 doi:10.20944/preprints201812.0209.v1

Article

Neural Population Coding and Approximations of Mutual Information for Discrete Variables

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 100846, China; whuang21@jhmi.edu

[2] Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD 21205, U.S.A.; kzhang4@jhmi.edu

* Correspondence: whuang21@jhmi.edu, kzhang4@jhmi.edu; Tel.: +1-443-287-5080

Version December 18, 2018 submitted to Preprints

Abstract: Information theory is widely used in various disciplines, and effective calculation of Shannon mutual information is typically not an easy task for many practical applications, including problems of neural population coding in computational and theoretical neuroscience. Asymptotic formulas based on Fisher information may provide accurate approximations to mutual information but this approach is restricted to continuous variables because the calculation requires derivatives with respect to the encoded variables. In this paper, we consider information-theoretic bounds and approximations 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, and these asymptotic formulas hold true for discrete variables as there is no requirement for differentiability. In particular, one of our approximation formulas has consistent performance and good 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 mutual information between the discrete variables or 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: Shannon mutual information; Kullback-Leibler divergence; Rényi divergence; Chernoff divergence; approximation; discrete stimuli; neural population codes

1. Introduction

Information theory is a powerful tool with widespread applications in diverse fields such as neuroscience, machine learning, and information and communication technology [1–7]. However, in many applications it is often notoriously difficult to effectively calculate Shannon mutual information [8]. Various approximation methods have been proposed to approximate the mutual information, such as those based on asymptotic expansion [9–13], k-nearest neighbor [14] and minimal spanning trees [15]. Another approach is to simplify the calculations by approximations based on information-theoretic bounds, such as the Cramér-Rao lower bound [16] and the van Trees' Bayesian Cramér-Rao bound [17].

An efficient approach for estimating the mutual information based on asymptotic approximation has attracted the attention of researchers in recent years [18–23]. For encoding of continuous variables, asymptotic relations between mutual information and Fisher information have been presented by several researchers [18–21]. Recently Huang and Zhang [23] proposed a more precise approximation formula which remains very accurate for high-dimensional variables. Unfortunately this approach does not generalize to discrete variables because 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 [22] 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.

In this paper, we present several information metrics to approximate the mutual information for discrete variables. While some input variables or stimuli are naturally continuous, such as movement direction, luminance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

level, and the pitch of a tone, other stimuli are naturally discrete, such as the types of molecules in olfaction, distinct visual objects, identity of faces, and words in human speech. For definiteness, in this paper we will frame our questions in the context of neural population coding, where some physical variables are encoded by eliciting responses from many neurons or a population of neurons. Nonetheless, our mathematical results are quite general and should be applicable to any large input-output system that satisfies specific conditions as it scales up. For example, independent sampling from repeated trials can be mathematically equivalent to having a population of neurons with independent responses.

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 mutual information in the limit of a large population or a large sample size. In particular, our approximation formula I_e (see Eq. 10) is valid for both discrete variables and continuous variables. The approximation formulas I_d and I_D (see Eq. 15 and 18) are applicable only to discrete variables, and they are more convenient to calculate than mutual information in some situations. Finally, we use numerical simulations to confirm the validity of our approximation formulas by a comparison against the values of mutual information obtained 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 [23]:

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

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

We denote the Kullback-Leibler divergence as

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

and denote Rényi divergence [24] of order $\beta+1$ as

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

Here $\beta D_{\beta}\left(\mathbf{x}||\hat{\mathbf{x}}\right)$ is equivalent to Chernoff divergence of order $\beta+1$ [25]. It is well known that $D_{\beta}\left(\mathbf{x}||\hat{\mathbf{x}}\right)\rightarrow D\left(\mathbf{x}||\hat{\mathbf{x}}\right)$ in the limit $\beta\rightarrow 0$.

We define

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

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

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

where in $I_{\beta,\alpha}$ we have $\beta\in(0,1)$ and $\alpha\in(0,\infty)$ and assume $p\left(\mathbf{x}\right)>0$ 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. By the definitions above in (9)–(11), we have

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

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

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

Furthermore, we define

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

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

where

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

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

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

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

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

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

2.2. Theorems

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

Theorem 1. *The mutual information I has the following bounds:*

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

Theorem 2. *The following inequalities are satisfied,*

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

where $\beta_1 = e^{-1}$ and according to Eq. 11,

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

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,*

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

 doi:10.20944/preprints201812.0209.v1

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

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

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

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:

