

Method for computing the optimal signal distribution and channel capacity

E. G. Shapiro,^{1,2} D. A. Shapiro,^{1,2,*} and S. K. Turitsyn^{2,3}

¹*Institute of Automation and Electrometry, Novosibirsk 630090, Russia*

²*Novosibirsk State University, Novosibirsk 630090, Russia*

³*Aston Institute of Photonic Technologies, Aston University, B4 7ET Birmingham, UK*

[*shapiro@iae.nsk.su](mailto:shapiro@iae.nsk.su)

Abstract: An iterative method for computing the channel capacity of both discrete and continuous input, continuous output channels is proposed. The efficiency of new method is demonstrated in comparison with the classical Blahut – Arimoto algorithm for several known channels. Moreover, we also present a hybrid method combining advantages of both the Blahut – Arimoto algorithm and our iterative approach. The new method is especially efficient for the channels with a priori unknown discrete input alphabet.

© 2015 Optical Society of America

OCIS codes: (060.2330) Fiber optics communications; (060.4510) Optical communications.

References and links

1. C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal* **27**, 379–423 (1948).
2. S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory* **18**, 14–20 (1972).
3. R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory* **18**, 460–473 (1972).
4. C. Chang and L. D. Davisson, "On calculating the capacity of an infinite-input finite (infinite)-output channel," *IEEE Trans. Inf. Theory* **34**, 1004–1010 (1988).
5. J. Huang and S. P. Meyn, "Characterization and computation of optimal distributions for channel coding," *IEEE Trans. Inf. Theory* **51**, 2336–2351 (2005).
6. P. O. Vontobel, A. Kavcic, D. M. Arnold, and H. A. Loeliger, "A generalization of the Blahut-Arimoto algorithm to finite-state channels," *IEEE Trans. Inf. Theory* **54**, 1887–1918 (2008).
7. C.-I. Chang and S. C. Fan, "The capacity of discrete-time memoryless Rayleigh-fading channels," *Information and Computation* **86**, 1–13 (1990).
8. G. Matz and P. Duhamel, "Information geometric formulation and interpretation of accelerated Blahut-Arimoto-type algorithms," in "Information Theory Workshop, 2004. IEEE," (2004), pp. 66–70.
9. A. Mecozzi and M. Shtaif, "On the capacity of intensity modulated systems using optical amplifiers," *IEEE Photon. Technol. Lett.* **13**, 1029–1031 (2001).
10. K.-P. Ho, "Exact evaluation of the capacity for intensity-modulated direct-detection channels with optical amplifier noises," *IEEE Photon. Technol. Lett.* **17**, 858–860 (2005).
11. K. S. Turitsyn and S. K. Turitsyn, "Nonlinear communication channels with capacity above the linear Shannon limit," *Opt. Lett.* **37**, 3600–3602 (2012).
12. M. A. Sorokina and S. K. Turitsyn, "Regeneration limit of classical Shannon capacity," *Nature Communications* **5**, 3861 (2014).
13. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran* (Cambridge University Press, Cambridge — New York, 1992).
14. G. Golub and C. F. Van Loan, *Matrix Computations, 3rd edition* (Johns Hopkins University Press, Baltimore, MD, 1996).
15. C. Shannon, "Probability of error for optimal codes in a Gaussian channel," *The Bell System Technical Journal* **38**, 611–656 (1959).
16. F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, *NIST Handbook of Mathematical Functions* (Cambridge University Press, New York, 2010).

17. I. C. Abou-Faycal, M. D. Trott, and S. Shamai (Shitz), "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Inf. Theory* **47**, 1290–1301 (2001).
18. P. P. Mitra and J. B. Stark, "Recent advances in coherent optical communication," *Nature* **411**, 1027–1030 (2001).
19. R.-J. Essiambre, G. J. Foschini, G. Kramer, and P. J. Winzer, "Capacity limits of information transport in fiber-optic networks," *Phys. Rev. Lett.* **101**, 163901 (2008).
20. R. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," *J. Lightwave Technol.* **28**, 662–701 (2010).
21. M. A. Sorokina, S. Sygletos, and S. K. Turitsyn, "Optimization of cascaded regenerative links based on phase sensitive amplifiers," *Opt. Lett.* **38**, 4378–4381 (2013).
22. A. D. Ellis, J. Zhao, and D. Cotter, "Approaching the non-linear Shannon limit," *J. Lightwave Technol.* **28**, 423–433 (2010).
23. D. J. Richardson, "Filling the light pipe," *Science* **330**, 327–328 (2010).
24. A. S. Holevo and V. Giovannetti, "Quantum channels and their entropic characteristics," *Rep. Prog. Phys.* **75**, 046001 (2012).
25. V. Giovannetti, S. Lloyd, L. Maccone, and J. H. Shapiro, "Electromagnetic channel capacity for practical purposes," *Nat. Photon.* **7**, 834–838 (2013).
26. 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* **50**, 2257–2270 (2004).
27. C. Hager, A. Graell i Amat, A. Alvarado, and E. Agrell, "Constellation optimization for coherent optical channels distorted by nonlinear phase noise," in "Global Communications Conference (GLOBECOM), 2012 IEEE," (2012), pp. 2870–2875.
28. E. Agrell, A. Alvarado, G. Durisi, and M. Karlsson, "Capacity of a nonlinear optical channel with finite memory," *J. Lightwave Technol.* **16**, 2862–2876 (2014).

1. Introduction

Calculation of Shannon capacity of communication channels is a challenging mathematical problem that is important for design of practical systems. In optical communications complexity of this problem is increased due to nonlinear properties of fibre channels and memory introduced by fibre dispersion. Numerical methods developed in the past for computation of channel capacity face various new challenges and have to be modified and further improved. In this work we propose two new methods for numerical calculation of the Shannon capacity. We focus here on the explaining how to constructively use the new approaches rather than on fine mathematical details and proofs.

