستهٔ X، آنتروپی بیشین است اگر و تنها اگر برای $k = 1, \ldots, N$، داشته باشیم $p_k = \frac{1}{N}$ و در این حالت $H(X) = log_2 N$ [۳]. اطلاع متقابل[۵] متغیرهای تصادفی X و Y که هر یک تعداد متناهی از مقادیر مجزا را اختیار می‌کند به‌صورت $I(X,Y) = H(X) + H(Y) - H(X,Y)$ تعریف می‌شود، که $H(X,Y)$ آنتروپی (X,Y) است [۸]. از این رو $I(X,Y) \geq 0$ و شرط لازم و کافی برای برقراری تساوی، استقلال X و Y است [۸]. کمیّت $I(X,Y)$ مقدار اطلاع به‌دست‌آمده راجع به متغیر تصادفی X بر اساس مشاهدات متغیر تصادفی Y را اندازه می‌گیرد [۳]. اگر $\mathcal{P}_k$ معرف خانواده توزیع‌های شرطی $\{p_{j|k}, j = 1, \ldots, N\}$ باشد که در آن $p_{j|k} = P(X = x_j | Y = y_k)$ است، میانگین آنتروپی شرطی X به‌شرط Y به‌صورت $H(X|Y) = \sum_{k=1}^{N} q_k H(\mathcal{P}_k)$ تعریف می‌شود و می‌توان نشان داد [۳] که

$$H(X|Y) = H(X,Y) - H(Y) \geq 0,$$
$$I(X,Y) = H(X) - H(X|Y),$$
$$I(X,Y) \leq \min(H(X), H(Y)).$$

برای متغیرهای تصادفی $\{X_i : i = 1, \ldots, N\}$ که تعداد متناهی از مقادیر مجزا را اختیار می‌کنند و $N > 2$، داریم $H(X_1, \ldots, X_N) \leq \sum_{k=1}^{N} H(X_k)$ و شرط لازم و کافی برای برقراری تساوی آن است که X_1، ...، X_N مستقل باشند [۳]. اطلاع متقابل فراهم‌آمده توسط متغیرهای تصادفی X_1، ...، X_N به‌صورت

$$I(X_1, \ldots, X_N) = H(X_1, \ldots, X_N) - \sum_{k=1}^{N} H(X_k)$$

تعریف می‌شود. شرط لازم و کافی برای استقلال X_1، ...، X_N آن است که اطلاع متقابل آنها صفر باشد یا به‌طور معادل برای هر $k = 1, \ldots, N-1$، داشته باشیم $I((X_1, \ldots, X_k), X_{k+1}) = 0$ [۳]. اگر متغیرهای تصادفی مستقل $X_1, \ldots, X_N$ و متغیر تصادفی Z تعداد متناهی از مقادیر مجزایی را اختیار کنند، ثابت می‌شود [۸]:

$$\sum_{k=1}^{N} I(X_k, Z) \leq I((X_1, \ldots, X_k), Z).$$

تعریف ۲.۱. [۳] متغیرهای تصادفی X و Y و Z از یک زنجیر مارکوف می‌آیند و می‌نویسند $X \to Y \to Z$ هرگاه

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

با استفاده از تعریف بالا نتیجهٔ زیر حاصل می‌شود:

نتیجه ۳.۱. [۳] اگر X و Y و Z متغیرهای تصادفی باشند، داریم:

۱. $X \to Y \to Z$ اگر و تنها اگر X و Z به‌شرط Y مستقل باشند.

۲. $X \to Y \to Z$ نتیجه می‌دهد $Z \to Y \to X$ و بنا بر این می‌توان نوشت $X \leftrightarrow Y \leftrightarrow Z$.

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

گزارهٔ زیر، یکی از نامساوی‌های مهم در نظریهٔ اطلاع با نام نامساوی پردازش داده‌ها است که برای معرفی مفهوم بسندگی از دیدگاه نظریهٔ اطلاع به آن نیاز داریم.

گزارهٔ ۴.۱. (نامساوی پردازش داده‌ها) [۳] اگر $X \to Y \to Z$، آن‌گاه $I(X,Y) \geq I(X,Z)$.

با استفاده از گزارهٔ ۴.۱ حقایق زیر به دست می‌آیند [۳]:

۱. اگر $Z = f(Y)$، آن‌گاه $I(X,Y) \geq I(X, f(Y))$.

۲. اگر $X \to Y \to Z$، آن‌گاه $I(X,Y) \geq I(X,Y|Z)$.

[۳] Kullback–Leibler distance

[۴] cross entropy

[۵] mutual information

فرض کنید $\{P_\theta\}$ خانواده‌ای از توزیع‌ها مربوط به متغیر تصادفی X باشد که مقادیر خود را در مجموعهٔ شمارای $\mathcal{X}$ اختیار می‌کند و Θ فضای پارامتر باشد. تابع $T : \mathcal{X} \to \mathcal{T}$ را یک آماره از داده‌ها در $\mathcal{X}$ نامند. در این حالت $\theta \to X \to T(X)$ و در پی آن با به کار گیری گزارهٔ ۴.۱ داریم:

