Conditions for a Monotonic Channel Capacity

Ieee Transactions on Communications (ISSN: 0090-6778)

Citation for the published paper:
Agrell, E. (2015) "Conditions for a Monotonic Channel Capacity". Ieee Transactions on Communications, vol. 63(3), pp. 738-748.

http://dx.doi.org/10.1109/tcomm.2014.2381247

(article starts on next page)

Conditions for a Monotonic Channel Capacity

Erik Agrell, *Senior Member, IEEE*

***Abstract*—Motivated by results in optical communications, where the performance can degrade dramatically if the transmit power is sufficiently increased, the channel capacity is characterized for various kinds of memoryless vector channels. It is proved that for all static point-to-point channels, the channel capacity under an equal-power constraint is a nondecreasing function of power. As a consequence, maximizing the mutual information over all input distributions with a certain power is for such channels equivalent to maximizing it over the larger set of input distributions with upperbounded power. The channel coding theorem is formally proved for an equal-power constraint. For interference channels such as optical wavelength-division multiplexing systems, the primary channel capacity is always nondecreasing with power if all interferers transmit with identical distributions as the primary user. Also, if all input distributions in an interference channel are optimized jointly, then the achievable sum-rate capacity is again nondecreasing. The results generalize to the channel capacity as a function of a wide class of costs, not only power.**

***Index Terms*—Achievable rate, capacity–cost function, channel capacity, mutual information, nonlinear distortion, optical communications, Shannon limit.**

I. INTRODUCTION

IN THE MOST cited paper in the history of information theory [1], Shannon in 1948 proved that with adequate coding, reliable communication is possible over a noisy channel, as long as the rate does not exceed a certain threshold, called the *channel capacity.* He provided a mathematical expression for the channel capacity of any point-to-point channel, based on its statistical properties. The expression is given as the supremum over all possible input distributions of a quantity later called the *mutual information* [2], [3]. The channel capacity is often studied as a function of a cost, such as the transmit power. More specifically, the capacity–cost function is defined as the supremum of the mutual information over all input distributions whose cost is either *equal* to a given constant or *upperbounded* by a constant—the convention differs between disciplines. We will return to the distinction between the two definitions at the end of this section.

For the *additive white Gaussian noise* (AWGN) channel, the channel capacity is known exactly [1, Sec. 24], [4, Ch. 9]. In recent years, the problem of calculating or estimating the channel capacity of more complicated channels has received a lot of attention (see surveys in [5]–[8]). Due to the absence of exact analytical solutions and the computational intractability of optimizing over all possible input distributions, most investigations of the channel capacity of non-AWGN channels rely on bounding techniques and asymptotic analysis.

This work was supported in part by the Swedish Foundation for Strategic Research (SSF) under grant RE07-0026 and the Swedish Research Council (VR) under grants 2007-6223 and 2013-5271. E. Agrell is with the Dept. of Signals and Systems, Chalmers Univ. of Technology, SE-41296 Göteborg, Sweden (e-mail: agrell@chalmers.se).

The main motivation for this paper comes from the type of nonlinear distortion encountered in fiber-optical communications. In contrast to linear channels, an optical fiber has the peculiar property that the performance of conventional communication systems degrades if the signal amplitude is increased beyond a certain level [9]–[14]. This phenomenon is well-known from experiments and simulations, and can also be explained theoretically. The lightwave propagation in an optical fiber is governed by a nonlinear differential equation, the *nonlinear Schrödinger equation* or, if polarization effects are considered, the *Manakov equation*. These equations include a nonlinear distortion term, whose amplitude is proportional to the cubed signal amplitude. At high enough signal amplitudes, this nonlinear distortion dominates the other terms in the differential equation, effectively drowning the signal.

Similarly, one might expect that the nonlinear distortion would force the mutual information and channel capacity down to zero at sufficiently high power, and in the past two decades, many results have been published in optical communications to support this conjecture [6]–[8], [15]–[32]. Already in 1993, Splett *et al.* modeled the interference from four-wave mixing in a wavelength-division multiplexing (WDM) system as an AWGN component, under some conditions on the noise and dispersion, in what might have been the first study ever of the channel capacity of a nonlinear optical link [33]. The variance of this AWGN depends nonlinearly on the transmit power, which is assumed equal on all wavelengths. Similar nonlinear channel models have been rediscovered, modified, and further analyzed in [10]–[12], [16], [20], [31], [34]–[36]. Due to the signal-dependent noise, their channel capacities are not monotonic: As the transmit power (or signal-to-noise ratio) increases, the channel capacity increases towards a peak and then decreases again as the power is further increased. Other channel models with signal-dependent AWGN were presented in [30], [32] and have similar nonmonotonic channel capacities. An essential assumption, explicit or implicit, in the derivation of these AWGN-based models is that the transmitted signal consists of independent, identically distributed symbols. This assumption is valid in uncoded transmission systems, but not in the presence of error-correction coding, since coding introduces correlation between symbols. Using a model derived under certain conditions on the transmitted signal is particularly risky in channel capacity calculations, since the channel capacity is by definition the maximum achievable rate using *any* transmission scheme—including those for which the constrained model is not valid.

A continuous-time channel model for cross-phase modulation (XPM) was presented by Mitra and Stark [15]. Although no discrete-time XPM model was obtained, they showed that the channel capacity of the XPM channel model is lowerbounded by the capacity of a signal-dependent AWGN

channel, and that this lower bound is nonmonotonic. They further conjectured that the true channel capacity would have a similar nonmonotonic behavior as its lower bound. Many variants of the Mitra–Stark lower bound have been presented in recent years, often along with the conjecture that the true channel capacity is also nonmonotonic [18], [20], [23], [28], [37]. This conjecture was disproved in the zero-dispersion case by Turitsyn *et al.* [38], who showed that the lower bound based on the AWGN channel [17] is very far from the true channel capacity, and that the channel capacity in fact grows logarithmically with power under certain conditions.

Another type of lower bound on channel capacity is obtained by fixing the input distribution and calculating the mutual information [7], [22], [27], [29], [39], [40] or by optimizing the mutual information over a subset of all possible input distributions [7], [25], [26], [29], [30]. All these lower bounds consistently show a nonmonotonic behavior, decreasing towards zero after a peak at a finite power, and the conjecture that the channel capacity would have a similar nonmonotonic behavior as its lower bounds is often repeated.

We believe that the results cited above, while mathematically correct, do not fully exploit the potential of capacity-achieving coding over nonlinear optical channels. We prove in this work mathematically that for a wide class of channel models, the capacity is a monotonic function (nondecreasing but not necessarily strictly increasing) of the transmit power. This property holds for any *static* channel model, defined as one whose channel law does not change depending on which input distribution it is combined with. The results are extended to a wide class of cost functions and to three specific multiuser scenarios.

The presented results hold regardless of whether the capacity–cost function is defined by maximizing over all input distributions with exactly the given cost or with an upperbounded cost. The proofs are developed assuming the former definition, and they are all trivial for the latter. An interesting consequence of the nondecreasing channel capacity is that the two definitions of the capacity–cost function are fully equivalent.

II. CHANNEL CAPACITY AND COST

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be real, n-dimensional vectors, representing the input and output, resp., of a discrete-time memoryless communication channel. Their respective domains, or alphabets, are denoted by $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^n$. The joint distribution $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ for $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ can be factorized as $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$, where $f_{\boldsymbol{X}}$ is the *input distribution* (which is in practice determined by the modulation format) and $f_{\boldsymbol{Y}|\boldsymbol{X}}$ is the *channel law.* We denote the *mutual information* between $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $I(\boldsymbol{X};\boldsymbol{Y})$, while $I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$ denotes a *conditional mutual information.* The *entropy* and *conditional entropy* are denoted by $H(\boldsymbol{X})$ and $H(\boldsymbol{X}|\boldsymbol{Z})$, resp., and the *differential entropy* and *conditional differential entropy* are denoted by $h(\boldsymbol{X})$ and $h(\boldsymbol{X}|\boldsymbol{Z})$, resp.

Using error-correction coding, codewords of N input symbols are selected from a codebook $\mathcal{C} \subseteq \mathcal{X}^N$. The *rate* of transmission, in bits per symbol, is $\log_2 |\mathcal{C}|/N$. The codewords can be transmitted with arbitrarily small error probability if the codewords are sufficiently long and the rate is sufficiently small. Such a rate is called an *achievable rate*, and the supremum of all achievable rates, over all possible codes and block lengths, is defined as the *operational channel capacity* or simply *capacity* [1, Sec. 1, 14].

