Metabolic cost of neuronal information in an empirical stimulus-response model

Lubomir Kostal* and Petr Lansky
Institute of Physiology of the Czech Academy of Sciences, Videnska 1083, 14220 Prague 4, Czech Republic

Mark D. McDonnell
*Institute for Telecommunications Research, University of South Australia, Building W,
Mawson Lakes Campus, Mawson Lakes Boulevard, Mawson Lakes SA, 5095, Australia*

The limits on maximum information that can be transferred by single neurons may help us to understand how sensory and other information is being processed in the brain. According to the efficient-coding hypothesis [8], neurons are adapted to the statistical properties of the signals to which they are exposed. In this paper we employ methods of information theory to calculate, both exactly (numerically) and approximately, the ultimate limits on reliable information transmission for an empirical neuronal model. We couple information transfer with the metabolic cost of neuronal activity and determine the optimal information-to-metabolic cost ratios. We find that the optimal input distribution is discrete with only six points of support, both with and without a metabolic constraint. However, we also find that many different input distributions achieve mutual information close to capacity, which implies that the precise structure of the capacity-achieving input is of lesser importance than the value of capacity.

Keywords: Information capacity, Metabolic cost, Stimulus-response curve

1. INTRODUCTION

One of the primary goals of neuroscience is to understand exactly how neurons process and convey information [58, 66]. Theoretical approaches to this problem have received significant attention over the past few decades [57, 58], with the earliest information-theoretic studies appearing relatively shortly after Shannon published his mathematical theory of information and communication [30, 56]. Since then, optimality conditions on information transfer in neuronal populations or single neuronal models have been frequently considered, leading to the idea of *efficient coding* [3, 4, 8], often providing inspiring results under varying assumptions. Among others, the following scenarios were analyzed in detail

- stochastic integrate-and-fire class of models and temporal spiking codes in Kostal [36], Suksompong and Berger [68],

- Poisson-type neuronal models in Johnson and Goodman [31], Nikitin *et al.* [54],

- Gaussian distribution of response variability (given each stimulus) in de Ruyter van Steveninck and Laughlin [67]

- deterministic response model in Laughlin [42]

- vanishing amplitude of the response variability in Brunel and Nadal [12], McDonnell and Stocks [49]

- weakly tuned systems (large response variability) in Kostal and Lansky [37]

- empirical neuronal models with rate or temporal coding schemes in Ikeda and Manton [29].

The input-output properties of neurons are usually characterized by their stimulus-response (S-R) function, i.e. by plotting the firing frequency (or any other measurable neuron response) versus the corresponding stimulus intensity, such as in Fig. 1. The goal of many information theoretic approaches is to identify stimulus characteristics that maximize information transfer between stimuli and responses. In a classical paper, Laughlin [42] applied an information-theoretic reasoning for a deterministic neuron, resulting in the optimality condition that the stimulus distribution function is proportional to the S-R curve derivative (see Fig. 1a).

In a more realistic neural system, however, the noise (i.e., the response variability within and across trials) is considered an integral part of the process (see, e.g., Lansky and Sacerdote [41], Shadlen and Newsome [63], Tuckwell [70]). The presence of response variability decreases the availability of simple closed-form solutions to the information-maximization problem [19, 23, 71], and also affects the character of the "optimal" input signal [23, 26, 40]. Under certain conditions, relatively uncomplicated approximate methods can be used, typically for monotonic stimulus-response curves. The asymptotic relation between mutual information and Fisher information [9, 59] has been employed for the analysis of optimality conditions in the setting of large signal-to-noise ratios [12, 49]—see Fig. 1b. The information-optimizing input probability density is then proportional to the square root of the Fisher information. The "opposite" situation, the small signal-to-noise ratio scenario (Fig. 1c), has been recently investigated in Kostal [35], Kostal and Lansky [37].

Numerical optimization of the information-theoretic quantities is the only option left in the general case (Fig. 1d). Specialized algorithms [18, 23] are usually required, and consequently studies on realistic neuronal models with stimulus-
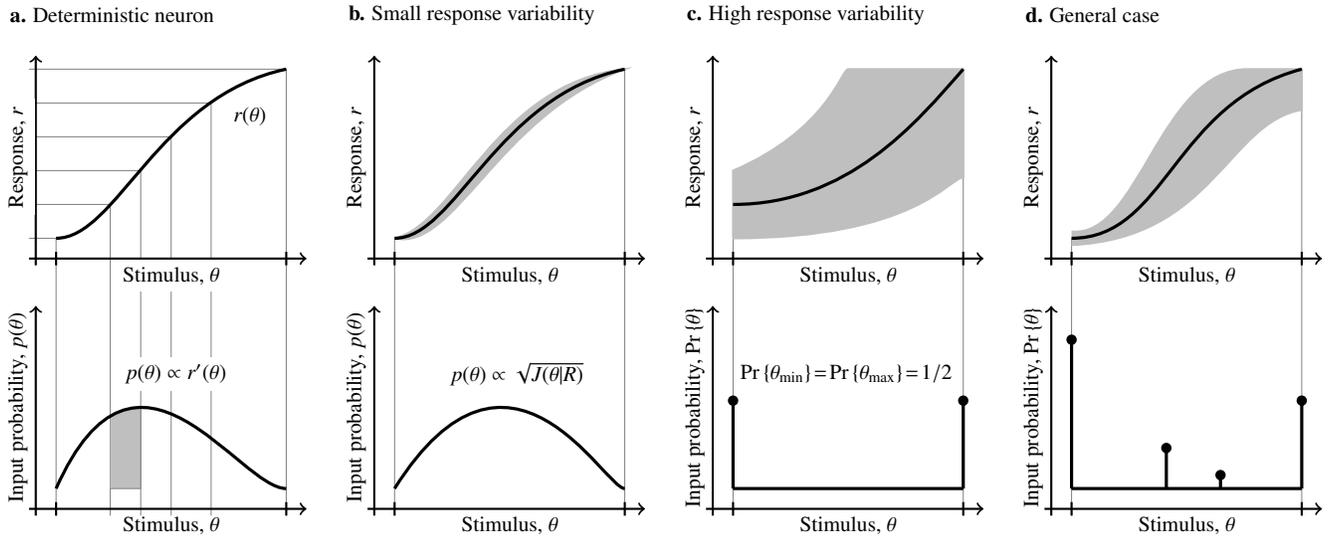
* E-mail: kostal@biomed.cas.cz

Figure 1. Illustration of typical stimulus-response relationships investigated in literature on the information-optimality conditions in neurons. The mean response together with its standard deviation and the corresponding information-maximizing input distributions are shown. The classical deterministic scenario (**a.**) has been employed frequently since Laughlin [42], although the lack of response variability may be viewed as a drawback. If the neuron responds with small response variability (possibly stimulus-dependent), the low-noise approximation (**b.**) results in continuous optimal input probability densities [9, 12, 49]. On the other hand, for neurons weakly tuned to the variation in stimulus (**c.**), the high-noise approximation yields binary equiprobable input distribution [35, 36]. The general (i.e., with no assumptions on the strength and character of response variability) and also the most realistic case (**d.**), appears less frequently [29, 68], mainly due to difficulties associated with calculation of the information capacity. The optimal input distribution is then usually discrete, with finite number of points of support. In this paper we compare both low- and high-noise approximations with the exact (numerical) solutions, while the effect of neuronal metabolic activity on the information transfer is taken into account.

dependent response variability are infrequent [46, 68, 74].