$$I(\theta, X) \geq I(\theta, T(X)). \qquad (۱)$$

اکنون این سؤال مطرح می‌شود که آیا می‌توان آماره‌ای انتخاب کرد که همهٔ اطلاعات X راجع به θ را حفظ کند؟ آمارهٔ T برای خانوادهٔ $\{P_\theta\}$ (یا برای θ) بسنده است هرگاه برای هر $t \in \mathcal{T}$، توزیع شرطی $f_\theta(x|t) = P_\theta(X = x|T(X) = t)$ مستقل از θ باشد. به‌عبارت دیگر، این توزیع شرطی تحت تغییر θ پایا است [۱۱]. این تعریف نشان می‌دهد همهٔ اطلاعاتی که در مشاهدهٔ x راجع به پارامتر θ است، در آمارهٔ $T(x)$ خلاصه شده است. طبق قضیهٔ تجزیهٔ فیشر-نیمن [۱۱] داریم $f_\theta(x) = f(x|t)f_\theta(t)$، یا به‌ازای $x \in \mathcal{X}$ و $s \in \Theta$،

$$P(X = x|\theta = s) = P(X = x|T(X) = T(x)) \times P(T(X) = T(x)|\theta = s).$$

از این رو به‌ازای هر $x \in \mathcal{X}$ و $s \in \Theta$،

$$\begin{aligned} &P(\theta = s, X = x|T(X) = T(x)) \\ &= \frac{P(\theta = s, X = x)}{P(T(X) = T(x))} \\ &= \frac{P(X = x|\theta = s)P(\theta = s)}{P(T(X) = T(x))} \\ &= \frac{P(X = x|T(X) = T(x))P(T(X) = T(x)|\theta = s)P(\theta = s)}{P(T(X) = T(x))} \\ &= \frac{P(X = x, T(X) = T(x))}{P(T(X) = T(x))} \frac{P(T(X) = T(x), \theta = s)}{P(T(X) = T(x))} \\ &= P(\theta = s|T(X) = T(x))P(X = x|T(X) = T(x)), \end{aligned}$$

و از این رو بسندگی $T(X)$ معادل با این حقیقت است که برای هر توزیع θ روی Θ، θ و X به‌شرط $T(X)$ از هم مستقل هستند. با استفاده از این حقیقت و با به کار گیری قسمت (۱) و (۲) از نتیجهٔ ۳.۱، $\theta \to T(X) \to X$ و در پی آن با استفاده از نامساوی پردازش داده‌ها داریم:

$$I(\theta, T(X)) \geq I(\theta, X). \qquad (۲)$$

علاوه بر این از استقلال شرطی θ و X به‌شرط $T(X)$ مشاهده می‌شود:

$$\begin{aligned} &\frac{P(\theta = s, X = x)}{P(T(X) = T(x))} \\ &= \frac{P(\theta = s, T(X) = T(x))}{P(T(X) = T(x))} \frac{P(X = x)}{P(T(X) = T(x))}, \end{aligned}$$

و بالاخره $P(\theta = s|X = x) = P(\theta = s|T(X) = T(x))$ که معادلهٔ اخیر این حقیقت را تصدیق می‌کند که اگر به‌جای X از آمارهٔ بسندهٔ $T(X)$ استفاده شود اطلاعی راجع به پارامتر θ از دست نمی‌رود.

در پایان با استفاده از نامساوی‌های (۱) و (۲) برای یک آمارهٔ بسندهٔ دلخواه T داریم:

$$I(\theta, X) = I(\theta, T(X)). \qquad (۳)$$

ذکر این نکته ضروری است که بسندگی نمی‌تواند گسترش یابد؛ به این مفهوم که با توجه به رابطهٔ (۱) اطلاع متقابل داده‌ها و مدل به‌هیچ‌وجه نمی‌تواند توسط پردازش داده‌های نمونه افزایش یابد.

به‌طور معادل با استفاده از تعریف بسندگی می‌توان گفت که شرط لازم و کافی برای این‌که آمارهٔ T بسنده باشد آن است که تابع $g : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$ وجود داشته باشد به‌طوری که به‌ازای هر $x \in \mathcal{X}$ و $t \in \mathcal{T}$ و $\theta \in \Theta$،

$$g(x, t) = f_\theta(x|t). \qquad (۴)$$

برای درک و فهم بیشتر بسندگی از دیدگاه نظریهٔ اطلاع، مثال زیر می‌تواند مفید باشد.

مثال ۵.۱. [۳] فرض کنید $X_1, \ldots, X_n \overset{i.i.d.}{\sim} ber(\theta)$ که $\theta \in (0, 1)$. بنا بر این،

$$f_\theta(\mathbf{x}) = \theta^{T(\mathbf{x})}(1 - \theta)^{n - T(\mathbf{x})},$$
$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X} = \{0, 1\}^n,$$

که در آن $T(\mathbf{x}) = \sum_{i=1}^n x_i$ تعداد ۱ها در $\mathbf{x}$ است. در این حالت،

$$f_\theta(x|t) = P_\theta(\mathbf{X}=\mathbf{x}|T=t) = \begin{cases} \dfrac{1}{\binom{n}{t}}, & T(\mathbf{x})=t \\ 0, & \text{در غیر این‌صورت} \end{cases} \quad (5)$$

