SECUENCIAS TÍPICAS EN EL ANÁLISIS DE CAPACIDAD DE UN CANAL DE COMUNICACIÓN DIGITAL

Trabajo de Grado

María del Pilar Ramos Huila

Director: Víctor Manuel Quintero Florez

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telecomunicaciones
Grupo de Radio e Inalámbricas - GRIAL
Sistemas de Comunicaciones Móviles e Inalámbricos
Popayán, 2021

SECUENCIAS TÍPICAS EN EL ANÁLISIS DE CAPACIDAD DE UN CANAL DE COMUNICACIÓN DIGITAL

María del Pilar Ramos Huila

Trabajo de grado presentado a la Facultad de Ingeniería
Electrónica y Telecomunicaciones de la
Universidad del Cauca como requisito parcial para optar al título de:
Ingeniera en Electrónica y Telecomunicaciones

Director: Víctor Manuel Quintero Florez

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telecomunicaciones
Grupo de Radio e Inalámbricas - GRIAL
Sistemas de Comunicaciones Móviles e Inalámbricos
Popayán, 2021

AGRADECIMIENTOS

Agradezco a mi madre Felipa Camayo y a mi padre Jesús Collo por decidir ser mis padres y ayudarme a crecer, a mis hermanas y a mis sobrinas y sobrinos por su paciencia. A mis amigas y amigos, que estuvieron presentes en el desarrollo de mi carrera, y que con su compañía solidificaron mi carácter para trabajar en grupo, y reconocer la importancia de nuestro camino como ingenieros. Estoy profundamente agradecida con las personas que han estado en los momentos más difíciles, para brindarme su apoyo, y ayudarme a culminar mi carrera. Agradezco también al ingeniero Harold Romo quien me animo a no desistir en mi interés por un estudio teórico y me vinculo al ingeniero Víctor Quintero quien me acompaño en la responsabilidad que asumí de desarrollar este trabajo de grado.

RESUMEN

El presente trabajo de grado de investigación es un estudio en torno a la obra de Claude Elwood Shannon (1916-2001)[1], quien sentó las bases matemáticas y técnicas del desarrollo de la teoría de la información. C. E. Shannon en *"A Mathematical Theory of Communication"* [1] identificó y propusó las relaciones estadísticas que subyacen a los sistemas de comunicación, profundizando en el estudio de la transmisión de información realizado por H. Nyquist [2] y R. Hartley [3], y estableciendo los límites fundamentales de todo sistema de comunicación.

La teoría de la información estableció un modelo para los sistemas de comunicación y los conceptos que permiten medir la información y la velocidad de su transmisión a través de un canal de comunicación. El modelo está formado por cinco elementos: la fuente de información, el transmisor, el canal, el receptor y el destino, los cuales están representados por entidades abstractas con un soporte matemático.

La entropía de la información es una medida de la incertidumbre con la que una fuente genera información[2]. Existe un límite fundamental para la compresión de información en una fuente, el cual esta en términos de la entropía de la información. Este límite es establecido por el teorema de codificación de fuente.

Las secuencias típicas es un concepto de la teoría de la información que define la regularidad en la aleatoriedad de la información, que se obtiene considerando la Ley de los Grandes Números (LLN, *Law of Large Numbers*) de la teoría de la probabilidad, y que se analiza a través de la Propiedad de Equipartición Asintótica

[1] Claude Elwood Shannon (1916 - 2001) fue un matemático e ingeniero eléctrico estadounidense, quien estableció los fundamentos teóricos para los circuitos digitales y la teoría de la información.

[2] El significado de la palabra *información* en esta teoría, se desvincula del sentido común que esta tiene, el cual no corresponde con el concepto abstracto en ingeniería.

(AEP, *Asymtotic Equipartition Property*). La propiedad permite definir el conjunto de secuencias que tiene la mayor probabilidad de ser generadas en un transmisor o codificador.

La información mutua es una medida de la dependencia mutua entre los procesos estocásticos de entrada y salida del canal. Mide la reducción de la incertidumbre sobre la información de entrada al canal dada la información a la salida del canal. Existe un límite fundamental para la cantidad de información a transmitir en un sistema de comunicación, el cual es conocido como la capacidad. Este límite esta en función de la información mutua y es establecido por el teorema de codificación de canal, el cual considera las limitaciones físicas de los componentes del sistema de comunicación. Para la obtención de la capacidad del canal de comunicación es importante la aplicación del concepto de secuencias típicas y la propiedad de equipartición asintótica (AEP).

El planteamiento que realizo Shannon estimuló la investigación de matemáticos e ingenieros en un amplio campo disciplinar de la ciencia y de la ingeniería que involucra e incluye la matemática, la física, la teoría de la codificación, la criptografía, el aprendizaje de máquinas (*machine learning*) y la teoría de la información cuántica, entre otras, que extienden y profundizan el estudio y la aplicación de la teoría de la información.

TABLA DE CONTENIDO

TABLA DE CONTENIDO

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ACRÓNIMOS

AEP	*Asymptotic Equipartition Property* - Propiedad de Equipartición Asintótica
BSC	*Binary Symmetric Channel* - Canal Binario Simétrico
i.i.d.	*Independent and Identically Distributed* - Independientes e Idénticamente Distribuidas
LLN	*Law of Large Numbers* - Ley de los Grandes Números
PMF	*Probability Mass Function* - Función de Distribución de Probabilidad
SLLN	*Strong Law of Large Numbers* - Ley de los Grandes Números Fuerte
WLLN	*Weak Law of Large Numbers* - Ley de los Grandes Números Débil
PCM	*Pulse Code Modulation* - Modulación por Pulsos Codificados
VOCODER	*Voice Coder* - Codificador de Voz
FM	*Frequency Modulation* - Modulación de Frecuencia

NOTACIÓN

Los conjuntos se denotan por letras mayúsculas caligráficas: $\mathcal{X}, \mathcal{Y}$. El número de elementos en el conjunto $\mathcal{X}$ es denotado por: $|\mathcal{X}|$. Las variables aleatorias se indican con letras mayúsculas: X y Y. La realización de la variable aleatoria X le corresponde un resultado denotado por x, que pertenece al conjunto $\mathcal{X}$, llamado alfabeto de la variable aleatoria. En el análisis se consideran únicamente variables aleatorias discretas. La función de distribución de probabilidad (pmf) de la variable aleatoria discreta X se denota por: p_X. La probabilidad de x se denota por $p_X(x)$, que corresponde a:

$$p_X(x) = \Pr\{X = x\}. \tag{1}$$

El soporte de $p_X(x)$ es $\mathcal{S}_X$, donde $\mathcal{S}_X = \{x \in \mathcal{X} : p_X(x) > 0\}$. Considerando dos variables aleatorias X y Y, la función de distribución de probabilidad (pmf) conjunta de (X, Y) se denota por $p_{XY}(x, y)$ y la función de distribución de probabilidad (pmf) condicional de Y dado X por $p_{Y|X}(y|x)$. El valor esperado con respecto a la variable aleatoria X se denota por el operador $\mathbb{E}_X[\cdot]$ y la varianza de la variable aleatoria X se denora por el operador $Var[\cdot]$.

Sea $n \in \mathbb{N}$, una secuencia de n variables aleatorias independientes e idénticamente distribuidas (i.i.d.) se representa por un vector de dimensión n denotado por $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$, y la realización correspondiente es denotada por $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T \in \mathcal{X}^n$. La pmf conjunta se define como:

$$p_{\mathbf{X}}(\mathbf{x}) = \Pr\{\mathbf{X} = \mathbf{x}\} = \prod_{i=1}^{n} p_X(x_i). \tag{2}$$

Se asume que la función logaritmo log es en base 2.

El símbolo $\lceil a \rceil$ es a redondeado al entero superior más cercano, i.e, el entero más pequeño que es más grande o igual que a. Y $\lfloor a \rfloor$ es a redondeado al entero inferior más cercano, i.e., el entero más grande que es más pequeño o igual a a.

INTRODUCCIÓN

El trabajo de grado en modalidad de investigación presenta un análisis introductorio de la teoría de la información y los conceptos básicos como entropía, secuencias típicas, información mutua y capacidad, con el objetivo de ser expuestos de manera clara y didáctica.

La teoría de la información es la ciencia de los sistemas abstractos de comunicación, y su objeto es la transmisión confiable y óptima de la información a través de un canal de comunicación. Shannon fundo las bases de la teoría de la información en 1948, con su artículo *"A Mathematical Theory of Communication"* [1], en el cual definió el problema de la comunicación de la siguiente manera: *"El problema fundamental de la comunicación es el de reproducir en un punto, exacta o aproximadamente un mensaje seleccionado en otro punto".*

La definición anterior se puede entender del siguiente modo: tomados dos puntos, uno considerado la fuente y el otro el destino, cómo lograr que un mensaje generado en la fuente sea replicado en el destino exacta o aproximadamente. Dada la comunicación en un escenario real, el mensaje transmitido puede no ser reconocible en el destino dadas las limitaciones del sistema o por diversas variables en el canal de comunicación.

Aunque la comunicación interesa a todo ser humano, los mecanismos técnicos y tecnológicos que dan solución a este problema y que han extendido la red de comunicación dentro y fuera de la tierra, son el objeto de diseño teórico y tecnológico que científicos e ingenieros desarrollan y perfeccionan, dando lugar a los modernos sistemas de comunicación digital.

La comprensión de la base teórica de los sistemas de comunicación, es el elemento que guía el desarrollo de este trabajo de investigación, el cual se fundamenta en la teoría de la probabilidad y la teoría de la transmision de H. Nyquist [2, 4] y R. Hartley [3], quienes sentaron las bases para la teoría de la información, relacionando conceptos de diferentes campos disciplinares: la ingeniería, la ciencia, la matemática y la computación.

A partir de los fundamentos precedentes a la teoría de la información, se quiere responder a las preguntas ¿cómo se ha originado?, ¿cuáles fueron sus causas?, y ¿qué es lo que realmente significa?, preguntas que científicos como Schrödinger [5] entre otros[3], se plantearon para explicar los fundamentos de la ciencia moderna.

Conceptos fundamentales como entropía se abordan desde su definición inicial en la termodinámica y su posterior redefinición en la teoría de la información. La información se estudia desde los planteamientos de H. Nyquist [2], R. Hartley [3] y Shannon [1] quienes la caracterizan como información digital. Otro de los conceptos aquí analizado es el de las secuencias típicas, el cual es un pilar fundamental en el análisis del comportamiento aleatorio de la información, y que permite estimar el comportamiento en el régimen asintótico de las secuencias generadas por una fuente de información discreta, y definir para el canal discreto sin ruido y con ruido una velocidad máxima de transmisión de información o capacidad.

Shannon en *"A Mathematical Theory of Communication"* [1] postula dos teoremas como métodos de análisis de sistemas de comunicación: el teorema de codificación de fuente y el teorema de codificación de canal, en los cuales introduce los conceptos entropía, secuencias típicas, información mutua y capacidad de canal. En el teorema de codificación de fuente, la entropía es la medida de información, y determina la máxima compresión de la representación de la información sin pérdidas. El teorema de codificación de canal establece un límite para la cantidad de información a transmitir en un sistema de comunicación afectado por ruido con una probabilidad de error arbitrariamente pequeña y cercana a cero, es decir, que se puede garantizar la transmisión de la información de manera confiable si la velocidad de generación de información de la fuente es menor que la capacidad del canal.

[3]Bertrand Rusell [6], James Jeans [7].

Para explicar el teorema de la teoría de la información, llamado *teorema de codificación de canal*, se presenta el análisis inicial que propone Shannon en [1], y se considera el desarrollo teórico posterior de diferentes autores para formalizar la demostración de este teorema [8–10]. Shannon realiza la demostración del teorema afirmando que el teorema se cumple para un código que debe existir dentro de un grupo de códigos, es decir, sin especificar una solución en particular, dada la prueba de que existe tal código que satisface el teorema, lo cual abrió un campo de investigación en la búsqueda de esos códigos que intenten alcanzar ese límite fundamental.

Este trabajo de investigación amplia la visión de cómo están unidas las invenciones teóricas y técnicas (relacionadas a los problemas de la comunicación), a través de la descripción matemática o probabilística de los elementos físicos que componen un sistema de comunicación.

El documento consiste de cuatro partes, que se describen a continuación:

- Parte I. Se inicia con la pregunta acerca de la comunicación, considerada esta como una actividad común en todos los seres vivos, seguida del análisis de los conceptos como entropía e información desde las teorías previas a la teoría de la información: física, matemática y teoría de la transmisión [2–4], expuesta en el primer capítulo:
  - **Capítulo 1.** Se plantea el objeto de estudio de la teoría de la información, el cual es el proceso de comunicación, realizando la diferenciación de la comunicación en todos los seres vivos y de la comunicación humana, con sus variantes y procesos sociales que determinan el desarrollo técnico y el avance tecnológico. Se realiza un perfil epistemológico del término entropía, dado que este concepto en la teoría de la información es fundamental para el planteamiento del teorema fundamental de la comunicación, con antecedentes en la termodinámica por R. Clausius [11] y en la mecánica estadística con L. Boltzmann [12], que para la aclaración teórica son planteados desde su conceptualización física. Se introducen los fundamentos que dieron paso a una teoría general del proceso de comunicación. Entre ellas, los avances en la teoría del cálculo y la probabilidad, el concepto de límite fundamental, y los desarrollos técnicos de

las comunicaciones previos a las investigaciones de Shannon realizado por H. Nyquist [2], R. Hartley [3] y N. Wiener [13, 14], quienes introdujeron términos como *intelligence* y *information*. Con este capítulo se examinan las fuentes [2, 3, 13] que fueron sintetizadas por Shannon.

- Parte II. Se desarrollan los conceptos definidos en la teoría de información para el sistema de comunicación discreto y sin memoria: entropía, secuencias típicas, la Propiedad de Equipartición Asintótica (AEP, *Asymptotic Equipartition Property*) y la capacidad de canal. Se presentan ejemplos que acompañan la definición de estos conceptos y transmiten las ideas que subyacen a ellos. Se plantean los dos teoremas fundamentales de la teoría de la información y se desarrolla la prueba matemática del teorema de codificación de canal para el sistema de comunicación discreto con ruido: la parte directa (*achievable*) y la parte inversa (*converse*). La segunda parte esta compuesta por los siguientes capítulos:

  - **Capítulo 2.** Se centra en introducir los conceptos que desarrolla Shannon para definir la teoría de la información, entre ellos, la definición de capacidad de canal para el sistema discreto sin ruido, entropía y el teorema de codificación de fuente. La definición de capacidad de canal en [1], se desarrolla para el sistema discreto sin ruido del teletipo y la telegrafía, y se presenta el cálculo que realizó Shannon para encontrar la capacidad de estos sistemas, mediante el conteo de secuencias. Se desarrolla el concepto de la entropía empírica, la entropía teórica y de las secuencias típicas que permiten profundizar en los métodos que utilizo Shannon para plantear sus dos teoremas. Estos se basan en la convergencia en probabilidad, en la Ley de los Grandes Números (LLN, *Law of Large Numbers*) y la AEP.
  - **Capítulo 3**. Se desarrollan los conceptos de tipicalidad conjunta de vectores de variables aleatorias discretas, y la tipicalidad en sentido fuerte.
  - **Capítulo 4.** Se introducen los conceptos de información mutua, secuencias típicas conjuntas, y se desarrolla el teorema de capacidad de canal de un sistema de comunicación discreto con ruido. Se presenta una idea general intuitiva de la prueba y el posterior desarrollo de la prueba directa (*achievability*) y la prueba inversa (*converse*) del teorema. Aquí se

introduce el concepto de probabilidad de error y se define la capacidad de canal como la maximización de la información mutua en función de las distribuciones de probabilidad de la fuente de información.

- Parte III. Se presentan las conclusiones del trabajo de investigación en el capítulo 5.
- Parte IV. Se desarrolla la formalización matemática de los diferentes conceptos que se abordan en el desarrollo de todo el documento.
  - Apéndice A. Se presentan los conceptos básicos de la teoría de la probabilidad.
  - Apéndice B. Se presenta la definición de entropía con sus diferentes propiedades y respectivas pruebas.
  - Apéndice C. Se desarrolla la solución de las ecuaciones de diferencias finitas, mediante el método de cálculo de la ecuación característica y de sus raíces reales.
  - Apéndice D. Se presentan los diferentes tipos de convergencia en probabilidad, la Ley de los Grandes Números Débil (WLLN, *Weak Law of Large Numbers*) y Ley de los Grandes Números Fuerte (SLLN, *Strong Law of Large Numbers*), con su respectiva demostración. Se presentan también las desigualdades más importantes en la teoría de la información necesarias para el desarrollo de la prueba del teorema de codificación de canal.

Capítulo 1

FUNDAMENTOS PREVIOS

1.1. MECANISMOS DEL PROCESO DE COMUNICACIÓN

La comunicación es un proceso presente en la naturaleza de todos los seres vivos. Las diferentes criaturas que existen en el mundo han desarrollado sus propias formas de comunicación. Entre ellas, la comunicación a través de mecanismos químicos, de emisión de frecuencias particulares, y el intercambio de información entre los diferentes procesos naturales como la fotosíntesis, o entre la membrana y el núcleo por medio de la electricidad.

Los individuos humanos basan su comunicación en diferentes formas de lenguaje y esta se extiende a todos los desarrollos científicos y técnicos que han construido y construirán. Actualmente, la exigencia de un mundo globalizado demanda una comunicación efectiva a larga distancia, por lo cual se estudian e investigan en profundidad todas las formas y mecanismos de comunicación existentes.

¿En qué consiste el proceso de la comunicación humana? Diferentes respuestas se han dado a esta pregunta, con las que se han diseñado teorías desde las ciencias humanas, tales como: la lingüística, la teoría de la acción comunicativa o la filosofía positivista del lenguaje. No obstante, el desafío de responder a esta pregunta

vincula directamente a las ciencias matemáticas y a los desarrollos tecnológicos que acrecientan las explicaciones y las respuestas prácticas que necesita un mundo globalizado, desarrollando teorías y procedimientos que garantizan una explicación y una solución técnica a la comunicación.

El término comunicación amplía su significado cuando se propone una solución científico técnica de los procesos comunicativos. Por lo cual, en realidad, se está frente a los problemas de transmisión de información y los mecanismos que conllevan a la circulación, al movimiento, a los medios y a los canales por donde se envían los mensajes. La comunicación es entendida como un proceso mediante el cual se organizan y se clasifican conjuntos cada vez más amplios de datos. La exigencia de almacenamiento, procesamiento e intercambio de estos, desafía al ingenio humano, el cual asume el reto de dar una solución científica y técnica a este problema cada vez más creciente.

El intento de los pueblos originarios por lograr la comunicación fue resuelto utilizando herramientas que extendieron su capacidad comunicativa, por ejemplo, en las Islas Canarias el silbido fue utilizado para comunicarse a largas distancias y permitió establecer una conversación entre dos puntos distantes, ya que era similar a la lengua propia, tradición que se mantiene hasta esta época [15]. El sonido de los tambores fue otra herramienta que extendió la voz humana, utilizada principalmente en el África [16]. Los mensajes más primitivos en las tribus de nativos americanos tales como: llamados a estar atentos, o señales de que algo estaba sucediendo, lo indicaban mediante señales de humo, por ejemplo, dos nubes de humo indicaban que todo estaba bien y tres significaban peligro o ayuda; además de estas, podía haber otros mensajes con una codificación más exclusiva entre fuente y destino. En China también fueron utilizadas las señales de humo y de fuego en las noches para dar advertencias sobre la cercanía de un enemigo, y el mensaje era retransmitido de la misma forma a lo largo de la Gran Muralla, lo cual alertaba de forma inmediata [17]. En Grecia la señal de humo se codificaba con un sistema de diez antorchas, separadas en dos grupos como se observa en la Fig. 1.1 (el primer grupo indicaba la fila y el segundo grupo la columna). El sistema fue diseñado por el historiador griego Polibio[1], el cual permitió comunicar palabras y oraciones a largas distancias [18]. En

[1]Polibio de Megalópolis (200 a. C.-118 a. C.) fue un historiador griego.

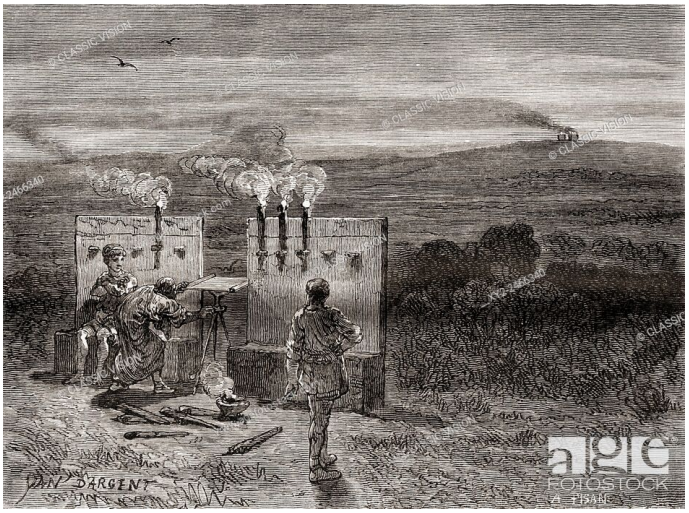
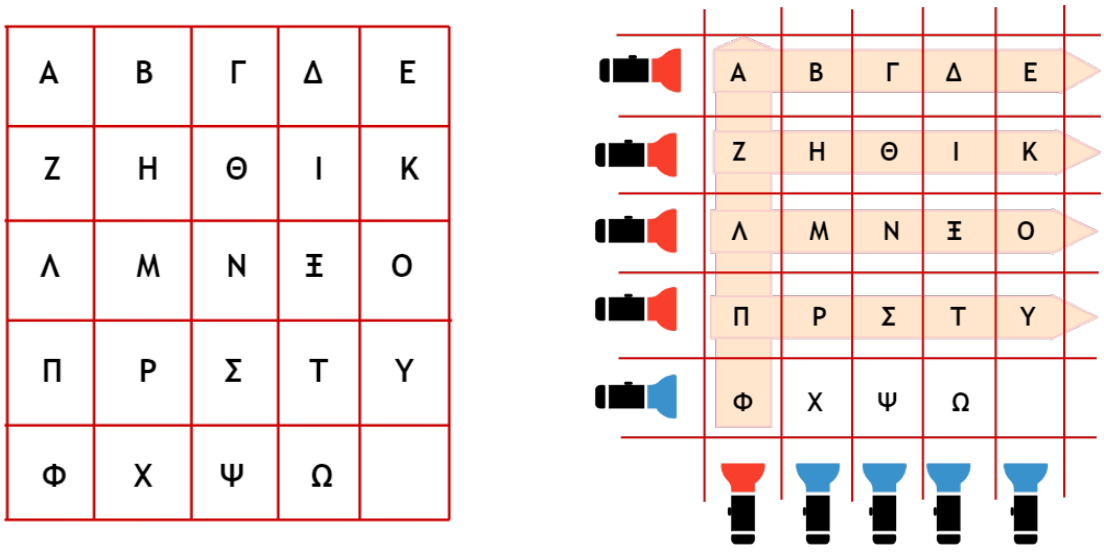


Figura 1.1: Sistema de Polibio. Tomado de: From Les Merveilles de la Science, published c. 1870.

la Fig. 1.2b las antorchas encendidas están representadas de color rojo, y las azules indican que las antorchas están apagadas. El código de Pi se halla encendiendo cuatro antorchas del primer grupo y una sola del segundo.

La mensajería era otro método de comunicación a distancia en diferentes civilizaciones antiguas. En Egipto y Grecia las palomas ofrecían en tiempos de guerra una comunicación más veloz, por su sentido de orientación y la capacidad de volar largas distancias, llevando mensajes cruciales atados a sus patas; de los musulmanes se dice que establecieron un servicio postal de palomas en todo Oriente Medio [17, 19]. Su papel a lo largo de la historia de la humanidad ha quedado grabada en tablillas de la antigua Sumeria [20]. Además de las palomas, la mensajería también fue realizada por humanos, por ejemplo, el imperio Persa utilizaba un mecanismo conocido como sistema de relevos, el cual consistía de un sistema de caminos, que diferentes jinetes iban recorriendo por tramos pasándose el mensaje desde el origen hasta el destino, siendo un sistema utilizado solo por el rey, en el cual la velocidad era relevante debido a que los caminos estaban despejados por decreto. Los Incas contaban con mensajeros especializados llamados Chasquis, los cuales se caracterizaban por ser atletas y conocer muy bien los caminos que facilitaban el encuentro con otro Chasqui, y de esta manera completar la entrega del mensaje [21].



(a) Cuadro de Polibio (b) Código de Π

Figura 1.2: Codificación de Π

Los ejemplos anteriores sirven para ilustrar e introducir que, desde épocas antiguas, existían algunos problemas técnicos propios de toda comunicación, entre otros, el medio a través del cual se transportaba el mensaje, las rutas o canales y la velocidad. La comunicación ha estado ligada siempre a la relación tiempo y espacio, porque se trata del envío del mensaje de un lugar a otro en un tiempo determinado, lo cual expresa la necesidad de fusionar el espacio y tiempo, hasta conquistar el ideal de todo proceso de comunicación, que consiste en la simultaneidad entre emisión y recepción [1].

En los albores de la humanidad, la información era transmitida de manera oral, lo que limitaba la reproducción y conservación de los mensajes. La escritura compensó de alguna manera este problema y la imprenta en el siglo XV, dio paso a la mecanización de la reproducción y conservación de información, a través de los libros y la difusión escrita. Quizá la escritura es el comienzo de una respuesta técnica al problema de la comunicación [22].

Aunque el problema fundamental era resolver el desarrollo técnico de la comunicación sobre bases científicas, no fue sino hasta el siglo XVIII cuando se planteó la posibilidad de usar la electricidad estática para la transmisión de mensajes usando

hilos conductores, como lo expresa el artículo: "*An Expeditious Method of Conveying Intelligence*" [23], el cual apareció en Escocia en 1753 de un autor que se conserva en el anonimato bajo el seudónimo C.M., el cual propone un sistema de telegrafía eléctrica compuesto de una fuente de electricidad, un dispositivo de procesamiento de la información a transmitir, cables conductores, y un mecanismo receptor para detectar la información transmitida, garantizando una alta velocidad de la transmisión de información gracias a las propiedades de las señales eléctricas, lo cual fue un anticipo al esquema de comunicación propuesto por Shannon [19].

Se hace necesario comentar algunos aspectos interesantes en [23], y es el uso de la expresión *expeditious method*, que expone uno de los temas y variables de la comunicación que acompaña toda la existencia del intercambio de información humana desde sus inicios: la necesidad de encontrar formas más rápidas y eficientes de transmitir el mensaje de un lugar a otro. Del mismo modo, *conveying* es el elemento intrínseco en todo proceso de comunicación. De esta manera, el título describe el futuro planteamiento del problema que la ciencia y la ingeniería buscan resolver. Y por último, comunicación e información entendidas como el intercambio que hacen los humanos de sus necesidades espirituales, psicológicas, fisiológicas, políticas, culturales y económicas, entre otras, en suma, lo que en [23] llama: *Intelligence*.

Si bien experimentar el fenómeno de la electricidad para esa época era un entretenimiento [24], la fabulosa expectativa que creó la experiencia eléctrica dio una respuesta alternativa a las necesidades socioeconómicas, políticas y militares, cuyas fuerzas permitieron la implementación del telégrafo óptico diseñado por Claude Chappe[2] en 1793 en Francia, en plena Revolución Francesa, el cual se observa en la Fig. 1.3, y del telégrafo eléctrico diseñado por Samuel Morse[3] en 1837 en Estados Unidos, estableciendo un enlace por línea física desde Washington hasta Baltimore de alrededor de 64 Km. Paralelamente se desarrollaron los estudios teóricos de la electricidad, la electroquímica y el electromagnetismo, que dieron bases sólidas a la telegrafía eléctrica e impulsaron otros desarrollos técnicos como la telefonía en 1876 y la telegrafía inalámabrica a través de ondas de radio en 1897. Todas las formas de expresión humana fueron encontrando medios técnicos de difusión gracias a estas

[2]Claude Chappe (1763 - 1805) fue un inventor francés.
[3]Samuel Finley Breese Morse (1791 - 1872) fue un pintor e inventor estadounidense.

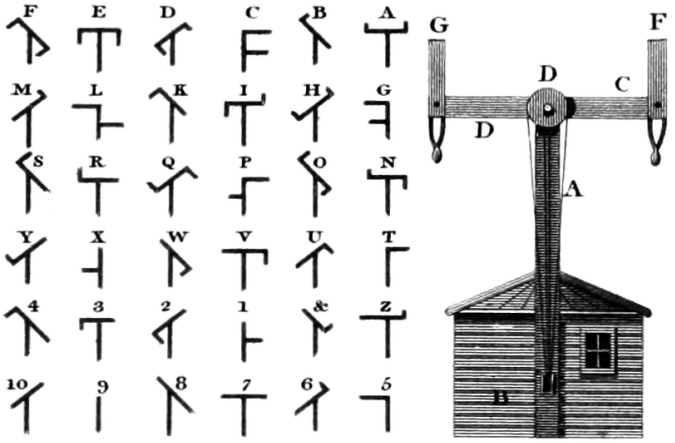


Figura 1.3: Telégrafo Óptico. Tomado de: John Farey, Jr., Public domain, via Wikimedia Commons HTML

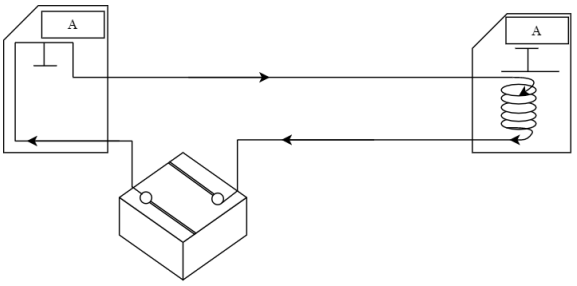
investigaciones y experimentaciones [22, 25].

Existe una correspondencia entre el desarrollo teórico, el avance técnico/científico y el poderío económico que es importante resaltar a la hora de pensar las razones que llevan a las naciones a proteger las investigaciones tecnológicas, compitiendo por la supremacía científico técnica. Se hace necesario enunciar que las investigaciones mencionadas se distribuyen en países como: Reino Unido, Francia, Suecia y Estados Unidos, sin tener en cuenta los desarrollos paralelos que pudieron haber ocurrido en los países euro-orientales y asiáticos.

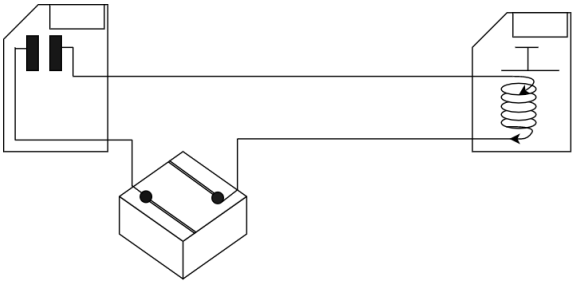
Toda respuesta técnica o tecnológica a necesidades humanas tiende a masificarse, por lo cual en poco tiempo las redes de telegrafía se implantaron y se extendieron en cada país y se impuso como recurso comunicativo en el mundo entero. Esto explica el por qué en desarrollos posteriores de los medios y canales de transmisión de información, rápidamente se extendieron cables de comunicación entre los continentes, como aconteció con el cable submarino en 1851, el cual conectaba el Reino Unido y Francia a través del Canal de la Mancha, e impulsó el desarrollo de cables trasatlánticos. Espacio, tiempo y frecuencia son conceptos determinantes a la hora de pensar las comunicaciones en un mundo que cada vez demanda un mayor número de conexiones. La telegrafía, el telegráfono, fueron respuestas tempranas a estas necesidades globales [22].

Figura 1.4: Teletipo.

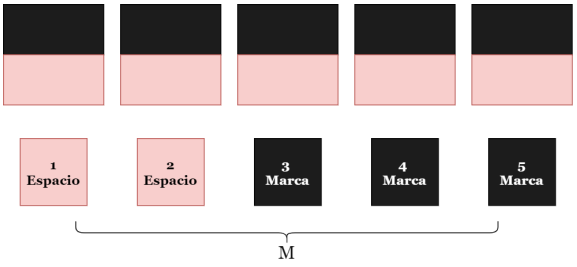
El siglo XX irrumpe con el desarrollo de la electrónica, con lo cual se habla ya de las tecnologías de la comunicación que relegan al pasado los importantes pero insuficientes progresos previos de los sistemas de comunicaciones, en términos tecnológicos. Por ejemplo, el teletipo desarrollado en 1912, expresó una nueva modalidad de transmisión y recepción de mensajes en un sistema de comunicación punto a punto. El sistema de teletipo consistía de una máquina de transmisión, como la que se observa en la Fig. 1.4, que producía pulsos eléctricos por la acción mecánica que sigue a la presión de teclas similares a las de una máquina de escribir, los cuales eran enviados a través de una línea telegráfica hasta una máquina de recepción, la cual convertía los pulsos eléctricos en una acción mecánica que imprimía letras o figuras en papel. El pulso eléctrico se puede describir como un flujo de corriente a través de la línea telegráfica. Cada letra o caracter era codificado utilizando el código de Baudot, el cual estaba definido por una combinación de cinco pulsos eléctricos de igual duración, con lo cual se podían representar hasta 32 caracteres. Debido a la necesidad de sincronización entre las máquina de transmisión y recepción, el código fue modificado para incluir caracteres de control que delimitaran el inicio y final de cada transmisión. Este código se observa en la Fig. 1.6. Posteriormente, el teletipo fue unido a equipos de perforación de cintas de papel, lo cual abrió la posibilidad de grabar los mensajes en la cinta antes de pasar por la línea telegráfica, para luego enviar las señales automáticamente al introducir la cinta en la máquina de trans-



(a) Cuando el contacto en la línea de la señal entre los dos dispositivos está cerrado, la corriente fluye y se dice que el pulso eléctrico es una marca.



(b) Cuando el contacto en la línea de la señal entre los dos dispositivos está abierto, la corriente no fluye y se dice que el pulso eléctrico es un espacio.



(c) Código de M. Cada carácter consiste de cinco pulsos eléctricos que pueden ser marcas o espacios.

Figura 1.5: Representación del Teletipo.

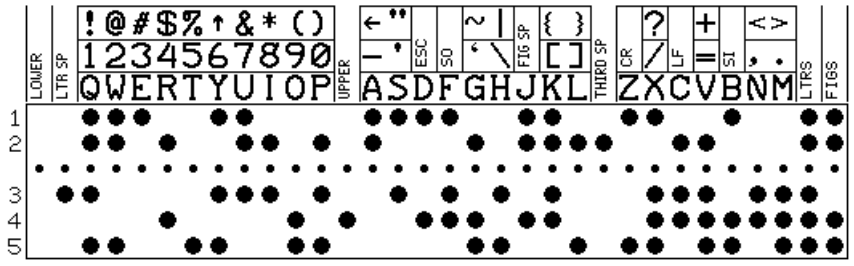


Figura 1.6: Código de Baudot, Alfabeto Internacional de Telegrafía Nº 2, ITA2.

misión. Las cintas perforadas también se generaron en recepción. El uso de cintas perforadas permitió una transmisión más rapida y eficiente de los mensajes debido a la automatización que el sistema lograba, en comparación a la impresión y repetición de forma manual en las estaciones teletipo previas [22, 26]. Con el teletipo y demás dispositivos de comunicación se establecieron redes basadas en la conmutación electrónica. A mediados del siglo XX la humanidad ya contaba con satélites y comunicaciones digitales, y en poco tiempo, la computación y posteriormente el internet dieron un impulso sin precedentes a lo que se llama la tercera ola de la revolución industrial [22].

El estudio de los fenómenos naturales como el electromagnetismo, vinculados a desarrollos paralelos y confluentes de la matemática como el cálculo y las probabilidades, han ido fortaleciendo la relación teoría y práctica, cuya manifestación más relevante es la ingeniería, que ha tenido la tarea de resolver los problemas y brindar las soluciones a partir de los adelantos promovidos desde la física y la matemática, que encuentran en Shannon el referente más importante en el campo de la teoría de las comunicaciones obtenida en un diálogo permanente entre matemática, probabilidad, termodinámica e ingeniería.

1.2. CONCEPTO DE ENTROPÍA

"I propose to call the magnitude S the entropy of the body, from the greek word ($\tau\rho\omega\pi\eta$) transformation. I have intentionally formed the word entropy so as to be as similar as possible to the word energy"

Rudolf Clausius

La ciencia se ha desarrollado a partir del método científico en el que se busca la información, se establecen hipótesis, comparaciones, comprobaciones y experimentaciones que confrontan las teorías propuestas y conducen a la exposición de teorías, intuiciones e ideas. En el proceso de producción de conocimientos, los investigadores relacionan diferentes disciplinas científicas y construyen conceptos nuevos de dos formas: a partir de nociones y categorías de uso común o que están inscritas en el desarrollo de otras disciplinas. En este último caso, el investigador al notar afinidades en las problemáticas, importa conceptos de otras ciencias que le pueden servir para explicar fenómenos distintos, aplicados en campos nuevos de la investigación científica y que requiere de la invención de nuevos conceptos que engloben semánticamente los fenómenos estudiados. En el primer caso, cuando las categorías y las nociones a las que recurre el investigador hacen uso de palabras con significados comunes, se debe tener en cuenta que las relaciones en que se sitúa la palabra, constituyen la imposición de un uso distinto a la definición común del término, asignándole a la palabra un significado no común sino conceptual.

En el caso de la palabra $\varepsilon\nu\tau\rho\omega\pi\alpha$[4] de origen griego, del ático antiguo, cuyos significados comunes son: girar, dar vuelta, o transformar, el análisis semántico que delinea los matices que el término asume o los diversos significados que los investigadores le proponen se llama perfil epistemológico [24], el cual se desarrolla brevemente en esta sección.

Fue Rudolf Clausius[5] quien introdujo la palabra entropía en el contexto del análisis

[4]Entropía.

[5]Rudolf Julius Emmanuel Clausius (1822 -1888) fue un físico y matemático alemán, y uno de

de la ciencia Física, específicamente en la problemática científica de la disciplina que estudia el calor: la termodinámica. Clausius comenzó tomando la palabra en su sentido etimológico, eligiendo como sinónimo de la palabra entropía, el verbo transformar. Este sentido entrañó dos aspectos relevantes para Clausius: primero, transformar es mutar; segundo, toda mutación tiene una causa que la produce. Para Clausius, la relevancia de la palabra transformación era su vínculo con el concepto fundamental de la Física que es la energía. ¿Pero de qué tipo de energía habló Clausius? Según sus análisis cuando se produce un movimiento o un cambio cualquiera fuera del tipo que fuere en un sistema por la aplicación de alguna forma de energía sobre este, una parte de la energía es utilizada para el fin propuesto i.e., el trabajo útil, y al mismo tiempo, se genera una liberación de energía, i.e., una energía no utilizada en el proceso de transformación de un tipo de energía en otro, a la cual Clausius denominó entropía [11, 27, 28].

Clausius tomó la palabra entropía con su significado original, pero sutilmente, a partir de su significado *transformar*, dio un viraje desviando el significado y contextualizó un nuevo sentido en el campo de la energía, del trabajo y del movimiento. Cuando un científico realiza tales movimientos, que fijan nuevos conceptos de términos comunes variando su sentido original, para poner la palabra al servicio de nombrar un fenómeno -físico en este caso- que conceptualiza y enriquece el conjunto de nombres que constituyen la teoría científica.

Posteriormente en la mecánica estadística, Ludwig Boltzmann desarrolló otro aspecto de la entropía [12], el cual se expondrá más adelante. Surge la siguiente pregunta en el contexto de la teoría de la información: ¿Cómo ocurre el cambio conceptual que lleva a la palabra entropía del análisis físico a un concepto fundamental de la teoría de la información? Para responder a esta pregunta objeto de la presente sección, se hace necesario describir detenidamente lo que Clausius intentó decir cuando habla de entropía, dado que esa descripción permite realizar comparaciones con los usos que se hizo posteriormente del mismo término.

los fundadores de la termodinámica. Entre sus aportes, se encuentra la enunciación del segundo principio de la termodinámica.

1.2.1. Segunda Ley de la Termodinámica

La invención de las máquinas a vapor o máquinas térmicas representó un momento de avance técnico, el cual dio soporte a la revolución industrial, y se basaban en el fenómeno térmico de transferencia de calor y transformación en trabajo. Entre las diferentes invenciones que se dieron, la máquina a vapor de James Watt, es reconocida por ser más práctica y económica en el consumo de energía. Este avance motivo las siguientes preguntas: ¿Cómo maximizar el desempeño de las máquinas térmicas?, ¿Cuál era el proceso más adecuado en la conversión de calor en trabajo? y, ¿Cuál sería la mejor sustancia para está conversión?. Las ideas básicas de la ciencia que abordarón el problema de la generación de trabajo mediante calor fueron establecidas por S. Carnot[6]. La conceptualización del calor Q como una forma de energía, lo cual dio origen al estudio de la energía térmica y su transformación en energía mecánica o viceversa, fue realizada gracias a los trabajos teórico-experimentales del Conde de Rumford, B.Thomson y J. Prescott Joule.

La termodinámica clásica define tres importantes leyes en el estudio del modelo continuo de la materia[7], es decir, no entra el análisis atómico-molecular, lo cual es estudiado por la mecánica estadística. Clausius desarrolló la segunda ley y el concepto de entropía. Para llegar al planteamiento de Clausius de la entropía es necesario presentar la primera ley de la termodinámica, la cual aplica el principio de conservación de energía[8], relacionando la energía térmica interna de un sistema U, el calor Q y el trabajo[9] W, que intervienen en el sistema, lo cual se define matemáticamente como:

$$\Delta U = Q - W, \tag{1.1}$$

donde, ΔU es el cambio de energía interna (transformación de energía), Q el calor total transferido al sistema y W es el trabajo neto realizado por el mismo, como se observa en la Fig. 1.7.

[6]Nicolas Léonard Sadi Carnot (1796 - 1832) fue un físico e ingeniero francés pionero en el estudio de la termodinámica. Se le reconoce hoy como el fundador o padre de la termodinámica.

[7]El modelo continuo de la materia hace referencia al estudio de la materia, tal que, al dividir sus partes no cambian sus propiedades, este modelo fue posteriormente descartado [29].

[8]La energía no se crea ni se destruye solo se transforma.

[9]Trabajo: energía que se transfiere mediante una fuerza que genera movimiento.

