Optimal population coding by noisy spiking neurons

Gašper Tkačik[1], Jason Prentice, Vijay Balasubramanian, Elad Schneidman

Supporting Information

[1]To whom correspondence should be addressed. E-mail: gtkacik@sas.upenn.edu

1 Results for 2 neurons

Here we compute the noise entropy, output entropy and information for two neurons $\sigma_{1,2}$ exposed to a binary input distribution, where $h_{1,2} = \pm 1$:

$$\begin{aligned} P_h(-1,-1) &= P_h(1,1) = \frac{1}{4}(1+\alpha) \\ P_h(-1,1) &= P_h(1,-1) = \frac{1}{4}(1-\alpha) \end{aligned} \tag{1}$$

In this distribution, the inputs to both neurons are zero mean, $\langle h_1 \rangle = \langle h_2 \rangle = 0$, unit variance, and the correlation between the inputs $\mathrm{Cov}(h_1, h_2) = \alpha$.

For two neurons, the conditional distribution of the outputs is given by Eq (2) of the main text:

$$P(\sigma_1, \sigma_2 | h_1, h_2) = \frac{1}{Z} e^{\beta(h_1\sigma_1 + h_2\sigma_2 + J\sigma_1\sigma_2)}, \tag{2}$$

$$Z = e^{\beta(h_1+h_2+J)} + e^{\beta(-h_1-h_2+J)} \tag{3}$$

$$+ \; e^{\beta(h_1-h_2-J)} + e^{\beta(-h_1+h_2-J)}. \tag{4}$$

This distribution depends on the coupling J and the neural reliability parameter β. For output entropy, $S[P(\sigma_1, \sigma_2)]$, we first evaluate $P(\sigma_1, \sigma_2)$:

$$\begin{aligned} P(\sigma_1,\sigma_2) &= \sum P_h(\vec{h}) P(\sigma_1, \sigma_2 | \vec{h}) \\ &= \frac{1}{4}(1+\alpha) \frac{e^{\beta(\sigma_1+\sigma_2+J\sigma_1\sigma_2)} + e^{\beta(-\sigma_1-\sigma_2+J\sigma_1\sigma_2)}}{e^{\beta(2+J)} + e^{\beta(-2+J)} + 2e^{-\beta J}} + \\ &+ \frac{1}{4}(1-\alpha) \frac{e^{\beta(\sigma_1-\sigma_2+J\sigma_1\sigma_2)} + e^{\beta(-\sigma_1+\sigma_2+J\sigma_1\sigma_2)}}{e^{\beta(2-J)} + e^{\beta(-2-J)} + 2e^{\beta J}}. \end{aligned} \tag{5}$$

With the distributions Eq (2,5) in hand, we can compute the output entropy:

$$S_{\mathrm{out}}[P(\sigma_1,\sigma_2)] = -\sum_{\sigma_{1,2}=\pm 1} P(\sigma_1,\sigma_2) \log_2 P(\sigma_1,\sigma_2), \tag{6}$$

and the noise entropy

$$S_{\mathrm{noise}} = -\sum_{h_{1,2}=\pm 1} P_h(h_1,h_2) \sum_{\sigma_{1,2}=\pm 1} P(\sigma_1,\sigma_2|h_1,h_2) \log_2 P(\sigma_1,\sigma_2|h_1,h_2), \tag{7}$$

by simply summing over 4 and 16 terms, respectively. The mutual information $I(\sigma_{1,2}; h_{1,2})$ is given by the difference between the signal and noise entropies, i.e. $I(\sigma_{1,2}; h_{1,2}) = S_{\mathrm{out}} - S_{\mathrm{noise}}$, and depends on the parameters α, β and J. For each choice of the input correlation α and neural reliability β, we can look for the optimal coupling $J^*(\alpha, \beta)$, as shown in Fig. 2 of the main paper.

In the case of Gaussian inputs, averaging over P_h analytically is not easy, so we evaluate the averages numerically by sampling from P_h (with 2000 samples for each choice of α

and β). Figure 1 shows how the signal and noise entropies behave as β changes from low- to high-neural-reliability regime ($\beta = 0.5, 1, 2$) with Gaussian inputs. We clearly see how, in the case of 2 neurons, the optimal system will move from noise- to output-entropy dominated regime as β is increased.