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

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

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

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. *To see how condition (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 $N^{-1}D(\mathbf{x}|\hat{\mathbf{x}}) \ge \eta$ for $\hat{\mathbf{x}} \in \bar{\mathcal{X}}_\omega(\mathbf{x})$. Now condition (33) is satisfied because*

$$\begin{aligned}
&\det(\mathbf{G}(\mathbf{x}))^{1/2} \int_{\bar{\mathcal{X}}_\varepsilon(\mathbf{x})} p(\hat{\mathbf{x}}) \exp(-D(\mathbf{x}||\hat{\mathbf{x}}))\, d\hat{\mathbf{x}} \\
&\le \det(\mathbf{G}(\mathbf{x}))^{1/2} \int_{\bar{\mathcal{X}}_\varepsilon(\mathbf{x})} p(\hat{\mathbf{x}}) \exp(-\hat{\eta}(\varepsilon) N)\, d\hat{\mathbf{x}} \\
&= O\left(N^{K/2} e^{-\hat{\eta}(\varepsilon)N}\right),
\end{aligned} \tag{39}$$

where by assumption we can find an $\hat{\eta}(\varepsilon) > 0$ for any given $\varepsilon \in (0,\omega)$. Condition (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 [26]:

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

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 [27] 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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

Denoting the Chernoff information [8] as

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

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

$$\begin{aligned} & I_{\beta,\alpha}-H(X) \\ & \leq h_{c}=-\sum_{m=1}^{M}p\left(\mathbf{x}_{m}\right)\ln\left(\sum_{\hat{m}=1}^{M}\frac{p\left(\mathbf{x}_{\hat{m}}\right)}{p\left(\mathbf{x}_{m}\right)}\exp\left(-C\left(\mathbf{x}_{m}||\mathbf{x}_{\hat{m}}\right)\right)\right) \qquad (45) \\ & \leq h_{d}=-\sum_{m=1}^{M}p\left(\mathbf{x}_{m}\right)\ln\left(\sum_{\hat{m}=1}^{M}\frac{p\left(\mathbf{x}_{\hat{m}}\right)}{p\left(\mathbf{x}_{m}\right)}\exp\left(-\beta_{m}D_{\beta}\left(\mathbf{x}_{m}||\mathbf{x}_{\hat{m}}\right)\right)\right). \qquad (46) \end{aligned}$$

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 (50), (46), (47) and (48) we have*

$$\begin{aligned} \max_{\beta\in(0,\,1)}I_{\beta}+\frac{K}{2}\ln\frac{4}{e}+O\left(N^{-1}\right) & \leq I_{e}=I+O\left(N^{-1}\right) \\ & \leq h_{d}+H(X)\leq I_{u}. \end{aligned} \tag{49}$$

Therefore, from (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 finite N. First of all, **Theorem 1** and **Theorem 2** tell us that the true mutual information I and I_e lie between $I_{\beta,\alpha}$ and I_u or between $I_{\beta_1,1}$ and I_u; that is, $I_{\beta,\alpha}\leq I\leq I_u$, and $I_{\beta_1,1}\leq I_e\leq I_u$. On the other hand, from (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 (31) for large but finite N, we have

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

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 definition (18), substitution of $p(\mathbf{x}_{\hat{m}}) \simeq p(\mathbf{x}_m)$ into (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 when 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 small.

The theoretical discussion above tells us that I_e and I_d are effective approximations to true mutual information I in the case 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 follow a Poisson distribution [23]. The mean responses or the tuning curve of neuron n, $n \in \{1, 2, \cdots, N\}$ is described by the tuning function $f(x; \theta_n)$, which takes the form of a rectified linear function:

$$f(x; \theta_n) = \max(0, x - \theta_n), \tag{54}$$

where $x \in [-T, T]$, is the stimulus with $T = 10$, and the centers $\theta_1, \theta_2, \cdots, \theta_N$ of the N neurons are uniformly distributed on interval $[-T, T]$, i.e., $\theta_n = (n-1)\, d - T$ with $d = 2T/(N-1)$ when $N \geq 2$, and $\theta_n = 0$ when $N = 1$. We suppose the discrete stimulus x has M possible values with $M = 5$, and these values are evenly spaced from $-T$ to T, namely, $x \in \mathcal{X} = \{x_m : x_m = 2(m-1)\, T/(M-1) - T, m = 1, 2, \cdots, M\}$.

In the first example, 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 curve $f(x; \theta)$ with the center $\theta = 0$.

