Preprint Accepted for publication in *BioSystems* (2013) **112**: 265–275

Information capacity and its approximations under metabolic cost in a simple homogeneous population of neurons

Lubomir Kostal[a)] and Petr Lansky
Institute of Physiology of the Czech Academy of Sciences, Videnska 1083, 14220 Prague 4, Czech Republic

We calculate and analyze the information capacity-achieving conditions and their approximations in a simple neuronal system. The input-output properties of individual neurons are described by an empirical stimulus-response relationship and the metabolic cost of neuronal activity is taken into account. The exact (numerical) results are compared with a popular "low-noise" approximation method which employs the concepts of parameter estimation theory. We show, that the approximate method gives reliable results only in the case of significantly low response variability. By employing specialized numerical procedures we demonstrate, that optimal information transfer can be near-achieved by a number of different input distributions. It implies that the precise structure of the capacity-achieving input is of lesser importance than the value of capacity. Finally, we illustrate on an example that an innocuously looking stimulus-response relationship may lead to a problematic interpretation of the obtained Fisher information values.

Keywords: Information capacity, Metabolic cost, Neuronal population

1. INTRODUCTION

The *information theory* (Gallager, 1968) provides an attractive methodology for the theoretical approach to the problem of information processing in neuronal systems, see, e.g., works of Brunel and Nadal (1998); Borst and Theunissen (1999); McDonnell, Ikeda, and Manton (2011); Stein (1967); Fairhall *et al.* (2001); Farkhooi, Müller, and Nawrot (2011); Wiener and Richmond (1999); Rieke *et al.* (1997); Kostal (2012) among many others. Of particular interest are usually the optimality conditions under which the information between stimuli and responses is maximized (Atick, 1992; Bialek and Owen, 1990; Laughlin, 1981, 1996; McDonnell and Stocks, 2008; Kostal, Lansky, and Rospars, 2008). The motivation stems from the *efficient-coding* hypothesis (Barlow, 1961; Atick, 1992), which states that the sensory neurons are adapted, in the information optimality sense, to the statistical properties of the signals to which they are naturally exposed.

Although it is universally accepted that neurons communicate using series of action potentials (spike trains) via chemical and electrical synapses, the exact structure of *neuronal code* is not yet fully resolved (Shadlen and Newsome, 1994; Stein, Gossen, and Jones, 2005). For the illustration we adopt the classical *frequency coding* scheme, where the information sent along axon is encoded in the number of spikes per observation time window (the firing rate) (Adrian, 1928). In most sensory systems, the firing rate increases, generally nonlinearly, with increasing stimulus intensity (Kandel, Schwartz, and Jessel, 1991). Any information possibly encoded in the temporal structure of the spike train is, however, ignored.

Characterization of the input-output properties of neurons, as well as of the neuronal models, is commonly done by so-called frequency (input-output) transfer functions in which the output is plotted against the strength of the input signal (e.g., stimulus intensity) (Lansky, Rodriguez, and Sacerdote, 2004; Carandini, 2004). The output is usually the frequency of firing, but it can be the level of any variable of interest, e.g., first-spike latency, channel conductance, receptor potential or its peak amplitude (Kostal, Lansky, and Rospars, 2008; Laughlin, 1981; Gremiaux *et al.*, 2012; McDonnell *et al.*, 2012). The transfer function is usually presented as a single curve, relating the mean (or the average of experimental measurements) response to each stimulus intensity. Since the response firing frequency often varies, apparently randomly at least to some degree, both within and across trials (Shadlen and Newsome, 1998; Stein, Gossen, and Jones, 2005), the curve is sometimes accompanied by standard deviations (Borst and Theunissen, 1999; Mountcastle, Poggio, and Werner, 1963). However, The complete descriptor of the input-output relationship under this scenario would be the full probability distribution of responses for each possible stimulus intensity.

Neurons also use significant amount of energy for the spiking activity, thus energy usage should be coupled to considerations about the efficiency of neuronal information transfer (Laughlin, de Ruyter van Steveninck, and Anderson, 1998; Levy and Baxter, 1996, 2002; Kostal, Lansky, and McDonnell, 2013). Attwell and Laughlin (2001) used anatomic and physiologic data to analyze the metabolic cost of different components of excitatory signaling, their results are employed in this paper when considering optimal information vs. metabolic cost ratios.

The goal of this paper is to extend the effort started in Kostal, Lansky, and McDonnell (2013). We analyze single neurons and their groups, where the group activity is the sum of individual neuronal activities. In particular, we identify the regions of validity and assess the precision of the low-noise approximate expressions (based on the concept of Fisher information) in dependence on the number of neurons. By employing specialized numerical procedures we show, that optimal information transfer can be near-achieved by a number

a)E-mail: kostal@biomed.cas.cz

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. Information capacity and capacity-cost function

As mentioned in Introduction, the complete input-output properties of the neuron (under the assumption of frequency coding) are described by the probability distribution of responses for each possible stimulus value. For the purpose of this paper we thus identify the conditional probability density function (p.d.f.) $f(r|\theta)$ with the *statistical neuronal model*, where θ is the stimulus strength (the "intensity" or just the "label" of stimulus feature), and r denotes the response firing rate. Probability distributions employed in this paper may be continuous or discrete, the specific type follows from the text.

Information about some particular stimulus, θ, based on observing the response r, can be defined as (Gallager, 1968)

$$I(\theta;r) = \ln \frac{\phi(\theta|r)}{p(\theta)} = \ln \frac{f(r|\theta)}{f(r)}, \tag{1}$$

where $p(\theta)$ is the p.d.f. over the stimulus ensemble and $\phi(\theta|r)$ is the p.d.f. describing the relative frequencies of possible stimuli intensities conditioned that response r was observed. The definition above can be justified intuitively, since the most informative stimulus-response (S-R) pairs are those, in which the response r can be used to "identify" θ with high *a posteriori* to *a priori* probability ratio. Since the p.d.f. $\phi(\theta|r)$ is usually unknown, the Bayes' law is used to obtain the last equality in Eq. (1), where $f(r) = \int_\Theta f(r|\theta)p(\theta)\,d\theta$.

The mean value of $I(\theta;r)$ taken over all S-R pairs in the stimulus ensemble Θ (described by the p.d.f. $p(\theta)$) and the response ensemble $R \sim f(r)$ (described by the p.d.f. $f(r)$), is denoted as the *mutual information*, $I(\Theta;R)$,

$$I(\Theta;R) = \int_\Theta \int_R I(\theta;r) f(r|\theta)p(\theta)\,dr\,d\theta. \tag{2}$$

The maximal information that can be transferred is called the information *capacity*,

$$C = \max_{p(\theta)} I(\Theta;R), \tag{3}$$

where the maximum is taken over all possible input distributions.

If the noise in information transfer is substantially low, there exists a lower bound, $I_{\text{low}}(\Theta;R)$, on the mutual information from Eq. (2), employed e.g., in Bernardo (1979); Brunel and Nadal (1998); Clarke and Barron (1990); Kostal, Lansky, and McDonnell (2013); McDonnell and Stocks (2008). One neat feature of the low-noise approximation lies in the fact, that it relates the concept of mutual information with the concept of Fisher information known from the estimation theory (Lehmann and Casella, 1998; Kay, 1993). We provide a brief informal derivation as follows; see e.g., Brunel and Nadal (1998) for details. The true stimulus intensity, θ, is identified (i.e., estimated) from the responses based on an estimator, $\hat{\theta}(r)$. The data-processing inequality (McEliece, 2002) states, that mutual information between the ensemble of stimuli, Θ, and the ensemble of their estimates, $\hat{\Theta}$, cannot be greater that mutual information between stimuli and responses,

$$I(\Theta;\hat{\Theta}) \le I(\Theta;R). \tag{4}$$

Mutual information, $I(\Theta,\hat{\Theta})$, can be equivalently written as (Gallager, 1968)

$$I(\Theta;\hat{\Theta}) = h(\hat{\Theta}) - \int_\Theta p(\theta)h(\hat{\Theta}|\theta)\,d\theta, \tag{5}$$

where $h(\hat{\Theta})$ is the differential entropy (informally, the uncertainty; see Gallager (1968) for details) of the unconditional distribution of the estimator ensemble, and $h(\hat{\Theta}|\theta)$ is the "uncertainty" of the estimator ensemble given some particular stimulus intensity θ. If the estimator $\hat{\theta}(r)$ is not biased (its mean value corresponds to θ), then there exists a lower (Cramer-Rao) bound on the mean squared error of the estimator (Pitman, 1979),

$$\int_R (\hat{\theta}(r) - \theta)^2 f(r|\theta)\,dr \ge \frac{1}{J(\theta|R)}, \tag{6}$$

where $J(\theta|R)$ is the Fisher information,

$$J(\theta|R) = \int_R \left[\frac{\partial \ln f(r|\theta)}{\partial\theta}\right]^2 f(r|\theta)\,dr. \tag{7}$$