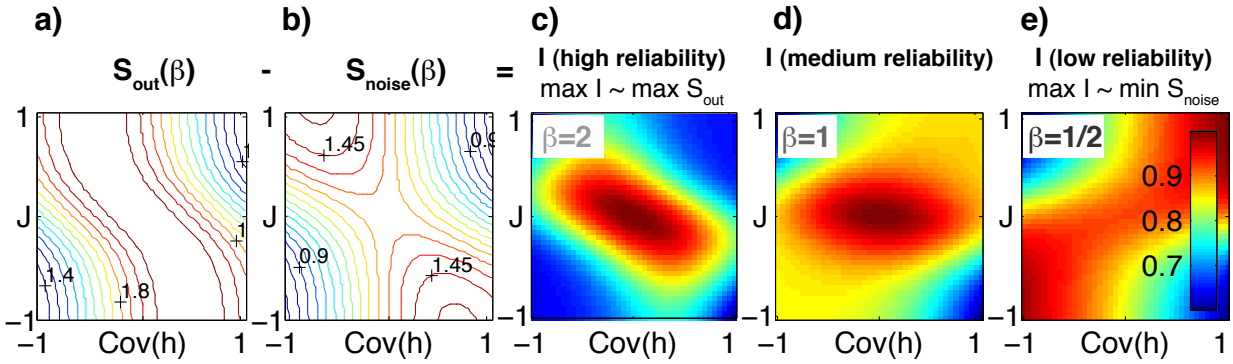

Figure 1: Results for two coupled neurons with Gaussian correlated inputs. **a)** The contour plot of output entropy, $S_{\rm out}$, as a function of the input covariance α on the x-axis, and the coupling J on y-axis, at $\beta = 1$. **b)** The contour plot of noise entropy, $S_{\rm noise}$, as a function of the input covariance $\alpha = {\rm Cov}(h)$ and the coupling J, at $\beta = 1$. The mutual information between the state of the two neurons and the value of the two inputs is given by the difference $S_{\rm out} - S_{\rm noise}$. Panels **c), d), e)** show the mutual information (in color), normalized to the maximum at each β, for three different values of $\beta = 2, 1, 0.5$, respectively. Red hues indicate high information values and thus a good choice for the optimal $J(\alpha)$. At high neural reliability (large β) in **c)**, the output entropy term dominates over small noise entropy, and high information is achieved for those values of J where the output entropy in **a)** is large. At intermediate reliability ($\beta = 1$) in **d)** the optimal J^* has to balance between output and noise entropy contributions of comparable size. At low reliability (small β) the noise entropy dominates, and the maximal information is achieved for those values of J where the noise entropy in **b)** is low.

2 Networks of neurons exposed to binary stimuli

To generalize our results for a system of 2 interacting neurons (see main text) to a small network composed of N neurons, we maximized the information with respect to couplings $\{h_i, J_{ij}\}$, when the network received either binary or Gaussian stimuli. The results for Gaussian stimuli are presented in the main text; here we present full results for the binary case.

We considered a network of N neurons responding to an input ensemble of K equiprobable N-bit binary patterns chosen randomly from the set of 2^N such patterns. For $N \lesssim 10$ it remained possible to numerically choose couplings h_i^0 and J_{ij} that maximized information about the input ensemble represented in network responses. Since the K patterns $\vec{h}^\mu$

($\mu = 1 \cdots K$) were chosen randomly, the correlation between inputs to different neurons is

$$\mathrm{Cov}(\vec{h}) = \frac{1}{K} \sum_{\mu} h_i^{\mu} h_j^{\mu}. \tag{8}$$

For any fixed K these correlations are weak, but non-zero, and thus the information-maximizing network could tune J_{ij} to exploit the input correlations and build an optimal K-way discriminator, i.e. a neural representation that best separates the inputs. Adapting to this stimulus ensemble is a difficult task for the network, because any correlations present in the stimulus are solely a result of having a small number of patterns K, and these correlations decay as K increases.

We numerically optimized the network interactions to maximize information, and found that when neural reliability β is low, the optimal network represents more information than the uncoupled ($J_{ij} = 0$) network (Fig. 2a). The relative increase is greater for larger networks. The increase in information is largest when the number of input patterns (K) is relatively small, but increasing with the size of the of the network (N) – this is reminiscent of the slow growth of the storage capacity of Hopfield networks, an analogy that we will return to below. When neural reliability β is high, there is less advantage to tuning the interactions (Fig. 2b) because, in the absence of noise, binary neurons have the capacity to encode complete information about an ensemble of binary inputs. Hence, as neural reliability increases, the neurons tend to become independent encoders. For any β, neurons also tend to become independent encoders as the number of input patterns K increases. This is because when K is large, the small-sample induced correlations between the patterns decay. Thus the network cannot use these correlations to improve information transmission. Input decorrelation is never an optimal strategy.

