PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Review

**Cite this article:** Agrell E, Alvarado A, Kschischang FR. 2016 Implications of information theory in optical fibre communications. *Phil. Trans. R. Soc. A* **374**: 20140438.
http://dx.doi.org/10.1098/rsta.2014.0438

Accepted: 9 September 2015

One contribution of 14 to a discussion meeting issue 'Communication networks beyond the capacity crunch'.

**Subject Areas:**
electrical engineering, applied mathematics

**Keywords:**
channel capacity, information theory, mutual information, optical communications, Shannon theory

**Author for correspondence:**
Erik Agrell
e-mail: agrell@chalmers.se

# Implications of information theory in optical fibre communications

Erik Agrell[1,2], Alex Alvarado[2] and Frank R. Kschischang[3]

[1]Department of Signals and Systems, Chalmers University of Technology, 41296 Gothenburg, Sweden
[2]Optical Networks Group, Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, UK
[3]Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada M5S 3G4

EA, 0000-0003-0685-451X; AA, 0000-0002-2172-3051; FRK, 0000-0002-4274-1785

Recent decades have witnessed steady improvements in our ability to harness the information-carrying capability of optical fibres. Will this process continue, or will progress eventually stall? Information theory predicts that all channels have a limited capacity depending on the available transmission resources, and thus it is inevitable that the pace of improvements will slow. However, information theory also provides insights into how transmission resources should, in principle, best be exploited, and thus may serve as a guide for where to look for better ways to squeeze more out of a precious resource. This tutorial paper reviews the basic concepts of information theory and their application in fibre-optic communications.

## 1. Introduction

We live in a connected society where digital information is continuously exchanged across the globe. The vast majority of this information is carried by optical fibres for at least part of their journey. The development of long-haul fibre-optic communications is a fascinating story of invention—recounted in part in [1]—in which various technological advances (the development of single-mode fibre, efficient laser transmitters and modulators, optical amplifiers, wavelength-division multiplexing schemes, coherent signal detection and digital signal processing)

THE ROYAL SOCIETY PUBLISHING

have, over time, yielded steady improvements in the information-carrying capacity (bit per second) in commercial systems. Can this progress continue indefinitely, or are optical fibres ultimately limited in their capacity to carry information? If so, how close are we to achieving this ultimate limit?

These questions fall within the domain of a mathematical and engineering discipline called *information theory*, founded in the seminal paper of Shannon [2]. Conceived as a 'mathematical theory of communication', one branch of information theory answers questions about the trade-off between the rate at which the transmitter can send information and the reliability with which the receiver can decode the received signal. Remarkably, in his celebrated 'channel coding theorem', Shannon showed that the trade-off between rate and reliability is not smooth, but is discontinuous: at transmission rates below a fundamental quantity—the channel capacity—any (arbitrarily high) reliability (or, equivalently, any arbitrarily low probability of error) is, in principle, achievable using sophisticated coding and decoding schemes, whereas at transmission rates above the channel capacity, arbitrarily high reliability is impossible, no matter how sophisticated the transmitter and receiver are. Information theory thus seems to be precisely the right tool with which to establish the ultimate information-carrying capability of optical fibres.

So, what is the capacity of an optical fibre? Unfortunately, the answer is quite subtle and, to date, open. Subtleties emerge for a number of reasons. For example, in practical long-haul transmission, a high optical intensity has to be transmitted in order to overcome the accumulated loss over many kilometres of fibre. At such high intensities, the optical fibre becomes a nonlinear medium, and even the simplest mathematical channel model involves a complicated stochastic partial differential equation (the so-called generalized nonlinear Schrödinger equation and variants thereof) which challenges (and so far defies) information-theoretic analysis. Because it accepts a waveform at its input and produces a waveform at its output, the optical fibre channel is a so-called waveform channel. Unfortunately, except for certain special cases (in which a waveform channel can be converted into a completely equivalent 'discrete-time channel'), no tractable general information-theoretic analysis is known. Furthermore, commercial optical fibre systems are often operated as 'networks', with the interference-causing signals of different users multiplexed and demultiplexed at various geographical locations. The question of the capacity of optical fibre networks then becomes a question of 'multiuser information theory', for which precise capacities are generally unknown, even in relatively simple situations. Finally, the answer to the question of determining the capacity of a fibre channel depends not only on the physical medium (the fibre itself), but also crucially on which physical devices (such as amplifiers and multiplexers) are used along the transmission path, as well as how the transmitter and receiver are implemented. The past 15 years have witnessed an intense research in the area of optical fibre capacity. For extensive literature surveys, see [3–6] and [7, §VI].

The purpose of this paper is to provide a tutorial introduction to information theory as it is (or might be) applied in understanding fundamental limits on the information-carrying capabilities of optical fibres. In §2, we introduce some of the basic terminology, and give precise formulations of the questions that information theory intends to answer. In §3, we provide some of the intuition that underlies Shannon's channel coding theorem. In §4, we turn to the problem of determining the capacity of waveform channels, starting with the classical additive white Gaussian noise channel (AWGN), and then considering various formulations of the optical fibre channel. In §5, we speculate on ways in which information-theoretic insights may translate into new architectures that more fully exploit the information-carrying capability of optical fibres. Finally, in §6, we point out that the application of information theory to long-haul fibre-optic communications is far from straightforward, leaving open many interesting questions and challenges for future research.

## 2. Channel capacity: the maximum data rate

The word 'capacity' has both a colloquial and a mathematical meaning. When we talk about the 'capacity' of a storage drive in daily life, we mean the amount of data that it can store, and when

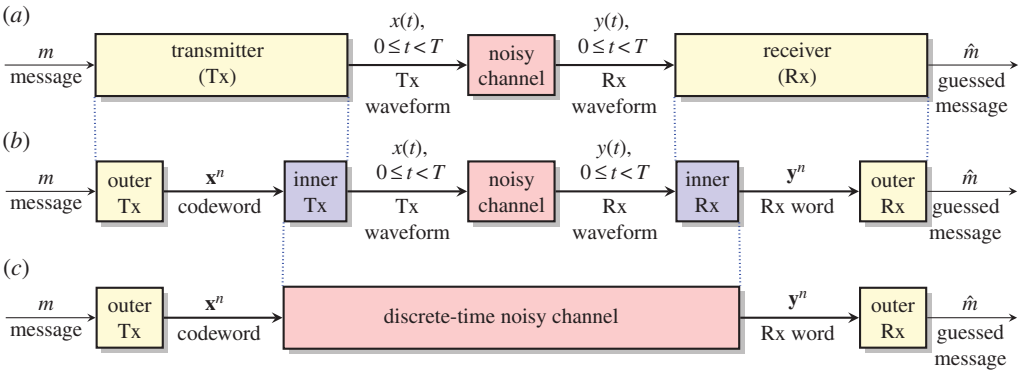


**Figure 1.** (*a*) A communication link, consisting of a transmitter, a waveform channel and a receiver. The purpose is to transmit a message $m$ in such a way that it can be reliably guessed (decoded) by the receiver. In (*b*), the transmitter is split into two units, an 'outer Tx' that maps the message $m$ into a discrete-time codeword $\mathbf{x}^n$, and an 'inner Tx' that maps the codeword to a waveform $x(t)$ of duration at most $T$. The receiver (Rx) is modelled using the reverse two blocks. In (*c*), the inner Tx, the channel and inner Rx are considered as a single unit with discrete-time input and output. (Online version in colour.)

we talk about the capacity of a communication system, we usually mean the amount of data that can be transported through it per time unit. When consumers or service providers scream for more capacity, it should be understood in this latter sense: there is a need for transporting even higher data volumes per time unit, or to more remote locations, or to a larger number of simultaneous users, or all of these.

Naively, increasing the data traffic volume should be ridiculously simple: can we not make some adjustment to a setting in the transmitting device, and cause bits to be pushed into the channel even faster than before? Unfortunately, without proper adjustment in the receiver, this increase is useless or even detrimental: more bits are transmitted than before, but if they are received in error, the received data are generally useless.

This trade-off between the rate of transmission and the reliability was discovered by Shannon, who defined mathematically the *capacity* of a transmission medium or a *channel*. The capacity is the maximum data rate that can be transmitted through the channel at an *arbitrarily small error probability*. This idea, which will be stated precisely in §2*a*, was revolutionary when it was presented in 1948 [2]. It means, in plain words, that reliable communication, *at positive rates of transmission*, is possible over unreliable channels. If the channel has a low quality, owing to noise, distortion, interference or any other physical impairments, a low rate has to be applied, but the data can, nevertheless, be transmitted virtually error-free. As an extreme example, the Voyager deep-space probes transmit photos and measurement data reliably to earth over billions of kilometres, using only a 22 W radio transmitter, thanks to a very low data rate (of the order of 160 bit/s).