Next, assume that an efficient estimator exists, i.e., the inequality in Eq. (6) becomes equality. Since the maximum possible value of entropy for a given variance σ^2 is known to be $\ln\sqrt{2\pi e\sigma^2}$ (Gallager, 1968), it follows by employing Eq. (6) that

$$h(\hat{\Theta}|\theta) \le \ln\sqrt{\frac{2\pi e}{J(\theta|R)}}. \tag{8}$$

In the low-noise limit the estimator is peaked about its true value, so it holds $h(\hat{\Theta}) \doteq h(\Theta) = \int_\Theta p(\theta)\ln p(\theta)\,d\theta$, and thus Eqns. (4) and (8) can be combined to give $I(\Theta;R) \ge I_{\text{low}}(\Theta;R)$,

$$I_{\text{low}}(\Theta;R) = -\int_\Theta p(\theta)\ln\left[p(\theta)\sqrt{\frac{2\pi e}{J(\theta|R)}}\right]d\theta, \tag{9}$$

where the approximation becomes tight as the response variability decreases (Brunel and Nadal, 1998; Rissanen, 1996).

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

$$C_{\text{low}} = \ln\frac{\int_\Theta \sqrt{J(\theta|R)}\,d\theta}{\sqrt{2\pi e}}, \tag{10}$$

with optimizing p.d.f. $p(\theta) \propto \sqrt{J(\theta|R)}$ (also known as the Jeffrey's prior (Bernardo, 1979)). Note, that the Euler-Lagrange

method leading to Eq. (10) requires $p(\theta)$ to be differentiable, confined to finite stimulus range and also strictly positive over this range.

Denote the metabolic cost of neuronal response r as $v(r)$. The actual form of $v(r)$ depends on the response character considered (firing rate, latency, ...). The cost of neuronal activity evoked by stimulus θ is thus

$$w(\theta) = \int_R f(r|\theta)v(r)\,dr. \tag{11}$$

The *mean* metabolic cost, W_p, associated with the stimulus ensemble described by p.d.f. $p(\theta)$ is

$$W_p = \int_\Theta w(\theta)p(\theta)\,d\theta. \tag{12}$$

The *capacity-cost* function (McEliece, 2002), $C(W)$, gives the information capacity under the additional constraint that the average metabolic cost W_p does not exceed some selected value W,

$$C(W) = \max_{p(\theta),W_p\le W} I(\Theta;R). \tag{13}$$

In many cases of interest the (unconstrained) capacity C is achieved at some finite value of W, which we denote as $W^\dagger$, i.e., $C = C(W)$ for all $W \ge W^\dagger$.

The optimal balance between the information capacity and metabolic cost is given by the *capacity per unit cost* (Verdu, 1990),

$$C^* = \max_W \frac{C(W)}{W}. \tag{14}$$

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. (14), if the solution exists. Applying the Euler-Lagrange method on Eq. (9), while taking the metabolic cost into account, gives the following low-noise approximation to the optimal p.d.f.

$$p(\theta) = \frac{1}{Z}\sqrt{\frac{J(\theta|R)}{2\pi e}}\exp[\lambda_W w(\theta)], \tag{15}$$

where Z is the normalization factor and λ_W is the Lagrange multiplier associated with the average metabolic cost. Detailed derivation of Eq. (15) is presented in Kostal, Lansky, and McDonnell (2013). Both Z and λ_W must be found numerically. The approximation to the capacity-cost function evaluated at W is obtained by maximizing C_{low} for all $W_p \le W$.

Since Eqns. (3), (13) and (14) can rarely be solved in a closed form, we present an efficient exact (numerical) optimization procedure (Huang and Meyn, 2005) usable for all general cases of practical interest (see details in the Appendix). The low-noise approximation requires the mean $E(R|\theta)$ of the S-R model $f(r|\theta)$ to be monotonic as a function of θ, plus some minor technical conditions on the character of the response variability (Brunel and Nadal, 1998; Rissanen, 1996). However, the original Shannon's theory was proposed under rather broad assumptions (Gallager, 1968), and thus the presented numerical methods are valid for a much broader class of S-R relationships (sigmoidal or not), see (Huang and Meyn, 2005) for some nonrestrictive technical conditions.

2.2. Stimulus-response characteristics

In this paper we employ an empirical stochastic S-R relationship described in Lansky, Pokora, and Rospars (2008). The average, $m(\theta)$, and the standard deviation, $\sigma(\theta)$, of the evoked firing rate are given by

$$m(\theta) = \frac{49.5}{1+\exp(3.5-\theta)}, \tag{16}$$

$$\sigma(\theta) = 8.75\exp\left[-\left(\frac{\theta-6.5}{4.8}\right)^2\right] + 29.9\exp\left[-\left(\frac{\theta-6.1}{1.1}\right)^2\right], \tag{17}$$

with θ in range $[0.6, 4.6]$.

It follows from Eq. (2), that in order to calculate the information-theoretic quantities, the full form of the distribution $f(r|\theta)$ must be known, while the empirical model describes only the mean value, Eq. (16), and the standard deviation, Eq. (17). To proceed on, we select the *gamma distribution* as a suitable model for probabilistic description of neuronal firing rates given stimulus intensity (Ikeda and Manton, 2009; Pawlas *et al.*, 2008),

$$f(r|\theta) = r^{k(\theta)-1}\frac{\exp(-r/s(\theta))}{\Gamma[k(\theta)]s(\theta)^{k(\theta)}}, \tag{18}$$

where $\Gamma(x)$ is the gamma function, the shape, $k(\theta)$, and scale, $s(\theta)$, parameters are related to $m(\theta)$ and $\sigma(\theta)$ as

$$k(\theta) = \frac{r^2(\theta)}{\sigma^2(\theta)} \tag{19}$$

$$s(\theta) = \frac{\sigma^2(\theta)}{m(\theta)}. \tag{20}$$

Eqns. (19) and (20) follow from the basic relationship between the scale/shape parameters and the mean/variance of the gamma distribution (Johnson, Kotz, and Balakrishnan, 1994). The S-R relationship is shown in Fig. 1a.

For the purpose of this paper we assume that the metabolic cost of neuronal signalling is proportional to the firing rate, as supported by both theoretical (e.g., Balasubramanian and Berry (2002)) and experimental studies (e.g., Attwell and Laughlin (2001)). In other words we have $v(r) = \kappa r$, where the constant of proportionality, $\kappa = 7.1\times10^8$ ATP molecules, describes the metabolic cost of a single action potential (Attwell and Laughlin, 2001). Eq. (11) is thus reduced to

$$w(\theta) = \kappa m(\theta). \tag{21}$$

We also consider a group n independent neurons, each following the S-R model described by Eq. (18). We assume that

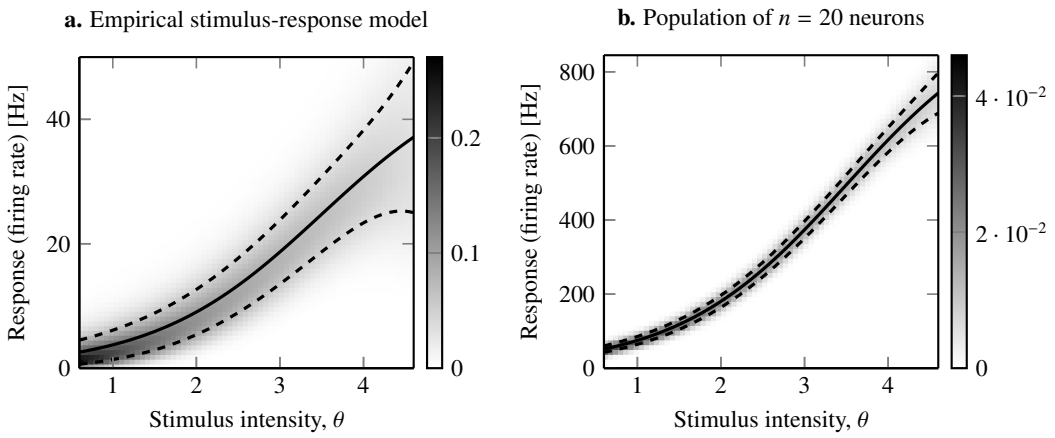

Figure 1. Stimulus-response characteristics. Empirical model was fitted by Lansky, Pokora, and Rospars (2008). Mean stimulus-response curve (solid) and standard deviation (dashed) is indicated. The probability distribution of responses given each stimulus is assumed to follow the gamma distribution, its values are indicated by shades of gray (**a.**). Analogous stimulus-response characteristics for a group of $n = 20$ neurons is shown in **b.**

