Timothy Chung

COMP70103

Statistical Information Theory

AUTUMN 2024

Contents

1
Introduction to Statistical Information Theory

1.1 *What is Information?*

The first transistor was invented in 1947 at Bell Laboratories (Bell Labs), a research lab owned by AT&T. It was the result of a collaboration between three scientists: John Bardeen, Walter Brattain, and William Shockley. The transistor, a small device capable of amplifying and switching electronic signals, replaced the bulky and less efficient vacuum tubes that were commonly used in radios and televisions.

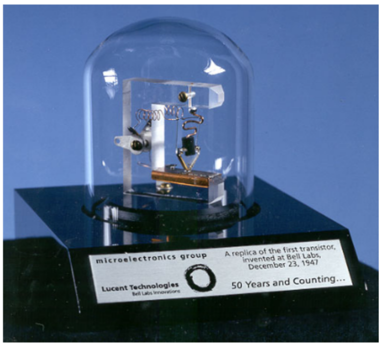


Figure 1.1: A replica of the first transistor, invented at Bell Labs in 1947.

Transistors could send messages to each other – but what exactly is the ‘information’ being transmitted?

In very early times, this information was identified with a physical object that carried it, e.g. a letter or carrier piegeon. It was later then carried by sound (military drums), and sometimes light (semaphores), and then electronic waves (radio).

Figure 1.2: Drummer boys during the Civil War. Before radios and modern communication devices, verbal commands were difficult to hear on noisy battlefields. Drummer boys used rhythmic beats to communicate orders such as advance, retreat, or charge.

It was Claude Shannon who turned the concept of information (which was a qualitative intuition) into a quantitative, mathematical construct that allows for optimisability.

1.2 *Enter Information Theory*

Fundamental Problem of Communication **Definition 1.2.1**

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. – Claude Shannon, 1948

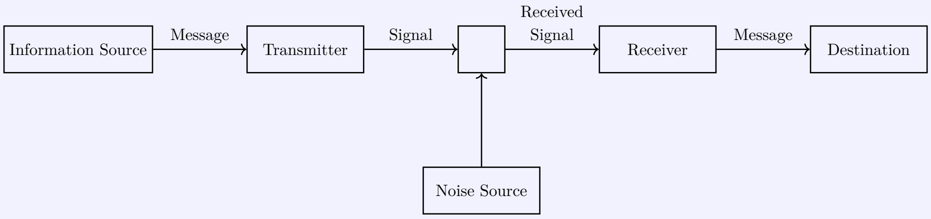


Figure 1.3: Semaphore signalling is a method of visual communication using flags or lights. The system was invented in the 1790s by Claude Chappe, and was used for ships to communicate without radios. Semaphore flag signalling is still taught and used today in some maritime operations and by certain organisations like the Boy Scouts.

Upon successful completion of this module you will be able to:

- Explain the key properties of information metrics in terms of communication principles.
- Calculate these metrics in common probability distributions.
- Compare different metrics and coding schemes on real-world data.
- Analyse the mathematical connections between data transmission and model fitting.
- Design principled data analysis plans with appropriate statistical criteria.

Part I: Source Coding (Compression)

- Information and probability
- Entropy and compression
- Source coding theorem

Part II: Channel Coding (Transmission)

- Mutual information and multivariate probability
- Channel codes
- Channel coding theorem

Part III: Statistics and Advanced Topics

- Model comparison and inference
- Estimating information from data
- Information decomposition

Main Textbook **Reference 1.2.1**

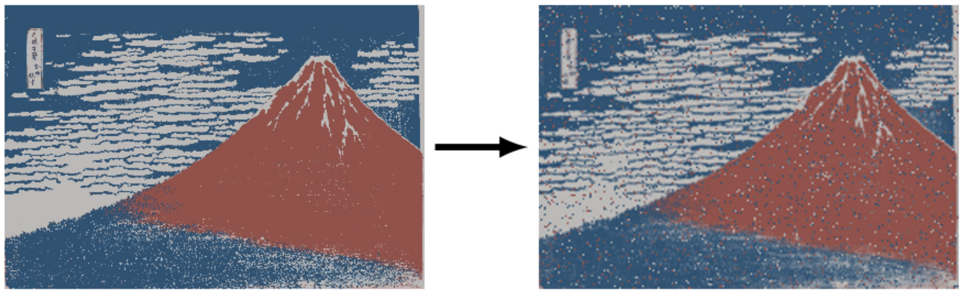
Information Theory, Inference, and Learning Algorithms, by David MacKay.
`http://www.inference.phy.cam.ac.uk/mackay/itila`

Additional Books **Reference 1.2.2**

Elements of Information Theory, by T. Cover & J. Thomas.
Mathematics for Machine Learning, by M. Deisenroth et al.
Information Theory: A Tutorial Introduction, by J. Stone.

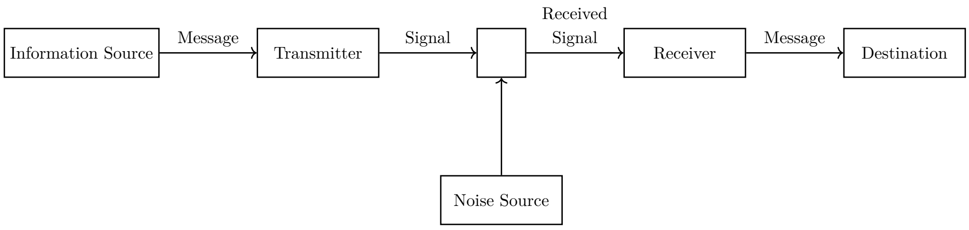
1.3 *Sending a Message Reliably*

- Alice goes to Japan and wants to send Bob a picture.
- Her connection is bad so it corrupts the image.



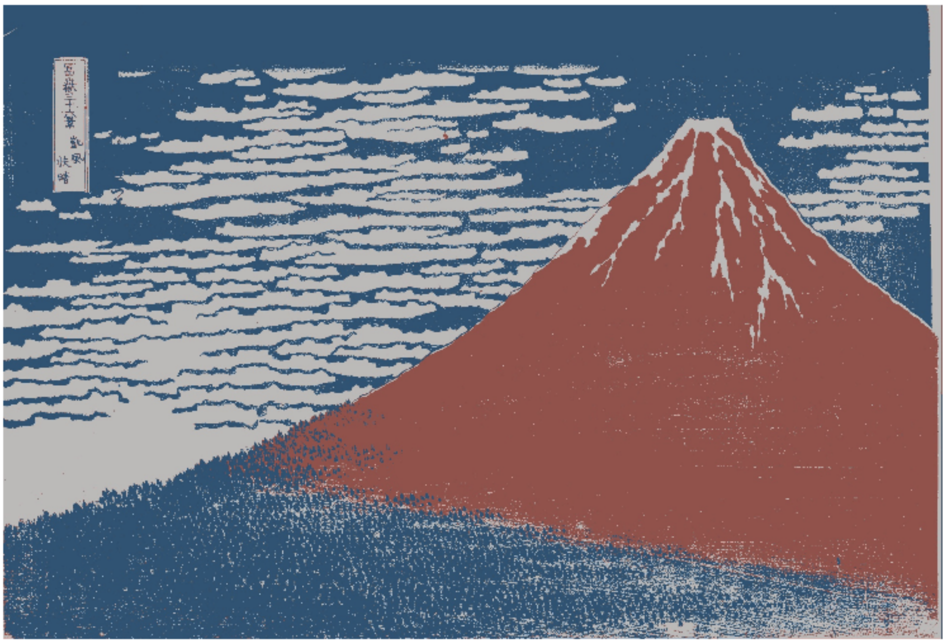
- **A physical solution:**
  - Using a more reliable device.
  - Moving to a zone with better signal.
  - Communicating via cable.
- **A system solution:**

- Build a system that can **detect** and **correct** errors. This is what we want.



1.4 A System Solution Example

1.4.1 Source Coding



We can build a map (code) of symbols to the data (red, blue, white). Since we are working with electronics, the symbols we are allowed to transmit are (0,1), a binary.

A good code should be: **Definition 1.4.1**

1. Short (i.e. few bits per pixel).
2. Easily decodable (i.e. easy to invert).

1. **Naive method: one-hot encoding.**
   - On average, 3 bits per pixel.

$$\begin{array}{rcl} \text{Blue} & \longrightarrow & 010 \\ \text{Red} & \longrightarrow & 001 \\ \text{White} & \longrightarrow & 100 \end{array}$$

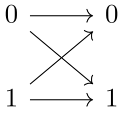
2. **Let's do better: Let more frequent symbols get shorter words (Huffman coding).**
   - Assume that blue pixels occur most often, followed by red, and then white.
   - On average, 1.56 bits per pixel.

$$\begin{array}{rcl} \text{Blue} & \longrightarrow & 0 \\ \text{Red} & \longrightarrow & 10 \\ \text{White} & \longrightarrow & 11 \end{array}$$

1.4.2 Channel Coding

- Alice encodes the image into 0's and 1's and sends it over a channel.
- The channel has a probability f of flipping the message.

We can assign the probabilities with the following equations: [1]



$$P(y = 0 \mid x = 0) = 1 - f; \qquad P(y = 0 \mid x = 1) = f;$$
$$P(y = 1 \mid x = 0) = f; \qquad P(y = 1 \mid x = 1) = 1 - f$$

- One way we can reduce the error rate by flipping is the repeat the code, sending the same message multiple times.

Source sequence (s)	**Transmitted sequence (t)**
0	000
1	111

- The received sequence ($\mathbf{r}$) is transmission ($\mathbf{t}$) plus noise ($\mathbf{n}$):

$$\begin{array}{c|ccccccc} s & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ t & \overbrace{000} & \overbrace{000} & \overbrace{111} & \overbrace{000} & \overbrace{111} & \overbrace{111} & \overbrace{000} \\ n & 000 & 001 & 000 & 000 & 101 & 000 & 000 \\ r & 000 & 001 & 111 & 000 & 010 & 111 & 000 \end{array}$$

- When we receive the transmitted signal, we require a **decoder** to reconstruct the source signal.
- The optimal decoder is Bayes' rule:

$$\hat{s} = \arg\max_s P(s \mid r_1 r_2 r_3) = \arg\max_s \frac{P(r_1 r_2 r_3 \mid s) P(s)}{P(r_1 r_2 r_3)}$$

- The optimal decoder is just the majority vote.
- Probability of making a **bit error** is:

$$p_b = \text{probability of two flips} + \text{probability of three flips} = 3f^2 + f^3$$

- The error probability has decreased from $\mathcal{O}(f)$ to $\mathcal{O}(f^2)$ at the cost of reducing the communication rate by 3.