## (a) Codebook, capacity and channel

To see how it is possible to transmit data reliably over an unreliable channel, consider the schematic in figure 1*a*. A message $m$, consisting of a block of $k$ bits, is encoded into a waveform $x(t)$ with duration $T$ or less. There are $M = 2^k$ possible messages, each of which corresponds to a unique waveform $x(t)$, which can be real- or complex-valued. The set of all $M$ waveforms is called the *codebook*. After transmission over the channel, a distorted version, $y(t)$, of $x(t)$ is observed by the receiver over the interval $0 \le t < T$. The data rate $R$ in bit/s that is achieved by this scheme is equal to the number of bits sent by the transmitter normalized by the observation time at the receiver, namely $R = k/T = (\log_2 M)/T$.

The receiver, upon observing $y(t)$, tries to guess which of the waveforms in the codebook might have been transmitted and outputs the message $\hat{m}$ corresponding to this waveform. If the received

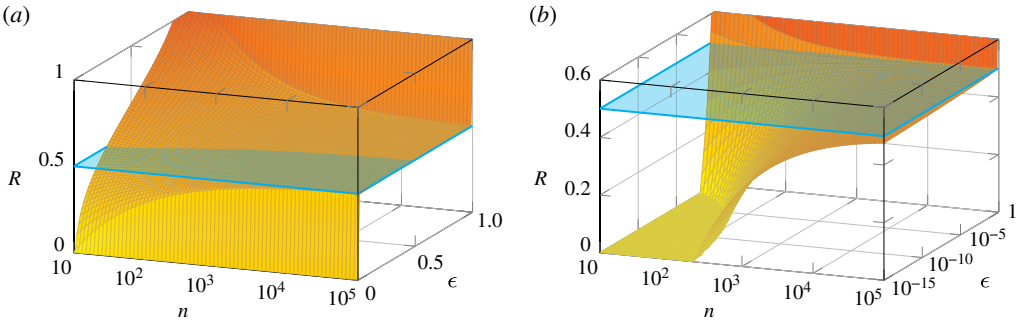


**Figure 2.** Estimate of maximum achievable rate $R$ for a discrete-time channel with $C_{\mathrm{dt}} = \frac{1}{2}$, as a function of error probability $\epsilon$ and block length $n$. (Detailed formulae are given in appendix.) (*a*) For each $0 < \epsilon < 1$, the maximum achievable rate $R$ approaches $C_{\mathrm{dt}}$ as $n \to \infty$. The horizontal plane shows the capacity. (*b*) The low-$\epsilon$ portion of (*a*) zoomed in. (Online version in colour.)

waveform is not overly distorted or corrupted by noise, and if the waveforms in the codebook are reasonably well separated in relation to the distortion, then the receiver's guess will be correct, producing $\hat{m} = m$.

The channel capacity in bit/s can now be mathematically defined as follows. We first decide on a given non-zero error probability threshold $\epsilon$ that can be accepted and a receiver observation duration $T$. Then, we search for the largest codebook such that an ideal receiver, upon observing $y(t)$ over the interval $0 \le t < T$, can correctly guess the transmitted message $m$ with probability no less than $1 - \epsilon$. Disregarding the facts that this search process may be prohibitively complex and that the ideal receiver may be unknown, the size of this optimal codebook is denoted by $M^*$ and the corresponding rate by $R = (\log_2 M^*)/T$. This quantity is the *maximum achievable rate* at duration $T$ and error probability $\epsilon$, and it exists theoretically for any $T > 0$ and $0 < \epsilon < 1$, even if we usually do not know how to compute it.

Next, we let $T$ increase, still for a given non-zero $\epsilon$, which means that $M^*$ also increases. Depending on the type of channel and the value of $\epsilon$, the ratio $R$ may increase or decrease with increasing $T$, but in any case it converges to a limit, $C = \lim_{T\to\infty} R$. This limit defines the capacity of the channel. One might expect that $C$ would decrease as $\epsilon$ decreases, but the limit is the same, regardless of $\epsilon$, as long as it is not 0 or 1 [2, theorem 12]. The qualitative behaviour of $R$ is well represented by figure 2 in §2b, if $n$ is replaced by $T$, for more or less any channel. A threshold phenomenon is evident at high $T$: every rate below capacity is achievable even at very low error probability (imagine a horizontal plane at a level $R < 0.5$ in figure 2), whereas a rate above capacity (a plane at $R > 0.5$) implies an error probability near 1.

To summarize, the channel capacity in bit/s of any waveform channel $x(t),\ 0 \le t < T \to y(t),\ 0 \le t < T$ is defined as

$$C = \lim_{T\to\infty} \frac{\log_2 M^*}{T}, \tag{2.1}$$

where $M^*$ is the maximum size of a codebook for which the probability of $\hat{m} \neq m$ is no larger than $\epsilon$, for any given $0 < \epsilon < 1$. By choosing $T$ large enough, transmission is possible at any data rate below $C$ at an arbitrarily small non-zero error probability, whereas this is impossible above $C$. However, although Shannon predicted the existence of codebooks for reliable transmission at data rates below capacity, he did not give a recipe for the construction of such codebooks. This has been an active area of research for 65 years [8].

The channel capacity $C$ is sometimes given as a function of a transmission parameter, if the transmitter is constrained to select waveforms $x(t)$ that satisfy certain conditions. In this case, the maximization that gives $M^*$ in (2.1) is over a constrained set of codebooks. A common example is the power-constrained capacity $C(P)$, where $P$ denotes the maximum transmit power of any waveform $x(t)$, averaged over $t$. In this case, all waveforms $x(t)$ in the codebook are required to

satisfy $\int_0^T |x(t)|^2 \, \mathrm{d}t \leq PT$. Constraints on the bandwidth, peak power and/or block length (delay) may also apply.

In contrast to the colloquial meaning of the word 'capacity' discussed in the beginning of this section, the channel capacity as defined in (2.1) is a mathematical concept, and hence it applies to mathematical channels. In order to make information-theoretic statements about a real-world, physical channel, a *channel model* is needed, which captures the essence of the physical channel's behaviour and gives it a mathematically precise formulation, taking into account the random nature of the signal propagation. A good channel model should be both physically realistic and tractable for mathematical analysis, which is sometimes difficult to achieve. For many types of copper-wired and wireless communication links, such models exist and are generally accepted, being relatively accurate for a wide range of transmission conditions, yet simple enough to allow information-theoretic analysis.

Unfortunately, as described in §4b, fibre-optical channel modelling is more complicated. The fibre-optical channel does not only include *random effects,* owing to the noise contributed by optical amplifiers, but also *nonlinear effects,* owing to the so-called Kerr nonlinearity in the fibre, which kicks in at high optical intensity. The relation between the optical channel's input $x(t)$ and output $y(t)$ is usually modelled by a stochastic nonlinear partial differential equation (the generalized nonlinear Schrödinger equation). For example, the Kerr nonlinearity implies that if the signal level of the input is doubled, unlike many models for copper-wired and wireless channels, the statistical distribution of the output changes in a more intricate way than by pure scaling.

Another feature of nonlinear channels is *bandwidth expansion* (or, in rare cases, bandwidth contraction). This means that if the transmitted signal is confined to a certain frequency band, the received signal may include frequency components outside this band. This phenomenon is important in practice, as it influences the spectrum allocation and receiver design. It does not influence the channel capacity $C$ in (2.1), which is measured in bit/s, but it has a strong influence on the maximum *spectral efficiency*, which is the channel capacity per unit bandwidth, measured in bit/s/Hz. If $W$ is the bandwidth under consideration, then the spectral efficiency $C/W$ will be different depending on whether $W$ is defined at the transmitter, receiver or elsewhere, because these bandwidths are in general different. There are also several mathematically inequivalent ways to define bandwidth, but this is outside the scope of this article.

The peculiarities of the nonlinear optical channel will be further discussed in §4b.