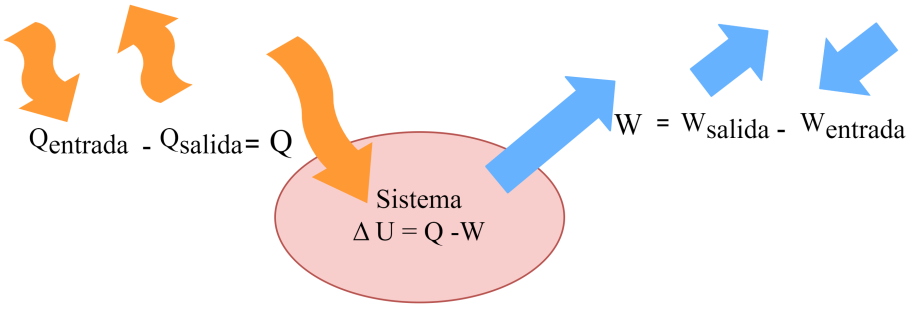


Figura 1.7: Primera Ley de la Termodinámica.

De (1.1) se derivan dos argumentos, primero, que todas las formas de energía son intercambiables, por tanto, todas pueden utilizarse para realizar trabajo, y segundo, que no es posible convertir toda la energía disponible en trabajo.

Clausius en *The Mechanical Theory of Heat* [11] desarrolló el estudio de la eficiencia de las máquinas térmicas, para lo cual, analizó la dirección del intercambio de calor en los procesos espontáneos, planteando el siguiente axioma: "el calor no puede pasar por sí mismo de un cuerpo frío a un cuerpo caliente". A partir de este axioma se pueden diferenciar dos tipos de procesos: procesos irreversibles y procesos reversibles. Los primeros no dependen de la dirección, ya que solo ocurren en un sentido, por ejemplo, según el argumento de Clausius, un objeto frío en contacto con uno caliente nunca se volverá más frío, transfiriendo calor al objeto caliente y haciéndolo más caliente. La primera ley de la termodinámica definida en (1.1) no interviene en el sentido de la conversión de las formas de energía, afirmando que el aumento en una forma de energía debe estar asociado a la disminución de otra forma de energía. Si se considera una piedra rodando, sus energías cinética, potencial o elástica se convierten en trabajo o calor disipado. En el caso contrario, si la piedra está en reposo, no es posible que se enfríe de manera espontánea convirtiendo ese calor desalojado en movimiento (energía cinética) [30].

La ley que describe está prohibición en la naturaleza fue desarrollada por Clausius, considerando dos ciclos de Carnot[10] que funcionan en el mismo rango de tempera-

[10]Es un ciclo ideal reversible entre dos fuentes de calor, por ejemplo, el motor térmico realiza un ciclo de Carnot si intercambia calor entre dos fuentes con diferentes temperaturas. Los procesos cíclicos de Carnot son siempre reversibles lo cual permite la máxima transformación de calor en trabajo.

tura, uno como motor térmico[11] y el otro como refrigerador[12] , donde el refrigerador consume el trabajo que proporciona el motor térmico. En su análisis compara las cantidades de calor Q que se intercambian, de lo cual concluyó lo siguiente: "la eficiencia de un ciclo de Carnot es máxima y universal". Máxima porque ningún ciclo en el mismo rango de temperatura tiene una eficiencia mayor que un ciclo de Carnot, y universal porque trabajando con cualquier sustancia produce la misma eficiencia. Del análisis de la eficiencia para el ciclo en Carnot de un gas ideal, Clausius encontró que [27]:

$$\left.\frac{Q}{T}\right|_{motor} = \left.\frac{Q}{T}\right|_{refrigerador}, \tag{1.2}$$

donde, la relación Q/T no cambia a través del ciclo, la cual fue definida como entropía y denotada como S por Clausius. La variación de la entropía ΔS para un proceso reversible, es:

$$\Delta S = \frac{Q}{T}, \tag{1.3}$$

donde ΔS es válida solo para procesos reversibles tales como los representados por el ciclo de Carnot. El desarrollo de Clausius independizó los ciclos de Carnot, y probó que para un proceso arbitrario, la velocidad de cambio de la entropía, denotado por dS, satisface la desigualdad:

$$dS \geq \frac{dQ}{T}, \tag{1.4}$$

donde, dQ es la variación infinitesimal del calor bajo la convención de que, si $dQ > 0$, el sistema absorbe calor (que se convertirá en trabajo), y si es $dQ < 0$, el sistema cede calor (desalojado), con referencia a la temperatura absoluta T. La relación en (1.4) es conocida como la segunda ley de la termodinámica. En [11], la naturaleza reversible del proceso entre dos estados de equilibrio es definida por la igualdad en (1.4), de lo contrario el proceso es irreversible. Así, el diferencial de entropía en un proceso reversible, es:

$$dS = \frac{dQ}{T}, \tag{1.5}$$

donde, dQ es el calor transferido y T la temperatura neta del sistema. Por lo tanto,

[11]Es un sistema que usa la transferencia de calor para hacer trabajo, tales como: motores de gasolina y Diesel, motores jet, turbinas de vapor.

[12]Es un motor térmico que funciona al revés. Se produce calor y es aplicado trabajo, para que del cuerpo frío halla transferencia de calor al cuerpo caliente.

lo siguiente es valido:

$$\oint dS = 0. \tag{1.6}$$

donde, la integración se realiza sobre un ciclo del proceso.

En un proceso irreversible,

$$dS > \frac{dQ}{T}. \tag{1.7}$$

Para concluir, la entropía S y la energía interna U dependen solamente del estado del sistema, es decir, de sus condiciones iniciales y finales, y no de la forma de transición de un estado a otro. En consecuencia, la entropía es una propiedad de estado. De (1.6) la entropía de un proceso reversible no afecta la entropía total del universo. Sin embargo, los procesos naturales son todos irreversibles, y por tanto de (1.7) la entropía no puede disminuir y todos los procesos aportan al aumento de la entropía total. La segunda ley de la termodinámica se puede expresar de la siguiente forma: *la entropía total de un sistema aumenta o permanece constante en cualquier proceso; nunca disminuye.* En contraste con la energía que es un cantidad conservativa la entropía siempre crece. Además, mientras la energía es la habilidad para realizar un trabajo, la entropía es la medida de cuánta energía no es disponible para realizar un trabajo [28].

1.2.2. Mecánica Estadística

Ludwig Boltzmann[13] en un intento por comprender y ampliar el enfoque macroscópico y mecánico de la segunda ley de la termodinámica, definida por variables como trabajo, energía, calor, entre otros, centró su análisis retornando a las concepciones atomísticas[14] desarrolladas por los filósofos griegos, tales como Demócrito, Epicuro y Leucipo. Estos filósofos distinguían el mundo en dos órdenes : el mundo de la apariencia y el mundo que subyace a todos los fenómenos, constituido por el movimiento de los átomos.

[13]Ludwig Boltzmann (1844-1906) fue un físico austriaco, fundador de la mecánica estadística.

[14]La atomística, consideraba la realidad de los átomos no solo como artificios teóricos en la explicación de algunas leyes físicas o químicas, sino que afirma que existen como entidades físicas reales aunque inobservables [31].

Boltzmann está entre los primeros físicos que afrontaron la comprensión del micromundo subyacente al mundo macroscópico, y fue él, quien examinó con detenimiento, que no era suficiente la comprensión de los fenómenos usando medidas extensivas -que determinaban el entendimiento del mundo a partir de cálculos y sistemas de medición deterministas- obtenidas con mucho esfuerzo y trabajo por los físicos, dado que se necesitaba comprender la inmensa variabilidad y la multiplicidad de partículas en movimiento que se atraían y se rechazaban entre sí a nivel atómico: ¿Cómo comprender aquello que se presenta como un caos? La clave fue el concepto de entropía establecido por Clausius considerandolo como un fenómeno microscópico, como una medida del orden y el desorden, la cual solo podría ser obtenida desde la concepción de un cálculo aproximado y no determinista, como lo es el cálculo probabilístico.

En [11] se resolvió la cuestión de la transferencia de calor de un cuerpo a otro, y de la conversión del calor en trabajo en un ciclo termodinámico, considerando la relación que existe entre calor y energía, como se explicó anteriormente. Boltzmann se dedicó al análisis microscópico, para lo cual caracterizó los gases, primero, como constituidos por átomos, cuyos movimientos y trayectorias eran indescriptibles desde la mecánica clásica, y segundo, desarrolló el análisis matemático en términos de la distribución de probabilidad de velocidad de las moléculas de un gas, la cual actualmente se conoce como la distribución de velocidades de Maxwell-Boltzmann[15], que describe el equilibrio térmico de un gas [33]. Luego planteó la hipótesis ergódica, hasta llegar al planteamiento del teorema de Boltzmann y el teorema H, que proporciona la equivalencia en la mecánica estadística de la segunda ley de la termodinámica.

De este estudio, se observa que las variables macroscópicas se caracterizan por la multitud de movimientos que presentan las partículas microscópicas y aunque no se presenta el análisis matemático de estos teoremas, para dar más concreción al tema de este capítulo, se delinearán las ideas que introducen el método de análisis probabilístico en el concepto de entropía [33, 34].

La mecánica estadística introduce conceptos como macroestado y microestado, para

[15]El concepto de distribución de velocidades hace referencia a que las moleculas en un gas se mueven con cierta velocidad (energía cinética), la cual es proporcional a la temperatura del gas. A una temperatura ambiente (de $300°K$), por ejemplo, la velocidad de las partículas de Nitrógeno del aire, tienen una velocidad promedio de 422 m/s [32].

Caras	Sellos
5	0
4	1
3	2
2	3
1	4
0	5

Tabla 1.1: Número de posibles combinaciones.

comprender a que se refieren, se analiza la siguiente pregunta: ¿cuáles son los posibles resultados de lanzar cinco monedas? Si el lanzamiento de cada moneda puede resultar en cara C o sello S, los posibles resultados generales con respecto al número de caras y sellos se registran en la Tabla 1.1. Los resultados de la Tabla 1.1, se denominan macroestados. Un macroestado es una propiedad general, la cual no especifica los detalles del sistema, en el ejemplo, no dice en que orden ocurrieron las caras y los sellos. En este sistema de cinco monedas hay seis macroestados. Algunos macroestados son más probables que otros. En la Tabla 1.2 se presentan las formas en que cada macroestado puede ocurrir. Cada secuencia es llamada un microestado, el cual, es una detallada descripción de cada macroestado del sistema [28, 34].

El macroestado de $3C$ y $2S$ puede ocurrir de 10 distintas formas, igual que el macroestado $2C$ y $3S$, los cuales son más probables que el macroestado de $5C$ o $5S$, ya que ocurren de manera conjunta en 20 de 32 posibles casos. A diferencia de $5C$ o $5S$, los cuales siendo más ordenados, es decir todos caras o todos sellos, son los menos probables, ya que solo ocurren de manera conjunta en 2 de 32 posibles casos. Esto es posible deducirlo si cada microestado tiene igual probabilidad de ocurrir. En el ejemplo, cada moneda debe estar equilibrada, lo que significa que la probabilidad de que suceda una cara es igual a la probabilidad de que suceda un sello. En el caso de un sistema aleatorio, los microestados se asumen equiprobables para que el razonamiento sea correcto. Si el número de lanzamientos de la moneda se incrementa, por ejemplo, en 100 veces, el número de microestados del macroestado $100C$ es 1, pero el número de microestados de $50C$ y $50S$ es 1×10^{29}, lo cual indica que es mucho más probable encontrar una configuración más desordenada.

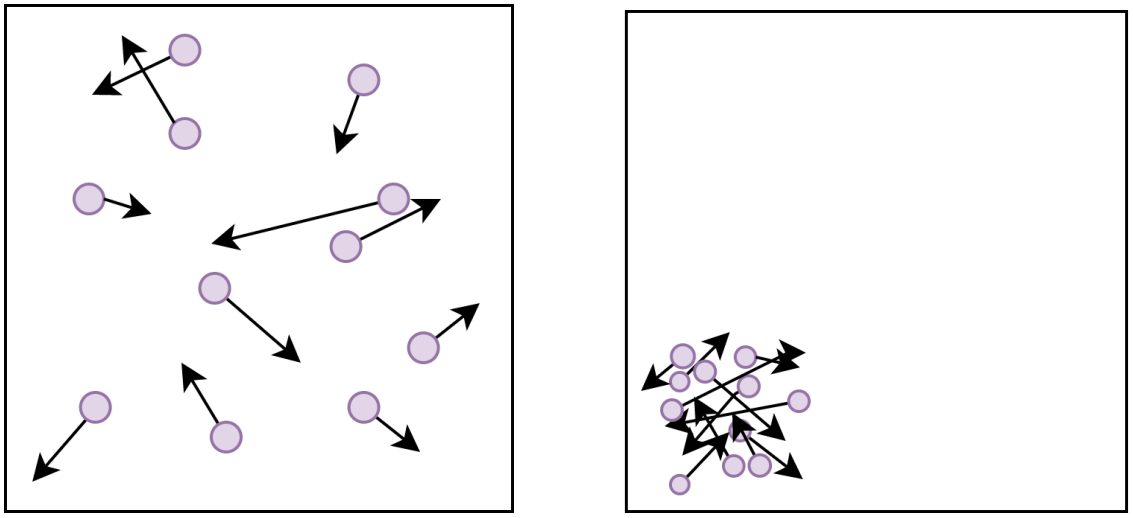
Para Boltzmann, la entropía se convertía en una medida probabilística y propusó

Macroestado	Microestados Individuales	Número de Microestados
5C, 0S	CCCCC	1
4C, 1S	CCCCS, CCCSC, CCSCC, CSCCC, SCCCC	5
3C, 2S	CCCSS, CCSSC, CSSCC, SSCCC, SCCCS, SCCSC, SCSCC, CSCCS, CSCSC, CSSCC	10
2C, 3S	SSSCC, SSCCS, SCCSS, CCSSS, CSSSC, CSSCS, CSCSS, SCSSC, SCSCS, SCCSS	10
1C, 4S	CSSSS, SCSSS, SSCSS, SSSCS, SSSSC	5
0C, 5S	SSSSS	1
		Total 32

Tabla 1.2: Lanzamiento de cinco monedas.

el estudio de la evolución de la función de distribución de probabilidad que define el movimiento de moleculas en un determinado gas, la cual corresponde a una variable aleatoria o macroestado. El macroestado de un gas está determinado por sus propiedades macroscópicas, tales como: volumen, presión y temperatura. Y sus microestados corresponden con la descripción detallada de las posiciones y velocidades de sus átomos. El ejemplo anterior es congruente con el fenoméno físico, considerando el número de lanzamientos de la moneda exponencialmente grande, por ejemplo, si es equivalente al número de átomos en una pequeña muestra de aire, se tienen 2.7×10^{19} átomos en 1 cm^3 de un gas ideal a una presión de 1 atm[16] y a una temperatura de $0°C$, pero en este caso, no interesa si cae cara o sello, sino la velocidad promedio de los átomos [28].

[16]La unidad de la presión es la atmósfera, la cual equivale a la presión que ejerce la atmósfera terrestre al nivel del mar.



(a) Configuración más probable. (b) Configuración menos probable.

Figura 1.8: Configuración de un grupo de moléculas en un contenedor.

El equilibrio se concibe en un sistema físico como el macroestado más probable de un gas, y se da por sentado que, la evolución de un estado a otro tiene un sentido: del menos probable al más probable. Para un gas, las condiciones más probables son descritas por la distribución de velocidades Maxwell-Boltzmann en direcciones aleatorias, como se observa en la Fig. 1.8a, en contraste con un tipo de macroestado más definido como se observa en la Fig. 1.8b donde todos las partículas se encuentran en una esquina, lo cual aunque es posible es improbable[17].

De esta manera, definiendo un sistema con un número finito de estados, se busca encontrar entre ellos, aquel que represente el equilibrio del sistema, i.e., el macroestado que contiene el número mayor de microestados, cuya distribución en el espacio aislado, corresponde a la máxima probabilidad.

Sea N el número de partículas, V el volumen y U la energía interna, valores constantes. Sea Ω la probabilidad de que un macroestado ocurra. El macroestado de máxima probabilidad $\Omega_{N,V,U}$ corresponde al estado de equilibrio del sistema [12]. Boltzmann propusó que la entropía de un sistema S en un estado (macroestado) es

[17]La evidencia de que los átomos estaban en constante movimiento fue radical con la teoría de Einstein del movimiento Browniano, la cual sería además sin contradicción la evidencia de lo que dedujo Maxwell y Boltzmann [34].

proporcional al logaritmo natural de la probabilidad del número de microestados de un macroestado, i.e.,

$$S = k_B \ln \Omega, \tag{1.8}$$

donde k_B es la constante de Boltzmann. Por lo tanto, la entropía se maximiza con el macroestado de máxima probabilidad, lo cual indica que la condición de desorden es sinónimo de alta entropía. Las fuerzas externas pueden forzar a un sistema a disminuir su entropía, pero en la medida que avanza el tiempo los átomos se dispersan y no retornan a su estado inicial. Por ejemplo, cuando se limpia el polvo de una habitación en la mañana, se encuentra en contraste el estado del sistema que es la acumulación de polvo, con el orden o limpieza que se quiere establecer, por lo tanto, al aplicar a través de la escoba una fuerza que cambia el sistema a un estado "más organizado" intenta forzar la entropía para que esta disminuya, sin embargo, en el transcurso del día, espontáneamente a la habitación retorna el polvo y el desorden, y el sistema retorna a su estado más probable [35]. Boltzmann probó que esta expresión para S es equivalente a la definición obtenida por Clausius en (1.3) y por tanto, equivalente a la segunda ley de la termodinámica [34].

1.3. FUENTES INTEGRANTES DE LA TEORÍA DE LA INFORMACIÓN

La teoría de la información es la síntesis de una serie de desarrollos científicos-técnicos, que Shannon incorporó congruentemente, haciendo converger las teorías modernas, tales como: la física termodinámica, el calculo infinitesimal, la mecánica estadística y la teoría de la probabilidad, y los desarrollos técnicos del momento, tales como: la telefonía, la telegrafía, el sistema de Modulación por Pulsos Codificados (PCM, *Pulse Code Modulation*), los sistemas Codificadores de Voz (VOCODER, *Voice Coder*), la televisión, la Modulación de Frecuencia (FM, *Frequency Modulation*), entre otros desarollos del siglo XX. En ese mismo periodo, la mecánica cúantica se desarrollaba para sentar nuevas bases de la comprensión de la realidad, como el

principio de incertidumbre desarrollado W. Heisenberg[18].

Se hará una síntesis de la confluencia del cálculo y la probabilidad, entendiendo el primero, como la matemática que inaugura la concepción probabilística de la realidad y la probabilidad como el intento humano por pensar la realidad, medirla y obrar en consonancia y en concordancia entre el pensamiento y los fenómenos físicos. Shannon es consciente de la necesidad de una teoría fundamentada científicamente que resuelva técnicamente la cuestión de la transmisión de información. Para ello, recoge los trabajos de L. Boltzmann, A. Markov[19] y A. Kolmogorov[20] en los que la física entra en relación vinculante con la probabilidad. Para tener una idea general del vínculo entre el cálculo y la probabilidad, se hará una descripción del fundamento histórico de la teoría de límites, además de señalar que la matemática que fundamenta el desarrollo de la teoría física es el cálculo, y actualmente, el cálculo estadístico.

1.3.1. Cálculo Probabilístico

En la época clásica griega se plantearon la mayoría de problemas que ha asumido la ciencia de los últimos siglos. Así, los problemas típicos del cálculo infinitesimal, tales como: el cálculo de perímetros, áreas y volúmenes de las figuras elementales (círculos, esferas, conos, pirámides, etc.) habían sido establecidos y resueltos de alguna manera, por Arquímedes[21] con el Método de Recubrimientos y Eudoxo[22] por el Método de Exhaustación [36]. Los métodos generales y analíticos de resolución de estos problemas, sin embargo, se desarrollaron a partir del siglo XVII por Leibniz[23]

[18]Werner Karl Heisenberg (1901 - 1976) fue un físico teórico alemán y uno de los fundadores de la mecánica cuántica y premio nobel de Física en 1932.

[19]Andréi Andréievich Markov (1856-1922) fue un matemático ruso, quien desarrolló la teoría de procesos estocásticos.

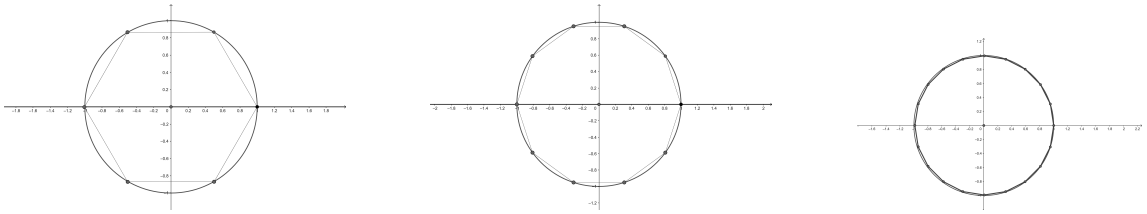
[20]Andrey Nikolayevich Kolmogorov (1903 - 1987) fue un matemático ruso, quien realizo grandes aportes a la teoría de la probabilidad, la teoría de la información y la teoría de números.

[21]Arquímedes (287 a.C. - 212 a.C.) fue un matemático griego.

[22]Eudoxo de Cnido (apróx. 390-337 a.C.) filósofo y matemático griego miembro de la Academia de Platón, su idea principal se expresa como: toda magnitud finita puede ser agotada mediante la substracción de una cantidad determinada.

[23]Gottfried Wilhelm Leibniz (1646–1716) fue un filósofo, matemático, lógico, teólogo, jurista, bibliotecario y político alemán , famoso por desarrollar el cálculo integral y diferencial.

Figura 1.9: Aproximación al círculo por el Método de Eudoxio



(a) $n = 6$ (b) $n = 10$ (c) $n = 20$

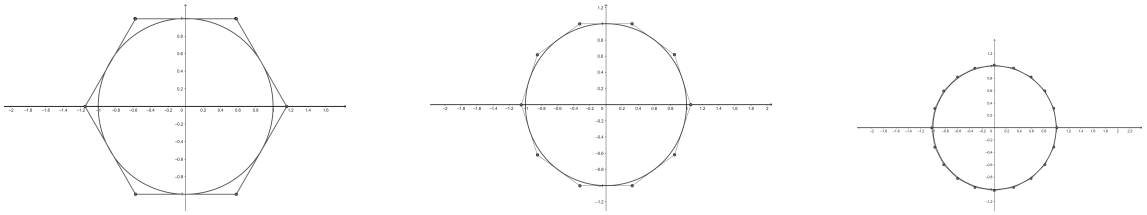
Figura 1.10: Polígonos inscritos en un círculo de radio unitario.

y Newton[24] creando el Cálculo Infinitesimal.

El método de exhaustación, aplicado para calcular el perímetro de un círculo consiste en inscribir y circunscribir polígonos regulares, de tal manera, que el perímetro de cierto polígono sea aproximado al perímetro de la circunferencia.

En la Fig. 1.9 se observa este método. El triángulo inscrito y circunscrito se encuentran muy lejos del perímetro del circulo a diferencia del octágono el cual se aproxima a la circunferencia. El paso del triángulo al octágono, se obtiene por el incremento del número de lados del polígono. Al aplicar este método a un círculo de radio unitario,

[24]Isaac Newton (1642–1727) fue un físico, teólogo, inventor, alquimista y matemático inglés, quien independientemente de Leibniz desarrolló el cálculo integral y diferencial. Es reconocido además por el desarrolló de las leyes de la mecánica clásica.



(a) $n = 6$ (b) $n = 10$ (c) $n = 20$

Figura 1.11: Polígonos circunscritos en un círculo de radio unitario.

Perímetro	n = 6	n = 10	n = 20
Polígono Inscrito	6.015	6.181	6.214
Polígono Circunscrito	6.917	6.499	6.364

Tabla 1.3: Aproximación a la circunferencia con un polígono de n lados.

dado que se conoce que la medida de la circunferencia es $2\pi r = 6.283$, se grafican las aproximaciones de diferentes polígonos regulares de n lados. En en la Fig. 1.10 se observan los polígonos inscritos y en la Fig. 1.11 los polígonos circunscritos . Los valores del perímetro para cada n se presentan en la Tabla 1.3 , en la cual se observa que se aproxima al valor del perímetro de la circunferencia con un margen de error. Este método es conocido como la cuadratura del círculo y es la base, como lo afirmó Leibniz, del cálculo de límites [37]. La idea de este método consiste en realizar *sucesivas* aproximaciones para encontrar el valor más próximo de la circunferencia. Con el método no se obtiene una coincidencia entre el perímetro del polígono y la circunferencia, sólo una aproximación. El valor exacto se resuelve con el cálculo infinitesimal, realizando el "paso al límite", lo que implica el incremento del número de lados indefinidamente. Este hecho abrió paso a la la teoría de límites [36, 37].

¿Qué relación se puede observar entre la concepción de Eudoxo y la teoría probabilística?, la respuesta es que el método de Eudoxo expresa el fundamento de la ciencia probabilística, que nace del entendimiento que todo concepto matemático acerca de la realidad no hace otra cosa que medir que tanto se aproxima el conocimiento humano a la explicación de ella, cuestión que se obtiene encontrando las tendencias o convergencias en el límite.

De ahi que del análisis de la Seccion 1.2, al abrirse en profundidad y en amplitud el mundo subyacente de los átomos y sus movimientos, la comprensión de estos sólo puede hacerse probabilísticamente, es decir, emergen conceptos probabilísticos, tales como, la entropía y la incertidumbre.

1.3.1.1. Ley de los Grandes Números

Los problemas de la teoría clásica de la probabilidad fueron planteados en el siglo XVII por matemáticos, físicos y astrónomos, entre ellos, B. Pascal[25], P. Fermat[26], C. Huygens[27]. La teoría de la probabilidad se consolidó con el tratado de J. Bernoulli[28], titulado *Ars Conjectandi*[29], y publicado en 1713, en el cual se postularon la Ley de los Grandes Números (LLN, *Law of Large Numbers*) y el teorema del límite central.

Los desarrollos teóricos que se lograron estaban basados en experimentos aleatorios sencillos. La generalización fue realizada posteriormente gracias a los aportes de P. Laplace [30], S. Poisson[31] y P. Chebyshev[32], siendo este último quien generalizó la LLN en 1867 a secuencias de variables aleatorias independientes con varianza finita [38].

Hay dos formas de la LLN: la Ley del los Grandes Números Débil (WLLN, *Weak Law of Large Numbers*) y la Ley del los Grandes Números Fuerte (SLLN, *Strong Law of Large Numbers*) , las cuales se diferencian por el tipo de convergencia que aplican, como se describe a continuación.

Sea una secuencia de n variables aleatorias $X_1, X_2, \ldots, X_n$ independientes e idénti-

[25]Blaise Pascal (1623-1662) fue un filósofo, matemático, físico y religioso francés, fundador de la moderna teoría de la probabilidad.

[26]Pierre de Fermat (1601-1655) fue un matemático francés, quien es conocido como el fundador de la teoría de los números. Compartía correspondencia con B. Pascal discutiendo los fundamentos de la teoría de la probabilidad.

[27]Christiaan Huygens (1629-1695) fue un matemático, astrónomo y físico alemán, fundador de la teoría de ondas de la luz. El también compartía correspondencia con B. Pascal sobre problemas matemáticos.

[28]Jaques Bernoulli (1654-1705) fue un matemático suizo, quien introdujo los primeros principios del cálculo de variaciones

[29]"*The Art of Conjecturing*", contiene la teoría de permutaciones y combinaciones y también los conocidos números de Bernoulli.

[30]Pierre Simón, marquis de Laplace (1749-1827) fue un matemático, astrónomo y físico quien es conocido por sus investigaciones en la estabilidad del sistema solar. Demostró la utilidad de la probabilidad para interpretar datos científicos.

[31]Siméon-Denis Poisson (1781-1840) fue un matemático francés, quien desarrolló la distribución que lleva su nombre, y realizo contribuciones a la ley de los grandes números.

[32]Pafnuty Chebyshev (1821-1894) fue un matemático ruso, quien es conocido por su trabajo en la teoría de los números primos. Desarrolló una desigualdad básica de la teoría de la probabilidad, la cual permite demostrar la ley de los grandes números.

camente distribuidas (i.i.d.), con media muestral M_n definida como:

$$M_n = \frac{X_1 + X_2 + ... + X_n}{n}, \tag{1.9}$$

donde, M_n es otra variable aleatoria, y sea $\epsilon \in \mathbb{R}$, con $\epsilon > 0$ y arbitrariamente pequeño, la WLLN establece que:

$$\Pr\{|M_n - \mu| \geq \epsilon\} = \Pr\left\{\left|\frac{X_1 + X_2 + ... + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0, \quad n \to \infty. \tag{1.10}$$

donde, μ es el valor esperado de una variable aleatoria genérica X i.e., $\mu = \mathbb{E}_X[X]$.

La WLLN afirma que la media muestral M_n converge en probabilidad a la media teórica μ a medida que el número de variables aleatorias crece, i.e., que la mayor parte de la distribución de la variable aleatoria M_n se concentra en un intervalo $[\mu-\epsilon, \mu+\epsilon]$ alrededor de μ con una alta probabilidad a medida que n crece. Dado que $n \to \infty$, esta probabilidad converge a 1. La Ley débil, dice también que la divergencia de M_n a μ es posible pero con mínima probabilidad, ya que tiende a cero mientras n tienda a infinito. Esta ley, sin embargo, no da información concluyente con respecto a las divergencias posibles, lo cual si es claro en la ley fuerte [39].

La SLLN relaciona igual que la WLLN la media muestral M_n de variables aleatorias i.i.d. y la media teórica μ en otro tipo de convergencia. La SLLN establece que:

$$\Pr\left\{\lim_{n\to\infty} \frac{X_1 + X_2 + \ldots + X_n}{n} = \mu\right\} = 1. \tag{1.11}$$

De acuerdo a la ley fuerte M_n converge a μ con probabilidad 1 [39].

1.3.1.2. Cadenas de Markov

La WLLN fue demostrada para variables aleatorias independientes. A. Markov demostró que la convergencia se cumple también para variables aleatorias dependientes, para lo cual estudió los experimentos que pueden generarlas, e ideó una estructura matemática que describía tales variables conocida como cadenas de Markov.

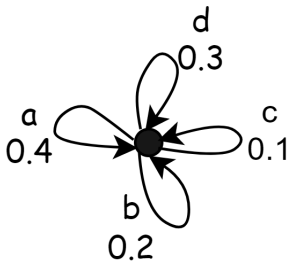


Figura 1.12: Gráfico de la Cadena de Markov de un estado.

Las cadenas de Markov son cadenas homogéneas simples, que están descritas por un conjunto de estados y una matriz de transición, en el caso de variables aleatorias i.i.d., por ejemplo, sea $\mathcal{X} = \{a, b, c, d\}$ las transiciones de los elementos de $\mathcal{X}$ son independientes y ocurren de acuerdo a su probabilidad conduciendo a un único estado, como se observa en la Fig. 1.12. Markov aplicó las cadenas al estudio de la estructura estadística del poema *"Eugene Onegin"*[33] en el que mediante un trabajo manual descubrió que la probabilidad estacionaria de las vocales es $p = 0.432$, la probabilidad de que de una vocal siga otra vocal es $p_1 = 0.128$, y la probabilidad de que una vocal siga a una consonante es $p_2 = 0.663$. El proceso estocástico que resulta de las cadenas del poema modelan su lectura o escritura, no como una simple secuencia de letras sin sentido, sino como la repetición artificial del proceso de escritura gracias a su estructura estadística. Con esta idea, Markov introdujó adicionalmente un análisis que considera la dependencia de procesos aleatorios a la evolución temporal, el cual es un aspecto fundamental para definir una cadena de Markov [40, 41].

Los procesos de Markov fueron aplicados por Shannon para describir matemáticamente la fuente de información, con el fin de aprovechar el conocimiento que puede ser obtenido de la estructura estadística del mensaje. Así, la fuente discreta genera un mensaje, representado por una secuencia de símbolos, en la cual cada símbolo es seleccionado en función de las probabilidades de transición que dependen, en general, de los símbolos anteriores. De esta manera, Shannon sugirió que cualquier fuente de información es un proceso de Markov, lo cual, permite la aproximación al lenguaje natural por medio de una serie de lenguajes artificiales simples [1, 42].

El estudio de la estructura estadística del lenguaje ha sido de gran utilidad en los

[33]Novela clásica rusa escrita por Aleksander Pushkin en 1833.

sistemas de comunicación para definir códigos más eficientes en la transmisión de información, por ejemplo, la distribución estadística en su forma más elemental, es la base del código de Morse, el cual utilizaba puntos y líneas para codificar las letras del alfabeto inglés, y con un simple punto representa la letra de mayor ocurrencia en inglés, i.e., *e*, mejorando notablemente la velocidad de tranmisión de información.

1.3.2. Teoría de la Transmisión

Desde el siglo XIX se buscaba desarrollar y optimizar los sistemas de transmisión de señales de información, entre ellos los sistemas de telegrafía y telefonía. La investigación se sustentó en las contribuciones de inventores, físicos y matemáticos[34] que aportaron al desarrollo de los sistemas de telecomunicaciones. Sin embargo, es en el siglo XX donde se observan avances relevantes, los cuales ocurrieron inicialmente en los Laboratorios Bell de la compañía AT&T, en donde grandes científicos e ingenieros ampliaron las posibilidades no solo en la transmisión de información sino en el procesamiento y almacenamiento de datos. En los Laboratorios Bell se desarrollaron inventos como el transistor, el láser, el teléfono de tonos, el procesador digital de señales, entre otros, y además, la investigación tecnológica por parte de algunos ingenieros eléctricos y matemáticos fue orientada a la formulación matemática de los límites fundamentales de los sistemas de comunicación para la transmisión de señales de telegrafía y futuras señales [22].

En 1924 se iniciaron las investigaciones de estos límites. Harry Nyquist[35] publicó su artículo titulado *Certain Factors Affecting Telegraph Speed* [2], en el que considera las variables que determinan la velocidad de transmisión, buscando "la selección de la mejor forma de onda que permitiera obtener la máxima velocidad de transmisión de datos, sin demasiada interferencia o afectación por ruido, y la selección apropiada de códigos que permitieran transmitir una máxima cantidad de inteligencia (*intelligence*)". Nyquist utilizó la palabra *intelligence*, que se podría traducir como información. Nyquist publicó en 1928 el artículo: *Certain Topics in Telegraph*

[34]Heinrich Hertz 1887, Emile Berliner 1888, Mihajlo Pupin 1899, Ferdinand Graf von Zeppelin 1900, John A. Fleming 1904, entre otros.

[35]Harry Nyquist (1889 – 1976) ingeniero electrónico sueco, quien realizó grandes aportes a la teoría de la información.

Transmission Theory [4], en el cual profundizó los temas propuestos anteriormente y presentó el teorema de muestreo de señales analógicas. Nyquist es reconocido por sus diferentes contribuciones inventivas y teóricas en el campo de la ingeniería de las comunicaciones.

En 1928, Ralph Hartley[36] publicó el artículo: *Transmission of Information* [3], en el cual planteó la forma de encontrar un parámetro para comparar la capacidad de varios sistemas de comunicación. La capacidad según Hartley es una medida de la información, la cual corresponde a la habilidad que tiene el sistema de recepción para distinguir la secuencia de símbolos transmitida a través de un medio que puede distorsionar la señal en un menor o mayor grado. En términos generales, la información la definió como un concepto aislado de factores psicológicos, e intentó encontrar la definición técnica que permitiera la medida cuantitativa de la información.

Hartley desarrolló su análisis a partir del proceso de comunicación discreto. La comunicación la describió como el proceso establecido por un emisor que selecciona un símbolo particular y forma una secuencia de símbolos que contiene un mensaje, el cual es dirigido a un receptor por un mecanismo de transmisión. El mensaje se representa por símbolos, los cuales codifican palabras, o representan los símbolos punto y línea del sistema de telegrafía. El interés en [3] es la forma en que se introdujo la medida de información, la cual se describe brevemente.

Sea s el número de símbolos que utiliza el sistema y n el número de selecciones que se realizan, por lo tanto se tienen en total s^n posibles secuencias. Hartley planteó dos argumentos. El primer argumento considera s^n como medida de información, pero dado un s fijo y a medida que avanza la comunicación, n se incrementa, por lo tanto, la cantidad de información crece exponencialmente. Sin embargo, la información que esta relacionada a la capacidad de identificar un mensaje en el receptor, no crece exponencialmente con el número de símbolos n por secuencia. Dado que esta hipótesis no brindó el resultado esperado, Hartley propusó el segundo argumento: *la medida de información es proporcional al número de selecciones*:

$$H = Kn,$$

[36]Ralph Vinton Lyon Hartley (1888 – 1970) investigador de la electrónica estadounidense, quien contribuyo a la fundamentación de la teoría general de la información.

donde, H es cantidad de información, n el número de selecciones y K la constante que depende del número de símbolos s de cada selección. Hartley consideró dos sistemas con números de símbolos s_1 y s_2, constantes K_1 y K_2, y número de selecciones de símbolos n_1 y n_2, respectivamente, y supuso que, si el número de secuencias posibles es igual, independiente de n_1 y n_2, entonces la cantidad de información es la misma para ambos sistemas. De lo cual encontró que $K = k_0 \log s$, donde k_o tiene un valor arbitrario y es el mismo para los dos sistemas, lo que se puede omitir haciendo la base logarítmica arbitraria. Hartley concluyó que,

$$H = n \log s \tag{1.12}$$

$$= \log s^n. \tag{1.13}$$

Así, la cantidad de información es el logaritmo del número posible de secuencias, y el valor numérico depende de la base del logaritmo que se utilice. Nyquist y Hartley son los científicos anteriores al desarrollo de la teoría de Shannon, a los que este reconoce por sus aportes a la teoría de la información, del mismo modo, la convergencia de diferentes conceptos de la termodinámica, de la teoría de la probabilidad y de la mecánica estadística. En el capítulo siguiente se dedicará el estudio a la síntesis teórico técnica, que se promulga bajo el nombre de teoría de la información.

Capítulo 2

TEORIA DE LA INFORMACION

Al pensar la teoría de la información propuesta por Shannon se hace necesario referenciar a N. Wiener[1], como un reconocimiento a su labor teórica y práctica, por su participación en el desarrollo de la teoría de las comunicaciones: bien sea porque fue contemporáneo a Shannon; porque fue una fuente directa y de primera mano; o simplemente, porque al mismo tiempo que Shannon desarrollaba su propuesta teórica y técnica, Wiener[2] lo hacía también por su propia cuenta [13, 14].

La teoría de la información fue crucial para el mundo, por dos razones importantes: la primera, esta relacionada al hito histórico de la segunda guerra mundial, la cual aceleró diversos procesos investigativos, entre ellos la urgente necesidad de mejorar los sistemas de comunicaciones, la encriptación de la información, la velocidad de transmisión y la seguridad; en segundo lugar, en Estados Unidos las compañías de investigación y desarrollo científico impulsaron el estudio de los sistemas de comunicación y su implementación a gran escala. Shannon trabajó para los Laboratorios

[1]Norbert Wiener (1894 – 1964) fue un matemático y filósofo estadounidense, quien desarrollo la ciencia de la cibernética.

[2]Wiener y Shannon consideraban que el modelo matemático del sistema de comunicación era descrito con precisión por la teoría de la probabilidad. De tal manera, que el mensaje o señal a transmitir y el ruido que alteraba el mensaje eran de naturaleza estadística. Sin embargo, Wiener afirmaba que la señal solo podía ser procesada después de que el ruido la había alterado, a diferencia de Shannon que considera que la señal podía ser procesada antes y después de su transmisión a través de un canal ruidoso [43].

Bell[3] desde 1941, cuando aún EEUU no había entrado en guerra, pero en Europa estaba en pleno apogeo. Sus investigaciones abarcaron el espectro de la guerra [44] y la postguerra, que era un momento todavía más significativo porque implicaba la carrera por la supremacía de las naciones bajo la tutela del desarrollo tecnológico, cuando se desató otra forma de guerra competitiva entre los EEUU y la entonces Unión Soviética, llamada guerra fría [45].

2.1. MODELO GENERAL

Los problemas de la comunicación como se planteó en el primer capítulo, son en esencia los mismos, pero las formas técnicas de resolverlos cambian fundamentalmente y se perfeccionan, no ya espontáneamente, sino como consecuencia directa de la formulación teórica y del desarrollo técnico. De esta manera Shannon definió en términos teóricos el problema de la comunicación y desarrolló la teoría de la información siguiendo el principio básico de todo aprendizaje, que es acercarse al problema y resolverlo de la manera más sencilla, similar a la forma de pensar los problemas matemáticos y geométricos, calculando lo desconocido usando lo conocido o regular.

Así Shannon ideó un modelo general en el que se establecen abstracciones matemáticas de las entidades físicas que participan en el proceso de la comunicación. El modelo general de los sistemas de comunicación se observa en la Fig. 2.1, en el cual las entidades que lo conforman son: la fuente de información, el transmisor, el canal de comunicación, el receptor y el destino.

1. La fuente de información produce los mensajes que serán enviados al destino. La naturaleza de los mensajes es de dos tipos: discreta o continua. Los mensajes discretos son secuencias de símbolos seleccionados de un conjunto finito y discreto de símbolos. Los mensajes continuos varían en función del tiempo o del espacio formando un conjunto infinito y continuo de valores, por ejemplo, las señales de audio, para los sistemas de radio y telefonía, o las señales de televisión las cuales dependen de coordenadas espaciales y del tiempo.

[3]Fundada en 1925 por la empresa AT&T.

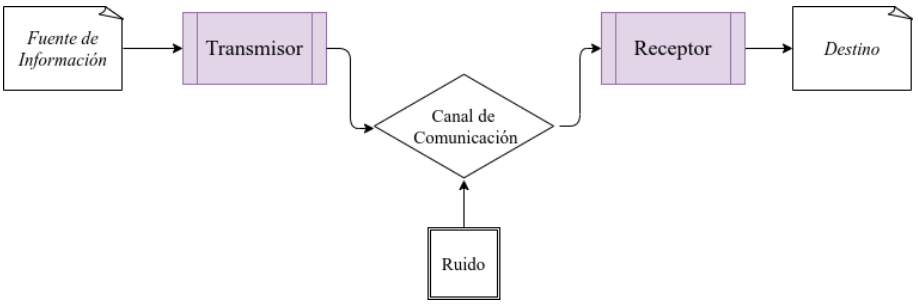


Figura 2.1: Diagrama General del Sistema de Comunicación.

2. El transmisor opera sobre el mensaje para producir una señal adecuada para su transmisión a través del canal. Entre las operaciones que realiza el transmisor se encuentran la conversión de un tipo de señal en otra, y los procesos de modulación y codificación, entre otras.

3. El canal es el medio utilizado para la transmisión desde un punto (transmisor) hasta otro punto (receptor).

4. El receptor realiza las operaciones inversas del transmisor, por ejemplo, la demodulación y la decodificación, con el fin de reconstruir el mensaje original a partir de la señal recibida.

5. El destino es a quien se dirige el mensaje.

El objetivo del modelo general del sistema de comunicación es replicar el mensaje en el destino, tomando ventaja de la codificación y a pesar de la afectación que impone el ruido sobre la secuencia de símbolos o señal a transmitir [1].