at any time instant all neurons are subject to the same stimulus value, θ, and the response, r, of the population is the sum of individual responses (firing rates). Such setup results in a relative decrease of the amount of noise in the population response with increasing size of the population, since the mean response of the population is $nm(\theta)$ while its standard deviation $\sqrt{n}\sigma(\theta)$, where $m(\theta)$ and $\sigma(\theta)$ are described by Eqns. (16) and (17). We examine in detail the case of $n = 20$ neurons. The S-R curve and its standard deviation are shown in Fig. 1b, suggesting that the low-noise approximation might be applicable. The metabolic cost associated with stimulus θ is $nw(\theta)$. The full model of the population response, given by p.d.f. $f(r|\theta)$, can be easily constructed since the sum of n independent and identically gamma-distributed random variables with shape parameter k and scale parameter s is again a gamma-distributed random variable with the same scale parameter and shape parameter equal to nk (Johnson, Kotz, and Balakrishnan, 1994).

3. RESULTS AND DISCUSSION

The theoretical methods described in the previous section are now applied in this section on the empirical neuronal model described by Eqns. (16)–(20).

3.1. Single neuron

Fig. 2a compares the optimal input distributions for achieving the capacity, C, and capacity per unit cost, C^*. 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 5 points with non-zero probability. The distributions differ in the exact stimulus intensity values of non-zero probability (except for $\theta_{\min}$ and $\theta_{\max}$ which are stable, although $\Pr\{\theta_{\max}\} \doteq 0.001$ for the C^*-achieving input distribution). During the numerical calculations we observed, 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 (2001) and related theoretical considerations found in Wu and Verdu (2010)). Simultaneously, the low-noise approximation (Fig. 2b) does not resemble the exact input distribution, confirmed by the approximate capacity value, $C_{\text{low}} = 0.88$ bits. It is worth noting, that employing uniform input distribution results in $I(\Theta; R) = 1.25$ bits, substantially higher than C_{low}. Obviously, the studied empirical S-R relationship does not fall within the low-noise category. In Fig. 3 the ratio $C(W)/W$, is shown for both exact (numerical) and approximate methods and the optimal cost, $W^* = 0.42 \times 10^{10}$, is indicated. The value of $C(W^*) = 0.83$ bits offers the best balance between the ultimate limit on reliable information transfer, versus the induced metabolic cost of neuronal activity.

3.2. A small group of independent neurons

The standard deviation (compared to the mean response) is not negligible for the empirical S-R model considered, see Fig. 1a, thus we are obviously straining the low-noise approximation beyond its range of validity. A natural way to reduce the amount of "noise" is to consider the summed response of a group of independent neurons. Fig. 4 shows the information-optimality conditions. The capacity is $C = 3.12$ bits, first achieved at $W^\dagger = 23.12 \times 10^{10}$ ATP molecs/s, capacity at optimal cost is $C(W^*) = 1.81$ bits, with $W^* = 6.67 \times 10^{10}$ ATP molecs/s. The optimal balance thus states that by operating at the 29 % of the required metabolic cost to achieve capacity it is possible to achieve 58 % of the capacity. The low-

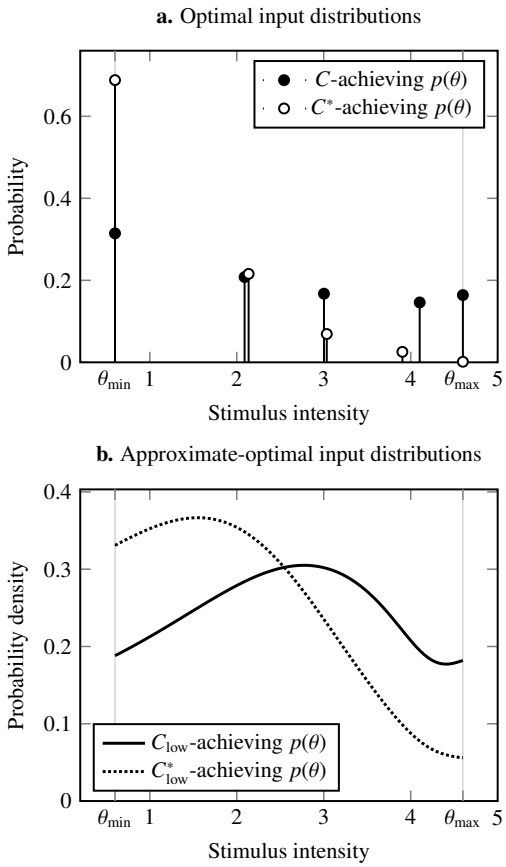

Figure 2. Optimal input distributions achieving the capacity, C, and capacity per unit cost, C^* for the empirical stimulus-response model from Fig. 1a. The exact optimal input distributions achieving C resp. C^* are shown in **a**), the low-noise approximation in **b**). The exact optimal input distributions are both discrete with 5 points of support. The continuous-valued low-noise p.d.f. approximation does not describe the exact solution well.

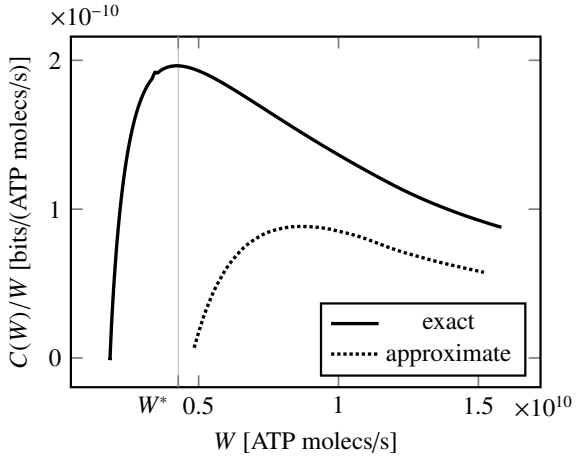

Figure 3. Ratio of the capacity-cost function to the metabolic cost of neuronal activity for the empirical S-R model from Fig. 1a. The optimal balance between information capacity and metabolic cost, C^*, is given by the maximum of the $C(W)/W$ ratio, which occurs at the optimal average metabolic cost $W^* = 0.42 \times 10^{10}$, yielding $C(W^*) = 0.83$ bits. Low-noise approximation is shown for comparison, but performs poorly, since the selected model does not have sufficiently small response variability.

noise approximation to the capacity-cost function becomes more precise (Fig. 4c, compare with Fig. 3) with $C_{\text{low}} = 2.97$ bits, while the approximation to the capacity at optimal cost, $C(W^*)$, is underestimated and the optimal cost overestimated. Fig. 4b resp. 4d compares the exact and low-noise approximation to the optimal input probability distributions. We compare the cumulative distribution functions in order to better visualize the degree of correspondence between the approximate and exact result. Although the exact input distributions are discrete and the approximate ones are continuous, we observe an increase in their "similarity" when compared to the single-neuron case (Fig. 2).

The employed numerical algorithm allows to control the precision of the result, therefore it is possible to compare input distributions that near-achieve capacity. For an input distribution resulting in mutual information $I = I(\Theta; R)$ we define the relative precision as $1 - I/C$. Fig. 5 shows three different input distributions which achieve the capacity within 1 % precision (*cf.* Fig. 4b). The differences among the distributions are apparent in the number of points of support, in stimulus intensities with non-zero probabilities, and the assigned probabilities. We conclude, that even for S-R relationships with very low response variability, the exact form of capacity-achieving input distribution is of limited importance.

3.3. Arbitrary number of neurons

The dependence of the capacity on the number of neurons, n, in the population is shown in Fig. 6 together with its low-noise approximation. Although the low-noise approximation curve almost passes through the exact solutions, its values for each n are slightly shifted. The increase of both C and C_{low} is approximately logarithmic with n, especially for large n. The relative error of the approximation, $1 - C_{\text{low}}/C$, starts at 37 % for $n = 1$ and decreases fast to 5 % at $n = 22$, continuing in a slow decrease to 1 % at $n = 274$. We conclude, that the low-noise approximation is valid for models with indeed low signal-to-noise ratio (*cf.* Fig. 1b), similar conclusions can be drawn from examples shown in Kostal (2010).