بنا بر این $\theta \to T(\mathbf{X}) \to (X_1,\dots,X_n)$ از یک زنجیر مارکوف می‌آید و لذا $T(\mathbf{X})$ برای θ بسنده است. از دیدگاه دیگر نیز می‌توان بسندگی T را نتیجه گرفت. از آن‌جا که برای یک $\theta \in (0,1)$ و $t \in \mathcal{T} = \{0,\dots,n\}$ دلخواه، همهٔ $\mathbf{x}$ها با t تا ۱ و $n-t$ تا صفر دارای احتمال‌های برابر هستند و تعداد چنین $\mathbf{x}$ ها برابر با ترکیب $\binom{n}{t}$ است، رابطهٔ (۵) در معادلهٔ (۴) صدق می‌کند که در آن $g(\mathbf{x},t)$ توزیع یکنواخت روی همهٔ $\mathbf{x}$ هایی است که دقیقاً t تا ۱ دارند و بنا بر این $T(\mathbf{X})$ یک آمارهٔ بسنده برای θ است.

یک آمارهٔ بسنده ممکن است حاوی اطلاعاتی باشد که مناسب نیست و برای رفع این مشکل از آمارهٔ بسندهٔ مینیمال استفاده می‌شود که این آماره اطلاعات در داده‌ها راجع به پارامتر را به‌طور ماکسیمال فشرده می‌کند. آمارهٔ بسندهٔ $T(X)$ را برای θ بسندهٔ مینیمال گویند هرگاه تابعی از هر آمارهٔ بسندهٔ دیگر مثل $U(X)$ باشد [۱۱]. بنا بر این $\theta \to T(X) \to U(X) \to X$ که در این حالت اطلاعات حاوی θ در نمونه بیشترین فشردگی را دارد و همچنین

$$I(X,T(X)) \le I(X,U(X)), \quad (6)$$

که این نامساوی نشان می‌دهد از بین همهٔ توابع بسنده، بسندگی مینیمال دارای حد اقل اطلاع متقابل روی نمونهٔ X است. به‌کمک روابط (۳) و (۶) در حالتی که $(X,Y) \sim p(x,y)$ باشد، می‌توان قسمت وابستهٔ X نسبت به Y را تعریف کرد. در این حالت متغیر تصادفی T را به‌گونه‌ای پیدا می‌کنیم که:

۱. $T \to X \to Y$ از یک زنجیر مارکوف بیاید.

۲. $I(X,T)$ کمینه باشد (مینیمال بودن)، در حالی که $I(Y,T)$ بیشینه باشد (بسندگی)، که این دو حقیقت معادل با کمینه کردن تابع لاگرانژ $\mathcal{L}[p(t|x)] = I(X,T) - \beta I(Y,T)$ نسبت به شرط مارکوف است.

متغیر تصادفی T را که در شرایط بالا صدق می‌کند، گره اطلاع [۶] بین X و Y گویند. حل مسئله گره اطلاع به معادلات زیر منجر می‌شود [۲۰]:

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} \exp(-\beta D(p(y|x) \parallel p(y|t))),$$

که در آن،

$$p(t) = \sum_x p(t|x)p(x)$$

$$p(y|t) = \sum_x p(y|x)p(x|t)$$

$$Z(x,\beta) = \sum_t p(t) \exp(-\beta D(p(y|x) \parallel p(y|t))),$$

$$D(p(y|x) \parallel p(y|t)) = E_{p(y|x)}\left[\log \frac{p(y|x)}{p(y|t)}\right].$$

۲ ارتباط زنجیرهای مارکوف و بسندگی از دیدگاه نظریهٔ اطلاع

عنصر اصلی فرمول‌بندی خاصیت مارکوف بر حسب آنتروپی و اطلاع، استفاده از دیدگاه تابع بسنده است. ایدهٔ اصلی بسندگی یک متغیر تصادفی برای یک متغیر تصادفی دیگر از رابطهٔ (۳) گرفته شده است.

تعریف ۱.۲. [۱۶] اگر X و Y متغیرهای تصادفی روی یک فضای احتمال $(\Omega,\mathcal{F},P)$ باشند به‌طوری که $I(X,Y) < \infty$، و g یک تابع با مقدار حقیقی و بورل اندازه‌پذیر باشد، متغیر تصادفی $g(X) = g \circ X$ را یک تابع بسنده از متغیر تصادفی X برای متغیر تصادفی Y گویند هرگاه $I(g(X),Y) = I(X,Y)$. به‌عبارت دیگر $g(X)$ برای Y بسنده است هرگاه $g(X)$ شامل همهٔ اطلاعات متغیر تصادفی Y باشد که توسط متغیر تصادفی X فراهم می‌آیند.

قضیهٔ زیر، یک شرط لازم و کافی برای بسندگی یک آماره برای یک متغیر تصادفی دیگر ارائه می‌کند.