When the neurons are independent and exposed to binary inputs of zero mean, i.e. $P(h_i = 1) = P(h_i = -1) = 0.5$, we can compute the information conveyed by the single neuron:

$$\begin{aligned}
I_1(\sigma; h) &= S[p(\sigma)] - \langle S[p(\sigma|h)] \rangle_{p(h)} && (9)\\
&= 1 - [\beta \langle E \rangle + \log Z] \,/\, \log 2 && (10)\\
&= 1 - [-\partial_\beta \log Z + \log Z] \,/\, \log 2 && (11)\\
&= 1 - [-\beta \tanh \beta + \log(2 \cosh \beta)] \,/\, \log 2, && (12)
\end{aligned}$$

where in Eq (10) we used the thermodynamic identity that relates the free energy, F, the partition function Z, the expected value of the "energy" $E = -\sum_i (h_i + h_i^0)\sigma_i - 1/2 \sum_{ij} J_{ij}\sigma_i\sigma_j$, and the conditional entropy S: $\beta F = -\log Z = \beta \langle E \rangle - S$, and the fact that the output entropy of the neuron will optimally be 1 bit. Fig. 3 displays the match between the results of this calculation with the numerical optimization solution for large K where independence between neurons is expected.

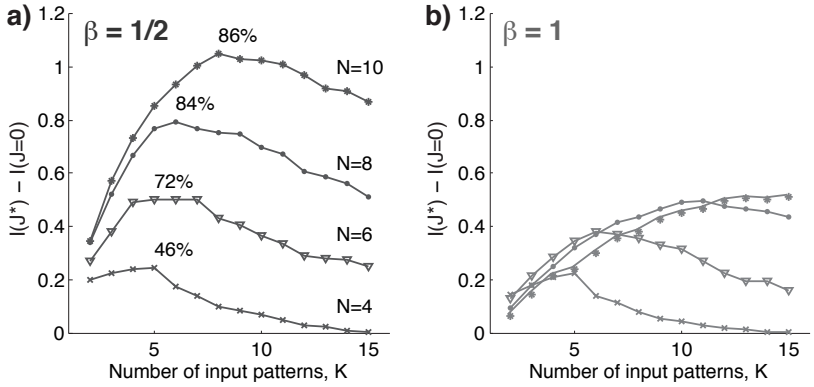

Figure 2: Information transmission in small networks with binary input distributions. **a)** The increased capacity $I_{J^*}(\{\sigma_i\}\,;\vec{h}) - I_{J=0}(\{\sigma_i\}\,;\vec{h})$ of optimal vs uncoupled networks, for $N = 4, 6, 8, 10$ neurons (crosses, triangles, dots, stars), as a function of the number K of equiprobable, randomly chosen, binary input patterns. The numbers at the peaks denote the relative increase in information capacity at that point. As K is increased, the input correlations decay and the optimal solution approaches the independent solution. The neural reliability parameter is $\beta = 1/2$. **b)** The same plot for $\beta = 1$ shows that the capacity of the optimal network increases less (although the baseline capacity of the uncoupled network is larger), and the peak is reached at larger K.

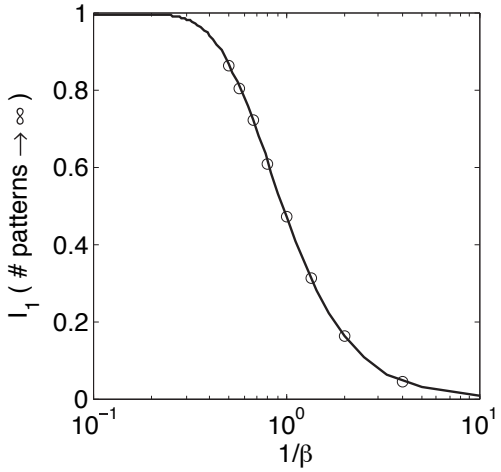

Figure 3: Line: The information per neuron, I_1, in the uncoupled case, as a function of $1/\beta$ (result from Eq (12); dots: information per neuron in optimized network with $N = 10$ neurons exposed to $K = 10^4$ random binary input patterns). Because $K \gg N$, the optimized solution is incapable of yielding more information than the independent solution.

3 Metastable states for binary inputs, and relation to Hopfield networks

In this section we present results related to the emergence of maximum likelihood (ML) states for networks responding to binary inputs. These results complement Figs. 4,5 of

the main paper that discuss Gaussian inputs.

A pattern that has locally maximal likelihood (ML), or a "meta-stable state" as it is known in statistical physics, is a configuration of variables $\mathcal{G}_\mu = \{\sigma_i\}_\mu$, such that a flip of any of the σ_i would decrease that pattern's likelihood (or equivalently, increase its energy in the analogous Ising model). ML patterns are defined when no stimulus is present, i.e. with respect to the likelihood given by optimal parameters $\mathbf{g} = \left\{h_i^{*,0}, J_{ij}^*\right\}$ and zero-input, $h_i(s) = 0$.