Los sistemas de comunicación se encuentran en un ambiente natural y social, el cual introduce errores o interferencias en la transmisión de información. Por ejemplo, factores meteorológicos como el viento, el agua, la temperatura, afectan de manera imprevista el sistema. Factores humanos como el espionaje son otras formas de interferencia que perturban la comunicación. En general, estos factores se relacionan con el ruido, el cual se define como: el conjunto de todos los elementos extraños que, no siendo parte del modelo general de la comunicación, limitan el cumplimiento de su objetivo.

En el modelo de la Fig. 2.1 se deben diferenciar, por la naturaleza de la fuente y de la señal a transmitir, los sistemas de comunicación discretos y continuos. Además, el sistema de comunicación se puede analizar considerando o no el ruido. En este capítulo se introducen los conceptos de medida de información, secuencias típicas y capacidad del canal, definidos para el sistema de comunicación discreto sin ruido.

2.2. PROBABILIDAD, INFORMACIÓN Y ENTROPÍA

2.2.1. Probabilidad

La realidad que abarca la totalidad de fenómenos naturales y artificiales se presenta al sentido común humano como un conjunto de situaciones aleatorias disímiles, i.e., como una realidad caótica e inexplicable, que sólo puede ser desentrañada por la ciencia y la técnica, intentando romper la apariencia de caos, estableciendo razones y causas que expliquen los fenómenos aleatorios.

En la comunicación se distinguen dos fenómenos aleatorios. En primer lugar, la comunicación es necesaria porque entre un emisor que envía un mensaje a un receptor, el mensaje es desconocido por el receptor, por ejemplo, transmitir el resultado de lanzar una moneda, el receptor conoce que hay dos resultados posibles: cara o sello, pero no sabe cuál se transmitió. En un diálogo entre dos personas, las cuales se transmiten sus ideas, formuladas de manera independiente, el escucha no conoce con certeza lo que va a expresar la persona que es cuestionada, por lo tanto, se está a la espera de su respuesta y de sus planteamientos, es decir, hay incertidumbre en los mensajes. En segundo lugar, ocurre otro fenómeno intrínseco a la comunicación, el cual es debido a los factores que interfieren en el canal de comunicación, en el último ejemplo, si las dos personas se encuentran en un parque, los sonidos de la calle tales como: personas conversando en la cercanía, perros ladrando, una máquina funcionando hacen que quien está escuchando no entienda el mensaje debido a estas interferencias producidas al azar [46].

La ciencia física establece leyes que fijan la repetición y los ciclos de los fenómenos físicos, para aprehender la variabilidad y multiplicidad existente, fijando lo invariable en lo variable, con conceptos que nacen del intento de medir, lo cual para la física e ingeniería son formas de conocer. En la física moderna estas medidas se llaman leyes naturales de los fenómenos físicos, que parecen determinantes, pero que poco a poco se han ido convirtiendo en aproximaciones o leyes de tendencia de la realidad o leyes límite [37, 46, 47].

La ciencia y la tecnología tienen como tareas construir modelos matematicos y sistemas de medición que definan la tendencia subyacente a la variabilidad aleatoria. Shannon hizo lo uno y lo otro en la teoría de la información: utilizó la probabilidad para encontrar *la convergencia de situaciones aleatorias*, y definió la entropía como la medida de esa probabilidad o de esa tendencia. El análisis probabilístico no intenta definir una verdad absoluta sobre los fenómenos[4], lo que intenta es acercarse a la incertidumbre que subyace a toda realidad, y cuyo conocimiento son las aproximaciones al límite de la misma [1, 46, 48, 49].

En el Apendice A, se encuentran los conceptos básicos de la teoría de la probabilidad que son la base del desarrollo propuesto por Shannon, y que permiten abordar los problemas de la comunicación.

2.2.2. Información y Entropía

La concepción de información lejos de ser teorética es técnica. Para Shannon información es una medida que permite estimar el comportamiento de la señal o mensaje producido por la fuente de información, la cual puede ser entendida también como el amplio número de posibilidades de generación de señales desde la fuente de información, sin importar su significado. La construcción abstracta de la información surge a partir de la pregunta: ¿existe una medida de la información? o ¿cómo cuantificar la información producida por la fuente? y ¿cuál es la unidad de medida de la información? Las respuestas a estas preguntas fueron dadas por Shannon, quien desarrolló los conceptos matemáticos de: medida de información y entropía. Antes

[4]Estudia y mide la realidad no determinista.

de presentar la definición formal, se introducen las ideas que subyacen a estos dos conceptos.

En la India aproximadamente en el siglo II a.C. se planteó el siguiente problema en la farmacopea ayurvédica: existen seis sabores medicinales[5], el problema a resolver es encontrar: ¿cuántas posibles combinaciones pueden generarse con estos seis sabores tomando uno a la vez, dos a la vez, y así hasta tener los seis sabores juntos? [50].

La solución de este problema se presenta en la Fig. 2.2, en la cual los sabores son representados por a, b, c, d y f. Para contar las combinaciones posibles, se pregunta, ¿se añade a? y las respuestas sí o no dan paso para b, c y así sucesivamente. Se realizan 6 preguntas con respuestas de si o no. El diagrama de árbol de la Fig. 2.2 muestra todas las posibles combinaciones o secuencias que se pueden formar como resultado de añadir o no un sabor. El número total de secuencias posibles es $64 = 2^6$, incluyendo la secuencia en la que ningún sabor se añade, por lo tanto, el número real de combinaciones es 63 [50, 51].

Si se diseña un sistema de comunicación para el cual, la fuente de información es el conjunto de combinaciones o secuencias de sabores posibles y el canal transmite un sabor o un resultado de las posibles combinaciones, surge la pregunta: ¿cómo medir la información al seleccionar una secuencia de sabores del conjunto de combinaciones posibles?

Sea $\mathcal{A}$ el conjunto de combinaciones y A la variable aleatoria que representa la selección de un elemento de $\mathcal{A}$. Una forma de medir la información es contar el número de preguntas necesarias para identificar un elemento de $\mathcal{A}$. Por lo tanto, si el número total de secuencias, i.e., $|\mathcal{A}|$, es conocido, se considera nombrar cada elemento con una palabra binaria, i.e., una secuencia de ceros y unos tal que: el '1' indica la presencia de un sabor en la combinación y el '0' lo contrario, por ejemplo, en la Fig. 2.3 se observa que en la combinacion de sabores $\{a, b, d, f\}$ la palabra binaria correspondiente es 110101. La longitud de la palabra es definida por: $\log |\mathcal{A}|$ bits[6], donde la magnitud bits surge porque la base logarítmica es binaria, en general,

[5] *Madhura, amla, lavana, bitter, pungent, kasaya*: dulce, ácido, salado, amargo, picante, astringente.

[6] Dígitos binarios (*Binary Digits*).

para las variables aleatorias discretas lo anterior es acorde a la Definición 2.1.

Definición 2.1 (*The essential bit content* [8])**.** El número total de bits que permite representar o identificar cada uno de los estados o elementos de una variable aleatoria es:

$$H_0(A) \triangleq \log |\mathcal{A}|. \tag{2.1}$$

De la Definición 2.1, la medida de información es $\log |\mathcal{A}| = 6$ bits, para el problema planteado. Esta medida corresponde al número de bits necesarios para identificar una combinación de sabores, y es al mismo tiempo el número de preguntas binarias realizadas para obtener el número total de secuencias posibles, i.e, $2^6 = 64$ o lo que es lo mismo, el número de preguntas necesarias para identificar una secuencia de $\mathcal{A}$.

En el problema planteado, la selección que realiza el sistema de un elemento de $\mathcal{A}$ es equiprobable, i.e., todos los elementos tienen igual probabilidad de ser seleccionados. Por lo tanto, la Definición 2.1, no incluye ningún factor probabilístico y solo introduce la cuantificación en bits de la información.

El siguiente problema es plantear el carácter probabilístico de la medida de información, el cual surge de la observación de los eventos más ocurrentes o seguros con respecto a eventos improbables (en los cuales hay mayor incertidumbre) en el proceso de comunicación.

En otro ejemplo, se quiere determinar la medida de información que producen dos cartas, las cuales son escritas utilizando los caracteres del alfabeto del idioma que le corresponden, por ejemplo, si es el alfabeto español se cuenta con 27 letras. La primera carta es generada por una máquina en la cual, cada letra ocurre aleatoriamente con la misma probabilidad que cualquier otra letra del alfabeto, y la segunda, es escrita por una persona.

De acuerdo al ejemplo anterior, una aproximación a la medida de información de la primera carta, corresponde al número de preguntas que permiten determinar cada letra seleccionada, dado que todas las letras son equiprobables, entonces, para identificar una letra se requieren $\log 27 = 4.75$ bits por letra. Si por ejemplo, la carta tiene un total de 25 letras, la medida de información del texto seria proporcional

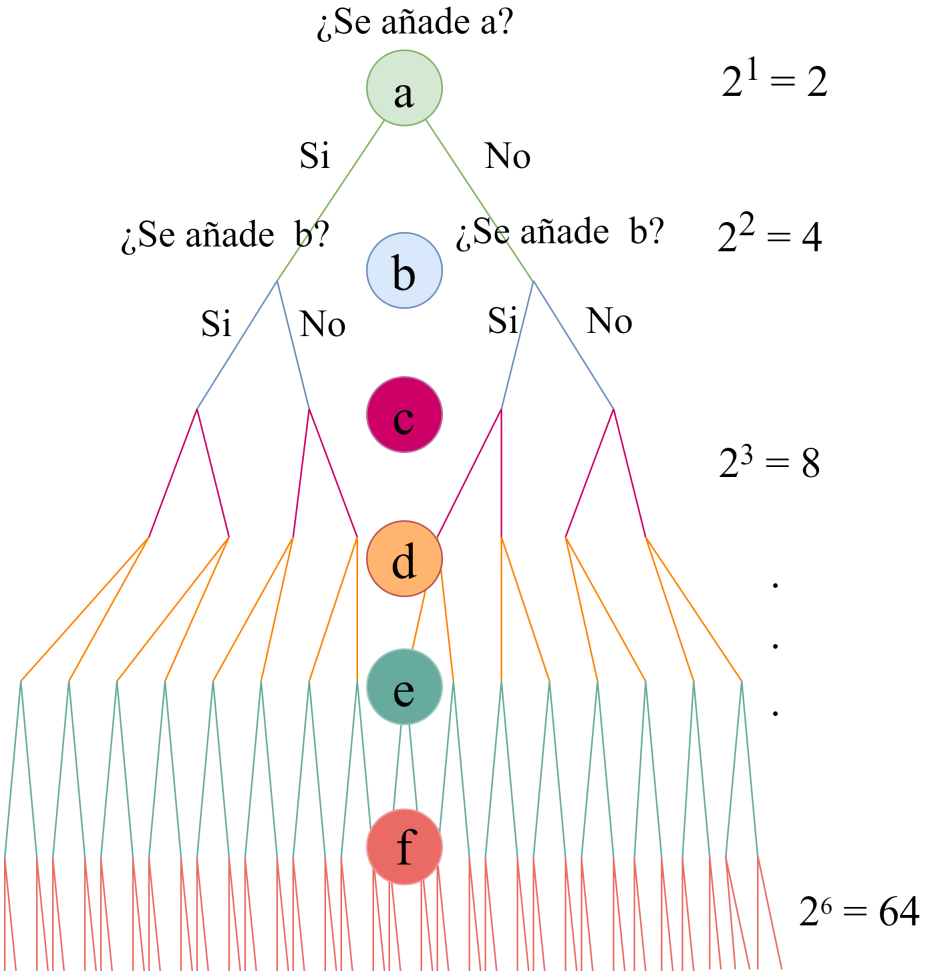


Figura 2.2: Combinatoria de los seis sabores ayurvédicos.

a $25 \times 4.75 = 118.87$ bits, i.e., se requiere realizar un total de 118.87 preguntas en promedio para identificar las letras que componen el escrito, donde por cada letra se realizan alrededor de 4.75 preguntas binarias.

Sin embargo, en la segunda carta el lenguaje ofrece una estructura que hace de unas letras más ocurrentes que otras, por lo tanto, se puede deducir las letras del escrito más probables con mayor rapidez, si se conocen las probabilidades de las letras del alfabeto en cuestión.

La Fig. 2.4a [52] y la Fig. 2.4b [8] presentan la frecuencia de ocurrencia de las letras,

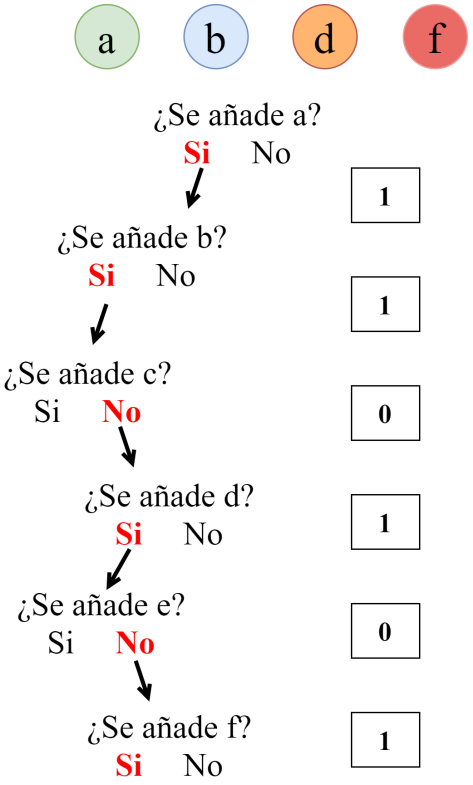


Figura 2.3: Un resultado.

sin considerar signos de puntuación, del texto "La Regenta[7]" en español y en inglés, respectivamente. Se observa en la Fig. 2.4 que en el español la letra a y en el inglés la letra e ocurren con mayor frecuencia. Por lo cual, si se quiere de forma similar transmitir el texto transformado en secuencias de ceros y unos, la longitud de ceros y unos que representa la a o la e debería ser mínima, ya que son las letras más probables, pero, para letras como la k y w en el español la longitud debería tener un mayor número de bits, lo cual se puede entender, porque se requiere de un mayor número de preguntas para identificar un elemento menos probable.

En general, los diversos alfabetos humanos tienen asociada una estructura estadística que permite plantear una codificación para la transmisión de mensajes con ahorro de espacio y tiempo, al asignar un código a las letras o grupos de letras que depende de tal estructura estadística, con el fin de aprovechar tal comportamiento para

[7]La Regenta es la primera novela de Leopoldo Alas Clarín, publicada en dos tomos en 1884 y 1885 [52].

desarrollar un sistema de comunicación óptimo.

En consecuencia, la medida de información del mensaje en la segunda carta está determinada por la estructura estadística de los elementos del mensaje, i.e., que la Definición 2.1 no es suficiente, porque se quiere medir ¿cuánta información aportan los eventos más probables (o más ocurrentes) y los eventos menos probables (o posibles pero improbables)?

La medida de información cuantifica el grado de incertidumbre de los símbolos o letras que componen los mensajes generados por la fuente de información, lo cual se encuentra expresado en la primera parte de la Definición 2.2.

Definición 2.2 (*Shannon information content and Entropy* [8])**.** Sea X una variable aleatoria, la cual toma valores en el alfabeto $\mathcal{X} = \{a_1, a_2, \ldots, a_n\}$ y su función distribución de probabilidad (pmf, *probability mass function*) p_X, se denota para cada $p_X(x = a_i)$ como p_i, con $p_i \geq 0$, y $\sum_i^n p_i = 1$. Por lo tanto,

- La medida de información se define para cada $x \in \mathcal{S}_X$, como:

$$\iota(x) \triangleq -\log p_i, \tag{2.2}$$

  cuya unidad de medida es el bit por símbolo dado que la base del logaritmo utilizada es dos. La función $\iota(x)$ es una nueva variable aleatoria que cuantifica la cantidad de información proporcionada por el evento $X = x$.

- La entropía de la variable aleatoria X se define como:

$$H(X) = -\sum_{a_i \in \mathcal{S}_X}^{n} p_i \log p_i. \tag{2.3}$$

  cuya unidad de medida es el bit. La entropía $H(X)$ es una función que depende solamente de $p_X(x)$. $H(X) = H(p_1, p_2, \ldots, p_n) = 0$ si y solo si, un elemento tiene probabilidad uno (no incertidumbre) por lo tanto no hay información. La entropía además tiene su valor máximo cuando $p_i = \frac{1}{n}$ con $i \in \{1, 2, \ldots, n\}$ (distribución equiprobable) i.e., $H(X) = \log n$ donde n es la cardinalidad del conjunto, i.e., $n = |\mathcal{X}|$.

La entropía de la variable X también puede ser definida como el promedio ponderado de la medida de información i.e., el valor esperado de una variable aleatoria equivalente a la medida de información:

$$H(X) = \mathbb{E}_X[\iota(x)]. \tag{2.4}$$

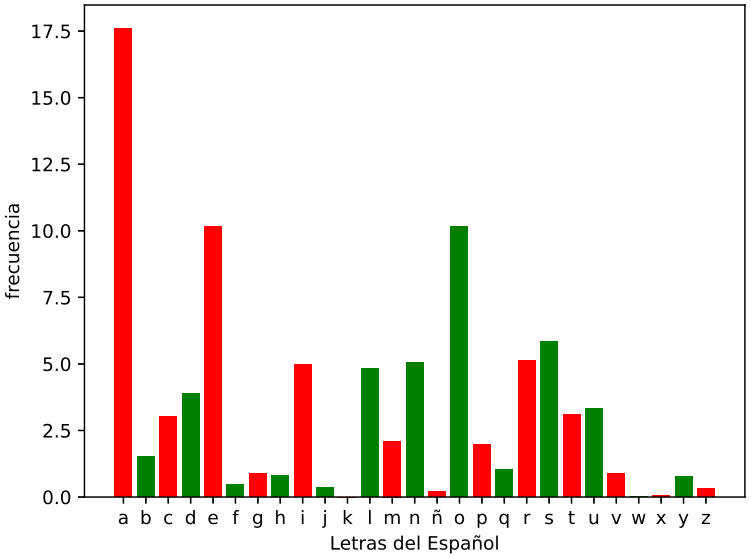
La medida de información $\iota(x)$ de una realización x se comporta como se observa en la Fig. 2.5, y muestra que la información es cero, cuando la probabilidad es uno, es decir, cuando no hay incertidumbre y por lo tanto el resultado no aporta información, en cambio, una probabilidad cercana a cero (un evento extraño o de rara ocurrencia), aporta una mayor cantidad de información. En la Tabla 2.1, se encuentra la medida de información para cada una de las letras del alfabeto español del ejemplo anterior, donde se observa que la letra a tiene una medida de información de 2.51 bits y la letra k una medida de información de 15.02 bits. Así un resultado incierto requiere de mayor información para eliminar la incertidumbre, de la misma manera, de un resultado cierto que no representa ningún tipo de incertidumbre, no se requiere información para conocer de su ocurrencia [53].

La entropía definida en (2.3) se calcula para el ejemplo con los datos de la Tabla 2.1, de donde se obtiene que la entropía sobre la distribución de probabilidad del alfabeto español es $H(X) = 3.60$ bits. Este resultado para el texto la "La Regenta" indica el promedio de preguntas que permiten identificar las letras de este texto según su estructura estadística, el cual comparado con el número de preguntas requeridas para identificar una letra de la primera carta (4.75 bits) es menor. Shannon encontró que $H(X)$ es el límite fundamental de preguntas promedio necesarias para identificar un resultado de un experimento aleatorio.

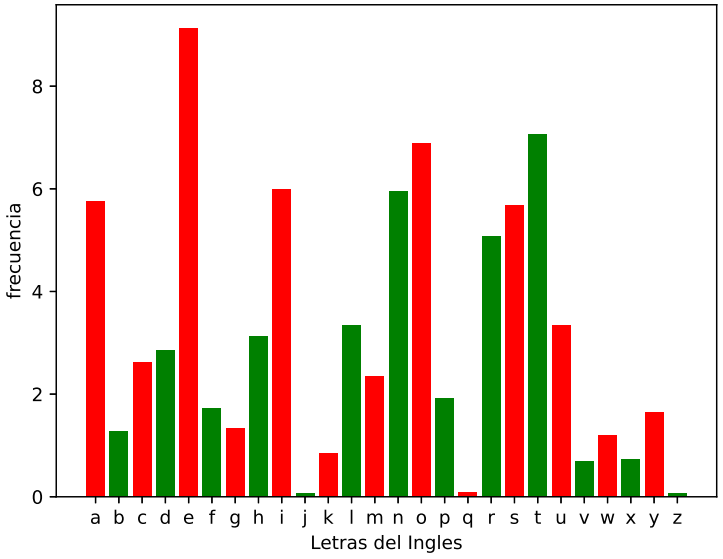
La entropía de la variable aleatoria binaria X denotada por $H_2(X)$, con alfabeto $\mathcal{X} = \{0, 1\}$ y con pmf: $p_X(0) = 1 - p_X(1) = p$ y $0 \leq p \leq 1$, por la segunda parte de la Definición 2.2 es:

$$H_2(X) = -p \log p - (1 - p) \log(1 - p). \tag{2.5}$$

La Fig. 2.6 presenta la entropía de una variable aleatoria binaria $H_2(X)$ de (2.5), donde, la entropía es igual a cero si la probabilidad p es nula o es igual a uno i.e., que



(a) Frecuencias de las letras en el Español.



(b) Frecuencias de las letras en el Inglés.

Figura 2.4: Frecuencias de ocurrencia de letras.

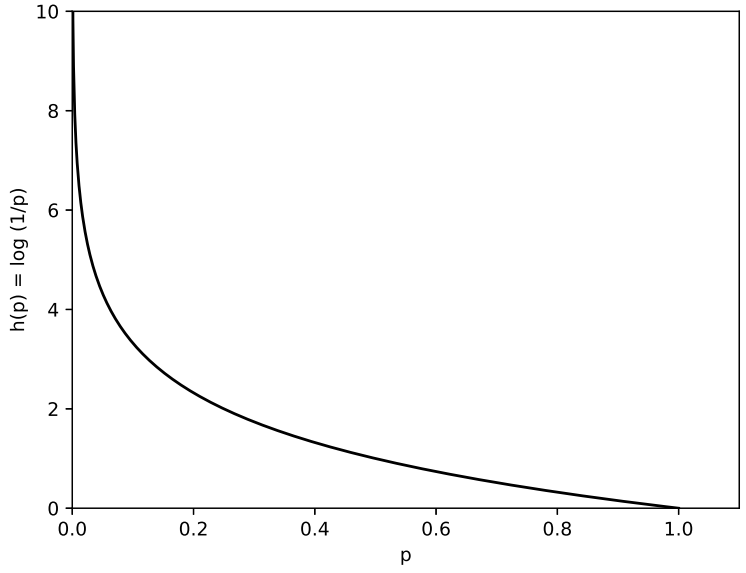


Figura 2.5: Medida de Información.

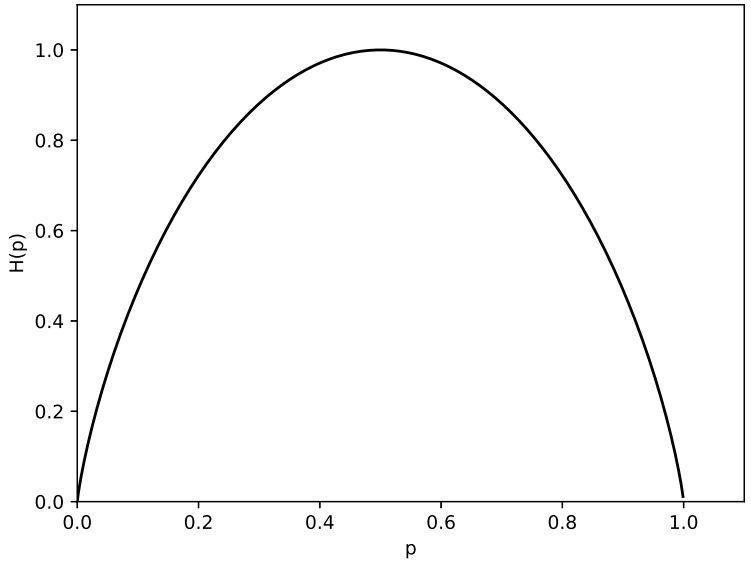


Figura 2.6: Entropía de la variable aleatoria binaria X.

estos estados no aportan información (no se requiere información para determinar ese estado). Como se mencionó en la Definición 2.2, la entropía es máxima cuando los resultados de la variable aleatoria son equiprobables. Por lo tanto, la entropía puede ser vista como la medida del grado de incertidumbre de un proceso estocástico.

2.2.2.1. Medida de Información para Variables Aleatorias Independientes

Sean X y Y dos variables aleatorias (discretas) i.i.d. con realizaciones x y y, respectivamente. Por la definición de independencia de variables aleatorias, la pmf conjunta puede ser expresada de la siguiente manera:

$$p_{XY}(x, y) = p_X(x)p_Y(y). \tag{2.6}$$

La cantidad de información necesaria para identificar los resultados de manera conjunta de dos variables aleatorias se incrementa. Al aplicar en (2.2) la pmf conjunta de X y Y, la medida de información conjunta requerida es:

$$\iota(x, y) = -\log p_{XY}(x, y) = -\log p_X(x)p_Y(y) = -\log p_X(x) - \log p_Y(y), \tag{2.7}$$

En palabras, al observar las variables aleatorias independientes x y y de manera conjunta, la medida de información es igual a la suma de la medida de información de x más la medida de información de y, como si cada una se hubiese observado separadamente. La medida de información de x, y es:

$$\iota(x, y) = \iota(x) + \iota(y) \quad \text{si } x \text{ y } y \text{ son independientes.} \tag{2.8}$$

La entropía también es aditiva para las variables aleatorias independientes:

$$H(X, Y) = H(X) + H(Y), \tag{2.9}$$

lo cual sigue inmediatamente de la Definición 2.2.

i	a_i	p_i	$\iota(a_i)$
1	a	0.1759	2.51
2	b	0.0154	6.02
3	c	0.0302	5.04
4	d	0.0387	4.69
5	e	0.1016	3.29
6	f	0.0048	7.70
7	g	0.0087	6.84
8	h	0.0079	6.98
9	i	0.0497	4.33
10	j	0.0037	8.07
11	k	0.00003	15.02
12	l	0.0481	4.37
13	m	0.0210	5.57
14	n	0.0504	4.31
15	ñ	0.0020	8.96
16	o	0.1016	3.29
17	p	0.0196	5.67
18	q	0.0102	6.62
19	r	0.0514	4.28
20	s	0.0584	4.09
21	t	0.0310	5.01
22	u	0.0332	4.91
23	v	0.0089	6.81
24	w	0.0001	13.28
25	x	0.0007	10.48
26	y	0.0078	7.00
27	z	0.0032	8.28

Tabla 2.1: Distribución de Probabilidad del Alfabeto Español del texto la Regenta.

2.3. SISTEMA DE COMUNICACIÓN DISCRETO SIN RUIDO

El sistema de comunicación discreto es aquel en el que tanto el mensaje producido por la fuente como la señal que se transmite por el canal son secuencias de símbolos discretos. El canal recibe también el adjetivo de discreto. Aquí se analizan los sistemas discretos más sencillos que fueron la base del desarrollo de esta teoría: el teletipo y la telegrafía.

Se hace necesario, por lo tanto, examinar el significado de canal discreto. Con esta expresión Shannon describe el sistema mediante el cual, partiendo de un conjunto finito de símbolos elementales $s_1, s_2 \ldots, s_n$ con duración temporal t_i en segundos para cada s_i, la cual puede ser distinta para cada símbolo, se realizan selecciones sucesivas formando una secuencia de símbolos a medida que el tiempo avanza, la cual es transmitida de un punto a otro. El sistema de transmisión envía la información con una cierta velocidad, la cual se considera en el caso más sencillo como la tasa de bits o velocidad de transmisión de datos a la que la fuente emite símbolos por unidad de tiempo y se denota por R. En general, en la comunicación digital los símbolos estan representados por caracteres binarios “0”, o “1”, llamados bits, y por lo tanto, R se mide en bits por unidad de tiempo.

El sistema de comunicación discreto con un conjunto finito de símbolos elementales, debe estar diseñado para transmitir todas las posibles secuencias o un subconjunto de ellas. El número de secuencias elegidas se postulan como las señales a transmitir a través del canal discreto.

2.3.1. Entropía de la Fuente Discreta

Implícitamente se han planteado los problemas que Shannon definió: ¿cómo describir matemáticamente una fuente de información? y ¿cuánta información en bits por segundo se produce en una fuente determinada?

La fuente de información discreta produce los mensajes como secuencias de símbolos.

Por lo tanto, la fuente de información se define por un conjunto finito de símbolos y un conjunto de probabilidades que corresponden a la probabilidad de ocurrencia de cada símbolo. Por ejemplo, dado que el conjunto de letras del español ocurren en un texto con determinada probabilidad como ya se expuso en la sección anterior (2.2.2), el lenguaje escrito es una fuente de información discreta y la selección de una letra es una variable aleatoria, por lo tanto, las selecciones sucesivas de letras que forman un escrito se representan como una secuencia de variables aleatorias o un vector de variables aleatorias.

Las variables aleatorias pueden ser dependientes o independientes estadísticamente (ver Apendice A). La realidad es un conjunto de fenómenos correlacionados, pero la comprensión de ellos se realiza considerando solo ciertos aspectos y no la totalidad. De esta manera, en el caso de variables aleatorias dependientes, la fuente de información es determinada por un conjunto de probabilidades de transición de un símbolo a otro (o probabilidades condicionales) y un conjunto de estados. En el caso de variables aleatorias i.i.d, solo hay un estado, por la propiedad de independencia estadística de la realización de cada variable aleatoria y el conjunto de probabilidades solo corresponde a la probabilidad de ocurrencia del símbolo en cuestión.

Así, la fuente de información esta caracterizada como un proceso ergódico[8] y por la secuencia de variables aleatorias $\{X_m, m \in \mathbb{Z}^+ \text{ y } m \geq 1\}$ indexadas por el índice m, donde las variables aleatorias X_m tienen similar distribución de probabilidad p_X y alfabeto $\mathcal{X}$; X denota la variable aleatoria genérica, con entropía $H(X)$.

Sea un conjunto finito de estados posibles de la fuente de información $s_1, s_2, \cdots, s_i$, con $i \in \{1, 2..., n\}$, y un conjunto de probabilidades de transición $p_{j|i}(j|i)$ denotada $p_i(j)$ para cada estado s_i, que define la probabilidad de pasar del estado i al estado j. Para hallar la entropía de la fuente de información se define la entropía de cada estado H_i, la cual se calcula, por la pmf $p_i(j)$, de acuerdo a la Definición 2.2, y como se presenta a continuación:

$$H_i = -\sum_{j=1}^{n} p_i(j) \log p_i(j) \tag{2.10}$$

[8]Cada una de las secuencias generadas por el proceso ergódico tiene las mismas propiedades estadísticas.

La probabilidad de ocurrencia del estado s_i es p_i, por tanto, la entropía H de la fuente es:

$$H = \sum_{i=1}^{n} p_i H_i. \tag{2.11}$$

En caso de que se trate de variables aleatorias independientes, donde p_i es la probabilidad del símbolo i, y no como las descritas anteriormente, símbolos sucesivos dependientes, el cálculo de la entropía se simplifica en (2.12):

$$H = -\sum_{i \in \mathcal{S}_X}^{n} p_i \log p_i. \tag{2.12}$$

En el Teorema 3 en [1] se plantea el análisis de la probabilidad de las secuencias generadas por la fuente de información $p_{\mathbf{X}}(\mathbf{x})$ denotada como p, con respecto a la entropía genérica de la fuente $H(X)$, la cual está definida para X en (2.3). El Teorema define a partir del comportamiento de la probabilidad de las secuencias para grandes valores de n, una división de todas las secuencias posibles que pueden ser generadas por la fuente de información en dos conjuntos, los cuales se analizan a continuación.

Al considerar una fuente de información con m símbolos, que genera una larga secuencia de símbolos de longitud n, por la Ley de los Grandes Números (LLN, *Law of Large Numbers*) hay una alta probabilidad de tener alrededor de $p_1 n$ ocurrencias del primer símbolo, $p_2 n$ ocurrencias del segundo símbolo, y así sucesivamente, donde $p_1, p_2, \ldots, p_m$ son las probabilidades de cada símbolo, respectivamente. La probabilidad p de la secuencia en particular será aproximadamente:

$$p \simeq p_1^{p_1 n} p_2^{p_2 n} \cdots p_m^{p_m n}. \tag{2.13}$$

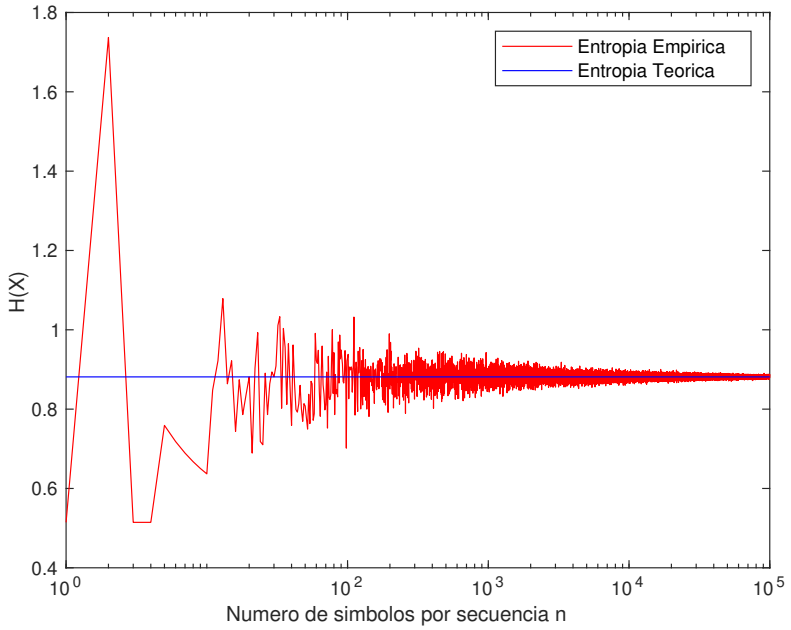


Figura 2.7: Entropía Experimental de una secuencia binaria con $p_X(0) = 0.7$.

de (2.13) se puede obtener que:

$$\log p \simeq \log p_1^{p_1 N} p_2^{p_2 N} \cdots p_n^{p_n N} \tag{2.14}$$

$$= p_1 N \log p_1 + p_2 N \log p_2 + \cdots + p_n N \log p_n$$

$$= N \sum_{i=1}^{n} p_i \log p_i \tag{2.15}$$

$$= -NH. \tag{2.16}$$

Así se obtiene,

$$H \simeq -\frac{1}{N} \log p \tag{2.17}$$

La expresión en (2.17), es llamada *entropía empírica* de una secuencia típica [10] [54], es una aproximación a las definiciones teóricas en (2.11) y (2.12) y permitió establecer el concepto probabilístico de secuencias típicas. En la Fig. 2.7 [54] se observa que a medida que n es lo suficientemente grande la entropía experimental se aproxima a la entropía teórica, como lo establece el Teorema 2.1.

Teorema 2.1 (Teorema 3 en [1]). *Sean $\epsilon > 0$ y $\delta > 0$, es factible encontrar un n_0, de tal forma que las secuencias de longitud $n \geq n_0$ se dividen en dos conjuntos:*

1. *Un conjunto para el cual la probabilidad total es menor que ϵ.*
2. *El resto, forma un conjunto que satisface la desigualdad,*

$$\left| -\frac{1}{n}\log p - H \right| < \delta. \tag{2.18}$$

Este teorema divide el conjunto total de las secuencias posibles en dos: el conjunto no-típico y el conjunto típico. La prueba de este teorema la da Shannon en el Apéndice 3 de [1], utilizando la Ley Fuerte de los Grandes Números (SLLN, *Strong Law of Large Numbers*) presentada en el Apéndice D. Sin embargo, dado que algunos autores afirman la poca rigurosidad matemática en la demostración del teorema, los lectores pueden encontrar una demostración detallada del Teorema 2.1 y los siguientes teoremas en [1], que generalizan la teoría de Shannon [53, 55]. Para profundizar no sólo en el análisis matemático sino en la comprensión de este teorema, del que nace la idea de secuencias típicas, se abordará desde diferentes puntos de vista que han analizado teóricamente el tema, entre ellos [8–10, 53, 54, 56, 57].

2.3.2. Capacidad del Canal

Surge la pregunta respecto al canal discreto, ¿cómo se puede medir la capacidad del canal para transmitir información? y, ¿qué entender por capacidad?

Para hallar la medida de la capacidad de transmisión de información C, se considera el canal del teletipo cuyo funcionamiento se describe en la Fig. 1.5 [26], sistema que utilizaba el código Baudot, como se expuso en la Sección 1.1.

El sistema del teletipo permite transmitir 32 símbolos, codificados en 5 pulsos eléctricos o 5 bits, i.e., $\log 32$, asumiendo todos los símbolos equiprobables. Para hallar la medida de la capacidad de transmisión de información C del sistema en bits por segundo, se parte del numero bits por símbolos y se multiplica por la cantidad de

símbolos a transmitir por unidad de tiempo. Esto es,

$$C = 5\frac{\text{bits}}{\text{símbolo}} \times n\frac{\text{símbolos}}{\text{segundo}} = 5n \quad \text{bits/segundo} \tag{2.19}$$

La capacidad del canal teletipo $5n$, es la máxima velocidad de transmisión de datos que puede ser alcanzada. En el caso del teletipo, no hay ninguna restricción en las secuencias o posibles combinaciones de símbolos. Sin embargo, una generalización debe contar con los sistemas "para los que existen diferentes longitudes de símbolos y restricciones en las secuencias permitidas" [1].

En la teoría de la información se establecen las posibilidades y las limitaciones fundamentales de los sistemas de comunicación mediante la definición de conceptos, tales como: entropía y capacidad de canal [8], donde la entropía mide la información promedio generada por la fuente de información con respecto a la distribución de probabilidad de la variable aleatoria de entrada al canal, y la capacidad del canal, es el límite de transmisión de datos que el sistema impone independiente de la velocidad de producción de símbolos en la fuente de información. La Definición 2.3 del parámetro C asociado al canal discreto, determina la máxima velocidad de transmisión de datos para cualquier sistema de comunicación discreto sin ruido.

Definición 2.3 (Definición 1 en [1])**.** La capacidad de una canal discreto sin ruido C es la máxima tasa a la cual el canal discreto puede transmitir información, y es definida por:

$$C = \lim_{t\to\infty} \frac{\log N(t)}{t}, \tag{2.20}$$

donde, $N(t)$ es el número de secuencias permitidas de duración t. Para los sistemas discretos sin ruido, el límite en general es un número finito.

El Lema 2.2 es una forma de validar la Definición 2.3.

Lema 2.2. *Para todo* $t \in \mathbb{R}$ *y para todo* $(t_1, t_2, \ldots, t_n) \in \mathbb{R}^n$ *existe un vector* $\mathbf{a} = (a_1, a_2, \ldots, a_n) \in \mathbb{N}^n$ *tal que,*

$$t = a_1t_1 + a_2t_2 + \ldots + a_nt_n + \delta \quad con \quad \delta < \text{mín}\{t_i\}. \tag{2.21}$$

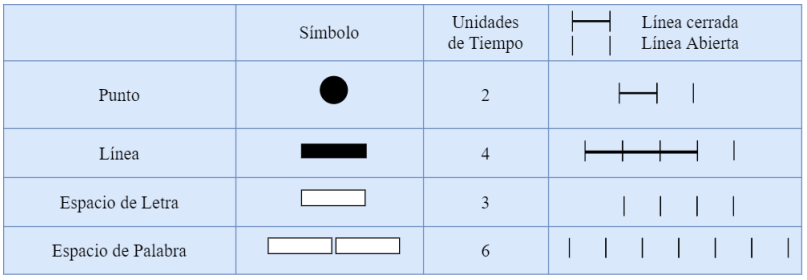


Figura 2.8: Símbolos del telégrafo.

Para todo $t > \max\{t_i\}$ *donde* $i \in \{1, 2, \ldots, n\}$

$$N(t) = N(t - t_1) + N(t - t_2) + \cdots + N(t - t_n). \tag{2.22}$$

La Definición 2.3 propone un análisis que resuelva el conteo del número total de secuencias posibles en un tiempo t, para lo cual se aplica el Lema 2.2, el cual es una aplicación del cálculo de recurrencias lineales [58]. La solución analítica a (2.22) se obtiene por el cálculo de la ecuación característica y de las raíces reales, y se selecciona la solución real mayor, ya que define el comportamiento asintótico del número total de secuencias.

A continuación, se analiza el canal de telegrafía y se calcula la capacidad por la Definición 2.3 y el Lema 2.2.

El sistema de telegrafía utiliza los símbolos, punto, línea, espacio de letra y espacio de palabra, los cuales se obtienen por la codificación en línea abierta y línea cerrada del sistema de comunicación, de acuerdo a una cierta duración, como se observa en la Fig. 2.8. Estos símbolos primarios permiten definir un conjunto de símbolos utilizados para codificar los mensajes, limitados por ciertas restricciones, que consisten en que dos espacios no pueden ser consecutivos, lo cual se representa gráficamente como una transición de símbolos entre dos estados en la Fig. 2.9, y por lo tanto, los símbolos de la fuente se convierten en los que se presentan en la Fig. 2.10, con sus respectivos tiempos.

Con los datos de la Fig. 2.10 se tiene un conjunto de símbolos $\mathcal{W} = \{s_1, s_2, s_3, s_4, s_5, s_6\}$

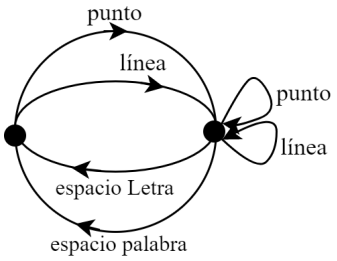


Figura 2.9: Representación gráfica de las restricciones de los símbolos del telégrafo. Tomada de "A Mathematical Theory of Communications".