قضیهٔ ۲.۲. [۱۶] متغیرهای تصادفی X و Y با تکیه‌گاه‌های به‌ترتیب $\{x_1,\dots,x_N\}$ و $\{y_1,\dots,y_M\}$ و تابع g با مقادیر

[۶] information bottleneck

حقیقی و بورل اندازه‌پذیر را در نظر بگیرید. شرط لازم و کافی برای آن‌که $g(X)$ یک تابع بسنده برای Y باشد آن است که توزیع احتمال شرطی Y به‌شرط X فقط به مقدار $g(X)$ بستگی داشته باشد؛ یعنی اگر به‌ازای $i = ۱,\ldots,N$ و $j = ۱,\ldots,N$ داشته باشیم $g(x_i) = g(x_j)$، آن‌گاه

$$P(Y = y_k|X = x_i) = P(Y = y_k|X = x_j). \quad (۷)$$

به‌عبارت دیگر X و Y متغیرهای تصادفی مستقل هستند هنگامی که مقدار $g(X)$ ثابت باشد، یعنی برای هر z که $P(g(X) = z) > ۰$ باشد، داریم:

$$P(X = x_j, Y = y_k|g(X) = z) =$$
$$P(X = x_j|g(X) = z)P(Y = y_k|g(X) = z). \quad (۸)$$

اثبات. تعریف کنید

$$\mathcal{P} = \{p_۱,\ldots,p_N\}$$

که $p_j = P(X = x_j)$،

$$\mathcal{Q} = \{q_۱,\ldots,q_M\}$$

که $q_k = P(Y = y_k)$ و

$$\mathcal{R} = \{r_{jk}, \quad j = ۱,\ldots,N, \quad k = ۱,\ldots,M\}$$

که در آن

$$r_{jk} = P(X = x_j, \quad Y = y_k).$$

فرض کنید $\{z_۱,\ldots,z_s\}$ مجموعهٔ مقادیر مجزا از برد تابع g باشد که مقادیر خود را روی مجموعهٔ $\{x_۱,\ldots,x_N\}$ اختیار می‌کنند و قرار دهید $g(x_j) = z_l$، که $j \in D_l$ یک افراز از مجموعهٔ $\{۱,\ldots,N\}$ است. به‌علاوه برای $j \in D_l$ تعریف کنید $v(x_j) = l$ و همچنین $t_{lk} = P(g(X) = z_l, Y = y_k)$ و $t_l = P(g(X) = z_l)$. اعداد

$$u_{jk} = \frac{t_{v(x_j)k}.p_j}{t_{v(x_j)}}, \quad j = ۱,\ldots,N, \quad k = ۱,\ldots,M,$$

از یک توزیع احتمال $\mathcal{U} = \{u_{jk}\}$ می‌آیند؛ زیرا

$$\begin{aligned}\sum_{j=۱}^{N}\sum_{k=۱}^{M} u_{jk} &= \sum_{j=۱}^{N}\sum_{k=۱}^{M}\frac{t_{lk}.p_j}{t_l}\\ &= \sum_{j=۱}^{N}\sum_{k=۱}^{M}\frac{P(g(X) = z_l, Y = y_k)P(X = x_j)}{P(g(X) = z_l)}\\ &= \sum_{j=۱}^{N} P(X = x_j)\left[\frac{\sum_{k=۱}^{M} P(g(X) = z_l, Y = y_k)}{P(g(X) = z_l)}\right]\\ &= ۱.\end{aligned}$$

با توجه به این حقیقت که $\mathcal{R} = \mathcal{U}$ معادل با این است که به‌ازای هر $j = ۱,\ldots,N$ و $k = ۱,\ldots,M$، داشته باشیم $r_{jk} = u_{jk}$، پس $\frac{r_{jk}}{p_j} = \frac{t_{v(x_j)k}}{t_{v(x_j)}}$. بنا بر این اگر $g(x_i) = g(x_j)$ باشد، از این‌که $v(x_i) = v(x_j)$ داریم:

$$\frac{r_{ik}}{p_i} = \frac{t_{v(x_i)k}}{t_{v(x_i)}} = \frac{t_{v(x_j)k}}{t_{v(x_j)}} = \frac{r_{jk}}{p_j},$$

یا به‌طور معادل

$$\frac{P(X = x_i, Y = y_k)}{P(X = x_i)} = \frac{P(X = x_j, Y = y_k)}{P(X = x_j)},$$

که همان رابطهٔ (۷) است.

از این رو ثابت کردیم رابطهٔ (۷) هم‌ارز با $\mathcal{R} = \mathcal{U}$ است. پس با استفاده از گزارهٔ ۱.۱ و تساوی

$$I(X,Y) - I(g(X),Y) = D(\mathcal{R} \parallel \mathcal{U}),$$

شرط لازم و کافی برای $I(X,Y) = I(g(X),Y)$ آن است که $\mathcal{R} = \mathcal{U}$ باشد. بنا بر این رابطهٔ (۷) برقرار است اگر و تنها اگر $I(g(X),Y) = I(X,Y)$ باشد، که همان تعریف بسندگی $g(X)$ نسبت به متغیر تصادفی Y است. به‌علاوه اگر $g(x_j) = z_l$، آن‌گاه $\{X = x_j\} \subseteq \{g(X) = z_l\}$ و در پی آن

$$\begin{aligned}&P(X = x_j, Y = y_k|g(X) = z_l)\\ &= \frac{P(X = x_j, Y = y_k, g(X) = z_l)}{P(g(X) = z_l)}\\ &= \frac{P(X = x_j, Y = y_k)}{P(g(X) = z_l)}\\ &= \frac{P(X = x_j)P(Y = y_k|X = x_j)}{P(g(X) = z_l)}\\ &= \frac{P(X = x_j, g(X) = z_l)}{P(g(X) = z_l)}P(Y = y_k|X = x_j)\\ &= P(X = x_j|g(X) = z_l)P(Y = y_k|X = x_j),\end{aligned}$$

و لذا با مقایسهٔ تساوی بالا با رابطهٔ (۸) می‌توان گفت به‌شرط $g(x_j) = z_l$ رابطهٔ (۸) معادل است با

$$P(Y = y_k|X = x_j) = P(Y = y_k|g(X) = z_l).$$

بنا بر این (۷) و (۸) معادل هستند. □

با توجه به تعریف ۲.۱، می‌توان یک تعریف معادل برای زنجیر مارکوف ارائه کرد.

