Scope and Limitations of the Nonlinear Shannon Limit

Marco Secondini *Member, IEEE*, Enrico Forestieri *Member, IEEE*

Abstract—The concept and significance of the so called nonlinear Shannon limit is reviewed and its relation to the channel capacity is analyzed from an information theory point of view. It is shown that this is a limit (if at all) holding only for conventional detection strategies. Indeed, it should only be considered as a limit to the information rate that can be achieved with a given modulation/detection scheme. By virtue of some simple examples and theoretical results it is also shown that, using the same approximated models commonly adopted for deriving the nonlinear Shannon limit, the information rate can be arbitrarily increased by increasing the input power. To this aim, the validity of some popular approximations to the output distribution is also examined to show that their application outside the scope for which they were devised can lead to pitfalls. To the best of our belief, the existence of a true nonlinear Shannon limit has still not been demonstrated and the problem of determining the channel capacity of a fiber-optic system in the presence of Kerr nonlinearities is still an open issue.

Index Terms—Optical fiber communication, information rates, channel capacity, communication system nonlinearities, wavelength division multiplexing.

I. INTRODUCTION

One of the most important results of the pioneering work of Shannon was demonstrating that a reliable communication over a noisy channel is possible, provided that the transmission rate is less than a characteristic quantity determined by the statistical properties of the channel, which he named *channel capacity* [2]. Conversely, he also showed that this is not possible when this quantity is exceeded. Other than laying the foundations of what is now called *information theory* and establishing such a general result, Shannon also derived a specific closed-form expression for the capacity of an AWGN channel [3], which increases logarithmically with the signal-to-noise ratio (SNR). Hence, for such a channel, the ultimate theoretical limit for the transmission rate is determined only by the available signal power. Shannon's work has been since extended and generalized both to account for a broader class of channels and to obtain closed-form capacity expressions for other specific channels (see [4] and references therein).

In this context, the optical fiber channel is a major challenge due to its peculiar nonlinear behavior. At low signal powers it essentially behaves like an AWGN channel; in fact, modern optical systems can achieve spectral efficiencies close to the channel capacity per unit bandwidth when working in such a linear regime [5]. However, contrarily to the AWGN case,

when the signal power increases, the ensuing nonlinearities impair conventional systems to the point that they cease to work at higher powers. This naturally poses questions about the impact of nonlinearities on the capacity of a fiber-optic channel and whether an ultimate limit to the capacity exists when increasing the signal power. Answering these fundamental questions is of paramount importance for the development of future optical communication networks [6].

A variety of studies and facts hint at the possibility that such a finite limit to the fiber-optic channel capacity actually exists, so much that a name for it was coined—the *nonlinear Shannon limit* (NSL) [5]. Moreover, both analytical expressions and simulation techniques for computing this limit can be found in the literature; many experimental systems can approach it, but no one was able to demonstrate a system able to overstep it; and its possible existence scares telecom operators with the threat of a looming capacity crunch [7]. Yet, as we will show in this paper, nobody has ever proved that such a limit exists.

To the best of our knowledge, a finite capacity limit for a fiber-optic channel was predicted for the first time in [8].[1] This pioneering work anticipated all others by several years and already contained the essential elements of the NSL and of the currently most used channel model: the Gaussian noise (GN) model [9]. Maybe for being ahead of time and not easily accessible, it went seemingly unnoticed for almost two decades. A similar result was then independently rediscovered in [10]–[13], exploiting different nonlinear fiber models (the first three based on Feynman's path integral, the latter on Volterra series). The NSL name was then introduced several years later in [5], with explicit reference to the formula presented in [10]. Even though based on different approximate models, all of these studies share the same hypotheses for computing the NSL, namely, joint Gaussianity of input and output, and absence of channel memory. These hypotheses are explicitly mentioned in [10]–[12] and implied in the Pinsker formalism used in [13] and in the Shannon formula for the AWGN channel used in [8]. The most popular paper on the fiber-optic channel capacity was published later and is based on an entirely different approach: the channel capacity is numerically estimated through an accurate simulation of the channel and partial optimization of the input distribution (uniform rings) and of the output metrics (bivariate Gaussian) [14]. The importance of [14] is twofold; on one hand it provides an accurate and detailed description of optical systems and networks to be analyzed from an information theory point of view; on the other hand it assures that the obtained

M. Secondini and E. Forestieri are with TeCIP Institute, Scuola Superiore Sant'Anna, 56124 Pisa, Italy, and with National Laboratory of Photonic Networks, CNIT, 56124 Pisa, Italy (email: marco.secondini@sssup.it; forestieri@sssup.it).

[1]This work was developed in mathematical detail in the doctoral thesis (in German) available at http://dx.doi.org/10.14279/depositonce-5080.

estimate accounts for all possible relevant fiber effects, also pointing out their relative weight. The dominant effect turns out to be the inter-channel nonlinearity (also considered in [8], [10]–[13]) and this result is in substantial agreement with the previous ones. Other relevant papers on the channel capacity are [15], [16]. The first one, supplemented in [17], deepens and makes explicit the GN model concept used in [8], drawing the consequent conclusions in terms of channel capacity, also extended to account for a dual polarization. The second one develops a detailed analytical model of the nonlinear interference based on a regular perturbation method, producing results in agreement with [14].

All of the aforementioned results are in substantial agreement, seemingly leaving no escape from the NSL. However, in recent years the NSL concept has been criticized and it was suggested that the capacity problem is far from being solved. Some evidences to this effect, whose importance is often neglected, can be identified in the same papers mentioned above. In [10] it is said that the joint Gaussianity hypothesis leads only to a lower bound for the capacity of the considered channel model, but then it is conjectured that the final result mirrors, at least qualitatively, the actual capacity value (so that it presumably reaches a maximum and then decreases). In [14] it is explained that the adopted methodology does not guarantee obtaining the actual channel capacity, as the optimization of the input distribution and the fitting of the output one is performed on a limited set—in particular excluding any time correlation—potentially neglecting some possible information-rate gain. The possible exploitation of the long time coherence of the interchannel interference, aimed at mitigating its effects and improving the lower bound on capacity, is explicitly suggested in [18] and then demonstrated in [19], with a further improvement obtained in [20], [21]. The analysis in [18]–[21] does not consider a network environment, in which time coherence might be significantly reduced.

There are also some theoretical arguments against the NSL: in [22] it was demonstrated that the capacity of a static discrete-time channel, even when defined using an equality constraint on the average power, cannot decrease when the input power increases, as instead predicted by the NSL. For example, in [23] and [24] it is shown that ever-increasing bounds on the capacity (per symbol) can be obtained in the presence of, respectively, nonlinear phase-noise and four-wave mixing in non-dispersive fibers, exactly in the cases where the typical approach used to compute the NSL would provide the usual behavior with a finite maximum. As a matter of fact, all papers that take into account realistic fiber models only obtain lower bounds (or approximations) on the capacity with a finite maximum. By contrast, the tightest known upper bound for a realistic fiber-optic channel model equals the capacity of the AWGN channel [25] and, therefore, increases indefinitely with power. This does not allow to draw conclusions about the impact of fiber nonlinearities on channel capacity, whose value is essentially unknown at higher powers.

The main objective of this work is to discuss the significance of the so-called NSL, rejecting its *hard* interpretation as a true capacity limit and presenting an alternative *soft* interpretation, which leaves the channel capacity problem open. To this aim,

the information theoretical meaning of the NSL and its relation with channel capacity are investigated. As the derivation of the NSL usually entails some specific assumptions in terms of modulation and detection—mainly, joint Gaussianity and uncorrelation of input and output samples—the accuracy and optimality of these assumptions is also investigated. At a theoretical level, the main message of this work is that the NSL (as it has been computed so far) should not be interpreted as the ultimate capacity limit of the optical fiber channel. Rather, it should only be considered as a lower bound, as already claimed in several previous works [6], [22], [23], [26]–[28], whose tightness still can (and should) be improved. However, the scope of this work is not limited to a theoretical analysis. The fallacy of some Gaussian assumptions, the practical need of more accurate models for evaluating the capacity, and the actual possibility to go beyond the NSL are also investigated through some examples. Large improvements on the information rates are obtained in some simplified cases and smaller (but still significant) ones in more realistic cases. Only limitations due to fiber Kerr nonlinearity are investigated. Other physical effects, such as fiber melting, may actually limit channel capacity. Those effects, however, will become relevant only when (and if) a practical way to go significantly beyond the actual limitations imposed by Kerr nonlinearity will be found.