To evaluate the precision of the approximation formulas, we employed Monte Carlo (MC) simulation to approximate mutual information I [23]. 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

$$I_{MC}^* = \frac{1}{j_{\max}} \sum_{j=1}^{j_{\max}} \ln\left(\frac{p(\mathbf{r}_j | x_j)}{p(\mathbf{r}_j)}\right), \tag{55}$$

where

$$p(\mathbf{r}_j) = \sum_{m=1}^{M} p(\mathbf{r}_j | x_m) p(x_m). \tag{56}$$

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

$$I_{std} = \sqrt{\frac{1}{i_{\max}} \sum_{i=1}^{i_{\max}} \left(I_{MC}^i - I_{MC}\right)^2}, \tag{57}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

where

$$I_{MC}^{i} = \frac{1}{j_{\max}} \sum_{j=1}^{j_{\max}} \ln \left(\frac{p(\mathbf{r}_{\Gamma_{j,i}} | x_{\Gamma_{j,i}})}{p(\mathbf{r}_{\Gamma_{j,i}})} \right), \tag{58}$$

$$I_{MC} = \frac{1}{i_{\max}} \sum_{i=1}^{i_{\max}} I_{MC}^{i}, \tag{59}$$

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 compare I_{MC} with I_e, I_d and I_D, which are illustrated in Figure 1b–d. Here we define the relative error of approximation, e.g., for I_e, as

$$DI_e = \frac{I_e - I_{MC}}{I_{MC}}, \tag{60}$$

and the relative standard deviation

$$DI_{std} = \frac{I_{std}}{I_{MC}}. \tag{61}$$

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

$$p(x_m) = Z^{-1} \exp\left(-\frac{x_m^2}{2T^2} \right), m = 1, 2, \cdots, M, \tag{62}$$

where Z is the normalization constant. The results are illustrated in Figure 2.

Next, we changed each tuning curve $f(x; \theta_n)$ to a step function or Heaviside function:

$$f(x; \theta_n) = \begin{cases} 1; x \geq \theta_n \\ 0; x < \theta_n \end{cases}. \tag{63}$$

Figure 3 and Figure 4 show the results under the same conditions of Figures 1 and 2 except for the shape of the tuning curves.

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 obtained by Monte Carlo method. All these approximations were extremely accurate when $N > 100$. The saturation of information for large N is due to the fact that completely distinguishing all $M = 5$ stimuli requires at most $\log_2 5 = 2.32$ bits of information. It is not possible to gain more information than this amount regardless of how many neurons are used in the population. For relatively small values of N, we found that I_D tended to be less accurate than I_d and I_e. In these simulations, I_d and I_e were about equally good although I_e is more versatile and could potentially be used for continuous stimuli as well.

4. Conclusions

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

We propose to approximate the mutual information with Kullback-Leibler divergence and provide several formulas, including I_e in Eq. 10 or 13, I_d in Eq. 15 and I_D in Eq. 18. The approximation I_e works not only

