Approximations of Shannon Mutual Information for Discrete Variables with Applications to Neural Population Coding

Wentao Huang [1,2,*] **and Kechen Zhang** [2,*]

[1]   Key Laboratory of Cognition and Intelligence and Information Science Academy of China Electronics Technology Group Corporation, Beijing 100086, China

[2]   Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

*   Correspondence: wnthuang@gmail.com (W.H.); kzhang4@jhmi.edu (K.Z.); Tel.: +1-443-204-0536(W.H.); +1-410-955-3538(K.Z.)

Abstract: Although Shannon mutual information has been widely used, its effective calculation is often difficult for many practical problems, including those in neural population coding. Asymptotic formulas based on Fisher information sometimes provide accurate approximations to the mutual information but this approach is restricted to continuous variables because the calculation of Fisher information requires derivatives with respect to the encoded variables. In this paper, we consider information-theoretic bounds and approximations of the mutual information based on Kullback–Leibler divergence and Rényi divergence. We propose several information metrics to approximate Shannon mutual information in the context of neural population coding. While our asymptotic formulas all work for discrete variables, one of them has consistent performance and high accuracy regardless of whether the encoded variables are discrete or continuous. We performed numerical simulations and confirmed that our approximation formulas were highly accurate for approximating the mutual information between the stimuli and the responses of a large neural population. These approximation formulas may potentially bring convenience to the applications of information theory to many practical and theoretical problems.

Keywords: neural population coding; mutual information; Kullback–Leibler divergence; Rényi divergence; Chernoff divergence; approximation; discrete variables

---

1. Introduction

Information theory is a powerful tool widely used in many disciplines, including, for example, neuroscience, machine learning, and communication technology [1–7]. As it is often notoriously difficult to effectively calculate Shannon mutual information in many practical applications [8], various approximation methods have been proposed to estimate the mutual information, such as those based on asymptotic expansion [9–13], *k*-nearest neighbor [14], and minimal spanning trees [15]. Recently, Safaai et al. proposed a copula method for estimation of mutual information, which can be nonparametric and potentially robust [16]. Another approach for estimating the mutual information is to simplify the calculations by approximations based on information-theoretic bounds, such as the Cramér–Rao lower bound [17] and the van Trees' Bayesian Cramér–Rao bound [18].

In this paper, we focus on mutual information estimation based on asymptotic approximations [19–24]. For encoding of continuous variables, asymptotic relations between mutual information and Fisher information have been presented by several researchers [19–22]. Recently, Huang and Zhang [24] proposed an improved approximation formula, which remains accurate for high-dimensional variables. A significant advantage of this approach is that asymptotic approximations are sometimes very useful in analytical studies. For instance, asymptotic approximations allow us to prove that the

optimal neural population distribution that maximizes the mutual information between stimulus and response can be solved by convex optimization [24]. Unfortunately this approach does not generalize to discrete variables since the calculation of Fisher information requires partial derivatives of the likelihood function with respect to the encoded variables. For encoding of discrete variables, Kang and Sompolinsky [23] represented an asymptotic relationship between mutual information and Chernoff information for statistically independent neurons in a large population. However, Chernoff information is still hard to calculate in many practical applications.

Discrete stimuli or variables occur naturally in sensory coding. While some stimuli are continuous (e.g., the direction of movement, and the pitch of a tone), others are discrete (e.g., the identities of faces, and the words in human speech). For definiteness, in this paper, we frame our questions in the context of neural population coding; that is, we assume that the stimuli or the input variables are encoded by the pattern of responses elicited from a large population of neurons. The concrete examples used in our numerical simulations were based on Poisson spike model, where the response of each neuron is taken as the spike count within a given time window. While this simple Poisson model allowed us to consider a large neural population, it only captured the spike rate but not any temporal structure of the spike trains [25–28]. Nonetheless, our mathematical results are quite general and should be applicable to other input–output systems under suitable conditions to be discussed later.

In the following, we first derive several upper and lower bounds on Shannon mutual information using Kullback–Leibler divergence and Rényi divergence. Next, we derive several new approximation formulas for Shannon mutual information in the limit of large population size. These formulas are more convenient to calculate than the mutual information in our examples. Finally, we confirm the validity of our approximation formulas using the true mutual information as evaluated by Monte Carlo simulations.

2. Theory and Methods

2.1. Notations and Definitions

Suppose the input $\mathbf{x}$ is a K-dimensional vector, $\mathbf{x} = (x_1, \cdots, x_K)^T$, which could be interpreted as the parameters that specifies a stimulus for a sensory system, and the outputs is an N-dimensional vector, $\mathbf{r} = (r_1, \cdots, r_N)^T$, which could be interpreted as the responses of N neurons. We assume N is large, generally $N \gg K$. We denote random variables by upper case letters, e.g., random variables X and R, in contrast to their vector values $\mathbf{x}$ and $\mathbf{r}$. The mutual information between X and R is defined by

$$I = I(X; R) = \left\langle \ln \frac{p(\mathbf{r}|\mathbf{x})}{p(\mathbf{r})} \right\rangle_{\mathbf{r},\mathbf{x}}, \tag{1}$$

where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$, $\mathbf{r} \in \mathcal{R} \subseteq \mathbb{R}^N$, and $\langle \cdot \rangle_{\mathbf{r},\mathbf{x}}$ denotes the expectation with respect to the probability density function $p(\mathbf{r}, \mathbf{x})$. Similarly, in the following, we use $\langle \cdot \rangle_{\mathbf{r}|\mathbf{x}}$ and $\langle \cdot \rangle_{\mathbf{x}}$ to denote expectations with respect to $p(\mathbf{r}|\mathbf{x})$ and $p(\mathbf{x})$, respectively.

If $p(\mathbf{x})$ and $p(\mathbf{r}|\mathbf{x})$ are twice continuously differentiable for almost every $\mathbf{x} \in \mathcal{X}$, then for large N we can use an asymptotic formula to approximate the true value of I with high accuracy [24]:

$$I \simeq I_G = \frac{1}{2} \left\langle \ln \left(\det \left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e} \right) \right) \right\rangle_{\mathbf{x}} + H(X), \tag{2}$$

which is sometimes reduced to

$$I \simeq I_F = \frac{1}{2} \left\langle \ln \left(\det \left(\frac{\mathbf{J}(\mathbf{x})}{2\pi e} \right) \right) \right\rangle_{\mathbf{x}} + H(X), \tag{3}$$

where det $(\cdot)$ denotes the matrix determinant, $H(X) = -\langle \ln p(\mathbf{x}) \rangle_{\mathbf{x}}$ is the stimulus entropy,

$$\mathbf{G}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) + \mathbf{P}(\mathbf{x}), \tag{4}$$

$$\mathbf{P}(\mathbf{x}) = -\frac{\partial^2 \ln p(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T}, \tag{5}$$

and

$$\mathbf{J}(\mathbf{x}) = -\left\langle \frac{\partial^2 \ln p(\mathbf{r}|\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} \right\rangle_{\mathbf{r}|\mathbf{x}} = \left\langle \frac{\partial \ln p(\mathbf{r}|\mathbf{x})}{\partial \mathbf{x}} \frac{\partial \ln p(\mathbf{r}|\mathbf{x})}{\partial \mathbf{x}^T} \right\rangle_{\mathbf{r}|\mathbf{x}} \tag{6}$$

is the Fisher information matrix.

We denote the Kullback–Leibler divergence as

$$D(\mathbf{x}||\hat{\mathbf{x}}) = \left\langle \ln \frac{p(\mathbf{r}|\mathbf{x})}{p(\mathbf{r}|\hat{\mathbf{x}})} \right\rangle_{\mathbf{r}|\mathbf{x}}, \tag{7}$$

and denote Rényi divergence [29] of order $\beta + 1$ as

$$D_\beta(\mathbf{x}||\hat{\mathbf{x}}) = \frac{1}{\beta} \ln \left\langle \left(\frac{p(\mathbf{r}|\mathbf{x})}{p(\mathbf{r}|\hat{\mathbf{x}})} \right)^\beta \right\rangle_{\mathbf{r}|\mathbf{x}}. \tag{8}$$

Here, $\beta D_\beta(\mathbf{x}||\hat{\mathbf{x}})$ is equivalent to Chernoff divergence of order $\beta + 1$ [30]. It is well known that $D_\beta(\mathbf{x}||\hat{\mathbf{x}}) \to D(\mathbf{x}||\hat{\mathbf{x}})$ in the limit $\beta \to 0$.

We define

$$I_u = -\langle \ln \langle \exp(-D(\mathbf{x}||\hat{\mathbf{x}})) \rangle_{\hat{\mathbf{x}}} \rangle_{\mathbf{x}}, \tag{9}$$

$$I_e = -\left\langle \ln \left\langle \exp\left(-e^{-1} D(\mathbf{x}||\hat{\mathbf{x}})\right) \right\rangle_{\hat{\mathbf{x}}} \right\rangle_{\mathbf{x}}, \tag{10}$$

$$I_{\beta,\alpha} = -\left\langle \ln \left\langle \exp\left(-\beta D_\beta(\mathbf{x}||\hat{\mathbf{x}}) + (1-\alpha) \ln \frac{p(\mathbf{x})}{p(\hat{\mathbf{x}})}\right) \right\rangle_{\hat{\mathbf{x}}} \right\rangle_{\mathbf{x}}, \tag{11}$$

where in $I_{\beta,\alpha}$ we have $\beta \in (0,1)$ and $\alpha \in (0,\infty)$ and assume $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$.

In the following, we suppose $\mathbf{x}$ takes M discrete values, $\mathbf{x}_m$, $m \in \mathcal{M} = \{1, 2, \cdots, M\}$, and $p(\mathbf{x}_m) > 0$ for all m. Now, the definitions in Equations (9)–(11) become

$$I_u = -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln \left(\sum_{\hat{m}=1}^{M} \frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \exp(-D(\mathbf{x}_m||\mathbf{x}_{\hat{m}})) \right) + H(X), \tag{12}$$

$$I_e = -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln \left(\sum_{\hat{m}=1}^{M} \frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \exp\left(-e^{-1} D(\mathbf{x}_m||\mathbf{x}_{\hat{m}})\right) \right) + H(X), \tag{13}$$

$$I_{\beta,\alpha} = -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln \left(\sum_{\hat{m}=1}^{M} \left(\frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \right)^\alpha \exp(-\beta D_\beta(\mathbf{x}_m||\mathbf{x}_{\hat{m}})) \right) + H(X). \tag{14}$$

Furthermore, we define

$$I_d = - \sum_{m=1}^{M} p(\mathbf{x}_m) \ln \left(1 + \sum_{\hat{m} \in \mathcal{M}_m^u} \frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \exp \left(-e^{-1} D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right) \right) \right) + H(X), \tag{15}$$

$$I_u^d = - \sum_{m=1}^{M} p(\mathbf{x}_m) \ln \left(1 + \sum_{\hat{m} \in \mathcal{M}_m^u} \frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \exp \left(-D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right) \right) \right) + H(X), \tag{16}$$

$$I_{\beta,\alpha}^d = - \sum_{m=1}^{M} p(\mathbf{x}_m) \ln \left(1 + \sum_{\hat{m} \in \mathcal{M}_m^\beta} \left(\frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \right)^\alpha \exp \left(-\beta D_\beta\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right) \right) \right) + H(X), \tag{17}$$

$$I_D = - \sum_{m=1}^{M} p(\mathbf{x}_m) \ln \left(1 + \sum_{\hat{m} \in \mathcal{M}_m^u} \exp \left(-e^{-1} D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right) \right) \right) + H(X), \tag{18}$$

where

$$\check{\mathcal{M}}_m^\beta = \left\{ \hat{m} : \hat{m} = \operatorname*{arg\,min}_{\check{m} \in \mathcal{M} - \hat{\mathcal{M}}_m^\beta} D_\beta\left(\mathbf{x}_m || \mathbf{x}_{\check{m}}\right) \right\}, \tag{19}$$

$$\check{\mathcal{M}}_m^u = \left\{ \hat{m} : \hat{m} = \operatorname*{arg\,min}_{\check{m} \in \mathcal{M} - \hat{\mathcal{M}}_m^u} D\left(\mathbf{x}_m || \mathbf{x}_{\check{m}}\right) \right\}, \tag{20}$$

$$\hat{\mathcal{M}}_m^\beta = \left\{ \hat{m} : D_\beta\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right) = 0 \right\}, \tag{21}$$

$$\hat{\mathcal{M}}_m^u = \left\{ \hat{m} : D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right) = 0 \right\}, \tag{22}$$

$$\mathcal{M}_m^\beta = \check{\mathcal{M}}_m^\beta \cup \hat{\mathcal{M}}_m^\beta - \{m\}, \tag{23}$$

$$\mathcal{M}_m^u = \check{\mathcal{M}}_m^u \cup \hat{\mathcal{M}}_m^u - \{m\}. \tag{24}$$

Here, notice that, if $\mathbf{x}$ is uniformly distributed, then by definition I_d and I_D become identical. The elements in set $\check{\mathcal{M}}_m^\beta$ are those that make $D_\beta\left(\mathbf{x}_m || \mathbf{x}_{\check{m}}\right)$ take the minimum value, excluding any element that satisfies the condition $D_\beta\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right) = 0$. Similarly, the elements in set $\check{\mathcal{M}}_m^u$ are those that minimize $D\left(\mathbf{x}_m || \mathbf{x}_{\check{m}}\right)$ excluding the ones that satisfy the condition $D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right) = 0$.