The main goal of this paper is to show how to provide exact (i.e., numerical) and approximate solutions to the problem of information maximization. As an example we consider an empirical S-R relationship based on classical experiments [52], and the precision of the approximate expressions is checked. Furthermore, since neurons consume significant amounts of energy during their activity [51], the metabolic cost is taken into account when identifying the optimal information versus metabolic cost performance. We correspondingly extend both low- and high-noise approximations to take into account the metabolic cost constraints. By employing specialized numerical procedures, we show that optimal information transfer can be near-achieved by a number of different input distributions, which implies that the precise structure of the capacity-achieving input is of lesser importance than the value of capacity.

2. METHODS

2.1. Stimulus-response characteristics obtained from experimental data

As already mentioned, the input-output properties of neurons are commonly characterized by the stimulus-response (S-R) function, $r(\theta)$, in which the mean output response, r, is plotted against the stimulus intensity, θ. The output is usually the spiking frequency, but it can be the level of any variable of interest, e.g., first-spike latency, channel conductance, receptor potential [27, 39, 42, 72], or even the mean frequency or peak power of an oscillatory random signal [48]. The S-R curves are usually monotonically increasing, and often of sigmoidal shape [13, 40, 49]. They are usually presented as a single curve, describing the mean response, but less frequently they are accompanied by standard deviations to account for the response variability.

For the purpose of this paper we selected an empirical S-R relationship, obtained from experimental data on thalamic neurons of macaque monkeys. The precise experimental setup is not important for this study, we treat the model as the basic case of stimulus-response relationship in order to illustrate the information-theoretic methodology. For each value of stimulus intensity, θ, the response neuronal firing rate, $r(\theta)$ (number of spikes in a time window 200–1000 ms long) was measured. It was found, that a majority of neurons examined in the experiment satisfy the S-R relationship of the form [52]

$$r(\theta) = a\,(\theta - \theta_0)^b + c, \tag{1}$$

where the constants a, b, c, θ_0 vary among neurons, here we employ the particular fit with values $a = 5.51, b = 0.632, c = 10, \theta_0 = 0$ and the minimal and maximal values of θ are 0 and 100 respectively. In order to describe the variability of responses in dependence on θ [52], we chose the generic
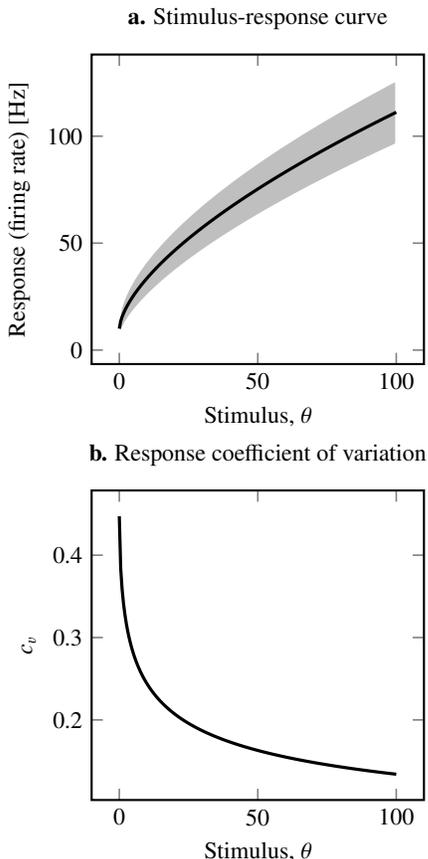
a. Stimulus-response curve

b. Response coefficient of variation

Figure 2. Simulus-response characteristics based on experimental data. The model was obtained by fitting experimental measurements coming from thalamic neurons of macaques [52]. Mean stimulus-response curve (solid) and standard deviation are indicated (**a**). Co-efficient of variation describes the relative strength of the response variability with respect to the mean response value (**b**). Note that the response variability is neither vanishing nor too high (*cf.* Fig. 1).

Poisson-like behavior, i.e., the variance of spike count (frequency) being proportional to the mean [70]

$$\sigma(\theta) = \sqrt{\alpha r(\theta)}, \tag{2}$$

where $\sigma(\theta)$ is the standard deviation of the responses and the constant of proportionality is selected to be $\alpha = 2$ The S-R relationship is shown in Fig. 2a.

The response coefficient of variation,

$$c_v = \frac{\sigma(\theta)}{r(\theta)}, \tag{3}$$

determines the relative response variability with respect to its mean value (Fig. 2b), and can be considered to be the inverse of the "signal-to-noise ratio" [62]. By visual inspection we conclude that the investigated S-R model neither is in the "low-noise" (Fig. 1b) nor the "high-noise" (Fig. 1c) category, rather, it is a generic member of the "general" category (Fig. 1d). We will make this inference more precise by comparing the approximate methods with exact solutions.

We stress at this point, that the main goal of this paper is to demonstrate and interpret the information-theoretic methodology; the particular choice of the S-R model is of secondary importance.

2.2. Metabolic cost of neuronal activity

Neurons use significant amount of energy for the spiking activity, and thus energy usage should be coupled to considerations about the efficiency of neuronal information transfer [6, 43–45, 51, 61]. Attwell and Laughlin [5] used anatomic and physiologic data to analyze the metabolic cost of different components of excitatory signaling. They showed that the signaling-related energy consumption increases linearly with spiking frequency. The linearity of the dependence over a wide range of firing frequencies is also confirmed by numerical studies on biologically relevant single-compartment neuronal models [7].

In the empirical S-R model given by Eqns. (1) and (2), the response is the neuronal firing rate (number of action potentials per second). The signaling-related metabolic cost, $w(\theta)$, of neuronal activity associated with stimulus intensity θ can thus be defined as

$$w(\theta) = \kappa r(\theta), \tag{4}$$

where the constant of proportionality, $\kappa = 7.1 \times 10^8$ ATP molecules, describes the metabolic cost of a single action potential [5].

Consider an ensemble of different stimulus intensities, each occurring with different relative frequency (i.e., probability). Such an ensemble is described by the probability density function (p.d.f.), $p(\theta)$. The *average* metabolic cost, W, associated with the stimulus ensemble is then expressed as

$$W_p = \int_\Theta w(\theta) p(\theta) \, d\theta, \tag{5}$$

where the integral is over all possible stimulus intensities.

2.3. Information capacity and capacity-cost function

Under the generally accepted view that the task of a sensory neuron is to communicate the value of its stimulus to its target neurons, the logical next task is to find the maximum amount of information that can be conveyed [68]. To accomplish the task, the neuron is treated as an information channel, see Dimitrov and Miller [20], Johnson and Goodman [31], Kostal, Lansky, and Rospars [39], Laughlin [42], Mc-Donnell and Stocks [49], Stein [65], de Ruyter van Stevenick and Laughlin [67], Suksompong and Berger [68] among others.