For small enough networks of N neurons, such ML patterns can be found by exhaustively enumerating the likelihood of all 2^N patterns and examining their likelihood under single σ_i flips. We numerically searched for optimal networks of $N = 10$ neurons at neural reliability parameter $\beta = 1/2$, exposed to $K = 10$ equiprobable binary inputs. On average, we found on the order of 10 ML patterns (average ~ 7.7 over 30 generated networks) for each network.

For binary inputs, these ML patterns, arising purely as a result of the optimized parameters $\mathbf{g}^* = \left\{\vec{h}^{*0}, J_{ij}^*\right\}$, correspond well to the patterns of the input $\vec{h}^t$. E.g., Fig. 4 plots the overlaps (scalar products) of the 10 ML patterns, $\mathcal{G}_\mu, \mu = 1 \ldots 10$, found in one particular optimized network, with the $K = 10$ input patterns that that network received: 8 of the 10 input patterns overlap perfectly with one of the ML states, i.e. $|\mathcal{G}_\mu \cdot \vec{h}^t| = 1$. This result is typical of other optimization runs.

As a result, when a particular stimulus $\vec{h}^t$ is applied, the state of the network very reliably collapses onto a single output pattern (which is predominantly a ML pattern that has a high overlap with the applied stimulus). This is indicated by the histogram of $S[P(\{\sigma_i\})|\vec{h}]$ (with mean at 0.18 bits) in Fig. 4. When we look at the full distribution of output patterns, $P(\{\sigma_i\})$, a striking difference between the optimal (J^*) vs uncoupled ($J = 0$) networks emerges, consistent with the observations made in Fig. 4 of the main paper: the distribution of output binary patterns $P(\{\sigma_i\})$ is localized more tightly around the ML patterns in optimal networks. Effectively, strong optimal couplings *exclude* certain patterns from the repertoire of binary words and thus prevent noise from driving the system there, thus making use of only a small subset of the full 2^N binary patterns to carry information.

This result implies that when the network has adapted to the particular stimulus ensemble $P_h(\vec{h})$, it has "stored" (some of) the presented patterns in the stimulus into the structure of its optimal coupling constants, $\mathbf{g}^* = \left\{\vec{h}^{*0}, J_{ij}^*\right\}$. Correspondingly, the likelihood (or energy) landscape is no longer uniform, as in the case of uncoupled ($J = 0$) networks, but contains ~ 10 ML patterns, surrounded by their local "basins of attraction"; see Fig. 4 of the main paper.

The notion of the "basin of attraction" around a maximal likelihood (ML) pattern can be made precise by introducing a so-called zero-temperature Monte Carlo (ZTMC) dynamics: given any binary pattern $\{\sigma_i\}$, flip the σ_i in a random sequence until the likelihood cannot be increased further. Then, the system must have settled into one of the ML patterns, $\mathcal{G}_\mu$. As a consequence, any of the 2^N possible binary patterns can be mapped to some ML pattern. All patterns that the dynamics takes into the ML pattern

$\mathcal{G}_\mu$ are said to constitute the basin of attraction of $\mathcal{G}_\mu$. The number of distinct patterns that are mapped by dynamics into $\mathcal{G}_\mu$ is called the "size of the basin of attraction" of $\mathcal{G}_\mu$.

Thus we examined the sizes of the basins of attraction of ML patterns in 30 optimal networks of $N = 10$ neurons responding to binary stimuli. For those ML patterns that had perfect overlaps with the patterns from the stimulus set, the basins were large, peaking at about 90 patterns. For ML patterns that had imperfect overlaps with the patterns from the stimulus set, the basins had much smaller sizes, peaking at less than 10 patterns; see Fig. 4.

Such behavior is strongly reminiscent of Hopfield associative memory network. We recall that in a Hopfield net, the "memories" $\vec{h}^t$ are given. A coupling matrix, $J^H = 1/N \sum_t \vec{h}^t * \vec{h}^{t,T}$ is then explicitly constructed, and it is shown that such system, evolving under the zero-temperature Monte Carlo (ZTMC) dynamics, will converge to one of the memories, if the initial state has a big enough overlap with the stored memory. The Hopfield network is a dynamical Ising-like network, in which the dynamics maps partial memories to complete memories (recall). When viewed as an Ising-like model in equilibrium, such a network stores the memories as ML patterns by construction.

In contrast, in our case the matrix J_{ij} has not been constructed by hand to contain a chosen set memories as ML patterns. Rather, such a structure has emerged as a consequence of information optimization in the low neural reliability (low β) regime. Since it is a consequence of optimization, our network does not suffer from complete storage capacity breakdown as more and more patterns are added – rather, the excess information capacity peaks when the stimulus ensemble contains $K^* \sim N$ patterns, and then gracefully degrades to the performance of the uncoupled ($J = 0$) network (see Fig. 2). Although we cannot numerically explore the scaling of the excess information beyond $N \sim 10$ neurons, it seems reasonable to assume that the $K^* \sim O(N)$ scaling of the peak performance, found for the Hopfield network memorizing random binary patterns, holds for our case as well.