2.2. Theorems

In the following, we state several conclusions as theorems and prove them in Appendix.

Theorem 1. *The mutual information I is bounded as follows:*

$$I_{\beta,\alpha} \leq I \leq I_u. \tag{25}$$

Theorem 2. *The following inequalities are satisfied,*

$$I_{\beta_1,1} \leq I_e \leq I_u \tag{26}$$

where $I_{\beta_1,1}$ is a special case of $I_{\beta,\alpha}$ in Equation (11) with $\beta_1 = e^{-1}$ so that

$$I_{\beta_1,1} = - \left\langle \ln \left\langle \exp \left(-\beta_1 D_{\beta_1}\left(\mathbf{x} || \hat{\mathbf{x}}\right) \right) \right\rangle_{\hat{\mathbf{x}}} \right\rangle_{\mathbf{x}}. \tag{27}$$

Theorem 3. *If there exist $\gamma_1 > 0$ and $\gamma_2 > 0$ such that*

$$\beta D_\beta\left(\mathbf{x}_m || \mathbf{x}_{m_1}\right) \geq \gamma_1 \ln N, \tag{28}$$

$$D\left(\mathbf{x}_m || \mathbf{x}_{m_2}\right) \geq \gamma_2 \ln N, \tag{29}$$

for discrete stimuli $\mathbf{x}_m$*, where* $m \in \mathcal{M}$*,* $m_1 \in \mathcal{M} - \mathcal{M}_m^{\beta}$ *and* $m_2 \in \mathcal{M} - \mathcal{M}_m^u$*, then we have the following asymptotic relationships:*

$$I_{\beta,\alpha} = I_{\beta,\alpha}^d + O\left(N^{-\gamma_1}\right) \leq I \leq I_u = I_u^d + O\left(N^{-\gamma_2}\right) \tag{30}$$

and

$$I_e = I_d + O\left(N^{-\gamma_2/e}\right). \tag{31}$$

Theorem 4. *Suppose* $p(\mathbf{x})$ *and* $p(\mathbf{r}|\mathbf{x})$ *are twice continuously differentiable for* $\mathbf{x} \in \mathcal{X}$*,* $\|q'(\mathbf{x})\| < \infty$*,* $\|q''(\mathbf{x})\| < \infty$*, where* $q(\mathbf{x}) = \ln p(\mathbf{x})$ *and* $'$ *and* $''$ *denote partial derivatives* $\partial/\partial\mathbf{x}$ *and* $\partial^2/\partial\mathbf{x}\partial\mathbf{x}^T$*, and* $\mathbf{G}_{\gamma}(\mathbf{x})$ *is positive definite with* $\left\|N\mathbf{G}_{\gamma}^{-1}(\mathbf{x})\right\| = O(1)$*, where* $\|\cdot\|$ *denotes matrix Frobenius norm,*

$$\mathbf{G}_{\gamma}(\mathbf{x}) = \gamma\left(\mathbf{J}(\mathbf{x}) + \mathbf{P}(\mathbf{x})\right), \tag{32}$$

$\gamma = \beta(1-\beta)$ *and* $\beta \in (0,1)$*. If there exist an* $\omega = \omega(\mathbf{x}) > 0$ *such that*

$$\det\left(\mathbf{G}(\mathbf{x})\right)^{1/2} \int_{\bar{\mathcal{X}}_{\varepsilon}(\mathbf{x})} p(\hat{\mathbf{x}}) \exp\left(-D\left(\mathbf{x}\|\hat{\mathbf{x}}\right)\right) d\hat{\mathbf{x}} = O\left(N^{-1}\right), \tag{33}$$

$$\det\left(\mathbf{G}_{\gamma}(\mathbf{x})\right)^{1/2} \int_{\bar{\mathcal{X}}_{\varepsilon}(\mathbf{x})} p(\hat{\mathbf{x}}) \exp\left(-\beta D_{\beta}\left(\mathbf{x}\|\hat{\mathbf{x}}\right)\right) d\hat{\mathbf{x}} = O\left(N^{-1}\right), \tag{34}$$

for all $\mathbf{x} \in \mathcal{X}$ *and* $\varepsilon \in (0,\omega)$*, where* $\bar{\mathcal{X}}_{\omega}(\mathbf{x}) = \mathcal{X} - \mathcal{X}_{\omega}(\mathbf{x})$ *is the complementary set of* $\mathcal{X}_{\omega}(\mathbf{x}) = \left\{\check{\mathbf{x}} \in \mathbb{R}^K : (\check{\mathbf{x}} - \mathbf{x})^T \mathbf{G}(\mathbf{x})(\check{\mathbf{x}} - \mathbf{x}) < N\omega^2\right\}$*, then we have the following asymptotic relationships:*

$$I_{\beta,\alpha} \leq I_{\gamma_0} + O\left(N^{-1}\right) \leq I \leq I_u = I_G + K/2 + O\left(N^{-1}\right), \tag{35}$$

$$I_e = I_G + O\left(N^{-1}\right), \tag{36}$$

$$I_{\beta,\alpha} = I_{\gamma} + O\left(N^{-1}\right), \tag{37}$$

where

$$I_{\gamma} = \frac{1}{2} \int_{\mathcal{X}} p(\mathbf{x}) \ln\left(\det\left(\frac{\mathbf{G}_{\gamma}(\mathbf{x})}{2\pi}\right)\right) d\mathbf{x} + H(X) \tag{38}$$

and $\gamma_0 = \beta_0(1-\beta_0) = 1/4$ *with* $\beta_0 = 1/2$*.*

Remark 1. *We see from Theorems 1 and 2 that the true mutual information I and the approximation* I_e *both lie between* $I_{\beta_1,1}$ *and* I_u*, which implies that their values may be close to each other. For discrete variable* $\mathbf{x}$*, Theorem 3 tells us that* I_e *and* I_d *are asymptotically equivalent (i.e., their difference vanishes) in the limit of large N. For continuous variable* $\mathbf{x}$*, Theorem 4 tells us that* I_e *and* I_G *are asymptotically equivalent in the limit of large N, which means that* I_e *and I are also asymptotically equivalent because* I_G *and I are known to be asymptotically equivalent [24].* □

Remark 2. *To see how the condition in Equation (33) could be satisfied, consider the case where* $D(\mathbf{x}\|\hat{\mathbf{x}})$ *has only one extreme point at* $\hat{\mathbf{x}} = \mathbf{x}$ *for* $\hat{\mathbf{x}} \in \mathcal{X}_{\omega}(\mathbf{x})$ *and there exists an* $\eta > 0$ *such that* $N^{-1}D(\mathbf{x}|\hat{\mathbf{x}}) \geq \eta$ *for* $\hat{\mathbf{x}} \in \bar{\mathcal{X}}_{\omega}(\mathbf{x})$*. Now, the condition in Equation (33) is satisfied because*

$$\det\left(\mathbf{G}(\mathbf{x})\right)^{1/2} \int_{\bar{\mathcal{X}}_{\varepsilon}(\mathbf{x})} p(\hat{\mathbf{x}}) \exp\left(-D\left(\mathbf{x}\|\hat{\mathbf{x}}\right)\right) d\hat{\mathbf{x}}$$

$$\leq \det\left(\mathbf{G}(\mathbf{x})\right)^{1/2} \int_{\bar{\mathcal{X}}_{\varepsilon}(\mathbf{x})} p(\hat{\mathbf{x}}) \exp\left(-\hat{\eta}(\varepsilon)N\right) d\hat{\mathbf{x}}$$

$$= O\left(N^{K/2} e^{-\hat{\eta}(\varepsilon)N}\right), \tag{39}$$

where by assumption we can find an $\hat{\eta}(\varepsilon) > 0$ for any given $\varepsilon \in (0, \omega)$. The condition in Equation (34) can be satisfied in a similar way. When $\beta_0 = 1/2$, $\beta_0 D_{\beta_0}(\mathbf{x}||\hat{\mathbf{x}})$ is the Bhattacharyya distance [31]:

$$\beta_0 D_{\beta_0}(\mathbf{x}||\hat{\mathbf{x}}) = -\ln\left(\int_{\mathcal{R}} \sqrt{p(\mathbf{r}|\mathbf{x})p(\mathbf{r}|\hat{\mathbf{x}})}d\mathbf{r}\right), \tag{40}$$

and we have

$$\mathbf{J}(\mathbf{x}) = \left.\frac{\partial^2 (D(\mathbf{x}||\hat{\mathbf{x}}))}{\partial\hat{\mathbf{x}}\partial\hat{\mathbf{x}}^T}\right|_{\hat{\mathbf{x}}=\mathbf{x}} = \left.\frac{\partial^2 (4\beta_0 D_{\beta_0}(\mathbf{x}||\hat{\mathbf{x}}))}{\partial\hat{\mathbf{x}}\partial\hat{\mathbf{x}}^T}\right|_{\hat{\mathbf{x}}=\mathbf{x}} = \left.\frac{\partial^2 (8H_l^2(\mathbf{x}||\hat{\mathbf{x}}))}{\partial\hat{\mathbf{x}}\partial\hat{\mathbf{x}}^T}\right|_{\hat{\mathbf{x}}=\mathbf{x}}, \tag{41}$$

where $H_l(\mathbf{x}||\hat{\mathbf{x}})$ is the Hellinger distance [32] between $p(\mathbf{r}|\mathbf{x})$ and $p(\mathbf{r}|\hat{\mathbf{x}})$:

$$H_l^2(\mathbf{x}||\hat{\mathbf{x}}) = \frac{1}{2}\int_{\mathcal{R}} \left(\sqrt{p(\mathbf{r}|\mathbf{x})} - \sqrt{p(\mathbf{r}|\hat{\mathbf{x}})}\right)^2 d\mathbf{r}. \tag{42}$$

By Jensen's inequality, for $\beta \in (0, 1)$ we get

$$0 \le D_\beta(\mathbf{x}||\hat{\mathbf{x}}) \le D(\mathbf{x}||\hat{\mathbf{x}}). \tag{43}$$

Denoting the Chernoff information [8] as

$$C(\mathbf{x}||\hat{\mathbf{x}}) = \max_{\beta \in (0,1)} \left(\beta D_\beta(\mathbf{x}||\hat{\mathbf{x}})\right) = \beta_m D_{\beta_m}(\mathbf{x}||\hat{\mathbf{x}}), \tag{44}$$

where $\beta D_\beta(\mathbf{x}||\hat{\mathbf{x}})$ achieves its maximum at β_m, we have

$$
\begin{aligned}
I_{\beta,\alpha} - H(X) & \\
\le h_c = -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln &\left(\sum_{\hat{m}=1}^{M} \frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \exp\left(-C(\mathbf{x}_m||\mathbf{x}_{\hat{m}})\right)\right) \tag{45} \\
\le h_d = -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln &\left(\sum_{\hat{m}=1}^{M} \frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \exp\left(-\beta_m D_\beta(\mathbf{x}_m||\mathbf{x}_{\hat{m}})\right)\right). \tag{46}
\end{aligned}
$$