Within the framework of information theory, the *information* about some particular stimulus, θ, from observing the response, r, depends on the stimulus ensemble described by the p.d.f. $p(\theta)$ as in [23, p. 16]

$$i(\theta; r) = \ln \frac{\phi(\theta|r)}{p(\theta)}, \tag{6}$$

where $\phi(\theta|r)$ is the p.d.f. describing the relative frequencies of possible stimuli intensities for observed r. Intuitively, the most informative stimulus-response pairs are those, in which the response r can be used to "identify" θ with high specificity, i.e., with high *a posteriori* to *a priori* probability ratio. The fact that the stimulus ensemble is described probabilistically by $p(\theta)$ can also be (informally) justified, since during stimulation, stimulus values cannot be known beforehand (from the "point of view" of the neuron).

The p.d.f. $\phi(\theta|r)$ is usually difficult to calculate, therefore Bayes' law is used to write Eq. (6) equivalently as

$$i(\theta; r) = \ln \frac{f(r|\theta)}{f(r)}, \qquad (7)$$

where the (unconditional) response p.d.f. is $f(r) = \int_\Theta f(r|\theta)p(\theta)\, d\theta$. The conditional p.d.f. $f(r|\theta)$, describing relative frequencies of different responses evoked by stimulus θ, is related to the S-R curve (Fig. 1). In fact, $f(r|\theta)$ can be thought of as the *complete* descriptor of the model since both the S-R curve, $r(\theta) = \int_R r f(r|\theta)\, dr$, and (analogously) the standard deviation, $\sigma(\theta)$, can be calculated from it.

The average value of $i(\theta; r)$ taken over all stimulus-response pairs is denoted as the *mutual information*, $I(\Theta; R)$, between stimuli and responses,

$$I(\Theta; R) = \int_\Theta \int_R i(\theta; r) f(r|\theta) p(\theta)\, dr\, d\theta. \qquad (8)$$

If the logarithm in Eq. (7) is changed to base 2, then $I(\Theta; R)$ gives the average number of *bits* per response (i.e., per single channel use) that can be transmitted.

The maximum rate of reliable information transmission, denoted as information *capacity*, C, is then

$$C = \max_{p(\theta)} I(\Theta; R), \qquad (9)$$

where the maximum is taken over all possible input distributions defined over the allowed stimulus range. The *capacity-cost* function [50], $C(W)$, gives the information capacity under the additional constraint that the average metabolic cost W_p given by Eq. (5) does not exceed some selected value W,

$$C(W) = \max_{p(\theta), W_p \le W} I(\Theta; R), \qquad (10)$$

Among the properties of $C(W)$ defined by Eq. (10) are, that $C(W)$ is a non-decreasing function of W and that $C = C(W)$ for W growing without any bounds. However, the capacity in Eq. (9) is in many cases achieved already at some finite value of W, denoted here as $W^\dagger$.

The optimal balance between the information capacity and metabolic cost is given by the *capacity per unit cost* [71],

$$C^* = \max_W \frac{C(W)}{W}. \qquad (11)$$

The inverse value, $1/C^*$, can be interpreted as the minimal possible cost of a reliably transmitted bit. Furthermore, we define the *capacity at optimal cost*, $C(W^*)$, where W^* is the optimal cost solving Eq. (11), if the solution exists.

Eqns. (9)–(11) can rarely be solved in a closed form. However, approximate as well as specialized numerical methods exist. In the following we describe two approximate methods valid for the case of high and low response variability (Fig. 1), and an efficient optimization procedure [28] usable for all general cases of practical interest.

2.3.1. Approximate methods

The advantage of approximations lies in the fact that they allow us to investigate, although locally and perhaps under a restrictive scenario, the effect of individual parameters of the problem on the optimality conditions. The effect of changing parameter values might be difficult to assess by employing only numerical methods.

High-noise approximation. For very low signal-to-noise ratios (Fig. 1c), the capacity-achieving stimulus distribution is well approximated by a binary distribution, with equiprobable masses located at the minimal, resp. the maximal allowed stimulus intensities ($\theta_{\min}$ resp. $\theta_{\max}$), as shown under different assumptions in Huang and Meyn [28] and Kostal [35]. The lower bound on the true capacity, $C \ge C_{\text{bin}}$, can thus be written as

$$C_{\text{bin}} = \frac{1}{2} \int_R i(\theta_{\min}; r) f(r|\theta_{\min})\, dr +$$
$$+ \frac{1}{2} \int_R i(\theta_{\max}; r) f(r|\theta_{\max})\, dr, \qquad (12)$$

where $f(r)$ in the definition of $i(\theta; r)$ in Eq. (7), is $f(r) = [f(r|\theta_{\min}) + f(r|\theta_{\max})]/2$, and the integrals in Eq. (12) are over all possible responses. Despite the simplicity of the binary approximation, Eq. (12), the difference $C - C_{\text{bin}}$ is usually negligible for $C < 0.8$ bits, see Kostal [35] for more details and examples.

Both approaches establishing Eq. (12) in Huang and Meyn [28] and Kostal [35] are somewhat technical, but Eq. (12) can be heuristically justified as follows. If the response variability is too large, then by increasing the stimulus repertoire (i.e., by adding any third value to the two possibilities $\theta_{\min}$, $\theta_{\max}$) we also increase the uncertainty of correct stimulus estimation so much, that the reliability of information that can be transmitted decreases.

a. Low-noise approximation. If the noise in information transfer is low (Fig. 1b), there exists a lower bound, $I_{\text{low}}(\Theta; R)$, on the mutual information from Eq. (8), given as (see, e.g., Brunel and Nadal [12] for details)

$$I_{\text{low}}(\Theta; R) = - \int_\Theta p(\theta) \ln \left[p(\theta) \sqrt{\frac{2\pi e}{J(\theta|R)}} \right] d\theta, \qquad (13)$$

where $J(\theta|R)$ is the Fisher information obtained about a specific stimulus value, $\Theta = \theta$, given an observed response distribution, R:

$$J(\theta|R) = \int_R \left[\frac{\partial \ln f(r|\theta)}{\partial \theta} \right]^2 f(r|\theta)\, dr. \qquad (14)$$

The low-noise bound $I_{\text{low}}(\Theta; R)$ has been employed in Bernardo [9], Brunel and Nadal [12], Clarke and Barron [16], McDonnell and Stocks [49], Nadal and Bonnasse-Gahot [53] among others.

The standard Euler-Lagrange method can be applied to maximize $I_{\text{low}}(\Theta; R)$ with respect to $p(\theta)$, yielding the *low-noise* approximation to the capacity, $C \geq C_{\text{low}}$,

$$C_{\text{low}} = \ln \frac{\int_\Theta \sqrt{J(\theta|R)}\, d\theta}{\sqrt{2\pi e}}. \tag{15}$$

with optimizing density $p(\theta) \propto \sqrt{J(\theta|R)}$ (also known as Jeffrey's prior [9]). This optimal solution can also be obtained by rewriting Eq. (13) as a Kullback-Leibler entropy, and noting that such a quantity is always non-negative [49]. The optimization of Eq. (13) will be shown in more detail with additional metabolic cost constraint in Section 3.1. Note that the Euler-Lagrange method leading to Eq. (15) requires $p(\theta)$ to be differentiable, confined to a finite stimulus range and also strictly nonzero over this range.