We consider a problem of numerical computation of the maximum of the mutual information functional introduced by Shannon [1]:

$$C = \max_{P(x)} \left\{ \iint dx dy P(y|x) P(x) \log_2 \left[\frac{P(y|x)}{\int dx' P(y|x') P(x')} \right] \right\}, \quad (1)$$

here $P(y|x)$ is the given conditional probability of detecting output signal y for an input signal x ; both variables x and y may be discrete or continuous; here we consider only memoryless channels for simplicity of presentation and comparison with known methods, however, we would like to stress that our method can be also extended to channels with finite memory; $P(x)$ is the probability density function of the input signal x . Consider $P(y|x)$ to be a twice differentiable function of x . When variable x' is discrete $\int dx' \dots$ is changed by the summation over the discrete support. The maximal value of functional (1) should be found under the condition of a bounded from above average power S :

$$\int x^2 P(x) dx \leq S. \quad (2)$$

The Shannon functional optimization problem has analytical solutions only for a very limited class of communication channels and, as a rule, requires numerical calculations.

The most known and widely used numerical method in this field is the Blahut – Arimoto algorithm (BAA) [2, 3] that was originally developed for finite discrete alphabets of both x and

y. The BAA has a classical status and plays a fundamental role in numerical computations of capacity for various communication channels (see e.g. [4–6] and references therein). In this work we consider a general case of continuous-input, continuous output channel. In general, distribution $P(x)$ that maximizes the mutual information (the functional presented in (1)) for such channels can be continuous or discrete or mixed (including both discrete and continuous parts). The discrete distribution formally can be presented using the continuous description as a set of Dirac delta functions with different coefficients. For channels with continuous input, continuous output the original BAA was modified [7,8]. Our aim here is to introduce a new numerical method of channel capacity evaluation and demonstrate its advantage compared to the BAA for several channels considered below. The new point used in our method in application to continuous input, continuous output channels is an adaptive iterative tuning of the discrete approximations of continuous signal x .

The proposed method is a technical tool for optimizing the mutual information with respect to variable x . As distinct from BAA we do not consider the input alphabet as a fixed set of values, but something to be determined through numerical optimization. Generally, a wide class of distributions can provide the value of functional (1) close to its limit [5]. The proposed method makes it possible to narrow down this class.

To illustrate the proposed relaxation method and to make a comparison with the BAA we consider three channels: (i) the classical Gaussian linear channel, just to set a reference point, (ii) the Rician channel that is known in optical communications as intensity modulation direct detection channel [9, 10], and (iii) recently introduced (in the context of optical communications) nonlinear continuous input, continuous output regenerative channel. The latter includes cascaded interleaving of additive noise with nonlinear signal transformation $x \rightarrow x + \alpha \sin \beta x$, where α, β are given parameters. Such nonlinear transformation (that effectively restricts the signal corruption by noise) can lead to channel capacity exceeding the capacity for the corresponding linear additive white Gaussian noise channel [11, 12]. However, this is possible only for an optimal input signal alphabet making this an interesting and challenging optimization problem that we use for testing the new method and its comparison to BAA. The choice of channels for comparison is not critical and we demonstrate that both the classical BAA and our method lead to exactly the same results in all three cases, however, with different computational complexity.

The paper is organized as follows. Section 2 describes the proposed method of computing the maximum of mutual information functional. Section 3 presents the numerical results for three channels chosen for comparison. In Section 3.1 the relaxation and the BAA methods are briefly compared with the analytical formula for the linear Gaussian channel. In Section 3.2 we compare the new method with the BAA for the Rice distribution channel whose capacity was first numerically found in [9, 10]. Section 3.3 is focused on the nonlinear noise squeezing channel comparative study by both methods. In Section 4 the difference between the two methods is explained and the hybrid approach, combining BAA and the relaxation method is tested, showing excellent performance. Finally, Section 5 presents our conclusions. We take out the more rigorous presentment of the numerical procedure into the Appendix.

2. Relaxation method

Definite integrals over a finite segment $[a, b]$ can be approximated by the Riemann sums

$$\int_a^b f(x)g(x) dx \approx \sum_{i=1}^{L-1} f(x_i)g(x_i)\Delta_i, \quad (3)$$

$$x_1 = a, \quad x_L = b, \quad \Delta_i = x_{i+1} - x_i,$$

where $f(x), g(x)$ are the continuous functions, L is the number of intervals. Arbitrary accuracy of the numerical approximation is controlled by the smallness of Δ_i . Denote $p_i = g(x_i)\Delta_i$, then

$$\sum_{i=1}^{L-1} f(x_i)g(x_i)\Delta_i = \int_a^b f(x) \sum_{i=1}^{L-1} p_i \delta(x-x_i) dx,$$

where $\delta(x)$ is the Dirac delta function. The numbers p_i can be chosen rational, i.e. there exist integer positive numbers $n_1, n_2, \dots, n_{L-1}, n_1 + \dots + n_{L-1} = M$ such that

$$\left| p_1 - \frac{n_1}{M} \right| + \left| p_2 - \frac{n_2}{M} \right| + \dots + \left| p_{L-1} - \frac{n_{L-1}}{M} \right| \leq \varepsilon$$

for arbitrary ε while M being sufficiently large.

Let the values of x_i are equidistant: $\Delta_1 = \dots = \Delta_{L-1} = \Delta$. Then one can present the function $g(x)$ as follows:

$$g(x) \approx \sum_{i=1}^{L-1} \frac{n_i}{M} \delta(x-x_i) \equiv \sum_{j=1}^M \frac{1}{M} \delta(x-x'_j).$$

We replace the set of intermediate points x_1, \dots, x_{L-1} by new primed set (x'_1, \dots, x'_M) , where each point x_i enters the new set several times with multiplicity n_i :

$$x'_1 = \dots = x'_{n_1} = x_1, x'_{n_1+1} = \dots = x'_{n_1+n_2} = x_2, \dots, x'_{n_1+\dots+n_{L-1}+1} = \dots = x'_M = x_{L-1}.$$

Then the factor $1/M$ in the sum becomes independent of index j . Underline that the approximation of function $g(x)$ as a comb of delta-functions is valid only under integral (3). After this approximation the integral turns into the sum.