- **Shannon's Channel Coding Theorem** shows that for rates $R < C$ (channel capacity), error-free communication is possible with arbitrarily small bit error probability p_b.
- As shown in Figure 1.4, the *achievable* region corresponds to rates where noise can be corrected with proper coding, while the *non-achievable* region corresponds to rates $R > C$, where reliable communication is impossible.
- **Key idea**: By using longer codewords and random coding, we can reduce p_b significantly, even in noisy channels, for rates less than C.

[1] This is known as the binary symmetric channel. Another view on probabilities:

x	y	$P(y \mid x)$
0	0	$1 - f$
0	1	f
1	0	f
1	1	$1 - f$

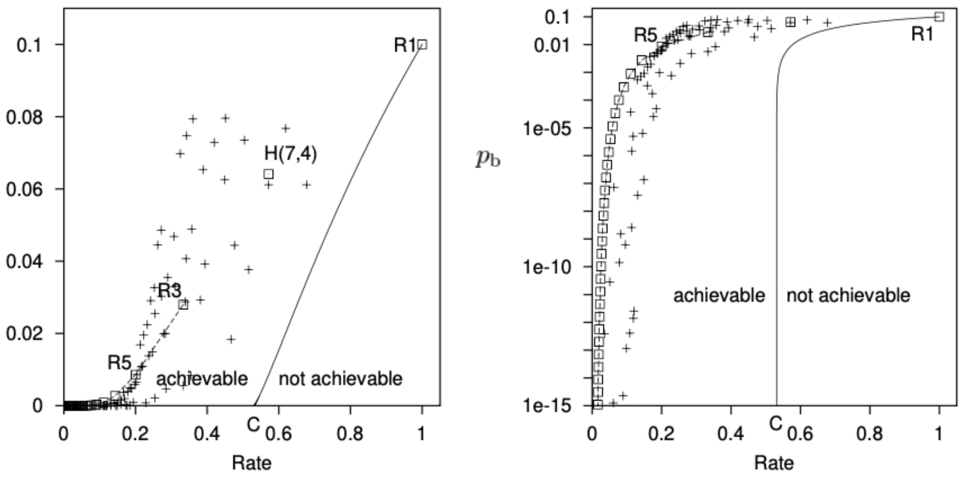


Figure 1.4: Achievable regions for error-free communication.

- **Trade-off**: Lower rates allow more bits for error correction, enabling error-free transmission at the cost of reducing the communication rate.

Key takeaways **Intuition 1.4.1**

Information theory is *the art of redundancy*:

- removing redundancy when it hurts (compression),
- and adding redundancy when it helps (channel coding).

Information theory tells you what a statistical model *can and cannot do*.

- Information and statistics are *two sides of the same coin*.

1.5 Recap on Math

1.5.1 Sets and Functions

- A **set** is a mathematical word for ‘a collection of things’.

Example Sets

Weekdays: $W := \{\text{Mon, Tue, Wed, Thu, Fri, Sat, Sun}\}$.
Weekend: $E := \{\text{Sat, Sun}\}$.

- **Common set operations**:
  - **Membership**: $s \in S \iff s$ is an element of S.
  - **Subset**: $R \subset S \iff$ all elements in R are in S.
  - **Complement**: S^c is the set of things not in S.
  - **Cardinality**: $|S|$ is the number of elements in S.

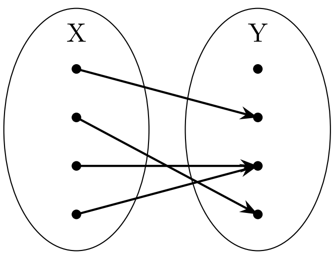
All fun and games until...
$$R = \{S \mid S \notin S\}$$
This is the Russell’s Paradox, which posits that self-referencing sets create contradictions.

Functions

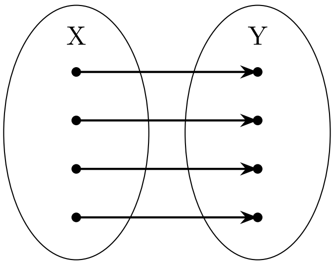
- A **function** $f : X \to Y$ assigns an element of Y to each element of X.
- A function is **injective** if $f(x_1) = f(x_2) \implies x_1 = x_2$.
- A function is **bijective** if it’s injective and $\forall y \in Y$, $\exists x \in X$ such that $f(x) = y$.

Bijective functions are *invertible*, meaning you can find a unique element in X for each element in Y.

Non-injective function



Bijective function



1.5.2 Logarithms

- The **logarithm** of a number x, with base b, is the power to which b must be raised to equal x. That is:
$$x = b^{\log_b(x)}$$
- In simpler terms, the logarithm is the inverse of the exponential function. For example, $\log_2(8) = 3$ because $2^3 = 8$.
- Unless otherwise stated:
  - $\log$ means $\log_2$
  - $\ln$ means $\log_e$

Logarithms have many useful properties, such as:

$$\begin{aligned}
\log_b(xy) &= \log_b(x) + \log_b(y) && \text{(product rule)}\\
\log_b(x^n) &= n\log_b(x) && \text{(power rule)}\\
\log_b(x) &= \frac{\log_a(x)}{\log_a(b)} && \text{(change of base formula)}\\
\log_b(1) &= 0 && \text{(logarithm of 1)}
\end{aligned}$$

Note 1.5.1 Potential Pitfall

$\log_b(x+y)$ is in general *not* simplifiable.

Example 1.5.1 (log(8))

$\log(8) = 3$, since $2^3 = 8$.

Example 1.5.2 (log(1/8))

$\log(1/8) = -3$, since $\log(1/8) = \log(8^{-1}) = -\log(8)$.

Example 1.5.3 (ln(2))

$\ln(2) \approx 0.69$ – you will be seeing this a lot!

1.5.3 Concavity and Convexity

A function is **convex** *iff* a line between two points does not cross the function. Mathematically, this is represented as:

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

for all $x_1, x_2 \in \mathbb{R}$ and $\lambda \in [0,1]$.

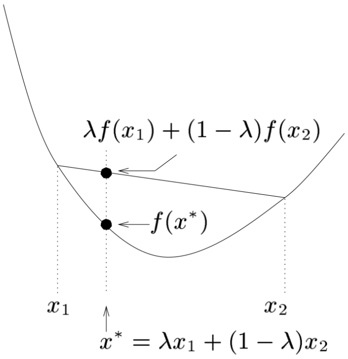

$\lambda f(x_1) + (1-\lambda) f(x_2)$
$f(x^*)$
x_1
x_2
$x^* = \lambda x_1 + (1-\lambda) x_2$

Figure 1.5: A convex function.

Similarly, a **concave** function is one where the inequality is reversed.

Note 1.5.2 Concave or Convex?

The terms convex and concave can be a bit confusing. We will specify:

Concave-up / Convex-down ($\cup$) vs Concave-down / Convex-up ($\cap$)

In the grand scheme of things, convex refers to $\cup$ and concave refers to $\cap$.

Personally, the whole up-down mess is more a semantic issue, with elementary calculus textbooks using ‘concave up’ and ‘concave down’ to describe ‘convex’ and ‘concave’ functions respectively. For a longer thread: Math StackExchange.

To prove a function is convex, you can use the following approaches:

- **Option 1:** Show the second derivative does not change sign.
- **Option 2:** Check if the function has *convexity-preserving operations*:
  - **Affine mapping:** If $f(x)$ is convex, then $f(Ax+b)$ is also convex.
  - **Non-negative weighted sum:** If $f_i(x)$ are convex and $w_i \geq 0$, then $\sum_i w_i f_i(x)$ is convex.
  - Many more.

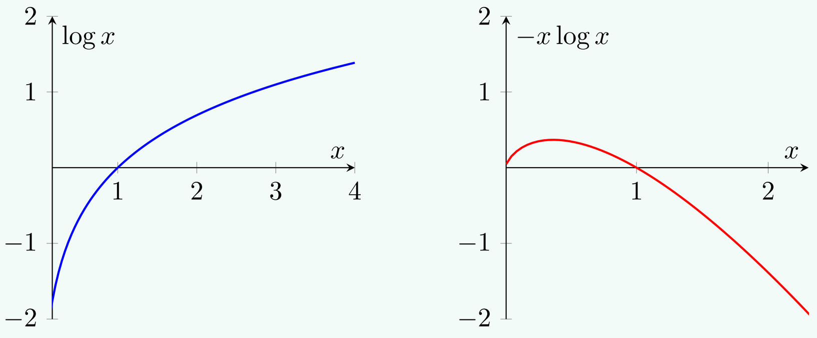
Convex Optimization by Boyd & Vandenberghe **Reference 1.5.1**

Convex Optimization is a highly recommended textbook for further reading.

1.5.4 Concavity and Convexity

Example 1.5.4 (Example: Concavity of the Logarithm)

The log and $-x \log x$ are concave-down ($\cap$) functions.



1.5.5 Limits

(ϵ, δ) Limit for 1D Functions **Definition 1.5.1**

A function $f(x)$ tends to L as x tends to p, written as:

$$\lim_{x \to p} f(x) = L$$

This means that for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all x, if $0 < |x - p| < \delta$, then $|f(x) - L| < \epsilon$.
The (ϵ, δ)-definition provides a rigorous way of understanding limits by ensuring that $f(x)$ gets arbitrarily close to L when x is sufficiently close to p, within the range controlled by δ.