Shannon's channel coding theorem [1, Sec. 13, 23], [4, Sec. 7.7, 9.1] states that the operational channel capacity is equal to the *information channel capacity*, which is defined as the supremum of the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ between the channel input and output, where the supremum is taken over all input distributions $f_{\boldsymbol{X}}$. The capacity-achieving distribution may be continuous or discrete [41], [42].[1]

In this work, the channel capacity is characterized as a function of some kind of cost. Closely following the definitions in [43], [44, Sec. 3.3], we define the *cost function* $b(\boldsymbol{x})$ as a deterministic, real, nonnegative function of an input symbol $\boldsymbol{x} \in \mathcal{X}$. The cost of a codeword $\boldsymbol{c} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is defined as $b(\boldsymbol{c}) = (b(\boldsymbol{x}_1) + \cdots + b(\boldsymbol{x}_N))/N$. Let $\tilde{M}(N,p,\beta)$ be the size of the largest codebook $\tilde{\mathcal{C}}$ such that (i) $b(\boldsymbol{c}) \leq \beta$ for all $\boldsymbol{c} \in \tilde{\mathcal{C}}$ and (ii) each codeword can be decoded with an error probability not larger than p. The *capacity–cost function* is defined as [4, Sec. 7.5, 9] [44, Sec. 3.3]

$$\tilde{C}(\beta) \triangleq \lim_{p \to 0} \lim_{N \to \infty} \frac{\log_2 \tilde{M}(N,p,\beta)}{N}. \tag{1}$$

The channel coding theorem with an upperbounded cost now states that [45, Sec. 7.3], [44, Sec. 3.3]

$$\tilde{C}(\beta) = \sup_{f_{\boldsymbol{X}} \in \tilde{\Omega}(\beta)} I(\boldsymbol{X};\boldsymbol{Y}), \tag{2}$$

where $\tilde{\Omega}(\beta)$ is the set of all distributions $f_{\boldsymbol{X}}$ over $\mathcal{X}$ such that $\mathbb{E}[b(\boldsymbol{X})] \leq \beta$. It is well known that the channel capacity, as defined above, is nondecreasing with β [24], [44, Sec. 3.3]. This follows from (2) and the fact that $\tilde{\Omega}(\beta) \supseteq \tilde{\Omega}(\beta')$ for all $\beta \geq \beta'$.

In this paper, we focus on another type of cost constraint. Instead of upperbounding the cost of the codewords as in the previous paragraph, the codewords are all required to have the same exact cost. This scenario has been touched upon in the past [46], [47], but not received as rigorous information-theoretic treatment as the bounded-cost constraint. Formally, let $M(N,p,\beta)$ be the size of the largest codebook $\mathcal{C}$ such that (i) $b(\boldsymbol{c}) = \beta$ for all $\boldsymbol{c} \in \mathcal{C}$ and (ii) each codeword can be decoded with an error probability not larger than p. In analogy with (1), the capacity–cost function is defined as

$$C(\beta) \triangleq \lim_{p \to 0} \lim_{N \to \infty} \frac{\log_2 M(N,p,\beta)}{N}. \tag{3}$$

It is also possible to define the information capacity with an equality constraint, analogous to the right-hand side of (2), as

$$C_{\mathrm{i}}(\beta) \triangleq \sup_{f_{\boldsymbol{X}} \in \Omega(\beta)} I(\boldsymbol{X};\boldsymbol{Y}), \tag{4}$$

where $\Omega(\beta)$ is the set of all distributions $f_{\boldsymbol{X}}$ over $\mathcal{X}$ such that $\mathbb{E}[b(\boldsymbol{X})] = \beta$. This quantity has been analyzed and characterized extensively in optical communications (e.g., [15],

[1] With a slight abuse of notation, we also include distributions that have no probability density function [4, Sec. 8.5].

[20], [25], [21, eq. (11.5)]) and also considered in wireless communications [48], which partly motivates this work. The interest in $C_{\rm i}(\beta)$ comes from an implicit assumption that a channel coding theorem would exist also with an equal-cost constraint, i.e., that $C(\beta) = C_{\rm i}(\beta)$. This relation, while intuitively reasonable, was to our knowledge never formally proven, which may cast some doubts on the operational interpretation of any results based on $C_{\rm i}(\beta)$. In the next section, it will be shown that $C_{\rm i}(\beta)$ is nondecreasing with β and, as a consequence thereof, that indeed $C(\beta) = C_{\rm i}(\beta)$.

An implicit assumption for (4) is that $I(\boldsymbol{X};\boldsymbol{Y})$ is calculated from the same channel law $f_{\boldsymbol{Y}|\boldsymbol{X}}$ for all $f_{\boldsymbol{X}} \in \Omega(\beta)$ and all $\beta > 0$. In other words, the channel remains the same regardless of which codebook is used. This is a standard assumption in information theory [45, Sec. 4.2], and it is not considered to restrict generality. Channel laws with this property are formally defined as *static* in the next section.

Readers with an information theory background have probably only encountered static channel models and may not see the need to define a name for channels with this property. In the optical communications literature, however, channel models $f_{\boldsymbol{Y}|\boldsymbol{X}}$ that change with $f_{\boldsymbol{X}}$ have been proposed frequently. Consider for example the well-known Gaussian noise model for fiber-optical links without dispersion compensation [10]–[12], [31]–[36]. This nonstatic model is in its simplest form given by the channel law

$$f_{Y|X}(y|x) = \frac{1}{\pi(\sigma_0^2 + \eta \mathsf{P}^3)} e^{-\frac{|y-x|^2}{\sigma_0^2+\eta\mathsf{P}^3}}, \tag{5}$$

where X and Y are the complex channel input and output, resp., σ_0^2 and η are two constant link parameters, and $\mathsf{P} = \mathbb{E}[|X|^2]$. This is an AWGN channel, whose noise variance depends on P and hence on $f_{\boldsymbol{X}}$.[2] Its information capacity, obtained by Shannon's standard formula [1, Sec. 24], [4, Sec. 9.1], is commonly given as [31], [33], [36]

$$C(\mathsf{P}) = \log_2\left(1 + \frac{\mathsf{P}}{\sigma_0^2 + \eta\mathsf{P}^3}\right). \tag{6}$$

This function, which clearly decreases to zero at high power P, exemplifies the nonmonotonic behavior of the capacity of certain optical (nonstatic) channel models. We advise that such channel models, while unarguably accurate in uncoded systems [11], [36], should be used with caution in information-theoretic analysis. First, it is not clear whether (4) has any operational meaning in terms of maximum achievable rates for nonstatic channels such as (5). Shannon's channel coding theorem, in its standard memoryless form, assumes that the channel law operates on each symbol $\boldsymbol{X}$ independently, which is not the case if $f_{\boldsymbol{Y}|\boldsymbol{X}}$ changes with $f_{\boldsymbol{X}}$. And second, such models are questionable from a physical viewpoint, as they imply an infinite channel memory [50], [49]. Only static channel models will be considered further in this paper.

III. POINT-TO-POINT CHANNELS

In this section, we are concerned with a discrete-time, memoryless vector channel between a single transmitter and a single receiver, formally defined as follows.

[2] A *static* model for a similar channel as (5) was given in [49, Eq. (13)].

Definition 1: A *static* point-to-point channel is a memoryless relationship $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$ between vectors $\boldsymbol{X} \in \mathcal{X}$ and $\boldsymbol{Y} \in \mathcal{Y}$, which is a function of $\boldsymbol{y}$ and $\boldsymbol{x}$ but does not change with $f_{\boldsymbol{X}}$.

Such a relationship can represent a continuous-time band-limited channel by sampling the transmitted and received waveforms at the Nyquist rate [1, Sec. 23], and it can represent channels with an arbitrarily long (finite) memory by choosing the dimension n much larger than the channel memory [45, Sec. 4.6], [51]. The dimensions may also, in addition to time, represent frequency (wavelength), space, polarization, lightwave modes, or all of these. Hence, the theory applies to a wide variety of channels in different applications.

The capacity is commonly studied as a function of the transmit power, which is obtained by setting $b(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ for all $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^n$. The results in this paper hold not only for transmit power but also more generally for any unbounded cost function, according to the following definition.

Definition 2: An *unbounded cost function* $b(\boldsymbol{x})$ over a domain $\mathcal{X}$ is a real, nonnegative function such that for any given $b_0 \geq 0$, there exists a vector $\boldsymbol{x} \in \mathcal{X}$ for which $b(\boldsymbol{x}) = b_0$.

The main result for point-to-point channels is the following theorem, which implies that the channel capacity will either increase indefinitely or converge to a finite value as the cost increases, depending on the channel. However, it cannot have a peak for any channel or any cost. Despite its simple nature, it has to our knowledge not been stated before.

Theorem 1 (Monotonic Channel Capacity): Let $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$ be a static point-to-point channel defined on $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$. Let $b(\boldsymbol{x})$ be an unbounded cost function on $\mathcal{X}$. Then $C_{\rm i}(\beta)$ is a nondecreasing function of β.