The primary task of the network in the present paper is to *transmit* information, by mapping its inputs to the equilibrium distribution of output patterns. Hopfield networks, on the other hand, are usually discussed as examples of computation, that is, irreversible dynamical processes. Interestingly, if our networks, which maximize information in their equilibrium response, are allowed to evolve under ZTMC (Hopfield) dynamics, they would also implement dynamic associative recall, by mapping arbitrary inputs to stored ML patterns – i.e., the memorized inputs. Equivalently, in the stimulus-free condition, the equilibrium distribution of responses in our networks is strongly peaked at the "memories," reflecting the stimuli to which the network has adapted.

Finally, in our setup, ML patterns can and do appear even when the inputs are not binary, for example, when they are drawn from the Gaussian distribution. This provides an extension of the notion of (Hopfield-net-like) "memory" as a ML response pattern, to continuous inputs.

4 Creating Gaussian stimulus ensembles

To explore the optimal networks exposed to correlated Gaussian inputs, we needed to create an ensemble $P_h(\vec{h})$ of $N = 10$-dimensional correlated Gaussian processes and draw samples from it. Assuming that the mean values $\langle h_i \rangle = 0$, this ensemble is fully characterized by the spectrum of 10 real and positive eigenvalues d_i, and an orthonormal rotation matrix P, such that the covariance matrix C of the samples $\vec{h}$ can be decomposed as:

$$C = PDP^T, \tag{13}$$

where $D = \mathrm{diag}(d_i)$. For each covariance matrix, we choose d_i to be exponentially distributed with a decay constant $1/\log(2)$. The rotation matrices P were picked in the following way: we first created a random symmetric matrix A whose elements were independently drawn from the normal distribution with zero mean and unit variance. Then we decomposed A into diagonal basis, $A = PQP^T$, and extracted the resulting matrix P as a random choice for a rotation matrix. Having drawn the spectrum of eigenvalues D and the rotation matrix P, we could then generate the required Gaussian inputs $\vec{h}$ for each network optimization. Lastly, we subtracted a constant from each h_i to fix $\langle h_i \rangle = 0$, and normalized each h_i so that the variance of each h_i was 1. To illustrate the results of the described procedure, Fig. 5 shows the spectrum of eigenvalues of covariance matrices of stimuli.

This procedure guarantees that: **i)** each neuron always receives inputs with the same statistics (gaussian distributed, zero mean, unit variance), so all the "interesting" structure is in the covariance matrix; **ii)** the choice of the eigenvalue spectrum ensures that the variance along different principal axis is significantly different, so that the distribution is not spherically symmetric and there is opportunity for the network to decorrelate (if all d_i were the same, the stimulus would have been decorrelated by construction).

5 Optimal values for biases h_0^*

When the firing rate is not constrained (see next section), the values of the biases h_0 are selected by the optimization procedure so that the neurons fire about half of the time, i.e. $\langle \sigma_i \rangle \sim 0$. This ensures that the output entropy can be maximized, by maximizing the first term in Eq (6) of the main paper. To check this, we histogram the mean firing rates across all optimal networks with $N = 10$ neurons, exposed to Gaussian stimuli, at reliability values ranging from $\beta = 1/5 \ldots 5$. Figure 6 shows that the mean firing probability stays close to 50% ($\langle \sigma_i \rangle = -0.005$, $\mathrm{Std}(\sigma_i) = 0.11$). We can further examine, in the mean-field approximation, the effective stimulus-averaged firing bias acting on each neuron, $h_i^{eff} = \langle h_i^t \rangle + h_i^0 + \sum_j J_{ij} \langle \sigma_j \rangle$. To insure that the neurons are close to spiking with probability 0.5, this average effective firing bias should be close to zero. For each neuron in each optimal network we therefore plot the optimal bias, h_i^0, against the contribution from other neurons in the network, $\sum_j J_{ij} \langle \sigma_j \rangle$. There is a clear negative correlation over the whole range of reliability values β, indicating that indeed the optimal h_i^0 has been

chosen to realign the operating point of each neuron to the center of its tanh nonlinearity. The anti-correlation is only aproximate because the networks are not necessarily in a regime where the mean-field approximation is precise.

6 Optimizing information at constrained firing rate

To examine the effects of information optimization while keeping the average firing rates of neurons pegged to a chosen value, we use the method of constrained optimization using Lagrange multipliers. We form the functional:

$$\mathcal{L}(\left\{h_i^0, J_{ij}\right\}, \lambda) = I(\{\sigma_i\}; \vec{h}) - \lambda \frac{1}{N} \sum_{i=1}^{N} \langle \sigma_i \rangle. \tag{14}$$

Here, the mutual information $I(\{\sigma_i\}; \vec{h})$ is the same as in Eq (5) of the main paper, but we have added a term that constrains the mean firing rate of the neurons; λ serves as the Lagrange multiplier.