Returning to integral (1) we denote the probability distribution $g(x) = P(x)$ and replace by the comb. In this way the problem reduces to the search of extremum for mutual information functional

$$F = \sum_{k=1}^M p_k \int dy P(y|x_k) \log_2 \frac{P(y|x_k)}{\sum_{j=1}^M p_j P(y|x_j)}. \quad (4)$$

The normalization of the probability density function is included automatically when $p_k = 1/M$.

The bounded power condition

$$\int x^2 P(x) dx = S \quad (5)$$

is taken into account by the Lagrange multipliers method. Hereafter we suppose the average power S to be a fixed constant. Consider an auxiliary functional

$$\Phi = F + \lambda \left(S - \sum_{k=1}^M p_k x_k^2 \right), \quad (6)$$

where λ is the Lagrange multiplier.

Instead of a functional (6) it is possible to maximize any functional $\tilde{\Phi} = \Phi + \gamma S$, where γ is an auxiliary constant. Since S is constant problems of extremum for both functional Φ and $\tilde{\Phi}$ are equivalent. At the same time with $\tilde{\Phi}$ we acquire an additional parameter γ and exploit it to provide the diagonal dominance in the matrix, and then simplify the numerical procedures. The point of maximum of functional Φ is determined from equations $\partial \tilde{\Phi} / \partial x_i = 0$. As follows from (6), the set reduces to

$$\frac{\partial \tilde{\Phi}}{\partial x_k} = \frac{\partial}{\partial x_k} (F + \gamma S) - 2\lambda p_k x_k = 0, \quad k = 1, \dots, M. \quad (7)$$

It is the set of transcendental algebraic equations. The role of γ -term is shifting of all the eigenvalues. It is especially important if some eigenvalues are close to zero and then the matrix inversion procedure is ill-conditioned.

Let us describe the relaxation procedure. Denote $F_k = \partial F / \partial x_k$, $\Phi_k = F_k - 2\lambda p_k x_k$ and introduce the additional parameter t (“time”). The variables will be considered to be functions of parameter t : $x_k = x_k(t)$, $k = 1 \dots, M$; $\lambda = \lambda(t)$. In place of solving the algebraic set of transcendental Eqs. (7) we solve a set of linear differential equations

$$\frac{d\Phi_k}{dt} = -\Phi_k, \quad k = 1, \dots, M. \quad (8)$$

When we multiply (8) by Φ_k and sum up all the equations over index k , we see that the norm $\|\Phi_k\|^2 \rightarrow 0, t \rightarrow \infty$. Then its solution converges to the solution of the set (7) with increasing “time” t . The relaxation is exponential $\|\Phi_k\|^2 \sim e^{-2t}$ and the solution becomes negligible at $t \gtrsim 1$. Set (8) can be solved faster by iterations. In relaxation method [13] we replace ordinary differential Eqs. (9) by approximate finite-difference equations. We search for the steady-state solution, the details of relaxation process itself are not interesting for our purpose. Then calculating the stationary solution we may take a long step τ .

Denote the “velocities” $dx_k/dt = w_k$. From (7) we get

$$\sum_{l=1}^M \frac{\partial \tilde{\Phi}_k}{\partial x_l} w_l = -\Phi_k, \quad k = 1, \dots, M. \quad (9)$$

The solution of differential Eqs. (8) tends to zero, then the solution of Eq. (9) tends to the solution of transcendental Eqs. (7).

Further stages of iteration procedure are:

- 0° Prestore the initial values x_1, \dots, x_M .
- 1° Find w_k from linear set (9).
- 2° Suppose $x_k(t + \tau) = x_k(t) + \tau w_k$, where τ is the step.
- 3° Repeat 1° and 2° until the solution gets the target accuracy.

Solving set (9) we use Gauss–Seidel method of successive iterations. We vary parameter γ in order to provide the diagonal dominance in the corresponding matrix. The matrix with diagonal dominance is well-conditioned then the iteration procedure is stable and the convergence is faster [14]. The number of required operations is $O(M^2)$ per an iteration.

3. Numerical results

For the purpose of numerical modeling the continuous functions are usually replaced by stepwise functions defined at a finite number of intervals. Thus for approximation of continuous function in a finite domain the set of its values is sufficient in finite number of points. Calculations for continuous channels then can be executed by the same methods as for discrete channels.

3.1. The Gaussian channel

To test the new method and compare it to the BAA we first apply it to the well-known linear Gaussian channel. For linear additive white Gaussian noise channel with noise dispersion N the conditional probability is given by:

$$P(y|x) = \frac{1}{\sqrt{\pi N}} \exp \left[-\frac{(x-y)^2}{N} \right]. \quad (10)$$

The optimal distribution for this case is continuous and also Gaussian [15]. For a given value of S it is

$$P(x) = \frac{1}{\sqrt{2\pi S}} \exp\left[-\frac{x^2}{2S}\right]. \quad (11)$$

Both the computed capacity and the probability distribution $P(x)$ are in a good agreement with the well-known analytical results. For the relaxation method, the calculations require 5 – 10 iterations for each S . The noise dispersion is $N = 0.36$, the step is $\tau = 25$, the number of steps for x is $M = 800$. The difference between numerical result and analytical formula is observed in 4th digit only. We would like to point out that for this example the BAA requires only 1 – 2 iterations then both the methods give the result for the probability distribution almost immediately. We include the example of the linear Gaussian channel for a fair comparison, just to show that the BAA is optimal for this particular classical channel compared to the relaxation method. However, as we will demonstrate below for other types of channels with discrete optimal distribution of x the new method is more efficient compared to the BAA.