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

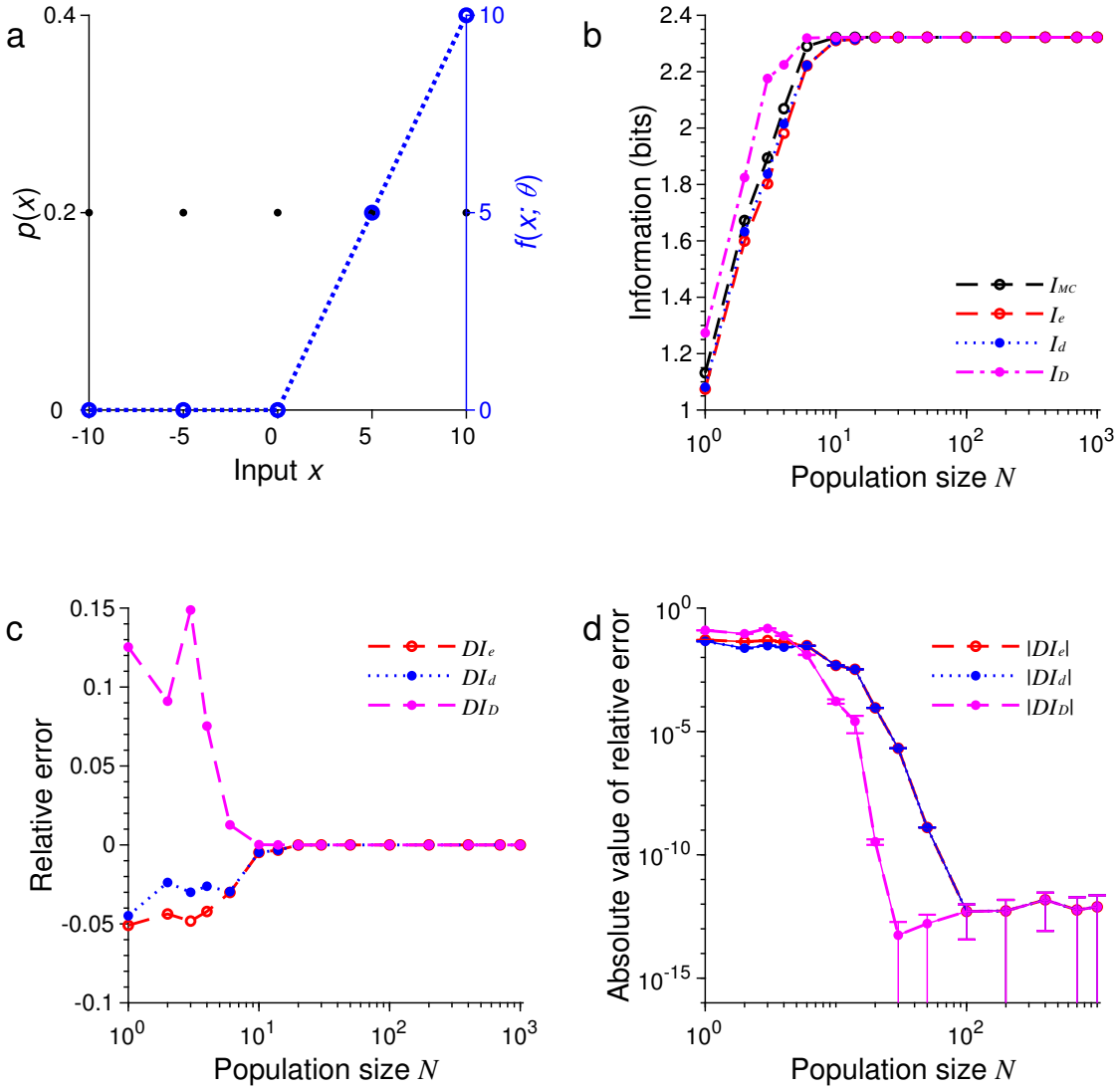

Figure 1. A comparison of approximations I_e, I_d and I_D against I_{MC} obtained by Monte Carlo method for one-dimensional discrete stimuli. (**a**) Discrete uniform distribution of the stimulus $p(x)$ (black dots) and the rectified linear tuning curve $f(x;\theta)$ with center $\theta = 0$ (blue dashed 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 panel b. (**d**) The absolute values of the relative errors $|DI_e|$, $|DI_d|$ and $|DI_D|$ as in panel c, with error bars showing standard deviations of repeated trials.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