Using the same binary and Gaussian stimulus ensembles as in the unconstrained case presented in the paper, we numerically optimized the couplings $\mathbf{g} = \{h_i^0, J_{ij}\}$ for $N = 10$ neurons, while varying the multiplier λ across a range of values. At each value of λ, this allows us to find the optimal couplings at a particular value of the mean firing rate, $m = 1/N \sum_{i=1}^{N} \langle \sigma_i \rangle$, that depends parametrically on λ.

By thus scanning through a range of values for λ, we can construct parametric plots of $I(\vec{h}; \{\sigma_i\})$ as a function of m, shown in Fig 7 – these curves are also known as rate-distortion (RD) plots in information theory. Consistently with the results of the main paper, for both binary and Gaussian inputs, the optimal solution provides an improvement of $\sim$ 10-50% when there is no firing rate constraint. In the binary case the improvement decreases with increasing neural reliability (β), from $\sim$ 50% at $\beta = 0.5$ to negligible improvement at $\beta = 2$; in the Gaussian case the improvement remains even at high reliability β, because the network is able to decorrelate the stimuli.

For both types of stimuli, as the firing rate is constrained, the improvement in information decreases in absolute magnitude (see Fig 7), but the information capacity also decreases because the set of output binary patterns is more and more biased towards silence ($\{-1\}$ pattern), causing the output entropy (and therefore information) to fall. Nevertheless, optimization still provides benefits across a range of firing rates.

Since the Lagrange multiplier only constraints the mean firing rate of the population, we can ask about the distribution of rates within that population. Distributions averaged over many solutions show that the networks receiving binary inputs tend to maximize information at constrained rate by designating some neurons to remain silent, while utilizing others fully. This results in bimodal distributions of firing rates, as seen in Fig 7. In contrast, in the Gaussian case the firing rate distribution is a single-peaked function. We checked that the qualitative shapes of the distributions are not a consequence of averaging over the distributions found in many optimized networks, but rather reflect the shape of the firing rate distributions in each optimized network.