3.2. Rice channel

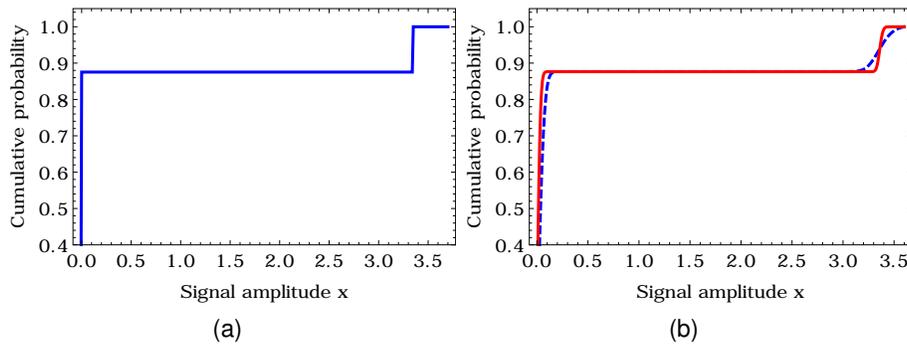


Fig. 1. An integral of the optimal probability distribution function for IMDD channel: $V(x) = \int_0^x P(x') dx'$, computed: (a) by relaxation method after 500 iterations, (b) by BAA after 1000 iterations (dashed curve) and 20000 iterations (solid).

The Rice channel is defined by the transformation $y = |x + \eta|$ between the transmitted x and received signal y , where η is the complex Gaussian noise. For instance, under assumption of negligible fibre nonlinearity, this formula describes the intensity modulation direct detection (IMDD) channel in optical communications when information coding (of the complex envelope field) is applied only over the signal power (intensity) without using optical phase. In this case η is the complex Gaussian term generated by the amplified spontaneous emission. The statistical properties of the Rice channel are given by the following distribution [9, 10]:

$$P(y|x) = \frac{y}{\sigma^2} \exp\left[-\frac{x^2 + y^2}{2\sigma^2}\right] I_0\left(\frac{xy}{\sigma^2}\right), \quad (12)$$

here σ^2 is the dispersion, I_0 is the modified Bessel function of zero order [16]. It is known [10, 17] that the optimal input signal distribution for the continuous Rice channel has at least a single discrete point at $x = 0$. For relatively small S the optimal alphabet is the binary format [10]. Starting from a binary signal format at small S , the number of discrete points is increased with S . We compare here the BAA and new relaxation method for $\sigma = 1$ considering only discrete variant of the Rice channel. Figure 1(a) shows the optimal probability distribution

function calculated by the relaxation method at $M = 401, S = 1.4$. The probability density in this case is $P(x) = p_1 \delta(x - x_1) + p_2 \delta(x - x_2), x_1 = 0, p_1 = 0.875, x_2 = 3.35, p_2 = 0.125$. Figure 1(a) is plotted after 500 iterations. Figure 1(b) shows the same probability distribution function calculated by the BAA after 20000 iterations (solid curve) and after 1000 iterations (dashed). From these curves it is seen that the new method requires less number of iterations especially for piece-wise continuous functions. After relaxation method the resultant distribution function consists immediately of “stairs”. The iterations are necessary only to refine their heights and positions. The BAA blurs the fronts and the stair shape becomes sharper only with increasing the number of iterations. This example shows that the relaxation method has certain advantages over the BAA when applied to non-Gaussian channels.

3.3. Nonlinear transformation channel

Next we consider optimization of the mutual information functional over input signal distributions for a recently introduced nonlinear channel [11, 12]. The following analysis, in general, can be applied to the n -dimensional vectors y and x representing output and input signals, correspondingly. However, without loss of generality, we examine here one-dimensional channel, and will use in what follows the notations x and y for one-dimensional (real) continuous input, continuous output, respectively. The transmission system presents a channel with R cascaded identical nonlinear elements interleaved with the standard linear additive white Gaussian noise spans (in-between nonlinear elements). In the absence of the nonlinear elements, this channel corresponds to the classical linear white Gaussian noise channel with the corresponding accumulated dispersion. As explained in detail in [11, 12], the signal evolution in such system can be described in terms of the stochastic map — a discrete version of the Langevin equation for stochastic processes:

$$y_k = T(y_{k-1}) + \eta_k, \quad k = 1, \dots, R, \quad y_0 = x + \eta_0, y = y_R.$$

Here k is the discrete index and T is the transfer function of the noise squeezing nonlinear element. The term η_k models the Gaussian noise with zero mean and the dispersion given by N_k added at k -th node

Here we consider a simple example with single nonlinear element placed in-between two linear Gaussian channels. The first output $y_0 = x + n_0$ is the input signal x plus the first additive Gaussian noise n_0 . The average value of the noise is zero $\bar{n}_0 = 0$, its dispersion is $\overline{n_0^2} = N_0$. Here the bar denotes the averaging. Next, the y_0 acts as an input for the nonlinear element. The nonlinear element transforms the signal according to the formula [12]:

$$y_1 = y_0 + \alpha \sin(\beta y_0). \quad (13)$$

Then the Gaussian noise is added to the signal $y_1 = y_{01} + n_1, \bar{n}_1 = 0, \overline{n_1^2} = N_1$. The conditional probability distribution for such a channel is

$$P(y|x) = \int \exp \left[-\frac{(y - z - \alpha \sin \beta z)^2}{N_1} - \frac{(x - z)^2}{N_0} \right] \frac{dz}{2\pi\sqrt{N_0 N_1}}. \quad (14)$$

At $\alpha = 0$ the channel is Gaussian. The average value of the noise is zero, its overall dispersion is $N = N_0 + N_1$. The optimal distribution in this case is given by Eq. (11). Below we assume $N_0 = N_1 = 0.36, \beta = 3$.

For linear Gaussian channel the optimal distribution of input signal is a continuous function. There exists also a number of discrete realizations of input alphabet providing the Shannon functional being arbitrary close to the maximum. Therefore, there is no problem of choosing the

optimal alphabet. When the optimal input distribution $P(x)$ is discrete, the situation is changed dramatically. The choice of input alphabet in this case is more difficult.