Proof: We will show that for any given pair of costs $\beta \geq \beta' \geq 0$, $C_{\rm i}(\beta) \geq C_{\rm i}(\beta')$. Let, for any $0 < \epsilon \leq 1$,

$$\beta'' \triangleq \beta' + \frac{\beta - \beta'}{\epsilon}. \tag{7}$$

We define a time-sharing random symbol $\boldsymbol{X} \in \mathcal{X}$ given an auxiliary binary random variable Q such that

$$\boldsymbol{X} \triangleq \begin{cases} \boldsymbol{X}', & Q = 0, \\ \boldsymbol{X}'', & Q = 1, \end{cases} \tag{8}$$

where $\Pr\{Q = 1\} = \epsilon$ and the distributions of $\boldsymbol{X}' \in \mathcal{X}$ and $\boldsymbol{X}'' \in \mathcal{X}$ satisfy $\mathbb{E}[b(\boldsymbol{X}')] = \beta'$ and $\mathbb{E}[b(\boldsymbol{X}'')] = \beta''$, resp. Such distributions exist, by assumption, for any costs $\beta', \beta'' \geq 0$. Thus

$$\begin{aligned} \mathbb{E}[b(\boldsymbol{X})] &= (1-\epsilon)\mathbb{E}[b(\boldsymbol{X}')] + \epsilon\mathbb{E}[b(\boldsymbol{X}'')] \\ &= (1-\epsilon)\beta' + \epsilon\beta'' \\ &= \beta. \end{aligned} \tag{9}$$

Because $Q \to \boldsymbol{X} \to \boldsymbol{Y}$ is a Markov chain, the mutual information can be bounded as

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &\geq I(\boldsymbol{X};\boldsymbol{Y}|Q) \\ &= (1-\epsilon)I(\boldsymbol{X};\boldsymbol{Y}|Q=0) + \epsilon I(\boldsymbol{X};\boldsymbol{Y}|Q=1) \\ &\geq (1-\epsilon)I(\boldsymbol{X};\boldsymbol{Y}|Q=0) \\ &= (1-\epsilon)I(\boldsymbol{X}';\boldsymbol{Y}'), \end{aligned} \tag{10}$$

where the first inequality follows from [4, eq. (2.122)] and $\boldsymbol{Y}'$ is defined as the channel output when the input is $\boldsymbol{X}'$. This inequality holds for any $0 < \epsilon \leq 1$ and any distributions $f_{\boldsymbol{X}'} \in \Omega(\beta')$ and $f_{\boldsymbol{X}''} \in \Omega(\beta'')$. Choosing $f_{\boldsymbol{X}'}$ as a capacity-achieving distribution in $\Omega(\beta')$, the right-hand side of (10) becomes $(1-\epsilon)C_{\mathrm{i}}(\beta')$. Thus,

$$\begin{aligned} C_{\mathrm{i}}(\beta) &\geq I(\boldsymbol{X};\boldsymbol{Y}) \\ &\geq (1-\epsilon)C_{\mathrm{i}}(\beta'). \end{aligned} \tag{11}$$

If now $C_{\mathrm{i}}(\beta) < C_{\mathrm{i}}(\beta')$, then (11) would yield a contradiction in the range $0 < \epsilon < 1 - C_{\mathrm{i}}(\beta)/C_{\mathrm{i}}(\beta')$. Hence, $C_{\mathrm{i}}(\beta) \geq C_{\mathrm{i}}(\beta')$. □

Intuitively, an input distribution with a nondecreasing mutual information for a given channel can be constructed by combining two parts, a high-probability part at a moderate cost, which does not vary much as the overall average cost is changed, and a low-probability part, a "satellite," which absorbs the whole increase in average cost by moving away from the other part [13], [14]. As $\epsilon \to 0$, the input distribution $f_{\boldsymbol{X}}$ becomes more and more like $f_{\boldsymbol{X}'}$, while the average cost $\mathbb{E}[b(\boldsymbol{x})]$ remains at β because the lower cost $\beta' < \beta$ is balanced by another cost $\beta'' \gg \beta$.

The nondecreasing nature of the information capacity can be exploited to establish a channel coding theorem with equality cost constraint as follows. Even if Shannon and other information theorists may have been aware of the theorem, we have not seen it in print.

Theorem 2 (Coding Theorem): For any static point-to-point channel $f_{\boldsymbol{Y}|\boldsymbol{X}}$,

$$C(\beta) = \sup_{f_{\boldsymbol{X}} \in \Omega(\beta)} I(\boldsymbol{X};\boldsymbol{Y}). \tag{12}$$

Proof: The theorem will be proved in two steps. First, it is shown that the operational channel capacities (1) and (3) are the same and second, that the information capacity $\sup I(\boldsymbol{X};\boldsymbol{Y})$ is the same regardless of whether the optimization is over $\tilde{\Omega}(\beta)$ or $\Omega(\beta)$. The theorem then follows from the regular channel coding theorem (2).

For the first step, we use the relation

$$M(N,p,\beta) \leq \tilde{M}(N,p,\beta) \leq M(N+1,p,\beta), \tag{13}$$

where the first inequality is trivial from the definitions of M and $\tilde{M}$, whereas the second was proved by Shannon [46, pp. 649–651], [47, eq. (195)], who added an $(N+1)$th symbol to every codeword in a codebook $\mathcal{C}$ with codeword length N and codeword cost at most β, to obtain a codebook $\tilde{\mathcal{C}}$ with codeword length $N+1$ and cost exactly β. Taking the logarithm of all three parts of (13), dividing by N, and letting $N \to \infty$ proves via (1) and (3) that $C(\beta) \leq \tilde{C}(\beta) \leq C(\beta)$, in other words $C(\beta) = \tilde{C}(\beta)$.

For the second step, the information capacity (2) is written as

$$\tilde{C}(\beta) = \sup_{\beta' \leq \beta} C_{\mathrm{i}}(\beta'), \tag{14}$$

which by Theorem 1 is equal to $C_{\mathrm{i}}(\beta)$. Combining the two steps, $C(\beta) = \tilde{C}(\beta) = C_{\mathrm{i}}(\beta)$ and (12) follows. □

This means not only that a channel coding theorem holds for an equal-cost constraint but also that the two channel capacities (1) and (3) are equivalent. The cost-limited channel capacity $\tilde{C}(\beta)$ is achieved by an input distribution $f_{\boldsymbol{X}}$ for which the cost equals the maximum allowed value β. A practical interpretation is that when designing a capacity-achieving code for a nonlinear channel, it suffices to consider only codes for which all codewords have the same cost β.

IV. INTERFERENCE CHANNELS

We consider a discrete-time, memoryless interference channel with k users, each with the purpose of transmitting a message from a transmitter to a receiver [44, Ch. 6], for example an optical WDM system. The input and output are denoted by $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$, resp., for $i = 1, \ldots, k$. The ith receiver attempts to recover $\boldsymbol{X}_i$ based on $\boldsymbol{Y}_i$, without knowledge of $\boldsymbol{Y}_j$ for $j \neq i$. The statistics of the received vectors is given by the conditional distribution $f_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_k|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k}$, which does not change with the cost. Independent data is transmitted by each user, and the joint *a priori* input distribution $f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k}$ is therefore equal to the product of the marginal distributions $f_{\boldsymbol{X}_1} \cdots f_{\boldsymbol{X}_k}$. All input distributions $f_{\boldsymbol{X}_i}$ are known to all users. From the viewpoint of user i, all interfering input symbols $\boldsymbol{X}_j$ for $j \neq i$ are assumed to be independent between channel uses. This assumption, which is conventional in optical communications, is valid if the codebook of user j is not known to user i or if user j transmits uncoded data.

Three scenarios, or *behavioral models* [52], are considered in the following subsections. The aim in the first two scenarios is to determine the maximum achievable rate of the primary user, referred to as user 1, while treating the signals from the other users $\boldsymbol{X}_2, \ldots, \boldsymbol{X}_k$ as (nonlinear) noise. The received vectors $\boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_k$ are unknown at receiver 1 and the channel can be represented by the conditional distribution $f_{\boldsymbol{Y}_1|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k}$. The third and last scenario represents joint optimization of $f_{\boldsymbol{X}_1}, \ldots, f_{\boldsymbol{X}_k}$, considering the full interference channel model $f_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_k|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k}$. The point-to-point case extends straightforwardly to the first and third case (Sec. IV-A and IV-C), whereas the second case requires a somewhat more elaborate treatment (Sec. IV-B). There also exist behavioral models, not treated in this paper, for which the capacity is not monotonic [52].

The following lemma about conditional mutual information will be useful in Sec. IV-B.

Lemma 3: For any $\boldsymbol{X}$ and $\boldsymbol{Y}$, and any discrete $\boldsymbol{Z}$,

$$|I(\boldsymbol{X};\boldsymbol{Y}) - I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})| \leq H(\boldsymbol{Z}). \tag{15}$$

Proof: By the chain rule for mutual information,

$$I(\boldsymbol{X};\boldsymbol{Y},\boldsymbol{Z}) = I(\boldsymbol{X};\boldsymbol{Y}) + I(\boldsymbol{X};\boldsymbol{Z}|\boldsymbol{Y}), \tag{16}$$

$$I(\boldsymbol{X};\boldsymbol{Y},\boldsymbol{Z}) = I(\boldsymbol{X};\boldsymbol{Z}) + I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}). \tag{17}$$

Eliminating $I(\boldsymbol{X};\boldsymbol{Y},\boldsymbol{Z})$ and rearranging terms,

$$|I(\boldsymbol{X};\boldsymbol{Y}) - I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})| = |I(\boldsymbol{X};\boldsymbol{Z}) - I(\boldsymbol{X};\boldsymbol{Z}|\boldsymbol{Y})|. \tag{18}$$

Since $\boldsymbol{Z}$ is discrete by assumption, the right-hand side can be upperbounded using

$$0 \leq I(\boldsymbol{X};\boldsymbol{Z}) \leq H(\boldsymbol{Z}), \tag{19}$$

$$0 \leq I(\boldsymbol{X};\boldsymbol{Z}|\boldsymbol{Y}) \leq H(\boldsymbol{Z}|\boldsymbol{Y}) \leq H(\boldsymbol{Z}), \tag{20}$$

which completes the proof. □

A. Fixed Interference Distributions

Suppose that the input distributions $f_{\boldsymbol{X}_2},\ldots,f_{\boldsymbol{X}_k}$ are fixed and do not change even if $f_{\boldsymbol{X}_1}$ would change. From the viewpoint of the primary user, the interference caused by the other users can be included in the channel model. Since the conditional distribution $f_{\boldsymbol{Y}_1|\boldsymbol{X}_1}$ in this case depends on the distributions $f_{\boldsymbol{X}_2},\ldots,\boldsymbol{X}_k$, but not on $f_{\boldsymbol{X}_1}$, Theorem 1 applies and the capacity–cost function for the primary user is nondecreasing. This scenario was called *behavioral model (a)* in [52].