2.3.2. Numerical methods

The solution to the problems described by Eqns (9) and (10) lies in finding the maximizing input probability density, $p(\theta)$, for the non-linear functional $I(\Theta; R)$. For the purpose of numerical implementation, however, the entire function $p(\theta)$ must be represented by a finite set of numbers. Basically, either the function $p(\theta)$ is evaluated on a set of grid points (possibly non-equidistant) $\{\theta_1, \theta_2, \ldots, \theta_k\}$, or various sophisticated particle-filter methods are employed [18]. The latter approach is computationally very intensive, and its advantages can be considered as minor when a sufficiently dense input grid is available. Furthermore, nearly all information-optimal input densities are known to be discrete (e.g., composed of Dirac's delta functions, Fig. 1d), see, e.g., Chan, Hranilovic, and Kschischang [14], Huang and Meyn [28], Ikeda and Manton [29], McDonnell and Flitney [47], Nikitin *et al.* [54], Tchamkerten [69].

By employing the input grid of size k, the problem of finding the capacity is reduced to finding the maximum of a function in a k-dimensional space. Furthermore, since $I(\Theta; R)$ is concave in $p(\theta)$, the optimization involves no local maxima and the solution, if it exists, is unique. Employing powerful (i.e., gradient-based) optimization methods is, however, not efficient in a highly-dimensional space, since specialized numerical procedures for finding the capacity have been proposed. The classical Blahut-Arimoto algorithm [10] requires non-trivial extensions to incorporate continuous-valued responses and metabolic cost constraints to determine the capacity-cost function.

In this paper we employ the cutting-plane algorithm [28, 33], which is applicable under more general circumstances than the Blahut-Arimoto algorithm. Among the advantages are the reported faster convergence, easy addition of input constraints, provided they are expressed as expectations over $p(\theta)$, and control over the numerical precision of the result.

The principle of the cutting-plane algorithm is the representation of a non-linear optimization problem as a sequence of converging linear programming problems.

We also note that, as discussed by Chiang and Boyd [15], Komninakis, Vandenberghe, and Wesel [34], McDonnell and Flitney [47], the problem of finding the capacity achieving input distribution can be solved using convex optimisation techniques [11]. After discretizing the input and output variables, we were therefore able to verify the solutions we found with the cutting-plane algorithm using CVX, which is a package for specifying and solving convex problems [25].

Finally we note, that both the low- and high-noise approximations require the S-R curves to be sigmoidal, such as shown in Fig. 1, plus some minor technical conditions on the character of the response variability [12, 35]. However, the original Shannon's theory was proposed under rather broad assumptions [23], and thus the presented numerical methods are valid for a much broader class of S-R relationships (sigmoidal or not), see [28] for some nonrestrictive technical conditions.

3. RESULTS

3.1. Theoretical considerations

It follows from Eq. (8), that in order to calculate the information-theoretic quantities, the full form of the density $f(r|\theta)$ must be known, while our empirical model describes only the mean value, Eq. (1), and the standard deviation, Eq. (2). To proceed, we select the *gamma distribution* as a suitable model for probabilistic description of neuronal firing rates given stimulus intensity [29, 55],

$$f(r|\theta) = r^{k(\theta)-1} \frac{\exp(-r/s(\theta))}{\Gamma[k(\theta)]s(\theta)^{k(\theta)}}, \tag{16}$$

where $\Gamma(x)$ is the gamma function, the shape, $k(\theta)$, and scale, $s(\theta)$, parameters are related to Eqns. (1)–(2) as

$$k(\theta) = \frac{r^2(\theta)}{\sigma^2(\theta)}, \tag{17}$$

$$s(\theta) = \frac{\sigma^2(\theta)}{r(\theta)}. \tag{18}$$

The shape of the gamma density for small values of c_v, see Eq. (3) and Fig. 2b, may resemble the shape of a Gaussian distribution [32]. We also repeated the calculations presented below with a more heavy-tailed model of $f(r|\theta)$ (the lognormal distribution), affecting the results marginally. Thus we conclude that the particular choice of gamma density is not restrictive.

First, we extend the high-noise approximation from Eq. (12) to account for the metabolic cost in Eq. (5), in order to approximate the capacity-cost function in Eq. (10). Several ways to proceed are possible, probably the simplest is to keep the input two-point distribution supported at $\theta_{\min}$ and $\theta_{\max}$, and vary the probability to satisfy the cost constraint $W_p \leq W$. The capacity approximation evaluated at $W = W_p$

can therefore be written as

$$C_{\text{bin}} = (1 - \gamma) \int_R i(\theta_{\min}; r) f(r|\theta_{\min}) \, dr +$$
$$+ \gamma \int_R i(\theta_{\max}; r) f(r|\theta_{\max}) \, dr, \qquad (19)$$

where $0 < \gamma < 1$ is the relative weight, depending on both the average cost W_p and the stimulus-related cost $w(\theta)$,

$$\gamma = \frac{W_P - w(\theta_{\min})}{w(\theta_{\max}) - w(\theta_{\min})}, \qquad (20)$$

and the unconditional response p.d.f. required to calculate $i(\theta, r)$ is $f(r) = (1 - \gamma) f(r|\theta_{\min}) + \gamma f(r|\theta_{\max})$. Note that the maximum of $C_{\text{bin}}(W_p)$ occurs at $\gamma = 1/2$. The high-noise approximation to the capacity cost function evaluated at point W is then obtained by simple maximization of $C_{\text{bin}}(W_p)$ over all $W_p \leq W$.

Second, we extend the low-noise approximation from Eq. (13), to account for the metabolic cost, again first by examining the approximation to capacity evaluated at $W = W_p$. The problem is approached by employing the Euler-Lagrange formalism [73], i.e, by finding a solution to the equation (assuming differentiability of $p(\theta)$ and $p(\theta) > 0$ for the entire stimulus range)

$$\frac{\partial L}{\partial p} - \frac{d}{dx} \frac{\partial L}{\partial p'} = 0. \qquad (21)$$

The Lagrangian L includes Eq. (13), the normalization constraint $\int_\Theta p(\theta) \, d\theta = 1$ and the cost constraint, as

$$L = I_{\text{low}}(\Theta; R) + \lambda_1 \left[\int_\Theta p(\theta) \, d\theta - 1 \right] +$$
$$+ \lambda_W \left[\int_\Theta w(\theta) p(\theta) \, d\theta - W_p \right], \qquad (22)$$

where λ_1, λ_W are the associated Lagrange multipliers. Since the Lagrangian does not depend on the derivative of the input density, p', Eq. (21) reduces to the algebraic form, yielding the optimal density to be

$$p(\theta) = \sqrt{\frac{J(\theta|R)}{2\pi e}} \exp[\lambda_1 - 1 + \lambda_W w(\theta)], \qquad (23)$$

which reduces to the Jeffrey's prior, $p(\theta) \propto \sqrt{J(\theta|R)}$, for stimulus limited to finite range and no metabolic cost constraint ($\lambda_W = 0$). Substituting Eq. (23) into Eq. (13) gives the low-noise capacity approximation at $W = W_p$ as

$$C_{\text{low}} = 1 - \lambda_1 - \lambda_w W_p, \qquad (24)$$

and the approximation to the capacity-cost function evaluated at W is obtained by maximizing Eq. (24) for all $W_p \leq W$. Unfortunately, no closed-form solution for λ_1, λ_W can be obtained, their values must be determined numerically by substituting Eq. (23) into the two constraint conditions (normalization and metabolic cost).