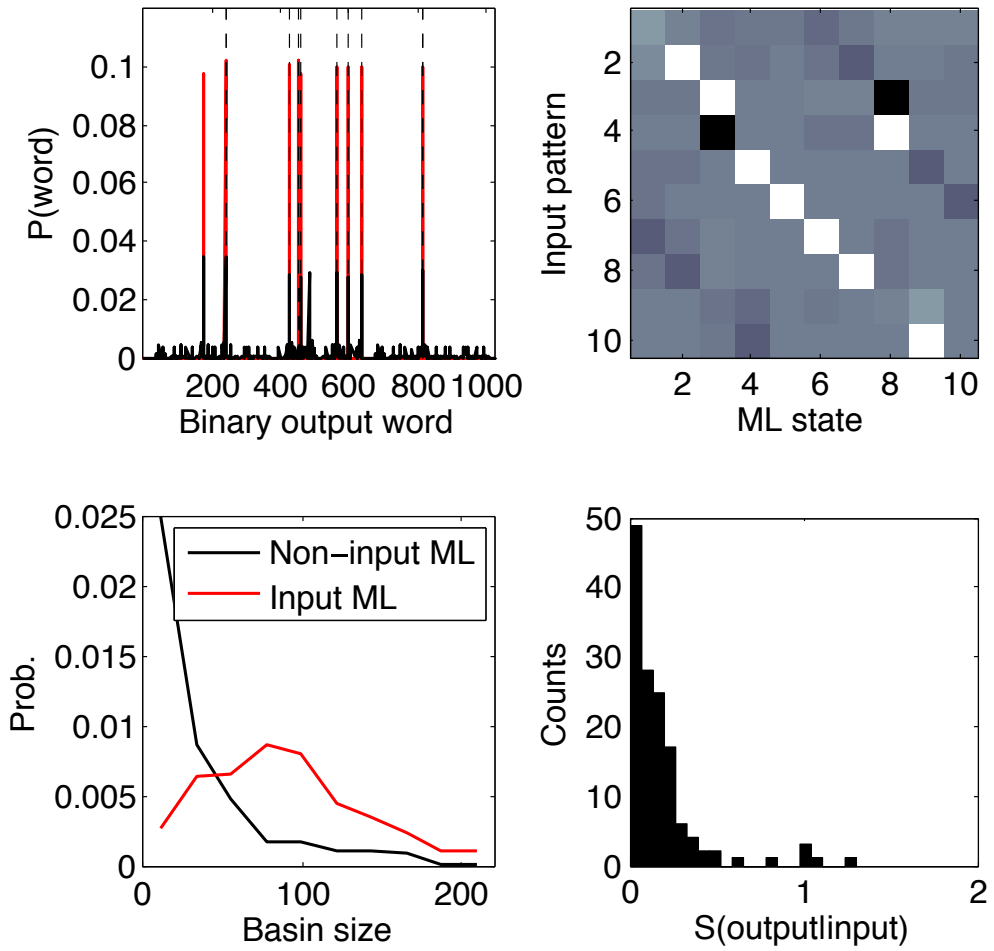

Figure 4: The analysis of network states of $N = 10$ neurons at $\beta = 1/2$, exposed to $K = 10$ equiprobable binary input patterns. *Upper left.* An example of the probability, $P(\{\sigma_i\})$, of network states for an independent ($J = 0$, black) and optimized (J^*, red) network, for each of the 2^N possible configurations $\{\sigma_i\}$. Black dashed lines show the ML patterns of the network under no-stimulus condition. Because of the couplings, the optimal network (red) has a much smaller probability of fluctuating into one of the states that do not correspond well to a ML pattern. *Upper right.* A comparison of ML patterns of an example network with its $K = 10$ possible input patterns. Shown is the overlap (scalar product between the input and ML patterns, see text) in color code, where dark black or bright white show perfect (anti-)overlap. For 8 out of 10 patterns the network has "memorized" the input as a ML pattern. *Lower left.* The analysis of the basins of attraction of the ML patterns, for 30 optimized networks. For each ML pattern, we count how many of the possible 2^N network states belong to that basin. Plotted is the distribution of basin sizes for ML patterns that perfectly overlap with an input (in red), and states that have less-than-perfect overlap (in black). *Lower right.* The distribution of conditional entropies for 30 optimized networks, when an input pattern is presented. The conditional entropy peaks close to 0, meaning that even when individual neurons are unreliable the network interactions drive inputs into stable output patterns, albeit at the cost of reducing output entropy.

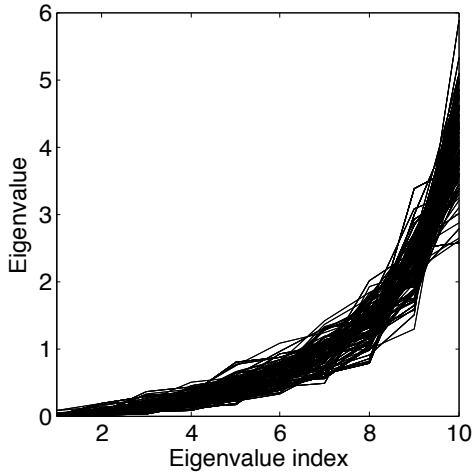

Figure 5: The eigenvalue spectra computed from 450 sets of Gaussian stimuli drawn from $P_h(\vec{h})$ as described in the text; these stimuli were used to optimize $N = 10$ networks at 15 different values of neural reliability parameter β, with 30 replicate optimizations for every β.

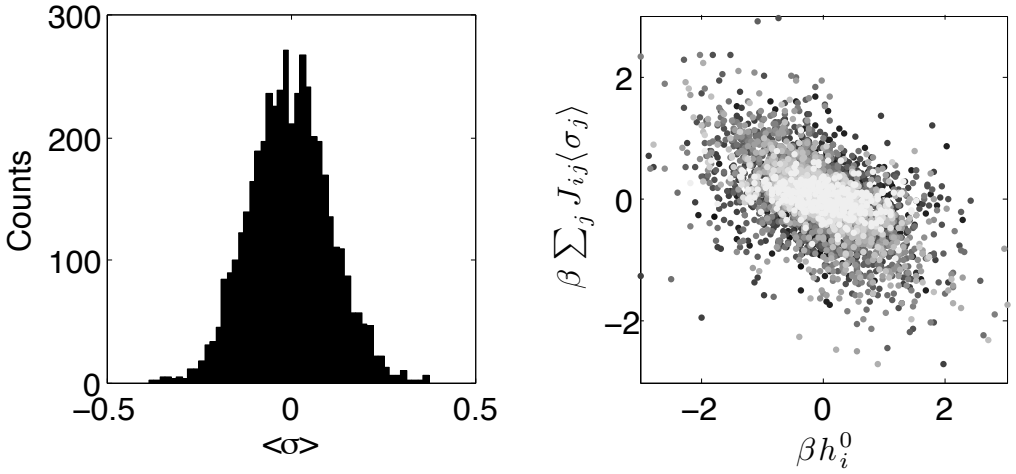

Figure 6: *Left.* The distribution of firing rates in networks of $N = 10$ neurons, exposed to Gaussian stimuli, pooled across a range of 15 values for β, $\beta = 1/5 \ldots 5$ (with 30 optimized networks for each value of β). *Right.* The anti-correlation between the optimal value for the bias acting on neuron σ_i, h_i^0, and the mean-field effect of other neurons σ_j, $j \neq i$. Each point represents a single neuron from one of the optimal networks; shown are 30 networks across a range of 15 values for β, from $\beta = 1/5$ (dark gray) to $\beta = 5$ (light gray). We checked that for every separate value of β, the correlation is significantly negative.