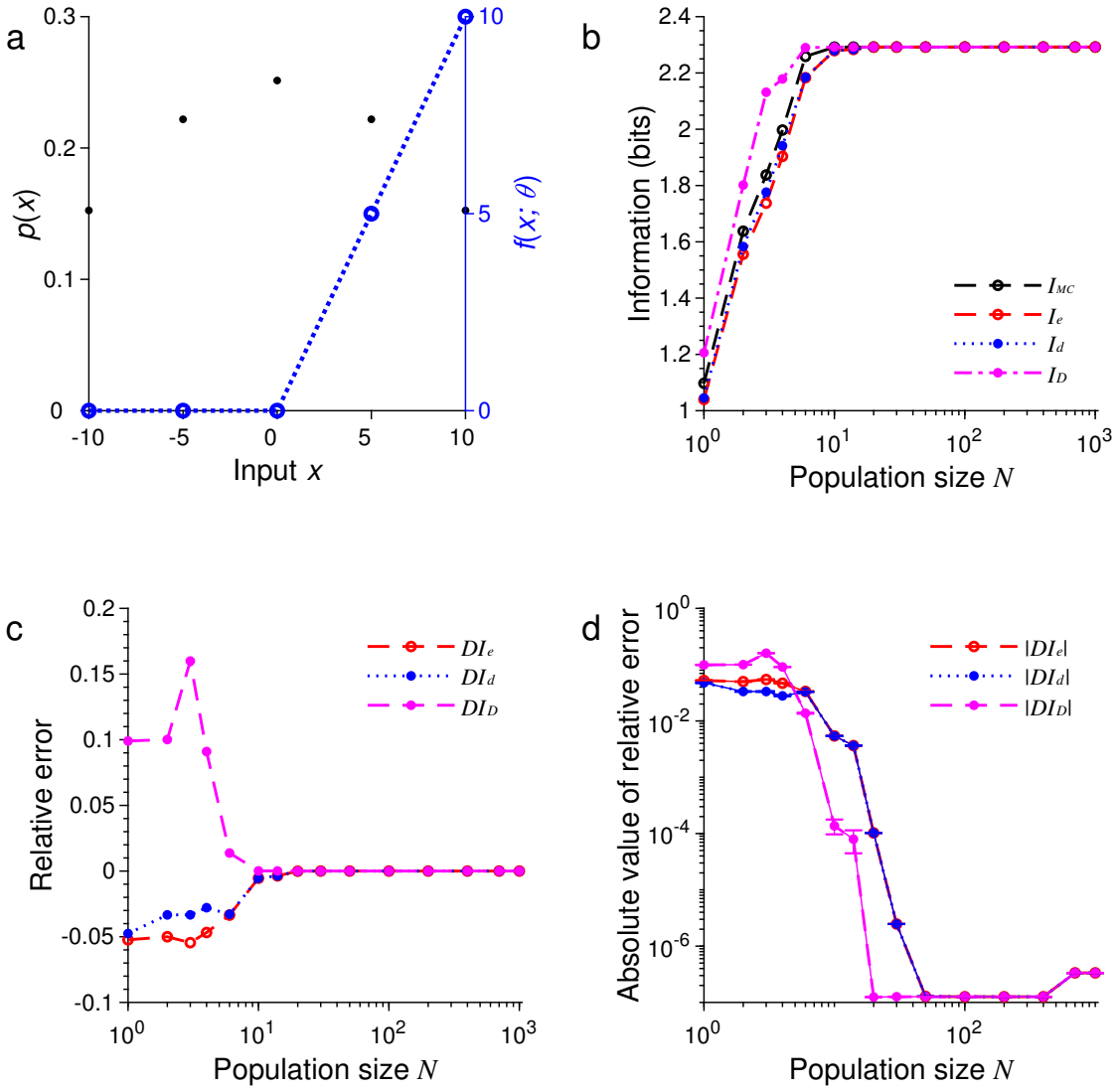

Figure 2. A comparison of approximations I_e, I_d and I_D against I_{MC}. The situation is identical to that in Figure 1 except that the stimulus distribution $p(x)$ is peaked rather flat (black dots in panel a).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