تعریف ۳.۲. دنبالهٔ $\{X_n\}_{n\in\mathbb{N}}$ از متغیرهای تصادفی با مقادیر حقیقی روی فضای احتمال $(\Omega,\mathcal{F},P)$ را که تعداد متناهی از مقادیر را اختیار می‌کند زنجیر مارکوف گسسته‌پارامتر گویند هرگاه برای $n\in\mathbb{N}$ داشته باشیم

$$I((X_1,\ldots,X_n),X_{n+1})=I(X_n,X_{n+1}).$$

نتیجهٔ زیر، ارتباط بین بسندگی و زنجیر مارکوف را بیان می‌کند.

نتیجه ۴.۲. تعریف ۲ بیان‌کنندهٔ این حقیقت است که برای یک زنجیر مارکوف گسسته‌پارامتر اطلاعی که راجع به X_{n+1} بر اساس مشاهدات $X_1,\ldots,X_n$ به دست می‌آید برابر با اطلاعی است که راجع به X_{n+1} بر اساس فقط مشاهدهٔ آخر، یعنی X_n، به دست می‌آید و از این رو X_n که به‌صورت یک تابع از $X_1,\ldots,X_n$ است برای هر $n\in\mathbb{N}$ یک تابع بسنده برای X_{n+1} می‌باشد. به‌طور معادل،

$$P(X_{n+1}=x_{n+1}|X_1,\ldots,X_n)=P(X_{n+1}=x_{n+1}|X_n).$$

۳ بسندگی احتمالاتی و بسندگی الگوریتمی

در ابتدا تعریفی برای آمارهٔ بسندهٔ احتمالاتی ارائه می‌کنیم. برای این منظور با نوشتن رابطهٔ (۴) به‌صورت

$$log\frac{1}{g(x,t)}=log\frac{1}{f_\theta(x|t)}, \qquad (۹)$$

می‌توان تعریف زیر را ارائه کرد.

تعریف ۱.۳. تابع $T:\mathcal{X}\to\mathcal{T}$ را آمارهٔ بسندهٔ احتمالاتی برای $\{f_\theta\}$ گویند هرگاه تابع $g:\mathcal{X}\times\mathcal{T}\to\mathbb{R}$ وجود داشته باشد به‌طوری که برای هر $\theta\in\Theta$، $x\in\mathcal{X}$ و $t\in\mathcal{T}$ در شرط (۹) صدق کند. (برای راحتی قرارداد می‌کنیم $log\frac{1}{0}=\infty$)

برای به دست آوردن نسخهٔ امید ریاضی تعریف بسندگی احتمالاتی، ثابت می‌شود تعریف‌های زیر هم‌ارز با تعریف ۱.۳ هستند.

تعریف ۲.۳. تابع $T:\mathcal{X}\to\mathcal{T}$ را آمارهٔ بسندهٔ احتمالاتی برای $\{f_\theta\}$ گویند هرگاه برای هر $\theta\in\Theta$ داشته باشیم:

$$\sum_x f_\theta(x)log(\frac{1}{f_\theta(x|T(x))})=\sum_x f_\theta(x)log(\frac{1}{g(x,T(x))}).$$

تعریف ۳.۳. تابع $T:\mathcal{X}\to\mathcal{T}$ را آمارهٔ بسندهٔ احتمالاتی برای $\{f_\theta\}$ گویند هرگاه برای هر توزیع پیشین $\pi(\theta)$ روی Θ داشته باشیم:

$$\sum_{\theta,x}\pi(\theta,x)\log(\frac{1}{f_\theta(x|T(x))})=\sum_{\theta,x}\pi(\theta,x)\log(\frac{1}{g(x,T(x))}).$$

برای اثبات این حقیقت که تعریف ۱.۳، تعریف ۲.۳ را نتیجه می‌دهد فرض کنید به‌ازای هر $\theta\in\Theta$ و $x\in\mathcal{X}$ و $t\in\mathcal{T}$ رابطهٔ (۹) برقرار باشد. با گرفتن امید ریاضی از طرفین رابطهٔ (۹)، تعریف ۲.۳ ثابت می‌شود. برای اثبات عکس، فرض کنید برای هر $\theta\in\Theta$ تعریف ۲.۳ برقرار باشد. با تعریف $f_\theta(t)=\sum_{y\in\mathcal{X}:T(y)=t}f_\theta(y)$ و اضافه کردن

$$\sum_x f_\theta(x)log\frac{1}{f_\theta(T(x))},$$

به هر دو طرف تساوی، تعریف ۲.۳ را می‌توان دوباره به‌صورت زیر بازنویسی کرد:

$$\sum_x f_\theta(x)log\frac{1}{f_\theta(x)}=\sum_x f_\theta(x)log\frac{1}{g_\theta(x)}, \qquad (۱۰)$$

که $g_\theta(x)=f_\theta(T(x))g(x,T(x))$. با توجه به تعریف ۱.۱ رابطهٔ (۱۰) زمانی برقرار است که برای هر $x\in\mathcal{X}$ داشته باشیم $g_\theta(x)=f_\theta(x)$ و در پی آن تعریف ۱.۳ نتیجه می‌شود.