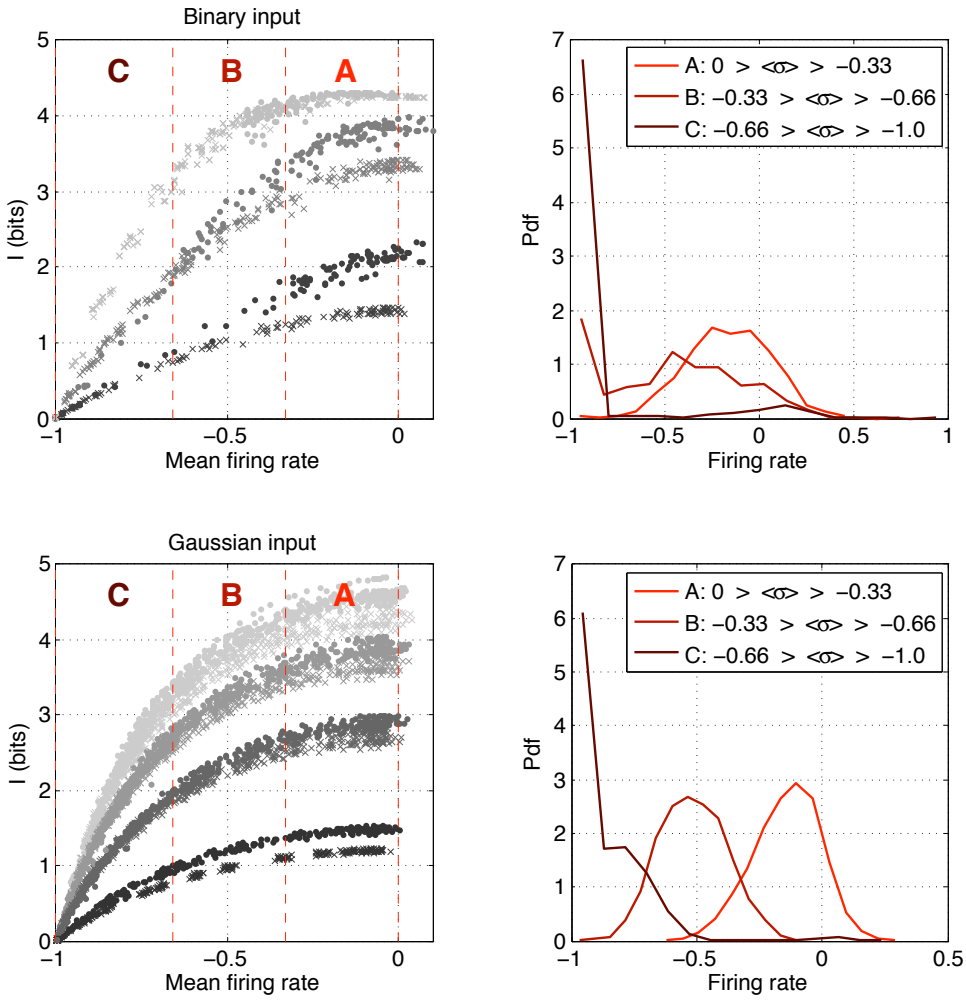

Figure 7: Optimization of information transmission with a mean firing rate constraint in a population of $N = 10$ neurons. *Upper left.* The capacity-rate curve for binary input distribution: crosses show independent ($J = 0$) solutions, dots show optimal J^* solutions; grayscale corresponds to $\beta = 1/2, 1, 2$ (grayscale, dark-to-light). Each symbol represents one network optimized for a draw of $K = 20$ equiprobable binary inputs. *Upper right.* The average distribution of firing rates of individual neurons measured over an interval of mean rates. The intervals are indicated between dashed lines in the upper left panel and in the inset in the upper right panel. The firing rate distributions are pooled over all values of β. With a strong constraint towards silence (interval C), the distribution of rates within the population is bimodal (some neurons stay almost completely silent). *Lower row.* The same analysis for the Gaussian input distribution, at neural reliability values $\beta = 1/2, 1, 1.5, 2$ (grayscale, dark-to-light). There are two notable differences from the binary case. First, even at high reliability (light gray), the optimal solutions still yield higher information transmission than the independent solutions. Second, the distribution of firing rates is closer to unimodal.