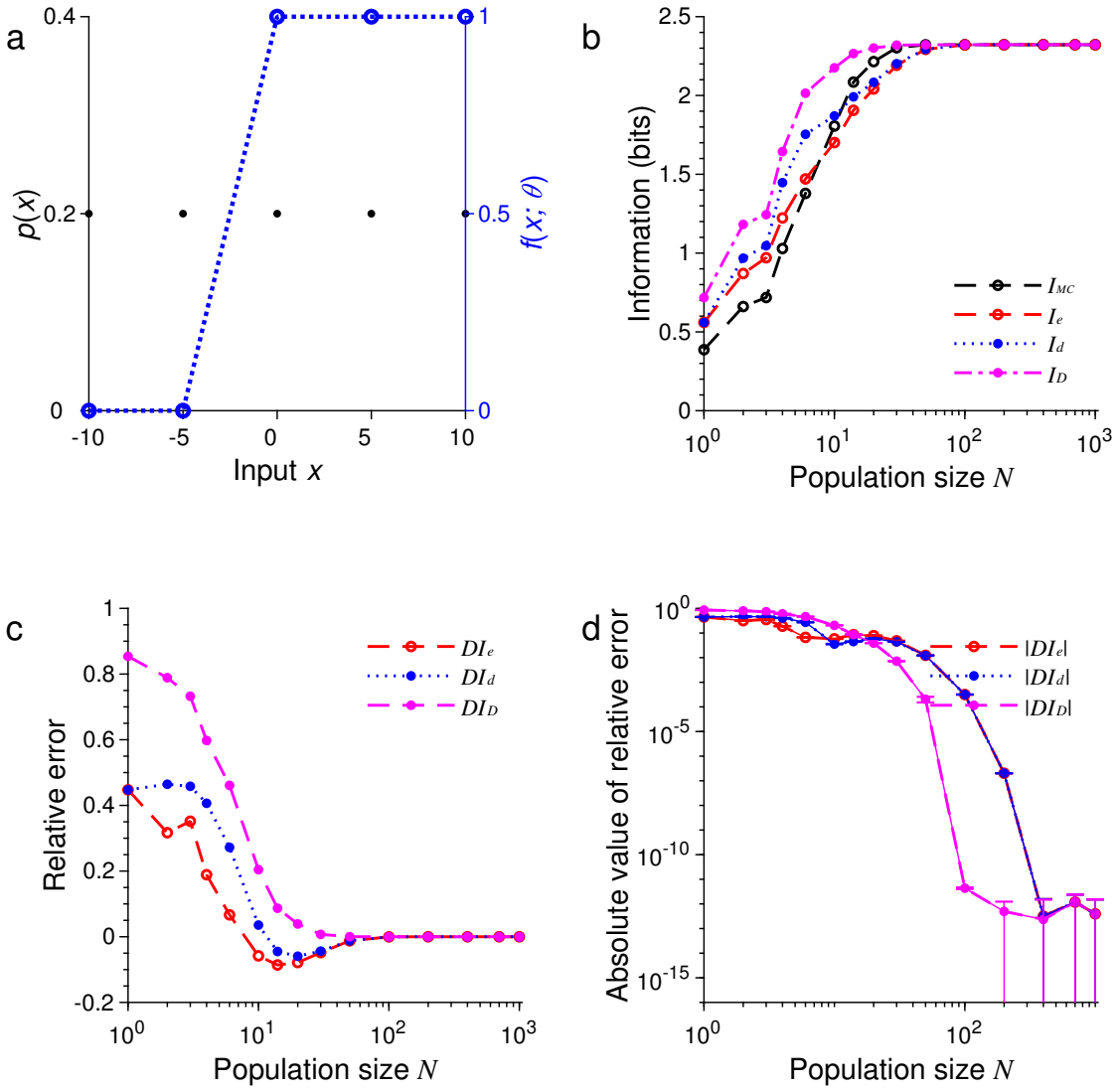

Figure 3. A comparison of approximations I_e, I_d and I_D against I_{MC}. The situation is identical to that in Figure 1 except for the shape of the tuning curve (blue dashed lines in panel a).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