برای اثبات معادل بودن تعریف ۲ و تعریف ۱.۳ از خطی بودن امید ریاضی و این‌که $p(\theta,x)=\pi(\theta)f_\theta(x)$ استفاده می‌شود.

پیچیدگی کولموگوروف[۷] (اطلاع یا آنتروپی الگوریتمی) $K(x)$، برای اوّلین بار در مراجعی نظیر [۹، ۲، ۱۹] به‌عنوان طول کوتاه‌ترین توصیف دودویی مؤثر از x معرفی شد. به‌عبارت دیگر $K(x)$، طول کوتاه‌ترین برنامهٔ دودویی برای محاسبهٔ رشتهٔ x از یک کامپیوتر همگانی نظیر ماشین تورینگ[۸]است [۹]. ماشین تورینگ یک مدل ریاضی برای وسیله‌ای است که می‌تواند داده‌ها را بر یک نوار ذخیره‌سازی قابل کنترل بخواند و بنویسد. مثلاً $K(x)$ می‌تواند به‌عنوان طول کوتاه‌ترین برنامهٔ کامپیوتری باشد که x چاپ و سپس برنامهٔ متوقف شود. این برنامهٔ کامپیوتری می‌تواند با یکی از زبان‌های فراگیر کامپیوتری نظیر

[۷] Kolmogorov complexity

[۸] Turing machine

C، جاوا[۹] و $LISP$[۱۰] نوشته شود و در هر یک از آنها نتیجهٔ برنامه برای مقدار $K(x)$ در یک ثابت که به x بستگی ندارد متفاوت است. به‌طور مشابه پیچیدگی کولموگوروف شرطی $K(x|y)$ طول کوتاه‌ترین برنامه برای محاسبهٔ x است، به‌شرطی که y یک ورودی جانبی (مثل نوار مغناطیسی، فلاپی دیسک، فلش مموری و لوح فشرده) باشد [۶]. برای بررسی دقیق‌تر موضوع، فرض کنید T_1، T_2، ... شمارش استاندارد همهٔ ماشین‌های تورینگ باشد و ϕ_1، ϕ_2، ... شمارش توابع متناظر که توسط یک ماشین تورینگ مخصوص محاسبه شده است. بنا بر این T_i، مقدار ϕ_i را محاسبه می‌کند. مجموعهٔ S شامل x با ویژگی $K(x) = K(S) + log|S| + O(1)$ را بهینه نامند. یک آمارهٔ بسندهٔ الگوریتمی x، کوتاه‌ترین برنامه برای یک مجموعهٔ S حاوی x است که بهینه باشد، یعنی در شرط معادلهٔ بالا صدق کند. یک آمارهٔ بسندهٔ الگوریتمی با مجموعهٔ بهینهٔ S را مینیمال گویند هرگاه هیچ مجموعهٔ بهینه‌ای مثل S' وجود نداشته باشد که $K(S') < K(S)$. برای بهتر درک کردن مفهوم مجموعهٔ بهینه، مثال زیر می‌تواند مفید باشد.

مثال ۴.۳. فرض کنید $k \in \{0, 1, \ldots, n\}$ باشد و x یک رشته به‌طول n با تعداد k تا ۱ باشد. این x می‌تواند به‌عنوان نتیجهٔ پرتاب یک سکه با اریبی $p = \frac{k}{n}$ در نظر گرفته شود. توصیف دوقسمتی از x را می‌توان به‌صورت تعداد ۱ها در x برای اوّلین بار (k) که اندیس در شرط $j \leq log|S|$ صدق می‌کند و x در یک مجموعهٔ S حاوی رشته‌های به طول n با k تا ۱ است، در نظر گرفت. این مجموعهٔ بهینه است، زیرا $K(x|n) = K(S) + log|S|$.

اکنون به ارتباط بسندگی الگوریتمی و بسندگی احتمالاتی می‌پردازیم. برای این منظور، تعریف زیر ارائه می‌شود.

تعریف ۵.۳. یک آمارهٔ دنباله‌ای، تابعی مانند $S : \{0,1\}^* \to 2^{\{0,1\}^*}$ است (که در آن

$$\{0,1\}^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\}$$

و ε لغت تهی (حرفی نداشته باشد) است) به‌طوری که برای هر n و $x \in \{0,1\}^n$ داریم:

۱) $S(x) \subseteq \{0,1\}^n$،

۲) $x \in S(x)$،

۳) به‌ازای هر n، مجموعهٔ $\{s : \exists\, x \in \{0,1\}^n : S(x) = s\}$ یک افراز از $\{0,1\}^n$ باشد.