Capacity per unit cost and the optimal cost in dependence on the population size are shown in Fig. 7. The value of C^* decreases with n, Fig. 7a. From this point of view, single-neuron case is the most efficient one (although such a conclusion depends on the model and especially on the definition of the metabolic cost). Also note that the capacity generally tends to infinity with growing number of neurons. The low-noise approximation performs worse in determining C^* than in determining C. Fig. 7b shows the optimal cost normalized per neuron, W^*/n, since the plot of W^* vs. n would be dominated

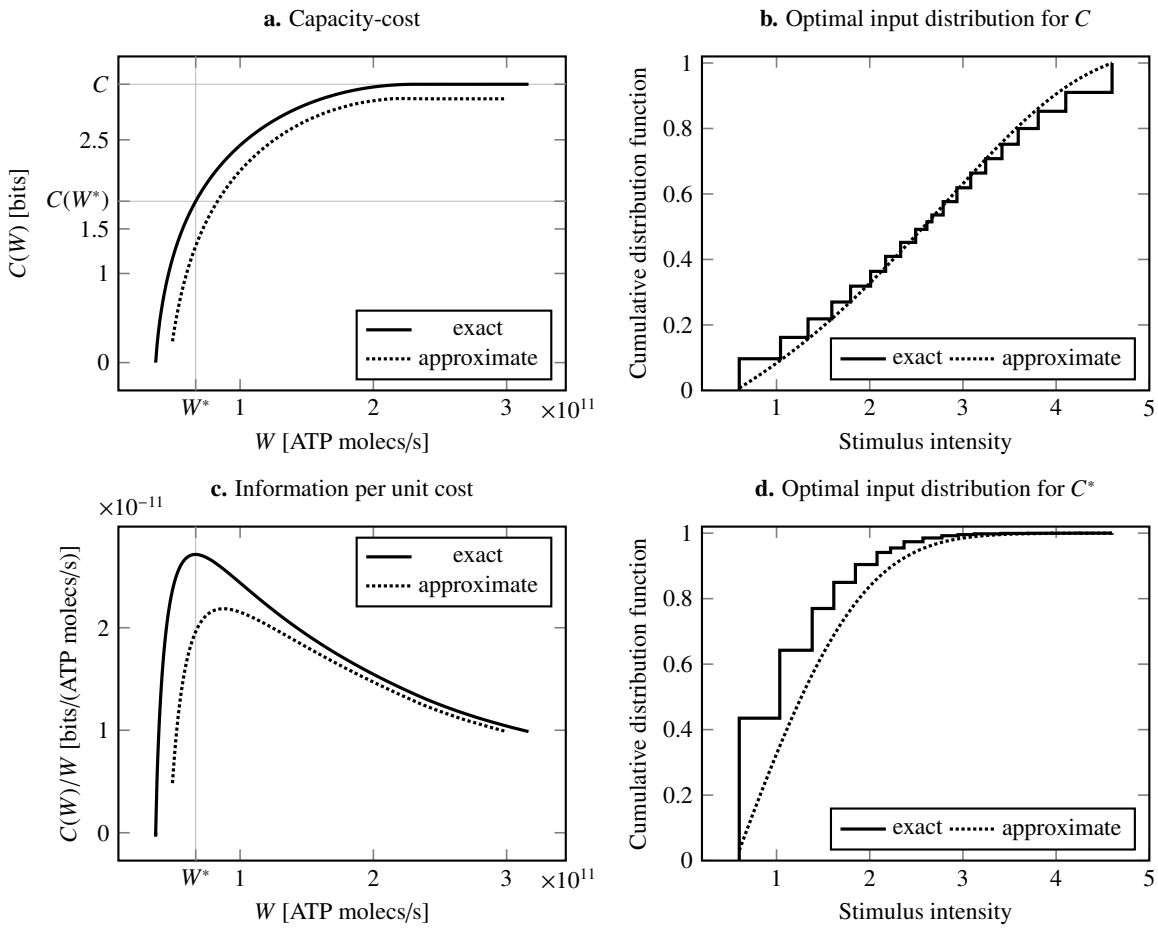

Figure 4. Information-optimality conditions for the homogeneous population of 20 neurons. Each neuron follows the empirical stimulus-response model from Fig. 1a. The low-noise approximation to the capacity-cost function (**a.**) performs better than in the single-neuron case (i.e., compare Fig. 3 with panel **c.** here). The comparison between exact and approximate optimal input distributions achieving C resp. C^* are shown in panels **b** resp. **d** by employing the cumulative distribution functions. Although the exact input distributions are discrete (with 20 (**b.**) resp. 17 (**d.**) points of support) while the approximate ones are continuous, we observe a similarity in the overall shapes of cumulative distributions. Even better correspondence can be achieved by considering a larger population (see Fig. 10).

by the linear contribution of n. We see, that the average optimal cost per neuron also decreases with increasing n, the low-noise approximation overestimates the true values. Fig. 7c shows the increase of the capacity at optimal cost, $C(W^*)$, with increasing population size (the capacity, C, is added for comparison).

The ratio $C(W^*)/C \doteq 0.6$ is almost independent on n for the population sizes examined ($n = 1,\ldots,300$), while the ratio of optimal cost, W^*, to the capacity-achieving cost, $W^\dagger$, decreases from 0.37 for a single neuron, to 0.25 for $n = 300$ (see Fig. 8a where n increases along the curve from right to left). The ratio $C(W^*)/C$ shows a minimum around medium-sized populations ($n = 20,\ldots,70$), however, the effect is small. The capacity and capacity at unit cost per neuron are monotonicaly decreasing in Fig. 8b. If these ratios are taken as indicators of optimal population size then it implies that the most efficient scenario is the single neuron case. Adding other neurons increases the capacity (Fig. 7c) but not sufficiently fast. The same conclusion can be reached for the capacity per unit cost per neuron, since C^* is already decreasing with n (Fig. 7a).

The variability of the low-noise approximation of optimal input distributions, for both capacity and capacity per unit cost, are shown in Fig. 9 for selected sizes of neuronal population. While the approximations to the capacity-achieving input p.d.f. are almost independent of the population size, the capacity per unit cost-achieving p.d.f. approximations depend on n more substantially.

Finally, we "imitate" the conditions required for the derivation of Eq. (15), namely the assumption that $p(\theta)$ is nonzero over the stimulus range. Thus we reduce the size, k, of the input grid in the numerical procedure while maintaining a very low response variability ($n = 274$ neurons). The goal here is to investigate the correspondence between exact and approxi-

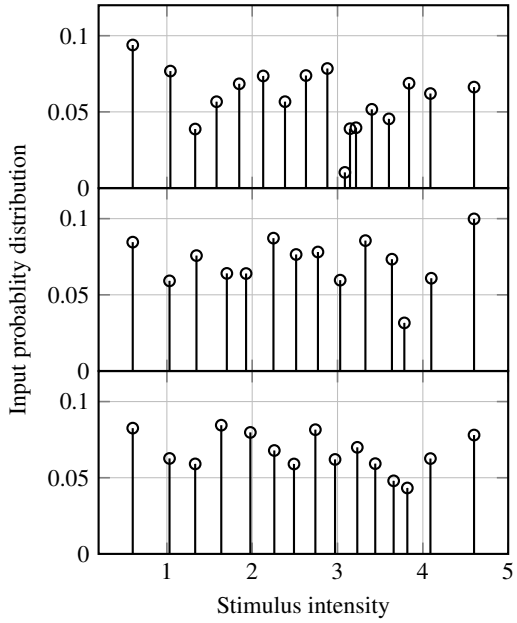

Figure 5. Examples of input distributions that near-achieve capacity for a population of $n = 20$ neurons. Mutual information resulting from these distributions falls within 1 % from the true capacity, *cf.* Fig. 4b. The distributions differ in the number and stimulus intensities with non-zero probabilities, and the assigned probabilities. The choice of different input distributions which near-achieve capacity within a given precision increases with the signal-to-noise ratio (the number of neurons in the population).