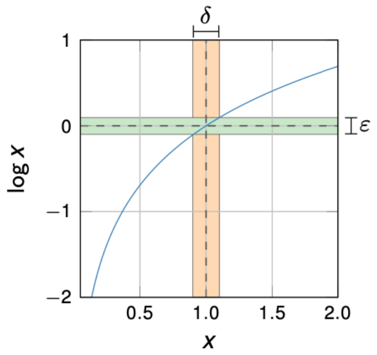


Figure 1.6: Visual representation of the (ϵ, δ)-definition for limits using $\log x$.

Example 1.5.5 (Example: $\lim_{x \to 1} \ln(x) = 0$)

To make $\ln(x)$ within ϵ of 0, x must be within δ of 1.

- The closer x is to 1, the closer $\ln(x)$ is to 0.
- We choose δ to control how close x must be to 1.

Infinite Limit Definition **Definition 1.5.2**

A function $f(x)$ tends to ∞ as x approaches p, written as:

$$\lim_{x \to p} f(x) = \infty$$

This holds if, for every $N > 0$, there exists a $\delta > 0$ such that whenever $0 < |x - p| < \delta$, we have $f(x) > N$. The closer x is to p, the larger $f(x)$ becomes, growing beyond any pre-specified threshold N.

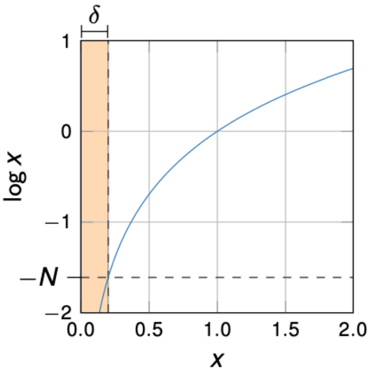


Figure 1.7: Visualisation of the infinite limit behaviour using $\log x$ as x approaches 0.

Example 1.5.6 (Example: $\lim_{x \to 0^+} \ln(x) = -\infty$)

To make $\ln(x)$ smaller than $-N$, x must be within δ of 0 from the right.

- As x approaches 0 from the positive side, $\ln(x)$ decreases without bound.

Limit at Infinity **Definition 1.5.3**

A function $f(x)$ tends to L as x tends to ∞, written as:

$$\lim_{x \to \infty} f(x) = L$$

This means that for every $\epsilon > 0$, there exists a constant $c > 0$ such that whenever $x > c$, we have $|f(x) - L| < \epsilon$. In other words, as x becomes arbitrarily large, $f(x)$ gets closer and closer to L.

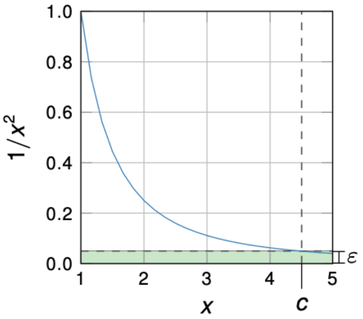


Figure 1.8: Graph showing the limit at infinity using $1/x^2$.

Example 1.5.7 (Example: $\lim_{x \to \infty} \frac{1}{x^2} = 0$)

To make $\frac{1}{x^2}$ within ϵ of 0, x must be larger than some constant c.

- The larger x is, the closer $\frac{1}{x^2}$ gets to 0.

1.5.6 Random Variables

A **random variable** is a mathematical concept used to represent any outcome that can take multiple values due to inherent randomness. Simply put, it is a ”thing” that can take different values for reasons we may not fully know or control.

Random Variable Definition **Definition 1.5.4**

A random variable X is a measurable function $X : \Omega \to E$, where:

- Ω is the **sample space**, representing all possible outcomes of an experiment.
- E is the **event space**, representing all possible measurements associated with the outcomes.
- The random variable assigns a measurable outcome from Ω to a corresponding value in E.

This formal definition ensures that random variables are mathematically well-defined and that probabilities can be consistently assigned to the different values the variable might take.

Example 1.5.8 (Examples of Random Variables)

- A **coin flip**, which can land heads or tails. Here, the random variable could represent the outcome (heads or tails).
- The **result of an election** before votes are cast. The random variable might represent the number of votes for each candidate.
- The **questions on an exam**. The random variable could model which questions are chosen randomly from a pool of possibilities.

Ingredients of a Random Variable **Definition 1.5.5**

To define a random variable, we need three key components:

- The **sample space** (Ω): the set of all possible outcomes of an experiment.
- The **event space** (E): the set of all possible measurements or values associated with each outcome.
- The **probability function** (P): assigns a probability to each outcome in the sample space. This ensures that the likelihood of different outcomes can be quantified.

Example 1.5.9 (Example: Sequence of Coin Flips)

Consider an experiment where we toss three coins and count the number of heads:

- The sample space is $\Omega = \{H, T\}^3$, representing all possible combinations of heads and tails.
- The event space is $E = \{0, 1, 2, 3\}$, where each value corresponds to the number of heads observed.
- The random variable X is the **heads-counting function**, mapping outcomes to the number of heads:

$$X(ttt) = 0, \quad X(tht) = 1, \quad X(thh) = 2, \quad X(hht) = 2, \ldots$$

One important requirement of any probability function associated with random variables is that the probabilities must be non-negative and must sum to 1.

$$\forall \omega \in \Omega, \quad P(\omega) \geq 0 \quad \text{and} \quad P(\Omega) = 1$$

This ensures that every possible outcome has a valid probability assigned to it, and that the total probability over all possible outcomes is 1, representing certainty that some outcome in Ω will occur.

1.5.7 Types of Random Variables

Random variables can be classified based on the nature of the event space E:

- **Discrete:** If E is countable and finite, we call the random variable *discrete*. This means that the set of possible values that the random variable can take is finite or countable.
- **Continuous:** If $E \subseteq \mathbb{R}^d$, the random variable is *continuous*. This means that the possible values lie in a continuous range, such as all real numbers between certain bounds.

We often refer to the event space E as the **alphabet** of the random variable X, and denote it as $\mathcal{X}$. This alphabet represents the set of all possible values that the random variable can take.

Notation Overview:
To simplify our discussions about random variables, we will use the following notation:

Symbol	**Meaning**
X (upper case)	Random variable
x (lower case)	Particular value (or realisation) of X
$\mathcal{X}$ (squiggly)	Alphabet (possible values of x)
$X \sim p(x)$	X is distributed according to p
$p(X = x_i), p(x_i), p_i$	Probability of event x_i

Table 1.1: Notation for Random Variables

Key Points to Remember: **Intuition 1.5.1**

- $\mathcal{X}$ is the set of all possible outcomes (values) of a random variable X. It can either be discrete or continuous.
- X represents the random variable as a whole, while x represents a specific realisation or value that X can take.
- The distribution $p(x)$ tells us how likely it is for X to take each value in $\mathcal{X}$.

1.5.8 Probability Distributions

Random variables can be described by their **probability distributions**. The type of probability distribution depends on the nature of the random variable X:

- **Discrete:** For a discrete random variable, we use the **probability mass function** (PMF). The total probability for all possible values of X must sum to 1:

$$\sum_{x \in \mathcal{X}} P(X = x) = 1$$

- **Continuous:** For a continuous random variable, we use the **probability density function** (PDF). The integral of the PDF over the entire range of possible values must equal 1:

$$\int_{\mathcal{X}} p(x)\, dx = 1$$

Both of these are often referred to as the **probability distribution function** (PDF) in a general sense, though it should be clear from the context whether we mean a probability mass function (discrete) or a probability density function (continuous).

Note 1.5.3 Probability Higher than 1?

For **density** functions, it is possible for the function value to be greater than 1 at certain points. However, the integral over any interval $a \leq x \leq b$ must still be less than or equal to 1:

$$\int_a^b p(x)\, dx \leq 1$$

This ensures that the total probability remains valid, even if the function exceeds 1 at specific values of x.

1.5.9 Expected Value

The **expected value** of a random variable X is the long-run average value that X takes when considering its probability distribution. For a given random variable $X \sim p(x)$ and a function f, the expected value of $f(X)$ under the distribution p is:

$$\mathbb{E}_p[f(X)] := \sum_{x \in \mathcal{X}} p(x) f(x)$$

When the probability distribution is clear from context, we can omit p and simply write the expectation as $\mathbb{E}[f(X)]$.

- **Mean of X:** The mean is the expected value of the random variable itself:

$$\mathbb{E}[X]$$

- **Variance of X:** The variance measures the spread of the random variable around its mean:

$$\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2$$

1.5.10 Probability Distributions

Theorem 1.5.1 Jensen's Inequality

Let f be a convex function and let X be a random variable with a probability distribution $p(x)$. Jensen's inequality states that:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Proof (for the Discrete Case):

Let X be a discrete random variable taking values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, where $0 \leq p_i \leq 1$ and $\sum_{i=1}^n p_i = 1$.
We will prove the inequality by induction on n, the number of possible values of X.

Base Case ($n = 1$):

When $n = 1$, X takes a single value x_1 with probability $p_1 = 1$. Then:

$$\mathbb{E}[f(X)] = f(x_1), \quad \mathbb{E}[X] = x_1, \quad \text{so} \quad \mathbb{E}[f(X)] = f(\mathbb{E}[X])$$

Thus, the inequality holds with equality.

Inductive Step:

Assume that Jensen's inequality holds for any discrete random variable taking $n-1$ values.
Consider X taking n values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$.
The expected value of X is:

$$\mathbb{E}[X] = p_1 x_1 + \sum_{i=2}^{n} p_i x_i \quad \text{where } p_1 \in [0, 1]$$

Define $P = \sum_{i=2}^{n} p_i = 1 - p_1$ and $\mu = \frac{1}{P} \sum_{i=2}^{n} p_i x_i$. Then:

$$\mathbb{E}[X] = p_1 x_1 + P\mu \quad \text{where } p_1, P \in [0, 1]$$

By the convexity of f and knowing that $p_1 + P = 1$, we have:

$$\begin{aligned} f(\mathbb{E}[X]) &= f(p_1 x_1 + P\mu) \\ &\leq p_1 f(x_1) + P f(\mu) \quad \text{(by convexity of } f\text{)} \end{aligned}$$

By the induction hypothesis, applied to the $n - 1$ values $x_2, x_3, \ldots, x_n$ with adjusted probabilities $\frac{p_i}{P}$ for $i = 2, \ldots, n$, we have:

$$\sum_{i=2}^{n} \frac{p_i}{P} f(x_i) \leq f\left(\frac{1}{P} \sum_{i=2}^{n} p_i x_i\right) = f(\mu)$$

Multiplying both sides by P:

$$\sum_{i=2}^{n} p_i f(x_i) \leq P f(\mu)$$

Adding $p_1 f(x_1)$ to both sides:

$$\sum_{i=1}^{n} p_i f(x_i) \leq p_1 f(x_1) + P f(\mu)$$

From the earlier inequality:

$$f(\mathbb{E}[X]) \leq p_1 f(x_1) + P f(\mu) \leq \sum_{i=1}^{n} p_i f(x_i) = \mathbb{E}[f(X)]$$

Thus, the inequality holds for n values.