B. Equal Distributions

In this section and the next, we consider scenarios where the distributions of all users are governed by a *network controller,* by assigning modulation formats and power levels to all users. While the next section discusses the case of joint optimization over all user distributions, we assume in this section that all users apply the same input distribution $f_{\boldsymbol{X}}$, or linearly rescaled versions thereof. Hence, the joint distribution of the users is

$$f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k) = \prod_{i=1}^{k} \alpha_i^n f_{\boldsymbol{X}}(\alpha_i \boldsymbol{x}_i), \tag{21}$$

for some given constants $\alpha_1,\ldots,\alpha_k$.

An important special case is $\alpha_1 = \cdots = \alpha_k = 1$, which makes all marginal distributions $f_{\boldsymbol{X}_1}, f_{\boldsymbol{X}_2},\ldots,f_{\boldsymbol{X}_k}$ identical. This special case, called *behavioral model (c)* in [52], was studied in, e.g., [7], [26]. The power scaling via $\alpha_1,\ldots,\alpha_k$ provides additional degrees of freedom.

The network controller may wish to select the distribution $f_{\boldsymbol{X}}$ such that the achievable rate of any single channel, say, channel 1, is maximized. Hence, we define the *constrained information capacity* of the primary channel as

$$C_1(\beta) \triangleq \sup I(\boldsymbol{X}_1;\boldsymbol{Y}_1), \tag{22}$$

where the supremum is taken over all distributions of the form (21), with $f_{\boldsymbol{X}} \in \Omega(\beta)$. Clearly, the mutual information $I(\boldsymbol{X}_1;\boldsymbol{Y}_1)$ is an achievable rate for this channel, for any $f_{\boldsymbol{X}}$, and hence $C_1(\beta)$ is an achievable rate. In the context of optical communications, C_1 was studied in [7], [26], [32]. By analogy with the point-to-point channel, it might be tempting to interpret C_1 as the *maximum* achievable rate under certain conditions; however, we believe that no such claims can be made without a precisely stated coding theorem for the interference channel, which is presently lacking.

Theorem 4: The constrained information capacity $C_1(\beta)$ is a nondecreasing function of $\beta > 0$, for any interference channel $f_{\boldsymbol{Y}_1|\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_k}$.

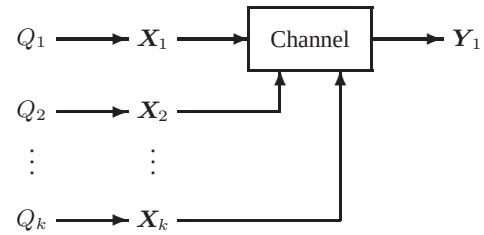

Fig. 1. An interference channel with time-sharing inputs, analyzed in Sec. IV-B. The primary channel $\boldsymbol{X}_1 \to \boldsymbol{Y}_1$ is affected by interference from the other inputs $\boldsymbol{X}_2,\ldots,\boldsymbol{X}_k$, which are all independent.

Proof: Let $f_{\boldsymbol{X}'} \in \Omega(\beta')$ be a distribution[3] for which the supremum in (22) is attained, i.e., $C_1(\beta') = I(\boldsymbol{X}_1;\boldsymbol{Y}_1)$, at some cost $\beta' \geq 0$. We will show that $C_1(\beta) \geq C_1(\beta')$ for any $\beta \geq \beta'$.

For any given $\beta \geq \beta'$ and $0 < \epsilon \leq 1$, let

$$\beta'' \triangleq \beta' + \frac{\beta - \beta'}{\epsilon} \tag{23}$$

and let $f_{\boldsymbol{X}''}$ be any distribution over $\mathcal{X}$ with $\mathbb{E}[b(\boldsymbol{X}'')] = \beta''$. We now define a time-sharing random vector $\boldsymbol{X}$ given an auxiliary binary random variable Q such that

$$\boldsymbol{X} \triangleq \begin{cases} \boldsymbol{X}', & Q = 0, \\ \boldsymbol{X}'', & Q = 1, \end{cases} \tag{24}$$

where $\Pr\{Q = 1\} = \epsilon$. This vector satisfies

$$\begin{aligned} \mathbb{E}[b(\boldsymbol{X})] &= (1-\epsilon)\mathbb{E}[b(\boldsymbol{X}')] + \epsilon\mathbb{E}[b(\boldsymbol{X}'')] \\ &= (1-\epsilon)\beta' + \epsilon\beta'' \\ &= \beta. \end{aligned} \tag{25}$$

As illustrated in Fig. 1, the joint distribution $f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k}$ is generated by an analogous k-fold time-sharing method, using the auxiliary variables $Q_1,\ldots,Q_k$. These variables have the same distribution as Q and are independent. They control the input symbols $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k$ such that $\boldsymbol{X}_i = \boldsymbol{X}_i'$ if $Q_i = 0$ and $\boldsymbol{X}_i = \boldsymbol{X}_i''$ if $Q_i = 1$, where

$$f_{\boldsymbol{X}_i'}(\boldsymbol{x}) = \alpha_i^n f_{\boldsymbol{X}'}(\alpha_i \boldsymbol{x}), \tag{26}$$

$$f_{\boldsymbol{X}_i''}(\boldsymbol{x}) = \alpha_i^n f_{\boldsymbol{X}''}(\alpha_i \boldsymbol{x}) \tag{27}$$

for $i = 1,\ldots,k$. Obviously, the time-sharing symbols $\boldsymbol{X}_i$ jointly follow the desired distribution (21).

The mutual information of the primary channel can be bounded as

$$\begin{aligned} I(\boldsymbol{X}_1;\boldsymbol{Y}_1) &\geq I(\boldsymbol{X}_1;\boldsymbol{Y}_1|Q_1) && (28) \\ &\geq I(\boldsymbol{X}_1;\boldsymbol{Y}_1|Q_1,Q_2,\ldots,Q_k) \\ &\quad - H(Q_2,\ldots,Q_k), && (29) \end{aligned}$$

where (28) holds because $Q_1 \to \boldsymbol{X}_1 \to \boldsymbol{Y}_1$ is a Markov chain and (29) follows by setting $\boldsymbol{Z} = [Q_2,\ldots,Q_k]$ in Lemma 3. The first term of the right-hand side of (29) can be bounded

[3]Or, more precisely, a sequence of distributions.

as

$$\begin{aligned} &I(\boldsymbol{X}_1;\boldsymbol{Y}_1|Q_1,Q_2,\ldots,Q_k) \\ &\quad= \sum_{(q_1,\ldots,q_k)\in\{0,1\}^k} \Pr\{Q_1=q_1,\ldots,Q_k=q_k\} \\ &\qquad\qquad \cdot I(\boldsymbol{X}_1;\boldsymbol{Y}_1|Q_1=q_1,\ldots,Q_k=q_k) \\ &\quad\geq \Pr\{Q_1=\cdots=Q_k=0\} \\ &\qquad \cdot I(\boldsymbol{X}_1;\boldsymbol{Y}_1|Q_1=\cdots=Q_k=0) \\ &\quad= (1-\epsilon)^k I(\boldsymbol{X}_1';\boldsymbol{Y}_1') \\ &\quad= (1-\epsilon)^k C_1(\beta'). \end{aligned} \tag{30}$$

The second term of the right-hand side of (29) is

$$\begin{aligned} H(Q_2,\ldots,Q_k) &= \sum_{i=2}^{k} H(Q_i) \\ &= (k-1)H_2(\epsilon), \end{aligned} \tag{31}$$

where $H_2(u) \triangleq -u\log_2 u - (1-u)\log_2(1-u)$. Combining (22), (29), (30), and (31) yields

$$\begin{aligned} C_1(\beta) &= \sup_{f_{\boldsymbol{X}}\in\Omega(\beta)} I(\boldsymbol{X}_1;\boldsymbol{Y}_1) \\ &\geq \sup_{0<\epsilon\leq 1} \left[(1-\epsilon)^k C_1(\beta') - (k-1)H_2(\epsilon)\right] \\ &= \lim_{\epsilon\to 0} \left[(1-\epsilon)^k C_1(\beta') - (k-1)H_2(\epsilon)\right] \\ &= C_1(\beta'), \end{aligned} \tag{32}$$

which completes the proof. □

Intuitively, the proof relies on constructing a "satellite distribution" [13] for $\boldsymbol{X}$, where the "satellite," denoted by $\boldsymbol{X}''$ in (24), carries a much higher cost than $\boldsymbol{X}'$ and occurs with lower probability.