## (b) Discrete-time channel models

For practical reasons, the transmitter in figure 1 is often divided into two parts, as illustrated in figure 1*b*. The first part, here called the *outer transmitter,* converts the message $m$ to a sequence[1] of $n$ discrete-time symbols $\mathbf{x}^n$, where each symbol $x_i$ is a real or complex number, or a vector of numbers[2] and the second part, the *inner transmitter,* converts the sequence of symbols to a waveform $x(t)$. The corresponding parts can be defined on the receiver side, also illustrated in figure 1*b*. As we shall see in §3, this subdivision facilitates information-theoretic analysis, and it also represents the design of digital communication systems, where the functionalities of the inner and outer parts are often carried out by separate hardware or software components.

In most, if not all, commercially deployed communication systems, the inner transmitter is a linear filter (pulse shaper) and the inner receiver includes a linear (matched) filter followed by sampling. This set-up is provenly optimal for the band-limited AWGN channel, which is the most

[1]Notation convention: a sequence such as $x_1, x_2, \ldots, x_n$ is denoted as $\mathbf{x}^n$. Similarly, $y_1, \ldots, y_n$ is denoted as $\mathbf{y}^n$, $X_1, \ldots, X_n$ is denoted as $\mathbf{X}^n$, etc. Random variables are uppercase (e.g. $X$) and a single realization thereof lowercase ($x$).

[2]The outer transmitter is usually constructed by concatenating a binary error-correcting code and a bit-to-symbol mapper.

common linear channel model and will be formally defined in §4a, but is not the most general set-up and not necessarily optimal for nonlinear or non-Gaussian channels.

If the inner transmitter and receiver are kept fixed while the outer parts are optimized, it is convenient to regard the inner parts as part of the channel, as in figure 1*c*. This creates a so-called *discrete-time channel*, whose input and output are both a sequence of symbols, with input symbols selected from some set $\mathcal{X}$ and outputs produced in some set $\mathcal{Y}$. Usually, the inner transmitter accepts symbols at some fixed *symbol rate* $R_s$. Such a discrete-time channel is typically described by a probabilistic channel law $p_{\mathbf{Y}^n|\mathbf{X}^n}(\mathbf{y}^n \mid \mathbf{x}^n)$, giving the probability that the inner receiver produces a particular output sequence $\mathbf{y}^n$ when $\mathbf{x}^n$ is passed to the inner transmitter.

The capacity of a discrete-time channel $\mathbf{x}^n \to \mathbf{y}^n$ can be defined in analogy with (2.1). As before, we fix a non-zero error probability $\epsilon$ that can be accepted and a sequence length $n$. Based on the message $m$, the transmitter chooses one length-$n$ sequence from a predefined codebook for transmission. Let $M^*$ denote the largest size of the codebook such that the receiver can recover the transmitted message with error probability no larger than $\epsilon$. Then, the capacity of a discrete-time channel, in bit/symbol or bit/channel use, is defined as

$$C_{\mathrm{dt}} = \lim_{n\to\infty} \frac{\log_2 M^*}{n}. \tag{2.2}$$

In analogy with (2.1), the limit exists for any given $0 < \epsilon < 1$, and is the same for any $\epsilon$ in this range. This behaviour is exemplified in figure 2, where the maximum achievable rate approaches a threshold function at high $n$ (right end of the graphs). Most channels, in continuous as well as discrete time, exhibit similar behaviour, possibly with rescaled axes.

The discrete-time channel capacity $C_{\mathrm{dt}}$, multiplied by the symbol rate $R_s$, lower bounds the capacity $C$ of the underlying waveform (continuous-time) channel, because there are fewer degrees of freedom in the transmitter and receiver of figure 1*c* than in figure 1*a*.

## (c) Multiuser information theory

Although we have so far focused on systems with just a single transmitter and a single receiver, realistic data transmission systems must account for situations in which there are multiple transmitters and multiple receivers. For example, in fibre-optic transmission, signals of different users are often multiplexed at different wavelengths for transmission over the same physical fibre in a so-called *wavelength-division multiplexing* (WDM) system. Because of channel nonlinearities, these signals, even though they are propagating centred at different wavelengths, interact with each other, creating interference. The study of fundamental limits in such scenarios is the domain of *multiuser information theory*, a topic of enormous contemporary research [9].

While there are many variations in possible set-ups, an important example of a multiuser communication scenario relevant to fibre-optic communication is the so-called *K-user interference channel*, in which we have $K$ different transmitters and $K$ different receivers, with each receiver interested in correct decoding of the message sent by its corresponding transmitter. The channel model is usually given, so that the signal transmitted by all other users affects the signal received by the intended receiver. If the rate of reliable information transmission by the first transmitter-receiver pair is denoted as $R_1$, that of the second pair by $R_2$, and so on, then one can define a so-called *capacity region* as (the closure of) the set of simultaneously achievable $(R_1, R_2, \ldots, R_K)$ tuples. The capacity region then indicates all possible trade-offs between the transmission rates of the $K$ users using all manner of transmission strategies.

Unfortunately, although one has fairly general inner and outer bounds that apply in certain situations, capacity regions are known precisely only in certain special situations. For example, it is unknown for the seemingly straightforward case of the $K$-user interference channel, even when $K = 2$ and the simple AWGN channel (no memory or nonlinearity) is considered. The application of multiuser information theory to optical WDM systems will be discussed in §5.

# 3. The capacity of discrete-time channels

## (a) Memoryless channels

Here, we show how capacity is computed for the important special case of a discrete-time memoryless channel. Such a channel might be induced (at least in certain situations) by the combination of the inner transmitter, waveform channel and inner receiver depicted in figure 1*b*, resulting in a channel that accepts sequences of symbols from some alphabet $\mathcal{X}$ and produces symbols from some alphabet $\mathcal{Y}$ as shown in figure 1*c*. For the sake of simplicity, we will focus initially on the special case where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets, discussing more general alphabets later.

The term *memoryless* means that the output $Y_i$ of the channel at discrete time $i$, given the channel input at time $i$, is independent of channel inputs and outputs at all other times. The channel law for a memoryless (and time-invariant) channel is given by

$$p_{\mathbf{Y}^n|\mathbf{X}^n}(\mathbf{y}^n \mid \mathbf{x}^n) = \prod_{i=1}^{n} p_{Y|X}(y_i \mid x_i), \tag{3.1}$$

where $p_{Y|X}(y \mid x)$ is a fixed function (independent of $i$), giving the probability of receiving symbol $y \in \mathcal{Y}$ when symbol $x \in \mathcal{X}$ is sent.

As explained in §2b, the capacity of a discrete-time channel is given in (2.2) as a limit taken as the number of discrete channel uses, $n$, grows large. To compute the limit in the case of a memoryless channel, we will exploit the law of large numbers.

## (b) Typical sets

The (weak) law of large numbers (LLN) captures the intuitive notion that the empirical mean associated with a large number of independent statistical trials should be very near the expected value. More precisely, suppose that $X_1, \ldots, X_n$ is a sequence of independent identically distributed (i.i.d.) random variables with an expected value of $\mu$. The arithmetic mean

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \tag{3.2}$$

is then itself a random variable with an expected value of $\mu$. For a large $n$, one would be surprised to find that $\bar{X}_n$ takes a value very far from $\mu$, and indeed, the LLN states that, for every positive number $\epsilon$,

$$\Pr[|\bar{X}_n - \mu| > \epsilon] \to 0 \quad \text{as } n \to \infty. \tag{3.3}$$

Now let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables taking values in a finite alphabet $\mathcal{X}$. For any $x \in \mathcal{X}$, let $p_X(x)$ denote the probability that the random variable $X_i$ takes on value $x$. This probability does not depend on $i$ as a consequence of the assumption that $X_1, \ldots, X_n$ are identically distributed. The probability that any particular sequence $\mathbf{x}^n$ occurs, i.e. the probability that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, denoted as $p_{\mathbf{X}^n}(\mathbf{x}^n)$, is then

$$p_{\mathbf{X}^n}(\mathbf{x}^n) = \prod_{i=1}^{n} p_X(x_i). \tag{3.4}$$

Taking logarithms (to the base 2, say) and scaling by $-1/n$ in (3.4) allows us to apply the LLN, since then

$$-\frac{1}{n}\log_2 p_{\mathbf{X}^n}(\mathbf{x}^n) = -\frac{1}{n}\sum_{i=1}^{n} \log_2 p_X(x_i) \to \mathbb{E}[-\log_2 p_X(X)] \quad \text{as } n \to \infty, \tag{3.5}$$

where $\mathbb{E}$ denotes expectation. The right-hand side of (3.5) is usually denoted as $H(X)$ and called the *entropy* associated with the probability mass function (PMF) $p_X$. Explicitly,

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x).$$

The choice of the base of logarithm is arbitrary, but conventionally the base 2 logarithm is chosen, in which case the entropy is measured in units of bit/symbol.