1.5.11 The Normal Distribution

Definition: Normal (or Gaussian) Distribution **Definition 1.5.6**

The probability density function (PDF) for the d-dimensional normal distribution with mean μ and covariance Σ, denoted by $\mathcal{N}(\mu, \Sigma)$, is:

$$p(x; \mu, \Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

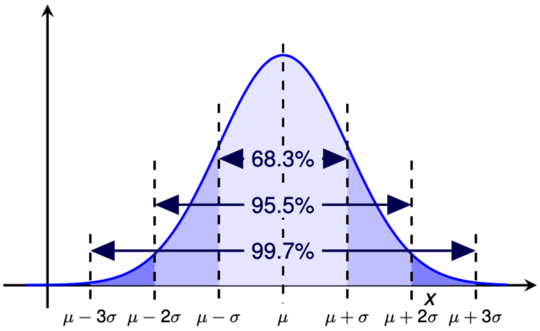


Figure 1.9: The Normal Distribution curve, with highlights of the percentages of data within 1σ, 2σ, and 3σ from the mean.

1.6 Summary

- **Course aims:**
  - Information theory as a unified language for statistics, communications, and geometry.
  - Useful for studying the limitations and possibilities of statistical models.
- **Relevant mathematical concepts:**
  - Probability distributions and random variables.
  - Convexity/concavity and Jensen's inequality.

2

Entropy and Its Interpretations

Consider a random variable X taking values in an alphabet $\mathcal{X} = \{x_1, \ldots, x_l\}$.

Definition: Information Content **Definition 2.0.1**

The *information content* (also known as surprise or surprisal) of a realisation x is defined as:

$$h(x) := \log \frac{1}{p(x)}$$

Definition: Entropy **Definition 2.0.2**

The *entropy* of a random variable X is given by:

$$H(X) := \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$

Entropy represents the expected value of the information content (or surprisal) across all possible realisations of X:

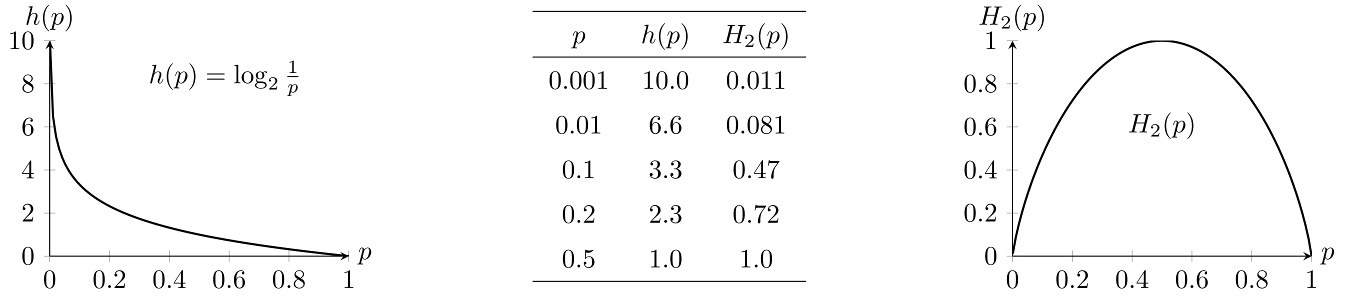
$$H(X) = \mathbb{E}[h(X)] = \sum_{x} p(x)h(x)$$

2.1 Entropy of a Biased Coin

Let X be a coin flip from a biased coin with alphabet $\mathcal{X} = \{H, T\}$ and probabilities $(p, 1-p)$. The entropy of X is given by the binary entropy function:

This is the expected value of surprisal when flipping a biased coin.

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



p	$h(p)$	$H_2(p)$
0.001	10.0	0.011
0.01	6.6	0.081
0.1	3.3	0.47
0.2	2.3	0.72
0.5	1.0	1.0

Figure 2.1: Unlikely outcomes provide more information – but if the outcomes of the coin are skewed towards one side, the overall entropy decreases, because the extra information gained by the rarer outcome is diminished by the expectation, where it doesn't happen often enough, reducing the total expected surprise.

Matehmatically, a biased coin flip X is just the **Bernoulli distribution** with parameter p:

$$X \sim \mathcal{B}(p)$$

Note 2.1.1 The units of $h(x)$

The units of $h(x)$ depend on the base of the log– $\log_2$ is for bits, $\log_{10}$ is for dits, and $\log_e$ is for nats.

Terminology of 'Bit' **Definition 2.1.1**

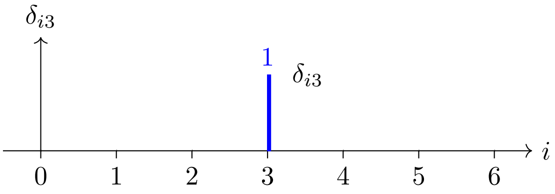
We denote two things with **bit**:

1. The units of information content, taken with $\log_2$.
2. A random variable with alphabet $\{0, 1\}$.

2.2 Properties of Entropy in Discrete Distributions

- The entropy of discrete distributions is non-negative.
- The general bound on discrete entropy is $0 \leq H(X) \leq \log |\mathcal{X}|$.
- Entropy is minimised for the Kronecker delta distribution, (i.e. $p_i = 1, p_{j \neq i} = 0$). For example, take the Kronecker delta function δ_{i3}:

$$\delta_{i3} = \begin{cases} 1 & \text{if } i = 3 \\ 0 & \text{otherwise} \end{cases}$$



Calculating entropy for this distribution:

$$H(X) = \left(1 \cdot \log 1 + \sum_{i \neq 3} 0 \cdot \log 0 \right) = 0$$

- Entropy is maximised for the uniform distribution

Impossible Events– What if $p_j = 0$? **Intuition 2.2.1**

The mathematical convention is to treat $0 \log 0 = 0$ – impossible events do not change entropy.

2.3 Independent Random Variables

What happens if I flip two coins separately? Consider two independent flips X and Y. Recall for independent RVs, $P(X, Y) = P(X)P(Y)$. The joint entropy of X and Y is given by:

$$\begin{aligned} h(x, y) &= -\log P(x, y) \\ &= -\log \big(P(x)P(y)\big) \\ &= -\log P(x) - \log P(y) \\ &= h(x) + h(y) \end{aligned}$$

Similarly for entropy, note that **entropy is additive for independent random variables:**

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} P(x,y)h(x,y) \\
&= \sum_{x,y} P(x)P(y)h(x,y) + \sum_{x,y} P(x)P(y)h(y) \\
&= \underbrace{\sum_y \cancel{P(y)}}_{1} \cdot \sum_x P(x)h(x) + \underbrace{\sum_x \cancel{P(x)}}_{1} \cdot \sum_y P(y)h(y) \\
&= H(X) + H(Y).
\end{aligned}$$

Shannon Axioms **Definition 2.3.1**

The surprisal of an event with probability p, $i(p)$ must satisfy:

1. Certain events are unsurprising: $i(1) = 0$.
2. Less probable events are more surprising: $\frac{di}{dp} \leq 0$.
3. Independent events yield the sum of their surprisals:

$$i(p \cdot q) = i(p) + i(q).$$

The only function satisfying these axioms is the negative logarithm:

$$i(p) = h(p) = -\log_b(p)$$

Note: There are many equivalent sets of axioms for entropy.

2.4 Information Gain

2.4.1 A Number Guessing Game

In the game *Who's Who*, one player picks a character, and the other asks yes-or-no questions to guess their identity, such as:

- Are they wearing glasses?
- Do they have a moustache?

To explore a simpler variation, consider a game called *Sixty-Three*:

- One player selects an integer x from 0 to 63.
- The other player asks questions to guess x.

Optimal Questions: One possible set of optimal questions:

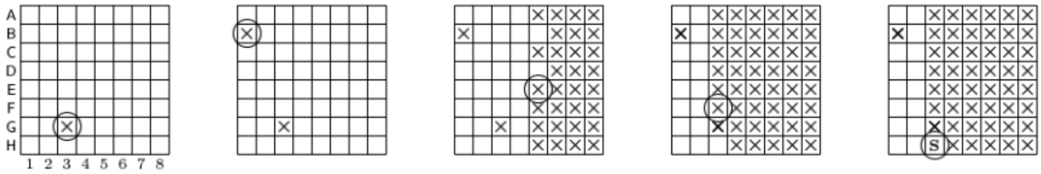
1. Is $x \geq 32$?
2. Is $x \mod 32 \geq 16$?
3. …
4. Is $x \mod 2 = 1$?

This strategy defines a map $\{0, \ldots, 63\} \rightarrow \{0,1\}^6$, where each output bit is the answer to a specific question. For example, $x = 35$ corresponds to 100011.
If x is uniformly distributed, each answer provides 1 bit of information, calculated as $h = -\log(1/2) = 1$ bit. Thus, the total information gained is 6 bits.
Information content corresponds to the length of the binary encoding of x.

2.4.2 A Submarine Guessing Game

- I pick one of 64 squares to hide a submarine.
- In each round, you fire at one square, resulting in a ‘hit’ or ‘miss.’

(The difference with *sixty-three* is that you always fire to just one square.)