The numerical simulation shows that even for relatively small parameter α the capacity-achieving input distribution for this channel is discrete with corresponding discrete optimal input alphabet. In the new method we do need to exploit the discrete input distribution. We would like also to emphasize that the (optimal) input alphabet is not known a priori. In this sense, we cannot even a priori claim the considered channel to be discrete or continuous. For comparison purpose, the numerical computations are realized by the BAA and by the relaxation method. The values x_1, \dots, x_M obtained at $\alpha = 0$ are used as initial first-step estimates in the calculations.

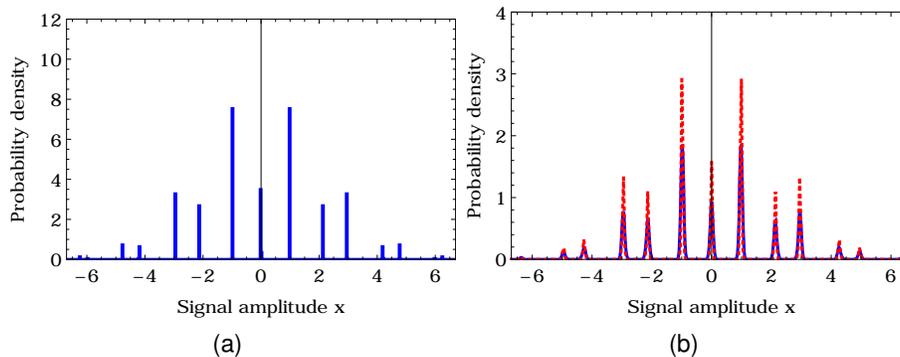


Fig. 2. Optimal probability density $P(x)$ for the nonlinear regenerative channel: (a) calculated by the relaxation method, (b) by BAA after 50000 iterations (dashed line) and 150000 iterations (solid). Parameters of the modeling are $S = 5$, $\alpha = 0.05$.

Below we analyze the evolution of discrete distribution with increasing nonlinearity α . Figure 2(a) shows the optimal function $P(x)$ calculated by the relaxation method after 1000 iterations at $\alpha = 0.05$. The step is $\tau = 15$, the stabilization parameter is $\gamma = 10$. The plot is a histogram with x -step $\Delta = 0.0294$ and the number of points $M = 800$. Figure 2(b) shows the same distribution density calculated by the BAA. The required number of operations for BAA to achieve comparable result is 50 – 150 times larger.

Figure 3(a) shows the optimal function $P(x)$ for higher parameter $\alpha = 0.1$ calculated by the relaxation method after 1000 iterations. Figure 3(b) shows the same distribution density calculated by the BAA. It is seen that using BAA even 20000 iterations are insufficient to include values $x = 0, x = \pm 2.2$ into the desired alphabet. The relaxation method finds out these values using much less number of iterations.

Further increase in nonlinearity α improves the convergence of both the methods since the optimal alphabet consists of less number of symbols at higher α . The plots for $\alpha = 0.15$ calculated by the new method after 500 iterations are shown in Fig. 4(a). Figure 4(b) shows the density calculated by BAA after 5000 and 20000 iterations. The optimal distribution density $P(x)$ for $\alpha = 0.25$ is not shown since it coincides with that at $\alpha = 0.15$. Capacity C increases and then decreases with parameter α , therefore the optimal value exists. Figure 5 demonstrates that the noise squeezing leading to capacity increase over the capacity of the linear white Gaussian noise channel can be achieved with α varied over wide interval $\alpha = 0.3 - 0.7$.

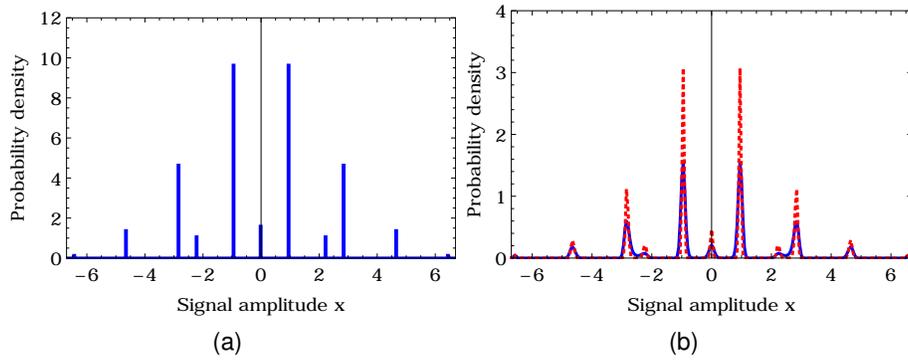


Fig. 3. Optimal probability density $P(x)$ for the regeneration channel: (a) calculated by the relaxation method (a), by BAA after 2000 iterations (dots), 10000 iterations (dashed line) and 20000 iterations (solid). Parameters of the modeling are $S = 5, \alpha = 0.1$.

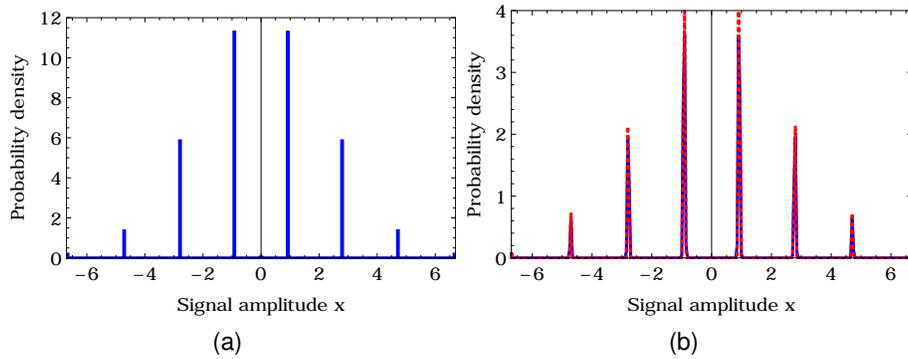


Fig. 4. Optimal probability density $P(x)$ for the regenerative channel: (a) calculated by the relaxation method, (b) by BAA after 5000 iterations (solid line) and 20000 iterations (dotted). Parameters of the modeling are $S = 5, \alpha = 0.15$.