The paper is organized as follows. The NSL is introduced in Section II, discussing its information theoretical meaning, its relation to channel capacity, and the role played by Gaussian assumptions and the GN model in its computation. The accuracy of typical Gaussian assumptions in the context of capacity evaluation is investigated in Section III. Achievable information rates beyond the NSL and capacity bounds for single- and dual-polarization fiber channels are shown in Section IV. Conclusions are finally drawn in Section V.

II. CHANNEL CAPACITY, NONLINEAR SHANNON LIMIT, AND GN MODEL: HOW ARE THEY RELATED?

The capacity $\mathcal{C}$ of a channel is defined as the maximal rate at which information can be reliably (i.e., with arbitrarily low error probability) transmitted through the channel [2]. From an information theoretic point of view, a generic discrete-time channel[2] is characterized by an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$, and the conditional distribution $p(\mathbf{y}_N|\mathbf{x}_N)$, where $\mathbf{x}_N = (x_1, x_2, \ldots, x_N)$, with $x_i \in \mathcal{X}$, and $\mathbf{y}_N = (y_1, y_2, \ldots, y_N)$, with $y_i \in \mathcal{Y}$, are realizations of the input $\mathbf{X}_N = (X_1, X_2, \ldots, X_N)$ and corresponding output $\mathbf{Y}_N = (Y_1, Y_2, \ldots, Y_N)$ random vector processes. For ergodic processes, the average mutual information rate between input and output is defined as

$$I(\mathbf{X}; \mathbf{Y}) = \lim_{N \to \infty} \frac{1}{N} I(\mathbf{X}_N; \mathbf{Y}_N) \qquad (1)$$

where

$$I(\mathbf{X}_N; \mathbf{Y}_N) = E\left\{\log_2 \frac{p(\mathbf{Y}_N|\mathbf{X}_N)}{p(\mathbf{Y}_N)}\right\} \qquad (2)$$

[2]In the following, random variables are denoted by upper-case symbols, their realizations by the corresponding lower-case symbols, and vectors by bold symbols.

is the mutual information between $\mathbf{X}_N$ and $\mathbf{Y}_N$, $E\{\cdot\}$ being the expectation operator and $p(\mathbf{y}_N) = \int p(\mathbf{y}_N|\mathbf{x}_N)p(\mathbf{x}_N)d\mathbf{x}_N$ the output distribution. The capacity of the channel is

$$\mathcal{C} = \lim_{N\to\infty} \frac{1}{N} \sup_{\mathbf{X}_N} I(\mathbf{X}_N; \mathbf{Y}_N) \qquad (3)$$

where the supremum $\sup_{\mathbf{X}_N}$ is taken with respect to all input distributions $p(\mathbf{x}_N)$ satisfying a specific constraint (usually on the average power) [29].

In general, only the capacity of some specific channels can be evaluated analytically, the additive white Gaussian noise (AWGN) channel being perhaps the most notable example [2]. A discrete-time AWGN channel is characterized by the input-output relation[3] $y_k = x_k + n_k$, where n_k are realizations of i.i.d. variables with a zero-mean proper complex Gaussian distribution [30]—often referred to as circularly-symmetric complex Gaussian (CSCG)—with variance P_n. Given a constraint on the average signal power P_s, the capacity of the AWGN channel is achieved when also the input symbols are i.i.d. CSCG and has the well known expression $\mathcal{C} = \log_2(1 + P_s/P_n)$ [3]. As the capacity achieving input distribution and the corresponding output distribution are zero-mean, this capacity can also be expressed in terms of the input variance $\sigma_x^2 = P_s$, output variance $\sigma_y^2 = P_s + P_n$, and covariance $\sigma_{xy} = P_s$ as[4]

$$\mathcal{C} = \log_2(1 + P_s/P_n) = \log_2\left(\frac{\sigma_x^2\sigma_y^2}{\sigma_x^2\sigma_y^2 - |\sigma_{xy}|^2}\right) \qquad (4)$$

In this case, it is apparent that, since the noise power P_n is fixed and independent of the signal, the channel capacity grows unbounded with signal power P_s. This means that the channel capacity is limited only if the available resources (power) are limited (e.g., due to economic or technological constraints).

When turning our attention to the optical fiber channel, the picture is more complicated. In this case, even the accurate computation of (1) is practically unfeasible, since an exact analytical expression for $p(\mathbf{y}_N|\mathbf{x}_N)$ is unknown, while its numerical estimation, which must be performed in a $4N$-dimensional space, is practically limited to very small values of N. On the other hand, convergence of (1) and (3) to their actual limit can be expected when N is of the order of the channel memory, which can be many hundreds of symbols in typical dispersion-unmanaged systems. For the same reasons, also the optimization of $p(\mathbf{x}_N)$ in (3) cannot be performed analytically and, when resorting to numerical approaches, it is limited to small values of N. Nevertheless, as discussed in Section I, it is a widespread opinion that the capacity of the optical fiber channel has a peculiar dependence on signal power and is actually limited by the NSL. This is usually explained by invoking the cubic dependence of nonlinear interference on signal power; when the latter increases, the former increases even faster, such that the channel capacity
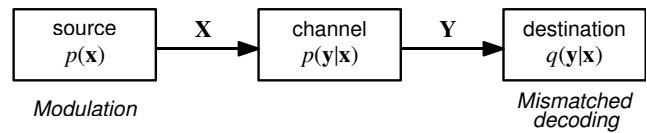

Fig. 1. Discrete-time channel with mismatched decoding.

reaches its maximum value at some optimum launch power and then decreases.

In order to understand what is the real meaning of the NSL and its relation to channel capacity, we consider the more general case, depicted in Fig. 1, of a discrete-time channel with input process $\mathbf{X}$ and output process $\mathbf{Y}$ and of a detector that takes maximum-a-posteriori-probability (MAP) decisions based on a mismatched channel law $q(\mathbf{y}_N|\mathbf{x}_N) \neq p(\mathbf{y}_N|\mathbf{x}_N)$. This is the typical situation in those cases, such as the optical fiber one, in which the actual channel law is unknown. Following [31], we define the *achievable information rate* (AIR) with mismatched decoding metric $q(\mathbf{y}_N|\mathbf{x}_N)$ as[5]

$$I_q(\mathbf{X}; \mathbf{Y}) \triangleq \lim_{N\to\infty} \frac{1}{N} E\left\{\log \frac{q(\mathbf{y}_N|\mathbf{x}_N)}{\int p(\mathbf{x}_N)q(\mathbf{y}_N|\mathbf{x}_N)d\mathbf{x}_N}\right\} \quad (5)$$

With respect to the average mutual information rate (1), (5) is obtained by replacing the actual channel law $p(\mathbf{y}_N|\mathbf{x}_N)$ with an arbitrary mismatched law $q(\mathbf{y}_N|\mathbf{x}_N)$, while the expectation is still taken with respect to the actual joint input-output distribution $p(\mathbf{y}_N|\mathbf{x}_N)p(\mathbf{x}_N)$ induced by the input distribution and the actual channel law. The AIR (5) has some interesting properties [31], [32]: it is achievable on the actual channel with source probability $p(\mathbf{x}_N)$ and a MAP detection matched to $q(\mathbf{y}_N|\mathbf{x}_N)$; it can be easily computed through numerical simulations in almost any practical case, without an explicit knowledge of the channel law; its maximization over any possible detection law (obtained for $q(\mathbf{y}_N|\mathbf{x}_N) = p(\mathbf{y}_N|\mathbf{x}_N)$) provides the average mutual information rate (1); and its maximization over both $p(\mathbf{x}_N)$ and $q(\mathbf{y}_N|\mathbf{x}_N)$ provides the channel capacity $\mathcal{C}$. Thus, $I_q(\mathbf{X}; \mathbf{Y}) \leq I(\mathbf{X}; \mathbf{Y}) \leq \mathcal{C}$. In the context of optical fiber communications, the AIR defined as in (5) has been explicitly used, for instance, in [33]–[37].