Setting $\mu = H(X)$ in (3.3) yields, for any positive number $\epsilon$,

$$\Pr\left[\left|-\frac{1}{n}\sum_{i=1}^{n}\log_2 p_X(X_i) - H(X)\right| > \epsilon\right] \to 0 \quad \text{as } n \to \infty. \tag{3.6}$$

For fixed $\epsilon$, this motivates one to define a *typical set* of length-$n$ sequences as

$$\mathcal{A}_\epsilon^{(n)} = \left\{\mathbf{x}^n \mid \left|-\frac{1}{n}\sum_{i=1}^{n}\log_2 p_X(x_i) - H(X)\right| \le \epsilon\right\}. \tag{3.7}$$

By taking a sufficiently large $n$, the probability that a random i.i.d. sequence $\mathbf{X}^n$ yields an element in the typical set approaches arbitrarily close to unity. Intuitively this means that, for large $n$, outcomes *not* in the typical set occur only with vanishing probability.

If $\mathbf{x}^n \in \mathcal{A}_\epsilon^{(n)}$, from (3.1) and the definition (3.7), each typical sequence occurs with a probability lying in a bounded range

$$2^{-n(H(X)+\epsilon)} \le p_{\mathbf{X}^n}(\mathbf{x}^n) \le 2^{-n(H(X)-\epsilon)}. \tag{3.8}$$

Because the total probability of any set of sequences cannot exceed unity, the typical set cannot be too large. Denoting the cardinality of the typical set as $|\mathcal{A}_\epsilon^{(n)}|$, we have

$$|\mathcal{A}_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}. \tag{3.9}$$

On the other hand, for a sufficiently large $n$, we know that the total probability of $\mathcal{A}_\epsilon^{(n)}$ can be made to exceed $1-\epsilon$, which implies that

$$|\mathcal{A}_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\epsilon)} \quad \text{when } n \text{ is sufficiently large.} \tag{3.10}$$

Thus, we see that the typical set cannot be too small either. Very roughly, when $\epsilon$ is small and $n$ is large, (3.8)–(3.10) tell us that the typical set contains about $2^{nH(X)}$ sequences each of probability 'near' $2^{-nH(X)}$, accounting for very nearly all of the probability mass. The entropy $H(X)$ can, therefore, be interpreted as the growth-rate factor of the size of the typical set of sequences of length $n$ associated with a PMF $p_X$.

A typical set can similarly be defined for i.i.d. sequences of continuous random variables with common probability density function (PDF) $p_X(x)$, by taking

$$\mathcal{A}_\epsilon^{(n)} = \left\{\mathbf{x}^n \mid \left|-\frac{1}{n}\sum_{i=1}^{n}\log_2 p_X(x_i) - h(X)\right| \le \epsilon\right\},$$

where

$$h(X) = -\int_{-\infty}^{\infty} p_X(x)\log_2 p_X(x)\,\mathrm{d}x \tag{3.11}$$

denotes the so-called *differential entropy* associated with the PDF $p_X$. The LLN implies that the probability of the typical set approaches unity as $n \to \infty$. In analogy to (3.9)–(3.10), the *volume*, $\mathrm{Vol}(\mathcal{A}_\epsilon^{(n)})$, of the typical set can be shown to be bounded as

$$(1-\epsilon)2^{n(h(X)-\epsilon)} \le \mathrm{Vol}(\mathcal{A}_\epsilon^{(n)}) \le 2^{n(h(X)+\epsilon)},$$

where the lower bound applies when $n$ is sufficiently large. Thus, the differential entropy $h(X)$ can be interpreted as the growth-rate factor of the volume of the typical set of sequences of length $n$ associated with a PDF $p_X$.

## (c) Mutual information and channel capacity

Without attempting to be absolutely rigorous, using only these rough properties of typical sets, we will now consider information transmission over a discrete memoryless channel. We assume the channel has input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and channel law $p_{Y|X}(y \mid x)$, giving the

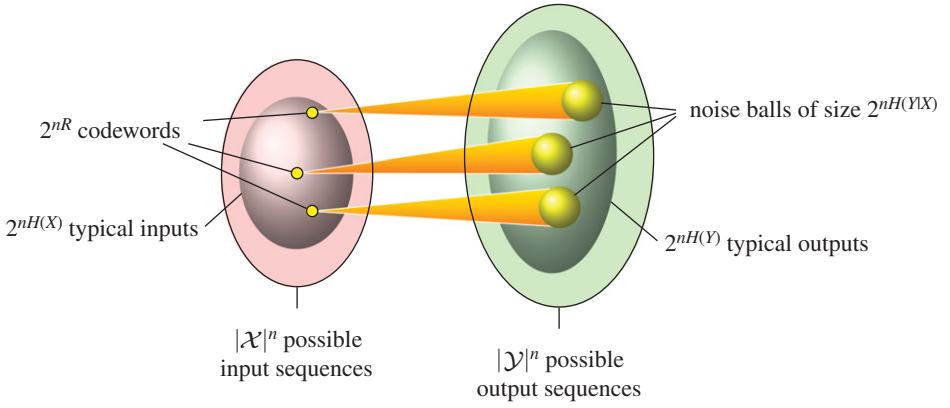


**Figure 3.** The relationships between typical sets, noise balls and transmitted and received codewords. (Online version in colour.)

probability of observing symbol $y \in \mathcal{Y}$ at the channel output when symbol $x \in \mathcal{X}$ is transmitted. We will assign a PMF $p_X$ to the transmitted symbols, and assume that the transmitter is constrained to the transmission of typical sequences only. When a sequence $\mathbf{x}^n$ is transmitted, and assuming $n$ is sufficiently large, the set of output sequences observed by the receiver are confined to a typical set of size about $2^{nH(Y|X)}$, where $H(Y \mid X)$ denotes the so-called *conditional entropy* of $Y$ given $X$, given as

$$H(Y \mid X) = -\sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X}(y \mid x) \log_2(p_{Y|X}(y \mid x)). \tag{3.12}$$

We think of this set as a 'noise ball' or 'uncertainty ball,' in the set of possible outputs, associated with the transmitted sequence $\mathbf{x}^n$. The set of probable received sequences (corresponding to all possible transmitted sequences) is itself a typical set containing about $2^{nH(Y)}$ elements, where $H(Y)$ is the entropy associated with the PMF $p_Y(y) = \sum_{x \in \mathcal{X}} p_X(x) p_{Y|X}(y \mid x)$. Our goal is to design a codebook of transmitted sequences with the property that there is little probability of overlap between the noise balls corresponding to different codewords. In this case, each received sequence is highly likely to fall within the noise ball associated with just one codeword (the transmitted one), and thus a low probability of error can be achieved if the decoder simply produces the associated codeword. The relationships among these various sets are illustrated in figure 3.

To maximize the transmission rate, we would like to design a codebook with as many codewords of possible, yet with the property that the noise balls corresponding to different codewords are essentially non-overlapping. Ideally, we might hope that the noise balls completely *partition* the set of typical output sequences. Assuming that the transmitted and received sequences are of length $n$, because the set of typical channel output sequences contains about $2^{nH(Y)}$ elements, and each noise ball contains about $2^{nH(Y|X)}$ elements, we would certainly not hope for more than about

$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y)-H(Y|X))}$$

codewords. The quantity $H(Y) - H(Y \mid X)$ is a fundamental quantity in information theory called the *mutual information*, $I(X;Y)$, between discrete random variables $X$ and $Y$, and it is given as

$$I(X;Y) = H(Y) - H(Y \mid X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \log_2 \left( \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \right). \tag{3.13}$$

If one replaces 'cardinality' with 'volume', an identical argument applies for memoryless channels whose input and output are continuous random variables. In this case, the ratio of the volume of the typical output sequences to the volume of a noise ball gives a bound on the size of a decodable codebook expressed in terms of the mutual information $I(X;Y)$ between continuous

random variables $X$ and $Y$, defined as

$$I(X;Y) = h(Y) - h(Y \mid X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{X,Y}(x,y)\log_2\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right)\mathrm{d}y\,\mathrm{d}x. \tag{3.14}$$

Thus, regardless of whether the random variables are discrete or continuous, we would not expect a codebook with codewords of length $n$ to have more than $2^{nI(X;Y)}$ codewords while achieving reliable decoding; or, put another way, we would expect that $R \leq I(X;Y)$. Indeed, it is possible to show that if a codebook of more than $2^{nR}$ codewords of length $n$ is used, then the probability of error of any decoder cannot be made arbitrarily small, i.e. the probability of error is in that case strictly bounded away from zero.