Restricciones	Símbolos	Unidades de Tiempo
Punto	●	2
Línea	▬	4
Punto - Espacio de Letra	● □	5
Punto - Espacio de Palabra	● □ □	8
Línea - Espacio de Letra	▬ □	7
Línea - Espacio de Palabra	▬ □ □	10

Figura 2.10: Restricciones sobre las secuencias de símbolos.

con los tiempos respectivos $t_1 = 2, t_2 = 4, t_3 = 5, t_4 = 7, t_5 = 8, t_6 = 10$, y se obtiene:

$$t = 2a_1 + 4a_2 + 5a_3 + 7a_4 + 8a_5 + 10a_6 + \delta, \quad \delta < 2. \tag{2.23}$$

El cálculo de capacidad de este sistema consiste en contar el número de secuencias de duración $t < \text{máx}\{t_i\}$ y aplicar el Lema 2.2, esto es:

$$N(t) = N(t-2) + N(t-4) + N(t-5) + N(t-7) + N(t-8) + N(t-10). \tag{2.24}$$

El número de secuencias $N(t)$ para $t < \text{máx}\{t_i\} = 10$ se expresan en la Tabla 2.2. A partir de estos resultados se puede calcular $N(t)$ para valores de t lo suficientemente grandes utilizando los resultados anteriores. Dado el crecimiento exponencial del número de secuencias posibles a medida que t crece, en la Fig. 2.11a y en la

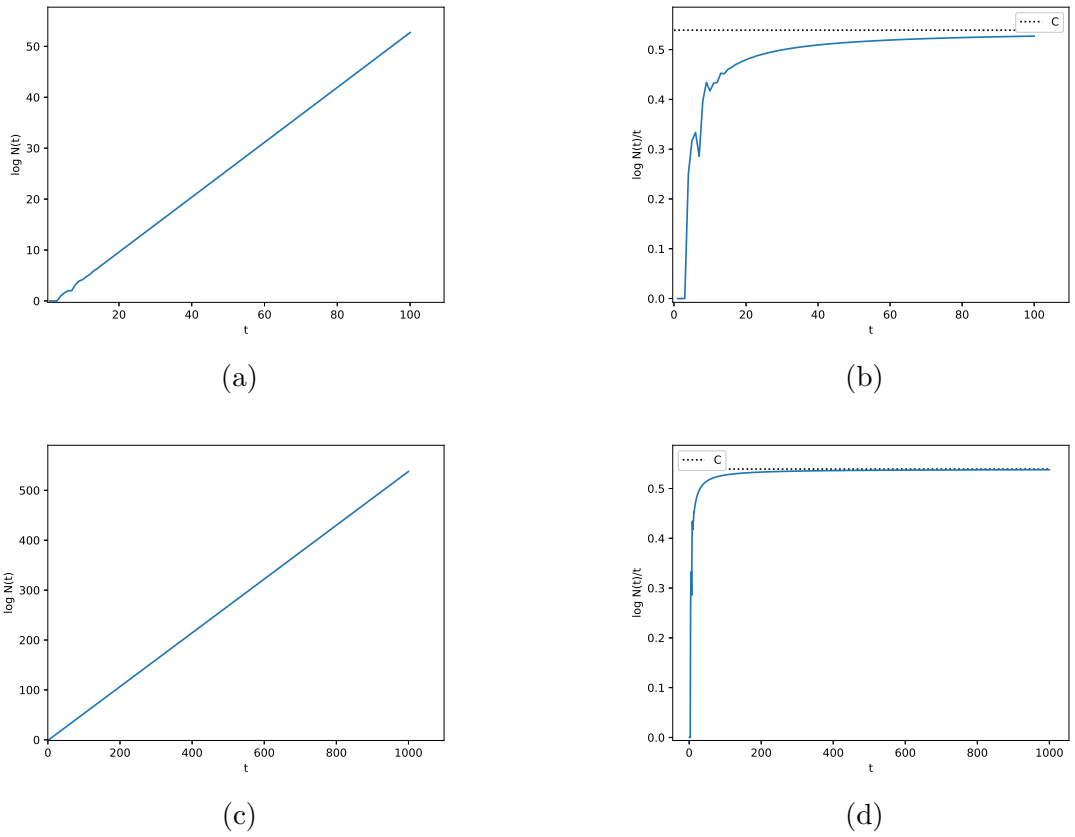


Figura 2.11: Capacidad del sistema de Telegrafía.

Fig. 2.11c se presenta el comportamiento lineal del $\log N(t)$ con respecto a t, para $t = 100$ y $t = 1000$, respectivamente. Y para los mismos valores de t, la relación $\frac{\log N(t)}{t}$ en la Fig. 2.11b y en la Fig. 2.11d muestran el valor al cual converge la capacidad representada por la línea punteada y obtenida por la Definición 2.3, y la cual es aproximadamente 0.53 bits por segundo, la cual se calcula analíticamente a continuación.

La forma analítica de resolver (2.24) es aplicando el análisis de recurrencias lineales, que se presenta en el Apéndice C. La ecuación característica de (2.24) es:

$$1 = a^{10} + a^8 + a^7 + a^5 + a^4 + a^2. \tag{2.25}$$

t	**a**	δ	Secuencias	$N(t)$	$\log N(t)/t$
1	**0**	1	-	-	-
2	{(1,0,0,0,0,0) }	0	s_1	1	0
3	{(1,0,0,0,0,0) }	1	s_1	1	0
4	{(2,0,0,0,0,0), (0,1,0,0,0,0) }	$0, 0$	s_1s_1 s_2	2	0.25
5	{(2,0,0,0,0,0), (0,1,0,0,0,0), (0,0,1,0,0,0) }	$1, 1, 0$	s_1s_1 s_2 s_3	3	0.316
6	{(3,0,0,0,0,0), (1,1,0,0,0,0), (0,0,1,0,0,0) }	$0, 0, 1$	$s_1s_1s_1$ s_1s_2 s_2s_1 s_3	4	0.333
7	{(3,0,0,0,0,0), (1,1,0,0,0,0), (0,0,0,1,0,0) }	$1, 1, 0$	$s_1s_1s_1$ s_1s_2 s_2s_1 s_4	4	0.285
8	{(4,0,0,0,0,0), (2,1,0,0,0,0), (1,0,1,0,0,0), (0,2,0,0,0,0), (0,0,0,1,0,0), (0,0,0,0,1,0) }	$0, 0, 1,$ $0, 1, 0$	$s_1s_1s_1s_1$ $s_1s_1s_2$ $s_1s_2s_1$ $s_2s_1s_1$ s_1s_3 s_3s_1 s_2s_2 s_4 s_5	9	0.396
9	{(4,0,0,0,0,0), (2,1,0,0,0,0), (1,0,1,0,0,0), (0,2,0,0,0,0), (0,0,0,0,1,0), (2,0,1,0,0,0), (0,1,1,0,0,0), (1,0,0,1,0,0) }	$1, 1, 1,$ $1, 1, 0,$ $0, 0$	$s_1s_1s_1s_1$ $s_1s_1s_2$ $s_1s_2s_1$ $s_2s_1s_1$ s_1s_3 s_3s_1 s_2s_2 s_5 $s_1s_1s_3$ $s_1s_3s_1$ $s_3s_1s_1$ s_2s_3 s_3s_2 s_1s_4 s_4s_1	15	0.434

Tabla 2.2: Número de Secuencias de duración t

Las raíces reales de (2.25) son:

$$a_1 = -0.8627, \quad a_2 = 0.688 \tag{2.26}$$

Y la capacidad es por (C.18):

$$C = 0.539 \quad \text{bits por segundo.} \tag{2.27}$$

2.4. SECUENCIAS TÍPICAS

En la Sección 2.3.1 se introdujó el Teorema 2.1 el cual establece la división en dos conjuntos de todas las secuencias posibles generadas por un proceso estocástico: el conjunto típico y el conjunto no típico, lo cual formalizado plantea para una larga secuencia de variables aleatorias i.i.d. la Propiedad de Equipartición Asintótica (AEP, *Asymptotic Equipartition Property*), la cual tiene una analogía con la ley de los grandes números (LLN) en teoría de la probabilidad. La LLN tiene dos formas: la ley débil de los grandes números (WLLN) y la ley fuerte de los grandes números (SLLN). En teoría de la información, el equivalente es la *AEP* débil y la *AEP* fuerte, las cuales permiten plantear y demostrar los teoremas fundamentales de la teoría de la información.

2.4.1. Propiedad de Equipartición Asintótica débil

La fuente de información se considera como una secuencia de variables aleatorias $\{X_m, m \in \mathbb{Z}^+ \text{ y } m \geq 1\}$, donde las variables aleatorias X_m son i.i.d., cada una con similar distribución de probabilidad p_X y alfabeto $\mathcal{X}$; X denota la variable aleatoria genérica, con entropía $H(X)$.

Teorema 2.3 (Teorema 5.1 en [10]). *Sea* $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T \in \mathcal{X}^n$ *, con pmf conjunta:*

$$p_{\mathbf{X}}(\mathbf{x}) = p_X(x_1)p_X(x_2)\cdots p_X(x_n). \tag{2.28}$$

La AEP débil de $p_{\mathbf{X}}(\mathbf{X})$ establece la convergencia en probabilidad:

$$-\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X}) \to H(X), \tag{2.29}$$

cuando $n \to \infty$, i.e., para cualquier $\epsilon > 0$ y n suficientemente grande, se cumple:

$$\Pr\left\{\left|-\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X}) - H(X)\right| < \epsilon\right\} \geq 1-\epsilon. \tag{2.30}$$

Prueba: (Lema 60 en [54]) Sea Y una variable aleatoria discreta definida como:

$$Y = -\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X}) \tag{2.31}$$

Se calcula el valor esperado de la variable aleatoria Y, esto es:

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}\left[-\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X})\right] \\
&\overset{(a)}{=} -\frac{1}{n}\sum_{\mathbf{x}\in\mathcal{S}_{\mathbf{X}}^n} p_{\mathbf{X}}(\mathbf{x})\log p_{\mathbf{X}}(\mathbf{x}) \\
&\overset{(b)}{=} -\frac{1}{n}\sum_{m=1}^{n}\sum_{\mathbf{x}\in\mathcal{S}_{\mathbf{X}}^n} p_{\mathbf{X}}(\mathbf{x})\log p_X(x_m) \\
&\overset{(c)}{=} -\frac{1}{n}\sum_{m=1}^{n}\left[\sum_{x_1\in\mathcal{S}_X} p_X(x_1)\cdots\sum_{x_n\in\mathcal{S}_X} p_X(x_n)\right]\log p_X(x_m) \\
&= -\frac{1}{n}\sum_{m=1}^{n}\left[\sum_{x_1\in\mathcal{S}_X} p_X(x_1)\cdots\sum_{x_m\in\mathcal{S}_X} p_X(x_m)\cdots\sum_{x_n\in\mathcal{S}_X} p_X(x_n)\right]\log p_X(x_m) \\
&= -\frac{1}{n}\sum_{m=1}^{n}\sum_{x_m\in\mathcal{S}_X} p_X(x_m)\log p_X(x_m) \\
&= \frac{1}{n}\sum_{m=1}^{n} H(X) \\
&= \frac{1}{n}nH(X) \\
&= H(X),
\end{aligned} \tag{2.32}$$

donde,

(a) Resulta de aplicar el operador valor esperado a la transformación del vector de variables aleatorias $g(\mathbf{X}) = Y$:

$$\mathbb{E}[g(\mathbf{X})] = \sum_{x_1,x_2,\ldots,x_n \in \mathcal{S}_X^n} p_{\mathbf{X}}(x_1, x_2, \ldots, x_n) g(x_1, x_2, \ldots, x_n). \tag{2.33}$$

(b) Resulta de aplicar el logaritmo a la pmf conjunta dada en (2.28) de variables aleatorias independientes, la cual corresponde a una productoria, de lo cual se obtiene:

$$\begin{aligned} \log p_{\mathbf{X}}(\mathbf{x}) &= \log p_X(x_1) p_X(x_2) \ldots p_X(x_n) \\ &= \log p_X(x_1) + \log p_X(x_2) + \ldots + \log p_X(x_n) \\ &= \sum_{m=1}^{n} \log p_X(x_m) \end{aligned} \tag{2.34}$$

(c) De la pmf conjunta $p_{\mathbf{X}}(\mathbf{x})$ definida en (2.28), se obtiene que al aplicar la sumatoria sobre todas las realizaciones del vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ es equivalente a la sumatoria individual (por independencia estadística) de cada realización en el soporte $\mathcal{S}_X$, esto es:

$$\sum_{\mathbf{x} \in \mathcal{S}_{\mathbf{X}}^n} p_{\mathbf{X}}(\mathbf{x}) = \sum_{x_1 \in \mathcal{S}_X} p_X(x_1) \sum_{x_2 \in \mathcal{S}_X} p_X(x_2) \cdots \sum_{x_n \in \mathcal{S}_X} p_X(x_n), \tag{2.35}$$

en la cual se asume cada sumatoria igual a uno:

$$\sum_{x_1 \in \mathcal{S}_X} p_X(x_1) = 1, \quad \sum_{x_2 \in \mathcal{S}_X} p_X(x_2) = 1, \quad \ldots \sum_{x_n \in \mathcal{S}_X} p_X(x_n) = 1 \tag{2.36}$$

Por otro lado, la varianza $\mathrm{Var}[Y]$:

$$\begin{aligned} \mathrm{Var}[Y] &= \mathrm{Var}\left[-\frac{1}{n} \log p_{\mathbf{X}}(\mathbf{X}) \right] \\ &\overset{(a)}{=} \frac{1}{n^2} \mathrm{Var}\big[\log p_{\mathbf{X}}(\mathbf{X}) \big] \\ &\overset{(b)}{=} \frac{1}{n^2} \sum_{m=1}^{n} \mathrm{Var}[\log p_{X_n}(X_n)] \end{aligned}$$

$$= \frac{1}{n}\text{Var}[\log p_X(X)], \tag{2.37}$$

donde,

(a) Resulta de aplicar el operador de varianza a la transformación del vector de variables aleatorias.

(b) De forma similar que en (2.34), la sumatoria de logaritmos se debe a la independencia estadística de las variables aleatorias.

De la desigualdad de Chebyshev (ver Apéndice D), para cualquier $a > 0$, lo siguiente es válido:

$$\Pr\{|Y - \mathbb{E}[Y]| \geq a\} \leq \frac{\text{Var}[Y]}{a^2}. \tag{2.38}$$

Lo cual aplicando (2.31), (2.32) y (2.37) es,

$$\Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X}) - H(X)\right| \geq a\right\} \leq \frac{1}{a^2 n}\text{Var}[\log p_X(X)]. \tag{2.39}$$

Se observa que la variable aleatoria X tiene un valor esperado y varianza finito, por lo tanto, la expresión del lado derecho de (2.39) es finita. Así, para todo $\epsilon' > 0$, existe un n suficientemente grande, de modo que:

$$\Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X}) - H(X)\right| \geq a\right\} \leq \epsilon'. \tag{2.40}$$

Finalmente,

$$\begin{aligned}
&\Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X}) - H(X)\right| < a\right\} \\
&= 1 - \Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X}) - H(X)\right| \geq a\right\} &(2.41)\\
&\geq 1 - \epsilon'. &(2.42)
\end{aligned}$$

Por lo tanto, para todo $\epsilon > 0$, existe un n suficientemente grande, tal que:

$$\Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{X}) - H(X)\right| < \epsilon\right\} \geq 1 - \epsilon. \tag{2.43}$$

$\mathbf{x}$	r	$-\log p_{\mathbf{X}}(\mathbf{x})$
.....	0	51.45
1111111111111111111111111111111111111111111111111111111111111111111111111111	100	173.69
........11.....111.....1.1...1.1...1..1.....1...111....1..1.1..1....111...1.1111111	30	**88.12**
11.....1.....1.....1.....11.....1.....11.....1.....1.....1.....1...11..1.....11.....	17	72.23
1.1.1.....1111.....1.....1.....1...1.....1....1.....1.....1.....1.....	15	69.79
.1...1.....111.....11.....1.....11....1..1.....1.	13	67.34
.....1..1.....1...1..1....1.....1.....1.....1.....1.....1.	11	64.9
.....1.....1.....1.....1.....1.....1.....1.....1.....1.....1.....	9	62.45
...1.....1....1.....1.....1.....1.....1..1.....	7	60.01

Tabla 2.3: Algunos resultados de $\mathbf{X}$, con $p_X(1) = 0.3$ y $n = 100$.

Lo que completa la prueba. ■

El Teorema 2.3 expresa que la entropía $H(X)$ se puede estimar con una probabilidad muy cercana a uno, tanto como se quiera, a partir de una secuencia lo suficientemente grande de variables aleatorias X i.i.d., como se obtiene en (2.43), lo cual implica que las secuencias posibles se dividen en dos clases: aquéllas en las que la magnitud de ocurrencias de los distintos símbolos converge en probabilidad a la función de distribución de probabilidad de X, y el resto. La primera clase se denominan *secuencias típicas*, y estás tienen una probabilidad conjunta cercana a uno a medida que n crece, como se observa en la parte derecha de (2.43). Las demás secuencias ocurren con una probabilidad arbitrariamente pequeña, como se obtiene en (2.40). Las secuencias típicas tienen también la característica de que la probabilidad de cada secuencia es aproximadamente "igual", de lo cual proviene el nombre de equipartición asintótica.

2.4.2. Tipicalidad

En el Teorema 2.3 la división en dos clases de todas las secuencias está determinado por n. ¿Por qué ayuda el incremento de n? Para resolver la pregunta se observa lo que sucede a la distribución de probabilidad binomial en el siguiente ejemplo [8].

Sea $\mathcal{X} = \{0, 1\}$ y sea $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T \in \mathcal{X}^n$ una secuencia de variables aleatorias i.i.d. con $p_X(1) = p_1 = 0.3$. La entropía de la variable aleatoria binaria

genérica X es $H_2(X) = 0.8812$. La probabilidad de una secuencia $\mathbf{x}$ que contiene r unos y $n - r$ ceros es:

$$p_{\mathbf{X}}(\mathbf{x}) = p_1^r(1 - p_1)^{n-r}. \tag{2.44}$$

El número de secuencias binarias con r unos es:

$$N(r) = \binom{n}{r}. \tag{2.45}$$

Luego, sea $R \in \mathbb{N}$, una variable aleatoria que representa el número de unos en una secuencia binaria de n símbolos. La probabilidad de generación de una secuencia binaria de longitud n con r unos es:

$$p_R(r) = \binom{n}{r} p_1^r(1 - p_1)^{n-r}. \tag{2.46}$$

El valor esperado de R es $\mathbb{E}[R] = np_1$, la varianza es $\sigma^2 = np_1(1-p_1)$, y la desviación estándar es $\sigma = \sqrt{np_1(1 - p_1)}$. De manera que, el número de unos promedio se encuentra en un rango de: $r \approx np_1 \pm \sqrt{np_1(1 - p_1)}$.

En la Tabla 2.3 se representan nueve secuencias de $\mathbf{X}$ con r unos y $n - r$ ceros, dado $n = 100$, y se calcula la medida de información de una realización del vector, i.e., $\iota(\mathbf{x}) = -\log p_{\mathbf{X}}(\mathbf{x})$. Para $r = 0$ (dada la pmf de X) la medida de información es el valor mínimo de incertidumbre y para $r = 100$ es el valor máximo, por lo tanto, estos valores definen el rango de incertidumbre que producen las secuencias generadas por la pmf de $\mathbf{X}$. Para $r = 30$ se obtiene la entropía de la secuencia de variables aleatorias, i.e., $H(\mathbf{X}) = nH_2(X) = 88.12$ bits, el resto de resultados que se presentan tienen una medida de información cercana a la entropía de $\mathbf{X}$.

En la Tabla 2.4 se presentan las gráficas de las funciones $N(r)$, $p_{\mathbf{X}}(\mathbf{x})$, $-\log p_{\mathbf{X}}(\mathbf{x})$ y $p_R(r)$ para $n = 100$ y $n = 1000$, con el fin de observar el comportamiento de las diferentes funciones. De forma analítica se calcula el número de unos promedio [8].

Si n es 100, se obtienen los siguientes datos:

$$\mathbb{E}[R] = np_1 = 30 \tag{2.47}$$

$$\sigma = \sqrt{np_1(1 - p_1)} = 4.5. \tag{2.48}$$

Por lo tanto,

$$r \approx 30 \pm 4.5. \tag{2.49}$$

Si $n = 1000$, entonces,

$$\mathbb{E}[R] = 300 \tag{2.50}$$

$$\sigma = 14.5. \tag{2.51}$$

Por lo tanto,

$$r \approx 300 \pm 14.5. \tag{2.52}$$

Se observa que a medida que n crece, el rango de los valores más probables para r crece de acuerdo a (2.49) y (2.52), no obstante, la desviación estándar indica que a medida que n crece, la distribución de probabilidad de r se encuentra en mayor concentración debido a que σ crece en función de la $\sqrt{n}$. Lo anterior se observa gráficamente en la Tabla 2.4, en la primera y última fila. La gráfica de $-\log p_{\mathbf{X}}(\mathbf{x})$ presenta además el valor medio $H(\mathbf{X}) = nH_2(X)$, el cual es 88.12 y 881.2 para $n = 100$ y $n = 1000$, respectivamente. El Teorema 2.3 establece que existe un conjunto típico formado por las secuencias para las cuales $-\log p_{\mathbf{X}}(\mathbf{x})$ es muy cercano a $nH_2(X)$.

2.4.3. Conjunto Típico

De la AEP débil se deriva la división de todas las secuencias posibles en dos conjuntos, como ya se mencionó anteriormente. Aquí se presenta formalmente la regla que define que una secuencia pertenece al conjunto típico, asignando el resto al conjunto no típico, entre ellas, las secuencias más probables, y se desarrollan las propiedades de las secuencias del conjunto típico, fundamentales para establecer el límite en la codificación de fuente.

Definición 2.4. Sea $X \in \mathcal{X}$ una variable aleatoria genérica con pmf $p_X(x)$ y sea $p_{\mathbf{X}}(\mathbf{x})$ la pmf conjunta del vector de variables aleatorias $\mathbf{X}$ con dimension n definida en (2.28). Para cualquier $\epsilon > 0$ arbitrariamente pequeño, el conjunto de secuencias típicas con respecto a p_X es el conjunto de secuencias $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$

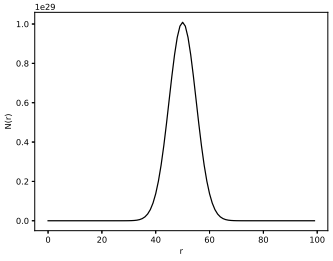
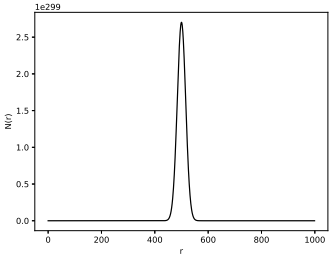
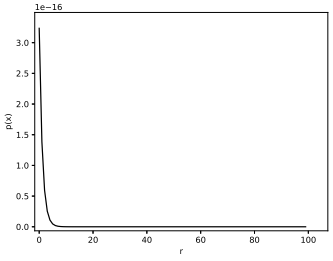
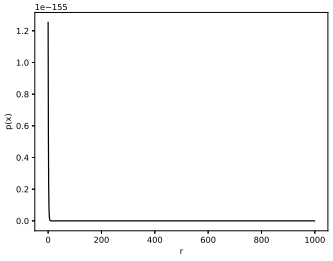
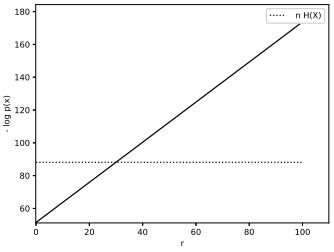
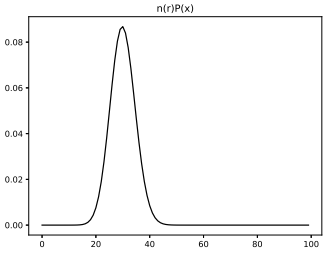
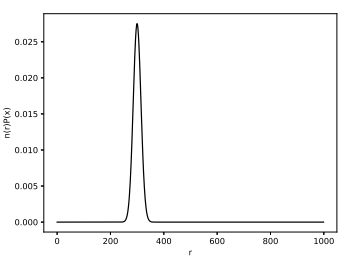
	N=100	N=1000
$N(r)$		
$p_{\mathbf{X}}(\mathbf{x})$		
$-\log p_{\mathbf{X}}(\mathbf{x})$		
$p_R(r)$		

Tabla 2.4: Representación de la Distribución Binomial.

denotado por $\mathcal{T}_X^{n,\epsilon}$, el cual es:

$$\mathcal{T}_X^{n,\epsilon} \triangleq \left\{ \mathbf{x} \in \mathcal{X}^n : \left| -\frac{1}{n} \log p_{\mathbf{X}}(\mathbf{x}) - H(X) \right| < \epsilon \right\}. \tag{2.53}$$

Ya que $\mathcal{T}_X^{n,\epsilon}$ depende de n, de ϵ y de la pmf $p_X(x)$ de la variable aleatoria X, se puede afirmar que: cualquiera sea el valor de ϵ, se obtiene un conjunto típico que contiene la mayor parte de la probabilidad a medida que n aumenta, lo cual se probó en el Teorema 2.3 (AEP débil). El conjunto típico por la Definición 2.4 y el Teorema 2.3 tiene las siguientes propiedades:

Lema 2.4 (Teorema 5.3 en [10]). *Sea $\mathcal{T}_X^{n,\epsilon}$ el conjunto de secuencias típicas con respecto a p_X y $\epsilon > 0$, por lo tanto, se satisface lo siguiente:*

1. *Si el vector de realizaciones $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{T}_X^{n,\epsilon}$, entonces:*

$$2^{-n(H(X)+\epsilon)} \leq p_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)}. \tag{2.54}$$

2. *Para n lo suficientemente grande,*

$$\Pr\{\mathcal{T}_X^{n,\epsilon}\} \triangleq \sum_{\mathbf{x} \in \mathcal{T}_X^{n,\epsilon}} p_{\mathbf{X}}(\mathbf{x}) \geq 1 - \epsilon. \tag{2.55}$$

3. *Para n lo suficientemente grande,*

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |\mathcal{T}_X^{n,\epsilon}| \leq 2^{n(H(X)+\epsilon)}. \tag{2.56}$$

Prueba:

- Prueba de (2.54): Es obtenida directamente de la Definición 2.4.
- Prueba de (2.55): Dado que (2.30) se cumple para n lo suficientemente grande y por la Definición 2.4, la probabilidad del conjunto típico es:

$$\Pr\{\mathcal{T}_X^{n,\epsilon}\} = \Pr\left\{ \left| -\frac{1}{n} \log p_{\mathbf{X}}(\mathbf{X}) - H(X) \right| < \epsilon \right\} \geq 1 - \epsilon. \tag{2.57}$$

- Prueba de (2.56): A partir de la probabilidad total del espacio muestral de secuencias, se aplica (2.55) y la cota inferior de (2.54), obteniendo lo siguiente:

$$1 = \sum_{\mathbf{x}\in\mathcal{X}^n} p_{\mathbf{X}}(\mathbf{x}) \tag{2.58}$$

$$\geq \sum_{\mathbf{x}\in\mathcal{T}_X^{n,\epsilon}} p_{\mathbf{X}}(\mathbf{x}) \tag{2.59}$$

$$\geq \sum_{\mathbf{x}\in\mathcal{T}_X^{n,\epsilon}} 2^{-n(H(X)+\epsilon)} \tag{2.60}$$

$$= |\mathcal{T}_X^{n,\epsilon}| 2^{-n(H(X)+\epsilon)}. \tag{2.61}$$

Para n suficientemente grande (2.61), implica:

$$|\mathcal{T}_X^{n,\epsilon}| \leq 2^{n(H(X)+\epsilon)}. \tag{2.62}$$

De forma similar, haciendo uso de (2.55) se obtiene lo siguiente:

$$1 - \epsilon \leq \sum_{\mathbf{x}\in\mathcal{T}_X^{n,\epsilon}} p_{\mathbf{X}}(\mathbf{x}) \tag{2.63}$$

$$\leq \sum_{\mathbf{x}\in\mathcal{T}_X^{n,\epsilon}} 2^{-n(H-\epsilon)} \tag{2.64}$$

$$= |\mathcal{T}_X^{n,\epsilon}| 2^{n(H-\epsilon)}. \tag{2.65}$$

Para n suficientemente grande (2.65) implica:

$$|\mathcal{T}_X^{n,\epsilon}| \geq (1-\epsilon) 2^{n(H-\epsilon)}, \tag{2.66}$$

lo cual completa la prueba de (2.56) y la prueba del Lema 2.4. ■

En el Lema 2.4 se desarrollan las propiedades del conjunto típico obtenidas de la AEP débil. La primera propiedad en (2.54) define un intervalo de probabilidad de ocurrencia típico, en el cual los elementos del conjunto típico son todos aquellos con una probabilidad de ocurrencia muy cercana a $2^{-nH(X)}$. La segunda propiedad en (2.55) establece que la probabilidad del conjunto típico es muy cercana a uno, i.e., las secuencias típicas hacen parte de un conjunto con una alta probabilidad de

ocurrencia a medida que n crece, o con n suficientemente grande. La tercera propiedad en (2.56) es muy importante porque permite contar (de forma aproximada) el número de elementos que conforman el conjunto típico, el cual es aproximadamente $2^{nH(X)}$. La relación de la cardinalidad del conjunto típico con respecto al conjunto total de posibles secuencias indica que, en general el número de secuencias típicas es muy pequeño, como se observa en (2.67), en el cual la relación tiende a cero [10]:

$$\frac{|\mathcal{T}_X^{n,\epsilon}|}{|\mathcal{X}|^n} \simeq \frac{2^{n(H(X)+\epsilon)}}{2^{n\log|\mathcal{X}|}} = 2^{-n(\log|\mathcal{X}|-H(X)-\epsilon)} \to 0, \tag{2.67}$$

cuando n tiende a infinito y siempre que $H(X) < \log|\mathcal{X}|$. Sin embargo, aunque es insignificante la cardinalidad del conjunto típico con respecto al conjunto total de posibles secuencias, la probabilidad del conjunto típico hace que las secuencias típicas sean altamente probables, es decir, que el conjunto de secuencias típicas determina el comportamiento promedio de selección de secuencias para n suficientemente grande [10].

Las afirmaciones anteriores acerca del conjunto típico se exponen en el siguiente ejemplo. En un experimento aleatorio se utilizan tres canicas: dos de color blanco y una de color negro, las cuales se ubican dentro de una caja. El experimento consiste en sacar una canica de la caja, anotar su color y devolverla. Al repetir 5 veces el experimento (sin realizar ninguna modificación) son posibles $2^5 = 32$ diferentes combinaciones de canicas, las cuales se presentan en la Fig. 2.12 [56]. Las probabilidades de ocurrencia de las diferentes secuencias estan definidas por la distribución binomial y se presentan en la última columna de la Fig. 2.12.

Sea X la variable aleatoria que representa la selección de una canica del alfabeto $\mathcal{X} = \{blanca, negra\}$, y su pmf p_X es para $p_X(x = blanca) = p_b = 2/3$ y $p_X(x = negra) = p_n = 1/3$. La entropía de X se calcula como la entropía binaria $H_2(X)$ definida en (2.5), la cual es:

$$\begin{aligned} H_2(X) &= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \\ &= 0.918. \end{aligned} \tag{2.68}$$

Al repetir el experimento n veces se genera una secuencia de resultados de longitud

	Secuencia	Probabilidad					Típica
1	● ● ● ● ●	1/3	1/3	1/3	1/3	1/3	0.0041
2	● ● ● ● ○	1/3	1/3	1/3	1/3	2/3	0.0082
3	● ● ● ○ ●	1/3	1/3	1/3	2/3	1/3	0.0082
4	● ● ● ○ ○	1/3	1/3	1/3	2/3	2/3	0.0165
5	● ● ○ ● ●	1/3	1/3	2/3	1/3	1/3	0.0082
6	● ● ○ ● ○	1/3	1/3	2/3	1/3	2/3	0.0165
7	● ● ○ ○ ●	1/3	1/3	2/3	2/3	1/3	0.0165
8	● ● ○ ○ ○	1/3	1/3	2/3	2/3	2/3	0.0329 *
9	● ○ ● ● ●	1/3	2/3	1/3	1/3	1/3	0.0082
10	● ○ ● ● ○	1/3	2/3	1/3	1/3	2/3	0.0165
11	● ○ ● ○ ●	1/3	2/3	1/3	2/3	1/3	0.0165
12	● ○ ● ○ ○	1/3	2/3	1/3	2/3	2/3	0.0329 *
13	● ○ ○ ● ●	1/3	2/3	2/3	1/3	1/3	0.0165
14	● ○ ○ ● ○	1/3	2/3	2/3	1/3	2/3	0.0329 *
15	● ○ ○ ○ ●	1/3	2/3	2/3	2/3	1/3	0.0329 *
16	● ○ ○ ○ ○	1/3	2/3	2/3	2/3	2/3	0.0658 *
17	○ ● ● ● ●	2/3	1/3	1/3	1/3	1/3	0.0082
18	○ ● ● ● ○	2/3	1/3	1/3	1/3	2/3	0.0165
19	○ ● ● ○ ●	2/3	1/3	1/3	2/3	1/3	0.0165
20	○ ● ● ○ ○	2/3	1/3	1/3	2/3	2/3	0.0329 *
21	○ ● ○ ● ●	2/3	1/3	2/3	1/3	1/3	0.0165
22	○ ● ○ ● ○	2/3	1/3	2/3	1/3	2/3	0.0329 *
23	○ ● ○ ○ ●	2/3	1/3	2/3	2/3	1/3	0.0329 *
24	○ ● ○ ○ ○	2/3	1/3	2/3	2/3	2/3	0.0658 *
25	○ ○ ● ● ●	2/3	2/3	1/3	1/3	1/3	0.0165
26	○ ○ ● ● ○	2/3	2/3	1/3	1/3	2/3	0.0329 *
27	○ ○ ● ○ ●	2/3	2/3	1/3	2/3	1/3	0.0329 *
28	○ ○ ● ○ ○	2/3	2/3	1/3	2/3	2/3	0.0658 *
29	○ ○ ○ ● ●	2/3	2/3	2/3	1/3	1/3	0.0329 *
30	○ ○ ○ ● ○	2/3	2/3	2/3	1/3	2/3	0.0658 *
31	○ ○ ○ ○ ●	2/3	2/3	2/3	2/3	1/3	0.0658 *
32	○ ○ ○ ○ ○	2/3	2/3	2/3	2/3	2/3	0.1317

Figura 2.12: Experimento de la caja.

n, la cual ocurre con cierta probabilidad. Sea $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ el vector de variables aleatorias que representa la selección de una secuencia del conjunto total de posibles secuencias, y sea $p_{\mathbf{X}}$ la pmf conjunta. Definida la entropía de la variable aleatoria genérica X, se busca encontrar el conjunto típico de acuerdo a la Definición 2.4, con $\epsilon = 0.138$ y $n = 5$ utilizando (2.54), obteniendo:

$$0.026 \leq p_{\mathbf{X}}(\mathbf{x}) \leq 0.067. \tag{2.69}$$

Este rango define las secuencias que pertenecen al conjunto típico, las cuales son señaladas con asterisco en la Fig. 2.12. Al sumar las probabilidades marcadas se obtiene una probabilidad total del conjunto típico de 0.658, lo cual significa que aproximadamente 2/3 de la probabilidad total corresponde al conjunto típico, y 1/3 de la probabilidad corresponde a las secuencias del conjunto no típico.

En el ejemplo anterior, la relación entre la cardinalidad del conjunto típico (15 elementos) con respecto a la cardinalidad del conjunto total de secuencias (32 elementos) es 0.47, i.e., el tamaño del conjunto típico corresponde aproximadamente a la mitad del conjunto total. ¿Qué sucede con el incremento de n? Si se considera un número de repeticiones mayor, por ejemplo, $n = 500$ y con $\epsilon = 0.046$, el número de secuencias del conjunto típico es aproximadamente $2^{nH(X)} = 2^{459}$. La relación de la cardinalidad del conjunto de secuencias típicas con respecto a la cardinalidad del conjunto total de secuencias, es:

$$\frac{|\mathcal{T}_X^{n,\epsilon}|}{|\mathcal{X}|^n} \simeq \frac{2^{459}}{2^{500}} = 2^{-41} \approx 10^{-13}. \tag{2.70}$$

En la Tabla 2.5 se presenta la relación (2.67) para diferentes valores de n, en la cual se logra observar que el tamaño del conjunto típico con respecto al conjunto total de secuencias es insignificante a medida que n crece. Para concluir hay que resaltar que, aunque la probabilidad total del conjunto típico es muy cercana a uno, las secuencias más probables no pertenecen al conjunto típico. Las secuencias típicas definidas por la AEP débil son aquellas para las cuales la entropía experimental es muy cercana a la entropía teórica (ver Sección 2.3.1), por lo tanto, de las secuencias más probables se puede mostrar que en general son secuencias no típicas, debido a que su entropía experimental esta muy lejos de la entropía teórica. En el ejemplo,

n	$\lvert\mathcal{T}_X^{n,\epsilon}\rvert$	$\lvert\mathcal{T}_X^{n,\epsilon}\rvert/\lvert\mathcal{X}\rvert^n$
100	$2^{91.8}$	10^{-3}
500	2^{459}	10^{-13}
1000	2^{918}	10^{-25}

Tabla 2.5: Relación de la cardinalidad del conjunto típico con respecto al conjunto total para n.

la entropía experimental de la secuencia más probables es:

$$\begin{aligned} -\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{x}) =& -\frac{1}{5}\log p_X(x_1)p_X(x_2)\dots p_X(x_5) \\ =& -\frac{1}{5}\log 0.1317 \\ =& 0.5849, \end{aligned} \tag{2.71}$$

la cual es diferente del resultado mostrado en (2.68), por lo tanto la secuencia más probable no es típica.

De la Definición 2.4 y de (2.67) se puede afirmar que el conjunto típico es un subconjunto de $\mathcal{X}^n$ con un muy bajo número de elementos con respecto al conjunto total de posibles secuencias, al cual en conjunto le corresponde una alta probabilidad. Por lo tanto, surge la siguiente pregunta ¿es el subconjunto más pequeño de alta probabilidad? La respuesta implica demostrar que no existe otro subconjunto más pequeño de alta probabilidad o que un subconjunto de alta probabilidad tiene el mismo número de elementos que el conjunto típico.

Sea X una variable aleatoria con pmf p_X y alfabeto $\mathcal{X}$. Sea $\mathcal{A}_\delta$ el subconjunto de $\mathcal{X}$ al cual le corresponde una probabilidad similar que la del conjunto típico, establecido en la Definición 2.5.

Definición 2.5 (*The smallest δ-sufficient subset* en [8])**.** El subconjunto δ-suficiente más pequeño de $\mathcal{X}$ denotado por $\mathcal{A}_\delta$, con $0 < \delta < 1$, satisface:

$$\Pr\{x \in \mathcal{A}_\delta\} \geq 1 - \delta. \tag{2.72}$$

De la Definición 2.5, si x no pertenece a $\mathcal{A}_\delta$, entonces:

$$\Pr\{x \notin \mathcal{A}_\delta\} < \delta. \tag{2.73}$$

El valor de δ define el subconjunto $\mathcal{A}_\delta$, que se obtiene de organizar los elementos de $\mathcal{X}$ en orden decreciente de probabilidad, y se suma la probabilidad de los elementos hasta que esta sea mayor o igual a $1-\delta$. Los elementos del subconjunto estan descritos por la Definición 2.6.

Definición 2.6 (Definición en [8])**.** El número total de bits que permiten representar o identificar los elementos de $\mathcal{A}_\delta$ es:

$$H_\delta(X) = \log |\mathcal{A}_\delta|. \tag{2.74}$$

Para $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ un vector de variables aleatorias i.i.d. con pmf $p_{\mathbf{X}}$ y alfabeto $\mathcal{X}^n$, la Definición 2.5 de conjuntos típicos de alta probabilidad se reescribe como:

Definición 2.7 (Definición conjunto más pequeño en [9])**.** Para cada $n \in \mathbb{Z}^+$ sea $\mathcal{A}_\delta^{(n)}$ el subconjunto de $\mathcal{X}^n$ más pequeño, para el cual se cumple:

$$\Pr\{\mathbf{x} \in \mathcal{A}_\delta^{(n)}\} \geq 1 - \delta. \tag{2.75}$$

Si tal conjunto existe distinto del conjunto típico, debe tener una intersección significativa con el conjunto típico, y por lo tanto debe tener aproximadamente el mismo número de elementos [9], lo cual se establece en el Teorema 2.5.

Teorema 2.5 (Teorema 3.3.1 en [9])**.** *Sea* $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ *un vector de variables aleatorias i.i.d. con pmf* $p_{\mathbf{X}}$*. Para* $\delta < \frac{1}{2}$ *y* $\delta' > 0$*, si* $\Pr\{\mathcal{A}_\delta^{(n)}\} > 1 - \delta$ *entonces,*

$$\frac{1}{n} \log |\mathcal{A}_\delta^{(n)}| > H(X) - \delta', \tag{2.76}$$

para un valor de n suficientemente grande.

Prueba: Sean $\mathcal{T}_X^{n,\epsilon}$ el conjunto típico dado por la Definición 2.4 y $\mathcal{A}_\delta^{(n)}$ dado por la Definición 2.5, tal que

$$\Pr\{\mathbf{x} \in \mathcal{T}_X^{n,\epsilon}\} \geq 1 - \epsilon \tag{2.77}$$

$$\Pr\{\mathbf{x} \in \mathcal{A}_\delta^{(n)}\} \geq 1 - \delta. \tag{2.78}$$

La probabilidad del subconjunto $\mathcal{A}_\delta^{(n)}$ se puede escribir como:

$$\Pr\{\mathbf{x} \in \mathcal{A}_\delta^{(n)}\} = \Pr\{\mathbf{x} \in \mathcal{A}_\delta^{(n)} \cap \mathcal{T}_X^{n,\epsilon}\} + \Pr\{\mathbf{x} \in \mathcal{A}_\delta \cap \overline{\mathcal{T}_X^{n,\epsilon}}\} \tag{2.79}$$

donde $\overline{\mathcal{T}_X^{n,\epsilon}} = \{\mathbf{x} \notin \mathcal{T}_X^{n,\epsilon}\}$ denota el complemento de $\mathcal{T}_X^{n,\epsilon}$. De (2.79) se obtiene:

$$\Pr\{\mathbf{x} \in \mathcal{A}_\delta^{(n)} \cap \mathcal{T}_X^{n,\epsilon}\} = \Pr\{\mathbf{x} \in \mathcal{A}_\delta^{(n)}\} - \Pr\{\mathbf{x} \in \mathcal{A}_\delta \cap \overline{\mathcal{T}_X^{n,\epsilon}}\} \tag{2.80}$$

$$\geq 1 - \epsilon - \delta, \tag{2.81}$$

donde, el primer término a la derecha del igual en (2.80) se obtiene de (2.78); y el segundo término a la derecha del igual en (2.80) se obtiene considerando que, la intersección de dos conjuntos tiene como máximo valor de probabilidad la máxima probabilidad de uno de los conjuntos, i.e., $\Pr\{\mathbf{x} \in \overline{\mathcal{T}_X^{n,\epsilon}}\} < \epsilon$ por lo tanto $\Pr\{\mathbf{x} \in \mathcal{A}_\delta \cap \overline{\mathcal{T}_X^{n,\epsilon}}\} < \epsilon$.