Quite interestingly, an analysis of the literature reveals that the NSL, as defined in [5], [10], but also as computed in many other publications [13], [14], [16], is just an instance of (5), computed with specific input distributions and mismatched channel laws (see Section I for a detailed discussion). The convergence of results that can be found in the literature about the NSL has induced the belief that the actual channel capacity is very close to the NSL or, at least, follows the same trend. Such a convergence, however, is due to the fact—not always explicitly stated—that all those results, though obtained through different approaches, are related to the computation of the same quantity (5) with similar assumptions (usually i.i.d. input symbols with CSCG or uniform-ring distribution and detection metrics optimized for a Gaussian memoryless channel). Therefore, the NSL, exactly as the AIR (5), is only a

---

[3] A channel characterized by the input-output relation $y_k = ax_k + b + n$, with a and b complex constants (known at the receiver), will be also referred to as AWGN channel as it can be easily reduced to it by considering the transformed output $y_k' = (y_k - b)/a$.

[4] Written in this form, (4) gives also the true channel capacity for any *correlated* Gaussian channel.

[5] In [31], this quantity is actually referred to as the *auxiliary-channel lower bound* to the information rate.

lower bound to channel capacity, whose tightness is not known a priori.

A special case of (5)—referred to as the Gaussian AIR in the sequel—is obtained when considering i.i.d. CSCG input variables and a detector matched to an AWGN channel (see footnote 3) with the same input-output covariance matrix of the real channel

$$I_G(\mathbf{X}; \mathbf{Y}) = \log_2 \left(\frac{\sigma_x^2 \sigma_y^2}{\sigma_x^2 \sigma_y^2 - |\sigma_{xy}|^2} \right) \qquad (6)$$

This result shows that the Gaussian AIR can be computed for *any* channel by simply estimating a covariance matrix. Remarkably, the expression of the Gaussian AIR equals the (second) expression of the channel capacity for AWGN channels in (4). This means that the same expression can be used (and, in fact, it is commonly used) for any channel, but with different meanings: it gives the true channel capacity for AWGN channels, while it provides only a lower bound for generic channels. For instance, the information theoretical limits computed in [8], [10], [15], [16], [18] are all obtained by using (6), though computing the covariance matrix with different approaches.

A final topic, which is somewhat related to the concept of NSL and which we want to discuss in this section, is the GN model. The importance of the GN model and its relevance to the present discussion lie in its accuracy in predicting the performance of real systems[6] and in its simplicity, which allows the derivation of a closed-form expression of the channel capacity that is in substantial agreement with the NSL [15]. In the GN model, because of the accumulated dispersion, all co-propagating channels of a WDM system become CSCG after a short distance, regardless of their input distribution. Hence, the adopted modulation format is irrelevant to the generation of nonlinear interference, which is always generated by the interaction of CSCG processes. Such nonlinear interference is, in turn, modeled as an additive CSCG process, approximately white over the signal bandwidth. In practice, supposing all WDM signals independent realizations of the same process with same average power P_s, this leads to a discrete-time AWGN model with a noise power increasing with signal power as $P_n = P_{\text{ASE}} + \eta P_s^3$. Here P_{ASE} is the power of the accumulated amplified-spontaneous-emission (ASE) noise and η a parameter depending on system configuration. The *exact* capacity of such a channel is given by (4) and is achieved with i.i.d. CSCG inputs. The cubic dependence of P_n on P_s makes channel capacity achieve its maximum at some optimum power and then decrease [15]. In other words, the GN model leaves no escape from the NSL: changing the modulation format would not reduce nonlinear interference; and no compensation or detection strategy could mitigate the impact of AWGN. This, coupled with the many successes of the GN model, has strengthened the belief that the capacity of the optical fiber channel is actually limited by the NSL. If no others, there is at least one pitfall in this reasoning. Indeed,

[6]An enhanced version of the GN model has been introduced in [38] to correct some inaccuracies of the original model—highlighted for instance in [16],[37],[39]—without significant changes to the basic features of the model.

the GN model (and its extensions) is just an approximation, whose accuracy has been verified only for a limited set of input distributions (usually with i.i.d. samples) and conventional (usually memoryless) detection strategies. This, according to (5), makes it only a lower bound to the channel capacity. On the other hand, capacity evaluation entails optimization of (5) over *any* possible input distribution and *any* possible detection metric, jointly accounting for a possible large number of input and output symbols N. This is out of the scope of validity of the GN model.

When considering *waveform* channels (typical physical channels, such as optical fibers, copper cables, free-space, and so on, usually take waveforms as input and produce waveforms as output), a typical quantity of interest is the capacity per unit bandwidth (spectral efficiency). In this case, the problem is usually divided in two steps: first, a discrete-time version of the channel is found, in which band-limited input and output waveforms are completely represented by a discrete set of input and output symbols and related by a conditional distribution, as in Fig. 1; then, the capacity of such a discrete channel is computed (in bit/symbol) according to (3) and converted to spectral efficiency (in bit/s/Hz) by accounting for the symbol rate and bandwidth limitation. In this way, using the sampling theorem, Shannon extended his work from discrete-time to continuous-time channels [3]. This approach, already fraught with difficulties in the AWGN channel [40], becomes even more complicated in the nonlinear optical fiber channel, in which signal bandwidth is not a conserved quantity during propagation. Also this point will be shortly addressed in the following section.

III. THE GAUSSIAN FALLACY

Gaussian approximations play a fundamental role in the derivation of the nonlinear Shannon limit. In this section, we critically review some common Gaussian assumptions and approximations that are typically made in the context of optical fiber communications and show, through some simple examples, that they are not always fully justified from a theoretical point of view and may often lead to erroneous results. Note that we do not question the accuracy of the approximated models and results that rely on those assumptions and can be found in the literature (e.g., the GN model and the nonlinear Shannon limit). Nor do we deny their relevance in terms of achievable performance in conventional systems. However, we believe that those results have been sometimes attributed excessive generality, as if they could provide the ultimate answer to the capacity problem. In fact, this is not the case and the capacity problem is still open.

A. Effect of Dispersion on a Non-Gaussian Process

Often, it is argued that a large amount of chromatic dispersion can turn any input process (e.g., a QPSK signal) into a Gaussian output process by virtue of the central limit theorem. This is, for instance, the underlying hypothesis in the well known GN model. Indeed, it can be demonstrated that the marginal distribution of *each* output sample converges to a Gaussian distribution as the accumulated dispersion increases,

regardless of the input distribution (at least for typical signals employed in communications) [41]. This, however, does not imply the Gaussianity of the output process, as in this case the *joint* distribution of all its samples has to be Gaussian. In fact, the following proposition holds.