C. Joint Optimization

In the third and last scenario, we assume that the system includes a mechanism to optimize the transmission schemes of all users jointly, for example via a central network controller. As in the previous two scenarios, the transmitters and receivers are still *operated* separately, in the sense that the transmitters and receivers do not exchange information about their respective signals.[4]

Let R_i be an achievable rate for the transmitter–receiver pair $i = 1,\ldots,k$ and let $\boldsymbol{R} \triangleq (R_1,\ldots,R_k)$ be a vector of rates that can be *simultaneously* achieved over the interference channel, with arbitrarily small error probability. The *capacity region* $\mathscr{C}(\boldsymbol{\beta})$, where $\boldsymbol{\beta} \triangleq (\beta_1,\ldots,\beta_k)$, is defined as the closure of the set of all achievable rate vectors $\boldsymbol{R}$ when every codeword used by user $i = 1,\ldots,k$ has the exact cost β_i [44, Sec. 4.1, 6.1]. While no analytical expression is known for the capacity region of general interference channels [44, Ch. 6], the monotonicity can be established via the following theorem.

[4]If data instead is jointly encoded over all transmitted signals $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k$ and jointly decoded based on all received signals $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_k$, then the channel is equivalent to a high-dimensional point-to-point channel and Theorem 1 applies.

Theorem 5: Let $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_k)$ and $\boldsymbol{\beta}' = (\beta_1',\ldots,\beta_k')$ be two cost vectors such that $\beta_i \geq \beta_i' \geq 0$ for $i = 1,\ldots,k$. Then their capacity regions satisfy $\mathscr{C}(\boldsymbol{\beta}) \supseteq \mathscr{C}(\boldsymbol{\beta}')$.

Proof: Let, for any $0 < \epsilon \leq 1$,

$$\boldsymbol{\beta}'' \triangleq \boldsymbol{\beta}' + \frac{\boldsymbol{\beta}-\boldsymbol{\beta}'}{\epsilon}. \tag{33}$$

Let $\boldsymbol{R}'$ and $\boldsymbol{R}''$ be achievable rate vectors at costs $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}''$, resp. By time sharing [4, Sec. 15.3.3], [44, Sec. 4.4], the rate

$$(1-\epsilon)\boldsymbol{R}' + \epsilon\boldsymbol{R}'' \geq (1-\epsilon)\boldsymbol{R}' \tag{34}$$

is achievable at cost

$$(1-\epsilon)\boldsymbol{\beta}' + \epsilon\boldsymbol{\beta}'' = \boldsymbol{\beta}. \tag{35}$$

The capacity region $\mathscr{C}(\boldsymbol{\beta})$ thus includes all rate vectors of the form $(1-\epsilon)\boldsymbol{R}'$, where $\boldsymbol{R}'$ is achievable at cost $\boldsymbol{\beta}'$ and ϵ is an arbitrarily small positive number. Since the capacity region by definition is the *closure* of all achievable rate vectors [44, Sec. 4.1, 6.1], $\mathscr{C}(\boldsymbol{\beta})$ also includes $\lim_{\epsilon\to 0}(1-\epsilon)\boldsymbol{R}' = \boldsymbol{R}'$. In conclusion, $\boldsymbol{R}' \in \mathscr{C}(\boldsymbol{\beta})$ for all $\boldsymbol{R}' \in \mathscr{C}(\boldsymbol{\beta}')$, which implies $\mathscr{C}(\boldsymbol{\beta}) \supseteq \mathscr{C}(\boldsymbol{\beta}')$. □

The capacity region is a k-dimensional object, and it varies as a function of the k-dimensional vector $\boldsymbol{\beta}$. The following two corollaries exemplify how linear combinations of the achievable rates change when the cost is varied linearly.

Corollary 6: If the cost is varied along a line as

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \mu\boldsymbol{\Delta}, \tag{36}$$

where all components of $\boldsymbol{\beta}_0$ and $\boldsymbol{\Delta}$ are nonnegative, then all achievable rates $R_1,\ldots,R_k$ are nondecreasing functions of $\mu \geq 0$, and the achievable sum rate $R_1 + \cdots + R_k$ is also a nondecreasing function of $\mu \geq 0$.

Corollary 7: If all transmitters obey the same cost constraint $\beta_1 = \cdots = \beta_k = \beta$, then all achievable rates $R_1,\ldots,R_k$ are nondecreasing functions of β.

V. NUMERICAL EXAMPLE

In this section, examples are given for mutual information and channel capacity as functions of the transmit power, for a simple nonlinear channel. The studied channel is chosen mainly for its simplicity, because evaluating the channel capacity is numerically possible only for very low-dimensional, memoryless channels, which unfortunately excludes more realistic channel models.

A. A Nonlinear Channel

We consider a very simple channel with nonlinear distortion and additive noise, represented as

$$Y = a(X) + Z, \tag{37}$$

where X and Y are the input and output of the channel, resp., $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $a(\cdot)$ is a given deterministic function, and Z is white Gaussian noise with zero mean and variance σ_Z^2. For a given channel input x, the channel law is given by the conditional probability density function (pdf)

$$f_{Y|X}(y|x) = \frac{1}{\sigma_Z} f_G\left(\frac{y-a(x)}{\sigma_Z}\right), \tag{38}$$

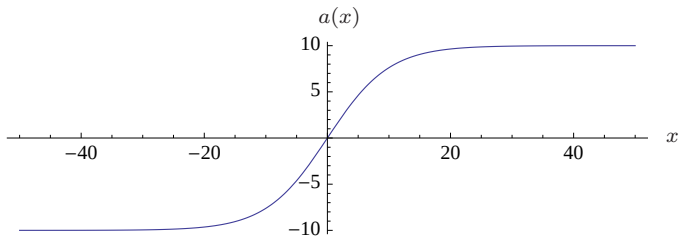

Fig. 2. A simple example of nonlinear distortion, given by (40) for $a_{\max} = 10$. The channel is essentially linear for small $|x|$ and binary for large $|x|$.

where $f_G(x) \triangleq (1/\sqrt{2\pi})\exp(-x^2/2)$ is the zero-mean, unit-variance Gaussian pdf. Since $f_{Y|X}(y|x)$ is Gaussian for any x, the conditional entropy is [1, Sec. 20], [4, Sec. 8.1]

$$h(Y|X) = \frac{1}{2}\log_2 2\pi e\sigma_Z^2. \tag{39}$$

For a given input distribution f_X, the output distribution f_Y is obtained by marginalizing the joint distribution $f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x)$, and the mutual information is calculated as $I(X;Y) = h(Y) - h(Y|X)$.

In this example, we select $a(x)$ in (37) as a smooth clipping function

$$a(x) = a_{\max}\tanh\left(\frac{x}{a_{\max}}\right), \tag{40}$$

where $a_{\max} > 0$ sets an upper bound on the output. The hyperbolic tangent is commonly used to model nonlinear amplifiers [53], [54] and similar characteristics, albeit biased, can model a light-emitting diode in intensity-modulated optical systems [55]. If the instantaneous channel input X has a sufficiently high magnitude compared with $a_{\max}$, the channel is essentially binary. For X close to zero, on the other hand, the channel approaches a linear AWGN channel.

The channel parameters are $a_{\max} = 10$ and $\sigma_Z = 1$ throughout this section. The function $a(x)$ in (40), which represents the nonlinear part of the channel (37), is shown in Fig. 2. Since the channel law $f_{Y|X}(y|x)$ given by (38) and (40) depends on x, y, and $a_{\max}$ but nothing else, the channel is static according to Definition 1 and Theorem 1 applies.

B. Mutual Information

The mutual information $I(X;Y)$ is evaluated by numerical integration, as a function of the average transmit power $\mathsf{P} = \mathbb{E}[X^2]$. No optimization over input distributions is carried out. The input distribution $f_X(x)$ is constructed from a given unit-power distribution $g(x)$, rescaled to the desired power P as $f_X(x) = \alpha g(\alpha x)$, where $\alpha = 1/\sqrt{\mathsf{P}}$. The results are presented in Fig. 3 for three continuous input pdfs $f_X(x)$: zero-mean Gaussian, zero-mean uniform, and single-sided exponential,

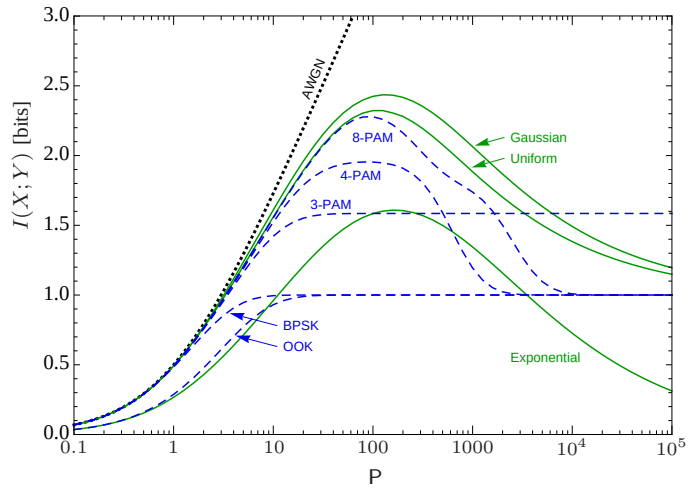

Fig. 3. Mutual information for the nonlinear channel in (37) with $a_{\max} = 10$ and $\sigma_Z = 1$ with various continuous (solid) and discrete (dashed) input distributions. The discrete distributions have uniform probabilities and equal spacing. The AWGN channel capacity is included for reference (dotted).

defined as, respectively,

$$f_{X_1}(x) = \frac{1}{\sqrt{\mathsf{P}}} f_G\left(\frac{x}{\sqrt{\mathsf{P}}}\right), \tag{41}$$

$$f_{X_2}(x) = \begin{cases} \frac{1}{2\sqrt{3\mathsf{P}}}, & -\sqrt{3\mathsf{P}} \le x \le \sqrt{3\mathsf{P}}, \\ 0, & \text{elsewhere}, \end{cases} \tag{42}$$

$$f_{X_3}(x) = \begin{cases} \sqrt{\frac{2}{\mathsf{P}}}e^{-x\sqrt{2/\mathsf{P}}}, & x \ge 0, \\ 0, & x < 0. \end{cases} \tag{43}$$

At asymptotically low power P, the channel is effectively an AWGN channel. In this case, the mutual information is governed by the mean value of the input distribution, according to [56]. All zero-mean input distributions achieve approximately the same mutual information, which approaches the AWGN channel capacity. The asymptotic mutual information for the exponential distribution, whose mean is $\sqrt{\mathsf{P}/2}$, is half that achieved by zero-mean distributions.