دنباله‌ای از توابع جرم احتمال $\{f_\theta^{(n)}\}_{n\in\mathbb{N}}$ را که $f_\theta^{(n)}$ تابع جرم احتمال روی $\{0,1\}^n$ است و توزیع کناری f_θ روی اوّلین قطعه n بیتی است، در نظر بگیرید. آمارهٔ دنباله‌ای S را تقریباً بسنده برای θ از دیدگاه احتمالاتی-فردی گویند هرگاه توابع $g^{(1)}$ و $g^{(2)}$ و ثابت c وجود داشته باشند به‌قسمی که به‌ازای هر θ و n و $x \in \{0,1\}^n$ داشته باشیم:

$$|log\frac{1}{f_\theta^{(n)}(x|S(x))} - log\frac{1}{g^{(n)}(x|S(x))}| < \varepsilon,$$

و S را تقریباً بسنده برای θ از دیدگاه احتمالاتی-امید ریاضی گویند هرگاه توابع $g^{(1)}$، $g^{(2)}$، ... و ثابت c' وجود داشته باشند به‌طوری که به‌ازای هر θ و n داشته باشیم:

$$\Big|\sum_{x\in\{0,1\}^n} f_\theta^{(n)}(x)\Big[log\frac{1}{f_\theta^{(n)}(x|S(x))} - log\frac{1}{g^{(n)}(x|S(x))}\Big]\Big| \leq c'. \qquad (11)$$

آمارهٔ دنباله‌ای S را از دیدگاه الگوریتمی، بسنده گویند هرگاه ثابت c وجود داشته باشد که برای هر n و $x \in \{0,1\}^n$، برنامهٔ مولّد $S(x)$ یک آمارهٔ بسندهٔ الگوریتمی برای x (نسبت به ثابت c) باشد؛ یعنی

$$K(S(x)) + log|S(x)| \leq K(x) + c.$$

در قضیهٔ زیر نشان می‌دهیم که آمارهٔ بسندهٔ الگوریتمی، آمارهٔ بسندهٔ احتمالاتی نیز هست.

قضیهٔ ۶.۳. اگر S یک آمارهٔ دنباله‌ای باشد که از لحاظ الگوریتمی بسنده است، به‌ازای هر θ که $K(f_\theta) < \infty$، یک ثابت c وجود دارد که برای هر n، نامساوی (۱۱) برقرار است و در آن $g^{(n)}$ توزیع یکنواخت است؛ یعنی

$$g^{(n)}(x|s) = \begin{cases} \frac{1}{|\{x \in \{0,1\}^n : S(x) = s\}|}, & S(x) = s \\ 0, & \text{در غیر این‌صورت} \end{cases}$$

[۹] Java

[۱۰] locator identifier separation protocol

بنا بر این اگر $sup_{\theta\in\Theta}K(f_\theta)<\infty$ باشد، S یک آمارهٔ تقریباً بسنده برای θ از دیدگاه احتمالاتی-امید ریاضی است، که g توزیع یکنواخت است.

اثبات. طبق تعریف بسندگی الگوریتمی، یک ثابت c وجود دارد به‌طوری که به‌ازای هر θ و n داریم:

$$\sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)[K(S(x))+log|S(x)|] \le \sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)K(x)+c. \qquad (۱۲)$$

اکنون با ثابت فرض کردن θ، که $K(f_\theta)<\infty$، به‌ازای یک $c_\theta\approx K(f_\theta)$ و هر n خواهیم داشت:

$$۰ \le \sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)K(x) - \sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)log\frac{۱}{f_\theta(x)} \le c_\theta, \qquad (۱۳)$$

که دلیل نامساوی اوّلی، گزارهٔ ۱.۱ است. دلیل نامساوی دوم این است که با توجه به $K(f_\theta)<\infty$، کد شانون-فانو[۱۱] می‌تواند توسط یک برنامهٔ کامپیوتری با یک اندازهٔ ثابت انجام شود که به n بستگی ندارد.

با توجه به (۱۲) و (۱۳) برای هر n داریم:

$$\sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)log\frac{۱}{f_\theta(x)} \le \sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)[K(S(x))+log|S(x)|] \le \sum_x f_\theta^{(n)}(x)log\frac{۱}{f_\theta(x)}+c_\theta.$$

برای $s\subseteq\{۰,۱\}$، نماد

$$f_\theta^{(n)}(s)=\sum_{y\in\mathcal{X}:S(y)=s} f_\theta^{(n)}(y)$$

را در نظر می‌گیریم. با توجه به فرضیات قسمت ۳ از تعریف ۵.۳ داریم:

$$\sum_{s:\exists x\in\{۰,۱\}^n:S(x)=s} f_\theta^{(n)}(s)=۱,$$

که $f_\theta^{(n)}(s)$ تابع جرم احتمال روی S شامل مجموعهٔ مقادیر آمارهٔ S است که می‌تواند دنباله‌ای به طول n را اختیار کند. بنا بر این با استفاده از گزارهٔ ۱.۱ داریم:

$$\sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)K(S(x)) \ge \sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)log\frac{۱}{f_\theta^{(n)}(S(x))}.$$

توجه کنید که برای هر n داریم:

$$\sum_{x\in\{۰,۱\}} f_\theta^{(n)}(x)[log\frac{۱}{f_\theta^{(n)}(S(x))}+log\frac{۱}{f_\theta^{(n)}(x|S(x))}] = \sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)log\frac{۱}{f_\theta(x)}, \qquad (۱۴)$$