By Theorem 4,

$$\max_{\beta \in (0,1)} I_{\beta,\alpha} = I_{\gamma_0} + O\left(N^{-1}\right), \tag{47}$$

$$I_{\gamma_0} = I_G - \frac{K}{2}\ln\frac{4}{e}. \tag{48}$$

If $\beta_m = 1/2$, then, by Equations (50), (46), (47) and (48), we have

$$\max_{\beta \in (0,1)} I_\beta + \frac{K}{2}\ln\frac{4}{e} + O\left(N^{-1}\right) \le I_e = I + O\left(N^{-1}\right)$$

$$\le h_d + H(X) \le I_u. \tag{49}$$

Therefore, from Equations (45), (46) and (49), we can see that I_e and I are close to $h_c + H(X)$. □

2.3. Approximations for Mutual Information

In this section, we use the relationships described above to find effective approximations to true mutual information I in the case of large but finite N. First, Theorems 1 and 2 tell us that the true mutual information I and its approximation I_e lie between lower and upper bounds given by:

$I_{\beta,\alpha} \leq I \leq I_u$ and $I_{\beta_1,1} \leq I_e \leq I_u$. As a special case, I is also bounded by $I_{\beta_1,1} \leq I \leq I_u$. Furthermore, from Equation (2) and (36) we can obtain the following asymptotic equality under suitable conditions:

$$I = I_e + O\left(N^{-1}\right). \tag{50}$$

Hence, for continuous stimuli, we have the following approximate relationship for large N:

$$I \simeq I_e \simeq I_G. \tag{51}$$

For discrete stimuli, by Equation (31) for large but finite N, we have

$$I \simeq I_e \simeq I_d = -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln\left(1 + \sum_{\hat{m} \in \mathcal{M}_m^u} \frac{p(\mathbf{x}_{\hat{m}})}{p(\mathbf{x}_m)} \exp\left(-e^{-1} D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right)\right)\right) + H(X). \tag{52}$$

Consider the special case $p(\mathbf{x}_{\hat{m}}) \simeq p(\mathbf{x}_m)$ for $\hat{m} \in \mathcal{M}_m^u$. With the help of Equation (18), substitution of $p(\mathbf{x}_{\hat{m}}) \simeq p(\mathbf{x}_m)$ into Equation (52) yields

$$
\begin{aligned}
I \simeq I_D &= -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln\left(1 + \sum_{\hat{m} \in \mathcal{M}_m^u} \exp\left(-e^{-1} D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right)\right)\right) + H(X) \\
&\simeq -\sum_{m=1}^{M} p(\mathbf{x}_m) \sum_{\hat{m} \in \mathcal{M}_m^u} \exp\left(-e^{-1} D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right)\right) + H(X) \\
&= I_D^0
\end{aligned}
\tag{53}
$$

where $I_D^0 \leq I_D$ and the second approximation follows from the first-order Taylor expansion assuming that the term $\sum_{\hat{m} \in \mathcal{M}_m^u} \exp\left(-e^{-1} D\left(\mathbf{x}_m || \mathbf{x}_{\hat{m}}\right)\right)$ is sufficiently small.

The theoretical discussion above suggests that I_e and I_d are effective approximations to true mutual information I in the limit of large N. Moreover, we find that they are often good approximations of mutual information I even for relatively small N, as illustrated in the following section.

3. Results of Numerical Simulations

Consider Poisson model neuron whose responses (i.e., numbers of spikes within a given time window) follow a Poisson distribution [24]. The mean response of neuron n, with $n \in \{1, 2, \cdots, N\}$, is described by the tuning function $f(x; \theta_n)$, which takes the form of a Heaviside step function:

$$f(x; \theta_n) = \begin{cases} A, & \text{if } x \geq \theta_n, \\ 0, & \text{if } x < \theta_n, \end{cases} \tag{54}$$

where the stimulus $x \in [-T, T]$ with $T = 10$, $A = 10$, and the centers $\theta_1, \theta_2, \cdots, \theta_N$ of the N neurons are uniformly spaced in interval $[-T, T]$, namely, $\theta_n = (n-1)d - T$ with $d = 2T/(N-1)$ for $N \geq 2$, and $\theta_n = 0$ for $N = 1$. We suppose that the discrete stimulus x has $M = 21$ possible values that are evenly spaced from $-T$ to T, namely, $x \in \mathcal{X} = \{x_m : x_m = 2(m-1)T/(M-1) - T, m = 1, 2, \cdots, M\}$. Now, the Kullback–Leibler divergence can be written as

$$D(x_m || x_{\hat{m}}) = f(x_m; \theta_n) \log\left(\frac{f(x_m; \theta_n)}{f(x_{\hat{m}}; \theta_n)}\right) + f(x_{\hat{m}}; \theta_n) - f(x_m; \theta_n). \tag{55}$$

Thus, we have $\exp\left(-e^{-1}D\left(x_m||x_{\hat{m}}\right)\right) = 1$ when $f\left(x_m;\theta_n\right) = f\left(x_{\hat{m}};\theta_n\right)$, $\exp\left(-e^{-1}D\left(x_m||x_{\hat{m}}\right)\right) = \exp\left(-e^{-1}A\right)$ when $f\left(x_m;\theta_n\right) = 0$ and $f\left(x_{\hat{m}};\theta_n\right) = A$, and $\exp\left(-e^{-1}D\left(x_m||x_{\hat{m}}\right)\right) = 0$ when $f\left(x_m;\theta_n\right) = A$ and $f\left(x_{\hat{m}};\theta_n\right) = 0$. Therefore, in this case, we have

$$I_e = I_d. \tag{56}$$

More generally, this equality holds true whenever the tuning function has binary values.

In the first example, as illustrated in Figure 1, we suppose the stimulus has a uniform distribution, so that the probability is given by $p(x_m) = 1/M$. Figure 1a shows graphs of the input distribution $p(x)$ and a representative tuning function $f(x;\theta)$ with the center $\theta = 0$.

To assess the accuracy of the approximation formulas, we employed Monte Carlo (MC) simulation to evaluate the mutual information I [24]. In our MC simulation, we first sampled an input $x_j \in \mathcal{X}$ from the uniform distribution $p(x_j) = 1/M$, then generated the neural responses $\mathbf{r}_j$ by the conditional distribution $p(\mathbf{r}_j|x_j)$ based on the Poisson model, where $j = 1, 2, \cdots, j_{\max}$. The value of mutual information by MC simulation was calculated by

$$I_{MC}^* = \frac{1}{j_{\max}} \sum_{j=1}^{j_{\max}} \ln\left(\frac{p(\mathbf{r}_j|x_j)}{p(\mathbf{r}_j)}\right), \tag{57}$$

where

$$p(\mathbf{r}_j) = \sum_{m=1}^{M} p(\mathbf{r}_j|x_m)p(x_m). \tag{58}$$

To assess the precision of our MC simulation, we computed the standard deviation of repeated trials by bootstrapping:

$$I_{std} = \sqrt{\frac{1}{i_{\max}} \sum_{i=1}^{i_{\max}} \left(I_{MC}^i - I_{MC}\right)^2}, \tag{59}$$

where

$$I_{MC}^i = \frac{1}{j_{\max}} \sum_{j=1}^{j_{\max}} \ln\left(\frac{p(\mathbf{r}_{\Gamma_{j,i}}|x_{\Gamma_{j,i}})}{p(\mathbf{r}_{\Gamma_{j,i}})}\right), \tag{60}$$

$$I_{MC} = \frac{1}{i_{\max}} \sum_{i=1}^{i_{\max}} I_{MC}^i, \tag{61}$$

and $\Gamma_{j,i} \in \{1, 2, \cdots, j_{\max}\}$ is the (j,i)-th entry of the matrix $\mathbf{\Gamma} \in \mathbb{N}^{j_{\max} \times i_{\max}}$ with samples taken randomly from the integer set $\{1, 2, \cdots, j_{\max}\}$ by a uniform distribution. Here, we set $j_{\max} = 5 \times 10^5$, $i_{\max} = 100$ and $M = 10^3$.

For different $N \in \{1, 2, 3, 4, 6, 10, 14, 20, 30, 50, 100, 200, 400, 700, 1000\}$, we compared I_{MC} with I_e, I_d and I_D, as illustrated in Figure 1b–d. Here, we define the relative error of approximation, e.g., for I_e, as

$$DI_e = \frac{I_e - I_{MC}}{I_{MC}}, \tag{62}$$

and the relative standard deviation

$$DI_{std} = \frac{I_{std}}{I_{MC}}. \tag{63}$$
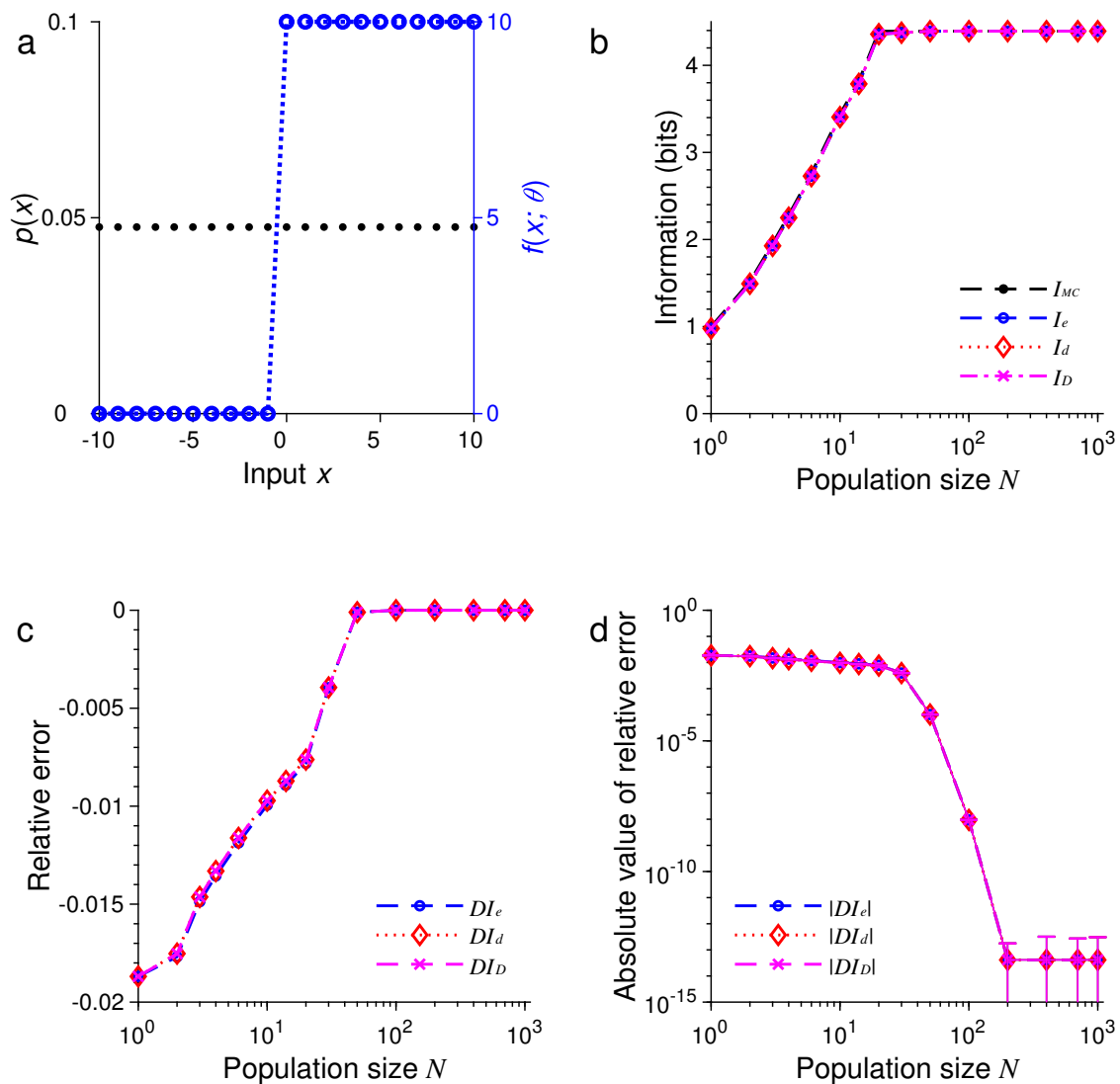
Figure 1. A comparison of approximations I_e, I_d and I_D against I_{MC} obtained by Monte Carlo method for one-dimensional discrete stimuli. (**a**) Discrete uniform distribution of the stimulus $p(x)$ (black dots) and the Heaviside step tuning function $f(x;\theta)$ with center $\theta = 0$ (blue dashed lines) ; (**b**) The values of I_{MC}, I_e, I_d and I_D depend on the population size or total number of neurons N ; (**c**) The relative errors DI_e, DI_d and DI_D for the results in (b) ; (**d**) The absolute values of the relative errors $|DI_e|$, $|DI_d|$ and $|DI_D|$ as in (c), with error bars showing standard deviations of repeated trials.