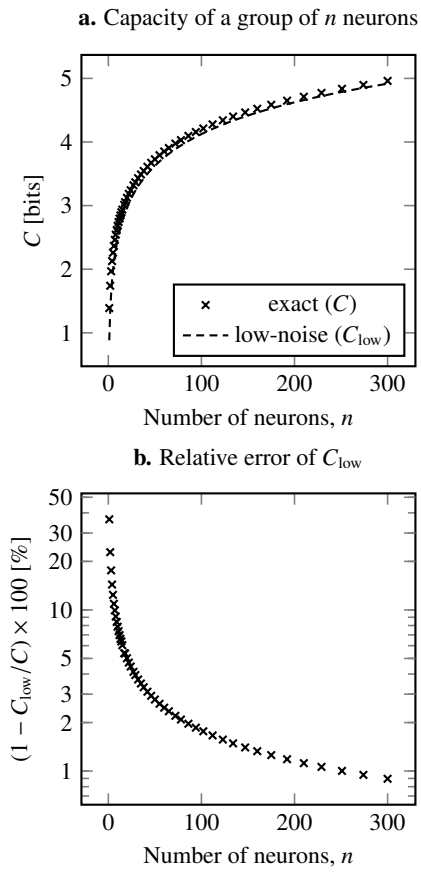

Figure 6. Comparison of exact and approximate capacity results. Both the capacity and its approximation grow approximately logarithmically with increasing number of neurons (**a**). The relative error (**b**) of the approximation starts at 37 % for a single neuron and drops fast to less than 5 % for 22 neurons (*cf.* Fig. 1b and 4). The relative error then decreases slowly, being less than 1 % for a group of 274 neurons.

mate optimal input distributions, since we expect that most of the stimulus intensities in the undersampled grid will be assigned non-zero probabilities. This procedure requires a bit of fine-tuning, though, since too severe undersampling would result in nearly uniform probabilities. The results are shown in Fig. 10 for a sufficiently high signal-to-noise ratio scenario (population size $n = 274$) and the input grid containing 50 and 70 equidistant points. The numerically obtained values of probabilities at the undersampled grid points, $\Pr\{\theta_i\}$, were used to calculate "histogram" for easier comparison with the low-noise input p.d.f. approximation. (The histogram values were calculated as $H(\theta_i) = \Pr\{\theta_i\}/\Delta$ for $i = 1,\ldots,k$, where Δ is the distance between grid points.) The low-noise approximation matches the undersampled capacity-achieving input distribution, with mild underestimation at the end points (Fig. 10a). It is worth noting, that the absolute difference between the true capacity and "undersampled" capacity is only 0.02 bits, which further illustrates the wide choice of near-capacity achieving input distributions in the high signal-to-noise ratio situation. The underestimation of the small stimulus intensity probabilities is more apparent for the capacity per unit cost calculations (Fig. 10b), but overall the agreement between the continuous p.d.f. and the undersampled approximation is good.

3.4. Non-regular Fisher information example

In this section we construct a simple S-R relationship to demonstrate a "counter-intuitive" behavior of the low-noise approximation, for an apparently (mathematically) well-behaving mixture model. Analogous models are common when describing, e.g., neuronal coding of odorant mixtures (Rospars *et al.*, 2008) or the spiking activity of bursting neurons (Bhumbra, Inyushkin, and Dyball, 2004; DeBusk *et al.*, 1997; Tuckwell, 1988), with one distribution responsible for the "slow" firing and the other for the "fast" firing regime. In our model we do not aim for a realistic neuronal description, thus we choose a mixture of two Gaussians for convenience (the response in Eq. (22) is thus not the firing rate). It is possible to construct more realistic variant of Eq. (22) at the expense of cumbersome notation, but with no impact on our main conclusion.

Consider the stimulus intensity range $0 \le \theta \le 1$ (in arbi-

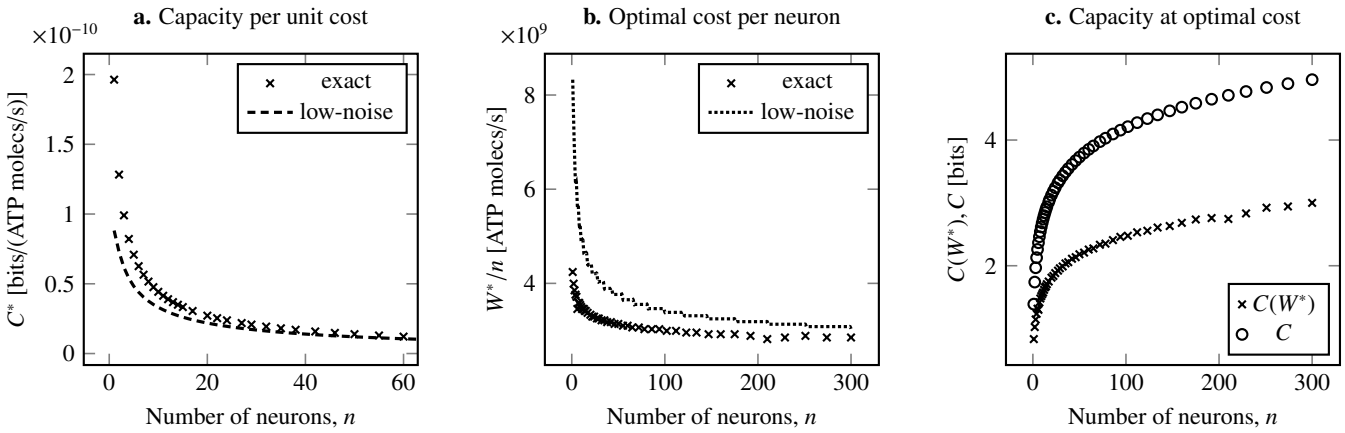

Figure 7. Comparison of exact and approximate capacity per unit cost in dependence on population size. The maximum information per unit cost, C^*, decreases with the size of neuronal population (**a**), making the single-neuron case most "efficient". The low-noise approximation is less precise than in the case of capacity (Fig. 6). The optimal metabolic cost W^* is plotted normalized per neuron (**b**) to show its relative decrease with population size, otherwise its dependence on n would be almost linear. The low-noise approximation severely overestimates W^*/n. The capacity at optimal cost, $C(W^*)$, is shown together with C in (**c**).

trary units) and the following S-R relationship

$$f(r|\theta) = (1-\theta^2)g(r-\theta^2;\sigma_1) + \theta^2 g(r-\theta^2;\sigma_2), \quad (22)$$

where $g(r;\sigma)$ is the probability density of a Gaussian distribution with zero mean and variance σ^2,

$$g(r;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{r^2}{2\sigma^2}\Big). \quad (23)$$

The model in Eq. (22) describes a smooth change of the response variability distribution from one Gaussian to the other with increasing stimulation, along the mean S-R curve $r(\theta) = \theta^2$.

Fisher information defined by Eq. (7) cannot be found in a closed form for the model in Eq. (22). It is, however, possible to evaluate $J(\theta|R)$ at the extremal points of the stimulus range by noting that $f(r|\theta)$ is differentiable, thus we can interchange the limits $\theta \to 0$ and $\theta \to 1$ with integration in Eq. (7). After some algebraic manipulation, and by employing the substitution $r \leftarrow (r-1)$ we obtain (the prime denotes derivative with respect to r)

$$J(0|R) = 0, \quad (24)$$

$$J(1|R) = \int_R \Big\{4\frac{[g'(r;\sigma_2)]^2}{g(r;\sigma_2)} + 4g(r;\sigma_2) - 8g(r;\sigma_1) - \\ - 8g'(r;\sigma_2) + 8\frac{g'(r;\sigma_2)g(r;\sigma_1)}{g(r;\sigma_2)} + 4\frac{g(r;\sigma_1)}{g(r;\sigma_2)}\Big\}\, dr. \quad (25)$$

By employing that $\int_R g(r;\sigma) = 1$, $\int_R g'(r;\sigma) = 0$ and the Gaussian character of $g(r;\sigma)$, the Eq. (25) can be further simplified to yield

$$J(1|R) = \frac{4}{\sigma_2^2} - 4 + 4\int_{-\infty}^{\infty} \frac{\sigma_2}{\sigma_1^2\sqrt{2\pi}} \exp\Big[\Big(\frac{1}{2\sigma_2^2} - \frac{1}{\sigma_1^2}\Big)r^2\Big]\, dr. \quad (26)$$

The integral in Eq. (26) is divergent for $\sigma_1 \geq \sqrt{2}\sigma_2$, otherwise it equals $\sigma_2^2/(\sigma_1\sqrt{2\sigma_2^2 - \sigma_1^2})$. Since $f(r|\theta)$ is continuously differentiable in θ we conclude, that for $\sigma_1 \geq \sqrt{2}\sigma_2$ the $J(\theta|R)$ is a continuous function of θ, starting from zero at $\theta = 0$ and approaching infinity as $\theta \to 0$. We stress at this point, that $f(r|\theta)$ in Eq. (22) satisfies the assumptions which are usually mentioned in applications (Kay, 1993), i.e., continuous differentiability in θ, the independence of the support on θ and automatic existence of the unbiased estimator. Zero or divergent values of Fisher information, however, cannot be ignored. One might be tempted to conclude, based on Eq. (6), that stimuli intensities near $\theta = 0$ cannot be estimated at all (infinite mean squared error), while intensities near $\theta = 1$ can be estimated with arbitrarily high precision. In fact, different assumptions are broken in this case, preventing us to apply the Cramer-Rao bound, and correspondingly the low-noise approximation.