On the other hand, suppose that we select a transmission rate $R = I(X;Y) - \epsilon$, where $\epsilon$ is a small positive number. For any $n$, the total cardinality (or volume) of the union of all the noise balls associated with a codebook of $2^{nR}$ length-$n$ codewords is at most $2^{nR} \cdot 2^{nH(Y|X)}$, accounting for a fraction of at most

$$f = \frac{2^{nR} \cdot 2^{nH(Y|X)}}{2^{nH(Y)}} = 2^{-n\epsilon}$$

of the cardinality (or volume) of the set of typical output sequences. As $n \to \infty$, this fraction approaches zero, suggesting that a placement of essentially non-overlapping noise balls becomes increasingly more feasible with increasing block length. Indeed, as Shannon proved in [2] in his celebrated 'random coding argument,' a completely random placement is highly likely, when $n$ is sufficiently large, to produce a configuration with low average error probability. By expurgating a constant fraction of the codewords with the worst individual probability of error (which has a negligible impact on the rate), the worst-case probability of error can also be made to approach zero.

In summary, we see that at transmission rates $R$ greater than the mutual information $I(X;Y)$, the probability of error cannot be made to approach zero, while at transmission rates $R$ less than $I(X;Y)$, any arbitrarily small error probability can, in principle, be achieved by choosing the block length $n$ to be sufficiently large. Thus for discrete memoryless channels, when the channel input symbols are chosen according to a probability mass (or density) function $p_X(x)$, the maximum achievable rate of reliable information transmission is the mutual information $I(X;Y)$.

Because the mutual information depends on $p_X(x)$, the channel capacity for discrete memoryless channels is equal to the maximum mutual information that can be achieved over all possible input distributions, i.e.

$$C = \max_{p_X} I(X;Y). \tag{3.15}$$

It is important to note that (3.15) is not the *definition* of capacity; it is an expression that gives the capacity for memoryless channels. In particular, this so-called single-letter expression does not hold for general channels with memory. The multiletter generalization

$$C = \lim_{n\to\infty} \sup_{p_{\mathbf{X}^n}} \frac{1}{n} I(\mathbf{X}^n;\mathbf{Y}^n)$$

holds for so-called *information stable* channels with memory. While this latter formula gives the capacity of many channels of practical interest, it is possible to create mathematical models for channels for which even this formula fails to hold; see [10] for a discussion and development of a capacity formula that applies to even more exotic channel models with memory.

All of these capacity formulae apply to waveform channels whenever such channels are completely equivalent to some discrete-time channel model (for example, via projection on a countable orthogonal basis). For nonlinear waveform channels such as the optical-fibre channel, it appears difficult to achieve such an equivalence, and, therefore, one typically must resort to making approximations and assumptions (e.g. the assumption that waveforms are essentially bandwidth-limited).

# 4. The capacity of waveform channels

## (a) The additive white Gaussian noise channel

The most well-studied channel in information theory is the AWGN channel. This is for two reasons. First, it accurately describes the propagation on wired and wireless links under certain conditions; and second, it is one of the few waveform channels whose capacity is known exactly.

In the AWGN channel, the relationship between the input $x(t)$ and output $y(t)$, which are both complex processes, is

$$y(t) = x(t) + w(t), \tag{4.1}$$

where $w(t)$ is a complex-valued, zero-mean, white Gaussian random process with power spectral density $N_0$ W/Hz, which is assumed constant.[3] White Gaussian statistics for $w(t)$ in (4.1) is a good model for thermal noise (Johnson–Nyquist noise) introduced by electronic components in the receiver in any finite band of practical interest. Moreover, in optical communications, white Gaussian noise models well the amplified spontaneous emission noise introduced by optical amplifiers.

Assume that the complex waveform $x(t)$ is limited to a bandwidth $W$, i.e. its spectrum is zero outside $[f, f+W]$ for some $f$, and also has a limited power $P$. For any $P$ and $W$, the power- and bandwidth-constrained channel capacity in bit/s of the complex AWGN channel in (4.1) is [2, theorem 17]

$$C(P,W) = W \log_2 \left(1 + \frac{P}{N_0 W}\right), \tag{4.2}$$

where $N_0 W$ is the power of the complex noise $z(t)$ in the signal bandwidth and $P/(N_0 W)$ is the *signal-to-noise ratio* (SNR). This capacity is achieved by choosing $x(t)$ to be a Gaussian random process with power spectral density $P/W$ over the band $[f, f+W]$, and zero otherwise. The expression in (4.2), though sometimes called 'the Shannon capacity', applies specifically to the channel capacity of the AWGN channel. Other channels may have a larger or smaller capacity, depending on their particular characteristics.

The capacity $C(P,W)$ in (4.2) represents the maximum number of bits per second that can be reliably be transmitted through the channel (4.1), when $x(t)$ is power- and bandwidth-constrained. To increase the capacity, one can increase the bandwidth $W$, the power $P$, or both. If the bandwidth is fixed and the transmitted power is increased, then the capacity $C(P,W)$ tends to infinity, but it grows only logarithmically with power. On the other hand, if the power is fixed and the bandwidth increases, the capacity will never exceed $C(P) = \lim_{W\to\infty} C(P,W) = P \log_2 e/N_0$ bit/s. These two cases highlight the fact that when bandwidth is available, it is a good idea to spread the power over the whole bandwidth rather than only using a small part of it.

We will now use the band-limited AWGN channel in (4.1) as the noisy channel in figure 1*b* to exemplify two fundamental principles regarding achievable rates of continuous- and discrete-time channel models. First, there exist multiple discrete-time channels that correspond to the same waveform channel, depending on the choices for the inner transmitter and receiver. These discrete-time channels can have different capacities, of which the highest is equal to the capacity of the underlying waveform channel. Second, there exist multiple transmission schemes for the same discrete-time channel (figure 1*c*), depending on the choices for the outer transmitter and receiver. These schemes can have different mutual information, of which the highest is equal to the channel capacity of the discrete-time channel.

To elaborate on the first point, we design a waveform $x(t)$ from a sequence of complex numbers $\mathbf{x}^n$ as

$$x(t) = \sum_{i=1}^{n} x_i \operatorname{sinc}\left(\frac{t}{T_s} - i\right), \tag{4.3}$$

[3]'White' means in this context that $w(t)$ had autocorrelation function $\mathbb{E}[w(t)w^*(t')] = N_0\delta(t-t')$, where $\delta(t)$ is the Dirac delta function and $(\cdot)^*$ denotes complex conjugate.

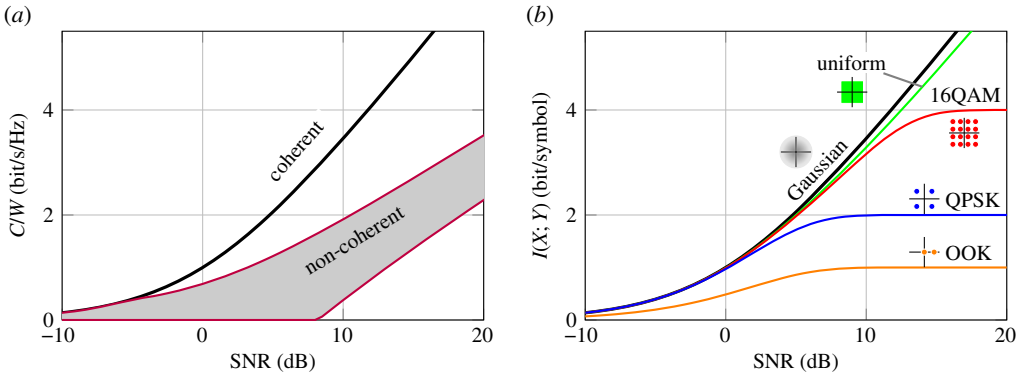


**Figure 4.** (*a*) The channel capacity of two discrete-time channels based on the same waveform channel. The highest of these, which gives the capacity of (4.5), is equal to the capacity of the waveform channel (4.1) with coherent detection. If a suboptimal non-coherent detector is applied, the capacity (in the shaded region) is less. (*b*) The mutual information $I(X;Y)$ of the discrete-time channel (4.5) for selected continuous and discrete input distributions. The highest of these represents the capacity of the channel. (Online version in colour.)