Luego,

$$1 - \epsilon - \delta \leq \Pr\{\mathbf{x} \in \mathcal{A}_\delta^{(n)} \cap \mathcal{T}_X^{n,\epsilon}\} \tag{2.82}$$

$$= \sum_{\mathbf{x} \in \mathcal{A}_\delta^{(n)} \cap \mathcal{T}_X^{n,\epsilon}} p_{\mathbf{X}}(\mathbf{x}) \tag{2.83}$$

$$\overset{(a)}{\leq} \sum_{\mathbf{x} \in \mathcal{A}_\delta^{(n)} \cap \mathcal{T}_X^{n,\epsilon}} 2^{-n(H(X)-\epsilon)} \tag{2.84}$$

$$= |\mathcal{A}_\delta^{(n)} \cap \mathcal{T}_X^{n,\epsilon}| 2^{-n(H(X)-\epsilon)} \tag{2.85}$$

$$\overset{(b)}{\leq} |\mathcal{A}_\delta^{(n)}| 2^{-n(H(X)-\epsilon)}, \tag{2.86}$$

donde,

(a) De la propiedad (2.54), i.e., $p_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)}$ y considerando nuevamente

el argumento de que la intersección de dos conjuntos tiene como máximo valor de probabilidad, la máxima probabilidad de uno de los conjuntos.

(b) El número de elementos de la intersección de conjuntos es como máximo el número de elementos de uno de los conjuntos.

Por lo tanto, de (2.86) se obtiene lo siguiente:

$$|\mathcal{A}_\delta^{(n)}| \geq (1-\epsilon-\delta)2^{n(H(X)-\epsilon)}. \tag{2.87}$$

Aplicando la función logaritmo a (2.87), y considerando que la función logaritmo es una función creciente, lo siguiente es valido:

$$\log|\mathcal{A}_\delta^{(n)}| \geq \log(1-\epsilon-\delta) + nH(X) - n\epsilon. \tag{2.88}$$

Haciendo que $\delta' = -\frac{1}{n}\log(1-\epsilon-\delta)+\epsilon$, se obtiene lo siguiente:

$$\frac{1}{n}\log|\mathcal{A}_\delta^{(n)}| \geq H(X) - \delta', \tag{2.89}$$

y esto completa la prueba. ■

Así del Teorema 2.5 se concluye que el subconjunto $\mathcal{A}_\delta^{(n)}$ y $\mathcal{T}_X^{n,\epsilon}$ tienen aproximadamente el mismo tamaño, i.e., que el conjunto típico es el conjunto más pequeño de alta probabilidad.

La AEP débil y el conjunto de secuencias típicas con sus propiedades permiten plantear en el sistema de comunicación la compresión de datos, de acuerdo a la estructura probabilística de las secuencias generadas por la fuente de información.

2.5. CODIFICACIÓN DE FUENTE

En el modelo del sistema de comunicación discreto punto a punto que se presenta en la Fig. 2.1, se mostró que los mensajes producidos por la fuente de información discreta tienen asociada una estructura estadística, tal que, permite definir una

métrica de la incertidumbre o de la información producida por la fuente, denominada entropía. La fuente de información discreta definida en la Sección 2.3.1, se puede ver como un generador de secuencias de variables aleatorias i.i.d. $\{X_m, m \in \mathbb{Z}^+ \text{ y } m \geq 1\}$ indexadas por el índice m, es un conjunto ordenado, por lo tanto, la fuente de información se considera sin memoria y discreta en el tiempo.

Al realizar el experimento de lanzar una moneda n veces, el resultado del lanzamiento i-ésimo es la variable aleatoria X_i con alfabeto $\mathcal{X} = \{cara, sello\}$, el cual es independiente de previos y posteriores resultados. Ahora, se quiere describir el resultado del experimento sin errores usando símbolos binarios, por lo tanto, se codifican los resultados: *cara* como un '0' y *sello* como un '1'. Luego el resultado del experimento es codificado en una secuencia de n símbolos binarios, llamada palabra código (*codeword*). Si la moneda es equilibrada, los resultados son equiprobables, es decir, cada palabra código tiene una probabilidad igual a 2^{-n}. En el caso general, donde los resultados no son equiprobables, la codificación puede ser diseñada de tal manera que, los resultados más probables sean codificados en palabras código de menor longitud de símbolos y los resultados menos probables en palabras código de mayor longitud. La entropía $H(X)$ mide la cantidad de incertidumbre de la variable aleatoria X, por lo tanto, es una métrica fundamental en la codificación de los datos producidos por la fuente de información, como se estudia a continuación [10].

La fuente de información tiene asociada otra propiedad y es la redundancia, la cual definida informalmente es el uso repetitivo de símbolos en los mensajes, por ejemplo, de sílabas o palabras enteras, en un libro, o un discurso, las cuales permiten que se pueda recuperar un mensaje incluso cuando ocurren pérdidas de palabras, mediante su predicción a partir del contexto del mensaje.

El considerar la fuente de información discreta es fundamental en la teoría de la información, ya que su carácter estadístico permite encontrar la codificación adecuada que remueve la redundancia del mensaje, haciendo más eficiente el proceso de transmisión, i.e., el proceso de compresión de datos de la fuente. Shannon desarrolló los conceptos básicos que le permitieron establecer el teorema de codificación de fuente para canales sin ruido (Teorema 9 en [1]).

Teorema 2.6 (Teorema 9 en [1]). *Sea una fuente de información con entropía H bits por símbolo y capacidad C bits por segundo asociada al canal. Es posible codificar la salida de la fuente de tal forma que la velocidad de transmisión promedio sea $\frac{C}{H} - \epsilon$ símbolos por segundo, sobre el canal, donde ϵ es arbitrariamente pequeño. No es posible transmitir a una velocidad de transmisión mayor que $\frac{C}{H}$.*

Este teorema determina la velocidad promedio de transmisión de información de un código fuente, el cual se puede aproximar a $\frac{C}{H}$ pero no puede ser mayor, y la entropía no puede exceder la capacidad del canal al considerar un sistema de comunicación sin ruido. La demostración proporcionada por Shannon de la parte inversa del teorema es obtenida haciendo uso del argumento de que la entropía en la entrada del canal es igual a la entropía de la fuente, y la parte directa del teorema fue demostrada de dos formas: la primera derivada de la AEP débil, y la segunda presentando un método de codificación [1].

El Teorema 2.6 es el planteamiento original de Shannon, sin embargo, la forma actual en la que se plantea el teorema es distinta. Aunque el concepto de entropía de fuente como límite fundamental es el mismo, hay dos versiones del teorema: la primera en la que se hace referencia a las cotas inferior y superior de la longitud promedio de las palabras código; y la segunda en la que se formaliza el teorema desde la velocidad de transmisión de información promedio. Así, la teoría de la información diseña los códigos de compresión de datos y define los límites fundamentales que permiten una transmisión de información segura y confiable.

Si la fuente de información produce n mensajes con igual probabilidad, el código fuente más sencillo asigna palabras código de longitud constante $H_0(X)$ (si el código es creado con símbolos binarios), el cual corresponde a la cantidad de bits requeridos para representar cada resultado. En el ejemplo presentado en la Sección 2.2.2, las secuencias de sabores les corresponde un secuencia de símbolos binarios de longitud 6 bits. Si en cambio, las probabilidades de los mensajes son distintas, se asignan palabras código de longitud variable l, de modo que, el código fuente es la transformación de un elemento del conjunto de mensajes a un elemento asociado en el conjunto de palabras código, conocido como libro código (*codebook*), i.e., un mapeo del conjunto total de mensajes al conjunto total de palabras código, lo cual se

formaliza en la Definición 2.8.

Definición 2.8 (Definición 4.1 en [10])**.** Sea $\mathcal{X}$ el alfabeto de la variable aleatoria X y $\mathcal{D}^*$ el conjunto de todas las secuencias de longitud finita formadas por los símbolos del alfabeto código D-ario. El código fuente $\mathcal{C}$ asociado a la variable aleatoria X, es $\mathcal{C} : \mathcal{X} \to \mathcal{D}^*$, donde, $\mathcal{C}(x)$ denota la palabra código que corresponde a x y $l(x)$ su longitud.

Definición 2.9 (Definición longitud promedio en [9])**.** La longitud promedio L de un código fuente $\mathcal{C}$ para una variable aleatoria X con pmf p_X es:

$$L = \sum_{x \in \mathcal{S}_X}^{|\mathcal{D}^*|} p_X(x) l(x). \tag{2.90}$$

En general, el sistema de comunicación busca minimizar la longitud promedio L del código, de esta manera, las longitudes más cortas se asignarán a los mensajes más probables y las longitudes más largas se asignarán a los mensajes menos probables. Este argumento ha sido previamente utilizado en los sistemas de comunicación modernos, por ejemplo, en el código Morse la letra e del inglés se codifica con un punto. La longitud del código no puede reducirse arbitrariamente, ya que en la transmisión del mensaje codificado, el cual esta formado por palabras código concatenadas, se debe garantizar en recepción una decodificación no ambigua del mensaje [9].

En la codificación de fuente existen diferentes códigos, los cuales se pueden encontrar en [8,10,57], sin embargo, dado cualquier código, su longitud promedio está determinada por el límite de compresión de información, la cual se obtiene por la AEP débil. Este límite es formalizado en el Teorema 9 en [1] y es conocido como el teorema de codificación de fuente, el cual plantea la longitud promedio de las palabras código en la compresión de datos. En sí misma, la AEP débil establece un código fuente para una secuencia de variables aleatorias, de la cual se obtiene el límite fundamental.

Sea $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ un vector de variables aleatorias discretas i.i.d., con pmf conjunta $p_{\mathbf{X}}$, y las realizaciones del vector se denotan por $\mathbf{x} \in \mathcal{X}^n$. Por la AEP débil, el conjunto de secuencias $\mathcal{X}^n$ se divide en dos conjuntos: el conjunto típico $\mathcal{T}_X^{n,\epsilon}$ y el conjunto no típico. Para diseñar un esquema de codificación y definir el conjunto de

palabras código, se ordenan las secuencias[9]. El orden establecido asigna un índice a cada secuencia, de modo que, cada elemento del conjunto típico es descrito por el índice que identifica la secuencia en el conjunto. Los índices que ordenan los conjuntos típico y no típico, forman un código, por lo tanto, cada índice hace parte de la palabra código. Para diferenciar las secuencias se asigna un prefijo de un bit a la palabra código binaria: un cero para las secuencias del conjunto típico, y un uno para las secuencias del conjunto no típico. Finalmente, se determina el número de bits necesarios para representar cada palabra código, i.e., su longitud [9].

- Las palabras código que corresponden a los elementos del conjunto típico tendrán una longitud $l_1 \leq n(H(X)+\epsilon)+2$, debido a que $|\mathcal{T}_X^{n,\epsilon}| \leq 2^{n(H(X)+\epsilon)}$ por (2.56), más un bit extra ya que $n(H(X)+\epsilon)$ puede no ser una cantidad entera, y el segundo bit extra se utiliza para diferenciar los conjuntos.
- De forma similar, las palabras del código correspondientes a los elementos del conjunto no típico tendrán una longitud $l_2 \leq n\log|\mathcal{X}|+2$.

De lo anterior, se obtiene un esquema de codificación para todas las secuencias $\mathbf{x} \in \mathcal{X}^n$. Sea $l(\mathbf{x})$ la longitud de la palabra código correspondiente a $\mathbf{x}$. Si n es suficientemente grande para que $\Pr\{\mathcal{T}_X^{n,\epsilon}\} \geq 1-\epsilon$, el valor esperado de la longitud del código está acotado por la siguiente expresión:

$$\begin{aligned}
\mathbb{E}[l(\mathbf{X})] &= \sum_{\mathbf{x}\in\mathcal{X}^n} p_{\mathbf{X}}(\mathbf{x})l(\mathbf{x}) && (2.91)\\
&\overset{(a)}{=} \sum_{\mathbf{x}\in\mathcal{T}_X^{n,\epsilon}} l_1 p_{\mathbf{X}}(\mathbf{x}) + \sum_{\mathbf{x}\notin\mathcal{T}_X^{n,\epsilon}} l_2 p_{\mathbf{X}}(\mathbf{x}) && (2.92)\\
&\overset{(b)}{\leq} \sum_{\mathbf{x}\in\mathcal{T}_X^{n,\epsilon}} (n(H(X)+\epsilon)+2)p_{\mathbf{X}}(\mathbf{x}) + \sum_{\mathbf{x}\notin\mathcal{T}_X^{n,\epsilon}} (n\log|\mathcal{X}|+2)p_{\mathbf{X}}(\mathbf{x}) && (2.93)\\
&= \Pr\{\mathcal{T}_X^{n,\epsilon}\}(n(H(X)+\epsilon)+2) + (1-\Pr\{\mathcal{T}_X^{n,\epsilon}\})(n(\log|\mathcal{X}|)+2) && (2.94)\\
&\leq n(H(X)+\epsilon) + \epsilon n\log|\mathcal{X}| + 2 && (2.95)\\
&\overset{(c)}{=} n(H(X)+\epsilon'), && (2.96)
\end{aligned}$$

donde,

[9]El cual puede ser: lexicográfico o numérico.

(a) La sumatoria sobre el conjunto $\mathcal{X}^n$ se parte en dos sumas: la del conjunto típico y la del conjunto no típico.

(b) Se aplica las cotas superiores para $l_1(\mathbf{x})$ y $l_2(\mathbf{x})$ como fue descrita previamente.

(c) Y $\epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2}{n}$ es arbitrariamente pequeña con una elección apropiada de ϵ y n.

Este resultado junto con las consideraciones de esta sección prueban el teorema de codificación de fuente, expresado en una de sus formas en el Teorema 2.7, el cual establece una cota superior de la longitud promedio de las palabras código.

Teorema 2.7 (Teorema 3.2.1 en [9]). *Sea* $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ *un vector de variables aleatorias i.i.d. que representa la salida de una fuente de información con entropía* $H(X)$*, y sea* $\epsilon > 0$*, existe un código que permite mapear secuencias* $\mathbf{x}$ *de longitud* n *en secuencias binarias tales que el mapeo es uno a uno y la longitud promedio tiene una cota superior dada por:*

$$\mathbb{E}\left[\frac{1}{n} l(\mathbf{X})\right] \leq H(X) + \epsilon, \tag{2.97}$$

para n *suficientemente grande.*

En el Teorema 2.7 se afirma la existencia de un código fuente conocido como código instantáneo el cual implica una decodificación única. El valor máximo que puede tener la longitud promedio es $nH(X)$ bits.

La expresión completa de la longitud promedio L es [10]:

$$H(X) \leq \frac{L}{n} \leq H(X) + \frac{1}{n}. \tag{2.98}$$

En [9, 10, 57] se encuentra la demostración completa de las cotas de la longitud promedio.

La compresión de datos es de dos tipos: con pérdidas (*lossy compression*) y sin pérdidas (*lossless compression*). En el primer tipo de compresión se toleran pérdidas de

$\delta = 0$		$\delta = 1/16$	
x	$\mathcal{C}(x)$	x	$\mathcal{C}(x)$
a	000	a	00
b	001	b	01
c	010	c	10
d	011	d	11
e	100	e	-
f	101	f	-
g	110	g	-
h	111	h	-

Tabla 2.6: Código fuente binario.

información, las cuales pueden surgir por códigos ambiguos que generan errores en la decodificación. El párametro δ denota la probabilidad de fallo, si esta probabilidad se hace lo suficientemente pequeña la compresión con pérdidas puede resultar útil. Un compresor sin pérdidas debe tener asociado un código que aún removiendo la redundancia de la información, permita recuperar la información original de la secuencia codificada [8].

Un esquema simple que describe la codificación de un alfabeto fuente en un código de compresión de datos de los dos tipos, se desarrolla en el siguiente ejemplo:

Sea $\mathcal{X} = \{a, b, c, d, e, f, g, h\}$ con pmf $p_X(x) = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right\}$. La codificación de $\mathcal{X}$ se obtiene al asignar una palabra código de $\mathcal{C}$ con alfabeto $D = \{0, 1\}$ a cada elemento. El número total de bits que representan sin pérdidas los elementos de $\mathcal{X}$ por (2.1) es 3 bits, i.e., un código fuente formado por 8 palabras código, el cual se presenta en la segunda columna de la Tabla 2.6.

El parámetro $\delta = 0$ indica que el código de x denotado por $\mathcal{C}(x)$ es asignado a todos los elementos que son posibles resultados de la variable aleatoria X. Se puede observar que la probabilidad del subconjunto $\mathcal{B} = \{e, f, g, h\}$, i.e., $\Pr\{\mathcal{B}\} = \frac{1}{16}$, corresponde al 6.25 % de la probabilidad total, así, si se decide correr el riesgo de ignorar estos resultados, es decir que sea $\delta = \frac{1}{16}$, se obtiene un subconjunto $\mathcal{A}_\delta = \{a, b, c, d\}$ de $\mathcal{X}$ tal que la cantidad de bits necesaria para representar los elementos del nuevo subconjunto es 2 bits, lo cual significa que solo son necesarios dos dígitos binarios para asignar un código a los elementos del subconjunto $\mathcal{A}_\delta$, como se observa en la

cuarta columna de la Tabla 2.6.

En la parte izquierda de la Tabla 2.6 se ha asignado un código de longitud constante para cada elemento de $\mathcal{X}$, la compresión de datos de la fuente de información es sin pérdidas. Si el alfabeto fuente es redefinido, de tal manera que asume una probabilidad de fallo, el código para el nuevo subconjunto $\mathcal{A}_\delta$ de $\mathcal{X}$ es una compresión de datos que tolera pérdidas, i.e. que si ocurre un resultado ignorado, por ejemplo g, no es posible decodificar ese elemento en la recepción.

Si se considera una secuencia de n lanzamientos de una moneda donde se codifican los resultados: *cara* como un '0' y *sello* como un '1', los resultados serían secuencias de $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, donde $x_n \in \mathcal{X} = \{0, 1\}$, con probabilidades $p_X(0) = 0.9$ y $p_X(1) = 0.1$. Las secuencias $\mathbf{x}$ *más probables* tienen en su mayoría $0s$. La probabilidad de que la secuencia $\mathbf{x}$ que contiene r unos y $n - r$ ceros es dada por (2.44) y la entropía de la variable aleatoria genérica X es $H(X) = 0.4689$ bits. Se va a analizar el subconjunto $\mathcal{A}_\delta$ dado por la Definición 2.5 para $n = 4$ y $n = 10$. En el caso de $n = 4$, se tiene $|\mathcal{X}|^4 = 16$ y $H_0(\mathbf{X}) = 4$ bits. En la Tabla 2.7 se presentan las probabilidades para secuencias que contienen r unos, así, si $r = 2$, el número de secuencias con dos unos es 6 cada una con probabilidad 0.0081. La probabilidad de fallo δ, corresponde a la suma de probabilidades de las secuencias o resultados que se van a ignorar, así de acuerdo a la Definición 2.7, si $\delta = 0$ cada secuencia se mapea a una palabra código, el total de secuencias sería 16 y la longitud constante de las palabras código sería 4 bits; si $\delta = 0.0001$ implica que la secuencia con $r = 4$ se ignoraría, y el subconjunto $\mathcal{A}_{0.0001}$ tendría una cardinalidad de 15 elementos y $H_\delta(\mathbf{X}) = \log 15 = 3.9$ bits; si $\delta = 0.001$ implica que la secuencia con $r = 4$ y las secuencias con $r = 3$ se ignorarían, y el subconjunto $\mathcal{A}_{0.001}$ tendría una cardinalidad de 11 elementos y $H_\delta(\mathbf{X}) = \log 11 = 3.45$ bits. De esta manera, $H_\delta(\mathbf{X})$ disminuye a medida que el número de elementos del conjunto disminuye y la probabilidad de fallo aumenta.

En la Fig. 2.13 se presentan los subconjuntos de acuerdo al valor que toma δ, donde cada posible secuencia es organizada de menor a mayor probabilidad y en la Fig. 2.14 se presentan los distintos valores o pasos que toma $H_\delta(\mathbf{X})$ en función del riesgo o probabilidad de fallo δ. El comportamiento que se observa para $n = 4$ y $n = 10$, es

r	nCr	$p_{\mathbf{X}}(\mathbf{x})$	δ	$\log p_{\mathbf{X}}(\mathbf{x})$
0	1	0.6561	0.34	-0.6
1	4	0.0729	0.12	-3.7
2	6	0.0081	0.01	-6.9
3	4	0.0009	0.001	-10.1
4	1	0.0001	0	-13.2

Tabla 2.7: Probabilidad de las secuencias con r unos.

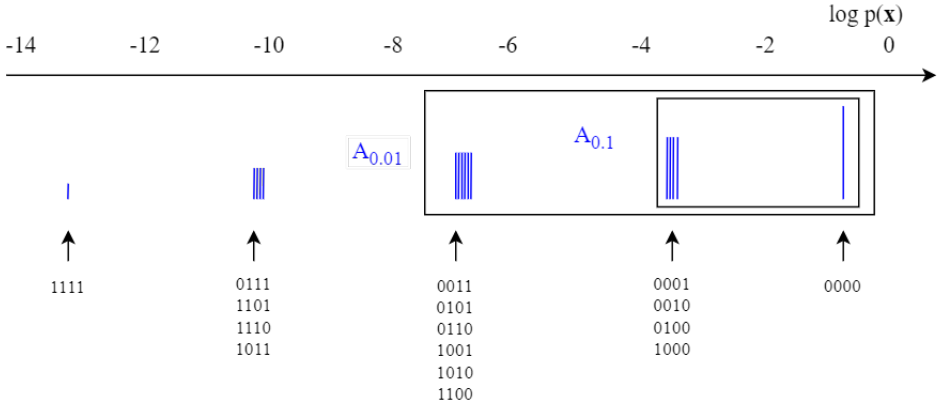


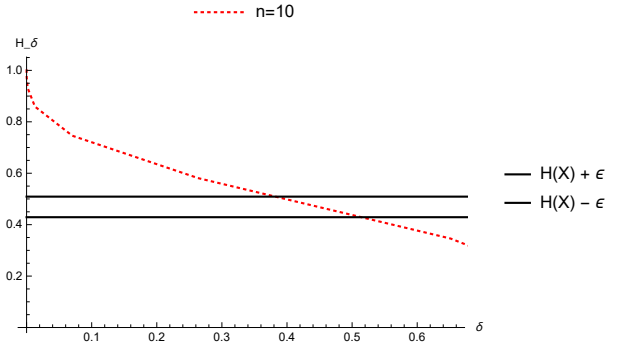
Figura 2.13: Subconjunto $\mathcal{A}_\delta$.

una fuerte dependencia de $H_\delta(\mathbf{X})$ con respecto a δ. Sin embargo, a medida que n es suficientemente grande, se puede encontrar que $H_\delta(\mathbf{X})$ se vuelve casi independiente de δ, i.e., para $0 < \delta < 1$ $H_\delta(\mathbf{X})$ es casi uniforme con valores muy cercanos a $nH(X)$, lo cual se observa en la Fig. 2.14c y es demostrado en [8]. Lo anterior es establecido en el siguiente teorema.

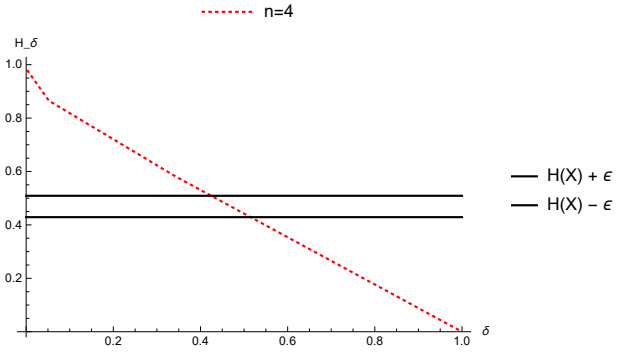
Teorema 2.8 (Teorema 4.1 en [8]). *Sea X una variable aleatoria con entropía $H(X) = H$ bits. Sea $\mathbf{X}$ un vector de variables aleatorias i.i.d. de dimensión n con variable aleatoria genérica X y sea $\epsilon > 0$ y $0 < \delta < 1$, luego existe un entero positivo n_0 tal que para $n > n_0$ se cumple:*

$$\left| \frac{1}{n} H_\delta(\mathbf{X}) - H \right| < \epsilon. \tag{2.99}$$

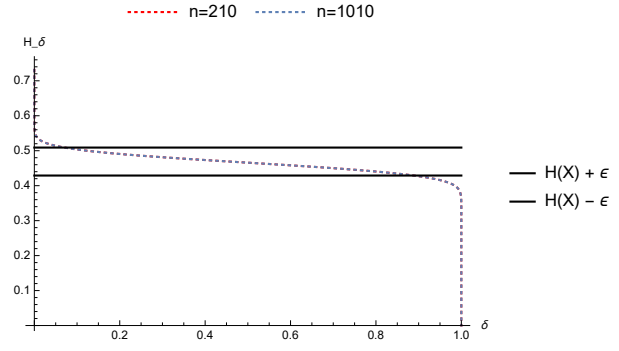
El Teorema 2.8 establece que la codificación de una fuente descrita por $\mathbf{X}$ tolera una pequeña probabilidad de error δ, tal que es posible la compresión de datos



(a)



(b)



(c)

Figura 2.14: H_δ vs δ.

reduciendo como mínimo la longitud de las palabras código en $nH(X)$ bits. Incluso si la probabilidad de error es significativa, es posible comprimir sin pérdidas de información en $nH(X)$ bits a medida que $n \to \infty$. Si se reduce la longitud, de tal manera que sea menor que $nH(X)$ bits, es casi seguro que se pierda información a medida que $n \to \infty$.

Capítulo 3

SECUENCIAS TÍPICAS CONJUNTAS

El concepto de tipicalidad es extendido a conjuntos de vectores de variables aleatorias i.i.d., de manera que, la tipicalidad vista previamente, es definida de forma similar para los vectores $\mathbf{X}$ y $\mathbf{Y}$ con variables aleatorias genéricas X y Y, respectivamente y la pmf conjunta $p_{\mathbf{XY}}$.

La tipicalidad del vector $\mathbf{X}$ y la tipicalidad conjunta de $(\mathbf{X}, \mathbf{Y})$ tiene dos formas: en sentido débil y en sentido fuerte, a las cuales les corresponde el conjunto típico débil y el conjunto típico fuerte. Estos conceptos son útiles en el desarrollo de la demostración del teorema de codificación de canal para el canal con ruido, como se presenta más adelante.

3.1. TIPICALIDAD CONJUNTA DÉBIL

En el Teorema 2.3 se estableció la AEP débil para el vector $\mathbf{X}$ de variables aleatorias discretas con pmf $p_{\mathbf{X}}$, y se definió el conjunto típico $\mathcal{T}_X^{n,\epsilon}$ con $\epsilon > 0$ y arbitrariemente pequeño. De manera similar, se desarrolla a continuación la tipicalidad conjunta débil de los vectores $\mathbf{X}$ y $\mathbf{Y}$, donde la entropía conjunta de las variables aleatorias

genéricas es denotada por $H(X,Y)$.

Lema 3.1 (Lema 62 en [54]). *Sean $\mathcal{X}$ y, $\mathcal{Y}$ dos conjuntos contables y X y Y dos variables aleatorias con pmf conjunta p_{XY}. Sean $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T \in \mathcal{X}^n$ y $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T \in \mathcal{Y}^n$ dos vectores de variables aleatorias discretas i.i.d. de dimensión n cuya pmf conjunta es:*

$$p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = \prod_{k=1}^{n} p_{XY}(x_k, y_k). \tag{3.1}$$

Luego, para cualquier $\epsilon > 0$ arbitrariamente pequeño y n suficientemente grande, se cumple:

$$\Pr\left\{ \left| -\frac{1}{n} \log p_{\mathbf{XY}}(\mathbf{X}, \mathbf{Y}) - H(X,Y) \right| < \epsilon \right\} \geq 1 - \epsilon. \tag{3.2}$$

Prueba: (Lema 62 en [54]) Sea Z una variable aleatoria discreta definida como:

$$Z = -\frac{1}{n} \log p_{\mathbf{XY}}(\mathbf{X}, \mathbf{Y}). \tag{3.3}$$

El valor esperado y la varianza de Z, $\mathbb{E}[Z]$ y $\mathrm{Var}[Z]$, respectivamente, se obtienen de forma similar que el valor esperado y la varianza de Y, definida en (2.31), esto es:

$$\mathbb{E}[Z] = H(X,Y), \tag{3.4}$$

$$\mathrm{Var}[Z] = \frac{1}{n} \mathrm{Var}[\log p_{XY}(X,Y)]. \tag{3.5}$$

De la desigualdad de Chebyshev (ver Apéndice D), para cualquier $a > 0$, lo siguiente es válido:

$$\Pr\{|Z - \mathbb{E}[Z]| \geq a\} \leq \frac{\mathrm{Var}[Z]}{a^2}. \tag{3.6}$$

Lo cual aplicando (3.3), (3.4) y (3.5) en (3.6) produce:

$$\Pr\left\{ \left| -\frac{1}{n} \log p_{\mathbf{XY}}(\mathbf{X}, \mathbf{Y}) - H(X,Y) \right| \geq a \right\} \leq \frac{1}{a^2 n} \mathrm{Var}[\log p_{XY}(X,Y)]. \tag{3.7}$$

Se observa que el valor esperado y la varianza de las variables aleatorias del vector $Z = (X,Y)$ son finitos, por lo tanto, la expresión del lado derecho de (3.7) es finita.

Así, para cualquier $\epsilon' > 0$, siempre existe un n suficientemente grande, de modo que:

$$\Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{XY}}(\mathbf{X},\mathbf{Y}) - H(X,Y)\right| \geq a\right\} \leq \epsilon'. \tag{3.8}$$

Finalmente, se observa que:

$$\begin{aligned}
&\Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{XY}}(\mathbf{X},\mathbf{Y}) - H(X,Y)\right| < a\right\} \\
&\quad = 1 - \Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{XY}}(\mathbf{X},\mathbf{Y}) - H(X,Y)\right| \geq a\right\} &(3.9)\\
&\quad \geq 1 - \epsilon'. &(3.10)
\end{aligned}$$

Por lo tanto, para todo $\epsilon > 0$, siempre existe un n suficientemente grande tal que:

$$\Pr\left\{\left| -\frac{1}{n}\log p_{\mathbf{XY}}(\mathbf{X},\mathbf{Y}) - H(X,Y)\right| < \epsilon\right\} \geq 1 - \epsilon. \tag{3.11}$$

Lo que completa la prueba. ■

La tipicalidad conjunta expresa la convergencia en probabilidad de una secuencia conjunta de variables aleatorias X y Y i.i.d. a la entropía conjunta $H(X,Y)$, siempre que n sea suficientemente grande.

3.1.1. Conjunto Típico Conjunto

Del Lema 3.1 se obtiene una división en dos conjuntos del total de las secuencias obtenidas de forma conjunta de los vectores $\mathbf{X}$ y $\mathbf{Y}$, estos son: un subconjunto nombrado conjunto típico conjunto (*joint typical set*) y su complemento. El conjunto típico conjunto define las secuencias típicas conjuntas del alfabeto $\mathcal{X} \times \mathcal{Y}$.

Definición 3.1. Sean $X \in \mathcal{X}$ y $Y \in \mathcal{Y}$ dos variables aleatorias genéricas con pmf conjunta p_{XY}. Sean $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ y $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ dos vectores de variables aleatorias discretas de dimensión n cuya distribución de probabilidad

conjunta es:

$$p_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) = \prod_{k=1}^{n} p_{XY}(x_k, y_k), \tag{3.12}$$

Luego, para cualquier $\epsilon > 0$ arbitrariamente pequeño, el conjunto de secuencias típicas conjuntas denotado por $\mathcal{T}_{XY}^{n,\epsilon}$ con respecto a p_{XY} es:

$$\begin{aligned}\mathcal{T}_{XY}^{n,\epsilon} \triangleq \Bigg\{ (\mathbf{x},\mathbf{y}) \in (\mathcal{X}\times\mathcal{Y})^n : & \left| -\frac{1}{n}\log(p_{\mathbf{X}}(\mathbf{x})) - H(X) \right| < \epsilon, \\ & \left| -\frac{1}{n}\log(p_{\mathbf{Y}}(\mathbf{y})) - H(Y) \right| < \epsilon, \quad \text{y} \\ & \left| -\frac{1}{n}\log(p_{\mathbf{XY}}(\mathbf{x},\mathbf{y})) - H(X,Y) \right| < \epsilon \Bigg\}. \end{aligned} \tag{3.13}$$

Se observa que si, $(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}^{n,\epsilon}$, entonces, $\mathbf{x} \in \mathcal{T}_{X}^{n,\epsilon}$ y $\mathbf{y} \in \mathcal{T}_{Y}^{n,\epsilon}$. Es decir, que $\mathbf{x}$ es una secuencia típica de $p_{\mathbf{X}}$ y $\mathbf{y}$ es una secuencia típica de $p_{\mathbf{Y}}$, entonces, $(\mathbf{x},\mathbf{y})$ es una secuencia típica conjunta de $p_{\mathbf{XY}}$. El conjunto típico conjunto por la Definición 3.1 y el Lema 3.1 tiene las siguientes propiedades.

Lema 3.2. *Sea $\mathcal{T}_{XY}^{n,\epsilon}$ el conjunto de secuencias típicas conjuntas con respecto a p_{XY} y $\epsilon > 0$, por lo tanto, se satisface lo siguiente:*

1. *Si el vector de realizaciones $(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}^{n,\epsilon}$, entonces:*

$$2^{-n(H(X,Y)+\epsilon)} \leq p_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \leq 2^{-n(H(X,Y)-\epsilon)} \tag{3.14}$$

$$2^{-n(H(X|Y)+2\epsilon)} \leq p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) \leq 2^{-n(H(X|Y)-2\epsilon)}. \tag{3.15}$$

2. *Para n suficientemente grande,*

$$\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}_{XY}^{n,\epsilon}} p_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \geq 1-\epsilon \tag{3.16}$$

3. *Para n suficientemente grande*

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |\mathcal{T}_{XY}^{n,\epsilon}| \leq 2^{n(H(X,Y)+\epsilon)}. \tag{3.17}$$

Prueba:

- Prueba de (3.14): Es obtenida directamente de la Definición 3.1.
- Prueba de (3.15): Por la Definición 3.1, se cumple que $\mathbf{x} \in \mathcal{T}_X^{n,\epsilon}$ y $\mathbf{y} \in \mathcal{T}_Y^{n,\epsilon}$ y por lo tanto:

$$2^{-n(H(X)+\epsilon)} \leq p_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)} \quad \text{y,} \tag{3.18}$$

$$2^{-n(H(Y)+\epsilon)} \leq p_{\mathbf{Y}}(\mathbf{y}) \leq 2^{-n(H(Y)-\epsilon)}. \tag{3.19}$$

  La probabilidad condicional definida como: $p_{X|Y}(x|y) = p_{XY}(x,y)/p_Y(y)$ de X dado Y. Al multiplicar la desigualdad (3.14) por el inverso de (3.19), se obtiene lo siguiente:

$$2^{-n(H(X|Y)+2\epsilon)} \leq p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) \leq 2^{-n(H(X|Y)-2\epsilon)}, \tag{3.20}$$

  donde $H(X|Y) = H(X,Y) - H(Y)$.

- Prueba de (3.16): Del Lema 3.1 para $\epsilon > 0$ se obtiene directamente (3.16).
- Prueba de (3.17): A partir de la probabilidad total del espacio muestral de secuencias, se aplica (3.16) y la cota inferior de (3.14), obteniendo lo siguiente:

$$1 = \sum_{(\mathbf{x},\mathbf{y}) \in (\mathcal{X}\times\mathcal{Y})^n} p_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \tag{3.21}$$

$$\geq \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}^{n,\epsilon}} p_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \tag{3.22}$$

$$\geq \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}^{n,\epsilon}} 2^{-n(H(X,Y)+\epsilon)} \tag{3.23}$$

$$= |\mathcal{T}_{XY}^{n,\epsilon}| 2^{-n(H(X,Y)+\epsilon)}. \tag{3.24}$$

  Para n suficientemente grande, implica:

$$|\mathcal{T}_{XY}^{n,\epsilon}| \leq 2^{n(H(X,Y)+\epsilon)}. \tag{3.25}$$

  De forma similar, haciendo uso de (3.17) se obtiene lo siguiente:

$$1 - \epsilon \leq \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{XY}^{n,\epsilon}} p_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \tag{3.26}$$

$$\leq \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}_{XY}^{n,\epsilon}} 2^{-n(H(X,Y)-\epsilon)} \tag{3.27}$$

$$= |\mathcal{T}_{XY}^{n,\epsilon}| 2^{-n(H(X,Y)-\epsilon)}. \tag{3.28}$$

Para n suficientemente grande (3.28) implica:

$$|\mathcal{T}_{XY}^{n,\epsilon}| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}, \tag{3.29}$$

lo cual completa la prueba de (3.17) y la prueba del Lema 3.2. ■

Se observa que, las propiedades del conjunto típico conjunto son similares a las propiedades del conjunto típico obtenidas por la AEP débil. Así, del subconjunto $\mathcal{T}_{XY}^{n,\epsilon}$ de $\mathcal{X}\times\mathcal{Y}$ se afirma que, la probabilidad de los elementos del conjunto típico conjunto tienen una probabilidad de ocurrencia muy cercana a $2^{-nH(X,Y)}$ por (3.14); la probabilidad del conjunto típico conjunto es muy cercana a uno, con n suficientemente grande por (3.16); y por último, el número de elementos del conjunto típico conjunto es aproximadamente $2^{nH(X,Y)}$ por (3.17).

3.2. TIPICALIDAD FUERTE

La tipicalidad débil requiere que la entropía empírica de una secuencia sea cercana a la entropía teórica, sin embargo, una noción más fuerte de tipicalidad se basa en que la frecuencia relativa de cada posible resultado sea cercana a la probabilidad correspondiente [10]. A partir de esta idea, se construye la definición del conjunto típico fuerte para $\mathbf{X}$.

3.2.1. Conjunto Típico Fuerte

Se considera la secuencia de variables aleatorias discretas $\{X_m, m \in \mathbb{Z}^+ \text{ y } m \geq 1\}$, donde las variables aleatorias X_m son i.i.d., cada una con similar distribución de

probabilidad p_X y alfabeto $\mathcal{X}$. X denota la variable aleatoria genérica con entropía $H(X)$.

Definición 3.2 (Definición 6.1 en [10])**.** Sea $X \in \mathcal{X}$ una variable aleatoria con pmf p_X y sea $\mathbf{X}$ el vector de variables aleatorias discretas de dimension n con pmf conjunta $p_{\mathbf{X}}$. El conjunto típico fuerte denotado $\mathcal{W}_X^{n,\delta}$ con respecto a la pmf p_X es el conjunto de secuencias $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$, tal que, $N(x;\mathbf{x}) = 0$ para $\mathbf{x} \notin \mathcal{S}_X$, y para el cual se cumple:

$$\sum_{x \in \mathcal{S}_X} \left| \frac{1}{n} N(x;\mathbf{x}) - p_X(x) \right| \leq \delta, \tag{3.30}$$

donde, $N(x;\mathbf{x})$ es el número de ocurrencias de x en la secuencia $\mathbf{x}$ y δ es un número real positivo arbitrariamente pequeño. Las secuencias en $\mathcal{W}_X^{n,\delta}$ son llamadas secuencias típicamente fuertes δ (*strongly δ-typical sequences*).

3.2.2. Propiedad de Equipartición Asintótica Fuerte

El conjunto típico fuerte $\mathcal{W}_X^{n,\delta}$ comparte propiedades similares con el conjunto típico débil $\mathcal{T}_X^{n,\epsilon}$, a continuación se presenta el la AEP en sentido fuerte.

Lema 3.3 (Teorema 6.2 en [10])**.** *Existe un $\eta > 0$, tal que $\eta \to 0$ mientras $\delta \to 0$, y lo siguiente se cumple:*

1. *Si $\mathbf{x} \in \mathcal{W}_X^{n,\delta}$, entonces:*

$$2^{-n(H(X)+\eta)} \leq p_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}. \tag{3.31}$$

2. *Para n suficientemente grande,*

$$\Pr\{\mathbf{x} \in \mathcal{W}_X^{n,\delta}\} \geq 1 - \delta. \tag{3.32}$$

3. *Para n suficientemente grande,*

$$(1-\delta)2^{n(H(X)-\eta)} \leq |\mathcal{W}_X^{n,\delta}| \leq 2^{n(H(X)+\eta)}. \tag{3.33}$$

Prueba: ▪ Prueba de (3.31): Para $\mathbf{x} \in \mathcal{W}_X^{n,\delta}$, la pmf conjunta $p_{\mathbf{X}}$ se obtiene del producto de probabilidades de los elementos de $x \in \mathcal{X}$, tales que $p_X(x) > 0$ y $N(x;\mathbf{x})$ es el número de ocurrencias en la secuencias $\mathbf{x}$, i.e.,

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{x \in \mathcal{S}_X} p_X(x)^{N(x;\mathbf{x})}. \tag{3.34}$$

A partir de (3.34) aplicando la función logaritmo, se obtiene lo siguiente:

$$\begin{aligned}
&\log p_{\mathbf{X}}(\mathbf{x}) \\
&= \sum_{x \in \mathcal{S}_X} N(x;\mathbf{x}) \log p_X(x) \\
&= \sum_{x \in \mathcal{S}_X} (N(x;\mathbf{x}) - np_X(x) + np_X(x)) \log p_X(x) \\
&= n \sum_{x \in \mathcal{S}_X} p_X(x) \log p_X(x) - n \sum_{x \in \mathcal{S}_X} \left(\frac{1}{n} N(x;\mathbf{x}) - p_X(x) \right) (-\log p_X(x)) \\
&= -n \left[H(X) + \sum_{x \in \mathcal{S}_X} \left(\frac{1}{n} N(x;\mathbf{x}) - p_X(x) \right) (-\log p_X(x)) \right]. \qquad (3.35)
\end{aligned}$$

Dado que $\mathbf{x} \in \mathcal{W}_X^{n,\delta}$, se cumple que:

$$\sum_{x \in \mathcal{S}_X} \left| \frac{1}{n} N(x;\mathbf{x}) - p_X(x) \right| \leq \delta, \tag{3.36}$$

y al aplicar valor absoluto al segundo término de (3.35), implica lo siguiente:

$$\begin{aligned}
&\left| \sum_{x \in \mathcal{S}_X} \left(\frac{1}{n} N(x;\mathbf{x}) - p_X(x) \right) (-\log p_X(x)) \right| \\
&\leq \sum_{x \in \mathcal{S}_X} \left| \frac{1}{n} N(x;\mathbf{x}) - p_X(x) \right| (-\log p_X(x)) \qquad (3.37) \\
&\leq -\log \big(\min p_X(x) \big) \sum_{x \in \mathcal{S}_X} \left| \frac{1}{n} N(x;\mathbf{x}) - p_X(x) \right| \qquad (3.38) \\
&\leq -\delta \log \big(\min p_X(x) \big) \qquad (3.39) \\
&= \eta, \qquad (3.40)
\end{aligned}$$

donde, (3.37) se obtiene por la desigualdad triangular y el factor $-(\log p_X(x))$ sale del valor absoluto por ser positivo; en (3.38) se elige el valor máximo de $-(\log p_X(x))$, que se obtiene del mínimo valor en la pmf p_X; (3.39) se obtiene al aplicar la desigualdad de (3.36); y finalmente $\eta = -\delta \log\big(\min p_X(x)\big) > 0$. De (3.40) se obtiene:

$$-\eta \leq \sum_{x \in \mathcal{S}_X} \left(\frac{1}{n} N(x; \mathbf{x}) - p_X(x) \right) (-\log p_X(x)) \leq \eta. \tag{3.41}$$

Lo siguiente se obtiene de (3.35):

$$-\eta \leq -\frac{1}{n} \log p_{\mathbf{X}}(\mathbf{x}) - H(X) \leq \eta \tag{3.42}$$

$$H(X) - \eta \leq -\frac{1}{n} \log p_{\mathbf{X}}(\mathbf{x}) \leq H(X) + \eta \tag{3.43}$$

$$-n(H(X) + \eta) \leq \log p_{\mathbf{X}}(\mathbf{x}) \leq -n(H(X) - \eta), \tag{3.44}$$

o lo que es equivalente,

$$2^{-n(H(X)+\eta)} \leq p_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}, \tag{3.45}$$

donde, $\eta \to 0$ mientras $\delta \to 0$. Esto completa la prueba.