Closed form expression for the Fisher information, $J(\theta|R)$, can be found in terms of the S-R curve, $r(\theta)$, and its standard deviation, $\sigma(\theta)$. Substituting Eqns. (17) and (18) into Eq. (16), and inserting the resulting gamma-density into Eq. (14), leads to

$$J(\theta|R) = \frac{1}{\sigma^6(\theta)} \Big\{ 8r(\theta)\sigma^3(\theta)r'(\theta)\sigma'(\theta) -$$
$$- 3\sigma^4(\theta)[r'(\theta)]^2 - 4r^2(\theta)\sigma^2(\theta)[\sigma'(\theta)]^2 +$$
$$+ 4r^2(\theta)[\sigma(\theta)r'(\theta) - r(\theta)\sigma'(\theta)]^2 \times$$
$$\times \Psi'\Big([r(\theta)]^2/\sigma^2(\theta)\Big) \Big\}, \qquad (25)$$

where $r'(\theta), \sigma'(\theta)$ are the derivatives with respect to θ, $\Psi(x)$ is the digamma function and $\Psi'(x)$ its derivative [2]. The Eq. (25) was obtained by realizing that the integral in Eq. (14) for the gamma density can be expanded into a sum of integrals of the following three types (up to multiplicative factors), $\int_R e^{-r} r^\beta \, dr = \Gamma(\beta+1)$ or $\int_R e^{-r} r^\beta \ln r \, dr = \Gamma(\beta+1)\Psi(\beta+1)$ or $\int_R e^{-r} r^\beta \ln^2 r \, dr = \Gamma(\beta+1)[\Psi^2(\beta+1) + \Psi'(\beta+1)]$, where various functions of $r(\theta)$ and $\sigma(\theta)$ and their derivatives appear instead of β and the multiplicative factors. Although many terms cancel at the end, the exact calculation is long and tedious and therefore we do not repeat it here. Nonetheless, we verified Eq. (25) by inserting various expressions for $r(\theta), \sigma(\theta)$ and comparing the closed-form results in Eq. (25) against numerical integration of Eq. (14). Note, that $J(\theta|R)$ can be zero at points for which a small change in a stimulus intensity does not result in a change in the shape of the density $f(r|\theta)$ (locally). For example, $J(\theta_c|R) = 0$ if for some particular stimulus value θ_c holds simultaneously $r'(\theta_c) = 0$ and $\sigma'(\theta_c) = 0$.

Finally, we connect the low-noise optimal input density approximation, Eq. (23), with the optimal input density for the deterministic case in Fig. 1a. Assume, that the S-R relationship is such, that the response variability is *independent* of the stimulus intensity, i.e., the variability is constant along the whole S-R curve. It follows, that $f(r|\theta)$ must be of the form

$$f(r|\theta) = g(r - r(\theta)), \qquad (26)$$

where r is the response, $r(\theta)$ the S-R curve and $g(r)$ is a p.d.f. with zero mean. Substituting Eq. (26) into Eq. (14), gives (provided that $r(\theta)$ is continuously differentiable)

$$J(\theta|R) = \left(\frac{dr(\theta)}{d\theta}\right)^2 \int_R \left(\frac{d \ln g(r)}{dr}\right)^2 g(r) \, dr, \qquad (27)$$

after employing the fact that the integral in Eq. (14) is shift-invariant [17]. Since the integral in Eq. (27) does not depend on θ, the optimal density in Eq. (23) can be written as $p(\theta) \propto r'(\theta)$, when no cost constraint is applied. We see that the classical deterministic result (Fig. 1a, Laughlin [42]) is applicable also to S-R models with small stimulus-independent noise. Note however the ambiguity of the optimality conditions for the deterministic case from the information-theoretic point of view, since any input density with continuous support achieves infinite mutual information [17].

3.2. Information-optimality conditions

Fig. 3a shows the complete capacity-cost function, with the capacity, $C = 1.78$ bits, and capacity at optimal cost, $C(W^*) = 1.16$ bits, indicated. The value of $C(W^*)$ offers the best balance between the ultimate limit on reliable information transfer, versus the induced metabolic cost of neuronal activity. The capacity is first achieved (within a small tolerance) at the average cost $W^\dagger = 3.9 \times 10^{10}$ ATP molecules per second, while the optimal cost is $W^* = 1.68 \times 10^{10}$ ATP molecs/s. In other words, the optimal balance says that while operating only at 41 % of the required metabolic cost to achieve capacity, it is possible to achieve 65 % of the capacity.

Fig. 3b compares the optimal input distribution for achieving capacity and its low-noise approximation (the trivial high noise approximation is not shown). Sufficiently dense input grid has been used, $k = 300$, to ensure high precision. The results confirm the discrete character of the optimal input distribution being supported only at 6 points with non-zero probability (the positions of these points have been optimized according to the method proposed by Huang and Meyn [28]). On the other hand, the low-noise approximation does not resemble the exact input distribution, confirmed by the approximate capacity value, $C_{\text{low}} = 1.32$ bits. It is worth noting, that employing a uniform input distribution results in $I(\Theta; R) = 1.29$ bits, substantially higher than C_{low}. Obviously, the studied empirical S-R relationship does not fall within the low-noise category. The detailed analysis of the low-noise approximation precision in dependence on the response variability will be published in Kostal and Lansky [38].

In Fig. 3c the ratio $C(W)/W$, required to calculate the capacity per unit cost C^*, shows its relative "sharpness" about the optimal value. Decreasing the average cost from W^* results in a more dramatic decrease of the optimality conditions than by the increase from W^*.

Fig. 3d, similarly to Fig. 3b, compares the exact and low-noise approximation to the input probability distribution achieving the optimal information-cost ratio C^*. Again, the low-noise approximation does not describe the situation well.

The employed numerical algorithm allows to control the numerical precision of the result, therefore it is possible to compare input distributions that near-achieve capacity. For an input distribution resulting in mutual information $I_p = I(\Theta; R)$ we define the relative precision as $1 - I_p/C$. Fig. 4 shows three different input distributions which achieve the capacity within 1 % precision (cf. Fig. 3b). Although the distributions are discrete with 6 points of support, there are differences in the precise stimulus intensity values (except for $\theta_{\min} = 0$ and $\theta_{\max} = 100$ which are stable), and their probabilities in the region $\theta < 50$. We observe that near-capacity (and near-capacity per unit cost) can be achieved by a set of different distributions, and that the convergence towards the exact (stable) solution can be relatively slow (confirming observations of Abou-Faycal, Trott, and Shamai [1]).

4. DISCUSSION

Channel capacity gives the upper bound on *reliable* information transfer, e.g., the neuron can be theoretically used, by appropriately changing the stimulus values, to transmit 100 bits of information, at C bits/response on average, and these 100 bits can be recovered exactly from the responses (with arbitrarily small probability of error). Shannon's channel coding theorem ensures that such a procedure (coding and decoding operation) exists [23]. The complexity of such operations can be sometimes reduced greatly in the presence of feedback (the dependence of the current stimulus on past responses) or by including some fidelity criterion on the approximate reconstruction of the true sensory input [23, 24]. We have not included these possibilities in our current effort. Rather, we take the traditional point of view in computational neuroscience and maximize just the mutual information between stimuli and responses [4, 39, 46, 65].