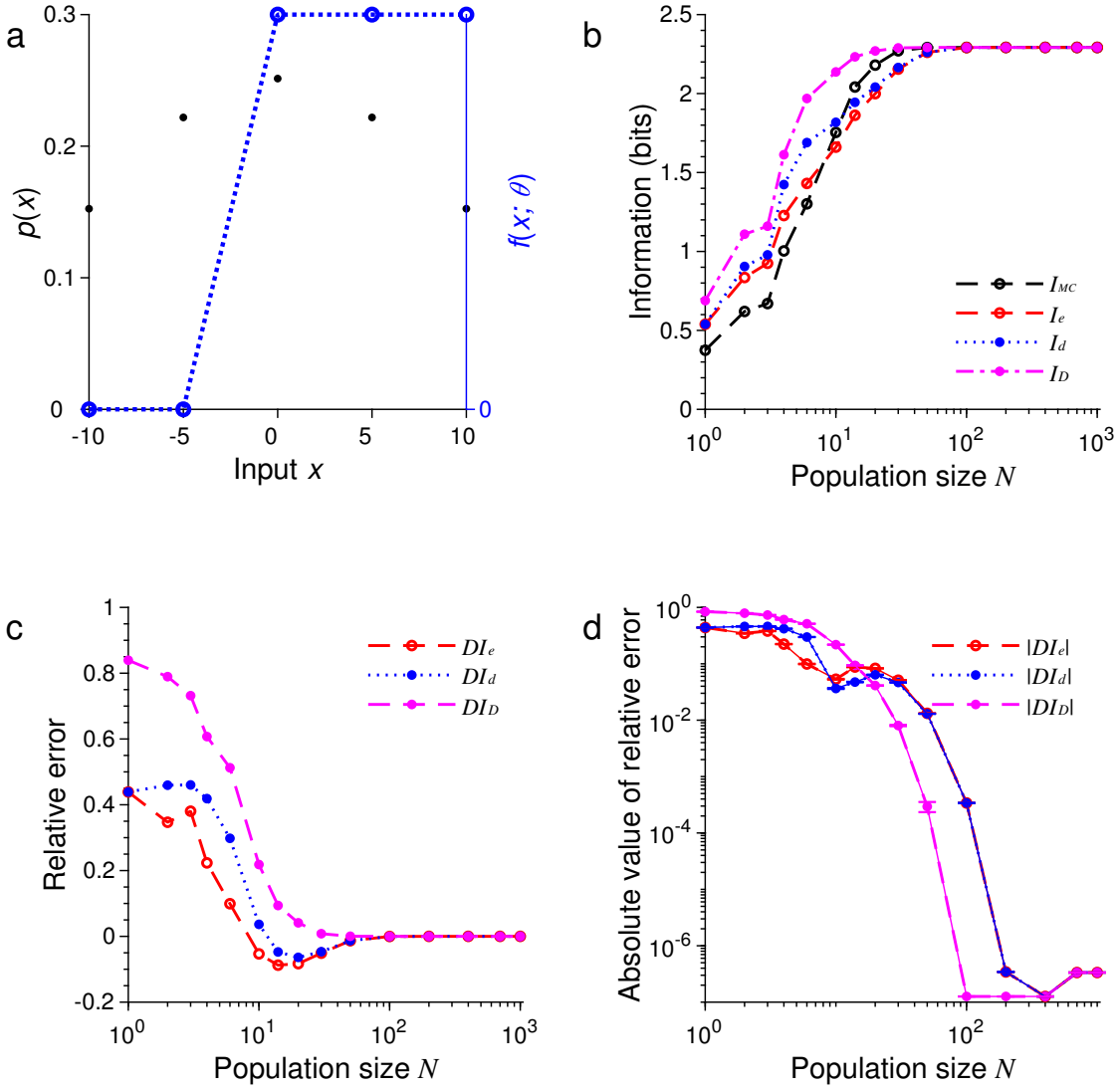

Figure 4. A comparison of approximations I_e, I_d and I_D against I_{MC}. The situation is identical to that in Figure 1 except that the stimulus distribution $p(x)$ is peaked rather flat (black dots in panel a) and that the shape of the tuning curve is different (blue dashed lines in panel a).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

for discrete stimuli but also for continuous stimuli. The formula for I_e is well justified theoretically and its performance was excellent in our numerical simulations. Overall I_e is our most accurate and versatile approximation formula although in some cases I_d and I_D are more convenient to calculate.

Our numerical experimental results show that the three approximations I_e, I_d and I_D were very accurate even when the population size N is relatively small. 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. In conclusion we recommend using I_e and I_d as approximations for Shannon mutual information. Of these two, I_e is the universal formula and I_d is restricted only to discrete variables.

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

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 partially 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

$$\begin{aligned} I_{\beta,\alpha} &= -\left\langle \ln\left(\int_{\mathcal{X}} \left\langle \frac{p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})}{p^{\beta}(\mathbf{r}|\mathbf{x})\, p^{\alpha}(\mathbf{x})} \right\rangle_{\mathbf{r}|\mathbf{x}} d\hat{\mathbf{x}} \right) \right\rangle_{\mathbf{x}} + H(X) \\ &\le -\left\langle \left\langle \ln\left(\int_{\mathcal{X}} \frac{p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})}{p^{\beta}(\mathbf{r}|\mathbf{x})\, p^{\alpha}(\mathbf{x})} d\hat{\mathbf{x}} \right) \right\rangle_{\mathbf{r}|\mathbf{x}} \right\rangle_{\mathbf{x}} + H(X) \end{aligned} \tag{A1}$$

and

$$\begin{aligned} & -\left\langle \left\langle \ln\left(\int_{\mathcal{X}} \frac{p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})}{p^{\beta}(\mathbf{r}|\mathbf{x})\, p^{\alpha}(\mathbf{x})} d\hat{\mathbf{x}} \right) \right\rangle_{\mathbf{r}|\mathbf{x}} \right\rangle_{\mathbf{x}} + H(X) - I \\ &= \left\langle \left\langle \ln\left(\int_{\mathcal{X}} \frac{p(\mathbf{r},\hat{\mathbf{x}})}{p(\mathbf{r},\mathbf{x})} d\hat{\mathbf{x}} \right) \left(\int_{\mathcal{X}} \frac{p^{\beta}(\mathbf{r}|\hat{\mathbf{x}})p^{\alpha}(\hat{\mathbf{x}})}{p^{\beta}(\mathbf{r}|\mathbf{x})\, p^{\alpha}(\mathbf{x})} d\hat{\mathbf{x}} \right)^{-1} \right\rangle_{\mathbf{r}|\mathbf{x}} \right\rangle_{\mathbf{x}} \\ &\le \ln \int_{\mathcal{R}} p(\mathbf{r}) \frac{\int_{\mathcal{X}} p^{\beta}(\mathbf{r}|\mathbf{x})\, p^{\alpha}(\mathbf{x}) d\mathbf{x}}{\int_{\mathcal{X}} p^{\beta}(\mathbf{r}|\hat{\mathbf{x}}) p^{\alpha}(\hat{\mathbf{x}}) d\hat{\mathbf{x}}} d\mathbf{r} \\ &= 0. \end{aligned} \tag{A2}$$