For the second example, we only changed the probability distribution of stimulus $p(x_m)$ while keeping all other conditions unchanged. Now, $p(x_m)$ is a discrete sample from a Gaussian function:

$$p(x_m) = Z^{-1} \exp\left(-\frac{x_m^2}{2\sigma^2}\right), m = 1, 2, \cdots, M, \tag{64}$$

where Z is the normalization constant and $\sigma = T/2$. The results are illustrated in Figure 2.
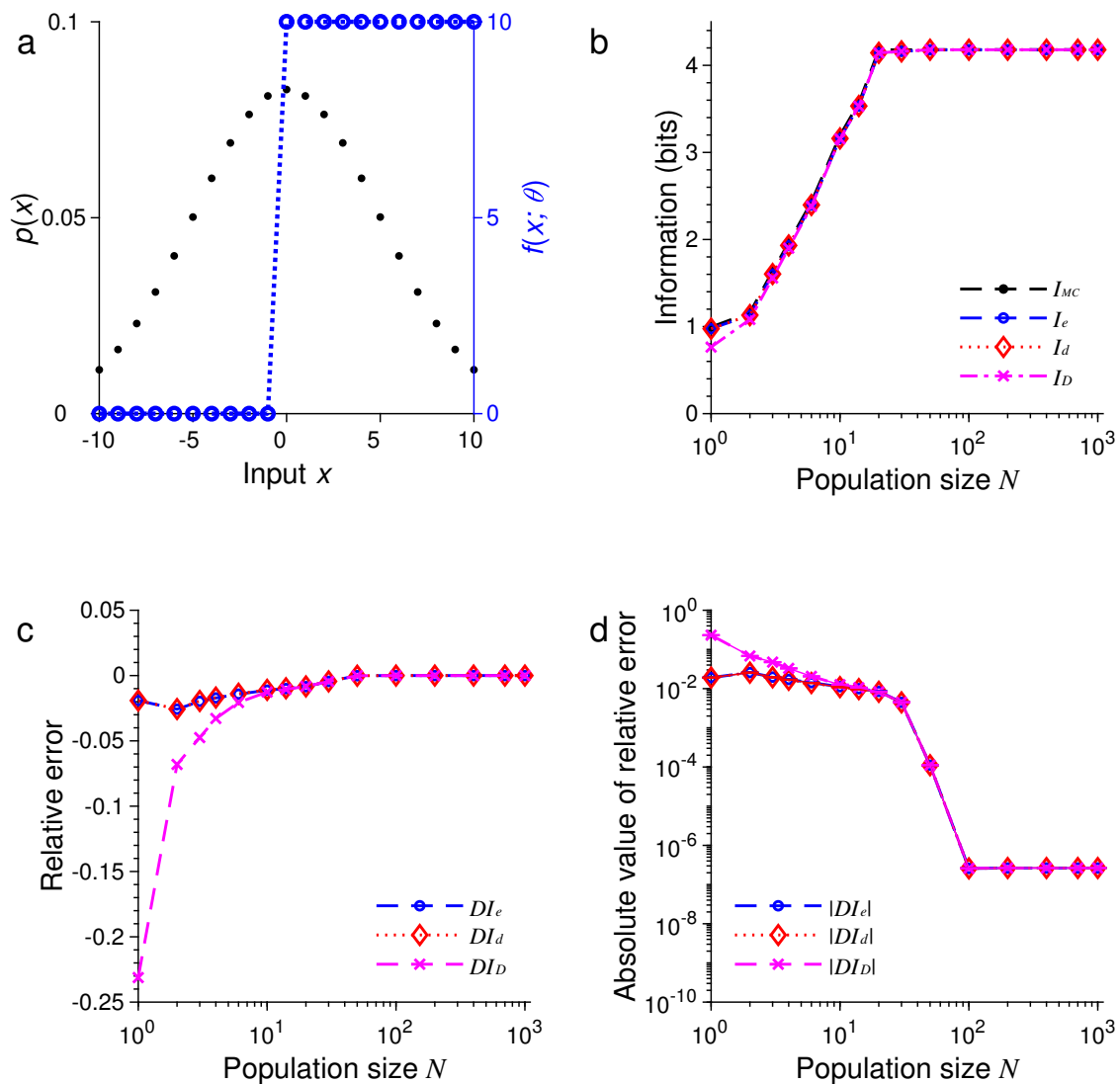
Figure 2. A comparison of approximations I_e, I_d and I_D against I_{MC}. The situation is identical to that in Figure 1 except that the stimulus distribution $p(x)$ is peaked rather flat (black dots in (a)). (**a**) Discrete Gaussian-like distribution of the stimulus $p(x)$ (black dots) and the Heaviside step tuning function $f(x;\theta)$ with center $\theta = 0$ (blue dashed lines); (**b**) The values of I_{MC}, I_e, I_d and I_D depend on the population size or total number of neurons N; (**c**) The relative errors DI_e, DI_d and DI_D for the results in (b); (**d**) The absolute values of the relative errors $|DI_e|$, $|DI_d|$ and $|DI_D|$ as in (c), with error bars showing standard deviations of repeated trials.

Next, we changed each tuning function $f(x;\theta_n)$ to a rectified linear function:

$$f(x;\theta_n) = \max(0, x - \theta_n), \tag{65}$$

Figures 3 and 4 show the results under the same conditions of Figures 1 and 2 except for the shape of the tuning functions.
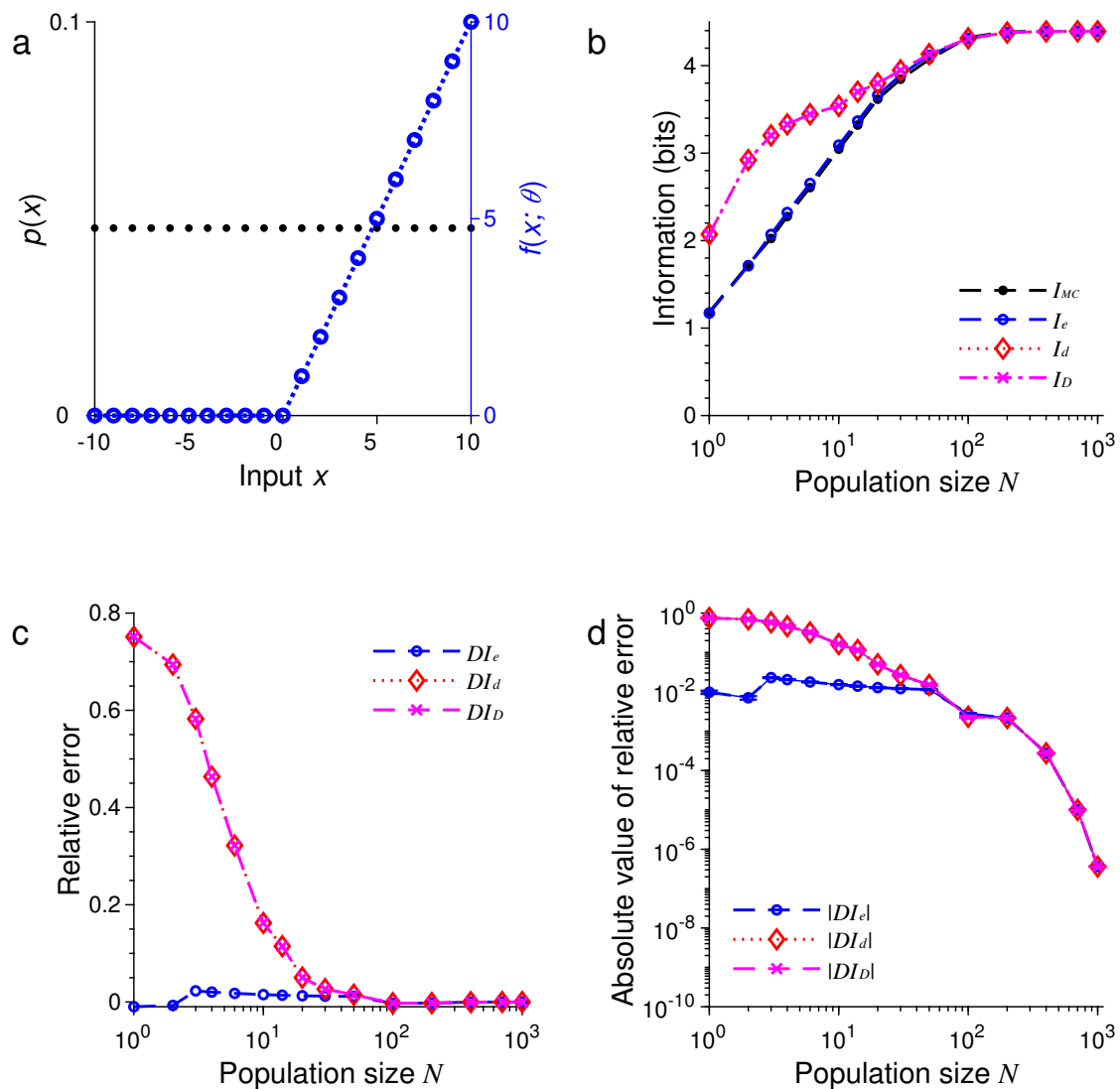
Figure 3. A comparison of approximations I_e, I_d and I_D against I_{MC}. The situation is identical to that in Figure 1 except for the shape of the tuning function (blue dashed lines in (a)). (**a**) Discrete uniform distribution of the stimulus $p(x)$ (black dots) and the rectified linear tuning function $f(x;\theta)$ with center $\theta = 0$ (blue dashed lines); (**b**) The values of I_{MC}, I_e, I_d and I_D depend on the population size or total number of neurons N; (**c**) The relative errors DI_e, DI_d and DI_D for the results in (b); (**d**) The absolute values of the relative errors $|DI_e|$, $|DI_d|$ and $|DI_D|$ as in (c), with error bars showing standard deviations of repeated trials.