In this paper, we treated the neuron as a *discrete time* information channel. That is, the stimulus is assumed to be "constant" for a certain amount of time, rather than changing continuously with time. While the "discrete-time" assumption is consistent with the experimental protocol of Mountcastle, Poggio, and Werner [52] it is not clear, whether the assumption holds also for naturally occurring stimuli. Fortunately, the assumption that the channel is discrete-time is not too restrictive, since most continuous-time channels can be viewed as discrete-time with very little loss [50].

Another question is the effect of channel (neuron) *memory*, when the current response depends not only on the current stimulus, but also on past stimulations and responses [23, 35]. In this paper, we do not take the effect of channel memory into account. Broadly stated, the empirical S-R relationship is *static* in the sense that the possible effects of, e.g., neuronal adaptation [22] or realistic temporal response properties of neurons and their effect of information transmission, are not investigated here. Unlike the case when the stimulus statistics can be assumed, (e.g., in Fairhall *et al.* [21]) the mathematical treatment of mutual information maximization for "non-static" neuronal models (information channels) turns out to be rather difficult. Exact calculation of capacity for general channels with memory is currently not possible, in certain cases approximate results for extremal noise behavior can be given [35]. Numerical lower and upper bounds can be calculated [60], though. The calculation of these bounds would represent a significant extension (in complexity) of the algorithms employed in this paper. In the experimental setup of Mountcastle, Poggio, and Werner [52] each stimulus intensity was presented for the duration of (usually) 1 second, and therefore the effect of stimulus history on the current response can be considered minimal (if any).

We found that the capacity-achieving input distribution is discrete with only 6 points of support. It is known that for many channels the information-optimizing input density is supported only at a finite number of points (e.g., it is discrete, composed of Dirac's delta functions), see Fig. 1d. The only non-trivial information channel models with continuous optimizing input densities (we are aware of) are the additive
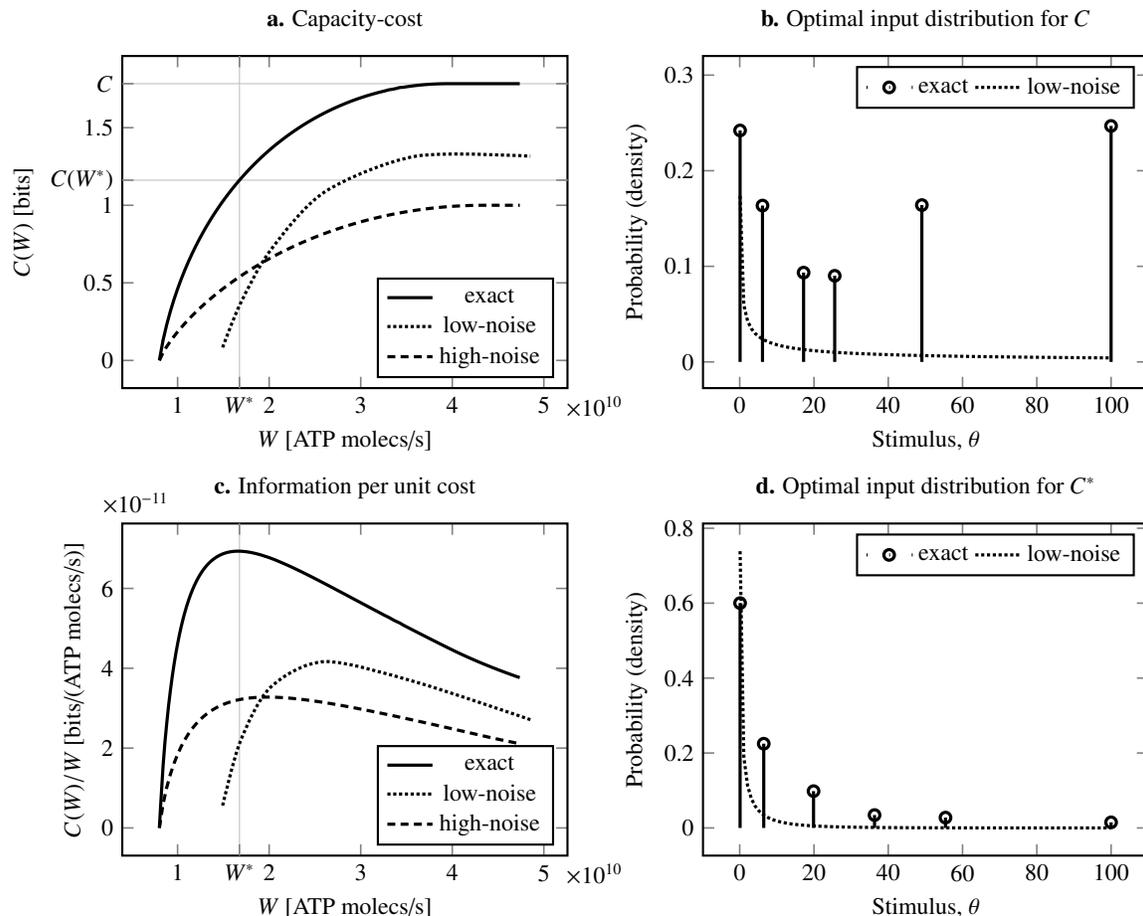
a. Capacity-cost

b. Optimal input distribution for C

c. Information per unit cost

d. Optimal input distribution for C^*

Figure 3. Information-optimality conditions for the empirical stimulus-response model from Fig. 2. The capacity-cost function (**a**) gives the ultimate limit on reliable information transfer (in bits) such, that the metabolic cost of neuronal activity does not exceed the selected value W, its maximum is the information capacity $C = 1.78$ bits. Low- and high-noise approximations are shown for comparison, but both perform poorly, since the selected model falls neither within the high- nor the low-noise category. The optimal balance between information capacity and metabolic cost, C^*, is given by the maximum of the $C(W)/W$ ratio (**c**), which occurs at the optimal average metabolic cost W^*. The optimal input distributions achieving C resp. C^* are shown in panels **b** resp. **d**. In both cases, the optimal input distribution is discrete with 6 points of support. Although each distribution's support set is qualitatively similar, they are clearly not the same. The continuous-valued low-noise density approximation does not describe the exact solution well.

white Gaussian noise linear channel with unbounded input range subject to a variance constraint, and Gallagher's phase channel [23, p. 556]. Unfortunately, we are not aware of an intuitively clear justification of capacity-achieving distribution discreteness. For more complicated channel models it is difficult to prove that the optimal input density is discrete moreover, the proofs of discreteness in particular cases are non-constructive [64, 69]. In this paper we do not attempt to prove the discreteness of the capacity-achieving distribution formally, rather, we conclude that the input distribution is discrete by observing that the majority of probabilities of the input grid points in the numerical algorithm are suppressed to zero.