- Prueba de (3.32): Se reescribe el número de ocurrencias de x en $\mathbf{x}$, como la sumatoria del número de veces que $X_k = x$, i.e.:

$$N(x; \mathbf{X}) = \sum_{k=1}^{n} B_k(x), \tag{3.46}$$

donde $B_k(x) = 1$ si $X_k = x$, y $B_k(x) = 0$ si $X_k \neq x$. Luego $B_k(x)$ con $k \in \mathbb{Z}^+$ son variables aleatorias i.i.d. cuya pmf p_X esta definida por:

$$\Pr\{B_k(x) = 1\} = p_X(x), \text{ y } \Pr\{B_k(x) = 0\} = 1 - p_X(x). \tag{3.47}$$

El valor esperado $\mathbb{E}[B_k(x)]$ es:

$$\mathbb{E}[B_k(x)] = (1 - p_X(x)) \times 0 + p_X(x) \times 1 = p_X(x). \tag{3.48}$$

Por la LLN débil, para cualquier $\delta > 0$ y para cualquier $x \in \mathcal{X}$, se cumple lo siguiente:

$$\Pr\left\{\left|\frac{1}{n}\sum_{k=1}^{n} B_k(x) - p_X(x)\right| > \frac{\delta}{|\mathcal{X}|}\right\} < \frac{\delta}{|\mathcal{X}|}, \tag{3.49}$$

para n suficientemente grande. Luego se obtiene,

$$\Pr\left\{\left|\frac{1}{n}N(x;\mathbf{X}) - p_X(x)\right| > \frac{\delta}{|\mathcal{X}|} \text{ para algún } x\right\} \tag{3.50}$$

$$= \Pr\left\{\left|\frac{1}{n}\sum_{k=1}^{n} B_k(x) - p_X(x)\right| > \frac{\delta}{|\mathcal{X}|} \text{ para algún } x\right\} \tag{3.51}$$

$$= \Pr\left\{\bigcup_x \left\{\left|\frac{1}{n}\sum_{k=1}^{n} B_k(x) - p_X(x)\right| > \frac{\delta}{|\mathcal{X}|}\right\}\right\} \tag{3.52}$$

$$\leq \sum_x \Pr\left\{\left|\frac{1}{n}\sum_{k=1}^{n} B_k(x) - p_X(x)\right| > \frac{\delta}{|\mathcal{X}|}\right\} \tag{3.53}$$

$$< \sum_x \frac{\delta}{|\mathcal{X}|} \tag{3.54}$$

$$= \delta, \tag{3.55}$$

donde, en (3.53) se ha considerado que la union $\Pr\{A \cup B\} \leq \Pr\{A\} + \Pr\{B\}$. Dado que

$$\sum_{x \in \mathcal{S}_X} \left|\frac{1}{n}N(x;\mathbf{x}) - p_X(x)\right| > \delta, \tag{3.56}$$

lo cual implica:

$$\left|\frac{1}{n}N(x;\mathbf{x}) - p_X(x)\right| > \frac{\delta}{|\mathcal{X}|} \text{ para algún } x \in \mathcal{X}. \tag{3.57}$$

Por lo tanto, lo siguiente puede ser obtenido:

$$\Pr\ \{\ \mathbf{X} \in \mathcal{W}_X^{n,\delta}\} = \Pr\left\{ \sum_x \left| \frac{1}{n} N(x;\mathbf{X}) - p_X(x) \right| \leq \delta \right\} \tag{3.58}$$

$$= 1 - \Pr\left\{ \sum_x \left| \frac{1}{n} N(x;\mathbf{X}) - p_X(x) \right| > \delta \right\} \tag{3.59}$$

$$\geq 1 - \Pr\left\{ \left| \frac{1}{n} N(x;\mathbf{X}) - p_X(x) \right| > \frac{\delta}{|\mathcal{X}|} \text{ para algún } x \in \mathcal{X} \right\} \tag{3.60}$$

$$> 1 - \delta, \tag{3.61}$$

donde, (3.60) es utilizado aplicando (3.57) en (3.59), y (3.61) se obtiene de (3.55). Lo cual completa la prueba.

- Prueba de (3.33): Esta prueba se obtiene de las propiedades 1 y 2.

Esto completa la prueba del Lema 3.3 ■

La tipicalidad fuerte es más potente y flexible que la tipicalidad débil como una herramienta para el desarrollo de los teoremas en problemas sin memoria, pero puede ser solo utilizada para variables aleatorias con alfabeto finito. Se dice que es más fuerte en el sentido de que la tipicalidad fuerte implica la tipicalidad débil, pero no al contrario [10].

Sea X una variable aleatoria discreta con alfabeto $\mathcal{X} = \{0, 1, 2\}$ y pmf: $p_X(0) = 0.5$, $p_X(1) = 0.25$ y $p_X(2) = 0.25$. Sea $\mathbf{X}$ el vector de variables aleatorias i.i.d. de dimensión n y $q(i)$ la frecuencia relativa de ocurrencia del símbolo $i \in \mathcal{X}$, en la secuencia $\mathbf{x}$, i.e., $\frac{1}{n}N(i;\mathbf{x})$. Ahora se observa la tipicalidad de la secuencia $\mathbf{x}$: para que $\mathbf{x}$ sea típica débil, la entropía empírica debe ser muy cercana a la entropía teórica, esto es:

$$-\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{x}) = -q(0)\log 0.5 - q(1)\log 0.25 - q(2)\log 0.25 \tag{3.62}$$

$$= -0.5\log 0.5 - 0.25\log 0.25 - 0.25\log 0.25 \tag{3.63}$$

$$= H(X). \tag{3.64}$$

Al seleccionar $q(i) = p(i)$ para todo i, la secuencia $\mathbf{x}$ es típica débil. Si en cambio, se escoge: $q(0) = 0.5$, $q(1) = 0.5$ y $q(2) = 0$, la secuencia $\mathbf{x}$ continua siendo típicamente

débil con respecto a p_X, pero la condición de tipicalidad fuerte no se satisface, debido a que la frecuencia relativa de ocurrencia de cada símbolo i es $q(i)$, la cual no es igual a $p(i)$ para $i \in \{1, 2\}$ [10].

Capítulo 4

CAPACIDAD DEL CANAL DIGITAL DISCRETO

El modelo del sistema de comunicación presentado en la Fig. 2.1 corresponde a un proceso estocástico definido por variables aleatorias discretas. El análisis se enfocó en la medida de información $\iota(x)$ y el promedio ponderado de la medida de información de todos los posibles resultados o medida de incertidumbre promedio de la variable aleatoria X, denotada por $H(X)$. La fuente de información fue modelada como una secuencia de variables aleatorias i.i.d. y se consideró el canal de comunicación discreto sin ruido. El concepto de secuencias típicas permitió establecer que una transmisión punto a punto sin ruido la codificación de la información esta determinada por la entropía como límite fundamental establecido por el teorema de codificación de fuente en [1]. Ahora, se quiere analizar y representar el efecto del ruido sobre el canal de comunicación y examinar el límite fundamental que propone el teorema de codificación de canal con respecto a la velocidad de transmisión de información.

En el sistema de comunicación discreto sin ruido, la entropía de la fuente de información definida para variables aleatorias i.i.d. en (2.12), es igual a la entropía de la salida del canal $H(Y)$, i.e., $H(Y) = H(X)$, ya que no hay interferencia o cambios que alteren la medida de información. Sin embargo, considerar el sistema de comunicación sin ruido es una idealización, por lo cual, la introducción del ruido en el canal de comunicación requiere una abstracción matemática adecuada, i.e.,

medir la información de la secuencia de variables aleatorias a la entrada del canal con respecto a la secuencia de variables aleatorias a la salida del canal o viceversa. Lo anterior implica la observación de los fénomenos físicos correlacionados, en los cuales cualquier suceso está vinculado o es afectado por el medio que lo rodea. De esta consideración se extiende el concepto de medida de información y se introduce el concepto de ruido en el canal, con el cual se acerca a un modelo de comunicación más real.

El análisis del sistema de comunicación con ruido requiere de la extensión de las propiedades que satisface la entropía para conjuntos de vectores de variables aleatorias discretas.

4.1. ENTROPÍA CONDICIONAL E INFORMACIÓN MUTUA

En la Definición 2.2 se observó que la entropía $H(X)$ de la variable aleatoria discreta X, es un funcional de la pmf p_X, de forma similar, las entropías que se definen a continuación, son funcionales de la función de distribución de probabilidad en cuestión.

Definición 4.1 (Definición 10 en [54])**.** Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos contables y X, Y dos variables aleatorias con pmf conjunta p_{XY}. La entropía conjunta de X y Y denotada por $H(X,Y)$ es:

$$H(X,Y) = -\sum_{(x,y)\in\mathcal{S}_X\times\mathcal{S}_Y} p_{XY}(x,y)\log p_{XY}(x,y). \tag{4.1}$$

En el caso de variables aleatorias independientes la entropía conjunta esta dada por (2.9). Otra forma de escribir la entropía conjunta es:

$$H(X,Y) = \mathbb{E}\big[\log p_{XY}(X,Y)\big] \tag{4.2}$$

Definición 4.2 (Definición 11 en [54])**.** Sean $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$, n conjuntos contables y sea $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ el vector de variables aleatorias de dimensión n con

pmf conjunta $p_{\mathbf{X}}$, la entropía conjunta de $\mathbf{X}$, denotada por $H(\mathbf{X})$, es:

$$H(\mathbf{X}) = -\sum_{\mathbf{x}\in\mathcal{S}_{\mathbf{X}}} p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}). \tag{4.3}$$

La entropía conjunta del vector $\mathbf{X}$ puede ser escrito como:

$$H(\mathbf{X}) = -\mathbb{E}\big[\log p_{\mathbf{X}}(\mathbf{X})\big]. \tag{4.4}$$

La entropía de un vector de variables aleatorias $X_1, X_2, \ldots X_n$ mutuamente independientes, satisface lo siguiente:

$$H(\mathbf{X}) = \sum_{k=1}^{n} H(X_k). \tag{4.5}$$

4.1.1. Entropía Condicional

La entropía condicional de X dado Y denotada por $H(X|Y)$ es una medida de la cantidad de información promedio necesaria para identificar la variable aleatoria X, dada la observación de la variable aleatoria Y.

Definición 4.3 (Definición 2.15 en [10])**.** Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos finitos contables y X, Y dos variables aleatorias con pmf conjunta p_{XY}. La entropía condicional de X dado $Y = b_k$, es la entropía de la pmf condicional $p_{X|Y}(x|y = b_k)$:

$$H(X|y = b_k) = -\sum_{x\in\mathcal{X}} p_{X|Y}(x|y = b_k) \log p_{X|Y}(x|y = b_k). \tag{4.6}$$

Definición 4.4 (Definición en 2.15 [10])**.** Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos finitos contables y X, Y dos variables aleatorias con pmf conjunta p_{XY}. La entropía de X dado Y es:

$$H(X|Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{XY}(x, y) \log p_{X|Y}(x|y). \tag{4.7}$$

La entropía condicional se puede reescribir de la definición de la pmf cojunta y de

la Definición 4.3, de donde se obtiene lo siguiente:

$$H(X|Y) = -\sum_{(x,y)\in\mathcal{S}_X\times\mathcal{S}_Y} p_Y(y)p_{X|Y}(x|y)\log p_{X|Y}(x|y) \tag{4.8}$$

$$= \sum_{y\in\mathcal{S}_Y} p_Y(y)\left[-\sum_{x\in\mathcal{S}_X} p_{X|Y}(x|y)\log p_{X|Y}(x|y)\right] \tag{4.9}$$

$$= \sum_{y\in\mathcal{S}_Y} p_Y(y)H(X|y), \tag{4.10}$$

donde $H(X|y)$ es la entropía de X condicionada para un valor fijo de $Y = y$. Otra forma de escribir la entropía condicional es:

$$H(X|Y) = -\mathbb{E}\big[\log p_{X|Y}(X|Y)\big] \tag{4.11}$$

Definición 4.5. La entropía marginal de X, es el nombre dado a la entropía de X, $H(X)$, con respecto a la entropía conjunta $H(X,Y)$.

De la entropía condicional de X dado Y, se obtiene una reducción en la entropía de la variable aleatoria X o de la variable aleatoria que se observa por el conocimiento de la realización de la otra variable aleatoria.

Lema 4.1 (Lema 38 en [54])**.** *Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos finitos contables y X, Y dos variables aleatorias con pmf conjunta p_{XY}, por lo tanto, la siguiente desigualdad se satisface:*

$$H(X|Y) \leq H(X), \tag{4.12}$$

donde se cumple la igualdad si y solo si las variables aleatorias X y Y son independientes.

Prueba: De (4.11), y de la definición de probabilidad condicional, se obtiene:

$$H(X|Y) = -\mathbb{E}_{XY}\left[\log\left(\frac{p_X(X)p_{Y|X}(Y|X)}{p_Y(Y)}\right)\right] \tag{4.13}$$

$$= -\mathbb{E}_X[\log p_X(X)] - \mathbb{E}_{XY}\left[\log\left(\frac{p_{Y|X}(Y|X)}{p_Y(Y)}\right)\right] \tag{4.14}$$

$$= H(X) - \mathbb{E}_{XY}\left[\log\left(\frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right)\right] \tag{4.15}$$

$$= H(X) + \mathbb{E}_{XY}\left[\log\left(\frac{p_X(X)p_Y(Y)}{p_{XY}(X,Y)}\right)\right] \tag{4.16}$$

$$\leq H(X) + \log\left(\mathbb{E}_{XY}\left[\left(\frac{p_X(X)p_Y(Y)}{p_{XY}(X,Y)}\right)\right]\right) \tag{4.17}$$

$$= H(X) + \log\left(\sum_{(x,y)\in\mathcal{S}_X\times\mathcal{S}_Y} p_X(X)p_Y(Y)\right) \tag{4.18}$$

$$= H(X), \tag{4.19}$$

donde (4.17) se obtiene de la aplicación de la desigualdad de Jensen en (2.14b).

Si las variables aleatorias X y Y son independientes, de (4.15) se obtiene lo siguiente:

$$H(X|Y) = H(X) + \mathbb{E}_{XY}\left[\log\left(\frac{p_X(X)p_Y(Y)}{p_X(X)p_Y(Y)}\right)\right] \tag{4.20}$$

$$= H(X). \tag{4.21}$$

Lo cual completa la prueba del Lema 4.1. ■

La entropía conjunta se puede definir en términos de la entropía condicional, por medio de la aplicación de la regla de la cadena en el Lema 4.2.

Lema 4.2 (Proposición 2.16 en [10])**.** *La entropía conjunta, la entropía condicional y la entropía marginal están relacionadas por:*

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{4.22}$$

Prueba: De (2.4), (4.2) y (4.11) se obtiene lo siguiente:

$$\begin{aligned} H(X,Y) &= -\mathbb{E}_{XY}\big[\log p_{XY}(X,Y)\big] && (4.23)\\ &= -\mathbb{E}_{XY}\left[\log\left(\frac{p_X(X)p_{XY}(X,Y)}{p_X(X)}\right)\right] && (4.24)\\ &= -\mathbb{E}_X\big[\log p_X(X)\big] - \mathbb{E}_{XY}\left[\log\left(\frac{p_{XY}(X,Y)}{p_X(X)}\right)\right] && (4.25)\\ &= H(X) + H(Y|X). && (4.26) \end{aligned}$$

En (4.24), si $p_X(x) \neq 0$. De forma similar si $p_Y(y) \neq 0$, se obtiene:

$$H(X,Y) = H(Y) + H(X|Y). \tag{4.27}$$

Lo cual completa la prueba. ■

El Lema 4.2, afirma que la incertidumbre de X y Y es la incertidumbre de X más la incertidumbre de Y dado X.

4.1.2. Información Mutua

La información mutua entre las variables aleatorias X y Y, es la cantidad promedio de información acerca de una de las variables aleatorias proporcionada por la ocurrencia de la otra variable aleatoria, y se denota por $I(X;Y)$.

Definición 4.6 (Definición 13 en [54])**.** Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos finitos contables y, X y Y dos variables aleatorias con pmf conjunta p_{XY}, la información mutua entre X y Y, denotada por $I(X;Y)$, es:

$$I(X;Y) = \sum_{x,y\in\mathcal{S}_X\times\mathcal{S}_Y} p_{XY}(x,y)\log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right). \tag{4.28}$$

La información mutua entre las variables aleatorias X y Y también se puede escribir

como:

$$I(X;Y) = \mathbb{E}_{XY}\left[\log\left(\frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right)\right] \tag{4.29}$$

$$= \mathbb{E}_{XY}\left[\log\left(\frac{p_{Y|X}(Y|X)}{p_Y(Y)}\right)\right] \tag{4.30}$$

$$= \mathbb{E}_{XY}\left[\log\left(\frac{p_{X|Y}(X|Y)}{p_X(X)}\right)\right]. \tag{4.31}$$

El siguiente lema presenta algunas propiedades útiles de la información mutua.

Lema 4.3 (Lema 43 en [54])**.** *Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos finitos contables, y X y Y dos variables aleatorias con pmf conjunta p_{XY}, entonces se satisface lo siguiente:*

$$I(X;Y) = I(Y;X), \tag{4.32}$$

$$I(X;Y) = H(X) - H(X|Y), \tag{4.33}$$

$$I(X;Y) = H(Y) - H(Y|X), \tag{4.34}$$

$$I(X;Y) \geq 0, \tag{4.35}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{4.36}$$

$$I(X;X) = H(X). \tag{4.37}$$

Prueba:

- Prueba de (4.32): Se obtiene directamente de la Definición 4.6.
- Prueba de (4.33): Desarrollando (4.31) se obtiene:

$$I(X;Y) = -\mathbb{E}_X\big[\log p_X(X)\big] + \mathbb{E}_{XY}\big[\log p_{X|Y}(X|Y)\big] \tag{4.38}$$

$$= H(X) - H(X|Y), \tag{4.39}$$

  esto completa la prueba de (4.33).

- Prueba de (4.34): Desarrollando (4.30) se obtiene:

$$I(X;Y) = -\mathbb{E}_Y\big[\log p_Y(Y)\big] + \mathbb{E}_{XY}\big[\log p_{Y|X}(X|X)\big] \tag{4.40}$$
$$= H(Y) - H(Y|X), \tag{4.41}$$

  esto completa la prueba de (4.34).

- Prueba de (4.35): De (4.33) y del Lema 4.1 se obtiene:

$$H(X|Y) \leq H(X) \tag{4.42}$$
$$H(X) - H(X|Y) \geq H(X) - H(X) \tag{4.43}$$
$$I(X;Y) \geq 0. \tag{4.44}$$

  Si X y Y son independientes, $p_{XY}(X,Y) = p_X(X)p_Y(Y)$, de (4.29) se obtiene:

$$I(X;Y) = \mathbb{E}_{XY}\left[\log\left(\frac{p_X(X)p_Y(Y)}{p_X(X)p_Y(Y)}\right)\right] \tag{4.45}$$
$$= \mathbb{E}_{XY}[\log 1] \tag{4.46}$$
$$= 0, \tag{4.47}$$

  esto completa la prueba de (4.35)

- Prueba de (4.36): De (4.33) y del Lema 4.2 se obtiene:

$$I(X;Y) = H(X) - H(X|Y) \tag{4.48}$$
$$= H(X) - [-H(Y) + H(X,Y)] \tag{4.49}$$
$$= H(X) + H(Y) - H(X,Y), \tag{4.50}$$

  esto completa la prueba de (4.36).

- Prueba de (4.37): Sea Y una variable aleatoria idéntica a X, i.e., $Y = X$.

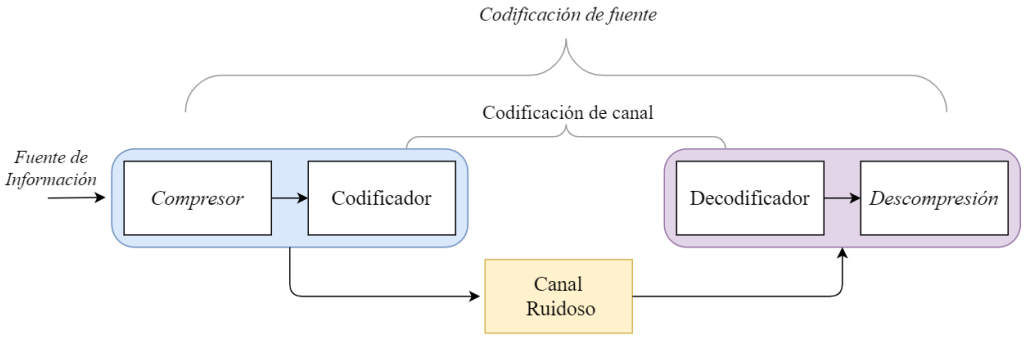


Figura 4.1: Diagrama de Bloques de la Codificación

De (4.29):

$$I(X;X) = \mathbb{E}_{XY}\left[\log\left(\frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right)\right], Y = X \tag{4.51}$$

$$= \mathbb{E}_X\left[\log\left(\frac{p_X(X)}{p_X(X)p_X(X)}\right)\right] \tag{4.52}$$

$$= \mathbb{E}_X\left[\log\left(\frac{1}{p_X(X)}\right)\right] \tag{4.53}$$

$$= -\mathbb{E}_X[\log p_X(X)] \tag{4.54}$$

$$= H(X), \tag{4.55}$$

esto completa la prueba de (4.37).

Lo cual completa la prueba del Lema 4.3. ■

4.2. SISTEMA DE COMUNICACIÓN DISCRETO CON RUIDO

En el sistema de comunicación discreto con ruido, la señal que se transmite es modificada por el canal, esto es, que la señal que se recibe no es la misma señal que fue emitida por el transmisor. La modificación que introduce el canal, se denomina

ruido, el cual produce en la señal cambios en cada transmisión, de manera que, la señal recibida es una función de la señal transmitida y del ruido.

El ruido posee una estructura estadística altamente aleatoria. Por lo cual es representado por un proceso estocástico definido de acuerdo al sistema de comunicación. El ruido afecta el mensaje durante la transmisión a través del canal. El objetivo de la comunicación es que a pesar del ruido, el receptor recupere el mensaje original con total certeza, lo cual no es posible si en el transmisor el mensaje no se ha procesado adecuadamente para protegerlo del ruido. La codificación es la condición previa que permite reconstruir el mensaje original en el receptor [8].

Las operaciones realizadas por el transmisor a los datos generados por la fuente de información, se pueden dividir en dos etapas de codificación: codificación de fuente y, la codificación de canal, como se observa en la Fig. 4.1. En la codificación de fuente, el objetivo es representar la salida de la fuente de información por una secuencia de dígitos binarios, y la pregunta se enfoca en cuántos bits por unidad de tiempo se requieren para caracterizar la salida de cualquier modelo de fuente de información. La codificación de canal se ocupa de que la transmisión de las secuencias de datos a través de un canal con ruido sea confiable [57].

El modelo general de la Fig. 2.1 se vuelve a presentar, pero ahora descrito por las funciones de distribucion de probabilidad de variables aleatorias dependientes, como se observa en la Fig. 4.2. La fuente de información es descrita por un vector $\mathbf{X}$ de variables aleatorias i.i.d. de dimensión n; el canal Q es descrito por una matriz de probabilidades de transición que modelan el efecto del ruido que actúa sobre el vector de entrada del canal, convirtiéndolo en un nuevo vector a la salida del canal; en el punto de recepción, las secuencias están definidas por un vector $\mathbf{Y}$ de dimensión n, las cuales están correlacionadas con las secuencias de entrada, porque de lo contrario es imposible la comunicación.

El problema de la comunicación plantea las siguientes preguntas: ¿es posible una comunicación libre de errores a través de un canal con ruido?, ¿cuáles son las posibilidades o limitaciones de tal sistema de comunicación?, y ¿cuál es el objetivo de la codificación de canal? [1].

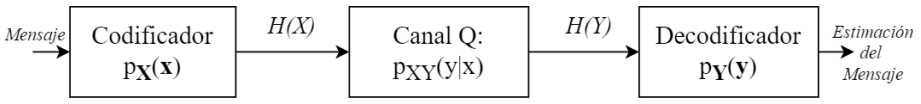


Figura 4.2: Diagrama del Canal con Ruido.

En este capítulo, se asume que los mensajes han pasado a través de un compresor de datos, de manera que, las secuencias de variables aleatorias que representan el mensaje no tienen redundancia. La codificación de canal, añade una redundancia especial diseñada para hacer que la señal recibida sea decodificable [10]. A continuación, se desarrollan los conceptos probabilísticos que permiten estudiar el sistema de comunicación discreto con ruido.

Definición 4.7 (Definición 7.1 en [10])**.** Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos finitos contables y X y Y dos variables aleatorias discretas con pmf conjunta p_{XY}, correspondientes a la entrada y salida del canal de comunicación, respectivamente. La matriz de transición de $\mathcal{X}$ a $\mathcal{Y}$ esta dada por los elementos de la pmf condicional $p_{Y|X}(y|x)$. Un canal discreto es un sistema de una única entrada $X = x$, con $x \in \mathcal{X}$ y una única salida $Y = y$, con $y \in \mathcal{Y}$ definido por:

$$p_{XY}(x, y) = p_X(x)p_{X|Y}(x|y), \tag{4.56}$$

para todo $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Previamente se expresó que la señal recibida se puede considerar como una función de la señal de entrada y del ruido. A partir de esta idea, la Definición 4.7 se puede plantear alternativamente de la siguiente forma:

Definición 4.8 (Definición en [10])**.** Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos finitos contables. Sea X una variable aleatoria con alfabeto $\mathcal{X}$ y una matriz de transición de $\mathcal{X}$ a $\mathcal{Y}$ descrita por $p_{Y|X}(y|x)$. Se define una nueva variable aleatoria Z_x con $\mathcal{Z}_x = \mathcal{Y}$ para todo $x \in \mathcal{X}$, tal que:

$$\Pr\{Z_x = y\} = p_{Y|X}(y|x), \tag{4.57}$$

para todo $y \in \mathcal{Y}$. Asumiendo que Z_x, para todo $x \in \mathcal{X}$ son mutuamente independientes y tambien independientes de X, se puede definir la variable aleatoria:

$$Z = \{Z_x : x \in \mathcal{X}\}, \tag{4.58}$$

llamada *ruido variable*. Se observa que Z es independiente de X. Ahora se define una variable aleatoria que toma valores en $\mathcal{Y}$, como:

$$Y = Z_x \quad \text{si} \quad X = x. \tag{4.59}$$

De acuerdo a lo anterior, Y es una función de X y Z. Luego para $x \in \mathcal{X}$ tal que $\Pr\{X = x\} > 0$, se obtiene:

$$\begin{aligned} \Pr\{X = x, Y = y\} &= \Pr\{X = x\} \Pr\{Y = y | X = x\} && (4.60)\\ &= \Pr\{X = x\} \Pr\{Z_x = y | X = x\} && (4.61)\\ &= \Pr\{X = x\} \Pr\{Z_x = y\} && (4.62)\\ &= \Pr\{X = x\} p_{Y|X}(y|x). && (4.63) \end{aligned}$$

Es decir, se obtiene 4.56 de la Definición 4.7, donde (4.61) se obtiene al considerar que Z_x es independiente de X. Suponiendo X y Y como las variables de entrada y salida, se obtiene la definición de canal discreto.

La Definición 4.8 se resume en la siguiente definición:

Definición 4.9 (Definición 7.2 en [10])**.** Sean $\mathcal{X}$, $\mathcal{Y}$ y $\mathcal{Z}$ tres conjuntos contables finitos. Sea la función $\alpha : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$, y Z una variable aleatoria con alfabeto $\mathcal{Z}$, llamada ruido variable. Un canal discreto (α, Z) es un sistema de una única entrada y salida, y alfabeto de entrada $\mathcal{X}$ y alfabeto de salida $\mathcal{Y}$. El ruido Z es independiente de X. Para cualquier variable aleatoria de entrada X, la variable aleatoria de salida Y es definida por:

$$Y = \alpha(X, Z). \tag{4.64}$$

El canal discreto transmite una única entrada. El canal puede ser usado repetidamente, para lo cual se representa el tiempo discreto por el índice i, donde $i \in \mathbb{Z}^+, i \geq 1$, e indica la aplicación o uso i-ésimo del canal. En el caso de estudio, cada uso del canal es independiente del resultado anterior. Si el canal discreto es determinado por sucesivas realizaciones independientes se le llama: canal discreto sin memoria, el cual se define a continuación.

Definición 4.10 (Definición 7.4 en [10])**.** Un canal discreto sin memoria Q, es una secuencia de réplicas de un canal genérico discreto. Estos canales discretos son indexados por un tiempo discreto i, donde $i \in \mathbb{Z}^+, i \geq 1$. La transmisión a través del canal se supone instantánea. Sean X_i y Y_i la entrada y salida del canal en el tiempo i, respectivamente. La función de distribución de probabilidad de la salida dependerá solamente de la entrada en el tiempo i y es condicionalmente independiente de las entradas previas del canal o salidas, expresado matemáticamente como:

$$p_{XY}(Y_i = y, Xi = x) = p_X(X_i = x)p_{Y|X}(y|x). \tag{4.65}$$

4.2.1. Capacidad del Canal

Dado un canal de comunicación discreto sin memoria, el problema a considerar es: ¿cómo garantizar una transmisión de información confiable? Se analiza el siguiente ejemplo, que conduce al concepto que determina la transmisión de información de manera óptima a través de un canal discreto con ruido.

Se supone que hay dos símbolos posibles para la transmisión, 0 y 1, y la velocidad de transmisión de datos es 1000 símbolos por segundo con probabilidades $p_0 = p_1 = \frac{1}{2}$, es decir, la fuente de información está produciendo datos a una velocidad de 1000 bits por segundo. Si el efecto del ruido en este canal, es de tal forma que 1 de cada 10 bits es recibido incorrectamente (un 0 como 1, o un 1 como 0). Esto es 100 bits equívocos de 1000 bits por segundo, o que el 10 % de símbolos recibidos son incorrectos. ¿Cuál es la velocidad de transmisión de información?

1. La velocidad de transmisión es naturalmente menor que 1000 bits por segundo. Si se resta el número esperado de errores por segundo, la velocidad de transmisión se podría considerar como 900 bits por segundo. Pero esto no es correcto, porque el receptor no conoce donde han ocurrido los errores. Es decir, la velocidad de transmisión de información es tal que, en el receptor se puede recuperar todo el contenido de la información transmitida.

2. Ahora, el caso extremo es en el que el ruido es tan grande que los símbolos en el receptor son completamente independientes de los símbolos transmitidos, lo

que corresponde a un nivel de ruido máximo sobre la señal transmitida. Los símbolos 0 y 1 en el receptor tendrán una probabilidad igual a 1/2 independiente del símbolo transmitido. De manera aleatoria, la mitad de los símbolos recibidos serían correctos. Y se afirmaría que: el sistema transmite 500 bits por segundo. En realidad, no se transmite ninguna información porque se obtendría el mismo comportamiento prescindiendo del canal y lanzando una moneda en el punto de recepción.

La pregunta que subyace a este ejemplo es ¿cuánta información se puede comunicar a través de un canal confiablemente? y ¿cuál es la velocidad de transmisión de información? Lo que interesa es encontrar las formas de codificar la señal transmitida de manera que todos los bits que se comunican se recuperen con una probabilidad de error insignificante.

Con respecto a la pregunta del ejemplo: ¿cuál es la velocidad de transmisión de información?, se realiza un análisis de los enunciados 1 y 2. Se considera la entropía en bits por segundo como la medida a la que la fuente produce información. Una forma de encontrar la velocidad de transmisión de información, surge de la siguiente idea: la resta que se considera en la primera parte del problema es un indicio de que la velocidad de transmisión de información de la señal recibida, es menor que la velocidad de transmisión de información transmitida, y la diferencia es la medida de información promedio pérdida que se produce en la transmisión, lo cual también se expresa como la incertidumbre que genera la recepción de una señal, con respecto a lo que realmente ha sido transmitido [1].

A partir de la Definición 2.2 de entropía como medida de incertidumbre, se puede intuir que la entropía condicional de la señal transmitida dada la observación de la señal recibida se relaciona con la incertidumbre o la información pérdida. De esta manera, la velocidad de transmisión de información se obtendría restando de la velocidad de producción de información de la fuente, i.e., la entropía de la fuente[1], la entropía condicional de la fuente de información dada la señal recibida. Esta velocidad corresponde con la Definición 4.6 de información mutua que se obtiene en

[1] O de la entrada a la canal.

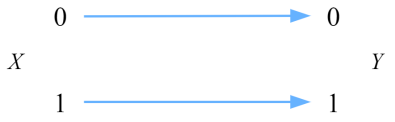


Figura 4.3: Canal Binario sin Ruido.

(4.33) y en (4.34):

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

En el ejemplo que se consideró, la entropía de la fuente es $H(X) = 1$ bit por símbolo o 1000 bits por segundo. Si un 0 es recibido la probabilidad *a posteriori* que un 0 fue transmitido es 0.9 y la probabilidad que un 1 fue transmitido es 0.1. Con estos datos, se va a calcular la entropía condicional y la información mutua.

$$H(X|Y) = -[0.9 \log 0.9 + 0.1 \log 0.1] \tag{4.66}$$

$$= 0.47 \quad \text{bits por símbolo.} \tag{4.67}$$

Esto es también, 470 bits por segundo. Luego, el sistema está transmitiendo a una velocidad de transmisión de datos de:

$$1000 - 470 = 530 \quad \text{bits por segundo.} \tag{4.68}$$

Algunos modelos de canales útiles se presentan a continuación.

Definición 4.11. Canal binario sin ruido: sean $\mathcal{X} = \{0, 1\}$ y $\mathcal{Y} = \{0, 1\}$ los alfabetos de las variables aleatorias X y Y, respectivamente. en este canal la entrada es duplicada en la salida; tal como se observa en la Fig. 4.3. El canal está definido por la matriz de transición:

$$p_{Y|X}(y|x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.69}$$

Definición 4.12. Canal binario simétrico (BSC, *Binary Simmetric Channel*): sean $\mathcal{X} = \{0, 1\}$ y $\mathcal{Y} = \{0, 1\}$ los alfabetos de las variables aleatorias X y Y, respectivamente. Sea $\eta > 0$ el parámetro que corresponde a la probabilidad de error. El canal

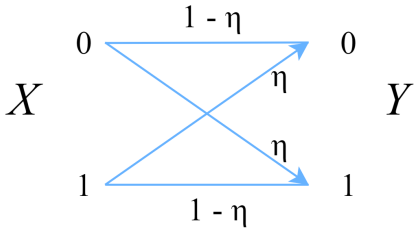


Figura 4.4: Canal Binario Simétrico (BSC).

BSC con entrada X y salida Y esta definido por la matriz de transición:

$$p_{Y|X}(y|x) = \begin{pmatrix} 1-\eta & \eta \\ \eta & 1-\eta \end{pmatrix}. \tag{4.70}$$

El canal BSC está representado en el diagrama de transición de la Fig. 4.4.

La información mutua del canal BSC, se obtiene de (4.34):

$$I(X;Y) = H(Y) - H(Y|X) \tag{4.71}$$

$$= H(Y) - \sum_{x \in \mathcal{S}_X} p_X(x) H(Y|x). \tag{4.72}$$

Y dado que:

$$H(Y|x) = -\sum_{y \in \mathcal{S}_Y} p_{Y|X}(y|x) \log p_{Y|X}(y|x) \tag{4.73}$$

$$= -[p_{Y|X}(0|x) \log p_{Y|X}(0|x) + p_{Y|X}(1|x) \log p_{Y|X}(1|x)]. \tag{4.74}$$

$H(Y|x)$ para $x \in \mathcal{X}$ se denota como $H_2(\eta)$ debido a que su forma coincide con la entropía binaria definida en (2.5), con probabilidad $p = \eta$, esto es:

$$H_2(\eta) = -[\eta \log \eta + (1-\eta) \log(1-\eta)] \tag{4.75}$$

Así, la información mutua es:

$$I(X;Y) = H(Y) - \sum_{x \in \mathcal{S}_X} p_X(x) H_2(\eta) \tag{4.76}$$

$$= H(Y) - H_2(\eta) \tag{4.77}$$

$$\leq 1 - H_2(\eta), \tag{4.78}$$

donde $H(Y) = 1$ si la pmf de Y es uniforme (máximo valor de entropía o de incertidumbre), lo cual se cumple si p_X es uniforme, es decir, $p_X(0) = p_X(1) = 1/2$ en el canal BSC.

Por lo tanto, la cota superior de (4.78) denotada por C para del canal BSC es:

$$C = 1 - H_2(\eta) \quad \text{bits por símbolo.} \tag{4.79}$$

La capacidad C del canal BSC para $0 \leq \eta \leq 1$ se grafica en la Fig. 4.5; se observa que C alcanza su máximo valor cuando $\eta = 0$ y $\eta = 1$, y el valor mínimo se obtiene cuando $\eta = 0.5$. Cuando $\eta = 0$, la capacidad es la velocidad de transmisión de datos máxima a través del canal, i.e., que los datos pueden ser comunicados de manera confiable, porque no hay lugar a error o a pérdida de información. Cuando $\eta = 1$, lo mismo puede ser obtenido, realizando un decodificación que invierta los bits recibidos. Por lo tanto, los canales binarios permiten recuperar confiablemente los bits transmitidos si la probabilidad de error es nula o si la probabilidad de error es uno. Cuando $\eta = 0.5$, la salida del canal es independiente de la entrada del canal, y debido a esto, ninguna información puede ser comunicada a través del canal [10]. A partir de este análisis, se presenta la siguiente definición.

Definición 4.13. La capacidad de una canal discreto sin memoria con probabilidad de error aleatoria y cercana a cero, se define como:

$$C \triangleq \max_{p_X(x)} I(X;Y), \tag{4.80}$$

donde, X y Y son la entrada y la salida respectivamente del canal discreto, y el máximo es tomado sobre todas las distribuciones de entrada p_X. La capacidad se mide en bits por símbolo de entrada al canal.

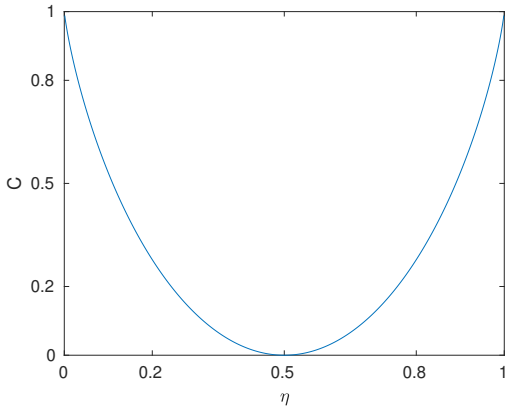


Figura 4.5: Capacidad del Canal BSC.

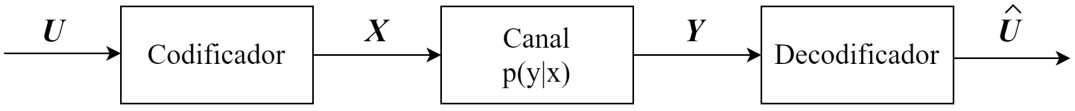


Figura 4.6: Codificación de Canal

La capacidad del canal es la velocidad de transmisión de datos, la cual se obtiene al seleccionar la distribución de entrada al canal que maximiza la información mutua. Su unidad de medida, se expresa también en *bits por uso de canal*, es decir, el número promedio de bits que se intentan transmitir a través del canal por uso de canal.

4.3. CODIFICACIÓN DE CANAL

La codificación de canal se enfoca en el problema de transmitir confiablemente un mensaje que es perturbado durante la transmisión por ruido. La capacidad de canal es la métrica fundamental asociada al canal de comunicación, la cual representa, la velocidad de transmisión de datos máxima con una probabilidad de error arbitrariamente pequeña.

En la Fig. 4.6 se presenta el esquema de codificación de canal en el cual, la entrada del canal es modelada por el vector $\mathbf{X}$ de variables aleatorias discretas, y la salida del canal por el vector $\mathbf{Y}$, con variables aleatorias genéricas X y Y, respectivamente: sea $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ el conjunto de posibles mensajes, y U la variable aleatoria

discreta que representa la selección de un mensaje, por lo tanto $\mathbf{X}$ es el resultado de la codificación del mensaje a transmitir u, y de $\mathbf{Y}$ se obtiene la estimación del mensaje transmitido, denotado por $\hat{u}$, lo cual se formaliza en la Definición 4.14.

Definición 4.14. El código de canal (n, K) para un canal sin memoria discreto y asociado a la Fig. 4.6, caracterizado por $(\mathcal{X}, p_{Y|X}(y|x), \mathcal{Y})$, está definido por:

1. Un conjunto de mensajes $\mathcal{U} = \{1, 2, \ldots, 2^K\}$.
2. Una función de codificación $f : \mathcal{U} \to \mathcal{X}^n$, la cual genera palabras código (*codewords*) $f(1), f(2), \ldots, f(2^K)$ en $\mathcal{X}^n$, de longitud n. El conjunto de palabras código es llamado el diccionario (*codebook*).
3. Una función de decodificación $g : \mathcal{Y}^n \to \hat{\mathcal{U}}$, donde $\hat{\mathcal{U}} = \{0, 1, 2, \ldots, 2^K\}$ es la estimación que se obtiene del conjunto de mensajes $\mathcal{U}$. El elemento 0 se utiliza para indicar un error.

La cardinalidad del conjunto de mensajes, corresponde al número de palabras código $|\mathcal{U}| = 2^K$, el cual es un número entero. El número de bits que especifican cada palabra código se obtiene de aplicar la función logaritmo en base dos, i.e., $K = \log_2 |\mathcal{U}|$, lo cual no es necesariamente un número entero[2].