where $\mathrm{sinc}(v) = \sin(\pi v)/(\pi v)$ and $T_\mathrm{s} = 1/R_\mathrm{s} = 1/W$ is the time between transmission of two symbols $x_i$. This is a *linear modulator*, which we use as the inner transmitter in figure 1*b*. We consider two different inner receivers, the *coherent* and *non-coherent* receiver, which are both based on the *matched filter* output

$$r(t) = \frac{1}{T_\mathrm{s}} \int_{-\infty}^{\infty} y(\tau) \mathrm{sinc}\left(\frac{\tau - t}{T_\mathrm{s}}\right) \mathrm{d}\tau. \tag{4.4}$$

In the coherent receiver, the output $y_i$ is the complex sampled filter output $y_i = r(iT_\mathrm{s})$ for $i = 1, \ldots, n$, whereas in the non-coherent receiver, the phase information is lost, and the output is only the magnitude $|r(iT_\mathrm{s})|$. It can be shown that the coherent receiver in combination with the inner transmitter (4.3) and the AWGN waveform channel (4.1) generates the complex *discrete-time AWGN channel*

$$y_i = x_i + w_i, \tag{4.5}$$

where $w_i$ are complex independent Gaussian random variables with zero mean and variance $\sigma_w^2 = \mathbb{E}[|w_i|^2] = N_0/T_\mathrm{s} = N_0 W$. The variance of the transmitted symbol $x_i$ is $\sigma_x^2 = P$ for all $i$.

The capacities of these two discrete-time channels are shown in figure 4*a*, where the shaded region indicates that the capacity of the non-coherent channel is known by means of upper and lower bounds (see appendix) but not exactly. Despite the fact that they both communicate over the same waveform channel, their capacities are quite different. Interestingly, the capacity of the complex discrete-time AWGN channel is $\log_2(1 + \sigma_x^2/\sigma_w^2)$ bit/symbol, which corresponds to $(1/T_\mathrm{s}) \log_2(1 + \sigma_x^2/\sigma_w^2)$ bit/s. This is exactly the same expression as (4.2), which shows that this combination of transmitter and receiver is, indeed, optimal for the continuous-time AWGN channel. This optimality can be understood by means of Nyquist's sampling theorem, which states that any complex waveform band-limited to $[f, f + W]$ can be completely described by its samples taken at rate $W$. Hence, if the waveform is time-limited to a large time interval $T$, then $n = WT$ complex numbers are enough to completely describe the waveform. In other words, there is a one-to-one correspondence between $x(t)$, $0 \le t < T = nT_\mathrm{s}$ in figure 1*b* and $x_1, \ldots, x_n$ in figure 1*c*, and analogously for $y(t)$, which means that the continuous- and discrete-time AWGN channels are equivalent and have the same capacity. For general channels, however, there exists no such equivalence between continuous- and discrete-time models.

We now turn to the second principle, namely that different transmission schemes for the same channel have different maximum achievable rates. Using the discrete-time AWGN channel (4.5) as a case study, the mutual information (3.13) and (3.14) is evaluated in figure 4*b* for a variety of

continuous and discrete input distributions. For the finite input alphabets $\mathcal{X}$, which all correspond to well-known digital modulation formats, the mutual information converges to $\log_2 |\mathcal{X}|$ as the SNR increases, whereas the it grows unboundedly in the case of continuous input distributions. The highest mutual information is at all SNRs obtained for the circular Gaussian distribution, and this mutual information is, indeed, equal to the highest capacity in figure 4*a*. This shows that the capacity-achieving input distribution for this discrete-time channel is circular Gaussian.

## (b) The optical fibre channel

Because information theory is a mathematical science, it needs a mathematical description of the channel. The most common model of lightwave propagation in an optical fibre is given by the *generalized nonlinear Schrödinger equation*. It describes how the optical field $U(t,z)$ changes with time $t$ and with the distance $z$ from the fibre's end, according to the partial differential equation

$$\frac{\partial U(t,z)}{\partial z} = -\alpha U(t,z) - j\frac{\beta_2}{2}\frac{\partial^2 U(t,z)}{\partial t^2} + j\gamma |U(t,z)|^2 U(t,z) + W(t,z), \tag{4.6}$$

where $\alpha$, $\beta_2$ and $\gamma$ are fibre parameters that characterize the loss, dispersion and nonlinearity, respectively, and $j = \sqrt{-1}$. The term $W(t,z)$ represents random noise, uncorrelated in $t$ and $z$, which is added in optical amplifiers and detectors. If the input to a fibre of length $L$ is denoted by $X(t) = U(t,0)$ and its output by $Y(t) = U(t,L)$, then (4.6) can represent the noisy channel in figure 1*a*,*b*. The model has been demonstrated to be very accurate and it can be adapted to a wide range of scenarios, including different amplification schemes and dual-polarization transmission. Unfortunately, no explicit relation between $X(t)$ and $Y(t)$ is known. The equation cannot be solved analytically except in a few special cases.

The capacity of the optical fibre channel is not known. This is a property that it shares with most other real-world channels. The standard approach in such cases is to sandwich the capacity between lower and upper bounds. If such bounds can be derived, and if they are reasonably close to each other, then reliable conclusions can be drawn about the capacity, and the results often give insights into how to design efficient transmission schemes for the channel in question.

Lower and upper bounds have different nature and are derived by different mathematical techniques. A lower bound describes what is possible, whereas an upper bound describes what is impossible. Any transmission scheme (i.e. any choice of transmitter and receiver in figure 1*a*) gives a lower bound, if the error probability is sufficiently low. Other lower bounds can be obtained by studying other transmission schemes. For example, it is sufficient to study any single pair of inner transmitter and receiver in figure 1*b*. Therefore, there exists a rich literature about lower bounds on the capacity of optical channels [4,7,11,12]. Without any ambitions to plot any of these bounds exactly, which would confine the study to a specific system set-up, the general behaviour of these lower bounds is illustrated by the coloured curves in figure 5. Even though most of these lower bounds have a peak at some power, their envelope does not: at low transmit power, when nonlinear distortion is negligible, it increases following the capacity of the AWGN channel, and at higher power it flattens out. Capacity can never decrease with power, which is obvious from the definition of $P$ as a *maximum* power in §2a.

In contrast, upper bounds cannot be based on isolated transmission schemes—they predict that rates above a certain value can never be achieved for the considered channel, with neither present methods nor any methods that may devised in the future. Thus, an upper bound derived for a specific discrete-time channel is not necessarily valid for the underlying waveform channel. Figure 5 includes the only upper bound on the capacity of the fibre-optic channel that we are aware of, which was proved only recently [13]. In plain words, it states that the capacity is no larger than the capacity of an AWGN channel (4.2) with the same amount of noise, the same transmit power and the same bandwidth.[4] The presence of loss, dispersion and nonlinearity in equation (4.6) cannot increase capacity.

[4]Care must be exercised when defining the bandwidth for a nonlinear system, as it may vary with the propagation distance, see §2a.

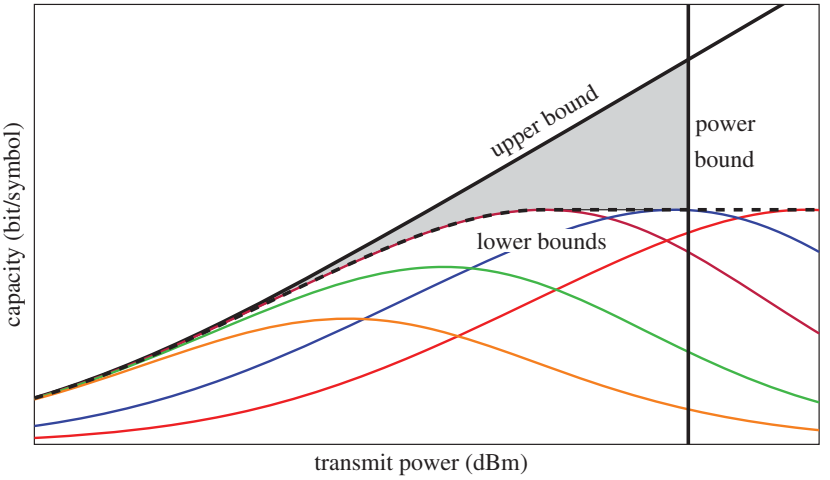


**Figure 5.** General behaviour of upper and lower bounds on the capacity of fibre-optic channels. An upper bound on the average transmit power is also shown, which confines the capacity to a triangular region (shaded). It is not known where in this region the true capacity lies. (Online version in colour.)