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

- **Scenario 1:** You hit the submarine with the first question.
  - You obtain $\log(64) = 6$ bit.
- **Scenario 2:** You miss 32 times, then hit.
  - The first miss gives you $\log(64/63) = 0.0227$ bit.
  - By the 32^{nd} miss, you have

    $$\log(64/63)+\log(63/62)+\cdots+\log(33/32) = 0.0227+0.0230+\cdots+0.0444 = 1 \text{ bit.}$$

  - The hit gives you $\log(32) = 5$ bit, for a total of 6 bit.
- **General case:** You miss $64 - n$, then hit:

$$\underbrace{\log\frac{64}{63} + \log\frac{63}{62} + \cdots + \log\frac{n+1}{n}}_{\text{misses}} + \underbrace{\log\frac{n}{1}}_{\text{hit}} = \log 64 = 6 \text{ bit}$$

- **Conclusion:** Regardless of when you hit, you always get 6 bit.

Conclusions **Intuition 2.4.1**

- For uniform distributions, $h(x)$ corresponds to the length of the **binary representation** of x (i.e., number of yes/no questions).
- $h(x)$ is a measure of the **“intrinsic”** information content of x, regardless of how that information is obtained.
- Information is given by **probability mass exclusions** (Hartley 1928).

The goal of experiments is to **reduce uncertainty** about something by excluding possibilities. The takeaway is that you should design experiemtns with evenly probable outcomes to maximise information gain, as shown in Figure 2.1.

2.5 Entropy in Continuous Distributions

We have explored entropy in discrete RVs with finite alphabets, we then extend this to continuous RVs– by discretising the variable's continous domain.

- We have a random variable $X \in \mathbb{R}$ with PDF $f(x)$.
- We use bins of width Δ to get a discrete variable X^Δ with

$$p_i = \int_{(i-\frac{1}{2})\Delta}^{(i+\frac{1}{2})\Delta} f(x)\,dx = f(x_i)\Delta$$

- Now we take $H(X^\Delta)$ as $\Delta \to 0$:

$$\begin{aligned} H(X^\Delta) &= -\sum p_i \log p_i \\ &= -\sum f(x_i)\Delta \log(f(x_i)\Delta) \\ &= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta \\ &= \underbrace{-\sum \Delta f(x_i) \log f(x_i)}_{\text{Riemann integral}} \quad \underbrace{-\log \Delta}_{\text{Divergent term}} \end{aligned}$$

- Oh no, $\lim_{\Delta \to 0} \log \Delta = -\infty$, so $H(X^\Delta)$ diverges for any $f(x)$.
- A foolproof strategy is to ignore $\log \Delta$ anyway, and define the **differential entropy** as:

Definition: Differential entropy **Definition 2.5.1**

The *differential entropy* of a continuous RV X with PDF $f(x)$ is given by

$$H(X) := -\int f(x) \log f(x)\,dx$$

Note 2.5.1 Warning

The $\log \Delta$ will come back to haunt us! Differential entropy lacks many interesting properties of discrete entropy.

Disclaimer: We may use the term 'entropy' for continuous RVs too.

2.5.1 Properties of Differential Entropy

Take an example of differential entropy in uniform distributions: Let X be a uniform random variable in the interval of length a.

$$X \sim \mathcal{U}([0,a]) \quad \text{i.e.} \quad p(x) = \frac{1}{a}$$

$$H(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a}\,dx = -\log \frac{1}{a} \int_0^a \frac{dx}{a} = \log a$$

Conclusions: We see that Differential entropy:

- Can be negative, e.g. for $a < 1$.
- Grows with the volume of the distribution ($2^{H(X)} = a > 0$)

2.5.2 Surprisal in Gaussian Distributions

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a 1D Gaussian random variable, i.e.

$$p(x) = \left(\sigma\sqrt{2\pi}\right)^{-1} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Then, surprisal is just the negative *ln* of the PDF:

$$h(x) = \ln\left(\sigma\sqrt{2\pi}\right) + \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

This expression is equivalent to scaled mean squared error (MSE):

$$h(x) = \ln\sigma + \frac{1}{2\sigma^2}\text{MSE}(x, \mu) + C.$$

If we predict x with a PDF $\mathcal{N}(\mu, \sigma^2)$, we can lower the surprisal by:

- Reducing the bias (i.e., moving μ closer to x),
- (Sometimes) increasing the precision (i.e., increasing σ).

Modelling x

We want to reduce surprisal to ensure we can compactly encode x more efficiently. In this example, we want to model x with the Gaussian distribution. Reducing surprisal is then equivalent to reducing the prediction error, or the mean squared error (MSE).

Try to imagine a distribution of x and then we play around with the parameters of a Gaussian distribution to make it fit the distribution of x better.

If we find a more accurate μ for our model, we then reduce the MSE. And if our μ is already quite close to the true value of x, increasing precision (making reducing σ to make the distribution more narrow, centring closer around μ) can also sometimes reduce surprisal. However, if μ is too far from the true value of x, increasing precision can sometimes increase surprisal.

Intuition 2.5.1

2.5.3 Entropy in Gaussian Distributions

We show that Entropy has a closed-form expression for Gaussian Distributions:
To calculate the entropy of a D-dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$,

$$p(x) = |2\pi\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

The entropy $H(X)$ is given by:

$$H(X) = -\int_{-\infty}^{+\infty} \mathcal{N}(x; \mu, \Sigma) \ln \mathcal{N}(x; \mu, \Sigma)\, dx.$$

Expanding, we get:

$$H(X) = \frac{1}{2}\mathbb{E}[\ln|2\pi\Sigma|] + \frac{1}{2}\mathbb{E}[(x-\mu)^\top\Sigma^{-1}(x-\mu)].$$

For the first term, $\mathbb{E}[\ln|2\pi\Sigma|] = \ln|2\pi\Sigma|$.
For the second term:

$$\mathbb{E}\left[(x-\mu)^\top\Sigma^{-1}(x-\mu)\right] = \text{tr}\left(\Sigma^{-1}\mathbb{E}\left[(x-\mu)(x-\mu)^\top\right]\right) = \text{tr}(\Sigma^{-1}\Sigma) = D.$$

Substituting back, we get:

$$H(X) = \frac{1}{2}\ln|2\pi\Sigma| + \frac{1}{2}D = \frac{1}{2}\ln\left(|2\pi\Sigma|e^D\right).$$

Overall:

$$H(X) = \frac{1}{2}\ln|2\pi e\Sigma|.$$

2.5.4 Takeaways of Entropy

- Entropy quantifies the **average information content** of an observation x.
- Entropy serves as a **generalised variance** (e.g., measures predictability of X).
- For discrete distributions:
  - Entropy is **bounded**: $0 \leq H(X) \leq \log|\mathcal{X}|$.
  - Related to the **number of yes/no questions** needed to determine x.
- For continuous distributions, **differential entropy** is defined, but can be negative and lacks some properties of discrete entropy.

2.6 Source Coding Theorem

Definition: Code and Code Length **Definition 2.6.1**

Given a random variable X with alphabet $\mathcal{X}$ and an alphabet $\mathcal{D}$, a **code** is a mapping

$$C : \mathcal{X} \to \mathcal{D}^*,$$

where $\mathcal{D}^*$ is the set of all finite-length strings of symbols in $\mathcal{D}$.
The quantity $\ell(x)$ is the **code length** of $C(x)$, and $L = \mathbb{E}[\ell(x)]$ the **average code length**.

- For each symbol x, the string $C(x) \in \mathcal{D}^*$ is called a **codeword**.
- When $|\mathcal{D}| = 2$, C is a **binary code**. When $|\mathcal{D}| = d$, C is a **d-ary code**.

Here are two example codes for the alphabet $\mathcal{X} = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$.

Binary	
x	$C(x)$
a	00
b	01
c	10
d	11

Ternary	
x	$C(x)$
a	0
b	1
c	2
d	200

2.6.1 Coding for the Uniform Distribution

- We've seen that (in uniform distributions), entropy corresponds to the number of **yes/no** questions needed to guess x.

$$\begin{array}{rcl rcl} 35 & \Rightarrow & 100011 & 6 & \Rightarrow & 000110 \\ 0 & \Rightarrow & 000000 & 17 & \Rightarrow & 010001 \\ 42 & \Rightarrow & 101010 & \dots & \Rightarrow & \dots \end{array}$$

- This forms a **binary code** for X with length

$$L = H(X) = \log |\mathcal{X}|.$$

- More generally, $\log |\mathcal{X}|$ is referred to as the **raw bit content** of X.

Note 2.6.1 Warning: Ceiling and the "Extra Bit"

When $|\mathcal{X}|$ is not a power of two, we may need an "extra bit" to encode symbols. You may see this written as $L = \lceil \log |\mathcal{X}| \rceil$ or $L = \log |\mathcal{X}| + 1$.

2.6.2 Non-uniform Distributions

What happens when the distribution isn't uniform?

- **Example**: compressing Wikipedia, which has alphabet $\mathcal{X} = \mathcal{E} \cup \mathcal{U}$:
  - $\mathcal{E}$ is the English alphabet, $|\mathcal{E}| = 26$;
  - $\mathcal{U} = \{!, @, \#, \$, \%, -, \&, *, (,)\}$ includes some unicode characters.
- Assume that 98% of Wikipedia content comes from $\mathcal{E}$, and 2% from $\mathcal{U}$.
- A naive binary code of $\mathcal{X}$ would require $\lceil \log |\mathcal{X}| \rceil = \lceil \log 36 \rceil = 6$ bits.
- **But**... if we ignore $\mathcal{U}$, we can encode in $\lceil \log |\mathcal{E}| \rceil = \lceil \log 26 \rceil = 5$ bits.

Note 2.6.2 Reduction in Code Length

We have **reduced the code length** with only **2% error**, simply by refusing to encode the minority of rarely-used symbols. And this generally would not have a major impact on the readability of the text– most information could still be conveyed.

2.6.3 Law of Large Numbers

The Law of Large Numbers states that as the number of independent, identically distributed (i.i.d.) random variables increases, their sample average converges to the expected value. This result forms a cornerstone of probability and statistics, establishing that with a large enough sample size, we can expect the sample mean to approximate the population mean closely.