Proposition 1. *Given a stochastic input process and a linear dispersive fiber, the output process is Gaussian if and only if the input process is Gaussian, regardless of the amount of accumulated dispersion.*

Proof: Let $\{x_k\}$ be the samples, taken at the Nyquist rate $1/T$, of a band-limited process at the input of a linear dispersive fiber with transfer function $H(f) = \exp(-j2\pi\beta_2 f^2 L)$, with β_2 the group velocity dispersion (GVD) parameter and L the length of the fiber. Let $\{y_k\}$ be the corresponding samples at the fiber output, taken at the same rate $1/T$. As $|H(f)| = 1$, the output power spectral density and the corresponding Nyquist rate remain unchanged. Letting

$$h_i = \int_{-1/2T}^{1/2T} H(f)e^{j2\pi f i T} df \tag{7}$$

be the samples of the impulse response of the fiber, we can express the output samples as a linear combination of the input ones

$$y_k = \sum_i h_i x_{k-i} \tag{8}$$

Thus, if $\{x_k\}$ are jointly Gaussian, then also the output samples $\{y_k\}$ are jointly Gaussian, as they are obtained from $\{x_k\}$ through the linear transformation (8) [42]. On the other hand, noting that fiber propagation can be reversed by applying the inverse transfer function $H^{-1}(f) = H^*(f)$, we can also express the input samples as a linear combination of the output ones

$$x_k = \sum_i h_{-i}^* y_{k-i} \tag{9}$$

Thus, we can also prove by contradiction that if $\{x_k\}$ are not jointly Gaussian, then $\{y_k\}$ cannot be jointly Gaussian. In fact, if they were, any linear combination of $\{y_k\}$ should be Gaussian [42], including the input samples (9), which, by hypothesis, are not. ∎

An alternative proof can be constructed also by noting that the linear transformation (8) preserves both the power spectral density and the (differential) entropy rate of the propagating process. This implies not only that a non-Gaussian process cannot be turned into a Gaussian one, but also that it does not even get closer (in terms of entropy) to it.

A consequence of Proposition 1 is that the interfering channels in a dispersion-uncompensated WDM system cannot be simply modeled as Gaussian processes, as assumed by the GN-model. In fact, the generated nonlinear interference does depend on modulation format. For instance, a PSK signal generates a weaker interference compared to a QAM signal or a Gaussian process, even in dispersion-uncompensated links [16], [37]. This effect, not accounted for by the simple GN model, is instead explicated by more accurate models such as [16], [37]–[39]. As we will show, this is only part of the story, as the statistical properties of the input signals affect not only the amount of nonlinear interference, but also its statistical

properties, with important implications in terms of channel capacity.

B. Effect of Nonlinearity on a Gaussian Process

Proposition 1 makes clear that dispersion, alone, cannot turn a non-Gaussian process into a Gaussian one. Nevertheless, it is often argued that, if the input process is CSCG (e.g., ASE noise), then the output process will remain CSCG even in the presence of Kerr nonlinearity.[7] However, this is true only if the fiber is linear or if the process is white (non band limited). In fact, the following proposition holds.

Proposition 2. *Given a band-limited CSCG input process and a nonlinear fiber, the output process is not Gaussian.*

Proof: Let x_1 and x_2 be two generic samples of the input process, with joint CSCG distribution [30]

$$p(x_1, x_2) = \frac{1}{\pi^2 A} \exp\left(-\frac{\sigma_2^2|x_1|^2 + \sigma_1^2|x_2|^2 - 2\Re\{\sigma_{12}x_1^* x_2\}}{A}\right) \tag{10}$$

where $\sigma_1^2 = E\{|X_1|^2\}$, $\sigma_2^2 = E\{|X_2|^2\}$, $\sigma_{12} = E\{X_1 X_2^*\}$, and $A = \sigma_1^2\sigma_2^2 - |\sigma_{12}|^2$. Let $y_k = x_k \exp(-j\gamma L|x_k|^2)$, $k = 1, 2$ be the corresponding samples at the output of a nonlinear non-dispersive fiber ($\gamma L > 0$, with γ the nonlinear parameter). By using the fundamental theorem for the transformation of random variables [42] and observing that the Jacobian of the transformation is unitary, their distribution is

$$p(y_1, y_2) = \frac{1}{\pi^2 A} \exp\left(-\frac{1}{A}\left(\sigma_2^2|y_1|^2 + \sigma_1^2|y_2|^2\right.\right. \tag{11}$$
$$\left.\left. - 2\Re\{\sigma_{12}y_1^* y_2 e^{j\gamma L(|y_2|^2 - |y_1|^2)}\}\right)\right)$$

Therefore, the output samples are jointly Gaussian if and only if the corresponding input samples are uncorrelated ($\sigma_{12} = 0$). If the input process is white, then all its samples are uncorrelated and the output samples are jointly Gaussian. On the other hand, if the input process is not white, some of its samples are correlated, meaning that at least some of the output samples are not jointly Gaussian. Finally, using Proposition 1, this conclusion can be extended to a nonlinear dispersive fiber. ∎

As an example, we consider the propagation of a CSCG process with raised-cosine power spectral density with roll-off factor 0.2, power P and (equivalent low-pass) bandwidth $1/2T$ through a nonlinear non-dispersive fiber. Fiber and signal parameters are chosen to have an average nonlinear phase rotation $\gamma LP = 2$. Fig. 2 shows a contour plot of the joint distribution, obtained through Monte Carlo simulations, of the real parts of the output samples $y(t)$ and $y(t + \tau)$, where $\tau = T/4$ in Fig. 2(a) and $\tau = T$ in Fig. 2(b). As expected from Proposition 2, for $\tau = T/4$ the joint distribution is not CSCG as the input samples are correlated, while for $\tau = T$ it is CSCG as the input samples are uncorrelated.

---

[7]Note that CSCG processes usually play an important role in the determination of channel capacity and related bounds as they have the maximum differential entropy for a given power spectral density [25], [30].
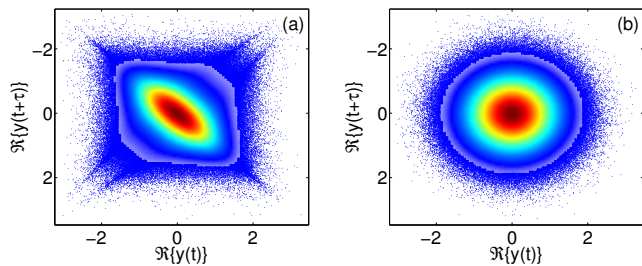
Fig. 2. Joint distribution of the real parts of $y(t)$ and $y(t + \tau)$, for (a) $\tau = T/4$ and (b) $\tau = T$.

C. Spectral Broadening of a Gaussian Process

When considering the capacity (per unit bandwidth) of the optical fiber *waveform* channel, an important issue, which is often neglected, is spectral broadening. In fact, most systems are usually designed to operate in a weakly nonlinear regime, assuming that spectral broadening is negligible. This assumption is indeed reasonable when considering WDM systems with conventional modulation and detection as, in this case, inter-channel nonlinearity limits system performance before spectral broadening and signal-noise interaction become relevant [14], [37]. On the other hand, in single-channel systems, spectral broadening may become a relevant effect and be the main cause of spectral efficiency limitations [43]. For instance, it has been shown that at zero dispersion the capacity (per symbol) of the fiber channel grows unbounded with power, as signal-noise interaction affects only the phase of the received signal [23], [44]. However, because of spectral broadening, the same conclusion cannot be drawn about spectral efficiency [44], which is usually the quantity of interest in a waveform channel. Moreover, even in WDM systems, as we will show later, inter-channel nonlinearity may not be as relevant as it may appear, if properly modeled. For all these reasons, we expect spectral broadening to become an important issue to be investigated.

Spectral broadening can be used as an instructive example to show how Gaussian assumptions, besides being unjustified from a theoretical point of view (Propositions 1 and 2), may also lead to completely erroneous conclusions when used as approximations or working hypotheses.

For the sake of simplicity, we limit the analysis to the case of a nonlinear non-dispersive fiber, as in Proposition 2, considering a stationary process $u(z, t)$ that evolves according to the nonlinear equation $\partial u / \partial z = -j\gamma |u|^2 u$. The corresponding evolution equation for the autocorrelation function $R_u(z, \tau) \triangleq E\{u(z, t)u^*(z, t - \tau)\}$ is

$$\frac{\partial R_u}{\partial z} = E\left\{\frac{\partial u}{\partial z} v^* + u \frac{\partial v^*}{\partial z}\right\} = -j\gamma E\{|u|^2 uv^* - |v|^2 v^* u\} \tag{12}$$

where we have defined $v(z, t) \triangleq u(z, t - \tau)$ and omitted the dependence on time and space. If we assume that the process u (and therefore v) is CSCG and remains such during propagation (therefore neglecting Proposition 2), then the fourth-order moments in (12) can be expressed in terms of the autocorrelation function itself by exploiting Isserlis' theorem

[42]. In this case, using the stationarity and properness of u and v, (12) reduces to

$$\partial R_u / \partial z = 0 \tag{13}$$

which states that the autocorrelation function (and, therefore, the power spectral density) of the process does not change during propagation.