There is a sizeable gap between the lower and upper bounds in figure 5. The gap increases with power, which may give the impression that capacity might, indeed, be infinite in some cases. Unfortunately, there is a third bound, which comes not from information theory but from physics: If the transmit power is increased too high, the fibre will heat up and eventually melt. Because the core of a modern optical fibre is very thin, of the order of 1 µm, this so-called *fibre fuse* [14] does not require a huge amount of power. Hence, the capacity is not only bounded from above and below, but also from the right. It is still an open question where in the grey region the true capacity lies.

The picture gets more complicated when one considers a whole network of connections rather than a single point-to-point link. An optical network consists of a mesh of nodes connected with fibres. Each fibre carries many parallel signals on different wavelengths, which interfere with each other during propagation. The connections are routed from source to destination via several intermediate nodes, where the signals are separated in *wavelength-selective switches*. This set-up creates a topology that connects geographically separated nodes with each other in the same configurable network.

The maximum achievable rates in optical networks have been analysed only under certain simplifying assumptions. The most common approach is to consider the capacity of a single point-to-point link in the network, assuming certain *behavioural models* for the transmission on the interfering channels [12,15,16]. Another approach, which is better aligned with conventional information theory, is to study the set of rates that can be simultaneously achieved over a set of interfering connections in the network [17–19].

## 5. Architectural implications of information theory

Although we have so far argued that information theory is an indispensable tool for the *analysis* of communication systems, we now argue that it is also eminently suitable for system *design*. By understanding the fundamental limits associated with any given transmission strategy, system builders can allocate development effort appropriately, converging to a design with the most suitable trade-off between data transmission performance and implementation complexity in any given application. Importantly, by understanding fundamental limits, system designers 'know when to stop'; they understand the futility of attempting to improve systems beyond the best achievable performance.

Of course, as indicated in §4b, our information-theoretic understanding of the fibre-optic channel is far from complete, and much further investigation needs to be done. As our understanding of the fundamental limits (and the transmission schemes that approach them) matures, we expect—or, more accurately, we speculate—that new approaches for information transmission over the fibre-optic channel will emerge.

At present, with the notable exception of soliton systems (which exploit the specific peculiarities of the nonlinear Schrödinger equation [20]), the design of transmission strategies for the fibre-optic channel has largely been guided by our understanding of transmission schemes for the AWGN channel. Widely used linear matched-filter receivers—which are known to provide a sufficient statistic for detection at the output of the AWGN channel—are not necessarily the optimal processor at the output of a nonlinear channel. In AWGN channels, capturing and processing energy outside the frequency band of the transmitted signal is useless; in fibre channels, owing to the nonlinearity, such processing may be helpful as out-of-band signals are correlated with the in-band signal of interest.

Information theory is expected to help system designers understand the impact on achievable transmission rates of replacing expensive optical components and devices with less efficient and less expensive alternatives. In the terminology of §2b, such devices are often part of the 'inner transmitter' and 'inner receiver'. It is often the case that an appropriate adjustment to the 'outer transmitter' and 'outer receiver'—an improved error-correcting code, say—can overcome or partially offset the additional system impairments created by use of a less-than-ideal optical component. We would expect that the potentially significant cost reductions that may be achievable using this approach will drive research and development in this direction.

WDM creates user subchannels centred at different carrier frequencies (wavelengths), analogously to the method by which the signals from different radio or television broadcasters are kept separated. Unfortunately, this linear multiplexing technique inevitably suffers from interference (crosstalk) between the channels owing to channel nonlinearity, which appears to limit the data transmission rates that can be achieved in such systems. A recently proposed nonlinear frequency division multiplexing (NFDM) approach [21–24] exploits the properties of the so-called nonlinear Fourier transform, to decompose the ideal nonlinear Schrödinger channel into parallel noninteracting subchannels, much as the ordinary (linear) Fourier transform decomposes a linear time-invariant channel into non-interacting subchannels. (A single-channel version of the idea of modulating information in this manner was termed 'eigenvalue communication' in [25].) At least in principle, it may be possible to multiplex the signals of different users in different bands in the nonlinear spectral domain. In contrast with WDM systems, in which crosstalk occurs even in ideal noise-free systems, because nonlinearly multiplexed subchannels are completely noninteracting in the absence of noise, we speculate that the amount of interaction between channels will be smaller than in WDM systems in the low-noise regime. At present, this idea is far from practical, because there are no known physical devices (apart from synthesis by digital algorithms) that can achieve such multiplexing with the same convenience as the linear superposition of multiple modulated laser sources operating at different wavelengths. Furthermore, deviations from ideality (in particular, loss, noise, imperfections in waveform synthesis) will have a deleterious effect that is at present only poorly understood. Certainly, further investigation of NFDM is warranted.

In applying the results of information theory, system designers must be cautious to ensure that all constraints and costs are reflected in the information-theoretic model. For example, the usual analysis of the capacity of the AWGN channel considers only the cost of the transmitter power $P$ and the bandwidth $W$, neglecting, for example, the power expended in the operation of the receiver. While such a model is certainly appropriate in situations (like long-distance wireless communication) where the transmitter power is the dominant component in the total system power budget, this may not be the regime of greatest interest in long-haul optical communications, where significant power is consumed in the operation of the receiver. To operate near the channel capacity requires long codes and complicated power-hungry decoding algorithms; thus, as suggested by recent information-theoretic analyses [26], when minimizing

total power consumption it may be beneficial to operate a system at some gap from the channel capacity.

Multiuser information theory, as described briefly in §2c, is also likely to play an important role in the design of future optical communication systems, as it studies techniques by which different users can coordinate and cooperate to achieve certain communications objectives. In $K$-user linear multiuser interference channels, the recently proposed concept of *interference alignment* [27] is an intriguing concept with a potential for application to optical-fibre systems. Here, each user partitions its set of available degrees-of-freedom into two bins: one bin to collect the signal of interest, and one bin to collect the interference. Transmissions are carefully coordinated, so that all $K-1$ signals not wanted by any particular receiver are aligned to fall within the second bin, whereas the desired signal falls into the first bin. In the context of the fibre-optic channel, one such alignment strategy—termed 'interference focusing'—has been proposed in [6,18,19].

As optical fibre system designs have increasingly come to afford ever more sophisticated digital signal-processing and error-correcting decoding algorithms in the receiver, design choices made previously to simplify processing are being revisited. A primary example is the development of single-mode fibre, which greatly simplifies equalization and signal processing in long-haul systems. On the other hand, multimode or few-mode fibre gives the possibility of achieving a substantial enhancement in information-carrying capacity, at the expense of devices that allow for coupling of different signals simultaneously into several modes at the transmitter, and detecting and processing these modes at the receiver. Unlike the similar conventional linear multiantenna transmission systems used in wireless communications, nonlinearity is expected to be a dominant consideration in such multimode systems. Establishing fundamental information-theoretic limits on such systems remains an open problem.

## 6. Conclusion

Information theory builds upon mathematical models of communication systems to establish fundamental limits on their information-carrying capability. Information theory guides system designers to find efficient strategies with which to exploit a given set of transmission resources. Even though the exact channel capacity, which is the maximum achievable data rate for a given channel, is not known exactly, bounds and estimates are available, which give important insights into system designs. By optimizing the available resources to maximize the capacity of the channel, or bounds thereon, designers can 'future proof' their systems: even if they choose to operate these systems far from fundamental limits, they know that sophisticated coding and signal processing techniques, in principle, exist that can move the operating point to a more efficient regime. While models of fibre-optic channels have so far defied exact information-theoretic analyses, substantial progress continues to be made, and the insights obtained are likely to inform system designs for many years to come.

Competing interests. We declare we have no competing interests.

Funding. The research was supported by the Swedish Research Council under grant nos. 2012-5280 and 2013-5271, the Engineering and Physical Sciences Research Council (EPSRC) through the project UNLOC (EP/J017582/1) and the Natural Sciences and Engineering Research Council of Canada.

## Appendix A. Mathematical toolbox

This appendix provides mathematical expressions for some of the capacity estimates and bounds that are illustrated in this paper.