Combining (A1) and (A2), we immediately get the lower bound in (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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

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

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

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

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

where λ is the Lagrange multiplier and

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

Setting $\frac{\partial \tilde{\Phi}[q(\hat{\mathbf{x}})]}{\partial q(\hat{\mathbf{x}})} = 0$ and using the constraint (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 (1), (A5) and (A9), we get the upper bound in (25). This completes the proof of **Theorem 1**.□

Appendix A.2. Proof of Theorem 2

It follows from (43) that

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

where $\beta_1 = e^{-1}$ and $\alpha_1 = 1$. We immediately get (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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2018 doi:10.20944/preprints201812.0209.v1

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

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

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

Combining (A11), (A13) and **Theorem 1**, we get the lower bound in (30). In a manner similar to the above, we can get the upper bound in Eqs. (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 (25) can be written as

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

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

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

where

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

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

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

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

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

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

 doi:10.20944/preprints201812.0209.v1

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, $\int_{\hat{\mathcal{Y}}_\varepsilon} \mathbf{y}^T\mathbf{v}_1\phi_\varepsilon\left(\mathbf{y}\right) d\mathbf{y} = \mathbf{0}$,

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

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

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

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

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

136 for some constant $\delta > 0$.

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

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

 doi:10.20944/preprints201812.0209.v1

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

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

and

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

It follows from (A15) and (A31) that

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

Note that

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

Now we have

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

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

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

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

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

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

Notice that $0 < \gamma = \beta\left(1-\beta\right) \leq 1/4$ 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 (25), (A36) and (A37) yields the lower bound in (35).

The proof of Eq. (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 *5th International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1611.01886*; 2017.
7. Huang, W.; Huang, X.; Zhang, K. Information-theoretic interpretation of tuning curves for multiple motion directions. In *Information Sciences and Systems (CISS), 2017 51st Annual Conference on.*; 2017.
8. Cover, T.M.; Thomas, J.A. *Elements of Information, 2nd Edition*; Wiley-Interscience: New York, 2006.
9. Miller, G.A. Note on the bias of information estimates. *Information theory in psychology: Problems and methods* **1955**, *2*, 100.
10. Carlton, A. On the bias of information estimates. *Psychological Bulletin* **1969**, *71*, 108.
11. Treves, A.; Panzeri, S. The upward bias in measures of information derived from limited data samples. *Neural Computation* **1995**, *7*, 399–407.
12. Victor, J.D. Asymptotic bias in information estimates and the exponential (Bell) polynomials. *Neural Computation* **2000**, *12*, 2797–2804.
13. Paninski, L. Estimation of entropy and mutual information. *Neural computation* **2003**, *15*, 1191–1253.
14. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Physical review E* **2004**, *69*, 066138.
15. Khan, S.; Bandyopadhyay, S.; Ganguly, A.R.; Saigal, S.; Erickson III, D.J.; Protopopescu, V.; Ostrouchov, G. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E* **2007**, *76*, 026209.
16. Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society* **1945**, *37*, 81–91.
17. Van Trees, H.L.; Bell, K.L. *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*; John Wiley: Piscataway, 2007.
18. Clarke, B.S.; Barron, A.R. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **1990**, *36*, 453–471.
19. Rissanen, J.J. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory* **1996**, *42*, 40–47.
20. Brunel, N.; Nadal, J.P. Mutual information, Fisher information, and population coding. *Neural Comput.* **1998**, *10*, 1731–1757.
21. Sompolinsky, H.; Yoon, H.; Kang, K.J.; Shamir, M. Population coding in neuronal systems with correlated noise. *Phys. Rev. E* **2001**, *64*, 051904.
22. Kang, K.; Sompolinsky, H. Mutual information of population codes and distance measures in probability space. *Phys. Rev. Lett.* **2001**, *86*, 4958–4961.
23. Huang, W.; Zhang, K. Information-theoretic bounds and approximations in neural population coding. *Neural computation* **2018**, *30*, 885–944.
24. Rényi, A. On measures of entropy and information. Fourth Berkeley Symposium on Mathematical Statistics and Probability. The Regents of the University of California, University of California Press, 1961, pp. 547–561.
25. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics* **1952**, *23*, 493–507.
26. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99–109.
27. Beran, R. Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* **1977**, *5*, 445–463.