This result is in apparent contrast with common experience and is due to the wrong assumption that the process remains Gaussian during propagation. As demonstrated in Proposition 2, this is true only for white processes. Remarkably, while the Gaussian approximations employed to compute (6) are pessimistic in terms of capacity (per symbol), they might become optimistic in terms of spectral efficiency if used to model spectral broadening.

D. AIR and Capacity of a Discrete-Time Regular-Perturbation (or Volterra-Series) WDM Channel Model

Inter-channel nonlinearity, and in particular XPM, is the dominant effect that, in WDM systems, gives rise to the NSL [5], [10], [13], [14], which is usually derived by neglecting the other (minor) effects. In this section we show that, though the Gaussian AIR (6) is actually limited by XPM, channel capacity is not. To this aim, among many different XPM models available in the literature, we select a simple model based on a first-order regular perturbation approach [16], [39] (and equivalent to a third-order Volterra series) as it is commonly employed for performance evaluation, provides a good accuracy in conventional systems, and allows to construct a simple proof of our claim. Considering only two WDM channels for simplicity, this model leads to the following discrete-time channel

$$y_k = x_k + \sum_{\ell, m, n} c_{\ell, m, n} w_{k-m} w_{k-n}^* x_{k-\ell} + n_k \tag{14}$$

where x_k and y_k are, respectively, the input and output samples of the channel of interest at discrete time k; w_k is the input sample of the interfering channel at time k; n_k are the ASE noise samples (i.i.d. CSCG random variables) with variance $E\{|N_k|^2 = P_n\}$; and $\{c_{\ell, m, n}\}$, with $\ell, m, n = 0, \ldots, N - 1$, are some generic channel coefficients (causality is assumed for simplicity of notation but is not required to derive the following results). Note that (14) does not account for other nonlinear effects such as intra-channel nonlinearity, four wave mixing, signal-noise interaction, and spectral broadening. These effects, however, have a negligible impact on the performance of conventional WDM systems and, as discussed before, are often neglected in the computation of the NSL. Finally, we make the additional assumptions (typical in optical networks and corresponding to the behavioral model (c) described in [24]) that $\{x_k\}$ and $\{w_k\}$ are independently drawn from the same alphabet with equal distribution and same average power $E\{|X_k|^2\} = E\{|W_k|^2\} = P_s$ (fairness and independence among channels) and that $\{w_k\}$ are unknown to the receiver.

In the following, we state and discuss the following result, whose proof is deferred to the Appendix.
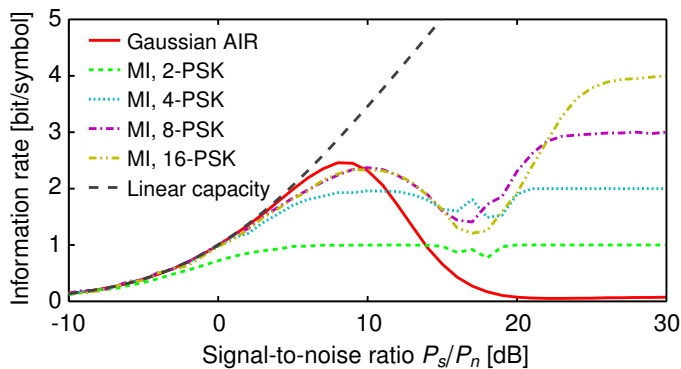
Fig. 3. AIR with Gaussian i.i.d. symbols and Gaussian detection metrics or with M-PSK symbols and optimum (symbol-by-symbol) detection for the discrete-time channel (14). The capacity of the AWGN channel is also shown for comparison.

Theorem 3. *The capacity of the discrete-time WDM channel (14) grows unbounded with power. By contrast, its Gaussian AIR has a finite maximum.*

The theorem is proved for the two-channel WDM model (14) for simplicity of notation, but could be easily extended to many WDM channels. The key factors of the demonstration are the use of a discrete input alphabet (M-PSK in particular, but different alphabets would also work) and the use of the optimum (matched) detector, which treats the interference from the other channel with its correct (discrete) statistics rather than as a Gaussian noise. This allows to prove that, for any alphabet size M, the error probability can be made arbitrarily small as the signal-to-noise ratio $P_s/P_n \to \infty$, meaning that an information rate of $\log_2 M$ bits per symbol can be achieved. The proof is fairly simple for the regular-perturbation model (14), but we conjecture that a similar (perhaps more complicated) proof could be constructed for different models (e.g., based on a logarithmic perturbation [18], [37]).

As an example, we consider a channel like (14), with two WDM signals and memory $N = 3$, whose 27 coefficients $\{c_{\ell mn}\}$ have been independently drawn from a CSCG distribution with variance $P_n/100$. Fig. 3 compares the AIR with Gaussian inputs and Gaussian detection metrics (A.13) with the mutual information (MI) in (A.17) for various M-PSK input alphabets. According to Theorem 3, while the AIR has a finite maximum at some intermediate SNR and decreases to a very small value at large SNR, the MI for the M-PSK alphabet always saturates to its maximum value $\log_2 M$ at high SNR. Similar results (not shown here) have been obtained for different channel realizations $\{c_{\ell mn}\}$, memory N, and alphabet size M. Note that similar results have been obtained in [24] for a different WDM channel model, in which it is assumed that four-wave mixing dominates over all other nonlinear effects and that dispersion is negligible (memoryless channel). This further supports the notion that inter-channel interference, alone, does not bound the capacity of WDM systems to a finite value. Note also that channel model (14) assumes a synchronization (or predictable drift) among WDM channels and neglects other nonlinear effects, such as signal-noise interaction and spectral broadening. Therefore, we argue that the search for a possible capacity limitation induced by nonlinearity (in some sense, a "true" NSL) should be based on the presence of those additional effects and their interplay with inter-channel interference.

IV. BEYOND THE NONLINEAR SHANNON LIMIT

Theorem 3 proves and Fig. 3 exemplifies that the capacity of WDM systems is not limited to a finite value by the combined effect of XPM and ASE noise, as the NSL suggests. However, Theorem 3 and Fig. 3 do not account for other effects such as signal-noise interaction and spectral broadening, which, though usually negligible with respect to XPM, might still be the cause of a capacity limitation. Moreover, though the apparent capacity limitation induced by XPM can be theoretically avoided by using proper detection metrics (see the Appendix), this might be practically unfeasible for large channel memory. This section addresses both those issues by investigating the AIR improvements that can be obtained when considering more realistic amplified multi-span WDM systems and computationally feasible (though expensive) detection strategies.

Two different cases are considered: a single-polarization optical fiber channel, in which signal propagation is governed by the nonlinear Schrödinger equation, and a dual-polarization optical fiber channel, in which propagation is governed by the Manakov equation [45]. Both channels consist of 1000 km of standard single-mode fiber (attenuation $\alpha = 0.2\,\mathrm{dB/km}$, dispersion $\beta_2 = -21.67\,\mathrm{ps}^2/\mathrm{km}$, nonlinear coefficient $\gamma = 1.27\,\mathrm{W}^{-1}\mathrm{km}^{-1}$) and ideal distributed amplification (distributed gain exactly compensating for attenuation and unit spontaneous emission factor). In both cases, the AIR values are obtained by considering five Nyquist-WDM channels (1 symbol/s/Hz)—each with bandwidth $B = 50\,\mathrm{GHz}$ and modulated by i.i.d. CSCG symbols (on one or two polarizations)—and different detection strategies. The propagation of the WDM signal corrupted by the ASE noise injected by optical amplifiers is accurately modeled by using the split-step Fourier method, such that all nonlinear effects are included in the simulation. The expectation in (5) is estimated on the central channel by averaging over many random realizations. Capacity lower and upper bounds are obtained by using the non-decreasing lower bound [22] and the AWGN upper bound [25], respectively.