Finally, we let the tuning function $f(x;\theta_n)$ have a random form:

$$f(x;\theta_n) = \begin{cases} B, & \text{if } x \in \theta_n = \{\theta_n^1, \theta_n^2, \cdots, \theta_n^K\}, \\ 0, & \text{otherwise}, \end{cases} \tag{66}$$

where the stimulus $x \in \mathcal{X} = \{1, 2, \cdots, 999, 1000\}$, $B = 10$, the values of $\{\theta_n^1, \theta_n^2, \cdots, \theta_n^K\}$ are distinct and randomly selected from the set $\mathcal{X}$ with $K = 10$. In this example, we may regard $\mathcal{X}$ as a list of natural objects (stimuli), and there are a total of N sensory neurons, each of which responds only to K randomly selected objects. Figure 5 shows the results under the condition that $p(x)$ is a uniform distribution. In Figure 6, we assume that $p(x)$ is not flat but a half Gaussian given by Equation (64) with $\sigma = 500$.
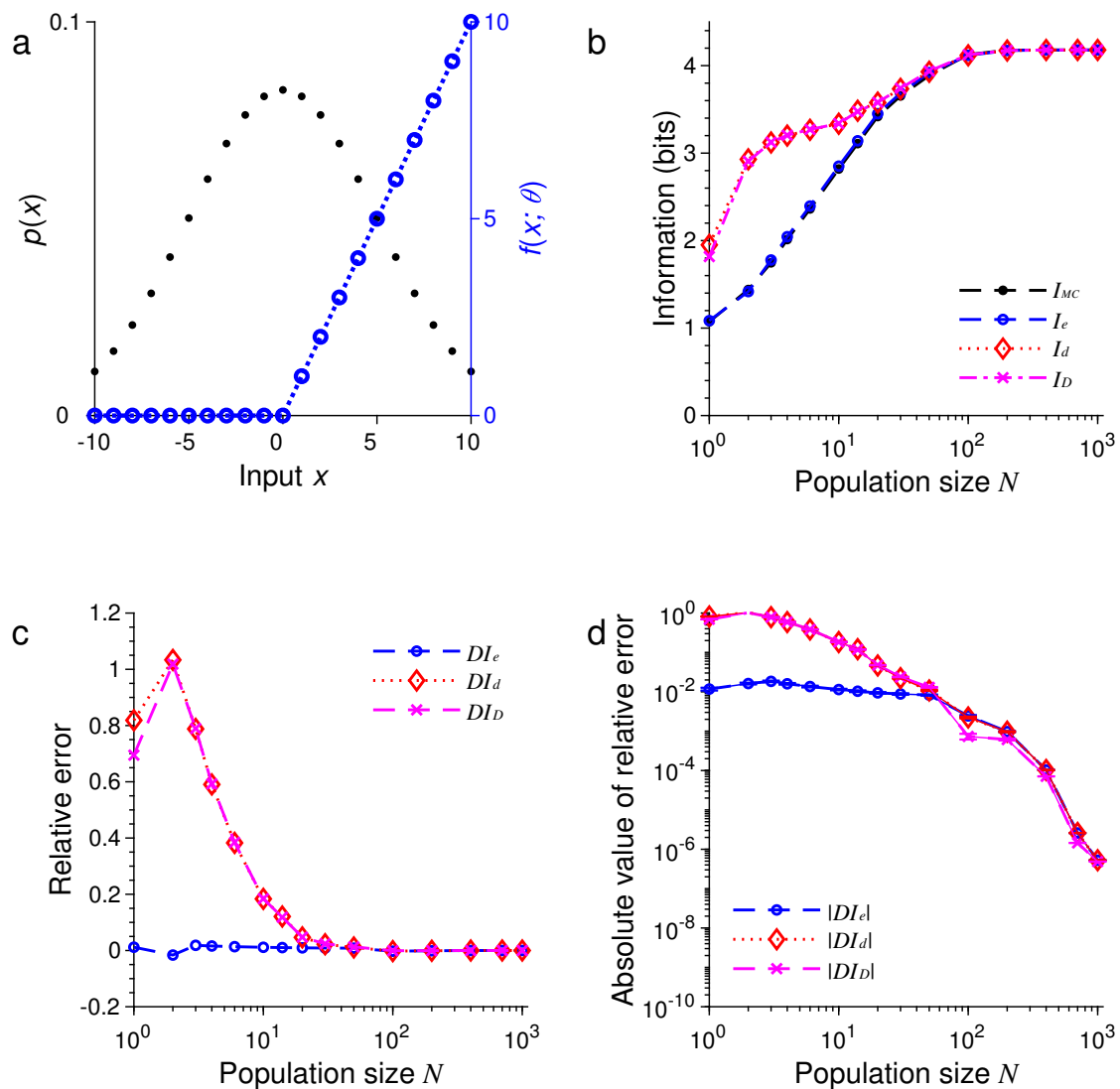
Figure 4. A comparison of approximations I_e, I_d and I_D against I_{MC}. The situation is identical to that in Figure 3 except that the stimulus distribution $p(x)$ is peaked rather flat (black dots in (a)). (**a**) Discrete Gaussian-like distribution of the stimulus $p(x)$ (black dots) and the rectified linear tuning function $f(x;\theta)$ with center $\theta = 0$ (blue dashed lines); (**b**) The values of I_{MC}, I_e, I_d and I_D depend on the population size or total number of neurons N; (**c**) The relative errors DI_e, DI_d and DI_D for the results in (b); (**d**) The absolute values of the relative errors $|DI_e|$, $|DI_d|$ and $|DI_D|$ as in (c), with error bars showing standard deviations of repeated trials.

In all these examples, we found that the three formulas, namely, I_e, I_d and I_D, provided excellent approximations to the true values of mutual information as evaluated by Monte Carlo method. For example, in the examples in Figures 1 and 5, all three approximations were practically indistinguishable. In general, all these approximations were extremely accurate when $N > 100$.

In all our simulations, the mutual information tended to increase with the population size N, eventually reaching a plateau for large enough N. The saturation of information for large N is due to the fact that it requires at most $\log_2 M$ bits of information to completely distinguish all M stimuli. It is impossible to gain more information than this maximum amount regardless of how many neurons are used in the population. In Figure 1, for instance, this maximum is $\log_2 21 = 4.39$ bits, and in Figure 5, this maximum is $\log_2 1000 = 9.97$ bits.
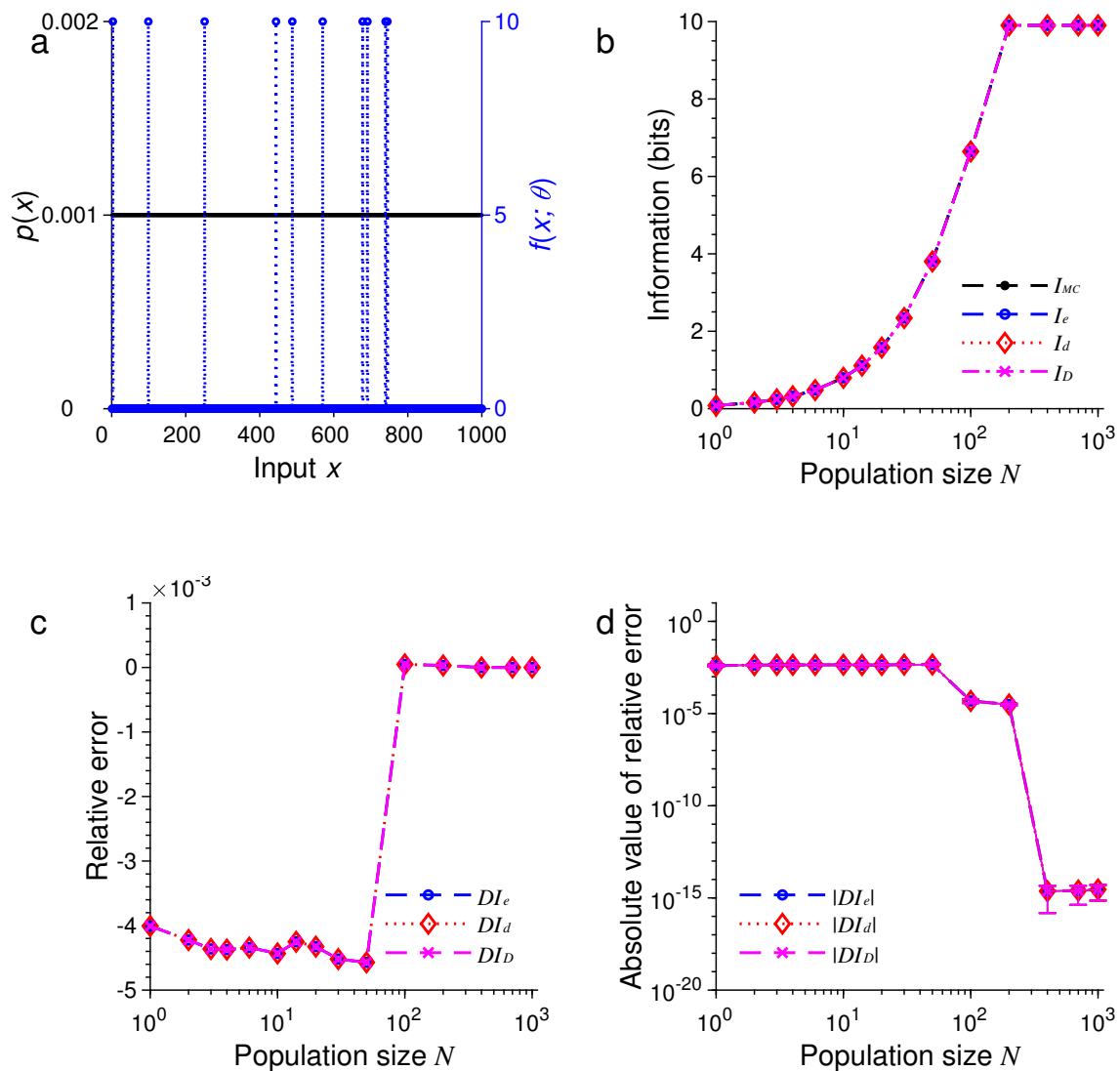
Figure 5. A comparison of approximations I_e, I_d and I_D against I_{MC}. The situation is similar to that in Figure 1 except that the tuning function is random (blue dashed lines in (a)); see Equation (66). (**a**) Discrete uniform distribution of the stimulus $p(x)$ (black dots) and the random tuning function $f(x; \theta)$; (**b**) The values of I_{MC}, I_e, I_d and I_D depend on the population size or total number of neurons N; (**c**) The relative errors DI_e, DI_d and DI_D for the results in (b); (**d**) The absolute values of the relative errors $|DI_e|$, $|DI_d|$ and $|DI_D|$ as in (c), with error bars showing standard deviations of repeated trials.