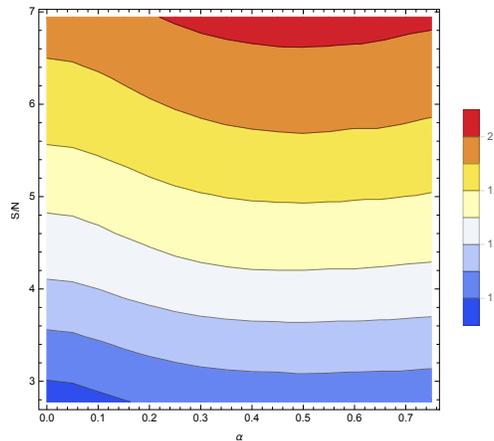


Fig. 5. The channel capacity C/B as a function of nonlinearity α and signal-to-noise ratio S/N . The edge curve $\alpha = 0$ corresponds to the Shannon limit for linear real channel.

4. Hybrid method

Let us compare the methods using the number of iterations as the optimization parameter. At $S = 5, \alpha = 0.15$ the maximal probability $p = 0.27$ corresponds to $x = \pm 0.95$ (these values are displayed in Fig. 4). Consider the vicinity of the points $x = \pm 0.95$. The probability to find x within interval $(x - \Delta/2, x + \Delta/2)$ is given by formula

$$P(x, \Delta) = \int_{x-\Delta/2}^{x+\Delta/2} p(x') dx'.$$

The greater is interval Δ the better it is the accuracy of determining the probability p . At the same time, the accuracy of determining position x decreases with increasing Δ . Figure 6 presents the optimal probability P as a function of iteration number (more precisely the decimal logarithm of the iteration number). One can see that the number of iteration is much less when the relaxation method is employed.

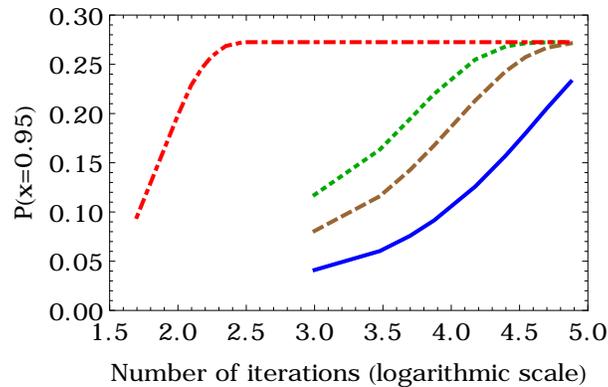


Fig. 6. The comparison of methods. Probability P of symbol $x = 0.95$ as a function of the number of iterations (\log_{10} scale) calculated by the BAA at $\Delta = 0.3$ (solid line), 0.2 (dashed), 0.1 (dotted). The upper curve depicts the result of relaxation method at $\Delta = 0.001$.

Let us point out the main difference of the relaxation method from that by Blahut – Arimoto algorithm [2]. The BAA fixes the values $x_1 < x_2 < \dots < x_L$ and determines the optimal probabilities p_1, p_2, \dots, p_L of these values. In the relaxation method all the values x_i are “moving” (more precisely “running”) to increase the mutual information at fixed probabilities p_i . The initial values p_1, \dots, p_L may be chosen equal to $1/L$ or some other initial set may be used. For example, it is possible to use a combination of both methods — a hybrid capacity computation method — to find the discrete distribution $P(x) = \sum_{i=1}^K p_i \delta(x - x_i)$ faster. Let the optimal points x_1, \dots, x_K and corresponding probabilities p_1, \dots, p_K are unknown. We assume the trial values for a dense grid of fixed points $\tilde{x}_1, \dots, \tilde{x}_K$ and initial probabilities $\tilde{p}_1, \dots, \tilde{p}_K$. The BAA reveals the optimal probabilities including some zero probabilities. The optimal points x_i are accurate within the step of initial grid. To accelerate the search, first one makes n Blahut – Arimoto iterations, and then exclude the zero probabilities and corresponding points from the set $\tilde{x}_1, \dots, \tilde{x}_L$. The result can be used as initial values in relaxation method.

The optimization problem of mutual information for the channel transmitting a finite set of symbols is reduced to the determination of the optimal alphabet sequence $\{x_1, \dots, x_L\}$ and calculation of the best probabilities $p_i > 0$ for each x_i . The best alphabet found by BAA is chosen from a domain $[a, b]$. The initial values $a = x_1 < x_2 < \dots < x_L = b$ are assumed with some step Δ . The less is step Δ the better it is the accuracy of calculated optimal values. The

probabilities p_i of values x_i occur to be nonzero for closest values to given x_i . It is demonstrated by Figs. 1(b) and 2(b). As the number of iterations increases, the function $P(x)$ becomes more similar to the sum $\sum_k p_k \delta(x - x_k)$.

However, the number of iterations required to get the comb of delta-functions is by 1 – 2 orders greater than in the suggested relaxation method. The advantage of new method is the initial form of function $P(x)$ which already consists of delta-function “lumps”. During the calculations the positions x_i move to provide the optimal value of functional (1). The BAA calculations broaden the lumps. That is why the greater number of iterations is needed for a good approximation.

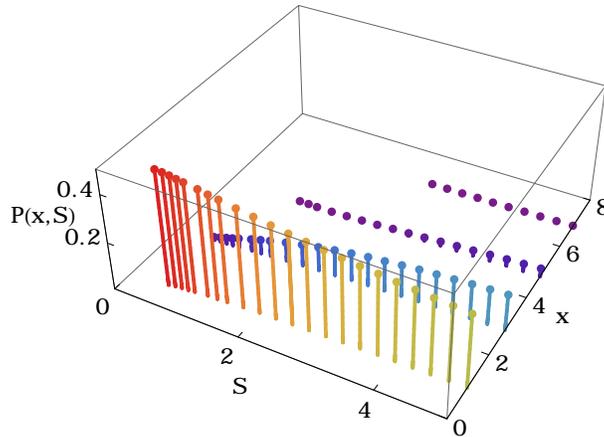


Fig. 7. Probability distribution $P(x)$ as a function of energy S at $\alpha = 0.25$.