Detailed comparison of our results with similar studies (and also among those studies) is difficult due to different scenarios

and assumptions. In an extensive numerical study Wiener and Richmond [74] employed experimental data to calculate the capacity of visual V1 complex cells of monkeys. They used both the rate and temporal coding schemes (modulation of firing intensity within the response time-window for each stimulation). More recently, Ikeda and Manton [29] numerically calculated the information capacity of both rate and simple temporal codes in a neuronal model with gamma distribution of interspike intervals. Furthermore, Johnson and Goodman [31] estimated the capacity of Poissonian neurons and their populations. Their results can be interpreted as giving the upper bound on capacity of any neuronal model, where the stimulus is encoded into the point-process intensity; for realistic firing rates the upper bound on capacity is about 2 bits per spike. The assumptions of Johnson and Goodman [31] are
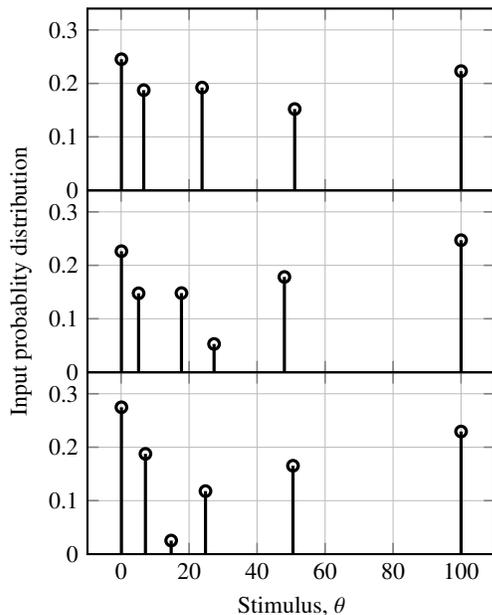
Figure 4. Input distributions that near-achieve capacity for the stimulus-response model from Fig. 2. Mutual information resulting from these distributions falls within 1% from the true capacity. Notice especially the differences for smaller stimulus intensities ($\theta < 50$). See also Fig. 3b.

above (namely time-continuity of the channel communication), the disadvantage is that such a scenario is less amenable to direct or approximate calculations for non-Poisson neuronal models and various stimulus constraints. We conclude, that our results are numerically compatible with those of, e.g., Ikeda and Manton [29], Johnson and Goodman [31], Wiener and Richmond [74], that is the capacity is of the order of 1 bit per channel use.

5. CONCLUSIONS AND FINAL REMARKS

We calculated, both exactly and approximately, the information capacities, capacities per unit cost and their respective optimizing input distributions, for an empirical model of a neuron.

Whether the true information transfer in neurons is close to capacity is, however, a completely different question that we do not seek to answer in this paper.

different and more general than in the rest of the papers cited

[1] Abou-Faycal, I. C., Trott, M. D., and Shamai, S., "The capacity of discrete-time memoryless Rayleigh-fading channels," IEEE Trans. Inf. Theory **47**, 1290–1301 (2001).

[2] Abramowitz, M. and Stegun, I. A., *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1965).

[3] Alexander, R. M., *Optima for Animals* (Princeton Univ. Press, Princeton, NJ, 1996).

[4] Atick, J. J., "Could information theory provide an ecological theory of sensory processing?" Netw. Comput. Neural Syst. **3**, 213–251 (1992).

[5] Attwell, D. and Laughlin, S. B., "An energy budget for signaling in the grey matter of the brain," J. Cereb. Blood Flow Metab. **21**, 1133–1145 (2001).

[6] Baddeley, R., Abbott, L. F., Booth, M. C. A., Sengpiel, F., Freeman, T., Wakeman, E. A., and Rolls, E. T., "Responses of neurons in primary and inferior temporal visual cortices to natural scenes," Proc. Roy. Soc. B **264**, 1775–1783 (1997).

[7] Balasubramanian, V. and Berry, M. J., "A test of metabolically efficient coding in the retina," Netw. Comput. Neural Syst. **13**, 531–552 (2002).

[8] Barlow, H. B., "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, edited by W. Rosenblith (MIT Press, Cambridge, 1961) pp. 217–234.

[9] Bernardo, J. M., "Reference posterior distributions for Bayesian inference," J. Roy. Stat. Soc. B **41**, 113–147 (1979).

[10] Blahut, R., "Computation of channel capacity and rate-distortion functions," IEEE Trans. Inf. Theory **18**, 460–473 (1972).

[11] Boyd, S. and Vandenberghe, L., *Convex Optimization* (Cambridge University Press, 2004).

[12] Brunel, N. and Nadal, J.-P., "Mutual information, Fisher information, and population coding," Neural Comput. **10**, 1731–1757 (1998).

[13] Carandini, M., "Amplification of trial-to-trial response variability by neurons in visual cortex," PLoS Biol. **2**, e264 (2004).

[14] Chan, T. H., Hranilovic, S., and Kschischang, F. R., "Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs," IEEE Trans. Inf. Theory **51**, 2073–2088 (2005).

[15] Chiang, M. and Boyd, S., "Geometric programming duals of channel capacity and rate distortion," IEEE Transactions on Information Theory **50**, 245–258 (2004).

[16] Clarke, B. S. and Barron, A. R., "Information-theoretic asymptotics of Bayes methods," IEEE Trans. Inf. Theory **36**, 453–471 (1990).

[17] Cover, T. M. and Thomas, J. A., *Elements of Information Theory* (John Wiley and Sons, Inc., New York, 1991).

[18] Dauwels, J., "Numerical Computation of the Capacity of Continuous Memoryless Channels," in *Proceedings of the 26th symposium on information theory in the Benelux*, edited by J. Cardinal, N. Cerf, and O. Delgrange (WIC, Brussels, Belgium,

2005) pp. 221–228.

[19] Davis, M., "Capacity and cutoff rate for Poisson-type channels," IEEE Trans. Inf. Theory **26**, 710–715 (1980).

[20] Dimitrov, A. G. and Miller, J. P., "Neural coding and decoding: communication channels and quantization," Netw. Comput. Neural Syst. **12**, 441–472 (2001).

[21] Fairhall, A. L., Lewen, G. D., Bialek, W., and de Ruyter van Steveninck, R. R., "Efficiency and ambiguity in an adaptive neural code," Nature **412**, 787–792 (2001).

[22] Farkhooi, F., Müller, E., and Nawrot, M. P., "Adaptation reduces variability of the neuronal population code," Phys Rev E **83**, 050905 (2011).

[23] Gallager, R. G., *Information Theory and Reliable Communication* (John Wiley and Sons, Inc., New York, USA, 1968).

[24] Gastpar, M., Rimoldi, B., and Vetterli, M., "To code, or not to code: Lossy source-channel communication revisited," IEEE Trans. Inf. Theory **49**, 1147–1158 (2003).

[25] Grant, M. and Boyd, S., "Cvx: Matlab software for disciplined convex programming (web page and software)," (2009), http://stanford.edu/ boyd/cvx, June.

[26] Greenwood, P. E. and Lansky, P., "Optimal signal estimation in neuronal models," Neural Comput. **17**, 2240–2257 (2005).

[27] Gremiaux, A., Nowotny, T., Martinez, D., Lucas, P., and Rospars, J.-P., "Modelling the signal delivered by a population of first-order neurons in a moth olfactory system," Brain Res. **1434**, 123–135 (2012).

[28] Huang, J. and Meyn, S. P., "Characterization and computation of optimal distributions for channel coding," IEEE Trans. Inf. Theory **51**, 2336–2351 (2005).

[29] Ikeda, S. and Manton, J. H., "Capacity of a single spiking neuron channel," Neural Comput. **21**, 1714–1748 (2009).

[30] Jacobson, H., "The informational capacity of the human ear," Science **112**, 143 (1950).