For relatively small values of N, we found that I_D tended to be less accurate than I_e or I_d (see Figures 5 and 6). Our simulations also confirmed two analytical results. The first one is that $I_d = I_D$ when the stimulus distribution is uniform; this result follows directly from the definitions of I_d and I_D and is confirmed by the simulations in Figures 1, 3, and 5. The second result is that $I_d = I_e$ (Equation (56)) when the tuning function is binary, as confirmed by the simulations in Figures 1, 2, 5, and 6. When the tuning function allows many different values, I_e can be much more accurate than I_d and I_D, as shown by the simulations in Figures 3 and 4. To summarize, our best approximation formula is I_e because it is more accurate than I_d and I_D, and, unlike I_d and I_D, it applies to both discrete and continuous stimuli (Equations (10) and (13)).
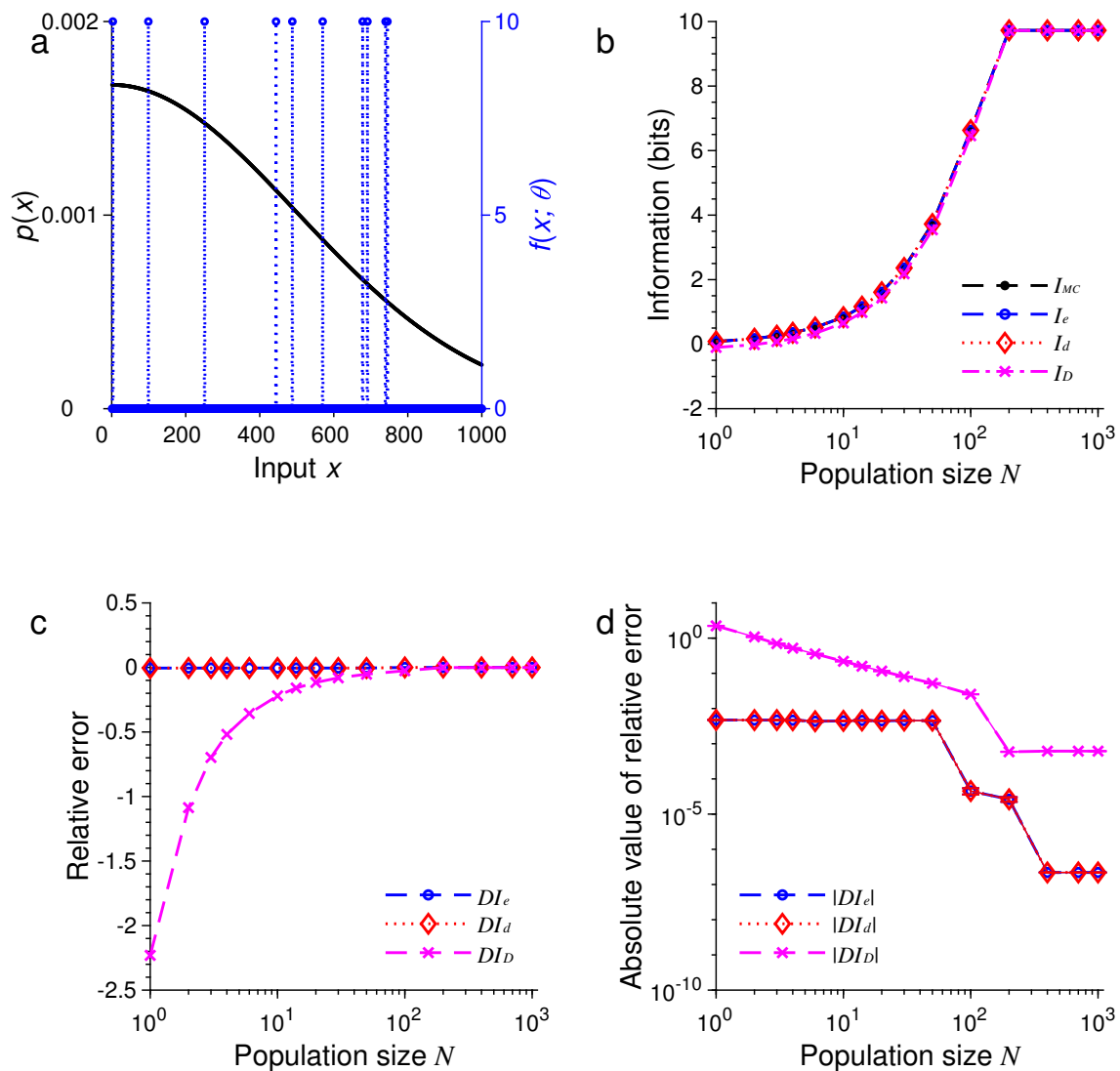
Figure 6. A comparison of approximations I_e, I_d and I_D against I_{MC}. The situation is identical to that in Figure 5 except that the stimulus distribution $p(x)$ is not flat (black dots in (a)). **(a)** Discrete Gaussian-like distribution of the stimulus $p(x)$ (black dots) and the random tuning function $f(x;\theta)$; **(b)** The values of I_{MC}, I_e, I_d and I_D depend on the population size or total number of neurons N; **(c)** The relative errors DI_e, DI_d and DI_D for the results in (b); **(d)** The absolute values of the relative errors $|DI_e|$, $|DI_d|$ and $|DI_D|$ as in (c), with error bars showing standard deviations of repeated trials.

4. Discussion

We have derived several asymptotic bounds and effective approximations of mutual information for discrete variables and established several relationships among different approximations. Our final approximation formulas involve only Kullback–Leibler divergence, which is often easier to evaluate than Shannon mutual information in practical applications. Although in this paper our theory is developed in the framework of neural population coding with concrete examples, our mathematical results are generic and should hold true in many related situations beyond the original context.

We propose to approximate the mutual information with several asymptotic formulas, including I_e in Equation (10) or Equation (13), I_d in Equation (15) and I_D in Equation (18). Our numerical experimental results show that the three approximations I_e, I_d and I_D were very accurate for large population size N, and sometimes even for relatively small N. Among the three approximations, I_D

tended to be the least accurate, although, as a special case of I_d, it is slightly easier to evaluate than I_d. For a comparison of I_e and I_d, we note that I_e is the universal formula, whereas I_d is restricted only to discrete variables. The two formulas I_e and I_d become identical when the responses or the tuning functions have only two values. For more general tuning functions, the performance of I_e was better than I_d in our simulations.

As mentioned before, an advantage of of I_e is that it works not only for discrete stimuli but also for continuous stimuli. Theoretically speaking, the formula for I_e is well justified, and we have proven that it approaches the true mutual information I in the limit of large population. In our numerical simulations, the performance of I_e was excellent and better than that of I_d and I_D. Overall, I_e is our most accurate and versatile approximation formula, although, in some cases, I_d and I_D are slightly more convenient to calculate.

The numerical examples considered in this paper were based on an independent population of neurons whose responses have Poisson statistics. Although such models are widely used, they are appropriate only if the neural responses can be well characterized by the spike counts within a fixed time window. To study the temporal patterns of spike trains, one has to consider more complicated models. Estimation of mutual information from neural spike trains is a difficult computational problem [25–28]. In future work, it would be interesting to apply the asymptotic formulas such as I_e to spike trains with small time bins each containing either one spike or nothing. A potential advantage of the asymptotic formula is that it might help reduce the bias caused by small samples in the calculation of the response marginal distribution $p(\mathbf{r}) = \sum_\mathbf{x} p(\mathbf{r}|\mathbf{x})p(\mathbf{x})$ or the response entropy $H(R)$ because here one only needs to calculate the Kullback–Leibler divergence $D(\mathbf{x}||\hat{\mathbf{x}})$, which may have a smaller estimation error.

Finding effective approximation methods for computing mutual information is a key step for many practical applications of the information theory. Generally speaking, Kullback–Leibler divergence (Equation (7)) is often easier to evaluate and approximate than either Chernoff information (Equation (44)) or Shannon mutual information (Equation (1)). In situations where this is indeed the case, our approximation formulas are potentially useful. Besides applications in numerical simulations, the availability of a set of approximation formulas may also provide helpful theoretical tools in future analytical studies of information coding and representations.

As mentioned in the Introduction, various methods have been proposed to approximate the mutual information [9–16]. In future work, it would be useful to compare different methods rigorously under identical conditions in order to asses their relative merits. The approximation formulas developed in this paper are relatively easy to compute for practical problems. They are especially suitable for analytical purposes; for example, they could be used explicitly as objective functions for optimization or learning algorithms. Although the examples used in our simulations in this paper are parametric, it should be possible to extend the formulas to nonparametric problem, possibly with help of the copula method to take advantage of its robustness in nonparametric estimations [16].

Author Contributions: W.H. developed and proved the theorems, programmed the numerical experiments and wrote the manuscript. K.Z. verified the proofs and revised the manuscript.

Funding: This research was supported by an NIH grant R01 DC013698.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Proofs

Appendix A.1. Proof of Theorem 1

By Jensen's inequality, we have

$$I_{\beta,\alpha} = -\left\langle \ln \left(\int_{\mathcal{X}} \left\langle \frac{p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})}{p^{\beta}(\mathbf{r}|\mathbf{x})p^{\alpha}(\mathbf{x})} \right\rangle_{\mathbf{r}|\mathbf{x}} d\hat{\mathbf{x}} \right) \right\rangle_{\mathbf{x}} + H(X)$$

$$\leq -\left\langle \left\langle \ln \left(\int_{\mathcal{X}} \frac{p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})}{p^{\beta}(\mathbf{r}|\mathbf{x})p^{\alpha}(\mathbf{x})} d\hat{\mathbf{x}} \right) \right\rangle_{\mathbf{r}|\mathbf{x}} \right\rangle_{\mathbf{x}} + H(X) \tag{A1}$$

and

$$-\left\langle \left\langle \ln \left(\int_{\mathcal{X}} \frac{p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})}{p^{\beta}(\mathbf{r}|\mathbf{x})p^{\alpha}(\mathbf{x})} d\hat{\mathbf{x}} \right) \right\rangle_{\mathbf{r}|\mathbf{x}} \right\rangle_{\mathbf{x}} + H(X) - I$$

$$= \left\langle \left\langle \ln \left(\int_{\mathcal{X}} \frac{p(\mathbf{r},\hat{\mathbf{x}})}{p(\mathbf{r},\mathbf{x})} d\hat{\mathbf{x}} \right) \left(\int_{\mathcal{X}} \frac{p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})}{p^{\beta}(\mathbf{r}|\mathbf{x})p^{\alpha}(\mathbf{x})} d\hat{\mathbf{x}} \right)^{-1} \right\rangle_{\mathbf{r}|\mathbf{x}} \right\rangle_{\mathbf{x}}$$

$$\leq \ln \int_{\mathcal{R}} p(\mathbf{r}) \frac{\int_{\mathcal{X}} p^{\beta}(\mathbf{r}|\mathbf{x})p^{\alpha}(\mathbf{x})d\mathbf{x}}{\int_{\mathcal{X}} p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})d\hat{\mathbf{x}}} d\mathbf{r}$$

$$= 0. \tag{A2}$$

Combining Equations (A1) and (A2), we immediately get the lower bound in Equation (25).

In this section, we use integral for variable $\mathbf{x}$, although our argument is valid for both continuous variables and discrete variables. For discrete variables, we just need to replace each integral by a summation, and our argument remains valid without other modification. The same is true for the response variable $\mathbf{r}$.

To prove the upper bound, let

$$\Phi[q(\hat{\mathbf{x}})] = \int_{\mathcal{R}} p(\mathbf{r}|\mathbf{x}) \int_{\mathcal{X}} q(\hat{\mathbf{x}}) \ln \left(\frac{p(\mathbf{r}|\mathbf{x})q(\hat{\mathbf{x}})}{p(\mathbf{r}|\hat{\mathbf{x}})p(\hat{\mathbf{x}})} \right) d\hat{\mathbf{x}} d\mathbf{r}, \tag{A3}$$

where $q(\hat{\mathbf{x}})$ satisfies

$$\begin{cases} \int_{\mathcal{X}} q(\hat{\mathbf{x}})d\hat{\mathbf{x}} = 1 \\ q(\hat{\mathbf{x}}) \geq 0 \end{cases}. \tag{A4}$$

By Jensen's inequality, we get

$$\Phi[q(\hat{\mathbf{x}})] \geq \int_{\mathcal{R}} p(\mathbf{r}|\mathbf{x}) \ln \left(\frac{p(\mathbf{r}|\mathbf{x})}{p(\mathbf{r})} \right) d\mathbf{r}. \tag{A5}$$

To find a function $q(\hat{\mathbf{x}})$ that minimizes $\Phi[q(\hat{\mathbf{x}})]$, we apply the variational principle as follows:

$$\frac{\partial \tilde{\Phi}[q(\hat{\mathbf{x}})]}{\partial q(\hat{\mathbf{x}})} = \int_{\mathcal{R}} p(\mathbf{r}|\mathbf{x}) \ln \left(\frac{p(\mathbf{r}|\mathbf{x})q(\hat{\mathbf{x}})}{p(\mathbf{r}|\hat{\mathbf{x}})p(\hat{\mathbf{x}})} \right) d\mathbf{r} + 1 + \lambda, \tag{A6}$$

where λ is the Lagrange multiplier and

$$\tilde{\Phi}[q(\hat{\mathbf{x}})] = \Phi[q(\hat{\mathbf{x}})] + \lambda \left(\int_{\mathcal{X}} q(\hat{\mathbf{x}})d\hat{\mathbf{x}} - 1 \right). \tag{A7}$$

Setting $\frac{\partial \tilde{\Phi}[q(\hat{\mathbf{x}})]}{\partial q(\hat{\mathbf{x}})} = 0$ and using the constraint in Equation (A4), we find the optimal solution

$$q^*(\hat{\mathbf{x}}) = \frac{p(\hat{\mathbf{x}}) \exp\left(-D\left(\mathbf{x}||\hat{\mathbf{x}}\right)\right)}{\int_{\mathcal{X}} p(\check{\mathbf{x}}) \exp\left(-D\left(\mathbf{x}||\check{\mathbf{x}}\right)\right) d\check{\mathbf{x}}}. \tag{A8}$$

Thus, the variational lower bound of $\Phi\left[q(\hat{\mathbf{x}})\right]$ is given by

$$\Phi\left[q^*(\hat{\mathbf{x}})\right] = \min_{q(\hat{\mathbf{x}})} \Phi\left[q(\hat{\mathbf{x}})\right] = -\ln\left(\int_{\mathcal{X}} p(\hat{\mathbf{x}}) \exp\left(-D\left(\mathbf{x}||\hat{\mathbf{x}}\right)\right) d\hat{\mathbf{x}}\right) d\mathbf{x}. \tag{A9}$$