Fig. 4 reports AIRs and capacity bounds for the single-polarization channel. The four AIR curves correspond to four different detection strategies. The first one (solid line) accounts for linear propagation effects through dispersion compensation and for nonlinear effects only by an increased noise variance, as prescribed by the GN model [15]. The second one (dot-dashed line) mitigates intra-channel nonlinear effects by performing a single-channel digital backpropagation (DBP) on the received samples. The third one (dotted line) mitigates also inter-channel nonlinearity by assuming that it generates a time-varying inter-symbol interference (ISI) that can be partly
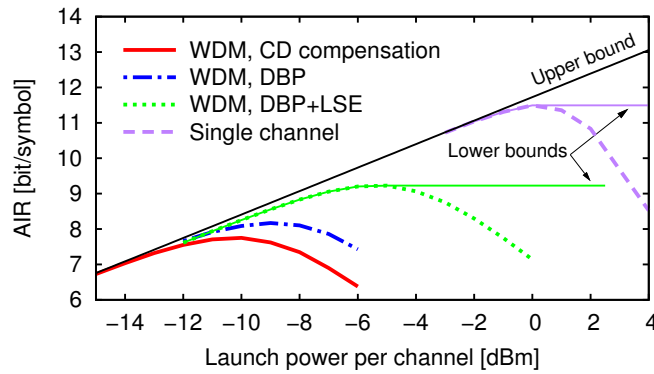
Fig. 4. Achievable information rates and capacity bounds on the single-polarization optical fiber channel. Information rates can be read also as spectral efficiencies in bit/s/Hz.
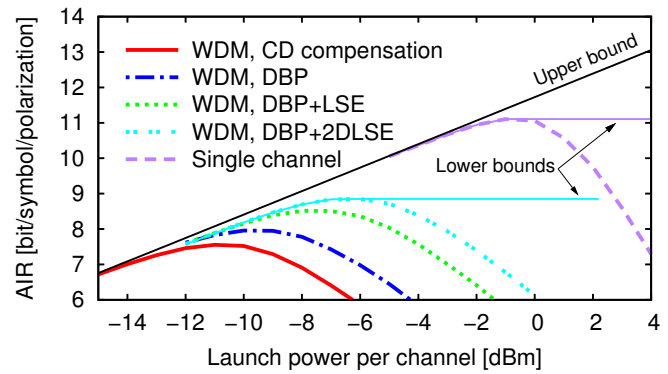

Fig. 5. Achievable information rates and capacity bounds on the dual-polarization optical fiber channel. Information rates can be read also as spectral efficiencies in bit/s/Hz/polarization.

compensated for by using a least-square equalizer (LSE) [20].[8] Finally, the fourth one (dashed line) fully compensates for inter-channel nonlinearity by performing a multi-channel DBP and is mainly limited by signal-noise interaction (which is not further addressed in this work but whose impact can also be mitigated [46], [47]). Note that, while the first three cases do assume a bandwidth constraint for demodulation (only the signal bandwidth B is available), the fourth one does not. These two different assumptions basically define two different discrete-time channels. Finally, two capacity lower bounds (with or without bandwidth constraint) and a capacity upper bound are plotted.

Fig. 5 reports the AIRs and capacity bounds (per polarization) obtained for the dual-polarization channel when using the same approaches considered in the previous case. Moreover, it shows an additional AIR curve obtained by considering a two-dimensional LSE for the mitigation of inter-channel nonlinearity, which is more suitable for the Manakov propagation equation.

Results indicate that the NSL, as obtained with a fixed input distribution (Gaussian) and a mismatched detection strategy (Gaussian), is not, in fact, a true limit. For instance, higher information rates can be achieved with improved detection strategies based on more accurate models for intra- and inter-channel nonlinearity. Unfortunately, the AIR gains obtained by using the LSE in links with distributed amplification are lower [20] as, in this case, the coherence time of the XPM effect reduces significantly [18]. However, the non-decreasing lower bounds of Fig. 4 and Fig. 5—theoretically achievable by satellite constellations [22], i.e., by complementing the input distribution with a satellite point with proper probability and power—clearly show that evaluating channel capacity by using approximate models (though reasonably accurate for conventional modulations) can lead to misleading results, and

strongly suggest that a Gaussian input distribution is highly sub-optimal at high powers. This does not tell us how or if we can go beyond the known lower bounds, but is a stimulus to look for non-conventional modulation and detection, knowing that conventional approaches are not optimum and that the capacity limit lies somewhere between the saturating lower bounds and the ever-increasing upper bounds shown in Fig. 4 and Fig. 5.

V. CONCLUSIONS

The concept of NSL has become very popular in recent years. It has been independently computed by different approaches, with a good agreement between analytical and numerical results. And, in fact, the NSL has proved to be an important practical limit for conventional systems, fostering the belief that, in analogy to the classical Shannon limit, it constitutes an ultimate limit for reliable communication over optical fiber channels. However, theory does not support such a belief, as the NSL is obtained with specific constraints on modulation and detection, whose optimality is not proved. In particular, Gaussian assumptions and approximations about input and output processes are often involved in the computation of the NSL. Those assumptions and approximations are definitely of great practical value and have proved very accurate in modeling conventional systems. Yet, they are not justified at a theoretical level and, when used outside the scope for which they were devised, can lead to pitfalls and inconsistencies. For instance, we have shown that fiber propagation turns a Gaussian process into a non-Gaussian one, rather than the opposite. And that the non-Gaussianity of the propagating process is fundamental to explain spectral broadening in optical fibers.

Though the discrepancy between the definition of channel capacity and the methodology adopted to compute the NSL—as well as between conventional models and reality—can be easily demonstrated, it is still unclear whether it is only of theoretical interest or does have a practical value. In fact, three different degrees of discrepancy can be observed depending on the considered scenario. First, a fundamental discrepancy is found when considering a channel model that,

---

[8]The use of transmitted symbols for the adaptation of the LSE equalizer, though reasonable for the practical computation of performance, does not comply with the requirements of the data processing inequality [29]. This makes the corresponding curves an approximation, rather than a rigorous lower bound. However, some preliminary results indicate that the lower bounds obtained by using the multi-carrier approach described in [21] and decoding metrics optimized for a phase-noise channel practically equal the approximated results presented here.

though simplified to allow an analytical study, accounts for the essential effects in the determination of the NSL—i.e., accumulated ASE noise and XPM. In this case, while Gaussian assumptions about modulation and detection metrics lead to the conventional NSL with a finite maximum, the use of a discrete modulation alphabet and proper detection metrics allows to arbitrarily increase the information rate by increasing signal power. Then, a less fundamental but still significant discrepancy is found when considering a fully realistic numerical model of a link with ideal distributed amplification. In this case, the ever-increasing information rate obtained for the simplified model can neither be analytically proved (due to the complexity of the model) nor numerically verified (due to the long channel memory). Still, a significant AIR gain can be obtained with respect to the conventional NSL simply by using more accurate detection metrics that account for channel memory and the non-additive nature of XPM. Finally, a smaller but still measurable discrepancy is found by using the same approach in links with lumped amplification, in which the AIR gain reduces with the span loss. Nevertheless, the tightest known lower and upper bounds to the capacity of realistic optical fiber channels diverge at high power, leaving the capacity problem still open and the hope for further AIR improvements alive.

Some other issues, concerning the very definition of the optical fiber channel and its capacity, have not been discussed. Those issues—such as the representation of a nonlinear waveform channel through a discrete-time model; the need of consistent definitions of bandwidth and spectral efficiency; and the definition and analysis of multi-user and network scenarios—have received little attention so far, in the belief that they are not relevant compared to the major effects that determine the NSL. We argue that, should the latter be effectively overcome, those issues will become of paramount importance.