The mutual information curves for all three input pdfs reach a peak around $\mathsf{P} = 100$, when a large portion of the input samples still fall in the linear regime of the channel. When the transmit power P is further increased, the mutual information decreases towards a value slightly less than 1 bit asymptotically for the zero-mean input pdfs and 0 for the exponential input. The asymptotes are explained by the fact that at high enough power, almost all input samples fall in the nonlinear regime, where the channel behaves as a 1-bit noisy quantizer. The same argument shows that for *any* continuous distribution, the asymptotical mutual information is less than 1 bit.

Similar results for various discrete input distributions are also included in Fig. 3. The studied one-dimensional constellations are on–off keying (OOK), binary phase-shift keying (BPSK), and m-ary pulse amplitude modulation (m-PAM). The constellation points are equally spaced and the input samples X are chosen uniformly from these constellations.

The mutual information for m-PAM constellations with $m \geq 4$ exhibits the same kind of peak as the continuous distributions in Fig. 3; indeed, a uniform distribution over equally spaced m-PAM approaches the continuous uniform distribution as $m \to \infty$.

Similarly to the continuous case, the mutual information for zero-mean discrete input distributions approach the AWGN channel capacity as $\mathsf{P} \to 0$. Half this channel capacity is achieved by the OOK input, which has the same mean value $\sqrt{\mathsf{P}/2}$ as the exponential input above. The asymptotics when $\mathsf{P} \to \infty$ depends on whether the input distribution includes a nonzero probability mass at $X = 0$. If not, the channel again acts like a 1-bit quantizer and the asymptotic mutual information is slightly less than 1. For distributions with a probability mass at $X = 0$, here exemplified by 3-PAM, the channel asymptotically approaches a ternary-output noisy channel whose possible outputs are not only $Y = \pm a_{\max} + Z$ but also $Y = 0 + Z$. Hence, for any input distribution (discrete, continuous, or mixed), the mutual information is upperbounded by $\log_2 3 = 1.58$.

In conclusion, this particular channel has the property that the mutual information for any input distribution approaches a limit as $\mathsf{P} \to \infty$, and this limit is upperbounded by $\log_2 3$. It might seem tempting to conclude that the channel capacity, which is the supremum of all mutual information curves, would behave similarly. However, as we shall see in the next section, this conclusion is not correct, because the limit of a supremum is in general not equal to the supremum of a limit. Specifically, the asymptotical channel capacity is $\lim_{\mathsf{P}\to\infty} C(\mathsf{P}) = \lim_{\mathsf{P}\to\infty} \sup_g I(X;Y)$, which is not equal to $\sup_g \lim_{\mathsf{P}\to\infty} I(X;Y) \leq \log_2 3$.

C. Channel Capacity

The standard method to calculate the channel capacity of a discrete memoryless channel is by the *Arimoto–Blahut algorithm* [57], [58], [4, Sec. 10.8], [59, Ch. 9]. It has been extended to continuous-input, continuous-output channels in [60] and furthermore to cost-constrained inputs in [61]. The idea in [61] is to represent distributions by lists of samples, so-called *particles.* A particle-based input distribution has the form

$$f_X(x) = \sum_{i=1}^{s} w_i \delta(x - c_i), \tag{44}$$

where $\delta(\cdot)$ is the Dirac delta function, s is the number of particles, $\boldsymbol{c} = (c_1, \ldots, c_s)$ are the particles, and $\boldsymbol{w} = (w_1, \ldots, w_s)$ are the probabilities, or weights, associated with each particle. If s is large enough, any distribution can be represented in the form (44) with arbitrarily small error. With this representation,

$$f_Y(y) = \sum_{i=1}^{s} \frac{w_i}{\sigma_Z} f_G\left(\frac{y - a(c_i)}{\sigma_Z}\right), \tag{45}$$

which yields $h(Y)$, and thereby $I(X;Y)$, by numerical integration.

Since $h(Y|X)$ is constant, the capacity is obtained by maximizing $h(Y)$ subject to constraints on the total probability and power. This problem is in general nonconvex. In [61], the optimization is done by *alternating optimization* [59, Sec. 9.1], first finding $\boldsymbol{w}$ for a given $\boldsymbol{c}$ using the Arimoto–Blahut algorithm and then finding $\boldsymbol{c}$ for a given $\boldsymbol{w}$ using a gradient search, and so on. Here, we apply gradient search techniques for both steps. The objective is to maximize the Lagrangian function

$$L(\boldsymbol{c}, \boldsymbol{w}, \lambda_1, \lambda_2) \triangleq h(Y) + \lambda_1 \left(\sum_{i=1}^{s} w_i - 1\right) + \lambda_2 \left(\sum_{i=1}^{s} w_i c_i^2 - \mathsf{P}\right), \tag{46}$$

where the Lagrange multipliers λ_1 and λ_2 are determined to maintain the constraints $\sum_i w_i = 1$ and $\sum_i w_i c_i^2 = \mathsf{P}$ during the optimization process. The gradients of L with respect to $\boldsymbol{c}$ and $\boldsymbol{w}$ are calculated, and a steepest descent algorithm (or more accurately, "steepest ascent") is applied to maximize L. In each iteration, a step is taken in the direction of either of the two gradients.[5] The step size is determined using the *golden section method* [62, Sec. 10.4]. Several initial values $(\boldsymbol{c}, \boldsymbol{w})$ were tried. The number of particles s was heuristically chosen by doubling its value until the obtained channel capacity changed by less than 0.01. This convergence criterion was satisfied at $s = 16$ in all cases.

The topography of L as a function of $\boldsymbol{c}$ and $\boldsymbol{w}$ turned out to include vast flat fields, where a small step has little influence on L. This made the optimization numerically challenging. No suboptimal local maxima were found for the studied channel and constraints, although for nonlinear channels in general, the mutual information as a function of the input distribution may have multiple maxima.[6]

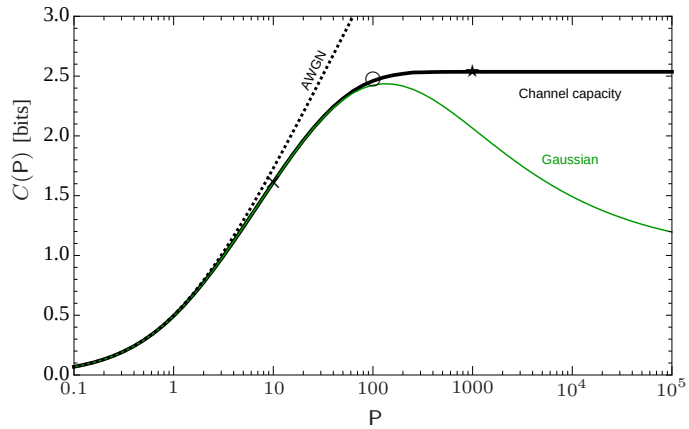

Fig. 4. Channel capacity for the same channel (thick solid), compared with the mutual information of the Gaussian distribution in Fig. 3 (thin solid) and the AWGN channel capacity (dotted). Even though most mutual information curves decrease, the channel capacity does not, thus supporting Theorem 1. The three markers refer to distributions in Fig. 5.

[5]Moving in the direction of the joint gradient turned out to be less efficient, because for small and large P, the numerical values of $\boldsymbol{c}$ and $\boldsymbol{w}$ are not of the same order of magnitude.

[6]An exception occurs when the constellation points $\boldsymbol{c}$ are fixed and the only constraint is $\sum w_i = 1$. In this special case, the mutual information is a concave function of $\boldsymbol{w}$ for any channel [4, Sec. 2.7, 7.3] and there is thus a unique maximum.