- Let $Y^n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the mean of n i.i.d. random variables $X_1, \dots, X_n$, with

$$\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] = \mu \quad \text{and} \quad \mathrm{Var}(X_i) = \cdots = \mathrm{Var}(X_n) = \sigma^2.$$

- By direct calculation, it can be shown that:

$$\mathbb{E}[Y^n] = \mu \quad \text{and} \quad \mathrm{Var}(Y^n) = \frac{\sigma^2}{n}.$$

- As $n \to \infty$, Y^n has **vanishing variance**, meaning Y^n becomes increasingly close to μ.

Theorem 2.6.1 Weak Law of Large Numbers

Let $Y^n = \frac{1}{n}\sum_{i=1}^{n} X_i$. As $n \to \infty$, Y^n converges in probability to μ:

$$Y^n \xrightarrow{P} \mu, \quad \text{that is,} \quad \lim_{n\to\infty} \Pr\left(|Y^n - \mu| < \varepsilon\right) = 1$$

for any $\varepsilon > 0$.

3
Reference

Theorem 3.0.1 Theorem Name

This is the statement of the theorem.

```
\thm{Theorem Name}{This is the statement of the theorem.}
```

Corollary 3.0.1 Corollary Name

This is the statement of the corollary.

```
\cor{Corollary Name}{This is the statement of the corollary.}
```

Lemma 3.0.1 Lemma Name

This is the statement of the lemma.

```
\lem{Lemma Name}{This is the statement of the lemma.}
```

Claim 3.0.1 Claim Name

This is the statement of the claim.

```
\clm{Claim Name}{This is the statement of the claim.}
```

Example 3.0.1 (Example Name)

This is the explanation of the example.

```
\ex{Example Name}{This is the explanation of the example.}
```

Note 3.0.1 Side Note Box

This is a side note.

```
\sn{Side Note Box}{This is a side note.}
```

Note 3.0.1 Side Note Box

This is a smaller side note.

```
\sns{Side Note Box}{This is a
smaller side note.}
```

This is a block of highlighted text

```
\hl{This is a block of highlighted text}
```

This is a block of highlighted text that's smaller.

```
\hls{This is a block of
highlighted text that's
smaller.}
```

Definition Title **Definition 3.0.2**

This is an example definition.

```
\defb{Definition Title}{This is an example definition.}
```

Small Def Title

Example Text

```
\defsb{Small Def Title}{Example
Text}
```

Definition 3.0.1

Extra Title ***Non-Examinable* 3.0.2**

This is an example box with extra information.

```
\extrab{Extra Title}{This is an example box with extra
information.}
```

Small Title

Example Text

```
\extrasb{Small Title}{Example
Text}
```

***Non-Examinable* 3.0.1**

Example Title **Example Q 3.0.2**

This is an example question.
`\egb{Example Title}{This is an example question.}`

Answer here
`\tcblower Answer here`

Q1c - 2018 **Exam Q 3.0.2**

This is an example exam question.
`\examb{1c}{2018}{This is an example exam question.}`

Reference Title **Reference 3.0.2**

This is an example reference to source material.
`\refb{Reference Title}{This is an example reference to source material.}`

Intuition Title **Intuition 3.0.2**

This is an example of an intuitive explanation.
`\intuitb{Intuition Title}{This is an example of an intuitive explanation.}`

The front matter of a book refers to all of the material that comes before the main text. The following table from shows a list of material that appears in the front matter of *The Visual Display of Quantitative Information*, *Envisioning Information*, *Visual Explanations*, and *Beautiful Evidence* along with its page number. Page numbers that appear in parentheses refer to folios that do not have a printed page number (but they are still counted in the page number sequence).

	Books			
Page content	*VDQI*	*EI*	*VE*	*BE*
Blank half title page	(1)	(1)	(1)	(1)
Frontispiece[1]	(2)	(2)	(2)	(2)
Full title page	(3)	(3)	(3)	(3)
Copyright page	(4)	(4)	(4)	(4)
Contents	(5)	(5)	(5)	(5)
Dedication	(6)	(7)	(7)	7
Epigraph	–	–	(8)	–
Introduction	(7)	(9)	(9)	9

The design of the front matter in Tufte's books varies slightly from the traditional design of front matter. First, the pages in front matter are traditionally numbered with lowercase roman numerals (*e.g.*, i, ii, iii, iv, ...). Second, the front matter page numbering sequence is usually separate from the main matter page numbering. That is, the page numbers restart at 1 when the main matter begins. In contrast, Tufte has enumerated his pages with arabic numerals that share the same page counting sequence as the main matter.

Small Title

Example Text
`\egsb{Small Title}{Example Text}`

Answer
`\tcblower Answer`

Example Q 3.0.1

Q1c - 2018

This is an example exam question, smaller.
`\examsb{1c}{2018}{This is an example exam question, smaller.}`

Exam Q 3.0.1

Reference Title

This is an example reference to source material.
`\refsb{Reference Title}{This is an example reference to source material.}`

Reference 3.0.1

Intuition Title

This is an example of an intuitive explanation, smaller.
`\intuitsb{Intuition Title}{This is an example of an intuitive explanation, smaller.}`

Intuition 3.0.1

[1] The contents of this page vary from book to book. In *VDQI* this page is blank; in *EI* and *VE* this page holds a frontispiece; and in *BE* this page contains three epigraphs.

On the side note of...

yeah

Non-Examinable **3.0.3**

There are also some variations in design across Tufte's four books. The page opposite the full title page (labeled "frontispiece" in the above table) has different content in each of the books. In *The Visual Display of Quantitative Information*, this page is blank; in *Envisioning Information* and *Visual Explanations*, this page holds a frontispiece; and in *Beautiful Evidence*, this page contains three epigraphs. The dedication appears on page 6 in *VDQI* (opposite the introduction), and is placed on its own spread in the other books. In *VE*, an epigraph shares the spread with the opening page of the introduction.
None of the page numbers (folios) of the front matter are expressed except in *BE*, where the folios start to appear on the dedication page.

THE FULL TITLE PAGE of each of the books varies slightly in design. In all the books, the author's name appears at the top of the page, the title it set just above the center line, and the publisher is printed along the bottom margin. Some of the differences are outlined in the following table.

Feature	*VDQI*	*EI*	*VE*	*BE*
Author				
Typeface	serif	serif	serif	sans serif
Style	italics	italics	italics	upright, caps
Size	24 pt	20 pt	20 pt	20 pt
Title				
Typeface	serif	serif	serif	sans serif
Style	upright	italics	upright	upright, caps
Size	36 pt	48 pt	48 pt	36 pt
Subtitle				
Typeface	–	–	serif	–
Style	–	–	upright	–
Size	–	–	20 pt	–
Edition				
Typeface	sans serif	–	–	–
Style	upright, caps	–	–	–
Size	14 pt	–	–	–
Publisher				
Typeface	serif	serif	serif	sans serif
Style	italics	italics	italics	upright, caps
Size	14 pt	14 pt	14 pt	14 pt

THE TABLES OF CONTENTS in Tufte's books give us our first glimpse of the structure of the main matter. *The Visual Display of Quantitative Information* is split into two parts, each containing some number of chapters. His other three books only contain chapters—they're not broken into parts.

3.1 Typefaces

Tufte's books primarily use two typefaces: Bembo and Gill Sans. Bembo is used for the headings and body text, while Gill Sans is used for the title page and opening epigraphs in *Beautiful Evidence*.
Since neither Bembo nor Gill Sans are available in default LaTeX installations, the

Tufte-LaTeX document classes default to using Palatino and Helvetica, respectively. In addition, the Bera Mono typeface is used for `monospaced` type.
The following font sizes are defined by the Tufte-LaTeX classes:

LaTeX size	Font size	Leading	Used for
`\tiny`	5	6	sidenote numbers
`\scriptsize`	7	8	–
`\footnotesize`	8	10	sidenotes, captions
`\small`	9	12	quote, quotation, and verse environments
`\normalsize`	10	14	body text
`\large`	11	15	B-heads
`\Large`	12	16	A-heads, TOC entries, author, date
`\LARGE`	14	18	handout title
`\huge`	20	30	chapter heads
`\Huge`	24	36	part titles

Table 3.1: A list of LaTeX font sizes as defined by the Tufte-LaTeX document classes.

3.2 Headings

Tufte's books include the following heading levels: parts, chapters,[2] sections, subsections, and paragraphs. Not defined by default are: sub-subsections and subparagraphs.

Heading	Style	Size
Part	roman	24/36×40 pc
Chapter	italic	20/30×40 pc
Section	italic	12/16×26 pc
Subsection	italic	11/15×26 pc
Paragraph	italic	10/14

Table 3.2: Heading styles used in *Beautiful Evidence*.

Paragraph Paragraph headings (as shown here) are introduced by italicized text and separated from the main paragraph by a bit of space.

3.3 Environments

The following characteristics define the various environments:

Environment	Font size	Notes
Body text	10/14×26 pc	
Block quote	9/12×24 pc	Block indent (left and right) by 1 pc
Sidenotes	8/10×12 pc	Sidenote number is set inline, followed by word space
Captions	8/10×12 pc	

Table 3.3: Environment styles used in *Beautiful Evidence*.

[2] Parts and chapters are defined for the `tufte-book` class only.

Column 1	**Column 2**	**Column 3**
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.	Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.	Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.	Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.	Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.	Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.	Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.

Table 3.4: Example table with limited column widths

4
On the Use of the `tufte-book` Document Class

The Tufte-LaTeX document classes define a style similar to the style Edward Tufte uses in his books and handouts. Tufte's style is known for its extensive use of sidenotes, tight integration of graphics with text, and well-set typography. This document aims to be at once a demonstration of the features of the Tufte-LaTeX document classes and a style guide to their use.

4.1 Page Layout

4.1.1 Headings

This style provides A- and B-heads (that is, `\section` and `\subsection`), demonstrated above.
If you need more than two levels of section headings, you'll have to define them yourself at the moment; there are no pre-defined styles for anything below a `\subsection`. As Bringhurst points out in *The Elements of Typographic Style*,[1] you should "use as many levels of headings as you need: no more, and no fewer." The Tufte-LaTeX classes will emit an error if you try to use `\subsubsection` and smaller headings.