APPENDIX A

In this appendix we prove Theorem 3 and give an expression of the mutual information for an M-PSK input alphabet.

Proof of Theorem 3: Defining the normalized variables

$$\tilde{X}_k \triangleq \frac{X_k}{\sqrt{P_s}}, \ \tilde{W}_k \triangleq \frac{W_k}{\sqrt{P_s}}, \ \tilde{N}_k \triangleq \frac{N_k}{\sqrt{P_n}}, \ \tilde{Y}_k \triangleq \frac{Y_k}{\sqrt{P_s}} \quad \text{(A.1)}$$

to make explicit their dependence on the signal power P_s and noise power P_n, we rewrite (14) as

$$\tilde{y}_k = \tilde{x}_k + P_s \xi(\tilde{x}_k, \mathbf{s}_k) + \sqrt{P_n/P_s}\, \tilde{n}_k \quad \text{(A.2)}$$

where $\mathbf{s}_k \triangleq (\tilde{x}_{k-1}, \ldots \tilde{x}_{k-N+1}, \tilde{w}_k, \ldots, \tilde{w}_{k-N+1})$ is a state vector, N is the memory of the channel, and

$$\xi(\tilde{x}_k, \mathbf{s}_k) = \sum_{\ell,m,n} c_{\ell,m,n} \tilde{w}_{k-m} \tilde{w}^*_{k-n} \tilde{x}_{k-\ell} \quad \text{(A.3)}$$

is the normalized interference term. We consider i.i.d. symbols drawn from a normalized M-PSK alphabet

$$\tilde{x}_k, \tilde{w}_k \in \mathcal{X} = \{\chi_i\}_{i=1}^M \quad \text{(A.4)}$$

with $\chi_i = \exp(j2\pi i/M)$, such that the state vector may take $N_s = M^{2N-1}$ different values

$$\mathbf{s}_k \in \mathcal{S} = \{\boldsymbol{\sigma}_i\}_{i=1}^{N_s} \quad \text{(A.5)}$$

with uniform probability. Moreover, it can be verified that for any alphabet size M

$$\min_{i \neq j} |\chi_i - \chi_j| = 2\sin(\pi/M) \quad \text{(A.6)}$$

$$\max_{i,j} |\chi_i - \chi_j| \leq 2, \quad \text{(A.7)}$$

equality in the last expression holding only when M is even.

Given two distinct elements of the input alphabet, χ_i and χ_j, and an arbitrary pair of possible state vectors, $\boldsymbol{\sigma}_a$ and $\boldsymbol{\sigma}_b$, the distance between the corresponding noise-free output symbols can be written as

$$d_{ijab} = |\chi_i - \chi_j + P_s [\xi(\chi_i, \boldsymbol{\sigma}_a) - \xi(\chi_j, \boldsymbol{\sigma}_b)]| \quad \text{(A.8)}$$

If the difference of the interference terms is zero, then

$$d_{ijab} = |\chi_i - \chi_j| \geq 2\sin(\pi/M) \quad \text{(A.9)}$$

otherwise, using the reverse triangle inequality, we can write

$$d_{ijab} \geq P_s |\xi(\chi_i, \boldsymbol{\sigma}_a) - \xi(\chi_j, \boldsymbol{\sigma}_b)| - |\chi_i - \chi_j|$$
$$\geq P_s |\xi(\chi_i, \boldsymbol{\sigma}_a) - \xi(\chi_j, \boldsymbol{\sigma}_b)| - 2 \quad \text{(A.10)}$$

which, given that $|\xi(\chi_i, \boldsymbol{\sigma}_a) - \xi(\chi_j, \boldsymbol{\sigma}_b)| > 0$ by hypothesis, can be made as large as desired by increasing P_s. Therefore, for P_s large enough, we have

$$d_{\min} \triangleq \min_{i \neq j,\, a,b} d_{ijab} \geq \min_{i \neq j} |\chi_i - \chi_j| = 2\sin(\pi/M) \ \text{(A.11)}$$

If we consider an optimum symbol-by-symbol detector that selects the transmitted symbol minimizing the Euclidean distance with respect to the received sample y_k, the error probability at high power P_s can be upper bounded as

$$P_e \leq \Pr\left\{ \sqrt{P_n/P_s}|\tilde{N}_k| > d_{\min}/2 \right\} = \exp\left(-P_s d_{\min}^2/4P_n\right)$$
$$\leq \exp\left(-P_s \sin^2(\pi/M)/P_n\right) \quad \text{(A.12)}$$

which proves that, for any size M of the modulation alphabet, the error probability can be made as small as desired by increasing the SNR P_s/P_n. Therefore, the information rate can be arbitrarily increased by increasing the alphabet size M and the SNR. On the other hand, the Gaussian AIR (6) can be obtained by considering i.i.d. CSCG input samples in (A.2). After some simple but long calculations, we obtain

$$I_G = \log_2\left(\frac{P_n + P_s + c_1 P_s^2 + c_2 P_s^3}{P_n + c_3 P_s^3} \right) \quad \text{(A.13)}$$

with

$$c_1 = 2\sum_m \Re\{c_{0mm}\}, \quad c_2 = \left| \sum_{\ell,m} c_{\ell mm} \right|^2 + \sum_{\ell,m,n} |c_{lmn}|^2,$$
$$c_3 = c_2 - \left| \sum_m c_{0mm} \right|^2 \quad \text{(A.14)}$$

It can be easily verified that $c_2 \geq c_3 \geq 0$, the last equality holding only when all channel coefficients vanish (Gaussian channel). Thus, the AIR (A.13) takes only finite values and

$$\lim_{P_s \to \infty} I_G = \log_2(c_2/c_3) \quad \text{(A.15)}$$

is also finite (but for the Gaussian channel case).

∎

According to (A.2) and omitting subscript k, the conditional distribution of the output sample $\tilde{y}$ given the input symbol $\tilde{x}$ and state vector $\mathbf{s}$ is Gaussian and can be written as

$$p(\tilde{y}|\tilde{x}, \mathbf{s}) = \frac{P_s}{\pi P_n} \exp\left(-\frac{P_s}{P_n}|\tilde{y} - \tilde{x} - P_s\xi(\tilde{x}, \mathbf{s})|^2\right) \quad \text{(A.16)}$$

Thus, for a discrete input alphabet of size M, the (symbol-by-symbol) mutual information between the input and output variables X and Y (or their normalized version) is

$$I(X;Y) = \log_2 M + \frac{1}{MN_s}\sum_{\tilde{x}\in\mathcal{X}}\sum_{\mathbf{s}\in\mathcal{S}}$$
$$\int_{-\infty}^{\infty} p(\tilde{y}|\tilde{x}, \mathbf{s})\log_2\frac{\sum_{\mathbf{s}'\in\mathcal{S}} p(\tilde{y}|\tilde{x}, \mathbf{s}')}{\sum_{x'\in\mathcal{X}}\sum_{\mathbf{s}'\in\mathcal{S}} p(\tilde{y}|x', \mathbf{s}')}d\tilde{y} \quad \text{(A.17)}$$

REFERENCES

[1] M. Secondini and E. Forestieri, "The limits of the nonlinear shannon limit," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2016, paper Th3D.1.

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

[3] ——, "Communication in the presence of noise," *Proc. IRE*, vol. 31, pp. 10–21, Jun. 1949.

[4] S. Verdú, "Fifty years of Shannon theory," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2057–2078, 1998.

[5] A. D. Ellis, Z. Jian, and D. Cotter, "Approaching the non-linear Shannon limit," *J. Lightwave Technol.*, vol. 28, no. 4, pp. 423–433, Feb. 15 2010.

[6] E. Agrell, M. Karlsson *et al.*, "Roadmap of optical communications," *J. Optics*, vol. 18, no. 6, p. 063002, 2016. [Online]. Available: http://stacks.iop.org/2040-8986/18/i=6/a=063002

[7] A. Chraplyvy, "Plenary paper: The coming capacity crunch," in *Proc. Europ. Conf. Opt. Commun. (ECOC)*, Sept 2009.