Figure 2 is based on the *binary symmetric channel* with crossover probability $p$, which is a discrete-time memoryless channel with input and output alphabets $\mathcal{X}=\mathcal{Y}=\{0,1\}$ and the channel law $p_{Y|X}(y\,|\,x)=p$ if $y\neq x$. In [28], it is estimated that a code of length $n$ can have as

many as $M^*$ codewords and still achieve an error probability $\epsilon$ over this channel, with $p \in (0, \frac{1}{2})$, where

$$\log_2 M^* = nC(p) - \sqrt{nV(p)}Q^{-1}(\epsilon) + \tfrac{1}{2}\log_2 n + \mathcal{O}(1)$$

in which $Q(x) = \frac{1}{2}\mathrm{erfc}(x/\sqrt{2})$ is the standard Gaussian tail $Q$ function, $C(p) = 1 + p\log_2 p + (1-p)\log_2(1-p)$ is the channel capacity and $V(p) = p(1-p)(\log_2(1/p-1))^2$ is the so-called channel dispersion (not to be confused with the fibre dispersion in §4b). The curves of figure 2 were obtained for $p = 0.11$ by converting $M^*$ to a rate and setting the $\mathcal{O}(1)$ term to zero. The maximum achievable rate as a function of $n$ and $\epsilon$ for other channels behaves qualitatively the same, displaying the same threshold phenomenon.

The non-coherent AWGN capacity in figure 4*a* was characterized by the upper bound in [29, eqn (42)] and the lower bound in [30, eqn (13)]. These bounds are not very tight. Stronger bounds exist but are in general more complex to evaluate [29,30]. The continuous input distributions in figure 4*b* are a complex circular Gaussian distribution and a complex uniform distributions over a square. Their mutual information was calculated for the real and imaginary parts separately as $I(X;Y) = h(Y) - h(Y \mid X)$, see (3.14), where $h(Y)$ was computed by numerically integrating (3.11) and $h(Y \mid X) = \frac{1}{2}\log_2 \pi e \sigma_w^2$. The discrete distributions are 16-ary quadrature amplitude modulation (QAM), quaternary phase shift keying (QPSK), and on–off keying (OOK). In all three cases, the constellations are scaled to the desired power and probabilities are uniform on the constellation points. The mutual information of these constellations was calculated by means of Gauss–Hermite quadratures according to [31, §4.5].

## References


1. Hecht J. 2004 *City of light: the story of fiber optics*, 2nd edn. New York, NY: Oxford University Press.
2. Shannon CE. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423/623–656. (doi:10.1002/j.1538-7305.1948.tb01338.x)
3. Kahn JM, Ho KP. 2004 Spectral efficiency limits and modulation/detection techniques for DWDM systems. *IEEE J. Sel. Top. Quantum Electron.* **10**, 259–272. (doi:10.1109/JSTQE.2004.826575)
4. Agrell E, Alvarado A, Durisi G, Karlsson M. 2014 Capacity of a nonlinear optical channel with finite memory. *J. Lightw. Technol.* **32**, 2862–2876. (doi:10.1109/JLT.2014.2328518)
5. Agrell E. 2015 Conditions for a monotonic channel capacity. *IEEE Trans. Commun.* **63**, 738–748. (doi:10.1109/TCOMM.2014.2381247)
6. Ghozlan H, Kramer G. 2015 Models and information rates for multiuser optical fiber channels with nonlinearity and dispersion. (http://arxiv.org/abs/1503.03124)
7. Essiambre RJ, Kramer G, Winzer PJ, Foschini GJ, Goebel B. 2010 Capacity limits of optical fiber networks. *J. Lightw. Technol.* **28**, 662–701. (doi:10.1109/JLT.2009.2039464)
8. Costello Jr DJ, Forney Jr GD. 2007 Channel coding: the road to channel capacity. *Proc. IEEE* **95**, 1150–1177. (doi:10.1109/JPROC.2007.895188)
9. El Gamal A, Kim YH. 2011 *Network information theory*. Cambridge, UK: Cambridge University Press.
10. Verdú S, Han TS. 1994 A general formula for channel capacity. *IEEE Trans. Inf. Theory* **40**, 1147–1157. (doi:10.1109/18.335960)
11. Mitra PP, Stark JB. 2001 Nonlinear limits to the information capacity of optical fibre communications. *Nature* **411**, 1027–1030. (doi:10.1038/35082518)
12. Secondini M, Forestieri E, Prati G. 2013 Achievable information rate in nonlinear WDM fiber-optic systems with arbitrary modulation formats and dispersion maps. *J. Lightw. Technol.* **31**, 3839–3852. (doi:10.1109/JLT.2013.2288677)
13. Kramer G, Yousefi MI, Kschischang FR. 2015 Upper bound on the capacity of a cascade of nonlinear and noisy channels. In *Proc. IEEE Inf. Theory Workshop* (*ITW*), *Jerusalem, Israel, 26 April–1 May*.
14. Kashyap R. 2013 The fiber fuse—from a curious effect to a critical issue: A 25th year retrospective. *Opt. Express.* **21**, 6422–6441. (doi:10.1364/OE.21.006422)

15. Agrell E, Karlsson M. 2013 WDM channel capacity and its dependence on multichannel adaptation models. In *Proc. Opt. Fiber Commun. Conf.* (*OFC*), p. OTu3B.4. *Anaheim, CA, 17–21 March*.
16. Agrell E, Karlsson M. 2015 Influence of behavioral models on multiuser channel capacity. *J. Lightw. Technol.* **33**, 3507–3515. (doi:10.1109/JLT.2015.2438951)
17. Taghavi MH, Papen GC, Siegel PH. 2006 On the multiuser capacity of WDM in a nonlinear optical fiber: coherent communication. *IEEE Trans. Inf. Theory* **52**, 5008–5022. (doi:10.1109/TIT.2006.883540)
18. Ghozlan H, Kramer G. 2010 Interference focusing for mitigating cross-phase modulation in a simplified optical fiber model. In *Proc. IEEE Int. Symp. Inf. Theory* (*ISIT*), pp. 2033–2037. *Austin, TX, 13–18 June*.
19. Ghozlan H, Kramer G. 2011 Interference focusing for simplified optical fiber models with dispersion. In *Proc. IEEE Int. Symp. Inf. Theory* (*ISIT*), pp. 376–379. *Saint Petersburg, Russia, 31 July–5 Aug*.
20. Mollenauer LF, Gordon JP. 2006 *Solitons in optical fibers: fundamentals and applications*. Burlington, MA: Elsevier Academic Press.
21. Yousefi MI, Kschischang FR. 2014 Information transmission using the nonlinear Fourier transform: parts I–III. *IEEE Trans. Inf. Theory* **60**, 4312–4369. (doi:10.1109/TIT.2014.2321143)
22. Meron E. 2012 Aspects of communications in the optical fiber channel. PhD dissertation, Tel Aviv University, Faculty of Engineering, Israel.
23. Prilepsky JE, Derevyanko SA, Turitsyn SK. 2013 Nonlinear spectral management: linearization of the lossless fiber channel. *Opt. Exp.* **21**, 24 344–24 367. (doi:10.1364/OE.21.024344)
24. Prilepsky JE, Derevyanko SA, Blow KJ, Gabitov I, Turitsyn SK. 2014 Nonlinear inverse synthesis and eigenvalue division multiplexing in optical fiber channels. *Phys. Rev. Lett.* **113**, 013901. (doi:10.1103/PhysRevLett.113.013901)
25. Hasegawa A, Nyu T. 1993 Eigenvalue communication. *J. Lightw. Technol.* **11**, 395–399. (doi:10.1109/50.219570)
26. Grover P, Woyach KA, Sahai A. 2011 Towards a communication-theoretic understanding of system-level power consumption. *IEEE J. Sel. Areas Commun.* **29**, 1744–1755. (doi:10.1109/JSAC.2011.110922)
27. Cadambe VR, Jafar SA. 2008 Interference alignment and degrees of freedom of the *k*-user interference channel. *IEEE Trans. Inf. Theory* **54**, 3425–3441. (doi:10.1109/TIT.2008.926344)
28. Polyanskiy Y, Poor HV, Verdú S. 2010 Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory* **56**, 2307–2359. (doi:10.1109/TIT.2010.2043769)
29. Katz M, Shamai (Shitz) S. 2004 On the capacity-achieving distribution of the discrete-time noncoherent and partially coherent AWGN channels. *IEEE Trans. Inf. Theory* **50**, 2257–2270. (doi:10.1109/TIT.2004.834745)
30. Durisi G. 2012 On the capacity of the block-memoryless phase-noise channel. *IEEE Commun. Lett.* **16**, 1157–1160. (doi:10.1109/LCOMM.2012.060812.120312)
31. Szczecinski L, Alvarado A 2015 *Bit-interleaved coded modulation: fundamentals, analysis and design*. Chichester, UK: John Wiley & Sons.