The BAA computing gives a smooth distribution in principle, but the channel input probabilities after running BAA could become a good initial condition for the relaxation method. Let us use the combined technique to nonlinear regenerative channel. An example calculated by the hybrid method is shown in Fig. 7. The initial positions are chosen in the nodes of uniform grid. Initial 200 iterations are executed by BAA. The step is $\Delta = 0.02503$, $x \in [-10, 10]$, the number of points is $M = 800$. Then 40 iterations are carried out by the relaxation method. The results of BAA iterations are exploited as initial values of probabilities. The main ridge is placed at $x \approx 1$. Its position x and height P (shown by steams) are slightly variable with S . Additional ridges burn with increasing S . Their heights are growing with S but remain sufficiently small, especially at large x . That means the higher number of symbols in optimal input alphabet with increasing energy. The negative part is not shown in Fig. 7, since it repeats the positive part $P(-x, S) = P(x, S)$.

Another example of the hybrid calculation is the optimal Rice probability distribution at $S = 5$. In this case the binary format is not the better from the capacity point of view. The distribution $P(x)$ is obtained by the BAA method after 20000 iterations. They are shown in Fig. 8 by the solid line and do not change after the additional 20000 BAA iterations. The maximum points are $x \approx 0, 3.4, 5.75, 7.9$. The peaks are broadened with increasing x . The new relaxation method is applied to complete the calculation. The equidistant grid from BAA method $x_i = H(i - 1)/n$ is used as the sequence of initial values. Here $H = 20$ is the length of segment, $n = 801$, $\Delta x = 0.025$. The output of BAA calculations are exploited as an initial values of probability $q_i = P(x_i)\Delta x$. The variation of probability distribution is shown in Fig. 8 by dashed line and circles. It is evident that the distribution converges to the discrete distribution. The peaks at $x = 0, 3.44, 5.77$ are clearly distinguishable.

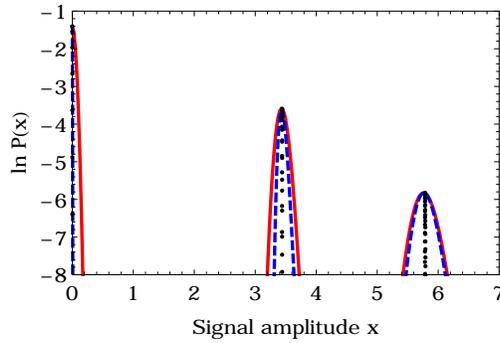


Fig. 8. Probability distribution logarithm $\ln P(x)$ for the Rice channel: initial distribution obtained by BAA (solid curve), after 10 iterations by new method (dashed), after 400 iterations (circles). The number of nodes is 801, the step $\Delta x = 0.025$ in segment $x \in [0, 20]$.

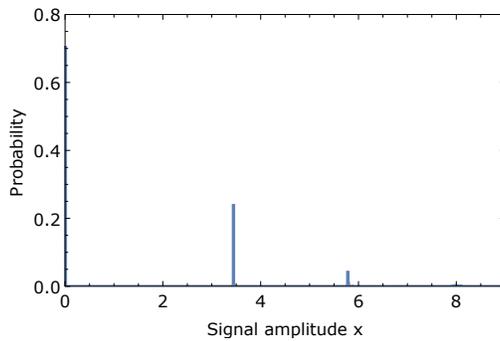


Fig. 9. The histogram of probabilities for the Rice channel calculated by 20000 BAA and 400 relaxation method iterations.

To present the result it is convenient to plot the histogram. The segment of x variation $[0, H]$ is divided by K half-intervals: $[0, \delta_1) \cup \dots \cup [\delta_{K-1}, H)$. Then we determine how many points hit the interval and sum up the probabilities of these points p_i . We obtain the probabilities that are not negligible. The accuracy of position is determined by corresponding the half-interval length. The histogram is shown in Fig. 9.

Note that the relaxation method enables one to get a local maximum of the mutual information functional. When the alphabet is fixed the maximum is unique. When the optimal input alphabet is not known there are several local maxima. The search of global maximum is a separate intricate problem. For this problem the combination of new method with BAA into novel hybrid method seems to be a promising approach. For each initial condition BAA would help to refine the points, and then the probabilities could be computed by the relaxation method. To make sure that we get the global maximum at first we calculate with small accuracy, choose the appropriate initial condition, and then recalculate the maximum with better accuracy.

The modern accelerated methods for calculating the capacity [5, 8] involve two limitations: (i) of signal peak power $x \in [-W, W]$ and (ii) of average power (2). The limitation of average power in the memoryless channels is necessary, since without this limitation the capacity could become unrestrictedly large with growing S . Like in paper [10], we consider the problem with limitation (5). The strict limitation enables one to avoid the additional limitation of the signal peak power.

5. Conclusions

Analysis of capacity of fiber-optic communication channels has attracted a great deal of interest in recent years (see e.g. [18–23] and references therein). This interest is directly linked to the major challenge faced by the modern optical communication systems — necessity to cope with the fast growing information traffic that requires corresponding increase of the networks capacity. The problem of computation of the Shannon (channel) capacity is a classical problem of the information theory. Only a short time ago it was recognized that the Shannon capacity of the nonlinear fiber channels has not yet been fully understood [18–21]. The famous Shannon result for capacity [1] is derived for a linear channel with the additive Gaussian noise. However, fiber channels are nonlinear, their properties are changed with increase of signal power. Methods for calculation of the Shannon capacity of complex nonlinear channels with memory, such as, fiber channels are not yet well developed. The key characteristic of the channel, conditional probability of receiving certain output y provided that signal x was sent, is not known for nonlinear fiber channels. Current estimates of the lower bound of capacity are based on direct computation of mutual information for different modulation formats for varying input signal power [19, 20]. This approach allows one to predict and describe certain many important features of the information transmission, but, of course, this does not answer the question about the maximum possible error-free transmission rate. More efforts are needed in development of novel and enhancing the existing methods of capacity computation for nonlinear fiber channels.