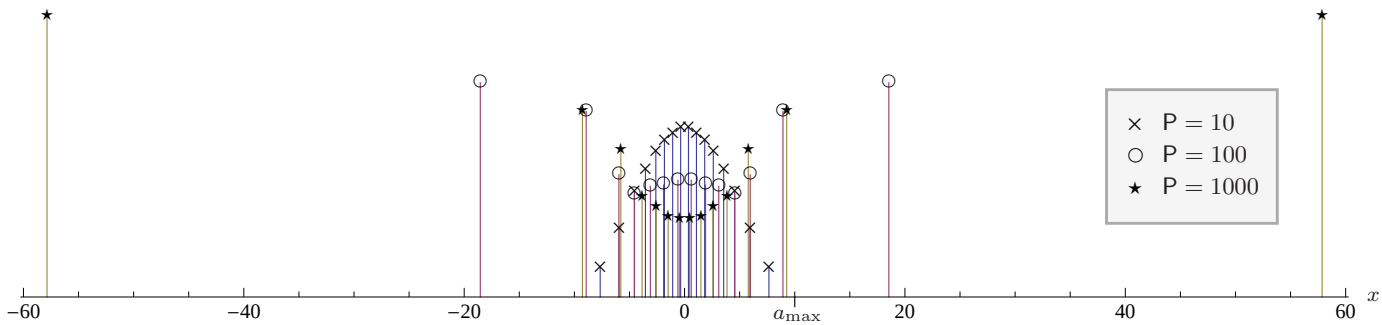

Fig. 5. Discrete approximations of the capacity-achieving input distributions for $\mathsf{P} = 10$, 100, and 1000.

This channel capacity, numerically obtained by the above method, is shown in Fig. 4 for the studied channel (thick solid curve). As promised by Theorem 1, the curve differs from most mutual information curves by not having a peak at any P. The channel capacity follows the mutual information of the Gaussian distribution closely until around $\mathsf{P} = 100$. However, while the Gaussian case attains its maximum mutual information $I(X;Y) = 2.44$ bits/symbol at $\mathsf{P} = 130$ and then begins to decrease, the channel capacity continues to increase towards its asymptote $\lim_{\mathsf{P}\to\infty} C(\mathsf{P}) = 2.54$ bits/symbol. The fact that the capacity curve rises somewhat over the peak and not only flattens out is encouraging for future work on capacity-achieving coding for more realistic nonlinear channel.

This asymptotical channel capacity can be explained as follows. Define the random variable $A \triangleq a(X)$. Since $a(\cdot)$ is a continuous, strictly increasing function, there is a one-to-one mapping between $X \in (-\infty, \infty)$ and $A \in (-a_{\max}, a_{\max})$. Thus $I(X;Y) = I(A;Y)$, where $Y = A + Z$. This represents a standard discrete-time AWGN channel whose input A is subject to a peak power constraint. The capacity of a peak-power-constrained AWGN channel was bounded already in [1, Sec. 25] and computed numerically in [41], where it was also shown that the capacity-achieving distribution is discrete. The asymptote in Fig. 4, which is 2.54 bits/symbol or, equivalently, 1.76 nats/symbol, agrees perfectly with the amplitude-constrained capacity in [41, Fig. 2] for $a_{\max}/\sigma_Z = 10$.

Some almost capacity-achieving input distributions are shown in Fig. 5, numerically optimized as described above. For $\mathsf{P} = 10$, the optimized discrete input distribution is essentially a nonuniformly sampled Gaussian pdf, and the obtained channel capacity, 1.61, has the same value as the mutual information of a continuous Gaussian pdf, shown in Fig. 3. For $\mathsf{P} = 100$ and 1000, the distribution is more uniform in the range where the channel behaves more or less linearly, which for this channel is approximately at $-a_{\max}/2 < x < a_{\max}/2$, with some high-power outliers in the nonlinear range $|x| > a_{\max}$. In all cases, increasing the number of particles s from what is shown in Fig. 5 does not increase the mutual information significantly, from which we infer that these discrete input distributions perform practically as well as the best discrete or continuous input distributions for this channel.

Although the capacity-achieving distributions would look quite different for other types of nonlinear channels, a general observation can be made from Fig. 5: Even at high average power, the input should consist of samples with moderate power, for which the channel is good, most of the time. The high average power is achieved by a single particle having a very large power; thus, the capacity-achieving distribution is a satellite distribution [13]. This single particle, or satellite, corresponds to $\boldsymbol{X}''$ and $\boldsymbol{X}_1''$ in the proofs of Theorems 1 and 4, resp., which as $\epsilon \to 0$ have high cost (power) and low probability.

VI. SUMMARY AND CONCLUSIONS

It was proved that the channel capacity is a nondecreasing function of a cost (such as transmit power) in the following cases:

- Point-to-point memoryless vector channels $f_{\boldsymbol{Y}|\boldsymbol{X}}$ that do not change with the input distribution $f_{\boldsymbol{X}}$.
- Interference channels where all users, except the one of interest, transmit data from fixed input distributions.
- Interference channels where all users transmit data from the same (optimized) distribution.
- Interference channels where the distributions of all users are optimized jointly.

The mutual information may be decreasing with cost in all these cases, but not the channel capacity in Shannon's sense.

In contrast, there are numerous examples in the literature where the channel capacity, or numerical approximations thereof, has a peak at a certain cost, after which it decreases towards zero [6]–[8], [15]–[33]. These examples all pertain to one of the following cases:

- Point-to-point channels that change depending on the transmitter settings, typically as a function of the transmit power [50].
- Interference channels where the transmission scheme of one user (the one of interest) is optimized while the other users satisfy the same power constraint by pure amplification [52].

A practical interpretation is that when designing codes for nonlinear channels under the constraint of a maximum average power, it suffices to consider codes in which all codewords satisfy the power constraint with equality. This is in contrast to previous works in optical communications, which often assumed the existence of an optimal (finite) power. Further

research is needed to show whether the new approach is just a way to achieve the same rates as before at a higher power, or if it may lead to significantly increased achievable rates.

ACKNOWLEDGMENTS

The author is indebted to A. Alvarado, P. Bayvel, L. Beygi, G. Durisi, T. Eriksson, R.-J. Essiambre, C. Häger, M. Karlsson, J. Karout, G. Kramer, F. R. Kschischang, E. Telatar, and S. K. Turitsyn for helpful discussions, criticism, and comments on early versions of this work. The presentation was greatly improved thanks to detailed feedback from the Editor S. Hranilovic and four insightful reviewers.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July, Oct. 1948.

[2] ——, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. IT-2, no. 3, pp. 8–19, Sept. 1956.

[3] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Trans. Inform. Theory*, vol. IT-2, no. 4, pp. 102–108, Dec. 1956.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.

[5] M. Katz and S. Shamai (Shitz), "On the capacity-achieving distribution of the discrete-time noncoherent and partially coherent AWGN channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2257–2270, Oct. 2004.

[6] J. M. Kahn and K.-P. Ho, "Spectral efficiency limits and modulation/detection techniques for DWDM systems," *IEEE J. Sel. Topics Quantum Electron.*, vol. 10, no. 2, pp. 259–272, Mar./Apr. 2004.

[7] R.-J. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," *J. Lightw. Technol.*, vol. 28, no. 4, pp. 662–701, Feb. 2010.

[8] R. I. Killey and C. Behrens, "Shannon's theory in nonlinear systems," *J. Modern Opt.*, vol. 58, no. 1, pp. 1–10, Jan. 2011.

[9] A. Mecozzi, "Limits to long-haul coherent transmission set by the Kerr nonlinearity and noise of the in-line amplifiers," *J. Lightw. Technol.*, vol. 12, no. 11, pp. 1993–2000, Nov. 1994.

[10] A. Bononi, P. Serena, N. Rossi, E. Grellier, and F. Vacondio, "Modeling nonlinearity in coherent transmissions with dominant intrachannel-four-wave-mixing," *Opt. Expr.*, vol. 20, no. 7, pp. 7777–7791, Mar. 2012.

[11] A. Carena, V. Curri, G. Bosco, P. Poggiolini, and F. Forghieri, "Modeling of the impact of nonlinear propagation effects in uncompensated optical coherent transmission links," *J. Lightw. Technol.*, vol. 30, no. 10, pp. 1524–1539, May 2012.

[12] L. Beygi, E. Agrell, P. Johannisson, M. Karlsson, and H. Wymeersch, "A discrete-time model for uncompensated single-channel fiber-optical links," *IEEE Trans. Commun.*, vol. 60, no. 11, pp. 3440–3450, Nov. 2012.

[13] E. Agrell and M. Karlsson, "Satellite constellations: Towards the nonlinear channel capacity," in *Proc. IEEE Photon. Conf. (IPC)*, Burlingame, CA, Sept. 2012.

[14] C. Häger, A. Graell i Amat, A. Alvarado, and E. Agrell, "Constellation optimization for coherent optical channels distorted by nonlinear phase noise," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Anaheim, CA, Dec. 2012, pp. 2870–2875.

[15] P. P. Mitra and J. B. Stark, "Nonlinear limits to the information capacity of optical fibre communications," *Nature*, vol. 411, pp. 1027–1030, June 2001.

[16] J. B. Stark, P. Mitra, and A. Sengupta, "Information capacity of nonlinear wavelength division multiplexing fiber optic transmission line," *Opt. Fiber Technol.*, vol. 7, no. 4, pp. 275–288, Oct. 2001.

[17] J. Tang, "The Shannon channel capacity of dispersion-free nonlinear optical fiber transmission," *J. Lightw. Technol.*, vol. 19, no. 8, pp. 1104–1109, Aug. 2001.