IN HIS LATER BOOKS,[2] Tufte starts each section with a bit of vertical space, a non-indented paragraph, and sets the first few words of the sentence in SMALL CAPS. To accomplish this using this style, use the `\newthought` command:

```
\newthought{In his later books}, Tufte starts...
```

4.2 Sidenotes

One of the most prominent and distinctive features of this style is the extensive use of sidenotes. There is a wide margin to provide ample room for sidenotes and small figures. Any `\footnote`s will automatically be converted to sidenotes.[3] If you'd like to place ancillary information in the margin without the sidenote mark (the superscript number), you can use the `\marginnote` command.
The specification of the `\sidenote` command is:

This is a margin note. Notice that there isn't a number preceding the note, and there is no number in the main text where this note was written.

```
\sidenote[⟨number⟩][⟨offset⟩]{Sidenote text.}
```

Both the ⟨*number*⟩ and ⟨*offset*⟩ arguments are optional. If you provide a ⟨*number*⟩ argument, then that number will be used as the sidenote number. It will change of the number of the current sidenote only and will not affect the numbering sequence of subsequent sidenotes.
Sometimes a sidenote may run over the top of other text or graphics in the margin space. If this happens, you can adjust the vertical position of the sidenote by providing a dimension in the ⟨*offset*⟩ argument. Some examples of valid dimensions are:

```
1.0in    2.54cm    254mm    6\baselineskip
```

If the dimension is positive it will push the sidenote down the page; if the dimension is negative, it will move the sidenote up the page.

[1] **Bringhurst2005**

[2] **Tufte2006**

[3] This is a sidenote that was entered using the `\footnote` command.

While both the ⟨*number*⟩ and ⟨*offset*⟩ arguments are optional, they must be provided in order. To adjust the vertical position of the sidenote while leaving the sidenote number alone, use the following syntax:

```
\sidenote[][⟨offset⟩]{Sidenote text.}
```

The empty brackets tell the `\sidenote` command to use the default sidenote number.
If you *only* want to change the sidenote number, however, you may completely omit the ⟨*offset*⟩ argument:

```
\sidenote[⟨number⟩]{Sidenote text.}
```

The `\marginnote` command has a similar *offset* argument:

```
\marginnote[⟨offset⟩]{Margin note text.}
```

4.3 References

References are placed alongside their citations as sidenotes, as well. This can be accomplished using the normal `\cite` command.[4]
The complete list of references may also be printed automatically by using the `\bibliography` command. (See the end of this document for an example.) If you do not want to print a bibliography at the end of your document, use the `\nobibliography` command in its place.
To enter multiple citations at one location,[5] you can provide a list of keys separated by commas and the same optional vertical offset argument: `\cite{Tufte2006,Tufte1990}`.

```
\cite[⟨offset⟩]{bibkey1,bibkey2,...}
```

4.4 Figures and Tables

Images and graphics play an integral role in Tufte's work. In addition to the standard `figure` and `tabular` environments, this style provides special figure and table environments for full-width floats.
Full page–width figures and tables may be placed in `figure*` or `table*` environments. To place figures or tables in the margin, use the `marginfigure` or `margintable` environments as follows (see figure 4.1):

Figure 4.1: This is a margin figure. The helix is defined by $x = \cos(2\pi z)$, $y = \sin(2\pi z)$, and $z = [0, 2.7]$. The figure was drawn using Asymptote (`http://asymptote.sf.net/`).

```
\begin{marginfigure}
    \includegraphics{helix}
    \caption{This is a margin figure.}
    \label{fig:marginfig}
\end{marginfigure}
```

The `marginfigure` and `margintable` environments accept an optional parameter ⟨*offset*⟩ that adjusts the vertical position of the figure or table. See the "Sidenotes" section above for examples. The specifications are:

```
\begin{marginfigure}[⟨offset⟩]
    ...
\end{marginfigure}

\begin{margintable}[⟨offset⟩]
    ...
\end{margintable}
```

Figure **??** is an example of the `figure*` environment and figure **??** is an example of the normal `figure` environment.

[4] The first paragraph of this document includes a citation.

[5] **Tufte2006**, **Tufte1990**

As with sidenotes and marginnotes, a caption may sometimes require vertical adjustment. The \caption command now takes a second optional argument that enables you to do this by providing a dimension ⟨*offset*⟩. You may specify the caption in any one of the following forms:

```
\caption{long caption}
\caption[short caption]{long caption}
\caption[][⟨offset⟩]{long caption}
\caption[short caption][⟨offset⟩]{long caption}
```

A positive ⟨*offset*⟩ will push the caption down the page. The short caption, if provided, is what appears in the list of figures/tables, otherwise the "long" caption appears there. Note that although the arguments ⟨*short caption*⟩ and ⟨*offset*⟩ are both optional, they must be provided in order. Thus, to specify an ⟨*offset*⟩ without specifying a ⟨*short caption*⟩, you must include the first set of empty brackets `[]`, which tell `\caption` to use the default "long" caption. As an example, the caption to figure **??** above was given in the form

```
\caption[Hilbert curves...][6pt]{Hilbert curves...}
```

Table 4.1 shows table created with the `booktabs` package. Notice the lack of vertical rules—they serve only to clutter the table's data.

Margin	Length
Paper width	8½ inches
Paper height	11 inches
Textblock width	6½ inches
Textblock/sidenote gutter	⅜ inches
Sidenote width	2 inches

Table 4.1: Here are the dimensions of the various margins used in the Tufte-handout class.

OCCASIONALLY LaTeX will generate an error message:

```
Error: Too many unprocessed floats
```

LaTeX tries to place floats in the best position on the page. Until it's finished composing the page, however, it won't know where those positions are. If you have a lot of floats on a page (including sidenotes, margin notes, figures, tables, etc.), LaTeX may run out of "slots" to keep track of them and will generate the above error.

LaTeX initially allocates 18 slots for storing floats. To work around this limitation, the Tufte-LaTeX document classes provide a \morefloats command that will reserve more slots.

The first time `\morefloats` is called, it allocates an additional 34 slots. The second time `\morefloats` is called, it allocates another 26 slots.

The `\morefloats` command may only be used two times. Calling it a third time will generate an error message. (This is because we can't safely allocate many more floats or LaTeX will run out of memory.)

If, after using the `\morefloats` command twice, you continue to get the `Too many unprocessed floats` error, there are a couple things you can do.

The \FloatBarrier command will immediately process all the floats before typesetting more material. Since `\FloatBarrier` will start a new paragraph, you should place this command at the beginning or end of a paragraph.

The \clearpage command will also process the floats before continuing, but instead of starting a new paragraph, it will start a new page.

You can also try moving your floats around a bit: move a figure or table to the next page or reduce the number of sidenotes. (Each sidenote actually uses *two* slots.)

After the floats have placed, LaTeX will mark those slots as unused so they are available for the next page to be composed.

4.5 Captions

You may notice that the captions are sometimes misaligned. Due to the way LaTeX's float mechanism works, we can't know for sure where it decided to put a float. Therefore, the Tufte-LaTeX document classes provide commands to override the caption position.

Vertical alignment To override the vertical alignment, use the `\setfloatalignment` command inside the float environment. For example:

```
\begin{figure}[btp]
    \includegraphics{sinewave}
    \caption{This is an example of a sine wave.}
    \label{fig:sinewave}
    \setfloatalignment{b}% forces caption to be bottom-aligned
\end{figure}
```

The syntax of the `\setfloatalignment` command is:

```
\setfloatalignment{⟨pos⟩}
```

where ⟨*pos*⟩ can be either `b` for bottom-aligned captions, or `t` for top-aligned captions.

Horizontal alignment To override the horizontal alignment, use either the `\forceversofloat` or the `\forcerectofloat` command inside of the float environment. For example:

```
\begin{figure}[btp]
    \includegraphics{sinewave}
    \caption{This is an example of a sine wave.}
    \label{fig:sinewave}
    \forceversofloat% forces caption to be set to the left of the float
\end{figure}
```

The `\forceversofloat` command causes the algorithm to assume the float has been placed on a verso page—that is, a page on the left side of a two-page spread. Conversely, the `\forcerectofloat` command causes the algorithm to assume the float has been placed on a recto page—that is, a page on the right side of a two-page spread.

4.6 Full-width text blocks

In addition to the new float types, there is a `fullwidth` environment that stretches across the main text block and the sidenotes area.

```
\begin{fullwidth}
  Lorem ipsum dolor sit amet...
\end{fullwidth}
```

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.7 Typography

4.7.1 Typefaces

If the Palatino, Helvetica, and `Bera Mono` typefaces are installed, this style will use them automatically. Otherwise, we'll fall back on the Computer Modern typefaces.

4.7.2 Letterspacing

This document class includes two new commands and some improvements on existing commands for letterspacing.

When setting strings of ALL CAPS or SMALL CAPS, the letterspacing—that is, the spacing between the letters—should be increased slightly.[6] The `\allcaps` command has proper letterspacing for strings of FULL CAPITAL LETTERS, and the `\smallcaps` command has letterspacing for SMALL CAPITAL LETTERS. These commands will also automatically convert the case of the text to upper- or lowercase, respectively.

The `\textsc` command has also been redefined to include letterspacing. The case of the `\textsc` argument is left as is, however. This allows one to use both uppercase and lowercase letters: THE INITIAL LETTERS OF THE WORDS IN THIS SENTENCE ARE CAPITALIZED.

4.8 Document Class Options

The `tufte-book` class is based on the LaTeX `book` document class. Therefore, you can pass any of the typical book options. There are a few options that are specific to the `tufte-book` document class, however.

The `a4paper` option will set the paper size to A4 instead of the default US letter size.

The `sfsidenotes` option will set the sidenotes and title block in a sans serif typeface instead of the default roman.

The `twoside` option will modify the running heads so that the page number is printed on the outside edge (as opposed to always printing the page number on the right-side edge in `oneside` mode).

The `symmetric` option typesets the sidenotes on the outside edge of the page. This is how books are traditionally printed, but is contrary to Tufte's book design which sets the sidenotes on the right side of the page. This option implicitly sets the `twoside` option.

The `justified` option sets alldocclsoptdef and right). The default is to set the text ragged right. The body text of Tufte's books are set ragged right. This prevents needless hyphenation and makes it easier to read the text in the slightly narrower column.