First, zero value of Fisher information is generally interpreted as that there is no *unbiased* estimator for the problem in question (Pitman, 1979), therefore Cramer-Rao bound in the form of Eq. (6) does not apply. The variant of Cramer-Rao inequality for biased estimators can be used (Brunel and Nadal, 1998), however, the problem is that the dependence of bias on θ is usually not known beforehand.

Second, the condition $J(1|R) = \infty$ lower-bounds the mean square error of the estimator by zero (provided that the unbiased estimator exists), making the Cramer-Rao clearly not achievable. Thus, one may view the Cramer-Rao bound as a trivial inequality (mean square error is greater than zero). However, it turns out that the model from Eq. (22) does not satisfy additional requirements for the Cramer-Rao bound to be applicable. As follows from the logic of the proof of the Cramer-Rao inequality (Pitman, 1979, Chapter 5), the bound can hold only if both the density $f(r|\theta)$ and the estimator $\hat{\theta}(r)$ satisfy certain conditions. Since the form of the estimator

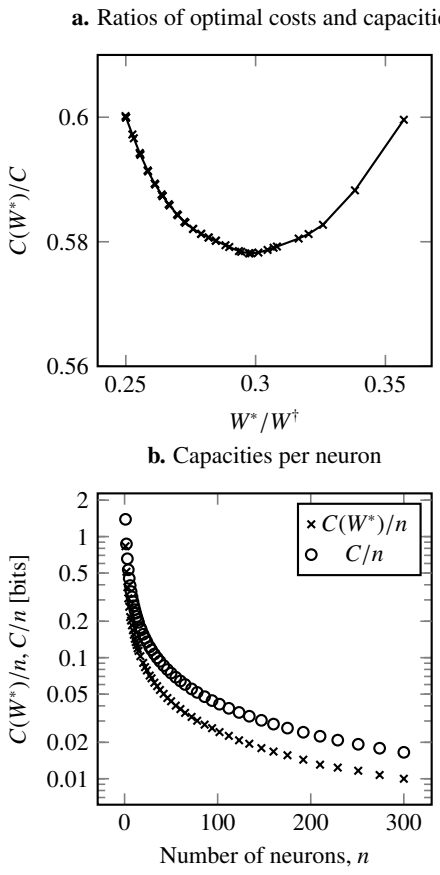

Figure 8. Ratios of capacity and cost values at information-optimal conditions. The ratio of capacity at optimal cost to capacity vs. the ratio of optimal to capacity-achieving metabolic cost is shown in **a**. The neuronal population size increases along the curve from the right to the left. Note that while the ratio $W^*/W^\dagger$ decreases with increasing n, the ratio of capacities stays relatively the same, around 0.6, with mild minimum for medium-sized populations ($n = 20, \ldots, 70$). The values of $C(W^*)$ and C per neuron can be taken as indicators of optimal population size for these quantities. The most "efficient" is thus the single-neuronal case.

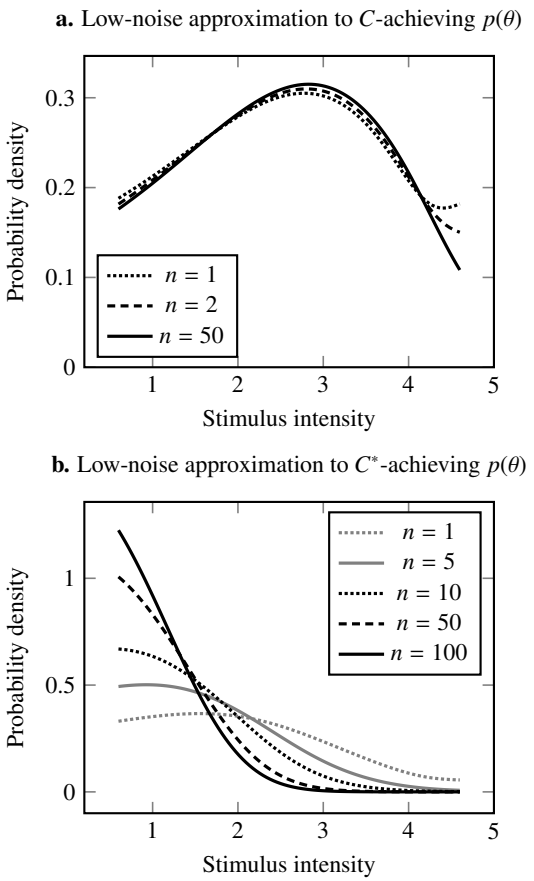

Figure 9. Low-noise approximations of capacity and capacity per unit cost achieving input densities. The capacity-achieving p.d.f. approximation shows only minimal dependence on the population size, shapes for $n > 50$ are visually not distinguishable (**a**). The capacity per unit cost-achieving p.d.f. approximations depend on n more significantly (**b**), since the relative optimal cost per neuron decreases with increasing n (see also Fig. 7b).

is usually not known beforehand, $f(r|\theta)$ must satisfy additional constraints (Pitman, 1979). It follows, that for some θ_0, the function $[f'(r|\theta)]^2/f(r|\theta_0)$ (the prime denotes derivative with respect to θ) must be bounded and integrable (in r), at least in the infinitesimal neighborhood of θ_0 and also at $\theta = \theta_0$. The density model in Eq. (22) does not meet these conditions. We note, that it is stated also in Brunel and Nadal (1998), that in cases where Fisher information diverges the low-noise approximation is not applicable.

To show exactly what are the differences between the true (numerical) and the approximate (invalid) capacity calculation, we present the results of the low-noise approximation application in Fig. 11, for a sufficiently high signal to noise ratio (Fig. 11a), $\sigma_1 = 0.02$, $\sigma_2 = 0.01$. The exact capacity-achieving input distribution is discrete, shown in Fig. 11b, and $C = 4.31$ bits. The undersampled capacity-achieving solution (mutual information equal to 4.30 bits), analogous to Fig. 10, is shown in Fig. 11c. In order to numerically evaluate the low-noise input density approximation from Eq. (15), the integration of $\sqrt{J(\theta|R)}$ had to be stopped at $\theta = 1 - 10^{-5}$. The value of C_{low} in Eq. (10) then varied between 3.6 to 4.04 bits, depending on the method employed (simple trapezoidal or adaptive (R Development Core Team, 2008)). The correspondence between the undersampled numerical solution and the low-noise approximation is quite poor, especially near the extremal points of the stimulus range (as expected).

4. CONCLUSIONS

In this paper we calculated the information capacities, capacities per unit cost and their respective optimizing input distributions, for an empirical model of neuron and a group of neurons. We identified regions of validity and precision of the low-noise approximation with respect to exact numerical

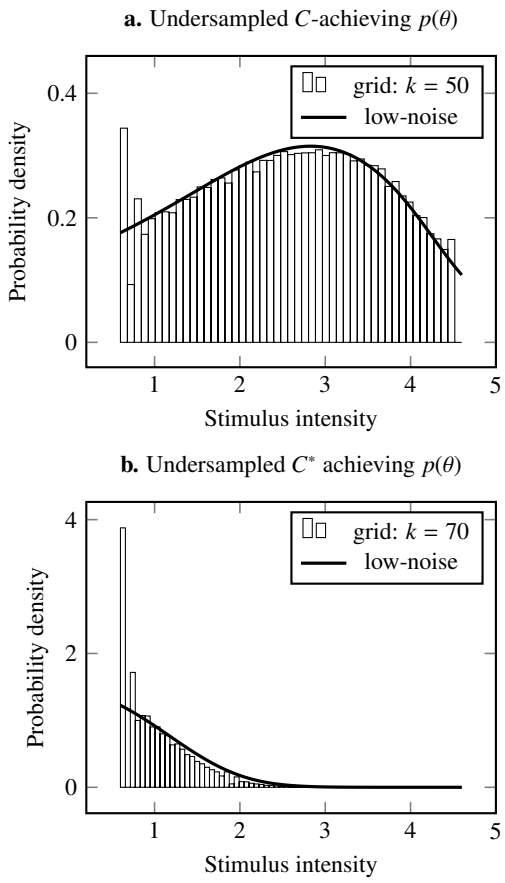

Figure 10. Comparison of low-noise approximations to optimal input densities with undersampled numerical solutions. The population of 274 neurons corresponds to a scenario with very high signal-to-noise ratio. The input grid size was reduced to 50 and 70 points to obtain non-zero probabilities over the whole grid in order to imitate the assumption of the low-noise input p.d.f. approximation in Eq. (15). The agreement between the low-noise approximation and undersampled input grid calculations is very good, with mild discrepancy for small and high stimulus intensities. The underestimation of numerically obtained probabilities is more apparent in the comparison between capacity per unit cost-achieving input distributions.

solutions. We found out, that the low-noise approximation requires relatively high signal-to-noise ratio scenario in order to give relevant results, and that the approximation performs less favorably when the capacity per unit cost is to be determined. This conclusion should be taken into account whenever the approximation is applied. We demonstrated, by employing numerical optimization procedures, that the capacity can be near-achieved by a variety of different input distributions. From this point of view, the numerical value of capacity is more "relevant" than the exactly determined optimal probability distribution. We also demonstrated on an innocuously looking S-R relationship, that the low-noise approximation may not be valid and provided reasons why it is so.