و دو بخشی را که x توسط اوّلین رمزگذاری $S(x)$ با استفاده از $log\frac{۱}{f_\theta^{(n)}(S(x))}$ بیت و سپس توسط $log|S(x)|$ بیت رمزنگاری می‌شود در نظر بگیرید. با توجه به گزارهٔ ۱.۱ این کد باید کارایی کمتری نسبت به کد شانون-فانو با طول $log\frac{۱}{f_\theta(x)}$ داشته باشد. بنا بر این با توجه به (۱۴) برای هر n داریم:

$$\sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)log|S(x)| \ge log\sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)log\frac{۱}{f_\theta^{(n)}(x|S(x))}.$$

بنا بر این یک c' وجود دارد که برای هر n داشته باشیم:

$$|\sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)log|S(x)| - \sum_{x\in\{۰,۱\}^n} f_\theta^{(n)}(x)log\frac{۱}{f_\theta^{(n)}(x|S(x))}| \le c'.$$

□

۴ نتیجه‌گیری

در این مقاله با تعریف مفهوم بسندگی یک آماره نسبت به یک متغیر تصادفی دیگر از دیدگاه نظریهٔ اطلاع، یک ارتباط بین مفهوم بسندگی و زنجیرهای مارکوف نتیجه می‌شود. ذکر این نکته جالب است که در حالت خاص که یک آماره نسبت به یک متغیر تصادفی ثابت (که در این‌جا مقدار پارامتری را اختیار می‌کند که باید براورد شود) بسنده باشد همان مفهوم بسندگی معمولی نتیجه می‌شود. علاقه‌مندان با در نظر گرفتن پارامتر به‌عنوان یک متغیر تصادفی می‌توانند این مفهوم را در حالت بیزی گسترش دهند، که می‌تواند شروعی برای تحقیقات آینده باشد. در بخش آخر مقاله مفهوم بسندگی الگوریتمی مطرح می‌شود و ثابت می‌شود که همان مفهوم بسندگی احتمالاتی را نتیجه می‌دهد.

[۱۱] Shannon-Fano code

مراجع

[1] Basharin, G. P., Langville, A. N. and Naumov, V. A. (2004). The life and work of A. A. Markov. *Linear Algebra and its Applications*, **386**, 3-26.

[2] Chaitin, G. J. (1969). On the length of programs for computing finite binary sequences: statistical considerations. *journal of Association for Computing Machinery*, **16**, 145-159.

[3] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. 2nd edition, Wiley.

[4] Fushing, H., Wang, H., Vander Waal, K., McCowan, B. and Koehl, P. (2013). Multi-scale clustering by building a robust and self correcting. *Ultrametric Topology on Data Points, PloS one*, **8**.

[5] Fushing, H., Kevin, F. and Cho-Jui, H. (2016). Machine learning meliorates computing and robustness in discrete combinatorial optimization problems. *Frontiers in Applied Mathematics and Statistics*, **2** , 1-8.

[6] Gacs, P., Tromp, J. and Vitanyi, P. M. B. (2001). Algorithmic statistics. *IEEE Transactions on Information Theory*, **47**, 2443-2463.

[7] Ghurye, S. G. (1968). Information and sufficient sub-fields. *Annals of Mathematical Statistics*, **39**, 2056-2066.

[8] Gray, R. M. (2009). *Entropy and information theory.* Springer.

[9] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, **1**, 1-7.

[10] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86.

[11] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses.* 3rd edition, Springer, New York.

[12] Li, M. and Vitanyi, P. M. B. (1997). *An introduction to Kolmogorov complexity and its applications.* 2nd ed. New York: Springer-Verlag.

[13] Longest, K. C. and Vaisey, S. (2008). fuzzy: A program for performing qualitative comparative analyses (QCA) in stata. *The Stata Journal*, **8**, 79-104.

[14] Markov, A. A. (1907). Issledovanie zamechatel'nogo sluchaya zavisimyh ispytanij. *Izvestiya Akademii Nauk, SPb, VI seriya*, **1**(**93**), 61-80.

[15] Markoff, A. (1910). Recherches sur un cas remarquable d'épreuves dépendantes. *Acta mathematica*, **33**(**1**), 87-104.

[16] Pop-Stojanovic, Z. R. (2006). A classroom note: entropy, information, and Markov property. *Teaching Mathematics*, **9**, 1-12.

[17] Ramu, P., Qu, X., Youn, B., Haftka R. T., and Choi, K. (2006). Safety factors and inverse reliability measures. *International Journal of Reliability and Safety*, **1**, 1-2.

[18] Shannon, C. E. (1948). The mathematical theory of communication. *The Bell System Technical Journal*, **27**, 623-656.

[19] Solomonoff. R. J. (1964). A formal theory of inductive inference, part 1 and part 2. *Information and computation*, **7**, 1–22.

[20] Tishby, N., Pereira, F. C. and Bialek, W. (1999). The information Bottleneck method. in proceeding of the 37th Annual Allerton Conference on Communication, *Control and Computing*, 368-377.

[21] Vereshchagin, N. K. (2009). *Algorithmic minimal sufficient statistic Revisited*. In CiE (pp. 478-487).

[22] Vereshchagin, N. (2016). Algorithmic minimal sufficient statistics: a new approach. *Theory of Computing Systems*, **58**, 463-481.

[23] Von Hilgers, P. and Langville, A. N. (2006). *The Five Greatest Applications of Markov Chains*. Proceedings of the Markov Anniversary Meeting. Boson Press.

[24] Wallace, C. and Freeman, P. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society*, **49**, 240–251.