There are interesting related research directions: capacity analysis in specially designed nonlinear channels (e.g. with signal regeneration or noise squeezing elements) [11, 12] that, in principle, would allow capacity to be higher than in the linear AWGN channel and studies of capacity of quantum information channels [24, 25]. The numerical methods presented here are relevant to these research directions and we anticipate that application of our approach will lead to new results in these areas too. In this paper we applied new method for computation of the Shannon capacity of nonlinear noise squeezing channel and confirmed that its capacity is greater than the AWGN channel capacity (see Fig. 5). We have found that the optimal distribution becomes discrete even at small parameter α .

In conclusion, the relaxation method is proposed for maximization of the Shannon mutual information functional (1) both for discrete and continuous input, continuous output channels. The method is compared with the well-known Blahut – Arimoto algorithm for Gaussian and Rician channels and a nonlinear channel with a cascade of noise squeezing elements. The advantage of new method is demonstrated, especially for the problems with the initially unknown discrete optimal alphabet or if the optimal input alphabet is continuous, but we approximate capacity using discrete inputs. We demonstrated also the advantage of using the new hybrid method combining BAA and the relaxation approach. The presented new methods can be applied to more complex channels, including recently introduced models (see e.g. [21, 26–28] and references therein).

A. Appendix. Numerical optimization

Let us state the numerical approximate optimization as a lemma.

Lemma: Denote

$$G[P(x)] = \int \int dx dy P(x|y)P(x) \log_2 \left[\frac{P(y|x)}{\int ds P(y|s)P(s)} \right].$$

Assume that (a) there exists a finite maximum value C of functional G ; (b) for each given

accuracy δ the discrete sequence $\{p_1, \dots, p_N\}$ and the set of points $\{\tilde{x}_1, \dots, \tilde{x}_N\}$ exist such that

$$P(x) = \sum_{i=1}^N p_i \delta(x - \tilde{x}_i), \quad \sum_{i=1}^N p_i = 1$$

satisfies inequality

$$|C - G[P(x)]| < \delta.$$

Then

$$\lim_{M \rightarrow \infty} \left[\max_{\{x_1, \dots, x_M\}} \sum_{k=1}^M p_k \int dy P(y|x_k) \log_2 \frac{P(y|x_k)}{\sum_{j=1}^M p_j P(y|x_j)} \right] = C$$

Here $p_j = 1/M, j = 1, \dots, M$.

Proof: Denote

$$P_M(x) = \sum_{i=1}^M \frac{1}{M} \delta(x - x_i).$$

Specify ε and show that $\exists M_0$ such that at $M \geq M_0$

$$\left| \max_{\{x_1, \dots, x_M\}} G[P_M(x)] - C \right| \leq \varepsilon.$$

By assumption there exists a set of points $\{\tilde{x}_1, \dots, \tilde{x}_N\}$ and probabilities $\{p_1, \dots, p_N\}$ such that

$$|C - G[P(x)]| \leq \varepsilon/2, \quad P(x) = \sum_{i=1}^N p_i \delta(x - x_i).$$

Since functional $G[P(x)]$ is a continuous function of variables p_1, \dots, p_N , there exist $\Delta_1 = (p_1 - \delta_1, p_1 + \delta_1), \dots, \Delta_N = (p_N - \delta_N, p_N + \delta_N)$, the neighborhoods of p_1, \dots, p_N , correspondingly, such that

$$|G[P'(x)] - G[P(x)]| \leq \varepsilon/2, \quad P'(x) = \sum_{i=1}^N p'_i \delta(x - \tilde{x}_i)$$

as soon as $p'_i \in \mathcal{D}_i$, i.e. $|p'_i - p_i| < \delta_i$.

Let M be a positive integer such that each interval Δ_i includes several points of the form m/M , where m is a natural number, to the right and to the left from point p_i . Note that the interval Δ_i includes no less than $|\Delta_i| M - 1$ points of the form m/M . Choose points

$$\frac{m'_1}{M} \in \Delta_1, \dots, \frac{m'_N}{M} \in \Delta_N$$

and

$$\frac{m''_1}{M} \in \Delta_1, \dots, \frac{m''_N}{M} \in \Delta_N$$

such that $m'_i/M \leq p_i \leq m''_i/M$. Then

$$\sum_{i=1}^N \frac{m'_i}{M} \leq \sum_{i=1}^N p_i = 1 \leq \sum_{i=1}^N \frac{m''_i}{M}.$$

Since points $m'_i/M, (m'_i + 1)/M, \dots, m''_i/M \in \Delta_i$, there exist a natural numbers m_i , such that $m'_i \leq m_i \leq m''_i$ and $\sum_{i=1}^N m_i = M$. Denote $\tilde{p}_k = m_k/M$. Since $\tilde{p}_k \in \Delta_k, k = 1, \dots, N$, then

$$|G[\tilde{P}(x)] - G[P(x)]| \leq \varepsilon/2, \quad \tilde{P}(x) = \sum_{k=1}^N \tilde{p}_k \delta(x - x_k).$$

It is obvious that

$$\max_{\{x_1, \dots, x_M\}} G[P_M(x)] \geq G[\tilde{P}(x)].$$

Since

$$\left| C - G[\tilde{P}(x)] \right| \leq \left| C - G[P(x)] \right| + \left| G[P(x)] - G[\tilde{P}(x)] \right| \leq \varepsilon$$

and

$$C \geq \max_{\{x_1, \dots, x_M\}} G[P_M(x)] \geq G[\tilde{P}(x)].$$

We get

$$\left| C - \max_{\{x_1, \dots, x_M\}} G[P_M(x)] \right| \leq \varepsilon,$$

then the lemma is valid. Q.E.D.

Acknowledgments

This work is supported by Grant of Ministry of Education and Science of the Russian Federation (agreement No. 14.B25.31.0003), the EPSRC project UNLOC, the Government program of the leading research schools NSH-4447.2014.2.