Finally we note, that we are convinced that the values of capacities are useful in the context of neuronal systems, since they provide an upper bound on information about stimulus intensities that can be transferred reliably, by whatever means available. How, and whether at all, the real neurons approach these limits represents an open problem which is beyond the scope of this paper.

Acknowledgements

This work was supported by the Institute of Physiology RVO:67985823, the Centre for Neuroscience P304/12/G069 and by the Grant Agency of the Czech Republic projects P103/11/0282 and P103/12/P558.

Appendix A: Numerical methods

In this paper we employ the cutting-plane algorithm (Huang and Meyn, 2005; Kelley, 1960), which is applicable under more general circumstances than the Blahut-Arimoto algorithm (e.g., continuously varying input and/or output, metabolic constraints), and also allows to control 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.

Assume, that $\mathcal{F}$ is a set of all input stimulus distributions defined over finite stimulus range, $[\theta_{\min}, \theta_{\max}]$. Additionally, it is required that $f(r|\theta)$ is sufficiently well behaved, see Huang and Meyn (2005) for not too restrictive assumptions. The *channel sensitivity* function $g_p(\theta)$, given the input p.d.f. $p(\theta)$, is defined as the Kullback-Leibler distance

$$g_p(\theta) = \int_R f(r|\theta) \ln \frac{f(r|\theta)}{f(r)}\, dr, \tag{A1}$$

thus Eq. (2) can be expressed as

$$I(\Theta; R) = \int_\Theta g_p(\theta) p(\theta)\, d\theta. \tag{A2}$$

It follows that for some selected $\Theta_0 \sim p_0(\theta) \in \mathcal{F}$ holds

$$I(\Theta_0; R) = \min_{p(\theta)\in\mathcal{F}} \int_\Theta g_p(\theta) p_0(\theta)\, d\theta, \tag{A3}$$

since

$$\begin{aligned} &\int_\Theta g_p(\theta) p_0(\theta)\, d\theta = \\ &= \int_\Theta \int_R p_0(\theta) f(r|\theta) \ln \left[\frac{f(r|\theta)}{f(r)} \frac{f_0(r)}{f_0(r)} \right] dr\, d\theta = \\ &= I(\Theta_0; R) + \int_R \ln \frac{f_0(r)}{f(r)} \int_\Theta p_0(\theta) f(r|\theta)\, d\theta\, dr. \end{aligned} \tag{A4}$$

where $f_0(r) = \int_\Theta f(r|\theta) p_0(\theta)\, d\theta$. The last term in Eq. (A4) is ≥ 0, since it can be written in the form of a Kullback-Leibler distance, which is known to be non-negative, and equal to zero only if $p(\theta) = p_0(\theta)$. The minimum is unique, and therefore it holds for $p(\theta)$: $\int_\Theta g_p(\theta) p_0(\theta)\, d\theta \geq I(\Theta, R)$.

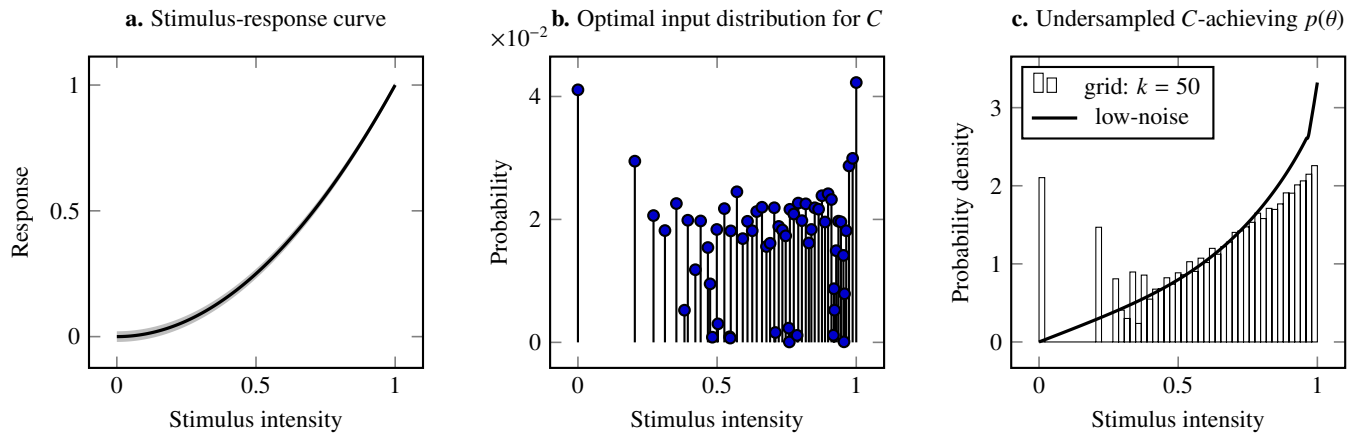

Figure 11. Stimulus-response model with non-regular Fisher information, in the regime of small response variability. The mean stimulus-response curve together with standard deviation is shown (**a**), the model follows Eq. (22) for $\sigma_1 = 0.02$ and $\sigma_2 = 0.01$. The capacity-achieving input distribution is discrete (**b**) with $C = 4.31$ bits. The correspondence between the undersampled numerical solution and the low-noise approximation is quite poor (**c**). The low-noise approximation is not valid, since the model does not meet additional Cramer-Rao regularity conditions (as explained in the main text).

The main idea behind expressing $I(\Theta_0; R)$ in Eq. (A3) as a minimization over a set of input distributions instead of calculating it directly is, that the minimization is *linear* in $p_0(\theta)$, i.e., Eq. (A3) can be written by using the inner (scalar) product $\langle \cdot, \cdot \rangle$ as

$$I(\Theta_0; R) = \min_{p(\theta)\in\mathcal{F}} \left\langle g_p(\theta), p_0(\theta) \right\rangle . \tag{A5}$$

The general problem of finding capacity in Eq. (3) is solved as follows. The algorithm is initialized with an arbitrary distribution $p_0(\theta) \in \mathcal{F}$. In the n-th iteration of the algorithm we have a set of n input distributions, $\{p_0(\theta), p_1(\theta), \ldots, p_{n-1}(\theta)\} \subset \mathcal{F}$, and by employing Eq. (A5) we have

$$I_n(\Theta; R) = \min_{0\le i\le n-1} \langle p(\theta), g_i(\theta) \rangle , \tag{A6}$$

where $g_i(\theta)$ is the sensitivity function given $p_i(\theta)$, $g_i \equiv g_{p_i}$. The next input distribution, $p_n(\theta)$, is obtained as a solution to the problem

$$p_n(\theta) = \arg\max_{p(\theta)} \{ I_n(\Theta; R) : p(\theta) \in \mathcal{F} \}. \tag{A7}$$

The optimization in Eq. (A7) can be expressed as a *linear programming* problem

$$\begin{aligned} &\text{maximize} && c \\ &\text{subject to} && \langle p(\theta), g_i(\theta) \rangle \ge c \quad \text{for } i = 0, \ldots n-1, \\ & && p(\theta) \in \mathcal{F}. \end{aligned} \tag{A8}$$

The standard k-dimensional maximization linear programming problem in variables $\mathbf{x}^{\mathrm{T}} = (x_1, x_2, \ldots, x_k)$, is formulated as

$$\begin{aligned} &\text{maximize} && \mathbf{d}^{\mathrm{T}}\mathbf{x} \\ &\text{subject to} && \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & && \mathbf{x} \ge 0, \end{aligned} \tag{A9}$$

where $\mathbf{A}$ is an $m \times k$ matrix of m linear constraints.