[18] A. G. Green, P. B. Littlewood, P. P. Mitra, and L. G. L. Wegener, "Schrödinger equation with a spatially and temporally random potential: Effects of cross-phase modulation in optical communication," *Phys. Rev. E*, vol. 66, no. 4, pp. 046 627-1–12, Oct. 2002.

[19] E. E. Narimanov and P. Mitra, "The channel capacity of a fiber optics communication system: perturbation theory," *J. Lightw. Technol.*, vol. 20, no. 3, pp. 530–537, Mar. 2002.

[20] L. G. L. Wegener, M. L. Povinelli, A. G. Green, P. P. Mitra, J. B. Stark, and P. B. Littlewood, "The effect of propagation nonlinearities on the information capacity of WDM optical fiber systems: cross-phase modulation and four-wave mixing," *Physica D: Nonlinear Phenomena*, vol. 189, no. 1–2, pp. 81–99, Feb. 2004.

[21] K.-P. Ho, *Phase-Modulated Optical Communication Systems*. New York, NY: Springer, 2005.

[22] I. B. Djordjevic, B. Vasic, M. Ivkovic, and I. Gabitov, "Achievable information rates for high-speed long-haul optical transmission," *J. Lightw. Technol.*, vol. 23, no. 11, pp. 3755–3763, Nov. 2005.

[23] J. Tang, "A comparison study of the Shannon channel capacity of various nonlinear optical fibers," *J. Lightw. Technol.*, vol. 24, no. 5, pp. 2070–2075, May 2006.

[24] M. H. Taghavi, G. C. Papen, and P. H. Siegel, "On the multiuser capacity of WDM in a nonlinear optical fiber: coherent communication," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5008–5022, Nov. 2006.

[25] R.-J. Essiambre, G. J. Foschini, G. Kramer, and P. J. Winzer, "Capacity limits of information transport in fiber-optic networks," *Phys. Rev. Lett.*, vol. 101, no. 16, pp. 163 901-1–4, Oct. 2008.

[26] T. Freckmann, R.-J. Essiambre, P. J. Winzer, G. J. Foschini, and G. Kramer, "Fiber capacity limits with optimized ring constellations," *IEEE Photon. Technol. Lett.*, vol. 21, no. 20, pp. 1496–1498, Oct. 2009.

[27] I. B. Djordjevic, H. G. Batshon, L. Xu, and T. Wang, "Coded polarization-multiplexed iterative polar modulation (PM-IPM) for beyond 400 Gb/s serial optical transmission," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, Los Angeles, CA, Mar. 2010, p. OMK2.

[28] A. D. Ellis, J. Zhao, and D. Cotter, "Approaching the non-linear Shannon limit," *J. Lightw. Technol.*, vol. 28, no. 4, pp. 423–433, Feb. 2010.

[29] B. P. Smith and F. R. Kschischang, "Future prospects for FEC in fiber-optic communications," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16, no. 5, pp. 1245–1257, Sept./Oct. 2010.

[30] B. Goebel, R.-J. Essiambre, G. Kramer, P. J. Winzer, and N. Hanik, "Calculation of mutual information for partially coherent Gaussian channels with applications to fiber optics," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5720–5736, Sept. 2011.

[31] G. Bosco, P. Poggiolini, A. Carena, V. Curri, and F. Forghieri, "Analytical results on channel capacity in uncompensated optical links with coherent detection," *Opt. Expr.*, vol. 19, no. 26, pp. B440–B449, Dec. 2011.

[32] A. Mecozzi and R.-J. Essiambre, "Nonlinear Shannon limit in pseudolinear coherent systems," *J. Lightw. Technol.*, vol. 30, no. 12, pp. 2011–2024, June 2012.

[33] A. Splett, C. Kurtzke, and K. Petermann, "Ultimate transmission capacity of amplified optical fiber communication systems taking into account fiber nonlinearities," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Montreux, Switzerland, Sept. 1993, pp. 41–44.

[34] P. Poggiolini, A. Carena, V. Curri, G. Bosco, and F. Forghieri, "Analytical modeling of nonlinear propagation in uncompensated optical transmission links," *IEEE Photon. Technol. Lett.*, vol. 23, no. 11, pp. 742–744, June 2011.

[35] P. Poggiolini, "The GN model of non-linear propagation in uncompensated coherent optical systems," *J. Lightw. Technol.*, vol. 24, no. 30, pp. 3875–3879, Dec. 2012.

[36] P. Poggiolini, G. Bosco, A. Carena, V. Curri, Y. Jiang, and F. Forghieri, "The GN-model of fiber non-linear propagation and its applications," *J. Lightw. Technol.*, vol. 32, no. 4, pp. 694–721, Feb. 2014.

[37] A. D. Ellis and J. Zhao, "Channel capacity of non-linear transmission systems," in *Impact of Nonlinearities on Fiber Optic Communications*, S. Kumar, Ed. New York, NY: Springer, 2011, ch. 13, pp. 507–538.

[38] K. S. Turitsyn, S. A. Derevyanko, I. V. Yurkevich, and S. K. Turitsyn, "Information capacity of optical fiber channels with zero average dispersion," *Phys. Rev. Lett.*, vol. 91, no. 20, pp. 203 901-1–4, Nov. 2003.

[39] K.-P. Ho and J. M. Kahn, "Channel capacity of WDM systems using constant-intensity modulation formats," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, Anaheim, CA, Mar. 2002, pp. 731–733.

[40] I. B. Djordjevic, "Codes on graphs, coded modulation and compensation of nonlinear impairments by turbo equalization," in *Impact of Nonlinearities on Fiber Optic Communications*, S. Kumar, Ed. New York, NY: Springer, 2011, ch. 12, pp. 451–505.

[41] J. G. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Inf. Control*, vol. 18, no. 3, pp. 203–219, Apr. 1971.

[42] M. C. Gursoy, H. V. Poor, and S. Verdú, "The noncoherent Rician fading channel—Part I: structure of the capacity-achieving input," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2193–2206, Sept. 2005.

[43] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 1019–1030, Sept. 1990.

[44] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, UK: Cambridge University Press, 2011.
[45] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: Wiley, 1968.
[46] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, no. 3, pp. 611–656, May 1959.
[47] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[48] G. Taricco and M. Elia, "Capacity of fading channel with no side information," *Electron. Lett.*, vol. 33, no. 16, pp. 1368–1370, July 1997.
[49] E. Agrell, A. Alvarado, G. Durisi, and M. Karlsson, "Capacity of a nonlinear optical channel with finite memory," *J. Lightw. Technol.*, vol. 32, no. 16, pp. 2862–2876, Aug. 2014.
[50] E. Agrell, "Nonlinear fiber capacity," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, London, UK, Sept. 2013.
[51] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.
[52] E. Agrell and M. Karlsson, "WDM channel capacity and its dependence on multichannel adaptation models," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, Anaheim, CA, Mar. 2013, p. OTu3B.4.
[53] M. I. Sobhy, E. A. Hosny, M. W. R. Ng, and E. A. Bakkar, "Nonlinear system and subsystem modeling in time domain," *IEEE Trans. Microwave Theory Tech.*, vol. 44, no. 12, pp. 2571–2580, Dec. 1996.
[54] K. Abdelfattah and B. Razavi, "Modeling op amp nonlinearity in switched-capacitor sigma-delta modulators," in *Proc. IEEE Custom Integr. Circ. Conf.*, 2006, pp. 197–200.
[55] H. Elgala, R. Mesleh, and H. Haas, "An LED model for intensity-modulated optical communication systems," *IEEE Photon. Technol. Lett.*, vol. 22, no. 11, pp. 835–837, June 2010.
[56] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1243, June 2002.
[57] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.
[58] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, July 1972.
[59] R. W. Yeung, *Information Theory and Network Coding*. New York, NY: Springer, 2008.
[60] C.-I. Chang and L. D. Davisson, "On calculating the capacity of an infinite-input finite (infinite) -output channel," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 1004–1010, Sept. 1988.
[61] J. Dauwels, "Numerical computation of the capacity of continuous memoryless channels," in *Proc. Symp. Inf. Theory Benelux*, Brussels, Belgium, May 2005, pp. 221–228.
[62] J. Lundgren, M. Rönnqvist, and P. Värbrand, *Optimization*. Lund, Sweden: Studentlitteratur, 2010.

PLACE PHOTO HERE

Erik Agrell (M'99–SM'02) received the Ph.D. degree in information theory in 1997 from Chalmers University of Technology, Sweden.

From 1997 to 1999, he was a Postdoctoral Researcher with the University of California, San Diego and the University of Illinois at Urbana-Champaign. In 1999, he joined the faculty of Chalmers University of Technology, where he is a Professor in Communication Systems since 2009. In 2010, he co-founded the Fiber-Optic Communications Research Center (FORCE) at Chalmers, where he leads the signals and systems research area. He is a Visiting Professor at University College London since 2014. His research interests belong to the fields of information theory, coding theory, and digital communications, and his favorite applications are found in optical communications.

Prof. Agrell served as Publications Editor for the IEEE Transactions on Information Theory from 1999 to 2002 and is an Associate Editor for the IEEE Transactions on Communications since 2012. He is a recipient of the 1990 John Ericsson Medal, the 2009 ITW Best Poster Award, the 2011 GlobeCom Best Paper Award, the 2013 CTW Best Poster Award, and the 2013 Chalmers Supervisor of the Year Award.