Ahora, es necesario incluir una medida probabilística del error producido cuando el mensaje u no coincide con la estimación $\hat{u}$.

Definición 4.15. Sea λ_u la probabilidad condicional de error, de $\hat{u}$ dado que el mensaje transmitido es u, definida como:

$$\lambda_u \triangleq \Pr\{\hat{U} \neq u | U = u\} = \sum_{\mathbf{y} \in \mathcal{Y}^N : g(\mathbf{y}) \neq u} \Pr\{\mathbf{Y} = \mathbf{y} | \mathbf{X} = f(u)\}. \tag{4.81}$$

A continuación se presentan las medidas de desempeño de un código de canal.

[2]Por sencillez, se ha seleccionado a lo largo del documento una codificación binaria, por lo cual utiliza la función logaritmo en base dos, en el caso general, para una codificación $d-$aria, simplemente se debe realizar el cambio de base, aplicando el logaritmo en base d.

Definición 4.16. La máxima probabilidad de error de un código (n, K) es definida como:

$$\lambda_{max} \triangleq \max \lambda_u, \tag{4.82}$$

para $u \in \mathcal{U}$.

Definición 4.17. La probabilidad de error promedio de un código (n, K) está definida como:

$$P_e \triangleq \Pr\{g(\mathbf{Y}) \neq U\} = \Pr\{\hat{U} \neq U\}. \tag{4.83}$$

De la Definición 4.17 se obtiene lo siguiente:

$$\begin{aligned} P_e &= \Pr\{\hat{U} \neq U\} && (4.84)\\ &= \sum_{u \in \mathcal{U}} \Pr\{U = u\} \Pr\{\hat{U} \neq U | U = u\} && (4.85)\\ &= \sum_{u \in \mathcal{U}} \frac{1}{2^K} \Pr\{\hat{U} \neq u | U = u\} && (4.86)\\ &= \frac{1}{2^K} \sum_{u \in \mathcal{U}} \lambda_u, && (4.87) \end{aligned}$$

donde se considera que la pmf de U es uniforme, i.e., $p_U(u) = \Pr\{U = u\} = \frac{1}{2^K}$. Por lo tanto, P_e es la media aritmética de λ_u, y se cumple que:

$$P_e \leq \lambda_{max}. \tag{4.88}$$

Definición 4.18. La tasa de codificación[3] R de un código de canal (n, K) es definida como:

$$R \triangleq \frac{\log 2^K}{n} = \frac{K}{n} \text{ bits por uso de canal.} \tag{4.89}$$

Definición 4.19. La tasa de codificación R se dice que es asintóticamente alcanzable (*achievable*) para un canal sin memoria discreto, si para cualquier $\epsilon > 0$ existe para un n suficientemente grande un código de canal (n, K), tal que:

$$\frac{\log 2^K}{n} > R - \epsilon \tag{4.90}$$

[3]Velocidad de tranmisión de datos.

y

$$\lambda_{max} < \epsilon. \tag{4.91}$$

Por sencillez, asintóticamente alcanzable se escribirá como una tasa de codificación R alcanzable. Una tasa de codificación R es alcanzable si existe una secuencia de códigos cuyas tasas de codificación se aproximan a R y cuyas probabilidades de error se aproximan a cero. El teorema de codificación de canal da una caracterización de las tasas de codificación alcanzables [10].

A continuación se presenta el teorema de codificación de canal, el cual es el resultado fundamental de la teoría de la información. Aquí se estudia el caso discreto, aunque el resultado general más conocido es en el caso continuo. Este teorema en este trabajo no es un punto de partida sino un punto de llegada en la comprensión de la teoría de la información, que se encuentra constituida y sintetizada en los teoremas de codificación de fuente y codificación de canal.

4.4. TEOREMA FUNDAMENTAL PARA EL CANAL CON RUIDO DISCRETO

Teorema 4.4 (Teorema 11 en [1]). *Sea un canal discreto con capacidad C y una fuente discreta con entropía H en bits por segundo. Si $H \leq C$ existe un sistema de codificación para el cual, la salida de la fuente puede ser transmitida sobre el canal con una probabilidad de error arbitrariamente pequeña (o una equivocación arbitrariamente pequeña). Si $H > C$ es posible codificar la fuente de tal forma que, la equivocación sea menor que $H - C + \epsilon$, donde ϵ es arbitrariamente pequeño. En esta última opción, no existe un método de codificación del cual se obtenga una equivocación menor que $H - C$.*

El Teorema 4.4 es conocido como el teorema de codificación de canal de un sistema de comunicación con ruido discreto. El significado del teorema se centra en el límite fundamental que impone la capacidad de canal, o velocidad de transmisión de datos C, la cual permite establecer en un sistema de comunicación con ruido, la transmisión

de información a través del canal discreto con una probabilidad de error tan pequeña como se quiera, realizando una codificación adecuada de los mensajes. Se cumple además, que lo anterior no es posible para cualquier velocidad de tranmisión de datos mayor a C. Ahora, con el fin de presentar de forma más explícita el teorema se reescribe en el Teorema 4.5.

Teorema 4.5. *Para un canal sin memoria discreto, todas las tasas de codificación de canal que se encuentran por debajo de la capacidad son alcanzables (achievable). Específicamente, para cada $R < C$, existe una secuencia de códigos (n, K) con una probabilidad de error máxima $\lambda_{max} \to 0$. Inversamente, cualquier secuencia de códigos (n, K) con $\lambda_{max} \to 0$ debe tener $R \leq C$.*

Shannon en [1] probó el Teorema 4.4 utilizando los conceptos de tipicalidad, y de prueba de existencia matemática, pero no de forma rigurosa. Después de 1948, la formalización matemática de la prueba fue desarrollada incialmente por A. Feinstein [59] de la parte directa y por R. Fano [43] de la parte inversa del teorema. La prueba que aquí se presenta de la parte directa (*achievability*) fue desarrollada en [8] y [9] y de la parte inversa (*converse*) en [10]. Se seleccionaron estos autores, debido a la sencillez y alcance de los argumentos matemáticos que utilizan.

4.4.1. Prueba Directa

La parte directa del teorema afirma que existe un código de canal que alcanza la tasa de codificación máxima, i.e., la capacidad de canal, para la cual, necesariamente se cumple que la probabilidad de error es arbitrariamente pequeña y cercana a cero. La prueba consiste en asumir que tal código de canal existe en un grupo de códigos, y se encuentra la probabilidad de error promedio sobre este grupo. Luego, se prueba que la probabilidad de error promedio puede ser menor que un ϵ, y se utiliza el siguiente argumento: si el promedio de un conjunto de números es menor que ϵ, debe existir por lo menos uno en el conjunto que sea menor que ϵ [1]. Lo cual se puede observar de manera intuitiva en la siguiente analogía:

En una guardería de bebés, hay un grupo de 100 bebés y se desea probar que hay un bebé con un peso menor a 10 Kg. Un intento de resolver el problema es tomar el

peso de cada bebé, lo cual puede ser complejo. El método de Shannon para resolver este problema es tomar a todos los bebés, y pesarlos sobre una gran báscula. Si se encuentra que su peso promedio es menor a 10 Kg, debe existir por lo menos un bebé que pesa menos de 10 Kg, aunque en realidad deben de ser más. El método de Shannon no garantiza revelar la existencia de cual niño esta bajo de peso, sino que existe uno. Por ejemplo, si se obtiene un peso grupal menor que 1000 Kg, entonces se puede afirmar de que existe un niño con un peso menor a 10 Kg [8].

De esta manera, Shannon calcula la probabilidad promedio de error de un grupo de códigos de canal con una tasa de codificación de información cercana a $I(X;Y)$ y encuentra que es pequeña, por lo tanto, deben existir códigos de canal para los cuales la probabilidad de error es muy pequeña.

Prueba: (Prueba en [8] y [9]). Se construye un sistema de codificación y decodificación, para un canal discreto sin memoria $p_{Y|X}(y|x)$, donde la entrada y salida del canal están descritas por los vectores de variables aleatorias $\mathbf{X}$ y $\mathbf{Y}$, respectivamente, y las pmf genéricas de entrada $p_X(x)$ y de salida $p_Y(y)$. Matemáticamente lo que se quiere mostrar es:

1. Para cada distribución de entrada $p_X(x)$ se demuestra la existencia de un código de canal, para un n suficientemente grande, tal que la velocidad de tranmisión de datos $I(X;Y)$ es alcanzable, o:
   - *a*) La tasa de codificación del código de canal es arbitrariamente cercana a $I(X;Y)$.
   - *b*) La probabilidad de error máxima λ_{max} es arbitrariamente pequeña y cercana a cero.
2. Luego, se elige la distribución de entrada $p_X(x)$ tal que maximice la información mutua, i.e., $\max I(X;Y) = C$, lo cual permite concluir que la velocidad de transmisión de datos C es alcanzable.

4.4.1.1. Codificación Aleatoria y Decodificación Típica Conjunta

Se propone el siguiente sistema de codificación y decodificación, con tasa de codificación R':

1. La entrada del canal es descrita por el vector $\mathbf{X}$ de variables aleatorias discretas i.i.d. con dimensión n, distribución de entrada genérica $p_X(x)$ y pmf conjunta, tal como se muestra a continuación:

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_X(x_i), \tag{4.92}$$

   y se define un código de canal aleatorio $(n, nR') = (n, K)$ denotado por $\mathcal{C}$ con tasa de codificación R', que produce $2^{nR'} = 2^K$ palabras código. Las 2^K palabras código de longitud n denotadas por $\mathbf{x}^{(s)} \in \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(2^K)}\}$ con $s \in \{1, 2, \ldots, 2^K\}$, se escriben como las filas de una matriz, tal como se muestra a continuación:

$$\mathcal{C} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ \vdots & \vdots & & \vdots \\ x_1^{(2^K)} & x_2^{(2^K)} & \cdots & x_n^{(2^K)} \end{bmatrix}. \tag{4.93}$$

   Cada entrada en esta matriz es generada i.i.d. de acuerdo a $p_X(x)$. Así, la probabilidad de generar un código particular es:

$$\Pr\{\mathcal{C}\} = \prod_{s=1}^{2^K} \prod_{i=1}^{n} p_X(x_i)^{(s)}. \tag{4.94}$$

2. El código es conocido tanto por el codificador como el decodificador.

3. Un mensaje s es seleccionado de acuerdo a una distribución uniforme, i.e., $\Pr\{s\} = 2^{-K}$, y se transmite la palabra código asociada, $\mathbf{x}^{(s)}$.

4. La secuencia de entrada al canal, $\mathbf{x} = \mathbf{x}^{(s)}$, se transmite a través del canal.

5. La secuencia recibida $\mathbf{y}$ es una realización del vector $\mathbf{Y}$ de variables aleatorias

discretas i.i.d., la cual describe la salida del canal con pmf condicional[4], como se muestra a continuación:

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(s)}) = \prod_{i=1}^{N} p(y_i|x_i^{(s)}). \tag{4.95}$$

6. El receptor decodifica la secuencia recibida $\mathbf{y}$, utilizando la decodificación típica conjunta[5], la cual hace que la secuencia $\mathbf{y}$ sea mapeada al mensaje estimado $\hat{s}$ si:

$$(\mathbf{x}^{(\hat{s})}, \mathbf{y}) \in \mathcal{T}_{XY}^{n,\epsilon}, \tag{4.96}$$

lo cual significa que $(\mathbf{x}^{(\hat{s})}, \mathbf{y})$ son secuencias típicas conjuntas, y que no existe $s' \neq s$ tal que,

$$(\mathbf{x}^{(\hat{s})}, \mathbf{y}) \in \mathcal{T}_{XY}^{n,\epsilon}, \tag{4.97}$$

lo cual a su vez significa que no existe otro $\hat{s}$ tal que sean secuencias típicas conjuntas. De lo contrario, si existe tal s' o si hay más de uno, se afirma que la estimación es errada, y $\mathbf{y}$ es decodificada en un mensaje constante $\hat{s} = 0$.

7. En la decodificación hay un error si $\hat{s} \neq s$. El evento de error, se define como $\mathcal{E} : \{\hat{s} \neq s\}$.

4.4.1.2. Análisis de la Probabilidad de Error

Dado que se utiliza la decodificación típica conjunta se encuentra que existen dos fuentes de error en este tipo de decodificación, las cuales son:

1. La salida del canal $\mathbf{y}$ no es típica conjunta con la palabra código transmitida $\mathbf{x}^{(s)}$.

2. Existen otras palabras código en $\mathcal{C}$ que son típicamente conjuntas con $\mathbf{y}$.

[4]La pmf condicional se obtiene al aplicar la independencia condicional, la cual es soportada por la característica del canal sin memoria.

[5]La decodificación típica conjunta es más fácil de analizar, aunque no es un algortimo óptimo de decodificación. A diferencia de la decodificación de máxima probabilidad, que por otro lado, es difícil de analizar [9].

Sea $\mathcal{E} = \{\hat{s} \neq s\}$ que denota el evento de error en el receptor. Dado el código de canal $\mathcal{C}$, se calcula la probabilidad de error promedio, sobre todas las palabras código de $\mathcal{C}$ y sobre un grupo de diccionarios, esto es:

$$\Pr\{\mathcal{E}\} = \sum_{\mathcal{C}} \Pr\{\mathcal{C}\} P_e\{\mathcal{C}\}, \tag{4.98}$$

lo cual es la suma de los productos de la probabilidad de error promedio de cada código P_e en la Definición 4.17, y la probabilidad del código $\Pr\{\mathcal{C}\}$.

El desarrollo siguiente consiste en encontrar, primero la probabilidad de error promedio $\Pr\{\mathcal{E}\}$ y mostrar que puede hacerse este valor más pequeño que un número deseado. Lo que permite deducir que existe al menos un código $\mathcal{C}$ cuya probabilidad de error sea también menor que este número pequeño.

Aplicando (4.87) en (4.98) y considerando que el conjunto de mensajes $\mathcal{S}$ tiene asociada una pmf uniforme, se obtiene lo siguiente:

$$\begin{aligned} \Pr\{\mathcal{E}\} &= \sum_{\mathcal{C}} \Pr\{\mathcal{C}\} P_e\{\mathcal{C}\} \\ &= \sum_{\mathcal{C}} \Pr\{\mathcal{C}\} \frac{1}{2^K} \sum_{s=1}^{2^K} \lambda_s(\mathcal{C}) \qquad (4.99) \\ &= \frac{1}{2^K} \sum_{s=1}^{2^K} \sum_{\mathcal{C}} \Pr\{\mathcal{C}\} \lambda_s(\mathcal{C}) \qquad (4.100) \end{aligned}$$

Por la símetria de la construcción del código, dado por la independencia de cada mensaje transmitido s, la probabilidad de error promedio sobre un grupo de códigos no depende del valor particular de s[6], de tal forma que, se puede tomar sin pérdida de generalidad que $s = 1$. De esta manera, la palabra código transmitida es $\mathbf{x}^{(1)}$, y

[6]La probabilidad de error de un mensaje s no depende de que s sea seleccionado, por lo tanto, se puede calcular para un valor arbitrario.

de (4.100) se obtiene lo siguiente:

$$\Pr\{\mathcal{E}\} = \frac{1}{2^K}\sum_{s=1}^{2^K}\sum_{\mathcal{C}}\Pr\{\mathcal{C}\}\lambda_s(\mathcal{C}) \tag{4.101}$$

$$= \Pr\{\mathcal{C}\}\lambda_1(\mathcal{C}) \tag{4.102}$$

$$= \Pr\{\mathcal{E}|s=1\}. \tag{4.103}$$

Se define el evento E_i en el cual la palabra código $\mathbf{x}^{(i)}$ y $\mathbf{y}$ son típicas conjuntas, donde $\mathbf{y}$ es el resultado de transmitir $\mathbf{x}^{(i)}$ a través del canal, lo cual se muestra a continuación:

$$E_i = \{(\mathbf{x}^{(i)}, \mathbf{y}) \in \mathcal{T}_{XY}^{n,\epsilon}\},\ i \in \{1, 2, \ldots, 2^K\}. \tag{4.104}$$

Dado el esquema de decodificación típica conjunta se genera un error, si E_1^c ocurre, i.e., el complemento de E_1, lo cual significa que las secuencias $\mathbf{x}^{(1)}$ y $\mathbf{y}$ no son típicas conjuntas; o si $E_2 \cup E_3 \cup \ldots \cup E_{2^K}$ ocurre, i.e., una o más palabras código erradas son típicas conjuntas con $\mathbf{y}$. Por lo tanto, se obtiene la siguiente definición equivalente para la probabilidad de error promedio de (4.103):

$$\Pr\{\mathcal{E}|s=1\} = \Pr\{E_1^c \cup E_2 \cup E_3 \cup \ldots \cup E_{2^K}|s=1\} \tag{4.105}$$

$$\leq \Pr\{E_1^c|s=1\} + \sum_{s=2}^{2^K}\Pr\{E_i|s=1\}, \tag{4.106}$$

donde la desigualdad en (4.106) se obtiene de la probabilidad de unión de eventos. Ahora, se debe considerar lo siguiente:

1. La probabilidad de que la entrada $\mathbf{x}^{(1)}$ y la salida $\mathbf{y}$ no sean típicas conjuntas se desvanece a medida que n sea suficientemente grande, por la segunda parte del Lema 3.2. Sea ϵ la cota superior de esta probabilidad, tal que $\epsilon \to 0$ cuando $n \to \infty$. Para cualquier ϵ se puede encontrar una longitud de palabra código $n(\epsilon)$ tal que,

$$\Pr\left\{E_1^c|s=1\right\} = \Pr\left\{\left(\mathbf{x}^{(1)}, \mathbf{y}\right) \notin \mathcal{T}_{XY}^{N,\epsilon}\right\} \leq \epsilon. \tag{4.107}$$

2. La probabilidad de que $\mathbf{x}^{(s')}$ y $\mathbf{y}$ sean típicas conjuntas, donde $s' \neq 1$ se obtiene aplicando las propiedades presentadas en Lema 3.2, y encontrando la probabilidad de que dos secuencias independientes $\mathbf{x}'$ y $\mathbf{y}'$ sean típicas conjuntas, esto es:

$$\begin{aligned}
\Pr\left\{(\mathbf{x}', \mathbf{y}') \in \mathcal{T}_{XY}^{n,\epsilon}\right\} &= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}_{XY}^{n,\epsilon}} p_{\mathbf{X}}(\mathbf{x}) p_{\mathbf{Y}}(\mathbf{y}) && (4.108)\\
&\leq |\mathcal{T}_{XY}^{n,\epsilon}| 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} && (4.109)\\
&\leq 2^{n(H(X,Y)-\epsilon)} 2^{-n(H(X)-\epsilon)-n(H(Y)-\epsilon)} && (4.110)\\
&= 2^{-n(H(x)+H(Y)-H(X,Y)-3\epsilon} && (4.111)\\
&= 2^{-n(I(X;Y)-3\epsilon)}. && (4.112)
\end{aligned}$$

Por lo tanto, la probabilidad de que $\mathbf{x}^{(i)}$ para $i \neq 1$ y $\mathbf{y}$ sean típicas conjuntas es menor o igual a $2^{-n(I(X;Y)-3\epsilon)}$.

Luego, aplicando (4.107), (4.112) en la probabilidad de error promedio de (4.106) y con $K = nR'$, se obtiene lo siguiente:

$$\begin{aligned}
\Pr\{\mathcal{E}\} &\leq \epsilon + \sum_{s=2}^{2^K} 2^{-n(I(X;Y)-3\epsilon)} && (4.113)\\
&= \epsilon + (2^K - 1) 2^{-n(I(X;Y)-3\epsilon)} && (4.114)\\
&\leq \epsilon + 2^{-n(I(X;Y)-R'-3\epsilon)} && (4.115)\\
&\leq 2\epsilon, && (4.116)
\end{aligned}$$

para, n suficientemente grande y $R' < I(X;Y) - 3\epsilon$. Por lo tanto, es posible seleccionar ϵ y n, de tal forma que la probabilidad de error promedio sobre todas las palabras código y diccionarios sea menor o igual a 2ϵ.

La prueba se completa con el siguiente análisis respecto a la selección de un código, para el cual se realizan las siguientes modificaciones:

1. Se selecciona la pmf p_X tal que, sea la distribución de entrada al canal óptima. Por lo tanto, si $R' < I(X;Y) - 3\epsilon$ se convierte en $R' < C - 3\epsilon$ o $R' < C$.

2. Ya que la probabilidad de error promedio sobre el grupo de códigos de canal es $\Pr\{\mathcal{E}\} \leq 2\epsilon$, debe existir por lo menos un código con probabilidad de error promedio $\Pr\{\mathcal{E}|\mathcal{C}\} = P_e(\mathcal{C}) < 2\epsilon$.

3. Para mostrar que no sólo la probabilidad de error promedio es pequeña, sino también la probabilidad máxima de error λ_{max}, se modifica el código descartando la peor mitad de las palabras código[7], las que tienen mayor probabilidad de producir errores. Las que quedan deben tener una probabilidad condicional de error menor que 4ϵ. Las palabras código restantes forman un nuevo código. Este nuevo código tiene $2^{nR'-1}$ palabras código, es decir, se ha reducido la velocidad de transmisión de información inicial de R' a $R' - \frac{1}{n}$ (una reducción insignificante si n es grande), y se alcanza una $\lambda_{max} < 4\epsilon$. El código resultante puede no ser el mejor código en velocidad de transmisión de datos y longitud de palabra código, pero aún es lo suficientemente bueno para demostrar el teorema de codificación de canal con ruido [8].

En conclusión, se ha diseñado un código de canal, con una velocidad de transmisión de datos $R' - \frac{1}{n}$, donde $R' < C - 3\epsilon$ y probabilidad de error máxima es menor a 4ϵ, lo cual prueba que cualquier velocidad de transmisión de datos puede ser lograble si es menor al valor de capacidad.

Esto prueba la parte directa del teorema. ■

4.4.2. Prueba Inversa

La parte inversa del teorema de codificación de fuente, establece que si R es una tasa de codificación alcanzable, se cumple que, para un código de canal (n, K), con n suficientemente grande y probabilidad de error máxima arbitrariamente pequeña y cercana a cero se debe satisfacer, que $R \leq C$.

En el desarrollo de esta prueba es importante definir el esquema de codificación de canal como una cadena de Markov. Las cadenas de Markov se presentan en el

[7]Este truco, se conoce como *expurgación*.

Apéndice A. Se considera un código de canal como se presenta en la Fig. 4.6 con palabras código de longitud n, formado por: U la variable aleatoria que describe la selección de un mensaje, X_i y Y_i para $1 \leq i \leq n$, la entrada y la salida en el tiempo i del canal, respectivamente, y $\hat{U}$ que corresponde al mensaje estimado. Debido a que estas variables aleatorias son generadas secuencialmente en el canal discreto, de acuerdo a alguna regla determinista[8] o probabilística[9] en el siguiente orden $U, X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n, \hat{U}$ o lo que es igual, la sucesión $U, \mathbf{X}, \mathbf{Y}, \hat{U}$, forman una cadena de Markov, lo cual se denota de la siguiente forma:

$$U \to \mathbf{X} \to \mathbf{Y} \to \hat{U}, \tag{4.117}$$

donde, la dependencia de $\mathbf{x}$ con respecto al mensaje u y $\mathbf{y}$ con respecto a $\hat{u}$, es determinista. La dependencia de $\mathbf{X}$ y $\mathbf{Y}$ es aleatoria. Así se cumple, que para todo $(u, \mathbf{x}, \mathbf{y}, \hat{u}) \in \mathcal{U} \times \mathcal{X}^n \times \mathcal{Y}^n \times \hat{\mathcal{U}}$ tal que $p_{\mathbf{X}}(\mathbf{x}) > 0$ y $p_{\mathbf{Y}}(\mathbf{y}) > 0$, lo siguiente:

$$q(u, \mathbf{x}, \mathbf{y}, \hat{u}) = p_U(u) \left(\prod_{i=1}^{n} p_{UX}(x_i|u) \right) \left(\prod_{i=1}^{n} p_{XY}(y_i|x_i) \right) p_{Y\hat{U}}(\hat{u}|\mathbf{y}) \tag{4.118}$$

$$= p_U(u) p_{UX}(\mathbf{x}|u) p_{XY}(\mathbf{y}|\mathbf{x}) p_{Y\hat{U}}(\hat{u}|\mathbf{y}). \tag{4.119}$$

donde $q = p_{UXY\hat{U}}$ denota la pmf conjunta y se ha aplicado independencia condicional. Se considera además el siguiente Lema:

Lema 4.6 (Lema 7.16 en [10])**.** *Sean* $\mathbf{X}$ *y* $\mathbf{Y}$ *los vectores de variables aleatorias a la entrada y salida del canal, respectivamente, ambos de dimensión* n*. Para un canal discreto sin memoria con un código de canal sin realimentación*[10]*, para cualquier* $n \geq 1$ *se cumple:*

$$I(\mathbf{X}; \mathbf{Y}) \leq nC. \tag{4.120}$$

Prueba: Para cualquier $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$, si $p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) > 0$, $p_{\mathbf{X}}(\mathbf{x}) > 0$, y la pmf

[8]Cuando la entrada del canal, no es afectada por ruido y de manera directa el resultado en la salida del canal determina completamente la entrada, y por lo tanto, el mensaje enviado.

[9]Cuando la entrada al canal y la salida del canal, son afectadas por una matriz de transición, en la cual se incluye la probabilidad de error o fallo producido por el ruido.

[10]La realimentación hace referencia a que el transmisor recibe una confirmación de parte del receptor del mensaje enviado, lo cual, permite que el transmisor evalue su proxima selección i.e., X_{i+1} de acuerdo, a los valores previos recibidos $Y_1, Y_2, \ldots, Y_i$. Sin embargo, lo que se utiliza es un canal, en el que su salida Y_i depende únicamente de la entrada X_i en el tiempo i.

condicional:

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p(y_i|x_i). \tag{4.121}$$

El valor esperado de (4.121) se obtiene como se muestra a continuación:

$$-\mathbb{E}\big[\log p_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})\big] = -\mathbb{E}\left[\log \prod_{i=1}^{n} p(Y_i|X_i)\right] = -\sum_{i=1}^{n} \mathbb{E}\big[\log p(Y_i|X_i)\big], \tag{4.122}$$

lo cual es por (4.11), la entropía condicional de la salida $\mathbf{Y}$ con respecto a la observación de la entrada $\mathbf{X}$, i.e., $H(\mathbf{Y}|\mathbf{X}) = \sum_{i=1}^{n} H(Y_i|X_i)$. Luego, por la Definición 4.6 aplicada a los vectores $\mathbf{X}$ y $\mathbf{Y}$, se obtiene lo siguiente:

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) && (4.123)\\ &= H(\mathbf{Y}) - \sum_{i=1}^{n} H(Y_i|X_i) && (4.124)\\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) && (4.125)\\ &= \sum_{i=1}^{n} I(X_i;Y_i) && (4.126)\\ &\leq nC && (4.127) \end{aligned}$$

donde, (4.124) se obtiene de (4.122), lo cual también se puede obtener de la definición de canal discreto considerando el canal sin memoria; y (4.125) es debido a que la entropía de un conjunto de variables aleatorias es menor que la suma de las entropías individuales que se establece generalizando (B.23); y (4.127) se obtiene de la Definición 4.13. Lo cual completa la prueba del lema. ■

Un primer acercamiento a la prueba de la parte inversa del teorema se alcanza, al considerar una transmisión de información a través de un canal discreto, para el cual la probabilidad de error del código de canal sea nula, entonces se obtiene que la tasa de codificación R, es siempre menor que C, i.e., $R \leq C$.

Sea el código (n, K) con $K = nR$ y la probabilidad de error igual a cero, de modo que, el mensaje de entrada u es determinado por la secuencia recibida $\mathbf{Y}$, i.e., $H(U|\mathbf{Y}) = 0$. Se asume que $\mathcal{U}$ tiene una pmf uniforme, i.e., $p_U(u) = \frac{1}{2^K}$ y por lo

tanto, $H(U) = K = nR$, y se obtiene lo siguiente:

$$nR = H(U) = H(U|\mathbf{Y}) + I(U;\mathbf{Y}) \tag{4.128}$$
$$= I(U;\mathbf{Y}) \tag{4.129}$$
$$\leq I(\mathbf{X},\mathbf{Y}) \tag{4.130}$$
$$\leq nC \tag{4.131}$$

donde, (4.128) es debido a (4.33); (4.130) se obtiene de la desigualdad de procesamiento de datos (D.34); y (4.131) se obtiene de aplicar el Lema 4.6. De esta forma, para cualquier código (n, K) con probabilidad de error nula, se cumple para todo $n \geq 1$ que, $R \leq C$, siendo C el límite fundamental para la velocidad de transmisión de datos.

Ahora, se presenta la prueba extendida a códigos de canal con probabilidad de error arbitrariamente pequeña. Aquí se utilizaran dos desigualdades, la desigualdad de procesamiento de datos y la desigualdad de Fano, las cuales se presentan en el Apéndice D.

Prueba: Para cualquier código de canal $(n, K) = (n, nR)$ la probabilidad de error máxima tiende a cero, i.e., $\lambda_{max} \to 0$ de la Definición 4.16, lo cual implica que la probabilidad de error promedio también tiende a cero, i.e., $P_e \to 0$. El esquema de codificación en 4.14 permite definir la siguiente cadena de Markov:

$$U \to \mathbf{X}(u) \to \mathbf{Y} \to \hat{U}, \tag{4.132}$$

para cualquier n. El conjunto de mensajes indexados por $u \in \mathcal{U}$ tiene una distribución uniforme, por lo tanto (4.87) se cumple, y se obtiene lo siguiente:

$$nR = H(U) \tag{4.133}$$
$$= H(U|\hat{U}) + I(U;\hat{U}) \tag{4.134}$$
$$\leq 1 + P_e nR + I(U;\hat{U}) \tag{4.135}$$
$$\leq 1 + P_e nR + I(\mathbf{X};\mathbf{Y}) \tag{4.136}$$
$$\leq 1 + P_e nR + nC, \tag{4.137}$$

donde, (4.133) se obtiene debido a que U tiene una pmf uniforme; (4.135) se obtiene de aplicar la desigualdad de Fano descrita en (D.42); (4.136) se obtiene de aplicar la desigualdad de procesamiento de datos definida en (D.34); y (4.137) se obtiene de aplicar el Lema 4.6.

Dividiendo (4.137) entre n, se obtiene lo siguiente:

$$R \leq \frac{1}{n} + P_e R + C. \tag{4.138}$$

Para n suficientemente grande, $n \to \infty$, los dos primeros términos a la derecha de la desigualdad (4.138) tienden a cero, y por lo tanto, se concluye que:

$$R \leq C. \tag{4.139}$$

Se puede reescribir (4.138), de la siguiente forma:

$$P_e \geq 1 - \frac{C}{R} - \frac{1}{nR}, \tag{4.140}$$

por lo tanto, esta expresión demuestra que si $R > C$, la probabilidad de error será próxima a uno, o estará lejos de ser pequeña. Por lo tanto, no es posible alcanzar una probabilidad de error arbitrarimente pequeña a tasas de codificación mayores que la capacidad.

Esto prueba la parte inversa del teorema. ■

Esta prueba es la prueba débil inversa del teorema de codificación de canal [9], [10].

Capítulo 5

CONCLUSIONES

5.1. CONCLUSIONES

Las secuencias típicas determinadas por la AEP es un concepto de la teoría de la información, muy útil y muy bello, en el sentido que permiten un desarrollo matemático consistente y sencillo de la teoría de la información, y su comprensión, obtenido a partir del desarrollo de la teoría de conjuntos, de la teoría de la probabilidad y de la aplicación de la LLN. De la aplicación del conjunto típico a un conjunto de mensajes descritos por un proceso estocástico en un sistema de comunicación, se obtiene un subconjunto muy reducido, en el cual se enfoca la atención. Muy bello, porque en esencia manifiesta la convergencia o regularidad asintótica del comportamiento de un proceso estocástico cuando el número de repeticiones tiende a ser muy grande, i.e., una visión regular del comportamiento aparentemente aleatorio, lo cual se observa en la Fig. 2.7, en el cual la medida empírica tiende a la medida teórica.

La LLN es un concepto básico en la teoría de la probabilidad, el cual se logra analizar de forma sencilla, aplicando dos desigualdades básicas, pero muy potentes: la desigualdad de Markov y la desigualdad de Chebyshev, las cuales ofrecen un método de análisis más simple, y que se encuentra directamente relacionado con el análisis de la AEP, las cuales son utilizadas para su demostración. En las referencias principales [9], [10], entre otras, se afirma que la AEP es a la teoría de la información,

lo que la LLN es a la teoría de la probabilidad, lo cual en parte se considera correcto, pero además de esta correspondencia, del estudio expuesto se observa que del análisis de la LLN se obtiene la AEP la cual extiende el estudio de la teoría de la probabilidad, y añade los conceptos de conjunto típico y sus propiedades. De igual manera que la LLN tiene dos formas: en sentido débil y en sentido fuerte, la AEP también tiene dos formas: la AEP débil, que define la convergencia en probabilidad de la entropía empírica (una variable aleatoria que converge a la entropía teórica), condición que es la regla de las secuencias que pertenecen al conjunto típico; y la AEP fuerte que define el conjunto de secuencias típicas para las cuales no solo se cumple la regla de la AEP débil, sino que se cumple que la frecuencia relativa es próxima a la probabilidad de cada posible resultado.

El concepto de secuencias típicas es fundamental en la teoría de la información, ya que permite distinguir dos conjuntos, entre ellos uno muy reducido que concentra la mayor probabilidad. Este descubrimiento junto con los conceptos de medida de información y entropía, son la base del planteamiento del teorema de codificación de fuente y del teorema de codificación de canal. En el segundo capítulo se analizó la entropía como el límite fundamental en la compresión de datos que se realiza sobre el conjunto típico, el cual a su vez está determinado por la entropía de una variable aleatoria genérica. La capacidad del canal discreto con ruido expuestó en el cuarto capítulo es una medida que se obtiene como una función de entropías, o en otras palabras, el límite máximo de la velocidad de transmisión de información es definido por la medida de información promedio de dos vectores de variables aleatorias que corresponden a la entrada y salida del canal. Así como en la entrada del canal, se determinó un conjunto de secuencias típicas, el cual es con una alta probabilidad, el conjunto de secuencias que se van a transmitir, a la salida del canal se encuentra a su vez, un conjunto de secuencias típicas recibidas, y también se obtiene el conjunto típico, las cuales están relacionadas con las secuencias de entrada. La definición de estos conjuntos, permiten aplicar al esquema de codificación el método de decodificación típica conjunta, lo cual permite un análisis probabilístico del error y encontrar que la probabilidad máxima de error es muy próxima a cero, cuando la tasa de codificación es menor a la capacidad. En términos técnicos implementar la decodificación típica conjunta es un proceso complejo, pero en términos teóricos es

una herramienta adecuada para la comprensión de la codificación de canal.

Shannon propone un tipo de codificación como solución y como problema, la codificación es un tema también amplio, y de ella se han obtenido los esquemas de codificación de los modernos sistemas de comunicación, entre ellos: códigos Hamming, códigos de corrección de errores, turbo códigos, códigos de baja densidad de paridad (LDPC, *Low Density Parity Check*). Dado que los esquemas de codificación tienen longitudes de bloque arbitrarias, el problema que plantea el diseño original es un problema de optimización de dimensión infinita (*single letter characterization*). Sin embargo, la solución óptima que caracteriza los límites fundamentales de la comunicación punto a punto, puede expresarse como la de un problema de optimización de dimensión finita. Además, para muchos canales y fuentes específicas, este problema de optimización de dimensión finita se puede resolver explícitamente de forma cerrada [60].

La teoría de la información es una teoría matemática principalmente soportada por la probabilidad, que abstrae la esencia de los problemas de los sistemas de comunicación, la cual se aplica en los sistemas de comunicación inalámbricos, aprendizaje autonómo, biología, medicina e información cuántica, entre muchos otros.

La teoría de la información no define directamente como se deben implementar los sistemas de comunicación. La teoría de información define los principios bajo los cuales estos se diseñan.

5.2. TRABAJOS FUTUROS

Este trabajo propusó, la investigación en el campo de la teoría de la información soportada en una sólida base teórica de los conceptos que la fundamentan, sin embargo, no incluye la pregunta por la actualidad de los problemas de la comunicación, lo cual queda como un trabajo posterior, el cual permitirá introducir al estudiante de pregrado de ingeniería electrónica y telecomunicaciones en el campo de grandes y excelentes teóricos de la teoría de la información, quienes han extendido el problema de la transmisión de información a la determinación teórica de los límites funda-

mentales para canales de múltiples transmisores y múltiples receptores (teoría de la información de redes), es decir, las redes de comunicación, los cuales se analizan como una extensión del análisis de los sistemas de comunicación punto a punto, y las comunicaciones inalámbricas. Adicionalmente se debe trabajar sobre las bases teóricas de la teoría de la información cuántica, lo cual permitirá establecer sus límites fundamentales.

En este trabajo se observa que las secuencias típicas son un método de análisis probabilístico, el cual permite determinar una regularidad asintótica en el caos que se observa de forma inmediata, siendo posible considerar su aplicación en situaciones en las que se requiere estimar un comportamiento regular de grandes cantidades de información, lo cual representa un potencial de análisis teórico para las actuales tecnologías, y con el cual se pueden encontrar y proponer nuevos retos teóricos que den pie al desarrollo de nuevas tecnologías.

A partir de este trabajo, en el que se analizaron principalmente conceptos teoricos de la teoría de la información fundada por Shannon, se sigue la propuesta de investigar la teoría y la practica de los modelos de comunicación más generales, es decir, las redes de comunicación, y la propuesta de encontrar los conceptos matematicos abstractos que tienen una relación directa con la aplicación y el diseño de los modernos sistemas de comunicación.

Bibliografía

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] H. Nyquist, "Certain factors affecting telegraph speed," *Transactions of the American Institute of Electrical Engineers*, vol. XLIII, pp. 412–422, January 1924.

[3] R. V. L. Hartley, "Transmission of information," *Bell System Technical Journal*, vol. 7, no. 3, pp. 535–563, 1928.

[4] H. Nyquist, "Certain topics in telegraph transmission theory," *Transactions of the American Institute of Electrical Engineers*, vol. 47, no. 2, pp. 617–644, 1928.

[5] E. Schrödinger, *La naturaleza y los griegos*. Aguilar, 1961.

[6] B. Russell, *A History of Western Philosophy [poor font]*. 1945.

[7] J. Jeans, *The Growth of Physical Science*. Cambridge Library Collection - Physical Sciences, Cambridge University Press, 2009.

[8] D. J. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, first ed., 2003.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing) (Hardcover)*. New Jersey: John Wiley & Sons, Inc., Hoboken, second ed., 2006.

[10] R. W. Yeung, *Information theory and network coding*. Springer, Boston, MA, 2008.

[11] R. Clausius and W. Browne, *The Mechanical Theory of Heat*. Macmillan, 1879.

[12] A. S. Mariano López de Haro, "Boltzmann y la segunda ley."

[13] N. Wiener, *Cybernetics: or the Control and Communication in the Animal and the Machine*. The MIT Press, 2 ed., 1965.

[14] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.

[15] "El silbo gomero, lenguaje silbado de la isla de la gomera (islas canarias)." https://ich.unesco.org/es/RL/el-silbo-gomero-lenguaje-silbado-de-la-isla-de-la-gomera-islas-canarias-00172, 2009.

[16] M. Kernan, "The talking drums." https://www.smithsonianmag.com/arts-culture/the-talking-drums-29197334/, 2000.

[17] Y. Eshchar, "Where there's smoke, there's a message." https://davidson.weizmann.ac.il/en/online/sciencepanorama/where-theres-smoke-theres-message: :text=Smoke, 2018.

[18] C. W. Bryant, "How do you send a smoke signal?." https://adventure.howstuffworks.com/survival/wilderness/how-to-send-smoke-signal.htm, 2008.

[19] R. Pike, "Book Review: The Early History of Data Networks by Gerard Holzmann and Bjorn Pehrson (IEEE Computer Society Press)," *SIGCOMM Comput. Commun. Rev.*, vol. 24, p. 107–108, Oct. 1994.

[20] N. T. Peter James, *Ancient inventions*. Ballantine Books, 1 ed., 1994.

[21] "Sistema vial en el tahuantinsuyo." https://historiaperuana.pe/periodo-autoctono/sistema-vial-tahuantinsuyo, 2020.

[22] A. Huurdeman, *The Worldwide History of Telecommunications*. Wiley - IEEE, Wiley, 2003.

[23] C. Morrison, "An Expeditious Method of Conveying Intelligence by Means of Electricity," 1753.

[24] G. Bachelard, *La Formación del Espíritu Científico.* Siglo XXI, 2000.

[25] J. B. Anderson and R. Johnnesson, *Introduction: First Ideas and Some History*, pp. 1–29. 2005.

[26] S. Hallas, "The teletype story," 2007.

[27] G. W. e. Andreas Greven, Gerhard Keller, *Entropy.* Princeton Series in Applied Mathematics, Princeton University Press, 2003.

[28] P. P. Urone and R. Hinrichs, *College Physics.* Houston, Texas: OpenStax, June 2012.

[29] M. González, "Historia y epistemología de las ciencias," *Enseñanza de las Ciencias*, p. 18, 2006.

[30] J. L. Fernández, "Segunda Ley de la Termodinámica."

[31] A. Archibald Browning Drysdale, "A short history of philosophy - Archibald B. D. Alexander."

[32] C. R. Nave, "What is the Maxwell-Boltzmann distribution? (article)."

[33] J. Uffink, "Boltzmann's Work in Statistical Physics," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, spring 2017 ed., 2017.

[34] E. Johnson, *Anxiety and the Equation: Understanding Boltzmann's Entropy.* Mit Press, 2018.

[35] G. Wilches Chaux, "la gestión del riesgo: una aproximación alternativa," Jan. 2010.

[36] J. Castaño, *Introducción al Cálculo Infinitesimal.* Serie Didactica de Matemática Contemporanea, Norma, 1962.

[37] R. Courant, H. Robbins, and L. Gala, *¿Qué es la matemática?: una exposición elemental de sus ideas y métodos.* Ciencia y técnica (Aguilar).: Matemáticas y estadística, Aguilar, 1967.

[38] J. E. Lightner, "A Brief Look at the History of Probability and Statistics," *The Mathematics Teacher*, vol. 84, no. 8, pp. 623–630, 1991. Publisher: National Council of Teachers of Mathematics.

[39] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*. Cambridge, Massachusetts, athena scientific ed., 2000.

[40] G. P. Basharin, A. N. Langville, and V. A. Naumov, "The life and work of A.A. Markov," *Linear Algebra and its Applications*, vol. 386, pp. 3 – 26, 2004. Special Issue on the Conference on the Numerical Solution of Markov Chains 2003.

[41] A. A. Markov, "An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains," *Science in Context*, vol. 19, no. 4, p. 591–600, 2006.