[31] Johnson, D. H. and Goodman, I. N., "Inferring the capacity of the vector Poisson channel with a Bernoulli model," Netw. Comput. Neural Syst. **19**, 13–33 (2008).

[32] Johnson, N., Kotz, S., and Balakrishnan, N., *Continuous Univariate Distributions, Vol. 1* (John Wiley & Sons, New York, 1994).

[33] Kelley, J. E., "The cutting-plane method for solving convex programs," J. Soc. Industrial App. Math. **8**, 703–712 (1960).

[34] Komninakis, C., Vandenberghe, L., and Wesel, R. D., "Capacity of the binomial channel or minimax redundancy for memoryless channels," in *Proc. IEEE International Symposium on Information Theory* (Washington DC, USA, 2001) p. 127.

[35] Kostal, L., "Information capacity in the weak-signal approximation," Phys. Rev. E **82**, 026115 (2010).

[36] Kostal, L., "Approximate information capacity of the perfect integrate-and-fire neuron using the temporal code," Brain Res. **1434**, 136–141 (2012).

[37] Kostal, L. and Lansky, P., "Information transfer with small-amplitude signals," Phys. Rev. E **81**, 050901(R) (2010).

[38] Kostal, L. and Lansky, P., "Information capacity and its approximations under metabolic cost in a simple homogeneous population of neurons," BioSystems **112**, 265–275 (2013).

[39] Kostal, L., Lansky, P., and Rospars, J.-P., "Efficient olfactory coding in the pheromone receptor neuron of a moth," PLoS Comput. Biol. **4**, e1000053 (2008).

[40] Lansky, P., Pokora, O., and Rospars, J.-P., "Classification of stimuli based on stimulus-response curves and their variability," Brain Res. **1225**, 57–66 (2008).

[41] Lansky, P. and Sacerdote, L., "The Ornstein-Uhlenbeck neuronal model with signal-dependent noise," Phys. Lett. A **285**, 132–140 (2001).

[42] Laughlin, S. B., "A simple coding procedure enhances a neuron's information capacity," Z. Naturforsch. **36**, 910–912 (1981).

[43] Laughlin, S. B., de Ruyter van Steveninck, R. R., and Anderson, J. C., "The metabolic cost of neural information," Nat. Neurosci. **1**, 36–41 (1998).

[44] Levy, W. B. and Baxter, R. A., "Energy efficient neural codes," Neural Comput. **8**, 531–543 (1996).

[45] Levy, W. B. and Baxter, R. A., "Energy-efficient neuronal computation via quantal synaptic failures," J. Neurosci. **22**, 4746–4755 (2002).

[46] Machens, C. K., Gollisch, T., Kolesnikova, O., and Herz, A. V. M., "Testing the efficiency of sensory coding with optimal stimulus ensembles," Neuron **47**, 447–456 (2005).

[47] McDonnell, M. D. and Flitney, A. P., "Signal acquisition via polarization modulation in single photon sources," Phys. Rev. E **80**, 060102(R) (2009).

[48] McDonnell, M. D., Mohan, A., Stricker, C., and Ward, L. M., "Input-rate modulation of gamma oscillations is sensitive to network topology, delays and short-term plasticity," Brain Research **1434**, 162–177 (2012).

[49] McDonnell, M. D. and Stocks, N. G., "Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations," Phys. Rev. Lett. **101**, 058103 (2008).

[50] McEliece, R. J., *The Theory of Information and Coding* (Cambridge University Press, Cambdridge, UK, 2002).

[51] Moujahid, A., d'Anjou, A., and Torrealdea, F. J., "Energy and information in Hodgkin-Huxley neurons," Phys. Rev. E **83**, 031912 (2011).

[52] Mountcastle, V. B., Poggio, G. F., and Werner, G., "The relation of thalamic cell response to peripheral stimuli varied over an intensive continuum," J. Neurophysiol. **26**, 807–834 (1963).

[53] Nadal, J.-P. and Bonnasse-Gahot, L., "Perception of categories: From coding efficiency to reaction times," Brain Res. **1434**, 47–61 (2012).

[54] Nikitin, A. P., Stocks, N. G., Morse, R. P., and McDonnell, M. D., "Neural population coding is optimized by discrete tuning curves," Phys. Rev. Lett. **103**, 138101 (2009).

[55] Pawlas, Z., Klebanov, L. B., Prokop, M., and Lansky, P., "Parameters of spike trains observed in a short time window," Neural Comput. **20**, 1325–1343 (2008).

[56] Quastler, H., *Essays on the Use of Information Theory in Biology* (University of Illinois Press, 1953).

[57] Quiroga, R. Q. and Panzeri, S., "Extracting information from neuronal populations: information theory and decoding approaches," Nat. Rev. Neurosci. **10**, 173–185 (2009).

[58] Rieke, F., de Ruyter van Steveninck, R., Warland, D., and Bialek, W., *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, 1997).

[59] Rissanen, J. J., "Fisher information and stochastic complexity," IEEE Trans. Inf. Theory **42**, 40–47 (1996).

[60] Sadeghi, P., Vontobel, P. O., and Shams, R., "Optimization of Information Rate Upper and Lower Bounds for Channels With Memory," IEEE Trans. Inf. Theory **55**, 663–688 (2009).

[61] Schreiber, S., Machens, C. K., Herz, A. V. M., and Laughlin, S. B., "Energy-efficient coding with discrete stochastic events," Neural Comput. **14**, 1323–1346 (2002).

[62] Schroeder, D. J., *Astronomical optics* (Academic Press, San Diego, USA, 1999).

[63] Shadlen, M. N. and Newsome, W. T., "The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding," J. Neurosci. **18**, 3870–3896 (1998).

[64] Smith, J. G., "The information capacity of amplitude-and variance-constrained sclar gaussian channels," Inform. Control

18, 203–219 (1971).

[65] Stein, R. B., "The information capacity of nerve cells using a frequency code," Biophys. J. **7**, 797–826 (1967).

[66] Stein, R. B., Gossen, E. R., and Jones, K. E., "Neuronal variability: noise or part of the signal?" Nat. Rev. Neurosci. **6**, 389–397 (2005).

[67] de Ruyter van Steveninck, R. R. and Laughlin, S. B., "The rate of information transfer at graded-potential synapses," Nature **379**, 642–644 (1996).

[68] Suksompong, P. and Berger, T., "Capacity analysis for integrate-and-fire neurons with descending action potential thresholds," IEEE Trans. Inf. Theory **56**, 838–851 (2010).

[69] Tchamkerten, A., "On the discreteness of capacity-achieving distributions," IEEE Trans. Inf. Theory **50**, 2773–2778 (2004).

[70] Tuckwell, H. C., *Introduction to Theoretical Neurobiology, Vol. 2* (Cambridge University Press, New York, 1988).

[71] Verdu, S., "On channel capacity per unit cost," IEEE Trans. Inf. Theory **36**, 1019–1030 (1990).

[72] Wainrib, G., Thieullen, M., and Pakdaman, K., "Intrinsic variability of latency to first-spike," Biol. Cyb. **103**, 43–56 (2010).

[73] Weinstock, R., *Calculus of variations* (Dover, New York, 1974).

[74] Wiener, M. C. and Richmond, B. J., "Using response models to estimate channel capacity for neuronal classification of stationary visual stimuli using temporal coding," J. Neurophysiol. **82**, 2861–2875 (1999).