Therefore, from Equations (1), (A5) and (A9), we get the upper bound in Equation (25). This completes the proof of Theorem 1. □

Appendix A.2. Proof of Theorem 2

It follows from Equation (43) that

$$
\begin{aligned}
I_{\beta_1,\alpha_1} &= -\left\langle \ln\left\langle \exp\left(-\beta_1 D_{\beta_1}\left(\mathbf{x}||\hat{\mathbf{x}}\right) + (1-\alpha_1)\ln\frac{p\left(\mathbf{x}\right)}{p\left(\hat{\mathbf{x}}\right)}\right)\right\rangle_{\hat{\mathbf{x}}}\right\rangle_{\mathbf{x}} \\
&\leq -\left\langle \ln\left\langle \exp\left(-e^{-1}D\left(\mathbf{x}||\hat{\mathbf{x}}\right)\right)\right\rangle_{\hat{\mathbf{x}}}\right\rangle_{\mathbf{x}} = I_e \\
&\leq -\left\langle \ln\left\langle \exp\left(-D\left(\mathbf{x}||\hat{\mathbf{x}}\right)\right)\right\rangle_{\hat{\mathbf{x}}}\right\rangle_{\mathbf{x}} = I_u,
\end{aligned}
\tag{A10}
$$

where $\beta_1 = e^{-1}$ and $\alpha_1 = 1$. We immediately get Equation (26). This completes the proof of Theorem 2. □

Appendix A.3. Proof of Theorem 3

For the lower bound $I_{\beta,\alpha}$, we have

$$
\begin{aligned}
I_{\beta,\alpha} &= -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln\left(\sum_{\check{m}=1}^{M}\left(\frac{p\left(\mathbf{x}_{\check{m}}\right)}{p\left(\mathbf{x}_m\right)}\right)^{\alpha} \exp\left(-\beta D_{\beta}\left(\mathbf{x}_m|\mathbf{x}_{\check{m}}\right)\right)\right) \\
&= -\sum_{m=1}^{M} p(\mathbf{x}_m) \ln\left(1 + d\left(\mathbf{x}_m\right)\right) + H\left(X\right),
\end{aligned}
\tag{A11}
$$

where

$$d\left(\mathbf{x}_m\right) = \sum_{\check{m}\in\mathcal{M}-\{m\}}\left(\frac{p\left(\mathbf{x}_{\check{m}}\right)}{p\left(\mathbf{x}_m\right)}\right)^{\alpha} \exp\left(-\beta D_{\beta}\left(\mathbf{x}_m|\mathbf{x}_{\check{m}}\right)\right). \tag{A12}$$

Now, consider

$$
\begin{aligned}
&\ln\left(1 + d\left(\mathbf{x}_m\right)\right) \\
&= \ln\left(1 + a\left(\mathbf{x}_m\right) + b\left(\mathbf{x}_m\right)\right) \\
&= \ln\left(1 + a\left(\mathbf{x}_m\right)\right) + \ln\left(1 + b\left(\mathbf{x}_m\right)\left(1 + a\left(\mathbf{x}_m\right)\right)^{-1}\right) \\
&= \ln\left(1 + a\left(\mathbf{x}_m\right)\right) + O\left(N^{-\gamma}\right),
\end{aligned}
\tag{A13}
$$

where

$$a\left(\mathbf{x}_m\right) = \sum_{\hat{m}\in\mathcal{M}_m^{\beta}}\left(\frac{p\left(\mathbf{x}_{\hat{m}}\right)}{p\left(\mathbf{x}_m\right)}\right)^{\alpha} \exp\left(-\beta D_{\beta}\left(\mathbf{x}_m||\mathbf{x}_{\hat{m}}\right)\right), \tag{A14a}$$

$$
\begin{aligned}
b\left(\mathbf{x}_m\right) &= \sum_{\check{m}\in\mathcal{M}-\mathcal{M}_m^{\beta}}\left(\frac{p\left(\mathbf{x}_{\check{m}}\right)}{p\left(\mathbf{x}_m\right)}\right)^{\alpha} \exp\left(-\beta D_{\beta}\left(\mathbf{x}_m||\mathbf{x}_{\check{m}}\right)\right) \\
&\leq N^{-\gamma_1}\sum_{\check{m}\in\mathcal{M}-\mathcal{M}_m^{\beta}}\left(\frac{p\left(\mathbf{x}_{\check{m}}\right)}{p\left(\mathbf{x}_m\right)}\right)^{\alpha} = O\left(N^{-\gamma_1}\right).
\end{aligned}
\tag{A14b}
$$

Combining Equations (A11) and (A13) and Theorem 1, we get the lower bound in Equation (30). In a manner similar to the above, we can get the upper bound in Equations (30) and (31). This completes the proof of Theorem 3. □

Appendix A.4. Proof of Theorem 4

The upper bound I_u for mutual information I in Equation (25) can be written as

$$
\begin{aligned}
I_u &= -\int_{\mathcal{X}} \left(\ln \int_{\mathcal{X}} p\left(\hat{\mathbf{x}}\right) \exp\left(-D\left(\mathbf{x}|\hat{\mathbf{x}}\right)\right) d\hat{\mathbf{x}} \right) p\left(\mathbf{x}\right) d\mathbf{x} \\
&= -\left\langle \ln \left(\int_{\mathcal{X}} \exp\left(\langle L(\mathbf{r}|\hat{\mathbf{x}}) - L(\mathbf{r}|\mathbf{x}) \rangle_{\mathbf{r}|\mathbf{x}} \right) d\hat{\mathbf{x}} \right) \right\rangle_{\mathbf{x}} + H\left(X\right).
\end{aligned}
\tag{A15}
$$

where $L\left(\mathbf{r}|\hat{\mathbf{x}}\right) = \ln\left(p\left(\mathbf{r}|\hat{\mathbf{x}}\right) p\left(\hat{\mathbf{x}}\right)\right)$ and $L\left(\mathbf{r}|\mathbf{x}\right) = \ln\left(p\left(\mathbf{r}|\mathbf{x}\right) p\left(\mathbf{x}\right)\right)$.

Consider the Taylor expansion for $L(\mathbf{r}|\hat{\mathbf{x}})$ around $\mathbf{x}$. Assuming that $L(\mathbf{r}|\hat{\mathbf{x}})$ is twice continuously differentiable for any $\hat{\mathbf{x}} \in \mathcal{X}_\omega(\mathbf{x})$, we get

$$
\begin{aligned}
&\langle L(\mathbf{r}|\hat{\mathbf{x}}) - L(\mathbf{r}|\mathbf{x}) \rangle_{\mathbf{r}|\mathbf{x}} \\
&= \mathbf{y}^T \mathbf{v}_1 - \frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\mathbf{y}^T \mathbf{G}^{-1/2}\left(\mathbf{x}\right) \left(\mathbf{G}(\check{\mathbf{x}}) - \mathbf{G}(\mathbf{x})\right) \mathbf{G}^{-1/2}\left(\mathbf{x}\right) \mathbf{y}
\end{aligned}
\tag{A16}
$$

where

$$
\mathbf{y} = \mathbf{G}^{1/2}\left(\mathbf{x}\right)\left(\hat{\mathbf{x}} - \mathbf{x}\right),
\tag{A17}
$$

$$
\mathbf{v}_1 = \mathbf{G}^{-1/2}\left(\mathbf{x}\right) q'\left(\mathbf{x}\right)
\tag{A18}
$$

and

$$
\check{\mathbf{x}} = \mathbf{x} + t\left(\hat{\mathbf{x}} - \mathbf{x}\right) \in \mathcal{X}_\omega(\mathbf{x}), \quad t \in (0,1).
\tag{A19}
$$

For later use, we also define

$$
\mathbf{v} = \mathbf{G}^{-1/2}\left(\mathbf{x}\right) l'\left(\mathbf{r}|\mathbf{x}\right)
\tag{A20}
$$

where

$$
l\left(\mathbf{r}|\mathbf{x}\right) = \ln p\left(\mathbf{r}|\mathbf{x}\right).
\tag{A21}
$$

Since $\mathbf{G}(\check{\mathbf{x}})$ is continuous and symmetric for $\check{\mathbf{x}} \in \mathcal{X}$, for any $\epsilon \in (0,1)$, there is a $\varepsilon \in (0,\omega)$ such that

$$
\left| \mathbf{y}^T \mathbf{G}^{-1/2}\left(\mathbf{x}\right) \left(\mathbf{G}(\check{\mathbf{x}}) - \mathbf{G}(\mathbf{x})\right) \mathbf{G}^{-1/2}\left(\mathbf{x}\right) \mathbf{y} \right| < \epsilon \|\mathbf{y}\|^2
\tag{A22}
$$

for all $\mathbf{y} \in \mathcal{Y}_\varepsilon$, where $\mathcal{Y}_\varepsilon = \left\{ \mathbf{y} \in \mathbb{R}^K : \|\mathbf{y}\| < \varepsilon\sqrt{N} \right\}$. Then, we get

$$
\begin{aligned}
&\ln \left(\int_{\mathcal{X}} \exp\left(\langle L(\mathbf{r}|\hat{\mathbf{x}}) - L(\mathbf{r}|\mathbf{x}) \rangle_{\mathbf{r}|\mathbf{x}} \right) d\hat{\mathbf{x}} \right) \\
&\geq -\frac{1}{2}\ln\left(\det\left(\mathbf{G}(\mathbf{x})\right)\right) + \ln \int_{\mathcal{Y}_\varepsilon} \exp\left(\mathbf{y}^T\mathbf{v}_1 - \frac{1}{2}\left(1+\epsilon\right)\mathbf{y}^T\mathbf{y} \right) d\mathbf{y}
\end{aligned}
\tag{A23}
$$

and with Jensen's inequality,

$$
\begin{aligned}
&\ln \int_{\mathcal{Y}_\varepsilon} \exp\left(\mathbf{y}^T\mathbf{v}_1 - \frac{1}{2}\left(1+\epsilon\right)\mathbf{y}^T\mathbf{y} \right) d\mathbf{y} \\
&\geq \ln \Psi_\varepsilon + \int_{\hat{\mathcal{Y}}_\varepsilon} \mathbf{y}^T\mathbf{v}_1 \phi_\varepsilon\left(\mathbf{y}\right) d\mathbf{y} \\
&= \frac{K}{2}\ln\left(\frac{2\pi}{1+\epsilon}\right) + O\left(N^{-K/2}e^{-N\delta}\right),
\end{aligned}
\tag{A24}
$$

where δ is a positive constant, $\int_{\hat{\mathcal{Y}}_\varepsilon} \mathbf{y}^T \mathbf{v}_1 \phi_\varepsilon(\mathbf{y})\, d\mathbf{y} = \mathbf{0}$,

$$\begin{cases} \phi_\varepsilon(\mathbf{y}) = \Psi_\varepsilon^{-1} \exp\left(-\frac{1}{2}(1+\epsilon)\mathbf{y}^T\mathbf{y}\right) \\ \Psi_\varepsilon = \int_{\hat{\mathcal{Y}}_\varepsilon} \exp\left(-\frac{1}{2}(1+\epsilon)\mathbf{y}^T\mathbf{y}\right) d\mathbf{y} \end{cases} \tag{A25}$$

and

$$\hat{\mathcal{Y}}_\varepsilon = \left\{ \mathbf{y} \in \mathbb{R}^K : |y_k| < \varepsilon\sqrt{N/K},\ k = 1, 2, \cdots, K \right\} \subseteq \mathcal{Y}_\varepsilon. \tag{A26}$$

Now, we evaluate

$$\begin{aligned} \Psi_\varepsilon &= \int_{\mathbb{R}^K} \exp\left(-\frac{1}{2}(1+\epsilon)\mathbf{y}^T\mathbf{y}\right) d\mathbf{y} - \int_{\mathbb{R}^K - \hat{\mathcal{Y}}_\varepsilon} \exp\left(-\frac{1}{2}(1+\epsilon)\mathbf{y}^T\mathbf{y}\right) d\mathbf{y} \\ &= \left(\frac{2\pi}{1+\epsilon}\right)^{K/2} - \int_{\mathbb{R}^K - \hat{\mathcal{Y}}_\varepsilon} \exp\left(-\frac{1}{2}(1+\epsilon)\mathbf{y}^T\mathbf{y}\right) d\mathbf{y}. \end{aligned} \tag{A27}$$

Performing integration by parts with $\int_a^\infty e^{-t^2/2} dt = \frac{e^{-a^2/2}}{a} - \int_a^\infty \frac{e^{-t^2/2}}{t^2} dt$, we find

$$\begin{aligned} \int_{\mathbb{R}^K - \hat{\mathcal{Y}}_\varepsilon} \exp\left(-\frac{1}{2}(1+\epsilon)\mathbf{y}^T\mathbf{y}\right) d\mathbf{y} &\leq \frac{\exp\left(-\frac{1}{2}(1+\epsilon)\varepsilon^2 N\right)}{\left((1+\epsilon)^2\varepsilon^2 N/(4K)\right)^{K/2}} \\ &= O\left(N^{-K/2} e^{-N\delta}\right), \end{aligned} \tag{A28}$$

for some constant $\delta > 0$.