[42] C. Bento, "Markov models and Markov chains explained in real life: probabilistic workout routine," Jan. 2021.

[43] R. M. Fano, *Transmission of Information: A Statistical Theory of Communication*. The MIT Press, 1 ed., 1961.

[44] N. J. A. Sloane and A. D. Wyner, *Communication Theory of Secrecy Systems - The material in this paper appeared originally in a confidential report .[A] Mathematical Theory of Cryptography"dated Sept. 1, 1945, which has now been declassified.*, pp. 84–143. 1993.

[45] R. Soni, J.;Goodman, *A Mind at Play: How Claude Shannon Invented the Information Age*. Simon Schuster, 2017.

[46] I. J. J.M. Wozencraft, *Principles of Communication Engineering*. John Wiley Sons Inc, 1966.

[47] W. Heisenberg, *La Imagen de la Naturaleza en la Física Actual*. Ediciones Orbis, 1988.

[48] A. D. W. Claude E. Shannon, N.J.A. Sloane, *Collected Papers of Claude E. Shannon*. IEEE Press, 1 ed., 1993.

[49] R. G. Gallager, "Claude e. shannon: a retrospective on his life, work, and impact," *IEEE Transactions on Information Theory*, vol. 47, pp. 2681–2695, Nov 2001.

[50] D. Wujastyk, "The combinatorics of tastes and humours in classical indian medicine and mathematics," *Journal of Indian Philosophy*, vol. 28, no. 5/6, pp. 479–495, 2000.

[51] B. Cruise, "Visual telegraphs (case study) (video)."

[52] A. Sánchez, "Frecuencia de las letras en castellano: "La Regenta" | Kriptópolis," Dec. 2013.

[53] A. Khinchin, *Mathematical Foundations of Information Theory*. Dover Books on Mathematics, Dover Publications, 1957.

[54] V. M. Q. Florez, *Noisy Channel-Output Feedback in the Interference Channel*. PhD thesis, University of Lyon, 2017.

[55] B. McMillan, "The basic theorems of information theory," *Ann. Math. Statist.*, vol. 24, pp. 196–219, 06 1953.

[56] *Information Theory and Coding: What Did Shannon Promise?*, ch. 5, pp. 150–210. John Wiley Sons, Ltd, 2006.

[57] R. G. Gallager, *Information Theory and Reliable Communication*. Wiley, 1st ed., 1968.

[58] E. Ruppert, "Solving recurrences," *Technical Report*, 2008.

[59] A. Feinstein, "A new basic theorem of information theory," *Transactions of the IRE Professional Group on Information Theory*, vol. 4, pp. 2–22, Sep. 1954.

[60] A. S. Avestimehr, S. N. Diggavi, C. Tian, and D. N. C. Tse, "An approximation approach to network information theory," *Foundations and Trends® in Communications and Information Theory*, vol. 12, no. 1-2, pp. 1–183, 2015.

[61] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. 1 (v. 1)*. John Wiley and Sons (WIE), 3rd ed., 1968.

Secuencias Típicas en el análisis de Capacidad de un Canal de Comunicación Digital

ANEXOS

Trabajo de Grado

María del Pilar Ramos

Director: Víctor Manuel Quintero Florez

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telecomunicaciones
Grupo de Radio e Inalámbricas - GRIAL
Sistemas de Comunicaciones Móviles e Inalámbricos
Popayán, 2021

Apéndice A

PROBABILIDAD

A.1. Conceptos Básicos de Probabilidad

Definición A.1 (Definición en [61])**.** Un *espacio de probabilidad* $(\Omega, \mathcal{B}, P)$, donde Ω es un conjunto no vacío llamado *espacio muestral*, $\mathcal{B}$ es un campo de Borel de los subconjuntos de Ω, y P es una función no-negativa definida para todo $A \in \mathcal{B}$ con la propiedad que $P(\Omega) = 1$ y

$$P\left\{ \bigcup_{n=1}^{\infty} A_n \right\} = \sum_{n=1}^{\infty} P\{A_n\}, \tag{A.1}$$

$A_n \in \mathcal{B}$, y los A_n's son disjuntos. P es llamada *medida de probabilidad*.

Por ejemplo: si $\Omega = \{\omega_1, \omega_2, \ldots\}$ es un conjunto contable finito, $\mathcal{B}$ es la colección de todos los subconjuntos de Ω, $(p_1, p_2, \ldots)$ es una secuencia de números no negativos cuya suma es 1, luego, la definición de $P\{A\} = \sum\{p_n : \omega_n \in A\}$ hace $(\Omega, \mathcal{B}, P)$ un espacio de probabilidad; es llamado espacio de probabilidad *discreto*.

Definición A.2. Una variable aleatoria X es una función que mapea Ω en algún conjunto $\mathcal{X}$, llamado el *rango de* X i.e., $X : \Omega \to \mathcal{X}$. Se asume, generalmente que $\mathcal{X}$ es un subconjunto de números reales.

Definición A.3. El valor esperado de la variable aleatoria X, con pmf p_X se define como:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x), \tag{A.2}$$

lo cual converge a un valor finito, si se cumple: $\sum_{x \in \mathcal{X}} |x| p_X(x) < \infty$.

Definición A.4. La varianza de la variable aleatoria X, con pmf $p_X(x)$, es el valor esperado de la variable aleatoria $(X - \mathbb{E}[X])^2$, el cual es:

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]. \tag{A.3}$$

Definición A.5 (Proceso Estocástico)**.** Un proceso estocástico es un modelo matemático de un experimento probabílistico que evoluciona en el tiempo y genera un secuencia de valores númericos. Donde cada valor número en la secuencia es modelado por una variable aleatoria. De forma no formal, es una secuencia de variables aleatorias.

Definición A.6 (Probabilidad Condicional)**.** Sean X y Y dos variables aleatorias, si se conoce el resultado de Y, tal que $p_Y(y) > 0$, esto proporciona un conocimiento parcial del valor que toma X y tiene asociada la pmf condicional $p_{X|Y}(x|y)$ de una variable aleatoria X con respecto a la variables aleatoria Y es:

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)} \tag{A.4}$$

Definición A.7 (Independencia de Variables Aleatorias)**.** Dos variables aleatorias X y Y son independientes, denotadas por $X \perp Y$, si:

$$p_{XY}(x, y) = p_X(x) p_Y(y), \tag{A.5}$$

para todo $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Para más de dos variables aleatorias, se distinguen dos tipos de independencia.

Definición A.8 (Independencia Mutua)**.** Para $n \geq 3$, las variables aleatorias $X_1, X_2, \ldots, X_n$

son mutuamente independientes si:

$$p_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = p_X(x_1)p_X(x_2)\ldots p_X(x_n), \tag{A.6}$$

para todo $x_1, x_2, \ldots, x_n$.

Definición A.9 (Independencia por parejas)**.** Para $n \geq 3$, las variables aleatorias $X_1, X_2, \ldots, X_n$, son independientes a pares si X_i y X_j son independientes para todo $1 \leq i \leq j \leq n$.

La independencia mutua implica la independencia por parejas, pero el contrario no es cierto.

Definición A.10 (Independencia Condicional)**.** Para variableas aleatorias X, Y y Z, X es independiente de Z condicionada por Y, y se denota como: $X \perp Z|Y$, si se cumple:

$$p_{XYZ}(x, y, z)p_Y(y) = p_{XY}(x, y)p_{YZ}(y, z), \tag{A.7}$$

para todo x, y y z, o equivalentemente:

$$p_{XYZ}(x, y, z) = \frac{p(x, y)p(y, z)}{p(y)} = p(x, y)p(z|y) \text{ si } p_Y(y) > 0 \tag{A.8}$$

y es cero de lo contrario.

Definición A.11 (Cadenas de Markov)**.** Para las variables aleatorias $X_1, X_2, \cdots, X_n$, donde $n \geq 3$, $X_1 \to X_2 \to \cdots \to X_n$ forman una cadena de Markov si:

$$\begin{aligned} p(x_1,&x_2, \cdots, x_n)p(x_2)p(x_3)\cdots p(x_{n-1}) \\ =&p(x_1, x_2)p(x_2, x_3)\cdots p(x_{n-1}, x_n). \end{aligned} \tag{A.9}$$

Para todo $x_1, x_2, \ldots, x_n$, o equivalentemente,

$$p(x_1,x_2, \cdots, x_n) = p(x_1, x_2)p(x_3|x_2)\cdots p(x_n|x_{n-1}), \tag{A.10}$$

si $p(x_2), p(x_3), \cdots, p(x_{n-1}) > 0$ y cero de lo contrario.

Se observa que la definición de independencia condicional expresada: $X \perp Z|Y$ es equivalente a una cadena de Markov $X \to Y \to Z$.

Apéndice B

MEDIDAS DE INFORMACIÓN

B.1. Entropía

La definición de entropía de una variable aleatoria discreta X, denotada por $H(X)$, es un funcional de la pmf p_X, que mide la cantidad de información promedio de X o su medida incertidumbre. La definición que se presenta en (2.3) es la única expresión posible de entropía. Si se considera el caso especial en el que p_X es uniforme se demuestra fácilmente el teorema de unicidad de $H(X)$ o $H(p_X)$; para ello se requiere denotar p_X para el conjunto contable $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ como un conjunto de valores de probabilidad $(p_1, p_2, \ldots, p_n)$ asociado a cada resultado de la variable aleatoria X; y que la entropía cumpla las siguientes propiedades:

Lema B.1. *La entropía $H(p_1, p_2, \ldots, p_n)$ cumple las siguientes propiedades:*

1. *Para cualquier n, la función $H(p_1, p_2, \ldots, p_n)$ tiene su máximo valor para $p_i = \frac{1}{n}$, con $i \in \{1, 2, \ldots, n\}$.*
2. *$H(X, Y) = H(X) + H(Y|X)$.*
3. *$H(p_1, p_2, \ldots, p_n, 0) = H(p_1, p_2, \ldots, p_n)$, es decir, añadir un evento o cualquier número de eventos improbables a una pmf no cambia la entropía.*

Teorema B.2. *Sea $H(p_1, p_2, \ldots, p_n)$ una función definida para cualquier entero n y para todos los valores $p_1, p_2, \ldots, p_n$ tales que $p_k \geq 0$ con $k \in \{1, 2, \ldots, n\}$, tal que, $\sum_{k=1}^{n} p_k = 1$. Si para cualquier n está función es continua con respecto a todos sus argumentos, y si las propiedades 1, 2 y 3 se cumplen, entonces:*

$$H(p_1, p_2, \ldots, p_n) = -\lambda \sum_{K=1}^{n} p_k \log p_k, \tag{B.1}$$

donde λ es una constante positiva[1]*.*

Prueba: Por brevedad se establece:

$$H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) = L(n). \tag{B.2}$$

Se debe mostrar que $L(n) = \lambda \log n$, donde λ es una constante positiva[2]. Por la aplicación de la propiedade 3 se obtiene lo siguiente:

$$L(n) = H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}, 0\right) \leq H\left(\frac{1}{n+1}, \frac{1}{n+1}, \ldots, \frac{1}{n+1}\right) = L(n+1), \tag{B.3}$$

por lo tanto, $L(n)$ es una función no decreciente de n[3]. Sean m y r enteros positivos. Considerar m conjuntos independientes $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m$, cada uno con r eventos equiprobables, la entropía de cada uno se obtiene a continuación:

$$H(\mathcal{S}_k) = H\left(\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r}\right) = L(r), \tag{B.4}$$

para $1 \leq k \leq m$. Por la propiedad 2, (generalizada a m dado que los conjuntos S_k son independientes) se obtiene lo siguiente:

$$H(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m) = \sum_{k=1}^{m} H(\mathcal{S}_k) = mL(r). \tag{B.5}$$

[1] Este teorema demuestra que la definición de la entropía es la única posible. Es un teorema de la singularidad o unicidad de la entropía.

[2] $L(n) = \lambda \sum_{i=1}^{n} p_i \log p_i$, todos los $p_i = \frac{1}{n}$, así: $L(n) = \lambda n \frac{1}{n} \log n = \lambda \log n$.

[3] Una función creciente de n. Se demuestra la monotonía creciente de la entropía.

EL producto de $\mathcal{S}_1\mathcal{S}_2 \ldots \mathcal{S}_m$ consiste de r^m eventos igualmente probables, así que su entropía es $L(r^m)$ y dado que $L(r) = \lambda \log r$, se obtiene lo siguiente:

$$\begin{aligned} L(r^m) &= \lambda \log r^m = m\lambda \log r \\ &= mL(r), \end{aligned} \tag{B.6}$$

y de forma similar, para cualquier otro par de enteros positivos n y s, se obtiene:

$$L(s^n) = nL(s). \tag{B.7}$$

Ahora, sean los números r, s y n dados arbitrariamente, y el número m determinado por las desigualdades:

$$r^m \leq s^n \leq r^{m+1}, \tag{B.8}$$

al aplicar el logaritmo a la desigualdad (B.8), se obtiene lo siguiente:

$$\log r^m \leq \log s^n < \log r^{m+1} \tag{B.9}$$

$$m \log r \leq n \log s < (m+1) \log r \tag{B.10}$$

$$\frac{m}{n} \leq \frac{\log s}{\log r} < \frac{m}{n} + \frac{1}{n}. \tag{B.11}$$

Por la monotonía de la función $L(n)$, se puede afirmar lo siguiente:

$$L(r^m) \leq L(s^n) \leq L(r^{m+1}), \tag{B.12}$$

aplicando (B.6) y (B.7) en la desigualdad (B.12), se obtiene lo siguiente:

$$mL(r) \leq nL(s) \leq (m+1)L(r), \tag{B.13}$$

por lo tanto,

$$\frac{m}{n} \leq \frac{L(s)}{L(r)} \leq \frac{m}{n} + \frac{1}{n}. \tag{B.14}$$

Finalmente,

$$\left| \frac{L(s)}{L(r)} - \frac{\log s}{\log r} \right| \leq \frac{1}{n}. \tag{B.15}$$

Ya que el lado izquierdo de la desigualdad (B.15) es independiente de m, y ya que n puede ser seleccionado arbitrariamente grande en el lado derecho, se obtiene lo siguiente:

$$\frac{L(s)}{\log s} = \frac{L(r)}{\log r}. \tag{B.16}$$

Lo cual, en vista de la arbitrariedad de r y s, significa que:

$$L(n) = \lambda \log n,$$

donde, λ es una constante. Por la monotonía de la función $L(n)$, se tiene que $\lambda \geq 0$. y se completa la prueba para el caso especial en que $p_k = 1/n$ $k \in \{1, 2, \ldots, n\}$. La prueba del caso más general, se puede seguir en [53]. ■

B.2. Propiedades de la Entropía

Lema B.3. *Sea $\mathcal{X}$ un conjunto contable y sea X una variable aleatoria con pmf p_X, entonces la entropía se encuentra acotada, de la siguiente forma:*

$$0 \leq H(X) \leq \log |\mathcal{X}|. \tag{B.17}$$

Prueba: La cota inferior de la entropía de la variable aleatoria X, se obtiene de considerar que para todo $x \in \mathcal{S}_X$, se cumple $0 < p_X(x) \leq 1$, i.e., $\frac{1}{p_X(x)} \geq 1$, por lo tanto:

$$\log \left(\frac{1}{p_X(x)} \right) \geq 0, \tag{B.18}$$

así, $H(X) \geq 0$. La cota superior de la entropía de la variable aleatoria X, se obtiene de (2.3):

$$H(X) = \mathbb{E}\left[\log \frac{1}{p_X(x)} \right] \tag{B.19}$$

$$\leq \log \mathbb{E}\left[\frac{1}{p_X(x)} \right] \tag{B.20}$$

$$= \log \sum_{x \in \mathcal{S}_X} 1 \tag{B.21}$$

$$= \log |\mathcal{X}|, \tag{B.22}$$

donde (B.20) se obtiene de aplicar la desigualdad de Jensen definida en (D.7). Así el valor máximo de la entropía se obtiene cuando la variable aleatoria X tiene una distribución de probabilidad uniforme, i.e., $p_X(x) = \frac{1}{|\mathcal{X}|}$ para todo $x \in \mathcal{S}_X$. Esto completa la prueba del Lema B.3. ■

Lema B.4. *Sean $\mathcal{X}$ y $\mathcal{Y}$ dos conjuntos contables y X y Y dos variables aleatorias con pmf conjunta p_{XY}, entonces se cumple:*

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

donde la igualdad se cumple si y solo si las variables aleatorias son independientes.

Prueba: De (4.2) se obtiene lo siguiente:

$$H(X,Y) = -\mathbb{E}\left[\log\left(\frac{p_X(X)p_Y(Y)p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right)\right] \tag{B.24}$$

$$= -\mathbb{E}[\log p_X(X)] - \mathbb{E}[\log p_Y(Y)] - \mathbb{E}\left[\log\left(\frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right)\right] \tag{B.25}$$

$$= H(X) + H(Y) + \mathbb{E}\left[\log\left(\frac{p_X(X)p_Y(Y)}{p_{XY}(X,Y)}\right)\right] \tag{B.26}$$

$$\leq H(X) + H(Y) + \log\left(\mathbb{E}\left[\frac{p_X(X)p_Y(Y)}{p_{XY}(X,Y)}\right]\right) \tag{B.27}$$

$$= H(X) + H(Y) + \log\left(\sum_{x,y \in \mathcal{S}_{XY}} p_X(X)p_Y(Y)\right) \tag{B.28}$$

$$= H(X) + H(Y), \tag{B.29}$$

donde (B.27) se obtiene de aplicar la desigualdad de Jensen. Si las variables aleatorias

son independientes, entonces $p_{XY}(x,y) = p_X(x)p_Y(y)$, así (B.26) se convierte en:

$$H(X,Y) = H(X) + H(Y) + \mathbb{E}\left[\log\left(\frac{p_X(X)p_Y(Y)}{p_X(X)p_Y(Y)}\right)\right] \tag{B.30}$$

$$= H(X) + H(Y). \tag{B.31}$$

Y esto completa la prueba del Lema B.4. ■

El siguiente lema es una generalización de la regla de la cadena para la entropía y la entropía condicional.

Lema B.5. *Sean $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ y $\mathcal{Y}$, $n+1$ conjuntos contables, y sea el vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ de variables aleatorias discretas de dimensión n, y sea Y una variable aleatoria con pmf conjunta $p_{\mathbf{X}}$, y $p_{\mathbf{X}Y}$, entonces se cumple lo siguiente:*

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \sum_{k=3}^{n} H(X_k|X_1, \ldots, X_{k-1}) \quad \textit{(B.32)}$$

$$H(X_1, \ldots, X_n|Y) = H(X_1|Y) + H(X_2|Y, X_1) + \sum_{k=3}^{n} H(X_k|Y, X_1, \ldots, X_{k-1}). \quad \textit{(B.33)}$$

Prueba: ▪ Prueba de (B.32): de (4.4) se obtiene lo siguiente:

$$H(\mathbf{X}) = -\mathbb{E}[p_{\mathbf{X}}(\mathbf{X})] \tag{B.34}$$

$$= -\mathbb{E}[\log(p_{X_1}(X_1)p_{X_2|X_1}(X_2|X_1)\ldots p_{X_n|X_1,X_2,\ldots,X_{n-1}}(X_n|X_1, X_2, \ldots, X_{n-1}))] \tag{B.35}$$

$$= -\mathbb{E}[\log p_{X_1}(X_1)] - \mathbb{E}[\log p_{X_2|X_1}(X_2|X_1)] - \ldots \tag{B.36}$$

$$- \mathbb{E}[\log p_{X_n|X_1,X_2,\ldots,X_{n-1}}(X_n|X_1, X_2, \ldots, X_{n-1})] \tag{B.37}$$

$$= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, X_2, \ldots, X_{n-1}), \tag{B.38}$$

y esto completa la prueba de (B.32).

- Prueba de (B.33): De (4.11) se obtiene lo siguiente:

$$
\begin{aligned}
H(\mathbf{X}|Y) = &- \mathbb{E}[\log p_{\mathbf{X}|Y}(\mathbf{X}|\mathbf{Y})] \\
&- \mathbb{E}\Big[\log \Big(p_{X_1|Y}(X_1|Y) p_{X_2|X_1Y}(X_2|X_1Y) \dots \\
&\quad p_{X_n|X_1,X_2,\dots,X_{n-1}Y}(X_n|X_1,X_2,\dots,X_{n-1}Y) \Big) \Big] \qquad \text{(B.39)} \\
&= \mathbb{E}\big[\log p_{X_1|Y}(X_1|Y) \big] - \mathbb{E}\big[\log p_{X_2|X_1Y}(X_2|X_1,Y) \big] - \dots \\
&- \mathbb{E}\big[\log p_{X_n|X_1,X_2,\dots,X_{n-1}Y}\big(X_n|X_1,X_2,\dots,X_{n-1}Y\big) \big] \qquad \text{(B.40)} \\
&= H(X_1|Y) + H(X_2|Y) + \dots \\
&\quad + H(X_n,|Y,X_1,X_2,\dots X_{n-1}), \qquad \text{(B.41)}
\end{aligned}
$$

y esto completa la prueba de (B.33). Lo cual completa la prueba del Lema B.5.

■

Apéndice C

DIFERENCIAS FINITAS

C.1. Recurrencias Lineales Homogéneas

Sean k y n enteros positivos, una relación de recurrencia lineal homogénea de grado k con coeficientes constantes, tiene la forma siguiente:

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + ... + c_k a_{n-k}, \tag{C.1}$$

donde c_i son constantes, con $i \in \{1, 2, \ldots, n-k\}$. Para que la definición de una relación de recurrencia este completa se deben conocer los k casos iniciales, que permiten especificar los valores de $a_0, a_1, ..., a_{k-1}$.

El Lema C.1 es una de las propiedades clave que hace que las recurrencias lineales homogéneas sean bastante fácil de resolver.

Lema C.1. *Si la secuencia a_n satisface* (C.1) *y a'_n es otra secuencia que satisface* (C.1)*, entonces $b_n = a_n + a'_n$ y $d_n = \alpha a_n$, son también secuencias que satisfacen* (C.1)*, con α constante.*

Prueba: Debido a que a_n y a'_n cumplen (C.1), se obtiene lo siguiente:

$$b_n = a_n + a'_n \tag{C.2}$$
$$= \big(c_1 a_{n-1} + c_2 a_{n-2} + ... + c_k a_{n-k}\big) + \big(c_1 a'_{n-1} + c_2 a'_{n-2} + ... + c_k a'_{n-k}\big) \tag{C.3}$$
$$= c_1(a_{n-1} + a'_{n-1}) + c_2(a_{n-2} + a'_{n-2}) + \ldots c_k(a_{n-k} + a'_{n-k}) \tag{C.4}$$
$$= c_1 b_{n-1} + c_2 b_{n-2} + \ldots + c_k b_{n-k}. \tag{C.5}$$

de forma similar, se obtiene lo siguiente:

$$d_n = \alpha a_n \tag{C.6}$$
$$= \alpha\big(c_1 a_{n-1} + c_2 a_{n-2} + ... + c_k a_{n-k}\big) \tag{C.7}$$
$$= c_1(\alpha a_{n-1}) + c_2(\alpha a_{n-2}) + \ldots + c_k(\alpha a_{n-k}) \tag{C.8}$$
$$= c_1 d_{n-1} + c_2 d_{n-2} + \ldots + c_k d_{n-k}. \tag{C.9}$$

Con esto se completa la prueba del Lema C.1. ■

Por lo tanto, del Lema C.1 se puede afirmar lo siguiente: si se encuentran algunas soluciones básicas a (C.1), entonces cualquier combinación lineal de ellas también será una solución de (C.1). Esto es útil, debido a que es relativamente fácil encontrar soluciones a (C.1), las cuales se pueden combinar sumando o multiplicándolas por constantes, con el fin de encontrar las soluciones particulares de los casos base que sean de interés. Por lo tanto, se desea establecer el método para encontrar las soluciones de (C.1).

Se considera analizar si las secuencias geométricas de la forma $a_n = r^n$ satisfacen (C.1), i.e., demostrar si cualquier r cumple lo siguiente:

$$r^n = c_1 r^{n-1} + c_2 r^{n-2} + ... + c_k r^{n-k}. \tag{C.10}$$

Se ubican todos los términos de (C.10) al lado izquierdo, como se muestra a continuación:

$$r^n - c_1 r^{n-1} - c_2 r^{n-2} - ... - c_k r^{n-k} = 0. \tag{C.11}$$

Ahora se dividen ambos lados de (C.11) entre r^{n-k}, y se obtiene lo siguiente:

$$r^k - c_1 r^{k-1} - c_2 r^{k-2} - ... - c_k r^0 = 0, \tag{C.12}$$

donde (C.12) es la ecuación carcaterística de (C.1). Por ejemplo, la ecuación característica de la recurrencia de Fibonacci $F_n = F_{n-1} + F_{n-2}$ es $r^2 - r - 1 = 0$. Lo anterior permite establecer el siguiente Lema:

Lema C.2. *Si r satisface* (C.12)*, entonces $a_n = r^n$ satisface* (C.1)*.*

Teorema C.3. *Si $r_1, r_2, ..., r_m$ satisfacen* (C.12)*, entonces para cualquier $\alpha_1, \alpha_2, ..., \alpha_m$ constantes, la secuencia $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n + ... + \alpha_m r_m^n$ cumple* (C.12)*.*

Prueba. Del Lema C.2, se sabe que r_i^n satisface (C.1), para cada i solución en (C.12); y por el Lema C.1, se sabe también que αr_i^n cumple (C.1), para cada i. Aplicando el Lema C.1 de nuevo, se obtiene: $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n + \ldots + \alpha_m r_m^n$ cumple (C.1). ■

Para observar los resultados previos del Teorema C.3 se presenta el siguiente ejemplo: la recurrencia Fibonacci definida como:

$$F_0 = 0 \tag{C.13}$$

$$F_1 = 1 \tag{C.14}$$

$$F_n = F_{n-1} + F_{n-2}, \text{ para } n \geq 2. \tag{C.15}$$

La ecuación característica de esta ecuación es: $r^2 - r - 1 = 0$. Esta recurrencia se resuelve para r usando la fórmula cuadrática, i.e., $r = \frac{1 \pm \sqrt{5}}{2}$. Por el Teorema C.3 se obtiene lo siguiente:

$$F_n = \alpha_1 \left(\frac{1 + \sqrt{5}}{2} \right)^n + \alpha_2 \left(\frac{1 - \sqrt{5}}{2} \right)^n, \tag{C.16}$$

expresión que satisface $F_n = F_{n-1} + F_{n-2}$, para cualquier valor de α_1 y α_2. Ahora, dados los valores iniciales de la recurrencia $F_0 = 0$ y $F_1 = 1$, es posible encontrar los valores particulares de α_1 y α_2. Se obtienen dos ecuaciones con dos incógnitas, que se resuelven con $\alpha_1 = \frac{1}{\sqrt{5}}$ y $\alpha_2 = -\frac{1}{\sqrt{5}}$, por lo tanto, se obtiene la expresión de

la recurrencia de Fibonacci completa, de la siguiente forma:

$$F_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n + -\frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n, \tag{C.17}$$

así, (C.17) satisface la recurrencia (C.1) por el Teorema C.3.

C.2. Número de Secuencias de Duración t

La capacidad es por la Definición 2.3, el logaritmo del número de secuencias de duración t, cuando t tiende a infinito. Para contar el número de secuencias se utiliza el concepto de ecuaciones de diferencias finitas, que representan una relación de recurrencia lineal y homogénea, la cual se expresa en (2.22). La cual tiene como solución una combinación lineal de las soluciones reales $\{x_0, x_1, \ldots\}$ de la ecuación característica, i.e., $x_0^t + x_1^t + \ldots$, donde x_0 es la solución real mayor, por lo tanto, se obtiene que la capacidad es:

$$C = \lim_{t\to\infty} \frac{\log X_0^t}{t} = \log X_0, \tag{C.18}$$

Para el sistema de telegrafía la ecuación de diferencias finitas es definida en (2.24), y su ecuación característica asociada es (2.25), de la cual se busca encontrar las raíces, para las cuales el Teorema C.3 afirma que son solución a la ecuación de recurrencias, la ecuación característica se reescribe:

$$a^{10} + a^8 + a^7 + a^5 + a^4 + a^2 - 1 = 0.$$

Utilizando el software Matlab se encuentran sus raíces reales, a saber: $a_1 = -0.862743$ y $a_2 = 0.688278$, por lo tanto, la solución de (2.24) es:

$$a_n = \alpha_1(-0.862743)^{-t} + \alpha_2(0.688278)^{-t}, \tag{C.19}$$

de (C.18), se obtiene que la capacidad para el sistema de telegrafía es el logaritmo de la raíz mayor de (2.25), i.e., $C = -\log a_2 = 0.539$ [1].

Apéndice D

CONVERGENCIA Y DESIGUALDADES

D.1. TIPOS DE CONVERGENCIA

D.1.1. Convergencia de una Secuencia Determinística

Definición D.1. Sea $a_1, a_2, \ldots, a_n$ una secuencia de números reales y sea a otro número real, se dice que la secuencia a_n converge a a, i.e., si,

$$\lim_{n\to\infty} a_n = a. \tag{D.1}$$

Si para cada $\epsilon > 0$ existe algún n_0, de modo que:

$$|a_n - a| \leq \epsilon \quad \text{para todo} \quad n \leq n_0, \tag{D.2}$$

donde ϵ corresponde al nivel de precisión, de que tan próximo esta a_n a a cuando n es suficientemente grande.

D.1.2. Convergencia en Probabilidad

Definición D.2. Sea $\{X_k\}, k \in \mathbb{Z}^+$ una secuencia de variables aleatorias (no necesariamente independientes) y sea X otra variable aleatoria, se dice que dicha sucesión converge en probabilidad a X, si para todo $\epsilon > 0$, se cumple:

$$\lim_{k\to\infty} \Pr\left\{|X_k - X| \leq \epsilon\right\} = 1. \tag{D.3}$$

D.2. DESIGUALDADES IMPORTANTES

D.2.1. Desigualdad de Jensen

Lema D.1. *Sea $f : (a, b) \to \mathbb{R}$, una función convexa. Sea $n \in \mathbb{N}$, $n \geq 2$. Dados los números $x_1, x_2, \ldots, x_n$, en el intervalo (a, b) y números no negativos $\lambda_1, \lambda_2, \lambda_n$, tales que:*

$$\lambda_1 + \lambda_2 + \ldots + \lambda_n = 1, \tag{D.4}$$

se obtiene lo siguiente:

$$f(\lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_n f(x_n). \tag{D.5}$$

o lo que es equivalente:

$$f\Big(\sum_{i=1}^{n} \lambda_i x_i\Big) \leq \sum_{i=1}^{n} \lambda_i f(x_i). \tag{D.6}$$

En caso de que f sea estrictamente convexa la desigualdad es estricta, excepto en los casos triviales. Para una función cóncava el resultado es similar, con la desigualdad en sentido distinto.

De forma similar, se encuentra la desigualdad de Jensen de variables aleatorias.

Lema D.2. *Sea X una variable aleatoria discreta, si $g(\cdot)$ es una función convexa, se cumple lo siguiente:*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \tag{D.7}$$

Prueba: Sea g un función convexa y si $0 \leq p \leq 1$, entonces para dos valores de la función, x, y, lo siguiente se cumple:

- $g(px + (1-p)y) \leq pg(x) + (1-p)g(y)$.
- Es dos veces diferenciable: $g''(x) \geq 0$.
- Para cualquier $c, x : g(x) \geq g(c) + g'(c)(x-c)$.

Lo cual se cumple para cualquier x y cualquier valor constante de c, así, que se considera una variable aleatoria X y $c = \mathbb{E}[X]$, por lo tanto, lo siguiente es valido:

$$g(X) \geq g(\mathbb{E}[X]) + g'(\mathbb{E}[X])(X - \mathbb{E}[X]), \tag{D.8}$$

ahora, se aplica el valor esperado a ambos lados de la desigualdad (D.8), y se obtiene lo siguiente:

$$\mathbb{E}[g(X)] \geq \mathbb{E}[g(\mathbb{E}[X])] + \mathbb{E}[g'(\mathbb{E}[X])]\mathbb{E}[(X - \mathbb{E}[X])] \tag{D.9}$$

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \tag{D.10}$$

donde, (D.10) se obtiene debido a que $\mathbb{E}[c] = c$, donde c es un número y $g(\mathbb{E}[X])$ es un número, y $\mathbb{E}[(X - \mathbb{E}[X])] = 0$. Lo cual completa la prueba del Lema D.2. ■

D.2.2. Desigualdad de Markov

Lema D.3. *Sea X una variable alearoria no negativa, y $a \in \mathbb{R}$, entonces:*

$$\Pr\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a} \text{ para todo } a > 0. \tag{D.11}$$

Prueba: Sea Y_a una variable aleatoria con $a > 0$, definida de la siguiente forma:

$$Y_a = \begin{cases} 0 & si \quad x < a \\ \\ a & si \quad x \geq a \end{cases}. \tag{D.12}$$

Por lo tanto, se obtiene lo siguiente:

$$Y_a \leq X \tag{D.13}$$

$$\mathbb{E}[Y_a] \leq \mathbb{E}[X], \tag{D.14}$$

donde, el valor esperado de Y_a es: $\mathbb{E}[Y_a] = a \Pr\{Y_a = a\} = a \Pr\{X \geq a\}$, o equivalentemente:

$$a \Pr\{X \geq a\} \leq \mathbb{E}[X] \tag{D.15}$$

$$\Pr\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a}. \tag{D.16}$$

Lo cual completa la prueba del Lema D.3. ■

D.2.3. Desigualdad de Chebyshev

Lema D.4. *Sea X una variable aleatoria con media $\mathbb{E}[X] = \mu$ y varianza $Var[X] = \sigma^2$, y $c \in \mathbb{R}$, se cumple la siguiente desigualdad:*

$$\Pr\left\{|X - \mu| \geq c\right\} \leq \frac{\sigma^2}{c^2}, \ \textit{para todo } c > 0. \tag{D.17}$$

Prueba: Se considera la variable aleatoria $(X - \mu)^2$, la cual es no negativa y se aplica la desigualdad de Markov con $a = c^2$.

$$\Pr\left\{|X - \mathbb{E}[X]| \geq c\right\} = \Pr\left\{(X - \mathbb{E}[X])^2 \geq c^2\right\} \tag{D.18}$$

$$\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{c^2} = \frac{\mathrm{Var}[X]}{c^2}. \tag{D.19}$$

Esto completa la prueba del Lema D.4. ■

D.2.4. Ley Débil de los Grandes Números

Definición D.3. Sea $\{X_n\}, n \in \mathbb{Z}^+$ una secuencia de variables aleatorias i.i.d., cada una con $\mathbb{E}[X] = \mu$ y Var[X] $= \sigma^2$, se define la media muestral como:

$$M_n = \frac{X_1 + X_2 + \ldots X_n}{n}. \tag{D.20}$$

El valor esperado de M_n es:

$$\begin{aligned}
\mathbb{E}[M_n] &= \mathbb{E}\left[\frac{X_1 + X_2 + \ldots X_n}{n}\right] && \text{(D.21)}\\
&= \frac{1}{n}\Big(\mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]\Big) && \text{(D.22)}\\
&= \frac{n\mu}{n} && \text{(D.23)}\\
&= \mu. && \text{(D.24)}
\end{aligned}$$

La varianza de M_n es:

$$\begin{aligned}
\mathrm{Var}[M_n] &= \mathrm{Var}\left[\frac{X_1 + X_2 + \ldots X_n}{n}\right] && \text{(D.25)}\\
&= \mathrm{Var}\left(\frac{X_1}{n}\right) + \ldots + \mathrm{Var}\left(\frac{X_n}{n}\right) && \text{(D.26)}\\
&= \frac{\sigma^2}{n^2} + \ldots + \frac{\sigma^2}{n^2} && \text{(D.27)}\\
&= \frac{\sigma^2}{n}. && \text{(D.28)}
\end{aligned}$$

Lema D.5. *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias i.i.d., con media μ. Para cada $\epsilon > 0$, se cumple lo siguiente:*

$$\Pr\{|M_n - \mu| \geq \epsilon\} = \Pr\left\{\left|\frac{X_1 + X_2 + ... + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0 \tag{D.29}$$

Prueba: La desigualdad de Chebyshev en (D.17), se aplica a la variable aleatoria

M_n. Para todo $\epsilon > 0$ se obtiene lo siguiente:

$$\Pr\{|M_n - \mu| \geq \epsilon\} \leq \frac{\text{Var}[X]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}, \tag{D.30}$$

donde la varianza se obtiene de (D.28). Luego, como $n \to \infty$, lo siguiente es valido:

$$\lim_{n\to\infty} \Pr\{|M_n - \mu| \geq \epsilon\} = 0. \tag{D.31}$$

Lo cual completa la prueba del Lema D.5. ■

D.2.5. Ley Fuerte de los Grandes Números

Lema D.6. *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias i.i.d. con media μ. Luego la secuencia de medias muestrales M_n converge a μ con probabilidad 1.*

$$\Pr\left\{ \lim_{n\to\infty} \left(\frac{X_1 + X_2 + \cdots + X_n}{n} \right) = \mu \right\} = 1. \tag{D.32}$$

D.2.6. Desigualdad de Procesamiento de Datos

El teorema de procesamiento de datos, establece que el *procesamiento de datos* puede solamente destruir la información.

Lema D.7. *Sean $\mathcal{W}, \mathcal{D}$ y $\mathcal{R}$ tres conjuntos contables y sean las variables aleatorias W, D y R, con pmf conjunta p_{WDR}, de modo que estas tres variables formen una cadena de Markov: $W \to D \to R$, i.e., la pmf conjunta p_{WDR} puede ser escrita como:*

$$p_{WDR}(w, d, r) = p_W(w) p_{D|W}(d|w) p_{R|D}(r|d). \tag{D.33}$$

La información promedio que trasnmite R sobre W, i.e., $I(W;R)$ es menor o igual a la información promedio que D transmite sobre W, i.e., $I(W;D)$, lo cual se expresa de la siguiente forma:

$$I(W;R) \leq I(W;D) \tag{D.34}$$

Prueba: Para cualquier conjunto de variables aleatorias X, Y y Z, la siguiente regla de la cadena se cumple:

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y). \tag{D.35}$$

Ahora, en el caso $w \to d \to r$, w y r son condicionalmente independientes dado d, asi que $I(W;R|D) = 0$. Usando la regla de la cadena dos veces:

$$I(W;D,R) = I(W;D) + I(W;R|D) \tag{D.36}$$
$$= I(W;D) \tag{D.37}$$

y

$$I(W;D,R) = I(W;R) + I(W;D|R). \tag{D.38}$$

Ya que $I(W;D|R) \geq 0$:

$$I(W;D) = I(W;R) + I(W;D|R) \tag{D.39}$$
$$I(W;D|R) = I(W;D) - I(W;R) \geq 0 \tag{D.40}$$
$$I(W;D) \geq I(W;R). \tag{D.41}$$

Esto completa la prueba del Lema D.7. ■

D.2.7. Desigualdad de Fano

En la transmisión de datos esta desigualdad establece una conexión entre una medida practica tradicional, la probabilidad de error, y una medida de información del efecto del canal sobre el ruido, y la equivocación o entropía condicional.

Lema D.8. *Sea $\mathcal{X}$ un conjunto contable y X, $\widehat{X}$ dos variables aleatorias con pmf conjunta $p_{X\widehat{X}}$, tal que para todo $(x,\widehat{x}) \in \mathcal{X}^2$, $p_{X|\widehat{X}} = p_{X|\widehat{X}}(x|\widehat{x})p_{\widehat{X}}(\widehat{x})$. Sea también $E = \mathbf{1}_{\{X \neq \widehat{X}\}}$, una variable aleatoria binaria con pmf p_E tal que $p = p_E(1) = 1 - p_E(0)$. Luego:*

$$H(X|\widehat{X}) \leq H(E) + p\log(|\mathcal{X} - 1|). \tag{D.42}$$

Prueba:

$$
\begin{aligned}
H(X|\widehat{X}) &= H(X|\widehat{X}) + H(E|X,\widehat{X}) && \text{(D.43)}\\
&= H(E,X|\widehat{X}) && \text{(D.44)}\\
&= H(E|\widehat{X}) + H(X|E,\widehat{X}) && \text{(D.45)}\\
&\leq H(E) + H(X|E,\widehat{X}) && \text{(D.46)}\\
&= H(E) + \sum_{\widehat{x}\in\mathcal{S}_{\widehat{X}}} p_{E,\widehat{X}}(0,\widehat{x}) H(X|E=0,\widehat{X}=\widehat{x}) && \text{(D.47)}\\
&\quad + p_{E\widehat{X}}(1,\widehat{x}) H(X|E=1,\widehat{X}=\widehat{x}) && \text{(D.48)}\\
&= H(E) + \sum_{\widehat{x}\in\mathcal{S}_{\widehat{X}}} p_{E\widehat{X}}(1,\widehat{x}) H(X|E=1,\widehat{X}=\widehat{x}) && \text{(D.49)}\\
&\leq H(E) + \sum_{\widehat{x}\in\mathcal{S}_{\widehat{X}}} p_{E\widehat{X}}(1,\widehat{x}) \log(|\mathcal{X}-1|) && \text{(D.50)}
\end{aligned}
$$

$$
\begin{aligned}
&= H(E) + \log(|\mathcal{X}-1|) \sum_{\widehat{x}\in\mathcal{S}_{\widehat{X}}} p_{E\widehat{X}}(1,\widehat{x}) && \text{(D.51)}\\
&= H(E) + p_E(1)\log(|\mathcal{X}-1|) && \text{(D.52)}\\
&= H(E) + p\log(|\mathcal{X}-1|), && \text{(D.53)}
\end{aligned}
$$

donde, (D.43) se obtiene debido a que el valor de la variable aleatoria E es conocido dado el conocimiento de las variables aleatorias X y $\widehat{X}$, es decir, $H(E|X,\widehat{X}) = 0$; (D.46) se obtiene condicionamiento no incrementa la entropía por el Lema 4.1; (D.47) se obtiene de aplicar (4.3) y la pmf condicional $p_{E|\widehat{X}} : 0,1 \times \mathcal{X} \to (0,1]$; (D.49) se obtiene de considera que si $E = 0$ el valor de la variable aleatoria X es conocido dado el conocimiento de la variable aleatoria $\widehat{X}$, es decir, $H(X|E=0,\widehat{X}=\widehat{x}) = 0$; (D.50) se obtiene del hecho que dado $E = 1$ y $\widehat{X} = \widehat{x}$, X puede tomar cualquier valor de $\mathcal{X}-1$ valores y se puede obtener un límite superior en la entropía suponiendo la pmf p_X esta distribuida uniformemente, i.e.,

$$H(X|E=1,\widehat{X}=\widehat{x}) \leq \log(|\mathcal{X}-1|), \qquad \text{(D.54)}$$

por el lema B.17; y (D.53) se obtiene dado que $p = \Pr[X \neq \widehat{X}] = p_E(1)$.

Esto completa la prueba del Lema D.8 [54].

■