The `bidi` option loads the `bidi` package which is used with XƎLaTeX to typeset bi-directional text. Since the `bidi` package needs to be loaded before the sidenotes and cite commands are defined, it can't be loaded in the document preamble.

The `debug` option causes the Tufte-LaTeX classes to output debug information to the log file which is useful in troubleshooting bugs. It will also cause the graphics to be replaced by outlines.

The `nofonts` option prevents the Tufte-LaTeX classes from automatically loading the Palatino and Helvetica typefaces. You should use this option if you wish to load your own fonts. If you're using XƎLaTeX, this option is implied (*i. e.*, the Palatino and Helvetica fonts aren't loaded if you use XƎLaTeX).

The `nols` option inhibits the letterspacing code. The Tufte-LaTeX classes try to load the appropriate letterspacing package (either pdfTeX's `letterspace` package

[6] **Bringhurst2005**

or the soul package). If you're using XƎLATEX with fontenc, however, you should configure your own letterspacing.

The notitlepage option causes \maketitle to generate a title block instead of a title page. The book class defaults to a title page and the handout class defaults to the title block. There is an analogous titlepage option that forces \maketitle to generate a full title page instead of the title block.

The notoc option suppresses Tufte-LATEX's custom table of contents (TOC) design. The current TOC design only shows unnumbered chapter titles; it doesn't show sections or subsections. The notoc option will revert to LATEX's TOC design.

The nohyper option prevents the hyperref package from being loaded. The default is to load the hyperref package and use the \title and \author contents as metadata for the generated PDF.

5
Customizing Tufte-LaTeX

The Tufte-LaTeX document classes are designed to closely emulate Tufte's book design by default. However, each document is different and you may encounter situations where the default settings are insufficient. This chapter explores many of the ways you can adjust the Tufte-LaTeX document classes to better fit your needs.

5.1 File Hooks

If you create many documents using the Tufte-LaTeX classes, it's easier to store your customizations in a separate file instead of copying them into the preamble of each document. The Tufte-LaTeX classes provide three file hooks: `tufte-common-local.tex`, `tufte-book-local.tex`, and `tufte-handout-local.tex`.

`tufte-common-local.tex` If this file exists, it will be loaded by all of the Tufte-LaTeX document classes just prior to any document-class-specific code. If your customizations or code should be included in both the book and handout classes, use this file hook.

`tufte-book-local.tex` If this file exists, it will be loaded after all of the common and book-specific code has been read. If your customizations apply only to the book class, use this file hook.

`tufte-common-handout.tex` If this file exists, it will be loaded after all of the common and handout-specific code has been read. If your customizations apply only to the handout class, use this file hook.

5.2 Numbered Section Headings

While Tufte dispenses with numbered headings in his books, if you require them, they can be anabled by changing the value of the `secnumdepth` counter. From the table below, select the heading level at which numbering should stop and set the `secnumdepth` counter to that value. For example, if you want parts and chapters numbered, but don't want numbering for sections or subsections, use the command:

```
\setcounter{secnumdepth}{0}
```

The default `secnumdepth` for the Tufte-LaTeX document classes is -1.

Heading level	Value
Part (in `tufte-book`)	-1
Part (in `tufte-handout`)	0
Chapter (only in `tufte-book`)	0
Section	1
Subsection	2
Subsubsection	3
Paragraph	4
Subparagraph	5

Table 5.1: Heading levels used with the `secnumdepth` counter.

5.3 Changing the Paper Size

The Tufte-LaTeX classes currently only provide three paper sizes: A4, B5, and US letter. To specify a different paper size (and/or margins), use the `\geometrysetup` command in the preamble of your document (or one of the file hooks). The full documentation of the `\geometrysetup` command may be found in the `geometry` package documentation.[1]

5.4 Customizing Marginal Material

Marginal material includes sidenotes, citations, margin notes, and captions. Normally, the justification of the marginal material follows the justification of the body text. If you specify the `justified` document class option, all of the margin material will be fully justified as well. If you don't specify the `justified` option, then the marginal material will be set ragged right.

You can set the justification of the marginal material separately from the body text using the following document class options: `sidenote`, `marginnote`, `caption`, `citation`, and `marginals`. Each option refers to its obviously corresponding marginal material type. The `marginals` option simultaneously sets the justification on all four marginal material types.

Each of the document class options takes one of five justification types:

`justified` Fully justifies the text (sets it flush left and right).

`raggedleft` Sets the text ragged left, regardless of which page it falls on.

`raggedright` Sets the text ragged right, regardless of which page it falls on.

`raggedouter` Sets the text ragged left if it falls on the left-hand (verso) page of the spread and otherwise sets it ragged right. This is useful in conjunction with the `symmetric` document class option.

`auto` If the `justified` document class option was specified, then set the text fully justified; otherwise the text is set ragged right. This is the default justification option if one is not explicitly specified.

For example,

```
\documentclass[symmetric,justified,marginals=raggedouter]{tufte-book}
```

will set the body text of the document to be fully justified and all of the margin material (sidenotes, margin notes, captions, and citations) to be flush against the body text with ragged outer edges.

THE FONT AND STYLE of the marginal material may also be modified using the following commands:

```
\setsidenotefont{⟨font commands⟩}
\setcaptionfont{⟨font commands⟩}
\setmarginnotefont{⟨font commands⟩}
\setcitationfont{⟨font commands⟩}
```

The `\setsidenotefont` sets the font and style for sidenotes, the `\setcaptionfont` for captions, the `\setmarginnotefont` for margin notes, and the `\setcitationfont` for citations. The ⟨*font commands*⟩ can contain font size changes (e.g., `\footnotesize`, `\Huge`, etc.), font style changes (e.g., `\sffamily`, `\ttfamily`, `\itshape`, etc.), color changes (e.g., `\color{blue}`), and many other adjustments.

If, for example, you wanted the captions to be set in italic sans serif, you could use:

[1] **pkg-geometry**

```
\setcaptionfont{\itshape\sffamily}
```

6
Compatibility Issues

When switching an existing document from one document class to a Tufte-LaTeX document class, a few changes to the document may have to be made.

6.1 Converting from `article` to `tufte-handout`

The following `article` class options are unsupported: `10pt`, `11pt`, `12pt`, `a5paper`, `b5paper`, `executivepaper`, `legalpaper`, `landscape`, `onecolumn`, and `twocolumn`. The following headings are not supported: `\subsubsection` and `\subparagraph`.

6.2 Converting from `book` to `tufte-book`

The following `report` class options are unsupported: `10pt`, `11pt`, `12pt`, `a5paper`, `b5paper`, `executivepaper`, `legalpaper`, `landscape`, `onecolumn`, and `twocolumn`. The following headings are not supported: `\subsubsection` and `\subparagraph`.

7
Troubleshooting and Support

7.1 Tufte-LaTeX Website

The website for the Tufte-LaTeX packages is located at http://code.google.com/p/tufte-latex/. On our website, you'll find links to our SVN repository, mailing lists, bug tracker, and documentation.

7.2 Tufte-LaTeX Mailing Lists

There are two mailing lists for the Tufte-LaTeX project:

Discussion list The `tufte-latex` discussion list is for asking questions, getting assistance with problems, and help with troubleshooting. Release announcements are also posted to this list. You can subscribe to the `tufte-latex` discussion list at http://groups.google.com/group/tufte-latex.

Commits list The `tufte-latex-commits` list is a read-only mailing list. A message is sent to the list any time the Tufte-LaTeX code has been updated. If you'd like to keep up with the latest code developments, you may subscribe to this list. You can subscribe to the `tufte-latex-commits` mailing list at http://groups.google.com/group/tufte-latex-commits.

7.3 Getting Help

If you've encountered a problem with one of the Tufte-LaTeX document classes, have a question, or would like to report a bug, please send an email to our mailing list or visit our website.
To help us troubleshoot the problem more quickly, please try to compile your document using the `debug` class option and send the generated `.log` file to the mailing list with a brief description of the problem.

7.4 Errors, Warnings, and Informational Messages

The following is a list of all of the errors, warnings, and other messages generated by the Tufte-LaTeX classes and a brief description of their meanings.

```
Error: \subparagraph is undefined by this class.
```

The `\subparagraph` command is not defined in the Tufte-LaTeX document classes. If you'd like to use the `\subparagraph` command, you'll need to redefine it yourself. See the "Headings" section on page 31 for a description of the heading styles availaboe in the Tufte-LaTeX document classes.

```
Error: \subsubsection is undefined by this class.
```

The `\subsubsection` command is not defined in the Tufte-LaTeX document classes. If you'd like to use the `\subsubsection` command, you'll need to redefine it yourself. See the “Headings” section on page 31 for a description of the heading styles availaboe in the Tufte-LaTeX document classes.

```
Error:  You may only call \morefloats twice.  See the
        Tufte-LaTeX documentation for other workarounds.
```

LaTeX allocates 18 slots for storing floats. The first time `\morefloats` is called, it allocates an additional 34 slots. The second time `\morefloats` is called, it allocates another 26 slots.
The `\morefloats` command may only be called two times. Calling it a third time will generate this error message. See page 33 for more information.

```
Warning:  Option ‘⟨class option⟩’ is not supported -- ignoring option.
```

This warning appears when you’ve tried to use ⟨*class option*⟩ with a Tufte-LaTeX document class, but ⟨*class option*⟩ isn’t supported by the Tufte-LaTeX document class. In this situation, ⟨*class option*⟩ is ignored.

```
Info:  The ‘symmetric’ option implies ‘twoside’
```

You specified the `symmetric` document class option. This option automatically forces the `twoside` option as well. See page 35 for more information on the `symmetric` class option.

7.5 Package Dependencies

The following is a list of packages that the Tufte-LaTeX document classes rely upon. Packages marked with an asterisk are optional.

- xifthen
- ifpdf*
- ifxetex*
- hyperref
- geometry
- ragged2e
- chngpage *or* changepage
- paralist
- textcase
- soul*
- letterspace*
- setspace
- natbib *and* bibentry
- optparams
- placeins
- mathpazo*
- helvet*
- fontenc
- beramono*
- fancyhdr
- xcolor
- textcomp
- titlesec
- titletoc