[8] A. Splett, C. Kurtzke, and K. Petermann, "Ultimate transmission capacity of amplified optical fiber communication systems taking into account fiber nonlinearities," in *Proc. Europ. Conf. Opt. Commun. (ECOC)*, vol. 2, 1993, pp. 41–44.

[9] P. Poggiolini, G. Bosco *et al.*, "The GN-model of fiber non-linear propagation and its applications," *J. Lightwave Technol.*, vol. 32, no. 4, pp. 694–721, Feb. 2014.

[10] P. P. Mitra and J. B. Stark, "Nonlinear limits to the information capacity of optical fiber communications," *Nature*, vol. 411, no. 6841, pp. 1027–1030, 28 Jun. 2001.

[11] A. G. Green, P. B. Littlewood *et al.*, "Schrödinger equation with a spatially and temporally random potential: Effects of cross-phase modulation in optical communication," *Phys. Rev. E*, vol. 66, no. 4, 2002.

[12] L. G. L. Wegener, M. L. Povinelli *et al.*, "The effect of propagation nonlinearities on the information capacity of WDM optical fiber systems: Cross-phase modulation and four-wave mixing," *Physica D*, vol. 189, no. 1, pp. 81–99, 2004.

[13] J. Tang, "The channel capacity of a multispan DWDM system employing dispersive nonlinear optical fibers and an ideal coherent optical receiver," *J. Lightwave Technol.*, vol. 20, no. 7, pp. 1095–1101, 2002.

[14] R.-J. Essiambre, G. Kramer *et al.*, "Capacity limits of optical fiber networks," *J. Lightwave Technol.*, vol. 28, no. 4, pp. 662–701, Feb. 2010.

[15] G. Bosco, P. Poggiolini *et al.*, "Analytical results on channel capacity in uncompensated optical links with coherent detection," *Opt. Exp.*, vol. 19, no. 26, pp. B438–B449, 16 Dec. 2011.

[16] A. Mecozzi and R.-J. Essiambre, "Nonlinear Shannon limit in pseudolinear coherent systems," *J. Lightwave Technol.*, vol. 30, no. 12, pp. 2011–2024, 15 Jun. 2012.

[17] G. Bosco, P. Poggiolini *et al.*, "Analytical results on channel capacity in uncompensated optical links with coherent detection: erratum," *Opt. Exp.*, vol. 20, no. 17, pp. 19 610–19 611, 2012.

[18] M. Secondini and E. Forestieri, "Analytical fiber-optic channel model in the presence of cross-phase modulation," *IEEE Photon. Technol. Lett.*, vol. 24, no. 22, pp. 2016 –2019, nov.15, 2012.

[19] R. Dar, M. Shtaif, and M. Feder, "New bounds on the capacity of the nonlinear fiber-optic channel," *Opt. Lett.*, vol. 39, no. 2, pp. 398–401, Jan 2014.

[20] M. Secondini and E. Forestieri, "On XPM mitigation in WDM fiber-optic systems," *IEEE Photon. Technol. Lett.*, vol. 26, no. 22, pp. 2252–2255, Nov 2014.

[21] D. Marsella, M. Secondini *et al.*, "A simple strategy for mitigating XPM in nonlinear WDM optical systems," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2015, paper Th4D.3.

[22] E. Agrell, "Conditions for a monotonic channel capacity," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 738–748, March 2015.

[23] K. S. Turitsyn, S. A. Derevyanko *et al.*, "Information capacity of optical fiber channels with zero average dispersion," *Phys. Rev. Lett.*, vol. 91, no. 20, Nov. 2003, paper 203901.

[24] E. Agrell and M. Karlsson, "Influence of behavioral models on multiuser channel capacity," *J. Lightwave Technol.*, vol. 33, no. 17, pp. 3507–3515, Sept 2015.

[25] G. Kramer, M. Yousefi, and F. Kschischang, "Upper bound on the capacity of a cascade of nonlinear and noisy channels," in *IEEE Information Theory Workshop (ITW)*, April 2015.

[26] E. Agrell, A. Alvarado *et al.*, "Capacity of a nonlinear optical channel with finite memory," *J. Lightwave Technol.*, vol. 32, no. 16, pp. 2862–2876, 2014.

[27] F. R. Kschischang, "Information-theoretic limits on coherent nonlinear optical-fiber communication," in *Opt. Fiber Commun. Conf. (OFC)*, 2015, paper W3K.1.

[28] E. Agrell, A. Alvarado, and F. R. Kschischang, "Implications of information theory in optical fibre communications," *Phil. Trans. R. Soc. A*, vol. 374, no. 2062, 2016.

[29] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[30] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.

[31] D. M. Arnold, H.-A. Loeliger *et al.*, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3498–3508, Aug. 2006.

[32] N. Merhav, G. Kaplan *et al.*, "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[33] I. Djordjevic, B. Vasic *et al.*, "Achievable information rates for high-speed long-haul optical transmission," *J. Lightwave Technol.*, vol. 23, no. 11, pp. 3755–3763, Nov. 2005.

[34] M. Franceschini, G. Bongiorni *et al.*, "Fundamental limits of electronic signal processing in direct-detection optical communications," *J. Lightwave Technol.*, vol. 25, no. 7, pp. 1742–1753, July 2007.

[35] G. Colavolpe, T. Foggi *et al.*, "Faster-than-Nyquist and beyond: how to improve spectral efficiency by accepting interference," *Opt. Exp.*, vol. 19, no. 27, pp. 26 600–26 609, Dec 2011.

[36] P. Serena, A. Bononi, and G. Colavolpe, "On the nonlinear capacity with memory of PS-QPSK and PDM-QPSK in WDM non-dispersion managed links," in *Proc. Europ. Conf. Opt. Commun. (ECOC)*, 2012, paper We.2.C.1.

[37] M. Secondini, E. Forestieri, and G. Prati, "Achievable information rate in nonlinear WDM fiber-optic systems with arbitrary modulation formats and dispersion maps," *J. Lightwave Technol.*, vol. 31, no. 23, pp. 3839–3852, Dec 2013.

[38] A. Carena, G. Bosco *et al.*, "EGN model of non-linear fiber propagation," *Opt. Exp.*, vol. 22, no. 13, pp. 16 335–16 362, Jun 2014.

[39] R. Dar, M. Feder *et al.*, "Properties of nonlinear noise in long, dispersion-uncompensated fiber links," *Opt. Exp.*, vol. 21, no. 22, pp. 25 685–25 699, 2013.

[40] D. Slepian, "On bandwidth," *Proc. IEEE*, vol. 64, pp. 292–300, 1976.

[41] A. Carena, V. Curri *et al.*, "Modeling of the impact of nonlinear propagation effects in uncompensated optical coherent transmission links," *J. Lightwave Technol.*, vol. 30, no. 10, pp. 1524–1539, May 2012.

[42] A. Papoulis, *Probability, random variables, and stochastic processes*. McGraw Hill, 1991.

[43] E. Forestieri and M. Secondini, "The nonlinear fiber-optic channel: Modeling and achievable information rate," *Proc. of Progress In Electromagnetics Research Symposium (PIERS) 2015*, pp. 1276–1283, 2015.

[44] M. I. Yousefi and F. R. Kschischang, "On the per-sample capacity of nondispersive optical fibers," *IEEE Trans. Inform. Theory*, vol. 11, no. 57, pp. 7522–7541, 2011.

[45] G. P. Agrawal, *Nonlinear Fiber Optics*, 3rd ed.   San Diego, CA: Academic, 2001.

[46] D. Marsella, M. Secondini, and E. Forestieri, "Maximum likelihood sequence detection for mitigating nonlinear effects," *J. Lightwave Technol.*, vol. 32, no. 5, pp. 908–916, Mar. 2014.

[47] N. Irukulapati, D. Marsella *et al.*, "Stochastic digital backpropagation with residual memory compensation," *J. Lightwave Technol.*, to appear.