Combining Equations (A15), (A23) and (A24), we get

$$I_u \leq \frac{1}{2}\left\langle \ln\left(\det\left(\frac{(1+\epsilon)}{2\pi}\mathbf{G}(\mathbf{x})\right)\right)\right\rangle_{\mathbf{x}} + H(X) + O\left(N^{-K/2} e^{-N\delta}\right). \tag{A29}$$

On the other hand, from Equation (A22) and the condition in Equation (33), we obtain

$$\begin{aligned} \int_{\mathcal{X}_\varepsilon(\mathbf{x})} &\exp\left(\langle L(\mathbf{r}|\hat{\mathbf{x}}) - L(\mathbf{r}|\mathbf{x})\rangle_{\mathbf{r}|\mathbf{x}}\right) d\hat{\mathbf{x}} \\ &\leq \det(\mathbf{G}(\mathbf{x}))^{-1/2} \int_{\mathbb{R}^K} \exp\left(\mathbf{y}^T\mathbf{v}_1 - \frac{1}{2}(1-\epsilon)\mathbf{y}^T\mathbf{y}\right) d\mathbf{y} \\ &= \det\left(\frac{1-\epsilon}{2\pi}\mathbf{G}(\mathbf{x})\right)^{-1/2} \exp\left(\frac{1}{2}(1-\epsilon)^{-1}\mathbf{v}^T\mathbf{v}_1\right) \end{aligned} \tag{A30}$$

and

$$\begin{aligned} \int_{\mathcal{X}} &\exp\left(\langle L(\mathbf{r}|\hat{\mathbf{x}}) - L(\mathbf{r}|\mathbf{x})\rangle_{\mathbf{r}|\mathbf{x}}\right) d\hat{\mathbf{x}} \\ &= \int_{\mathcal{X}_\varepsilon(\mathbf{x})} \exp\left(\langle L(\mathbf{r}|\hat{\mathbf{x}}) - L(\mathbf{r}|\mathbf{x})\rangle_{\mathbf{r}|\mathbf{x}}\right) d\hat{\mathbf{x}} + \int_{\mathcal{X} - \mathcal{X}_\varepsilon(\mathbf{x})} \exp\left(\langle L(\mathbf{r}|\hat{\mathbf{x}}) - L(\mathbf{r}|\mathbf{x})\rangle_{\mathbf{r}|\mathbf{x}}\right) d\hat{\mathbf{x}} \\ &\leq \det\left(\frac{1-\epsilon}{2\pi}\mathbf{G}(\mathbf{x})\right)^{-1/2}\left(\exp\left(\frac{\mathbf{v}^T\mathbf{v}_1}{2(1-\epsilon)}\right) + O\left(N^{-1}\right)\right). \end{aligned} \tag{A31}$$

It follows from Equations (A15) and (A31) that

$$
\left\langle \ln \left(\int_{\mathcal{X}} \exp \left(\langle L(\mathbf{r}|\hat{\mathbf{x}}) - L(\mathbf{r}|\mathbf{x}) \rangle_{\mathbf{r}|\mathbf{x}} \right) d\hat{\mathbf{x}} \right) \right\rangle_{\mathbf{x}}
$$
$$
\leq -\frac{1}{2} \left\langle \ln \left(\det \left(\frac{(1-\epsilon)}{2\pi} \mathbf{G}(\mathbf{x}) \right) \right) \right\rangle_{\mathbf{x}} + \frac{1}{2} (1-\epsilon)^{-1} \left\langle \mathbf{v}^T \mathbf{v}_1 \right\rangle_{\mathbf{x}} + O\left(N^{-1}\right). \tag{A32}
$$

Note that

$$
\left\langle \mathbf{v}^T \mathbf{v}_1 \right\rangle_{\mathbf{x}} = O\left(N^{-1}\right). \tag{A33}
$$

Now, we have

$$
I_u \geq \frac{1}{2} \left\langle \ln \left(\det \left(\frac{(1-\epsilon)}{2\pi} \mathbf{G}(\mathbf{x}) \right) \right) \right\rangle_{\mathbf{x}} + H(X) + O\left(N^{-1}\right). \tag{A34}
$$

Since ϵ is arbitrary, we can let it go to zero. Therefore, from Equations (25), (A29) and (A34), we obtain the upper bound in Equation (35).

The Taylor expansion of $h(\hat{\mathbf{x}}, \mathbf{x}) = \left\langle \left(\frac{p(\mathbf{r}|\hat{\mathbf{x}})}{p(\mathbf{r}|\mathbf{x})} \right)^\beta \right\rangle_{\mathbf{r}|\mathbf{x}}$ around $\mathbf{x}$ is

$$
h(\hat{\mathbf{x}}, \mathbf{x}) = 1 + \left\langle \frac{\beta}{p(\mathbf{r}|\mathbf{x})} \frac{\partial p(\mathbf{r}|\mathbf{x})}{\partial \mathbf{x}} \right\rangle_{\mathbf{r}|\mathbf{x}} (\hat{\mathbf{x}} - \mathbf{x}) +
$$
$$
(\hat{\mathbf{x}} - \mathbf{x})^T \left\langle \frac{\beta}{2p(\mathbf{r}|\mathbf{x})^2} \left((\beta-1) \frac{\partial p(\mathbf{r}|\mathbf{x})}{\partial \mathbf{x}} \frac{\partial p(\mathbf{r}|\mathbf{x})}{\partial \mathbf{x}^T} + p(\mathbf{r}|\mathbf{x}) \frac{\partial^2 p(\mathbf{r}|\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} \right) \right\rangle_{\mathbf{r}|\mathbf{x}} (\hat{\mathbf{x}} - \mathbf{x}) + \cdots
$$
$$
= 1 - \frac{\beta(1-\beta)}{2} (\hat{\mathbf{x}} - \mathbf{x})^T \mathbf{J}(\mathbf{x})(\hat{\mathbf{x}} - \mathbf{x}) + \cdots. \tag{A35}
$$

In a similar manner as described above, we obtain the asymptotic relationship (37):

$$
I_{\beta,\alpha} = I_\gamma + O\left(N^{-1}\right)
$$
$$
= \frac{1}{2} \int_{\mathcal{X}} p(\mathbf{x}) \ln \left(\det \left(\frac{\mathbf{G}_\gamma(\mathbf{x})}{2\pi} \right) \right) d\mathbf{x} + H(X). \tag{A36}
$$

Notice that $0 < \gamma = \beta(1-\beta) \leq 1/4$ and the equality holds when $\beta = \beta_0 = 1/2$. Thus, we have

$$
\det\left(\mathbf{G}_\gamma(\mathbf{x})\right) \leq \det\left(\mathbf{G}_{\gamma_0}(\mathbf{x})\right). \tag{A37}
$$

Combining Equations (25), (A36) and (A37) yields the lower bound in Equation (35).

The proof of Equation (36) is similar. This completes the proof of Theorem 4.   □

References

1. Borst, A.; Theunissen, F.E. Information theory and neural coding. *Nat. Neurosci.* **1999**, *2*, 947–57.
2. Pouget, A.; Dayan, P.; Zemel, R. Information processing with population codes. *Nat. Rev. Neurosci.* **2000**, *1*, 125–132.
3. Laughlin, S.B.; Sejnowski, T.J. Communication in neuronal networks. *Science* **2003**, *301*, 1870–1874.
4. Brown, E.N.; Kass, R.E.; Mitra, P.P. Multiple neural spike train data analysis: State-of-the-art and future challenges. *Nat. Neurosci.* **2004**, *7*, 456–461.
5. Bell, A.J.; Sejnowski, T.J. The "independent components" of natural scenes are edge filters. *Vision Res.* **1997**, *37*, 3327–3338.
6. Huang, W.; Zhang, K. An Information-Theoretic Framework for Fast and Robust Unsupervised Learning via Neural Population Infomax. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017; arXiv:1611.01886.

7. Huang, W.; Huang, X.; Zhang, K. Information-theoretic interpretation of tuning curves for multiple motion directions. In Proceedings of the 2017 51st Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 22–24 March 2017.

8. Cover, T.M.; Thomas, J.A. *Elements of Information*, 2nd ed.; Wiley-Interscience: New York, NY, USA, 2006.

9. Miller, G.A. Note on the bias of information estimates. *Inf. Theory Psychol.Probl. Methods* **1955**, *2*, 100.

10. Carlton, A. On the bias of information estimates. *Psychol. Bull.* **1969**, *71*, 108.

11. Treves, A.; Panzeri, S. The upward bias in measures of information derived from limited data samples. *Neural Comput.* **1995**, *7*, 399–407.

12. Victor, J.D. Asymptotic bias in information estimates and the exponential (Bell) polynomials. *Neural Comput.* **2000**, *12*, 2797–2804.

13. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253.

14. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138.

15. Khan, S.; Bandyopadhyay, S.; Ganguly, A.R.; Saigal, S.; Erickson, D.J., III; Protopopescu, V.; Ostrouchov, G. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys. Rev. E* **2007**, *76*, 026209.

16. Safaai, H.; Onken, A.; Harvey, C.D.; Panzeri, S. Information estimation using nonparametric copulas. *Phys. Rev. E* **2018**, *98*, 053302.

17. Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.

18. Van Trees, H.L.; Bell, K.L. *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*; John Wiley: Piscataway, NJ, USA, 2007.

19. Clarke, B.S.; Barron, A.R. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **1990**, *36*, 453–471.

20. Rissanen, J.J. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory* **1996**, *42*, 40–47.

21. Brunel, N.; Nadal, J.P. Mutual information, Fisher information, and population coding. *Neural Comput.* **1998**, *10*, 1731–1757.

22. Sompolinsky, H.; Yoon, H.; Kang, K.J.; Shamir, M. Population coding in neuronal systems with correlated noise. *Phys. Rev. E* **2001**, *64*, 051904.

23. Kang, K.; Sompolinsky, H. Mutual information of population codes and distance measures in probability space. *Phys. Rev. Lett.* **2001**, *86*, 4958–4961.

24. Huang, W.; Zhang, K. Information-theoretic bounds and approximations in neural population coding. *Neural Comput.* **2018**, *30*, 885–944.

25. Strong, S.P.; Koberle, R.; van Steveninck, R.R.D.R.; Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **1998**, *80*, 197.

26. Nemenman, I.; Bialek, W.; van Steveninck, R.D.R. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E* **2004**, *69*, 056111.

27. Panzeri, S.; Senatore, R.; Montemurro, M.A.; Petersen, R.S. Correcting for the sampling bias problem in spike train information measures. *J. Neurophysiol.* **2017**, *98*, 1064–1072.

28. Houghton, C. Calculating the Mutual Information Between Two Spike Trains. *Neural Comput.* **2019**, *31*, 330–343.

29. Rényi, A. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*; The Regents of the University of California, University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.

30. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **1952**, *23*, 493–507.

31. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99–109.

32. Beran, R. Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* **1977**, *5*, 445–463.