In the following we adapt the scheme in Eq. (A9) to the problem in Eq. (A8). First, we select k points of support of the input distribution, $\{\theta_1, \theta_2, \ldots, \theta_k\}$, in the allowed interval. These points are fixed during the course of the algorithm. We optimize with respect to the probabilities of these points, $\Pr\{\theta_i\}$, so that the allowed input distributions can be expressed as

$$p(\theta) = \sum_{i=1}^{k} \Pr\{\theta_i\}\, \delta(\theta - \theta_i), \tag{A10}$$

where $\delta(\cdot)$ is the Dirac delta function. Let the probability of θ_i in the j-th distribution, $p_j(\theta)$, be denoted as $p_{i,j}$, thus

$$p_j(\theta) = \sum_{i=1}^{k} p_{i,j}\, \delta(\theta - \theta_i). \tag{A11}$$

The probabilities of the points of support of the variable distribution $p(\theta)$ will be simply denoted as $p_1, \ldots, p_k$. We form $(k+1)$-dimensional vectors $\mathbf{x}$ and $\mathbf{d}$ as

$$\mathbf{x}^{\mathrm{T}} = (c, p_1, p_2, \ldots, p_k), \tag{A12}$$

$$\mathbf{d}^{\mathrm{T}} = (1, 0, 0, \ldots, 0), \tag{A13}$$

so that $\mathbf{d}^{\mathrm{T}}\mathbf{x} = c$. The required condition $\mathbf{x} \ge 0$ is thus not in contradiction with Eq. (A12).

Next, we express the matrix $\mathbf{A}$. We write the normalization condition $\sum_i p_{i,j} = 1$ by employing the $1 \times (k+1)$ matrix $\mathbf{A}_1$ as

$$\mathbf{A}_1 = (0, 1, 1, \ldots, 1), \tag{A14}$$

$$\mathbf{A}_1\mathbf{x} = 1, \tag{A15}$$

or equivalently by respecting the inequality condition in Eq. (A9) as

$$\mathbf{A}_1\mathbf{x} \leq 1, \tag{A16}$$
$$-\mathbf{A}_1\mathbf{x} \leq -1. \tag{A17}$$

In order to calculate $C(W)$ by means of the cutting-plane algorithm developed above, it suffices to introduce one more linear constraint,

$$\mathbf{A}_W = (0, w(\theta_1), w(\theta_2), \ldots, w(\theta_k)), \tag{A18}$$
$$\mathbf{A}_W\mathbf{x} \leq W. \tag{A19}$$

Next, we re-formulate the n conditions in the n-th iteration of the algorithm, $\langle p(\theta), g_i(\theta)\rangle \geq c$ for $i = 0, 1, \ldots, n-1$. Rewriting the condition as $c - \langle p(\theta), g_i(\theta)\rangle \leq 0$ yields

$$\mathbf{G}_i = (1, -g_i(\theta_1), -g_i(\theta_2), \ldots, -g_i(\theta_k)), \tag{A20}$$
$$\mathbf{G}_i\mathbf{x} \leq 0, \tag{A21}$$

where

$$g_i(\theta_j) = \int_R f(r|\theta_j) \ln \frac{f(r|\theta_j)}{\sum_{\ell=1}^{k} p_{\ell,i} f(r|\theta_\ell)}\, dr. \tag{A22}$$

The complete algorithm at n-th iteration can be summarized as follows.

1. *Initialization ($n = 0$).* Select $\{\theta_1, \ldots, \theta_k\}$ and $p_{j,0} = \Pr\{\theta_j\}$, for $j = 1 \ldots k$.

2. *Iteration ($n \geq 1$).* Form the $(3+n) \times k$ matrix $\mathbf{A}$ and vector $\mathbf{b}$,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ -\mathbf{A}_1 \\ \mathbf{A}_W \\ \mathbf{G}_0 \\ \vdots \\ \mathbf{G}_{n-1} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ W \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{A23}$$

   Solve the linear programming problem in Eq. (A9), use the solution $\{p_1^*, \ldots, p_k^*\}$ as the n-th distribution for the next iteration, $p_{j,n} = p_j^*$. Use the obtained value of c as the upper bound, and the value of $\langle p(\theta), g_p(\theta)\rangle$ as the lower bound to check the convergence, in particular (Huang and Meyn, 2005),

$$c_0 > c_1 > \cdots > c_n \to C \qquad \text{upper bound,} \tag{A24}$$
$$\langle g_n, p_n\rangle \to C \qquad \text{lower bound,} \tag{A25}$$
$$p_{i,n} \to p_i^* \qquad p^*(\theta). \tag{A26}$$

The relative precision of the solution was defined as $(c - \langle p(\theta), g_p(\theta)\rangle)/c$.

An extension of the algorithm (A8), that iteratively constructs the optimal input alphabet, has also been proposed (Huang and Meyn, 2005). For the case of capacity, we start with a 2-point grid at $\theta_{\min}$ and $\theta_{\max}$ and find the "capacity" for this undersampled grid. For the next cycle, a new grid point is placed at the such θ, for which the sensitivity function $g_n(\theta)$ attains maximum. In this way, it is possible to "build" an optimized input grid. Unfortunately, as the number of cycles increases, some already placed grid points are no longer assigned non-zero probabilities and can be removed. We found limited use for the grid-optimization procedure in the high signal-to-noise ratio, where, however, the exact positions of θ_i had very small impact (as discussed in the paper). Also note, that while maximization of mutual information is concave in the input probabilities, it is generally neither convex nor concave in the positions of grid points θ_i.

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).

Adrian, E. D., *Basis of Sensation* (W. W. Norton and Co., New York, 1928).

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

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).

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

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.

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

Bhumbra, G. S., Inyushkin, A. N., and Dyball, R. E. J., "Assessment of spike activity in the supraoptic nucleus," J. Neuroendocrinol. **16**, 390–397 (2004).

Bialek, W. and Owen, W. G., "Temporal filtering in retinal bipolar cells. elements of an optimal computation?" Biophys. J. **58**, 1227–1233 (1990).

Borst, A. and Theunissen, F. E., "Information theory and neural coding," Nature Neurosci. **2**, 947–958 (1999).

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

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

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

DeBusk, B. C., DeBruyn, E. J., Snider, R. K., Kabara, J. F., and Bonds, A. B., "Stimulus-dependent modulation of spike burst length in cat striate cortical cells," J. Neurophysiol. **78**, 199–213 (1997).

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).

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

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

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).

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

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

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

Kandel, E. R., Schwartz, J. H., and Jessel, T. M., *Principles of neural science* (Elsevier, New York, 1991).

Kay, S. M., *Fundamentals of statistical signal processing: estimation theory* (Prentice Hall, New Jersey, 1993).

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

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

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

Kostal, L., Lansky, P., and McDonnell, M. D., "Metabolic cost of neuronal information in an empirical stimulus-response model," Biol. Cybern. **107**, 355–365 (2013).

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).

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).

Lansky, P., Rodriguez, R., and Sacerdote, L., "Mean instantaneous firing frequency is always higher than the firing rate," Neural Comput. **16**, 477–489 (2004).

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

Laughlin, S. B., "Matched filtering by a photoreceptor membrane," Vision Res. **36**, 1529–1541 (1996).

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).

Lehmann, E. L. and Casella, G., *Theory of point estimation* (Springer Verlag, New York, 1998).

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

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

McDonnell, M. D., Ikeda, S., and Manton, J. H., "An introductory review of information theory in the context of computational neuroscience," Biol. Cybern. **105**, 55–70 (2011).

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).

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).

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

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).

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).

Pitman, E. J. G., *Some basic theory for statistical inference* (John Wiley and Sons, Inc., New York, 1979).

R Development Core Team,, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2008).

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

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

Rospars, J.-P., Lansky, P., Chaput, M., and Viret, P., "Competitive and non-competitive odorant interaction in the early neural coding of odorant mixtures," J. Neurosci. **28**, 2659–2666 (2008).

Shadlen, M. N. and Newsome, W. T., "Noise, neural codes and cortical organization," Curr. Opin. Neurobiol. **4**, 569–579 (1994).

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).

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

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).

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

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

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).

Wu, Y. and Verdu, S., "The Impact of Constellation Cardinality on Gaussian Channel Capacity," in *Forty-Eighth Annual Allerton Conference, University of Illinois at Urbana-Champaign* (2010) pp. 1–9.